Gaussian estimation:
Sequence and multiresolution models

Draft version, August 29, 2011

Iain M. Johnstone

Contents

(working) Preface

This is a book about some of the theory of nonparametric function estimation. The premise is that much insight can be gained even if attention is confined to a Gaussian sequence model

$$y_i = \theta_i + \epsilon z_i, \qquad i \in I, \tag{0.1}$$

where I is finite or countable, $\{\theta_i\}$ is fixed and unknown, $\{z_i\}$ are i.i.d. $N(0,1)$ noise variables and ϵ is a known noise level. If I is finite, this is an old friend, the multivariate normal means model, with independent co-ordinates and known variance. It is the centerpiece of parametric statistics, with many important, beautiful, and even surprising results whose influence extends well beyond the formal model into the practical, approximate world of data analysis.

It is perhaps not so obvious that the infinite sequence model could play a corresponding role in nonparametric statistics. For example, problems of nonparametric regression, density estimation and classification are typically formulated in terms of unknown functions, rather than sequences of parameters. Secondly, the additive white Gaussian noise assumption may seem rather remote.

There are several responses to these objections. First, the model captures many of the conceptual issues associated with non-parametric estimation, with a minimum of technical complication. For example, non-parametrics must grapple with the apparent impossibility of trying to estimate an infinite-dimensional object – a function – on the basis of a finite amount n of noisy data. With a calibration $\epsilon = 1/\sqrt{n}$, this challenge is plain to see in model (0.1). The broad strategy is to apply various methods that one understands in the multivariate normal model to finite submodels, and to argue that often not too much is lost by ignoring the (many!) remaining parameters.

Second, models and theory are always an idealisation of practical reality. Advances in size of datasets and computing power have enormously increased the complexity of both what we attempt to do in data analysis and the algorithms that we invent to carry out our goals. If one aim of theory is to provide clearly formulated, generalizable insights that might inform and improve our computational efforts, then we may need to accept a greater degree of idealisation in our models than was necessary when developing theory for the estimation of one, two or three parameters from modest numbers of observations.

Thirdly, it turns out that model (0.1) is often a reasonable approximation, in large samples, to other nonparametric settings. In parametric statistics, the central limit theorem and asymptotic normality of estimators extends the influence of multivariate normal theory to generalized linear models and beyond. In nonparametric estimation, it has long been observed that similar features are often found in spectrum, density and regression estimation.

Relatively recently, results have appeared connecting these problems to model (0.1) and thereby providing some formal support for these observations.

Model (0.1) and its justifications have been used and understood for decades, notably by Russian theoretical statisticians, led by I. A. Ibragimov and R. Z. Khasminskii. It was somewhat slower to receive wide discussion in the West. However, it received a considerable impetus when it was observed that (0.1) was a natural setting in which to understand the estimation of signals, functions and images in wavelet orthonormal bases. In turn, wavelet bases made it possible to give a linked theoretical and methodological account of function estimation that responded appropriately to spatial inhomogeneties in the data, such as (in an extreme form) discontinuities and cusps.

The goal of this book is to give an introductory account of some of the theory of estimation in the Gaussian sequence model that reflects these ideas.

Estimators are studied and compared using the tools of statistical decision theory, which for us means typically (but not always) comparison of mean squared error over appropriate classes of sets Θ supposed to contain the unknown vector θ. The best-worst-case or minimax principle is used, though deliberately more often in an approximate way than exactly. Indeed, we look for various kinds of approximate *adaptive* minimaxity, namely estimators that are able to come close to the minimax criterion simultaneously over a class of parameter sets. A basic theme is that the geometric characteristics of the parameter sets, which themselves often reflect assumptions on the *type* of smoothness of functions, play a critical role.

In the larger first part of the book, Chapters 1- 9, an effort is made to give "equal time" to some representative linear and non-linear estimation methods. Linear methods, of which kernel estimators, smoothing splines, and truncated series approaches are typical examples, are seen to have excellent properties when smoothness is measured in a sufficiently spatially uniform way. When squared error loss is used, this is geometrically captured by the use of hyperrectangles and ellipsoids. Non linear methods, represented here primarily by thresholding of data in a wavelet transform domain, come to the fore when smoothness of a less uniform type is permitted. To keep the account relatively self-contained, introductions to topics such as Gaussian decision theory, wavelet bases and transforms, and smoothness classes of functions are included. A more detailed outline of topics appears in Section 1.6 after an expanded introductory discussion. Starred sections contain more technical material and can be skipped on a first reading.

The second part of the book, Chapters 10– 15, is loosely organized as a tour of various types of asymptotic optimality in the context of estimation in the sequence model. Thus, one may be satisfied with optimality "up to log terms", or "up to constants" or "with exact constants". One might expect that as the demands on quality of optimality are ratcheted up, so are the corresponding assumptions, and that the tools appropriate to the task change. In our examples, intended to be illustrative rather than exhaustive, this is certainly the case. The other organizing theme of this second part is a parallel discussion of results for simple or "monoresolution" models (which need have nothing to do with wavelets) and conclusions specifically for multiresolution settings.

We often allow the noise level ϵ in (0.1) to depend on the index i –a small enough change to be easily accommodated in many parts of the theory, but allowing a significant expansion in models that are fairly directly convertible to sequence form. Thus, many linear inverse problems achieve diagonal form through a singular value or wavelet-vaguelette decompo-

sition, and problems with correlated Gaussian noise can be diagonalized by the principal compoent or Karhunen-Loève transformation.

Of course much is omitted. To explain some of the choices, we remark that the project began over ten years ago as an account of theoretical properties of wavelet shrinkage estimators based largely on work with David Donoho, Gérard Kerkyacharian and Dominique Picard. Much delay in completion ensued, due to other research and significant administrative distractions. This history has shaped decisions on how to bring the book to light after so much elapsed time. First among the choices has been to cast the work more as a graduate text and less as a current research monograph, which is hopefully especially apparent in the earlier chapters. Second, and consistent with the first, the book does not attempt to do justice to related research in recent years, including for example the large body of work on non-orthogonal regression, sparse linear models and compressive sensing. It is hoped, however, that portions of this book will provide helpful background for readers interested in these areas as well.

The intended readership, then, includes graduate students and others who would like an introduction to this part of the theory of Gaussian estimation, and researchers who may find useful a survey of a part of the theory. Helpful background for reading the book would be familiarity with mathematical statistics at the level of a first year doctoral course in the United States.

Acknowledgements [in progress]

This project has an absurdly long history and a corresponding list of debts of gratitude. The prehistory begins with a DMV seminar in March 1995 at Oberwolfach on wavelets in statistics, jointly with Dave Donoho, and June 1996 course at Kasteel de Berkct in the Netherlands organized by Piet Groeneboom.

The transition from LaTeX slides to blackboard exposition marks the true beginning of the book, and I am grateful to Lucien Birgé, Olivier Catoni and Pascal Massart for the invitation to give an advanced course at the École Normale Supérieure in Paris in Spring of 1998, and for the scientific and personal welcome extended by them and by Gérard Kerkyacharian, Dominique Picard and Alexander Tsybakov.

I warmly thank my coauthors: particularly Dave Donoho, with whom much of the wavelets in statistics work began, and repeat offenders Gérard Kerkyacharian, Dominique Picard and Bernard Silverman, as well as our friends Felix Abramovich, Yoav Benjamini, Jeff Hoch, Brenda MacGibbon, Alan Stern, and the late Marc Raimondo, who is sorely missed.

For encouragement and thoughtful comments on the manuscript, I'm greatly indebted to Felix Abramovich, Peter Bickel, Larry Brown, Emmanuel Candès, Shingchang Kou, Yi Lin, Brenda MacGibbon, Stéphane Mallat, Boaz Nadler, Michael Nussbaum, and John Rice as well as to the (then) students in courses at Berkeley and Stanford – Ery Arias Castro, Arnab Chakraborty, Jiashun Jin, Arthur Lu, Zongming Ma, Charles Mathis, Debhashis Paul, Hualin Wang. Some very valuable suggestions came from anonymous reviewers commissioned by John Kimmel and Lauren Cowles.

For the final push, I wish to specially thank Tony Cai, whose encouragement to complete the book took the concrete form of insightful counsel along with organizing further helpful comments from our colleagues Weidong Liu, Mark Low, Lie Wang, Ming Yuan and Harry

Zhou. Michael Martin and Terry O'Neill at the Australian National University, and Marta Sanz at the University of Barcelona, hosted a sabbatical leave which enabled the challenging task of imposing final discipline on a protracted project.

Thanks also to the John Simon Guggenheim Memorial Foundation for a Fellowship during which the first draft was written, and to the National Science Foundation and National Institutes of Health, which have supported much of my own research and writing, and to the Australian National University and University of Barcelona which provided space and time for writing.

Chapter dependency graph. A heavy solid line indicates a more than incidental scientific dependence of the higher numbered chapter on the lower numbered one. A dotted line indicates a weaker formal dependence, perhaps at the level of motivation. A more specific indication of cross-chapter dependence at the level of sections can then be found below.

In the first part, Chapters 2 and 3 provide basic material for the book as a whole, while the decision theory of Chapter 4 is important for virtually everything that follows. The linear estimation results, Chapters 5 and 6, form one endpoint in themselves. Chapters 8 and 9 on thresholding and properties of wavelet shrinkage form the other main endpoint in Part I; the wavelet primer Chapter 7 prepares the way.

In the second part, with numbers shown in Courier font, there some independence of the chapters at a formal level: while they lean heavily on Part I, the groups $\{10\}, \{11, 12\}$ and $\{13, 14, 15\}$ can be read separately of one another. The first chapter in each of these three groups (for Ch. 10, the first half) does not require any wavelet/multiresolution ideas.

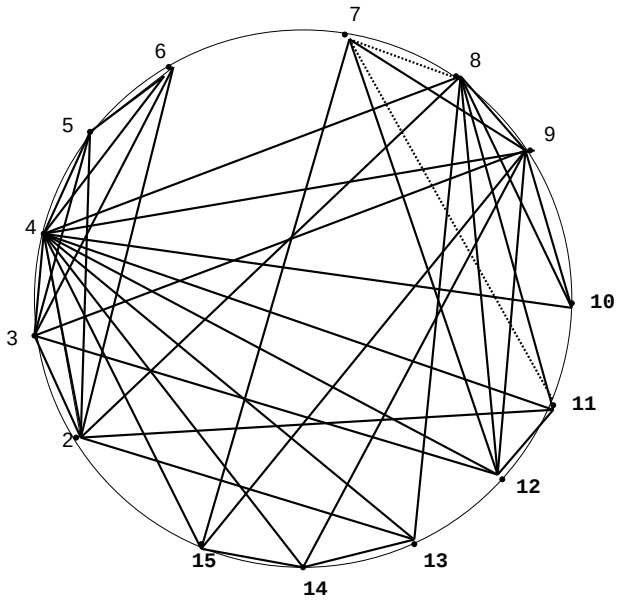

Figure 0.1 Chapter Dependencies

1

Introduction

> And hither am I come, a Prologue armed,... to tell you, fair beholders, that our play leaps o'er the vaunt and firstlings of those broils, beginning in the middle; starting thence away to what may be digested in a play. (Prologue, *Troilus and Cressida* William Shakespeare.)

The study of linear methods, non-linear thresholding and sparsity in the special but central setting of Gaussian data is enlightened by statistical decision theory. This overture chapter introduces these themes and the perspective to be adopted.

Section 1.1 begins with two data examples, in part to emphasize that while this is a theoretical book, the motivation for the theory comes from describing and understanding the properties of commonly used methods of estimation.

A first theoretical comparison follows in Section 1.2, using specially chosen cartoon examples of sparse signals. In order to progress from constructed cases to a plausible theory, Section 1.3 introduces, still in a simple setting, the formal structures of risk function, Bayes rules and minimaxity that are used throughout.

The signal in Gaussian white noise model, the main object of study, makes its appearance in Section 1.4, in both continuous and sequence forms, along with informal connections to finite regression models and spline smoothing estimators. Section 1.5 explains briefly why it is our guiding model; but it is the goal of the book to flesh out the story, and with some of the terms now defined, Section 1.6 provides a more detailed roadmap of the work to follow.

1.1 A comparative example

We use two real data examples to introduce and motivate some of the themes of the book. In the first case, (quasi-)linear methods of estimation seem more or less adequate, while in the second we see substantial improvement by the use of non-linear wavelet thresholding.

The temperature data. Figure 1.1 shows daily minimum temperatures Y_l in degrees Celsius recorded in Canberra, Australia in the leap year 2008. A smoother summary curve might be helpful to see the temperature trend shorn of day to day variability.

We might adopt as a (provisional, approximate) model

$$Y_l = f(t_l) + \sigma Z_l, \qquad l = 1, \ldots, n. \tag{1.1}$$

Here Y_l is the observed minimum temperature at a fixed time period t_l, here equally spaced, with $n = 366$, $f(t)$ is an unknown mean temperature function, while Z_l is a noise term,

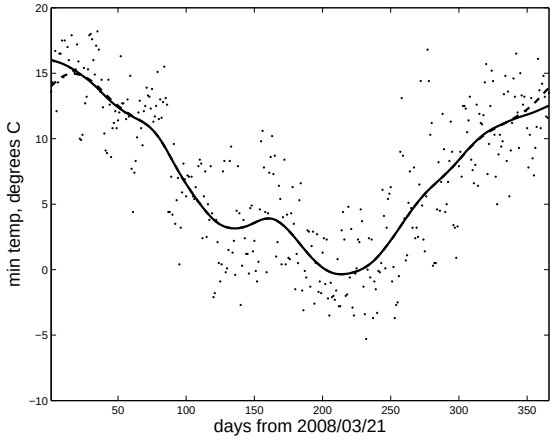

Figure 1.1 Spline smoothing of Canberra temperature data. Solid line: original spline fit, Dashed line: periodic spline

assumed to have mean zero, and variance one—since the standard deviation σ is shown explicitly.

Many approaches to smoothing could be taken, for example using local averaging with a kernel function or using local (linear) regression. Here we briefly discuss two versions of smoothing splines informally—Section 1.4 has formulas and a little more detail. The choice of splines here is merely for definiteness and convenience—what is important is that the estimators are *linear* in the data Y, and depend on a tuning or bandwidth parameter λ.

A least squares approach would seek an estimator $\hat{f}$ to minimize a residual sum of squares $S(f) = n^{-1}\sum_l [Y_l - f(t_l)]^2$. In nonparametric estimation, in which f is unconstrained, this would lead to an interpolation, $\hat{f}(t_l) = Y_l$, an overfitting which would usually be too rough to use as a summary. The spline approach brings in a penalty for roughness, for example $P(f) = \int (f'')^2$ in terms of the squared second derivative of f. The spline estimator is then chosen to minimize $S(f) + \lambda P(f)$, where the *regularization parameter* λ adjusts the relative importance of the two terms.

As both S and P are quadratic functions, it is not surprising (and verified in Section 1.4) that the minimizing $\hat{f}_\lambda$ is indeed linear in the data Y for a given value of λ. As λ increases from 0 to ∞, the solution will pass from rough (interpolating the data) to smooth (the linear least squares fit). A subjective choice of λ was made in Figure 1.1, but it is often desirable to have an "automatic" or data-driven choice specified by some algorithm.

Depending on whether one's purpose is to obtain a summary for a given year (2008) or to obtain an indication of an annual cycle, one may or may not wish to specifically require f and $\hat{f}_\lambda$ to be periodic. In the periodic case, it is natural to do the smoothing using Fourier series. If y_k and f_k denote the kth Fourier coefficient of the observed data and unknown function respectively, then the periodic linear spline smoother takes on the sim-

ple co-ordinatewise linear form $\hat{f}_k = y_k/(1 + \lambda w_k)$ for certain known constants w_k that increase with frequency like k^4.

Interestingly, in the temperature example, the periodic and nonperiodic fits are similar, differing noticeably only within a short distance of the year boundaries. This can be understood in terms of an 'equivalent kernel' form for spline smoothing, Section 3.5.

To understand the properties of linear estimators such as $\hat{f}_\lambda$, we will later add assumptions that the noise variables Z_l are Gaussian and independent. A probability plot of residuals in fact shows that these temperature data are reasonably close to Gaussian, though not independent, since there is a clear lag one sample autocorrelation. However the dependence appears to be short-range and appropriate adjustments for it could be made in a detailed analysis of this example.

The NMR data. Figure 1.2 shows a noisy nuclear magnetic resonance (NMR) signal sampled at $n = 2^J = 1024$ points. Note the presence both of sharp peaks and baseline noise. The additive regression model (1.1) might again be appropriate, this time with $t_l = l/n$ and perhaps with f substantially less smooth than in the first example.

The right hand panel shows the output of wavelet denoising. We give a brief description of the method using the lower panels of the figure—more detail is found in Chapter 7. The noisy signal is transformed, via an orthogonal discrete wavelet transform, into wavelet coefficients y_{jk}, organized by scale (shown vertically, from coarse level $j = 4$ to finest level $j = J - 1 = 9$) and by location, shown horizontally, with coefficients located at $k2^{-j}$ for $k = 1, \ldots, 2^j$. In this transform domain, we perform a hard thresholding

$$\hat{\theta}_{jk} = \begin{cases} y_{jk} & \text{if } |y_{jk}| > \hat{\sigma}\sqrt{2\log n}, \\ 0 & \text{otherwise} \end{cases}$$

to retain only the "large" coefficients, setting all others to zero. Here $\hat{\sigma}$ is a robust estimate of the error standard deviation[1]. The factor $\sqrt{2\log n}$ reflects the likely size of the largest of n independent zero mean standard normal random variables—Chapter 8 has a detailed discussion.

The thresholded coefficients, shown in the lower right panel, are then converted back to the time domain by the inverse discrete wavelet transform, yielding the estimated signal in the top right panel. The wavelet "denoising" seems to be remarkably effective at removing nearly all of the baseline noise, while preserving much of the structure of the sharp peaks.

By contrast, the spline smoothing approach cannot accomplish both these properties at the same time. The right panel of Figure 1.3 shows a smoothing spline estimate with an automatically chosen[2] value of λ. Evidently, while the peaks are more or less retained, the spline estimate has been unable to remove all of the baseline noise.

An intuitive explanation for the differing behaviors of the two estimates can be given using the idea of kernel averaging, in which a function estimate $\hat{f}(x) = \sum_l w_l(x)Y_l$ is obtained by averaging the data Y_l with a weight function

$$w_l(x) = h^{-1}K(h^{-1}(x - x_l)), \tag{1.2}$$

for a suitable kernel function K, usually non-negative and integrating to 1. The parameter

[1] using the median absolute deviation $MAD\{y_{J-1,k}\}/0.6745$, explained in Section 7.5

[2] chosen to minimize an unbiased estimate of mean squared error, Mallows C_L, explained in Section 6.4

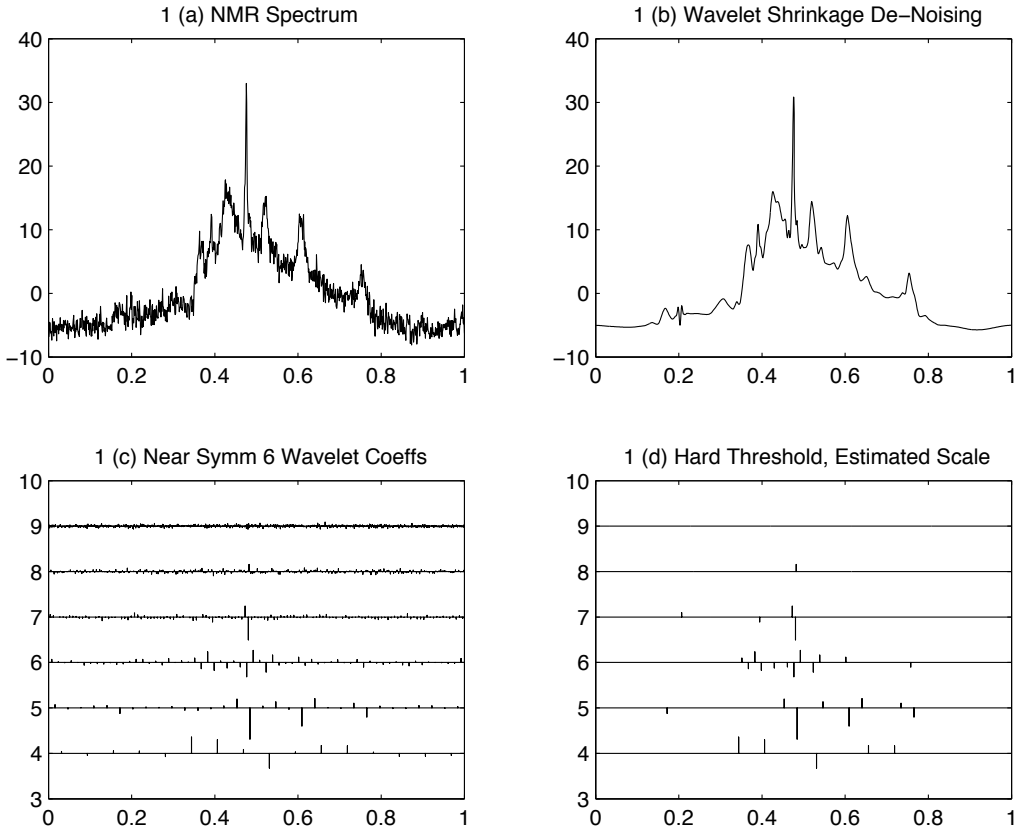

Figure 1.2 Wavelet thresholding of the NMR signal. Data originally via Chris Raphael from the laboratory of Andrew Maudsley, then at UCSF. Signal has $n = 1024$ points, discrete wavelet transform using Symmlet6 filter in `Wavelab`, coarse scale $L = 4$, hard thresholding with threshold $\hat{\sigma}\sqrt{2\log n}$ as in the text.

h is the "bandwidth", and controls the distance at over which observations contribute to the estimate at point x. The spline smoothing estimator, for equally spaced data, can be shown to have approximately this form, with a one-to-one correspondence between h and λ described in Chapter 6.4. A key property of the spline estimator is that the value of h does not vary with x.

By contrast, the kernel average view of the wavelet threshold estimate in Figure 1.2 shows that $h = h(x)$ depends on x strongly - the bandwidth is small in a region of sharp transients, and much larger in a zone of "stationary" behavior in which the noise dominates. This is shown schematically in Figure 1.3, but can be given a more precise form, as is done in Section 7.5.

One of the themes of this book will be to explore the reasons for the difference in performance of splines and wavelet thresholding in these examples. An important ingredient can be seen by comparing the lower panels in Figure 1.2. The true signal—assuming that we can speak of such a thing—appears to be concentrated in a relatively small number of

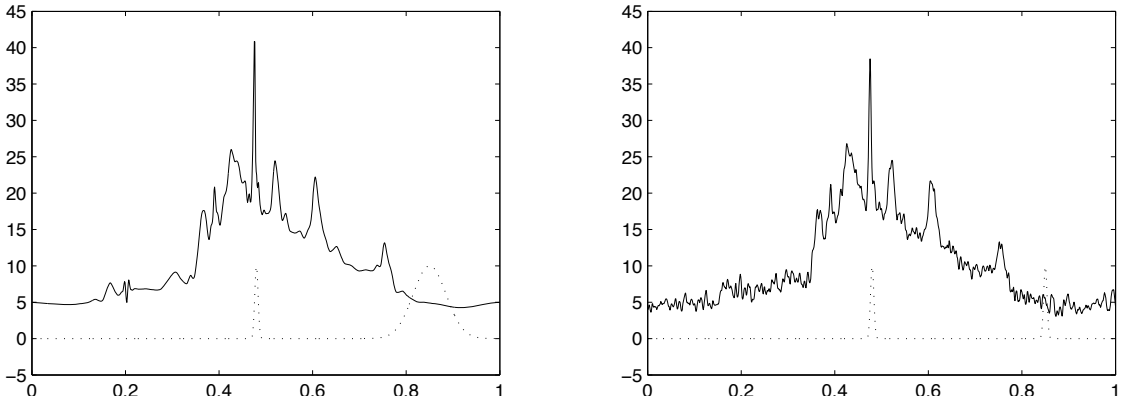

Figure 1.3 Schematic comparison of averaging kernels: The baseline dashed bell curves give qualitative indications of the size of the bandwidth h in (1.2), the equivalent kernel. In the left panel, corresponding to wavelet thresholding, the equivalent kernel depends on position, $h = h(x_l)$, whereas in the right panel, for spline smoothing, it is translation invariant.

wavelet coefficients, while the noise is scattered about globally and at an apparently constant standard deviation within and across levels. Thus the thresholding can literally clean out most of the noise while leaving the bulk of the signal energy, concentrated as it is in a few coefficients, largely undisturbed. This *sparsity of representation* of the signal in the wavelet transform domain is an essential property.

The example motivates a number of questions:

- *what are the properties of thresholding?* Can we develop expressions for, say, mean squared error and understand how to choose the value of the threshold?
- *when is it effective – e.g. better than linear shrinkage?* Can we compare the mean squared error of linear estimators and thresholding over various classes of functions, representing different amounts and types of smoothness?
- *what is the role of sparsity?* Can we develop quantitative measures of sparsity of representation and describe how they affect the possible mean squared error?
- *are optimality statements possible?* Can we identify assumptions on classes of functions for which it is possible to assert that linear, or threshold, estimators are, in an appropriate sense, nearly best?
- *are extensions to other settings possible?* Are there other nonparametric estimation problems, such as density estimation or linear inverse problems, in which similar phenomena appear?

Our goal will be to develop some theoretical definitions, tools and results to address these issues. A key technique throughout will be to use "sequence models", in which our methods, hypotheses and results are phrased in terms of the coefficients that appear when the function f is expanded in an orthogonal basis. In the NMR example, the (wavelet) coefficients are those in the bottom panels of Figure 1.2, while in the weather data, in the periodic form, they are the Fourier coefficients.

In the next section we turn to a first discussion of these questions in the simplest sequence model.

1.2 A first comparison of linear methods, sparsity and thresholding

We begin with a simple model, with an n-dimensional observation vector $y \sim N_n(\theta, \epsilon^2 I)$ with θ being the unknown mean and ϵ^2 the variance, assumed known.[3] We will study a sequence form of the model,

$$y_k = \theta_k + \epsilon z_k, \qquad z_k \overset{\text{i.i.d.}}{\sim} N(0,1). \tag{1.3}$$

which may be obtained by taking coefficients in *any* orthonormal basis. We might call this a "monoresolution" model when we wish to think of what is going on at a single level in the wavelet transform domain, as in the bottom panels of Figure 1.2.

Assume now that the $\{\theta_k\}$ are random, being drawn independently from a Gaussian prior distribution $N(0, \tau^2)$. The posterior distribution of θ_k given the data y is also Gaussian, and the Bayes estimator is given by the posterior mean

$$\hat{\theta}_k = \frac{\rho}{\rho + 1} y_k, \qquad \rho = \frac{\tau^2}{\epsilon^2}. \tag{1.4}$$

The constant ρ is the squared signal-to-noise ratio. The estimator, sometimes called the Wiener filter, is optimal in the sense of minimizing the posterior expected squared error.

This analysis has two important features. First, the assumption of a Gaussian prior distribution produces an optimal estimator which is a *linear* function of the data y. Second, the estimator does not depend on the choice of orthonormal basis: both the model (1.3) and the Gaussian prior are invariant under orthogonal changes of basis, and so the optimal rule has the same linear shrinkage in all coordinate systems.

In contrast, *sparsity* has everything to do with the choice of bases. Informally, "sparsity" conveys the idea that most of the signal strength is concentrated in a few of the coefficients. Thus a 'spike' signal $\gamma(1, 0, \ldots, 0)$ is much sparser than a 'comb' vector $\gamma(n^{-1/2}, \ldots, n^{-1/2})$ even though both have the same energy, or ℓ_2 norm: indeed these could be representations of the same vector in two different bases. In contrast, noise, almost by definition, is not sparse in any basis. Thus, among representations of signals in various bases, it is the ones that are sparse that will be most easily "denoised".

Figure 1.4 shows part of a reconstructed signal represented in two different bases: panel a) is a subset of 2^7 wavelet coefficients θ^W, while panel b) is a subset of 2^7 Fourier coefficients θ^F. Evidently θ^W has a much sparser representation than does θ^F.

The sparsity of the coefficients in a given basis may be quantified using ℓ_p norms [4]

$$\|\theta\|_p = \left(\sum_1^n |\theta_k|^p \right)^{1/p},$$

which track sparsity for $p < 2$, with smaller p giving more stringent measures. Thus, while

[3] The use of ϵ in place of the more common σ already betrays a later focus on "low noise" asympotics!

[4] in fact, only a *quasi*-norm for $p < 1$, Appendix C.1.

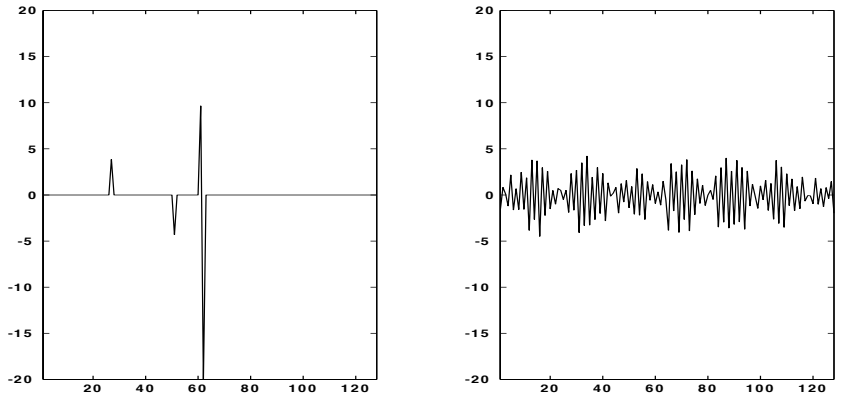

Figure 1.4 Panel (a): θ_k^W =level 7 of estimated NMR reconstruction g of Figure 1.2, while in panel (b): θ_k^F = Fourier coefficients of g at frequencies $65 \ldots 128$, both real and imaginary parts shown. While these do not represent exactly the same projections of f, the two overlap and $\|\theta^F\|_2 = 25.3 \approx 23.1 = \|\theta^W\|_2$.

the ℓ_2 norms of our two representations are roughly equal:

$$\|\theta^F\|_2 = 25.3 \approx 23.1 = \|\theta^W\|_2,$$

the ℓ_1 norm of the sparser representation θ^W is smaller by a factor of 6.5:

$$\|\theta^F\|_1 = 246.5 \gg 37.9 = \|\theta^W\|_1.$$

Figure 1.5 shows that the ℓ_p-norm level sets

$$\left\{\theta : \sum_1^n |\theta_k|^p \le C^p\right\}$$

become progressively smaller and clustered around the co-ordinate axes as p decreases. Thus, the only way for a signal in an ℓ_p ball to have large energy (i.e. ℓ_2 norm) is for it to consist of a few large components, as opposed to many small components of roughly equal magnitude. Put another way, among all signals with a given energy, the sparse ones are precisely those with small ℓ_p norm.

Thus, we will use sets $\{\|\theta\|_p \le C\}$ as quantitative models for *a priori* constraints that the signal θ has an approximately sparse representation in the given basis.

How might we exploit this sparsity information in order better to estimate θ : in other words, can we estimate θ^W better than θ^F? We quantify the quality of estimator $\hat{\theta}(y)$ using Mean Squared Error (MSE):

$$E\|\hat{\theta} - \theta\|^2 = \sum_{k=1}^n E(\hat{\theta}_k - \theta_k)^2. \tag{1.5}$$

Figure 1.6 shows an idealized case in which all θ_k are zero except for two spikes, each of size $1/2$. Assume, for simplicity here, that $\epsilon = \epsilon_n = 1/\sqrt{n}$ and that $p = C = 1$: it is thus

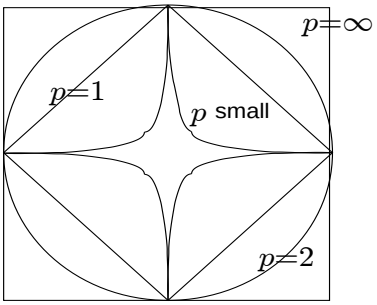

Figure 1.5 Contours of ℓ_p balls

supposed that $\sum_1^n |\theta_k| \le 1$. Consider the class of linear estimators $\hat{\theta}_c(y) = cy$, which have per co-ordinate variance $c^2\epsilon_n^2$ and squared bias $(1-c)^2\theta_k^2$. Consequently, the mean squared error (1.5)

$$MSE = \sum_1^n c^2\epsilon_n^2 + (1-c)^2\theta_k^2 = c^2 + (1-c)^2/2 = \begin{cases} 1 & c = 1 \\ 1/2 & c = 0. \end{cases}$$

The upper right panel shows the unbiased estimate with $c = 1$; this has no bias and only variance. The lower left panels shows $c = 0$ with no variance and only bias. The MSE calculation shows that no value of c leads to a linear estimate with much better error - the minimum MSE is 1/3 at $c = 1/3$. As an aside, if we were interested instead in the absolute, or ℓ_1 error $\sum_k |\hat{\theta}_k - \theta_k|$, we could visualize it using the vertical lines—again this is relatively large for all linear estimates.

In the situation of Figure 1.6, thresholding is natural. As is the preceding section, define the *hard threshold* estimator by its action on coordinates:

$$\hat{\theta}_{\lambda,k}(y) = \begin{cases} y_k & \text{if } |y_k| \ge \lambda\epsilon_n, \\ 0 & \text{otherwise.} \end{cases} \tag{1.6}$$

The lower right panel of Figure 1.6 uses a threshold of $\lambda\epsilon_n = 2.4\epsilon_n = 0.3$. For the particular configuration of true means θ_k shown there, the data from the two spikes pass the threshold unchanged, and so are essentially unbiased estimators. Meanwhile, in all other coordinates, the threshold correctly sets all coefficients to zero except for the small fraction of noise that exceeds the threshold.

In more detail, the mean squared error of thresholding is

$$E(\hat{\theta}_{\lambda,k} - \theta_k)^2 = E\{(y_k - \theta_k)^2, |y_k| \ge \lambda\epsilon_n\} + \theta_k^2 P\{|y_k| \le \lambda\epsilon_n\}. \tag{1.7}$$

If $\theta_k = 0$, we can write $y_k = \epsilon_n z_k$ with $z_k \sim N(0,1)$, and so the mean squared error is approximately

$$\epsilon_n^2 E\{z^2, |z| > \lambda\} \sim 2\epsilon_n^2 \lambda\phi(\lambda). \tag{1.8}$$

However, if θ_k is large relative to $\lambda\epsilon_n$, then the MSE is approximately $E(y_k - \theta_k)^2 = \epsilon_n^2$.

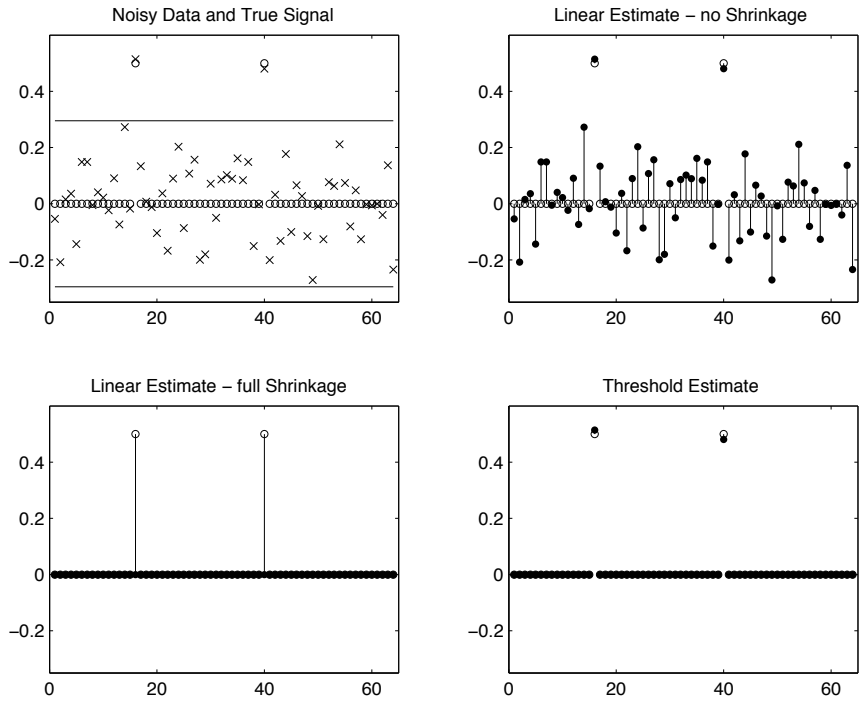

Figure 1.6 (a) Visualization of model (1.3): open circles are unknown values θ_k, crosses are observed data y_k. In the other panels, solid circles show various estimators $\hat{\theta}$, for $k = 1, \ldots, n = 64$. Horizontal lines are thresholds at $\lambda = 2.4\epsilon_n = 0.3$. (b) Vertical lines indicate absolute errors $|\hat{\theta}_{1,k} - \theta_k|$ made by leaving the data alone: $\hat{\theta}_1(y) = y$. (c) Corresponding absolute errors for the zero estimator $\hat{\theta}_0(y) = 0$. (d) Much smaller errors due to hard thresholding at $\lambda = 0.3$.

Hence, in the two spike setting,

$$\begin{aligned} E\|\hat{\theta}_\lambda - \theta\|^2 &\approx 2\epsilon_n^2 + 2(n-2)\epsilon_n^2 \lambda \phi(\lambda) \\ &\approx 2n^{-1} + 2\lambda\phi(\lambda) \approx 0.139 \end{aligned}$$

when $n = 64$ and $\lambda = 2.4$. This mean squared error is of course much better than for any of the linear estimators.

1.3 A game theoretic model and minimaxity

The skeptic will object that the configuration of Figure 1.6 was chosen to highlight the advantages of thresholding, and indeed it was! It is precisely to avoid the possibility of being misled by such reasoning from constructed cases that the tools of game theory have been adapted for use in statistics. A sterner and fairer test of an estimator is obtained by creating a statistical two person zero sum game or *statistical decision problem*. In our setting, this has the following rules:

(i) Player I ("the Statistician") is allowed to choose any estimator $\hat{\theta}(y)$, linear, threshold or of more complicated type.

(ii) Player II ("Nature") may choose a probability distribution π for θ subject only to the sparsity constraint that $E_\pi \|\theta\|_1 \leq 1$.

(iii) The payoff—the loss to the statistician— is calculated as the expected mean squared error of $\hat{\theta}(y)$ when θ is chosen according to π and then the observed data y is drawn from model (1.3): $y = \theta + \epsilon_n z$ for $z \sim N_n(0, I)$. Thus the expected loss, or *risk*, now averages over *both* θ *and* y:

$$B(\hat{\theta}, \pi) = E_\pi E_{y|\theta} \|\hat{\theta}(y) - \theta\|_2^2.$$

Of course, the Statistician tries to minimize the risk and Nature to maximize it.

Classical work in statistical decision theory (Wald, 1950; Le Cam, 1986), Chapter 4 and Appendix A, shows that the minimax theorem of von Neumann can be adapted to apply here, and that the game has a well defined value, the *minimax risk*:

$$R_n = \inf_{\hat{\theta}} \sup_{\pi} B(\hat{\theta}, \pi) = \sup_{\pi} \inf_{\hat{\theta}} B(\hat{\theta}, \pi). \tag{1.9}$$

An estimator $\hat{\theta}^*$ attaining the left hand infimum in (1.9) is called a *minimax* strategy or *estimator* for player I, while a prior distribution π^* attaining the right hand supremum is called *least favorable* and is an optimal strategy for player II. Schematically, the pair of optimal strategies $(\hat{\theta}^*, \pi^*)$ forms a *saddlepoint*, Figure 1.7: if Nature uses π^*, the best the Statistician can do is to use $\hat{\theta}^*$. Conversely, if the Statistician uses $\hat{\theta}^*$, the optimal strategy for Nature is to choose π^*.

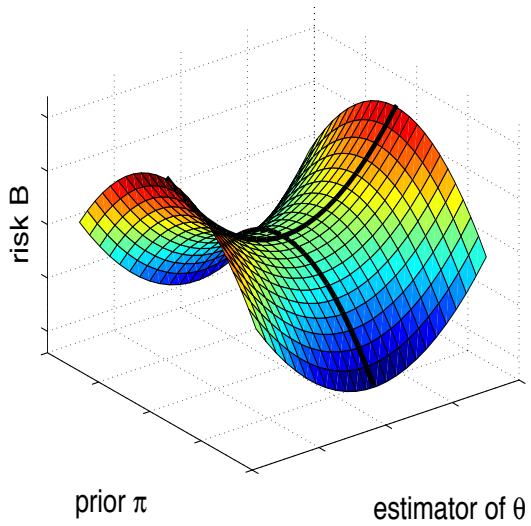

Figure 1.7 Left side lower axis: strategies π for Nature. Right side lower axis: strategies $\hat{\theta}$ for the Statistician. Vertical axis: payoff $B(\hat{\theta}, \pi)$ from the Statistician to Nature. The saddlepoint indicates a pair $(\hat{\theta}^*, \pi^*)$ of optimal strategies.

It is the *structure* of these optimal strategies, and their effect on the minimax risk R_n that is of chief statistical interest.

While these optimal strategies cannot be exactly evaluated for finite n, informative asymptotic approximations are available. Indeed, as will be seen in Section 13.4, an *approximately* least favorable distribution is given by drawing the individual coordinates θ_k independently from a *two point* distribution with

$$\theta_k = \begin{cases} \epsilon_n \sqrt{\log n} & \text{with probability } \alpha_n \doteq 1/\sqrt{n \log n} \\ 0 & \text{otherwise.} \end{cases} \tag{1.10}$$

This amounts to repeated tossing of a coin highly biased towards zero. Thus, in n draws, we expect to see a relatively small number, namely $n\alpha_n = \sqrt{n/\log n}$ of non-zero components. The size of these non-zero values is such that they are hard to distinguish from the larger values among the remaining, more numerous, $n - \sqrt{n/\log n}$ observations that are pure noise. Of course, what makes this distribution difficult for Player I, the Statistician, is that the *locations* of the non-zero components are random as well.

It can also be shown, Chapter 13, that an approximately minimax estimator for this setting is given by the hard thresholding rule described earlier, but with threshold given at least approximately by $\lambda_n = \epsilon_n \sqrt{\log(n \log n)}$. This estimate asymptotically achieves the minimax value

$$R_n \sim \sqrt{\log n/n}$$

for MSE. It can also be verified that no *linear* estimator can achieve a risk less than $1/2$ if Nature chooses a suitably uncooperative probability distribution for θ, Theorem 9.3 and (9.21). Compare Table 1.1.

In the setting of the previous section with $n = 64$ and $\epsilon_n = 1/\sqrt{n}$, we find that $\epsilon_n \sqrt{\log n} = 0.255$ and the expected non-zero number $n\alpha_n = 3.92$. Finally, the threshold $\epsilon_n \sqrt{\log(n \log n)} = .295$.

This—and any—statistical decision problem make a large number of assumptions, including values of parameters that typically are not known in practice. We will return later to discuss the virtues and vices of the minimax formulation. For now, it is perhaps the qualitative features of this solution that most deserve comment. Had we worked with simply a signal to noise constraint, $E_\pi \|\theta\|_2^2 \le 1$, say, we would have obtained a Gaussian prior distribution as being approximately least favorable and the linear Wiener filter (1.4) with $\epsilon_n^2 = \tau_n^2 = 1/n$ as an approximately minimax estimator. The imposition of a sparsity constraint $E_\pi \|\theta\|_1 \le 1$ reflects additional *a priori* information and yields great improvements in the quality of possible estimation, and produces optimal strategies that take us far away from Gaussian priors and linear methods.

1.4 The Gaussian Sequence Model

In this section we introduce the general sequence model, an extension of (1.3) that will be our main focus of study. The observed data are $y = (y_i)$ for i in a discrete index set $\mathcal{I}$ such as $\mathbb{N}$. It is assumed that the components y_i are statistically independent of one another, and follow Gaussian, or normal, distributions with unknown means θ_i and known positive

	Prior Constraint	
	traditional (ℓ_2)	*sparsity* (ℓ_1)
minimax estimator	linear	thresholding
least favorable π	Gaussian	sparse
minimax M.S.E.	$= 1/2$	$\sim \sqrt{\frac{\log n}{n}}$

Table 1.1 *Comparison of structure of optimal strategies in the monoresolution game under traditional and sparsity assumptions.*

variances $\epsilon\lambda_i$. Thus the sequence model may be written as

$$y_i = \theta_i + \epsilon\lambda_i z_i, \qquad z_i \overset{i.i.d}{\sim} N(0,1), \qquad i \in \mathcal{I}. \tag{1.11}$$

The index set will typically be a singleton, $\mathcal{I} = \{1\}$, finite $\mathcal{I} = \{1,\dots,n\}$, or infinite $\mathcal{I} = \mathbb{N}$. Multidimensional index sets, such as $\{1,\dots,n\}^d$ or $\mathbb{N}^d$ are certainly allowed, but will appear only occasionally. The scale parameter ϵ sets the level of the noise, and in some settings will be assumed to be small.

We give a first discussion of models motivating, or leading to, (1.11)—further examples and details are given in Chapters 2 and 3.

Nonparametric regression. In the previous two sections, θ was a vector with no necessary relation among its components. Now we imagine an unknown function $f(t)$. The independent variable t is thought of as low dimensional (1 for signals, 2 for images, 3 for volumetric fields etc.); indeed we largely confine attention to functions of a single variable, say time, in a bounded interval, say $[0,1]$. In a sampled-data model, we might have points $0 \le t_1 \le \cdots \le t_n \le 1$, and

$$Y_l = f(t_l) + \sigma Z_l, \qquad Z_l \overset{iid}{\sim} N(0,1). \tag{1.12}$$

This is the model for the two examples of Section 1.1 with the i.i.d. Gaussian assumption added.

We can regard Y, Z and $\mathbf{f} = (f(t_l))$ as vectors in $\mathbb{R}^n$ and bring in an arbitrary orthonormal basis $\{\boldsymbol{\varphi}_i\}$. For example, if the t_l were equally spaced, this might be the discrete Fourier basis of sines and cosines. In general, collecting the basis vectors as columns of a matrix $U = [\boldsymbol{\varphi}_1 \cdots \boldsymbol{\varphi}_n]$, we have

$$U^T Y = U^T \mathbf{f} + \epsilon U^T Z,$$

which becomes sequence model (1.11) on writing $y = U^T Y, \theta = U^T \mathbf{f}$ and $z = U^T Z$. Thus $y_k = \langle Y, \boldsymbol{\varphi}_k\rangle, \theta_k = \langle \mathbf{f}, \boldsymbol{\varphi}_k\rangle$ and so forth. [5] Here,

$$\langle \mathbf{f}, \mathbf{g}\rangle = n^{-1}\sum_{l=1}^{n} f(t_l)g(t_l)$$

[5] Our index convention: i for the sequence model and l for the time domain. We sometimes use k in place of i in some concrete settings, such as single wavelet resolution level or exact or approximate Fourier frequencies.

denotes the Euclidean inner product on $\mathbb{R}^n$, with corresponding norm $\|\cdot\|$.

We illustrate the reduction to sequence form with the smoothing spline estimator used in Section 1.1, and so we suppose that an estimator $\hat{f}$ of f in (1.12) is obtained by minimizing the penalized sum of squares $S(f) + \lambda P(f)$, or more explicitly

$$Q(f) = n^{-1} \sum_l [Y_l - f(t_l)]^2 + \lambda \int_0^1 (f'')^2. \tag{1.13}$$

The account here is brief; for much more detail see Green and Silverman (1994).

It turns out that a unique minimizer exists and belongs to the space S of "natural cubic splines" – twice continuously differentiable functions that are formed from cubic polynomials on each interval $[t_l, t_{l+1}]$ and are furthermore linear on the outlying intervals $[0, t_1]$ and $[t_n, 1]$. Equally remarkably, the space S has dimension exactly n, and possesses a special orthonormal basis, the *Demmler-Reinsch* basis. This basis consists of functions $\varphi_k(t)$—and associated vectors $\boldsymbol{\varphi}_k = (\varphi_k(t_l))$—that are simultaneously orthogonal both on the set of sampling points and on the unit interval:

$$\langle \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_k \rangle = \delta_{jk} \qquad \text{and} \qquad \int_0^1 \varphi_j'' \varphi_k'' = w_k \delta_{jk}. \tag{1.14}$$

The weights w_k are non-negative and increasing, indeed $w_1 = w_2 = 0$, so that the first two basis functions are linear. For $k \geq 3$, it can be shown that φ_k has $k-1$ sign changes, so that the basis functions exhibit increasing oscillation with k, and this is reflected in the values w_k for the roughness penalty. Because of this increasing oscillation with k, we may think of k as a frequency index, and the Demmler-Reinsch functions as forming a sort of Fourier basis that depends on the knot locations $\{t_i\}$.

This double orthogonality allows us to rewrite the criterion $Q(f)$, for $f \in S$, in terms of coefficients in the Demmler-Reinsch basis:

$$Q(\theta) = \sum_1^n (y_k - \theta_k)^2 + \lambda \sum_1^n w_k \theta_k^2.$$

The charm is that this can now readily be minimized term by term to yield the sequence model expression for the smoothing spline estimate $\hat{\theta}_{SS}$:

$$\hat{\theta}_{SS,k} = c_{\lambda k} y_k = \frac{1}{1 + \lambda w_k} y_k. \tag{1.15}$$

The estimator is thus linear in the data and operates *co-ordinatewise*. It achieves its smoothing aspect by shrinking the higher "frequencies" by successively larger amounts dictated by the increasing weights λw_k. In the original time domain,

$$\hat{\mathbf{f}} = \sum_k \hat{\theta}_{SS,k} \boldsymbol{\varphi}_k = \sum_k c_{\lambda k} y_k \boldsymbol{\varphi}_k. \tag{1.16}$$

There is no shrinkage on the constant and linear terms: $c_{\lambda 1} = c_{\lambda 2} = 1$, but for $k \geq 3$, the shrinkage factor $c_{\lambda k} < 1$ and decreases with increasing frequency. Large values of smoothing parameter λ lead to greater attenuation of the data, and hence greater smoothing in the estimate.

To represent the solution in terms of the original data, gather the basis functions into an $n \times n$ orthogonal matrix $U = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_n]/\sqrt{n}$. Then $\mathbf{Y} = \sqrt{n}Uy$ and $\mathbf{f} = \sqrt{n}U\theta$, and so

$$\hat{\mathbf{f}} = \sqrt{n}U\hat{\theta} = Uc_\lambda U'\mathbf{Y} = c_\lambda \mathbf{Y}. \qquad c_\lambda = \operatorname{diag}(c_{\lambda k}). \tag{1.17}$$

Notice that the change of basis matrix U does not depend on λ. Thus, many important aspects of the spline smoothing problem, such as the issue of choosing λ well from data, can be studied in the diagonal sequence form that the quasi-Fourier basis provides.

Software packages, such as `spline.smooth` in `R`, may use other bases, such as $B-$splines, to actually compute the spline estimate. However, because there is a unique solution to the optimization problem, the estimate computed in practice must coincide, up to numerical error, with (1.17).

We have so far emphasized structure that exists whether or not the points t_l are equally spaced. If, however, $t_l = l/n$ *and* it is assumed that f is periodic, then everything in the approach above has an explicit form in the Fourier basis—Section 3.4.

Continuous Gaussian white noise model. Instead of sampling a function at a discrete set of points, we might suppose that it can be observed—with noise!—throughout the entire interval. This leads to the central model to be studied in this book:

$$Y(t) = \int_0^t f(s)ds + \epsilon W(t), \qquad 0 \le t \le 1, \tag{1.18}$$

which we will sometimes write in an equivalent form, in terms of instantaneous increments

$$dY(t) = f(t)dt + \epsilon dW(t), \qquad 0 \le t \le 1. \tag{1.19}$$

The observational noise consists of a standard Brownian motion W, scaled by the known noise level ϵ. For an arbitrary square integrable function g on $[0, 1]$, we therefore write

$$\int_0^1 g(t)dY(t) = \int_0^1 g(t)f(t)dt + \epsilon \int_0^1 g(t)dW(t). \tag{1.20}$$

The third integral features a deterministic function g and a Brownian increment dW and is known as a Wiener integral. We need only a few properties of standard Brownian motion and Wiener integrals, which are recalled in Appendix C.8.

The function Y is observed, and we seek to recover the unknown function f, assumed to be square integrable: $f \in L_2[0, 1]$, for example using the integrated squared error loss

$$\|\hat{f} - f\|_{L_2}^2 = \int_0^1 (\hat{f} - f)^2.$$

To rewrite the model in sequence form, we may take any orthonormal basis $\{\varphi_i\}$ for $L_2[0, 1]$. Examples include the Fourier basis, or any of the classes of orthonormal wavelet bases to be discussed later. To set notation for the coefficients, we write

$$y_i = Y(\varphi_i) = \int_0^1 \varphi_i dY, \quad \theta_i = \langle f, \varphi_i \rangle = \int_0^1 f\varphi_i \quad z_i = W(\varphi_i) = \int_0^1 \varphi_i dW. \tag{1.21}$$

From the stationary and independent increments properties of Brownian motion, the Wiener

integrals z_i are Gaussian variables that have mean 0 and are uncorrelated:

$$\mathrm{Cov}(z_i, z_j) = E\Big[\int_0^1 \varphi_i\, dW \cdot \int_0^1 \varphi_j\, dW\Big] = \int_0^1 \varphi_i \varphi_j\, dt = \delta_{ij}.$$

[The Kronecker delta $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.] As a result, the continuous Gaussian model is entirely equivalent to the constant variance sequence model

$$y_i = \theta_i + \epsilon z_i \qquad \text{with} \quad z_i \overset{\text{iid}}{\sim} N(0,1). \tag{1.22}$$

The Parseval relation, (C.1), converts squared error in the function domain to the analog in the sequence setting:

$$\int_0^1 (\hat{f} - f)^2 = \sum_i (\hat{\theta}_i - \theta_i)^2. \tag{1.23}$$

Linking regression and white noise models. Heuristically, the connection between (1.12) and (1.18) arises by forming the partial sum process of the discrete data, now assumed to be equally spaced, $t_l = l/n$:

$$Y_n(t) \triangleq \frac{1}{n}\sum_1^{[nt]} Y_l = \frac{1}{n}\sum_1^{[nt]} f\left(\frac{l}{n}\right) + \frac{\sigma}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_1^{[nt]} Z_l. \tag{1.24}$$

The signal term is a Riemann sum approximating $\int_0^t f$, and the error term $n^{-\frac{1}{2}}\sum^{[nt]} Z_l$ converges weakly to standard Brownian motion as $n \to \infty$. Making the calibration $\epsilon = \epsilon(n) = \sigma/\sqrt{n}$, and writing $Y_{\epsilon(n)}$ for the process in (1.18), we see that, formally, the processes $Y_{\epsilon(n)}(t)$ and $Y_n(t)$ merge as $n \to \infty$. A formal statement and proof of this result is given in Chapter 3.10, using the notion of asymptotic equivalence of statistical problems, which implies closeness of risks for all decision problems with bounded loss. Here we simply observe that heuristically there is convergence of mean average squared errors. Indeed, for fixed functions $\hat{f}$ and $f \in L^2[0,1]$:

$$n^{-1}\|\hat{f} - f\|_{2,n}^2 = n^{-1}\sum_1^n [\hat{f}(i/n) - f(i/n)]^2 \to \int_0^1 [\hat{f} - f]^2.$$

Non white noise models. So far we have discussed only the constant variance subclass of models (1.11) in which $\lambda_i \equiv 1$. The scope of (1.11) is considerably broadened by allowing unequal $\lambda_i > 0$. Here we make only a few remarks, however, deferring further discussion and examples to Chapters 2 and 3.

When the index set I is finite, say $\{1, \ldots, n\}$, two classes of multivariate Gaussian models lead to (1.11):

(i) $Y \sim N(\theta, \epsilon^2\Sigma)$, by transforming to an orthogonal basis that diagonalizes Σ, so that (λ_i) are the eigenvalues of Σ, and

(ii) $Y \sim N(A\theta, \epsilon^2 I)$, by using the singular value decomposition of $A = \sum_i b_i u_i v_i^T$ and setting $y_i = b_i^{-1} Y_i$, so that $\lambda_i = b_i^{-1}$ are the inverse singular values.

When the index set I is countably infinite, case (i) corresponds to a Gaussian process with unknown mean function f and the sequence form is obtained from the Karhunen-Loève transform (Section 3.9). Case (ii) corresponds to observations in a linear inverse problem

with additive noise, $Y = Af + \epsilon Z$, in which we do not observe f but rather its image Af after the action of a linear operator A, representing some form of integration, smoothing or blurring. The conversion to sequence form is again obtained using a singular value decomposition, cf. Chapter 3.

1.5 Why study the sequence model?

While the sequence models (1.3) and (1.11) are certainly idealizations, there are several reasons why they repay detailed study.

(i) simplicity. By focusing on sequences of independent Gaussian variables, we can often do exact calculations. Generally, it turns out that all the issues are fundamental rather than merely technical. In parametric statistics, the analogy would be with study of the multivariate normal model after use of the central limit theorem and other asymptotic approximations.

(ii) depth. The model makes it possible to focus directly on important and profound phenomena, such as the Stein effect, in which maximum likelihood estimates of three or more mean parameters can be (often significantly) improved by shrinkage toward a point or subspace. Similarly, the "concentration of measure" phenomenon for product measures in high dimensional spaces (such as our Gaussian error distributions) plays an important role.

(iii) relevance. The sequence models and estimators used in them turn out to be close *enough* to actual methods to yield useful insights. Thus the contrast between linear estimators and thresholding is able to explain more or less fully some practically important phenomena in function estimation.

The finite dimensional multivariate normal model is the foundation of parametric statistical theory. For nonparametric statistics, the continuous signal in Gaussian white noise model, or its sequence version expressed in an orthonormal basis, plays an equivalent role. It first emerged in communications theory in work of Kotelnikov (1959). As Ibragimov and Has'minskii (1981); Ibragimov and Khas'minskii (1980, for example) have argued, the difficulties thrown up by the "signal+noise" model are essential rather than technical in nature.

1.6 Plan of the book

In the Canberra temperature and NMR data examples we saw that linear spline and nonlinear wavelet threshold estimators were respectively quite successful. The examples illustrate a basic point that, in function estimation, as elsewhere in statistics, an optimal or at least good choice of method will depend on the circumstances of the problem.

The theory to be developed in this book will formulate classes of assumptions under which linear estimators can perform well, and then move to circumstances in which coordinatewise thresholding is optimal, either in "monoresolution" or "multiresolution" settings.

The chapters are grouped into two parts. In the first, chapters 2–9 contain a sampling of material of broadest interest. In the second, Chapters 10 –15 then go into greater detail about optimality results for thresholding-type estimators in both 'monoresolution' and multiresolution models.

We use the ideas and tools of statistical decision theory, particularly Bayes rules and

minimaxity, throughout; introductory material appears in Chapters 2–4 and especially in Chapter 4. Chapters 5–6 focus primarily on optimality properties of linear estimators, especially using geometric properties of parameter spaces such as hyperrectangles and ellipsoids. Pinsker's theorem on the asymptotic optimality of linear rules over ellipsoids is discussed in Chapter 5. Chapter 6 introduces the notion of adaptive optimality—the ability of an estimaor to perform 'optimally' over a scale of parameter spaces without having to depend on *a priori* assumptions about parameters of those spaces. The James-Stein estimator is seen to lead to a class of adaptively minimax estimators that is quite similar to certain smoothing spline or kernel estimators that are commonly used in practice.

The focus then turns to the phenomena of sparsity and non-linear estimation via coordinatewise thresholding. To set the stage, Chapter 7 provides a primer on orthonormal wavelet bases and wavelet thresholding estimation. Chapter 8 focuses on the properties of thresholding estimators in the "sparse normal means" model: $y \sim N_n(\theta, \sigma^2 I)$ and the unknown vector θ is assumed to be sparse. Chapter 9 explores the consequences of these thresholding results for wavelet shrinkage estimation, highlighting the connection between sparsity, non-linear approximation and statistical estimation.

Part II is structured around a theme already implicit in Chapters 8 and 9: while wavelet bases are specifically designed to analyze signals using multiple levels of resolution, it is helpful to study initially what happens with thresholding etc. at a single resolution scale both for other applications, and before assembling the results across several scales to draw conclusions for function estimation.

Thus Chapters 10–14 are organized around two strands: the first strand works at a single or mono-resolution level, while the second develops the consequences in multiresolution models. Except in Chapter 10, each strand gets its own chapter. Three different approachs are explored—each offers a different tradeoff between generality, sharpness of optimality, and complexity of argument. We consider in turn

(i) optimal recovery and 'universal' thresholds (Ch. 10)
(ii) penalized model selection (Chs. 11, 12)
(iii) minimax-Bayes optimal methods (Chs. 13, 14)

The Epilogue, Chapter 15 has two goals. The first is to provide some detail on the comparison between discrete and continuous models. The second is to mention some recent related areas of work not covered in the text. The Appendices collect background material on the minimax theorem, functional classes, smoothness and wavelet decompositions.

1.7 Notes

Related books and monographs. The book of Ibragimov and Has'minskii (1981), along with their many research papers has had great influence in establishing the central role of the signal in Gaussian noise model. Textbooks on nonparametric estimation include Efromovich (1999) and Tsybakov (2008), which include coverage of Gaussian models but range more widely, and Wasserman (2006) which is even broader, but omits proofs.

Closer to the research level are the St. Flour courses by Nemirovski (2000) and Massart (2007). Neither are primarily focused on the sequence model, but do overlap in content with some of the chapters of this book. Ingster and Suslina (2003) focuses largely on hypoth-

esis testing in Gaussian sequence models. References to books focusing on wavelets and statistics are collected in the notes to Chapter 7.

Exercises

1.1 Let $y \sim N_n(\theta, \epsilon_n^2 I)$ and $\hat{\theta}_\lambda$ denote the hard thresholding rule (1.6). Let $r(\lambda, \theta_k; \epsilon_n) = E(\theta_{\lambda,k} - \theta_k)^2$ denote the risk (mean squared error) in a single co-ordinate.
(i) for the two point prior given in (1.10), express the Bayes risk $B(\hat{\theta}_\lambda, \pi) = E_\pi E_{y|\theta} \|\hat{\theta}_\lambda - \theta\|_2^2$ in terms of $\theta \to r(\lambda, \theta; \epsilon_n)$.
(ii) Using (1.7), derive the bound

$$r(\lambda, \mu\epsilon_n; \epsilon_n) \le (1 + \mu^2)\epsilon_n^2.$$

(iii) Using also (1.8), verify that

$$B(\hat{\theta}_\lambda, \pi) \le \sqrt{\log n / n} \cdot (1 + o(1)).$$

[This gives the risk for a 'typical configuration' of θ drawn from the least favorable prior (1.10). It does not yet show that the minimax risk R_n satisfies this bound. For a simple, but slightly suboptimal, bound see Theorem 8.1; for the actual argument, Theorems 13.6, 13.8 and 13.16].

2

The multivariate normal distribution

> We know not to what are due the accidental errors, and precisely because we do not know, we are aware they obey the law of Gauss. Such is the paradox. (Henri Poincaré, *The Foundations of Science.*)

Estimation of the mean of a multivariate normal distribution, $y \sim N_n(\theta, \sigma_0^2 I)$, is the elemental estimation problem of the theory of statistics. In parametric statistics it is sometimes plausible as a model in its own right, but more often occurs–perhaps after transformation–as a large sample approximation to the problem of estimating a finite dimensional parameter governing a smooth family of probability densities.

In nonparametric statistics, it serves as a building block for the study of the infinite dimensional Gaussian sequence model and its cousins, to be introduced in the next chapter. Indeed, a recurring theme in this book is that methods and understanding developed in the finite dimensional Gaussian location model can be profitably transferred to nonparametric estimation estimation.

It is therefore natural to start with some definitions and properties of the finite Gaussian location model for later use. Section 2.1 introduces the location model itself, and an extension to known diagonal covariance that later allows a treatment of certain correlated noise and linear inverse problem models.

Two important methods of generating estimators, regularization and Bayes rules, appear in Sections 2.2 and 2.3. Although both approaches can yield the same estimators, the distinction in point of view is helpful. Linear estimators arise from quadratic penalties/Gaussian priors, and the important conjugate prior formulas are presented. Non-linear estimators arise from ℓ_q penalties for $q < 2$, including the soft and hard thresholding rules, and from sparse mixture priors that place atoms at 0, Section 2.4.

Section 2.5 begins the comparative study of estimators through their mean squared error properties. The bias and variance of linear estimators are derived and it is shown that sensible linear estimators in fact *must* shrink the raw data. The James-Stein estimator explodes any hope that we can get by with linear methods, let alone the maximum likelihood estimator. Its properties are cleanly derived using Stein's unbiased estimator of risk; this is done in Section 2.6.

Soft thresholding consists of pulling each co-ordinate y_i towards, but not past, 0 by a threshold amount λ. Section 2.7 develops some of its properties, including a simple oracle inequality which already shows that thresholding outperforms James-Stein shrinkage on sparse signals, while James-Stein can win in other 'dense' settings.

Section 2.8 turns from risk comparison to probability inequalities on the tails of Lipschitz functions of a multivariate normal vector. This "concentration" inequality is often useful in high dimensional estimation theory; the derivation given has points in common with that of Stein's unbiased risk estimate.

Section 2.9 makes some remarks on more general linear models $Y = A\beta + \sigma e$ with correlated Gaussian errors e, and how some of these can be transformed to diagonal sequence model form.

2.1 Sequence models

The simplest finite white Gaussian sequence model has

$$y_i = \theta_i + \epsilon z_i, \qquad i = 1, \ldots, n. \tag{2.1}$$

Here (y_i) represents the observed data. The signal (θ_i) is unknown—there are n unknown parameters. The (z_i) are independent $N(0, 1)$ noise variables, and ϵ is the noise level, which for simplicity we generally assume to be known. The model is called *white* because the noise level ϵ is the same at all indices, which often represent increasing frequencies. Typically we will be interested in estimation of θ.

Equation (2.1) can also be written in the multivariate normal mean form $y \sim N_n(\theta, \epsilon^2 I)$ that is the central model for classical parametric statistical theory. We write $\phi_\epsilon(y - \theta) = \prod_i \phi_\epsilon(y_i - \theta_i)$ for the joint density of (y_i) with respect to Lebesgue measure. The univariate densities $\phi_\epsilon(y_i) = (2\pi\epsilon^2)^{-1/2} \exp\{-y_i^2/2\epsilon^2\}$. We put $\phi = \phi_1$ and $\Phi(y) = \int_{-\infty}^{y} \phi(s)ds$ for the standard normal density and cumulative distribution function.

Two generalizations considerably extend the scope of the finite sequence model. In the first, corresponding to indirect or inverse estimation,

$$y_i = \alpha_i \theta_i + \epsilon z_i, \qquad i = 1, \ldots, n, \tag{2.2}$$

the constants α_i are known and positive. In the second, relevant to correlated noise,

$$y_i = \theta_i + \epsilon \lambda_i z_i, \qquad i = 1, \ldots, n. \tag{2.3}$$

Here again the constants λ_i are known and positive. Of course these two models are equivalent in the sense that dividing by α_i in the former and setting $\lambda_i = 1/\alpha_i$ and $y_i' = y_i/\alpha_i$ yields the latter. In this sense, we may regard (2.3) as describing the general case. In Section 2.9, we review some Gaussian linear models that can be reduced to one of these sequence forms.

Among the issues to be addressed are

(i) we imagine (θ_i) to be "high dimensional". In particular, as ϵ decreases, the number of parameters $n = n(\epsilon)$ may increase. This makes the problem fundamentally *nonparametric*.
(ii) what are the effects of (α_i) or (λ_i), i.e. the consequences of indirect estimation, or correlated noise, on the ability to recover θ?
(iii) asymptotic behavior as $\epsilon \to 0$. This corresponds to a low-noise (or large sample size) limit.
(iv) optimality questions: can one describe bounds for minimum possible error of estimation and estimators that (more or less) achieve these bounds?

2.2 Penalized Least Squares, Regularization and thresholding

Two common, and related, methods of deriving and motivating estimators are via penalized least squares and via Bayes rules. We discuss the first here and the second in the next section.

We begin with model (2.2), which for a moment we write in matrix form $Y = A\theta + \epsilon z$, with $A = \mathrm{diag}(\alpha_i)$. The unbiased and least squares estimate of θ is found by minimizing $\theta \to \|Y - A\theta\|_2^2$. If θ is high dimensional, we may wish to *regularize* the solution by introducing a *penalty function* $P(\theta)$, and minimizing instead the penalized least squares criterion

$$Q(\theta) = \|Y - A\theta\|_2^2 + \lambda P(\theta).$$

Since A is diagonal, the "data term" is a sum of individual components and so it is natural to require that the penalty also be additive: $P(\theta) = \sum p_i(\theta_i)$, so that

$$Q(\theta) = \sum_i (y_i - \alpha_i\theta_i)^2 + \lambda p_i(\theta_i),$$

Two simple and commonly occurring penalty functions are *quadratic*: $P(\theta) = \sum \omega_i\theta_i^2$ for some non-negative constants ω_i, and q^{th} *power*: $P(\theta) = \|\theta\|_q^q = \sum_{i=1}^n |\theta_i|^q$.

The crucial *regularization parameter* λ determines the relative weight given to the sum of squared error and penalty terms: much more will be said about this later. As λ varies from 0 to $+\infty$, we may think of the penalized estimates $\hat\theta(\lambda)$ as forming a path from the roughest, least squares solution $\hat\theta(0) = (y_i/\alpha_i)$ to the smoothest solution $\hat\theta(\infty) = 0$.

Since $Q(\theta)$ has an additive structure, it can be minimized term by term, leading to a univariate optimization for each coefficient estimate $\hat\theta_i$. This minimization can be done explicitly in each of three important cases.

(i) ℓ_2 penalty: $p_i(\theta_i) = \omega_i\theta_i^2$. By differentiation, we obtain a co-ordinatewise linear shrinkage estimator of ridge type

$$\hat\theta_i(y) = \frac{\alpha_i}{\alpha_i^2 + \lambda\omega_i} y_i. \tag{2.4}$$

(ii) ℓ_1 penalty: $p(\theta_i) = 2|\theta_i|$. We take $\alpha_i \equiv 1$ here for convenience. Considering only a single co-ordinate and dropping subscripts i, we have

$$Q(\theta) = (y - \theta)^2 + 2\lambda|\theta|,$$

so that for $\theta \neq 0$,

$$\tfrac{1}{2}Q'(\theta) = -y + \theta + \lambda\,\mathrm{sgn}(\theta),$$

while at $\theta = 0$, the derivative is replaced by the subdifferential, the interval $[-y-\lambda, -y+\lambda]$. Consequently, the ℓ_1-penalized least squares estimate uses *soft thresholding* at threshold λ:

$$\hat\theta_\lambda(y) = \begin{cases} y - \lambda & y > \lambda \\ 0 & |y| \le \lambda \\ y + \lambda & y < -\lambda. \end{cases} \tag{2.5}$$

As evident from Figure 2.1, the estimator $\hat\theta_\lambda$ is characterized by a threshold zone $y \in$

$[-\lambda, \lambda]$, in which all data is set to 0, and by shrinkage toward 0 by a fixed amount λ whenever y lies outside the threshold zone: $|y| > \lambda$. The thresholding is called 'soft' as it is a continuous function of input data y. When applied to vectors $y = (y_i)$, it typically produces sparse fits, with many co-ordinates $\hat{\theta}_{\lambda,i} = 0$, with larger values of λ producing greater sparsity.

(iii) ℓ_0 penalty. $p(\theta_i) = I\{\theta_i \neq 0\}$. The total penalty counts the number of non-zero coefficients:

$$P(\theta) = \sum_i p(\theta_i) = \#\{i : \theta_i \neq 0\}.$$

Again considering only a single coordinate, and writing the regularization parameter as λ^2,

$$Q(\theta) = (y - \theta)^2 + \lambda^2 I\{\theta \neq 0\}.$$

By inspection,

$$\min_\theta Q(\theta) = \min\{y^2, \lambda^2\},$$

and the ℓ_0-penalized least squares estimate is given by *hard thresholding* at threshold λ:

$$\hat{\theta}_\lambda = \begin{cases} y & |y| > \lambda \\ 0 & |y| \leq \lambda. \end{cases} \tag{2.6}$$

This estimator 'keeps' or 'kills' the data y according as it lies outside or inside the threshold zone $[-\lambda, \lambda]$. Again $\hat{\theta}_\lambda$ produces sparse fits (especially for large λ), but with the difference that there is no shrinkage of retained coefficients. In particular, the estimate is no longer a continuous function of the data.

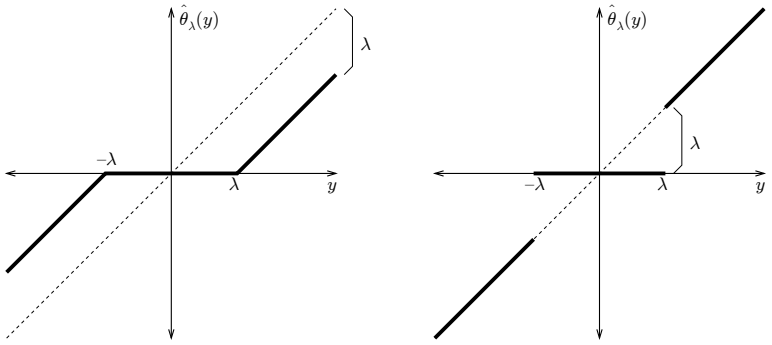

Figure 2.1 Left panel: soft thresholding at λ, showing threshold zone and shrinkage by λ towards 0 outside threshold zone. Dashed line is 45 degree line. Right panel: hard thresholding, with no shrinkage outside the threshold zone.

2.3 Priors, posteriors and Bayes estimates

We will make heavy use of the Bayesian machinery of priors and posteriors and of the decision theoretic ideas of loss functions and Bayes estimators. The ideas and notation are introduced informally here; some more detail is postponed to Chapter 4.

Suppose we have a prior probability distribution $\pi(d\theta)$ on $\mathbb{R}^n$, and a family of sampling distributions $P(dy|\theta)$, namely a collection of probability measures on the sample space $\mathcal{Y} = \mathbb{R}^n$ indexed by θ. Then there is a joint distribution πP on $\Theta \times \mathcal{Y}$ and two factorizations into marginal and conditional distributions:

$$\pi P(d\theta, dy) = \pi(d\theta)P(dy|\theta) = P_\pi(dy)\pi(d\theta|y).$$

Here $P_\pi(dy)$ is the marginal distribution of y and $\pi(d\theta|y)$ the posterior for θ given y.

Now suppose that all sampling distributions have densities with respect to Lebesgue measure, $P(dy|\theta) = p(y|\theta)dy$. Then the marginal distribution also has a density with respect to Lebesgue measure, $P_\pi(dy) = p(y)dy$, with

$$p(y) = \int p(y|\theta)\pi(d\theta),$$

and we arrive at *Bayes formula* for the posterior distribution

$$\pi(d\theta|y) = \frac{p(y|\theta)\pi(d\theta)}{p(y)}.$$

A *loss function* associates a loss $L(a, \theta) \geq 0$ with each pair (a, θ) in which $a \in \mathbb{R}^n$ denotes an action, or estimate, chosen by the statistician, and $\theta \in \mathbb{R}^n$ denotes the true parameter value. Typically $L(a, \theta) = w(a - \theta)$ is a function of $a - \theta$. Our chief examples here will be quadratic and $q-$th power losses:

$$w(t) = t^T Qt, \qquad w(t) = \|t\|_q^q = \sum_{i=1}^n |t_k|^q.$$

Here Q is assumed to be positive definite. Given a prior distribution π and observed data y, the *posterior expected loss* (or *posterior risk*)

$$E_y L(a, \theta) = \int L(a, \theta)\pi(d\theta|y)$$

is a function of a (and y). The *Bayes estimator* corresponding to loss function L is obtained by minimizing the posterior expected loss:

$$\hat{\theta}_\pi(y) = \text{argmin}_a E_y L(a, \theta). \tag{2.7}$$

For now, we assume that a unique minimum exists, and ignore measure theoretic niceties. Another, equivalent definition, is given in Chapter 4.

The *Bayes risk* of prior π is the expected value—with respect to the marginal distribution of y—of the posterior expected loss of $\hat{\theta}_\pi$:

$$B(\pi) = E_{P_\pi} E_y L(\hat{\theta}_\pi(y), \theta). \tag{2.8}$$

Example 1. Quadratic loss and posterior mean. Suppose that $L(a, \theta) = (a-\theta)^T Q(a-\theta)$ for some positive definite matrix Q. Then $a \to E_y L(a, \theta)$ has a unique minimum, given by the zero of

$$\nabla_a E_y L(a, \theta) = 2Q[a - E_y\theta],$$

and so the Bayes estimator for a quadratic loss function is just the posterior mean

$$\hat{\theta}_\pi(y) = E_y\theta = E(\theta|y). \tag{2.9}$$

Note, in particular, that this result does not depend on the value of $Q > 0$. The posterior expected loss of $\hat{\theta}_\pi$ is given by

$$E[L(\hat{\theta}_\pi, \theta)|y] = E[\theta - E(\theta|y)]^T Q[\theta - E(\theta|y)] = \mathrm{tr}[Q\mathrm{Cov}(\theta|y)].$$

Conjugate priors for the multivariate normal. Suppose that the sampling distributions $P(dy|\theta)$ is multivariate Gaussian $N_n(\theta, \Sigma)$ and that the prior distribution $\pi(d\theta)$ is also Gaussian: $N_n(\theta_0, T)$. Then the marginal distribution $P_\pi(dy)$ is $N(\theta_0, \Sigma + T)$ and the posterior distribution $\pi(d\theta|y)$ is also multivariate normal $N(\theta_y, \Sigma_y)$–this is the conjugate prior property. Perhaps most important are the formulas for the posterior mean and covariance matrix:

$$\theta_y = (\Sigma^{-1} + T^{-1})^{-1}(\Sigma^{-1}y + T^{-1}\theta_0), \qquad \Sigma_y = (\Sigma^{-1} + T^{-1})^{-1} \tag{2.10}$$

and the equivalent forms

$$\theta_y = T(T + \Sigma)^{-1}y + \Sigma(T + \Sigma)^{-1}\theta_0, \qquad \Sigma_y = T - T(T + \Sigma)^{-1}T. \tag{2.11}$$

Before the derivation, some remarks:

The posterior mean θ_y is a weighted average of the data y and the prior mean θ_0: the first formula shows that the weights are given by the data and prior *precision* matrices Σ^{-1} and T^{-1} respectively. The posterior precision Σ_y^{-1} is the sum of the prior and data precision matrices, and notably, does not depend on the data y! Hence, in this case, the Bayes risk (2.8) is just $B(\pi) = \mathrm{tr}Q\Sigma_y$.

In the important special case in which the prior mean $\theta_0 = 0$, then $\theta_y = Sy$ is a linear shrinkage rule, shrinking toward 0.

The quadratic regularization estimates discussed in the previous section can be interpreted as Bayes estimates for suitable priors. In the orthogonal setting ($A = I$), estimate (2.4) corresponds to posterior mean (2.10) for a prior $\theta \sim N(0, \lambda^{-1}\Omega^{-1})$ with $\Omega = \mathrm{diag}(\omega_i)$ and sampling variance $\Sigma = I$.

Proof Recall the basic formula for conditional distributions in the multivariate normal setting. Namely, if

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right]$$

with $\Sigma_{21} = \Sigma_{12}'$, then

$$\begin{aligned} y_1|y_2 &\sim N(\theta_{1|2}, \Sigma_{1|2}) \\ \theta_{1|2} &= \theta_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \theta_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

Apply this to the joint distribution that is implied by the assumptions on sampling distribution and prior, after noting that $\mathrm{Cov}(\theta, y) = T$,

$$\begin{pmatrix} \theta \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} \theta_0 \\ \theta_0 \end{pmatrix}, \begin{pmatrix} T & T \\ T & T + \Sigma \end{pmatrix}\right]$$

which yields formulas (2.11) for the posterior mean and variance, after noting that

$$I - T(T + \Sigma)^{-1} = \Sigma(T + \Sigma)^{-1}.$$

Formulas (2.10) may then be recovered by matrix algebra, using the identity

$$T - T(T + \Sigma)^{-1}T = (T^{-1} + \Sigma^{-1})^{-1}. \qquad \square$$

Exercise 2.1 gives an alternate derivation that leads directly to formulas (2.10).

Product priors and posteriors. Suppose that the components of the prior are independent, $\pi(d\theta) = \prod_i \pi_i(d\theta_i)$, *and* the sampling distributions are independent, each depending on only one θ_i, so that $P(dy|\theta) = \prod_i P(dy_i|\theta_i)$. Then the posterior distribution factorizes also:

$$\pi(d\theta|y) = \prod_i \pi(d\theta_i|y_i). \tag{2.12}$$

In this situation, then, calculations can be done co-ordinatewise, and are hence generally much simpler.

Additive Loss Functions take the special form

$$L(a, \theta) = \sum_i \ell(a_i, \theta_i). \tag{2.13}$$

Under the assumption of product joint distributions, we have just seen that the posterior distribution factorizes. In this case, the $k-$th component of the posterior expected loss

$$E_y \ell(a_i, \theta_i) = \int \ell(a_i, \theta_i)\pi(d\theta_i|y_i)$$

can be computed based on (a_i, y_i) alone. As a result, the posterior expected loss $E_y L(a, \theta)$ can be minimized term by term, and so the Bayes estimator

$$\hat{\theta}_\pi(y) = \text{argmin}_{(a_i)} E_y \sum_i \ell(a_i, \theta_i) = (\hat{\theta}_{\pi_i}(y_i)) \tag{2.14}$$

is *separable*: the $k-$th component of the estimator depends only on y_i.

Consider in particular the $q-$th power loss functions

$$L_q(a, \theta) = \sum_i |a_i - \theta_i|^q.$$

The preceding discussion on separability allows us to focus on a single co-ordinate, and

$$\hat{\theta}_{\pi_1}(y_1) = \text{argmin}_a \int |a - \theta_1|^q \pi(d\theta_1|y_1).$$

The posterior expected loss on the right side is strictly convex if $q > 1$. Some particular cases are familiar: $q = 2$ corresponds to the posterior mean, $q = 1$ to the posterior *median*, and $q = 0$ to the posterior *mode* (for discrete π). Indeed, for the case $q = 1$, recall the standard fact that $a \to \int |a - \theta| F(d\theta)$ is minimized at any median a_0, a point a_0 for which $F((-\infty, a_0]) \geq \frac{1}{2}$ *and* $F([a_0, \infty)) \geq \frac{1}{2}$.

To explain the expression $q = 0$, note that since $\lim_{q\to 0} |d|^q = I\{d \neq 0\}$, we may think of $L_0(a,\theta) = \#\{k : a_i \neq \theta_i\}$ as counting error. For a *discrete* prior, $\pi(d\theta) = \sum_1^r p_i \delta_{\theta_i}(d\theta)$, and we observe that

$$E[L_0(a_1,\theta_1)|y] = P(\theta_1 \neq a_1|y_1)$$

is minimized by choosing $a_1 = \text{argmax}_{\theta_i} P(\{\theta_i\}|y_1)$, in other words, the posterior mode—the most likely discrete value of θ given the observed data.

In the next section, we look at some examples involving the posterior median. For the remainder of this section, we return to squared error loss and consider the Gaussian sequence model.

Suppose, consistent with (2.3), that the sampling distributions of $y_i|\theta_i$ are independently $N(\theta_i,\sigma_i^2)$, for $i = 1,\dots,n$. Assume independent conjugate priors $\theta_i \sim N(\theta_{0i},\tau_i^2)$. This is just the diagonal form of the multivariate Gaussian model considered earlier. Putting $\Sigma = \text{diag}(\sigma_i^2)$ and $T = \text{diag}(\tau_i^2)$ into the earlier formulas (2.10)- (2.11) yields the marginal distribution $y_i \sim N(\theta_{0i},\sigma_i^2+\tau_i^2)$. The posterior law has $\theta_i|y_i \sim N(\theta_{y,k},\sigma_{y,k}^2)$, with the two formulas for the posterior mean given by

$$\theta_{y,k} = \frac{\sigma_i^{-2}y_i + \tau_i^{-2}\theta_{0i}}{\sigma_i^{-2}+\tau_i^{-2}} = \frac{\tau_i^2 y_i + \sigma_i^2\theta_{0i}}{\tau_i^2+\sigma_i^2}, \tag{2.15}$$

and the forms for the posterior variance being

$$\sigma_{y,k}^2 = \frac{1}{\sigma_i^{-2}+\tau_i^{-2}} = \frac{\tau_i^2\sigma_i^2}{\tau_i^2+\sigma_i^2}. \tag{2.16}$$

Thus, for example, the posterior mean

$$\theta_{y,k} \approx \begin{cases} \theta_{0,k} & \text{if } \sigma_i^2 \gg \tau_i^2, \\ y_i & \text{if } \tau_i^2 \gg \sigma_i^2, \end{cases}$$

corresponding to very concentrated and very vague prior information about θ respectively.

Remark on notation. Formulas are often simpler in the case of unit noise, $\epsilon = 1$, and we reserve a special notation for this setting: $X \sim N_n(\mu, I)$, or equivalently

$$x_i = \mu_i + z_i, \qquad z_i \overset{i.i.d.}{\sim} N(0,1), \tag{2.17}$$

for $i = 1,\dots,n$. It is usually easy to recover the formulas for general ϵ by rescaling.

Examples. 1. There is a useful analytic expression for the posterior mean in the Gaussian shift model $X \sim N_n(\mu, I)$. Writing $p = \pi \star \phi$ for the marginal density of x, we have

$$\hat{\mu}_\pi(x) = \int \mu\phi(x-\mu)\pi(d\mu)/p(x).$$

The standard Gaussian density satisfies

$$\frac{\partial}{\partial x_i}\phi(x) = -x_i\phi(x),$$

and so by rewriting $\mu = x + (\mu - x)$, we arrive at

$$\hat{\mu}_\pi(x) = x + \frac{\nabla p(x)}{p(x)} = x + \nabla \log p(x), \tag{2.18}$$

which represents the Bayes rule as the perturbation of the maximum likelihood estimator $\hat{\mu}_0(x) = x$ by a logarithmic derivative of the marginal density of the prior. [Remark on use.]

If the prior $\pi(d\mu) = \gamma(\mu)d\mu$ has a differentiable density that satisfies, for all μ,

$$\|\nabla \log \gamma(\mu)\| \le \Lambda, \tag{2.19}$$

then representation (2.18) shows that $\hat{\mu}_\pi(x)$ has *bounded shrinkage*: $\|\hat{\mu}_\pi(x) - x\| \le \Lambda$ for all x. Indeed, observing that $(\partial/\partial x_i)\phi(x-\mu) = -(\partial/\partial\mu_i)\phi(x-\mu)$, we have

$$(\partial p/\partial x_i)(x) = \int -(\partial\phi/\partial\mu_i)(x-\mu)\gamma(\mu)d\mu = \int (\partial\gamma/\partial\mu_i)\phi(x-\mu)d\mu$$

where we used (2.19) to conclude that $\gamma(\mu)\phi(x-\mu) \to 0$ as $\mu \to \infty$. Consequently,

$$\|\nabla \log p(x)\| \le \int \|\nabla \log \gamma(\mu)\|\phi(x-\mu)\gamma(\mu)d\mu \big/ p(x) \le \Lambda. \tag{2.20}$$

2. Discrete priors will play an important role at several points in this book. Here consider the simplest case, a symmetric two point prior concentrated on $\{-\tau, \tau\}$:

$$\pi_\tau = \tfrac{1}{2}(\delta_\tau + \delta_{-\tau}).$$

The posterior also concentrates on $\{-\tau, \tau\}$, but with posterior probabilities given by

$$\pi(\{\tau\}|x) = \frac{\frac{1}{2}\phi(x-\tau)}{\frac{1}{2}\phi(x-\tau) + \frac{1}{2}\phi(x+\tau)} = \frac{e^{x\tau}}{e^{x\tau} + e^{-x\tau}}, \tag{2.21}$$

so that

$$\pi(\{\tau\}|x) > \pi(-\tau|x) \quad \text{iff} \quad x > 0. \tag{2.22}$$

The posterior mean lies between $-\tau$ and $+\tau$:

$$\hat{\mu}_\tau(x) = E_\pi(\mu|x) = \tau \tanh \tau x, \tag{2.23}$$

and the posterior variance is found (try it!) to be

$$E\big[\big(\mu - E_\pi(\mu|x)\big)^2|x\big] = \frac{\tau^2}{\cosh^2 \tau x},$$

and the Bayes risk

$$B(\pi_\tau) = \tau^2 e^{-\tau^2/2} \int_{-\infty}^{\infty} \frac{\phi(x)dx}{\cosh \tau x}. \tag{2.24}$$

2.4 Sparse mixture priors and thresholding

A simple model for a sparse high dimensional vector is that its components are drawn i.i.d. from a distribution with a (large) atom at 0 and the remaining probability from a density on $\mathbb{R}\backslash\{0\}$. Such *sparse mixture priors* will occur in several later chapters. In this section, continuing our initial exploration of Bayes estimators, we explore some properties of a simple class of such priors, and focus in particular on the properties of the posterior *median* as a thresholding rule.

Consider then the sequence model (2.17) with noise level 1. Let the co-ordinates μ_i be distributed i.i.d as the 'sparse prior'

$$\pi(d\mu) = (1-w)\delta_0(d\mu) + w\gamma(\mu)d\mu. \tag{2.25}$$

Thus, with probability $1-w$, the mean μ_i is zero, while with probability w, the value is drawn from a probability density $\gamma(\mu)$ with respect to Lebesgue measure, which we assume to be *symmetric* and *unimodal*. The non-zero probability $w = \pi\{\mu \neq 0\}$ can take any value in $[0,1]$, but is small in sparse cases. The marginal density for $x = x_1$ is

$$p(x) = \int \phi(x-\mu)\pi(d\mu) = (1-w)\phi(x) + wg(x),$$

where the convolution $g(x) = \phi \star \gamma(x) = \int \phi(x-\mu)\gamma(\mu)d\mu$. The posterior density [1] is given by

$$\pi(\mu|x) = \begin{cases} (1-w)\phi(x)/p(x) & = 1-w(x) \quad \text{if } \mu = 0 \\ w\gamma(\mu)\phi(x-\mu)/p(x) & = w(x)\gamma(\mu|x) \quad \text{if } \mu \neq 0, \end{cases} \tag{2.26}$$

where the posterior non-zero probability $w(x) = P(\mu \neq 0|x) = wg(x)/p(x)$ and the conditional posterior $\gamma(\mu|x) = \gamma(\mu)\phi(x-\mu)/g(x)$. Expressing probabilities p in terms of *odds* $p/(1-p)$, we have

$$\frac{w(x)}{1-w(x)} = \frac{P(\mu \neq 0 \mid x)}{P(\mu = 0 \mid x)} = \frac{w}{1-w}\frac{g(x)}{\phi(x)}.$$

Thus the ratio of posterior to prior odds of nonzero μ equals the density ratio $g(x)/\phi(x)$.

We verify that this ratio is monotone increasing. Decompose $g(x)$ into parts $g_+(x)$ and $g_-(x)$ corresponding to integrals over $\mu > 0$ and $\mu < 0$ respectively. Thus, for example

$$(g_-/\phi)(x) = \int_{-\infty}^{0} e^{x\mu - \mu^2/2}\gamma(\mu)d\mu. \tag{2.27}$$

Since the prior density γ is symmetric about 0, this transforms to an integral over $(0,\infty)$ simply by changing the sign of x, and so

$$(g/\phi)(x) = 2\int_0^{\infty} \cosh(x\mu)e^{-\mu^2/2}\gamma(\mu)d\mu, \tag{2.28}$$

which increases from $(g/\phi)(0) < 1$ to $+\infty$ as x increases from 0 to $+\infty$. This accords with the natural intuition that the posterior odds of $\mu \neq 0$ should increase as the observed value x moves further away from 0—in either direction.

[1] here, to be precise, density means Radon-Nikodym derivative with respect to the (slightly non-standard) dominating measure $\nu(d\mu) = \delta_0(d\mu) + d\mu$.

We can now show that use of the sparse prior model (2.17) and (2.25) and the posterior median—Bayes estimate for ℓ_1 loss—implies the existence of a *threshold zone*, along with the other qualitative properties illustrated in Figure 2.2.

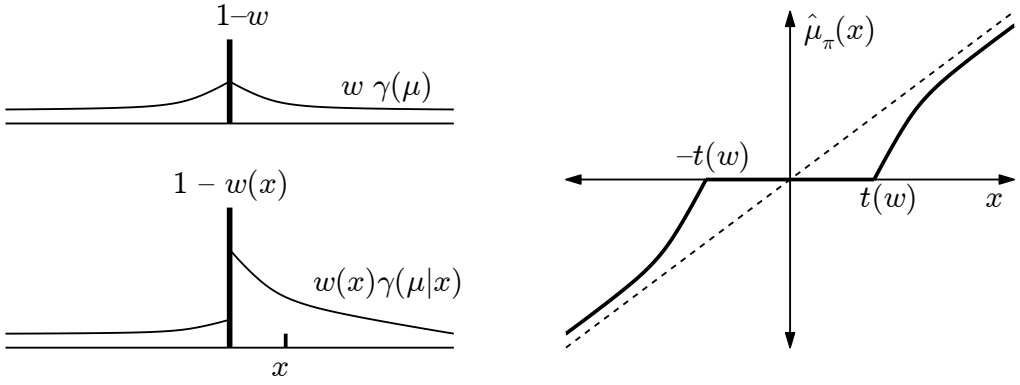

Figure 2.2 Top left: sparse prior with atom of probability $1 - w$ at 0, bottom left: posterior density after observing $x > 0$, with atom $1 - w(x)$ at 0. Right: posterior median estimator $\hat{\mu}_\pi(x)$ showing threshold zone $x \in [-t(w), t(w)]$.

Proposition 2.1 *Suppose that the prior has mixture form* (2.25) *for* $w > 0$ *and that the non-zero density* $\gamma(\mu)$ *is symmetric and unimodal. The posterior median* $\hat{\mu}_\pi(x)$ *is*

(a) monotone in x *and antisymmetric:* $\hat{\mu}(-x) = -\hat{\mu}(x)$,

(b) a shrinkage rule: $0 \le \hat{\mu}(x) \le x$ *for* $x \ge 0$,

(c) a threshold rule: there exists $t(w) > 0$ *such that* $\hat{\mu}(x) = 0$ *if and only if* $|x| \le t(w)$.

(d) Finally, the threshold $t(w)$*, as a function of* w*, is continuous and strictly decreasing from* $t = \infty$ *at* $w = 0$ *to* $t = 0$ *at* $w = 1$.

Proof (a). It is helpful at several points to work with odds and odds ratios, due to felicitous cancellations. Thus, if $\mu < \mu'$ and $x < x'$, then

$$\frac{\pi(\mu'|x')\pi(\mu|x)}{\pi(\mu|x')\pi(\mu'|x)} = \exp\{(\mu' - \mu)(x' - x)\} > 1.$$

Moving the denominator to the right side and integrating with respect to the dominating measure over $\mu \le m$ and $\mu' > m$, we obtain

$$P(\mu > m|x')P(\mu \le m|x) \ge P(\mu \le m|x')P(\mu > m|x).$$

Consequently $P(\mu \le m|x)$ is decreasing in x and so the posterior median $\hat{\mu}(x)$ is monontone increasing in x. The anti-symmetry of the posterior median is immediate from the symmetry of the prior and the Gaussian error density.

(b). From the first expression in (2.26) and the definition of g_-, we have

$$P(\mu < 0|x) = \frac{w g_-(x)}{(1 - w)\phi(x) + w g(x)}. \tag{2.29}$$

If $x > 0$, then $g_-(x) < g(x)/2$ using the symmetry of γ, and so $P(\mu < 0|x) < 1/2$ and

hence the posterior median $\hat{\mu}(x) \geq 0$. The posterior probability $P(\mu < x|x)$ may be written as a similar ratio, now with numerator $w \int_{-\infty}^{x} \phi(x-\mu)\gamma(\mu)d\mu$, and so a similar comparison argument shows that $P(\mu < x|x) \geq 1/2$, so that $\hat{\mu}(x) \leq x$.

(c). From the posterior density (2.26) we know that $P\{\mu = 0 \mid x\} > 0$ if the non-zero weight $w < 1$. Since by symmetry $P\{\mu < 0 \mid x = 0\} = P\{\mu > 0 \mid x = 0\}$, we conclude that

$$P\{\mu < 0 \mid x = 0\} < \tfrac{1}{2} < P\{\mu \leq 0 \mid x = 0\}$$

so that $\hat{\mu}_\pi(0) = 0$, which is also clear by reason of symmetry. More importantly, from (2.29), (2.27) and the monotonicity of $(g/\phi)(x)$, we see that both $x \to P\{\mu < 0 \mid x\}$ and $P\{\mu \leq 0 \mid x\}$ are continuous and strictly decreasing functions, and so the previous display remains valid on an *interval*: $-t(w) \leq x \leq t(w)$, which is the threshold zone property.

(d). From the analog of (2.29) for $P(\mu > 0|x)$, the threshold $t = t(w)$ satisfies

$$2wg_+(t) = (1-w)\phi(t) + wg(t).$$

Dividing by $w\phi(t)$, and rearranging in a manner analogous to (2.28) yields

$$w^{-1} = 1 + (g_+ - g_-)/\phi = 1 + 2\int_0^\infty \sinh(t\mu)e^{-\mu^2/2}\gamma(\mu)d\mu.$$

This equation shows that w is a continuous and strictly increasing function of t, from $w = 1$ at $t = 0$ to $w = 0$ at $t = \infty$. □

The tails of the prior density γ have an important influence on the amount of shrinkage of the posterior median. Exercise 2.3 outlines the proof of

Proposition 2.2 *Assume that the prior density has logarithmic derivative bounded by* Λ, (2.19). *Then the posterior median has bounded shrinkage: for all* x,

$$|\hat{\mu}(x; w) - x| \leq t(w) + \Lambda + 2. \tag{2.30}$$

Remark. The condition (2.19) implies, for $u > 0$, that $\log \gamma(u) \geq \log \gamma(0) - \Lambda u$ and so, for all u, that $\gamma(u) \geq \gamma(0)e^{-\Lambda|u|}$. Hence, for bounded shrinkage, the assumption requires the tails of the prior to be exponential or heavier. Gaussian priors do not satisfy (2.30), and indeed the shrinkage is then proportional to x for large x. Heuristically, this may be seen by arguing that the effect of the atom at 0 is negligible for large x, so that the posterior is essentially Gaussian, so that the posterior median equals the posterior mean, and is given, from (2.16) by

$$\tau^2 y/(\tau^2+1) = y - y/(\tau^2+1).$$

Concrete examples. Two priors which are suited to explicit numerical calculation in software are the Laplace density

$$\gamma_a(\mu) = \tfrac{1}{2}ae^{-a|\mu|},$$

which satisfies (2.19), and the quasi-Cauchy density

$$\gamma(\mu) = \frac{1}{\sqrt{2\pi}}\left[1 - \frac{|\mu|\tilde{\Phi}(|\mu|)}{\phi(\mu)}\right],$$

so named because the tails of the density decay like $1/\mu^2$. Here, as usual, $\tilde{\Phi}(t) = 1 - \Phi(t)$. It arises as a scale mixture of normals $\mu|\tau \sim N(0, \tau^{-1} - 1)$ with $\tau \sim \text{Beta}(\frac{1}{2}, 1)$.

For example, in the case of the Laplace density, the following formulas may be verified (Exercise 2.4 fills in some details.) First, define

$$\beta(x) = \frac{g(x)}{\phi(x)} - 1 = \frac{a}{2}\left[\frac{\Phi}{\phi}(x-a) + \frac{\tilde{\Phi}}{\phi}(x+a)\right] - 1.$$

Then, for the posterior median, using (**??**),

$$\hat{\mu}(x) = \max\{0, x - a - \Phi^{-1}(z_0)\}, \tag{2.31}$$

with $z_0 = a^{-1}\phi(x-a)[w^{-1} + \beta(x)]$. One can verify that as $x \to \infty$,

$$\beta(x) \sim \tfrac{1}{2}a/\phi(x-a), \qquad z_0 \sim \tfrac{1}{2}, \qquad \hat{\mu}(x) \sim x - a.$$

In particular, we see the bounded shrinkage property—for large x, the data is pulled down by about a. The threshold $t = t(w)$ and the weight $w = w(t)$ are related by

$$w(t)^{-1} = a(\Phi/\phi)(t-a) - \beta(t). \tag{2.32}$$

2.5 Mean squared error and linear estimators

We have described a large class of estimators that can be obtained using priors and regularization penalties and so it is natural to ask: how might we compare their properties? The simplest and most common approach is to study the mean squared error

$$r(\hat{\theta}, \theta) = E_\theta \|\hat{\theta} - \theta\|^2 = E_\theta \sum_{i=1}^{n} \left[\hat{\theta}_i(y) - \theta_i\right]^2.$$

Let us begin with the sequence model $y \sim N_n(\theta, \epsilon^2 I)$ and the class of *linear* estimators

$$\hat{\theta}_C(y) = Cy$$

for some $n \times n$ matrix C. The class of linear estimators includes smoothing splines, seen in Chapter 1, kernel estimators (Chapter 3) and other frequently used methods.

For any estimator $\hat{\theta}$, linear or not, the mean square error splits into variance and (squared) bias terms, yielding the *variance-bias decomposition*:

$$\begin{aligned} E\|\hat{\theta} - \theta\|^2 &= E\|\hat{\theta} - E\hat{\theta}\|^2 + \|E\hat{\theta} - \theta\|^2 \\ &= \text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta}). \end{aligned} \tag{2.33}$$

More specifically, since $\|\hat{\theta} - E\hat{\theta}\|^2 = \text{tr}(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T$, we have

$$\text{var}(\hat{\theta}) = \text{tr}[\text{Cov}(\hat{\theta})].$$

For linear estimators $\hat{\theta}_C$, clearly $\text{Cov}(Cy) = \epsilon^2 CC^T$ and so

$$\text{var}(\hat{\theta}_C) = \epsilon^2 \text{tr}\, CC^T = \epsilon^2 \text{tr}\, C^T C.$$

The bias $E\hat{\theta} - \theta = (C - I)\theta$, and hence the mean squared error becomes

$$r(\hat{\theta}_C, \theta) = \epsilon^2 \text{tr}\, C^T C + \|(I - C)\theta\|^2. \tag{2.34}$$

[Note that only second order distributional assumptions are used here, namely that $Ez = 0$ and $\mathrm{Cov}(z) = I$.]

The mean squared error is a quadratic function of θ, and the squared bias term is unbounded except when $C = I$. In this case $\hat{\theta}_I(y) = y$ is the *maximum likelihood estimator* (MLE) and is exactly unbiased for θ. The MSE of the MLE is constant,

$$r(\hat{\theta}_I, \theta) \equiv n\epsilon^2,$$

Thus, with linear estimators we already see the fundamental issue: there is no single estimator with uniformly best mean squared error, compare Figure 2.3.

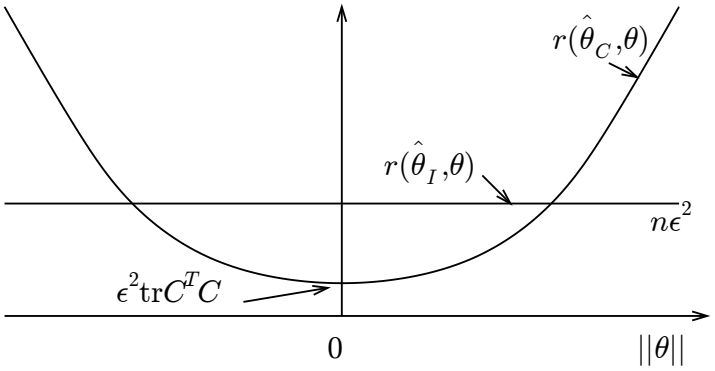

Figure 2.3 The mean squared error functions of linear estimators are quadratic, with smaller risk near 0, unless $C = I$, in which case the risk is constant. In particular, no single linear estimator is uniformly best for MSE. Plot assumes $C = cI$ for some c, so that $r(\hat{\theta}_C, \theta)$ is a function of $\|\theta\|$ only. A qualitatively similar picture holds for general C.

One way to exclude poor estimators is through the notion of *admissibility*. We say that estimator $\hat{\theta}$ is *inadmissible* if there exists another estimator $\hat{\theta}'$ such that $R(\hat{\theta}', \theta) \leq R(\hat{\theta}, \theta)$ for all θ, with strict inequality occurring for *some* θ. Such an estimator $\hat{\theta}'$ is said to *dominate* $\hat{\theta}$. And if no such dominating $\hat{\theta}'$ exists, then the original estimator $\hat{\theta}$ is called admissible. Admissibility itself is a rather weak notion of optimality–indeed, inadmissibility results are often of more interest than admissibility ones.

The most important (and surprising) fact about admissibilty is that the MLE $\hat{\theta}_I$ is **in**admissible exactly when $n \geq 3$. Indeed the positive part James-Stein estimator

$$\hat{\theta}^{JS+}(y) = \left(1 - \frac{(n-2)\epsilon^2}{\|y\|^2}\right)_+ y \tag{2.35}$$

dominates the MLE everywhere: $r(\hat{\theta}^{JS+}, \theta) < n\epsilon^2 = r(\hat{\theta}_I, \theta)$ for all $\theta \in \mathbb{R}^n, n \geq 3$. A short proof is given in the next section.

We can now describe a nice result on inadmissibility for linear estimators. We saw in §1.1 that smoothing splines shrink all frequencies except possibly for a low dimensional subspace on which no shrinkage occurs. In fact, all reasonable, i.e. admissible, linear estimators must behave in this way.

Theorem 2.3 *Suppose that $y \sim N_n(\theta, \epsilon^2 I)$. The linear estimator $\hat{\theta}_C(y) = Cy$ is admissible (for squared error loss) if and only if C*

(i) is symmetric,
(ii) has eigenvalues $0 \le \lambda_i(C) \le 1$, and
(iii) has at most two $\lambda_i(C) = 1$.

Proof We show only that each of these conditions is necessary for admissibility: if the condition fails we show how to construct a dominating estimator. (i) We use the notation $|A| = (A^T A)^{1/2}$ and the fact (Exercise 2.5) that $\mathrm{tr}A \le \mathrm{tr}|A|$, with equality only if A is symmetric, $A^T = A$.

Let D be defined via the identity $I - D = |I - C|$; clearly D is symmetric, and we use the variance-bias decomposition (2.33) to show that the MSE of $\hat{\theta}_D$ is everywhere better than that of $\hat{\theta}_C$ if C is not symmetric. Since

$$(I - D)^T (I - D) = |I - C|^2 = (I - C)^T (I - C),$$

the two estimators have the same (squared) bias. Turning to the variance terms, write

$$\mathrm{tr}\, D^T D = \mathrm{tr}\, I - 2\mathrm{tr}(I - D) + \mathrm{tr}(I - D)^T (I - D). \tag{2.36}$$

Comparing with the corresponding variance term for $\hat{\theta}_C$, we see that $\mathrm{tr}\, D^T D < \mathrm{tr}\, C^T C$ if and only if

$$\mathrm{tr}(I - D) = \mathrm{tr}|I - C| > \mathrm{tr}(I - C)$$

which occurs if and only if C fails to be symmetric.

(ii) As we may now assume that C is symmetric, we can find a decomposition $C = U \Lambda U^T$ with U orthogonal and $\Lambda = \mathrm{diag}(\lambda_i)$ containing the (real) eigenvalues of C. Now change variables to $\eta = U^T \theta$ and $x = U^T y \sim N(\eta, \epsilon^2 I)$. Since $E\|Cy - \theta\|^2 = E\|\Lambda - \eta\|^2$, we have

$$r(\hat{\theta}_C, \theta) = r(\hat{\eta}_\Lambda, \eta) = \sum_i \epsilon^2 \lambda_i^2 + (1 - \lambda_i)^2 \eta_i^2 = \sum_i r(\lambda_i, \eta_i).$$

Clearly, if any eigenvalue $\lambda_i \notin [0, 1]$, a strictly better MSE results by replacing λ_i by 1 if $\lambda_i > 1$ and by 0 if $\lambda_i < 0$.

(iii) Now suppose that $\lambda_1 = \ldots = \lambda_d = 1 > \lambda_i$ for $i > d \ge 3$, and let $x^d = (x_1, \ldots, x_d)$. We have noted that the positive part James-Stein estimator is everywhere better than $\hat{\eta}_I(x^d) = x^d$. So if we define a new estimator $\hat{\eta}$ to use $\hat{\eta}^{JS}$ on x^d and to continue to use $\lambda_i x_i$ for $i > d$, then

$$r(\hat{\eta}, \eta) = r(\hat{\eta}^{JS}, \eta^d) + \sum_{i>d} r(\lambda_i, \eta_i) < r(\Lambda, \eta),$$

and so $\hat{\eta}$ dominates $\hat{\eta}_\Lambda$ and hence $\hat{\theta}_C$.

For the converse, that conditions (i)-(iii) imply that $\hat{\theta}_C$ is admissible, see Cohen (1966). For the special case of the univariate MLE, see Remark 4.3 below. □

This still leaves a lot of linear estimators, to say nothing of the non-linear ones. To choose among the many admissible (and near admissible) rules, other criteria are needed. Thus, we

might also compare estimators by their maximum risk, seeking to find estimators whose maximum risk is as small as possible:

$$R_n = \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^n} E_\theta \|\hat{\theta} - \theta\|^2.$$

Here the infimum is taken over all estimators, linear or non-linear. We take up the systematic study of minimaxity in Chapter 4. For now, we mention the classical fact that the MLE $\hat{\theta}_I(y) = y$ is minimax:

$$R_n = n\epsilon^2 = \sup_{\theta \in \mathbb{R}^n} E_\theta \|y - \theta\|^2. \tag{2.37}$$

(This is proved, for example, using Corollary 4.9 and Proposition 4.15).

Remark 2.4 We digress briefly to record for later use some information about the smoothness of the risk functions of general estimators $\hat{\theta}$. For $y \sim N_n(\theta, \epsilon^2 I)$ and quadratic loss, the risk function $\theta \to r(\hat{\theta}, \theta)$ is analytic, i.e. has a convergent power series expansion, on the interior of the set on which it is finite. This follows, for example, from Lehmann and Romano (2005, Theorem 2.7.1), since $r(\hat{\theta}, \theta) = \int \|\hat{\theta}(y) - \theta\|^2 \phi_\epsilon(y - \theta) dy$ can be expressed in terms of Laplace transforms.

Mallows' C_L. There is a simple and useful unbiased estimate of the MSE of linear estimators $\hat{\theta}_C$. To derive it, observe that the residual $y - \hat{\theta}_C = (I - C)y$, and that the mean residual sum of squares (RSS) satisfies

$$E\|y - \hat{\theta}_C\|^2 = E\|(I - C)(\theta + \epsilon z)\|^2 = \epsilon^2 \mathrm{tr}\,(I - C)'(I - C) + \|(I - C)\theta\|^2. \tag{2.38}$$

Consequently the *C_L-statistic*, denoted here by U,

$$U(y) := \|y - \hat{\theta}_C\|^2 - n\epsilon^2 + 2\epsilon^2 \mathrm{tr}\, C$$

is found, by combining (2.38) and (2.34), to be an unbiased estimate of MSE:

$$E\{U\} = E\|\hat{\theta}_C - \theta\|^2.$$

Here is one application. If the matrix $C = C(\lambda)$ depends on a 'shrinkage' or 'bandwidth' parameter λ, and if ϵ^2 is known (or can be estimated), then one possibility is to choose λ to minimize the C_L estimate of MSE:

$$\begin{aligned} \hat{\lambda} &= \mathrm{argmin}_\lambda\, U_\lambda(y) \\ U_\lambda(y) &= \|y - C(\lambda)y\|^2 - n\epsilon^2 + 2\epsilon^2 \mathrm{tr}\, C(\lambda). \end{aligned} \tag{2.39}$$

When applied to orthogonal projections, the criterion is called Mallows' C_p.

2.6 The James-Stein estimator and Stein's Unbiased Risk Estimate

We have seen that Mallows' C_L provides an unbiased estimate of the risk of a linear rule $\hat{\theta}(y) = Cy$. In fact, there is a wide-ranging generalization: Stein (1981) gave a formula for an unbiased estimate of the mean squared error of a nearly arbitrary function of a multivariate Gaussian variate. Although the identity itself involves little more than integration by parts, it has proved powerful and influential.

Suppose that g is a nice function of a single variable $z \in \mathbb{R}$. Integration by parts and the rapid decay of the Gaussian density's tails show that

$$\int g(z)z\phi(z)dz = \int g(z)\big[-\frac{d}{dz}\phi(z)\big]dz = \int g'(z)\phi(z)dz.$$

If $Z \sim N_n(0, I)$ and $g : \mathbb{R}^n \to \mathbb{R}$, the formula becomes

$$E[Z_i g(Z)] = E[D_i g(Z)]. \tag{2.40}$$

Suppose now that g is vector valued, $g : \mathbb{R}^n \to \mathbb{R}^n$, that $X \sim N_n(\mu, I)$ and define the *divergence*

$$\nabla^T g = \sum_i D_i g_i = \sum_i \frac{\partial}{\partial x_i} g_i.$$

We may then rewrite the penultimate display as

$$E\,(X-\mu)^T g(X) = E\,\nabla^T g(X), \tag{2.41}$$

Regularity conditions do need attention here: some counterexamples are given below. It is, however, enough in (2.40) and (2.41) to assume that g is *weakly differentiable*: i.e. that g is absolutely continuous on all line segments parallel to the co-ordinate axes, and its partial derivatives (which consequently exist almost everywhere) are integrable on compact sets. Appendix C.15 gives the conventional definition of weak differentiability and the full proof of (2.41) and the following important consequence.

Proposition 2.5 *Suppose that $g : \mathbb{R}^n \to \mathbb{R}^n$ is weakly differentiable, that $X \sim N_n(\mu, I)$ and that for $i = 1, \ldots, n$, $E_\mu|X_i g_i(X)| + E|D_i g_i(X)| < \infty$. Then*

$$E_\mu\|X + g(X) - \mu\|^2 = E_\mu\{n + 2\nabla^T g(X) + \|g(X)\|^2\}. \tag{2.42}$$

Remarks. 1. The expression

$$U(x) = n + 2\nabla' g(x) + \|g(x)\|^2$$

is called *Stein's unbiased risk estimate* (SURE). In the particular case of a linear estimator $\hat{\mu}(x) = Cx$, it reduces to Mallows' C_L. Indeed $g(x) = (C-I)x$ and so $\nabla^T g(x) = \mathrm{tr}\, C - n$ and so

$$U(x) = -n + 2\mathrm{tr}\, C + \|(I-C)x\|^2.$$

2. Soft thresholding satisfies the weak differentiability condition. Indeed, writing $\hat{\mu}_S(x) = x + g_S(x)$, we see from (2.5) that

$$g_{S,i}(x) = \begin{cases} -\lambda & x_i > \lambda \\ -x_i & |x_i| \le \lambda \\ \lambda & x_i < -\lambda \end{cases} \tag{2.43}$$

is absolutely continuous as a function of each x_i, with derivative bounded by 1.

3. By contrast, hard thresholding has $\hat{\mu}_H(x) = x + g_H(x)$ with $g_{H,i}(x) = -x_i I\{|x_i| \le \lambda\}$, which is not even continuous, and so the unbiased risk formula cannot be applied.

4. Generalization to noise level ϵ and more generally to $Y \sim N_n(\theta, V)$ is straightforward (see Exercise 2.6).

The James-Stein estimate. For $X \sim N_n(\mu, I)$, the James-Stein estimator is defined by

$$\hat{\mu}^{JS}(x) = \left(1 - \frac{n-2}{\|x\|^2}\right)x, \tag{2.44}$$

and was used by James and Stein (1961) to give a more explicit demonstration of the inadmissibility of the maximum likelihood estimator $\hat{\mu}^{MLE}(x) = x$ in dimensions $n \geq 3$. [The MLE is known to be admissible for $n = 1, 2$, see e.g. Lehmann and Casella (1998, Ch. 5, Example 2.5 and Problem 4.5).] Later, Stein (1981) showed that the inadmissibility may be verified immediately from the unbiased risk formula (2.42). Indeed, if $n \geq 3$, $g(x) = -(n-2)\|x\|^{-2}x$ is weakly differentiable, and

$$D_i g(x) = -(n-2)\left(\frac{1}{\|x\|^2} - \frac{2x_i^2}{\|x\|^4}\right)$$

so that $\nabla^T g(x) = -(n-2)^2\|x\|^{-2}$ and so the unbiased risk estimator

$$U(x) = n - (n-2)^2\|x\|^{-2}.$$

Consequently

$$r(\hat{\mu}^{JS}, \mu) = n - (n-2)^2 E_\mu \|X\|^{-2}, \tag{2.45}$$

which is everywhere smaller than $r(\hat{\mu}^{MLE}, \mu) = E_\mu\|x - \mu\|^2 \equiv n$ *so long as* $n \geq 3$.

Remarks. 1. Where does the factor $n-2$ come from? A partial explanation: the estimator $\hat{\mu}(x) = (1 - \beta/\|x\|^2)x$ has unbiased risk estimate $U_\beta(x) = n - \{2\beta(n-2) - \beta^2\}/\|x\|^2$, and this quantity is minimized by the choice $\beta = n-2$. Note that $\beta = 2(n-2)$ has the same risk as the MLE. Also, we need $n \geq 3$ for finiteness of $E\|X\|^{-2}$, see (2.48) below.

2. The *positive part* James-Stein estimator

$$\hat{\mu}^{JS+}(x) = \left(1 - \frac{n-2}{\|x\|^2}\right)_+ x \tag{2.46}$$

has – necessarily – even better MSE than $\hat{\mu}^{JS}$ (Exercise 2.7).

The unbiased risk estimate leads to an informative bound on the mean squared error of the James-Stein rule.

Proposition 2.6 *If $X \sim N_n(\mu, I)$, then the James-Stein rule satisfies*

$$E_\mu\|\hat{\mu}^{JS} - \mu\|^2 \leq 2 + \frac{(n-2)\|\mu\|^2}{(n-2) + \|\mu\|^2}. \tag{2.47}$$

Proof For general μ, the sum of squares $\|X\|^2$ follows a non-central chisquared distribution with non-centrality parameter $\|\mu\|^2$. The non-central distribution may be realized as a mixture of central chi-squared distributions χ^2_{n+2N}, where N is a Poisson variate with mean $\|\mu\|^2/2$. (cf. e.g. Johnson and Kotz (1970, p. 132)). Recall also the formula

$$E\,[1/\chi^2_n] = 1/(n-2). \tag{2.48}$$

Hence, by conditioning first on N, and then using (2.48) and Jensen's inequality,

$$E[1/\chi^2_{n+2N}] = E[1/(n-2+2N)] \geq 1/(n-2+\|\mu\|^2).$$

Substituting into the unbiased risk formula (2.45), we obtain

$$r(\hat{\mu}^{JS}, \mu) \leq 2 + (n-2) - \frac{(n-2)^2}{n-2+\|\mu\|^2},$$

which yields the desired result after rearrangement. □

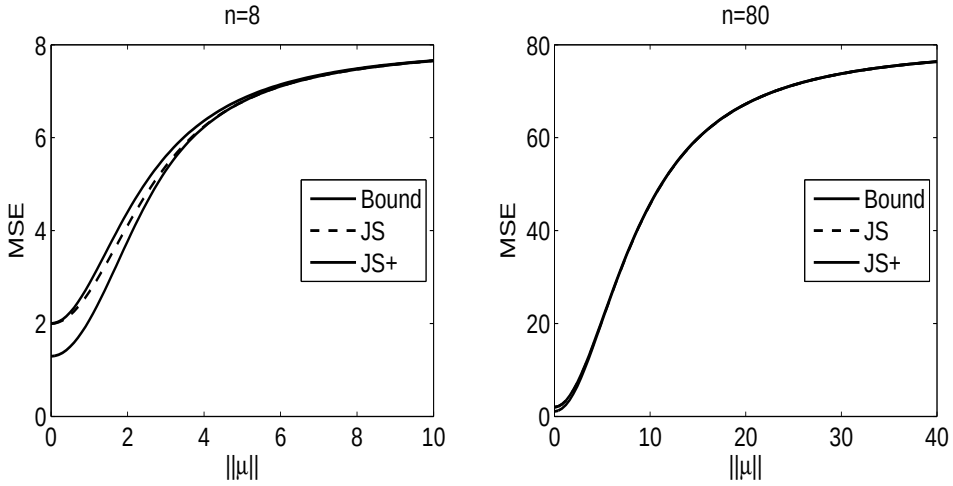

Figure 2.4 Exact risk functions of James-Stein rule $\hat{\mu}^{JS}$ (dashed) and positive part James-Stein $\hat{\mu}^{JS+}$ (solid) compared with upper bound from right side of (2.47). In the right panel ($n = 80$) the three curves are nearly indistinguishable.

Figure 2.4 illustrates several important aspects of the risk of the James-Stein estimator. First, the improvement offered by James-Stein relative to the MLE can be very large. For $\mu = 0$, we see from (2.45) and (2.48) that $r(\hat{\mu}^{JS}, 0) = 2$ while $r(\hat{\mu}^{MLE}, \mu) \equiv n$.

Second, the region of significant savings can be quite large as well. For $\|\mu\|^2 \leq \beta n$, the upper bound (2.47) is less than $(1+\beta n)/(1+\beta)$ so that, for example, if $\|\mu\|^2 \leq 4n$, then the savings is (roughly) at least 20 %.

Third, the improvement offered by the positive part estimator can be significant for both $\|\mu\|$ and n small, but otherwise the simple upper bound (2.47) gives a picture of the risk behavior that is accurate enough for most purposes.

Remarks. Exercise 2.9 provides details on the exact risk formulas for $\hat{\mu}^{JS+}$ used in Figure 2.4. It is known, e.g. Lehmann and Casella (1998, Example 5.7.3), that the positive part James-Stein rule cannot be admissible. While dominating estimators have been found, (Shao and Strawderman, 1994), the actual improvement over $\hat{\mu}^{JS+}$ seems not to be of practical importance.

Direct use of Jensen's inequality in (2.45) yields a bound inferior to (2.47), Exercise 2.8.

Corollary 2.7 *Let $\hat{\mu}_c(x) = cx$ be a linear shrinkage estimate. Then*

$$r(\hat{\mu}^{JS}, \mu) \leq 2 + \inf_c r(\hat{\mu}_c, \mu). \tag{2.49}$$

Proof The MSE of a linear shrinkage estimator $\hat{\mu}_c$ is

$$E\|cX - \mu\|^2 = c^2 n + (1-c)^2\|\mu\|^2. \tag{2.50}$$

In an idealized situation in which $\|\mu\|$ is known, the *ideal* shrinkage factor $c = c^{IS}(\mu)$ would be chosen to minimize this MSE, so that

$$c^{IS}(\mu) = \frac{\|\mu\|^2}{n + \|\mu\|^2}, \tag{2.51}$$

and

$$\inf_c r(\hat{\mu}_c, \mu) = \frac{n\|\mu\|^2}{n + \|\mu\|^2} \geq \frac{(n-2)\|\mu\|^2}{n - 2 + \|\mu\|^2}, \tag{2.52}$$

so that we need only refer to the preceding proposition. □

This is an example of an *oracle inequality*:

$$r(\hat{\mu}^{JS}, \mu) \leq 2 + r(\hat{\mu}^{IS}, \mu), \tag{2.53}$$

the risk of a bona fide estimator $\hat{\mu}^{JS}$ is bounded by the risk of the ideal estimator $\hat{\mu}^{IS}(x) = c^{IS}(\mu)x$, (unrealizable in practice, of course) plus an additive constant. In high dimensions, the constant 2 is small in comparison with the risk of the MLE, everywhere equal to n. On the other hand the bound (2.53) is sharp: at $\mu = 0$, the unbiased risk equality (2.45) shows that $r(\hat{\mu}^{JS}, 0) = 2$, while the ideal risk is zero.

The James-Stein estimator $\hat{\mu}^{JS}$ can be interpreted as an adaptive (quasi-) linear estimator. The ideal shrinkage constant $c^{IS}(\mu) = 1 - n/(n + \|\mu\|^2)$ and we can seek to estimate this using X. Indeed, $E\|X\|^2 = n + \|\mu\|^2$ and so $E\|X\|^{-2} \geq 1/(n + \|\mu\|^2)$, with approximate equality for large n. Consider therefore estimates of the form $\hat{c}(x) = 1 - \beta/\|x\|^2$ and note that we may determine β by observing that for $\mu = 0$, we have $E\hat{c} = 1 - \beta/(n-2) = 0$. Hence $\beta = n - 2$, and in this way, we recover precisely the James-Stein estimator.

For use in the next section, we record a version of (2.53) for arbitrary noise level.

Corollary 2.8 *Let $Y \sim N_n(\theta, \epsilon^2 I)$. The James-Stein estimate $\hat{\theta}^{JS+}(y)$ in* (2.35) *satisfies*

$$E\|\hat{\theta}^{JS+} - \theta\|^2 \leq 2\epsilon^2 + \frac{n\epsilon^2\|\theta\|^2}{n\epsilon^2 + \|\theta\|^2}.$$

2.7 Risk of soft thresholding

A brief study of the mean squared error properties of soft threshold estimators both illustrates some of the preceding ideas and allows for a first comparison of thresholding with James-Stein shrinkage. Chapter 8 has a more systematic discussion.

1°. Initially we adopt the unit noise setting, $X \sim N_n(\mu, I)$ and evaluate Stein's unbiased risk estimate for $\hat{\mu}_\lambda(x) = x + g_S(x)$, where the form of $g_S(x)$ for soft thresholding was given in (2.43). We have $(\partial g_{S,i}/\partial x_i)(x) = -I\{|x_i| \leq \lambda\}$ and so

$$\begin{aligned} E_\mu\|\hat{\mu}_\lambda(x) - \mu\|^2 &= E_\mu[U_\lambda(x)] \\ U_\lambda(x) &= n - 2\sum_1^n I\{|x_i| \leq \lambda\} + \sum_1^n x_i^2 \wedge \lambda^2. \end{aligned} \tag{2.54}$$

Since $U_\lambda(x)$ depends only on λ and the observed x, it is natural to consider minimizing $U_\lambda(x)$ over λ to get a threshold estimate $\hat{\lambda}_{SURE}$.

2°. Consider the one dimensional case with $X \sim N(\mu, 1)$. Let the (scalar) risk function $r_S(\lambda, \mu) = E_\mu[\hat{\mu}_\lambda(x) - \mu]^2$. By inserting the definition of soft thresholding and then changing variables to $z = x - \mu$, we obtain

$$r_S(\lambda, \mu) = \mu^2 \int_{-\lambda-\mu}^{\lambda-\mu} \phi(z)dz + \int_{\lambda-\mu}^{\infty} (z-\lambda)^2\phi(z)dz + \int_{-\infty}^{\lambda-\mu} (z+\lambda)^2\phi(z)dz.$$

Several useful properties follow from this formula. First, after some cancellation, one finds that

$$\frac{\partial}{\partial\mu} r_S(\lambda, \mu) = 2\mu\Phi([-\lambda-\mu, \lambda-\mu]) \le 2\mu, \tag{2.55}$$

which shows in particular that the risk function is monotone increasing for $\mu \ge 0$ (and of course is symmetric about $\mu = 0$).

Hence the risk increases from its value at $\mu = 0$,

$$r_S(\lambda, 0) = 2\int_{\lambda}^{\infty} (z-\lambda)^2\phi(z)dz \le e^{-\lambda^2/2}$$

(where the second inequality is Exercise 8.3) to its value at $\mu = \infty$,

$$r_S(\lambda, \infty) = 1 + \lambda^2,$$

(which follows, for example, by inspection of (2.54)). See Figure 2.5.

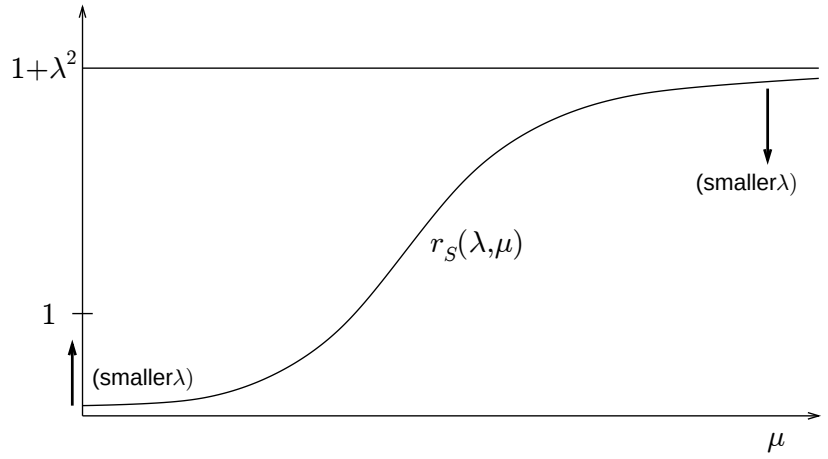

Figure 2.5 Qualitative behavior of risk function for soft thresholding. Arrows show how the risk function changes as the threshold λ is decreased.

3°. Some useful risk bounds are now easy consequences. Indeed, from (2.55) we have $r_S(\lambda, \mu) - r_S(\lambda, 0) \le \mu^2$. Using also the bound at ∞, we get

$$r(\lambda, \mu) \le r(\lambda, 0) + \min(1 + \lambda^2, \mu^2).$$

Making a particular choice of threshold, $\lambda_U = \sqrt{2\log n}$, and noting that $r(\lambda_U, 0) \le e^{-\lambda_U^2/2} = 1/n$, we arrive at

$$r(\lambda_U, \mu) \le (1/n) + (2\log n + 1)\min(\mu^2, 1).$$

Returning to noise level ϵ, and a vector observation $Y \sim N_n(\theta, \epsilon^2 I)$, and adding over the n coordinates, we can summarize our conclusions.

Lemma 2.9 *Let $Y \sim N_n(\theta, \epsilon^2 I)$ and $\hat{\theta}_\lambda$ denote soft thresholding with $\lambda = \epsilon\sqrt{2\log n}$. Then for all θ,*

$$E\|\hat{\theta}_\lambda - \theta\|^2 \le \epsilon^2 + (2\log n + 1)\sum_{i=1}^n \theta_i^2 \wedge \epsilon^2.$$

Comparison of James-Stein and thresholding. It is instructive to compare the bounds available for the mean squared error of James-Stein estimation and thresholding. Using the bound $\frac{1}{2}\min(a,b) \le ab/(a+b) \le \min(a,b)$, we find that the main term in the James-Stein bound Corollary 2.8 is

$$\frac{n\epsilon^2\|\theta\|^2}{n\epsilon^2 + \|\theta\|^2} \in [\tfrac{1}{2}, 1]\min(\sum \theta_i^2, n\epsilon^2).$$

For thresholding, looking at the main term in Lemma 2.9, we see that thresholding dominates (in terms of mean squared error) if

$$(2\log n)\sum_i \min(\theta_i^2, \epsilon^2) \ll \min\left(\sum \theta_i^2, n\epsilon^2\right).$$

For example, with $\epsilon = 1/\sqrt{n}$, and if θ is highly sparse, as for example in the case of a spike such as $\theta = (1, 0, \ldots, 0)$, then the left side equals $(2\log n)/n$ which is much smaller than the right side, namely 1.

Conversely, James-Stein dominates if all $|\theta_i|$ are nearly equal—recall, for example, the "comb" $\theta = \epsilon(1, \ldots, 1)$, where now the left side equals $(2\log n)\cdot n\epsilon^2$ which is now much *larger* than the right side, namely $n\epsilon^2 = 1$.

While thresholding has a smaller risk by a factor proportional to $\log n/n$ in our example, note that it can never be more than $O(\log n)$ worse than James-Stein, since $\sum \min(\theta_i^2, \epsilon^2) \le \min\left(\sum \theta_i^2, n\epsilon^2\right)$.

2.8 A Gaussian concentration inequality

A property of the multivariate normal model that finds frequent use in high dimensional estimation is the concentration of the distribution of Lipschitz functions. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be Lipschitz(L) if

$$|f(x) - f(y)| \le L\|x - y\|$$

for all $x, y \in \mathbb{R}^n$. Here $\|x\|$ is the usual Euclidean norm on $\mathbb{R}^n$. If f is differentiable, then we can take $L = \sup\|\nabla f(x)\|$.

Proposition 2.10 *If $Z \sim N_n(0, I)$, and $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz(L), then*

$$P\{f(Z) \geq Ef(Z) + t\} \leq e^{-t^2/(2L^2)}, \tag{2.56}$$

$$P\{f(Z) \geq Med f(Z) + t\} \leq \tfrac{1}{2} e^{-t^2/(2L^2)}. \tag{2.57}$$

Note that the dimension n plays a very weak role in the inequality, which is sometimes said to be "infinite-dimensional". The phrase "concentration of measure" refers at least in part to the fact that the distribution of a Lipschitz(1) function of n variables is concentrated about its mean, in the sense that the tails are no heavier than those of a *univariate* standard Gaussian, regardless of the value of n!

Some statistically relevant examples of Lipschitz functions include

(i) Order statistics. If $z_{(1)} \geq z_{(2)} \geq \cdots \geq z_{(n)}$ are the order statistics of a data vector z, then $f(z) = z_{(k)}$ has Lipschitz constant $L = 1$. The same is true for the absolute values $|z|_{(1)} \geq \cdots \geq |z|_{(n)}$. Section 8.10 has results on the maxima of Gaussian noise variates.

(ii) Ordered eigenvalues of symmetric matrices. Let A be an $n \times n$ symmetric matrix with eigenvalues $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$. If E is also symmetric, then (e.g. (Golub and Van Loan, 1996, p. 56 and 396))

$$|\lambda_k(A + E) - \lambda_k(A)| \leq \|E\|_F,$$

where $\|E\|_F^2 = \sum_{i,j} e_{i,j}^2$ denotes the square of the *Frobenius* norm, which is the Euclidean norm on $n \times n$ matrices. This is of statistical relevance, for example, if A is a sample covariance matrix, in which case $\lambda_1(A)$ is the largest principal component variance.

(iii) Orthogonal projections. If S is a linear subspace of $\mathbb{R}^n$, then $f(z) = \|P_S z\|$ has Lipschitz constant 1. If $\dim S = k$, then $\|P_S z\|^2 \stackrel{\mathcal{D}}{=} \chi^2_{(k)}$ and so

$$E\|P_S z\| \leq \{E\|P_S z\|^2\}^{1/2} = \sqrt{k}$$

and so the inequality implies

$$P\{\, \|P_S z\| \geq \sqrt{k} + t \,\} \leq e^{-t^2/2}. \tag{2.58}$$

These bounds play a key role in the oracle inequalities of Chapter 11.3.

(iv) Linear combinations of χ^2 variates. Suppose that $\alpha_i \geq 0$. Then $f(z) = (\sum \alpha_i z_i^2)^{1/2}$ is differentiable and Lipschitz: $\|\nabla f(z)\|^2 \leq \|\alpha\|_\infty$. Then a fairly direct consequence of (2.56) is the tail bound

$$P\{\sum \alpha_j (Z_j^2 - 1) > t\} \leq \exp\{-t^2/(32\|\alpha\|_1 \|\alpha\|_\infty)\} \tag{2.59}$$

for $0 < t \leq \|\alpha\|_1$ (Exercise 2.13). This is used for Pinsker's theorem in Chapter 5.4.

(v) Exponential sums. The function $f(z) = \log \sum_1^n \exp(\beta z_k)$ is Lipschitz(β). It appears, for example, in the study of Gaussian likelihood ratios of sparse signals, Section 13.5.

The two concentration inequalities of Proposition 2.10 have a number of proofs. We give an analytic argument for the first that builds on Stein's integration by parts identity (2.40). For the second, we shall only indicate how the result is reduced to the isoperimetric property of Gaussian measure—see e.g. Ledoux (2001) for a more complete discussion.

We begin with a lemma that bounds covariances in terms of derivatives.

Lemma 2.11 *Assume that $Y, Z \sim N_n(0, I)$ independently and set $Y_\theta = Y\cos\theta + Z\sin\theta$ for $0 \le \theta \le \pi/2$. Suppose that f and g are differentiable real valued functions on $\mathbb{R}^n$ of at most exponential growth. Then*

$$Cov\{f(Y), g(Y)\} = \int_0^{\pi/2} E[\nabla f(Y)^T \nabla g(Y_\theta)] \sin\theta d\theta. \tag{2.60}$$

Exponential growth means that $|f(y)| \le C \exp M|y|$ for some constants C and M.

An immediate corollary of (2.60) is the Gaussian Poincaré inequality:

$$\mathrm{Var} f(Y) \le E\|\nabla f(Y)\|^2.$$

Proof We may assume that $Eg(Y) = 0$, since replacing $g(y)$ by $g(y) - Eg(Y)$ changes neither side of the equation. Now, since Y and Z are independent, the covariance may be written $Ef(Y)[g(Y) - g(Z)]$. We exploit the path Y_θ from $Y_0 = Y$ to $Y_{\pi/2} = Z$, writing

$$g(Y) - g(Z) = -\int_0^{\pi/2} (d/d\theta) g(Y_\theta) d\theta.$$

We calculate $(d/d\theta)g(Y_\theta) = Z_\theta^T \nabla g(Y_\theta)$, where $Z_\theta = dY_\theta/d\theta = -Y\sin\theta + Z\cos\theta$. We arrive at

$$Ef(Y)[g(Y) - g(Z)] = -\int_0^{\pi/2} E[f(Y) Z_\theta^T \nabla g(Y_\theta)] d\theta. \tag{2.61}$$

The vectors Y_θ and Z_θ are independent and $N_n(0, I)$, being a rotation through angle θ of the original Y and Z, Lemma C.7. Inverting this rotation, we can write $Y = Y_\theta \cos\theta - Z_\theta \sin\theta$. Considering for now the ith term in the inner product in (2.61), we therefore have

$$\begin{aligned} E[f(Y) Z_{\theta,i} D_i g(Y_\theta)] &= E[f(Y_\theta\cos\theta - Z_\theta \sin\theta) Z_{\theta,i} D_i g(Y_\theta)] \\ &= -\sin\theta \cdot E[D_i f(Y) D_i g(Y_\theta)], \end{aligned}$$

where the second equality uses Stein's identity (2.40) applied to the $(n+i)$th component of the $2n$-dimensional spherical Gaussian vector (Y_θ, Z_θ). Adding over the n co-ordinates i and inserting into (2.61), we recover the claimed covariance formula. □

Proof of Concentration inequality (2.56)**.** This uses an exponential moment method. By rescaling and centering, we may assume that $L = 1$ and that $Ef(Y) = 0$. We will first show that for all $t > 0$,

$$E[e^{tf(Y)}] \le e^{t^2/2}. \tag{2.62}$$

Make the temporary additional assumption that f is differentiable and apply Lemma 2.11 to the functions f and $g = e^{tf}$. We have

$$E[\nabla f(Y)^T \nabla g(Y_\theta)] = tE[e^{tf(Y)} \nabla f(Y)^T \nabla f(Y_\theta)] \le tE e^{tf(Y)},$$

since differentiability and the Lipschitz bound on f entail $\|\nabla f\| \le 1$. Introduce the notation $e^{u(t)} = Ee^{tf(Y)}$, then differentiate with respect to t and use (2.60):

$$u'(t)e^{u(t)} = E[f(Y)e^{tf(Y)}] \le \int_0^{\pi/2} te^{u(t)} \sin\theta d\theta = te^{u(t)}.$$

Hence $u'(t) \le t$ for $t > 0$ and $u(0) = 0$, from which we get $u(t) \le t^2/2$ and so (2.62).

The assumption that f is differentiable can be removed by smoothing: the sequence $f_n = f \star \phi_{1/n}$ is Lipschitz(1) and converges to f a.e., so that (2.62) follows by Fatou's lemma.

Now we conclude by using Markov's inequality and (2.62). For each $t > 0$,

$$\begin{aligned} P(f(X) \geq u) &= P(e^{tf(X)} \geq e^{tu}) \\ &\leq e^{-tu} E e^{tf(X)} \leq e^{-tu+t^2/2}. \end{aligned}$$

The minimizing choice of t is $t = u$, and this yields our concentration inequality.

Remarks on (2.57)**.** Let P denote the probability measure corresponding to $Z \sim N_n(0, I)$. If A is a subset of $\mathbb{R}^n$ and $t > 0$, the dilation $A_t = \{z \in \mathbb{R}^n : d(z, A) < t\}$. The Gaussian isoperimetric inequality, e.g. Ledoux (2001, (2.9)), states that if A is a Borel set with $P(A) = \Phi(a)$ for some $a \in \mathbb{R}$, then $P(A_t) \geq \Phi(a + t)$ for every $t > 0$. [Thus the dilations of the half-plane $z_1 \leq 0$ have minimal Gaussian measure among all dilations of sets A of measure $\frac{1}{2}$].

In particular, if we take $A = \{z : f(z) \leq \mathrm{Med} f\}$, then $a = 0$ and if f is Lipschitz(1) , we have $A_t \subset \{z : f(z) \leq \mathrm{Med} f + t\}$. Consequently, using the isoperimetric inequality,

$$P(f(Z) > \mathrm{Med} f + t) \leq P(A_t^c) \leq \tilde{\Phi}(t) \leq \tfrac{1}{2} e^{-t^2/2},$$

where the final inequality is (2.68) in Exercise 2.11.

2.9 Some more general linear models

In this section we briefly describe some more general Gaussian models that can be reduced to sequence form, and review some approaches to regularization. As the emphasis is on sequence models, we do not discuss recent research areas such as the lasso or compressed sensing (see Chapter `Epilogue` for some references).

Some models that reduce to sequence form. A fairly general Gaussian linear model for estimation of means in correlated noise might be described in vector notation as $Y = A\beta + \sigma e$, or equivalently $Y \sim N(A\beta, \sigma^2 \Sigma)$. Some frequently occurring subclasses of this model can be reduced to one of the three sequence forms (2.1) - (2.3).

First, when $Y \sim N_n(\beta, \epsilon^2 I)$, one can take co-ordinates in *any* orthonormal basis $\{u_i\}$ for $\mathbb{R}^n$, yielding

$$y_i = \langle Y, u_i \rangle, \qquad \theta_i = \langle \beta, u_i \rangle, \qquad z_i = \langle z, u_i \rangle. \tag{2.63}$$

An essentially equivalent situation arises when $Y \sim N_n(A\beta, \sigma^2 I)$, and the matrix A is itself orthogonal: $A^T A = m I_n$. The columns of A might be orthogonal polynomials or other systems of functions, or orthogonal contrasts in the design of experiments, and so on. Specific examples include weighing designs, Hadamard and Fourier transforms (as in magnetic resonance imaging). The model can be put in the form (2.1) simply by premultiplying by $m^{-1} A^T$: define $y = m^{-1} A^T Y, z = m^{-1/2} A^T e$, and note especially the noise calibration $\epsilon = \sigma / \sqrt{m}$.

While this formulation appears parametric, formally it also covers the setting of nonparametric regression on a fixed equi-spaced design. Thus, the model

$$Y_l = f(l/n) + \sigma Z_l, \qquad l = 1, \ldots, n \tag{2.64}$$

with $Z_l \overset{i.i.d.}{\sim} N(0,1)$ becomes an example of (2.1) if one uses as design matrix an inverse discrete orthogonal wavelet (or Fourier) transform W^T to express $\mathbf{f} = (f(l/n)) = W^T\theta$. Thus here $A = W^T$. The components of y and θ are wavelet (or Fourier) coefficients of Y and $\mathbf{f}$ respectively.

If we drop the requirement (2.64) that the errors be normally distributed, keeping only the first and second moment requirements that Z have mean 0 and covariance I, then the same will be true of the transformed errors z. If the matrix W is in some sense 'dense', so that $z_i = \sum_k u_{ki} e_k$ has many non-zero terms of similar size, then by a central limit theorem for independent summands, the z_i will be approximately normally distributed.

Second, assume that $Y \sim N(A\beta, \epsilon^2 I)$, with A an $N \times M$ matrix. This can be converted into model (2.2) using the *singular value decomposition* $A = \sum_{i=1}^n \alpha_i u_i v_i^T$, where we assume that $\alpha_i > 0$ for $i = 1, \ldots, n = \text{rank}(A)$. We obtain

$$A\beta = \sum_i \alpha_i \theta_i u_i, \qquad \theta_i = \langle v_i, \beta \rangle,$$

so that $y_i = \langle Y, u_i \rangle = \langle A\beta, u_i \rangle + \epsilon \langle e, u_i \rangle = \alpha_i \theta_i + \epsilon z_i$ satisfies (2.2).

If one is specifically interested in the components of β, this transformation is not especially helpful. However, if the main focus is on the vector β, then the expansion $\beta = \sum \theta_i v_i$ may be useful, as can occur in the study of linear inverse problems, Chapter 3.

Interest in estimation of $\theta = A\beta$ can also arise in certain prediction problems. In the "in-sample" setting, one assesses a predictor $\hat{\theta} = A\hat{\beta}$ of a new observation vector $Y^* = A\beta + \sigma Z^*$ via the mean squared error $E\|A\hat{\beta} - Y^*\|^2 = E\|A(\hat{\beta} - \beta) - \sigma Z^*\|^2 = E\|\hat{\theta} - \theta\|^2 + N\sigma^2$.

Thirdly, assume that $Y \sim N(\beta, \epsilon^2 \Sigma)$, with covariance matrix Σ positive definite, with eigenvalues and eigenvectors

$$\Sigma u_i = \lambda_i^2 u_i, \qquad \lambda_i > 0,$$

so that with definitions (2.63), we recover the third sequence model (2.3), after noting that $\text{Cov}(y_i, y_j) = \epsilon^2 u_i \Sigma u_j = \epsilon^2 \lambda_i^2 \delta_{ij}$.

In the most general setting $Y \sim N(A\beta, \epsilon^2 \Sigma)$, however, a simple sequence version will typically only be possible if $A^T A$ and Σ have the same eigenvectors. This does occur, for example, if $A^T A$ and Σ are circulant matrices [2], and so are diagonalized by the discrete Fourier transform, (e.g. Gray (2006, Ch. 3)), or more generally if $A^T A$ and Σ commute.

Penalization and regularization. The least squares estimate of β is found by minimizing $\beta \to \|Y - A\beta\|_2^2$. If β is high dimensional, or if A has a smoothing character with many small singular values α_i, then the least squares solution for β is often ill-determined. See below for a simple example, and Section 3.8 for more in the setting of linear inverse problems.

A commonly used remedy is to *regularize* the solution by introducing a *penalty function* $P(\beta)$, and minimizing instead the penalized least squares criterion

$$Q(\beta) = \|Y - A\beta\|_2^2 + \lambda P(\beta).$$

Two simple and commonly occurring penalty functions are *quadratic*: $P(\beta) = \beta^T \Omega \beta$ for some non-negative definite matrix Ω, and q^{th} *power*: $P(\beta) = \|\beta\|_q^q = \sum_{i=1}^n |\beta_i|^q$. If P

[2] A matrix $\boldsymbol{C}$ is circulant if each row is obtained by cyclically shifting the previous row to the right by one; it is thus determined by its first row.

is strictly convex, or if P is convex and $A^TA > 0$, then Q is strictly convex and so the penalized criterion has at most one global minimum. Typically a minimum exists, and we denote it $\hat{\beta}(\lambda)$.

The *kernel* of the penalty, $\ker P = \{\beta : P(\beta) = 0\}$, typically consists of "very smooth" β. In our examples, if $\Omega > 0$ is positive definite, or if $q > 0$, then necessarily $\ker P = \{0\}$. More generally, if the penalty uses, say, squared second differences, then $P_2(\beta) = \sum_{i=2}^{n-1}(\beta_{i+1} - 2\beta_i + \beta_{i-1})^2$ and $\ker P_2 = \{\beta : \beta_k = c_0 + c_1 k,\ c_0, c_1 \in \mathbb{R}\}$ consists of linear functions.

The crucial *regularization parameter* λ determines the relative weight given to the sum of squared error and penalty terms: much more will be said about this later. As λ varies from 0 to $+\infty$, we may think of the penalized estimates $\hat{\beta}(\lambda)$ as forming a path from the roughest, least squares solution $\hat{\beta}(0) = \hat{\beta}_{LS}$ to the smoothest solution $\hat{\beta}(\infty)$ which necessarily belongs to $\ker P$.

We consider three especially important examples. First, the quadratic penalty $P(\beta) = \beta^T\Omega\beta$ is nice because it allows explicit solutions. The penalized criterion is itself quadratic:

$$Q(\beta) = \beta^T(A^TA + \lambda\Omega)\beta - 2Y^TA\beta + Y^TY.$$

Let us assume, for convenience, that at least one of A^TA and Ω is positive definite. In that case, $\partial^2 Q/\partial\beta^2 = 2(A^TA + \lambda\Omega)$ is positive definite and so there is a unique minimizer

$$\hat{\beta}(\lambda) = (A^TA + \lambda\Omega)^{-1}A^TY. \tag{2.65}$$

This is the classical *ridge regression* or *Tikhonov regularization* estimate, with *ridge* matrix Ω. For each λ, the estimate is a *linear* function $S(\lambda)Y$ of the data, with smoother matrix $S(\lambda) = (A^TA+\lambda\Omega)^{-1}A^T$. The trajectory $\lambda \to \hat{\beta}(\lambda)$ shrinks from the least squares solution $\hat{\beta}(0) = (A^TA)^{-1}A^TY$ down to $\hat{\beta}(\infty) = 0$.

Second, consider ℓ_1 penalties, which are used to promote sparsity in the solution. If the penalty is imposed after transformation to a sequence form such as (2.2) or (2.3), so that $P(\theta) = \sum|\theta_i|$, then the co-ordinatewise thresholding interpretation of Section 2.1 is available. When imposed in the original variables, so that $P(\beta) = \sum_1^n|\beta_i|$, the resulting estimator is known as the *lasso* – for **l**east **a**bsolute **s**election and **s**hrinkage **o**perator, introducted by Tibshirani (1996), see also Chen et al. (1998). There is no explicit solution, but the optimization problem is convex and many algorithms and a huge literature exists. See for example Büehlmann and van de Geer (2011) and Hastie et al. (2012).

Third, the ℓ_0 penalty $P(\beta) = \|\beta\|_0 = \#\{i : \beta_i \neq 0\}$ also promotes sparsity by penalizing the number of non-zero coefficients in the solution. As this penalty function is not convex, the solution is in general difficult to compute. However, in sufficiently sparse settings, the ℓ_0 and ℓ_1 solutions can coincide, and in certain practical settings, successful heuristics exist. (e.g. Donoho and Huo (2001), Candès and Romberg (2007)).

Example. Convolution furnishes a simple example of ill-posed inversion and the advantages of regularization. Suppose that $A = (a_{k-j}, 1 \le j,k \le n)$ so that $A\beta = a \star \beta$ represents convolution with the sequence (a_k). Figure 2.6 shows a simple example in which $a_0 = 1, a_{\pm1} = 1/2$ and all other $a_k = 0$. Although A is formally invertible, it is nearly singular, since for $\beta_{\text{osc}} = (+1,-1,+1,\ldots,\pm1)$, we have $A\beta_{\text{osc}} \doteq 0$, indeed the entries are exactly zero except at the boundaries. The instability of A^{-1} can be seen in the figure: the

left panel shows both $y = A\beta$ and $y' = A\beta + \sigma Z$ for a given signal β and a small added noise with $\sigma = .005$ and Z being a draw from $N_n(0, I)$. Although the observations y and y' are nearly identical, the least squares estimator $\hat{\beta}_{LS} = (A^T A)^{-1} A^T y = A^{-1} y$ is very different from $\hat{\beta}'_{LS} = A^{-1} y'$. Indeed A is poorly conditioned, its smallest singular value is $\alpha_n \doteq 0.01$, while the largest $\alpha_1 \doteq 2$.

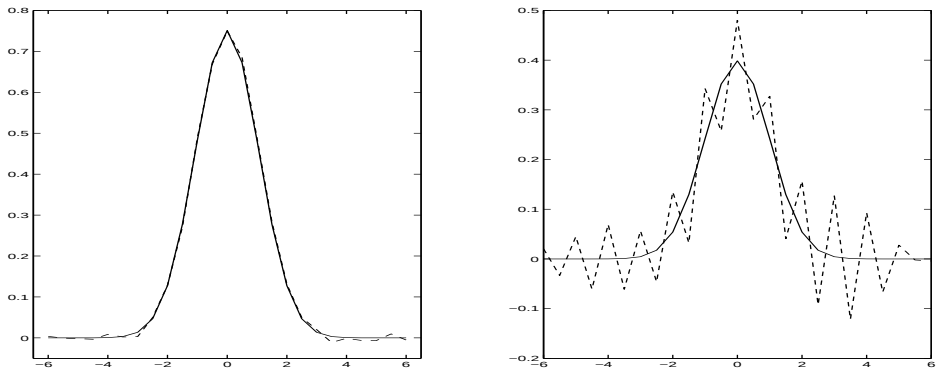

Figure 2.6 Left: Observed data $y = A\beta$, solid line, and $y' = A\beta + \sigma Z$, dashed line, for $\beta_l = \phi(t_l)$, the standard normal density with $t_l = (l/n) - 6$ and $n = 13, \sigma = 0.005$ and Z a draw from $N_n(0, I)$. Right: reconstructions $\hat{\beta}_{LS} = A^{-1} y$, dashed line, and regularized $\hat{\beta}(\lambda)$, solid line, from (2.65) with $\lambda = 0.01 = 2\epsilon = 2\sigma$.

Regularization with the squared second difference penalty P_2 removes the difficulty: with $\lambda = 0.01$, the reconstruction $\hat{\beta}(\lambda)$ from (2.65) is visually indistinguishable from the true β.

This may be understood in the sequence domain. If the banded matrices A and Ω are lightly modified in their bottom left and top right corners to be circulant matrices, then both are diagonalized by the (orthogonal) discrete Fourier transform, and in the Fourier coefficient domain, the effect of regularization is described by the co-ordinatewise formula (2.4). Indeed, substituting the frequency domain observation model $y_i = \alpha_i \theta_i + \epsilon z_i$, where here $\epsilon = \sigma$, we have

$$\hat{\theta}_i(y) = \frac{\alpha_i^2}{\alpha_i^2 + \lambda \omega_i} \theta_i + \frac{\epsilon \alpha_i}{\alpha_i^2 + \lambda \omega_i} z_i.$$

The sequence α_i decreases with increasing frequency i, while the regularizer constants ω_i increase. Thus at high frequencies, when $\lambda = 0$ the noise is amplified to $(\epsilon/\alpha_i) z_i$ (causing the jagged features in the figure), while when λ is positive ($= 2\epsilon$ in the figure), the term $\lambda \omega_i \gg \epsilon \alpha_i$ at high frequencies and the noise is successfully damped down.

2.10 Notes

Much of the material in this chapter is classical and can be found in sources such as Lehmann and Casella (1998).

§2. The connection between regularization with the ℓ_1 penalty and soft thresholding was exploited in Donoho et al. (1992), but is likely much older.

The soft thresholding estimator is also known as a "limited translation rule" by Efron (REF).

§3. Identity (2.18) is sometimes called Tweedie's formula (by Efron (2011) citing Robbins (1956)), and sometimes Brown's formula, for the extensive use made of it in Brown (1971).

§4. Priors built up from sparse mixture priors such as (2.25) are quite common in Bayesian variable selection problems. The connection with posterior median thresholding and most of the results of this section come from Johnstone and Silverman (2004a). Full details of the calculations for the Laplace and quasi-Cauchy examples may be found in Johnstone and Silverman (2005a, §6).

§5. Basic material on admissibility is covered in Lehmann and Casella (1998, Ch. 5). Inadmissibility of the MLE was established in the breakthrough paper of Stein (1956). The James-Stein estimator and positive part version were introduced in James and Stein (1961), for more discussion of the background and significance of this paper see Efron (1993). Theorem 2.3 on eigenvalues of linear estimators and admissibility is due to Cohen (1966). Mallows' C_L and its relative C_p are discussed in Mallows (1973). (CHECK)

§6. Stein (1981) presented the unbiased estimate of risk, Proposition 2.5 and, among much else, used it to give the quick proof of dominance of the MLE by the James Stein estimator presented here. The Stein identity characterizes the family of normal distributions: for example, if $n = 1$ and (2.41) holds for C^1 functions of compact support, then necessarily $X \sim N(\mu, 1)$, (Diaconis and Zabell, 1991).

Many other estimators dominating the MLE have been found–one classic paper is that of Strawderman (1971). There is a large literature on extensions of the James-Stein inadmissibility result to spherically symmetric distributions and beyond, one example is Evans-Stark (1996).

The upper bound for the risk of the James-Stein estimator, Proposition 2.6 and Corollary 2.7 are based on Donoho and Johnstone (1995).

We have also not discussed confidence sets – one entry point in to the literature is Hwang and Casella (1982) who show good properties for recentering the usual confidence set at the positive part James-Stein estimate.

§7. The unbiased risk estimate for soft thresholding was exploited in Donoho and Johnstone (1995), while Lemma 2.9 is from Donoho and Johnstone (1994a).

§8. The median version of the Gaussian concentration inequality (2.57) is due independently to Borell (1975) and Sudakov and Cirel′son (1974). The expectation version (2.56) is due to Cirel'son et al. (1976). Systematic accounts of the (not merely Gaussian) theory of concentration of measure are given by Ledoux (1996, 2001).

Our approach to the analytic proof of the concentration inequality is borrowed from Adler and Taylor (2007, Ch. 2.1), who in turn credit Chaumont and Yor (2003, Ch. 3.10), which has further references. The proof of Lemma 2.11 given here is lightly modified from Chatterjee (2009, Lemma 5.3) where it is used to prove central limit theorems by Stein's method. Tao (2011) gives a related but simpler proof of a weaker version of (2.56) with $\frac{1}{2}$ replaced by a smaller value C. An elegant approach via the semi-group of the Ornstein-Uhlenbeck process is described in Ledoux (1996, Ch. 2), this also involves an integration by parts formula.

Sharper bounds than (2.58) for the tail of χ^2 random variables are available (Laurent and Massart (1998), Johnstone (2001), Birgé and Massart (2001), `[CHECK!]`) . The constant 32 in bound (2.59) can also be improved to 8 by working directly with the chi-squared distribution.

Exercises

2.1 *(Gaussian priors.)* Suppose that $\theta \sim N_n(\theta_0, T)$ and that $y|\theta \sim N_n(\theta, I)$. Let $p(\theta, y)$ denote the joint density of (θ, y). Show that

$$-2 \log p(\theta, y) = \theta^T B \theta - 2\gamma^T \theta + r(y).$$

Identify B and γ, and conclude that $\theta|y \sim N(\theta_y, \Sigma_y)$ and evaluate θ_y and Σ_y.

2.2 Let F be an arbitrary probability distribution function on $\mathbb{R}$. A median of F is any point a_0 for which

$$F(-\infty, a_0] \geq \tfrac{1}{2} \qquad \text{and} \qquad F[a_0, \infty) \geq \tfrac{1}{2}.$$

Show (without calculus!) that

$$a \to M(a) = \int |a - \theta| dF(\theta)$$

is minimized at any median a_0.

2.3 *(Bounded shrinkage for the posterior median).* Establish Proposition 2.2, for example using the steps outlined below.

(a) Show using (2.19) that

$$\text{Odds}(\mu > c|X = x, \mu \neq 0) \geq \frac{\int_c^\infty e^{-\Lambda\mu}\phi(x-\mu)d\mu}{\int_{-\infty}^c e^{-\Lambda\mu}\phi(x-\mu)d\mu} \geq \frac{P(Z > -t-2)}{P(Z < -t-2)} \geq 3,$$

if $c = x - (\Lambda + t + 2)$ and Z is standard Gaussian.

(b) Show that

$$\text{Odds}(\mu \neq 0|X = x) \geq \frac{(g/\phi)(x)}{(g/\phi)(x-2)} \cdot \frac{1-w}{w}(g/\phi)(t) = \frac{(g/\phi)(x)}{(g/\phi)(x-2)}.$$

(c) Using (2.20), show that

$$\frac{(g/\phi)(x)}{(g/\phi)(x-2)} \geq \exp\int_{x-2}^x (t-\Lambda)dt \geq \exp(2t+2) \geq 2,$$

the last inequality holding if $x \geq t + \Lambda + 2$.

(d) Show that if $x \geq t + \Lambda + 2$, then $P(\mu \geq x - (t + \Lambda + 2)|X = x) \geq (3/4)(2/3) = 1/2$.

2.4 For the Laplace prior $\gamma_a(\mu) = \frac{1}{2}ae^{-a|\mu|}$, show that

$$g(x) = \tfrac{1}{2}a\exp(\tfrac{1}{2}a^2)\{e^{-ax}\Phi(x-a) + e^{ax}\tilde{\Phi}(x+a)\},$$
$$\tilde{\Gamma}(\mu|x) = \frac{e^{-ax}\tilde{\Phi}(\mu - x + a)}{e^{-ax}\Phi(x-a) + e^{ax}\tilde{\Phi}(x+a)}.$$

Use these expressions to verify the posterior median formula (2.31) and the threshold relation (2.32).

2.5 Let A be a square matrix and $|A| = (A^TA)^{1/2}$.

(i) Show how the polar decomposition $A = U|A|$, for suitable orthogonal U, can be constructed from the SVD of A.

(ii) Let (μ_i, e_i) be eigenvalues and eigenvectors of $|A|$. Show that $\text{tr}\, A \leq \text{tr}\,|A|$.

(iii) If equality holds in (ii), show that $Ae_i = |A|e_i$ for each i, and so that A must be symmetric.

2.6 (i) Suppose that $Y \sim N_d(\theta, V)$. For a linear estimator $\hat{\theta}_C(y) = Cy$, show that

$$r(\hat{\theta}_C, \theta) = \text{tr}\, C^TVC + \|(I-C)\theta\|^2.$$

(ii) If, in addition, $g : \mathbb{R}^n \to \mathbb{R}^n$ is smooth and satisfies $E\{|Y_i g_i(Y)| + |D_i g_j(Y)|\} < \infty$ for all i, j, show that

$$E_\theta\|Y + g(Y) - \theta\|^2 = E_\theta\{\text{tr}\, V + 2\text{tr}\,[VDg(Y)] + \|g(Y)\|^2\}. \tag{2.66}$$

2.7 Show that the positive part James-Stein estimator (2.46) has MSE no larger than the original James-Stein rule (2.44): $E\|\hat{\mu}^{JS+} - \mu\|^2 \leq E\|\hat{\mu}^{JS} - \mu\|^2$ for all $\mu \in \mathbb{R}^n$. .

2.8 Use Jensen's inequality in (2.45) to show that

$$r(\hat{\mu}_{JS}, \mu) \leq 4 + n\|\mu\|^2/(n + \|\mu\|^2).$$

2.9 *[Exact MSE for the positive part James-Stein estimator.]*
(i) Show that the unbiased risk estimator for $\hat{\mu}^{JS+}$ is

$$U(x) = \begin{cases} n - (n-2)^2 \|x\|^{-2}, & \|x\| > n-2 \\ \|x\|^2 - n, & \|x\| < n-2. \end{cases}$$

(ii) Let $F(t;k) = P(\chi^2_k \le t)$ and $\tilde{F}(t;k) = 1 - F(t;k)$. Show that for $t \ge 0$,

$$E[\chi^2_k, \chi^2_k \le t] = kF(t;k+2)$$
$$E[\chi^{-2}_k, \chi^2_k \le t] = (k-2)^{-1} F(t;k-2).$$

(iii) If $X \sim N_n(\mu, I)$, then let $K \sim \text{Poisson}(\|\mu\|^2/2)$ and $D = n + 2K$. Show that

$$r(\hat{\mu}^{JS}, \mu) = n - E_\mu (n-2)^2/(D-2)$$
$$r(\hat{\mu}^{JS+}, \mu) = n - E_\mu \Big\{ \frac{(n-2)^2}{D-2} \tilde{F}(n-2; D-2) + 2nF(n-2; D) - DF(n-2; D+2) \Big\}.$$

[which can be evaluated using routines for $F(t;k)$ available in many software packages.]

2.10 Suppose that $\epsilon = n^{-1/2}$ and $p < 2$. Compare the the large n behavior of the MSE of James-Stein estimation and soft thresholding at $\lambda = \epsilon\sqrt{2\log n}$ on the weak-ℓ_p-extremal sequences

$$\theta_k = k^{-1/p}, \qquad k = 1, \ldots, n.$$

2.11 *(Simple Gaussian tail bounds.)* (a) Let $\tilde{\Phi}(t) = \int_t^\infty \phi(s)ds$ and show that for $t > 0$,

$$\tilde{\Phi}(t) \le \phi(t)/t. \tag{2.67}$$

(b) By differentiating $e^{t^2/2}\tilde{\Phi}(t)$, show also that for $t > 0$,

$$\tilde{\Phi}(t) \le \tfrac{1}{2} e^{-t^2/2}. \tag{2.68}$$

2.12 *(Median and mean for maxima.)* If $Z \sim N_n(0, I)$ and M_n equals either $\max_i Z_i$ or $\max_i |Z_i|$, then use (2.56) to show that

$$|EM_n - \text{Med}M_n| \le \sqrt{2\log 2}. \tag{2.69}$$

(Massart (2007)).

2.13 *(Chi-squared tail bound.)* Use the inequality $(1+x)^{1/2} \ge 1 + x/4$ for $0 \le x \le 1$ to verify (2.59).

3

The infinite Gaussian sequence model

> It was agreed, that my endeavors should be directed to persons and characters supernatural, or at least romantic, yet so as to transfer from our inward nature a human interest and a semblance of truth sufficient to procure for these shadows of imagination that willing suspension of disbelief for the moment, which constitutes poetic faith. (Samuel Taylor Coleridge, *Biographia Literaria*, 1817)

For the first few sections, we focus on the infinite white Gaussian sequence model

$$y_i = \theta_i + \epsilon z_i \qquad\qquad i \in \mathbb{N}. \tag{3.1}$$

For some purposes and calculations this is an easy extension of the finite model of Chapter 2, while in other respects important new issues emerge. For example, the unbiased estimator $\hat{\theta}(y) = y$ has infinite mean squared error, and bounded parameter sets are no longer necessarily compact, with important consequences that we will see.

Right away, it must be remarked that we are apparently attempting to estimate an infinite number of parameters on the basis of what must necessarily be a finite amount of data. This calls for a certain suspension of disbelief which the theory attempts to reward.

Essential to the effort is some assumption that most of the θ_i are small in some sense. In this chapter we require θ to belong to an ellipsoid. In terms of functions expressed in a Fourier basis, this corresponds to mean-square smoothness. This and some consequences for mean squared error of linear estimators over ellipsoids are developed in Section **??**, along with a first rate of convergence result, for a truncation estimator that ignores all high frequency information.

We have seen already in the introductory Section 1.4 that (3.1) is equivalent to the continuous Gaussian white noise model. This connection, along with the heuristics also sketched there, allow us to think of this model as approximating the equispaced nonparametric regression model $Y_l = f(l/n) + \sigma Z_l$, compare (1.12). This opens the door to using (3.1) to gain insight into frequently used methods of nonparametric estimation. Thus, kernel and smoothing spline estimators are discussed in Sections 3.3 and 3.4 respectively, along with their bias and variance properties. In fact, a smoothing spline estimator is a kernel method in disguise and in the sequence model it is fairly easy to make this explicit, so Section 3.5 pauses for this detour.

Mean squared error properties return to the agenda in Section 3.6. The worst case MSE of a given smoothing spline over an ellipsoid (i.e. smoothness class) is calculated. This depends on the regularization parameter of the spline estimator, which one might choose to minimize

the worst case MSE. With this choice, standard rate of convergence results for smoothing splines can be derived.

The rest of the chapter argues that the splendid simplicity of the sequence model (3.1) actually extends to a variety of other settings. Two approaches are reviewed: transformation and approximation. The transformation approach looks at models that can be put into the independent Gaussian sequence form $y_i = \theta_i + \epsilon\lambda_i z_i$ for $i \in \mathbb{N}$ and known positive constants λ_i. This can be done for linear inverse problems with white Gaussian noise via the singular value decomposition, Section 3.8, and for processes with correlated Gaussian noise via the Karhunen-Loève transform (aka principal components), Section 3.9.

The approximation approach argues that with sufficient data, more concrete nonparametric function estimation problems such as density and spectral density estimation and flexible regression models "look like" the Gaussian sequence model. Methods and results can in principle, and sometimes in practice, be transferred from the simple white noise model to these more applications oriented settings. Section 3.10 gives a brief review of these results, in order to provide further motivation for our detailed study of the Gaussian sequence model in later chapters.

3.1 Parameter spaces and ellipsoids

We have seen in Chapter 1.4 that the Gaussian white noise model has continuous and discrete forms. The sequence form, (1.22), puts $y_i = \theta_i + \epsilon z_i$ for $i \in \mathbb{N}$ and $z_i \overset{i.i.d.}{\sim} N(0,1)$. The sample space is $\mathbb{R}^\infty$, with the Borel σ-field, and we denote the probability measure corresponding to $y = (y_i, i \in \mathbb{N})$ by P_θ. It follows from Kakutani's theorem, to be recalled in Section 3.7 below, that P_θ is either equivalent or orthogonal to the pure noise model P_0 in which $\theta = 0$. It is equivalent to P_0 if and only if $\theta \in \ell_2$. In that case, the likelihood ratio is given by

$$\frac{dP_\theta}{dP_0}(y) = \exp\left\{\frac{\langle y, \theta\rangle}{\epsilon^2} - \frac{\|\theta\|^2}{2\epsilon^2}\right\}.$$

The continuous form (1.18) puts $Y(t) = \int_0^t f(t)dt + \epsilon W(t), 0 \le t \le 1$. The sample space is taken to be $C[0,1]$ with the Borel σ-field, and we denote the probability measure corresponding to $\{Y(t), 0 \le t \le 1\}$ by P_f.

We will use squared error as the error measure, or loss function, in this chapter (except in Section 3.10). Thus $L\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_2^2 = \sum_{i\in\mathbb{N}}(\hat{\theta}_i - \theta_i)^2$ and the mean squared error

$$r(\hat{\theta}, \theta) = E_\theta L(\hat{\theta}(y), \theta) = E_\theta \|\hat{\theta}(y) - \theta\|_2^2.$$

This can be expressed in terms of functions and the continuous time domain using the Parseval relation (1.23) yielding $r(\hat{f}, f)$.

Suppose that θ is restricted to lie in a *parameter space* $\Theta \subset \ell_2$ and compare estimators through their worst case risk over Θ. Thus a particular importance attaches to the best possible worst-case risk, called the *minimax risk* over Θ:

$$R_N(\Theta) = R_N(\Theta, \epsilon) = \inf_{\hat{\theta}} \sup_{\theta\in\Theta} E_\theta L(\hat{\theta}(y), \theta). \tag{3.2}$$

The subscript "N" is a mnemonic for "non-linear" estimators, to emphasise that no restriction is placed on the class of estimators $\hat{\theta}$. One is often interested also in the minimax risk when the estimators are restricted to a particular class $\mathcal{E}$ defined by a property such as linearity. In such cases, we write $R_{\mathcal{E}}$ for the $\mathcal{E}$-minimax risk, under the assumption that the infimum in (3.2) is taken only over estimators in $\mathcal{E}$. Note also that we will often drop explicit reference to the noise level ϵ.

This is an extension of the notion of minimax risk over $\mathbb{R}^n$, introduced in Section 2.5. Indeed, in (3.2) we are *forced* to consider proper subsets Θ of $\ell_2(\mathbb{N})$. To see this, recall the classical minimax result quoted at (2.37), namely that $R_N(\mathbb{R}^n, \epsilon) = n\epsilon^2$. Since $\mathbb{R}^n \subset \ell_2(\mathbb{N})$ for each n, it is apparent that $R_N(\ell_2(\mathbb{N}), \epsilon) = \infty$, and in particular for *any* estimator $\hat{\theta}$

$$\sup_{\theta \in \ell_2(\mathbb{N})} E_\theta \|\hat{\theta} - \theta\|_2^2 = \infty. \tag{3.3}$$

Thus, a fundamental feature of non-parametric estimation is that some *a priori* restriction on the class of signals θ is required in order to make meaningful comparisons of estimators.

Fortunately, a great variety of such classes is available:

Lemma 3.1 *If Θ is compact in ℓ_2, then for ℓ_2 error, $R_N(\Theta, \epsilon) < \infty$.*

Proof Just consider the zero estimator $\hat{\theta}_0 \equiv 0$: then $\theta \to r(\hat{\theta}_0, \theta) = \|\theta\|_2^2$ is continuous on the compact Θ and so attains its maximum: $R_N(\Theta) \le \sup_\Theta r(\hat{\theta}_0, \theta) < \infty$. □

Two important classes of parameter spaces are the *ellipsoids* and *hyperrectangles*, defined respectively by

$$\Theta(a, C) = \{\theta : \sum_0^\infty a_k^2 \theta_k^2 \le C^2\}, \tag{3.4}$$

$$\Theta(\tau) = \{\theta : |\theta_k| \le \tau_k \text{ for all } k\} \tag{3.5}$$

We will see that each class can be used to encode different types of smoothness for functions $f \in L_2[0, 1]$. For now, we record criteria for compactness (the proofs are Exercise 3.1).

Lemma 3.2 *The ellipsoid $\Theta(a, C)$ is ℓ_2-compact if and only if $a_k > 0$ and $a_k \to \infty$.*
The hyperrectangle $\Theta(\tau)$ is ℓ_2-compact if and only if $\sum \tau_k^2 < \infty$.

In fact, Lemma 3.1 extends to sets of direct product form $\Theta = \mathbb{R}^r \times \Theta'$, where $r < \infty$ and Θ' is compact. The argument of the Lemma can also be extended to show that $R_N(\Theta, \epsilon) < \infty$ if $L(a, \theta) = w(\|a - \theta\|)$ with w continuous and Θ being $\|\cdot\|$-compact. At the same time, of course, compactness is not necessary for finiteness of the minimax risk, as (2.37) shows.

Ellipsoids and mean square smoothness. Consider the continuous form of the Gaussian white noise model (1.18). For integer $\alpha \ge 1$, let $f^{(\alpha)}$ denote the αth derivative of f and

$$\mathcal{F} = \mathcal{F}(\alpha, L) = \{f \in L_2[0, 1] : \int_0^1 [f^{(\alpha)}(t)]^2 dt \le L^2\}.$$

Historically, considerable interest focused on the behavior of the minimax estimation risk

$$R_N(\mathcal{F}, \epsilon) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} E \int_0^1 [\hat{f} - f]^2 \tag{3.6}$$

in the low noise limit as $\epsilon \to 0$. For example, what is the dependence on the parameters describing $\mathcal{F}$: namely (α, L)? Can one describe minimax estimators, and in turn, how do they depend on (α, L, ϵ)?

The parameter spaces $\mathcal{F}(\alpha, L)$ can be interpreted as ellipsoids in the orthonormal trigonometric basis for $L_2[0, 1]$. Let

$$\varphi_0(t) \equiv 1, \qquad \begin{cases} \varphi_{2k-1}(t) = \sqrt{2}\sin 2\pi k t & k = 1, 2, \ldots \\ \varphi_{2k}(t) \quad = \sqrt{2}\cos 2\pi k t. \end{cases} \tag{3.7}$$

Let $\Theta_2^\alpha(C)$ denote the ellipsoid (3.4) with semi-axes

$$a_0 = 0, \qquad a_{2k-1} = a_{2k} = (2k)^\alpha. \tag{3.8}$$

Lemma 3.3 *Suppose $\alpha \in \mathbb{N}$. Then $f \in \mathcal{F}(\alpha, C/\pi^\alpha)$ if and only if $\theta \in \Theta_2^\alpha(C)$.*

Proof Differentiation takes a simple form in the Fourier basis: for example for m even, $\varphi_{2k}^{(m)} = (2\pi k)^m \varphi_{2k}$, and collecting all such cases, we find that if $f = \sum \theta_k \varphi_k$, then

$$\int [f^{(m)}]^2 = \pi^{2m} \sum a_k^2 \theta_k^2,$$

from which the characterization follows immediately. □

The statistical importance of this result is that the functional minimax risk problem (3.6) is equivalent to a sequence space problem (3.2) under squared ℓ_2 loss. In the sequence space form, the parameter space is an ellipsoid. Its simple geometric form was exploited by Pinsker (1980) to give a complete solution to the description of minimax risk and estimators. We shall give Pinsker's solution in Chapter 5 as an illustration of tools that we will use for other parameter sets Θ in later chapters.

Remarks. 1. The ellipsoid representation (3.4)–(3.8) of mean-square smoothness extends to non-integer degrees of smoothness. Sometimes we put, more simply, just $a_k = k^\alpha$. Finiteness of $\sum k^{2\alpha}\theta_k^2$ can then be taken as a definition of finiteness of the Sobolev seminorm $\|f^{(\alpha)}\|_2$ even for non-integer α. Appendix B contains further details and references.

2. (3.8) shows that $\Theta(a, C)$ is actually not compact, for the trivial reason that $a_0 = 0$. However it does equal $\mathbb{R} \times \Theta'$ for Θ' compact.

3.2 Linear estimators and truncation

Linear estimators are simple and widely used, and so are a natural starting point for theoretical study. In practice they may take on various guises: kernel averages, local polynomial fits, spline smoothers, orthogonal series, Wiener filters and so forth. However in the sequence model, all such linear estimates can be written in the form $\hat{\theta}_C(y) = Cy$ for some matrix C, which when $I = \mathbb{N}$ has countably many rows and columns. It is therefore easy to extend the discussion of linear estimators in Section 2.5 to the infinite case. Thus the mean squared

error of $\hat{\theta}_C$ is still given by (2.38); one must only pay attention now to the convergence of infinite sums.

In particular, for $r(\hat{\theta}_C, \theta)$ to be finite, C needs to have finite Hilbert-Schmidt, or Frobenius, norm

$$\|C\|_{HS}^2 = \operatorname{tr} C^T C = \sum_{i,j=1}^{\infty} c_{ij}^2 < \infty. \tag{3.9}$$

Thus, C must be a bounded linear operator on ℓ_2 with square summable singular values. In particular, in the infinite sequence case, $C = I$ must be excluded, and so the bias term is necessarily unbounded over all of ℓ_2:

$$\sup_{\theta \in \ell_2} r(\hat{\theta}_C, \theta) = \infty, \tag{3.10}$$

as is expected anyway from the general result (3.3).

Familiar smoothing methods such as the Wiener filter and smoothing splines are linear shrinkers except possibly for a low dimensional subspace on which no shrinkage is done. Recall, for example, formula (1.15) for the smoothing spline estimator in the Demmler-Reinsch basis from Section 1.4, in which $w_1 = w_2 = 0$ and w_k increases for $k \geq 3$. This shrinks all co-ordinates but the first two.

In the infinite sequence model it is again true that linear estimators must shrink in all but at most two eigendirections. Indeed Theorem 2.3 extends to the infinite sequence model (3.1) in the most natural way: a linear estimator $\hat{\theta}_C(y) = Cy$ is admissible for squared error loss if and only if C is symmetric with finite Hilbert-Schmidt norm (3.9) and eigenvalues $\lambda_i(C) \in [0, 1]$ with at most two $\lambda_i(C) = 1$.

The proof for the inadmissibility part of Theorem 2.3 has the same structure in the infinite case, but the details of the first step are more intricate–for example, decomposition (2.36) is not directly useable). Mandelbaum (1984) gives the full argument.

Truncation estimators and Rates of Convergence

A particularly simple class of linear estimators is given by projection onto a subset of the co-ordinate axes: $(P_I y)_i = y_i$ if and only if $i \in I$. If the indices i correspond to frequency and the focus is on smoothing, it may be reasonable to restrict attention to nested sets of low frequencies $I_\nu = \{i : i \leq \nu\}$. We might call such a rule

$$\hat{\theta}_{\nu,i}(y) = \begin{cases} y_i & i \leq \nu, \\ 0 & i > \nu \end{cases}$$

a *truncation* estimator, as it discards frequencies above ν. *Caution*: a truncation estimator is quite different from a threshold estimator, e.g. (2.6)—the truncation estimator decides in advance, based on index i, and is linear, while the threshold estimator uses the data y_i and is *non*linear.

It is then natural to ask how to choose ν. One might try a minimax approach: suppose that a particular ellipsoid $\Theta(a, C)$ is given, and then find that value of ν which minimizes the maximum MSE over that ellipsoid. Using (2.34), we see that the MSE at a particular θ

arises from variance at low frequencies and from bias at high ones:

$$r(\hat{\theta}, \theta) = \sum_i E(\hat{\theta}_{\nu,i} - \theta_i)^2 = \nu\epsilon^2 + \sum_{i>\nu} \theta_i^2.$$

Only the bias term depends on θ, and for an ellipsoid (3.4) that bias is maximized by choosing θ_i to concentrate on the axis, or axes, of minimum squared half-width a_i^2. In particular, if $i \to a_i^2$ is increasing, the largest possible bias occurs at the lowest omitted frequency. Let e_ν denote a vector of zeros except for 1 in the ν^{th} slot. Then the maximizing $\theta_* = a_{\nu+1}^{-1} e_{\nu+1}$, so that

$$\bar{r}(\theta_\nu) := \sup_{\Theta(a,C)} r(\hat{\theta}_\nu, \theta) = \nu\epsilon^2 + C^2 a_{\nu+1}^{-2}.$$

Now specialize further to the mean-square smoothness classes in the trigonometric basis (3.7) in which the semi-axes a_i follow the polynomial growth (3.8). If we truncate at frequency k, then $\nu = 2k + 1$ (remember the constant term!) and

$$\bar{r}(\hat{\theta}_{2k}) = (2k+1)\epsilon^2 + C^2(2k+2)^{-2\alpha}.$$

As the cut-off frequency k increases, there is a trade-off of increasing variance with decreasing bias. The function is convex, and the optimal value is found by differentiation[1]:

$$2k_* + 2 = (2\alpha C^2/\epsilon^2)^{1/(2\alpha+1)}.$$

Substituting this choice into the previous display and introducing $r = 2\alpha/(2\alpha+1)$, we find

$$\begin{aligned}
\bar{r}_* &= \min_\nu \max_{\theta\in\Theta(a,C)} r(\hat{\theta}_\nu, \theta) \\
&= (2\alpha)^{1/(2\alpha+1)} C^{2(1-r)} \epsilon^{2r} - \epsilon^2 + C^2(2\alpha C^2/\epsilon^2)^{-r} \\
&\sim b_\alpha C^{2(1-r)} \epsilon^{2r},
\end{aligned}$$

as $\epsilon \to 0$, where the constant $b_\alpha = (2\alpha)^{1/(2\alpha+1)}(1 + 1/(2\alpha))$.

The calculation uncovers some important properties:

- the optimum cutoff frequency depends on the *signal to noise ratio* C/ϵ and the amount of smoothness α that is assumed—indeed k_* increases with C/ϵ and typically decreases with α.
- the 'rate of convergence' as $\epsilon \to 0$ is ϵ^{2r}. If one thinks of ϵ^2 as a proxy for inverse sample size $1/n$, then the rate becomes $r = 2\alpha/(2\alpha+1)$.
- the rate r increases with smoothness α: for twice differentiable functions $r = 4/5$, and r increases to 1 as $\alpha \nearrow \infty$.

3.3 Kernel Estimators

Kernel estimators form an important and widely used class in nonparametric regression and density estimation problems and beyond. We give a definition in the continuous Gaussian white noise model, discuss the connection with certain non-parametric regression settings,

[1] We ignore the fact that k_* should be an integer: as $\epsilon \to 0$, it turns out that using say $[k_*]$ would add a term of only $O(\epsilon^2)$, which will be seen to be negligible

and then begin to look at bias, variance and MSE properties. Finally, the sequence space form of a kernel estimator is derived in the Fourier basis.

A *kernel* $K(u)$ is a real valued, square integrable function with $\int K(u)du = 1$, *not necessarily* non-negative. The kernel is scaled to have *bandwidth* h

$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right).$$

Some common kernels include

$$K(t) = \begin{cases} (2\pi)^{-1/2}e^{-t^2/2} & \text{Gaussian} \\ (1/2)I_{[-1,1]}(t) & \text{Uniform} \\ (3/4)(1-t^2)I_{[-1,1]}(t) & \text{Quadratic/Epanechnikov} \\ (15/16)(1-t^2)^2 I_{[-1,1]}(t) & \text{Biweight.} \end{cases} \tag{3.11}$$

These are all symmetric and non-negative; all but the first also have compact support.

With observations $dY(t) = f(t) + \epsilon W(t), 0 \le t \le 1$, the kernel estimator of f is

$$\hat{f}_h(s) = \int_0^1 K_h(s-t)dY(t). \tag{3.12}$$

We interpret this using (1.20); for example we immediately have

$$E\,\hat{f}_h(s) = \int_0^1 K_h(s-t)f(t)dt, \qquad \mathrm{Var}\,\hat{f}_h(s) = \epsilon^2\int_0^1 K_h^2(s-t)dt. \tag{3.13}$$

From the first of formulas (3.13), one sees that $\hat{f}_h$ estimates a smoothed version of f given by convolution with the kernel of bandwidth h. The smaller the value of h, the more narrowly peaked is the kernel K_h and so the local average of f more closely approximates $f(s)$. One calls K_h an "approximate delta-function". Thus as h decreases so does the bias $E\,\hat{f}_h(s) - f(s)$, but inevitably at the same time the variance increases, at order $O(1/h)$.

From nonparametric regression to periodic kernels. To help in interpretation, we digress briefly to consider the nonparametric regression model $Y_l = f(t_l) + \sigma e_l$, for ordered t_l in $[0,1]$ and $l = 1,\dots,n$. A locally weighted average about s would estimate $f(s)$ via

$$\hat{f}(s) = \sum_l w_l(s)Y_l \Big/ \sum_l w_l(s). \tag{3.14}$$

A typical choice of weights might use a kernel $K(u)$ and set $w_l(s) = K_h(s-t_l)$. To pass heuristically from (3.14) to (3.12), we make two simplifying assumptions. First, suppose that the design points t_l are equally spaced, $t_l = l/n$, leading to the modified estimator

$$\tilde{f}_h(s) = n^{-1}\sum_{l=1}^n K_h(s-t_l)Y_l, \tag{3.15}$$

where the denominator in (3.14), namely $\sum_{l=1}^n K_h(s-l/n)$, has been approximated by an integral $n\int_0^1 K_h(s-u)du \doteq n$. Second, assume that f is *periodic* on $[0,1]$. Now we can imagine either extending the data periodically, $Y_l = Y_{l+jn}$, or equivalently, periodizing[2] the

[2] Note the difference between $\mathring{K}_h = (K_h)^\circ$, used here throughout, and $(\mathring{K})_h$.

kernel

$$\mathring{K}_h(t) = \sum_{m\in\mathbb{Z}} K_h(t-m). \tag{3.16}$$

and replacing K_h by $\mathring{K}_h$ in the definition of $\tilde{f}_h(s)$.

If f is periodic, $f(t) = f(t+m)$, we may write, using (3.16),

$$\int_0^1 \mathring{K}_h(s-t)f(t)dt = \int_{-\infty}^{\infty} K_h(s-u)f(u)du =: (K_h f)(u), \tag{3.17}$$

so that $K_h f$ refers to convolution over the whole line $\mathbb{R}$.

We would like to (mostly) ignore the periodization $K \to \mathring{K}$, and so we often assume

$$\text{supp}(K) \subset [-t_0, t_0], \quad \text{and} \quad h < 1/(2t_0), \tag{3.18}$$

so that

$$\text{if } t \in [-1/2, 1/2], \quad \text{then} \quad \mathring{K}_h(t) = K_h(t). \tag{3.19}$$

Let us summarize our conclusions regarding the mean and variance of a kernel estimator in the white noise model.

Lemma 3.4 *Given kernel K, let $\hat{f}_h = \mathring{K}_h \star Y = \int_0^1 \mathring{K}(\cdot - t)dY(t)$ denote convolution with the periodized kernel. We have*

$$E\hat{f}_h(s) = \int_{-\infty}^{\infty} K_h(s-t)f(t)dt = (K_h f)(s)$$

and, under assumption (3.18),

$$Var\hat{f}_h(s) = \epsilon^2 \int_{-\infty}^{\infty} K_h^2(t)dt = \epsilon^2 \|K_h\|_2^2.$$

Proof The first equality follows on combining (3.13) with (3.17). For the second, again start from (3.13) and observe that

$$\mathring{K}_h^2(u) = \Big[\sum_m K_h(u-m)\Big]^2 = \sum_m K_h^2(u-m)$$

where the second equality follows from (3.18), since for $m \neq m'$, the supports of $K_h(u-m)$ and $K_h(u-m')$ do not overlap. □

Global MSE. The global, or integrated, mean squared error of a kernel estimator also has a natural bias-variance decomposition. To describe it, we distinguish between L_2 norms $\|g\|_{2,I}$ on the observation interval $I = [0,1]$, and on all of $\mathbb{R}$, namely $\|g\|_2$. Then, under assumption (3.18),

$$E\|\hat{f}_h - f\|_{2,I}^2 = \frac{\epsilon^2}{h}\|K\|_2^2 + \|(I-K_h)f\|_{2,I}^2. \tag{3.20}$$

Notice the similarity of this mean squared error expression to (2.34) for a linear estimator in the sequence model. This is no surprise, given the sequence form of $\hat{f}_h$ to be described later in this section

Formula (3.20) is an immediate consequence of Lemma 3.4 and Fubini's theorem:

$$E\int_0^1 [\hat{f}_h(s) - f(s)]^2 = \int_0^1 \mathrm{Var}\,\hat{f}_h(s) + [E\,\hat{f}_h(s) - f(s)]^2 ds$$
$$= \epsilon^2 \|K_h\|_2^2 + \int_0^1 (K_h f - f)^2(s) ds.$$

Using also $\|K_h\|_2^2 = h^{-1}\|K\|_2^2$, we obtain (3.20). The result holds even without (3.18) if we replace K by $\mathring{K}$ on the right side.

$q-$th order kernels and bias reduction. A kernel is said to be of *$q-$th order* if it has vanishing moments of order 1 through $q-1$:

$$\mu_k = \int_{-\infty}^{\infty} v^k K(v) dv = \begin{cases} 1 & k = 0 \\ 0 & k = 1, \ldots, q-1 \\ q! c_q \neq 0 & k = q. \end{cases} \tag{3.21}$$

Observe that if K is symmetric about zero then necessarily $q \geq 2$. However, if K is symmetric and *non-negative*, then $c_2 > 0$ and so $q = 2$. We will see shortly that to obtain fast rates of convergence, kernels of order $q > 2$ are required. It follows that such kernels must necessarily have 'negative sidelobes'.

To see the bias reduction afforded by a $q-$th order kernel, assume that f has q continuous derivatives on $[0, 1]$. Then the Taylor series approximation to f at s takes the form

$$f(s - hv) = f(v) + \sum_{j=1}^{q-1} \frac{(-hv)^j}{j!} f^{(j)}(s) + \frac{(-hv)^q}{q!} f^{(q)}(s(v)),$$

for suitable $s(v)$ between $s - hv$ and s. The bias of $\hat{f}_h$ at s becomes

$$K_h f(s) - f(s) = \int K(v)[f(s - hv) - f(s)] dv = \frac{(-h)^q}{q!} \int v^q K(v) f^{(q)}(s(v)) dv \tag{3.22}$$

after using the vanishing moments (3.21).

As a result, the maximal bias of a $q-$th order kernel is uniformly $O(h^q)$:

$$\|K_h f - f\|_{\infty, I} = \sup_{0 \leq s \leq 1} |K_h f(s) - f(s)| \leq c h^q \|f^{(q)}\|_{\infty, I}.$$

Thus, other things being equal, (which they may not be, see Section 6.5), higher order kernels might seem preferable due to their bias reduction properties for smooth functions. [See Exercise 3.4 for an example of an infinite order kernel]. We will see this type of argument in studying the role of vanishing moments for wavelets in Chapter 7.

In summary, if K is a $q-$th order kernel, if (3.18) holds, and if f is C^q, then we have the local and global MSE expressions

$$E[\hat{f}_h(s) - f(s)]^2 = \frac{\epsilon^2}{h}\|K\|_2^2 + c_q^2 h^{2q} [D^q f(s)]^2 [1 + o(1)]$$
$$E\|\hat{f}_h - f\|_2^2 = \frac{\epsilon^2}{h}\|K\|_2^2 + c_q^2 h^{2q} \int (D^q f)^2 [1 + o(1)]. \tag{3.23}$$

Sequence space form of kernel estimators. Our kernel estimators are translation invariant, $K_h(s,t) = K_h(s-t)$, and so in the Fourier basis they should correspond to diagonal shrinkage. To describe this, let $\varphi_k(s)$ denote the trigonometric basis (3.7), and recall that the correspondence between the continuous model (1.18) and sequence form (3.1) is given by formulas (1.21) for y_k, θ_k etc.

Lemma 3.5 *Assume that the kernel $K(s)$ is symmetric. The sequence space form of the periodized kernel estimator $\hat{f}_h(s) = \int_0^1 \mathring{K}_h(s-t)dY(t)$ is given by*

$$\hat{\theta}_{h,k} = \widehat{K}(2\pi kh)y_k. \tag{3.24}$$

Thus the diagonal shrinkage constants in estimator $\hat{\theta}_h$ are given by the Fourier transform of kernel K, and their behavior for small bandwidths is determined by that of $\widehat{K}$ near zero[3]. Indeed, the $r-$th derivative of $\widehat{K}(\xi)$ at zero involves the $r-$th moment of K, namely $\widehat{K}^{(r)}(0) = (-i)^r \int t^r K(t)dt$. Hence an equivalent description of a $q-$th order kernel states that

$$\widehat{K}(\xi) = 1 - b_q\xi^q + o(\xi^q) \qquad \text{as} \quad \xi \to 0 \tag{3.25}$$

for some $b_q \neq 0$. Typically $b_q > 0$, reflecting the fact that the estimator usually shrinks coefficients toward zero.

For some of the kernels listed at (3.12), we have

$$\widehat{K}(\xi) = \begin{cases} e^{-\xi^2/2} & \text{Gaussian} \\ \sin\xi/\xi & \text{Uniform} \\ (3/\xi^2)(\sin\xi/\xi - \cos\xi) & \text{Quadratic/Epanechnikov.} \end{cases} \tag{3.26}$$

Proof We begin with the orthobasis of complex exponentials $\varphi_k^C(s) = e^{2\pi iks}$ for $k \in \mathbb{Z}$. The complex Fourier coefficients of the kernel estimator $\hat{f}_h$ are found by substituting (3.12) and interchanging orders of integration:

$$\int_0^1 \hat{f}_h(s)e^{-2\pi iks}ds = \int_0^1 \mathring{K}_h(s-t)e^{-2\pi ik(s-t)}ds \cdot \int_0^1 e^{-2\pi ikt}dY(t).$$

In other words, we have the diagonal form $\hat{\theta}_{h,k}^C = \gamma_{h,k}^C y_k^C$ for $k \in \mathbb{Z}$. Now using first the periodicity of $\mathring{K}_h$, and then its expression (3.16) in terms of K, we find that

$$\begin{aligned} \gamma_{h,k}^C &= \int_0^1 \mathring{K}_h(u)e^{-2\pi iku}du = \int_{-\infty}^{\infty} K_h(u)e^{-2\pi iku}du \\ &= \widehat{K_h}(2\pi k) = \widehat{K}(2\pi kh). \end{aligned} \tag{3.27}$$

Observe that since K is symmetric we have $\widehat{K}(-\xi) = \widehat{K}(\xi)$ and so $\gamma_{h,-k}^C = \gamma_{h,k}^C$.

It remains to convert this to the real trigonometric basis. The relation between Fourier coefficients $\{\hat{f}_k^C, k \in \mathbb{Z}\}$ in the complex exponential basis and real coefficients $\{f_k, k \geq 0\}$ in trigonometric basis (3.7) is given by

$$f_{2k} = (1/\sqrt{2})(f_k^C + f_{-k}^C), \qquad f_{2k-1} = (1/i\sqrt{2})(f_k^C - f_{-k}^C).$$

[3] The reader should note an unfortunate clash of established conventions: the hats in $\hat{\theta}, \hat{f}$ denoting estimators should not be confused with that in $\widehat{K}$ denoting Fourier transform!

The desired diagonal form (3.24) now follows from this and (3.27) since $\gamma^C_{h,-k} = \gamma^C_{h,k}$. □

3.4 Periodic spline estimators

Spline smoothing has become a popular technique in nonparametric regression, and serves as an important example of linear estimation in the Gaussian white noise model. As seen in Chapter 1.4, through the use of a particular orthonormal basis (Demmler-Reinsch) spline smoothing can be understood as a diagonal linear shrinkage method, even for unequally spaced regression designs. With an equally spaced design and Gaussian noise, the use of *periodic* splines allows a similar but more concrete analysis in the Gaussian sequence model. In particular, it is easy to derive an exact formula for the equivalent kernel in the large n, or small noise, limit. This discussion is a first illustration of how the Gaussian sequence model can provide concrete formulas for the "limiting objects" which strengthen understanding of similar finite sample settings. Much more information on spline theory, methods and applications may be found in the books by Wahba (1990), Hastie and Tibshirani (1990) and Green and Silverman (1994).

Suppose therefore that we observe

$$Y_l = f(l/n) + \epsilon Z_l, \qquad Z_l \overset{i.i.d.}{\sim} N(0,1)$$

for $l = 0, \ldots, n-1$. Since the observation points are equally spaced, we can use the Fourier basis (3.7). For convenience in notation, we consider only $n = 2m + 1$ odd. Let S now denote the linear space of trigonometric polynomials of degree m: $S_n = \{f : f(t) = \sum_{k=0}^{n-1} c_k \varphi_k(t), t \in [0,1]\}$.

The discrete sines and cosines will be $\boldsymbol{\varphi}_k = (\varphi_k(t_i))$, and the key point is that the double orthogonality relations (1.14) again hold, with now *explicit* weights

$$w_{2k-1} = w_{2k} = (2\pi k)^4. \tag{3.28}$$

We can now use the argument of Section 1.4. From the double orthogonality relations, for any function $f \in S_n$,

$$\begin{aligned} Q(f) &= n^{-1} \sum [Y_l - f(l/n)]^2 + \lambda \int f''^2 \\ &= \sum_{k=0}^{n-1} (y_k - \theta_k)^2 + \lambda \sum_{k=0}^{n-1} w_k \theta_k^2. \end{aligned}$$

So the minimizing periodic spline estimator has the form $\hat{\theta}_{PS,k} = c_{\lambda,k} y_k$ with an explicit formula for shrinkage at frequency k given by $c_{\lambda,0} = 1$ and

$$c_{\lambda,2k-1} = c_{\lambda,2k} = [1 + \lambda(2\pi k)^4]^{-1}$$

for $k \le m$. Thus the periodic spline problem has many of the qualitative features of general spline smoothing, along with a completely explicit description.

Remark. It is *not* true that the minimizer of $Q(f)$ over all functions lies in S, as was the case with cubic splines. The problem lies with *aliasing*: the fact that when $0 < r \le n$ and $l \in \mathbb{N}$, we have $\varphi_r = \varphi_{r+2ln}$ when restricted to $t_1, \ldots, t_n$.

	Finite model (3.31)	Infinite model (3.32)
Kernel	$\hat{f}_h = n^{-1}\sum_{i=1}^{n} K_h(t-t_i)Y_i$	$\hat{f}_h = \int_0^1 K_h(t-s)dY(s)$
		$\Updownarrow$
Spline	$\hat{f}_\lambda = \sum_0^{n-1} c_{\lambda k} y_k \varphi_k$	$\hat{f}_\lambda = \sum_0^{\infty} c_{\lambda k} y_k \varphi_k$

Table 3.1 *The analogy between spline smoothing and regression goes via versions of each method in the infinite sequence model.*

Similarly, the periodic spline estimate (1.16) in the finite model has a natural analogue in the infinite case. We define the smoothing spline estimate $\hat{\theta}_\lambda$ in the infinite sequence model as the minimizer of

$$\sum_1^\infty (y_k - \theta_k)^2 + \lambda \sum_1^\infty w_k \theta_k^2. \tag{3.29}$$

In general, the weights w_k should be positive and increasing. Just as in the finite case, the estimate $\hat{\theta}_\lambda$ has diagonal linear form,

$$\hat{\theta}_{\lambda,k}(y) = c_{\lambda k} y_k = (1+\lambda w_k)^{-1} y_k. \tag{3.30}$$

Note that a roughness penalty $P(f) = \int (D^m f)^2$ would correspond to weights $w_k = (2\pi k)^{2m}$, or simply to $w_k = k^{2m}$ if the constant factor is absorbed into λ.

We may interpret the m-th order smoothing spline as a Bayes estimator. Indeed, if the prior makes the co-ordinates θ_k independently $N(0, \tau_k^2)$ with $\tau_k^2 = bk^{-2m}$, then the posterior mean, according to (2.16), is linear with shrinkage factor

$$c_k = \frac{bk^{-2m}}{bk^{-2m} + \epsilon^2} = \frac{1}{1+\lambda k^{2m}},$$

after adopting the calibration $\lambda = \epsilon^2/b$. Section 3.9 interprets this prior in terms of $(m-1)$-fold integrated Brownian motion.

3.5 The Equivalent Kernel for Spline smoothing*.

Spline smoothing also has an interpretation in terms of local averaging which is not so apparent from its regularized least-squares formulation. This point of view comes out quite directly using sequence models. With this aim, we jump between the finite sequence model (2.64), namely

$$Y_i = f(i/n) + \sigma e_i, \qquad i = 1, \ldots, n \tag{3.31}$$

and the infinite sequence model (1.18) & (1.22), namely

$$\begin{aligned} Y_t &= \int_0^t f(s)ds + \epsilon W_t \qquad && t \in [0,1], \\ \Leftrightarrow \qquad y_k &= \theta_k + \epsilon z_k && k \in \mathbb{N} \end{aligned} \tag{3.32}$$

using the heuristics discussed around (1.24).

In terms of functions, the spline estimate is given by the series in the lower corner of Table 3.1.

We can now derive the kernel representation of the infinite sequence spline estimate. Substituting (1.21), $y_k = \int \varphi_k dY$ into $\hat{f}_\lambda = \sum_k c_{\lambda k} y_k \varphi_k$, we get

$$\hat{f}_\lambda(s) = \int C(s,t) dY(t), \qquad C(s,t) = \sum_0^\infty c_{\lambda k} \varphi_k(s) \varphi_k(t).$$

Now specialize to the explicit cubic weights for periodic splines in (3.28). Then $c_{\lambda,2k-1} = c_{\lambda,2k}$, and from (3.7) and the addition formula for sines and cosines,

$$\varphi_{2k-1}(s)\varphi_{2k-1}(t) + \varphi_{2k}(s)\varphi_{2k}(t) = 2\cos 2\pi k(s-t).$$

Hence the kernel $C(s,t)$ has translation form $K_\lambda(s-t)$, with formula

$$K_\lambda(s) = 1 + \sum_1^\infty \frac{2\cos 2\pi k s}{1 + \lambda(2\pi k)^4}.$$

But we can describe K_λ more explicitly! First, a definition: a function f on $\mathbb{R}$ can be made periodic with period 1 by *wrapping*: $g(t) = \sum_{j\in\mathbb{Z}} f(t+j)$.

Theorem 3.6 *The spline estimate with $\lambda = h^4$ has the kernel representation*

$$\hat{f}_\lambda(t) = \int_0^1 K_h(t-s) dY(s).$$

Here $K_h(t)$ is the wrapped version of $L_h(t) = (1/h)L(t/h)$. The equivalent kernel is given for $m = 1$ by $L(t) = (1/2)e^{-|t|}$ and for $m = 2$ by

$$L(t) = \frac{1}{2} e^{-|t|/\sqrt{2}} \sin\Big(\frac{|t|}{\sqrt{2}} + \frac{\pi}{4}\Big). \tag{3.33}$$

For general m, L is a $(2m)-$th order kernel, and is given by (3.36) *below.*

The kernel L_h has exponential decay, and is essentially negligible for $|t| \geq 8h$ for $m = 1, 2$ and for $|t| \geq 10h$ for $m = 3, 4$ compare Figure 3.1. The wrapped kernel K_h is therefore effectively identical with L_h on $[-\frac{1}{2}, \frac{1}{2}]$ when h is small: for example $h < 1/16$ or $h < 1/20$ respectively will do.

Thus in the infinite sequence model, periodic spline smoothing is identical with a particular kernel estimate. One may therefore interpret finite versions of periodic splines (and by analogy even B-spline estimates for unequally spaced data) as being approximately kernel smoothers. The approximation argument was made rigorous by Silverman (1984), who also showed that for unequally spaced designs, the bandwidth h varies with the fourth root of the design density.

Proof We may rewrite K_λ as

$$K_h(s) = \sum_{k\in\mathbb{Z}} \frac{e^{2\pi i k s}}{1 + (2\pi k h)^{2m}} = \sum_l L_h(s+l) \tag{3.34}$$

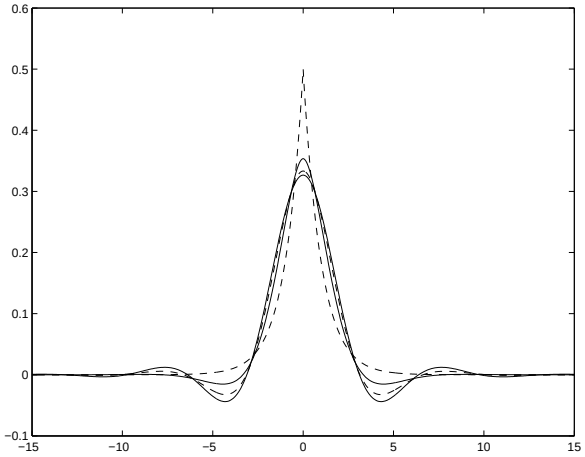

Figure 3.1 equivalent kernels for spline smoothing: dashed lines show $m = 1, 3$ and solid lines $m = 2, 4$. Only $m = 1$ is non-negative, the "side lobes" are more pronounced for increasing m.

where the second equality uses the Poisson summation formula (C.9) and implies that L_h has Fourier transform

$$\widehat{L_h}(\xi) = (1 + h^{2m}\xi^{2m})^{-1}. \tag{3.35}$$

We have $\widehat{L_h}(\xi) = \widehat{L}(h\xi)$—corresponding to the rescaling $L_h(t) = (1/h)L(t/h)$—and from Erdélyi et al. (1954, (Vol.??), p.10), with $r_k = (2k-1)\pi/(2m)$,

$$L(t) = (2m)^{-1} \sum_{k=1}^{m} e^{-|t| \sin r_k} \sin(|t| \cos r_k + r_k), \tag{3.36}$$

which reduces to the cited expressions for $m = 1$ and $m = 2$. □

Remark. Exercise 3.6 outlines a direct derivation via contour integration. Alternately, by successively differentiating (3.34), it is easily seen that

$$h^4 K_h^{(4)} + K_h = \sum_l \delta_l \tag{3.37}$$

where δ_l is the delta function at l. The solution of $h^4 L_h^{(4)} + L_h = \delta$ on $\mathbb{R}$ may be found by Fourier transformation, and yields the $m = 2$ case of (3.77), and then this is converted into a solution of (3.37) by periodization.

3.6 Spline Estimates over Sobolev Ellipsoids

So far we have said nothing about the mean squared error performance of the spline estimate, nor anything on the crucial question of how to choose the regularization parameter. These two issues are closely connected, and both depend on the smoothness of the function f being estimated. Our strategy here is to select convenient parameter spaces Θ, to evaluate the worst

case MSE of $\hat{\theta}_\lambda$ over Θ, and then to choose the value of λ that minimizes this maximum error. This yields information on the rate of convergence of $\hat{\theta}_\lambda$ to θ as $\epsilon \to 0$: we shall see that such rates of convergence, although crude tools, already yield useful information about estimators.

Maximum risk over ellipsoids. A general diagonal linear estimator with components $\hat{\theta}_k = c_k y_k$ has variance-bias decomposition

$$r(\hat{\theta}_c, \theta) = \epsilon^2 \sum_k c_k^2 + \sum_k (1 - c_k)^2 \theta_k^2.$$

The worst case risk over Θ has a corresponding form

$$\bar{r}(\hat{\theta}_c; \epsilon) = \sup_{\theta \in \Theta} r(\hat{\theta}_c, \theta) = \bar{V}(\epsilon) + \bar{B}^2(\Theta). \tag{3.38}$$

The max variance term $\bar{V}(\epsilon) = \epsilon^2 \sum_k c_k^2$ does not depend on Θ. On the other hand, the max bias term does not depend on the noise level ϵ. It does depend on Θ, but can be easily evaluated on ellipsoids.

Lemma 3.7 *Assume the homoscedastic white noise model $y_k = \theta_k + \epsilon z_k$. Let $\Theta = \Theta(a, C) = \{\theta : \sum a_k^2 \theta_k^2 \le C^2\}$ and $\hat{\theta}_c(y) = (c_k y_k)$. Then the maximum risk*

$$\bar{r}(\hat{\theta}_c; \epsilon) = \sup_{\theta \in \Theta} r(\hat{\theta}_c, \theta) = \epsilon^2 \sum_k c_k^2 + C^2 \sup_k a_k^{-2} (1 - c_k)^2.$$

Proof Make new variables $s_k = a_k^2 \theta_k^2 / C^2$ and note that the linear function $\sum d_k s_k$ is maximized over the non-negative simplex $\sum s_k \le 1$ by $\sup d_k$. Hence,

$$\bar{B}^2(\Theta) = C^2 \sup_k a_k^{-2} (1 - c_k)^2. \tag{3.39}$$

and the lemma follows from (3.38). □

The Variance-Bias Lemma. The next calculation occurs frequently enough that we record it here once and for all.

Lemma 3.8 (Variance-Bias) *The function $G(h) = vh^{-1} + bh^{2\beta}$, defined for $h \ge 0$ and positive constants v, b and β, has minimizing value and location*

$$G(h_*) = e^{H(r)} b^{1-r} v^r, \qquad h_* = r^{-1} e^{-H(r)} (v/b)^{1-r}.$$

The "rate" $r = 2\beta/(2\beta + 1)$, and $H(r) = -r \log r - (1-r)\log(1-r)$ is the binary entropy function.

For example, with kernel estimates based on a kernel K of order β, (3.23) shows that h can be thought of as a bandwidth and v as a variance factor (such as n^{-1} or ϵ^2), while b is a bias factor (for example involving $c(K, \beta) \int (D^\beta f)^2$.)

The proof is straightforward calculus, though the combination of the two terms in $G(h)$ to yield the multiplier $e^{H(r)}$ is instructive: the variance and bias terms contribute in the ratio 1 to $(2\beta)^{-1}$ at the optimum, so that in the typical case $\beta > \frac{1}{2}$, the bias contribution is the smaller of the two at the optimum h_*.

Aside on discretization approximations. Often a simpler expression results by replacing a sum by its (Riemann) integral approximation, or by replacing a minimization over non-negative integers by an optimization over a continuous variable in $[0, \infty)$. We use the special notation $\doteq$ to denote the approximate inequality in such cases. For example, the sum

$$S(\lambda) = \sum_{k=0}^{\infty} k^p (1 + \lambda k^q)^{-r} \doteq \kappa \lambda^{-\mu}, \qquad \mu = (p+1)/q, \tag{3.40}$$

with convergence if and only if $qr > p + 1$, and

$$\kappa = \kappa(p, r; q) = \int_0^{\infty} v^p (1 + v^q)^{-r} dv = \Gamma(r - \mu)\Gamma(\mu)/(q\Gamma(r)). \tag{3.41}$$

The approximation becomes an equality as $\lambda \to 0$, $S(\lambda)/\kappa\lambda^{-\mu} \to 1$.

For a minimization example, we observe that, if $0 < \alpha < \gamma$ and $\bar{\mu} = \alpha/\gamma$,

$$\bar{S}(\lambda) = \min_{k \in \mathbb{N}} \lambda k^{\alpha} + k^{\alpha - \gamma} \doteq \inf_{x > 0} \lambda x^{\alpha} + x^{\alpha - \gamma} = \bar{\kappa} \lambda^{1 - \bar{\mu}}, \tag{3.42}$$

with $\bar{\kappa} = e^{H(\alpha/\gamma)}$. The final equality uses, for example, the Variance-Bias lemma with $v = \lambda, h = x^{-\alpha}$, etc. Again we have asymptotic equality, $\bar{S}(\lambda)/\bar{\kappa}\lambda^{1-\bar{\mu}} \to 1$ as $\lambda \to 0$.

The errors in these discretization approximations[4] are quadratic in the size of the discretization step, and so can be expected often to be fairly small. Briefly, for the integral approximation, if G has, for example, $G(0) = 0$ and $\int_0^{\infty} |G''| < \infty$, then the difference between $\sum_{k=0}^{\infty} G(k\delta)\delta$ and $\int_0^{\infty} G(x)dx$ is $O(\delta^2)$, as follows from the standard error analysis for the trapezoid rule. Similarly, if G is C^2, then the difference between $\min_{k \in \mathbb{N}} G(k\delta)$ and $\inf_{x>0} G(x)$ is $O(\delta^2)$, as follows from the usual Taylor expansion bounds.

Spline estimators for fixed λ. We now specialize to shrinkage estimators

$$\hat{\theta}_{\lambda,k} = c_k y_k, \qquad c_k = (1 + \lambda k^{2m})^{-1}, \tag{3.43}$$

corresponding to roughness penalty $\int (D^m f)^2$, and to Sobolev ellipsoids

$$\Theta_2^{\alpha}(C) = \{\theta : \sum k^{2\alpha} \theta_k^2 \leq C^2\}. \tag{3.44}$$

Our plan is to use the preceding remarks about discretizations to obtain a simple formula for the worst case MSE for a spline estimator for given λ, and then in turn to optimize *that* to find the best (minimax) λ. The proofs are easy given the preparations we have made.

Proposition 3.9 *The worst case mean squared error for an m-th order spline estimate $\hat{\theta}_{\lambda}$ over a Sobolev ellipsoid $\Theta_2^{\alpha}(C)$ with $\alpha \leq 2m$ is*

$$\bar{r}(\hat{\theta}_{\lambda}; \epsilon) \doteq v_m \epsilon^2 \lambda^{-1/2m} + b_{\alpha m} C^2 \lambda^{2 \wedge (\alpha/m)}. \tag{3.45}$$

The constants v_m and $b_{\alpha m}$ appear in the proof. The worse case configuration is, approximately[5]*, given by $\theta^* = C k_*^{-2\alpha} e_{k_*}$, where*

$$k_* = \begin{cases} [(2m - \alpha)/\alpha]^{1/(2m)} \lambda^{-1/(2m)} & \text{if } \alpha \leq 2m, \\ 1 & \text{if } \alpha \geq 2m. \end{cases}$$

[4] Actually, we work in the reverse direction, from discrete to continuous!

[5] since k_* should be replaced by an integer, being whichever of $\lfloor k_* \rfloor$ or $\lceil k_* \rceil$ leads to the larger squared bias.

Remarks. 1. The exponent of λ in the bias term shows that high smoothness, namely $\alpha \geq 2m$, has no effect on the worst-case mean squared error.

2. The 'degrees of freedom' of the smoother $\hat{\theta}_\lambda = S_\lambda y$ is approximately

$$\operatorname{tr} S_\lambda = \sum_k c_k = \sum_k (1 + \lambda k^{2m})^{-1} \doteq c\lambda^{-1/(2m)}.$$

In the equivalent kernel of the Section 3.5, we saw that λ corresponded to h^{2m}, and so the degrees of freedom $\operatorname{tr} S_\lambda$ is approximately proportional to h^{-1}. In addition, if $\alpha \leq 2m$, $\operatorname{tr} S_\lambda$ is also proportional to the least favorable frequency k_*.

Proof For the variance term, use the integral approximation (3.40) with $q = 2m$:

$$\bar{V}(\epsilon) = \epsilon^2 \sum (1 + \lambda k^{2m})^{-2} \doteq v_m \epsilon^2 \lambda^{-1/2m},$$

where from (3.41) with $r = 2$ and $\mu = \mu_m = 1/(2m)$,

$$v_m = \mu_m \Gamma(2 - \mu_m)\Gamma(\mu_m) = (1 - \mu_m)/\operatorname{sinc}(\mu_m). \tag{3.46}$$

using Euler's reflection formula $\Gamma(1 - \mu)\Gamma(\mu) = \pi/\sin(\pi\mu)$, and the normalized sinc function $\operatorname{sinc}(x) = \sin(\pi x)/(\pi x)$. In the case $m = 2$ (cubic splines), $v_2 = 3\sqrt{2}\pi/16$.

For the squared bias term, note first that $1 - c_k = [1 + \lambda^{-1}k^{-2m}]^{-1}$, so that (3.39) becomes

$$\bar{B}^2 = C^2\lambda^2\{\inf_k \lambda k^\alpha + k^{\alpha-2m}\}^{-2}.$$

If $\alpha \geq 2m$, then $\bar{B}$ is maximized at $k_* = 1$, with $\bar{B} \sim C\lambda$, so that $b_{\alpha m} = 1$. If $\alpha \leq 2m$, then by differentiation, the minimum in $\bar{B}$ occurs at the claimed value of k_* and to evaluate the minimum value, apply (3.42) with $\gamma = 2m$ and $\bar{\mu} = \alpha/(2m)$ to obtain

$$\bar{B}^2 \doteq C^2\lambda^2(\bar{\kappa}\lambda^{1-\bar{\mu}})^{-2} = b_{\alpha m} C^2 \lambda^{\alpha/m},$$

with

$$b_{\alpha m} = e^{-2H(\alpha/2m)} = (2m)^{-2}\alpha^{\alpha/m}(2m - \alpha)^{2-\alpha/m}.$$

Note that $b_{\alpha m} = 1$ if $\alpha = 2m$. Combining the variance and bias terms yields (3.45). □

Minimax λ for the spline estimator. Our interest now turns to the value of λ that minimizes the maximum risk (3.45). This is called the *minimax* λ for the parameter space Θ.

Formula (3.45) shows that there is a variance-bias tradeoff, with small λ corresponding to small 'bandwidth' h and hence high variance and low bias, with the converse being true for large λ. To find the optimal λ, apply the Variance-Bias lemma with the substitutions

$$h = \lambda^{1/(2m)}, \qquad v = v_m\epsilon^2, \qquad b = b_{\alpha m}, \qquad \beta = 2m \wedge \alpha.$$

To summarize the results, define the *rate* $r(\alpha) = 2\alpha/(2\alpha + 1)$, and then set $r = r(\alpha \wedge 2m)$.

Theorem 3.10 *For periodic smoothing splines with weights $w_k = \lambda k^{2m}$, the minimax λ_* leads to*

$$\bar{r}(\hat{\theta}_\lambda; \epsilon) = \sup_{\Theta^\alpha(C)} r(\hat{\theta}_{\lambda_*}, \epsilon) \sim c_1(\alpha, m) C^{2(1-r)}\epsilon^{2r},$$

as $\epsilon \to 0$, with

$$\lambda_* \sim c_2(\alpha, m)(\epsilon^2/C^2)^{2m(1-r)}.$$

Remarks. 1. The rate of convergence $r = r(\alpha \wedge 2m)$ increases with α until $\alpha = 2m$, but does not improve further for functions with smoothness greater than α. We say that the rate *saturates* at $2m$, or that $r(2m)$ is a "speed limit" for m-th order splines.

2. In particular, for the typical choice $m = 2$, the rate of convergence saturates at speed limit $r(4) = 8/9$. If one uses a non-negative kernel, the 'generic' rate of convergence for a kernel estimator (at the optimal h) is $n^{-4/5} \asymp (\epsilon^2)^{4/5}$, at least for f with at least 2 continuous derivatives.

3. We will see in Chapter 5 that $r(\alpha)$ is the best possible rate of convergence, in the minimax sense, over $\Theta_2^\alpha(C)$. Thus m-th order splines can attain the optimal rate for all smoothness indices $\alpha \le 2m$.

An important points is that the optimal choice of λ_* needed to achieve this optimal rate depends on (C, α) (as well as m and ϵ^2). These values are unlikely to be known in practice, so the problem of *adaptation* consists, in this case, in finding estimators that achieve the optimal rate *without* having to specify values for C and α.

4. From the Variance-Bias lemma, one can identify the constants explicitly:

$$c_1(\alpha, m) = e^{H(r)} b_{\alpha m}^{1-r} v_m^r,$$
$$c_2(\alpha, m) = r^{-1} e^{-H(r)} (v_m / b_{\alpha m})^{1-r}.$$

5. If $\alpha = m$, then $b_{\alpha m} = 1/4$, which leads to the useful special case

$$\bar{r}(\hat{\theta}_{\lambda_*}; \epsilon) \sim e^{H(r)} (C^2/4)^{1-r} (v_m \epsilon^2)^r. \tag{3.47}$$

In particular, for cubic splines over ellipsoids of twice differentiable functions in mean square, we get that $\lambda_* \sim (v_2 \epsilon^2 / C^2)^{4/5}$. For a fixed function f, recall that $\int f''^2 = \pi^4 \sum a_k^2 \theta_k^2$. Thus, if f is known (as for example in simulation studies), and a reasonable value of λ is desired, one might set $C^2 = \pi^{-4} \int f''^2$ to arrive at the proposal

$$\lambda = \left(\frac{\pi}{2}\right)^4 \left(\frac{6\sqrt{2}\epsilon^2}{\int f''^2}\right)^{4/5}.$$

3.7 Non-white Gaussian sequence models

So far in this chapter, we have focused on the white infinite sequence model (3.1) and its cousins. Many of the methods of this book extend to a 'non-white' sequence model

$$y_i = \theta_i + \epsilon \lambda_i z_i, \qquad i \in \mathbb{N}. \tag{3.48}$$

where the z_i are again i.i.d. $N(0, 1)$, but the λ_i are known positive constants.

In the next two sections, we explore two large classes of Gaussian models which can be transformed into (3.48). These two classes parallel those discussed for the finite model in Section 2.9. The first, linear inverse problems, studies models of the form $Y = Af + \epsilon Z$, where Af is a linear operator, and the singular value decomposition (SVD) of A is needed to put the model into sequence form. The second, correlated data, considers models of the form $Y = f + \epsilon Z$, where Z is a correlated Gaussian process. In this setting, it is the Karhunen-Loève transform (KLT, also called principal component analysis) that puts matters into sequence form (3.48). The next two sections develop the SVD and KLT respectively, along with certain canonical examples that illustrate the range of possibilities for (λ_i).

When is model (3.48) well defined? We pause to recall the elegant Kakutani dichotomy for product measures (e.g. Williams (1991, Ch. 14), Durrett (2010, Ch. 5)). Let P and Q be probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$, absolutely continuous with respect to a probability measure λ. (For example, $\lambda = (P + Q)/2$.) Write $p = dP/d\lambda$ and $q = dQ/d\lambda$. The *Hellinger affinity*

$$\rho(P, Q) = \int \sqrt{pq}\, d\lambda \tag{3.49}$$

does not depend on the choice of λ. Now let $\{P_n\}$ and $\{Q_n\}$ be two sequences of probability measures on $\mathbb{R}$. Define product measures on sequence space $\mathbb{R}^\infty$, with the product Borel $\sigma-$field, by $P = \prod P_n$ and $Q = \prod Q_n$. The affinity behaves well for products: $\rho(P, Q) = \prod \rho(P_i, Q_i)$.

Kakutani's dichotomy says that if the components $P_n \sim Q_n$ for $n = 1, 2, \ldots$ then the products P and Q are either equivalent or orthogonal. And there is an explicit criterion:

$$P \sim Q \qquad \text{if and only if} \qquad \prod_{k=1}^{\infty} \rho(P_k, Q_k) > 0.$$

And when $P \sim Q$, the likelihood ratio dP/dQ is given by the product $\prod_{k=1}^{\infty} dP_k/dQ_k$.

The criterion is easy to apply for Gaussian sequence measures. A little calculation shows that the univariate affinity

$$\rho(N(\theta, \sigma^2), N(\theta', \sigma^2)) = \exp\{-(\theta - \theta')^2/(8\sigma^2)\}.$$

Let P_θ denote the product measure corresponding to (3.48). The dichotomy says that for two different mean vectors θ and θ', the measures P_θ and $P_{\theta'}$ are equivalent or orthogonal. [See Exercise ?? for an implication for statistical classification]. The product affinity

$$\rho(P_\theta, P_{\theta'}) = \exp\{-D^2/(8\epsilon^2)\}, \qquad D^2 = \sum_i (\theta_i - \theta_i')^2/\lambda_i^2. \tag{3.50}$$

Thus P_θ is absolutely continuous relative to P_0 if and only if $\sum \theta_i^2/\lambda_i^2 < \infty$, in which case the density is given in terms of the inner product $\langle \theta, x\rangle_\lambda = \sum \theta_i x_i/\lambda_i^2$ by

$$\frac{dP_\theta}{dP_0} = \exp\{\langle \theta, x\rangle_\lambda/\epsilon^2 - \|\theta\|_\lambda^2/(2\epsilon^2)\}.$$

Here θ_i/λ_i might be interpreted as the signal-to-noise ratio of the i-th co-ordinate.

We will again be interested in evaluating the quality of estimation of θ that is possible in model (3.48). An important question raised by the extended sequence model is the effect of the constants (λ_i) on quality of estimation—if λ_i increases with i, we might expect, for example, a decreased rate of convergence as $\epsilon \to 0$.

We will also be interested in the comparison of linear and non-linear estimators in model (3.48). For now, let us record the natural extension of formula (2.34) for the mean squared error of a linear estimator $\hat{\theta}_C(y) = Cy$. Let $\Lambda = \mathrm{diag}(\lambda_i)$, then

$$r(\hat{\theta}_C, \theta) = \epsilon^2 \mathrm{tr}\, C^T \Lambda C + \|(C - I)\theta\|^2. \tag{3.51}$$

Hellinger and L_1 distances. We conclude this section by recording some facts about

distances between (Gaussian) measures for use in Section 3.10 on asymptotic equivalence. A more systematic discussion may be found in Lehmann and Romano (2005, Ch. 13.1).

Let P_0 and P_1 be probability measures on $(\mathcal{X}, \mathcal{B})$ and ν a dominating measure, such as $P_0 + P_1$. Let p_0 and p_1 be the corresponding densities. The *Hellinger* distance $H(P_0, P_1)$ and L_1 or *total variation* distance between P_0 and P_1 are respectively given by

$$H^2(P_0, P_1) = \tfrac{1}{2}\int (\sqrt{p_0} - \sqrt{p_1})^2 d\nu,$$
$$\|P_0 - P_1\|_1 = \int |p_0 - p_1| d\nu.$$

Neither definition depends on the choice of ν. Expanding the square in the Hellinger distance, we have $H^2(P_0, P_1) = 1 - \rho(P_0, P_1)$, where ρ is the affinity (3.50). The Hellinger distance is statistically useful because the affinity behaves well for products (i.e. independence), as we have seen. The L_1 distance has a statistical interpretation in terms of the sum of errors of the likelihood ratio test between P_0 and P_1:

$$1 - \tfrac{1}{2}\|P_0 - P_1\|_1 = P_0(p_0 \le p_1) + P_1(p_1 < p_0).$$

The measures are related (Lehmann and Romano, 2005, Th. 13.1.2) by

$$H^2(P_0, P_1) \le \tfrac{1}{2}\|P_0 - P_1\|_1 \le [1 - \rho^2(P_0, P_1)]^{1/2}. \tag{3.52}$$

It is instructive to compute these distances when P_0 and $P_1 = P_\theta$ are Gaussian measures with means 0 and θ, and with common variances. Then $\rho(P_\theta, P_0)$ is given by (3.50) with $\theta' = 0$. To calculate the L_1 distance, observe that the likelihood ratio

$$p_1/p_0 = \exp(W - D^2/2), \qquad W = \sum_i x_i \theta_i / \sigma_i^2.$$

Under P_0 and P_θ respectively, $W \sim N(0, D^2)$ and $W \sim N(D^2, D^2)$ and we find

$$\|P_\theta - P_0\|_1 = 2[1 - 2\tilde{\Phi}(D/2)]. \tag{3.53}$$

We can now compare the quantities in (3.52) assuming that D is small. Indeed

$$H^2(P_\theta, P_0) \approx D^2/8, \quad \tfrac{1}{2}\|P_\theta - P_0\|_1 \approx 2\phi(0)D, \quad \text{and} \quad [1 - \rho^2(P_\theta, P_0)]^{1/2} \approx D.$$

In the continuous Gaussian *white* noise model ($\lambda_i \equiv 1$), we can re-interpret D^2 using Parseval's identity, so that $\|P_f - P_{\bar{f}}\|$ is given by (3.53) with

$$D = D_n(f) = \int_0^1 (f - \bar{f})^2. \tag{3.54}$$

3.8 Linear inverse problems

The continuous signal in Gaussian noise model led to a homoscedastic version of the basic sequence model (1.11). The more general form with unequal variances can arise when we do not observe f—ignoring the noise for now—but rather its image Af after the action of an operator A, representing some form of integration, smoothing or blurring. The recovery of f from the indirect observations Af is called an *inverse problem* and has a rich literature which

we barely touch. We consider only linear operators A and settings which lend themselves to expression in sequence model form.

We begin with an idealized extension of the continuous white noise model (1.18) and then pass to examples. Suppose, then, that the unknown function f is defined and square integrable on some domain $D \subset \mathbb{R}^d$.

The linear operator A is assumed to be bounded as a transformation from $\mathcal{H} = L_2(D, \mu_1)$ to $\mathcal{K} = L_2(U, \mu_2)$. Let the inner products on $\mathcal{H}$ and $\mathcal{K}$ be denoted $\langle \cdot, \cdot \rangle$ and $[\cdot, \cdot]$ respectively. The observations are given by

$$Y = Af + \epsilon Z, \tag{3.55}$$

a process on U, interpreted to mean that for any $\psi \in L_2(U)$, the observation is a functional

$$Y(\psi) = [Af, \psi] + \epsilon Z(\psi), \tag{3.56}$$

and $Z = \{Z(\psi)\}$ is a Gaussian process with mean zero and covariance function

$$\mathrm{Cov}(Z(\psi), Z(\psi')) = \int_U \psi \psi' \, d\mu_2.$$

The setting of *direct* estimation, in which $A = I$, is a special case in which $\mathcal{H} = \mathcal{K} = L_2[0, 1]$. With $\psi = I_{[0,t]}$, we write $Y(t)$ for $Y(\psi)$ and recover the signal in continuous white Gaussian noise model (1.18).

To arrive at the sequence form of (3.55), we employ the singular value decomposition (SVD) of the operator A. For definitions and details, see Appendix **??** and the references given there. Suppose that $A : \mathcal{H} \to \mathcal{K}$ is a compact linear operator between Hilbert spaces, with null space $N(A) = \{f \in \mathcal{H} : Af = 0\}$. The *singular value decomposition* of A consists of two sets of singular functions

(i) $\{\varphi_k\}$, an orthonormal set in $\mathcal{H}$ whose closed linear span equals the orthogonal complement of $N(A)$,
(ii) $\{\psi_k\}$, an orthonormal set in $\mathcal{K}$, and
(iii) singular values $b_k > 0$, such that

$$A\varphi_k = b_k \psi_k, \qquad A^* \psi_k = b_k \varphi_k.$$

From $[Af, \psi] = \langle f, A^* \psi \rangle$ and this last display, we have

$$[Af, \psi_k] = b_k \langle f, \psi_k \rangle. \tag{3.57}$$

Suppose now that A is one-to-one, so that $\{\varphi_k\}$ is an orthonormal basis for $\mathcal{H}$. Then we can use the "representer" equations (3.57) to express $f = \sum \langle f, \varphi_k \rangle \varphi_k$ in terms of quantities observable from (3.55), indeed

$$f = \sum b_k^{-1} [Af, \psi_k] \varphi_k.$$

From (3.56), $Y_k = Y(\psi_k) = [Af, \psi_k] + \epsilon Z(\psi_k)$ and so we get a sequence representation

$$Y_k = b_k \theta_k + \epsilon z_k. \tag{3.58}$$

As with the regression model, we set $y_k = Y_k / b_k$ and $\lambda_k = \epsilon / b_k$ to recover our basic sequence model (1.11). From this it is clear that the rate of variation inflation, i.e. the rate of decrease of b_k with k, plays a crucial role in the analysis.

Examples. (i) Differentiation. We observe $Y = g + \epsilon Z$ and seek to estimate the derivative $f = g'$. We can express g as the output of integration: $g(x) = Af(x) = \int_0^x f(s)ds$. We suppose that $\mathcal{H} = \mathcal{K} = L_{2,\text{per}}[0,1]$. Both sets of singular functions $\varphi_k = \psi_k$ are the complex exponentials $\psi_k(x) = \exp(2\pi ikx)$, $k \neq 0$, with singular values $b_k = 1/(2\pi ik) = O(k^{-1})$.

More generally, we might seek to recover $f = g^{(m)}$, so that g is the m-th iterated integral of f. In this case, the singular values $b_k = (2\pi ik)^{-m} = O(k^{-m})$ for $k \neq 0$.

(ii) Deconvolution. The smoothing operation consists of convolution with a known function b:

$$Af(x) = (b \star f)(x) = \int_0^1 b(x-t)f(t)dt,$$

and again the goal is to reconstruct f. The two-dimensional version is a natural model for image blurring.

In the easiest case for describing the SVD, when both f and b are periodic on $[0,1]$, we may again use the Fourier basis for $\mathcal{H} = \mathcal{K} = L_2[0,1]$, with $\varphi_k(x) = \psi_k(x) = e^{2\pi ix}$, and the singular values are the Fourier coefficients of b:

$$b_k = \int_0^1 b\psi_k,$$

since $(b \star f)_k = b_k f_k$.

If $b(x) = I\{|x| \le a\}$ is the "boxcar" blurring function, then $b_k = \sin(2\pi ka)/(\pi k)$, so that the singular values $b_k \approx O(k^{-1})$. For b smooth, say with r continuous derivatives, then $b_k = O(k^{-r})$ (e.g. Katznelson (1968, Ch. 1.4)).

(iii) The Abel equation $Af = g$ has

$$(Af)(x) = \frac{1}{\sqrt{\pi}} \int_0^x \frac{f(t)}{\sqrt{x-t}} dt$$

and goes back to Abel (1826), see Keller (1976) for an engaging elementary discussion and Gorenflo and Vessella (1991) for a list of motivating applications, including the Abel's original tautochrone problem.

To describe the singular value decomposition, let $\mathcal{H} = L_2[0,1]$ with $\{\varphi_k\}$ given by normalized Legendre polynomials $\varphi_k(x) = \sqrt{2n+1}P_k(1-2x)$. On the other side, let

$$\psi_n(x) = \sqrt{2/\pi}\sin(n+\tfrac{1}{2})\theta, \qquad x = \sin^2(\theta/2)$$

for $0 \le \theta \le \pi$. Setting $\tilde{\psi}_n(\theta) = \psi_n(x)$, the functions $\tilde{\psi}_n$ are orthonormal in $L_2[0,\pi]$ (and $\psi_n(x)$ can be expressed in terms of modified Jacobi polynomials $\sqrt{x}P_k^{1/2,-1/2}(1-2x)$, see (3.61) below). It is shown in Exercise 3.10 that $A\varphi_k = b_k\psi_k$ with singular values

$$b_k = (k+1/2)^{-1}.$$

Thus, in terms of decay of singular values, A behaves like half-order integration.

Remark. It is perhaps not immediately clear that A is a bounded linear operator on $L_2[0,1]$ (although of course it follows from the SVD). The kernel $A(s,t) = (s-t)^{-1/2}I\{s \ge t\}$ is

not square integrable on $[0,1]^2$, so the simplest criterion, finiteness of the Hilbert-Schmidt norm (C.5), doesn't apply. See the chapter Notes for further remarks.

(iii') Wicksell problem. Following Wicksell (1925) and Watson (1971), suppose that spheres are embedded in an opaque medium and one seeks to estimate the density of the sphere radii, p_S, by taking a planar cross-section through the medium and estimating the density p_O of the observed circle radii.

Assume that the centers of the spheres are distributed at random according to a homogeneous Poisson process. In Section 3.11 it is shown that p_O and p_S are related by

$$p_O(y) = \frac{y}{\mu}\int_y^b \frac{p_S(s)}{\sqrt{s^2-y^2}}ds, \qquad \mu = \int_0^b s p_S(s)ds. \tag{3.59}$$

We may put this into Abel equation form. Suppose, by rescaling, that $b = 1$ and work on the scale of squared radii, letting g be the density of $x = 1-y^2$ and p be the density of $t = 1-s^2$. Setting $\kappa = 2\mu/\sqrt{\pi}$, we get

$$g(x) = \frac{1}{2\mu}\int_0^x \frac{p(t)}{\sqrt{x-t}}dt = \frac{1}{\kappa}(Ap)(x).$$

Thus we can use observations on g and the SVD of A to estimate $f = p/\kappa$. To obtain an estimate of p we can proceed as follows. Since $\varphi_0 \equiv 1$ and p is a probability density, we have $\langle p, \varphi_0\rangle = 1$. Thus from (3.57)

$$1 = \kappa\langle f, \varphi_0\rangle = \kappa b_0^{-1}[Af, \psi_0]$$

and so $\kappa = b_0/[g,\psi_0]$ and hence

$$p = \kappa f = \sum_k \frac{b_0}{b_k}\frac{[g,\psi_k]}{[g,\psi_0]}\varphi_k$$

expresses p in terms of observable functions $[g,\psi_k]$.

(iv) Fractional order integration. For $\delta > 0$, let

$$(A_\delta f)(x) = \frac{1}{\Gamma(\delta)}\int_0^x \frac{f(t)}{(x-t)^{1-\delta}}dt = (f \star \Psi_\delta)(x) \tag{3.60}$$

where $\Psi_\delta(x) = x_+^{\delta-1}/\Gamma(\delta)$ and $x_+ = \max(x,0)$. Gel'fand and Shilov (1964, §5.5) explain how convolution with Ψ_δ and hence operator A_δ can be interpreted as integration of (fractional) order δ. Of course, $(A_1 f)(x) = \int_0^x f(t)dt$ is ordinary integration and $\delta = 1/2$ yields the Abel operator.

The SVD of A_δ can be given in terms of Jacobi polynomials $P_k^{a,b}(1-2x)$, Appendix **??** and Exercise 3.10:

$$\begin{aligned}
\varphi_k(x) &= \sqrt{2k+1}P_k(1-2x) && \text{on } L_2([0,1],dx)\\
\psi_k(x) &= g_{\delta,-\delta;k}^{-1}P_k^{\delta,-\delta}(1-2x) && \text{on } L_2([0,1],x^{-\delta}(1-x)^{-\delta}dx),\\
b_k &= (\Gamma(k-\delta+1)/\Gamma(k+\delta+1))^{1/2} \sim k^{-\delta} && \text{as } k\to\infty.
\end{aligned} \tag{3.61}$$

Thus, consistent with previous examples, the singular values decay at a rate corresponding to the order (integer or fractional) of integration.

(v) Heat equation. The classical one dimensional heat equation describes the diffusion of heat in a rod. If $u(x,t)$ denotes the temperature at position x in the rod at time t, then in appropriate units, u satisfies the equation

$$\frac{\partial}{\partial t}u(x,t) = \frac{\partial^2}{\partial x^2}u(x,t). \tag{3.62}$$

For our discussion here, we will assume that the initial temperature profile $u(x,0) = f(x)$ is unknown, and that the boundary conditions are periodic: $u(0,t) = u(1,t)$. We make noisy observations on the temperature in the rod at a time $T > 0$.

$$Y(x) = u(x,T) + \epsilon Z(x),$$

and it is desired to estimate the initial condition $f(x)$.

The heat equation (3.62) is a *linear* partial differential equation, having a unique solution which is a linear transform of the intial data f:

$$u(x,T) = (A_T f)(x).$$

This can be expressed in terms of the Gaussian heat kernel, but we may jump directly to the SVD of A_T by recalling that (3.62) along with the given boundary conditions can be solved by separation of variables. If we assume that the unknown, periodic f has Fourier sine expansion

$$f(x) = \sqrt{2}\sum_{k=1}^{\infty}\theta_k \sin \pi k x,$$

then

$$u(x,T) = \sqrt{2}\sum_{k=1}^{\infty}\theta_k e^{-\pi^2 k^2 T}\sin \pi k x.$$

Thus $\varphi_k(x) = \psi_k(x) = \sqrt{2}\sin \pi k x$, and the singular values $b_k = e^{-\pi^2 k^2 T}$.

The very rapid decay of b_k shows that the heat equation is extraordinarily ill-posed.

(vi) Radon transform and 2-d computed tomography (CT). In a two-dimensional idealization, this is the problem of reconstructing a function from its line integrals. Thus, let D be the unit disc in $\mathbb{R}^2$, and suppose that the unknown $f \in \mathcal{H} = L^2(D, \pi^{-1}dx)$.

A line at angle ϕ from the horizontal and distance s from the origin is given by $t \to (s\cos\phi - t\sin\phi, s\sin\phi + t\cos\phi)$ and denoted by $L_{s,\phi}$, compare Figure 3.2. The corresponding line integral is

$$\begin{aligned}(Af)(s,\phi) &= \text{Ave}\,[f|L_{s,\phi}\cap D] \\ &= \frac{1}{2\sqrt{1-s^2}}\int_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} f(s\cos\phi - t\sin\phi, s\sin\phi + t\cos\phi)dt.\end{aligned}$$

Here $(s,\phi) \in R = \{0 \le s \le 1, 0 \le \phi \le 2\pi\}$. The observations are noisy versions of the line integrals

$$Y(s,\phi) = Af(s,\phi) + \epsilon W(s,\phi), \qquad (s,\phi) \in R.$$

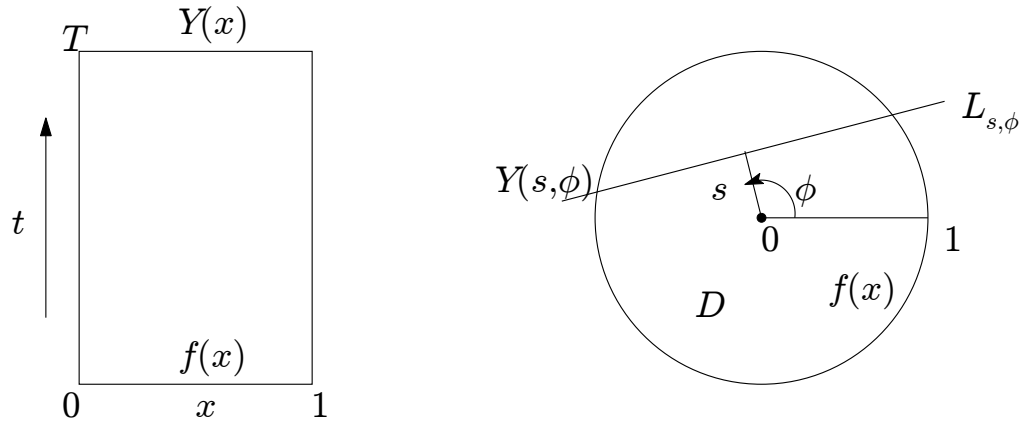

Figure 3.2 Left panel: domain for the heat equation. We observe $u(x,T)$ plus noise (top line) and wish to recover the initial data $f(x) = u(x,0)$, (bottom line). Right panel: domain for computed tomography example. We observe line integrals $(Af)(s,\phi)$ along lines $L_{s,\phi}$ plus noise, and wish to recover $f(x), x \in D$.

The SVD of A was derived in the optics and tomography literatures (Marr, 1974; Born and Wolf, 1975); we summarize it here as an example going beyond the Fourier basis. There is a double index set $N = \{(l,m) : m = 0,1,\ldots; l = m, m-2, \ldots, -m\}$, where m is the "degree" and l the "order". For $\nu = (l,m)$, the singular functions are given (Johnstone and Silverman, 1990) by

$$\varphi_\nu(r,\theta) = \sqrt{m+1}Z_m^{|l|}(r)e^{il\theta}, \qquad \psi_\nu(s,\phi) = U_m(s)e^{il\phi},$$

and the singular values $b_\nu = 1/\sqrt{m+1}$. Here $U_m(\cos\theta) = \sin(m+1)\theta/\sin\theta$ are Chebychev polynomials of the second kind, and the *Zernike* polynomials are characterized by the orthogonality relation $\int_0^1 Z_{k+2s}^k(r)Z_{k+2t}^k(r)rdr = ((k+2s+1)/2)\delta_{st}$..

The main point here is that the singular values b_ν decay slowly and so the reconstruction problem is only mildly ill-posed, consistent with the now routine use of CT scanners in medicine.

Remark. The model (3.55) and (3.58) is the natural infinite sequence model of $Y \sim N(A\theta, \epsilon^2 I)$. For an infinite sequence version of the covariance model $Y \sim N(\theta, \epsilon^2\Sigma)$, see the discussion of the Karhunen-Loève transform in Section 3.7.

3.9 Correlated noise

The Karhunen-Loève transform. Let $T = [a,b]$ or more generally, a compact set in $\mathbb{R}^d$. Suppose that $\{Z(t), t \in T\}$ is a zero mean Gaussian random process on an index set T. [That is, all finite-dimensional distributions $(Z(t_1),\ldots,Z(t_k))$ are Gaussian for all $(t_1,t_2,\ldots,t_k) \in T^k$ and positive integer k.] Assume also that Z is continuous in quadratic mean, or equivalently (Ash and Gardner, 1975, Ch 1.3) that the covariance function (or kernel)

$$R(s,t) = EZ(s)Z(t)$$

is jointly continuous in $(s,t) \in T^2$. The operator $Rf(s) = \int R(s,t)f(t)dt$ is nonnegative definite because it arises from a covariance kernel:

$$\langle Rf, f\rangle = \iint f(s)\mathrm{Cov}(Z(s), Z(t))f(t)dsdt = \mathrm{Var}\left(\int f(s)Z(s)ds\right) \geq 0.$$

Under these conditions it follows (Appendix C.3 has some details and references) that R is a compact operator on $L^2(T)$, and so it has, by the Hilbert-Schmidt theorem, a complete orthonormal basis $\{\varphi_n\}$ of eigenfunctions with eigenvalues $\lambda_n^2 \geq 0$,

$$\int R(s,t)\phi_i(t)dt = \lambda_i^2\phi_i(s), \qquad s \in T.$$

In addition, by Mercer's theorem, the series

$$R(s,t) = \sum \lambda_n^2\varphi_n(s)\varphi_n(t)$$

converges uniformly and in mean square on $T \times T$.

Define Gaussian variables (for i such that $\lambda_i > 0$)

$$z_i = \lambda_i^{-1}\int \varphi_i(t)Z(t)dt.$$

The z_i are i.i.d. $N(0,1)$: this follows from the orthonormality of eigenfunctions:

$$\mathrm{Cov}\left(\int \varphi_i Z, \int \varphi_j Z\right) = \int_{T\times T} \varphi_i R\varphi_j = \langle\varphi_i, R\varphi_j\rangle = \lambda_i^2\delta_{ij}. \tag{3.63}$$

The sum

$$Z(t) = \sum_i \lambda_i z_i \phi_i(t)$$

converges in mean-square on $L_2(T)$. Indeed, for a tail sum $r_{mn} = \sum_m^n \langle Z, \varphi_i\rangle\varphi_i$ we have, using (3.63), $Er_{mn}^2 = \sum_{i=m}^n \lambda_i^2\varphi_i^2(t) \to 0$ as $m, n \to \infty$ by Mercer's theorem.

If the eigenfunctions ϕ_i corresponding to $\lambda_i > 0$ are not complete, then we may add an orthonormal basis for the orthogonal complement of the closure of the range of R in $L_2(T)$ and thereby obtain an orthobasis for $L_2(T)$. Since R is symmetric, these ϕ_i correspond to $\lambda_i = 0$.

Now suppose that $Z(t)$ is observed with an unknown drift function added:

$$Y(t) = \theta(t) + \epsilon Z(t), \qquad t \in T.$$

If $\theta \in L_2(T)$, then we may take coefficients in the orthonormal set $\{\phi_i\}$:

$$y_i = \langle Y, \phi_i\rangle, \qquad \theta_i = \langle\theta, \phi_i\rangle,$$

to obtain exactly the sequence model (3.48). [Of course, co-ordinates corresponding to $\lambda_i = 0$ are observed perfectly, without noise.]

To summarize: for our purposes, the Karhunen-Loève transform gives (i) a diagonalization of the covariance operator of a mean-square continuous process, (ii) an example of the Gaussian sequence model, and (iii) a way to think about (and do computations with) Gaussian priors in the sequence model

Connection to Principal Components Analysis. The KLT is just the stochastic process

analog of finding the principal components of a sample covariance matrix. Indeed, suppose that the sample data is $\{x_{ij}\}$ for $i = 1, \ldots, n$ cases and $j = 1, \ldots, p$ variables. Let $\bar{x}_j = n^{-1} \sum_i x_{ij}$ denote the sample mean for variable j. Set $z_{ij} = x_{ij} - \bar{x}_j$ and make the correspondence $Z(\omega, t) = z_{ij}$, identifying the realization ω with i, and the "time" t with j. Then $R(t_1, t_2) = EZ(t_1)Z(t_2)$ corresponds to the an entry in the sample covariance matrix $S_{j_1 j_2} = n^{-1} \sum_i (x_{ij_1} - \bar{x}_{j_1})(x_{ij_2} - \bar{x}_{j_2})$.

Example: Integrated Wiener process priors.

The $m-1$-fold integrated Wiener process is defined by

$$Z_m^0(t) = \int_0^1 \frac{(t-u)_+^{m-1}}{(m-1)!} dW(u), \qquad t \in [0, 1].$$

The "free" Wiener process (so christened by Shepp (1966)) is derived from this with the aid of i.i.d standard Gaussian variables $\xi_0, \ldots, \xi_{m-1}$ independent of Z_m^0:

$$Z_m^\sigma(t) = \sigma \sum_{j=0}^{m-1} \xi_j \frac{t^j}{j!} + Z_m^0(t).$$

Most interesting is the case $m = 2$, since it corresponds to cubic smoothing splines:

$$Z_2^\sigma(t) = \sigma\xi_0 + \sigma\xi_1 t + \int_0^t (t-u) dW(u). \tag{3.64}$$

Wahba (1978, 1983, 1990) has advocated the use of Z_m^σ as a prior distribution for Bayesian estimation in the context of smoothing splines (actually, she recommends using $\sigma \to \infty$, for reasons that will be apparent. She showed (Wahba, 1990, Th. 1.5.3) that the smoothing spline based on the roughness penalty $\int (D^m f)^2$ arises as the limit of posterior means calculated from the Z_m^σ priors as $\sigma \to \infty$.)

This prior distribution has some curious features, so we explore its Karhunen-Loève transform now as preparation for later use. The key conclusion is that for each $\sigma \geq 0$, and in the $\sigma \to \infty$ limit, the eigenvalues satisfy

$$\lambda_i \sim 1/(\pi i)^m, \qquad \text{as } i \to \infty.$$

We discuss only the cases $m = 1, 2$ here.

However, it is simpler to discuss the $m = 1$ situation first, with $Z_1^\sigma(t) = \sigma\xi_0 + W(t)$, and covariance kernel $R_\sigma(s,t) = \mathrm{Cov}\,(Z_1^\sigma(s), Z_1^\sigma(t)) = \sigma^2 + s \wedge t$. The eigenvalue equation $R_\sigma \phi = \lambda^2 \phi$ becomes

$$\sigma^2 \int_0^1 \phi(t)dt + \int_0^s t\phi(t)dt + s \int_s^1 \phi(t)dt = \lambda^2 \phi(s). \tag{3.65}$$

Differentiating with respect to s yields

$$\int_s^1 \phi(t)dt = \lambda^2 \phi'(s) \tag{3.66}$$

and differentiating a second time yields the second order ordinary differential equation

$$-\phi(s) = \lambda^2 \phi''(s) \qquad 0 \leq s \leq 1. \tag{3.67}$$

	Boundary Conditions	Eigenvalues	Eigenfunctions
$\sigma = \infty$	$\phi'(0) = \phi'(1) = 0$	$\lambda_n^{-1} = n\pi$	$\sqrt{2}\cos n\pi t$
$\sigma = 0$	$\phi(0) = \phi'(1) = 0$	$\lambda_n^{-1} = (n + \frac{1}{2})\pi$	$\sqrt{2}\sin(n + \frac{1}{2})\pi t$
$0 < \sigma < \infty$	$\phi'(0) = \sigma^{-2}\phi(0)$, $\phi'(1) = 0$	$\lambda_n^{-1} \in (n\pi, (n + \frac{1}{2})\pi)$	$c_n \sin \lambda_n^{-1} t + \ldots$ $c_n \sigma^2 \lambda_n^{-1} \cos \lambda_n^{-1} t$
Periodic	$\phi(0) = \phi(1)$, $\phi'(0) = \phi'(1)$	$\lambda_{2n-1}^{-1} = \lambda_{2n}^{-1} = 2n\pi$	$\sqrt{2}\sin 2\pi n t$, $\sqrt{2}\cos 2\pi n t$

Table 3.2 *Effect of Boundary Conditions for the vibrating string equation*

The homogeneous equation $\lambda^2\phi'' + \phi = 0$ has two linearly independent solutions given by trigonometric functions

$$\phi(t) = a\sin(t/\lambda) + b\cos(t/\lambda). \tag{3.68}$$

The equations (3.65) and (3.66) impose boundary conditions which non-zero eigenfunctions must satisfy:

$$\phi'(1) = 0, \qquad \phi'(0) = \phi(0)/\sigma^2.$$

[The first condition is evident from (3.66) while the second follows by combining the two equations: $\lambda^2\phi'(0) = \int \phi = \lambda^2\phi(0)/\sigma^2$.]

Let us look first at the $\sigma \to \infty$ limit advocated by Wahba. In this case the boundary conditions become simply $\phi'(0) = \phi'(1) = 0$. Substituting into (3.68), the first condition implies that $a = 0$ and the second that $\sin(1/\lambda) = 0$. Consequently the eigenvalues and eigenfunctions are given by

$$\lambda_n = 1/n\pi, \qquad \phi_n(s) = \sqrt{2}\cos n\pi s, \qquad n = 1, 2, \ldots.$$

Equation (3.67) arises in traditional mathematical physics by separation of variables in the 'vibrating string' equation, e.g. Courant and Hilbert (1953, Sec. 5.3). The boundary condition $\phi'(1) = 0$ corresponding to the right end of the string being "free". In the case of the ordinary Wiener process ($\sigma = 0$), the left hand boundary condition becomes $\phi(0) = 0$, corresponding to the left end of the string being fixed at 0 – recall that $W(0) = 0$ almost surely. The condition for general σ, $\sigma^2\phi(0) = \phi'(0)$ corresponds to an 'elastically attached' endpoint.

Table 3.2 shows the eigenvalues λ_n and eigenfunctions corresponding to these various natural boundary conditions - all are easily derived from (3.68).

To describe the stochastic process, or "prior distribution" associated with *periodic* boundary conditions, recall that the Brownian Bridge $\tilde{W}(t) = W(t) - tW(1)$ satisfies $\tilde{W}(1) = \tilde{W}(0) = 0$ and has $\mathrm{Cov}(\tilde{W}(s), \tilde{W}(t)) = s \wedge t - st$. Proceeding as before, define a "free" Brownian Bridge

$$\tilde{Z}^\sigma(t) = \sigma\xi_0 + \tilde{W}(t),$$

and verify that it has covariance kernel $\tilde{R}_\sigma(s, t) = \sigma^2 + s \wedge t - st$. Equations (3.65) and (3.66) change in an obvious way, but the differential equation (3.67) remains the same. The

boundary conditions become

$$\phi(0) = \phi(1), \qquad \phi'(0) = \sigma^{-2}\phi(0) + \phi'(1),$$

and so the standard periodic boundary conditions and the usual sine and cosine eigenfunctions emerge from the $\sigma \to \infty$ limit.

In all cases summarized in Table 3.2, the eigenfunctions show increasing oscillation with increasing n, as measured by sign crossings, or frequency. This is a general phenomenon for such boundary value problems for second order differential equations (Sturm oscillation theorem - see e.g. Birkhoff and Rota (1969, Sec 10.7)). Note also that in the periodic case, the eigenvalues have multiplicity two – both sines and cosines of the given frequency – but in all cases the asymptotic behavior of the eigenvalues is the same: $\lambda_n^{-1} \sim n\pi$.

The analysis of the integrated Wiener prior (3.64), corresponding to cubic smoothing splines, then proceeds along the same lines, with most details given in Exercise 3.7 (see also Freedman (1999, Sec. 3)). The eigenvalue equation is a *fourth* order differential equation:

$$\phi(s) = \lambda^2 \phi^{(4)}(s).$$

This equation is associated with the vibrating *rod* (Courant and Hilbert, 1953, Secs IV.10.2 and V.4) – indeed, the roughness penalty $\int f''^2$ corresponds to the potential energy of deformation of the rod. It is treated analogously to the vibrating string equation. In particular, the (four!) boundary conditions for the $\sigma = \infty$ limit become

$$\phi''(0) = \phi'''(0) = 0, \qquad \phi''(1) = \phi'''(1) = 0,$$

corresponding to "free ends" at both limits.

3.10 Models with Gaussian limits*

Since the earliest days of nonparametric function estimation, striking similarities in large sample results – rates of convergence, distributional structure – have been observed in models as diverse as spectrum estimation, density estimation and nonparametric regression. In recent years, a rigorous expression of this phenomenon has been obtained used LeCam's notion of asymptotic equivalence of experiments. In each such case, a result exists stating that under certain regularity conditions on the unknown function f, in large samples, the model is asymptotically equivalent to the signal in Gaussian white noise model. Informally, this means that conclusions based on estimators, risk functions and asymptotic analysis in the white noise model can be carried over to corresponding estimators and risks in the other model sequence.

This section has two parts. In the first, we give the proof of the simplest case of the equivalence result of Brown and Low (1996a), which shows that nonparametric regression on $[0, 1]$ is asymptotically equivalent with the Gaussian white noise model. Some heuristics for this convergence were given in Chapter 1.4.

In the second part, essentially independent of the first, we give an informal, heuristic account of some of the other results in the growing list of equivalence results. The reader primarily interested in heuristcs can jump there directly.

Brown and Low's equivalence theorem

Outline of approach. We consider three statistical problems, each indexed by n, and having a common parameter space $f \in \Theta$.

$$(\mathcal{P}_n) \qquad dY_n(t) = f(t)dt + \sigma n^{-1/2} dW(t), \qquad 0 \le t \le 1, \tag{3.69}$$

$$(\bar{\mathcal{P}}_n) \qquad d\bar{Y}_n(t) = \bar{f}_n(t)dt + \sigma n^{-1/2} dW(t), \qquad 0 \le t \le 1, \tag{3.70}$$

$$(\mathcal{Q}_n) \qquad y_l = f(l/n) + \sigma z_l \qquad l = 1, \dots, n. \tag{3.71}$$

In problem $(\bar{\mathcal{P}}_n)$, the function $\bar{f}_n$ is a step function approximation to f, being piecewise constant on intervals $[(i-1)/n, i/n)$. We will define a distance $\Delta(\mathcal{P}_n, \mathcal{Q}_n)$ between statistical problems and show that it converges to zero in two steps. First, problems $\mathcal{P}_n$ and $\bar{\mathcal{P}}_n$ are on the same sample space, and so a convenient criterion in terms of L_1 distance shows that $\Delta(\mathcal{P}_n, \bar{\mathcal{P}}_n) \to 0$ under suitable conditions on Θ. Second, a reduction by sufficiency will show that in fact $\Delta(\bar{\mathcal{P}}_n, \mathcal{Q}_n) = 0$.

Before implementing this agenda, we need some definitions (due to Le Cam) to formalize the notion of distance between statistical problems. (See Le Cam (1986) and Le Cam and Yang (2000); also Nussbaum (2004) for an introduction and van der Vaart (2002) for historical perspective.)

Consider two regular[6] statistical problems $\mathcal{P}_0$ and $\mathcal{P}_1$ having sample spaces $\mathcal{Y}_0$ and $\mathcal{Y}_1$ but the *same* parameter space Θ. Let the two corresponding families of distributions be denoted by $\{P_{i,\theta}, \theta \in \Theta\}$ for $i = 0, 1$. To describe Le Cam's metric, we need to introduce risk functions. Let $\mathcal{A}$ be an action space and $L : \mathcal{A} \times \Theta \to [0, \infty)$ a loss function. The risk function of a (randomized) decision rule $\delta(A|y)$ is denoted by

$$r_L(\delta, \theta) = \iint L(a, \theta)\delta(da|y)P_\theta(dy), \tag{3.72}$$

compare (A.9) and the surrounding discussion for more detail. If $\delta(\cdot|y)$ is a point mass at $\hat{\theta}(y)$, then this definition reduces to (**??**).

The *deficiency* $\Delta_d(\mathcal{P}_0, \mathcal{P}_1)$ of $\mathcal{P}_0$ with respect to $\mathcal{P}_1$ is the smallest number $\epsilon \in [0, 1]$ such that for every arbitrary loss function L with $0 \le L(a, \theta) \le 1$ and every decision rule δ_1 in problem $\mathcal{P}_1$, there is a decision rule δ_0 in problem $\mathcal{P}_0$ such that $r_{0,L}(\delta_0, \theta) \le r_{1,L}(\delta_1, \theta) + \epsilon$ for all $\theta \in \Theta$. To obtain a distance on statistical problems, we symmetrize and set

$$\Delta(\mathcal{P}_0, \mathcal{P}_1) = \max\{\Delta_d(\mathcal{P}_0, \mathcal{P}_1), \Delta_d(\mathcal{P}_1, \mathcal{P}_0)\}.$$

The definition of distance is quite elaborate because it requires that performance in the two problems be similar regardless of the choice of estimand (action space) and measure of performance (loss function). In particular, since the loss functions need not be convex, randomized decision rules must be allowed (cf. (A.9)–(A.12) in Appendix A).

A simplification can often achieved when the problems have the same sample space.

Proposition 3.11 *If $\mathcal{Y}_0 = \mathcal{Y}_1$ and $\mathcal{P}_0$ and $\mathcal{P}_1$ have a common dominating measure ν, then*

$$\Delta(\mathcal{P}_0, \mathcal{P}_1) \le L_1(\mathcal{P}_0, \mathcal{P}_1),$$

[6] "Regular" means that it is assumed (a) that the sample spaces $\mathcal{Y}_i$ are complete separable metric spaces, equipped with the associated Borel σ-fields, and (b) that each family $\{P_{i,\theta}, \theta \in \Theta\}$ is dominated by a σ-finite measure. These assumptions hold for all cases we consider.

where the maximum L_1 distance is defined by

$$L_1(\mathcal{P}_0, \mathcal{P}_1) = \sup_{\theta\in\Theta} \int |p_{0,\theta}(y) - p_{1,\theta}(y)|\nu(dy).$$

Proof In the definition of deficiency, when the sample spaces agree, we can use the same decision rule in $\mathcal{P}_0$ as in $\mathcal{P}_1$, and if we write $\|L\|_\infty = \sup |L(a,\theta)|$, then from (3.72)

$$|r_{0,L}(\delta,\theta) - r_{1,L}(\delta,\theta)| \le \|L\|_\infty \int |p_{0,\theta}(y) - p_{1,\theta}(y)|\nu(dy).$$

Maximizing over θ shows that $r_{0,L}(\delta,\theta) \le r_{1,L}(\delta,\theta) + L_1(\mathcal{P}_0,\mathcal{P}_1)$. Repeating the argument with the roles of $\mathcal{P}_0$ and $\mathcal{P}_1$ reversed completes the proof. □

A sufficient statistic causes no loss of information in this sense.

Proposition 3.12 *Let $\mathcal{P}$ be a regular statistical problem with sample space $\mathcal{Y}$. Suppose that $S : \mathcal{Y} \to \mathcal{S}$ is a sufficient statistic, and let $\mathcal{Q} = \{Q_\theta; \theta \in \Theta\}$ denote the problem in which $S = S(Y)$ is observed. Then $\Delta(\mathcal{P},\mathcal{Q}) = 0$.*

Proof Since $S = S(Y)$ is sufficient for Y, there is a kernel $K(C|s)$ defined for (Borel) subsets $C \subset \mathcal{Y}$ such that $P_\theta(C) = \int K(C|s)Q_\theta(ds)$. This formalizes[7] the notion that the distribution of Y given S is free of θ. Given a decision rule δ for problem $\mathcal{P}$, we define a rule δ' for $\mathcal{Q}$ by $\delta'(A|s) = \int \delta(A|y)K(dy|s)$. By chasing the definitions, it is easy to verify, given a loss function L, that $r_L(\delta',\theta) = r_L(\delta,\theta)$. Hence $\Delta_d(\mathcal{Q},\mathcal{P}) = 0$. Since a rule for $\mathcal{Q}$ is automatically a rule for $\mathcal{P}$, we trivially have also $\Delta_d(\mathcal{P},\mathcal{Q}) = 0$, and hence $\Delta(\mathcal{P},\mathcal{Q}) = 0$. □

We are now ready to formulate and prove a special case of the Brown-Low theorem. Consider parameter spaces of Hölder continuous functions of order α. The case $0 < \alpha < 1$ is of most interest here—Appendix C gives the definitions for $\alpha \ge 1$. We set

$$\Theta_H^\alpha(C) = \{f \in C([0,1]) : |f(x) - f(y)| \le C|x-y|^\alpha, \text{for all } x, y \in [0,1]\}. \tag{3.73}$$

Theorem 3.13 *Let $\mathcal{P}_n$ and $\mathcal{Q}_n$ denote the continuous Gaussian white noise model* (3.69) *and the discrete regression model* (3.71) *respectively. Let the parameter space Θ for both models be the Hölder function class $\Theta_H^\alpha(C)$. Then, so long as $\alpha > 1/2$, the two problems are asymptotically equivalent:*

$$\Delta(\mathcal{P}_n, \mathcal{Q}_n) \to 0.$$

Proof We pursue the two step approach outlined earlier. Given a function $f \in \Theta_H^\alpha(C)$, define a piecewise constant step function approximation to it from the values $f(l/n)$. Set

$$\bar{f}_n(t) = f(l/n) \qquad \text{if } (l-1)/t \le t < l/n,$$

and put $\bar{f}_n(1) = f(1)$. [This type of interpolation from sampled values occurs again in Chapter 15.] As indicated at (3.70), let $\overline{\mathcal{P}}_n$ denote the statistical problem in which $\bar{f}_n$ is observed in continuous white noise. Since both $\mathcal{P}_n$ and $\overline{\mathcal{P}}_n$ have sample space $\mathcal{Y} = C([0,1])$ and are dominated, for example by P_0, the distribution of Y_n under $f = 0$, we have

[7] The existence of such a kernel, specifically a regular conditional probability distribution, is guaranteed for a regular statistical problem, see. e.g. Schervish (1995, Appendix B.3) or Breiman (1968).

$\Delta(\mathcal{P}_n, \overline{\mathcal{P}}_n) \le L_1(\mathcal{P}_n, \overline{\mathcal{P}}_n)$. The L_1 distance between P_f and $P_{\bar{f}_n}$ can be calculated fairly easily; indeed from (3.53) and (3.54),

$$\|P_f - P_{\bar{f}_n}\|_1 = 2[1 - 2\tilde{\Phi}(D_n(f)/2)],$$
$$D_n^2(f) = n\int_0^1 [\bar{f}_n(t) - f(t)]^2 dt.$$

From the Hölder assumption $|f(t) - f(l/n)| \le C|t - l/n|^\alpha$ for $t \in [(l-1)/n, l/n)$. [If $\alpha \ge 1$, it is enough to use $\alpha = 1$ and the Lipschitz property]. Consequently

$$D_n^2(f) \le n^2 C^2 \int_0^{1/n} u^{2\alpha} du = (2\alpha+1)^{-1} C^2 n^{1-2\alpha},$$

and this holds *uniformly* for all $f \in \Theta_H^\alpha(C)$. Hence $L_1(\mathcal{P}_n, \overline{\mathcal{P}}_n) \to 0$ so long as $\alpha > 1/2$.

For the second step, reduction by sufficiency, define

$$S_{n,l}(\bar{Y}_n) = n\left[\bar{Y}_n(l/n) - \bar{Y}_n((l-1)/n)\right], \qquad l = 1, \dots, n. \tag{3.74}$$

The variables $S_{n,l}$ are independent Gaussians with mean $f(l/n)$ and variance σ^2. Hence the vector $S_n = (S_{n,l})$ is an instance of statistical problem $\mathcal{Q}_n$. In addition, $S_n = S_n(\bar{Y}_n)$ is sufficient for $f \in \Theta$ in problem $\overline{\mathcal{P}}_n$, and so $\Delta(\overline{\mathcal{P}}_n, \mathcal{Q}_n) = 0$. Combining the two steps using the triangle inequality for metric Δ, we obtain $\Delta(\mathcal{P}_n, \mathcal{Q}_n) \to 0$. □

Remarks. 1. Let us describe how to pass from a procedure in one problem to a corresponding procedure in the other. Given a rule δ_n in regression problem $\mathcal{Q}_n$, we define a rule $\delta_n'(Y_n)$ in the white noise problem $\mathcal{P}_n$ simply by forming $S_n(Y_n)$ as in (3.74) and setting $\delta'(Y_n) = \delta_n(S_n)$. In the other direction we use the construction in the proof of Proposition 3.12. Given a rule δ_n in white noise problem $\mathcal{P}_n$, we define δ_n' in the regression problem by passing to problem $\overline{\mathcal{P}}_n$ (which has the same sample space as $\mathcal{P}_n$) and defining

$$\delta_n'(A|s_n) = E\left[\delta_n(A|\bar{Y}_n) \mid S_n(\bar{Y}_n) = s_n\right].$$

The conditional expectation is well defined as an estimator (free of f) by sufficiency, though of course it may in general be hard to evaluate. The evaluation is easy however in the case of a linear estimator $\delta_n(Y_n)(u) = \int_0^1 c(u,t)dY_n(t)$: one can check that

$$\delta_n'(S_n)(u) = \sum_{l=1}^n c_{nl}(u) S_{n,l}, \qquad c_{nl}(u) = \int_{(l-1)/n}^{l/n} c(u,t)dt.$$

2. Theorem 3.13 extends to a regression model with unequally spaced and heteroscedastic observations: instead of (3.71), suppose that $\mathcal{Q}_n$ becomes

$$y_l = f(t_{nl}) + \sigma(t_{nl}) z_l, \qquad l = 1, \dots, n.$$

If $t_{nl} = H^{-1}(l/(n+1))$ for a strictly increasing and absolutely continuous distribution function H and if $\sigma(t)$ is well-behaved, then after suitably modifying the definition (3.74), Brown and Low (1996a) show that $\mathcal{Q}_n$ is still asymptotically equivalent to $\mathcal{P}_n$.

3. An example shows that equivalence fails when $\alpha = 1/2$. Define $\epsilon_n(t) = \sqrt{t}$ on $[0, 1/(2n)]$ and then reflect it about $1/(2n)$ to extend to $[1/(2n), 1/n]$. Then extend ϵ_n by translation to each interval $[(l-1)/n, l/n]$ so as to obtain a tooth-like function on $[0,1]$

which is Hölder continuous with $\alpha = 1/2$, and for which $\sqrt{n}\int_0^1 \epsilon_n = \sqrt{2}/3$. Now consider estimation of the linear functional $Lf = \int_0^1 f(t)dt$. In problem $\mathcal{P}_n$, the normalized difference $\sqrt{n}(Y_n(1) - Lf) \sim N(0,1)$ exactly for all f and n. However, in model $\mathcal{Q}_n$, the observation vector $y = (y_l)$ has the same distribution whether $f = f_0 \equiv 0$ or $f = f_{1n} = \epsilon_n$, since $\epsilon_n(l/n) = 0$. Thus there can be no estimator $\delta_n(y)$ in $\mathcal{Q}_n$ for which $\sqrt{n}(\delta_n((y) - Lf) \to N(0,1)$ in distribution uniformly over $f \in \Theta_H^{1/2}(1)$, since $\sqrt{n}Lf_0 = 0$ while $\sqrt{n}Lf_{1n} = \sqrt{2}/3$.

Some other examples

Density Estimation. Suppose that $X_1, \ldots, X_n$ are drawn i.i.d. from an unknown density f supported on $[0,1]$. So long as f has Hölder smoothness greater than $1/2$, the experiment is asymptotically equivalent to

$$dY_t = f^{1/2}(t)dt + \tfrac{1}{2}n^{-1/2}dW_t, \qquad 0 \le t \le 1. \tag{3.75}$$

Nussbaum (1996). The appearance of the root density $f^{1/2}$ is related to the square root variance stabilizing transformation for Poisson data, which is designed to lead to the constant variance term. Note also that $f^{1/2}$ is square integrable with L_2 norm equal to 1!

Here is a heuristic argument, in the spirit of (1.24), that leads to (3.75). Divide the unit interval into $m_n = o(n)$ equal intervals of width $h_n = 1/m_n$. Assume also that $m_n \to \infty$ so that $h_n \to 0$. Write I_{kn} for the kth such interval, which at stage n extends from $t_k = k/m_n$ to t_{k+1}. First the 'Poissonization trick': draw a random number N_n of observations $X_1, \ldots, X_{N_n}$ of i.i.d. from f, with $N_n \sim \text{Poisson}(n)$. Then, because of the Poisson thinning property, the number of observations falling in the kth bin I_{kn} will be Poisson with mean $n\int_{I_{kn}} f \approx nf(t_k)h_n$. The square root transformation is variance stabilizing for the Poisson family and so $y_{kn} := \sqrt{N_n(I_{kn})} \sim N(\sqrt{f(t_k)nh_n}, 1/4)$ approximately for large n. Thus $y_k \approx \sqrt{f}(t_k)\sqrt{nh_n} + \frac{1}{2}e_{kn}$ with e_{kn} independent and approximately standard Gaussian. Now form a partial sum process as in (1.24), and premultiply by $\sqrt{h_n/n}$ to obtain

$$Y_n(t) = h_n^{1/2}n^{-1/2}\sum_1^{[m_n t]} y_{kn} \approx \sum_1^{[m_n t]} \sqrt{f}(t_k)h_n + (1/2)n^{-1/2}m_n^{-1/2}\sum_1^{[m_n t]} e_{kn}.$$

This makes it plausible that the process $Y_n(t)$, based on the density estimation model, merges in large samples with the Gaussian white noise process of (3.75).

A non-constructive proof of equivalence was given by Nussbaum (1996) under the assumption that f is $\alpha-$Hölder continuous for $\alpha > 1/2$, (3.73), and uniformly bounded below, $f(t) \ge \epsilon > 0$. A constructive argument was given by Brown et al. (2004) under a variety of smoothness conditions, including the Hölder condition with $\alpha > 1/2$. While the heuristic argument given above can be formalized for $\alpha > 1$, Brown et al. (2004) achieve $\alpha > 1/2$ via a conditional coupling argument that can be traced back to Komlós et al. (1975).

Nonparametric Generalized Linear Models. This is an extension of model (3.71) to errors drawn from an exponential family. Indeed count data with time varying Poisson intensities and dichotomous or categorical valued series with time varying cell probabilities occur naturally in practice (e.g. Kolaczyk (1997); Stoffer (1991)). We suppose that the densities in

the family may be written $P_\theta(dx) = p_\theta(x)\nu(dx)$ with $p_\theta(x) = e^{\theta U(x)-\psi(\theta)}$. Thus θ is the canonical parameter, $U(x)$ the sufficient statistic, $\nu(dx)$ the dominating measure on $\mathbb{R}$ and $\psi(\theta) = \log \int e^{\theta U(x)}\nu(dx)$ the cumulant generating function. (Lehmann and Casella (1998, Ch. 1) or Brown (1986) have more background on exponential families). All the standard examples – Poisson, Bernoulli, Gaussian mean, Gaussian variance, exponential – are included. We will describe a form of the equivalence result in the mean value parameterization, given by $\mu(\theta) = \psi'(\theta) = E_\theta U(X)$. Let $t_l = l/n$, $l = 1, \ldots, n$ and g be a sufficiently smooth function, typically with Hölder smoothness greater than 1/2. Assume that we have observations (t_i, X_i) in which X_i is drawn from $P_{\theta_i}(dx)$ with $\mu_i = \mu(\theta_i) = g(t_i)$. Recall that $\psi''(\theta) = \mathrm{Var}_\theta U(X)$, and let $V(\mu)$ be the *variance stabilizing* transformation for $\{P_\theta\}$ defined through $V'(\mu(\theta)) = 1/\sqrt{\psi''(\theta)}$. Then Grama and Nussbaum (1998) show that this experiment is asymptotically equivalent to

$$dY_t = V(g(t))dt + n^{-1/2}dW_t \qquad 0 \le t \le 1.$$

The Poisson case, with $V(\mu) = 2\sqrt{\mu}$, is closely related to the density estimation setting. For a second example, if X_l are independent $N(0, g(t_l))$, then we are in the Gaussian scale family and the corresponding exponential family form for $N(0, \sigma^2)$ has natural parameter $\theta = -1/\sigma^2$, mean parameter $\mu(\theta) = -1/(2\theta)$ and variance stabilising transformation $V(\mu) = 2^{-1/2}\log\mu$. So the corresponding white noise problem has $dY_t = 2^{-1/2}\log g(t) + n^{-1/2}dW_t$, for $t \in [0, 1]$.

Spectral density estimation. Suppose that $X^n = (X_1, \ldots, X_n)$ is a sample from a stationary Gaussian random process with mean zero and spectral density function $f(\xi)$ on $[-\pi, \pi]$, related to the covariance function $\gamma(k) = EX_jX_{j+k}$ via $f(\xi) = (2\pi)^{-1}\sum_{-\infty}^{\infty} e^{i\xi k}\gamma(k)$. Estimation of the spectral density f was the first nonparametric function estimation model to be studied asymptotically – see for example Grenander and Rosenblatt (1957).

Observe that $X^n \sim N(0, \Gamma_n(f))$ where the covariance matrix is Toeplitz: $\Gamma_n(f)_{jk} = \gamma(k-j)$. A classical approximation in time series analysis replaces the Toeplitz covariance matrix by a *circulant* matrix $\tilde{\Gamma}_n(f)$ in which the rows are successive shifts by one of a single periodic function on $\{0, 1, \ldots, n-1\}$. [8] The eigenvalues of a circulant matrix are given by the discrete Fourier transform of the top row, and so the eigenvalues of $\tilde{\Gamma}_n(f)$ are *approximately* $f(\xi_j)$ where ξ_j are equispaced points on $[-\pi, \pi]$. After an orthogonal transformation to diagonalize $\tilde{\Gamma}_n(f)$, one can say heuristically that the model $X^n \sim N(0, \Gamma_n(f))$ is approximately equivalent to

$$Z_j \sim N(0, f(\xi_j)), \qquad j = 1, \ldots, n.$$

This is the Gaussian scale model discussed earlier, and so one expects that both statistical problems will be asymptotically equivalent with

$$dZ_\xi = \log f(\xi) + 2\pi^{1/2}n^{-1/2}dW_\xi, \qquad \xi \in [-\pi, \pi]$$

for f in a suitable function class, such as the Hölder function class $\Theta^\alpha_H(C)$ on $[-\pi, \pi]$ with $\alpha > 1/2$ and restricted also to bounded functions $\epsilon \le f(\xi) \le 1/\epsilon$. Full proofs are given in Golubev et al. (2010).

[8] Indeed, set $\tilde{\gamma}_n(i) = \gamma(i)$ for $0 \le i \le (n-1)/2$, make $\tilde{\gamma}_n$ periodic by reflection about $n/2$ and define $\tilde{\Gamma}_n(f)_{jk} = \tilde{\gamma}_n(k-j)$.

This by no means exhausts the list of examples where asymptotic equivalence has been established; one might add random design nonparametric regression and estimation in diffusion processes. For further references see the bibliography of Carter (2011).

Some cautions are in order when interpreting these results. First, there are significant regularity conditions, for example concerning the smoothness of the unknown f. Thus, Efromovich and Samarov (1996) have a counterexample for estimation of $\int f^2$ at very low smoothness. Meaningful error measures for spectral densities may not translate into, say, squared error loss in the Gaussian sequence model. See Cai and Zhou (2009) for some progress with unbounded loss functions (and also the discussion in Chapter ??). Nevertheless, the asymptotic equivalence results lend further strength to the idea that the Gaussian sequence model is the fundamental setting for nonparametric function estimation, and that theoretical insights won there will have informative analogs in the more concrete practical problems of curve estimation.

3.11 Details

Derivation of Wicksell equation (3.59). Referring to Figure 3.3, suppose that the sampling plane is perpendicular to the horizontal axis. Let the true sphere radius be s. The sampling plane intersects the horizontal axis at $r = s\cos\theta$ and the radius of the circle seen in the vertical planar cross section is $y = s\sin\theta$.

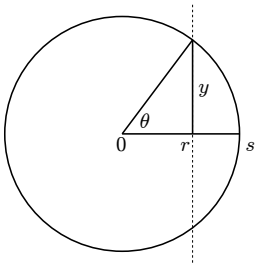

Figure 3.3 (two-dimensional projection of) three-dimensional sphere showing cut by sampling plane (dotted) perpendicular to horizontal axis

Observe that there is length-biased sampling: the sampling plane hits a sphere with probability proportional to its radius, so the density of sampled, or 'cut' sphere radii $p_C(s)$ is related to the true sphere radius density $p_S(s)$ by

$$p_C(s) = sp_S(s)/\mu \tag{3.76}$$

where $\mu = \int sp_S(s)ds$ is the normalization constant.

The distribution of observed radii decomposes over the sampled sphere density

$$P(Y \geq y) = \int_y^b P(Y \geq y|S = s)p_C(s)ds.$$

Conditional on $S = s$, the event $\{Y \geq y\} = \{s\sin\Theta \geq y\} = \{s\cos\Theta \leq \sqrt{s^2 - y^2}\}$. Now

$R = s\cos\Theta$ is uniformly distributed on $[0, s]$ by the homogeneous Poisson assumption, so it follows that $P(Y \geq y|S = s) = (1 - y^2/s^2)^{1/2}$. Consequently,

$$p_O(y) = -\frac{d}{dy}P(Y \geq y) = \int_y^b \frac{1}{\sqrt{1 - y^2/s^2}} \frac{y}{s^2} p_C(s)ds.$$

Substituting the biased sampling equation (3.76), we recover (3.59).

3.12 Notes

Defining Gaussian measures on infinite dimensional spaces is not completely straightforward and we refer to books by Kuo (1975) and Bogachev (1998) for complete accounts. For the sequence model (3.1) with $I = \mathbb{N}$, the subtleties can usually be safely ignored. For the record, as sample space for model (3.1), we take $\mathbb{R}^\infty$, the space of sequences in the product topology of pointwise convergence, under which it is complete, separable and metrizable. It is endowed with the Borel $\sigma-$field, and as dominating measure, we take $P_0 = P_{0,\epsilon}$, the centered Gaussian Radon measure (see Bogachev (1998, Example 2.3.5)) defined as the product of a countable number of copies of the $N(0, \epsilon^2)$ measure on $\mathbb{R}$.

For each $\theta \in \Theta = \ell_2(\mathbb{N})$, the measure P_θ with mean θ is absolutely continuous, indeed equivalent, to P_0, and has density

$$f_\theta(x) = dP_\theta/dP_0 = \exp\{\langle\theta, x\rangle/\epsilon^2 - \|\theta\|^2/2\epsilon^2\}.$$

Note that the random variable $\langle\theta, x\rangle$ appearing in the density has a $N(0, \|\theta\|^2)$ distribution under P_0 and in particular is finite P_0-almost surely.

In fact, it follows from the classical theorem of Kakutani (1948) on product measures that membership of θ in ℓ_2 is a necessary and sufficient condition for the distribution of x to be absolutely continuous with respect to that obtaining when $\theta = 0$, and further if $\theta \notin \ell_2$, then the two distributions are mutually singular.

Bogachev (1998, Theorem 3.4.4) shows that in a certain, admittedly weak, sense all infinite dimensional Gaussian measures are isomorphic to the sequence measure P_0.

One can formally extend the infinitesimal representation (1.19) to a compact set $D \subset \mathbb{R}^n$ if $t \to W_t$ is $d-$parameter Brownian sheet (Hida, 1980). If φ_i is an orthonormal basis for $L_2(D)$, then the operations (1.21) again yield data in the form of model (3.1).

Rice and Rosenblatt (1981) show that in the non-periodic case, the rate of convergence of the MSE is determined by the boundary behavior of f.

Speckman (1985).

References on data-determined choices of λ.

There is a large literature on the matching of posterior and frequentist probabilities in parametric models - the Bernstein-von Mises phenomenon. The situation is more complicated for non-parametric models. Some simple examples are possible with Gaussian sequence models and Gaussian priors—Johnstone (2010) develops three examples to illustrate some possibilities.

L_2 boundedness of the fractional integration operator A_δ is a consequence of classical results of Hardy and Littlewood (1928), see also Gorenflo and Vessella (1991, pp. 64–67).

Indeed, for $\delta \le 1/2$, the operator A_δ is bounded from $L_2[0,1]$ to $L_s[0,1]$ for a value $s = s(\delta) > 2$, while for $\delta > 1/2$, it is bounded from $L_2[0,1]$ to $C^{\delta-1/2}([0,1])$.

There is some discussion of orthogonal series methods in Hart (1997), though the emphasis is on lack-of-fit tests. Eubank (1999) has a focus on spline smoothing.

Exercises

3.1 *(Compactness criteria.)* Here ℓ_2 denotes square summable sequences with the norm $\|\theta\|^2 = \sum \theta_i^2$.
(a) The ellipsoid $\Theta = \{\theta : \sum_{k\ge1} a_k^2\theta_k^2 \le C^2\}$ is ℓ_2-compact if and only if $a_k > 0$ and $a_k \to \infty$.
(b) The hyperrectangle $\Theta = \prod_{k\ge1}[-\tau_k, \tau_k]$ is ℓ_2-compact if and only if $\sum_{k\ge1}\tau_k^2 < \infty$.

3.2 *(Equivalence of measures.)* Let P and Q be probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$, absolutely continuous with respect to a probability measure λ. (For example, $\lambda = (P+Q)/2$.) Write $p = dP/d\lambda$ and $q = dQ/d\lambda$. The *Hellinger affinity* $h(P,Q) = \int\sqrt{pq}\,d\lambda$ does not depend on the choice of λ. Let $\{P_n\}$ and $\{Q_n\}$ be two sequences of probability measures on $\mathbb{R}$. Define product measures on sequence space $\mathbb{R}^\infty$, with the product Borel $\sigma-$field, by $P = \prod P_n$ and $Q = \prod Q_n$. Then the celebrated theorem of Kakutani (1948) states that if $P_n \sim Q_n$ for $n = 1, 2, \ldots$ then P and Q are either equivalent or orthogonal. Moreover, $P \sim Q$ if and only if $\prod_{k=1}^\infty h(P_k, Q_k) > 0$. In case $P \sim Q$, $dP/dQ = \prod_{k=1}^\infty dP_k/dQ_k$.
(i) Taking all this as given, show first that

$$h(N(\theta_1,\sigma_1^2), N(\theta_2,\sigma_2^2) = \Big(\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}\Big)^{1/2}\exp\Big\{-\frac{(\theta_1-\theta_2)^2}{4(\sigma_1^2+\sigma_2^2)}\Big\}.$$

(ii) Suppose that z_i are i.i.d $N(0,1)$ for $i = 1, 2, \ldots$. Show that the measure P_θ corresponding to $y_i = \theta_i + \lambda_i z_i$ is absolutely continuous with respect to P_0 if and only if $\sum\theta_i^2/\lambda_i^2 < \infty$ and write down the likelihood ratio. `[WHAT IF θ = 0?]`
(iii) In the Gaussian sequence model $y_k = \theta_k + \epsilon z_k$, consider priors $\theta_k \sim N(0, \tau_k^2)$, independently with $\tau_k^2 = bk^{-2m}$. Under what conditions on m is the *marginal* distribution $P_\pi(dy)$ equivalent to $P_0(dy)$, the distribution conditional on $\theta = 0$?

3.3 *(Discrete orthogonality relations).* Let $\mathbf{e}_k$ denote the vector in $\mathbb{C}^n$ obtained by sampling the k-th complex exponential at $t_j = j/n$. Thus $\mathbf{e}_k = \{\exp(2\pi ikj/n), j = 0, 1, \ldots, n-1\}$. For $\mathbf{f}, \mathbf{g} \in \mathbb{C}^n$, use the usual inner product $\langle\mathbf{f},\mathbf{g}\rangle = \sum_1^n f_k\bar{g}_k$. Show that for $k, l \in \mathbb{Z}$,

$$\langle\mathbf{e}_k, \mathbf{e}_l\rangle = \begin{cases} n & \text{if } k-l \in n\mathbb{Z} \\ 0 & \text{otherwise.}\end{cases}$$

Turn now to the real case. For $k \ge 0$, let $\mathbf{c}_k = \{\cos(2\pi kj/n), j = 0, 1, \ldots, n-1\}$ and define $\mathbf{s}_k$ analogously using the $k-$th sine frequency. If $n = 2m+1$ is odd, then take $\{\mathbf{c}_0, \mathbf{s}_1, \mathbf{c}_1, \ldots \mathbf{s}_m, \mathbf{c}_m\}$ as the basis B_n for $\mathbb{R}^n$. If $n = 2m+2$ is even, then adjoin $\mathbf{c}_{n/2}$ to the previous set to form B_n. Show that the following orthogonality relations hold for basis vectors in B_n:

$$\langle\mathbf{c}_k,\mathbf{c}_l\rangle = \langle\mathbf{s}_k,\mathbf{s}_l\rangle = \frac{n}{2}\delta_{kl}, \qquad \langle\mathbf{c}_k,\mathbf{s}_l\rangle = 0,$$

with the exception of

$$\langle\mathbf{c}_0,\mathbf{c}_0\rangle = \langle\mathbf{c}_{n/2},\mathbf{c}_{n/2}\rangle = n,$$

where the last equation is only needed if n is even.
Hint. Derive the real relations from the complex by writing $\mathbf{e}_k = \mathbf{c}_k + i\mathbf{s}_k$ and using the complex orthogonality relations for pairs (k, l) and $(k, -l)$.

3.4 *(Infinite order kernels.)* Let $h_c(\xi) = 1/(|\xi| - c)^2$ and show that the function $e^{h_0(\xi)} I\{\xi \geq 0\}$ is C^∞. Define

$$\widehat{K}(\xi) = \begin{cases} 1 & \text{if } |\xi| \leq c \\ \exp\{-bh_1(\xi)\exp(-bh_c(\xi))\} & \text{if } c \leq |\xi| \leq 1 \\ 0 & \text{if } |\xi| \geq 1 \end{cases}$$

and show that $K(s) = (2\pi)^{-1} \int e^{is\xi} \widehat{K}(\xi) d\xi$ is a C^∞ kernel of infinite order (i.e. satisfies (3.21) with $q = \infty$) that decays faster than $|s|^{-m}$ for any $m > 0$. (McMurry and Politis, 2004)

3.5 *(Fourier transform of the equivalent kernel.)* The Fourier transform of an integrable function on $\mathbb{R}$ is defined by $\hat{f}(\xi) = \int_\infty^\infty f(x) e^{-i\xi x} dx$. If f is sufficiently nice, it may be recovered from the inversion formula $f(x) = (2\pi)^{-1} \int_\infty^\infty \hat{f}(\xi) e^{i\xi x} d\xi$. The Poisson summation formula (e.g. Mallat, page 28) states that under suitable conditions on f, [$(1 + x^2)(|f(x)| + |f'(x)| + |f''(x)|)$ bounded or the same condition on $\hat{f}$ will do], then

$$\sum_{k \in \mathbb{Z}} f(k) = \sum_{k \in \mathbb{Z}} \hat{f}(2\pi k).$$

Use the Poisson summation formula to give an alternate demonstration that the kernel $K_h(s) = 1 + 2\sum_1^\infty \frac{\cos 2\pi k s}{1 + (2\pi k h)^4}$ is the wrapped version $\sum_k L_h(s + k)$ of L_h, with Fourier transform given by

$$\widehat{L_h}(\xi) = [1 + \xi^4 h^4]^{-1}. \tag{3.77}$$

3.6 *(Evaluation of equivalent kernel.)* If $\alpha \in \mathbb{C}$ belongs to the upper half plane, show by contour integration that

$$\frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{e^{i\gamma x}}{x - \alpha} dx = \begin{cases} e^{i\alpha\gamma} & \text{if } \gamma > 0 \\ 0 & \text{if } \gamma < 0. \end{cases}$$

Use the partial fraction expansion

$$\prod_{k=1}^r (x - \beta_k)^{-1} = \sum_{k=1}^r c_k (x - \beta_k)^{-1}, \qquad 1/c_k = \prod_{j \neq k} (\beta_k - \beta_j),$$

to compute the equivalent kernel $L(t)$ given that $\hat{L}(\xi) = (1 + \xi^4)^{-1}$.

3.7 *(Wahba's prior for cubic splines.)* Show that

$$Z_2^\sigma(t) = \sigma\xi_1 + \sigma\xi_2 t + \int_0^t (t - u) dW(u),$$

the integrated (free) Wiener process, has covariance function

$$R_\sigma(s, t) = \sigma^2(1 + st) + R_0(s, t),$$
$$R_0(s, t) = \begin{cases} \frac{1}{2}s^2 t - \frac{1}{6}s^3 & 0 \leq s \leq t \\ \frac{1}{2}st^2 - \frac{1}{6}t^3 & 0 \leq t \leq s. \end{cases}$$

By differentiating the eigenvalue equation

$$\int_0^1 R_\sigma(s,t)\varphi(t)dt = \lambda^2\varphi(s)$$

four times, show that φ satisfies

$$\varphi(s) = \lambda^2\varphi^{(4)}(s),$$

with boundary conditions

$$\varphi''(0) = \sigma^{-2}\varphi'(0), \varphi'''(0) = \sigma^{-2}\varphi(0) \qquad \varphi''(1) = \varphi'''(1) = 0.$$

With $\sigma = 0$, show that the boundary conditions imply the equation $\cos\lambda^{-1/2}\cosh\lambda^{-1/2} = -1$ for the eigenvalues. In the $\sigma = \infty$ limit, show that the corresponding equation is $\cos\lambda^{-1/2}\cosh\lambda^{-1/2} = 1$. In either case, show that the eigenvalues satisfy, for large n

$$\lambda_n \sim \frac{1}{(n+\frac{1}{2})^2\pi^2} \sim \frac{1}{n^2\pi^2}.$$

Make plots of the first six eigenfunctions corresponding to the $\sigma = \infty$ limit.

3.8 *(Computational comparison.)* Consider two functions on $[0,1]$:

$$f_1(t) = \sin 4\pi t^2, \qquad f_2(t) = (e^{4t} - 1 - t)(1-t)^2,$$

and consider the model

$$Y_i = f(i/n) + \sigma z_i, \qquad z = 1, \ldots, n,$$

with $\sigma = 1$ and $z_i \sim N(0,1)$ chosen i.i.d. Let $\hat{f}_{SS,\lambda}$ and $\hat{f}_{PER,\lambda}$ denote the solutions to

$$\min Q(f) = n^{-1}\sum[Y_i - f(i/n)]^2 + \lambda\int_0^1 f''^2$$

among cubic splines and trignometric polynomials respectively. Note that $\hat{f}_{SS,\lambda}$ can be computed in `S-PLUS` using `smooth.spline()`. For $\hat{f}_{PER,\lambda}$, you'll need to use the discrete Fourier transform `fft()`, with attention to the real and imaginary parts. For λ, use the value suggested by the ellipsoid considerations in class:

$$\lambda = (\pi/2)^4(6\sqrt{2})^{4/5}(n\int f''^2)^{-4/5}.$$

Run experiments with $R = 100$ replications at $n = 50, 200$ and 1000 to compare the estimates $\hat{f}_{SS,\lambda}$ and $\hat{f}_{PER,\lambda}$ obtained for f_1 and f_2. Make visual comparisons on selected replications *chosen in advance*, as well as computing averages over replications such as

$$\frac{\text{ave}\,\|\hat{f}_{SS} - \hat{f}_{PER}\|_2^2}{\text{ave}\,\|\hat{f}_{SS} - f\|_2^2}.$$

3.9 Consider a slightly different family of shrinkage rules, to appear in Pinsker's theorem, and also indexed by a positive parameter:

$$\hat{\theta}_{\mu,k}(y) = (1 - k^m/\mu)_+ y_k, \qquad k \in \mathbb{N}.$$

Show that the maximum risk over a Sobolev ellipsoid $\Theta_2^\alpha(C)$ is approximated by

$$\bar{r}(\hat{\theta}_\mu;\epsilon) \sim \bar{v}_m\epsilon^2\mu^{1/m} + C^2\mu^{-2\min(\alpha/m,1)},$$

where

$$\bar{v}_m = 2m^2/(m+1)(2m+1).$$

If $\alpha = m$, show that the maximum MSE associated with the minimax choice of μ is given by

$$\bar{r}(\hat{\theta}_{\mu_*}; \epsilon) \sim e^{H(r)} C^{2-2r} (\bar{v}_m \epsilon^2)^r. \tag{3.78}$$

3.10 *(SVD for fractional integration.)* Let A_δ be the operator of fractional order integration (3.60). This exercise outlines the derivation of the singular value decomposition for a class of domain spaces, based on identites for Gauss' hypergeometric function and Jacobi polynomials that are recalled in Appendix **??**. Let $\rho_n(a, \delta) = \Gamma(a+n+1)/\Gamma(a+\delta+n+1) \sim n^{-\delta}$ as $n \to \infty$.
(a) Interpret identities (C.27) and (C.28) in terms of the operator A_δ and Jacobi polynomials:

$$A_\delta[w^a P_n^{a,b}(1-2w)](x) = \rho_n(a,\delta) x^{a+\delta} P_n^{a+\delta, b-\delta}(1-2x).$$

(b) Let $g_{a,b;n}$ denote the normalizing constants for Jacobi polynomials in (C.29); show that

$$\varphi_{a,b;n}(x) := g_{a,b;n}^{-1} w^a P_n^{a,b}(1-2x)$$

are orthonormal in $H^2_{-a,b} := L_2\big([0,1], x^{-a}(1-x)^b dx\big)$.
(c) Verify that the singular value decomposition of $A_\delta : H^2_{-a,b} \to H^2_{-a-\delta, b-\delta}$ is given by

$$\varphi_n = \varphi_{a,b;n}, \quad \psi_n = \varphi_{a+\delta, b-\delta; n}, \quad b_n^2 = \rho_n(a,\delta)\rho_n(b-\delta,\delta) \sim n^{-2\delta}, \quad n \to \infty.$$

(d) Set $a = 0$ to recover the SVD of A_δ as given in (3.61).
(e) Set $a = 0, \delta = 1/2$ and use the formula (Szegö, 1967, (4.1.8))

$$P_n^{1/2,-1/2}(x) = \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots 2n} \frac{\sin((2n+1)\theta/2)}{\sin(\theta/2)}, \quad x = \cos\theta$$

to recover the SVD of $A_{1/2}$ as given in Section 3.8 part (iii).

4

Gaussian decision theory

> In addition to those functions studied there are an infinity of others, and unless some principle of selection is introduced we have nothing to look forward to but an infinity of test criteria and an infinity of papers in which they are described. (G. E. P. Box, in *J. R. S. S. B. 19??*)

In earlier chapters we have formulated the Gaussian sequence model and indicated our interest in comparisons of estimators through their maximum risks, typically mean squared error, over appropriate parameter spaces. It is now time to look more systematically at questions of optimality.

Many powerful tools and theorems relevant to our purpose have been developed in classical statistical decision theory, often in far more general settings than used here. This chapter introduces some of these ideas, tailored for our needs. We focus on comparison of properties of estimators rather than the explicit taking of decisions, so that the name "decision theory" is here of mostly historical significance.

Our principle of selection—comparison, really—is minimaxity: look for estimators whose worst case risk is (close to) as small as possible for the given parameter space, often taken to encode some relevant prior information. This principle is open to the frequent and sometimes legitimate criticism that the worst case may be an irrelevant case. However, we aim to show that by appropriate choice of parameter space, and especially of *families* of parameter spaces, that sensible estimators emerge both blessed and enlightened from examination under the magnifying glass of the minimax prinicple.

A minimax estimator is exactly or approximately a Bayes estimator for a suitable "least favorable" prior. It is then perhaps not surprising that the properties of Bayes rules and risks play a central role in the study of minimaxity. Section 4.1 begins therefore with Bayes estimators, now from a more frequentist viewpoint than in Chapter 2. Section 4.2 goes more deeply than Chapter 2 into some of the elegant properties and representations that appear for squared error loss in the Gaussian model.

The heart of the chapter lies in the development of tools for evaluating, or approximating $R_N(\Theta)$, the minimax risk when the parameter is assumed to belong to Θ. Elementary lower bounds to minimax risk can often be derived from Bayes rules for priors supported on the parameter space, Section 4.3. For upper bounds and actual evaluation of the minimax risk, the minimax theorem is crucial. This is stated in Section 4.4, but an overview of its proof, even in this Gaussian setting, must be deferred to Appendix A.

Statistical independence and product structure of parameter spaces plays a vital role in

"lifting" minimax results from simpler component spaces to their products, as shown in Section 4.5.

A theme of this book is that conclusions about function estimation can sometimes be built up from very simple, even one dimensional, parametric constituents. As an extended example of the techniques introduced, we will see this idea at work in Sections 4.6 - 4.8. We start with minimaxity on a bounded interval in a single dimension and progress through hyperrectangles—products of intervals—to ellipsoids and more complex quadratically convex sets in $\ell_2(\mathbb{N})$.

Byproducts include conclusions on optimal (minimax) rates of convergence on Hölder, or uniform, smoothness classes, and the near mean square optimality of linear estimators over all quadratically convex sets.

A final Section 4.10 outlines a method for the exact asymptotic evaluation of minimax risks using classes of priors with appropriately simple structure. While this material is used on several later occasions, it can be omitted on first reading.

4.1 Bayes Estimators

In Section 2.3 we approached Bayes rules via calculations with the posterior distribution, for example using the posterior mean for squared error loss. In this chapter we largely adopt a different, though equivalent, approach, which considers instead the average of (frequentist) risk functions with respect to a prior distribution. Thus, if π is a probability distribution on $\ell_2(I)$, the *integrated risk* of an estimator $\hat{\theta}$ is defined by

$$\begin{aligned} B(\hat{\theta}, \pi) &= \int r(\hat{\theta}, \theta)\pi(d\theta) \\ &= E_\pi r(\hat{\theta}, \theta) = E_\pi E_\theta L(\hat{\theta}(y), \theta). \end{aligned} \tag{4.1}$$

An estimator $\hat{\theta}_\pi$ that minimizes $B(\hat{\theta}, \pi)$ for a fixed prior π is called a Bayes estimator for π, and the corresponding minimum value is called the *Bayes risk* $B(\pi)$; thus

$$B(\pi) = \inf_{\hat{\theta}} B(\hat{\theta}, \pi). \tag{4.2}$$

Of course $B(\pi) = B(\pi, \epsilon)$ also depends on the noise level ϵ, but again this will not always be shown explicitly.

Remark 4.1 One of the reasons for using integrated risks is that, unlike the ordinary risk function $\theta \to r(\hat{\theta}, \theta)$, the mapping $\pi \to B(\hat{\theta}, \pi)$ is *linear.* Representation (4.2) then shows that the Bayes risk $B(\pi)$ is a concave function of π.

The decidedly frequentist definition of Bayes estimators fortunately agrees with the subjectivist definition given at (2.7), under mild regularity conditions. We saw that the joint distribution πP of the pair (θ, y) may be decomposed two ways:

$$\pi P(d\theta, dy) = \pi(d\theta)P(dy|\theta) = P_\pi(dy)\pi(d\theta|y),$$

where $P_\pi(dy)$ is the marginal distribution of y and $\pi(d\theta|y)$ is the posterior distribution of

θ given y. The integrated risk of (4.1), which uses the first decomposition, may be written using the second, posterior decomposition as

$$B(\hat{\theta}, \pi) = E_{P_\pi} E_y L(\hat{\theta}(y), \theta).$$

Here, E_{P_π} denotes expectation with respect to the marginal distribution $P_\pi(dy)$ and E_y denotes expectation with respect to the posterior $\pi(d\theta|y)$. Thus one sees that $\hat{\theta}_\pi(y)$ is indeed obtained by minimizing the posterior expected loss (2.7), $\hat{\theta}_\pi(y) = \text{argmin}_a \; E_y L(a, \theta)$.

As seen in Chapter 2.3, this formula often leads to explicit expressions for the Bayes rules. In particular, if $L(a, \theta) = \|a - \theta\|_2^2$, the Bayes estimator is simply given by the mean of the posterior distribution, $\hat{\theta}_\pi(y) = E_\pi(\theta|y)$.

Uniqueness of the Bayes rule. The following sufficient condition is proved, for example, in Lehmann and Casella (1998, Corollary 4.1.4).

Proposition 4.2 *Suppose the loss function $L(a, \theta)$ is strictly convex in a. The Bayes estimator $\hat{\theta}_\pi$ is unique (a.e. P_θ for each θ) if both $B(\pi) < \infty$, and also a.s. P_π implies a.s. P_θ for each θ.*

Example. Univariate Gaussian. We revisit some earlier calculations to illustrate the two perspectives on Bayes risk. If $y|\theta \sim N(\theta, \epsilon^2)$ and the prior $\pi(d\theta)$ sets $\theta \sim N(0, \tau^2)$ then the posterior $\pi(d\theta|y)$ was found in Section 2.3 to be Gaussian with mean $\hat{\theta}_\pi(y) = \tau^2 y/(\tau^2+\epsilon^2)$ and posterior variance $\tau^2/(\tau^2+\epsilon^2)$ which is *linear* in y. From the frequentist perspective,

$$B(\pi_\tau) = \inf_{\hat{\theta}} B(\hat{\theta}, \pi_\tau),$$

and in the case of squared error loss, we know that the infimum exists among linear estimators $\hat{\theta}_c = cy$. Formula (2.34) showed that the risk of $\hat{\theta}_c$

$$r(\hat{\theta}_c, \theta) = c^2\epsilon^2 + (1-c)^2\theta^2$$

so that the integrated risk

$$B(\hat{\theta}_c, \pi_\tau) = c^2\epsilon^2 + (1-c)^2\tau^2.$$

Minimizing this over c yields the linear minimax choice $c_{\text{LIN}} = \tau^2/(\tau^2+\epsilon^2)$ as is of course expected from the posterior distribution calculation.

Remark 4.3 If $y|\theta \sim N(\theta, \epsilon^2)$, then the univariate MLE $\hat{\theta}_1(y) = y$ is admissible for squared error loss. For completeness, we indicate a proof. It suffices to take $\epsilon = 1$. The argument is by contradiction: supposing $\hat{\theta}_1$ inadmissible, we can find a dominating estimator $\tilde{\theta}$, whose risk function is necessarily continuous by Remark 2.4, so that there would exist $\delta > 0$ and an interval I of length $L > 0$ for which $r(\tilde{\theta}, \theta) \le 1 - \delta$ when $\theta \in I$. Now bring in the conjugate priors π_τ. From the example above, $1 - B(\pi_\tau) \sim \tau^{-2}$ as $\tau \to \infty$. However, the definition (4.1) of integrated risk implies that

$$1 - B(\tilde{\theta}, \pi_\tau) \ge \delta \pi_\tau(I) \sim c_0 \delta \tau^{-1}$$

as $\tau \to \infty$, with $c_0 = L/\sqrt{2\pi}$. Consequently, for τ large, we must have $B(\tilde{\theta}, \pi_\tau) < B(\pi_\tau)$, contradicting the very definition of the Bayes risk $B(\pi_\tau)$. Hence $\hat{\theta}_1$ must be admissible.

4.2 Bayes estimators for squared error loss

A number of formulas for Bayes estimators take especially convenient, even elegant, forms when squared error loss is used. Brown (1971) made remarkable use of the following simple identity.

Proposition 4.4 *Suppose that $L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_2^2$. For any estimator $\hat{\theta}$ and prior distribution $\pi(d\theta)$,*

$$B(\hat{\theta}, \pi) - B(\pi) = \int \|\hat{\theta} - \hat{\theta}_\pi\|^2 p. \tag{4.3}$$

Proof An alternative interpretation of the posterior mean arises by thinking of the Bayes risk in terms of the joint distribution πP_θ of (θ, y) :

$$B(\pi) = \inf\{E_{\pi P_\theta}\|\theta - \hat{\theta}(y)\|^2 : \hat{\theta}(y) \in L_2(P_\pi)\}.$$

Now $\hat{\theta}_\pi(y)$ can be viewed as the orthogonal projection of $\theta \in L_2(\pi P_\theta)$ on the closed linear subspace $L_2(P_\pi)$. Consequently, (4.3) is just the Pythagorean identity in $L_2(\pi P_\theta)$:

$$E_\pi E_\theta\|\theta - \hat{\theta}\|^2 = E_\pi E_\theta\|\theta - \hat{\theta}_\pi\|^2 + E_\pi E_\theta\|\hat{\theta}_\pi - \hat{\theta}\|^2. \qquad \square$$

Consider now the finite dimensional model $y \sim N_n(\theta, \epsilon^2 I)$. The posterior mean has representation

$$\hat{\theta}_\pi(y) = \int \theta\phi_\epsilon(y - \theta)\pi(d\theta) \Big/ \int \phi_\epsilon(y - \theta)\pi(d\theta). \tag{4.4}$$

Since $(\partial/\partial x_i)\phi_\epsilon(x) = -(x_i/\epsilon^2)\phi_\epsilon(x)$, we may write the numerator integrand as

$$\theta_i\phi_\epsilon(y - \theta) = y_i\phi_\epsilon(y - \theta) + \epsilon^2\frac{\partial}{\partial y_i}\phi_\epsilon(y - \theta),$$

leading to a representation of the Bayes estimator as a perturbation of the maximum likelihood rule

$$\hat{\theta}_\pi(y) = y + \epsilon^2\nabla \log p(y), \tag{4.5}$$

where the marginal density of y is $p(y) = \int \phi_\epsilon(y - \theta)\pi(d\theta)$, which is the convolution $\pi \star \Phi_\epsilon$.

Some Properties of Univariate Bayes Rules.

We apply Brown's identity and some facts about Fisher information, reviewed here and in Appendix **??** to obtain some useful bounds on Bayes risks. For the rest of this section, $n = 1$. If P is a probability measure on $\mathbb{R}$ with absolutely continuous density $p(x)dx$, the Fisher information is defined by

$$I(P) = \int \frac{p'(x)^2}{p(x)}dx.$$

This agrees with the definition of Fisher information for parametric families when $p(x; \theta) = p(x - \theta)$ is a location family. If $P_\tau(dx) = p(x/\tau)dx/\tau$ is a scaled version of p, then it is immediate that $I(P_\tau) = I(P_1)/\tau^2$.

The unbiased estimator $\hat{\theta}_0(y) = y$ has variance ϵ^2, and so $B(\hat{\theta}_0, \pi) = E_\pi E_\theta (y-\theta)^2 = \epsilon^2$, regardless of the prior π. Substituting $\hat{\theta}_0$ and formula (4.5) into (4.3), we have

$$\epsilon^2 - B(\pi) = \epsilon^4 \int \frac{p'(y)^2}{p(y)^2} p(y) dy.$$

Since p is the absolutely continuous density of the marginal distribution $\pi \star \Phi_\epsilon$, we arrive at a formula that is also sometimes called Brown's identity:

Corollary 4.5 *For $y \sim N(\theta, \epsilon^2)$ and squared error loss,*

$$B(\pi, \epsilon) = \epsilon^2 [1 - \epsilon^2 I(\pi \star \Phi_\epsilon)]. \tag{4.6}$$

Brown's identity (4.3) leads to an interesting formula for the directional or Gateaux derivative for the Bayes risk.

Lemma 4.6 *Given priors π_0 and π_1, let $\pi_t = (1-t)\pi_0 + t\pi_1$ for $0 \le t \le 1$. Then*

$$\frac{d}{dt} B(\pi_t)|_{t=0} = B(\hat{\theta}_{\pi_0}, \pi) - B(\pi_0). \tag{4.7}$$

Formula (4.7), which involves a "change of prior", should be compared with (4.3).

Proof Let $P_t = \Phi \star \pi_t$: since $I(P_t) < \infty$, the derivatives $p_t = (d/dy) P_t$ and p_t' exist for all y, and we put

$$\psi_0(y) = -(p_0'/p_0)(y) = y - \hat{\theta}_{\pi_0}(y),$$

where the final equality uses the Bayes estimator representation (4.5). From (4.6) and the derivative formula (C.16), we have

$$\frac{d}{dt} B(\pi_t)|_{t=0} = -\frac{d}{dt} I(P_t)|_{t=0} = \int [2\psi_0 p_1' + \psi_0^2 p_1] dy + I(P_0). \tag{4.8}$$

Observing that $p_1 = \phi \star \pi_1$ is the marginal density of π_1 and that $p_1'(y) = \int -(y-\theta)\phi(y-\theta)\pi_1(d\theta)$, we can write the previous integral as

$$\iint \left[-2(y - \hat{\theta}_{\pi_0})(y-\theta) + (y - \hat{\theta}_{\pi_0})^2\right] \phi(y-\theta)\pi_1(d\theta)dy$$
$$= -1 + E_{\pi_1} E_\theta (\theta - \hat{\theta}_{\pi_0})^2 = -1 + B(\hat{\theta}_{\pi_0}, \pi_1).$$

Recalling that $B(\pi_0) = 1 - I(P_0)$, we arrive at the formula (4.7). □

Now recall that Fisher information is bounded below by precision: for any distribution P,

$$I(P) \ge 1/\text{Var } P. \tag{4.9}$$

with equality if and only if P is Gaussian (see §4.11). Applying (4.9) to (4.6), we arrive at

Corollary 4.7

$$B(\pi, \epsilon) \le \frac{\epsilon^2 Var\ \pi}{\epsilon^2 + Var\ \pi}, \tag{4.10}$$

with equality if and only if π is Gaussian.

Finally, we give a lower bound for $B(\pi)$ that is sometimes easier to use than (4.6). It is essentially a version of the van Trees inequality (Van Trees, 1968) (see §4.11).

$$B(\pi, \epsilon) \geq \epsilon^2/(1 + \epsilon^2 I(\pi)). \tag{4.11}$$

Continuity of Bayes risks. The Fisher information representation of Corollary 4.5 can be used to show that the Bayes risk $B(\pi)$ is continuous in π. Note that definition (4.2) itself implies only upper semicontinuity for $B(\pi)$.

Lemma 4.8 *If π_n converges weakly to π, then $B(\pi_n) \to B(\pi)$.*

Proof It suffices to consider unit noise $\epsilon = 1$. Let $p_n(y) = \int \phi(y-\theta)\, d\pi_n$ and define $p(y)$ correspondingly from π. From (4.6), it is enough to show that

$$I(\pi_n \star \Phi) = \int \frac{p_n'^2}{p_n} \to \int \frac{p'^2}{p} = I(\pi \star \Phi). \tag{4.12}$$

Weak convergence says that $\int g\, d\pi_n \to \int g\, d\pi$ for every g bounded and continuous, and so p_n, p_n' and hence $p_n'^2/p_n$ converge respectively to p, p' and p'^2/p pointwise in $\mathbb{R}$. We construct functions G_n and G such that

$$0 \leq \frac{p_n'^2}{p_n} \leq G_n, \qquad 0 \leq \frac{p'^2}{p} \leq G,$$

and $\int G_n \to \int G$, and use the extended version of the dominated convergence theorem, Theorem C.6, to conclude (4.12). Indeed, from Brown's representation (4.5),

$$\frac{p_n'}{p_n}(y) = \hat{\theta}_{\pi_n}(y) - y = E_{\pi_n}[\theta - y|y],$$

and so $(p_n'/p_n)^2 \leq E_{\pi_n}[(\theta-y)^2|y]$, or equivalently

$$\left(\frac{p_n'}{p_n}\right)^2(y) \leq G_n(y) := \int (\theta-y)^2 \phi(y-\theta)\, \pi_n(d\theta).$$

A corresponding bound holds with π_n and p_n replaced by π and p and yields a bounding function $G(y)$. To complete the verification, note also that

$$\int G_n(y)\, dy = \iint (y-\theta)^2 \phi(y-\theta)\, dy\, \pi_n(d\theta) = 1 = \int G(y) dy. \qquad \square$$

Remark. The smoothing effect of the Gaussian density is the key to the convergence (4.12). Indeed, in general Fisher information is only lower semicontinuous: $I(\pi) \leq \liminf I(\pi_n)$. For a simple example in which continuity fails, take discrete measures π_n converging weakly to Φ, so that $I(\pi_n)$ is infinite for all n.

4.3 A lower bound for minimax risk

Recall from Section **??** the definition of the minimax risk over parameter set Θ:

$$R_N(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} r(\hat{\theta}, \theta).$$

There is an elementary, but very useful, lower bound for $R_N(\Theta)$ that may be derived using Bayes risks of priors supported in Θ. Indeed, if $\operatorname{supp}\pi \subset \Theta$, then

$$B(\hat{\theta}, \pi) = \int_\Theta r(\hat{\theta}, \theta)\pi(d\theta) \le \sup_{\theta\in\Theta} r(\hat{\theta}, \theta).$$

Minimizing over $\hat{\theta}$, we have

$$B(\pi) \le \inf_{\hat{\theta}} \sup_{\Theta} r(\hat{\theta}, \theta) = R_N(\Theta). \tag{4.13}$$

Define the worst-case Bayes risk over a collection $\mathcal{P}$ of probability measures as

$$B(\mathcal{P}) = \sup_{\pi\in\mathcal{P}} B(\pi). \tag{4.14}$$

Letting $\operatorname{supp}\mathcal{P}$ denote the union of all $\operatorname{supp}\pi$ for π in $\mathcal{P}$, we obtain the lower bound

$$\operatorname{supp}\mathcal{P} \subset \Theta \quad \Longrightarrow \quad R_N(\Theta) \ge B(\mathcal{P}). \tag{4.15}$$

Implicit in these remarks is a classical sufficient condition for minimaxity of an estimator $\hat{\theta}_0$: that there exist a sequence of priors π_n with $B(\pi_n) \to \bar{r} = \sup_\theta(\hat{\theta}_0, \theta)$. Indeed, from (4.13) we have $\bar{r} \le R_N(\Theta)$, which exactly says that $\hat{\theta}_0$ is minimax.

Corollary 4.9 *If $y|\theta \sim N(\theta, \epsilon^2)$, then $\hat{\theta}_1(y) = y$ is minimax for squared error loss. In addition, $\hat{\theta}_1$ is the unique minimax estimator.*

Proof Indeed, using the conjugate priors π_τ, we have $\bar{r}(\hat{\theta}_1) = \epsilon^2 = \lim_{\tau\to\infty} B(\pi_\tau)$. To establish uniqueness, suppose that $\hat{\theta}_1'$ is another minimax estimator with $P_\theta(\hat{\theta}_1 \ne \hat{\theta}_1') > 0$ for some and hence every θ. Then strict convexity of the loss function implies that the new estimator $\tilde{\theta} = (\hat{\theta}_1 + \hat{\theta}_1')/2$ satisfies, for all θ, $r(\tilde{\theta}, \theta) < (r(\hat{\theta}_1, \theta) + r(\hat{\theta}_1', \theta))/2 \le \epsilon^2$ which contradicts the admissibility of $\hat{\theta}_1$, Remark 4.3. □

Example 4.10 Bounded normal mean. Suppose that $y \sim N(\theta, 1)$ and that it is known *a priori* that $|\theta| \le \tau$, so that $\Theta = [-\tau, \tau]$. This apparently very special problem will be an important building block later in this chapter. We use the notation $\rho_N(\tau, 1)$ for the minimax risk $R_N(\Theta)$ in this case, in order to highlight the interval endpoint τ and the noise level, here equal to 1.

Let V_τ denote the prior on $[-\tau, \tau]$ having density $(3/(2\tau^3))(\tau - |\theta|)^2$; from the discussion above

$$\rho_N(\tau, 1) = \inf_{\hat{\theta}} \sup_{\theta\in[-\tau,\tau]} E(\hat{\theta} - \theta)^2 \ge B(V_\tau).$$

We use the van Trees inequality (4.11), along with $I(V_\tau) = I(V_1)/\tau^2$ to conclude that

$$\rho_N(\tau, 1) \ge \frac{1}{1 + I(V_\tau)} = \frac{\tau^2}{\tau^2 + I(V_1)}. \tag{4.16}$$

From this one learns that $\rho_N(\tau, 1) \nearrow 1$ as $\tau \to \infty$, indeed at rate $O(1/\tau^2)$. An easy calculation shows that $I(V_1) = 12$.

4.4 The Minimax Theorem

The minimax theorem of game and decision theory is a decisive tool in evaluating minimax risks, since it allows them to be calculated (or at least bounded) by finding the maximum Bayes risk over a suitable class of prior distributions. The resulting least favorable distribution and its associated Bayes estimator often give considerable insight into the estimation problem.

We state a version of the minimax theorem suited to the Gaussian sequence model. We defer to Appendix A a discussion of its assumptions and proof, and of its connections with the classical minimax theorems of game theory.

A function $f : T \to \mathbb{R}$ on a metric space T is lower semicontinuous at t if $f(t) \leq \liminf_{s \to t} f(s)$. The *action* a is typically an infinite sequence $a = (a_i) \in \mathbb{R}^\infty$. For technical reasons, we want to allow $a_i = \pm\infty$, and take the action space $\mathcal{A} = (\bar{\mathbb{R}})^\infty$, equipped with the topology of pointwise convergence: $a^n \to a$ if and only if $a_i^n \to a_i$ for each i.

Theorem 4.11 *Consider the Gaussian sequence estimation problem* (3.48) *and suppose that for each $\theta \in \ell_2(\mathbb{N}, \lambda)$ the loss function $L(a, \theta)$ is convex and lower semicontinuous in $a \in \mathcal{A}$. Let $B(\hat{\theta}, \pi)$ denote the integrated risk* (4.1)*. Let $\mathcal{P}$ be a convex set of probability measures on $\ell_2(\mathbb{N}, \lambda)$. Then*

$$\inf_{\hat{\theta}} \sup_{\pi \in \mathcal{P}} B(\hat{\theta}, \pi) = \sup_{\pi \in \mathcal{P}} \inf_{\hat{\theta}} B(\hat{\theta}, \pi) = B(\mathcal{P}) \tag{4.17}$$

A maximising π is called a least favorable distribution (with respect to $\mathcal{P}$).

Remarks 1. A pair $(\hat{\theta}^*, \pi^*)$ is called a *saddlepoint* if for all $\hat{\theta}$, and all $\pi \in \mathcal{P}$,

$$B(\hat{\theta}^*, \pi) \leq B(\hat{\theta}^*, \pi^*) \leq B(\hat{\theta}, \pi^*).$$

If a saddlepoint exists, then $\hat{\theta}^*$ is a Bayes rule for π^* (from the right side), and π^* is a least favorable distribution (since the left side implies $B(\pi) \leq B(\pi^*)$ for all π). See Figure 1.7. Proposition 4.13 below gives one setting in which a saddlepoint is guaranteed.

2. *Upper bound for $R_N(\Theta)$.* Let δ_θ denote a point probability mass concentrated at θ. Then we may rewrite $r(\hat{\theta}, \theta)$ as $B(\hat{\theta}, \delta_\theta)$. If Θ is a parameter space and $\mathcal{P}$ contains all point probability masses $\delta_\theta, \theta \in \Theta$, then clearly

$$\sup_{\theta \in \Theta} r(\hat{\theta}, \theta) \leq \sup_{\pi \in \mathcal{P}} B(\hat{\theta}, \pi),$$

and so minimizing over all estimators $\hat{\theta}$ and using the minimax theorem (4.17) gives an upper bound on minimax risk that we will use frequently:

$$R_N(\Theta) \leq B(\mathcal{P}). \tag{4.18}$$

The bound is useful because the Bayes-minimax risk $B(\mathcal{P})$ is often easier to evaluate than the minimax risk $R_N(\Theta)$. We can often show that the two are comparable in the low noise limit:

$$R_N(\Theta, \epsilon) \sim B(\mathcal{P}, \epsilon)$$

as $\epsilon \to 0$ (see Section 4.10).

3. In some cases, we may combine the lower and upper bounds (4.15) and (4.18). For example, if $\mathcal{P} = \mathcal{P}(\Theta) = \{\pi : \text{supp } \pi \subset \Theta\}$, then

$$R_N(\Theta) = B(\mathcal{P}(\Theta)). \tag{4.19}$$

Example 4.10 continued. In the bounded normal mean problem of the last section, we have $\Theta = [-\tau, \tau]$ and so

$$\rho_N(\tau, 1) = \sup\{B(\pi) \ : \ \text{supp}\, \pi \subset [-\tau, \tau]\}. \tag{4.20}$$

Remarks. 4. The weakening of continuity to lower semicontinuity seems necessary: even in dimension one with quadratic loss and $\epsilon = 1$, one checks that the (otherwise absurd) estimator $\hat{\theta}(y) = e^{y^2/4}/(1 + y)I\{y > 0\}$ has a risk function which is discontinuous at 0, but still lower semicontinuous. The assumption of lower semicontinuity allows all estimators to be included in statements such as (4.17).

5. It is easy to check that the loss functions $\|a - \theta\|_p^p$ are lower semicontinuous in a : if $a_i^{(n)} \to a_i^{(\infty)}$ for all i, then $\|a^{(\infty)} - \theta\|_p^p \le \liminf_n \|a^{(n)} - \theta\|_p^p$. See also Exercise 4.5.

Univariate Bayes Minimax Problems

Suppose that $\mathcal{P} \subset \mathcal{M}(\mathbb{R})$ is a convex set of probability measures. From the Fisher information representation (4.6).

$$B(\mathcal{P}) = \sup_{\pi \in \mathcal{P}} B(\pi) = 1 - \inf_{P \in \mathcal{P}^\star} I(P), \tag{4.21}$$

where $\mathcal{P}^\star = \{\Phi \star \pi, \pi \in \mathcal{P}\}$. We can now exploit properties of Fisher information $I(P)$, reviewed in Appendix **??**, to understand better the Bayes minimax problem $B(\mathcal{P})$. We take advantage also of the fact that convolution with the normal distribution makes every $P \in \mathcal{P}^\star$ smooth. The results find application in Sections 4.6, 8.7 and 13.2.

Let $\check{\pi}$ be the distribution of $-\theta$ when $\theta \sim \pi$; call $\mathcal{P}$ *symmetric* if $\pi \in \mathcal{P}$ implies $\check{\pi} \in \mathcal{P}$.

Proposition 4.12 *If $\mathcal{P} \subset \mathcal{M}(\mathbb{R})$ is convex and weakly compact, then there is a unique least favorable distribution $\pi_0 \in \mathcal{P}$. If $\mathcal{P}$ is symmetric, then so is π_0.*

Proof Since $B(\pi)$ is weakly upper semicontinuous on a weakly compact set $\mathcal{P}$, it attains its maximum at some π_0, and correspondingly $P_0 = \Phi \star \pi_0$ minimizes $I(P)$ over $\mathcal{P}^\star$. Since $p_0 = \phi \star \pi_0$ is positive on all of $\mathbb{R}$, we conclude from C.14 that P_0 is the unique minimizer of $I(P)$ on $\mathcal{P}^\star$, so that π_0 is also unique. Since $I(\check{\pi} \star \Phi) = I(\pi \star \Phi)$ for any π, we conclude from the uniqueness just shown that if $\mathcal{P}$ is symmetric, so must be π_0. □

Remark. For Section 8.7 we need an extension of Proposition 4.12. Let $\mathcal{P}_+(\mathbb{R})$ denote the collection of (sub-stochastic) measures π on $\mathbb{R}$ with $0 < \pi(\mathbb{R}) \le 1$, endowed with the topology of vague convergence, C.13. Then Proposition 4.12 also holds if $\mathcal{P} \subset \mathcal{P}_+(\mathbb{R})$ is convex and vaguely compact. The same proof works, since $I(P)$ is vaguely upper semicontinuous, and as $\pi_0(\mathbb{R}) > 0$, we have $p_0 > 0$ on all of $\mathbb{R}$.

Finally, we show that a least favorable distribution generates a saddle point in the Bayes minimax problem.

Proposition 4.13 *Let $\mathcal{P} \subset \mathcal{M}(\mathbb{R})$ be convex and suppose that $\pi_0 \in \mathcal{P}$ is least favorable. Then the corresponding Bayes rule $\hat{\theta}_{\pi_0}$ satisfies*

$$B(\hat{\theta}_{\pi_0}, \pi) \le B(\pi_0) \qquad \text{for all } \pi \in \mathcal{P}, \tag{4.22}$$

so that $(\hat{\theta}_{\pi_0}, \pi_0)$ is a saddle point for the Bayes minimax problem.

Proof If π_0 and $\pi_1 \in \mathcal{P}$ are given, then convexity of $\mathcal{P}$ says that $\pi_t = (1-t)\pi_0 + t\pi_1$ also belongs to $\mathcal{P}$ for $0 \le t \le 1$. Since $B(\pi)$ is concave on $\mathcal{P}$, a distribution π_0 is least favorable if and only if $(d/dt)B(\pi_t)|_{t=0} \le 0$ for each $\pi_1 \in \mathcal{P}$. The desired inequality (4.22) is now immediate from the Gateaux derivative formula (4.7). □

4.5 Product Priors and Spaces

Suppose that the coordinates θ_i of θ are gathered into groups: $\theta = (\theta_j, j \in J)$ for some finite or infinite set J. The θ_j may just be the individual components of θ, or they may consist of blocks of individual coefficients. For example, in a wavelet decomposition, we re-index the individual coordinates as θ_{jk} and θ_j may, for example, represent $(\theta_{jk}, k = 1, \dots, 2^j.)$

Suppose that the prior π makes the groups independent: $\pi(d\theta) = \prod_j \pi_j(d\theta_j)$. In (2.12) we saw that the posterior factorizes, and if in addition the loss function is additive, (2.13), then the Bayes rule is separable (2.14). In such cases, the risk functions are additive

$$r(\hat{\theta}_\pi, \theta) = \sum_j EL(\hat{\theta}_{\pi_j}(y_j), \theta_j) = \sum_j r(\hat{\theta}_{\pi_j}, \theta_j) \tag{4.23}$$

and in consequence, so are the Bayes risks

$$B(\pi) = \int r(\hat{\theta}_\pi, \theta)\pi(d\theta) = \sum_j B(\pi_j). \tag{4.24}$$

Independence is less favorable. Here is a trick that often helps in finding least favorable priors. Let π be an arbitrary prior, so that the θ_j are not necessarily independent. Denote by π_j the marginal distribution of θ_j. Build a new prior $\bar{\pi}$ by making the θ_j independent: $\bar{\pi} = \prod_j \pi_j$. This product prior is more difficult, as measured in terms of Bayes risk.

Lemma 4.14 $B(\bar{\pi}) \ge B(\pi)$.

Proof Because of the independence structure, the $\bar{\pi}-$posterior distribution of θ_j given y in fact depends only on y_j—compare (2.12). Hence the $\bar{\pi}$–Bayes rule is separable: $\hat{\theta}_{\bar{\pi},j}(y) = \hat{\theta}_{\pi_j}(y_j)$. From the additivity of losses and independence of components given θ, (4.23),

$$r(\hat{\theta}_{\bar{\pi}}, \theta) = \sum_j r(\hat{\theta}_{\bar{\pi},j}, \theta_j).$$

The π-average of the rightmost term therefore depends only the marginals π_j, so

$$\int r(\hat{\theta}_{\bar{\pi}}, \theta)\pi(d\theta) = \int r(\hat{\theta}_{\bar{\pi}}, \theta)\bar{\pi}(d\theta) = B(\bar{\pi}).$$

The left side is just $B(\hat{\theta}_{\bar{\pi}}, \pi)$, which is at least as large as $B(\pi)$ by definition. □

To see more intuitively why the product marginal prior $\bar{\pi}$ is harder then π, consider squared error loss: conditioning on all of y has to be better—lower variance—than conditioning on just y_j:

$$E_\pi[E_\pi(\theta_j|y) - \theta_j]^2 = E_\pi \mathrm{Var}\,(\theta_j|y)$$
$$\leq E_\pi \mathrm{Var}\,(\theta_j|y_j) = E_\pi[E_\pi(\theta_j|y_j) - \theta_j]^2.$$

Product Spaces. Suppose that $\Theta \subset \ell_2(I)$ is a product space $\Theta = \prod_{j\in J}\Theta_j$. The index j may refer to individual coordinates of $\ell_2(I)$, but in some cases each j may represent a cluster of coordinates (for example, in wavelet bases, all coefficients at a fixed scale.) If the loss function is additive and convex, then the minimax risk for Θ can be built from the minimax risk for each of the subproblems Θ_j.

Proposition 4.15 *Suppose that $\Theta = \prod_{j\in J}\Theta_j$ and $L(a,\theta) = \sum L_j(a_j,\theta_j)$. Suppose that $a_j \to L_j(a_j,\theta_j)$ is convex and lower semicontinuous for each θ_j. Then*

$$R_N(\prod_j \Theta_j, \epsilon) = \sum_j R_N(\Theta_j, \epsilon). \tag{4.25}$$

If $\theta_j^*(y_j)$ is separately minimax for each Θ_j, then $\theta^*(y) = \left(\theta_j^*(y_j)\right)$ is minimax for Θ.

Remarks: 1. There *is* something to prove here: among estimators $\hat{\theta}$ competing in the left side of (4.25), each coordinate $\hat{\theta}_j(y)$ may depend on *all* components $y_j, j \in J$. The result says that a minimax estimator need not exhibit such dependencies since $\theta_j^*(y)$ depends only on y_j.

2. The statement of this result does not involve prior distributions, and yet the simplest proof seems to need priors and the minimax theorem. A direct proof without priors is possible, but is more intricate.

Proof By the minimax theorem (4.11):

$$R_N(\Theta) = \sup\{B(\pi), \pi \in \mathcal{P}(\Theta)\},$$

where $\mathcal{P}(\Theta)$ denotes the collection of all probability measures supported in Θ. Given any such prior π, construct a new prior $\bar{\pi}$ as the product of the marginal distributions π_j of θ_j under π. Lemma 4.14 shows that $\bar{\pi}$ is more difficult than π : $B(\bar{\pi}) \geq B(\pi)$. Because of the product structure of Θ, each π_j is supported in Θ_j and $\bar{\pi}$ still lives on Θ. Thus the maximization can be restricted to priors with independent coordinates. Bayes risk is then additive, by (4.24), so the optimization can be term-by-term:

$$R_N(\Theta) = \sum_j \sup\{B(\pi_j) : \pi_j \in \mathcal{P}(\Theta_j)\} = \sum_j R_N(\Theta_j).$$

The verification that separately minimax $\theta_j^*(y_j)$ combine to yield a minimax $\theta^*(y)$ can now be left to the reader. □

4.6 Single Bounded Normal Mean

In this section and the next two, we confine attention to squared error loss.

If $y \sim N(\theta, \epsilon^2)$ and there is no constraint on θ, then we have seen, for example at (2.37),

that the minimax mean squared error for estimation of θ based on y equals the variance ϵ^2. Suppose now that θ is known to lie in a *bounded* interval of length 2τ, which without any real loss of generality we may assume to be centered about 0, so that we write $\Theta(\tau) = [-\tau, \tau]$. It is clear that any estimator $\hat{\theta}$, whether linear or not, can be improved simply by enforcing the interval constraint: if $\tilde{\theta} = [\hat{\theta}]_{-\tau}^{\tau} = \max\{\min\{\hat{\theta}, \tau\}, -\tau\}$, then $r(\tilde{\theta}, \theta) \leq r(\hat{\theta}, \theta)$. This section asks how much better is the nonlinear minimax risk

$$\rho_N(\tau, \epsilon) = \inf_{\hat{\theta}} \sup_{\theta \in [-\tau, \tau]} E_\theta(\hat{\theta} - \theta)^2 \tag{4.26}$$

than $\rho_N(\infty, \epsilon) = \epsilon^2$ and than the corresponding *linear* minimax risk $\rho_L(\tau, \epsilon)$ obtained by restricting $\hat{\theta}$ to linear estimators of the form $\hat{\theta}_c(y) = cy$?

Linear Estimators. Applying the variance-bias decomposition of mean squared error, (2.33), to a *linear* estimator $\hat{\theta}_c(y) = cy$, we obtain $E(\hat{\theta}_c - \theta)^2 = c^2\epsilon^2 + (1-c)^2\theta^2$. If the parameter is known to lie in a bounded interval $[-\tau, \tau]$, then the maximum risk occurs at the endpoints:

$$\sup_{\theta \in [-\tau, \tau]} E(\hat{\theta}_c - \theta)^2 = c^2\epsilon^2 + (1-c)^2\tau^2 = r(\hat{\theta}_c, \tau). \tag{4.27}$$

The minimax linear estimator is thus found by minimizing the quadratic function $c \to r(\hat{\theta}_c, \tau)$. It follows that

$$\rho_L(\tau, \epsilon) = \inf_c r(\hat{\theta}_c, \tau) = \frac{\epsilon^2\tau^2}{\epsilon^2 + \tau^2}. \tag{4.28}$$

The minimizer $c_* = \tau^2/(\epsilon^2 + \tau^2) \in (0, 1)$ and the corresponding minimax linear estimator

$$\hat{\theta}_{LIN}(y) = \frac{\tau^2}{\epsilon^2 + \tau^2} y. \tag{4.29}$$

Thus, if the prior information is that $\tau^2 \ll \epsilon^2$, then a large amount of linear shrinkage is indicated, while if $\tau \gg \epsilon^2$, then essentially the unbiased estimator is to be used.

Of course, $\hat{\theta}_{LIN}$ is also Bayes for a prior $\pi_\tau(d\theta) = N(0, \tau^2)$ and squared error loss. Indeed, from (2.16) we see that the posterior is Gaussian, with mean (4.29) and variance equal to the linear minimax risk (4.28)[1]. Note that this prior is *not* concentrated on $\Theta(\tau)$: only a moment statement is possible: $E_\pi\theta^2 = \tau^2$.

There is a simple but important scale invariance relation

$$\rho_L(\tau, \epsilon) = \epsilon^2\rho_L(\tau/\epsilon, 1). \tag{4.30}$$

Writing $\nu = \tau/\epsilon$ for the *signal-to-noise* ratio, we have

$$\rho_L(\nu, 1) = \nu^2/(1 + \nu^2) \sim \begin{cases} \nu^2 & \nu \to 0 \\ 1 & \nu \to \infty. \end{cases} \tag{4.31}$$

These results, however simple, are nevertheless a first quantitative indication of the importance of prior information, here quantified through ν, on possible quality of estimation.

[1] If $\hat{\theta}(y) = cy$ is a linear estimator that is Bayes for some prior $\pi(d\theta)$ under squared error loss, then it can be shown that the prior π is *necessarily* Gaussian. This property is a special case of a general phenomenon for exponential families: linear estimators are Bayes if and only if the prior comes from the conjugate prior family associated with that exponential family (Diaconis and Ylvisaker, 1979)

Projection Estimators. Orthogonal projections form an important and simple subclass of linear estimators. In one dimension the situation is almost trivial, with only two possibilities. Either $\hat{\theta}_0(y) \equiv 0$ with risk $r(\hat{\theta}_0, \theta) = \theta^2$—the pure bias case, or $\hat{\theta}_1(y) = y$, with risk $r(\hat{\theta}_1, \theta) = \epsilon^2$—the case of pure variance. Nevertheless, one can usefully define and evaluate the minimax risk over $\Theta = [-\tau, \tau]$ for projection estimators

$$\rho_P(\tau, \epsilon) = \inf_{c \in \{0,1\}} \sup_{\theta \in [-\tau,\tau]} E(\hat{\theta}_c - \theta)^2 = \min(\tau^2, \epsilon^2). \tag{4.32}$$

The choice is to "keep or kill": if the signal to noise ratio τ/ϵ exceeds 1, use $\hat{\theta}(y) = y$, otherwise use $\hat{\theta}(y) = 0$. The inequalities

$$\tfrac{1}{2}\min(\tau^2, \epsilon^2) \le \frac{\tau^2\epsilon^2}{\tau^2 + \epsilon^2} \le \min(\tau^2, \epsilon^2) \tag{4.33}$$

imply immediately that $\frac{1}{2}\rho_P(\tau, \epsilon) \le \rho_L(\tau, \epsilon) \le \rho_P(\tau, \epsilon)$, so that the best projection estimator is always within a factor of 2 of the best linear estimator.

Non-linear estimators. The non-linear minimax risk $\rho_N(\tau, \epsilon)$, (4.26), cannot be evaluated analytically in general. However the following properties are easy enough:

$$\rho_N(\tau, \epsilon) \le \rho_L(\tau, \epsilon), \tag{4.34}$$

$$\rho_N(\tau, \epsilon) = \epsilon^2 \rho_N(\tau/\epsilon, 1), \tag{4.35}$$

$$\rho_N(\tau, \epsilon) \text{ is increasing in } \tau, \tag{4.36}$$

$$\lim_{\tau \to \infty} \rho_N(\tau, \epsilon) = \epsilon^2. \tag{4.37}$$

Indeed (4.34) is plain since more estimators are allowed in the nonlinear competition, while (4.35) follows by rescaling, and (4.36) is obvious. Turning to (4.37), we recall that the classical result (2.37) says that the minimax risk for θ unconstrained to any interval, $\rho_N(\infty, \epsilon) = \epsilon^2$. Thus (4.37) asserts continuity as τ increases without bound—and this follows immediately from the example leading to (4.16): $\rho_N(\tau, 1) \ge \tau^2/(\tau^2 + I(V_1))$.

In summary so far, we have the bounds $\rho_N \le \rho_L \le \rho_P$, as illustrated in Figure 4.1, from which we might guess that the bounds are relatively tight, as we shall shortly see.

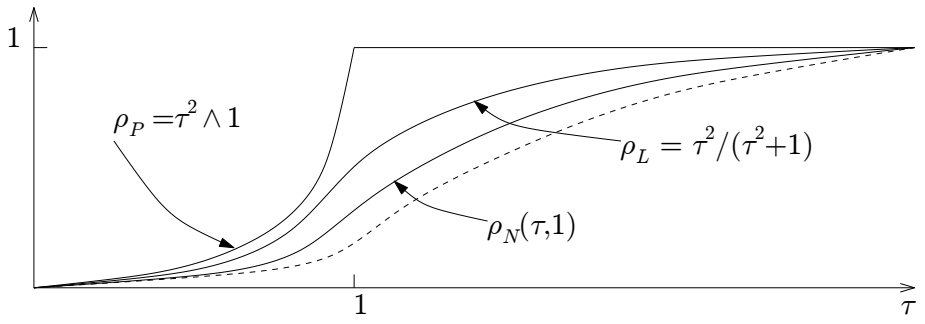

Figure 4.1 Schematic comparison of risk functions ρ_P, ρ_L and ρ_N, dotted line is the lower bound (4.16): $\rho_N(\tau, 1) \ge \tau^2/(\tau^2 + I(V_1)) = \tau^2/(\tau^2 + 12)$.

Near minimaxity of linear estimators.

In spite of the complex structure of non-linear minimax rules, it is remarkable that they do not, in this univariate setting, offer great improvements over linear estimators.

Theorem 4.16

$$\mu^* := \sup_{\tau,\epsilon} \frac{\rho_L(\tau,\epsilon)}{\rho_N(\tau,\epsilon)} \le 1.25. \tag{4.38}$$

Thus, regardless of signal bound τ and noise level ϵ, linear rules are within 25% of optimal for mean squared error. The bound $\mu^* < \infty$ is due to Ibragimov and Khas'minskii (1984). The extra work—some numerical—needed to obtain the essentially sharp bound 1.25 is outlined in Donoho et al. (1990) along with references to other work on the same topic.

Proof (partial.) We use projection estimators and the identity (2.24) for the two point priors $(1/2)(\delta_\tau + \delta_{-\tau})$ to give a short and instructive proof that $\mu^* \le 1/B(\pi_1)$. Numerical evaluation of the integral (2.24) shows the latter bound to be approximately 2.22.

First, it is enough to take $\epsilon = 1$, in view of the scaling invariances (4.30) and (4.35). We may summarize the argument by the inequalities:

$$\frac{\rho_L(\tau,1)}{\rho_N(\tau,1)} \le \frac{\tau^2 \wedge 1}{\rho_N(\tau,1)} \le \frac{1}{B(\pi_1)}. \tag{4.39}$$

Indeed, the first bound reflects a reduction to projection estimators, (4.32). For the second inequality, consider first $\tau \ge 1$, and use monotonicity (4.36) and the minimax risk lower bound (4.15) to obtain

$$\rho_N(\tau,1) \ge \rho_N(1,1) \ge B(\pi_1),$$

where $\pi_\tau = (1/2)(\delta_\tau + \delta_{-\tau})$ is the two point prior discussed above. Turning to $\tau \le 1$, we again note that $\rho_N(\tau,1) \ge B(\pi_\tau)$ and then from (2.24) that $\tau^2/B(\pi_\tau)$ is increasing in τ. □

An immediate corollary, using (4.28) and (4.33), is a bound for ρ_N:

$$(2\mu^*)^{-1} \min(\tau^2,\epsilon^2) \le \rho_N(\tau,\epsilon) \le \min(\tau^2,\epsilon^2). \tag{4.40}$$

The proof also gives sharper information for small and large τ: indeed, the linear minimax risk is then essentially equivalent to the non-linear minimax risk:

$$\mu(\tau) = \rho_L(\tau,1)/\rho_N(\tau,1) \to 1 \qquad \text{as } \tau \to 0,\infty. \tag{4.41}$$

Indeed, for small τ, the middle term of (4.39) is bounded by $\tau^2/B(\pi_\tau)$, which approaches 1, as may be seen from (2.24). For large τ, the same limit results from (4.37). Thus, as $\tau \to 0$, $\hat\theta_0(y) = 0$ is asymptotically optimal, while as $\tau \to \infty$, $\hat\theta(y) = y$ is asymptotically best. These remarks will play a role in the proof of Pinsker's theorem in the next chapter.

Least favorable priors are discrete.*

The fine structure of minimax rules is in general complicated, although some interesting and useful information is available. First, a property of analytic functions which plays a key role, both here and in Section 8.7.

Lemma 4.17 *Let ν be a probability measure and $K(\nu)$ the smallest interval containing the support of ν. Suppose that $r(\theta)$ is analytic on an open interval containing $K(\nu)$ and satsifies*

$$r(\theta) \le r_\nu = \int r(\theta')\nu(d\theta'), \qquad \theta \in K(\nu). \tag{4.42}$$

Then either $r(\theta)$ is constant on $K(\nu)$, or ν is a discrete measure whose support has no points of accumulation.

Proof Property (4.42) implies that the support of ν is contained in the set $\{\theta \in K(\nu) : r(\theta) = r_\nu\}$. Now we recall that if the set of zeros of an analytic function, here $r(\theta) - r_\nu$, has an accumulation point θ_0 inside its domain D, then it is identically zero on the connected component of D containing θ_0. □

Now to the minimax rules. Let $\bar{r}(\hat{\theta}) = \max_{|\theta|\le\tau} r(\hat{\theta}, \theta)$. Given a prior distribution π, let $M[\pi]$ denote the set of points where the Bayes rule for π attains its maximum risk:

$$M(\pi) = \left\{\theta \in [-\tau, \tau] \,:\, r(\hat{\theta}_\pi, \theta) = \bar{r}(\hat{\theta}_\pi)\right\}.$$

Proposition 4.18 *For the non-linear minimax risk $\rho_N(\tau, \epsilon)$ given by (4.26), a unique least favorable distribution π_τ exists and $(\hat{\theta}_\tau, \pi_\tau)$ is a saddlepoint. The distribution π_τ is symmetric, $supp(\pi_\tau) \subset M[\pi_\tau]$ and $M[\pi_\tau]$ is a finite set. Conversely, if a prior π satisfies $supp(\pi) \subset M[\pi]$ then $\hat{\theta}_\pi$ is minimax.*

Proof We apply Propositions 4.12 and 4.13 to the symmetric set $\mathcal{P}_\tau$ of probability measures supported on $[-\tau, \tau]$, which is weakly compact. Consequently a unique least favorable distribution $\pi_\tau \in \mathcal{P}_\tau$ exists, it is symmetric, and the corresponding Bayes rule $\hat{\theta}_\tau$ satisfies

$$r(\hat{\theta}_\tau, \theta) \le B(\pi_\tau) = \int r(\hat{\theta}_\tau, \theta)\pi_\tau(d\theta),$$

as we see by considering the point masses $\pi = \delta_\theta$ for $\theta \in [-\tau, \tau]$.

The risk function $\theta \to r(\hat{\theta}_\tau, \theta)$ is finite and hence analytic on $\mathbb{R}$, Remark 2.4 of Section 2.5, and not constant (Exercise 4.2). The preceding lemma shows that $\mathrm{supp}(\pi_\tau) \subset M(\pi_\tau)$, which can have no points of accumulation and (being also compact) must be a finite set.

Finally, if $\mathrm{supp}(\pi) \subset M[\pi]$, then $r(\hat{\theta}_\pi, \theta) = \bar{r}(\hat{\theta}_\pi)$ and so $\hat{\theta}_\pi$ must be minimax:

$$\bar{r}(\hat{\theta}_\pi) = B(\hat{\theta}_\pi, \pi) = \inf_{\hat{\theta}} B(\hat{\theta}, \pi) \le \inf_{\hat{\theta}} \bar{r}(\hat{\theta}).$$ □

In general, this finite set and the corresponding minimax estimator can only be determined numerically, see Kempthorne (1987); Donoho et al. (1990); Gourdin et al. (1994). Nevertheless, one can still learn a fair amount about these least favorable distributions. Since the posterior distribution of π_τ must also live on this finite set, and since the mean squared error of $\hat{\theta}_\tau$ must be everywhere less than ϵ^2, one guesses heuristically that the support points of π_τ will be spaced at a distance on the scale of the noise standard deviation ϵ.

For small τ, then, one expects that there will be only a small number of support points, and this was shown explicitly by Casella and Strawderman (1981). Their observation will be important for our later study of the least favorable character of sparse signal representations, so we outline the argument. Without loss of generality, set $\epsilon = 1$.

1. Proposition 4.18 says that the symmetric two point prior $\pi_\tau = (1/2)(\delta_\tau + \delta_{-\tau})$ is

minimax if $\{-\tau, \tau\} \subset M(\pi_\tau)$. For this two point prior, the posterior distribution and mean $\hat{\theta}_\tau$ were given in Chapter 2, (2.21) – (2.23), and we recall that the Bayes risk satisfies (2.24).

2. Since the posterior distribution concentrates on $\pm\tau$, one guesses from monotonicity and symmetry considerations that $M(\pi_\tau) \subset \{-\tau, 0, \tau\}$ for all τ. The formal proof uses a sign change argument linked to total positivity of the Gaussian distribution – see Casella and Strawderman (1981).

3. A second sign change argument shows that for $|\tau| < \tau_2$,

$$r(\hat{\theta}_\tau, 0) < r(\hat{\theta}_\tau, \tau).$$

Thus $\mathrm{supp}(\pi) = \{-\tau, \tau\} = M(\pi_\tau)$ and so $\hat{\theta}_\tau$ is minimax for $|\tau| < \tau_2$, and numerical work shows that $\tau_2 \doteq 1.057$.

This completes the story for symmetric two point priors. In fact, Casella and Strawderman go on to show that for $\tau_2 \le |\tau| < \tau_3$, an extra atom of the prior distribution appears at 0, and π_τ has the three-point form

$$\pi_\tau = (1-\alpha)\delta_0 + (\alpha/2)(\delta_\tau + \delta_{-\tau}).$$

This three point prior appears again in Chapters 8 and 13.

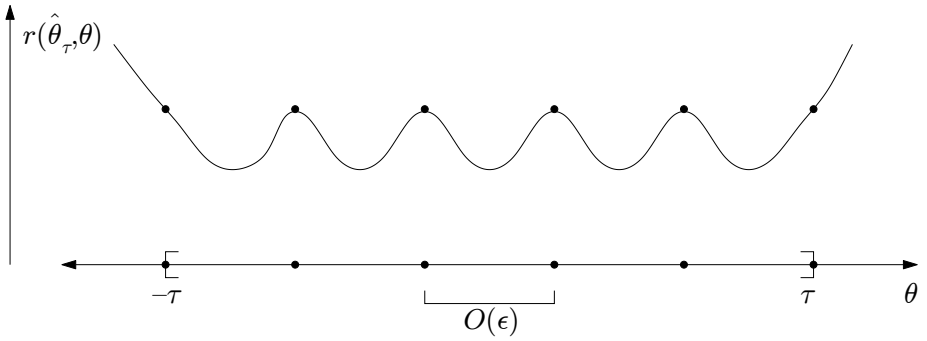

Figure 4.2 as the interval $[-\tau, \tau]$ grows, the support points of the least favorable prior spread out, and a risk function reminiscent of a standing wave emerges.

As $|\tau|$ increases, prior support points are added successively and we might expect a picture such as Figure 4.2 to emerge. Numerical calculations may be found in Gourdin et al. (1994). An interesting phenomenon occurs as τ gets large: the support points become gradually more spaced out. Indeed, if the least favorable distributions π_τ are rescaled to $[-1, 1]$ by setting $\nu_\tau(A) = \pi_\tau(\tau A)$, then Bickel (1981) derives the weak limit $\nu_\tau \Rightarrow \nu_\infty$, with

$$\nu_\infty(ds) = \cos^2(\pi s/2)ds, \tag{4.43}$$

for $|s| \le 1$, and shows that $\rho_N(\tau, 1) = 1 - \pi^2/\tau^2 + o(\tau^{-2})$ as $\tau \to \infty$.

4.7 Hyperrectangles

In this section, we 'lift' the results for intervals to hyperrectangles, and obtain some direct consequences for nonparametric estimation over Hölder classes of functions.

The set $\Theta \subset \ell_2(I)$ is said to be a hyperrectangle if

$$\Theta = \Theta(\tau) = \{\theta : |\theta_i| \le \tau_i \text{ for all } i \in I\} = \prod_i [-\tau_i, \tau_i].$$

For $\Theta(\tau)$ to be compact, it is necessary and sufficient that $\sum \tau_i^2 < \infty$, Example **??**. Algebraic and exponential decay provide natural examples for later use:

$$|\theta_k| \le Ck^{-\alpha}, \qquad k \ge 1, \alpha > 0, C > 0, \tag{4.44}$$

$$|\theta_k| \le Ce^{-ak}, \qquad k \ge 1, a > 0, C > 0. \tag{4.45}$$

We suppose that data y from the heteroscedastic Gaussian model (3.48) is observed, but for notational ease here, we set $\epsilon_i = \lambda_i \epsilon$, so that

$$y_i = \theta_i + \epsilon_i z_i, \qquad i \in I. \tag{4.46}$$

We seek to compare the linear and non-linear minimax risks $R_N(\Theta(\tau), \epsilon) \le R_L(\Theta(\tau), \epsilon)$. The notation emphasizes the dependence on scale parameter ϵ, for later use in asymptotics.

Proposition 4.15 says that the non-linear minimax risk over a hyperrectangle decomposes into the sum of the one-dimensional component problems:

$$R_N(\Theta(\tau), \epsilon) = \sum \rho_N(\tau_i, \epsilon_i). \tag{4.47}$$

Minimax *linear* estimators have a similar structure:

Proposition 4.19 *(i) If $\hat{\theta}_C(y) = Cy$ is minimax linear over hyperrectangles $\Theta(\tau)$, then necessarily C must be diagonal. (ii) Consequently,*

$$R_L(\Theta(\tau), \epsilon) = \sum_i \rho_L(\tau_i, \epsilon_i) \tag{4.48}$$

Before proving this, we draw an immediate and important consequence: by applying Theorem 4.16 term by term, $\rho_L(\tau_i, \epsilon_i) \le \mu^* \rho_N(\tau_i, \epsilon_i)$, it follows that the Ibragimov-Hasminski theorem lifts from intervals to hyperrectangles:

Corollary 4.20 *In model* (4.46)*,*

$$R_L(\Theta(\tau), \epsilon) \le \mu^* R_N(\Theta(\tau), \epsilon). \tag{4.49}$$

Proof of Proposition 4.19. First note that a diagonal linear estimator $\hat{\theta}_C(y) = (c_i y_i)$ has mean squared error of additive form:

$$r(\hat{\theta}_c, \theta) = \sum_i \epsilon_i^2 c_i^2 + (1 - c_i)^2 \theta_i^2. \tag{4.50}$$

Let $\bar{r}(\hat{\theta}_C) = \sup\{r(\hat{\theta}_C, \theta), \theta \in \Theta(\tau)\}$ and write $d(C) = \text{diag}(C)$ for the matrix obtained by setting the off-diagonal elements to 0. We show that this always improves the estimator over a hyperrectangle:

$$\bar{r}(\hat{\theta}_C) \ge \bar{r}(\hat{\theta}_{d(C)}). \tag{4.51}$$

Recall formula (3.51) for the mean squared error of a linear estimator. The variance term is easily bounded—with $\Delta = \mathrm{diag}(\epsilon_i^2)$, we have, after dropping off-diagonal terms,

$$\mathrm{tr}\, C^T \Delta C = \sum_{ij} c_{ij}^2 \epsilon_i^2 \geq \sum_i c_{ii}^2 \epsilon_i^2 = \mathrm{tr}\, d(C)^T \Delta d(C).$$

For the bias term, $\|(C-I)\theta\|^2$, we employ a simple but useful *random signs* technique. Let $\sigma \in \Theta(\tau) = \{(\pm\sigma_i)\}$ denote the *vertex set* of the corresponding hyperrectangle $\Theta(\sigma)$. Let π_σ be a probability measure that makes θ_i independently equal to $\pm\sigma_i$ with probability $1/2$. Then we may bound the maximum squared bias from below by an average, and then use $E\theta_i\theta_j = \sigma_i^2\delta_{ij}$ to obtain

$$\begin{aligned}\sup_{\theta\in V(\sigma)} \|(C-I)\theta\|^2 &\geq E\sum_{ij}(c_{ji}-\delta_{ji})^2\theta_i\theta_j \\ &= \sum_i (c_{ii}-1)^2\sigma_i^2 = \|(d(C)-I)\sigma\|^2.\end{aligned}$$

The risk of a diagonal linear estimator is identical at all the vertices of $V(\sigma)$—compare (4.50)—and so for all vertex sets $V(\sigma)$ we have shown that

$$\sup_{\theta\in V(\sigma)} r(\hat\theta_C, \theta) \geq \sup_{\theta\in V(\sigma)} r(\hat\theta_{d(C)}, \theta).$$

Since $\sigma \in \Theta(\tau)$ is arbitrary, we have established (4.51) and hence part (i).

Turning to part (ii), we may use this reduction to diagonal linear estimators to write

$$R_L(\Theta(\tau),\epsilon) = \inf_{(c_i)} \sup_{\theta\in\Theta(\tau)} \sum_i E(c_i y_i - \theta_i)^2.$$

Now, by the diagonal form $c_i y_i$ and the product structure of $\Theta(\tau)$, the infimum and the supremum can be performed term by term. Doing the supremum first, and using (4.27),

$$R_L(\Theta(\tau),\epsilon) = \inf_c r(\hat\theta_c, \tau). \tag{4.52}$$

Now minimizing over c, we get the right side of (4.49). □

Remarks. 1. It is evident from the proof that we only improve the maximum risk by restricting each c_i to the interval $[0,1]$.

2. For the admissibility result Theorem 2.3, all that was required was that a linear estimator be diagonal in *some* orthonormal basis. For minimaxity on a hyperrectangle $\Theta(\tau)$, which has product structure in a given basis, the estimator needs to be diagonal in *this* basis.

Hyperrectangles and discrete loss functions

Suppose again that $y_i \overset{\text{ind}}{\sim} N(\theta_i, \epsilon^2)$ for $i = 1,\dots,n$ and consider the product prior

$$\theta_i \overset{\text{ind}}{\sim} \tfrac{1}{2}(\delta_{\tau_i} + \delta_{-\tau_i}).$$

We take a brief break from squared error loss functions to illustrate the discussion of product priors, additive loss functions and posterior modes of discrete priors (cf. Section 2.3) in the

context of three related *discrete* loss functions

$$L_0(a,\theta) = \sum_i I\{a_i \neq \theta_i\},$$
$$N(a,\theta) = \sum_i I\{\operatorname{sgn} a_i \neq \operatorname{sgn} \theta_i\} \quad \text{and}$$
$$N_c(a,\theta) = I\{N(a,\theta) \geq c\}.$$

Here L_0 is counting error, while N counts *sign* errors and N_c, which is *not* additive, is the indicator of a tail event for N.

In each case, the Bayes rule for π, in accordance with (2.7), is found by minimizing, over a, the posterior expected loss. Since the prior has independent coordinates, so does the posterior, which is given by the noise level ϵ version of (2.21). Hence the distribution of θ_i given y is concentrated on $\pm\tau_i$, and by (2.22), it follows that for all three losses $E[L(a,\theta)|y]$ is minimized by the same Bayes rule

$$\hat{\theta}_{\pi,i}(y) = \tau_i \operatorname{sgn}(y_i),$$

and observe that

$$N(\hat{\theta}_\pi, \theta) = \sum_i I\{\operatorname{sgn} y_i \neq \operatorname{sgn} \theta_i\}$$

counts sign errors in the data.

Using the equivalent frequentist view of Bayes estimators, $B(\hat{\theta},\pi) \geq B(\hat{\theta}_\pi,\pi)$, cf. Section 4.1, we have therefore shown, using loss N_c as an example, that for all estimators $\hat{\theta}$, and in the joint distribution $\mathbb{P}$ of (θ, y), that

$$\mathbb{P}\{N(\hat{\theta},\theta) \geq c\} \geq \mathbb{P}\{N(\hat{\theta}_\pi,\theta) \geq c\}.$$

Consider now a hypercube situation, in which all $\tau_i \equiv \tau$. Then in the joint distribution $\mathbb{P}$, we have $N(\hat{\theta}_\pi,\theta) \overset{\mathcal{D}}{=} \mathrm{Bin}(n,\pi_1)$, where $\pi_1 = P\{N(\tau,\epsilon^2) < 0\} = \Phi(-\tau/\epsilon)$. Hence, for loss function N_c, the Bayes risk becomes a binomial probability tail event, $P\{\mathrm{Bin}(n,\pi_1) \geq c\}$.

These remarks will be used later for lower bounds in the optimal recovery approach to thresholding, Section 10.4.

Hyperrectangles and smoothness.

If (θ_i) represent the coefficients of a function f in an appropriate orthonormal basis, then the rate of decay of τ_i in a hyperrectangle condition can correspond to smoothness information about f. For periodic functions on $[0,1]$, the Fourier basis is natural. If f is C^α, in the sense of Hölder continuity (see Appendix C.16), then the Fourier coefficients satisfy (4.44) for some constant C (e.g. Katznelson (1968, p. 25) for α integer-valued and Zygmund (1959, p. 46) for $0 < \alpha < 1$.) However, the converse fails, so Fourier hyperrectangles do not exactly capture Hölder smoothness. On the other hand, a periodic function f is analytic if and only if there exist positive constants C and a so that (4.45) holds (e.g. Katznelson (1968, p. 26)). However, analyticity conditions are less often used in nonparametric theory than are constraints on a finite number of derivatives.

From this perspective, the situation is much better for wavelet bases, to be discussed in Chapter 7 and Appendix B, since Hölder smoothness is exactly characterized by hyperrectangle conditions, at least for non-integer α.

To describe this, we introduce doubly indexed vectors (θ_{jk}) and hyperrectangles of the form

$$\Theta^{\alpha}_{\infty}(C) = \{(\theta_{jk}) : |\theta_{jk}| \le C 2^{-(\alpha+1/2)j},\ j \in \mathbb{N}, k = 1, \dots, 2^j\}. \tag{4.53}$$

Let (θ_{jk}) for $j \ge 0$ and $k = 1, \dots, 2^j$ be the coefficients of f in an orthonormal wavelet basis for $L_2[0,1]$ of regularity $m > \alpha$. Then, according to Remark 9.5, f is $C^{\alpha}, \alpha \notin \mathbb{N}$, if and only if for some constant C, the coefficients $(\theta_{jk}) \in \Theta^{\alpha}_{\infty}(C)$ defined in (4.53). The subscript ∞ indicates that the bounds hold for all (j,k) and emphasizes that Hölder continuity measures uniform smoothness.

Proposition 4.21 *Assume a Gaussian white noise model $y_{jk} = \theta_{jk} + \epsilon z_{jk}$, with θ assumed to belong to a Hölder ball $\Theta^{\alpha}_{\infty}(C)$ defined at* (4.53)*. Then*

$$R_N(\Theta^{\alpha}_{\infty}(C), \epsilon) \asymp C^{2(1-r)}\epsilon^{2r}, \qquad r = 2\alpha/(2\alpha+1). \tag{4.54}$$

The notation shows the explicit dependence on both C and ϵ. The expression $a(\epsilon) \asymp b(\epsilon)$ means that there exist positive constants $\gamma_1 < \gamma_2$ depending only on α, but not on C or ϵ, such that for all ϵ, we have $\gamma_1 \le a(\epsilon)/b(\epsilon) \le \gamma_2$. The constants γ_i may not be the same at each appearance of $\asymp$.

While the wavelet interpretation is not needed to state and prove this result (which is why it can appear in this chapter!) its importance derives from the smoothness characterization. Indeed, this result exhibits the same rate of convergence as we saw for *mean square* smoothness, i.e. for Θ an ellipsoid. Note that here we also have a lower bound.

Proof Using (4.47), we can reduce to calculations based on the single bounded normal mean problem:

$$R_N(\Theta, \epsilon) = \sum_j 2^j \rho_N(C 2^{-(\alpha+1/2)j}, \epsilon).$$

Using (4.40), we have $\rho_N(\tau, \epsilon) = \gamma(\tau^2 \wedge \epsilon^2)$, where $\gamma \in [1/(2\mu^*), 1]$. So let $j_* \in \mathbb{R}$ be the solution of

$$C 2^{-(\alpha+1/2)j_*} = \epsilon.$$

For $j < j_*$, the variance term ϵ^2 is active in the bound for ρ_N, while for $j > j_*$ it is the squared bias term $C 2^{-(2\alpha+1)j}$ which is the smaller. Hence, with $j_0 = [j_*]$,

$$R_N(\Theta, \epsilon) \asymp \sum_{j \le j_0} 2^j \epsilon^2 + C^2 \sum_{j > j_0} 2^{-2\alpha j}.$$

These geometric sums are dominated by their leading terms, multiplied by constants depending only on α. Consequently,

$$R_N(\Theta, \epsilon) \asymp 2^{j_*}\epsilon^2 + C^2 2^{-2\alpha j_*} \asymp C^{2/(2\alpha+1)}(\epsilon^2)^{2\alpha/(2\alpha+1)}$$

which becomes (4.54) on substituting for r. □

4.8 Orthosymmetry and Hardest rectangular subproblems

Although the minimax structure of hyperrectangles is, as we have just seen, essentially straightforward, it is a key tool for obtaining deeper results on minimax risks for more general sets satisfying certain symmetry and convexity properties that we now define.

Θ is said to be *solid* and *orthosymmetric* if $\theta \in \Theta$ and $|\xi_i| \le |\theta_i|$ for all i implies that $\xi \in \Theta$ also. If a solid, orthosymmetric Θ contains a point τ, then the same is true for the entire hyperrectangle that it defines: $\Theta(\tau) \subset \Theta$.

Examples of solid orthosymmetric sets:

- Sets defined by the contours of symmetric increasing functions. Thus, if ψ is increasing on $\mathbb{R}^+$, then $\{\theta : \sum a_i \psi(\theta_i^2) \le 1\}$ is solid and orthosymmetric.
- ℓ_p *bodies*: defined by $\sum_i a_i^p |\theta_i|^p \le C^p$ for $p > 0$, and
- *Besov bodies*: defined by $\sum_j 2^{jsq}(\sum_k |\theta_{jk}|^p)^{q/p} \le C^q$ for $0 < p, q \le \infty$, Section 9.6.

Since Θ contains $\Theta(\tau)$ for each $\tau \in \Theta$, it is clear that $R_N(\Theta) \ge R_N(\Theta(\tau))$. Consequently, a simple but often useful lower bound to the non-linear minimax risk is obtained by restricting attention to the *hardest rectangular subproblem* of Θ:

$$R_N(\Theta) \ge \sup_{\tau \in \Theta} R_N(\Theta(\tau)). \tag{4.55}$$

For linear estimation, we first observe that according to the proof of Proposition 4.19, the maximum risk of any linear estimator $\hat{\theta}_C$ over any hyperrectangle can be reduced by discarding off-diagonal terms. Since this is true for every hyperrectangle and Θ is orthosymmetric, we must have

$$R_L(\Theta) = \inf_c \sup_{\theta \in \Theta} r(\hat{\theta}_c, \theta). \tag{4.56}$$

Here $\hat{\theta}_c(y) = (c_i y_i)$ denotes a diagonal linear estimator with $c \in \ell_2(\mathbb{N}, (\epsilon_i^2))$,

Quadratic convexity. To fully relate the linear minimax risk of Θ to that of the rectangular subproblems $\Theta(\tau)$, we need an extra convexity property. Θ is said to be *quadratically convex* if $\Theta_+^2 = \{(\theta_i^2) : \theta \in \Theta\}$ is convex. Examples include sets of the form $\{\theta : \sum a_i \psi(\theta_i^2) \le 1\}$ for ψ a convex function. This makes it clear that quadratic convexity is a stronger property than ordinary (linear) convexity. Particular examples include

- ℓ_p *bodies*: for $2 \le p \le \infty$, and
- *Besov bodies*: for $2 \le p \le q \le \infty$.

Just as in (4.55) the *linear* minimax risk over Θ is clearly bounded below by that of the hardest rectangular subproblem. However, for quadratically convex Θ, the *linear* difficulties are actually *equal*:

Theorem 4.22 *(Donoho et al., 1990) If Θ is compact, solid orthosymmetric and quadratically convex, then*

$$R_L(\Theta) = \sup_{\tau \in \Theta} R_L(\Theta(\tau)). \tag{4.57}$$

Combining (4.57), (4.49) and (4.55), we immediately obtain a large class of sets for which the linear minimax estimator is almost as good as the non-linear minimax rule.

Corollary 4.23 *If Θ is compact, solid orthosymmetric and quadratically convex, then $R_L(\Theta) \le \mu^* R_N(\Theta)$.*

This collection includes ℓ_p bodies for $p \ge 2$ – and so certainly ellipsoids, solid spheres, etc. and the Besov bodies just discussed.

Proof of Theorem 4.22. First we observe that (4.57) can be formulated as a minimax theorem. Indeed, (4.56) displays the left side as an inf sup. Turning to the right side of (4.57), and again to the proof of Proposition 4.19, we find from (4.52) that

$$\sup_{\tau\in\Theta} R_L(\Theta(\tau)) = \sup_{\tau\in\Theta}\inf_c r(\hat\theta_c,\tau).$$

To prove equality of (4.56) and the last display, we will apply the Kneser-Kuhn minimax theorem (Corollary A.4) with payoff function

$$f(c,s) = \sum_i c_i^2\epsilon_i^2 + (1-c_i)^2 s_i.$$

Note that $r(\hat\theta_c,\theta) = f(c,\theta^2)$ where $\theta^2 = (\theta_i^2)$. Clearly f is convex-concave – indeed, even linear in the second argument. By Remark 1 following Proposition 4.19, we may assume that the vector $c \in \ell_2(\mathbb{N},(\epsilon_i^2)) \cap [0,1]^\infty$, while $s \in \Theta_+^2 \subset \ell_1$. The latter set is convex by assumption and ℓ_1-compact by the assumption that Θ is ℓ_2-compact. Finally, $f(c,s)$ is trivially ℓ_1-continuous in s for fixed c in $[0,1]^\infty$. □

Example. Let $\Theta_{n,2}(C)$ denote an ℓ_2 ball of radius C in $\mathbb{R}^n$: $\{\theta : \sum_1^n \theta_i^2 \le C^2\}$. Theorem 4.22 says, in the homoscedastic case $\epsilon_i \equiv \epsilon$, that

$$R_L(\Theta_{n,2}(C),\epsilon) = \sup\{\epsilon^2 \sum_1^n \frac{\tau_i^2}{\epsilon^2+\tau_i^2} : \sum_1^n \tau_i^2 \le C^2\},$$

and since $s \to s/(1+s)$ is concave, it is evident that the maximum is attained at the vector with symmetric components $\tau_i^2 = C^2/n$. Thus,

$$R_L(\Theta_{n,2}(C),\epsilon) = n\epsilon^2 \cdot \frac{C^2}{n\epsilon^2 + C^2}, \tag{4.58}$$

which grows from 0 to the unrestricted minimax risk $n\epsilon^2$ as the signal-to-noise ratio $C^2/n\epsilon^2$ increases from 0 to ∞.

While the norm ball in infinite sequence space, $\Theta_2(C) = \{\theta \in \ell_2 : \|\theta\|_2 \le C\}$ is *not* compact, the preceding argument does yield the lower bound

$$R_L(\Theta_2(C),\epsilon) \ge C^2,$$

which already shows that no linear estimate can be uniformly consistent as $\epsilon \to 0$ over all of $\Theta_2(C)$. Section 5.5 contains an extension of this result.

Remark. We pause to preview how the various steps taken in this chapter and the next can add up to a result of some practical import. Let $\hat\theta_{SS,\lambda}$ denote the periodic smoothing spline with regularization parameter λ in the white noise model, Section 3.4. If it is agreed to compare estimators over the mean square smoothness classes $\Theta^\alpha = \Theta_2^\alpha(C)$, cf §**??**,

Remark 1, it will turn out that one cannot improve very much over smoothing splines from the worst-case MSE point of view.

Indeed, borrowing some results from the next chapter (§5.1, §5.2), the best mean squared error for such a smoothing spline satisfies

$$R_{SS}(\Theta^\alpha, \epsilon) = \inf_\lambda \sup_{\theta \in \Theta^\alpha} r(\hat{\theta}_{SS,\lambda}, \theta; \epsilon) \le (1 + c(\alpha, \epsilon))\, R_L(\Theta^\alpha, \epsilon),$$

along with the bound $\lim_{\epsilon \to 0} c(\alpha, \epsilon) \le 0.083$ if $\alpha \ge 2$. In combination with this chapter's result bounding linear minimax risk by a small multiple of non-linear minimax risk, Corollary 4.23, we can conclude that

$$R_{SS}(\Theta_2^\alpha(C), \epsilon) \le (1.10)(1.25)\, R_N(\Theta_2^\alpha(C), \epsilon)$$

for all $\alpha \ge 2$ and at least all sufficiently small ϵ. Thus even arbitrarily complicated non-linear esimators cannot have worst-case mean squared error much smaller than that of the relatively humble linear smoothing spline.

4.9 Correlated Noise*

For this section we consider a modification of Gaussian sequence model (3.1),

$$y_i = \theta_i + \epsilon z_i, \qquad i \in \mathbb{N}, \qquad \mathrm{Cov}(z) = \Sigma, \tag{4.59}$$

in which the components z_i may be correlated. This will be of interest in the later discussion of linear inverse problems with a wavelet-vaguelette decomposition, Chapter 12.

Make the obvious extensions to the definition of minimax risk among all non-linear and among linear estimators. Thus, for example, $R_N(\Theta, \Sigma) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} EL(\hat{\theta}(y), \theta)$ when y follows (4.59). The first simple result captures the idea that adding independent noise can only make estimation harder. Recall the non-negative definite ordering of covariance matrices or operators: $\Sigma' \prec \Sigma$ means that $\Sigma - \Sigma'$ is non-negative definite.

Lemma 4.24 *Consider two instances of model* (4.59) *with* $\Sigma' \prec \Sigma$. *Suppose that the loss function* $a \to L(a, \theta)$ *is convex. Then*

$$R_N(\Theta, \Sigma') \le R_N(\Theta, \Sigma), \qquad \textit{and} \qquad R_L(\Theta, \Sigma') \le R_L(\Theta, \Sigma).$$

Proof A conditioning argument combined with Jensen's inequality is all that is needed. Indeed, let y follow (4.59) and in parallel let $y' = \theta + \epsilon z'$ with $\mathrm{Cov}(z') = \Sigma'$. Since $\Sigma \succ \Sigma'$, we can find a zero mean Gaussian vector w with covariance $\Sigma - \Sigma'$, independent of z', so that $y = y' + w$. Let $\hat{\theta}(y)$ be an arbitrary estimator for noise Σ; we claim that

$$\tilde{\theta}(y') = E_\theta[\hat{\theta}(y)|y'] = E[\hat{\theta}(y' + w)|y']$$

has risk function at least as good as $\hat{\theta}(y)$. Indeed, using convexity of the loss function,

$$E_\theta L(\tilde{\theta}(y'), \theta) \le E_\theta E[L(\hat{\theta}(y' + w), \theta)|y'] = E_\theta E_\theta[L(\hat{\theta}(y), \theta)|y'] = E_\theta L(\hat{\theta}(y), \theta).$$

Since this holds for arbitrary $\hat{\theta}$, the statement for nonlinear minimax risk follows.

If $\hat{\theta}(y) = Cy$ is linear, then $\tilde{\theta}(y') = E[C(y' + w)|y'] = Cy'$ is also linear and so the preceding display also establishes the linear minimax inequality result. □

Corollary 4.25 *In white noise model* (3.1)*, if* $\epsilon' \le \epsilon$*, then* $R_N(\Theta, \epsilon') \le R_N(\Theta, \epsilon)$.

When the noise is independent in each coordinate and Θ is orthosymmetric, we have seen at (4.56) that the minimax linear estimator can be found among diagonal estimators. When the noise is correlated, however, diagonal estimation can be quite poor. First some notation: for covariance matrix Σ, let $\Sigma_d = \mathrm{diag}(\Sigma)$ be the diagonal matrix with entries taken from the diagonal of Σ. When considering only *diagonal* linear estimators, $\hat{\theta}_{c,i}(y) = c_i y_i$, let

$$R_{DL}(\Theta, \Sigma) = \inf_c \sup_{\theta \in \Theta} r(\hat{\theta}_c, \theta).$$

Of course, $R_{DL}(\Theta, \Sigma) = R_{DL}(\Theta, \Sigma_d)$ since R_{DL} involves only the variances of z. Finally, let the correlation matrix corresponding to Σ be

$$\rho(\Sigma) = \Sigma_d^{-1/2} \Sigma \Sigma_d^{-1/2}.$$

Proposition 4.26 *Suppose that* y *follows the correlated Gaussian model* (4.59)*. Let* $\lambda_{\min}$ *denote the smallest eigenvalue of* $\rho(\Sigma)$*. Suppose that* Θ *is orthosymmetric and quadratically convex. Then*

$$R_L(\Theta, \Sigma) \le R_{DL}(\Theta, \Sigma) \le \lambda_{\min}^{-1} R_L(\Theta, \Sigma).$$

If Σ is diagonal, then $\lambda_{\min} = 1$ and $R_{DL} = R_L$. This happens, for example, in the Karhunen-Loève basis, Section 3.9. If Σ is near-diagonal—in a sense to be made more precise in Chapter 12—then not much is lost with diagonal estimators. For general Σ, it can happen that $\lambda_{\min}$ is small and the bound close to sharp, see the example below.

Proof Only the right hand side bound needs proof. It is easily verified that $\Sigma \succ \lambda_{\min} \Sigma_d$ and that $\lambda_{\min} \le 1$ and hence using Lemma 4.24 that

$$R_L(\Theta, \Sigma) \ge R_L(\Theta, \lambda_{\min} \Sigma_d) \ge \lambda_{\min} R_L(\Theta, \Sigma_d).$$

By (4.56) in the independent co-ordinates model, $R_L(\Theta, \Sigma_d) = R_{DL}(\Theta, \Sigma_d)$. But as noted above, $R_{DL}(\Theta, \Sigma_d) = R_{DL}(\Theta, \Sigma)$. □

Example 4.27 Consider a p-variate "intra-class" correlation model in which $z_k = \sigma\eta + w_k$ is built from a common variable η and from w_k, all assumed to be i.i.d $N(0, 1)$. Then one checks that $\Sigma_{jk} = \sigma^2 + \delta_{jk}$ and then that $\lambda_{\min}(\rho(\Sigma)) = 1/(\sigma^2 + 1)$.

Suppose that $\sigma^2 > 1$, and $\tau^2 = \tau_p^2 \to \infty$ but $\tau_p^2 = o(p)$. Then for the hypercube $\Theta(\tau) = \{\theta = \sum_1^p \theta_k e_k : |\theta_k| \le \tau\}$, it can be shown, Exercise 4.11, that

$$R_{DL}(\Theta(\tau)) \sim (\sigma^2 + 1) R_L(\Theta(\tau)), \tag{4.60}$$

as $p \to \infty$, so that the bound of Proposition 4.26 is essentially sharp.

4.10 The Bayes Minimax Method*

In this section we outline a general strategy for asymptotic evaluation of minimax risks $R_N(\Theta)$ that will be useful in several settings.

We start with an upper bound, for fixed ϵ, which is easy after exploiting the minimax theorem. Suppose that $L(\theta, a)$ is convex in a for each $\theta \in \ell_2$. Let $\mathcal{M}$ be a convex collection

of probability measures on $\ell_2(I)$ containing Θ in the sense that point masses $\delta_\theta \in \mathcal{M}$ for $\theta \in \Theta$. Then, as we have seen at (4.18) and (4.17),

$$R_N(\Theta, \epsilon) \le B(\mathcal{M}, \epsilon) = \sup_{\pi \in \mathcal{M}} B(\pi). \tag{4.61}$$

We call the right side the *Bayes-minimax* risk. Often $\mathcal{M}$ is defined by constraints on marginal moments and in general $\mathcal{M}$ will *not* be supported on Θ. For example, if $\Theta(C)$ is the ellipsoid defined by $\sum a_i^2\theta_i^2 \le C^2$, then $\mathcal{M}(C) = \{\pi(d\theta) : \sum a_i^2 E_\pi \theta_i^2 \le C^2\}$.

The idea is that a judiciously chosen relaxation of the constraints defining Θ may make the problem easier to evaluate, and yet still be asymptotically equivalent to Θ as $\epsilon \to 0$.

The main task, then, is to establish that $R_N(\Theta, \epsilon) \sim B(\mathcal{M}, \epsilon)$ as $\epsilon \to 0$.

(a) Basic Strategy. Suppose that one can find a sequence ν_ϵ *supported in* Θ, that is nearly least favorable: $B(\nu_\epsilon) \sim B(\mathcal{M}, \epsilon)$. Then asymptotic equivalence would follow from the chain of inequalities

$$B(\nu_\epsilon) \le R_N(\Theta, \epsilon) \le B(\mathcal{M}, \epsilon) \sim B(\nu_\epsilon). \tag{4.62}$$

(b) Asymptotic Concentration. Often it is inconvenient to work directly with priors supported on Θ. Instead, one may seek a sequence $\pi_\epsilon \in \mathcal{M}$ that is both asymptotically least favorable, $B(\pi_\epsilon) \sim B(\mathcal{M}, \epsilon)$ and eventually *concentrates* on Θ:

$$\pi_\epsilon(\Theta) \to 1. \tag{4.63}$$

If one then constructs the conditioned prior $\nu_\epsilon = \pi_\epsilon(\cdot|\Theta)$ and additionally shows that

$$B(\nu_\epsilon) \sim B(\pi_\epsilon), \tag{4.64}$$

then asymptotic equivalence follows by replacing the last similarity in (4.62) by $B(\mathcal{M}, \epsilon) \sim B(\pi_\epsilon) \sim B(\nu_\epsilon)$.

There are significant details to fill in, which vary with the specific application. We try to sketch some of the common threads of the argument here, noting that some changes may be needed in each setting. There is typically a nested *family* of minimax problems with parameter space $\Theta(C)$ depending on C, so that $C < C'$ implies that $\Theta(C) \subset \Theta(C')$. Often, but not always, C will be a scale parameter: $\Theta(C) = C\Theta(1)$. We assume also that the corresponding prior family is similarly nested. Let $R(C, \epsilon) \le B(C, \epsilon)$ denote the frequentist and Bayes minimax risks over $\Theta(C)$ and $\mathcal{M}(C)$ respectively. We exploit the nesting structure by taking π_ϵ as the least favorable prior for $B(\gamma C, \epsilon)$ for some $\gamma < 1$. Although π_ϵ will typically not live on $\Theta(\gamma C)$, it often happens that it is asymptotically concentrated on the larger set $\Theta(C)$.

We now give some of the technical details needed to carry out this heuristic. The setting is ℓ_2 loss, but the argument can easily be generalized, at least to other norm based loss functions. Since C remains fixed, set $\Theta = \Theta(C)$. Let π_ϵ be a prior distribution with $B(\pi_\epsilon) \ge \gamma B(\gamma C, \epsilon)$ and $\pi_\epsilon(\Theta) > 0$. Set $\nu_\epsilon = \pi_\epsilon(\cdot|\Theta)$, and let $\hat{\theta}_{\nu_\epsilon}$ be the Bayes estimator of θ for the conditioned prior ν_ϵ. The issue is to relate $B(\nu_\epsilon)$ to $B(\pi_\epsilon)$. From the frequentist definition of Bayes risk $B(\pi_\epsilon) \le B(\hat{\theta}_{\nu_\epsilon}, \pi_\epsilon)$, and so

$$\begin{aligned} B(\pi_\epsilon) &\le E_{\pi_\epsilon}\{||\hat{\theta}_{\nu_\epsilon} - \theta||^2 | \Theta\}\pi_\epsilon(\Theta) + E_{\pi_\epsilon}\{||\hat{\theta}_{\nu_\epsilon} - \theta||^2, \Theta^c\} \\ &\le B(\nu_\epsilon)\pi_\epsilon(\Theta) + 2E_{\pi_\epsilon}\{||\hat{\theta}_{\nu_\epsilon}||^2 + ||\theta||^2, \Theta^c\}. \end{aligned}$$

Since also $B(\nu_\epsilon) \le R(C,\epsilon)$, on putting everything together, we have

$$\gamma B(\gamma C,\epsilon) \le B(\pi_\epsilon) \le R(C,\epsilon)\pi_\epsilon(\Theta) + 2E_{\pi_\epsilon}\{\|\hat\theta_{\nu_\epsilon}\|^2 + ||\theta||^2, \Theta^c\}.$$

In summary, we now have a lower bound for the minimax risk.

Lemma 4.28 *Suppose that for each $\gamma < 1$ one chooses $\pi_\epsilon \in \mathcal{M}(\gamma C)$ such that, as $\epsilon \to 0$,*

$$B(\pi_\epsilon) \ge \gamma B(\gamma C,\epsilon), \tag{4.65}$$

$$\pi_\epsilon(\Theta) \to 1, \tag{4.66}$$

$$E_{\pi_\epsilon}\{\|\hat\theta_{\nu_\epsilon}\|^2 + ||\theta||^2, \Theta^c\} = o(B(\gamma C,\epsilon)). \tag{4.67}$$

Then for each such γ,

$$R(C,\epsilon) \ge \gamma B(\gamma C,\epsilon)(1+o(1)). \tag{4.68}$$

Often the function $B(\gamma C,\epsilon)$ will have sufficient regularity that one can easily show

$$\lim_{\gamma\nearrow 1}\liminf_{\epsilon\to 0}\frac{B(\gamma C,\epsilon)}{B(C,\epsilon)} = 1. \tag{4.69}$$

See, for example, Exercise 4.6 for the scale family case. In general, combining (4.68) with (4.69), it follows that $R(C,\epsilon) \sim B(C,\epsilon)$.

Remark. Versions of this approach appear

1. in the discussion of Pinsker's theorem, where Θ is an ellipsoid, Chapter 5,
2. in estimation of η-sparse signals, where Θ is an ℓ_0-ball, Chapter 8,
3. and of *approximately* sparse signals, where Θ is an ℓ_p ball, Chapter 13,
4. and estimation of functions with spatial inhomogeneity, in which Θ is a Besov ball, Chapter 14.

4.11 Further details.

Proof of (4.9): We may of course suppose that $I(P) < \infty$, which entails that the density p of P exists and is absolutely continuous, and permits integration by parts in the following chain:

$$1 = \int p(y)dy = -\int (y-\mu)p'(y)dy \le \int (y-\mu)^2 p(y)dy \int [p'(y)]^2/p(y)dy,$$

with equality if and only if

$$(p'/p)(y) = (\log p)'(y) = c(y-\mu).$$

Proof of (4.11)*:* (taken from (Belitser and Levit, 1995)). The argument is of the same flavor as the Fisher information bound (4.9). Of course, by scaling arguments, we may reduce

to $\epsilon = 1$. Let $A = \hat{\theta}(y) - \theta$; and $B = (\partial/\partial\theta)[\log \phi(y-\theta)p(\theta)]$. Then by Fubini's theorem,

$$\begin{aligned} E_\pi E_\theta AB &= \int\int (\hat{\theta}(y) - \theta)\ \ (\partial/\partial\theta)[\log \phi(y-\theta)p(\theta)]\ \ \phi(y-\theta)p(\theta)dyd\theta \\ &= \int\int (\hat{\theta}(y) - \theta)\ \ (\partial/\partial\theta)[\phi(y-\theta)p(\theta)]\ \ d\theta dy \\ &= \int dy \int \phi(y-\theta)p(\theta)d\theta = 1. \end{aligned}$$

Now apply the Cauchy-Schwartz inequality: we have

$$E_\pi E_\theta A^2 = B(\hat{\theta}, \pi) \qquad \text{and} \qquad E_\pi E_\theta B^2 = 1 + I(\pi),$$

and now minimizing over $\hat{\theta}$ establishes (4.11). We note that improved bounds on the Bayes risk are given by Brown and Gajek (1990).

4.12 Notes

Aside: The celebrated paper of Brown (1971) uses (4.3) and (4.5) to show that statistical admissibility of $\hat{\theta}_\pi$ is *equivalent* to the recurrence of the diffusion defined by $dX_t = \nabla \log p(X_t)dt + 2dW_t$. In particular the classical and mysterious Stein phenomenon, namely the inadmissibility of the maximum likelihood estimator $\hat{\theta}(y) = y$ in exactly dimensions $n \geq 3$, is identified with the transience of Brownian motion in $\mathbb{R}^n, n \geq 3$. See also Srinivasan (1973).

Brown et al. (2006) gives an alternative proof of the Bayes risk lower bound (4.11), along with many other connections to Stein's identity (2.42).

The primary reference for the second part of this chapter is Donoho et al. (1990), where Theorems 4.16, 4.22 and 9.3 (for the case $\epsilon_i \equiv \epsilon$) may be found. The extension to the heteroscedastic setting given here is straightforward. The short proof of Theorem 4.16 given here relies on a minimax theorem; Donoho et al. (1990) give a direct argument.

`[J and MacGibbon?]` A Bayesian version of the I-H bound is given by Vidakovic and Dasgupta (1996), who show that the linear Bayes minimax risk for all symmetric and unimodal priors on $[-\tau, \tau]$ as at most 7.4% worse than the exact minimax rule. `[make exercise?]`

It is curious that the limiting least favorable distribution (4.43) found by Bickel (1981), after the transformation $x = \sin(\pi s/2)$, becomes $(2/\pi)\sqrt{1-x^2}dx$, the Wigner semi-circular limiting law for the (scaled) eigenvalues of a real symmetric matrix with i.i.d. entries (e.g. Anderson et al. (2010, Ch. 2)). Local repulsion—of prior support points, and of eigenvalues—is a common feature.

Least favorable distributions subject to moment constraints for the single normal mean with known variance were studied by Feldman (1991) and shown to be either normal or discrete.

Levit (1980, 1982, 1985) and Berkhin and Levit (1980) developed a more extensive theory of *second* order asymptotic minimax estimation of a d-dimensional Gaussian mean. Quite generally, they showed that the second order coefficient (here π^2), could be interpreted as twice the principal eigenvalue of the Laplacian (here $= -2d^2/dt^2$) on the fundamental domain (here $[-1, 1]$), with the asymptotically least favorable distribution having density the square of the principal eigenfunction, here $\omega(t) = \cos(\pi t/2)$. We do not delve further into this beautiful theory since it is essentially parametric in nature: in the nonparametric settings to be considered in these notes, we are still concerned with understanding the *first* order behaviour of the minimax risk with noise level ϵ or sample size n.

Exercises

4.1 *(Less noise is easier.)* Consider two versions of the sequence model: $y = \theta + \epsilon z$, and a lower noise version, $y' = \theta + \epsilon' z'$, where $\epsilon' < \epsilon$. Suppose that for each $\theta \in \ell_2(I)$, the loss function $L(a, \theta)$ is convex in a.

(a) Show that for each estimator $\hat{\theta}(y)$, there is an estimator $\hat{\theta}'(y')$ such that for all θ,

$$r(\hat{\theta}', \theta; \epsilon') \le r(\hat{\theta}, \theta; \epsilon).$$

(b) Conclude that for any parameter space Θ, and for any set of priors $\mathcal{P}$,

$$R_N(\Theta, \epsilon') \le R_N(\Theta, \epsilon),$$
$$B(\mathcal{P}, \epsilon') \le B(\mathcal{P}, \epsilon).$$

4.2 *(Qualitative features of risk of proper Bayes rules.)* Suppose that $y \sim N(\theta, \epsilon^2)$, that θ has a proper prior distribution π, and that $\hat{\theta}_\pi$ is the squared error loss Bayes rule.
(a) Show that $r(\hat{\theta}_\pi, \theta)$ cannot be constant for $\theta \in \mathbb{R}$. [Hint: Corollary 4.9.]
(b) If $E_\pi|\theta| < \infty$, then $r(\hat{\theta}_\pi, \theta)$ is at most quadratic in θ: there exist constants a, b so that $r(\hat{\theta}_\pi, \theta) \le a + b\theta^2$. [Hint: apply the covariance inequality (C.6) to $E_\pi[|\theta - x|\phi(\theta - x)]$.
(c) Suppose in addition that π is supported in a bounded interval I. Show that $P_\theta(\hat{\theta}_\pi \in I) = 1$ for each θ and hence that $r(\hat{\theta}_\pi, \theta)$ is unbounded in θ, indeed $r(\hat{\theta}_\pi, \theta) \ge c\theta^2$ for suitable $c > 0$.

4.3 *(Fisher information for priors on an interval.)* (a) Consider the family of priors $\pi_\beta(d\theta) = c_\beta(1 - |\theta|)^\beta$. For what values of β is $I(\pi_\beta) \le \infty$?
(b) What is the minimum value of $I(\pi_\beta)$?
(c) Show that ν_∞ in (4.43) minimizes $I(\pi)$ among probability measures supported on $[-1, 1]$.

4.4 *(Truncation of (near) least favorable priors.)* (a) Given a probability measure $\pi(d\theta)$ on $\mathbb{R}$, and M sufficiently large, define the restriction to $[-M, M]$ by $\pi^M(A) = \pi(A||\theta| \le M)$. Show that π^M converges weakly to π as $M \to \infty$.
(b) If π satisfies $\int |\theta|^p d\pi \le \eta^p$, show that π^M does also, for $M \ge \eta$.
(c) Given a class of probability measures $\mathcal{P}$ and $\gamma < 1$, show using Lemma 4.8 that there exists $\pi \in \mathcal{P}$ and M large so that $B(\pi^M) \ge \gamma B(\mathcal{P})$.

4.5 *(continuity properties of ℓ_p loss.)* Consider the loss function $L(a, \theta) = \|a - \theta\|_p^p$ as a function of $\theta \in \ell_2(\mathbb{N})$. Show that it is continuous for $p \ge 2$, while for $p < 2$ it is lower semi-continuous but not continuous.

4.6 *(Scaling bounds for risks.)* Consider $y = \theta + \epsilon z$ and squared error loss. Suppose that $\{\Theta(C)\}$ is a scale family of parameter spaces in $\ell_2(I)$, so that $\Theta(C) = C\Theta(1)$ for $C > 0$.
(a) Use the abbreviation $R(C, \epsilon)$ for (i) $R_N(\Theta(C); \epsilon)$, and (ii) $R_L(\Theta(C); \epsilon)$. In each case, show that if $C' \le C$ and $\epsilon' \le \epsilon$, then

$$R(C, \epsilon) \le (C/C')^2(\epsilon/\epsilon')^2 R(C', \epsilon'),$$

and that if $\mathcal{P}(C) = C\mathcal{P}(1)$ is a scale family of priors, that the same result holds for $B(C, \epsilon) = B(\mathcal{P}(C); \epsilon)$.
(b) Conclude that

$$\lim_{\gamma \to 1} \liminf_{\epsilon \to 0} \frac{B(\gamma C, \epsilon)}{B(C, \epsilon)} = 1.$$

4.7 *(Two point priors.)* Suppose that $y \sim N(\theta, 1)$, and consider the symmetric two point prior

$\pi_\tau^{(2)} = (1/2)(\delta_\tau + \delta_{-\tau})$. Show that for squared error loss,

$$\pi(\{\tau\}|y) = e^{\tau y}/(e^{\tau y} + e^{-\tau y}),$$
$$\hat{\theta}_\pi(y) = E(\theta|y) = \tau \tanh \tau y,$$
$$E[(\theta - \hat{\theta}_\tau)^2|y] = \tau^2/\cosh^2 \tau y,$$
$$B(\pi_\tau) = \tau^2 e^{-\tau^2/2} \int \frac{\phi(y)dy}{\cosh y\tau}.$$

4.8 *(Bounded normal mean theory for L_1 loss.)* Redo the previous question for $L(\theta, a) = |\theta - a|$. In particular, show that

$$\hat{\theta}_\pi(y) = \tau \text{ sgn } y, \qquad \text{and} \qquad B(\pi_\tau) = 2\tau\tilde{\Phi}(\tau),$$

where, as usual $\tilde{\Phi}(\tau) = \int_\tau^\infty \phi(s)ds$. In addition, show that

$$\mu^* = \sup_{\tau,\epsilon} \frac{\rho_L(\tau,\epsilon)}{\rho_N(\tau,\epsilon)} \le \frac{1}{B(\pi_1)} \doteq 1/.32 < \infty.$$

Hint: show that $\rho_L(\tau, 1) \le \rho_P(\tau, 1) = \min(\tau, \sqrt{2/\pi})$.

4.9 *(Continued.)* For L_1 loss, show that (a) $\rho_N(\tau,\epsilon) = \epsilon\rho_N(\tau/\epsilon, 1)$ is increasing in τ, and (b) $\lim_{\tau\to\infty} \rho_N(\tau,\epsilon) = \epsilon\gamma_0$, where $\gamma_0 = E_0|z| = \sqrt{2/\pi}$.
[Hint: for (b) consider the uniform prior on $[-\tau, \tau]$.]

4.10 *(Translation invariance implies diagonal Fourier optimality.)* Signals and images often are translation invariant. To make a simplified one-dimensional model, suppose that we observe, in the "time domain", $x_k = \gamma_k + \sigma\eta_k$ for $k = 1, \ldots, n$. To avoid boundary effects, assume that x, γ and η are extended to periodic functions of $k \in \mathbb{Z}$, that is $x(k + n) = x(k)$, and so on. Define the *shift* of γ by $(S\gamma)_k = \gamma_{k+1}$. The set Γ is called *shift-invariant* if $\gamma \in \Gamma$ implies $S\gamma \in \Gamma$. Clearly, then, $S^l\gamma \in \Gamma$ for all $l \in \mathbb{Z}$.
(a) Show that $\Gamma = \{\gamma : \sum_{k=1}^n |\gamma_k - \gamma_{k-1}| < C\}$ is an example of a shift-invariant set. Such sets are said to have bounded total variation.
Now rewrite the model in the discrete Fourier domain. Let $\mathrm{e} = e^{2\pi i/n}$ and note that the discrete Fourier transform $y = \mathcal{F}x$ can be written

$$y_k = \sum_{l=0}^{n-1} \mathrm{e}^{kl} x_l, \qquad k = 0, \ldots, n-1.$$

Similarly, let $\theta = \mathcal{F}\gamma, z = \mathcal{F}\eta$ and $\Theta = \mathcal{F}\Gamma$.
(b) Show that shift-invariance of Γ means that $\theta = (\theta_k) \in \Theta$ implies $M^l\theta = (\mathrm{e}^{lk}\theta_k) \in \Theta$ for $l \in \mathbb{Z}$. In particular, we have $\mathcal{F}S = M^{-1}\mathcal{F}$.
(c) Let $V(\tau) = \{M^l\tau, l \in \mathbb{Z}\}$ denote the *orbit* of τ under the action of M. By using a random shift (i.e. l chosen at random from $\{0, \ldots, n-1\}$), modify the random signs method to show that

$$\sup_{\theta\in V(\tau)} r(\hat{\theta}_{C^0,0}, \theta) \le \sup_{\theta\in V(\tau)} r(\hat{\theta}_{C,b}, \theta).$$

Thus, on a translation invariant set Γ, an estimator that is minimax among affine estimators must have diagonal linear form when expressed in the discrete Fourier basis.

4.11 *(Linear and diagonal minimax risk in intra-class model.)*
Consider the setting of Example 4.27.
(a) Show that in the basis of the Karhunen-Loève transform, the variances are

$$\varepsilon_1^2 = p\sigma^2 + 1, \qquad \varepsilon_k^2 = 1, \quad k \geq 2.$$

(b) Show that $R_L(\Theta(\tau)) = \sum_i \epsilon_i^2\tau^2/(\epsilon_i^2+\tau^2)$, and $R_{DL}(\Theta(\tau)) = p(1+\sigma^2)\tau^2/(1+\sigma^2+\tau^2)$.
(c) Derive conclusion (4.60).

5

Linear Estimators and Pinsker's Theorem

Compared to what an ellipse can tell us, a circle has nothing to say. (E. T. Bell).

Under appropriate assumptions, linear estimators have some impressive optimality properties. This chapter uses the optimality tools we have developed to study optimal linear estimators over ellipsoids, which as we have seen capture the notion of mean-square smoothness of functions. In particular, the theorems of Pinsker (1980) are notable for several reasons. The first gives an exact evaluation of the linear minimax risk in the Gaussian sequence model for quadratic loss over general ellipsoids in ℓ_2. The second shows that in the low noise limit $\epsilon \to 0$, the non-linear minimax risk is actually equivalent to the linear minimax risk: in other words, there exist linear rules that are asymptotically efficient. The results applies to ellipsoids generally, and thus to all levels of Hilbert-Sobolev smoothness, and also to varying noise levels in the co-ordinates, and so might be considered as a crowning result for linear estimation.

The linear minimax theorem can be cast as a simple Lagrange multiplier calculation, Section 5.1. Section 5.2 examines some examples: in the white noise, ellipsoids of mean square smoothness and of analytic function, leading to very different rates of convergence (and constants!). Fractional integration is used as an example of the use of the linear minimax theorem for inverse problems. Finally, a concrete comparison shows that the right smoothing spline is actually very close in performance to linear minimax rule.

Section 5.3 states the "big" theorem on asymptotic minimax optimality of linear estimators among *all* estimators in the low noise limit. In this section we give a proof for the white noise model with polynomial ellipsoid constraints – this allows a simplified argument in which Gaussian priors are nearly least favorable. The Bayes rules for these Gaussian priors are linear, and are essentially the linear minimax rules, which leads to the asymptotic efficiency.

Section 5.4 gives the proof for the more general case, weaving in ideas from Chapter 4 in order to combine the Gaussian priors with other priors needed for co-ordinates that have especially 'large' or 'small' signal to noise ratios.

The chapter concludes with a diversionary interlude, Section 5.5, that explains why the infinite sequence model requires a compactness assumption for even as weak a conclusion as consistency to be possible in the low noise limit.

5.1 Exact evaluation of linear minimax risk.

In this chapter we consider the non-white Gaussian sequence model, (3.48), which for now we write in the form

$$y_i = \theta_i + \sigma_i z_i, \qquad i \in \mathbb{N}. \tag{5.1}$$

Suppose that Θ is an ellipsoid in $\ell_2(\mathbb{N})$:

$$\Theta = \Theta(a, C) = \{\theta : \sum a_i^2 \theta_i^2 \le C^2\}. \tag{5.2}$$

A pleasant surprise is that there is an explicit solution for the minimax linear estimator over such ellipsoids.

Proposition 5.1 *Suppose that the observations follow sequence model* (5.1) *and that* Θ *is an ellipsoid* (5.2)*. Assume that* a_i *are positive and nondecreasing with* $a_i \to \infty$*. Then the minimax linear risk*

$$R_L(\Theta) = \sum_i \sigma_i^2 (1 - a_i/\mu)_+, \tag{5.3}$$

where $\mu = \mu(C)$ *is determined by*

$$\sum \sigma_i^2 a_i (\mu - a_i)_+ = C^2. \tag{5.4}$$

The linear minimax estimator is given by

$$\hat{\theta}_i^*(y) = c_i y_i = (1 - a_i/\mu)_+ y_i, \tag{5.5}$$

and is Bayes for a Gaussian prior π_C *having independent components* $\theta_i \sim N(0, \tau_i^2)$ *with*

$$\tau_i^2 = \sigma_i^2 (\mu/a_i - 1)_+. \tag{5.6}$$

Some characteristics of the linear minimax estimator (5.5) deserve note. Since the ellipsoid weights a_i are increasing, the shrinkage factors c_i decrease with i and hence downweight the higher "frequencies" more. In addition, there is a *cutoff* at the first index i such that $a_i \ge \mu$: the estimator is zero at frequencies above the cutoff. Finally, the optimal linear estimator depends on all the parameters $C, (\sigma_i)$, and (a_i)—as they vary, so does the optimal estimator. In particular, the least favorable distributions, determined by the variances τ_i^2 change with changing noise level.

Proof The set Θ is solid, orthosymmetric and quadratically convex. Since $\sup a_i = \infty$ it is also compact. Thus the minimax linear risk is determined by the hardest rectangular subproblem, and from Theorem 4.22,

$$R_L(\Theta) = \sup_{\tau \in \Theta} R_L(\Theta(\tau)) = \sup \Big\{ \sum_i \frac{\sigma_i^2 \tau_i^2}{\sigma_i^2 + \tau_i^2} \ : \ \sum a_i^2 \tau_i^2 \le C^2 \Big\}. \tag{5.7}$$

This maximum may be evaluated by forming the Lagrangian

$$\mathcal{L} = \sum_i \Big\{ \sigma_i^2 - \frac{\sigma_i^4}{\sigma_i^2 + \tau_i^2} \Big\} - \frac{1}{\mu^2} \sum_i a_i^2 \tau_i^2.$$

Simple calculus shows that the maximum is attained at τ_i^2 given by (5.6). The positive part

constraint arises because τ_i^2 cannot be negative. The Lagrange multiplier parameter μ is uniquely determined by the equation $\sum a_i^2 \tau_i^2 = C^2$, which on substitution for τ_i^2 yields (5.4). This equation has a unique solution since the left side is a continuous, strictly increasing function of μ. The corresponding maximum is then (5.3) and the *linear minimax estimator*, recalling (4.29), is given by $\hat{\theta}_i^*(y) = c_i y_i$ with

$$c_i = \frac{\tau_i^2}{\sigma_i^2 + \tau_i^2} = \left(1 - \frac{a_i}{\mu}\right)_+.$$

From this, it is evident that $\hat{\theta}^*$ is Bayes for a prior with independent $N(0, \tau_i^2)$ components. □

5.2 Some Examples

Sobolev Ellipsoids.

Consider the white noise case, $\sigma_k^2 \equiv \epsilon^2$. Return to the Hilbert-Sobolev parameter space in the trigonometric basis[1] considered in Section **??**, and with the ellipsoid (3.4) with $a_{2k} = a_{2k-1} = (2k)^\alpha$ for $\alpha > 0$, and write Θ as $\Theta_2^\alpha(C)$. Let us rewrite the condition (5.4) that determines μ_ϵ as

$$\mu \sum_{k \in N} a_k - \sum_{k \in N} a_k^2 = C^2/\epsilon^2. \tag{5.8}$$

To describe the summation set N, observe that the weights $a_k \approx k^\alpha$ with relative error at most $O(1/k)$ and so

$$N = N(\mu) = \{k \,:\, a_k < \mu\} \approx \{k \,:\, k < \mu^{1/\alpha}\}.$$

Setting $k_\mu = [\mu^{1/\alpha}]$, we have the integral approximations

$$\sum_{k \in N} a_k^p \doteq \sum_{k=1}^{k_\mu} k^{\alpha p} \doteq \frac{\mu^{p+1/\alpha}}{p\alpha + 1}.$$

Substituting into (5.8) and solving for μ_ϵ, we obtain

$$\mu_\epsilon^{1/\alpha} \doteq \left(\frac{(\alpha+1)(2\alpha+1)}{\alpha}\frac{C^2}{\epsilon^2}\right)^{1-r}, \tag{5.9}$$

where, in the usual rate of convergence notation, $r = 2\alpha/(2\alpha+1)$. We finally have

$$\begin{aligned} R_L(\Theta) &= \epsilon^2 \sum_{k \in N}\left(1 - \frac{a_k}{\mu}\right) \doteq \epsilon^2\left(k_\mu - \frac{1}{\mu}\frac{\mu^{1+1/\alpha}}{\alpha+1}\right) \\ &\doteq \frac{\alpha}{\alpha+1}\epsilon^2\mu^{1/\alpha} = \left(\frac{\alpha\epsilon^2}{\alpha+1}\right)^r\left((2\alpha+1)C^2\right)^{1-r} \\ &= P_r C^{2(1-r)}\epsilon^{2r}, \end{aligned} \tag{5.10}$$

[1] For concrete examples we index co-ordinates by k rather than i used in the general theory, in part to avoid confusion with $i = \sqrt{-1}$!

where the *Pinsker constant*

$$P_r = \left(\frac{\alpha}{\alpha+1}\right)^r (2\alpha+1)^{1-r} = \left(\frac{r}{2-r}\right)^r (1-r)^{r-1}.$$

Remarks. 1. The rate of convergence ϵ^{2r} depends on the assumed smoothness α: the greater the smoothness, the closer is the rate to the parametric rate ϵ^2.

2. The dependence on the scale C of the ellipsoid is also explicit: in fact, it might be written $C^2(\epsilon^2/C^2)^r$ to emphasise that the convergence rate r really applies to the (inverse) signal-to-noise ratio ϵ^2/C^2.

3. The shrinkage weights $w_k \doteq (1-k^\alpha/\mu)_+$ assign weight close to 1 for low frequencies, and cut off at $k \doteq \mu^{1/\alpha} \propto (C^2/\epsilon^2)^{1/(2\alpha+1)}$. Thus, the number of frequencies retained is an algebraic power of C/ϵ, decreasing as the smoothness α increases.

Fractional integration

We turn to an example of inverse problems that leads to increasing variances σ_i^2 in the sequence model. Consider the noisy indirect observations model

$$Y = Af + \epsilon Z, \tag{5.11}$$

introduced at (3.55), Here $A = I_\beta$ is the operator of $\beta-$fold integration, for β is a positive integer, applied to periodic functions in $L_2(0,1)$ with integral 0. Thus

$$I_1(f)(t) = \int_0^t f(s)ds,$$

$I_2 = I_1(I_1(f))$, etc.

In the trigonometric basis (3.7), it is easy to check, using the assumption $\int_0^1 f = 0$, that

$$I_1(\varphi_{2k}) = -(2\pi k)^{-1}\varphi_{2k-1}, \quad I_1(\varphi_{2k-1}) = (2\pi k)^{-1}\varphi_{2k}.$$

Writing, as in Section 1.4, $Y_i = \langle Y, \varphi_i\rangle, \theta_i = \langle f, \varphi_i\rangle$ and so on, we obtain the sequence form (3.58), with

$$|b_{2k}| = |b_{2k-1}| = (2\pi k)^{-\beta}.$$

Setting now $y_k = Y_k/b_k$, we recover model (5.1) with

$$\sigma_{2k} = \sigma_{2k-1} = (2k)^\beta \pi^\beta \epsilon.$$

Proposition 5.1 allows the evaluation of minimax mean squared error over ellipsoids $\Theta^\alpha(C)$ corresponding to the mean square smoothness condition $\int_0^1 (D^\alpha f)^2 \le C^2/\pi^{2\alpha}$. Calculation along the lines of Section 5.2 shows that a straightforward extension of (5.10) holds:

$$R_L(\Theta^\alpha(C),\epsilon) \sim P_{r,\beta} C^{2(1-r_\beta)} (\pi^\beta \epsilon)^{2r_\beta},$$

with $r_\beta = 2\alpha/(2\alpha+2\beta+1)$ and

$$P_{r,\beta} = \left(\frac{\alpha}{\alpha+2\beta+1}\right)^{r_\beta} \frac{(2\alpha+2\beta+1)^{1-r_\beta}}{2\beta+1}.$$

The index β of ill-posedness leads to a reduction in the rate of convergence from $r =$

$2\alpha/(2\alpha+1)$ in the direct case to $r_\beta = 2\alpha/(2\alpha+2\beta+1)$. When β is not too large, the degradation is not so serious.

Remark. When $\beta > 0$ is not an integer, an interpretation in terms of fractional integration is most natural when described in terms of the basis of complex exponentials $e_k(t) = e^{2\pi ikt}$ for $k \in \mathbb{Z}$. Indeed, if $f(t) \sim \sum c_k e_k(t)$, with $c_0 = 0$, and if we define

$$(I_\beta f)(t) \sim \sum_k \frac{c_k}{(ik)^\beta} e_k(t),$$

then Zygmund (2002, Vol. II, p. 135) shows that

$$(I_\beta f)(t) = \frac{1}{\Gamma(\beta)} \int_{-\infty}^t f(s)(t-s)^{\beta-1} ds.$$

Ellipsoids of analytic functions.

Return to the white noise setting $\sigma_k^2 \equiv \epsilon^2$. Again consider the trigonometric basis for periodic functions on $[0,1]$, but now with $a_{2k} = a_{2k-1} = e^{\alpha k}$, so that $\Theta(a,C) = \{\theta : \sum e^{2\alpha k}(\theta_{2k-1}^2 + \theta_{2k}^2) \le C^2\}$. Since the semiaxes decay exponentially with frequency, these ellipsoids contain only infinitely differentiable functions, which are thus much smoother than typical members of the Sobolev classes.

Remark. To interpret the exponential decay conditions, it may help to think of the periodic function $f(t)$ as a Fourier series in complex exponentials $f(t) = \sum_{-\infty}^{\infty} \zeta_k e^{2\pi ikt}$, where ζ_k is related to the real Fourier coefficients θ_{2k-1} and θ_{2k} as usual via $2\zeta_k = \theta_{2k-1} - i\theta_{2k}$ and $2\zeta_{-k} = \theta_{2k-1} + i\theta_{2k}$. Consider then the domain in which the function $g(z) = \sum_{-\infty}^{\infty} \zeta_k z^k$ of the complex variable $z = re^{2\pi it}$ remains analytic. On the unit circle $|z| = 1$, g reduces to our periodic function f. Now if $|\zeta_k| = O(e^{-\alpha|k|})$, then g is analytic in the annulus $A_\alpha = \{z : e^{-\alpha} < |z| < e^{\alpha}\}$ while a near converse also holds: if g is analytic in a domain containing $\overline{A_\alpha}$, then $|\zeta_k| = O(e^{-\alpha|k|})$. Thus, the larger the value of α, the greater the domain of analyticity.

We turn to interpretation of the linear minimax solution of Proposition 5.1. For given μ, the sum in (5.4) cuts off after $k(\mu) = [(2\alpha)^{-1}\log\mu]$, and so its evaluation involves geometric sums like $\sum_1^k e^{\alpha jp} \doteq c_{\alpha,p} e^{\alpha kp}$ for $p = 1$ and 2 which are, in contrast with the Sobolev case, dominated by a single leading term.

To solve for μ, set $\mu = e^{\alpha r}$ and note that the constraint (5.4) may be rewritten as

$$F(r) = \sum_k e^{\alpha k}(e^{\alpha r} - e^{\alpha k})_+ = C^2/(2\epsilon^2).$$

Restricting r to positive integers, we have $F(r) \doteq e^{2\alpha r}\gamma_\alpha$, with $\gamma_\alpha = c_{\alpha,1} - c_{\alpha,2} > 0$, from which we may write our sought-after solution as $\mu = \beta e^{\alpha r_0}$ for $\beta \in [1, e^\alpha)$ with

$$r_0 = \left[\frac{1}{2\alpha}\log\frac{C^2}{2\gamma_\alpha\epsilon^2}\right].$$

Now we may write the minimax risk (5.3) as $\epsilon \to 0$ in the form

$$R_L(\Theta,\epsilon) = 2\epsilon^2 \sum_{k=1}^{r_0} \left(1 - \beta^{-1} e^{-\alpha(r_0-k)}\right).$$

Thus it is apparent that the number of retained frequencies r_0 is logarithmic in signal to noise—as opposed to algebraic, in the Sobolev case—and the smoothing weights $c_k = 1 - \beta^{-1} e^{-\alpha(r_0 - k)}$ are very close to 1 except for a sharp decline that occurs near r_0. In particular, the minimax linear risk

$$R_L(\Theta, \epsilon) \sim 2\epsilon^2 r_0 \sim \frac{\epsilon^2}{\alpha} \log \epsilon^{-2}$$

is only logarithmically worse than the parametric rate ϵ^2, and the dependence on $\Theta(a, C)$ comes, at the leading order term, only through the analyticity range α and not via the scale factor C.

The minimax estimator compared with smoothing splines.

Still in the white noise setting, we return to the Sobolev ellipsoid setting to suggest that information derived from study of the minimax linear estimate and its asymptotic behavior is quite relevant to the smoothing spline estimates routinely computed in applications by statistical software packages. The following discussion is inspired by Carter et al. (1992).

We have seen in Chapter 3 that the Lagrange multiplier form of smoothing spline problem in the sequence model has form (3.29) with solution

$$\hat{\theta}^{SS}_{\lambda,k} = (1 + \lambda a_k^2)^{-1} y_k,$$

if we choose weights $w_k = a_k^2$ corresponding to the ellipsoid (5.2). This should be compared with the linear minimax solution of (5.5), namely

$$\hat{\theta}_{\mu,k} = (1 - a_k/\mu)_+ y_k.$$

If we make the identification $\lambda \leftrightarrow \mu^{-2}$, then the inequality $(1 + x^2)^{-1} \geq (1 - x)_+$ valid for positive x, shows that the spline estimate shrinks somewhat less in each frequency than the minimax rule.

Pursuing this comparison, we might contrast the worst case mean squared error of the Pinsker and smoothing spline estimates over Sobolev ellipsoids of smooth functions:

$$\bar{r}(\hat{\theta}_\lambda; \epsilon) = \sup_{\theta \in \Theta_2^\alpha(C)} r(\hat{\theta}_\lambda, \theta; \epsilon).$$

It is necessary to specify the order of smoothing spline: we take the weights equal to the (squared) ellipsoid weights: $w_k = a_k^2$, thus $w_{2k} - w_{2k-1} = (2k)^{2\alpha}$. When α is a non-negative integer m, this corresponds to a roughness penalty $\int (D^m f)^2$. We also need to specify the value of the regularization parameter λ to be used in each case. A reasonable choice is the optimum, or *minimax* value

$$\lambda_* = \operatorname*{argmin}_\lambda \bar{r}(\hat{\theta}_\lambda; \epsilon).$$

This is exactly the calculation done in Chapter 3 at (3.47) and (3.78) for the spline and minimax families respectively . [Of course, the result for the minimax family must agree with (5.10)!] In both cases, the solutions took the form

$$\lambda_* \sim (c_1 \epsilon^2 / C^2)^r, \qquad \bar{r}(\lambda_*, \epsilon) \sim c_2 e^{H(r)} C^{2(1-r)} \epsilon^{2r}, \tag{5.12}$$

with $r = 2\alpha/(2\alpha+1)$, and

$$c_1^{SS} = 2v_\alpha/\alpha, \qquad c_2^{SS} = v_\alpha^r/4^{1-r}, \qquad v_\alpha = (1-1/2\alpha)/\mathrm{sinc}\,(1/2\alpha),$$
$$c_1^{M} = \tfrac{1}{2}\bar{v}_\alpha/\alpha, \qquad c_2^{M} = \bar{v}_\alpha^r, \qquad \bar{v}_\alpha = 2\alpha^2/(\alpha+1)(2\alpha+1).$$

Thus the methods have the same dependence on noise level ϵ and scale C, with differences appearing only in the coefficients. We may therefore summarize the comparison through the ratio of maximum mean squared errors. Remarkably, the low noise smoothing spline maximal MSE turns out to be only negligibly larger than the minimax linear risk of the Pinsker estimate. Indeed, for $\Theta = \Theta_2^\alpha(C)$, using (5.12), we find that as $\epsilon \to 0$,

$$\frac{R_{SS}(\Theta,\epsilon)}{R_L(\Theta,\epsilon)} \sim \left(\frac{v_\alpha}{\bar{v}_\alpha}\right)^r \left(\frac{1}{4}\right)^{1-r} \doteq \begin{cases} 1.083 & \alpha = 2 \\ 1.055 & \alpha = 4 \\ \to 1 & \alpha \to \infty. \end{cases} \tag{5.13}$$

Similarly, we may compare the asymptotic choices of the smoothing parameter:

$$\frac{\lambda_{SS}}{\lambda_M} \sim \left(\frac{4v_\alpha}{\bar{v}_\alpha}\right)^r \doteq \begin{cases} 4.331 & \alpha = 2 \\ 4.219 & \alpha = 4 \\ \to 4 & \alpha \to \infty, \end{cases}$$

and so λ_{SS} is approximately four times λ_M and this counteracts the lesser shrinkage of smoothing splines noted earlier.

Furthermore, in the discrete smoothing spline setting of Section 3.4, Carter et al. (1992) present small sample examples in which the efficiency loss of the smoothing spline is even smaller than these asymptotic values. In summary, from the maximum MSE point of view, the minimax linear estimator is not so different from the Reinsch smoothing spline that is routinely computed in statistical software packages.

5.3 Pinsker's Asymptotic Minimaxity Theorem

We return to the general sequence model $y_i = \theta_i + \sigma_i z_i$, where, for asymptotic analysis, we introduce a small parameter ϵ via

$$\sigma_i = \epsilon\lambda_i.$$

We make two assumptions on the ellipsoid weights (a_i) and noise variances (σ_i^2):

(i) a_i are positive and nondecreasing with $\sup_i a_i = \infty$, and
(ii) as $\mu \to \infty$, the ratio

$$\eta^2(\mu) = \max_{a_i \le \mu} \sigma_i^2 \Big/ \sum_{a_i \le \mu/2} \sigma_i^2 \to 0. \tag{5.14}$$

Theorem 5.2 (Pinsker) *Assume that (y_i) follows the sequence model* (5.1) *with noise levels (σ_i). Let $\Theta = \Theta(a, C)$ be an ellipsoid* (5.2) *defined by weights (a_i) and radius $C > 0$. Assume that the weights satisfy conditions (i) and (ii). Then, as $\epsilon \to 0$,*

$$R_N(\Theta,\epsilon) = R_L(\Theta,\epsilon)(1+o(1)). \tag{5.15}$$

Thus the linear minimax estimator (5.5) *is asymptotically minimax among all estimators.*

Remarks. 1. The hardest rectangular subproblem results of Chapter **??** say that $R_L(\Theta;\epsilon) \le 1.25 R_N(\Theta;\epsilon)$, but this theorem asserts that, in the low noise limit, linear estimates cannot be beaten over ellipsoids, being fully efficient.

2. The condition that $\sup a_i = \infty$ is equivalent to compactness of Θ in ℓ_2. In Section 5.5, it is shown for the white noise model that if Θ is not compact, then $R_N(\Theta,\epsilon)$ does not even approach 0 as $\epsilon \to 0$.

3. Condition (ii) rules out exponential growth of σ_i^2, however it is typically satisfied if $\sigma_i^2 = \epsilon^2$ or grows polynomially with i.

4. Pinsker's proof is actually for an even more general situation. We aim to give the essence of Pinsker's argument in somewhat simplified settings.

General comments and heuristics for the proof

The approach is to construct a family of priors, indexed by ϵ, that has Bayes risk comparable to the minimax linear risk as $\epsilon \to 0$. Indeed, dropping explicit reference to Θ, we know from Chapter 4 that

$$R_L(\epsilon) \ge R_N(\epsilon) = \sup\{B(\pi) : \text{supp } \pi \subset \Theta\},$$

so that if we can construct a family of priors $\pi_\epsilon \subset \Theta$ for which

$$\liminf_{\epsilon\to 0} \; B(\pi_\epsilon)/R_L(\epsilon) \ge 1, \tag{5.16}$$

then it must be that $R_N(\epsilon) \sim R_L(\epsilon)$ as $\epsilon \to 0$.

We give first a proof under some relatively restricted conditions:

- (white noise) $\sigma_i \equiv \epsilon$,
- (polynomial growth) $b_1 i^\gamma \le a_i \le b_2 i^\gamma$ for positive constants b_1, b_2 and γ.

Pinsker's *linear* minimax theorem provides, for each ϵ, a Gaussian prior with independent co-ordinates $\theta_i \sim N(0,\tau_{i\epsilon}^2)$ where $\tau_{i\epsilon}^2 = \epsilon^2(\mu_\epsilon/a_i - 1)_+$ and μ_ϵ satisfies $\sum_i a_i(\mu_\epsilon - a_i) = C^2/\epsilon^2$. Since the sequence $(\tau_{i\epsilon}^2)$ maximizes (5.7), we might call this the least favorable *Gaussian* prior. It cannot be least favorable among all priors (in the sense of Section ??), for example because it is not supported on Θ. However, we will show that, under our restricted conditions, that a modification is indeed asymptotically concentrated on Θ, and implements the heuristics described above. The modification is made in two steps. First, define a Gaussian prior with slightly shrunken variances:

$$\pi_\epsilon^G : \;\; \theta_i \sim N\big(0,(1-\lambda_\epsilon)\tau_{i\epsilon}^2\big)\,,$$

with $\lambda_\epsilon \searrow 0$ to be specified. We will show that $\pi_\epsilon^G(\Theta) \to 1$ and so for the second step it makes sense to obtain a prior supported on Θ by conditioning

$$\pi_\epsilon(A) := \pi_\epsilon^G(A|\theta \in \Theta).$$

Comparing Gaussian and conditioned priors. We can do calculations easily with π_ϵ^G since it is Gaussian, but we are ultimately interested in $\pi_\epsilon(\cdot)$ and its Bayes risk $B(\pi_\epsilon)$. We need to show that they are close, which we expect because $\pi_\epsilon^G(\Theta) \approx 1$.

Let $\mathbb{E}$ denote expectation under the joint distribution of (θ, y) when $\theta \sim \pi_\epsilon^G$. Let $\hat{\theta}_\epsilon$ denote the Bayes rule for prior π_ϵ, so that $\hat{\theta}_\epsilon = \mathbb{E}[\theta|\Theta, y]$,

Lemma 5.3

$$(1-\lambda_\epsilon)R_L(\epsilon) \le B(\pi_\epsilon) + \mathbb{E}[\|\hat{\theta}_\epsilon - \theta\|^2, \Theta^c]. \tag{5.17}$$

Proof Since π_ϵ^G consists of co-ordinates θ_i independently distributed as $N(0, (1-\lambda_\epsilon)\tau_i^2)$, the Bayes risk is a sum of univariate terms:

$$B(\pi_\epsilon^G) = \sum \rho_L(\sqrt{1-\lambda_\epsilon}\tau_i, \epsilon) \ge (1-\lambda_\epsilon)\sum \rho_L(\tau_i, \epsilon) = (1-\lambda_\epsilon)R_L(\epsilon). \tag{5.18}$$

For any estimator $\hat{\theta}$,

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 = B(\hat{\theta}, \pi_\epsilon^G) \ge B(\pi_\epsilon^G) \ge (1-\lambda_\epsilon)R_L(\epsilon). \tag{5.19}$$

We can also decompose

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 = \mathbb{E}[\|\hat{\theta} - \theta\|^2|\Theta]\pi_\epsilon^G(\Theta) + \mathbb{E}[\|\hat{\theta} - \theta\|^2, \Theta^c].$$

If for $\hat{\theta}$ we take the Bayes rule for π_ϵ, namely $\hat{\theta}_\epsilon$, then by definition $\mathbb{E}[\|\hat{\theta}_\epsilon - \theta\|^2|\Theta] = B(\pi_\epsilon)$. Now, simply combine this with the two previous displays to obtain (5.17). □

We turn to a bound for the second term in (5.17).

Lemma 5.4 $\mathbb{E}[\|\hat{\theta}_\epsilon - \theta\|^2, \Theta^c] \le c\pi_\epsilon^G(\Theta^c)^{1/2}R_L(\epsilon)$.

Proof Define $a_{\min} = \min a_i$ and observe that on Θ, we have $\|\theta\|^2 \le a_{\min}^{-2}\sum a_i^2\theta_i^2 \le a_{\min}^{-2}C^2$. Then

$$\|\hat{\theta}_\epsilon\|^2 \le \mathbb{E}\left[\|\theta\|^2|\Theta, y\right] \le a_{\min}^{-2}C^2.$$

By contrast, on Θ^c, we have $C^2 < \sum a_i^2\theta_i^2$ and so

$$\|\hat{\theta}_\epsilon - \theta\|^2 \le 2\left[\|\hat{\theta}_\epsilon\|^2 + \|\theta\|^2\right] \le 2a_{\min}^{-2}\left[C^2 + \sum a_i^2\theta_i^2\right] \le 4a_{\min}^{-2}\sum a_i^2\theta_i^2,$$

with the result that

$$\mathbb{E}(\|\hat{\theta}_\epsilon - \theta\|^2, \Theta^c) \le 4a_{\min}^{-2}\sum a_i^2\mathbb{E}(\theta_i^2, \Theta^c). \tag{5.20}$$

Now use the Cauchy-Schwartz inequality, $\mathbb{E}(\theta_i^2, \Theta^c) \le (\mathbb{E}\theta_i^4)^{1/2}\pi_\epsilon^G(\Theta^c)^{1/2}$, along with $\mathbb{E}\theta_i^4 = 3\tau_i^4$ to see that the left side above is bounded by

$$4\sqrt{3}a_{\min}^{-2}\cdot\pi_\epsilon^G(\Theta^c)^{1/2}\cdot\sum a_i^2\tau_i^2.$$

Since $\sum a_i^2\tau_i^2 = R_L(\epsilon)$, the lemma is proved. □

Putting together the two lemmas, we have

$$B(\pi_\epsilon) \ge \left\{1 - \lambda_\epsilon - c\pi_\epsilon^G(\Theta^c)^{1/2}\right\}R_L(\epsilon),$$

and so it remains to show that for suitable $\lambda_\epsilon \to 0$, we also have $\pi_\epsilon^G(\Theta^c) \to 0$.

π_ϵ^G *concentrates on* Θ. Let $S = \sum a_i^2\theta_i^2$ so that under π_ϵ^G,

$$ES = (1-\lambda_\epsilon)\sum a_i^2\tau_i^2 = (1-\lambda_\epsilon)C^2, \qquad \text{and}$$
$$\operatorname{Var} S = 2(1-\lambda_\epsilon)^2\sum a_i^4\tau_i^4 \le 2(1-\lambda_\epsilon)^2C^2\max_i a_i^2\tau_i^2.$$

By the humble Chebychev inequality

$$\begin{aligned}\pi_\epsilon^G(\Theta^c) &\le P(S-ES > \lambda_\epsilon C^2) \le \lambda_\epsilon^{-2}C^{-4}\operatorname{Var} S \\ &\le 2(\lambda_\epsilon^{-1}-1)^2C^{-2}\max_i a_i^2\tau_i^2.\end{aligned} \tag{5.21}$$

Now $a_i^2\tau_i^2 = \epsilon^2 a_i(\mu_\epsilon - a_i) \le (\epsilon\mu_\epsilon/2)^2$. We use the polynomial growth assumption to bound μ_ϵ. Indeed, choose the largest integer k_ϵ such that $b_2(k_\epsilon+1)^\gamma \ge \mu_\epsilon/2$. Then $a_i \le \mu_\epsilon/2$ for all $i \le k_\epsilon$, and so

$$C^2/\epsilon^2 = \sum a_i(\mu_\epsilon - a_i)_+ \ge b_1(\mu_\epsilon/2)\sum_1^{k_\epsilon} i^\gamma.$$

Using an integral approximation and the definition of k_ϵ,

$$\sum_1^{k_\epsilon} i^\gamma \ge k_\epsilon^{\gamma+1}/(\gamma+1) \ge c_\gamma(\mu_\epsilon/(2b_2))^{1+1/\gamma},$$

where $c_\gamma > 0$ depends only on γ. Combining these two displays yields, with $\beta = 1/(2\gamma+1)$ and $c = c(b_1, b_2, \gamma)$,

$$(\epsilon\mu_\epsilon)^2 \le cC^{2(1-\beta)}\epsilon^{2\beta}.$$

In combination with (5.21), this shows that

$$\pi_\epsilon^G(\Theta^c) \le c\lambda_\epsilon^{-2}(\epsilon/C)^{2\beta} \to 0$$

if we choose, for example, $\lambda_\epsilon = \epsilon^{\beta/2}$.

5.4 General case proof*

There are three ways in which asymptotic equivalence of linear and non-linear estimates can occur. The first two are essentially univariate, and rely on the equivalence established at (4.41):

$$\frac{\rho_N(\tau,\epsilon)}{\rho_L(\tau,\epsilon)} \to 1 \qquad \text{as } \tau/\epsilon \to 0 \text{ or } \infty.$$

The third situation, covering intermediate values of τ/ϵ, exploits high-dimensionality in a critical way. It uses a Gaussian prior, for which the optimal estimator is linear. As we have seen in the special case considered in the last section, a concentration of measure property guarantees, as dimensionality grows, that such a prior is essentially supported on an appropriate ellipsoid.

Pinsker's proof handles the three modes simultaneously. The first step is to define a partition of indices $i \in \mathbb{N}$ into three sets N_s, N_g and N_b (with the mnemonics "**s**mall", "**g**aussian"

and "**b**ig"), with the co-ordinate signal-to-noise ratios $\tau_{i\epsilon}^2/\sigma_i^2$ determined by (5.6). The partition depends on a parameter $q > 1$ and declares that

$$i \in \quad N_s, \qquad N_g, \qquad N_b,$$

according as

$$\tau_{i\epsilon}^2/\sigma_i^2 \in \quad [0, q^{-1}], \qquad (q^{-1}, q), \qquad [q, \infty), \tag{5.22}$$

which is seen for $\tau_i^2/\sigma_i^2 = (\mu/a_i - 1)_+$ to be equivalent to

$$a_i \in \quad \Big[\frac{q\mu_\epsilon}{q+1}, \infty\Big], \quad \Big(\frac{\mu_\epsilon}{q+1}, \frac{q\mu_\epsilon}{q+1}\Big), \quad \Big(0, \frac{\mu_\epsilon}{q+1}\Big]. \tag{5.23}$$

Of course, the sets N_m, for $m \in \{s, g, b\}$, depend on ϵ and q.

Example: Sobolev ellipsoids (white noise case) continued. It turns out that each of the regimes "**b**", "**g**" and "**s**" occurs for a large range of indices i even in this canonical case. Indeed, recall from (5.9) that $\mu_\epsilon = c_\alpha(C/\epsilon)^{2\alpha/(2\alpha+1)}$. If we use the fact that $a_k \sim k^\alpha$, it is easy to see, for example, that

$$|N_g| \doteq \frac{q^{1/\alpha} - 1}{(q+1)^{1/\alpha}} c_\alpha^{1/\alpha} (C^2/\epsilon^2)^{1-r} \to \infty,$$

with similar expressions for $|N_b|$ and $|N_s|$ that also increase proportionally to $(C^2/\epsilon^2)^{1-r}$.

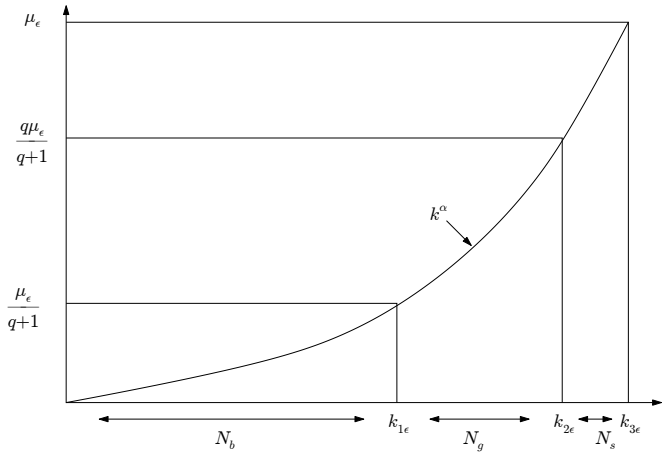

Figure 5.1 The "big", "gaussian" and "small" signal to noise regimes for Sobolev ellipsoids

Definition of priors $\pi = \pi(\epsilon, q)$. A key role is played by the minimax prior variances $\tau_{i\epsilon}^2$ found in Proposition 5.1. We first use them to build sub-ellipsoids Θ_s, Θ_b and $\Theta_g \subset \Theta$, defined for $m \in \{s, b, g\}$ by

$$\Theta_m = \Theta_m(\epsilon, q) = \{(\theta_i, i \in N_m) : \sum_{N_m} a_i^2\theta_i^2 \le \sum_{N_m} a_i^2\tau_{i\epsilon}^2\}.$$

Since $\sum a_i^2 \tau_{i\epsilon}^2 = C^2$, we clearly have $\Theta_s \times \Theta_g \times \Theta_b \subset \Theta$. We now define priors $\pi_{m\epsilon} = \pi_m(\epsilon, q)$ supported on Θ_m, see also Figure 5.2:

$\pi_{s\epsilon}$: for $i \in N_s$, set $\theta_i \overset{ind}{\sim} \pi_{\tau_i}$, the two point priors at $\pm\tau_i$,

$\pi_{b\epsilon}$: for $i \in N_b$, set $\theta_i \overset{ind}{\sim} \pi^V_{\tau_i}$, cosine priors on $[-\tau_i, \tau_i]$, with density $\tau_i^{-1}\cos^2(\pi\theta_i/2\tau_i)$,

$\pi_{g\epsilon}$: for $i \in N_g$, first define π^G, which sets $\theta_i \overset{ind}{\sim} N(0, (1-\lambda)\tau_i^2)$ for some fixed $\lambda \in (0,1)$. Then define $\pi_{g\epsilon}$ by conditioning:

$$\pi_g(A) = \pi^G(A|\theta \in \Theta_g).$$

While the "Gaussian" components prior π^G is not supported in Θ_g, for a suitable choice $\lambda = \lambda(\epsilon, q)$, we shall see that it *nearly* is, and so it makes sense to define π_g by conditioning. The full prior $\pi_\epsilon = \pi_{s\epsilon} \times \pi_{g\epsilon} \times \pi_{b\epsilon}$ and clearly π is supported on Θ.

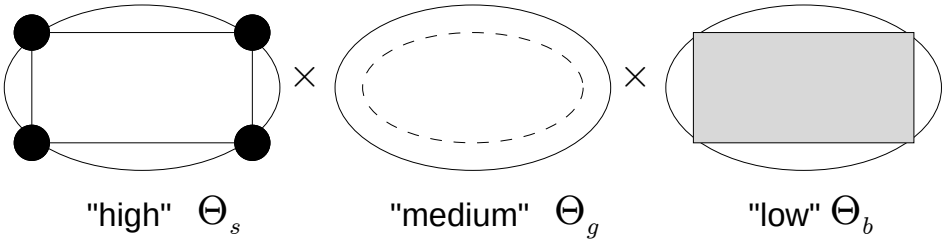

Figure 5.2 The "small" components prior is supported on the extreme points of a hyperrectangle in Θ_s; the "big" component prior lives on a solid hyperrectangle in Θ_b. The "Gaussian" components prior is mostly supported on Θ_g, cf. (5.30), note that the density contours do not match those of the ellipsoid.

Observe that the minimax risk $R_L(\epsilon) = R_s(\epsilon) + R_g(\epsilon) + R_b(\epsilon)$, where for $m = s, l, g$

$$R_m(\epsilon) = \sum_{i \in N_m} \rho_L(\tau_{i\epsilon}, \sigma_i).$$

We show that the priors $\pi_{m\epsilon} = \pi_m(\epsilon, q)$ have the following properties:

(i) $B(\pi_{s\epsilon}) \geq r_s(q^{-1/2})R_s(\epsilon)$ for all ϵ, and $r_s(q^{-1/2}) \to 1$ as $q \to \infty$,

(ii) $B(\pi_{b\epsilon}) \geq r_b(q^{1/2})R_b(\epsilon)$ for all ϵ, and $r_b(q^{1/2}) \to 1$ as $q \to \infty$, and

(iii) If $\delta > 0$ and $q = q(\delta)$ are given, and if $R_g(\epsilon) \geq \delta R_L(\epsilon)$, then for $\epsilon < \epsilon(\delta)$ sufficiently small, $B(\pi_{g\epsilon}) \geq (1-\delta)R_g(\epsilon)$.

Assuming these properties to have been established, we conclude the proof as follows. Fix $\delta > 0$ and then choose $q(\delta)$ large enough so that both $r_s(q^{-1})$ and $r_b(q) \geq 1 - \delta$. We obtain

$$B(\pi_{m\epsilon}) \geq (1-\delta)R_m(\epsilon), \qquad \text{for } m \in \{s, b\}. \tag{5.24}$$

Now, if $R_g(\epsilon) \geq \delta R_L(\epsilon)$, then the previous display holds also for $m = g$ and ϵ sufficiently small, by (iii), and so adding, we get $B(\pi_\epsilon) \geq (1-\delta)R_L(\epsilon)$ for ϵ sufficiently small. On the other hand, if $R_g(\epsilon) \leq \delta R_L(\epsilon)$, then, again using (5.24),

$$B(\pi_\epsilon) \geq (1-\delta)[R_b(\epsilon) + R_s(\epsilon)] = (1-\delta)[R_L(\epsilon) - R_g(\epsilon)] \geq (1-\delta)^2 R_L(\epsilon).$$

Either way, we establish (5.16), and are done.

Proofs for (i) and (ii). These are virtually identical and use the fact that two point and cosine priors are asymptotically least favorable as $\tau_i/\sigma_i \to 0$ and ∞ respectively. We tackle $B(\pi_{s\epsilon})$ first. For a scalar problem $y_1 = \theta_1 + \sigma_1 z_1$ with univariate prior $\pi(d\theta)$ introduce the notation $B(\pi, \sigma_1)$ for the Bayes risk. In particular, consider the two-point priors π_τ needed for the small signal case. By scaling, $B(\pi_\tau, \sigma) = \sigma^2 B(\pi_{\tau/\sigma}, 1)$, and the explicit formula (2.24) for $B(\pi_{\tau/\sigma}, 1)$ shows that when written in the form

$$B(\pi_\tau, \sigma) = \rho_L(\tau, \sigma) g(\tau/\sigma), \tag{5.25}$$

we must have $g(t) \to 1$ as $t \to 0$. Now, using this and (5.22) along with the additivity of Bayes risks,

$$B(\pi_{s\epsilon}) = \sum_{N_s} B(\pi_{\tau_i}, \sigma_i) = \sum_{N_s} g(\tau_{i\epsilon}/\sigma_i)\rho_L(\tau_{i\epsilon}, \sigma_i) \geq r_s(q^{-1/2}) R_s(\epsilon), \tag{5.26}$$

if we set $r_s(u) = \inf_{0 \leq t \leq u} g(t)$. Certainly $r_s(u) \to 1$ as $u \to 0$, and this establishes (i).

For the large signal case (ii), we use the cosine priors π_τ^V, and the Fisher information bound (4.16), so that the analog of (5.25) becomes

$$B(\pi_\tau^V, \sigma) \geq \rho_L(\tau, \sigma) h(\tau/\sigma),$$

with $h(t) = (t^2+1)/(t^2 + I(\pi_1^V)) \to 1$ as $t \to \infty$. The analog of (5.26), $B(\pi_{g\epsilon}) \geq r_b(q^{1/2}) R_b(\epsilon)$ follows with $r_b(q) = \inf_{t \geq q} h(t) \to 1$ as $t \to 1$.

Proof of (iii): This argument builds upon that given in the special white noise setting in the previous section. Let $\hat{\theta}_g = \mathbb{E}[\theta | \Theta_g, y]$ denote the Bayes rule for $\pi_{g\epsilon}$. With the obvious substitutions, the argument leading to (5.17) establishes that

$$(1-\lambda) R_g(\epsilon) \leq B(\pi_{g\epsilon}) + \mathbb{E}[\|\hat{\theta}_g - \theta\|^2, \Theta_g^c]. \tag{5.27}$$

Now we estimate $\mathbb{E}[\|\hat{\theta}_g - \theta\|^2, \Theta_g^c]$. Here and below, we abuse notation slightly by writing θ for $(\theta_i : i \in N_g)$, and similarly for y. We first record two properties of indices in the Gaussian range N_g: for $i, j \in N_g$,

$$q^{-1} < a_i/a_j < q, \qquad \tau_{i\epsilon}^2 \leq (1+q)\rho_L(\tau_{i\epsilon}, \sigma_i). \tag{5.28}$$

The first bound uses (5.23), as does the second after noting that $\tau_i^2/\rho_L(\tau_i, \sigma_i) = \mu/a_i$.

We now show an analog of Lemma 5.4:

$$\mathbb{E}\{\|\hat{\theta}_g - \theta\|^2, \Theta_g^c\} \leq c(q)\pi^G(\Theta_g^c)^{1/2} R_g(\epsilon). \tag{5.29}$$

We bound the left side of (5.29) exactly as in (5.20), using now that $a_i/a_{\min} \leq q$, and

obtaining instead the upper bound $4q^2 \sum_{N_g} \mathbb{E}[\theta_i^2, \Theta_g^c]$. Using Cauchy-Schwartz as before, and then the second part of (5.28), this in turn is bounded by

$$4\sqrt{3}q^2\pi^G(\Theta_g^c)^{1/2} \sum_{N_g} \tau_i^2 \leq c(q)\pi^G(\Theta_g^c)^{1/2} \sum_{N_g} \rho_L(\tau_{i\epsilon}, \sigma_i),$$

which is the desired bound (5.29).

We now show, by modifying the earlier Chebychev inequality argument, that

$$\pi^G(\Theta_g^c) \leq 2q(\lambda^{-1} - 1)^2 \eta^2(N_g), \tag{5.30}$$

where

$$\eta^2(N) = \max_{i \in N} \sigma_i^2 \Big/ \sum_{i \in N} \sigma_i^2.$$

The bound (5.30) reflects three necessary quantities, and hence shows why the method works. First q governs the signal to noise ratios $\tau_{i\epsilon}^2/\sigma_i^2$, while λ governs the 'slack' in the expectation ellipsoid. Finally $\eta^2(N_g)$ is a surrogate for the number of components $1/N_g$ in the unequal variance case. (Indeed, if all σ_i^2 are equal, this reduces to $1/|N_g|$).

Proof In the argument leading to (5.21), restrict the indices considered to N_g, define $C_g^2 = \sum_{i \in N_g} a_i^2 \tau_{i\epsilon}^2$, and conclude that

$$\pi^G(\Theta_g^c) \leq 2(\lambda_\epsilon^{-1} - 1)^2 C_g^{-2} \max_{i \in N_g} a_i^2 \tau_i^2.$$

From definition (5.6) of τ_i^2 and bounds (5.23) defining the Gaussian range N_g:

$$a_i^2 \tau_i^2 = \sigma_i^2 a_i (\mu - a_i)_+ \in \sigma_i^2 \mu^2 [q(q+1)^{-2}, 1/4],$$

and so

$$\frac{\max\, a_i^2 \tau_i^2}{\sum a_i^2 \tau_i^2} \leq \frac{(q+1)^2}{4q} \frac{\max\, \sigma_i^2}{\sum \sigma_j^2} \leq q\, \eta^2(N_g). \qquad \square$$

Inserting bound (5.30) we obtain

$$\mathbb{E}\{\|\hat{\theta}_g - \theta\|^2, \Theta_g^c\} \leq c(q)(\lambda_\epsilon^{-1} - 1)\eta(N_g) R_g(\epsilon). \tag{5.31}$$

We now use the hypothesis $R_g(\epsilon) \geq \delta R_L(\epsilon)$ to obtain a bound for $\eta(N_g)$. Indeed, using the definition of R_g and (5.7), we have

$$\begin{aligned} \sum_{N_g} \sigma_i^2 \geq R_g(\epsilon) \geq \delta R_L(\epsilon) &= \delta \sum \sigma_i^2 (1 - a_i/\mu)_+ \\ &\geq (\delta/2) \sum_{a_i \leq \mu/2} \sigma_i^2, \end{aligned}$$

and since (5.23) says that $N_g \subset \{i : a_i \leq \mu_\epsilon\}$,

$$\eta^2(N_g) = \max_{i \in N_g} \sigma_i^2 \Big/ \sum_{i \in N_g} \sigma_i^2 \leq (2/\delta)\eta^2(\mu_\epsilon).$$

Combining this last bound with (5.31), we obtain

$$\mathbb{E}[\|\hat{\theta}_g - \theta\|^2, \Theta_g^c] \le f(q,\lambda,\delta)\eta(\mu_\epsilon) R_g(\epsilon),$$

where $f(q,\lambda,\delta) = p(q)(\lambda^{-1}-1)\sqrt{2/\delta}$. We may now rewrite (5.27) to get

$$B(\pi_{g\epsilon}) \ge R_g(\epsilon)[1 - \lambda - f(q,\lambda,\delta)\eta(\mu_\epsilon)].$$

Observe that the condition (5.4), here with $\sigma_i^2 = \epsilon^2\lambda_i^2$, along with the assumption (i) that $a_i \nearrow \infty$ monotonically implies that $\mu_\epsilon \to \infty$ as $\epsilon \to 0$. Our assumption (ii) then implies that $\eta(\mu_\epsilon) \to 0$. Set $\lambda = \delta/2$ and note that for $\epsilon < \epsilon(\delta, q(\delta))$, we have $f(q(\delta),\lambda,\delta)\eta(\mu_\epsilon) < \delta/2$. This completes the proof of (iii).

Remark. The exponential bound (2.59) shows the concentration of measure more acutely. It is applied here to $z_i = \theta_i/\sigma_i$, with $\alpha_i = (1-\lambda)a_i^2\tau_i^2$ and $t = \lambda C_g^2 \le \|\alpha\|_1$ so long as $\lambda \le \frac{1}{2}$. If we set L equal to the right side of bound (5.30), then we get $\pi^G(\Theta_g^c) \le \exp\{-1/(4L)\}$. The latter is less than L when $L < 0.1$, so is certainly much stronger, but the Chebychev bound is enough for our purposes.

5.5 *Interlude:* Compactness and Consistency

This section, a digression, is included for variety, and because of the different methods used. We have seen from Pinsker's theorem that if an ellipsoid $\Theta(a)$ is compact, then $R_N(\Theta(a),\epsilon) \to 0$ as $\epsilon \to 0$. In fact, for quite general sets Θ, compactness is both necessary and sufficient for the existence of a uniformly consistent estimator, so long as we use the ℓ_2 norm to define both the error measure and the topology on Θ.

Theorem 5.5 *In the homoscedastic Gaussian sequence model* (3.1)*, assume that* Θ *is bounded in* $\ell_2(I)$*. Then as* $\epsilon \to 0$*,* $R_N(\Theta,\epsilon) \to 0$ *if and only if* $\overline{\Theta}$ *is compact.*

Of course, if $R_N(\Theta,\epsilon)$ does not converge to 0, then there exists $c > 0$ such that every estimator has maximum risk at least c regardless of how small the noise level might be. This again illustrates why it is necessary to introduce constraints on the parameter space in order to obtain meaningful results in nonparametric theory. In particular, there can be no uniformly consistent estimator on $\{\theta \in \ell_2(\mathbb{N}) : \|\theta\|_2 \le 1\}$, or indeed on any open set in the norm topology.

Because there are no longer any geometric assumptions on Θ, the tools used for the proof change: indeed methods from testing, classification and from information theory now appear. While the result involves only consistency and so is not at all quantitative, it nevertheless gives a hint of the role that covering numbers and metric entropy play in a much more refined theory (Birgé, 1983) that describes how the "massiveness" of Θ determines the possible rates of convergence of $R_N(\Theta)$.

A lower bound for misclassification error

Any method that chooses between a finite number m of alternative distributions necessarily has an error probability bounded below in terms of $\log m$ and the mutual separation of those distributions.

In detail, let $\{\theta_1, \ldots, \theta_m\}$ be a finite set, and $P_{\theta_1}, \ldots, P_{\theta_m}$ be a corresponding set of probability distributions on $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$. For convenience, assume that the P_{θ_i} are mutually absolutely continuous, and so have positive densities p_i with respect to some dominating measure ν. Then, the Kullback-Leibler divergence between two probability measures P and Q having densities p, q relative to ν is

$$K(P, Q) = \int \log \frac{dP}{dQ}\, dP = \int \log \frac{p}{q}\, p\, d\nu. \tag{5.32}$$

The following lower bound is a formulation by Birgé (1983, Lemma 2.7) of a lemma due to Ibragimov and Has'minskii (1981, pages 324-5).

Lemma 5.6 *With the above definitions, let $\hat{\theta} : \mathcal{Y} \to \{\theta_1, \ldots, \theta_m\}$ be an arbitrary estimator. Then*

$$ave_i\, P_{\theta_i}\{\hat{\theta} \neq \theta_i\} \geq 1 - \frac{ave_{i,j} K(P_{\theta_i}, P_{\theta_j}) + \log 2}{\log(m-1)}. \tag{5.33}$$

Remark. Both averages in inequality (5.33) can of course be replaced by maxima over i and (i, j) respectively.

Proof We first recall Fano's lemma from information theory (e.g. Cover and Thomas (1991, page 39)). Let θ be a random variable with distribution $P(\theta = \theta_i) = q_i$. The *conditional entropy* of θ given Y is defined by

$$H(\theta|Y) = -E \sum_i P(\theta = \theta_i|Y) \log P(\theta = \theta_i|Y),$$

where the expectation is taken over the Y marginal of the joint distribution of (θ, Y). Let $h(q) = -q \log q - (1-q)\log(1-q)$ be the binary entropy function. Write $p_e = P(\hat{\theta} \neq \theta)$ for the overall error probability when using estimator $\hat{\theta}$. Fano's lemma provides a lower bound for p_e:

$$h(p_e) + p_e \log(m-1) \geq H(\theta|Y).$$

To apply this, we choose the uniform distribution for θ: $q_i = 1/m$ for all i. Hence the marginal density of Y is just $\frac{1}{m}\sum_k p_k$ and the posterior probabilities $P(\theta = \theta_i|Y) = p_i / \sum_j p_j$. Consequently,

$$\begin{aligned} H(\theta|Y) &= -\frac{1}{m} \int \sum_i \frac{p_i}{\sum p_j} \log \frac{p_i}{\sum p_j} \sum_k p_k\, d\nu \\ &= \log m - \frac{1}{m} \sum_i \int p_i \log \frac{p_i}{\frac{1}{m}\sum_j p_j}\, d\nu. \end{aligned}$$

Now apply Jensen's inequality: $\log(\frac{1}{m}\sum p_j) \geq \frac{1}{m}\sum \log p_j$. Combine this with Fano's lemma and the bound $h(p_e) \leq \log 2$ to get

$$p_e \log(m-1) \geq \log m - \frac{1}{m^2} \sum_i \sum_j \int p_i \log \tfrac{p_i}{p_j}\, d\nu - \log 2.$$

Divide through by $\log(m-1)$ and insert definition (5.32) to yield the result. □

Necessity of compactness

For both parts of the proof, we use an equivalent formulation of compactness, valid in complete metric spaces, in terms of total boundedness: Θ is totally bounded if and only if for every δ, there is a finite set $\{\theta_i, \dots, \theta_m\}$ such that the open balls $B(\theta_i, \delta)$ of radius δ centered at θ_i cover $\overline{\Theta}$: so that $\overline{\Theta} \subset \cup_{i=1}^m B(\theta_i, \delta)$. Also, since Θ is bounded, it has a finite *diameter* $\Delta = \sup\{\|\theta_1 - \theta_2\| : \theta_1, \theta_2 \in \Theta\}$.

Let $\delta > 0$ be given. Since $R_N(\Theta, \epsilon) \to 0$, there exists a noise level ϵ and an estimator $\tilde{\theta}_\delta$ such that

$$E_{\theta,\epsilon}\|\tilde{\theta}_\delta - \theta\|^2 \le \delta^2/2 \qquad \text{for all } \theta \in \Theta. \tag{5.34}$$

Let Θ_δ be a finite and $2\delta-$*discernible* subset of Θ: each distinct pair θ_i, θ_j in Θ_δ satisfies $\|\theta_i - \theta_j\| > 2\delta$. From $\tilde{\theta}_\delta(y)$ we build an estimator $\hat{\theta}_\delta(y)$ with values confined to Θ_δ by choosing a closest $\theta_i \in \Theta_\delta$ to $\tilde{\theta}_\delta(y)$: of course, whenever $\hat{\theta}_\delta \ne \theta_i$, it must follow that $\|\tilde{\theta}_\delta - \theta_i\| \ge \delta$. Consequently, from Markov's inequality and (5.34), we have for all i

$$P_{\theta_i}\{\hat{\theta}_\delta \ne \theta_i\} \le P_{\theta_i}\{\|\tilde{\theta}_\delta - \theta_i\| \ge \delta\} \le \delta^{-2} E\|\tilde{\theta}_\delta - \theta_i\|^2 \le 1/2. \tag{5.35}$$

On the other hand, the misclassification inequality (5.33) provides a lower bound to the error probability: for the noise level ϵ Gaussian sequence model, one easily evaluates

$$K(P_{\theta_i}, P_{\theta_j}) = \|\theta_i - \theta_j\|^2/2\epsilon^2 \le \Delta^2/2\epsilon^2,$$

where Δ is the diameter of Θ, and so

$$\max_i P_{\theta_i}\{\hat{\theta}_\delta \ne \theta_i\} \ge 1 - \frac{\Delta^2/2\epsilon^2 + \log 2}{\log(|\Theta_\delta| - 1)}.$$

Combining this with (5.35) gives a uniform upper bound for the cardinality of Θ_δ:

$$\log(|\Theta_\delta| - 1) \le \Delta^2\epsilon^{-2} + 2\log 2.$$

We may therefore speak of a $2\delta-$discernible subset $\Theta_\delta \subset \Theta$ of *maximal* cardinality, and for such a set, it is easily checked that $\overline{\Theta}$ is covered by closed balls of radius 4δ centered at the points of Θ_δ. Since δ was arbitrary, this establishes that $\overline{\Theta}$ is totally bounded, and so compact.

Sufficiency of Compactness

Given $\delta > 0$, we will construct an estimator $\hat{\theta}_\epsilon$ such that $E_\theta\|\hat{\theta}_\epsilon - \theta\|^2 \le 20\,\delta^2$ on Θ for all sufficiently small ϵ. Indeed, compactness of Θ supplies a finite set $\Theta_\delta = \{\theta_1, \dots, \theta_m\}$ such that $\overline{\Theta} \subset \cup_{i=1}^m B(\theta_i, \delta)$, and we will take $\hat{\theta}_\epsilon$ to be the maximum likelihood estimate on the *sieve* Θ_δ. Thus we introduce the (normalized) log-likelihood

$$L(\theta) = \epsilon^2 \log dP_{\theta,\epsilon}/dP_{0,\epsilon} = \langle y, \theta\rangle - \tfrac{1}{2}\|\theta\|^2, \tag{5.36}$$

and the maximum likelihood estimate

$$\hat{\theta}_\epsilon = \arg\max_{\theta_i \in \Theta_\delta} L(\theta).$$

Since Θ has diameter Δ, we have for any $\theta \in \Theta$ the simple MSE bound

$$E_\theta \|\hat{\theta}_\epsilon - \theta\|^2 \le (4\delta)^2 + \Delta^2 \sum_{i:\|\theta_i - \theta\| \ge 4\delta} P_\theta\{\hat{\theta}_\epsilon = \theta_i\}. \tag{5.37}$$

We now show that the terms in the second sum are small when ϵ is small. Let $\theta \in \Theta$ be fixed, and choose a point in Θ_δ, renumbered to θ_1 if necessary, so that $\theta \in B(\theta_1, \delta)$. To have $\hat{\theta}_\epsilon = \theta_i$ certainly implies that $L(\theta_i) \ge L(\theta_1)$, and from (5.36)

$$L(\theta_i) - L(\theta_1) = \langle y - \tfrac{1}{2}(\theta_i + \theta_1), \theta_i - \theta_1 \rangle.$$

Substituting $y = \theta + \epsilon z$, putting $u = (\theta_i - \theta_1)/\|\theta_i - \theta_1\|$, and defining the standard Gaussian variate $Z = \langle z, u \rangle$, we find that $L(\theta_i) \ge L(\theta_1)$ implies

$$\epsilon Z \ge \langle \tfrac{1}{2}(\theta_i + \theta_1) - \theta, u \rangle \ge \tfrac{1}{2}\|\theta_i - \theta_1\| - \delta \ge \delta,$$

where in the second inequality we used $|\langle \theta_1 - \theta, u \rangle| \le \|\theta_1 - \theta\| < \delta$. Thus $P_\theta\{\hat{\theta}_\epsilon = \theta_i\} \le \tilde{\Phi}(\delta/\epsilon)$, and so from (5.37)

$$E_\theta \|\hat{\theta}_\epsilon - \theta\|^2 \le (4\delta)^2 + m\Delta^2 \tilde{\Phi}(\delta/\epsilon) \le 20\delta^2,$$

whenever ϵ is sufficiently small.

5.6 Notes and Exercises

Pinsker's paper inspired a considerable literature. Here we mention only two recent works which contain, among other developments, different proofs of the original result: Belitser and Levit (1995) and Tsybakov (1997), and the examples given in Sections 5.2– 5.2.

As noted in the proof of Theorem 4.22, identity (5.7) is itself a minimax theorem, indeed Pinsker gave a direct proof.

The consistency characterization, Theorem 5.5, is a special case of a result announced by Ibragimov and Has′minskiĭ (1977), and extended in Ibragimov and Khasminskii (1997).

Exercises

5.1 Consider a more general ellipsoid $\Theta = \{\theta : \theta^T A \theta \le C\}$ for a positive definite matrix A. Suppose $Y \sim N(\theta, \Sigma)$ and that A and Σ commute: $A\Sigma = \Sigma A$. Show that there is a linear transformation of Y for which the Pinsker theorems hold.
[The situation appears to be less simple if A and Σ do not commute – references?]

5.2 (*Non asymptotic bound for efficiency of smoothing splines.*) This exercise pursues the observations of Carter et al. (1992) that the efficiency of smoothing splines is even better for "non-asymptotic" values of ϵ.

(i) Revisit the proof of Proposition 3.9 and show that for $\alpha = m$ and the trigonometric basis

$$\bar{r}(\hat{\theta}_\lambda; \epsilon) \le v_\alpha \epsilon^2 \lambda^{-1/(2\alpha)} + (C^2/4)\lambda + \epsilon^2.$$

(ii) Revisit the evaluation of $R_L(\Theta, \epsilon)$ prior to (8.10) and show that

$$R_L(\Theta, \epsilon) \ge \frac{\alpha}{\alpha + 1} \epsilon^2 \mu_\epsilon^{1/\alpha}.$$

(iii) Let A be the set of values (α, δ) for which

$$\delta \sum_{k \ge 0} (\delta k)^\alpha [1 - (\delta k)^\alpha]_+ \le \int_0^1 v^\alpha (1 - v^\alpha) dv.$$

[It is *conjectured* that this holds for most or all $\alpha > 0, \delta > 0$]. Show that

$$\mu_\epsilon \geq \bar{\mu}_\epsilon = \left(\frac{(\alpha+1)(2\alpha+1)}{\alpha} \frac{C^2}{\epsilon^2} \right)^{\alpha/(2\alpha+1)}.$$

so long as $(\alpha, \bar{\mu}_\epsilon^{-1/\alpha}) \in A$.
(iv) Conclude that in these circumstances,

$$\frac{R_{SS}(\Theta;\epsilon)}{R_L(\Theta;\epsilon)} \leq e_\alpha + c_\alpha(\epsilon/C)^{2(1-r)}$$

for all $\epsilon > 0$. [Here e_α is the constant in the limiting efficiency (5.13).]

6

Adaptive Minimaxity over Ellipsoids

> However beautiful the strategy, you should occasionally look at the results. (Winston Churchill)

An estimator that is exactly minimax for a given parameter set Θ will depend, often quite strongly, on the details of that parameter set. While this is informative about the effect of assumptions on estimators, it is impractical for the majority of applications in which no single parameter set comes as part of the problem description.

In this chapter, we shift perspective in order to study the properties of estimators that can be defined without recourse to a fixed Θ. Fortunately, it turns out that certain such estimators can come close to being minimax over a whole *class* of parameter sets. We exchange exact optimality for a single problem for approximate optimality over a range of circumstances. The resulting 'robustness' is usually well worth the loss of specific optimality.

The example developed in this chapter is the use of the James-Stein estimator on *blocks* of coefficients to approximately mimick the behavior of linear minimax rules for particular ellipsoids.

The problem is stated in more detail for ellipsoids in Section 6.1. The class of linear estimators that are constant on blocks is studied in Section 6.2, while the blockwise James-Stein estimator appears in Section 6.3. The adaptive minimaxity of blockwise James-Stein is established; the proof boils down to the ability of the James-Stein estimator to mimick the ideal linear shrinkage rule appropriate to each block, as already seen in Section 2.6.

While the blockwise shrinkage approach may seem rather tied to the details of the sequence model, in fact it accomplishes its task in a rather similar way to kernel smoothers or smoothing splines in other problems. This is set out both by heuristic argument and in a couple of concrete examples in Section 6.4.

Looking at the results of our blockwise strategy (and other linear methods) on one of those examples sets the stage for the focus on non-linear estimators in following chapters: linear smoothing methods, with their constant smoothing bandwidth, are ill-equipped to deal with data with sharp transitions, such as step functions. It will be seen later that the adaptive minimax point of view still offers useful insight, but now for a different class of estimators (wavelet thresholding) and wider classes of parameter spaces.

Section 6.5 is again an interlude, containing some remarks on "fixed θ" versus worst case asymptotics and on superefficiency. Informally speaking, superefficiency refers to the possibility of exceptionally good estimation performance at isolated parameter points. In parametric statistics this turns out, fortunately, to be usually a peripheral issue, but examples

given here show that points of superefficiency are endemic in nonparametric estimation. The dangers of over-reliance on asymptotics based on a single θ are illustrated in an example where nominally optimal bandwidths are found to be very sensitive to aspects of the function that are difficult to estimate at any moderate sample size.

6.1 The problem of adaptive estimation

We again suppose that we are in the white noise Gaussian sequence model $y_i = \theta_i + \epsilon z_i$, and consider the family of ellipsoids corresponding to smoothness constraints $\int (D^\alpha f)^2 \le L^2$ on periodic functions in $L_2[0,1]$ when represented in the Fourier basis (3.7):

$$\Theta^\alpha(C) = \{\theta \in \ell_2 : \theta_0^2 + \sum_{l \ge 1} (2l)^{2\alpha}(\theta_{2l-1}^2 + \theta_{2l}^2) \le C^2\} \qquad \alpha, C > 0. \tag{6.1}$$

As we have seen in the previous chapter, Pinsker's theorem delivers a linear estimator $\hat\theta(\alpha, C, \epsilon)$, given by (5.5), which is minimax linear for all $\epsilon > 0$, and asymptotically minimax among *all* estimators as $\epsilon \to 0$.

As a practical matter, the constants (α, C) are generally unknown, and even if one believed a certain value (α_0, C_0) to be appropriate, there is an issue of robustness of MSE performance of $\hat\theta(\alpha_0, C_0, \epsilon)$ to misspecification of (α, C). One possible way around this problem is to construct an estimator family $\hat\theta_\epsilon^*$, whose definition does not depend on (α, C), such that if θ is in fact restricted to some $\Theta^\alpha(C)$, then $\hat\theta_\epsilon^*$ has MSE appropriate to that space:

$$\sup_{\theta \in \Theta^\alpha(C)} r(\hat\theta_\epsilon^*, \theta) \le c_\epsilon(\Theta) R_N(\Theta^\alpha(C), \epsilon) \qquad \text{as } \epsilon \to 0. \tag{6.2}$$

where $c_\epsilon(\Theta)$ is a bounded sequence. Write $\mathcal{T}_2$ for the collection of all ellipsoids $\{\Theta^\alpha(C) : \alpha, C > 0\}$. One then calls $\hat\theta_\epsilon$ *rate-adaptive*: it "learns" the right rate of convergence for all $\Theta \in \mathcal{T}_2$.

If $\hat\theta_\epsilon^*$ has the stronger property that $c_\epsilon(\Theta) \to 1$ for each $\Theta \in \mathcal{T}_2$ as $\epsilon \to 0$, then it is called *adaptively* asymptotically *minimax*: it gets the constant right as well! An adaptive minimax estimator sequence for Sobolev ellipsoids $\mathcal{T}_2$ was constructed by Efroimovich and Pinsker (1984), and this chapter presents their blockwise estimator approach, lightly modified with use of the James-Stein method. We will see that good non-asymptotic bounds are also possible, and that the dyadic-blocks James-Stein estimator is a plausible estimator for practical use in appropriate settings.

6.2 Blockwise Estimators

Consider the Gaussian sequence model, at first with abstract countable index set $\mathcal{I}$: $y_i = \theta_i + \epsilon z_i$. Suppose that $\mathcal{I}$ is partitioned into an ordered sequence of blocks B_j of finite cardinality n_j. We write y_j for the vector of coefficients $\{y_i, I \in B_j\}$, and similarly for θ_j, z_j etc.

In this chapter, we will mostly focus on the case that $\mathcal{I} = \mathbb{N}$, and suppose that the blocks are defined by an increasing sequence l_j :

$$B_j = \{l_j + 1, l_j + 2, \ldots, l_{j+1}\}, \qquad n_j = l_{j+1} - l_j. \tag{6.3}$$

Particular examples might include $l_j = j^\beta$ for some $\beta > 0$, or $l_j = [e^{\sqrt{j}}]$. However, we will devote particular attention to the case of *dyadic blocks.* in which $l_j = 2^j$, so that the jth block has cardinality $n_j = 2^j$.

In this case, we consider a variant of the ellipsoids (6.1) that is defined using weights that are constant on the dyadic blocks: $a_l \equiv 2^{j\alpha}$ if $l \in B_j$. The corresponding *dyadic Sobolev ellipsoids*

$$\Theta_D^\alpha(C) = \{\theta : \theta_1^2 + \sum_{j\geq 0} 2^{2j\alpha} \sum_{l\in B_j} \theta_l^2 \leq C^2\}. \tag{6.4}$$

Let $\mathcal{T}_{D,2}$ denote the class of such dyadic ellipsoids $\{\Theta_D^\alpha(C), \alpha, C > 0\}$.

The two approaches are norm-*equivalent*: write $\|\theta\|_{F,\alpha}^2$ for the squared norm appearing in (6.1) and $\|\theta\|_{D,\alpha}^2$ for that appearing in (6.4). It is easily seen that for all $\theta \in \ell_2$:

$$\|\theta\|_{D,\alpha} \leq \|\theta\|_{F,\alpha} \leq 2^\alpha \|\theta\|_{D,\alpha}. \tag{6.5}$$

Remark. For wavelet bases, weights that are constant on dyadic blocks are the natural way to represent mean-square smoothness–see Section 9.6. In this case, the index $I = (j,k)$, with $j \geq 0$ and $k \in \{1,\ldots,2^j\}$. The success of octave based thinking in harmonic analysis and wavelet methodology gives a distinguished status to the use of dyadic blocks vis-a-vis other choices, and we will focus most attention on this version. Note also that there are no parameters left unspecified, save the noise level ϵ, here assumed known.

Block diagonal linear estimators. This term refers to the subclass of diagonal linear estimators in which the shrinkage factor is constant within blocks: for all blocks j:

$$\hat{\theta}_{j,c_j}(y) = c_j y_j \qquad c_j \in \mathbb{R}.$$

The mean squared error on the jth block has a simple form

$$r(\hat{\theta}_{j,c_j}, \theta_j) = n_j \epsilon^2 c_j^2 + (1-c_j)^2 \|\theta_j\|^2.$$

The corresponding minimax risk among block linear estimators is then

$$R_{BL}(\Theta,\epsilon) = \inf_{(c_j)} \sup_{\Theta} \sum_j r_\epsilon(\hat{\theta}_{j,c_j}, \theta_j).$$

The minimax theorem for diagonal linear estimators, Theorem 4.22, and its proof have an immediate analog in the block case.

Proposition 6.1 *If Θ is compact, solid-orthosymmetric and quadratically convex, then*

$$R_{BL}(\Theta,\epsilon) = \sup_{\Theta} \inf_{(c_j)} \sum_j r_\epsilon(\hat{\theta}_{j,c_j}, \theta_j). \tag{6.6}$$

Proof As in the proof of Theorem 4.22, we apply the Kneser-Kuhn minimax theorem, this time with payoff function, for $c = (c_j)$ and $s = (s_i) = (\theta_i^2)$, given by

$$f(c,s) = \sum_j n_j \epsilon^2 c_j^2 + (1-c_j)^2 \sum_{i\in B_j} s_i. \qquad \square$$

To simplify (6.6), we adopt the ideal linear shrinkage interpretation from Section 2.6. Indeed the sum in (6.6) is minimized term by term, by the blockwise ideal shrinkage estimators given by (2.51) and with corresponding ideal MSE's given by (2.52). Thus

$$R_{BL}(\Theta,\epsilon) = \sup_{\Theta} \sum_j r_\epsilon(\hat{\theta}_j^{IS},\theta_j) = \sup_{\Theta} \sum_j \frac{n_j\epsilon^2\|\theta_j\|^2}{n_j\epsilon^2+\|\theta_j\|^2}. \tag{6.7}$$

Block Linear versus Linear. Clearly $R_L(\Theta,\epsilon) \le R_{BL}(\Theta,\epsilon)$. In two cases, more can be said:

(i) Call Θ *block symmetric* if Θ is invariant to permutations of the indices I within blocks. A variant of the argument in Section 4.7 employing random block-preserving permutations (instead of random signs) shows that if Θ is solid, ortho- and block- symmetric, then

$$R_L(\Theta,\epsilon) = R_{BL}(\Theta,\epsilon) \qquad \text{for all } \epsilon > 0. \tag{6.8}$$

The dyadic Sobolev ellipsoids $\Theta_D^\alpha(C)$ are block symmetric and so are an example for (6.8).

(ii) For general ellipsoids $\Theta(a,C)$ as in (5.2), and a block scheme (6.3), measure the *oscillation* of the weights a_l within blocks by

$$\mathrm{osc}(B_j) = \max_{l,l'\in B_j} \frac{a_l}{a_{l'}}.$$

It follows, Exercise 6.2, from the linear minimax risk formula (5.3) that if $a_k \to \infty$ and $\mathrm{osc}(B_j) \to 1$, then

$$R_L(\Theta,\epsilon) \sim R_{BL}(\Theta,\epsilon) \qquad \text{as } \epsilon \to 0. \tag{6.9}$$

In the Fourier ellipsoid case, (6.9) applies to all $\Theta(\alpha,C)$ if one uses blocks B_j defined by either $l_j = (j+1)^\beta$ for $\beta > 0$, or $l_j = e^{\sqrt{j}}$ – in either case $\mathrm{osc}\,(B_j) = (l_{j+1}/l_j)^\alpha \to 1$. The block sizes are necessarily subgeometric in growth: for dyadic blocks, $l_j = 2^j$, the condition fails: $\mathrm{osc}\,(B_j) \to 2^\alpha$.

6.3 Blockwise James Stein Estimation

We construct an estimator which on each block B_j applies the positive part James-Stein estimator (2.46):

$$\hat{\theta}_j^{JS}(y_j) = \Big(1 - \frac{(n_j-2)\epsilon^2}{\|y_j\|^2}\Big)_+ y_j. \tag{6.10}$$

A key benefit of the James-Stein estimate is the good bounds for its MSE: the consequences (2.49) and (2.52) of Proposition 2.6 show that when $n_j \ge 3$,

$$r_\epsilon(\hat{\theta}_j^{JS},\theta_j) \le 2\epsilon^2 + r_\epsilon(\hat{\theta}_j^{IS},\theta_j), \tag{6.11}$$

where we recall that the *ideal risk*

$$r_\epsilon(\hat{\theta}_j^{IS},\theta_j) = \frac{n_j\epsilon^2\|\theta_j\|^2}{n_j\epsilon^2+\|\theta_j\|^2}.$$

The full blockwise estimator, $\hat{\theta}^{BJS}$, is then defined by

$$\hat{\theta}_j^{BJS}(y) = \begin{cases} y_j & j < L \\ \hat{\theta}_j^{JS}(y_j) & L \le j < J_\epsilon \\ 0 & j \ge J_\epsilon \end{cases}. \tag{6.12}$$

For the 'earliest' blocks, specified by L, no shrinkage is performed. This may be sensible because the blocks are of small size ($n_j \le 2$), or are known to contain very strong signal, as is often the case if the blocks represent the lowest frequency components.

No blocks are estimated after J_ϵ. Usually J_ϵ is chosen so that $l_\epsilon = l_{J_\epsilon} = \epsilon^{-2}$, which equals the sample size n in the usual calibration. This restriction corresponds to not attempting to estimate, even by shrinkage, more coefficients than there is data.

It is now straightforward to combine earlier results to obtain risk bounds for $\hat{\theta}^{BJS}$ that will also show in many cases that it is asymptotically minimax.

Theorem 6.2 *In the homoscedastic white noise model, let $\hat{\theta}^{BJS}$ denote the block James-Stein estimator* (6.12).
(i) On dyadic blocks, for each $\Theta = \Theta_D^\alpha(C)$ with α and $C > 0$, the estimator $\hat{\theta}^{BJS}$ is adaptive minimax as $\epsilon \to 0$,

$$\sup_{\theta \in \Theta} r_\epsilon(\hat{\theta}^{BJS}, \theta) \sim R_N(\Theta, \epsilon). \tag{6.13}$$

(ii) For more general choices of blocks, assume that $osc(B_j) \to 1$ as $j \to \infty$ and that the block index J_ϵ in (6.12) *satisfies $J_\epsilon = o(\epsilon^{-\eta})$ for all $\eta > 0$. Then adaptive minimaxity* (6.13) *holds also for each $\Theta = \Theta^\alpha(C)$ with $\alpha, C > 0$.*

We see that, unlike the Pinsker linear minimax rule, which depended on ϵ, C and the details of the ellipsoid weight sequence (here α), the block James-Stein estimator has no adjustable parameters (other than the integer limits L and J_ϵ), and yet it can achieve asymptotically the exact minimax rate and constant for a range of values of C and α.

Proof We decompose the mean squared error by blocks,

$$r_\epsilon(\hat{\theta}^{BJS}, \theta) = \sum_j r_\epsilon(\hat{\theta}_j^{JS}, \theta_j)$$

and employ the structure of $\hat{\theta}^{BJS}$ given in (6.12). On low frequency blocks, $j < L$, the estimator is unbiased and contributes only variance terms $n_j \epsilon^2$ to MSE. On high frequency blocks, $j \ge J_\epsilon$, only a bias term $\|\theta_j\|^2$ is contributed. On the main frequency blocks, $L \le j < J_\epsilon$, we use the key bound (6.11). Assembling the terms, we find

$$r_\epsilon(\hat{\theta}^{BJS}, \theta) \le (l_L + 2J_\epsilon - 2L)\epsilon^2 + \sum_{j=L}^{J-1} r_\epsilon(\hat{\theta}_j^{IS}, \theta_j) + \sum_{l \ge l_\epsilon} \theta_l^2. \tag{6.14}$$

In view of (6.7), the first right-side sum is bounded above by the block linear minimax rule. Turning to the second sum, for any ellipsoid $\Theta(a, C)$ with $a_l \nearrow \infty$, define the

(squared) *maximal tail bias*

$$\Delta_\epsilon(\Theta) = \sup\left\{\sum_{l\ge l_\epsilon}\theta_l^2 : \sum a_l^2\theta_l^2 \le C^2\right\} = C^2 a_{l_\epsilon}^{-2}. \tag{6.15}$$

We therefore conclude that

$$\sup_\Theta r_\epsilon(\hat\theta^{BJS},\theta) \le (l_L + 2J_\epsilon)\epsilon^2 + R_{BL}(\Theta,\epsilon) + \Delta_\epsilon(\Theta). \tag{6.16}$$

Under either assumption (i) or (ii), we have as $\epsilon \to 0$ that

$$R_{BL}(\Theta,\epsilon) \sim R_L(\Theta,\epsilon) \sim R_N(\Theta,\epsilon),$$

where the first relation follows from (6.8) or (6.9) respectively, and the second relation follows from Pinsker's theorem.

Since the left side of (6.16) is, by definition, larger than $R_N(\Theta,\epsilon)$, we will be done if we show that the first and third right side terms in (6.16) are of smaller order than $R_N(\Theta,\epsilon) \asymp \epsilon^{2r}$ (with, as usual, $r = 2\alpha/(2\alpha+1)$).

For the first term, note that l_L is fixed, and that $J_\epsilon\epsilon^2 = o(\epsilon^{2-\eta})$ for each $\eta > 0$ by assumption (ii), which is also satisfied by $J_\epsilon = \log_2\epsilon^{-2}$ in the dyadic blocks case (i). Clearly we can choose η small enough that $\epsilon^{2-\eta} = O(\epsilon^{2r})$.

For the third term, since $a_l \asymp l^\alpha$ and $2\alpha > r$,

$$\Delta_\epsilon(\Theta) \le C^2 l_\epsilon^{-2\alpha} \asymp C^2(\epsilon^2)^{2\alpha} \ll \epsilon^{2r}. \qquad \square$$

For traditional Sobolev ellipsoids, dyadic blocks are too large, since with $a_l \sim l^\alpha$, osc $(B_j) \to 2^\alpha$, and so one has only rate adaptivity: $R_{BL}(\Theta,\epsilon) \le 2^{2\alpha}R_N(\Theta,\epsilon)(1+o(1))$. However part (ii) of the previous theorem shows that exact adaptation *can* be achieved with smaller block sizes, for which osc $B_j \to 1$. Thus $l_j = e^{\sqrt{j}}$ works, for example. However, the sequence $l_j = (j+1)^\beta$ is less satisfactory, since $l_{J_\epsilon} = \epsilon^{-2}$ implies that $J_\epsilon = \epsilon^{-2/\beta}$ and so $\epsilon^2 J_\epsilon$ is not $o(\epsilon^{2r})$ in the smoother cases, when $2\alpha + 1 \ge \beta$.

In fact, this last problem arises from the bound $2\epsilon^2$ in (6.11), and could be reduced by using a modified estimator $\hat\theta_j = (1 - \gamma\epsilon^2/\|y_j\|^2)y_j$ with $\gamma \in (n_j - 2, 2n_j]$. This reduces the error at zero to essentially a large deviation probability (see e.g. Brown et al. (1997), who use $\frac{3}{2}n_j$). However, in overall practical and MSE performance, the choice $n_j - 2$ has been preferred, and we have chosen to establish theoretical results for an estimator closer to that which one might use in practice.

Theorem 6.2 is an apparently more precise result than was established in 4.21 for Hölder classes, where full attention was not given to the constants. In fact the preceding argument goes through, since $\Theta_\infty^\alpha(C)$ defined in (4.53) satisfies all the required conditions, including block symmetry.

Remark. The original Efroimovich and Pinsker (1984) estimator set

$$\hat\theta_j = \left(1 - \frac{\lambda_j n_j \epsilon^2}{\|y_j\|^2}\right)_+ y_j, \qquad j \le J_\epsilon, \tag{6.17}$$

with $\lambda_j = 1 + t_j$ for $t_j > 0$. To prove adaptive minimaxity over a broad class of ellipsoids (5.2), they required in part that $n_{j+1}/n_j \to 1$ and $t_j \to 0$, but slowly enough that $\sum_j 1/(t_j^3 n_j) < \infty$. The Block James-Stein estimator (6.10) makes the particular choice

$\lambda_j = (n_j - 2)/n_j < 1$ and has the advantage that the oracle bound (6.11) deals simply with the events $\{\hat{\theta}_j = 0\}$ in risk calculations.

We will see later that the prescription (6.17) has also been used for block thresholding of wavelet coefficients, but now using larger values of λ_j, for example 4.505 in Cai (1999).

6.4 Comparing adaptive linear estimators

We now give some examples to make two points: first, that many linear smoothing methods, with their tuning parameter chosen from the data, behave substantially similarly, and second, that the Block James Stein shrinkage approach leads to one such example, whether conducted in blocks of Fourier frequencies or in a wavelet domain.

Consider the continuous Gaussian white noise model (1.18) or equivalently its sequence space counterpart (3.1) in the Fourier basis. Many standard linear estimators can be represented in this basis in the form

$$\hat{\theta}_k = \kappa(hk) y_k. \tag{6.18}$$

As examples, we cite

1. Weighted Fourier series. The function κ decreases with increasing frequency, corresponding to a downweighting of signals at higher frequencies. The parameter h controls the actual location of the "cutoff" frequency band.

2. Kernel estimators. We saw in Section 3.3 that in the time domain, the estimator has the form $\hat{\theta}(t) = \int h^{-1} K(h^{-1}(t-s)) dY(s)$, for a suitable kernel function $K(\cdot)$, typically symmetric about zero. The parameter h is the bandwidth of the kernel. The representation (6.18) follows after taking Fourier coefficients. Compare Lemma 3.5 and the examples given there.

3. Smoothing splines. We saw in Section 3.4 that the estimator $\hat{\theta}_k$ minimizes

$$\sum (y_k - \theta_k)^2 + \lambda^{2r} \sum k^{2r} \theta_k^2,$$

where the penalty term viewed in the time domain takes the form of a derivative penalty $\int (D^r f)^2$ for some integer r. In this case, $\hat{\theta}_k$ again has the representation (6.18) with $\kappa(\lambda k) = [1 + (\lambda k)^{2r}]^{-1}$.

In addition, many methods of choosing h or λ from the data y have been shown to be asymptotically equivalent to first order (see e.g. Härdle et al. (1988)) - these include cross validation, Generalized cross validation, Rice's method based on unbiased estimates of risk, final prediction error, Akaike information criterion. In this section we use a method based on an unbiased estimate of risk.

The point of the adaptivity result Theorem 6.2 however is that appropriate forms of the block James-Stein estimator should perform approximately as well as the best linear (or nonlinear) estimators, whether constructed by Fourier weights, kernels or splines, and without the need for an explicit choice of smoothing parameter from the data.

We will see this in examples below, but first we give an heuristic explanation of the close connection of these linear shrinkage families with the block James-Stein estimator (6.10). Consider a Taylor expansion of $\kappa(s)$ about $s = 0$. If the time domain kernel $K(t)$ corresponding to κ is even about 0, then the odd order terms vanish and $\kappa(s) = 1 + \kappa_2 s^2/2 +$

$\kappa_4 s^4/4! + \ldots$, so that for h small and a positive even integer q we have $\kappa(hk) = 1 - b_q h^q k^q$, compare (3.25).

Now consider grouping the indices k into blocks B_j - for example, dyadic blocks $B_j = \{k : 2^j < k \le 2^{j+1}\}$. Then the weights corresponding to two indices $k, \bar{k}$ in the same block are essentially equivalent: $k^{2r}/\bar{k}^{2r} \in [2^{-2r}, 2^{2r}]$ so that we may approximately write

$$\hat{\theta}_k \doteq (1 - c_j) y_k, \qquad k \in B_j. \tag{6.19}$$

Here c_j depends on h, but this is not shown explicitly, since we are about to determine c_j from the data y anyway.

For example, we might estimate c_j using an unbiased risk criterion, as described in Sections 2.5 and 2.6. Putting $C = (1 - c_j) I_{n_j}$ in the Mallows's C_L criterion (2.39) yields

$$U_{c_j}(y) = n_j \epsilon^2 - 2 n_j \epsilon^2 c_j + c_j^2 |y_j|^2. \tag{6.20}$$

[As noted below (2.42), this formula also follows from Stein's unbiased risk estimator applied to $\hat{\theta}_j(y) = y_j - c_j y_j$]. The value of c_j that minimizes (6.20) is $\hat{c}_j = n_j \epsilon^2 / \|y_j\|^2$, which differs from the James-Stein estimate (6.10) only in the use of n_j rather than $n_j - 2$.

Thus, many standard linear methods are closely related to the diagonal linear shrinkage estimator (6.19). In the figures below, we compare four methods:

1. *LPJS*: apply the James-Stein estimate (6.12) on each dyadic block in the Fourier frequency domain: $\hat{\theta}^{LPJS}(y) = (\hat{\theta}_j^{LPJS}(y_j))$. Dyadic blocking in the frequency domain is a key feature of Littlewood-Paley theory in harmonic analysis, hence the letters LP.
2. *WaveJS*: apply the James-Stein estimate (6.12) on each dyadic block in a wavelet coefficient domain: the blocks $y_j = (y_{jk}, k = 1, \ldots, 2^j)$.
3. *AutoSpline*: Apply a smoothing spline for the usual energy penalty $\int (f'')^2$ using a regularization parameter $\hat{\lambda}$ chosen by minimizing an unbiased estimator of risk.
4. *AutoTrunc*: In the Fourier frequency domain, use a cutoff function: $\hat{\kappa}(hl) = I\{l \le [h^{-1}]\}$ and choose the location of the cutoff by an unbiased risk estimator.

Implementation details. Let the original time domain data be $Y = (Y(l), l = 1, \ldots, N)$ for $N = 2^J$. The discrete Fourier transform (DFT), e.g. as implemented in MATLAB, sets

$$y(\nu) = \sum_{l=1}^{N} Y(l) e^{2\pi i (l-1)(\nu-1)/N}, \qquad \nu = 1, \ldots, N. \tag{6.21}$$

If the input Y is real, the output $y \in \mathbb{C}^N$ must have only N (real) free parameters. Indeed $y(1) = \sum_1^N Y(l)$ and $y(N/2+1) = \sum_1^N (-1)^l Y(l)$ are real, and for $r = 1, \ldots, N/2 - 1$, we have conjugate symmetry

$$y(N/2 + 1 + r) = \overline{y(N/2 + 1 - r)}. \tag{6.22}$$

Thus, to build an estimator, one can specify how to modify $y(1), \ldots, y(N/2+1)$ and then impose the constraints (6.22) before transforming back to the time domain by the inverse DFT.

1. (LPJS). Form dyadic blocks

$$y_j = \{\mathrm{Re}(y(\nu)), \mathrm{Im}(y(\nu)) : 2^{j-1} < \nu \le 2^j\}$$

for $j = 2, \ldots, J-1$. Note that $n_j = \#(y_j) = 2^j$. Apply the James Stein estimator (6.10) to each y_j, while leaving $y(\nu)$ unchanged for $\nu = 0, 1, 2$. Thus $L = 2$, and we take $\epsilon^2 = (N/2)\sigma^2$, in view of (6.35).

3. (Autospline). We build on the discussion of periodic splines in Section 3.4. There is an obvious relabeling of indices so that in the notation of this section, $\nu = 1$ corresponds to the constant term, and

each $\nu > 1$ to a pair of indices $2(\nu-1)-1$ and $2(\nu-1)$. Hence, linear shrinkage takes the form $\hat{\theta}_\lambda(\nu) = c_\nu(\lambda)y(\nu)$ with

$$c_\nu(\lambda) = [1+\lambda(\nu-1)^4]^{-1}.$$

Note that $c_\nu(\lambda)$ is real and is the same for the "cosine" and "sine" terms. We observe that $c_1(\lambda) = 1$ and decree, for simplicity, that $c_{N/2+1}(\lambda) = 0$. Then, on setting $d_\nu = 1 - c_\nu$ and applying Mallow's C_L formula (2.39), we get an unbiased risk criterion to be minimized over λ:

$$U(\lambda) = n + \sum_{\nu=2}^{N/2} d_\nu(\lambda)^2|y(\nu)|^2 - 4d_\nu(\lambda),$$

4. (AutoTruncate). The estimator that cuts off at frequency ν_0 is, in the frequency domain,

$$\hat{\theta}_{\nu_0}(\nu) = \begin{cases} y(\nu) & \nu \le \nu_0 \\ 0 & \nu > \nu_0. \end{cases}$$

Using Mallows C_p, noting that each frequency ν corresponds to *two* real degrees of freedom, and neglecting terms that do not change with ν_0, we find that the unbiased risk criterion has the form

$$U_{\nu_0}(y) \leftrightarrow 4\nu_0 + \sum_{\nu_0+1}^{N/2} |y(\nu)|^2, \qquad \nu_0 \in \{1,\dots,N/2\}.$$

2. (WaveJS). Now we use a discrete wavelet transform instead of the DFT. Anticipating the discussion in the next chapter, Y is transformed into wavelet coefficients $(y_{jk}, j = L,\dots,J-1, k = 1,\dots,2^j)$ and scaling coefficients $(\tilde{y}_{Lk}, k = 1,\dots,2^L)$. We use $L = 2, J = J_\epsilon$ and the Symmlet 8 wavelet, and apply Block James Stein to the blocks $y_j = (y_{jk} : k = 1,\dots,2^j)$, while leaving the scaling coefficients $\tilde{y}_L$ unchanged.

These are applied to two examples: (a) the minimum temperature data introduced in Section 1.1, and (b) a 'blocky' step function with simulated i.i.d. Gaussian noise added. The temperature data has correlated noise, so our theoretical assumptions don't hold exactly. Indeed, one cas see the different noise levels in each wavelet band (cf Chapter 7.5). We used an upper bound of $\hat{\sigma} = 5$ in all cases. Also, the underlying function is not periodic over this range and forcing the estimator to be so leads to somewhat different fits than in Figure 1.1; the difference is not central to the discussion in this section.

The qualitative similarity of the four smoothed temperature fits is striking: whether an unbiased risk minimizing smoothing parameter is used with splines or Fourier weights, or whether block James-Stein shrinkage is used in the Fourier or wavelet domains. The similarity of the linear smoother and block James-Stein fits was at least partly explained near (6.19).

The similarity of the Fourier and wavelet James-Stein reconstructions may be explained as follows. The estimator (6.19) is invariant with respect to orthogonal changes of basis for the vector $y_j = (y_k : k \in B_j)$. To the extent that the frequency content of the wavelets spanning the wavelet multiresolution space W_j is concentrated on a single frequency octave (only true approximately), it represents an orthogonal change of basis from the sinusoids belonging to that octave. The James-Stein estimator (6.10) is invariant to such orthogonal basis changes.

The (near) linear methods that agree on the temperature data also give similar, but now unsatisfactory, results on the 'Blocky' example. Note that none of the methods are effective at simultaneously removing high frequency noise *and* maintaining the sharpness of jumps and peaks.

It will be the task of the next few chapters to explain why the methods fail, and how wavelet thresholding can succeed. For now, we just remark that the blocky function, which evidently fails to be differentiable, does not belong to any of the ellipsoidal smoothing classes $\Theta_2^\alpha(C)$ for $\alpha \geq 1/2$ (based on the expectation that the Fourier coefficients decay at rate $O(1/k)$). Hence the theorems of this and the previous chapter do not apply to this example.

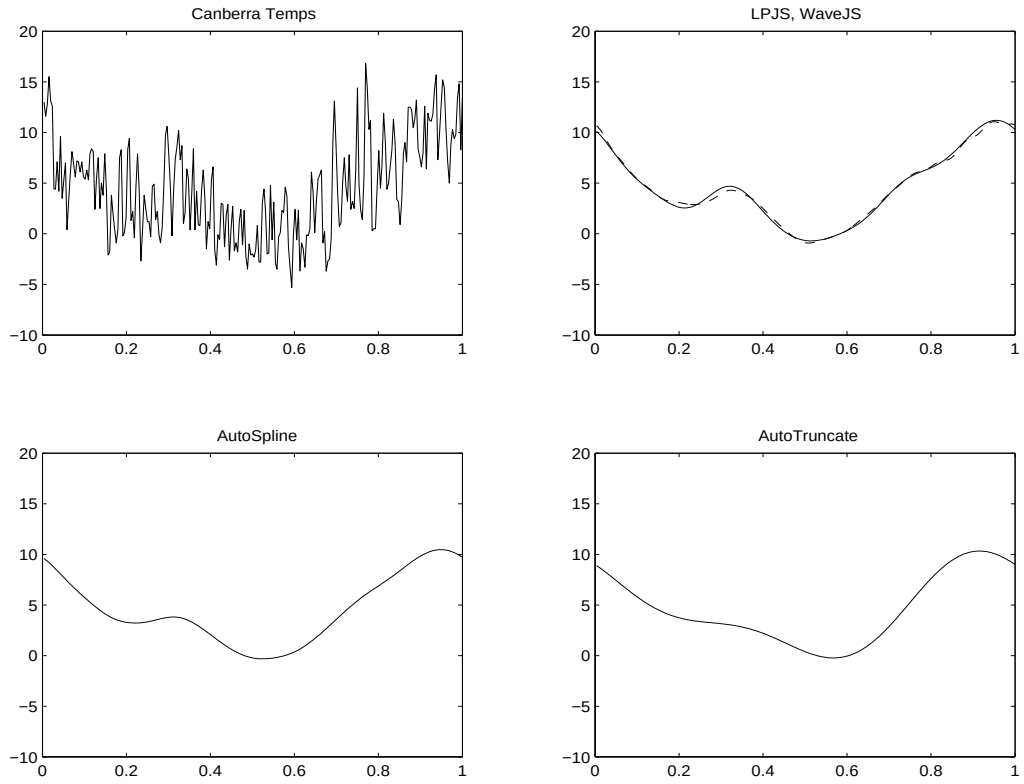

Figure 6.1 Top left: Canberra temperature data from Figure 1.1. Top right: block James-Stein estimates in the Fourier (solid) and wavelet (dashed) domains. Bottom panels: linear spline and truncation smoothers with bandwidth parameter chosen by minimizing an unbiased risk criterion.

6.5 *Interlude:* Superefficiency

This section looks at *one* of the motivations that underlies the use of worst-case and minimax analyses: a desire for a robust alternative to "fixed θ" asymptotics. In fixed θ asymptotics, the unknown function θ is kept fixed, and the risk behavior of an estimator sequence $\hat{\theta}_\epsilon$ is analysed as $\epsilon \to 0$. Asymptotic approximations might then be used to optimize parameters of the estimator – such as bandwidths or regularization parameters – or to assert optimality properties.

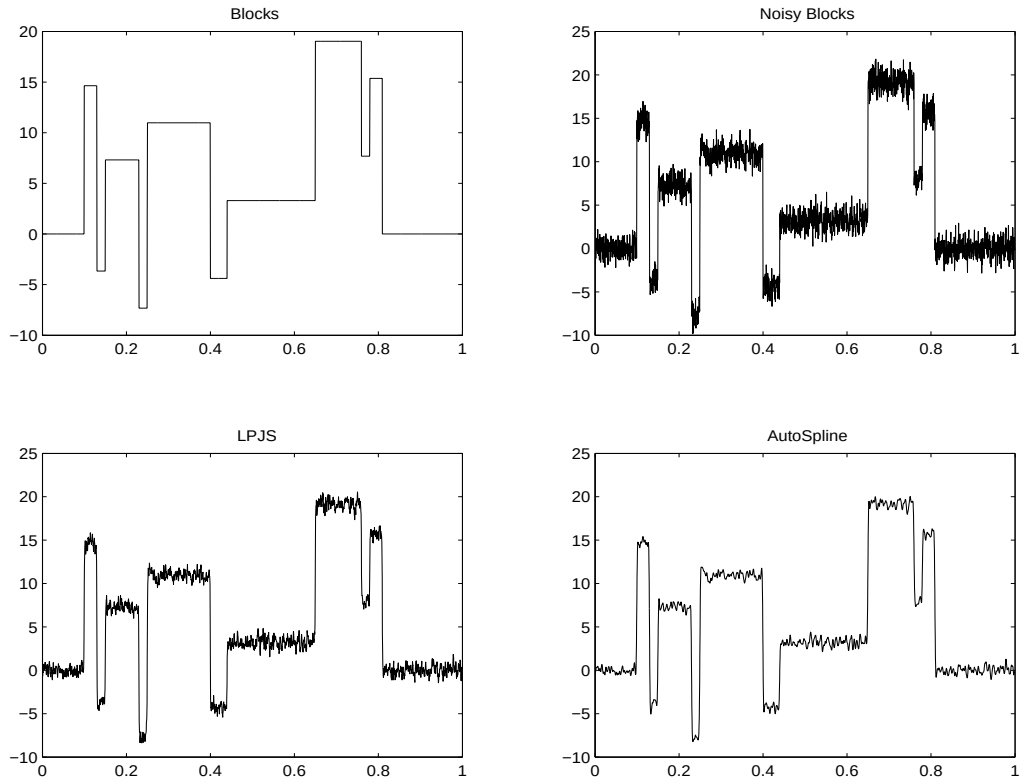

Figure 6.2 Top panels: A "blocky" step function with i.i.d Gaussian noise added, N = 2048. Bottom panels: selected reconstructions by block James-Stein and by smoothing spline (with data determined λ fail to remove all noise.

This mode of analysis has been effective in large sample analysis of finite dimensional models. Problems such as superefficiency are not serious enough to affect the practical implications widely drawn from Fisher's asymptotic theory of maximum likelihood.

In nonparametric problems with infinite dimensional parameter spaces, however, fixed θ asymptotics is more fragile. Used with care, it yields useful information. However, if optimization is pushed too far, it can suggest conclusions valid only for implausibly large sample sizes, and misleading for actual practice. In nonparametrics, superefficiency is more pervasive: even practical estimators can exhibit superefficiency at *every* parameter point, and poor behaviour in a neighbourhood of *any* fixed parameter point is a necessary property of *every* estimator sequence.

After reviewing Hodges' classical example of parametric superefficiency, we illustrate these points, along with concluding remarks about worst-case and minimax analysis.

Parametric Estimation: the Hodges example.

Suppose that $y \sim N(\theta, \epsilon^2)$ is a single scalar observation with ϵ small. A rather special case of Fisherian parametric asymptotics asserts that if $\hat{\theta}_\epsilon$ is an asymptotically normal and unbiased estimator sequence, $\epsilon^{-1}(\hat{\theta}_\epsilon - \theta) \stackrel{\mathcal{D}}{\to} N(0, v(\theta))$ when θ is true, *then* necessarily $v(\theta) \geq 1$. A consequence for mean squared error would then be that

$$\liminf_{\epsilon \to 0} \epsilon^{-2} E_\theta (\hat{\theta}_\epsilon - \theta)^2 = \liminf r_\epsilon(\hat{\theta}, \theta)/R_N(\Theta, \epsilon) \geq 1.$$

[For this subsection, $\Theta = \mathbb{R}$.] Hodges' counterexample modifies the MLE $\hat{\theta}(y) = y$ in a shrinking neighborhood of a single point:

$$\hat{\theta}_\epsilon(y) = \begin{cases} 0 & |y| < \sqrt{\epsilon} \\ y & \text{otherwise.} \end{cases}$$

Since $\sqrt{\epsilon} = \frac{1}{\sqrt{\epsilon}} \cdot \epsilon$ is many standard deviations in size, it is clear that *if* $\theta = 0$, this estimator has MSE equal to $2\epsilon^2 \int_{\epsilon^{-1/2}}^{\infty} y^2 \phi(y) dy << \epsilon^2$. On the other hand, if $\theta \neq 0$ and ϵ is small, and noting the rapid decay of the tails of the Gaussian distribution, then the interval $[-\sqrt{\epsilon}, \sqrt{\epsilon}]$ is essentially irrelevant to estimation of θ, and so

$$\epsilon^{-2} E_\theta (\hat{\theta}_\epsilon - \theta)^2 \to \begin{cases} 0 & \text{if } \theta = 0, \\ 1 & \text{otherwise,} \end{cases}$$

in clear violation of the Fisherian program. A fuller introduction to this and related superefficiency issues appears in Lehmann and Casella (1998, Section 6.2), Here we note two phenomena which are also characteristic of more general parametric settings:

(i) points of superefficiency are *rare*: in Hodges' example, only at $\theta = 0$. More generally, for almost all θ,

$$\liminf_{\epsilon \to 0} \frac{r_\epsilon(\hat{\theta}_\epsilon, \theta)}{R_N(\Theta, \epsilon)} \geq 1. \tag{6.23}$$

(ii) Superefficiency entails poor performance at nearby points. For Hodges' example, consider $\theta_\epsilon = \sqrt{\epsilon}/2$. Since the threshold zone extends $1/(2\sqrt{\epsilon})$ standard deviations to the right of θ_ϵ, it is clear that $\hat{\theta}_\epsilon$ makes a squared error of $(\sqrt{\epsilon}/2)^2$ with high probability, so $\epsilon^{-2} r(\hat{\theta}_\epsilon, \sqrt{\epsilon}/2) \doteq \epsilon^{-2}(\sqrt{\epsilon}/2)^2 \to \infty$. Consequently

$$\sup_{|\theta| \leq \sqrt{\epsilon}} \frac{r(\hat{\theta}_\epsilon, \theta)}{R_N(\Theta, \epsilon)} \to \infty. \tag{6.24}$$

LeCam, Huber and Hajek showed that more generally, superefficiency at θ_0 forces poor properties in a neighborhood of θ_0. Since broadly efficient estimators such as maximum likelihood are typically available with good risk properties, superefficiency has less relevance in parametric settings.

Remark. Hodges' estimator is an example of hard thresholding, to be discussed in some detail for wavelet shrinkage in non-parametric estimation. It is curious that the points of superefficiency that are unimportant for the one-dimensional theory become essential for sparse estimation of high dimensional signals.

Nonparametrics: Superefficiency everywhere

We return to the nonparametric setting, always in the Gaussian sequence model. Previous sections argued that the dyadic blocks James-Stein estimate (cf. (6.12) and Theorem 6.2(i) is a theoretically and practically promising method. Nevertheless, every fixed θ is a point of superefficiency in the sense of (6.23):

Proposition 6.3 *(Brown et al., 1997) Let $\Theta = \Theta^\alpha(C)$ be a Sobolev ellipsoid* (6.1)*. Then for every $\theta \in \Theta$,*

$$\frac{r_\epsilon(\hat{\theta}^{BJS}_\epsilon, \theta)}{R_N(\Theta, \epsilon)} \to 0. \tag{6.25}$$

Thus, if Θ corresponds to functions with second derivative ($m = 2$) having L_2 norm bounded by 1, say, then for *any* fixed such function, the blockwise James-Stein estimator has rate of convergence faster than $\epsilon^{8/5} \leftrightarrow n^{-4/5}$. Brown et al. (1997) also show that convergence cannot, in general, be very much faster – at best of logarithmic order in ϵ^{-1} – but the fixed θ rate is always slightly different from that of a natural minimax benchmark. Of course, in parametric problems, the rate of convergence is the same at almost all points.

Proof Fix $\Theta = \Theta^\alpha(C)$ and recall from (5.10) that $R_N(\Theta, \epsilon) \asymp \epsilon^{2r}$ as $\epsilon \to 0$, with $r = 2\alpha/(2\alpha + 1)$. A "fixed θ" bound for the risk of $\hat{\theta}^{BJS}$ follows from (6.14) : indeed, since $L = 2$ and $ab/(a + b) \le \min(a, b)$, we may write

$$r_\epsilon(\hat{\theta}^{BJS}, \theta) \le 2J_\epsilon\epsilon^2 + \sum_j \min(n_j\epsilon^2, \|\theta_j\|^2) + \sum_{l > \epsilon^{-2}} \theta_l^2.$$

The proof of Theorem 6.2 showed that the first and third terms were $o(\epsilon^{2r})$, uniformly over $\theta \in \Theta$. Consider, therefore, the second term, which we write as $R_1(\theta, \epsilon)$. For any j_ϵ, use the variance component below j_ϵ and the bias term thereafter:

$$R_1(\theta, \epsilon) \le 2^{j_\epsilon}\epsilon^2 + 2^{-2\alpha j_\epsilon} \sum_{j \ge j_\epsilon} 2^{2\alpha j} \|\theta_j\|^2.$$

To show that $R_1(\theta, \epsilon) = o(\epsilon^{2r})$, first fix a $\delta > 0$ and then choose j_ϵ so that $2^{j_\epsilon}\epsilon^2 = \delta\epsilon^{2r}$. [Of course, j_ϵ should be an integer, but there is no harm in ignoring this point.] It follows that $2^{-2\alpha j_\epsilon} = \delta^{-2\alpha}\epsilon^{2r}$, and so

$$\epsilon^{-2r} R_1(\theta, \epsilon) \le \delta + \delta^{-2\alpha} \sum_{j \ge j_\epsilon} 2^{2\alpha j} \|\theta_j\|^2 = \delta + o(1),$$

since the tail sum vanishes as $\epsilon \to 0$, for $\theta \in \Theta^\alpha(C)$. Since $\delta > 0$ is arbitrary, this shows that $R_1(\theta, \epsilon) = o(\epsilon^{2r})$ and establishes (6.25). □

The next result shows that for every consistent estimator sequence, and every parameter point $\theta \in \ell_2$, there exists a *shrinking* ℓ_2 neighborhood of θ over which the worst case risk of the estimator sequence is arbitrarily worse than it is at θ itself. Compare (6.24). In parametric settings, such as the Hodges example, this phenomenon occurs only for unattractive, superefficient estimators, but in nonparametric estimation the property is ubiquitous. Here,

neighborhood refers to balls in ℓ_2 norm: $B(\theta_0, \eta) = \{\theta : \|\theta - \theta_0\|_2 < \eta\}$. Such neighborhoods do not have compact closure in ℓ_2, and fixed θ asymptotics does not give any hint of the perils that lie arbitrarily close nearby.

Proposition 6.4 *Suppose that $\hat{\theta}_\epsilon$ is any estimator sequence such that $r_\epsilon(\hat{\theta}_\epsilon, \theta_0) \to 0$. Then there exists $\eta_\epsilon \to 0$ such that as $\epsilon \to 0$,*

$$\sup_{\theta \in B(\theta_0, \eta_\epsilon)} \frac{r_\epsilon(\hat{\theta}_\epsilon, \theta)}{r_\epsilon(\hat{\theta}_\epsilon, \theta_0)} \to \infty. \tag{6.26}$$

Remark. The result remains true if the neighborhood $B(\theta_0, \eta_\epsilon)$ is replaced by its intersection with any dense set: for example, the class of infinitely differentiable functions.

Proof Let $\gamma_\epsilon^2 = r_\epsilon(\hat{\theta}_\epsilon, \theta_0)$: we show that $\eta_\epsilon = \sqrt{\gamma_\epsilon}$ will suffice for the argument. The proof is a simple consequence of the fact that $\overline{B(1)} = \{\theta : \|\theta\|_2 \le 1\}$ is not compact (compare Theorem 5.5 or the example following Theorem 4.22), so that $R_N(B(1), \epsilon) \ge c_0 > 0$ even as $\epsilon \to 0$. All that is necessary is to rescale the estimation problem by defining $\bar{\theta} = \eta_\epsilon^{-1}(\theta - \theta_0)$, $\bar{y} = \eta_\epsilon^{-1}(y - \theta_0)$, $\bar{\epsilon} = \eta_\epsilon^{-1}\epsilon$, and so on. Then $\bar{y} = \bar{\theta} + \bar{\epsilon} z$ is an instance of the original Gaussian sequence model, and $\overline{B(\theta_0, \eta_\epsilon)}$ corresponds to the unit ball $\overline{B(1)}$. Rescaling the estimator also via $\hat{\bar{\theta}}_\epsilon(\bar{y}) = \eta_\epsilon^{-1}\big(\hat{\theta}_\epsilon(y) - \theta_0\big)$,

$$\gamma_\epsilon^{-2} E\|\hat{\theta}_\epsilon - \theta\|^2 = \eta_\epsilon^2 \gamma_\epsilon^{-2} E_{\bar{\epsilon}} \|\hat{\bar{\theta}}_\epsilon(\bar{y}) - \bar{\theta}\|^2,$$

and so, writing S_ϵ for the left side of (6.26), we obtain

$$S_\epsilon \ge \gamma_\epsilon^{-1} R_N(B(1), \epsilon) \ge c_0 \gamma_\epsilon^{-1} \to \infty. \qquad \square$$

Ultra-asymptotic bandwidth selection

Here is a "fixed-f" argument often encountered in asymptotics. Consider kernel estimators and the equispaced regression model discussed in Section 3.4. Using a qth order kernel, (3.21), in estimate $\hat{f}_h$, (3.12), leads to an approximate MSE expression, (3.23), of the form

$$r_a(h) = c_0(K)(nh)^{-1} + c_1(K) h^{2q} \int (D^q f)^2 \tag{6.27}$$

Then $r_a(h)$ is minimized at a bandwidth $h = h_n(f)$, and the minimum value $r_a(h_n(f))$ converges to zero at rate $n^{-2q/(2q+1)}$. Since $h_n(f)$ still depends on the unknown function f, the "plug-in" approach inserts a preliminary estimator $\tilde{f}_n$ of f, and uses $h_n(\tilde{f}_n)$ in the kernel estimate, such as (3.12) or (3.15). This approach goes back at least to Woodroofe (1970), for further references and discussion see Brown et al. (1997).

We study a version of this argument in the sequence model (3.1), which allows *exact* calculation of the small sample consequences of this asymptotic bandwidth selection argument. We use the Fourier basis with $\mathbb{Z}$ as index, so that positive integers l label cosine terms of frequency l and negative l label the sine terms, so that

$$f(t) = \sum_{l \ge 0} \theta_l \cos 2\pi l t + \sum_{l < 0} \theta_l \sin 2\pi l t \tag{6.28}$$

As in Section 3.3 and 6.4, represent a kernel estimator in the Fourier domain by diagonal shrinkage

$$\hat{\theta}_{h,l} = \kappa(hl) y_l, \tag{6.29}$$

where $\kappa(s) = \int e^{-ist} K(t) dt$ is the Fourier transform of kernel K. The $q-$th order moment condition becomes a statement about derivatives at zero, cf. (3.25). To simplify calculations, we use a specific choice of $q-$th order kernel:

$$\kappa(s) = (1 - |s|^q)_+. \tag{6.30}$$

For this kernel, the mean squared error of (6.29) can be written explicitly as

$$r_\epsilon(\hat{\theta}_h, \theta) = \sum_{|l| \le [h^{-1}]} \epsilon^2 (1 - |hl|^q)^2 + |hl|^{2q} \theta_l^2 + \sum_{|l| > [h^{-1}]} \theta_l^2. \tag{6.31}$$

Integral approximations to sums yield an asymptotic approximation to (6.31):

$$r_{a,\epsilon}(\hat{\theta}_h, \theta) = a_q \epsilon^2 h^{-1} + b_q(\theta) h^{2q},$$

which is exactly analogous to (6.27). Here $a_q = 4q^2(2q+1)^{-1}(q+1)^{-1}$, and $b_q(\theta) = \sum l^{2q} \theta_l^2$, is proportional to $\int (D^q f)^2$ when expressed in terms of f. In order that $b_q(\theta) < \infty$ for all q, we assume that f is infinitely differentiable. The asymptotically MSE-optimal bandwidth is found by minimizing $h \to r_{a,e}(\hat{\theta}_h, \theta)$. The Variance-Bias Lemma 3.8 gives

$$h_\epsilon = h_\epsilon(\theta) = \left[\frac{a_q \epsilon^2}{2q b_q(\theta)} \right]^{1/(2q+1)}, \tag{6.32}$$

and corresponding MSE

$$r_\epsilon(\hat{\theta}_{h_\epsilon(\theta)}, \theta) \sim c_q \left(2q b_q(\theta)\right)^{1/(2q+1)} \left(a_q \epsilon^2\right)^{2q/(2q+1)}, \tag{6.33}$$

with $c_q = 1 + (2q)^{-1}$. Thus the rate of convergence, $2q/(2q+1)$, reflects only the order of the kernel used and nothing of the properties of f. Although this already is suspicious, it would *seem*, so long as f is smooth, that the rate of convergence can be made arbitrarily close to 1, by using a kernel of sufficiently high order q.

However, this is an over literal use of fixed θ asymptotics – a hint of the problem is already suggested by the constant term in (6.33), which depends on $b_q(\theta)$ and could grow rapidly with q. However, we may go further and do exact MSE calculations with formula (6.31) using kernel (6.30). As specific test configurations in (6.28) we take

$$\theta_l = c(l_1, l_2) \begin{cases} |l|^{-3} & l \text{ even}, l \in [l_1, l_2] \\ |l|^{-3} & l \text{ odd}, -l \in [l_1, l_2] \\ 0 & \text{otherwise}, \end{cases} \tag{6.34}$$

and with $c(l_1, l_2)$ chosen so that a Sobolev 2nd derivative smoothness condition holds: $\sum l^4 \theta_l^2 = C^2$. Two choices are

$$\begin{aligned} &\text{(I)} \quad l_1 = 4, \quad l_2 = 20, \quad C = 60, \\ &\text{(II)} \quad l_1 = 4, \quad l_2 = 400, \quad C = 60. \end{aligned}$$

which differ only in the number of high frequency terms retained.

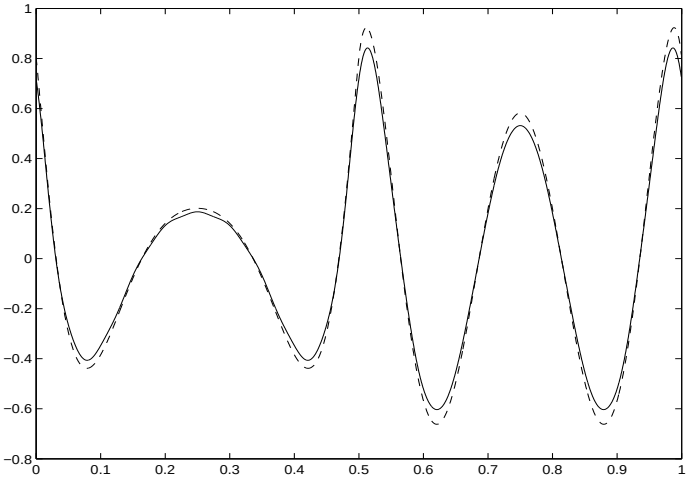

Figure 6.3 Two C^∞ functions, defined at (6.28) - (6.34). Solid line is θ^I, containing frequencies only through $l = 20$, dashed line is θ^{II}, with frequences up to $l = 400$.

Figure 6.4 shows the MSE $r_\epsilon(\hat{\theta}_{h_\epsilon(\theta^{II})}, \theta^{II})$ occasioned by using the $q-$th order optimal bandwidth (6.32) for $q = 2, 4, 8$ with exact risks calculated using (6.31). Clearly the 8th order kernel is always several times worse than the 2nd order kernel for $n = \epsilon^{-2}$ less than 10^6. The 4th order kernel will dominate $q = 2$ for n somewhat larger than 10^6, but $q = 8$ will dominate only at absurdly large sample sizes.

Figure 6.5 shows that the situation is not so bad in the case of curve I : because the higher frequencies are absent, the variance term in (6.31) is not so inflated in the $q = 8$ case.

However, with moderate noise levels ϵ, a test would not be able to discriminate beween θ^I and θ^{II}. This is an instance of the nearby instability of MSE seen earlier in this section.

We can also use (6.32) to compute the relative size of optimal bandwidths for the two functions, using $R_q = h_{\epsilon,q}(\theta_1)/h_{\epsilon,q}(\theta_2)$ as a function of q. Indeed, for $q = 2, 4, 8$, one computes that $R_q = 1, 2.6$ and 6.8.

Thus, at least for $q > 2$, both $h_\epsilon(\theta)$ and $r(\hat{\theta}_{h_\epsilon}, \theta)$ are very sensitive to aspects of the function that are difficult or impossible to estimate at small sample sizes. The fixed θ expansions such as (6.27) and (6.33) are potentially unstable tools.

Remarks. 1. Block James Stein estimation. Figures 6.4 and 6.5 also show the upper bounds (6.14) for the MSE of the dyadic blocks James-Stein estimator, and it can be seen that its MSE performance is generally satisfactory, and close to the $q = 2$ kernel over small sample sizes. Figure 6.6 compares the ratio $r_\epsilon(\hat{\theta}^{BJS}, \theta)/r_\epsilon(\hat{\theta}^q, \theta)$ of the Block JS mean squared error to the $q-$th order kernel MSE over a much larger range of $n = \epsilon^{-2}$. The James Stein MSE *bound* is never much worse than the MSE of the $q-$th order optimal bandwidth, and in many cases is much better.

2. Smoothness assumptions. Since θ^I and θ^{II} have finite Fourier expansions, they are certainly C^∞, but here they behave more like functions with about *two* square summable

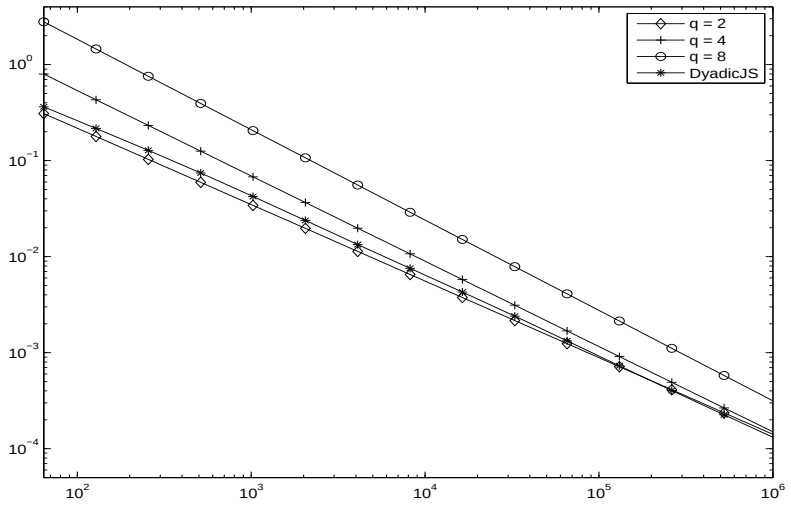

Figure 6.4 MSE of ideal bandwidth choice for θ^{II} : $r_\epsilon(\hat{\theta}_{h_\epsilon(\theta^{II})}, \theta^{II})$ resulting from $q-$th order optimal bandwidth (6.32) for $q = 2, 4, 8$ with exact risks calculated using (6.31). Also shown is the upper bound (6.14) for the risk of the dyadic blocks James Stein estimator (6.12).

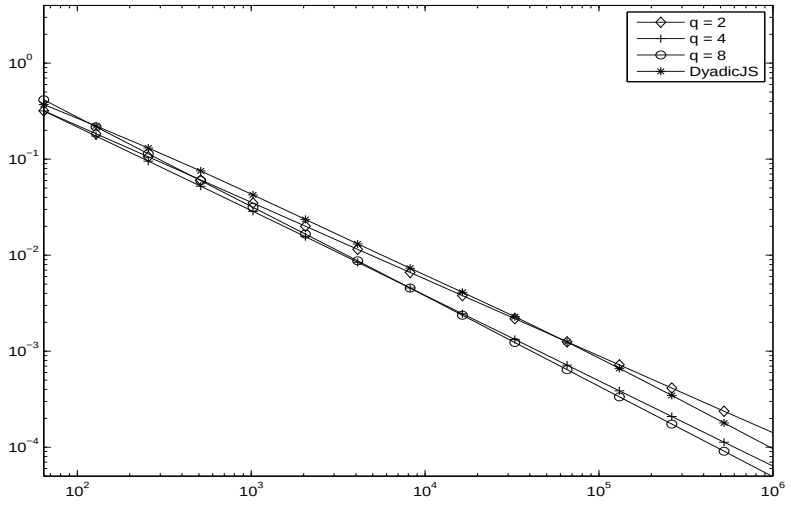

Figure 6.5 Corresponding plot of MSEs and James-Stein bound for ideal bandwidth choice for θ^I.

derivatives. Thus from the adaptivity Theorem 6.2, for α large, one expects that Block JS should eventually improve on the $q = 4$ and $q = 8$ kernels, and this indeed occurs in Figure 6.6 on the right side of the plot. However, the huge sample sizes show this "theoretical" to be impractical. Such considerations point toward the need for quantitative measures of

smoothness—such as Sobolev or Besov norms—that combine the *sizes* of the individual coefficients rather than qualitative hypotheses such as the mere existence of derivatives.

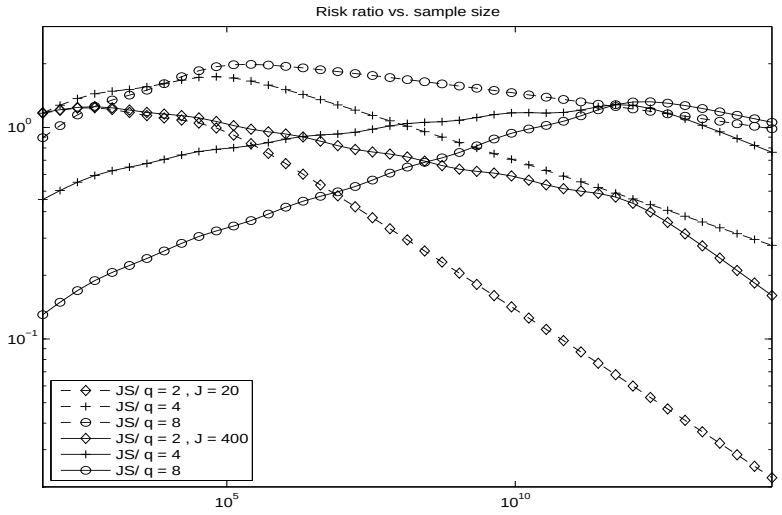

Figure 6.6 Ratio of James Stein MSE bound to actual MSE for kernels of order $q = 2, 4, 8$ at $\theta = \theta^I$ (dotted) and θ^{II} (solid) over a wide range of sample sizes $n = \epsilon^{-2}$.

3. Speed limits. There is a uniform version of (6.33) that says that over ellipsoids of functions with α mean-square derivatives, the uniform rate of convergence using the q-th order kernel is at best $(\epsilon^2)^{2q/(2q+1)}$, no matter how large α is. By contrast, the adaptivity results of Theorem 6.2 (and its extensions) for the block James-Stein estimate show that it suffers no such speed limit, and so might effectively be regarded as acting like an infinite order kernel. (Exercise 1 below has further details.)

Concluding discussion. Worst case analysis is, in a way, the antithesis of fixed θ analysis. The least favorable configuration—whether parameter point θ_ϵ or prior distribution π_ϵ—will generally change with noise level ϵ. This is natural, since the such configurations represent the "limit of resolution" attainable, which improves as the noise diminishes.

The choice of the space Θ to be maximized over is certainly critical, and greatly affects the least favorable configurations found. This at least has the virtue of making clearer the consequences of assumptions—far more potent in nonparametrics, even if hidden. It might be desirable to have some compromise in between the local nature of fixed θ asymptotics, and the global aspect of minimax analysis—perhaps in the spirit of the local asymptotic minimax approach used in parametric asyptotics. Nevertheless, if one can construct estimators that deal successfully with many least favorable configurations from the global minimax framework—as in the blockwise James-Stein constructions—then one can have some degree of confidence in such estimators for practical use in settings not too distant from the assumptions.

6.6 Discussion

[NEEDS REVISION.] **Visualizing least favorable distributions.** Pinsker's theorem gives an explicit construction of the asymptotically least favorable distribution associated with the ellipsoid $\Theta = \{\theta : \sum a_i^2\theta_i^2 \le C^2\}$: simply take independent variables $\theta_i \sim N(0, \tau_i^2)$, with τ_i given by (5.6). Recalling that the θ_i can be thought of as coefficients of the unknown function in an orthonormal basis $\{\varphi_i\}$ of $L_2[0,1]$, it is then instructive to plot sample paths from the random function

$$X(t) = \sum \theta_i \varphi_i(t).$$

Figure ??? shows two such sample paths, corresponding to smoothness $m = 1$ and $m = 2$ respectively (and with $\epsilon = 2^{-6}$ and $2^{-6.5}$ respectively). [In fact, the pictures were generated using a wavelet basis, and coefficient sequence $\bar{a}_l = 2^{m[\log_2 l]}$, but since $\bar{a}_l/a_l \in [2^{-m}, 1]$ relative to the trigonometric basis weight sequence (3.8), this has little influence on our qualitative conclusions – see Johnstone (1994) for details.]

Notice the spatial homogeneity of the sample paths – even though the smoothness of the paths is, as expected, very different in the two cases, the degree of oscillation within each figure is essentially constant as one moves from left to right in the domain of the function.

Challenges to the ellipsoid model.

Of course, not all signals of scientific interest will necessarily have this spatial homogeneity:

Ex: NMR spectrum of tryptophan in heavy water from DJHS

Ex: plethysmograph signal from Nason & Silverman

In each case, there are regions of great "activity" or "oscillation" in the signal, and other regions of relative smoothness.

Thus, by comparing sample paths from the Gaussian priors with the data examples, one naturally suspects that the ellipsoid model is not relevant in these cases, and to ask whether linear estimators are likely to perform near optimally (and in fact, they don't).

Another implicit challenge to the ellipsoid model and the fixed bandwidth smoothers implied by (5.5) begins to appear in the methodological and applied statistical literature at about the same time as Pinsker (1980). Cleveland (1979) investigates local smoothing, and Friedman and Stuetzle (1981), in describing the univariate smoother they constructed for projection pursuit regression say explicitly "the actual bandwidth used for local averaging at a particular value of (the predictor) can be larger or smaller than the average bandwidth. Larger bandwidths are used in regions of high local variablility of the response."

Commentary on the minimax approach. One may think of minimax decision theory as a strategy for evaluating the consequences of assumptions - the sampling model, loss function, and particularly the structure of the postulated parameter space Θ. The results of a minimax solution consist, of course, of the minimax value, the minimax strategy, the least favorable prior, and also, information gained in the course of the analysis.

A particular feature of least favorable distributions is that they indicate the "typical enemy" corresponding to the parameter space Θ chosen. The least favorable prior avoids both the arbitrariness of an effort to choose a single "representative" function, and yet focuses attention on elements of Θ that are "relevant" to the minimax problem. [Of course, whether

the "typical enemies" thus exhibited are *scientifically* relevant depends on the particular application.]

The minimax strategy is can be successful if the structure of Θ is intellectually and/or scientifically significant, *and* if it is possible to get close enough to a solution of the resulting minimax problem that some significant and interpretable structure emerges.

Pinsker's theorem is an outstanding success for the approach, since it yields an (asymptotically) sharp solution, along with the important structure of linear estimators, independent Gaussian least favorable priors, decay of shrinkage weights with frequency to a finite cutoff, and so on.

The clarity of the solution, paradoxically, also reveals some limitations of the result, or rather, of the formulation. The juxtaposition of the Pinsker priors and particular datasets suggests that for some scientific problems, one needs richer models of parameter spaces than ellipsoids (and their quadratically convex relatives.) This is one motivation for the introduction of Besov (and Triebel) bodies in Chapter 9.6 below.

6.7 Notes

van der Vaart (1997) gives a review of the history and proofs around superefficiency.

The exact risk analysis in § 6.5 is inspired by the study of density estimation in Marron and Wand (1992), which in turn cites Gasser and Müller (1984).

Of course, the density estimation literature also cautions against the use of higher order ($q > 2$) kernels due to these poor finite sample properties. We did not even attempt to consider the behavior of "plug-in" methods that attempt to estimate $h_\epsilon(\theta)$ – variability in the data based estimates of $h_\epsilon(\theta)$ would of course also contribute to the overall mean squared error. Loader (1999) provides a somewhat critical review of "plug-in" methods in the case $q = 2$.

While the choice $q = 8$ may seem extreme in the setting of traditional density estimation, it is actually standard to use wavelets with higher order vanishing moments – for example, the Daubechies Symmlet 8 discussed in Daubechies (1992, p. 198-199) or Mallat (1998, p. 252). Analogs of (6.27) and (6.33) for wavelet based density estimates appear in Hall and Patil (1993), though of course these authors do not use the expansions for bandwidth selection.

Exercises

6.1 *(Speed limits for q-th order kernels.)*

We have argued that in the Gaussian sequence model in the Fourier basis, it is reasonable to think of a kernel estimate with bandwidth h as represented by $\hat{\theta}_{h,l} = \kappa(hl)y_l$.

(a) Explain why it is reasonable to express the statement "K is a $q-$th order kernel," $q \in \mathbb{N}$, by the assumption $\kappa(s) = 1 - c_q s^q + o(s^q)$ as $s \to 0$ for some $c_q \neq 0$.

(b) Let $\Theta^\alpha(C) = \{\theta : \sum a_l^2 \theta_l^2 \leq C^2\}$ with $a_{2l-1} = a_{2l} = (2l)^\alpha$ be, as usual, an ellipsoid of $\alpha-$mean square differentiable functions. If K is a q-th order kernel in the sense of part (a), show that for each $\alpha > q$,

$$\inf_{h>0} \sup_{\theta \in \Theta^\alpha(C)} r_\epsilon(\hat{\theta}_h, \theta) \geq c(\alpha, q, C)(\epsilon^2)^{2q/(2q+1)}.$$

[Thus, for a second order kernel, the (uniform) rate of convergence is $n^{-4/5}$, even if we consider ellipsoids of functions with 10 or 10^6 derivatives. Since the (dyadic) block James Stein estimate has rate $n^{-2\alpha/(2\alpha+1)}$ over each $\Theta^\alpha(C)$, we might say that it corresponds to an infinite order kernel.]

6.2 *(Oscillation within blocks.)* Let $\Theta(a, C$ be an ellipsoid $\{(\theta_i) : \sum a_i^2\theta_i^2 \le C^2\}$. Assume that $a_i \nearrow \infty$. Let blocks B_j be defined as in (9.16) and the oscillation of a_i within blocks by

$$\mathrm{osc}(B_j) = \max_{l,l' \in B_j} \frac{a_l}{a_{l'}}.$$

Show that if $\mathrm{osc}(B_j) \to 1$ as $j \to \infty$ then

$$R_L(\Theta, \epsilon) \sim R_{BL}(\Theta, \epsilon) \qquad \text{as } \epsilon \to 0.$$

6.3 *(Block linear minimaxity.)* Show that if Θ is solid, orthosymmetric and block-symmetric, then

$$R_L(\Theta, \epsilon) = R_{BL}(\Theta, \epsilon) \qquad \text{for all } \epsilon > 0.$$

6.4 *(White noise in frequency domain).* Consider the discrete Fourier transform (6.21). Suppose in addition that the $Y(l)$ are i.i.d. mean zero, variance σ^2 variables and N is even. Show that

$$\mathrm{Var}(\mathrm{Re}(y(\nu))) = \mathrm{Var}(\mathrm{Im}(y(\nu))) = (N/2)\sigma^2. \tag{6.35}$$

6.5 *(Time domain form of kernel* (6.30)*).* Let $L(t) = \sin t/(\pi t)$. If, as in (6.30), $\kappa(s) = (1-|s|^q)_+$, show that the corresponding time domain kernel

$$K(t) = L(t) - (-i)^q L^{(q)}(t).$$

Make plots of K for $q = 2, 4$ and compare with Figure 3.1. Why is the similarity not surprising?

6.6 *(Exact risk details.)* This exercise records some details leading to Figures 6.3— 6.6.
(i) For vectors $x, X \in \mathbb{C}^N$, the inverse discrete Fourier transform `x = ifft(X)` sets $x(j) = N^{-1}\sum_{k=1}^N X(k)e^{-2\pi i(j-1)(k-1)/N}$, $j = 1, \ldots, N$. Suppose now that

$$X(1) = N\theta_0, \qquad \mathrm{Re}\, X(l+1) = N\theta_l, \quad \mathrm{Im}\, X(l+1) = N\theta_{-l}$$

for $1 \le l < N/2$ and $X(k) = 0$ for $k > N/2$. Also, set $t_j = j/N$, and verify that

$$\mathrm{Re}\, x(j) = f(t_{j-1}) = \theta_0 + \sum_{l=1}^{N/2} \theta_l \cos 2\pi l t_{j-1} + \theta_{-l} \sin 2\pi l t_{j-1}, \qquad j = 1, \ldots, N.$$

(ii) Consider the sequence model in the form $y_l = \theta_l + \epsilon z_l$ for $l \in \mathbb{Z}$. For the coefficients specified by (6.34) and below, show that risk function (6.31)

$$r(\hat\theta_h, \theta) = \epsilon^2 + 2\epsilon^2 \sum_1^{l_h} [1 - (hl)^q]^2 + h^{2q} C_{12}^2 \sum_{l=l_1}^{l_2 \wedge l_h} j^{2q-6} + C_{12}^2 \sum_{l_h+1}^{l_2} j^{-6},$$

where $l_h = [h^{-1}]$ and $C_{12}^2 = C^2/\sum_{l=l_1}^{l_2} j^{-2}$.
(iii) Introduce functions (which also depend on l_1, l_2 and C)

$$V(m, n; h, q) = \sum_{l=m}^{n} [1 - (hl)^q]^2, \qquad B(m, n; p) = C_{12}^2 \sum_{l=m\vee l_1}^{n \wedge l_2} j^{p-6},$$

and confirm that in terms of V and B,

$$b_q(\theta) = C_{12}^2 \sum_{l_1}^{l_2} j^{2q-6} = B(l_1, l_2; 2q)$$

$$r(\hat{\theta}_h, \theta) = \epsilon^2 + 2\epsilon^2 V(1, l_h; h, q) + h^{2q} B(1, l_h; 2q) + B(l_h + 1, l_2; 0).$$

The figures use a vector of values of ϵ^2 and hence of $h = h_\epsilon$ in (6.32) and l_h; these representations facilitate the vectorization of the calculations.

(iv) For the block James-Stein estimator, define blocks $y_b \leftrightarrow (y_l, 2^{b-1} < |l| \le 2^b)$, so that $n_b = 2^b$. Choose $n_\epsilon = \epsilon^{-2}$ so that $J_\epsilon = \log_2 n_\epsilon$ is an integer. Show that (6.14) becomes

$$r_\epsilon(\hat{\theta}^{BJS}, \theta) \le (2J_\epsilon + 1)\epsilon^2 + \sum_{b=2}^{J_\epsilon - 1} \frac{n_b B_b}{n_b + B_b n_\epsilon} + B_\epsilon,$$

where $B_b = B(2^{b-1} + 1, 2^b; 0)$ and $B_\epsilon = B(2^{J_\epsilon - 1} + 1, l_2; 0)$.

7

A Primer on Estimation by Wavelet Shrinkage

> When I began to look at what Meyer had done, I realized it was very close to some ideas in image processing. Suppose you have an image of a house. If you want to recognize simply that it is a house, you do not need most of the details. So people in image processing had the idea of approaching the images at different resolutions. (Stéphane Mallat, quoted in *New York Times*.)

When an image arrives on a computer screen over the internet, the broad outlines arrive first followed by successively finer details that sharpen the picture. This is the wavelet transform in action. In the presence of noisy data, and when combined with thresholding, this *multiresolution* approach provides a powerful tool for estimating the underlying object.

Our goal in this chapter is to give an account of some of the main issues and ideas behind wavelet thresholding as applied to equally spaced signal or regression data observed in noise. The purpose is both to give the flavor of how wavelet shrinkage can be used in practice, as well as provide the setting and motivation for theoretical developments in subsequent chapters. Both this introductory account and the later theory will show how the shortcomings of linear estimators can be overcome by appropriate use of simple non-linear thresholding. We do not attempt to be encyclopedic in coverage of what is now a large area, rather we concentrate on orthogonal wavelet bases and the associated multiresolution analyses for functions of a single variable.

The opening quote hints at the interplay between disciplines that is characteristic of wavelet theory and methods, and so is reflected in the exposition here.

Section 7.1 begins with the formal definition of a multiresolution analysis (MRA) of square integrable functions, and indicates briefly how particular examples are connected with important wavelet families. We consider decompositions of $L_2(\mathbb{R})$ and of $L_2([0,1])$, though the latter will be our main focus for the statistical theory.

This topic in harmonic analysis leads directly into a signal processing algorithm: the "two-scale" relations between neighboring layers of the multiresolution give rise in Section 7.2 to filtering relations which, in the case of wavelets of compact support, lead to the fast $O(n)$ algorithms for computing the direct and inverse wavelet transforms on discrete data.

Section 7.3 explains in more detail how columns of the discrete wavelet transform are related to the continuous wavelet and scaling function of the MRA, while Section 7.4 describes the changes needed to adapt to finite data sequences.

Finally in Section 7.5 we are ready to describe wavelet thresholding for noisy data using the discrete orthogonal wavelet transform of $n = 2^J$ equally spaced observations. The

'hidden sparsity' heuristic is basic: the wavelet transform of typical 'true' signals is largely concentrated in a few co-ordinates while the noise is scattered throughout, so thresholding will retain most signal while suppressing most noise.

How the threshold itself is set is a large question we will discuss at length. Section 7.6 surveys some of the approaches that have been used, and for which theoretical support exists. The discussion in these two sections is informal, with numerical examples. Corresponding theory is developed in later chapters.

7.1 Multiresolution analysis

This is not an *ab initio* exposition of wavelet ideas and theorems: some authoritative books include Meyer (1990), Daubechies (1992), Mallat (1998), and others listed in the chapter notes. Rather we present, without proofs, some definitions, concepts and results relevant to our statistical theory and algorithms. In this way, we also establish the particular notation that we use, since there are significantly different conventions in the literature.

It is a striking fact that the fast algorithms for *discrete* orthogonal wavelet transforms have their origin in change of basis operations on square integrable functions of a *continuous* variable. We therefore begin with the notion of a multiresolution analysis of $L_2(\mathbb{R})$. We concentrate on the univariate case, though the ideas extend to $L_2(\mathbb{R}^d)$. Constructions in the frequency domain play an important role, but these are largely deferred to a sketch in Appendix B.1 and especially the references given there.

Definition. A *multiresolution analysis* (MRA) of $L_2(\mathbb{R})$ is given by a sequence of closed subspaces $\{V_j, j \in \mathbb{Z}\}$ satisfying the following conditions:

(i) $V_j \subset V_{j+1}$,
(ii) $f(x) \in V_j$ if and only if $f(2x) \in V_{j+1}, \forall\, j \in \mathbb{Z}$,
(iii) $\cap_{j\in\mathbb{Z}} V_j = \{0\}, \qquad \overline{\cup_{j\in\mathbb{Z}} V_j} = L_2(\mathbb{R})$.
(iv) $\exists \varphi \in V_0$ such that $\{\varphi(x-k) : k \in \mathbb{Z}\}$ is an orthonormal basis (o.n.b) for V_0.

The function φ in (iv) is called the *scaling function* of the given MRA. Set $\varphi_{jk}(x) = 2^{j/2}\varphi(2^j x - k)$. One says that ϕ_{jk} has scale 2^{-j} and location $k2^{-j}$. Properties (ii) and (iv) imply that $\{\varphi_{jk}, k \in \mathbb{Z}\}$ is an orthonormal basis for V_j. The orthogonal projection from $L_2(\mathbb{R}) \to V_j$ is then

$$P_j f = \sum_k \langle f, \varphi_{jk}\rangle \varphi_{jk}.$$

The spaces V_j form an increasing sequence of approximations to $L_2(\mathbb{R})$: indeed property (iii) implies that $P_j f \to f$ in $L_2(\mathbb{R})$ as $j \to \infty$.

Example. Haar MRA. Set $I_{jk} = [2^{-j}k, 2^{-j}(k+1)]$. The "Haar multiresolution analysis" is defined by

$$V_j = \{f \in L_2(\mathbb{R}) : f|_{I_{jk}} = c_{jk}\}, \qquad \varphi = I_{[0,1]}.$$

Thus V_j consists of piecewise constant functions on intervals of length 2^{-j}, and $P_j f(x)$ is the average of f over the interval I_{jk} that contains x.

Example. Box spline MRA. Given $r \in \mathbb{N}$, set

$$V_j = \{f \in L_2 \cap C^{r-1} \text{and } f|_{I_{jk}} \text{ is a polynomial of degree } r\}.$$

If $r = 0$, this reduces to the Haar MRA. If $r = 1$, we get continuous, piecewise linear functions and if $r = 3$, cubic splines. For more on the construction of the scaling function φ, see Appendix B.1.

A key role in wavelet analysis is played by a pair of *two scale equations* and their associated discrete filter sequences. Given an MRA with scaling function φ, since $V_{-1} \subset V_0$, one may express $\varphi_{-1,0}$ in terms of $\varphi_{0,k}$ using the *two scale equation*

$$\frac{1}{\sqrt{2}}\varphi\big(\frac{x}{2}\big) = \sum_k h[k]\varphi(x-k). \tag{7.1}$$

The sequence $\{h[k]\}$ is called the *discrete filter* associated with φ, For the Haar MRA example, $h[0] = h[1] = 1/\sqrt{2}$.

Now take Fourier transforms, (C.7), of both sides: since $\widehat{\varphi_{0k}}(\xi) = e^{-ik\xi}\hat{\varphi}(\xi)$, the two scale equation has the reexpression

$$\hat{\varphi}(2\xi) = 2^{-1/2}\hat{h}(\xi)\hat{\varphi}(\xi), \tag{7.2}$$

where the *transfer function*

$$\hat{h}(\xi) = \sum h[k]e^{-ik\xi}.$$

The MRA conditions imply important structural constraints on $\hat{h}(\xi)$. These in turn lead to theorems describing how to construct scaling functions φ – some of these are reviewed, with references, in Appendix B.1.

Now we turn to the wavelets. Define the *detail subspace* $W_j \subset L_2$ as the orthogonal complement of V_j in V_{j+1}: $V_{j+1} = V_j \oplus W_j$. A candidate for a wavelet $\psi \in W_{-1} \subset V_0$ must satisfy its own two scale equation

$$\frac{1}{\sqrt{2}}\psi\big(\frac{x}{2}\big) = \sum_k g[k]\varphi(x-k). \tag{7.3}$$

Again, taking the Fourier transform of both sides and defining $\hat{g}(\xi) = \sum g_k e^{-ik\xi}$,

$$\hat{\psi}(2\xi) = 2^{-1/2}\hat{g}(\xi)\hat{\varphi}(\xi). \tag{7.4}$$

Define $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$. Suppose that it is possible to define ψ using (7.4) so that $\{\psi_{jk}, k \in \mathbb{Z}\}$ form an orthonormal basis for W_j. Then it follows `[proof ref needed?]` property (iii) of the MRA that the full collection $\{\psi_{jk}, (j,k) \in \mathbb{Z}^2\}$ forms an orthonormal basis for $L_2(\mathbb{R})$.

Thus we have decompositions

$$L_2(\mathbb{R}) = \oplus_{j\in\mathbb{Z}} W_j = V_J \oplus \bigoplus_{j\geq J} W_j,$$

for each J, with corresponding expansions

$$f = \sum_{j,k}\langle f, \psi_{jk}\rangle\psi_{jk} = \sum_k \langle f, \varphi_{Jk}\rangle\varphi_{Jk} + \sum_{j\geq J}\sum_k \langle f, \psi_{jk}\rangle\psi_{jk}.$$

The first is called a *homogeneous* expansion, while the second is said to be inhomogeneous since it combines only the detail spaces at scales finer than J.

A key heuristic idea is that for typical functions f, the wavelet coefficients $\langle f, \psi_{jk} \rangle$ are large only at low frequencies or wavelets located close to singularities of f. This heuristic notion is quantified in some detail in Section 9.6 and Appendix B.

Here is a simple result describing the wavelet coefficients of piecewise constant functions.

Lemma 7.1 *Suppose ψ has compact support $[-S, S]$ and $\int \psi = 0$. Suppose f is piecewise constant with d discontinuities. Then at level j at most $(2S-1)d$ of the wavelet coefficients $\theta_{jk} = \int f \psi_{jk}$ are non-zero, and those are bounded by $c2^{-j/2}$.*

Proof Let the discontinuities of f occur at $x_1, \ldots, x_d$. Since $\int \psi = 0$,

$$\theta_{jk} = \int f \psi_{jk} = 2^{-j/2} \int f(2^{-j}(t+k))\psi(t)dt$$

vanishes unless some x_i lies in the interior of $\mathrm{supp}(\psi_{jk})$. In this latter case, we can use the right hand side integral to bound $|\theta_{jk}| \le \|f\|_\infty \|\psi\|_1 2^{-j/2}$. The support of ψ_{jk} is $k2^{-j} + 2^{-j}[-S, S]$, and the number of k for which $x_i \in \mathrm{int}(\mathrm{supp}(\psi_{jk}))$ is at most $2S-1$. So the total number of non-zero θ_{jk} at level j is at most $(2S-1)d$. □

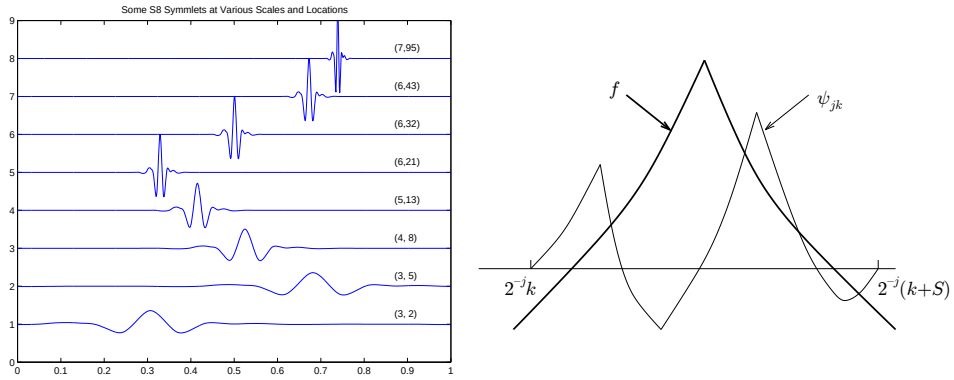

Figure 7.1 Left panel: Wavelets (from the Symmlet-8 family), the pair (j,k) indicates wavelet ψ_{jk}, at resolution level j and approximate location $k2^{-j}$. Right panel: Schematic of a wavelet ψ_{jk} of compact support "hitting" a singularity of function f.

The construction of some celebrated pairs (φ, ψ) of scaling function and wavelet is sketched, with literature references, in Appendix B.1. Before briefly listing some of the well known families, we discuss several properties that the pair (φ, ψ) might possess.

Support size. Suppose that the support of ψ is an interval of length S, say $[0, S]$. Then ψ_{jk} is supported on $k2^{-j} + 2^{-j}[0, S]$. Now suppose also that f has a singularity at x_0. The size of S determines the range of influence of the singularity on the wavelet coefficients $\theta_{jk}(f) = \int f \psi_{jk}$. Indeed, at level j, the number of coefficients that 'feel' the singularity at

x_0 is just the number of wavelet indices k for which supp ψ_{jk} covers x_0, which by rescaling is equal to S (or $S-1$ if x_0 lies on the boundary of suppψ_{jk}).

It is therefore in principle desirable to have small support for ψ and φ. These are in turn determined by the support of the filter h, by means of the two scale relations (7.1) and (7.3). For a filter $h = (h_k, k \in \mathbb{Z})$, its support is the smallest closest interval containing the non-zero values of h_k. For example, Mallat (1999, Chapter 7) shows that

(i) supp φ = supp h if one of the two is compact, and

(ii) if supp $\varphi = [N_1, N_2]$, then supp $\psi = [\frac{N_1-N_2+1}{2}, \frac{N_2-N_1+1}{2}]$.

Vanishing moments. The wavelet ψ is said to have r *vanishing moments* if

$$\int x^k \psi(x)dx = 0 \qquad k = 0, 1, \ldots r-1. \tag{7.5}$$

Thus ψ is orthogonal to all polynomials of degree $r-1$. As a result, the rate of decay of wavelet coefficients of a smooth function is governed by the number of vanishing moments of the wavelet ψ. For example, in Appendix B.1 we prove:

Lemma 7.2 *If f is C^α on $\mathbb{R}$ and ψ has $r \geq \lceil \alpha \rceil$ vanishing moments, then*

$$|\langle f, \psi_{jk} \rangle| \leq c_\psi C 2^{-j(\alpha+1/2)}.$$

If α is a positive integer, then the C^α assumption is just the usual notion that f has α continuous derivatives, and the constant $C = \|D^\alpha f\|_\infty / \alpha!$. For $\alpha > 0$ non-integer, we use the definition of Hölder smoothness, given in Appendix **??**. Note the parallel with the definition (3.21) of vanishing moments for an averaging kernel K, and with expression (3.22) for the approximation error of a qth order kernel.

Daubechies (1988) showed that existence of p vanishing moments for an orthogonal wavelet implied a support length for h, and hence for φ, ψ, of at least $2p-1$. Thus, for such wavelets, there is a tradeoff between short support and large numbers of vanishing moments. The resolution of this tradeoff is perhaps best made according to the context of a given application.

Regularity. Since $\hat{f}(x) = \sum \hat{\theta}_{jk} \psi_{jk}(x)$, the smoothness of $x \to \psi_{jk}(x)$ can impact the visual appearance of a reconstruction. However it is the number of vanishing moments that affects the size of wavelet coefficients at fine scales, at least in regions where f is smooth. So both properties are in general relevant. For the common wavelet families [to be reviewed below], it happens that regularity increases with the number of vanishing moments.

For wavelet bases, regularity of ψ implies that a corresponding number of moments vanish. We refer to Daubechies (1992, §5.5) for the proof of

Proposition 7.3 *If $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$ is an orthonormal basis for $L_2(\mathbb{R})$, and if ψ is C^r, with $\psi^{(k)}$ bounded for $k \leq r$ and $|\psi(x)| \leq C(1+|x|)^{-r-1-\epsilon}$, then $\int x^k \psi(x)dx = 0$ for $k = 0, \ldots, r$.*

Some wavelet families. The common constructions of instances of (φ, ψ) use Fourier techniques deriving from the two scale equations (7.2) and (7.4) and the filter transfer function $\hat{h}(\xi)$. Many constructions generate a family indexed by the number of vanishing moments p. For some further details see Appendix B.1, and wavelet texts, such as Mallat Ch.7.

Haar. The simplest and only rarely best: $\varphi = I_{[0,1]}$ and $\psi = I_{[0,1/2]} - I_{[1/2,1]}$. It has a single vanishing moment, and of course no smoothness.

Meyer. $\hat{\varphi}(\xi), \hat{\psi}(\xi)$ have compact support in frequency ξ, and so $\varphi(x)$ and $\psi(x)$ are C^∞, but do not have compact support in x – in fact they have only polynomial decay for large x. The wavelet has infinitely many vanishing moments.

Battle-Lemarié spline. These are wavelets derived from the spline MRA. The pair $\varphi(x), \psi(x)$ are polynomial splines of degree m and hence are C^{m-1} in x. They have exponential decay in x, and are symmetric (resp. anti-symmetric) about $x = 1/2$ for m odd (resp. even). The wavelet has $m + 1$ vanishing moments.

Compact support wavelets. Daubechies constructed several sets of compactly supported wavelets and scaling functions, indexed by the number of vanishing moments p for ψ.

(a) "Daubechies" family – the original family of wavelets D_{2p} in which ψ has minimum support length $2p - 1$, on the interval $[-p + 1, p]$. The wavelets are quite asymmetric, and have regularity that grows roughly at rate $0.2p$, though better regularity is known for small p – e.g. just over C^1 for $p = 3$.

(b) "Symmlet" family – another family with minimum support $[-p+1, p]$, but with filter h chosen so as to make ψ as close to symmetric (about $\frac{1}{2}$) as possible.

(c) "Coiflet" family – a family with p vanishing moments for ψ and also for φ:

$$\int \varphi = 1, \qquad \int t^k \varphi = 0, \quad 1 \le k < p.$$

This constraint forces a larger support length, namely $3p - 1$.

Wavelets on the interval $[0, 1]$.

In statistical applications, one is often interested in an unknown function f defined on an interval, say $I = [0, 1]$ after rescaling. Brutal extension of f to $\mathbb{R}$ by setting it to 0 outside I, or even more sophisticated extensions by reflection or folding, introduce a discontinuity in f or its derivatives at the edges of I.

If one works with wavelets of compact support (of length S, say), these discontinuities affect only a fixed number $2S$ of coefficients at each level j and so will often not affect the asymptotic behavior of global measures of estimation error on I. Nevertheless, both in theory and in practice, it is desirable to avoid such artificially created discontinuities. We refer here to two approaches that have been taken in the literature. [The approach of "folding" across boundaries, is dicussed in Mallat (1999, Sec. 7.5.2.).]

(i) Periodization. One restricts attention to *periodic* functions on I. Meyer (1990, Vol 1, Chapter III.11) shows that one can build an orthonormal basis for $L_{2,\mathrm{per}}(I)$ by periodization. Suppose that φ and ψ are nice orthonormal scaling and wavelet functions for $L_2(\mathbb{R})$ and define

$$\varphi^{\mathrm{per}}_{j,k}(x) = \sum_{\ell \in \mathbb{Z}} \varphi_{j,k}(x + \ell), \qquad \psi^{\mathrm{per}}_{j,k}(x) = \sum_{\ell \in \mathbb{Z}} \psi_{j,k}(x + \ell).$$

If φ, ψ have compact support, then for j larger than some j_1, these sums reduce to a single term for each $x \in I$. [Again, this is analogous to the discussion of periodization of kernels at (3.16) and (3.18)– (3.19).]

Define $V_j^{\text{per}} = \text{span}\,\{\varphi_{jk}^{\text{per}}, k \in \mathbb{Z}\}$, and $W_j^{\text{per}} = \text{span}\,\{\psi_{jk}^{\text{per}}, k \in \mathbb{Z}\}$: this yields an orthogonal decomposition

$$L_{2,per}(I) = V_L^{\text{per}} \oplus \bigoplus_{j \geq L} W_j^{\text{per}},$$

with $\dim V_j^{\text{per}} = \dim W_j^{\text{per}} = 2^j$ for $j \geq 0$. Meyer makes a detailed comparison of Fourier series and wavelets on $[0, 1]$, including remarkable properties such as uniform convergence of the wavelet approximations of any continuous function on $[0, 1]$.

(ii) Orthonormalization on $[0, 1]$ For non-periodic functions on $[0, 1]$, one must take a different approach. We summarize results of a construction described in detail in Cohen et al. (1993b), which builds on Meyer (1991) and Cohen et al. (1993a). The construction begins with a Daubechies pair (φ, ψ) having p vanishing moments and minimal support $[-p + 1, p]$. For j such that $2^j \geq 2p$ and for $k = p, \ldots, 2^j - p - 1$, the scaling functions $\varphi_{jk}^{\text{int}} = \varphi_{jk}$ have support contained wholly in $[0, 1]$ and so are left unchanged. At the boundaries, for $k = 0, \ldots, p - 1$, construct functions φ_k^L with support $[0, p + k]$ and φ_k^R with support $[-p - k, 0]$, and set

$$\varphi_{jk}^{\text{int}} = 2^{j/2}\varphi_k^L(2^j x), \qquad \varphi_{j,2^j-k-1}^{\text{int}} = 2^{j/2}\varphi_k^R(2^j(x-1)).$$

The $2p$ functions φ_k^L, φ_k^R are finite linear combinations of scaled and translated versions of the original φ and so have the same smoothness as φ. We can now define the multiresolution spaces $V_j^{\text{int}} = \text{span}\{\varphi_{jk}^{\text{int}}, k = 0, \ldots, 2^j - 1\}$. It is shown that $\dim V_j^{\text{int}} = 2^j$, and that they have two key properties:

(i) in order that $V_j^{\text{int}} \subset V_{j+1}^{\text{int}}$, it is required that the boundary scaling functions satisfy two scale equations. For example, on the left side

$$\frac{1}{\sqrt{2}}\varphi_k^L\left(\frac{x}{2}\right) = \sum_{l=0}^{p-1} H_{kl}^L \varphi_l^L(x) + \sum_{m=p}^{p+2k} h_{km}^L \varphi(x-m).$$

(ii) each V_j^{int} contains, on $[0, 1]$, all polynomials of degree at most $p - 1$.

Turning now to the wavelet spaces, W_j^{int} is defined as the orthogonal complement of V_j^{int} in V_{j+1}^{int}. Starting from a Daubechies wavelet ψ with support in $[-p + 1, p]$ and with p vanishing moments, construct ψ_k^L with support in $[0, p + k]$ and ψ_k^R with support in $[-p - k, 0]$ and define ψ_{jk}^{int} as for $\varphi_{jk}^{\text{int}}$ replacing $\varphi, \varphi_k^L, \varphi_k^R$ by ψ, ψ_k^L and ψ_k^R. It can be verified that $W_k^{\text{int}} = \text{span}\{\psi_{jk}^{\text{int}}, k = 0, \ldots, 2^{j-1}\}$ and that for each L with $2^L \geq 2p$,

$$L_2([0, 1]) = V_L^{\text{int}} \oplus \bigoplus_{j \geq L} W_j^{\text{int}}, \tag{7.6}$$

and hence $f \in L_2[0, 1]$ has an expansion

$$f(x) = \sum_{k=0}^{2^L-1} \beta_k \varphi_{Lk}^{\text{int}}(x) + \sum_{j \geq L} \sum_{k=0}^{2^j-1} \theta_{jk} \psi_{jk}^{\text{int}}(x)$$

with $\beta_k = \langle f, \varphi_{Lk}^{\text{int}} \rangle$ and $\theta_{jk} = \langle f, \psi_{jk}^{\text{int}} \rangle$. Note especially from property (ii) that since V_L^{int} contains polynomials of degree $\leq p - 1$, it follows that all ψ_{jk}^{int} have vanishing moments of order p.

7.2 The Cascade algorithm for the Discrete Wavelet Transform

A further key feature of wavelet bases is the availability of fast $O(N)$ algorithms for computing both the wavelet transform of discrete data and its inverse. This "cascade" algorithm is often derived, as we do below, by studying the structure of a multiresolution analysis of functions of a continuous real variable. In practice, it is used on finite data sequences, and the scaling function φ and wavelet ψ of the MRA are not used at all. This is fortunate, because the latter are typically only defined by limiting processes and so are hard to compute, compare (B.6) and (B.11). Thus there is a most helpful gap between the motivating mathematics and the actual data manipulations. Since our goal later is to give a theoretical account of the statistical properties of these data manipulations, our presentation here will try to be explicit about the manner in which discrete orthogonal wavelet coefficients in fact approximate their multiresolution relatives.

Suppose, then, that we have a multiresolution analysis $\{V_j\}$ generated by an orthonormal scaling function φ, and with detail spaces W_j generated by an orthonormal wavelet ψ so that the collection $\{\psi_{jk}, j, k \in \mathbb{Z}\}$ forms an orthonormal basis for $L_2(\mathbb{R})$.

Analysis and Synthesis operators. Consider a function $f \in V_j$. Let $a_j = \{a_j[k]\}$ denote the coefficients of f in the orthobasis $\mathcal{B}_j = \{\varphi_{jk}, k \in \mathbb{Z}\}$, so that

$$a_j[k] = \langle f, \phi_{jk} \rangle.$$

Since $V_j = V_{j-1} \oplus W_{j-1}$, we can also express f in terms of the basis

$$\mathcal{B}'_j = \{\varphi_{j-1,k}, k \in \mathbb{Z}\} \cup \{\psi_{j-1,k}, k \in \mathbb{Z}\}$$

with coefficients

$$a_{j-1}[k] = \langle f, \phi_{j-1,k} \rangle, \qquad d_{j-1}[k] = \langle f, \psi_{j-1,k} \rangle, \tag{7.7}$$

and mnemonics "a" for approximation and "d" for detail.

Since $\mathcal{B}$ and $\mathcal{B}'$ are orthonormal bases for the same space, the change of basis maps

$$\begin{aligned} A_j : &\quad a_j \to \{a_{j-1}, d_{j-1}\} && \text{("analysis")} \\ S_j : &\quad \{a_{j-1}, d_{j-1}\} \to a_j && \text{("synthesis")} \end{aligned}$$

must be orthogonal, and transposes of one another:

$$A_j A_j^T = A_j^T A_j = I, \qquad S_j = A_j^{-1} = A_j^T.$$

To derive explicit expressions for A_j and S_j, rewrite the two-scale equations (7.1) and (7.3) in terms of level j, in order to express $\varphi_{j-1,k}$ and $\psi_{j-1,k}$ in terms of φ_{jk}, using the fact that V_{j-1} and W_{j-1} are contained in V_j. Rescale by replacing x by $2^j x - 2k$ and multiply both equations by $2^{j/2}$. Recalling the notation $\varphi_{jk}(x) = 2^{j/2}\varphi(2^j x - k)$, we have

$$\varphi_{j-1,k}(x) = \sum_l h[l]\varphi_{j,2k+l}(x) = \sum_l h[l-2k]\varphi_{jl}(x). \tag{7.8}$$

The corresponding relation for the coarse scale wavelet reads

$$\psi_{j-1,k}(x) = \sum_l g[l]\varphi_{j,2k+l}(x) = \sum_l g[l-2k]\varphi_{jl}(x). \tag{7.9}$$

Taking inner products with f as in (7.7) yields the representation of A_j:

$$\begin{aligned} a_{j-1}[k] &= \sum_l h[l-2k]a_j[l] = Rh \star a_j[2k] \\ d_{j-1}[k] &= \sum_l g[l-2k]a_j[l] = Rg \star a_j[2k], \end{aligned} \tag{7.10}$$

where R denotes the *reversal* operator $Ra[k] = a[-k]$, and $\star$ denotes discrete convolution $a \star b[k] = \sum a[k-l]b[l]$. Introducing also the *downsampling* operator $Da[k] = a[2k]$, we could write, for example, $a_{j-1} = D(Rh \star a_j)$. Thus the analysis, or "fine-to-coarse" step $A_j : a_j \to (a_{j-1}, d_{j-1})$ can be described as "filter with Rh and Rg and then downsample".

Synthesis step S_j. Since $\varphi_{j-1,k} \in V_{j-1} \subset V_j$, we can expand $\varphi_{j-1,k}$ as $\sum_l \langle \varphi_{j-1,k}, \varphi_{jl} \rangle \varphi_{jl}$, along with an analogous expansion for $\psi_{j-1,k} \in W_{j-1} \subset V_j$. Comparing the coefficients (7.8) and (7.9) yields the identifications

$$\langle \varphi_{j-1,k}, \varphi_{jl} \rangle = h[l-2k], \qquad \langle \psi_{j-1,k}, \varphi_{jl} \rangle = g[l-2k].$$

Since $\varphi_{jl} \in V_j = V_{j-1} \oplus W_{j-1}$, we may use the previous display to write

$$\varphi_{jl} = \sum_k h[l-2k]\varphi_{j-1,k} + g[l-2k]\psi_{j-1,k}. \tag{7.11}$$

[Note that this time the sums are over k (the level $j-1$ index), not over l as in the analysis step!]. Taking inner products with f in the previous display leads to the synthesis rule

$$a_j[l] = \sum_k h[l-2k]a_{j-1}[k] + g[l-2k]d_{j-1}[k]. \tag{7.12}$$

To write this in simpler form, introduce the *zero-padding* operator $Za[2k] = a[k]$ and $Za[2k+1] = 0$, so that

$$a_j[l] = h \star Za_{j-1}[l] + g \star Zd_{j-1}[l].$$

So the sythesis or *coarse-to-fine* step $S_j : (a_{j-1}, d_{j-1}) \to a_j$ can be described as "zero-pad, then filter with h (and g), and then add".

Computation. If the filters h and g have length L, the analysis steps (7.10) each require L multiplys and adds to compute each coefficient. The synthesis step (7.12) similarly needs L multiplys and adds per coefficient.

The Cascade algorithm. We may represent the successive application of analysis steps beginning at level J and continuing down to a coarser level L by means of a cascade diagram

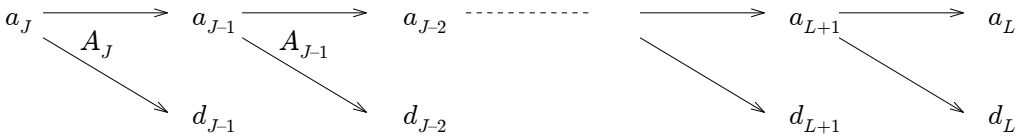

Figure 7.2 The cascade algorithm

Composition of each of these orthogonal transformations produces an orthogonal transformation $W = A_{L+1} \cdots A_{J-1} A_J$:

$$a_J \longleftrightarrow \{d_{J-1}, d_{J-2}, \ldots, d_L, a_L\}. \tag{7.13}$$

The forward direction is the analysis operator, given by the orthogonal discrete wavelet transform W. The reverse direction is the synthesis operator, given by its inverse, $W^T = S_J S_{J-1} \cdots S_{L+1}$.

W as a 'matrix'. W represents a change of basis from $V_J = \text{span}\{\varphi_{Jk}, k \in \mathbb{Z}\}$ to

$$V_L \oplus W_L \oplus \cdots \oplus W_{J-1} = \text{span}\{\{\varphi_{Lk}\} \cup \{\psi_{jk}\}, L \le j \le J-1, k \in \mathbb{Z}\}.$$

Define index sets $\mathcal{D} = \{I = (j,k) : L \le j \le J-1;\ k \in \mathbb{Z}\}$ and $\mathcal{A} = \{I = (L,k) : k \in \mathbb{Z}\}$. If we write $W = (W_{Ik})$ for $I \in \mathcal{D} \cup \mathcal{A}$ and $k \in \mathbb{Z}$, then we have

$$W_{Ik} = \begin{cases} \langle \psi_I, \varphi_{Jk} \rangle & I \in \mathcal{D} \\ \langle \varphi_{Lk'}, \varphi_{Jk} \rangle & I = (L,k') \in \mathcal{A}. \end{cases}$$

7.3 Discrete and Continuous Wavelets

Our goal now is to describe more explicitly how the vectors ψ_I are related to the $L_2(\mathbb{R})$ wavelets $\psi_{jk}(x) = 2^{j/2}\psi(2^j - k)$. For simplicity, we ignore boundary effects and remain in the setting of $\ell_2(\mathbb{Z})$.

The discrete filtering operations of the cascade algorithm make no explicit use of the wavelet ψ and scaling function φ. Yet they are derived from the multiresolution analysis generated by (φ, ψ), and it is our goal in this subsection to show more explicitly how the orthonormal columns of the discrete wavelet transform are approximations to the orthobasis functions φ_{jk} and ψ_{jk}.

Approximating φ and ψ from the filter cascade. So far, the cascade algorithm has been described implicitly, by iteration. We now seek a more explicit representation. Let $h^{(r)} = h \star Zh \star \cdots \star Z^{r-1}h$ and $g^{(r)} = h^{(r-1)} \star Z^{r-1}g$.

Lemma 7.4

$$a_{j-r}[k] = \sum_n h^{(r)}[n - 2^r k]\, a_j[n] = Rh^{(r)} \star a_j[2^r k].$$
$$d_{j-r}[k] = \sum_n g^{(r)}[n - 2^r k]\, a_j[n] = Rg^{(r)} \star a_j[2^r k].$$

This formula says that the 2^r-fold downsampling can be done at the end of the calculation if appropriate infilling of zeros is done at each stage. While not necessarily sensible in computation, this is helpful in deriving a formula.

To describe the approximation of φ and ψ it is helpful to consider the sequence of nested lattices $2^{-r}\mathbb{Z}$ for $r = 0, 1, \ldots$. Define functions $\varphi^{(r)}, \psi^{(r)}$ on $2^{-r}\mathbb{Z}$ using the r-fold iterated filters:

$$\varphi^{(r)}(2^{-r}n) = 2^{r/2}h^{(r)}[n], \qquad \psi^{(r)}(2^{-r}n) = 2^{r/2}g^{(r)}[n]. \tag{7.14}$$

Clearly $\varphi^{(0)}$ and $\psi^{(0)}$ are the original filters h and g, and we will show that $\phi^{(r)} \to \phi$, $\psi^{(r)} \to$

ψ in an appropriate sense. Indeed, interpret the function $\varphi^{(r)}$ on $2^{-r}\mathbb{Z}$ as a (signed) measure $\mu_r = \mu[\varphi^{(r)}]$ that places mass $2^{-r}\varphi^{(r)}(2^{-r}n)$ at $2^{-r}n$. Also interpret the function φ on $\mathbb{R}$ as the density with respect to Lebesgue measure of a signed measure $\mu = \mu[\varphi]$. Then weak convergence means that $\int f d\mu_r \to \int f d\mu$ for all bounded continuous functions f.

Proposition 7.5 *$\mu[\varphi^{(r)}]$ converges weakly to $\mu[\varphi]$ as $r \to \infty$.*

The left panel of Figure 7.3 illustrates the convergence for the Daubechies D4 filter. The proof of this and all results in this section is deferred to the end of the chapter.

We now describe the columns of the discrete wavelet transform in terms of these approximate scaling and wavelet functions. To do so, recall the indexing conventions $\mathcal{D}$ and $\mathcal{A}$ used in describing (W_{Ii}). In addition, for $x \in 2^{-(j+r)}\mathbb{Z}$, define

$$\varphi_{jk}^{(r)}(x) = 2^{j/2}\varphi^{(r)}(2^j x - k), \qquad \psi_{jk}^{(r)}(x) = 2^{j/2}\psi^{(r)}(2^j x - k). \tag{7.15}$$

Proposition 7.6 *Suppose that $N = 2^J$. The discrete wavelet transform matrix (W_{Ii}) with $I = (j,k)$ and $i \in \mathbb{Z}$ is given by*

$$W_{Ii} = \begin{cases} \langle \psi_I, \varphi_{Ji} \rangle = N^{-1/2}\psi_{jk}^{(J-j)}(i/N) & I = (jk) \in \mathcal{D}, \\ \langle \varphi_{Lk}, \varphi_{Ji} \rangle = N^{-1/2}\varphi_{Lk}^{(J-L)}(i/N) & I \in \mathcal{A}. \end{cases}$$

Thus, the Ith row of the wavelet transform matrix looks like $\psi_I^{(J-j)}$ (where $I = (j,k)$), and the greater the separation between the detail level j and the original sampling level J, the closer the corresponding function $\psi_{jk}^{(J-j)}$ is to the scaled wavelet $\psi_{jk}(x)$.

Cascade algorithm on sampled data. We have developed the cascade algorithm assuming that the input sequence $a_J[k] = \langle f, \varphi_{Jk} \rangle$. What happens if instead we feed in as inputs $a_J[k]$ a sequence of sampled values $\{f(k/N)\}$?

Suppose that f is a square integrable function on $2^{-J}\mathbb{Z} = N^{-1}\mathbb{Z}$. The columns of the discrete wavelet transform will be orthogonal with respect to the inner product

$$\langle f, g \rangle_N = N^{-1} \sum_{n \in \mathbb{Z}} f(N^{-1}n) g(N^{-1}n). \tag{7.16}$$

Proposition 7.7 *If $a_J[n] = N^{-1/2} f(N^{-1}n)$, and $N = 2^J$, then for $j \le J$,*

$$a_j[k] = \langle \varphi_{jk}^{(J-j)}, f \rangle_N, \qquad d_j[k] = \langle \psi_{jk}^{(J-j)}, f \rangle_N, \qquad k \in \mathbb{Z}. \tag{7.17}$$

Formulas (7.17) are an explicit representation of our earlier description that the sequences $\{a_j[k], k \in \mathbb{Z}\}$ and $\{d_j[k], k \in \mathbb{Z}\}$ are found from $\{a_J[k], k \in \mathbb{Z}\}$ by repeated filtering and downsampling. Formulas (7.17) suggest, without complete proof, that the iteration of this process is stable, in the sense that as $J - j$ increases (the number of levels of cascade between the data level J and the coefficient level j), the coefficients look progressively more like the continuous-time coefficients $\langle \varphi_{jk}, f \rangle$.

Table 7.1 highlights a curious parallel between the "continuous" and "discrete" worlds: the discrete filtering operations represented by the cascade algorithm, through the DWT matrix W, are the same in both cases!

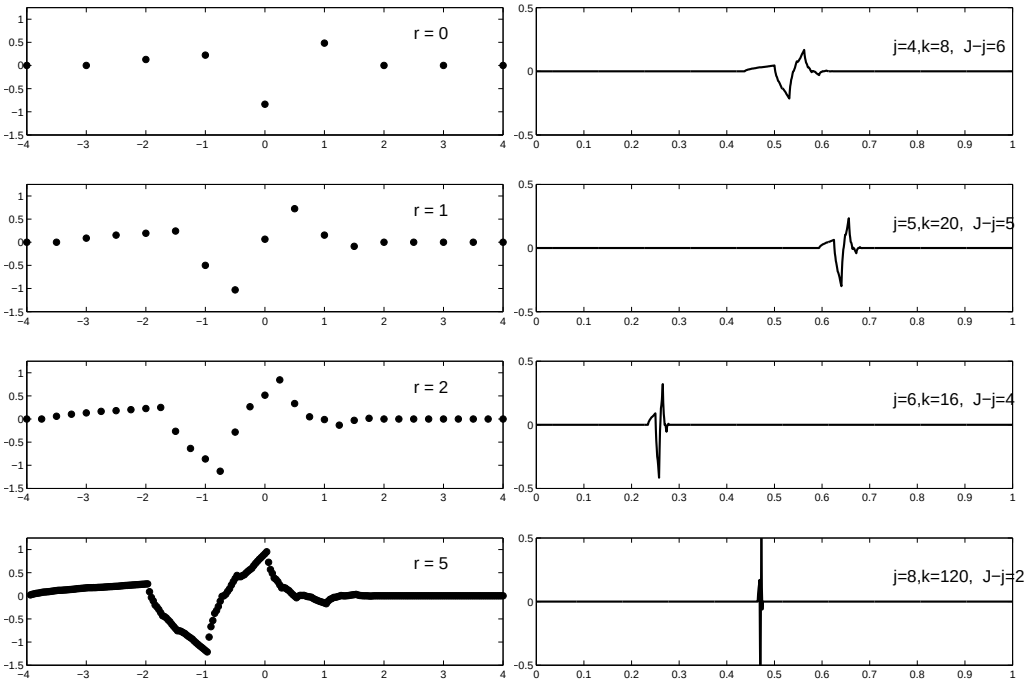

Figure 7.3 Left: The function $\varphi^{(r)}$ on $2^{-r}\mathbb{Z}$ for the Daubechies $D4$ filter for various values of r. Right: rows of the wavelet transform matrix, $N = 1024$, for the Daubechies D4 filter, showing scale j, location k and iteration number $J = j$.

Continuous world	Discrete World
$a_J[k] = \langle \varphi_{Jk}, f \rangle$	$a_J[k] = N^{-1/2} f(nN^{-1})$
$\downarrow$	$\downarrow$
$a_j[k] = \langle \varphi_{jk}, f \rangle$	$a_j[k] = \langle \varphi_{jk}^{(J-j)}, f \rangle_N$
$d_j[k] = \langle \psi_{jk}, f \rangle$	$d_j[k] = \langle \psi_{jk}^{(J-j)}, f \rangle_N$

Table 7.1 *Schematic comparing the orthogonal wavelet transform of functions $f \in L_2(\mathbb{R})$ with the discrete orthogonal wavelet transform of square summable sequences formed by sampling such functions on a lattice with spacing N^{-1}. The vertical arrows represent the outcome of $r = J - j$ iterations of the cascade algorithm in each case.*

7.4 Finite data sequences.

So far we have worked with infinite sequences a_j and $d_j \in \ell_2(\mathbb{Z})$. We turn to the action of the transform and its inverse on a *finite* data sequence a_J of length $N = 2^J$. It is now necessary to say how the boundaries of the data are treated. The transform W remains orthogonal so long as h is a filter generating an orthonormal wavelet basis, and either

(i) boundaries are treated periodically, or

(ii) we use boundary filters (e.g. Cohen et al. (1993b)) that preserve orthogonality.

In either case, the detail vectors d_j in (7.13) are of length 2^j, and the final approximation vector a_L is of length 2^L. The orthogonal transform is then "non-redundant", as it takes $N = 2^J$ coefficients a_J into $2^{J-1} + 2^{J-2} + \ldots + 2^L + 2^L = N$ coefficients in the transform domain. If h has B non-zero coefficients, then the computational complexity of both W and W^T is of order $2B(2^{J-1} + 2^{J-2} + \ldots + 2^L) \leq 2BN = O(N)$.

W maps a vector of data $y = (y_l, l = 1, \ldots, N)$ of length $N = 2^J$ into N wavelet coefficients $w = Wy$. Identifying y with a_J, we may identify w with $\{d_{J-1}, d_{J-2}, \ldots, d_L, a_L\}$. Compare again Figure 7.2. More specifically, we index $w = (w_I)$ with $I = (j, k)$ and

$$\begin{aligned} w_{jk} &= d_{jk} \qquad j = L, \ldots, J-1 \text{ and } k = 1, \ldots, 2^j \\ w_{L-1,k} &= a_{Lk} \qquad k = 1, \ldots, 2^L. \end{aligned}$$

With this notation, we may write $y = W^T w$ in the form

$$y = \sum w_I \boldsymbol{\psi}_I \tag{7.18}$$

with $\boldsymbol{\psi}_I$ denoting the columns of the inverse discrete wavelet transform matrix W^T. [The bolding is used to distinguish the *vector* $\boldsymbol{\psi}_I$ arising in the finite transform from the *function* $\psi_I \in L_2(\mathbb{R})$.] If we set $t_l = l/N$ and adopt the suggestive notation

$$\boldsymbol{\psi}_I(t_l) := \boldsymbol{\psi}_{I,l},$$

then we may write the forward transform $w = Wy$ in the form

$$w_I = \sum_l \boldsymbol{\psi}_I(t_l) y_l. \tag{7.19}$$

7.5 Wavelet shrinkage estimation

Basic model. Observations are taken at equally spaced points $t_l = l/n$, $l = 1, \ldots, n = 2^J$, and are assumed to satisfy

$$Y_l = f(t_l) + \sigma z_l, \qquad z_l \overset{i.i.d}{\sim} N(0, 1). \tag{7.20}$$

It is assumed, for now, that σ is known. The goal is to estimate f, at least at the observation points t_l. The assumption that the observation points are equally spaced is quite important whereas the specific form of the error model and knowledge of σ are less crucial.

Basic strategy. The outline is simply described. First, the *tranform* step, which uses a finite orthogonal wavelet transform W as described in the previous section. Second, a *processing* step in the wavelet domain, and finally an inverse transform, which is accomplished by W^T, since W is orthogonal.

$$\begin{array}{ccc} (y_l) & \xrightarrow{W} & (w_I) \\ & & \downarrow \eta \\ \big(\hat{f}(t_l)\big) & \xleftarrow{W^T} & (\hat{w}_I) \end{array} \tag{7.21}$$

Transform step. Being an orthogonal transform, W is non-redundant, and given $n = 2^J$ data values (y_l) in the "time" domain, produces n transform coefficients in the wavelet domain, by use of the cascade algorithm derived from a filter h, as described in Section 7.2.

The choice of filter h depends on a number of factors that influence the properties of the resulting wavelet, such as support length, symmetry, and number of vanishing moments (both for the wavelet and the scaling function). The tradeoffs between these criteria are discussed in Section 7.1 and in Mallat (1999, Chapter 7). Common choices in *Wavelab* include (boundary adjusted) versions of $D4$ or the symmlet $S8$.

Processing Step. Generally the estimated coefficients $\hat{w} = \eta(w)$ are found by the following recipe

$$\hat{w}_I = \begin{cases} \eta(w_I; t) & I \in \mathcal{D} \\ w_I & I \in \mathcal{A}. \end{cases}$$

Here $\eta(w_I; t)$ is a scalar function of the observed coefficient w_I, usually non-linear and depending on a parameter t. We say that η operates *co-ordinatewise*. Often, the parameter t is estimated, usually from all or some of the data at the same level as I, yielding the modified expression $\eta(w_I; t(w_j))$, where $I \in \mathcal{I}_j$. In some cases, the function η itself may depend on the coefficient index I or level j. Common examples include (compare Figure 2.1) *hard thresholding*:

$$\eta_H(w_I; t) = w_I I\{|w_I| \geq t\},$$

and *soft thresholding*:

$$\eta_S(w_I; t) = \begin{cases} w_I - t & w_I > t \\ 0 & |w_I| \leq t \\ w_I + t & w_I < -t. \end{cases}$$

These may be regarded as special cases of a more general class of *threshold shrinkage rules*, which are defined by the properties

odd:	$\eta(-x, t) = -\eta(x, t)$,
shrinks:	$\eta(x, t) \leq x$ if $x \geq 0$,
bounded:	$x - \eta(x, t) \leq t + b$ if $x \geq 0$, (some $b < \infty$),
threshold:	$\eta(x, t) = 0$ iff $\|x\| \leq t$.

Two examples (among many) of threshold shrinkage rules are provided by a) $\eta(x, t) = (1 - t^2/x^2)_+ x$ which arises in the study of the non-negative garrote Breiman (1995), and b) the posterior median, to be discussed further below.

Other choices for η, and methods for estimating t from data will be discussed in the next section. For now, we simply remark that James-Stein shrinkage, though not a co-ordinatewise thresholding method, also falls naturally into this framework:

$$\eta_{JS}(w_I; s(w_j)) = s(w_j) w_I,$$
$$s(w_j) = (1 - (2^j - 2)\sigma^2/|w_j|^2)_+.$$

While this estimator does threshold the entire signal to zero if the total energy is small

enough, $|w_j|^2 < (2^j - 2)\sigma^2$, it otherwise applies a common, data-determined *linear* shrinkage to all co-ordinates. When the true signal is sparse, this is less effective than thresholding, because either the shrinkage factor either causes substantial error in the large components, or fails to shrink the noise elements - it cannot avoid both problems simultaneously.

The estimator. Writing $\hat{f}$ for the vector $\left(\hat{f}(t_l)\right)$, we may summarize the estimation process as

$$\hat{f} = W^T \eta(Wy).$$

This representation makes the important point that the scaling and wavelet functions φ and ψ are not required or used in the calculation. So long as the filter h is of finite length, and the wavelet coefficient processing $w \to \hat{w}$ is $O(N)$, then so is the whole calculation.

Nevertheless, the iteration that occurs within the cascade algorithm generates approximations to the wavelet, as seen in Section 7.3. Thus, we may write the estimator more explicitly as

$$\begin{aligned} \hat{f}(t_l) &= \sum_I \eta_I(w)\boldsymbol{\psi}_I(t_l) \qquad (7.22) \\ &= \sum_{I\in\mathcal{A}} w_I\boldsymbol{\varphi}_I(t_l) + \sum_{I\in\mathcal{D}} \eta(w_I)\boldsymbol{\psi}_I(t_l), \end{aligned}$$

Thus, $\boldsymbol{\psi}_I = \psi_{jk}^{(J-j)}$ here is not the continuous time wavelet $\psi_{jk} = 2^{j/2}\psi(2^j \cdot -k)$, but rather the $(J-j)^{th}$ iterate of the cascade, after being scaled and located to match ψ_{jk}, compare (7.15).

The (I, l)th entry in the discrete wavelet transform matrix W is given by $\psi_{jk}^{(J-j)}(N^{-1}l)$ and in terms of the columns $\boldsymbol{\psi}_I$ of W, we have $y_l = \sum_I w_I\boldsymbol{\psi}_I(N^{-1}l)$.

First examples are given by the NMR data shown in Figure 1.2 and the simulated 'Bumps' example in Figure 7.4. The panels in Figure 1.2 correspond to the vertices of the processing diagram (7.21) (actually transposed!). The simulated example allows a comparison of soft and hard thresholding with the true signal and shows that hard thresholding here preserves the peak heights more accurately.

The thresholding estimates are

- simple, based on co-ordinatewise operations
- non-linear, and yet
- fast to compute ($O(n)$ time).

The appearance of the estimates constructed with the $\sqrt{2\log n}$ thresholds is

- noise free, with
- no peak broadening, and thus showing
- spatial adaptivity,

in the sense that more averaging is done in regions of low variability. Comparison with Figure 6.2 shows that linear methods fail to exhibit these properties.

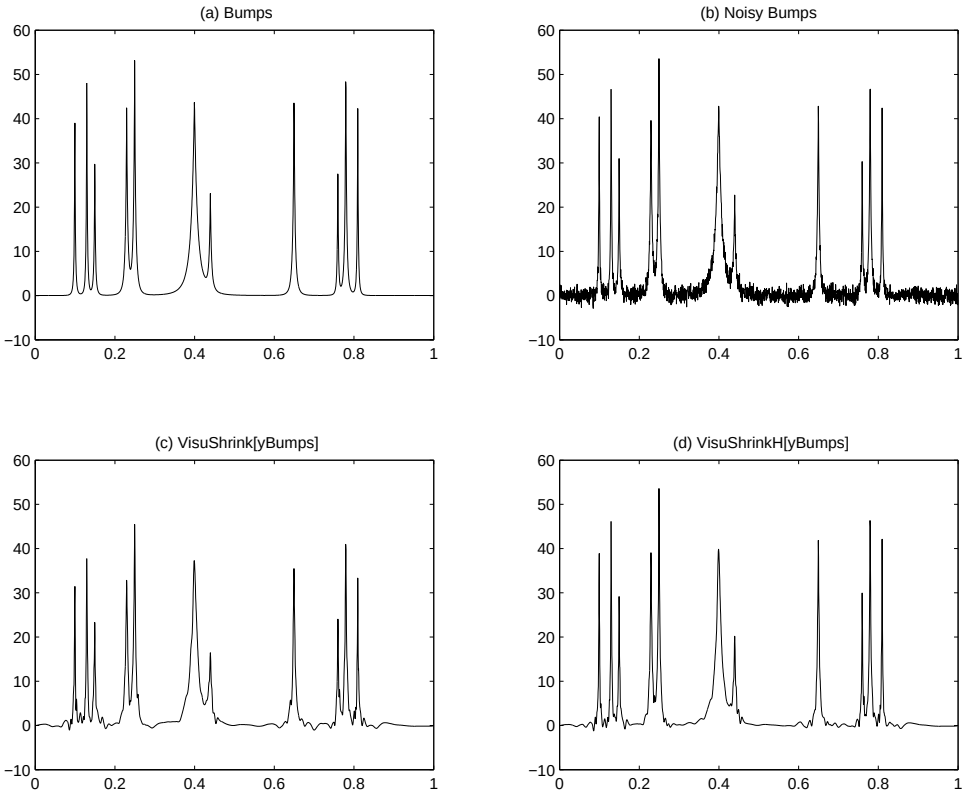

Figure 7.4 Panels (a), (b): artificial 'Bumps' signal constructed to resemble a spectrum, formula in Donoho and Johnstone (1994a), $\|f\|_N = 7$ and $N = 2048$ points. I.i.d. $N(0,1)$ noise added to signal, so signal to noise ratio is 7. Panels (c), (d): Discrete wavelet transform with Symmlet8 filter and coarse scale $L = 5$. Soft (c) and hard (d) thresholding with threshold $t = \sqrt{2\log n} \approx 3.905$.

The hidden sparsity heuristic. A rough explanation for the success of thresholding goes as follows. The model (7.20) is converted by the orthogonal wavelet transform into

$$w_I = \theta_I + \epsilon \tilde{z}_I, \qquad \epsilon = \sigma/\sqrt{n},\ \tilde{z}_I \overset{i.i.d}{\sim} N(0,1). \tag{7.23}$$

Since the noise is white (i.e. independent with constant variance) in the time domain, and the wavelet transform is orthogonal, the same property holds for the noise variables $\tilde{z}_I$ in the wavelet domain—they each contribute noise at level ϵ^2. On the other hand, in our sample signals, and more generally, it is often the case that the signal in the wavelet domain is *sparse,* i.e. its energy is largely concentrated in a few components. With concentrated signal and dispersed noise, a threshold strategy is both natural and effective, as we have seen in examples, and will see from a theoretical perspective in Chapter 9 and beyond. The sparsity of the wavelet representation may be said to be hidden, since it is not immediately apparent from the form of the signal in the time domain.

Estimation of σ. Assume that the signal is sparsely represented, and so most, if not all, data coefficients at the finest level are essentially pure noise. Since there are many (2^{J-1}) such coefficients, one can estimate σ^2 well using a robust estimator

$$\hat{\sigma}^2 = MAD\{w_{J-1,k}, k \in \mathcal{I}_{J-1}\}/0.6745,$$

which is not affected by the few coefficients which may contain large signal. Here MAD denotes the median absolute deviation (from zero). The factor 0.6745 is the population MAD of the standard normal distribution, and is used to calibrate the estimate.

Soft vs. Hard thresholding The choice of the threshold shrinkage rule η and the selection of threshold t are somewhat separate issues. The choice of η is problem dependent. For example, hard thresholding exactly preserves the data values above the threshold, and as such can be good for preserving peak heights (say in spectrum estimation), whereas soft thresholding forces a substantial shrinkage. The latter leads to smoother visual appearance of reconstructions, but this property is often at odds with that of good fidelity – as measured for example by average squared error between estimate and truth.

Correlated data. If the noise Z_l in (7.20) is stationary and correlated, then the wavelet transform has a decorrelating effect. (Johnstone and Silverman (1997) has both a heuristic and more formal discussion). In particular, the levelwise variances $\sigma_j^2 = \mathrm{Var}(w_{jk})$ are independent of k. Hence it is natural to apply *level-dependent* thresholding

$$\hat{w}_{jk} = \eta(w_{jk}, t_j).$$

For example, one might take $t_j = \hat{\sigma}_j \sqrt{2 \log n}$ with $\hat{\sigma}_j = MAD_k\{w_{jk}\}/0.6745$.

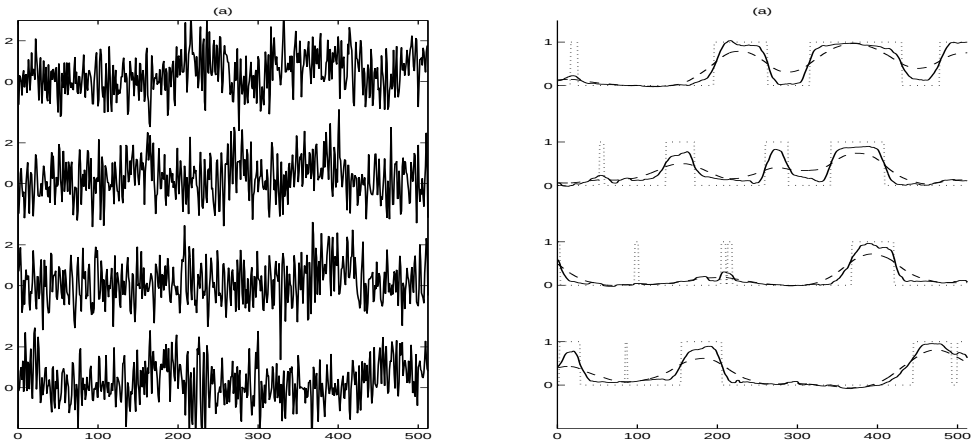

Figure 7.5 Ion channel data. Panel (a) sample trace of length 2048. Panel (b) Dotted line: true signal, Dashed line: reconstruction using (translation invariant) thresholding at $\hat{\sigma}_j \sqrt{2 \log n}$. Solid line: reconstruction using TI thresholding at data determined thresholds (a combination of SURE and universal).Further details in Johnstone and Silverman (1997).

Wavelet shrinkage as a spatially adaptive kernel method. We may write the result of

thresholding using (7.18) in the form

$$\hat{f}(t_l) = \sum_I \hat{w}_I \boldsymbol{\psi}_I(t_l) \qquad \hat{w}_I = c_I(y) w_I \tag{7.24}$$

where we have here written $\eta_I(w)$ in the "adaptive linear shrinkage" form $c_I(w) y_I$.

Inserting the wavelet transform representation (7.19) into (7.24) leads to a kernel representation for $\hat{f}(t_l)$:

$$\hat{f}(t_i) = \sum_I \sum_l c_I(y) \boldsymbol{\psi}_I(t_l) \boldsymbol{\psi}_I(t_i) y_l = \sum_l \hat{K}(t_i, t_l) y_l,$$

where the *kernel*

$$\hat{K}(s,t) = \sum_I c_I(y) \boldsymbol{\psi}_I(s) \boldsymbol{\psi}_I(t), \qquad s,t \in \{t_l = l/N\}. \tag{7.25}$$

The hat in this kernel emphasizes that it depends on the data through the coefficients $c_I(y)$. The individual component kernels $K_I(t,s) = \boldsymbol{\psi}_I(t)\boldsymbol{\psi}_I(s)$ have bandwidth $2^{-j}B$ where B is the support length of the filter h. Hence, one may say that the bandwidth of $\hat{K}$ at t_l is of order $2^{-j(t_l)}$, where

$$j(t_l) = \max\{j : c_I(y)\boldsymbol{\psi}_I(t_l) \neq 0, \text{ some } I \in \mathcal{I}_j\}.$$

In other words, t_l must lie within the support of a level j wavelet for which the corresponding data coefficient is not thresholded to zero. Alternatively, if a fine scale coefficient estimate $\hat{w}_{jk} \neq 0$, then there is a narrow effective bandwidth near $2^{-j}k$. Compare Figure 7.6. By separating the terms in (7.25) corresponding to the approximation set $\mathcal{A}$ and the detail set $\mathcal{D}$, we may decompose

$$\hat{K} = K_A + \hat{K}_D$$

where the approximation kernel $K_A(t_l, t_m) = \sum_k \boldsymbol{\varphi}_I(t_l)\boldsymbol{\varphi}_I(t_m)$ does not depend on the observed data y.

Exercise. With W the $N \times N$ discrete wavelet transform matrix, let $C = \text{diag}\,(c_I)$ be a diagonal matrix with entries c_I defined as above and let $\delta_l \in \mathbb{R}^N$ have zero entries except for a 1 in the l-th place. Show that the adaptive kernel at t_l, namely the vector $\hat{K}_l = \{\hat{K}(t_l, t_m)\}_{m=1}^N$, may be calculated using the wavelet transform via $\hat{K}_l = W^t C W \delta_l$.

Translation invariant versions. The discrete wavelet transform (DWT) is not shift invariant: the transform of a shifted signal is not the same as a shift of the transformed original. This arises because of the dyadic downsampling between levels that makes the DWT nonredundant. For example, the Haar transform of a step function with jump at $1/2$ has only one non-zero coefficient, whereas if the step is shifted to say, $1/3$, then there are $\log_2 N$ non-zero coefficients.

The transform, and the resulting threshold estimates, can be made invariant to shifts by multiples of N^{-1} by the simple device of averaging. Let S denote the operation of circular shifting by N^{-1}: $Sf(k/N) = f((k+1)/N)$, except for the endpoint which is wrapped around: $Sf(1) = f(1/N)$. Define

$$\hat{f}^{TI} = \text{Ave}_{1 \le k \le N}\, (S^{-k} \circ \hat{f} \circ S^k).$$

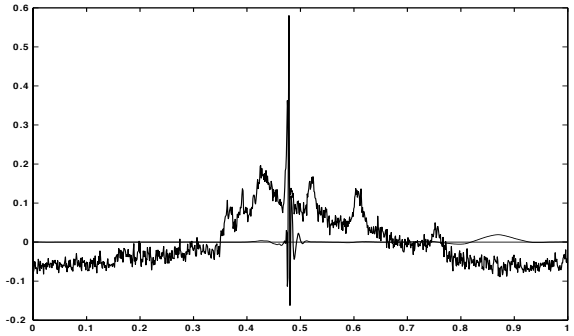

Figure 7.6 Produces figure showing spatially adaptive kernel features of hard thresholding as applied to the RaphaelNMR signal

The translation invariant (TI) estimator averages over all N shifts, and so would appear to involve at least $O(N^2)$ calculation. However, the proposers of this method, Coifman and Donoho (1995), describe how the algorithm can in fact be implemented in $O(N \log N)$ operations.

It can be seen from Figure 7.7 that the extra averaging implicit in $\hat{f}^{TI}$ reduces artifacts considerably. Experience in practice has generally been that translation invariant averaging improves the performance of virtually every method of thresholding, and its use is encouraged in situations where the $\log N$ computational penalty is not serious.

7.6 Choice of threshold.

We give only a partial discussion of this large topic here, and choose only among methods that have some theoretical support.

The key features of a threshold method are firstly, the existence of a *threshold zone* $[-t, t]$ in which all observed data is set to zero. This allows the estimator to exploit sparse signal representations by ensuring that the mean squared error is very small in the majority of co-ordinates in which the true signal is negligible.

Secondly the *tail behavior* of the estimate as $|x| \to \infty$ is also significant. More specifically, the growth of $x - \eta(x)$ – approaching zero or a constant or diverging – influences the bias properties of the estimate, particularly for large signal components.

Often, one may know from previous experience or subjective belief that a particular choice of threshold (say 3σ or 5σ) is appropriate. On the other hand, one may seek an *automatic* method for setting a threshold, and this will be the focus of subsequent discussion.

'Automatic' thresholding methods can be broadly divided into *fixed* versus *data-dependent*. "Fixed" methods set a threshold in advance of observing data. One may use a fixed number of standard deviations $k\sigma$, or a more conservative limit, such as the *universal* threshold $t = \sigma\sqrt{2\log n}$.

1. 'Universal' threshold $\lambda_n = \sqrt{2\log n}$. This is a fixed threshold method, and can be used with either soft or hard thresholding. If $Z_1, \ldots, Z_n$ are i.i.d. $N(0,1)$ variates, then it

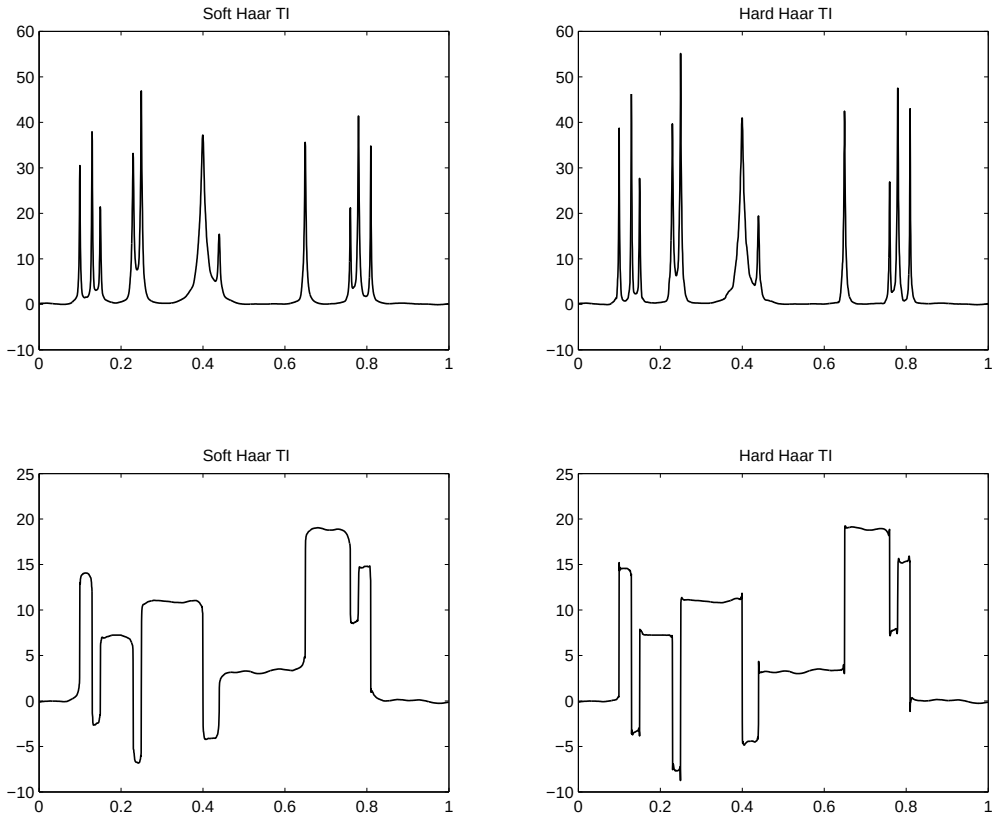

Figure 7.7 Figure to show SoftHaarTI-Blocks and HardHaarTI-Blocks with reference to CoDo paper for other test functions.

can be shown (compare (8.21)) that for $n \geq 2$,

$$P_n = P\{\max_{1 \leq i \leq n} |Z_i| > \sqrt{2 \log n}\} \leq \frac{1}{\sqrt{\pi \log n}}.$$

Similarly, it can be shown that the expected number of $|Z_i|$ that exceed the threshold will satisfy the same bound. Thus, for a wide range of values of n, including $64 = 2^6 \leq n \leq 2^{20}$, the expected number of exceedances will be between 0.15 and 0.25, so only in at most a quarter of realizations will *any* pure noise variables exceed the threshold.

[Since the wavelet transform is orthogonal,

$$P\{\hat{f}_n \equiv 0 | f \equiv 0\} = 1 - P_n \to 1.$$

Thus, with high probability, no "spurious structure" is declared, and in this sense, the universal threshold leads to a "noise free" reconstruction. [Note however that this does not mean that $\hat{f} = f$ with high probability when $f \neq 0$, since $\hat{f}$ is not linear in y.]

The price for this admirably conservative performance is that the method chooses large thresholds, which can lead to noticeable bias at certain signal strengths. This shows up in the

theory as extra logarithmic terms in the rate of convergence of this estimator, e.g. Theorem 10.10.

When combined with the soft thresholding non-linearity, the universal threshold leads to visually smooth reconstructions, but at the cost of considerable bias and relatively high mean squared error (cf. Donoho et al. (1995)).

2. False discovery rate (FDR) thresholding. This is a data dependent method for hard thresholding that is typically applied levelwise in the wavelet transform. Suppose that $y_i \sim N(\theta_i, \sigma^2)$ are independent, and form the order statistics of the magnitudes:

$$|y|_{(1)} \geq |y|_{(2)} \geq \ldots \geq |y|_{(n)}.$$

Fix the *false discovery rate* parameter $q \in (0, 1/2]$. Form quantiles $t_k = \sigma z(q/2 \cdot k/n)$. Let $\hat{k}_F = \max\{k : |y|_{(k)} \geq t_k\}$, and set $\hat{t_F} = t_{\hat{k}_F}$ and use this as the hard threshold

$$\hat{\theta}_k(y) = y_k I\{|y_k| \geq \hat{t}_F\}. \tag{7.26}$$

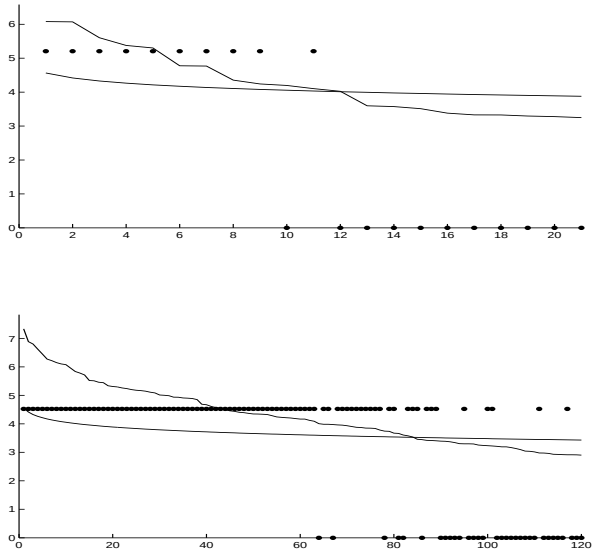

Figure 7.8 (a) 10 out of 10,000. $\mu_i = \mu_0 \doteq 5.21$ for $i = 1, \ldots, n_0 = 10$ and $\mu_i = 0$ if $i = 11, 12, \ldots, n = 10,000$. Data y_i from model (1.3), $\epsilon = 1$. Solid line: ordered data $|y|_{(k)}$. Solid circles: true unobserved mean value μ_i corresponding to observed $|y|_{(k)}$. Dashed line: FDR quantile boundary $t_k = z(q/2 \cdot k/n)$, $q = 0.05$. Last crossing at $\hat{k}_F = 12$ producing threshold $\hat{t}_F = 4.02$. Thus $|y|_{(10)}$ and $|y|_{(12)}$ are false discoveries out of a total of $\hat{k}_F = 12$ discoveries. The empirical false discovery rate $F\hat{D}R = 2/12$. (b) 100 out of 10,000. $\mu_i = \mu_0 \doteq 4.52$ for $i = 1, \ldots, n_0 = 100$; otherwise zero. Same FDR quantile boundary, $q = 0.05$. Now there are $\hat{k}_F = 84$ discoveries, yielding $\hat{t}_F = 3.54$ and $F\hat{D}R = 5/84$. (from Abramovich et al. (2006).)

The boundary sequence (t_k) may be thought of as a sequence of thresholds for $t-$statistics in model selection: the more variables (i.e. coefficients in our setting) enter, the easier it is for still more to be accepted (i.e. pass the threshold unscathed.)

As is shown in Abramovich et al. (2006), the FDR estimator has excellent mean squared error performance in sparse multinormal mean situations - for example being asymptotically adaptive minimax over ℓ_p balls. In addition (unpublished), it achieves the "right" rates of convergence over Besov function classes - thus removing the logarithmic terms present when the $\sqrt{2\log n}$ threshold is used.

However, the choice of q is an issue requiring further study – the smaller the value of q, the larger the thresholds, and the more conservative the threshold behavior becomes.

3. Stein's unbiased risk estimate (SURE) thresholding. This is a data dependent method for use with soft thresholding, again typically level by level. It has the special feature of allowing for certain kinds of correlation in the noise. Thus, assume that $y \sim N_n(\theta, V)$, and assume that the diagonal elements σ_{kk} of the covariance matrix are constant and equal to σ^2. This situation arises, for example, if in the wavelet domain, $k \to y_{jk}$ is a stationary process.

At (2.54), we derived the unbiased risk criterion for soft thresholding, and found that $E_\theta\|\hat{\theta} - \theta\|^2 = E_\theta \hat{U}(t)$, where (putting in the noise level σ^2)

$$\hat{U}(t) = \sigma^2 n + \sum_k y_k^2 \wedge t^2 - 2\sigma^2 \sum_k I\{|y_k| \le t\}.$$

Now set

$$\hat{t}_{SURE} = \operatorname*{argmin}_{0 \le t \le \sigma\sqrt{2\log n}} \hat{U}(t).$$

The criterion $\hat{U}(t)$ does not depend on details of the correlation (σ_{jk}, $j \ne k$) and so can be used in correlated data settings when the correlation structure is unknown, without the need of estimating it.

The minimization can be carried out in $O(n \log n)$ time.

The SURE estimate also removes logarithmic terms in the rates of convergence of wavelet shrinkage estimates over Besov classes (though a 'pretest' is needed in certain cases to complete the proofs).

4. Empirical Bayes. This data dependent method for levelwise thresholding provides a family of variants on soft and hard thresholding. Again assume an independent normal means model, $y_i = \theta_i + \sigma z_i$, with z_i i.i.d standard normal. As in Section 2.4, allow θ_i to independently be drawn from a mixture prior distribution π:

$$\theta_i \sim (1 - w)\delta_0 + w\gamma_a.$$

Here w is the probability that θ_i is non-zero, and $\gamma_a(d\theta)$ is a family of distributions with scale parameter $a > 0$, for example the double exponential

$$\gamma_a(d\theta) = (a/2)e^{-a|\theta|}d\theta.$$

Using L_1 loss $\|\hat{\theta} - \theta\|_1 = \sum_1^n |\hat{\theta}_i - \theta_i|$, it was shown in Section 2.4 that the Bayes rule for this prior is the *median* $\hat{\theta}_{EB}(y)$ of the posterior distribution of θ given y :

$$\hat{\theta}_{EB,i}(y) = \eta(y_i; w, a),$$

and that the posterior *median* η has *threshold* structure:

$$\eta(y; w, a) = 0 \quad \text{if} \quad |y| \le \sigma t(w, a),$$

while for large $|y|$, it turns out that $|y - \eta(y)| \sim \sigma a$.

The hyperparameters (w, a) can be estimated by maximizing the marginal likelihood of (w, a) given data (y_i). Indeed, the marginal of y_i

$$m(y_i|w,a) = \int \phi_\sigma(y_i - \theta_i)\pi(d\theta) = (1-w)\phi_\sigma(y_i) + w\int \phi_\sigma(y_i - \theta_i)\gamma_a(d\theta_i)$$

and the corresponding likelihood $\ell(w, a) = \prod_i m(y_i|w, a)$.

Theory shows that the method achieves the optimal rates of convergence, while simulations suggest that the method adapts gracefully to differing levels of sparsity at different resolution levels in the wavelet transform (Johnstone and Silverman, 2004a).

A numerical comparison. Table 7.2 is an extract from two larger tables in Johnstone and Silverman (2004a) summarizing results of a simulation comparison of 18 thresholding methods. The observations $x = \mu_0 I_S + z$ are of length 1000 with noise z_i being i.i.d. standard normal. The non-zero set S is a random subset of $\{1, \ldots, 1000\}$ for each noise realization, and each of three sizes $K = |S| = 5, 50, 500$ corresponding to 'very sparse', 'sparse' and 'dense' signals respectively. Four signal strengths $\mu_0 = 3, 4, 5$ and 7 were used, though only two are shown here. There are thus $3 \times 4 = 12$ configurations. One hundred replications were carried out for each of the values of K and μ_0, with the same 100,000 noise variables used for each set of replications.

Among the 18 estimators, we select here: 'Universal' or $\sqrt{2\log n} \approx 3.716$ thresholding, FDR thresholding with $q = 0.1$ and 0.01, SURE thresholding, and empirical Bayes thresholding respectively with $a = 0.2$ fixed and w estimated, and with (a, w) estimated, in both cases by marginal maximum likelihood.

For each method $\hat{\theta}_m$ and configuration θ_c, the average total squared error was recorded over the $n_r = 100$ replications:

$$r(\hat{\theta}_m.\theta_c) = n_r^{-1}\sum_1^{n_r} \|\hat{\theta}_m(\theta_c + z_r) - \theta_c\|^2.$$

Some results are given in Table 7.2 and the following conclusions can be drawn.

- hard thresholding with the universal threshold particularly with moderate or large amounts of moderate sized signal, can give disastrous results.
- Estimating the scale parameter a is probably preferable to using a fixed value, though it does lead to slower computations. In general, the automatic choice is quite good at tracking the best fixed choice, especially for sparse and weak signal.
- SURE is a competitor when the signal size is small ($\mu_0 = 3$) but performs poorly when μ_0 is larger, particularly in the sparser cases.
- If q is chosen appropriately, FDR can outperform exponential in some cases, but the choice of q is crucial and varies from case to case.

An alternative way to compare methods is through their *inefficiency*, which compares the risk of $\hat{\theta}_m$ for a given configuration θ_c with the best over all 18 methods:

$$\text{ineff}(\hat{\theta}_m, \theta_c) = 100\left[\frac{r(\hat{\theta}_m.\theta_c)}{\min_m r(\hat{\theta}_m.\theta_c)} - 1\right].$$

Table 7.2 *Average of total squared error of estimation of various methods on a mixed signal of length 1000.*

Number nonzero	5		50		500				
Value nonzero	3	5	3	5	3	5	**med**	**10th**	**max**
$a = 0.2$	38	18	299	95	1061	665	*18*	*30*	*48*
exponential	36	17	214	101	857	783	7	30	52
SURE	38	42	202	210	829	835	35	151	676
FDR q=0.01	43	26	392	125	2568	656	44	91	210
FDR q=0.1	40	19	280	113	1149	651	18	39	139
universal soft	42	73	417	720	4156	7157	529	1282	1367
universal hard	39	18	370	163	3672	1578	50	159	359

The inefficiency vector ineff$(\hat{\theta}_m)$ for a given method has 12 components (corresponding to the configurations θ_c) and Table 7.2 also records three upper quantiles of this vector: median, and 10th and 12th largest. Minimizing inefficiency has a minimax flavor—it turns out that the empirical Bayes methods have the best inefficiencies in this experiment.

5. Block Thresholding `[TO BE ADDED.]`

7.7 Further Details

Proof of Lemma 7.4. We write this out for a_j; there is a parallel argument for d_j. The argument is by induction. The case $r = 1$ is the analysis step (7.10). For general r, (7.10) gives

$$a_{j-r}[k] = Rh \star a_{j-r+1}[2k],$$

and using the induction hypothesis for $r - 1$, we obtain

$$\begin{aligned} a_{j-r}[k] &= \sum_l h[l-2k] \sum_n h^{(r-1)}[n - 2^{r-1}l]\, a_j[n] \\ &= \sum_n a_j[n] \sum_l h^{(r-1)}[n - 2^{r-1}l]\, h[l-2k]. \end{aligned}$$

Now $h[l-2k] = Z^{r-1}h[2^{r-1}l - 2^r k]$ and since $Z^{r-1}h[m] = 0$ unless $m = 2^{r-1}l$, and so the inner sum equals

$$\sum_m h^{(r-1)}[n-m] Z^{r-1}h[m - 2^r k] = h^{(r-1)} \star Z^{r-1}h[n - 2^r k] = h^{(r)}[n - 2^r k]. \qquad \square$$

Relating $h^{(r)}$ to φ. Recall that the scaling function φ was defined by the Fourier domain formula $\widehat{\varphi(\xi)} = \prod_{j=1}^{\infty} \frac{\hat{h}(2^{-j}\xi)}{\sqrt{2}}$. This suggests that we look at the Fourier transform of $h^{(r)}$. First note that the transform of zero padding is given by

$$\widehat{Zh}(\omega) = \sum_l e^{-il\omega} Zh[l] = \sum_k e^{-i2k\omega} h[k] = \hat{h}(2\omega),$$

so that $\widehat{h^{(r)}}(\omega) = \prod_{p=0}^{r-1} \hat{h}(2^p\omega)$. Making the substitution $\omega = 2^{-r}\xi$, we are led to define an r^{th} approximation to φ as a distribution $\varphi^{(r)}$ having Fourier transform

$$\widehat{\varphi^{(r)}}(\xi) = 2^{-r/2}\widehat{h^{(r)}}(2^{-r}\xi) = \prod_{j=1}^{r} \frac{\hat{h}(2^{-j}\xi)}{\sqrt{2}}. \tag{7.27}$$

We now verify that $\varphi^{(r)}$ can be thought of as a function (or more precisely, a measure) defined on $2^{-r}\mathbb{Z}$. Indeed, a discrete measure $\mu = \sum_n m[n]\delta_{2^{-r}n}$ supported on $2^{-r}\mathbb{Z}$ has Fourier transform

$$\hat{\mu}(\xi) = \int e^{-i\xi x}\mu(dx) = \sum_n m[n]e^{-i\xi 2^{-r}n} = \hat{m}(2^{-r}\xi).$$

Thus, the quantity $2^{-r/2}\widehat{h^{(r)}}(2^{-r}\xi)$ in (7.27) is the Fourier transform of a measure $\sum_n 2^{-r/2}h^{(r)}[n]\delta_{2^{-r}n}$. Secondly, a real valued function $g(2^{-r}n)$ defined on $2^{-r}\mathbb{Z}$ is naturally associated to the measure $\mu_g = \sum_n 2^{-r}g(2^{-r}n)\delta_{2^{-r}n}$, (the normalizing multiple 2^{-r} can be motivated by considering integrals of functions against μ_g). Combining these two remarks shows that $\varphi^{(r)}$ is indeed a function on $2^{-r}\mathbb{Z}$, with

$$2^{-r}\varphi^{(r)}(2^{-r}n) = 2^{-r/2}h^{(r)}[n]. \tag{7.28}$$

PROOF OF PROPOSITION 7.6. We first re-interpret the results of Lemma 7.4. Suppose $j < J$. Since $\varphi_{jk} \in V_J$, we have

$$\varphi_{jk} = \sum_n \langle \varphi_{jk}, \varphi_{Jn}\rangle \varphi_{Jn},$$

(and similarly for $\psi_{jk} \in W_j \subset V_J$.) If $f \in V_J$ and as before we set $a_j[k] = \langle f, \varphi_{jk}\rangle$, and $d_j[k] = \langle f, \psi_{jk}\rangle$, then by taking inner products with f in the previous display,

$$a_j[k] = \sum_n \langle \varphi_{jk}, \varphi_{Jn}\rangle a_J[n].$$

Replacing j with $J - r$ and comparing the results with those of the Lemma, we conclude that

$$\langle \varphi_{J-r,k}, \varphi_{Jn}\rangle = h^{(r)}[n - 2^r k], \qquad \langle \psi_{J-r,k}, \varphi_{Jn}\rangle = g^{(r)}[n - 2^r k].$$

Comparing the first of these with (7.28) and replacing $r = J - j$, we get

$$\langle \varphi_{jk}, \varphi_{Jn}\rangle = 2^{(j-J)/2}\varphi^{(J-j)}(2^{j-J}n - k) = N^{-1/2}\varphi_{jk}^{(J-j)}(n/N),$$

which is the second equation of Proposition 7.6. The first follows similarly.

PROOF OF PROPOSITION 7.7. Let $r = J - j$, so that $a_j = a_{J-r}$ and, using Lemma 7.4, $a_j[k] = \sum_n h^{(r)}[n - 2^r k]\, a_J[n]$. From (7.14),

$$h^{(r)}[n - 2^r k] = 2^{-r/2}\varphi^{(r)}(2^{-r}n - k) = N^{-1/2}\varphi_{jk}^{(r)}(N^{-1}n),$$

which implies that $a_j[k] = N^{-1}\sum_n \varphi_{jk}^{(r)}(N^{-1}n) f(N^{-1}n) = \langle \varphi_{jk}^{(J-j)}, f\rangle_N$. The argument for $d_j[k]$ is exactly analogous.

7.8 Notes

§1. Many expositions rightly begin with the continuous wavelet transform, and then discuss frames in detail before specialising to orthogonal wavelet bases. However, as the statistical theory mostly uses orthobases, we jump directly to the definition of multiresolution analysis due to Mallat and Meyer here in a unidimensional form given by Hernández and Weiss (1996):

1. Warning: many authors use the opposite convention $V_{j+1} \subset V_j$!

3. Conditions (i) - (iv) are not mutually independent -see Hernández and Weiss (1996).

Unequally spaced data? `[TC & LW: fill in!]`

More remarks on L_1 loss leading to posterior median.

Include Eisenberg example?

Topics not covered here: Extensions to other data formats: time series spectral density estimation, count data and Poisson estimation.

Books specifcially focused on wavelets in statistics include Ogden (1997), Vidakovic (1999), Jansen (2001) and Nason (2008). The emphasis in these books is more on describing methods and software and less on theoretical properties. Härdle et al. (1998) is a more theoretically oriented treatment of wavelets, approximation and statistical estimation, and has considerable overlap in content with the later chapters of this book, though with a broader focus than the sequence model alone.

8

Thresholding and Oracle inequalities

Less is more. (*Anon.*)

Oracle, n. something regarded as an infallible guide or indicator, esp. when its action is viewed as recondite or mysterious; a thing which provides information, insight, or answers. (Oxford English Dictionary)

Thresholding is very common, even if much of the time it is conducted informally, or perhaps most often, unconsciously. Most empirical data analyses involve, at the exploration stage, some sort of search for large regression coefficients, correlations or variances, with only those that appear "large", or "interesting" being retained for reporting purposes, or in order to guide further analysis.

For all its ubiquity, thresholding has received much less theoretical attention than linear estimation methods, such as those we have considered until now. This is perhaps due, in part, to the non-linearity that is inherent to thresholding: a scaled up version of the data does *not* always yield a proportionately scaled-up version of the estimate, since the very act of scaling up the data may put it over the retention threshold.

Consequently, the bias-variance decomposition cannot be used as directly for threshold estimators as for linear ones: one needs other features of the distribution of the data beyond first and second moments. The main concern of this chapter will therefore be to develop tools for analysing and understanding the mean squared error of soft and hard thresholding and its dependence on both the unknown mean and the threshold level.

Section 8.1 begins with a simple univariate mean squared error bound for hard thresholding. This is immediately used to show much faster rates of convergence over ℓ_1 balls in $\mathbb{R}^n$ than are possible with linear estimators.

A more systematic comparison of soft and hard thresholding begins in Section 8.2, with univariate upper and lower bounds for mean squared error that differ only at the level of constants. Soft thresholding is easier to study theoretically, but is not always better in practice.

Turning to data in n dimensions, we look at the properties of thresholding at $\epsilon\sqrt{2\log n}$, a value closely connected with the maximum (alsolute) value of n independent standard normal variates, here thought of as pure noise. Its mean squared error, for *any* signal θ in white Gaussian noise, is within a logarithmic factor of that achievable by an oracle who knows which co-ordinates exceed the noise level.

Without further information on the nature or size of θ, this logarithmic factor cannot be

improved. The demonstration of this, outlined in Section 8.5, benefits from the use of sparse two point priors, supported mostly on 0 but partly on a given $\mu > 0$. So these are studied first in Section 8.4, where they are set up precisely so that observed data near μ will still, in the posterior, be construed as most likely to have come from the atom at 0!

A simple class of models for a sparse signal says that at most a small number of coordinates can be non-zero, k out of n say, though we do not know *which* ones. The minimax risk for estimation of θ in such cases is studied in Sections 8.6–8.9, and is shown, for example, to be asymptotic to $2\epsilon_n^2 k_n \log(n/k_n)$ in the case $k_n \to \infty$ more slowly than n. Thresholding rules are asymptotically minimax in this case, and the upper bound is an easy consequence of earlier results in this chapter. The lower bound proceeds in two steps, through study of sparse priors in a univariate model, Section 8.7, followed by use of the Bayes-minimax method sketched in Chapter 4.

The highly sparse case, in which k_n remains bounded as n grows, has some special features. Section 8.9 looks at the case of a *single* spike and develops non-asymptotic upper and lower risk bounds, which also find use later in Chapter 11.

8.1 A crude MSE bound for hard thresholding.

Consider a single observation $y \sim N(\theta, \epsilon^2)$. The thresholding estimator may be written as $\hat{\theta}(y) = yI_E$ where E is the event $\{|y| > \lambda\epsilon\}$ on which y exceeds the threshold and is retained.

Denote the mean squared error of $\hat{\theta}$ by $r_H(\lambda, \theta) = E_\theta[yI_E - \theta]^2$. We construct two bounds for the mean squared error, according as the signal θ is smaller than the noise ϵ or not. It will be seen that this has the character of a bias *or* variance decomposition – since such a thing is of course not really possible, we are forced to accept extra terms, either additive or multiplicative, in the analogs of bias and variance.

Proposition 8.1 *If $y \sim N(\theta, \epsilon^2)$, there exists a constant M such that if $\lambda \geq 4$*

$$r_H(\lambda, \theta) \leq \begin{cases} M[\theta^2 + \lambda\phi(\lambda - 1)\epsilon^2] & \text{if } |\theta| \leq \epsilon \\ M\lambda^2\epsilon^2 & \text{if } |\theta| > \epsilon. \end{cases} \tag{8.1}$$

[As usual, ϕ denotes the standard normal density function.]

Proof Consider first the small signal case $|\theta| < \epsilon$. Arguing crudely,

$$E_\theta[yI_E - \theta]^2 \leq 2E_\theta y^2 I_E + 2\theta^2.$$

The first term is largest when $|\theta| = \epsilon$. In this case, if we set $x = y/\epsilon \sim N(1, 1)$ then

$$E_\theta y^2 I_E \leq \epsilon^2 \cdot 2\int_\lambda^\infty x^2\phi(x - 1)dx \leq 4\lambda\phi(\lambda - 1)\epsilon^2, \tag{8.2}$$

where we used the fact that for $y \geq 3$, $(y+1)^2\phi(y) \leq 2(y^2-1)\phi(y) = 2(d/dy)[-y\phi(y)]$.

In the large signal case, $|\theta| > \epsilon$, we use the relation $y = \theta + \epsilon z$ to analyse by cases, obtaining

$$yI_E - \theta = \begin{cases} \epsilon z & \text{if } |y| > \lambda\epsilon, \\ \epsilon z - y & \text{if } |y| \leq \lambda\epsilon, \end{cases}$$

so that in either case

$$(yI_E - \theta)^2 \le 2\epsilon^2(z^2 + \lambda^2).$$

Taking expectations gives the result, for example with $M = 8$. We have however de-emphasized the explicit constants (which will be improved later anyway in Lemma 8.5 and (8.17)) to emphasise the structure of the bound, which is the most important point here. □

Exercise 8.2 shows how the condition $\lambda > 4$ can be removed.

From the proof, one sees that when the signal is small, the threshold produces zero most of the time and the MSE is essentially the resulting bias plus a term for 'rare' errors which push the data beyond the threshold. When the signal is large, the data is left alone, and hence has variance of order ϵ, except that errors of order $\lambda\epsilon$ are produced about half the time when $\theta = \lambda\epsilon$!

Example 8.2 Let us see how (8.1) yields rough but useful information in an n-dimensional estimation problem. Suppose, as in the introductory example of Section 1.3, that $y \sim N_n(\theta, \epsilon_n^2, I)$ with $\epsilon_n = n^{-1/2}$ and that θ is assumed to be constrained to lie in an ℓ_1-ball $\Theta_{n,1} = \{\theta \in \mathbb{R}^n : \sum |\theta_i| \le 1\}$. On this set, the minimax risk for linear estimation equals $1/2$ (shown at (9.21) in the next chapter), but thresholding does much better. Let B_n be the set of "big" coordinates $|\theta_i| \ge \epsilon = n^{-1/2}$, and $S_n = B_n^c$. Clearly, when $\theta \in \Theta_{n,1}$, the number of big coordinates is relatively limited: $|B_n| \le n^{1/2}$. For the 'small' coordinates, $\theta_i^2 \le n^{-1/2}|\theta_i|$, so $\sum_{S_n} \theta_i^2 \le n^{-1/2}$. Now using (8.1)

$$\begin{aligned}\sum r_H(\lambda, \theta_i) &\le M \sum_{B_n} \lambda^2\epsilon^2 + M \sum_{S_n} [\theta_i^2 + \lambda\phi(\lambda - 1)\epsilon^2] \\ &\le M\lambda^2 n^{-1/2} + M[n^{-1/2} + \lambda\phi(\lambda - 1)].\end{aligned}$$

Choosing, for now, $\lambda = 1 + \sqrt{\log n}$, so that $\phi(\lambda - 1) = \phi(0)n^{-1/2}$, we finally arrive at

$$E\|\hat{\theta}_\lambda - \theta\|^2 \le M' \log n/\sqrt{n}.$$

While this argument does not give exactly the right rate of convergence, which is $(\log n/n)^{1/2}$, let alone the correct constant, compare (13.27) and Theorem 13.16, it already shows clearly that thresholding is much superior to linear estimation on the ℓ_1 ball.

8.2 Properties of Thresholding Estimators

In this section we consider two types of thresholding estimators $\hat{\delta}(x)$ in the simplest univariate case: $x \sim N(\mu, 1)$. We introduce some special notation for noise level $\epsilon = 1$.

Hard Thresholding.

$$\hat{\delta}_H(x, \lambda) = \begin{cases} x & |x| > \lambda \\ 0 & |x| \le \lambda. \end{cases} \tag{8.3}$$

Soft Thresholding.

$$\hat{\delta}_S(x, \lambda) = \begin{cases} x - \lambda & x > \lambda \\ 0 & |x| \le \lambda \\ x + \lambda & x < -\lambda. \end{cases} \tag{8.4}$$

Similarities. These two estimators are both *non-linear,* and in particular have in common the notion of a *threshold region* $|x| \leq \lambda$ in which no signal is estimated. Of course, hard thresholding is discontinuous, while soft thresholding is constructed to be continuous, which explains the names. Compare Figure 2.1. The threshold parameter in principle can vary over the entire range $[0, \infty]$, so the family includes the special linear estimators $\hat{\delta}(x, 0) = x$ and $\hat{\delta}(x, \infty) = 0$ that "keep" and "kill" the data respectively. In general, however, we will be interested in thresholds in the range between about 1.5 and a value proportional to the square root of log-sample-size. We now make some comments specific to each class.

Differences. Hard thresholding preserves the data outside the threshold zone, which can be important in certain applications, for example in denoising where it is desired to preserve as much as possible the heights of true peaks in estimated spectra. The mathematical consequence of the discontinuity is that the risk properties of hard thresholding are a little more awkward—for example the mean squared error is not monotonic increasing in $\mu \geq 0$. Hard thresholding also has the interesting property that it arises as the solution of a penalized least squares problem

$$\hat{\delta}_H(x, \lambda) = \arg\min_\mu (x - \mu)^2 + \lambda^2 I\{\mu \neq 0\}.$$

Indeed, when $\mu \neq 0$ the criterion has minimum value λ^2 when $\mu = x$ and when μ vanishes, the criterion equals x^2. Hard thresholding amounts to choosing the better of these two values.

Soft thresholding, on the other hand, shrinks the data towards 0 outside the threshold zone. The mean squared error function is now monotone in $\mu \geq 0$, and we will see later that the shrinkage aspect leads to significant smoothing properties in function estimation (e.g. Chapter 10). In practice, however, neither soft nor hard thresholding is universally preferable—the particular features of the application play an important role. The estimator that we call soft thresholding has appeared frequently in the statistics literature, for example Efron and Morris (1971), who term it a "limited-translation" rule. Soft thresholding also arises from a penalized least squares problem

$$\hat{\delta}_S(x, \lambda) = \arg\min_\mu (x - \mu)^2 + 2\lambda|\mu|,$$

as may be verified directly.

Notice that in the case of n-dimensional data, the same calculation can be conducted co-ordinatewise:

$$\hat{\delta}_S(x, \lambda) = \arg\min_{\mu \in \mathbb{R}^n} \sum_i (x_i - \mu_i)^2 + 2\lambda \sum_i |\mu_i|,$$

Since the penalty term is an ℓ_1 norm of μ, soft thresholding is sometimes also called the ℓ_1- *rule.* In the same vein, the corresponding penalty for hard thresholding in $\mathbb{R}^n$ is $\sum_i I\{\mu_i \neq 0\}$, which with slight abuse of notation might be called an ℓ_0 *penalty.*

Compromises. Many compromises between soft and hard thresholding are possible that appear in principle to offer many of the advantages of both methods: a threshold region for small x and exact or near fidelity to the data when x is large.

1) soft-hard thresholding (Gao and Bruce, 1997): This is a compromise between soft and

hard thresholding defined by

$$\hat{\delta}_{\lambda_1,\lambda_2}(x) = \begin{cases} 0 & \text{if } |x| \le \lambda_1 \\ \operatorname{sgn}(x)\frac{\lambda_2(|x|-\lambda_1)}{\lambda_2-\lambda_1} & \text{if } \lambda_1 < |x| \le \lambda_2 \\ x & \text{if } |x| > \lambda_2. \end{cases}$$

2) $\hat{\delta}(x) = (x - \lambda^2/x)_+$ suggested by Gao (1998) based on the "garotte" of Breiman (1995).

3) $\hat{\delta}(x)$ constructed as the posterior *median* for a prior distribution that mixes a point mass at zero with a Gaussian of specified variance (Abramovich et al., 1998).

While these and other proposals can offer useful advantages in practice, for these notes we concentrate on soft and hard thresholding, because of their simplicity and the fact that they encompass the main theoretical phenomena.

Soft thresholding.

The explicit risk function $r_S(\mu,\lambda) = E[\hat{\delta}_S(x,\lambda) - \mu]^2$ can be calculated by considering the various zones separately – explicit formulas are given in Section 8.11. Here we focus on qualitative properties and bounds. We first restate for completeness some results already proved in Section 2.7. Write $\Phi(A) = \int_A \phi(z)dz$ for the standard Gaussian measure of an interval A and let $I_\lambda = [-\lambda,\lambda]$. The risk function of soft thresholding is increasing:

$$\frac{\partial}{\partial\mu} r_S(\lambda,\mu) = 2\mu\Phi([I_\lambda - \mu]) \le 2\mu, \tag{8.5}$$

while

$$r(\lambda,\infty) = 1 + \lambda^2, \tag{8.6}$$

which shows the effect of the bias due to the shrinkage by λ, and

$$r_S(\lambda,0) = 2\int_\lambda^\infty (z-\lambda)^2\phi(z)dz \quad \begin{cases} \le e^{-\lambda^2/2} & (\text{all } \lambda) \\ \le 4\lambda^{-1}\phi(\lambda) & (\lambda \ge \sqrt{2}). \\ \sim 4\lambda^{-3}\phi(\lambda) & (\lambda \text{ large}). \end{cases} \tag{8.7}$$

(compare Exercise 8.3). A sharper bound is sometimes useful (also Exercise 8.3)

$$r_S(\lambda,0) \le 4\lambda^{-3}(1 + 1.5\lambda^{-2})\phi(\lambda), \tag{8.8}$$

valid for all $\lambda > 0$. The risk at $\mu = 0$ is small because errors are only made when the observation falls outside the threshold zone.

We summarize and extend some of these conclusions about the risk properties:

Lemma 8.3 *Let* $\bar{r}(\lambda,\mu) = \min\{r_S(\lambda,0) + \mu^2, 1 + \lambda^2\}$. *For all* $\lambda > 0$ *and* $\mu \in \mathbb{R}$,

$$\tfrac{1}{2}\bar{r}(\lambda,\mu) \le r_S(\lambda,\mu) \le \bar{r}(\lambda,\mu). \tag{8.9}$$

The risk bound $\bar{r}(\lambda,\mu)$ has the same qualitative flavor as the crude bound (8.1) derived earlier for hard thresholding, only now the constants are correct. In fact, the bound is sharp when μ is close to 0 or ∞. We may interpret $r_S(\lambda,0) + \mu^2$ as a "bias" term, adjusted for

risk at zero, and $1 + \lambda^2$ as a "variance" term, reflecting the risk for large μ. Figure 8.1 gives a qualitative picture of these bounds.

Proof Symmetry of the risk function means that we may assume without loss that $\mu \geq 0$. Write $r_\mu(\lambda, s) = (\partial/\partial\mu) r_S(\lambda, \mu)|_{\mu=s}$. By (8.5), the partial derivative $r_\mu \leq 2\mu$, and so

$$r_S(\lambda, \mu) - r_S(\lambda, 0) = \int_0^\mu r_\mu(\lambda, s) ds \leq \mu^2. \tag{8.10}$$

The upper bound follows from this and (8.6). For the lower bound, write $x = \mu + z$, and use the simple decomposition

$$E_\mu[\hat{\delta}_S(x, \lambda) - \mu]^2 \geq E[(z - \lambda)^2, z + \mu > \lambda] + \mu^2 P(z + \mu < \lambda). \tag{8.11}$$

If $\mu \leq \lambda$, the right side is bounded below by

$$E[(z - \lambda)^2, z > \lambda] + \mu^2/2 = (r_S(\lambda, 0) + \mu^2)/2,$$

using (8.7). If $\mu \geq \lambda$, then from monotonicity of the risk function, $r_S(\lambda, \mu) \geq r_S(\lambda, \lambda)$, and applying (8.11) at $\mu = \lambda$,

$$r_S(\lambda, \mu) \geq E[(z - \lambda)^2, z > 0] + \lambda^2/2 = \lambda^2 - 2\lambda\phi(0) + 1/2 \geq (\lambda^2 + 1)/2$$

with the last inequality valid if and only if $\lambda \geq \sqrt{8/\pi}$. In this case, the right sides of the last two displays both exceed $\bar{r}(\lambda, \mu)/2$ and we are done. The proof of the lower bound for $\lambda < \sqrt{8/\pi}$ is deferred to the Appendix. □

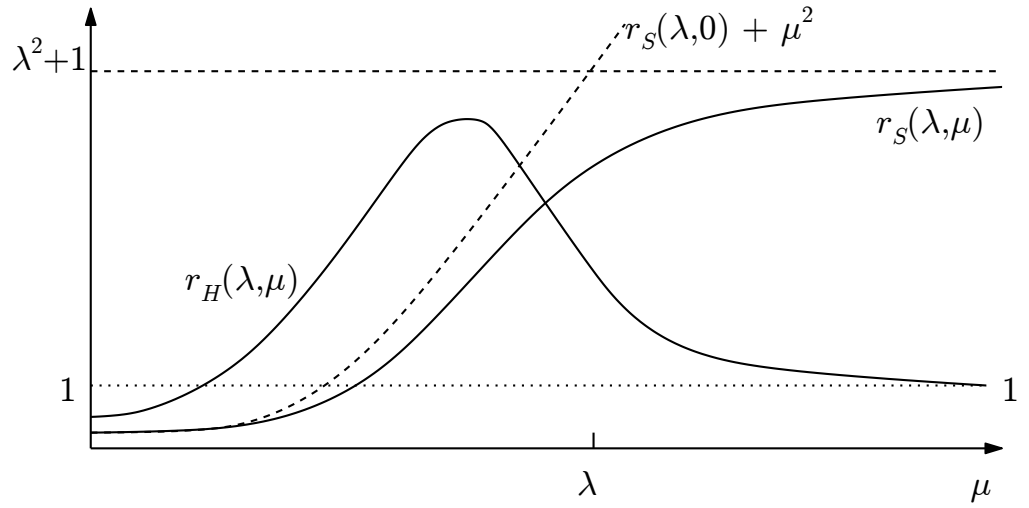

Figure 8.1 Schematic diagram of risk functions of soft and hard thresholding. Dashed lines indicate upper bounds for soft thresholding of Lemma 8.3.

Consequences of (8.9) are well suited to showing the relation between sparsity and quality of estimation. As was also shown in Section 2.7, using elementary properties of minima, one may write

$$r_S(\lambda, \mu) \leq r_S(\lambda, 0) + (1 + \lambda^2) \wedge \mu^2. \tag{8.12}$$

In conjunction with the bound $r_S(\lambda, 0) \le e^{-\lambda^2/2}$, (8.7), we arrive at

Corollary 8.4 *Suppose $y \sim N(\theta, \epsilon^2)$. Let $\delta > 0$ and $\lambda_\delta = \sqrt{2 \log \delta^{-1}}$. Then*

$$r_S(\lambda_\delta, \theta) \le \delta\epsilon^2 + (1 + 2\log\delta^{-1})(\theta^2 \wedge \epsilon^2). \tag{8.13}$$

Hard thresholding

The risk function is easily written in the form

$$r_H(\lambda, \mu) = \mu^2 \Phi(I_\lambda - \mu) + \int_{|z+\mu|>\lambda} z^2 \phi(z) dz. \tag{8.14}$$

The extreme values for small and large μ are:

$$\begin{aligned} r_H(\lambda, \infty) &= 1 \\ r_H(\lambda, 0) &= 2\int_\lambda^\infty z^2 \phi(z) dz = 2\lambda\phi(\lambda) + 2\tilde{\Phi}(\lambda) \sim 2\lambda\phi(\lambda), \end{aligned} \tag{8.15}$$

as $\lambda \to \infty$. Note that the value at ∞ reflects only variance and no bias, while the value at zero is small, though larger than that for soft thresholding due to the discontinuity at λ. However (8.14) also shows that there is a large risk near $\mu = \lambda$: for large λ:

$$r_H(\lambda, \lambda) \sim \lambda^2/2.$$

See Exercise 8.5 for more information near $\mu = \lambda$.

An analogue of the upper bound of Lemma 8.3 is available for hard thresholding. In this case, define

$$\bar{r}(\lambda, \mu) = \begin{cases} \min\{r(\lambda, 0) + 1.2\mu^2, 1 + \mu^2\} & 0 \le \mu \le \lambda \\ 1 + \mu^2 \tilde{\Phi}(\mu - \lambda) & \mu \ge \lambda, \end{cases}$$

and extend $\bar{r}$ to negative μ by making it an even function.

Lemma 8.5 *(a) For $\lambda > 0$ and $\mu \in \mathbb{R}$,*

$$(5/12)\bar{r}(\lambda, \mu) \le r_H(\lambda, \mu) \le \bar{r}(\lambda, \mu). \tag{8.16}$$

(b) The large μ component of $\bar{r}$ has the bound

$$\sup_{\mu \ge \lambda} \mu^2 \tilde{\Phi}(\mu - \lambda) \le \begin{cases} \lambda^2/2 & \text{if } \lambda \ge \sqrt{2\pi}, \\ \lambda^2 & \text{if } \lambda \ge 1. \end{cases}$$

Proof Again we assume without loss that $\mu \ge 0$. The upper bound for $\mu \ge \lambda$ is a direct consequence of (8.14). For $0 \le \mu \le \lambda$, the approach is as used for (8.10), but for the details of the bound $0 \le (\partial/\partial\mu) r_H(\lambda, \mu) \le 2.4\mu$, we refer to Donoho and Johnstone (1994a, Lemma 1). As a result we obtain, for $0 \le \mu \le \lambda$,

$$r_H(\lambda, \mu) \le r_H(\lambda, 0) + 1.2\mu^2.$$

The alternate bound, $r_H(\lambda, \mu) \le 1 + \mu^2$, is immediate from (8.14).

The lower bound is actually easier—by checking separately the cases $\mu \geq \lambda$ and $\mu \leq \lambda$, it is a direct consequence of an inequality analogous to (8.11):

$$E_\mu[\hat{\delta}_H(x,\lambda) - \mu]^2 \geq E[z^2, z + \mu > \lambda] + \mu^2 P(z + \mu < \lambda).$$

For part (b), set $\alpha = \mu - \lambda \geq 0$ and define $g(\alpha) = (\lambda + \alpha)^2 \tilde{\Phi}(\alpha)$. We have

$$g'(\alpha) = (\lambda + \alpha)\phi(\alpha)h(\alpha), \qquad h(\alpha) = 2\big(\tilde{\Phi}(\alpha)/\phi(\alpha)\big) - \lambda - \alpha,$$

and $h(0) = \sqrt{2\pi} - \lambda \leq 0$ if $\lambda \geq \sqrt{2\pi}$. Differentiation and the bound $\tilde{\Phi}(\alpha) \leq \phi(\alpha)/\alpha$ show that h is decreasing and hence negative on $[0,\infty)$, so that $g(\alpha) \leq g(0) = \lambda^2/2$. In the case where we only assume that $\lambda \geq 1$, we have $g(\alpha) \leq \lambda^2(1+\alpha)^2\tilde{\Phi}(\alpha) \leq \lambda^2$, as may be checked numerically, or by calculus. □

For use in later sections, we record some corollaries of the risk bounds. First, for $\lambda \geq 1$,

$$r_H(\lambda,\mu) \leq \begin{cases} r_H(\lambda,0) + 1.2\mu^2 & \mu \leq 1 \\ 1 + \lambda^2 & \mu > 1. \end{cases} \tag{8.17}$$

Second (Exercise 8.4)

$$r_H(\lambda,0) \leq \begin{cases} (2\lambda + \sqrt{2\pi})\phi(\lambda) & \text{all } \lambda > 0 \\ 4\lambda\phi(\lambda) & \lambda > 1. \end{cases} \tag{8.18}$$

Remark. In both cases, we have seen that the maximum risk of soft and hard thresholding is $O(\lambda^2)$. This is a necessary consequence of having a threshold region $[-\lambda,\lambda]$: if $\hat{\delta}(x)$ is any estimator vanishing for $|x| \leq \lambda$, then simply by considering the error made by estimating 0 when $\mu = \lambda$, we find that

$$E_\lambda(\hat{\delta}(x) - \lambda)^2 \geq \lambda^2 P_\lambda\{|x| \leq \lambda\} \approx \lambda^2/2 \qquad \text{for large } \lambda. \tag{8.19}$$

8.3 Thresholding in $\mathbb{R}^n$ and Oracle Inequalities

Let us turn now to the vector setting in which we observe n co-ordinates, $y_i = \theta_i + \epsilon z_i$, with as usual, z_i being i.i.d. $N(0,1)$. A leading example results from the discrete equispaced regression model (7.20) after applying a discrete orthogonal wavelet transform, compare (7.23).

Consider an estimator built from soft (or hard) thresholding applied co-ordinatewise, at threshold $\lambda_n = \epsilon\sqrt{2\log n}$:

$$\hat{\theta}^S_{\lambda_n,i} = \hat{\delta}_S(y_i, \epsilon\sqrt{2\log n}), \tag{8.20}$$

and let $\hat{\theta}^H_{\lambda_n}$ denote hard thresholding at the same level.

Remarks. 1. Here is one reason for the specific choice $\lambda_n = \sqrt{2\log n}$ (other choices will be discussed later.) We show that this threshold level is conservative, in the sense that

$$P\{\hat{\theta} = 0 | \theta = 0\} \to 1$$

as $n \to \infty$, so that with high probability, $\hat{\theta}$ does not assert the presence of "spurious structure". To verify this, note that if each y_i is distributed independently as $N(0, \epsilon^2)$, then the probability that no observation exceeds the threshold λ_n equals the extreme value probability

$$\pi_n = P\{\max_{i=1,\dots,n} |Z_i| \geq \sqrt{2\log n}\} = 1 - \left[1 - 2\tilde{\Phi}\left(\sqrt{2\log n}\right)\right]^n \leq \frac{1}{\sqrt{\pi \log n}}, \quad (8.21)$$

valid for $n \geq 2$ (see 3° in Appendix).

Table 8.1 compares the exact value π_n of the extreme value probability with the upper bound b_n given in (8.21). Also shown is the expectation of the number N_n of values Z_i that exceed the $\sqrt{2\log n}$ threshold. It is clear that the exceedance probability converges to zero rather slowly, but also from the expected values that the *number* of exceedances is at most one with much higher probability, greater than about 97%, even for n large. Compare also Exercise 8.6. And looking at the ratios b_n/π_n, one sees that while the bound b_n is not fully sharp, it does indicate the (slow) rate of approach of the exceedence probability to zero.

n	$\sqrt{2\log n}$	π_n	π_n^W	$\mathbb{E}N_n$	b_n
32	2.63	0.238	0.248	0.271	0.303
64	2.88	0.223	0.231	0.251	0.277
128	3.12	0.210	0.217	0.235	0.256
256	3.33	0.199	0.206	0.222	0.240
512	3.53	0.190	0.196	0.211	0.226
1024	3.72	0.182	0.188	0.201	0.214
2048	3.91	0.175	0.180	0.193	0.204
4096	4.08	0.169	0.174	0.186	0.196

Table 8.1 *For i.i.d. Gaussian noise: sample size n, threshold* $\sqrt{2\log n}$*, exceedance probability* π_n*, extreme value theory approximation* π_n^W *expected number of exceedances* $\mathbb{E}N_n$*, upper bound* b_n *of* (8.21)

The classical extreme value theory result Galambos (1978, p. 69) for the maximum of n i.i.d. $N(0, 1)$ variables Z_i, namely $M_n = \max_{i=1,\dots,n} Z_i$ states that

$$b_n^{-1}[M_n - a_n] \xrightarrow{\mathcal{D}} W, \qquad P(W \leq t) = \exp\{-e^{-t}\}, \quad (8.22)$$

where $a_n = \sqrt{2\log n} - (\log\log n + \log 4\pi)/(2\sqrt{2\log n})$ and $b_n = 1/\sqrt{2\log n}$. Section 8.10 has some more information on the law of M_n.

Here we are actually more interested in $\max_{i=1,\dots,n} |Z_i|$, but this is described quite well by M_{2n}. (Exercise 8.7 explains why). Thus the exceedance probability π_n might be approximated by $\pi_n^W = P(W \leq c_{2n})$ where $c_{2n} = (\sqrt{2\log n} - a_{2n})/b_{2n})$. Although the convergence to the extreme value distribution in (8.22) is slow, of order $1/\log n$ (e.g. Hall (1979), Galambos (1978, p. 140)). Table 8.1 shows the extreme value approximation to be better than the direct bound (8.21).

A non-asymptotic bound follows from the Tsirelson-Sudakov-Ibragimov bound Proposition 2.10 for a Lipschitz(1) function $f : \mathbb{R}^n \to \mathbb{R}$ of a standard Gaussian $n-$vector $Z \sim N_n(0, I)$:

$$P\{|f(Z) - Ef(Z)| \geq t\} \leq 2e^{-t^2/2}.$$

When applied to $f(z) = \max |z_i|$, this says that the tails of $\max |Z_i|$ are sub-Gaussian, while

the extreme value result in fact says that the limiting distribution has negligible variability around a_n.

Alan Miller's variable selection scheme. A method of Miller (1984, 1990) offers an interesting perspective on $\sqrt{2\log n}$ thresholding. Consider a traditional linear regression model

$$y = X\beta + \sigma^2 z,$$

where y has N components and X has $n < N$ columns $[x_1 \cdots x_n]$ and the noise $z \sim N_N(0, I)$. For convenience only, assume that the columns are centered and scaled: $x_i^t 1 = 0$ and $|x_i|^2 = 1$. Now create "fake" regression variables x_i^*, each as an independent random permutation of the entries in the corresponding column x_i. Assemble X and $X^* = [x_1^* \cdots x_n^*]$ into a larger design matrix $\tilde{X} = [X\ X^*]$ with coefficients $\tilde{\beta}^t = [\beta^t\ \beta^{*t}]$ and fit the enlarged regression model $y = \tilde{X}\tilde{\beta}$ by a forward stepwise method. Let the method stop just before the first 'fake' variable x_i^* enters the model. Since the new variables x_i^* are approximately orthonormal among themselves and approximately orthogonal to each x_i, the estimated coefficients $\hat{\beta}_i^*$ are essentially i.i.d. $N(0, 1)$, and so the stopping criterion amounts to "enter variables above the threshold given by $\max_{i=1,\ldots,n} |\hat{\beta}_i^*| \doteq \sqrt{2\log n}$.

Ideal Risk. Suppose that $y_i = \theta_i + \epsilon z_i,\ \ i = 1, \ldots n$, with, as usual z_i being i.i.d. $N(0, 1)$. Given a fixed value of θ, an *ideal* linear estimator $\theta_{c,i}^* = c_i^* y_i$ would achieve the best possible mean squared error among linear estimators for the given θ:

$$\min_{c_i} r(\theta_{c,i}^*, \theta) = \frac{\theta_i^2 \epsilon^2}{\theta_i^2 + \epsilon^2} \in [\tfrac{1}{2}, 1] \cdot \theta_i^2 \wedge \epsilon^2.$$

Because of the final bound, we might even restrict attention to the *ideal projection*, which chooses c_i from 0 or 1 to attain

$$\min_{c_i \in \{0,1\}} r(\theta_{c,i}^*, \theta) = \theta_i^2 \wedge \epsilon^2.$$

Thus the optimal projection choice $c_i(\theta)$ equals 1 if $\theta_i^2 \geq \epsilon^2$ and 0 otherwise, so that

$$\theta_i^*(y) = \begin{cases} y_i & \text{if } \theta_i^2 \geq \epsilon^2 \\ 0 & \text{if } \theta_i^2 \leq \epsilon^2. \end{cases}$$

One can imagine an "oracle", who has partial, but valuable, information about the unknown θ: for example, which co-ordinates are worth estimating and which can be safely ignored. Thus, with the aid of a "projection oracle", the best mean squared error attainable is the *ideal risk*:

$$\mathcal{R}(\theta, \epsilon^2) = \sum_i \theta_i^2 \wedge \epsilon^2,$$

[more correctly, $\mathcal{R}(\theta, \epsilon^2) = \sum_i (\theta_i^2 \wedge \epsilon^2)$.] In Chapter 9 we will discuss further the significance of the ideal risk, and especially its interpretation in terms of sparsity.

Of course, the statistician does not normally have access to such oracles, but we now show that it is nevertheless possible to mimick the ideal risk with threshold estimators, at least up to a precise logarithmic factor.

Proposition 8.6 *Suppose that* $y \sim N_n(\theta, I)$. *For the soft thresholding estimator* (8.20) *with* $\lambda_n = \epsilon\sqrt{2\log n}$,

$$E\|\hat{\theta}^S_{\lambda_n} - \theta\|_2^2 \le (2\log n + 1)\Big[\epsilon^2 + \sum_1^n \theta_i^2 \wedge \epsilon^2\Big]. \tag{8.23}$$

A similar result holds for $\hat{\theta}^H_{\lambda_n}$, *with the multiplier* $(2\log n + 1)$ *replaced by* $(2\log n + 1.2)$.

The factor $2\log n$ *is optimal without further restrictions on* θ, *as* $n \to \infty$,

$$\inf_{\hat{\theta}} \sup_{\theta\in\mathbb{R}^n} \frac{E\|\hat{\theta} - \theta\|^2}{\epsilon^2 + \sum_1^n \theta_i^2 \wedge \epsilon^2} \ge (2\log n)(1 + o(1)). \tag{8.24}$$

Results of this type help to render the idea of ideal risk statistically meaningful: a genuine estimator, depending only on available data, and not upon access to an oracle, can achieve the ideal risk $\mathcal{R}(\theta, \epsilon)$ up to the (usually trivial) additive factor ϵ^2 and the multiplicative factor $2\log n + 1$. In turn, the lower bound (8.9) shows that the ideal risk is also a lower bound to the mean squared error of thresholding, so that

$$\tfrac{1}{2}\mathcal{R}(\theta, \epsilon) \le E\|\hat{\theta}^S_{\lambda_n} - \theta\|_2^2 \le (2\log n + 1)[\epsilon^2 + \mathcal{R}(\theta, \epsilon)].$$

This logarithmic penalty can certainly be improved if extra constraints upon θ are added: for example that θ belong to some ℓ_p ball, weak or strong (Chapter 13). However, the lower bound (8.24) shows that the $2\log n$ factor is optimal for unrestricted θ, at least asymptotically.

Note that the upper bounds are non-asymptotic, holding for all $\theta \in \mathbb{R}^n$ and $n \ge 1$.

The upper bound extends trivially to correlated, heteroscedastic data, since it thresholding depends only on the univariate marginal distributions of the data. The only change is to replace ϵ^2 by ϵ_i^2, the variance of the ith coordinate, in the ideal risk, and to modify the additive factor to ave $_{1\le i\le n}\epsilon_i^2$. There is also a version of the lower bound under some conditions on the correlation structure: for details see Johnstone and Silverman (1997).

Proof *Upper bound.* For soft thresholding, a slightly stronger result was already established as Lemma 2.9. For hard thresholding, we use (8.17) to establish the bound, for $\lambda_n = \sqrt{2\log n}$

$$r_H(\lambda, \mu) \le (2\log n + 1.2)(n^{-1} + \mu^2 \wedge 1).$$

This is clear for $\mu > 1$, while for $\mu < 1$, one verifies that $r_H(\lambda, 0) = 2\lambda\phi(\lambda) + 2\tilde{\Phi}(\lambda) \le (2\log n + 1.2)n^{-1}$ for $n \ge 2$. Finally, add over co-ordinates and rescale to noise level ϵ.

Lower bound. The proof is deferred till Section 8.5, since it uses the sparse two point priors to be discussed in the next section. □

Bound (8.8) leads to a better risk bound for threshold $\lambda_n = \sqrt{2\log n}$ at 0 for $n \ge 2$

$$r(\lambda_n, 0) \le 1/(n\sqrt{\log n}). \tag{8.25}$$

Indeed, this follows from (8.8) for $n \ge 3$ since then $\lambda_n \ge 2$, while for $n = 2$ we just evaluate risk (8.60) numerically.

8.4 Sparse two point priors

We first study the curious properties of the two point prior

$$\pi_{\alpha,\mu} = (1-\alpha)\delta_0 + \alpha\delta_\mu, \qquad \mu > 0, \tag{8.26}$$

which we call a *sparse* prior in the case when α is small. The posterior distribution is also concentrated on $\{0, \mu\}$, and

$$P(\{\mu\}|x) = \frac{\alpha\phi(x-\mu)}{\alpha\phi(x-\mu) + (1-\alpha)\phi(x)} = \frac{1}{1+m(x)},$$

where the posterior probability ratio

$$m(x) = \frac{P(\{0\}|x)}{P(\{\mu\}|x)} = \frac{(1-\alpha)}{\alpha}\frac{\phi(x)}{\phi(x-\mu)} = \frac{(1-\alpha)}{\alpha}e^{-x\mu+\mu^2/2} \tag{8.27}$$

is decreasing in x: $m(x)/m(y) = e^{-\mu(x-y)}$.

The *posterior indifference point* is that value of x at which the posterior is indifferent between 0 and μ, so that $P(\{0\}|x) = P(\{\mu\}|x)$. We focus on the apparently peculiar situation in which this indifference point lies to the *right* of μ. Indeed, posterior equivalence corresponds to $m(x) = 1$, and writing $x = \mu + a$, it follows that μ and a are related by

$$\frac{\mu^2}{2} + a\mu = \log\frac{1-\alpha}{\alpha}. \tag{8.28}$$

Clearly a is positive so long as α is small enough that $\log(1-\alpha)/\alpha > \mu^2/2$.

Definition. The two point prior $\pi_{\alpha,\mu}$ has *sparsity* α and *overshoot* a if μ satisfies (8.28).

The prior probability on 0 is so large that even if x is larger than μ, but smaller than $\mu+a$, the posterior distribution places more weight on 0 than μ. [1] See Figure 8.2.

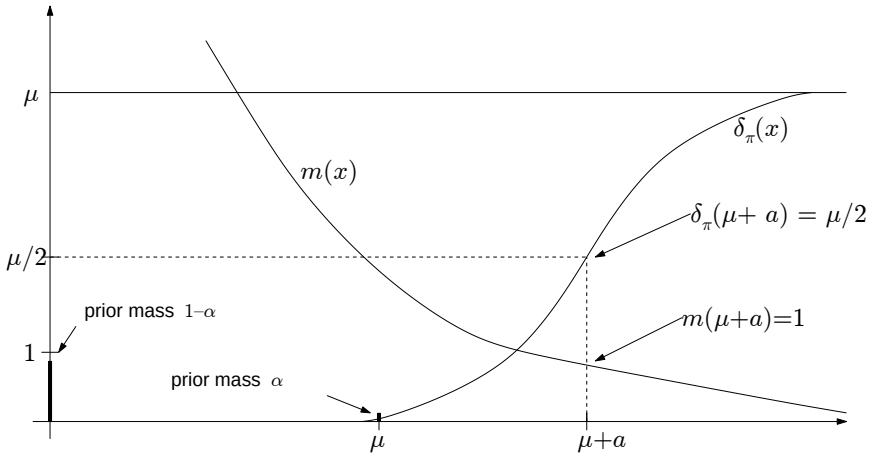

Figure 8.2 Two point priors with sparsity α and overshoot a: posterior probability ratio $m(x)$ and posterior mean $\delta_\pi(x)$

[1] Fire alarms are rare, but one may not believe that a ringing alarm signifies an actual fire without further evidence.

The Bayes rule for squared error loss, the posterior mean becomes

$$\delta_\pi(x) = \mu P(\{\mu\}|x) = \frac{\mu}{1 + m(x)}. \tag{8.29}$$

Substituting (8.28) into (8.27), we obtain $m(x) = \exp\{-\mu(x - \mu - a)\}$ and

$$\delta_\pi(\mu + z) = \frac{\mu}{1 + e^{-\mu(z-a)}}. \tag{8.30}$$

In particular, observe that $\delta_\pi(\mu)$ is small, and even $\delta_\pi(\mu + a) = \mu/2$ is far from μ.

To prepare for some asymptotics, note that if sparsity $\alpha < 1/2$ and overshoot a are given we may use (8.28) to specify $\mu(\alpha, a)$. Indeed, if $\mu_0(\alpha) = \mu(\alpha, 0) = \sqrt{2\log[(1-\alpha)/\alpha]}$, then

$$\mu(\alpha) = \mu(\alpha, a) = \sqrt{\mu_0^2(\alpha) + a^2} - a.$$

If we suppose further that $\alpha \to 0$ and $a = a(\alpha)$ is chosen so that $a(\alpha) = o(\mu_0(\alpha)) = o(\sqrt{2\log\alpha^{-1}})$, then $\mu(\alpha) \sim \mu_0(\alpha)$ and in particular,

$$\mu(\alpha) \sim \sqrt{2\log\alpha^{-1}}. \tag{8.31}$$

In this case, there is a simple and important asymptotic approximation to the Bayes risk of a sparse two point prior.

Lemma 8.7 *Let $\pi_{\alpha,\mu(\alpha)}$ have sparsity α and overshoot $a = (2\log\alpha^{-1})^\gamma$, for $0 < \gamma < 1/2$. Then, as $\alpha \to 0$,*

$$B(\pi_{\alpha,\mu(\alpha)}) \sim \alpha\mu^2(\alpha).$$

Proof By definition, we have

$$B(\pi_{\alpha,\mu(\alpha)}) = (1-\alpha)r(\delta_\pi, 0) + \alpha r(\delta_\pi, \mu(\alpha)). \tag{8.32}$$

Thus, a convenient feature of two point priors is that to study the Bayes risk, the frequentist risk function of δ_π only needs to be evaluated at two points. We give the heuristics first. When $\mu(\alpha)$ is large and the overshoot a is also large (though of smaller order), then (8.30) shows that for $x \sim N(\mu(\alpha), 1)$, the Bayes rule δ_π essentially estimates 0 with high probability, thus making an error of about μ^2. *A fortiori*, if $x \sim N(0,1)$, then δ_π estimates 0 (correctly) with even higher probability. More concretely, we will show that, as $\alpha \to 0$,

$$r(\delta_\pi, \mu(\alpha)) \sim \mu^2(\alpha), \qquad r(\delta_\pi, 0) = o(\alpha\mu^2(\alpha)). \tag{8.33}$$

Inserting these relations into the Bayes risk formula (8.32) yields the result.

The first relation is relatively easy to obtain. Using (8.30), we may write

$$r(\delta_\pi, \mu(\alpha)) = \mu^2(\alpha)\int_{-\infty}^{\infty} \frac{\phi(z)dz}{[1 + e^{\mu(z-a)}]^2} \sim \mu(\alpha)^2, \tag{8.34}$$

as $\alpha \to 0$, since the integral converges to 1 as both $\mu(\alpha)$ and $a(\alpha) \to \infty$ by the dominated convergence theorem. The second relation takes a little extra work, see the appendix. □

8.5 Optimality of $\sqrt{2\log n}$ risk bound

To establish the minimax lower bound (8.24), we set $\epsilon = 1$ without loss of generality and bring in a non-standard loss function

$$\tilde{L}(\hat{\theta}, \theta) = \frac{\|\hat{\theta} - \theta\|^2}{1 + \sum_i \theta_i^2 \wedge 1}.$$

Let $\tilde{r}(\hat{\theta}, \theta)$ and $\tilde{B}(\hat{\theta}, \pi) = \int \tilde{r}(\hat{\theta}, \theta)\pi(d\theta)$ respectively denote risk and integrated risk for the new loss function. By the usual arguments

$$\tilde{R}_N = \inf_{\hat{\theta}} \sup_{\theta} \tilde{r}(\hat{\theta}, \theta) \geq \inf_{\hat{\theta}} \sup_{\pi} \tilde{B}(\hat{\theta}, \pi) \geq \sup_{\pi} \tilde{B}(\pi),$$

(using only the elementary part of the minimax theorem). So we look for approximately least favorable distributions, and in particular construct π_n from i.i.d. draws from a sparse two-point prior π_{α_n,μ_n} with sparsity $\alpha_n = (\log n)/n$ and overshoot $a = (2\log \alpha_n^{-1})^{\gamma}$ for some $0 < \gamma \leq 1/2$. As seen in the previous section, this guarantees that

$$\mu_n = \mu(\alpha_n) \sim \sqrt{2\log \alpha_n^{-1}} \sim \sqrt{2\log n}.$$

Under n draws from $\pi_{\alpha,\mu}$, the number of nonzero θ_i has $N_n \sim \mathrm{Bin}(n, \alpha_n)$ with $EN_n = n\alpha_n = \log n$ and $\mathrm{Var}\, N_n \leq n\alpha_n = \log n$. Therefore $1 + \sum \theta_i^2 \wedge 1 = 1 + N_n \sim 1 + \log n$.

From these calculations and Lemma 8.7, we see that $B(\pi_n) \sim n\alpha_n\mu_n^2 \sim \log n \cdot 2\log n$, from which it is plausible that

$$\tilde{R}_n \geq \tilde{B}(\pi_n) \sim B(\pi_n)/(1 + \log n) \geq (2\log n)(1 + o(1)). \tag{8.35}$$

The remaining proof details are given in Section 8.11.

8.6 Minimax Risk for sparse vectors in $\mathbb{R}^n$

A natural measure of the sparsity of a vector $\theta \in \mathbb{R}^n$ is obtained by simply counting the number of nonzero components,

$$\|\theta\|_0 = \#\{\, i : \theta_i \neq 0 \,\}.$$

The subscript 0 acknowledges that this measure is sometimes called the ℓ_0-norm—somewhat inaccurately as it is not homogeneous.

The set of k-sparse vectors in $\mathbb{R}^n$ will be denoted by

$$\Theta_{n,0}(k) = \{\theta \in \mathbb{R}^n : \|\theta\|_0 \leq k\}, \tag{8.36}$$

though we often just abbreviate this as $\Theta_n(k)$. If $k \ll n$ and the components of θ represent pixel intensities then $\Theta_n(k)$, perhaps with an additional constraint that $\theta_i \geq 0$, models the collection of "nearly black" images (Donoho et al., 1992).

In the Gaussian white noise model

$$y_i = \theta_i + \epsilon z_i, \qquad i = 1, \ldots, n, \tag{8.37}$$

we might expect that knowledge that θ is sparse, or "nearly black", could be exploited to yield more accurate estimates. Thus, we might expect that the minimax risk

$$R_N(\Theta_n(k_n), \epsilon_n) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta_n(k_n)} E\|\hat{\theta} - \theta\|_2^2 \tag{8.38}$$

would be smaller, perhaps much smaller, than $R_N(\mathbb{R}^n, \epsilon_n) = n\epsilon_n^2$.

Theorem 8.8 *Assume model (8.37) and parameter space (8.36) with $k = k_n$. If $k_n/n \to 0$ with $k_n \to \infty$, then the minimax risk (8.38) satisfies*

$$R_N(\Theta_n(k_n), \epsilon_n) \sim 2\epsilon_n^2 k_n \log(n/k_n),$$

and the (soft or hard) thresholding estimator $\hat{\theta}_i(y) = \hat{\delta}(y_i, \epsilon_n\sqrt{2\log(n/k_n)})$ is asymptotically minimax.

Partial Proof. We establish the upper bound by using risk bounds for thresholding established in §8.2. For the lower bound we indicated heuristically how the result follows using sparse priors; the full proof is deferred to the next section.

By rescaling, it suffices to consider $\epsilon_n = 1$. For the upper bound, as in §8.2, let $r(\lambda, \mu)$ denote the univariate risk of thresholding at λ. Since $\mu \to r(\lambda, \mu)$ is for both soft and hard thresholding bounded by $1 + \lambda^2$ for all μ, and on $\Theta_n(k_n)$ at most k_n co-ordinates are non-zero, we have

$$\sum_1^n r(\lambda, \mu_i) \le (n - k_n) r(\lambda, 0) + k_n \sup_\mu r(\lambda, \mu) \le 4n\lambda^a \phi(\lambda) + k_n(1 + \lambda^2).$$

In the second inequality, the bound has exponent $a = -1$ for soft thresholding, (8.7), and exponent $a = 1$ for hard thresholding, (8.18). With $\lambda_n = \sqrt{2\log(n/k_n)}$, we have $\phi(\lambda_n) = \phi(0)k_n/n$, and so the first term on the right side is bounded by $c\lambda_n^a k_n$ and hence is of smaller order than the second term, which is itself asymptotically $2k_n \log(n/k_n)$, as claimed.

For the lower bound, again with $\epsilon_n = 1$, we have by the usual arguments

$$R_N(\Theta_n(k_n), \epsilon_n) \ge \sup\{B(\pi) \,:\, \text{supp}\, \pi \subset \Theta_n(k_n)\}.$$

For an approximately least favorable prior for $\Theta_n(k_n)$, take n i.i.d draws from a (univariate) sparse two point prior with sparsity $\alpha_n = k_n/n$ and overshoot $a = (2\log n/k_n)^{1/4}$. Of course, this is not quite right, as $N_n = \#\{i : \theta_i \ne 0\} \sim \text{Bin}(n, \alpha_n)$ has $EN_n = n\alpha_n = k_n$, so π_n doesn't quite concentrate on $\Theta_n(k_n)$. However, this can be fixed by setting $\alpha_n = \gamma k_n/n$ with $\gamma < 1$ and by further technical arguments set out in the next two sections and Exercise 8.12. Proceeding heuristically, then, we appeal to Lemma 8.7 to calculate

$$B(\pi_n) \sim n \cdot \alpha_n \mu^2(\alpha_n) \sim 2k_n \log(n/k_n),$$

since $\mu^2(\alpha_n) \sim 2\log \alpha_n^{-1} \sim 2\log n/k_n$. □

8.7 Sparse estimation—univariate model

In this section and the next we formalize the notion of classes of η-sparse signals and consider minimax estimation over such classes. We begin with a univariate model and then show how it leads to results for sparse estimation in $\mathbb{R}^n$. Suppose, therefore, that $Y = \theta + \epsilon Z$ with

$Z \sim N(0,1)$ and that θ is drawn from a distribution π which is non-zero with probability at most η. Thus let $\mathcal{P}(\mathbb{R})$ be the collection of probability measures on $\mathbb{R}$ and

$$\mathfrak{m}_0(\eta) = \{\, \pi \in \mathcal{P}(\mathbb{R}) :\ \pi(\{0\}) \geq 1-\eta \,\}.$$

Equivalently, $\mathfrak{m}_0(\eta)$ consists of those probability measures having a represention

$$\pi = (1-\eta)\delta_0 + \eta\nu, \tag{8.39}$$

where δ_0 is a unit point mass at 0 and ν an arbitrary probability distribution on $\mathbb{R}$. To avoid trivial cases, assume that $0 < \eta < 1$.

Given π, the integrated risk, using squared error loss, for an estimator $\hat{\theta}(y)$ of θ is then $B(\hat{\theta}, \pi) = E_\pi(\hat{\theta}(Y) - \theta)^2$. We study the Bayes minimax risk

$$\beta_0(\eta, \epsilon) = \inf_{\hat{\theta}} \sup_{\pi \in \mathfrak{m}_0(\eta)} B(\hat{\theta}, \pi) = \sup\{\, B(\pi) :\ \pi \in \mathfrak{m}_0(\eta) \,\},$$

where the second equality uses the minimax theorem 4.11. From the scale invariance

$$\beta_0(\eta, \epsilon) = \epsilon^2 \beta_0(\eta, 1),$$

it will suffice to study the unit noise quantity $\beta_0(\eta, 1)$ which we now write as $\beta_0(\eta)$. We first record properties of the least favorable distribution for fixed η and then look at the behavior of $\beta(\eta)$ as η varies.

Proposition 8.9 *Assume $0 < \eta < 1$. The Bayes minimax problem associated with $\mathfrak{m}_0(\eta)$ and $\beta_0(\eta)$ has a unique least favorable distribution π_η. The measure π_η is proper, symmetric and has countably infinite support with $\pm\infty$ as the only accumulation points.*

As the proof is a little intricate, it is postponed till the end of this section.

Proposition 8.10 *The univariate Bayes risk $\beta_0(\eta)$ is monotone increasing and continuous for $0 \leq \eta \leq 1$, with $\beta_0(\eta) \geq \eta$ and $\beta_0(1) = 1$. As $\eta \to 0$, the minimax risk*

$$\beta_0(\eta) \sim 2\eta \log \eta^{-1},$$

and an asymptotically minimax rule is given by soft thresholding at $\lambda = (2\log\eta^{-1})^{-1/2}$.

Proof First, monotonicity is obvious. For continuity suppose that $\pi_\eta = (1-\eta)\delta_0 + \eta\nu_\eta$ achieves the maximum in $\beta_0(\eta)$. The modified priors $\pi_{\omega,\eta} = (1-\omega\eta)\delta_0 + \omega\eta\nu_\eta$ converge weakly to π_η as $\omega \nearrow 1$. Consequently, from continuity of Bayes risks, Lemma 4.8,

$$\beta_0(\omega\eta) \geq B(\pi_{\omega,\eta}) \to B(\pi_\eta) = \beta_0(\eta),$$

and this is enough for continuity of the monotone β_0. When $\eta = 1$, there is no constraint on the priors, so by (4.20), $\beta_0(1) = \rho_N(\infty, 1) = 1$. Concavity implies that $B(1-\eta)\delta_0 + \eta\nu) \geq (1-\eta)B(\delta_0) + \eta B(\nu)$, and since $B(\delta_0) = 0$, maximizing over ν shows that $\beta_0(\eta) \geq \eta$.

For soft thresholding δ_λ, we have $r(\lambda, \mu) \leq 1 + \lambda^2$, compare Lemma 8.3. Since $\pi = (1-\eta)\delta_0 + \eta\nu$, we have

$$B(\delta_\lambda, \pi) = (1-\eta)r(\lambda, 0) + \eta \int r(\lambda, \mu)\nu(d\mu) \leq r(\lambda, 0) + \eta(1+\lambda^2).$$

For $\lambda = (2\log\eta^{-1})^{1/2}$, recall from (8.7) that $r(\lambda, 0) \sim 4\lambda^{-3}\phi(\lambda) = o(\eta)$, so that

$$\beta_0(\eta) \le B(\delta_\lambda, \pi) \le 2\eta\log\eta^{-1} + O(\eta).$$

For the lower bound we choose a sparse prior π as in Lemma 8.7 with sparsity η and overshoot $a = (2\log\eta^{-1})^{1/4}$. Then, from that lemma and (8.31), we obtain

$$\beta_0(\eta) \ge B(\pi_{\eta,\mu(\eta)}) \sim \eta\mu^2(\eta) \sim 2\eta\log\eta^{-1}. \qquad \square$$

Proof of Proposition 8.9 The set $\mathfrak{m}_0(\eta)$ is not weakly compact; instead we regard it as a subset of $\mathcal{P}_+(\mathbb{R})$, the substochastic measures on $\mathbb{R}$ with positive mass on $\mathbb{R}$, with the vague topology. Since $\mathfrak{m}_0(\eta)$ is then vaguely compact, we can apply Proposition 4.12 (via the remark immediately following it) to conclude the existence of a unique least favorable prior $\pi_\eta \in \mathcal{P}_+(\mathbb{R})$. Since $\eta < 1$, we know that $\pi_\eta(\mathbb{R}) > 0$. In addition, π_η is symmetric.

A separate argument is needed to show that π_η is proper, $\pi_\eta(\mathbb{R}) = 1$. Suppose on the contrary that $\alpha = 1 - \pi_\eta(\mathbb{R}) > 0$. From the Fisher information representation (4.6) and (4.21), we know that $P_0 = \Phi \star \pi_\eta$ minimizes $I(P)$ for P variying over $\mathfrak{m}_0(\eta)^\star = \{P = \Phi \star \pi : \pi \in \mathfrak{m}_0(\eta)\}$. We may therefore use the variational criterion in the form given at (C.17). Thus, let $P_1 = P_0 + \alpha\Phi \star \mu$ for an arbitrary (prior) probability measure μ on $\mathbb{R}$. Let the corresponding densities be p_1 and p_0, and set $\psi_0 = -p_0'/p_0$. Noting that $p_1 - p_0 = \alpha\phi \star \mu$, we may take $\mu = \delta_\theta$ for each $\theta \in \mathbb{R}$, and (C.17) becomes

$$E_\theta[-2\psi_0' + \psi_0^2] \le 0.$$

Stein's unbiased risk formula (2.42) applied to $d_{\pi_0}(x) = x - \psi_0(x)$ then shows that $r(d_{\pi_0}, \theta) \le 1$ for all θ. Since $d_0(x) = x$ is the *unique* minimax estimator of θ when $x \sim N(\theta, 1)$, Corollary 4.9, we have a contradiction and so it must be that $\pi_\eta(\mathbb{R}) = 1$.

As π_η is proper and least favorable, Proposition 4.13 yields a saddle point $(\hat\theta_{\pi_\eta}, \pi_\eta)$. Using the mixture representation (8.39), with ν_η corresponding to π_η, well defined because $\eta > 0$, we obtain from (4.22) applied to point masses $\nu = \delta_\theta$ that for all θ

$$r(\hat\theta_{\pi_\eta}, \theta) \le \int r(\hat\theta_{\pi_\eta}, \theta')\nu_\eta(d\theta').$$

In particular, $\theta \to r(\hat\theta_{\pi_\eta}, \theta)$ is uniformly bounded for all θ, and so is an analytic function on $\mathbb{R}$, Remark 2.4. It cannot be constant (e.g. Exercise 4.2) and so we can appeal to Lemma 4.17 to conclude that ν_η is a discrete measure with no points of accumulation in $\mathbb{R}$. The support of ν_η must be (countably) infinite, for if it were finite, the risk function of $\hat\theta_{\pi_\eta}$ would necessarily be unbounded (again, Exercise 4.2). $\square$

8.8 Minimax Bayes for sparse vectors in $\mathbb{R}^n$

We return to the n-dimensional setting and show that the gain due to sparsity can be expressed asymptotically in terms of the univariate Bayes risk $\beta_0(\eta)$.

Theorem 8.11 *Assume model (8.37) and parameter space (8.36) with $k = k_n \to \infty$. If $\eta_n = k_n/n \to \eta \ge 0$, then the minimax risk (8.38) satisfies*

$$R_N(\Theta_n(k_n), \epsilon_n) \sim n\epsilon_n^2\beta_0(k_n/n).$$

The proof uses the Bayes minimax method sketched in Chapter 4 with both upper and lower bounds derived in terms of priors built from i.i.d. draws from univariate priors in $\mathfrak{m}_0(\eta_n)$. As an intermediate, we need the class of priors supported on $\Theta_n(k_n)$,

$$\mathcal{M}_n(k_n) = \{\pi \in \mathcal{P}(\mathbb{R}^n) : \text{ supp } \pi \subset \Theta_n(k_n)\}$$

and the subclass $\mathcal{M}_n^e(k_n) \subset \mathcal{M}_n(k_n)$ of *exchangeable* priors.

The assumption that the number of non-zero terms $k_n \to \infty$ is essential for the Bayes minimax approach. If k_n remains bounded as n grows, we might say that we are in a "highly sparse" regime. The case of a single spike, $k_n = 1$, is considered in the next section, and a more general, related setting in Chapter 13.5.

Upper bound. This may be outlined in a single display,

$$R_N(\Theta_n(k_n), \epsilon_n) \le B(\mathcal{M}_n, \epsilon_n) \le B(\mathcal{M}_n^e, \epsilon_n) = n\epsilon_n^2 \beta_0(k_n/n). \tag{8.40}$$

For the details, recall that $B(\mathcal{M}, \epsilon) = \sup\{B(\pi), \pi \in \mathcal{M}\}$. The first inequality follows because $\mathcal{M}_n$ contains all point masses δ_θ for $\theta \in \Theta_n(k_n)$, compare (4.18). If we start with a draw from prior π and then permute the coordinates randomly with a permutation σ from the symmetric group S_n, we obtain a new, exchangeable prior $\pi^e = \text{ave}(\pi \circ \sigma, \sigma \in S_n)$. Concavity of the Bayes risk, Remark 4.1, guarantees that $B(\pi) \le B(\pi^e)$; this implies the second inequality.

The univariate marginal π_1 of an exchangeable prior $\pi \in \mathcal{M}_n^e(k_n)$ belongs to $\mathfrak{m}_0(k_n/n)$, and the independence trick of Lemma 4.14 says that if we make all coordinates independent with marginal π_1, then the product prior π_1^n is harder than π, so that

$$B(\pi) \le B(\pi_1^n) = n B(\pi_1).$$

Rescaling to noise level one and maximizing over $\pi_1 \in \mathfrak{m}_0(k_n/n)$, we obtain the equality in the third part of (8.40).

Lower Bound. We apply the Bayes minimax approach set out in Chapter 4.10, and in-particular in Lemma 4.28. The family of parameter spaces will be $\Theta_n(k)$, nested by k. The sequence of problems will be indexed by n, so that the noise level ϵ_n and sparsity k_n depend on n. We use the exchangeable classes of priors $\mathcal{M}_n^e$ defined above, with Bayes minimax risk given by $n\epsilon_n^2\beta_0(k_n/n)$, compare (8.40). We introduce the notation

$$B_n(\kappa, \epsilon_n) = n\epsilon_n^2 \beta_0(\kappa/n),$$

which is equally well defined for non-integer κ. For each $\gamma < 1$, then, we will construct a sequence of priors $\pi_n \in \mathcal{M}_n^e(\gamma k_n)$. Slightly different arguments are needed in the two cases $\eta_n \to \eta > 0$ and $\eta_n \to 0, k_n \to \infty$, but in both settings π_n is built from i.i.d. draws from a suitable one-dimensional distribution π_{1n}. With Θ_n denoting $\Theta_n(k_n)$, we will show that π_n has the properties

$$B(\pi_n) \ge \gamma B_n(\gamma k_n, \epsilon_n), \tag{8.41}$$

$$\pi_n(\Theta_n) \to 1, \tag{8.42}$$

$$E_{\pi_n}\{\|\hat{\theta}_{\nu_n}\|^2 + \|\theta\|^2, \Theta_n^c\} = o(B_n(\gamma k_n, \epsilon_n)) \tag{8.43}$$

where $\nu_n(\cdot) = \pi_n(\cdot|\Theta_n)$, and

$$\lim_{\gamma\to 1} \liminf_{n\to\infty} \frac{B_n(\gamma k_n, \epsilon_n)}{B_n(k_n, \epsilon_n)} = 1. \tag{8.44}$$

It then follows from Lemma 4.28 and the discussion after (4.69) that $R_N(\Theta_n(k_n), \epsilon_n) \geq B_n(k_n, \epsilon_n)(1 + o(1))$. In conjunction with the upper bound in (8.40) this will complete the proof of Theorem 8.11.

First suppose that $\eta_n \to \eta > 0$. By Proposition 8.10, we have $\beta_0(\eta_n) \to \beta_0(\eta)$. For $\gamma < 1$, we may choose M and a univariate prior $\pi_M \in \mathfrak{m}_0(\gamma\eta)$ with support contained in $[-M, M]$ and satisfying $B(\pi_M) \geq \gamma\beta_0(\gamma\eta)$, compare Exercise 4.4. The corresponding prior π_n in the noise level ϵ_n problem is constructed as $\theta_i = \epsilon_n\mu_i$, where $\mu_1, \ldots, \mu_n$ are i.i.d. draws from π_M. By construction and using $\beta_0(\eta_n) \sim \beta_0(\eta)$, we then have

$$B(\pi_n, \epsilon_n) \geq n\epsilon_n^2\gamma\beta_0(\gamma\eta) \sim \gamma B_n(\gamma k_n, \epsilon_n). \tag{8.45}$$

Since $\pi_M\{\mu_i \neq 0\} \leq \gamma\eta$, we may bound $\|\theta\|_0$ above stochastically by a Binomial$(n, \gamma\eta)$ variable, N_n say, so that

$$\pi_n\{\Theta_n^c\} \leq P\{N_n - EN_n > k_n - n\gamma\eta\} \to 0,$$

for example by Chebychev's inequality, since $\mathrm{Var}\, N_n \leq k_n \sim n\eta$.

For the technical condition (8.43), observe that under π_n, we have $\|\theta\|^2 \leq n\epsilon_n^2 M^2$ with probability one, so that the same is true for $\|\hat{\theta}_{\nu_n}\|^2$, and so the left side of (8.43) is bounded by $2nM\epsilon_n^2\pi_n(\Theta_n^c)$. On the other hand $B(\gamma k_n, \epsilon_n) \sim n\epsilon_n^2\beta_0(\gamma\eta)$, so that (8.43) also follows from $\pi_n(\Theta_n^c) \to 0$.

Property (8.44) follows from the continuity of $\beta_0(\gamma\eta)$ as $\gamma \to 1$.

The case $\eta_n \to 0, k_n \to \infty$ has a parallel argument, but makes essential use of sparse priors and the assumption that $k_n \to \infty$, as foreshadowed at the end of Section 8.6. For details, see Exercise 8.12.

8.9 Minimax risk for a single spike

Theorem 8.8 requires that the number k_n of spikes grow without bound as n increases. It is natural to enquire what happens when k_n remains bounded. In this section, we look at the simplest case, that of a single spike, $k_n = 1$. The multiple spike situation is similar, but with some non-trivial technical features, and so is postponed to Section 13.5.

Consider then the sparsest possible setting: signals in $\mathbb{R}^n$ with at most one non-zero co-ordinate. We suppose that the index of this nonzero co-ordinate is unknown and evaluate the cost of that ignorance in terms of minimax risk.

Our primary interest will lie with lower bounds, as earlier upper bounds apply easily here. To develop such lower bounds, define a 'bounded single spike' parameter set by

$$\Theta_n(\tau) = \{\theta \in \mathbb{R}^n : \theta = \gamma e_I \text{ for } I \in \{1, \ldots, n\}\,,\ |\gamma| \leq \tau\}. \tag{8.46}$$

Thus, $\Theta_n(\tau)$ is the union of n orthogonal needles, each corresponding to a $1-$dimensional bounded interval $[-\tau, \tau]$.

A natural candidate for least favorable prior on $\Theta_n(\tau)$ is the prior π_n obtained by choosing

an index $I \in \{1, \dots, n\}$ at random and then setting $\theta = \tau e_I$, for τ to be specified below. The posterior distribution of I is

$$p_{in}(y) = P(I = i|y) = \frac{\phi(y - \tau e_i)}{\sum_j \phi(y - \tau e_j)} = \frac{e^{\tau y_i}}{\sum_j e^{\tau y_j}}. \tag{8.47}$$

The posterior mean has components given, for example, by

$$\hat{\theta}_{\pi,1} = E(\theta_1|y) = \tau P(I = 1|y) = \tau p_{1n}(y).$$

The maximum *a posteriori* estimator $\hat{\theta}^{MAP}_{\pi,1} = \tau e_{\hat{I}}$, where $\hat{I} = \operatorname{argmax}_i P(I = i|y) = \operatorname{argmax}_i y_i$. The MAP estimator also has the property that if $\|\cdot\|$ is any ℓ_p norm, $0 < p \le \infty$, and $\hat{\theta}$ is any estimator,

$$P_\pi\{\|\hat{\theta} - \theta\| \ge \tau/2\} \ge P_\pi\{\|\hat{\theta}^{MAP}_\pi - \theta\| \ge \tau/2\} = P_\pi(\hat{I} \ne I). \tag{8.48}$$

By symmetry, and recalling that $y_i = \theta_i + z_i$,

$$\begin{aligned} P_\pi\{\hat{I} \ne I\} &= P_{\tau e_1}\{y_1 \ne \max_i y_i\} \\ &= P\{z_1 + \tau < \max_{j=2,\dots,n} z_j\} = P\{M_{n-1} - Z > \tau\} \end{aligned} \tag{8.49}$$

where $M_n := \max_{j=1,\dots,n} z_j$ is the maximum of n independent standard Gaussian variates and Z is another, independent, standard Gaussian.

From this single least favorable distribution, one can derive lower bounds of different flavors for various purposes. We illustrate three such examples, which will all find application later in the book. In two of the cases, we give the corresponding upper bounds as well.

First a bound that applies for the whole scale of ℓ_p error measures. It is phrased in terms of the probability of a large norm error rather than via an expected p-th power error—this is appropriate for the application to optimal recovery in Chapter 10.

Proposition 8.12 *Fix $\eta > 0$. There exist functions $\pi_\eta(n) \to 1$ as $n \to \infty$ such that for any $\tau_n \le \sqrt{(2-\eta)\log n}$ and all $p' > 0$*

$$\inf_{\hat{\theta}} \sup_{\Theta_n(\tau_n)} P_\theta\{\|\hat{\theta} - \theta\|_{p'} \ge \tau_n/2\} \ge \pi_\eta(n). \tag{8.50}$$

For the next two results the focus returns to squared error loss. The first of the two is an asymptotic evaluation of minimax risk in the style of Theorem 8.11, but appropriate to this single spike setting. Since $\Theta_n(\tau_n) \subset \Theta_{n,p}(\tau_n)$, it actually applies to all ℓ_p balls of small radius with $p < 2$. An extension of this result to the setting of a finite number of spikes is given as Theorem ??

Proposition 8.13 *Suppose $p < 2$ and fix $\eta > 0$. For $\tau_n \le \sqrt{2\log n}$ and squared error loss*

$$R_N(\Theta_n(\tau_n)) \sim R_N(\Theta_{n,p}(\tau_n)) \sim \tau_n^2.$$

The second MSE result is an example of a non-asymptotic bound, i.e. valid for all finite n. This prepares for further non-asymptotic bounds in Section 11.5. As might be expected, the non-asymptotic bounds are less sharp than their asymptotic cousins. In this setting, recall that for a single bounded normal mean in $[-\tau, \tau]$, Section 4.6 showed that the minimax risk satisfies

$$c_0(\tau^2 \wedge 1) \le \rho_N(\tau, 1) \le \tau^2 \wedge 1.$$

Proposition 8.14 *Suppose that $y \sim N_n(0, I)$. There exists $c_1 > 0$ such that for all $n \ge 2$,*

$$c_1[\tau^2 \wedge (1 + \log n)] \;\le\; R_N(\Theta_n(\tau)) \;\le\; (\log n)^{-1/2} + \tau^2 \wedge (1 + 2\log n).$$

Proof of Proposition 8.12. Since the spike prior π_n concentrates on $\Theta_n(\tau)$, we have $\sup_{\theta \in \Theta} P_\theta(A) \ge P_{\pi_n}(A)$, and hence from (8.48), the left side of (8.50) is bounded below by $P_\pi(\hat{I} \ne I)$. Now appeal to (8.49) and the hypothesis $\tau_n \le \sqrt{(2-\eta)\log n}$ to conclude that the minimax error probability is bounded below by

$$\pi_\eta(n) = P\{M_{n-1} - Z \ge \sqrt{(2-\eta)\log n}\}.$$

It is intuitively clear from (8.22) that $\pi_\eta(n) \to 1$ as $n \to \infty$ for fixed η. One possible formal argument, admittedly crude, goes as follows. Set $a = \sqrt{(2-\eta)\log n}$ and $a' = \sqrt{(2-\eta')\log n}$ for some $\eta' < \eta$. We have

$$P(M_{n-1} - Z \ge a) = P(M_{n-1} \ge a')P(Z \le a' - a).$$

For any $\eta' > 0$, we have $P(M_{n-1} \ge a') \to 1$, for example by (8.58) in the next section. A little algebra shows that $a' - a \ge \sqrt{2\gamma \log n}$ for some $\gamma(\eta, \eta') > 0$ and hence $P(Z \le a' - a) \to 1$ also.

Proof of Proposition 8.13. For the upper bound, simply consider the zero estimator $\hat{\theta}_0(y) \equiv 0$ whose MSE is just the squared bias $\sum \theta_i^2$. By a convexity argument, the maximum of $\sum_i \theta_i^2$ over $\sum |\theta_i|^p \le \tau^p$ equals τ^2 and is attained at (permutations of) a spike vector $\tau(1, 0, \ldots, 0)$. [A generalization of this argument appears in Chapter 10.3].

For the lower bound, we return to the spike prior π_n with $\theta = \tau_n e_I$. Compute the mean squared error Bayes risk by symmetry and then decomposing on the event $\{I = 1\}$,

$$\begin{aligned} B(\pi_n) &= n E_\pi E_\theta(\hat{\theta}_{\pi,1} - \theta_1)^2 \\ &= (n-1)E_\pi E_\theta[\hat{\theta}^2_{\pi,1} | I \ne 1] + E_{\tau_n e_1}(\hat{\theta}_{\pi,1} - \tau_n)^2. \end{aligned}$$

We use the earlier heuristic that the primary contribution to the Bayes risk comes from the error made by $\hat{\theta}_{\pi,1}$ when θ_1 in fact takes the rare value δ_n : this yields

$$B(\pi_n) \ge \tau_n^2 E_{\tau_n e_1}[p_{1n}(y) - 1]^2 \tag{8.51}$$

and our strategy will be to show that $p_{1n}(y) \to 0$ for all y and use the dominated convergence theorem to conclude that $B(\pi_n) \ge \tau_n^2(1 + o(1))$. Using (8.47), and noting under $P_{\tau_n e_1}$ that $y_1 = \tau_n + z_1$ and $y_j = z_j$ for $j \ge 2$, we arrive at

$$p_{1n}(y) = [1 + V_n W_{n-1}]^{-1},$$

where $W_{n-1} = (n-1)^{-1}\sum_2^n e^{\tau_n z_j - \tau_n^2/2}$ and $V_n = (n-1)e^{-\tau_n^2/2-\tau_n z_1}$.

Let $\lambda_n = \sqrt{2\log n}$ and assume initially that $\lambda_n - \tau_n \to \infty$. Then by Lemma 8.15 below, $W_n \xrightarrow{p} 1$ as $n \to \infty$. For V_n, observe that

$$\lambda_n^2 - \tau_n^2 - 2\tau_n z \geq (\lambda_n - \tau_n - z_+)(\lambda_n + \tau_n) \to \infty, \tag{8.52}$$

again because $\lambda_n - \tau_n \to \infty$. Consequently $V_n \to \infty$ for each fixed z_1 and so $p_{1n}(y) \to 0$, and $B(\pi_n) \geq \tau_n^2(1 + o(1))$.

If $\tau_n \leq \lambda_n$ but $\lambda_n - \tau_n \not\to \infty$, then we simply choose $\tau_n' \sim \tau_n$ for which $\lambda_n - \tau_n' \to \infty$ in the previous argument, to obtain $B(\pi_n) \geq \tau_n'^2(1 + o(1)) \sim \tau_n^2$, as required.

Proof of Proposition 8.14. For the upper bound, consider the maximum risk of soft thresholding at $\lambda_n = \sqrt{2\log n}$. Bound (8.12) says that

$$\sup_{\Theta_n(\tau)} r(\hat\theta_{\lambda_n}, \theta) \leq (n-1)r_S(\lambda_n, 0) + r_S(\lambda_n, \tau) \leq n r_S(\lambda_n, 0) + \tau^2 \wedge (1 + \lambda_n^2).$$

The upper bound now follows from (8.25).

For the lower bound, this time we seek a bound for $B(\pi_n)$ valid for all n. Introduce $\ell_n = \sqrt{1 + \log n}$ and $\tau_n = \tau \wedge \ell_n$. We start from (8.51) and note that on the event $E_n = \{y_1 \neq \max_j y_j\}$ we have $p_{1n}(y) \leq 1/2$ and so $B(\pi_n) \geq (\tau_n^2/4)P_{\tau_n e_1}(E_n)$. From (8.49),

$$P(E_n) \geq P\{Z < 0, M_{n-1} > \tau_n\} = \tfrac{1}{2}P\{M_{n-1} \geq \tau_n\} \geq \tfrac{1}{2}P\{M_{n-1} \geq \ell_n\}$$

We leave it as Exercise 8.8 to verify that $P(M_{n-1} > \ell_n) \geq c_0$ for $n \geq 2$.

The remaining lemma carries a small surprise.

Lemma 8.15 *Let $z_1, \ldots, z_n \overset{i.i.d}{\sim} N(0,1)$ and $\lambda_n = \sqrt{2\log n}$. If $\lambda_n - \tau_n \to \infty$, then*

$$W_n = n^{-1}\sum_1^n e^{\tau_n z_k - \tau_n^2/2} \xrightarrow{p} 1.$$

Proof Since $Ee^{\tau_n z_k} = e^{\tau_n^2/2}$, we have $EW_n = 1$, but the variance can be large: $\mathrm{Var}\, W_n = n^{-1}(e^{\tau_n^2} - 1) \leq e^{\tau_n^2 - \log n}$. When $\mathrm{Var}\, W_n \to 0$, we have $W_n \xrightarrow{p} 1$ by the usual Chebychev inequality argument, but this fails for τ_n near $\sqrt{2\log n}$. Instead, we pick $b_0 \in (1/2, 1)$ and for $\tau_n < \sqrt{b_0 \log n}$ use the simple Chebychev argument. However, for $\sqrt{b_0 \log n} \leq \tau_n \leq \lambda_n$ such that $\lambda_n - \tau_n \to \infty$, we turn to the triangular array form of the weak law of large numbers, recalled in Proposition C.10. Put $b_n = e^{\tau_n \lambda_n}$ and $X_{nk} = e^{\tau_n z_k}$ and consider the truncated variables $\bar X_{nk} = X_{nk} I\{|X_{nk}| \leq b_n\} = e^{\tau_n z_k} I\{z_k \leq \lambda_n\}$. A short calculation shows that for any r,

$$E\bar X_{nk}^r = Ee^{r\tau_n z} I\{z \leq \lambda_n\} = e^{r^2\tau_n^2/2}\Phi(\lambda_n - r\tau_n).$$

We verify the truncation assumptions (i) and (ii) of Proposition C.10. Indeed, $\sum_1^n P(X_{nk} > b_n) = n\tilde\Phi(\lambda_n) \leq 1/\lambda_n \to 0$, and from the prior display

$$b_n^{-2}\sum E\bar X_{nk}^2 = c_0\tilde\Phi(2\tau_n - \lambda_n)/\phi(2\tau_n - \lambda_n) \leq c_0/(2\tau_n - \lambda_n),$$

where $c_0 = \phi(0)$. If $\tau_n \geq \sqrt{b_0 \log n}$ with $b_0 > 1/2$, then $2\tau_n - \lambda_n \to \infty$ and condition (ii) holds. Now set $a_n = \sum_1^n E\bar{X}_{nk} = ne^{\tau_n^2/2}\Phi(\lambda_n - \tau_n)$. Proposition C.10 says that $\sum_1^n e^{\tau_n z_k} = a_n + o_p(b_n)$, or equivalently that

$$W_n = \Phi(\lambda_n - \tau_n) + o_p(b_n n^{-1} e^{-\tau_n^2/2}).$$

Now $b_n n^{-1} e^{-\tau_n^2/2} = \exp\{-(\lambda_n - \tau_n)^2/2\}$ and hence $W_n \xrightarrow{p} 1$ if $\lambda_n - \tau_n \to \infty$. □

Remark. The variable $\sum_{k=1}^n e^{\tau_n z_k}$ is the basic quantity studied in the *random energy model* of statistical physics, e.g. Mézard and Montanari (2009, Ch. 5), where it serves as a toy model for spin glasses. In the current notation, it exhibits a phase transition at $\tau_n = \lambda_n = \sqrt{2 \log n}$, with qualitatively different behavior in the "high temperature" ($\tau_n < \lambda_n$) and "low temperature" ($\tau_n > \lambda_n$) regimes.

8.10 The distribution of $M_n = \max Z_i$

Simple bounds follow from the concentration inequalities (2.56) and (2.57). Since $z \to \max z_i$ is a Lipschitz(1) function, we have for $t > 0$

$$\begin{aligned} P\{|M_n - \text{Med}M_n| \geq t\} &\leq e^{-t^2/2} \\ P\{|M_n - EM_n| \geq t\} &\leq 2e^{-t^2/2}. \end{aligned} \tag{8.53}$$

Both MedM_n and EM_n are close to $L_n = \sqrt{2 \log n}$. Indeed

$$|EM_n - \text{Med}M_n| \leq \sqrt{2 \log 2}, \tag{8.54}$$

$$L_n - 1 \leq \text{Med}M_n \leq L_n, \tag{8.55}$$

$$L_n - 1 - \sqrt{2 \log 2} \leq EM_n \leq L_n. \tag{8.56}$$

The bound (8.54) is Exercise 2.12. The right bound of (8.55) follows from (8.57) below, and for the left bound see Exercise 8.9. The right hand bound of (8.56) is Proposition C.9 and the left bound then follows from (8.54) and (8.55). Of course, asymptotic expressions for MedM_n and EM_n follow from the extreme value limit theorem (8.22).

In fact M_n is confined largely to a shrinking interval L_n of width $2 \log^2 L_n / L_n$. Indeed, arguing analogously to (8.21), we have for $n \geq 2$,

$$P\{M_n \geq L_n\} \leq 1/(\sqrt{2\pi} L_n). \tag{8.57}$$

while Exercise 8.9 shows that for $L_n \geq 3$,

$$P\{M_n \leq L_n - 2L_n^{-1} \log^2 L_n\} \leq \exp\{-\tfrac{1}{3} \exp(\log^2 L_n)\}. \tag{8.58}$$

Numerics. Finding the quantiles of M_n, defined by $P\{M_n \leq x_\alpha\} = \alpha$, is easily done, and yields the central columns in the table below. We abbreviate the lower bound by $\tilde{L}_n = L_n - 2L_n^{-1} \log^2 L_n$

n	$\tilde{L}_n$	$x_{.10}$	$x_{.50}$	$x_{.90}$	L_n
32	1.16	1.48	2.02	2.71	2.63
128	1.66	2.10	2.55	3.15	3.11
1024	2.31	2.84	3.20	3.71	3.72
4096	2.70	3.26	3.58	4.05	4.08

8.11 Appendix: Further details

2°. The mean squared error of a thresholding rule $\hat{\delta}(x,\lambda)$ (either hard or soft) is found by breaking the range of integration into regions $(-\infty,-\lambda),[-\lambda,\lambda]$ and (λ,∞) to match the thresholding structure. For example, with soft thresholding

$$\begin{aligned} r(\lambda,\mu) &= E_\mu[\hat{\delta}(x,\lambda)-\mu]^2 \qquad (8.59)\\ &= \int_{-\infty}^{-\lambda}(x+\lambda-\mu)^2\phi(x-\mu)dx + \int_{-\lambda}^{\lambda}\mu^2\phi(x-\mu)dx + \int_{\lambda}^{\infty}(x-\lambda-\mu)^2\phi(x-\mu)dx \end{aligned}$$

One obtains the following basic mean squared error formulas:

$$\begin{aligned} r_S(\lambda,\mu) &= 1+\lambda^2+(\mu^2-\lambda^2-1)[\Phi(\lambda-\mu)-\Phi(-\lambda-\mu)] \qquad (8.60)\\ &\quad -(\lambda-\mu)\phi(\lambda+\mu)-(\lambda+\mu)\phi(\lambda-\mu),\\ r_H(\lambda,\mu) &= \mu^2[\Phi(\lambda-\mu)-\Phi(-\lambda-\mu)]+\tilde{\Phi}(\lambda-\mu)+\tilde{\Phi}(\lambda+\mu) \qquad (8.61)\\ &\quad +(\lambda-\mu)\phi(\lambda-\mu)+(\lambda+\mu)\phi(\lambda+\mu) \end{aligned}$$

where ϕ and Φ denote the standard Gaussian density and cumulative distribution functions respectively, and $\tilde{\Phi}(x)=1-\Phi(x)$.

3°. *Proof of (8.21).* We have that $\pi_n = 1-(1-\delta)^n \le n\delta$, with

$$\delta = 2\tilde{\Phi}(\sqrt{2\log n}) \le \frac{2\phi(\sqrt{2\log n})}{\sqrt{2\log n}} = \frac{1}{n\sqrt{\pi\log n}}.$$

4°. *Proof of lower bound in Lemma 8.3 for* $0\le\lambda\le 2$. Let μ_λ be the solution in μ of $r(\lambda,0)+\mu^2 = 1+\lambda^2$. Since $r(\lambda,0)\le e^{-\lambda^2/2}<1$, (compare (8.7)), it is clear that $\mu_\lambda>\lambda$. For $\mu\le\mu_\lambda$ we may write, using (8.5),

$$R(\lambda,\mu)=\frac{r(\lambda,\mu)}{\bar{r}(\lambda,\mu)}=\frac{r(\lambda,0)+\int_0^\mu 2s\Phi(I_\lambda-s)ds}{r(\lambda,0)+\mu^2}.$$

We first verify that $R(\lambda,\mu)$ is decreasing in $\mu\le\mu_\lambda$. Indeed $\mu\to[c+f_1(\mu)]/[c+f_2(\mu)]$ is decreasing if both $f_1'(\mu)\le f_2'(\mu)$ and $(f_1/f_2)(\mu)$ is decreasing. The former condition is evident, while the latter follows by the rescaling $v=s/\mu$: for then $(f_1/f_2)(\mu)=2\int_0^1\Phi(I_\lambda-\mu v)dv$.

For $\mu\ge\mu_\lambda$, we also have $R(\lambda,\mu)\ge R(\lambda,\mu_\lambda)$ since $r(\lambda,\mu)\ge r(\lambda,\mu_\lambda)$ while $\bar{r}(\lambda,\mu)\equiv 1+\lambda^2$. Consequently, for all μ

$$R(\lambda,\mu)\ge r(\lambda,\mu_\lambda)/[r(\lambda,0)+\mu_\lambda^2],$$

and numerical evaluation for $0\le\lambda\le 2$ shows the right side to be bounded below by .516, with the minimum occurring for $\lambda\in[.73,.74]$.

5°. *Proof of second half of* (8.33). Combining (8.27) and (8.28), we have $m(x)=e^{\mu(\mu+a-x)}$. Using formula (8.29) for δ_π, then changing variables to $z=x-\mu-a$ and finally exploiting (8.28), we find that

$$(1-\alpha)E_0\delta_\pi^2=(1-\alpha)\mu^2\int\frac{\phi(x)dx}{[1+e^{\mu(\mu+a-x)}]^2}=\mu^2\alpha\phi(a)\int_{-\infty}^{\infty}\frac{e^{-(\mu+a)z-z^2/2}dz}{[1+e^{-\mu z}]^2}.$$

We now verify that

$$\frac{(1-\alpha)E\delta_\pi^2(z)}{\alpha\mu^2(\alpha)} \le \phi(a)\int_0^\infty e^{-(\mu+a)z-z^2/2}dz + \int_{-\infty}^\infty \frac{\phi(w)dw}{1+e^{\mu(w+a)}}. \tag{8.62}$$

Consider the integral first over $(0,\infty)$: we may replace the denominator by 1 to obtain the first term in (8.62). Over $(-\infty,0)$, we have $e^{-\mu z}/[1+e^{-\mu z}] \le 1$, and with $v=-z$ this part of the integral is bounded by

$$\mu^2\alpha\int_0^\infty \frac{\phi(v-a)dv}{1+e^{\mu v}},$$

which with $w=v-a$ leads to the second term in (8.62). By dominated convergence, both right hand side terms converge to zero as μ and $a\to\infty$.

6°. Proof of (8.35). This bears some resemblance to the technique used to complete the proof of Pinsker's theorem. We define an event $A_n=\{N_n \le n\alpha_n+\eta_n\}$ on which π_n concentrates, with $\eta_n=(\log n)^{2/3}$ say, then show desired behavior on A_n, and control what happens on A_n^c.

Let $\tilde\delta_n$ denote the Bayes rule for π_n with respect to loss $\tilde L_n$. Using $\mathbb{E}$ to denote expectation under the joint distribution of $\theta\sim\pi_n$ and x, we have

$$\begin{aligned}\tilde B(\pi_n) &= \mathbb{E}[L_n(\tilde\delta_n,\theta)/(1+N_n)] \ge \mathbb{E}[L_n(\tilde\delta_n,\theta),A_n]/[1+n\alpha_n+\eta_n]\\ &= (1+o(1))\mathbb{E}L_n(\tilde\delta_n,\theta)/(1+n\alpha_n) \qquad (\star)\\ &\ge (1+o(1))B(\pi_n)/(1+\log n)\sim 2\log n,\end{aligned}$$

where to justify the critical step $(\star)$, we must verify that

$$\mathbb{E}\|\tilde\delta-\theta\|^2, A_n^c] = o\{B(\pi_n)\} = o(\mu_n^2\log n).$$

We focus only on the trickier term $\mathbb{E}[\|\tilde\delta\|^2, A_n^c]$. Set $p(\theta)=1+N_n(\theta)$. Using by turns the conditional expectation representation for

$$\tilde\delta_{n,i}(x) = \frac{\mathbb{E}[\theta_i/p(\theta)|x]}{\mathbb{E}[1/p(\theta)|x]},$$

the Cauchy-Schwartz and Jensen inequalities, we find

$$\begin{aligned}\|\tilde\delta_n\|^2 &\le \frac{\mathbb{E}[\|\theta\|^2/p(\theta)|x]}{\mathbb{E}[1/p(\theta)|x]} \le \mathbb{E}[p(\theta)|x]\,\mathbb{E}[\|\theta\|^2/p(\theta)|x] \qquad \text{and}\\ \mathbb{E}\{\|\tilde\delta_n\|^2, A_n^c\} &\le \{\mathbb{E}p^4(\theta)\,\mathbb{P}^2(A_n^c)\,\mathbb{E}\|\theta\|^8/p^4(\theta)\}^{1/4}\\ &\le C\mu_n^2\,\mathbb{P}^{1/2}(A_n^c)\,\log n = o(\mu_n^2\log n),\end{aligned}$$

since $\|\theta\|^8 = N_n\mu_n^8$ and $\mathbb{E}N_n^p = O(\log^p n)$.

8.12 Notes

Problem. Find bounds sharper than (8.19) for the smallest maximum risk attainable by an estimator having threshold zone $[-\lambda,\lambda]$. Since such estimators form a convex set, this should be accessible via the minimax theorem and the formulas for posterior mean given in Chapter 4. The problem is a little reminiscent of bounding minimax risk subject to specified risk properties at a point (compare Bickel (1983)).

Exercises

8.1 *(Mill's ratio and Gaussian tails.)* The function $R(\lambda)=\tilde\Phi(\lambda)/\phi(\lambda)$ is sometimes called Mill's ratio. Show that the modified form

$$M(\lambda) = \frac{\lambda\tilde\Phi(\lambda)}{\phi(\lambda)} = \int_0^\infty e^{-v-v^2/(2\lambda^2)}dv,$$

and hence that $M(\lambda)$ is increasing from 0 at $\lambda = 0$ up to 1 at $\lambda = \infty$.
Define the l-th approximation to the Gaussian tail integral by

$$\tilde{\Phi}_l(\lambda) = \lambda^{-1}\phi(\lambda) \sum_{k=0}^{l} \frac{(-1)^k}{k!} \frac{\Gamma(2k+1)}{2^k \lambda^{2k}}.$$

Show that for each $k \geq 0$ and all $\lambda > 0$ that

$$\tilde{\Phi}_{2k+1}(\lambda) \leq \tilde{\Phi}(\lambda) \leq \tilde{\Phi}_{2k}(\lambda).$$

[Hint: induction shows that $(-1)^{l-1}[e^{-x} - \sum_0^l (-1)^k x^k/k! \geq 0$ for $x \geq 0$.]
As consequences, we obtain, for example, the bounds

$$\lambda^{-1}\phi(\lambda)(1-\lambda^{-2}) \leq \tilde{\Phi}(\lambda) \leq \lambda^{-1}\phi(\lambda), \tag{8.63}$$

and the expansion, for large λ,

$$\tilde{\Phi}(\lambda) \sim \lambda^{-1}\phi(\lambda)[1 - \lambda^{-2} + 3\lambda^{-4} - 15\lambda^{-6} + O(\lambda^{-8})]. \tag{8.64}$$

8.2 *(alternate hard threshold bound.)* Show how the proof of Proposition 8.1 can be modified so as to show that for all $\lambda > 0$,

$$r_H(\lambda,\theta) \leq \begin{cases} 2[\theta^2 + 2(\lambda+15)\phi(\lambda-1)\epsilon^2] & \text{if } |\theta| \leq \epsilon \\ 2(\lambda^2+1)\epsilon^2 & \text{if } |\theta| > \epsilon. \end{cases}$$

8.3 *(Risk of soft thresholding at 0.)* Let $z \sim N(0,1)$, and $r(\lambda,0) = E\hat{\delta}_S^2(z)$ denote the mean squared error of soft thresholding at $\lambda = 0$.
(a) Use (8.63) and (8.64) to show that

$$\begin{aligned} r(\lambda,0) &\leq 4\lambda^{-3}(1+1.5\lambda^{-2})\phi(\lambda) && \lambda > 0, \\ r(\lambda,0) &\sim 4\lambda^{-3}\phi(\lambda), && \lambda \to \infty \end{aligned}$$

(b) Conclude that $r(\lambda,0) \leq 4\lambda^{-1}\phi(\lambda)$ if, say, $\lambda \geq \sqrt{2}$.
(c) Let $\delta(\lambda) = e^{-\lambda^2/2} - r(\lambda,0)$. Use (8.63) to show that $\delta(\lambda) > 0$ for $\lambda \geq \lambda_0 = 2\phi(0)$.
(d) Show that $\delta(\lambda)$ is concave for $\lambda \in [0,1]$, and conclude that $r(\lambda,0) \leq e^{-\lambda^2/2}$ for all $\lambda \geq 0$.

8.4 Derive the following inequalities for hard thresholding, which are sharper than direct application of the bounds in (8.16):

$$\begin{aligned} r_H(\lambda,\lambda) &\geq (\lambda^2+1)/2, \\ r_H(\lambda,0) &\geq (2\lambda \vee \sqrt{2\pi})\phi(\lambda), \\ r_H(\lambda,0) &\leq (2\lambda + \sqrt{2\pi})\phi(\lambda) \\ r_H(\lambda,0) &\leq 2(\lambda + 1/\lambda)\phi(\lambda). \end{aligned}$$

(Birgé and Massart, 2001)

8.5 *(risk behavior near threshold.)* In the notation of Section 8.2, show that
(i) for soft thresholding, as $\lambda \to \infty$,

$$r_S(\lambda,\lambda) = \lambda^2 - \sqrt{2/\pi}\lambda + 1/2 + \tilde{\Phi}(2\lambda) \sim \lambda^2.$$

(ii) for hard thresholding, as $\lambda \to \infty$,

$$\begin{aligned} r_H(\lambda, \lambda - 2\sqrt{\log\lambda}) &= (\lambda - 2\sqrt{\log\lambda})^2 + O((\log\lambda)^{-1/2}), \\ r_H(\lambda, \lambda) &\sim \lambda^2/2, \\ r_H(\lambda, \lambda + 2\sqrt{\log\lambda}) &\le 1 + (2\pi\log\lambda)^{-1/2}. \end{aligned}$$

8.6 *(Number of exceedances of universal threshold.)* Let $N_n = \sum_1^n I\{|Z_i| \ge \sqrt{2\log n}\}$.
(a) If Z_i are i.i.d. $N(0,1)$, show that $N_n \sim \mathrm{Bin}(n, p_n)$ with $p_n = 2\tilde{\Phi}(\sqrt{2\log n})$.
(b) Show that $P(N_n \ge 2) \le (np_n)^2 \le 1/(\pi\log n)$.

8.7 *(Maximum of absolute values of Gaussian noise mimicks M_{2n}.)* Let $h_i, i = 1, \ldots$ be independent half-normal variates (i.e. $h_i = |Z_i|$ for $Z_i \sim N(0,1)$), and ϵ_i be independent ± 1 variates, independent of $\{h_i\}$. Let $Z_i = h_i\epsilon_i$ and T_n be the random time at which the number of positive ϵ_i reaches n. Show that the Z_i are independent standard normal and that

$$\max_{i=1,\ldots,n} |Z_i| \overset{\mathcal{D}}{=} \max_{i=1,\ldots,n} h_i = \max_{i=1,\ldots,T_n} Z_i = M_{T_n},$$

and that T_n is close to $2n$ in the sense that

$$(T_n - 2n)/\sqrt{2n} \Rightarrow N(0,1).$$

8.8 *(Lower bound for maximum of Gaussians.)* Let $z_i \overset{i.i.d.}{\sim} N(0,1)$ and $M_n = \max z_i$. Let $\ell_n = \sqrt{1 + \log n}$. Show that for some $c_1 > 0$, for all $n \ge 2$,

$$P(M_n \ge \ell_n) \ge c_1.$$

Hint. Use (8.63) and $(1-x)^m \le e^{-mx}$.

8.9 *(Left tail bound for M_n.)* Let $L_n = \sqrt{2\log n}$, and as above $M_n = \max z_i$.
(a) Show that $P\{M_n \le \lambda\} \le \exp\{-n\tilde{\Phi}(\lambda)\}$.
(b) Establish the left side of bound (8.55) by using (8.63) for $\lambda \ge 1/2$ to show that for $n \ge 4$,

$$n\tilde{\Phi}(L_n - 1) \ge (e/2)\sqrt{e/(2\pi)},$$

(c) Again use (8.63) to show that

$$P\{M_n \le L_n - 2L_n^{-1}\log^2(L_n)\} \le \exp\{-H(L_n)\exp(\log^2 L_n)\}$$

where $H(x) = \phi(0)(\lambda^{-1} - \lambda^{-3})\exp\left(\log^2 x - 2x^{-2}\log^4 x\right)$ and $\lambda = x - 2x^{-1}\log^2 x$. Verify numerically that $H(x) \ge 1/3$ for $x \ge 3$ and hence conclude (8.58).

8.10 *(Properties of Miller's selection scheme.)* Refer to Alan Miller's variable selection scheme, and assume as there that the columns are centered and scaled: $\langle x_i, 1\rangle = 0$ and $\langle x_i, x_i\rangle = 1$. Show that the permuted columns are approximately orthogonal to each other and to the original columns. More precisely, show that
(i) if $j \ne k$, then $\langle x_j^*, x_k^*\rangle$ has mean 0 and standard deviation $1/\sqrt{N-1}$, and
(ii) for *any* pair (j,k), similarly $\langle x_j^*, x_k\rangle$ has mean 0 and standard deviation $1/\sqrt{N-1}$.

8.11 *(Miller's selection scheme requires many components.)* Suppose that $x_1 = (1,-1,0)^T/\sqrt{2}$ and $x_2 = (0,-1,1)^T/\sqrt{2}$. Consider the random permutations x_1^* and x_2^* described in A. Miller's selection method. Compute the distribution of $\langle x_1^*, x_2^*\rangle$ and show in particular that it equals 0 with zero probability.

8.12 *(Lower bound in Theorem 8.11, sparse case)* Adopt the setting of Section 8.6. Suppose that $\eta_n \to 0$ and that $k_n \to 0$. Let $\gamma < 1$ be given, and build π_n from n i.i.d draws (scaled by ϵ_n) from the univariate sparse prior $\pi_{\gamma\eta_n}$ with sparsity $\gamma\eta_n$ and overshoot $(2\log(\gamma\eta_n)^{-1})^{1/4}$, compare Section 8.4. Show that

1. The number N_n of non-zero components in a draw from π_n is distributed as Binomial$(n, \gamma\eta_n)$, and hence that $\pi_n(\Theta_n) \to 1$ if and only if $k_n \to \infty$,
2. on Θ_n, we have $\|\theta\|^2 \leq \gamma^{-1}\epsilon_n^2\mu_n^2 EN_n$ `(define` μ_n`)`, and
3. for all y, show that $\|\hat{\theta}_{\nu_n}\|^2 \leq \gamma^{-1}\epsilon_n^2\mu_n^2 EN_n$.

As a result, verify that the sequence π_n satisfies conditions (8.41) – (8.44), and hence that $R_N(\Theta_n(k_n), \epsilon_n) \geq B_n(k_n, \epsilon_n)(1 + o(1))$.

9

Sparsity, adaptivity and wavelet thresholding

In this chapter, we explore various measures for quantifying sparsity and the connections among them. In the process, we will see hints of the links that these measures suggest with approximation theory and compression. We then draw consequences for adaptive minimax estimation, first in the single sequence model, and then in multiresolution settings.

In Section 9.1, traditional linear approximation is contrasted with a version of non-linear approximation that greedily picks off the largest coefficients in turn. Then a more explicitly statistical point of view relates the size of *ideal risk* to the non-linear approximation error. Thirdly, we look at the decay of individual ordered coefficients: this is expressed in terms of a weak-ℓ_p condition. The intuitively natural connections between these viewpoints can be formalized as an equivalence of (quasi-)norms in Section 9.2.

Consequences for estimation now flow quite directly. Section 9.3 gives a lower bound for minimax risk using hypercubes, and the oracle inequalities of the last chapter in terms of ideal risk combined with the quasi-norm equivalences lead to upper bounds for $\sqrt{2\log n}$ thresholding over weak ℓ_p balls that are only a logarithmic factor worse than the hypercube lower bounds. When $p < 2$, these are polynomially better rates than can be achieved by any linear estimator–this is seen in Section 9.5 using some geometric ideas from Section 4.8.

To interpret and extend these results in the setting of function estimation we need to relate sparsity ideas to smoothness classes of functions.

The fundamental idea may be expressed as follows. A function with a small number of isolated discontinuities, or more generally singularities, is nevertheless smooth "on average." If non-parametric estimation is being assessed via a global norm, then one should expect the rate of convergence of good estimators to reflect the average rather than worst case smoothness.

Thus, a key idea is the degree of uniformity of smoothness that is assumed, and this is measured in an L_p sense. Section 9.6 introduces this topic in more detail by comparing three examples, namely uniform ($p = \infty$), mean-squre ($p = 2$) and average ($p = 1$) smoothness conditions, and then working up to the definition of Besov classes as a systematic framework covering all the cases.

Focusing on the unit interval $[0, 1]$, it turns out that many Besov classes of smoothness α are contained in weak $\ell_{p(\alpha)}$ balls, Section 9.7. After some definitions for estimation in the continuous Gaussian white noise problem in Section 9.8, the way is paved for earlier results in this chapter to yield, in Section 9.9, broad adaptive near-minimaxity results for $\sqrt{2\log n}$ thresholding over Besov classes.

These results are for integrated mean squared error over all $t \in [0,]$; Section 9.10 shows

that the same estimator, and similar proof ideas, lead to rate of convergence results for estimating $f(t_0)$ at a single point t_0.

The final Section 9.11 gives an overview of the topics to be addressed in the second part of the book.

9.1 Approximation, Ideal Risk and Weak ℓ_p Balls

Non-linear approximation

Let $\{\psi_i, i \in \mathbb{N}\}$ be an orthonormal basis for $L_2[0,1]$, and consider approximating $f \in L_2[0,1]$ by a linear combination of basis functions from a subset $K \subset \mathbb{N}$:

$$P_K f = \sum_{i \in K} \theta_i \psi_i .$$

The coefficients $\theta_i = \langle f, \psi_i \rangle$, and we will not distinguish between f and the corresponding coefficient sequence $\theta = \theta[f]$. Again, using the orthonormal basis property, we have

$$\| f - P_K f \|_2^2 = \sum_{i \notin K} \theta_i^2 .$$

The operator P_K is simply orthogonal projection onto the subspace spanned by $\{\psi_i, i \in K\}$, and yields the best L_2 approximation of f from this subspace. In particular, P_K is linear, and we speak of best linear approximation.

Now consider the best *choice* of a subset K of size k: we have

$$c_k^2(f) = \inf \left\{ \| f - P_K f \|_2^2 : \#(K) \le k \right\},$$

or what is the same

$$c_k^2(\theta) = \inf \left\{ \sum_{i \notin K} \theta_i^2 : \#(K) \le k \right\}. \tag{9.1}$$

Let $|\theta|_{(1)} \ge |\theta|_{(2)} \ge \ldots$ denote the amplitudes of θ in decreasing order. Then $c_k^2(f)$ is what remains after choosing the k largest coefficients, and so

$$c_k^2(f) = c_k^2(\theta) = \sum_{l > k} |\theta|_{(l)}^2,$$

and we call $c_k(\theta)$ the *compression numbers* associated with $\theta = \theta[f]$.

Let $K_k(\theta)$ be the set of indices corresponding to the k largest magnitudes. Since $K_k(f)$ depends strongly on f, the best approximation operator $Q_k f = P_{K_k(\theta)} f$ is *non-linear*: $Q_k(f+g) \ne Q_k f + Q_k g$.

Thus the rate of decay of $c_k(\theta)$ with k measures the rate of non-linear approximation of f using the best choice of k functions from the basis. To quantify this, define a sequence (quasi)norm

$$|\theta|_{c,\alpha}^2 = \sup_{k \ge 0} k^{2\alpha} \sum_{l > k} |\theta|_{(l)}^2,$$

with the convention that $k^{2\alpha} = 1$ when $k = 0$. In other words, $|\theta|_{c,\alpha} = C$ means that $(\sum_{l>k} |\theta|_{(i)}^2)^{1/2} \le C k^{-\alpha}$ for all k and that C is the smallest constant with this property.

So far, the index set has been $\mathbb{N}$. The expression (9.1) for $c_k^2(\theta)$ is well defined for any finite or countable index set I, and hence so is $|\theta|_{c,\alpha}^2$, if the supremum is taken over $k = 0, 1, \ldots, |I|$.

Ideal Risk

Return to estimation in a Gaussian white sequence model

$$y_i = \theta_i + \epsilon z_i, \qquad i \in I,$$

thought of, as usual, as the coefficients of the continuous Gaussian white noise model (1.18) in the orthonormal basis $\{\psi_i\}$.

Suppose that $K \subset I$ indexes a finite subset of the variables and that P_K is the corresponding orthogonal projection. The variance-bias decomposition of MSE is given by

$$E_\epsilon \| P_K y - f \|^2 = \#(K)\epsilon^2 + \| P_K f - f \|^2.$$

The 'ideal' subset minimizes the MSE for estimating f,

$$\begin{aligned} \mathcal{R}(f, \epsilon) &:= \inf_{K \subset \mathbb{N}} E_\epsilon \| P_K y - f \|^2 && (9.2) \\ &= \inf_k \Big\{ k\epsilon^2 + \inf_{K:\#(K)=k} \| P_K f - f \|^2 \Big\} && (9.3) \\ &= \inf_k \big\{ k\epsilon^2 + c_k^2(\theta) \big\}. && (9.4) \end{aligned}$$

The second and third forms show the connection between ideal estimation and non-linear approximation, and hint at the manner in which approximation theoretic results have a direct implication for statistical estimation.

Write $S_k = k\epsilon^2 + c_k^2(\theta)$ for the best MSE for model size k. The differences

$$S_k - S_{k-1} = \epsilon^2 - |\theta|_{(k)}^2$$

are increasing with k, and so the minimum value of $k \to S_k$ occurs as $k \to |\theta|_{(k)}^2$ 'crosses' the level ϵ^2, or more precisely, at the index k given by

$$N(\epsilon) = N(\theta, \epsilon) = \#\{i : |\theta_i| \geq \epsilon\}, \tag{9.5}$$

Compare Figure [ADD IN] . [in approximation theory, this is called the distribution function of $|\theta|$, a usage related to, but not identical with the standard statistical term.]

It is thus apparent that, in an orthonormal basis, the ideal subset estimation risk coincides with our earlier notion of ideal risk (Section 8.3):

$$\mathcal{R}(f, \epsilon) = \mathcal{R}(\theta, \epsilon) = \sum \theta_i^2 \wedge \epsilon^2$$

The ideal risk measures the intrinsic difficulty of estimation in the basis $\{\psi_i\}$. Of course, it is attainable only with the aid of an oracle who knows $\{i : |\theta_i| \geq \epsilon\}$.

In addition, (9.4) and (9.5) yield the decomposition

$$\mathcal{R}(\theta, \epsilon) = N(\epsilon, \theta)\epsilon^2 + c_{N(\epsilon)}^2(\theta). \tag{9.6}$$

Thus, the ideal risk is small precisely when both $N(\epsilon)$ and $c_{N(\epsilon)}$ are. This has the following interpretation: suppose that $N(\epsilon, \theta) = k$ and let $K_k(\theta)$ be the best approximating set of

size k. Then the ideal risk consists of a variance term $k\epsilon^2$ corresponding to estimation of the k coefficients in $K_k(\theta)$ and a bias term $c_k^2(\theta)$ which comes from not estimating all other coefficients. Because the oracle specifies $K_k(\theta) = \{i : |\theta_i| > \epsilon\}$, the bias term is as small as it can be for any projection estimator estimating only k coefficients.

The rate of decay of $\mathcal{R}(\theta, \epsilon)$ with ϵ measures the rate of estimation of θ (or $f[\theta]$) using the ideal projection estimator for the given basis. Again to quantify this, we define a second sequence norm

$$|\theta|_{IR,r}^2 = \sup_{\epsilon>0} \epsilon^{-2r} \sum_i \theta_i^2 \wedge \epsilon^2.$$

In other words, $|\theta|_{IR,r} = B$ means that $\mathcal{R}(\theta, \epsilon) \le B^2\epsilon^{2r}$ for all $\epsilon > 0$, and that B is the smallest constant for which this is true.

Identity (9.6) says that good estimation is possible precisely when θ compresses well in basis $\{\psi_i\}$, in the sense that both the number of large coefficients $N(\epsilon)$ and the compression number $c_{N(\epsilon)}^2$ are small. Proposition 9.1 below uses (9.6) to show that the compression number and ideal risk sequence quasinorms are equivalent.

Weak ℓ_p and Coefficient decay

A further natural measure of the "compressibility" of θ is the rate at which the *individual* magnitudes $|\theta_i|$ decay. More formally, we say that $\theta = (\theta_i, i \in I) \in w\ell_p$, if the *decreasing rearrangement* $|\theta|_{(1)} \ge |\theta|_{(2)} \ge \dots$ satisfies, for some C,

$$|\theta|_{(l)} \le C l^{-1/p}, \qquad l = 1, \dots, |I|,$$

and we set $\|\theta\|_{w\ell_p}$ equal to the smallest such C. Thus

$$\|\theta\|_{w\ell_p} = \max_k k^{1/p} |\theta|_{(k)}.$$

Here $\|\theta\|_{w\ell_p}$ is a *quasi-norm* rather than a norm, since instead of the triangle inequality, it satisfies only

$$\|\theta + \theta'\|_{w\ell_p}^p \le 2^p(\|\theta\|_{w\ell_p}^p + \|\theta'\|_{w\ell_p}^p), \qquad (p > 0). \tag{9.7}$$

See 3° below for the proof, and also Exercise 9.1. We write $w\ell_p(C)$ for the (quasi)norm ball of radius C, or $w\ell_{n,p}(C)$ if we wish to emphasize that $I = \{1, \dots, n\}$.

Smaller values of p correspond to faster decay for the components of θ. We will be especially interested in cases where $p < 1$, since these correspond to the greatest sparsity.

We note some relations satisfied by $w\ell_p(C)$.

1°. $\ell_p(C) \subset w\ell_p(C)$. This follows from

$$[k^{1/p}|\theta|_{(k)}]^p \le k \cdot (1/k) \sum_1^k |\theta|_{(l)}^p \le \|\theta\|_{\ell_p}^p.$$

2°. $w\ell_p \subset \ell_{p'}$ for all $p' > p$, since if $\theta \in w\ell_p$, then

$$\sum_1^\infty |\theta|_{(k)}^{p'} \le C^{p'} \sum_1^\infty k^{-p'/p} = C^p \zeta(p'/p).$$

3°. A plot of $N(\theta,\epsilon)$ versus ϵ shows that the maximum of $\epsilon \to \epsilon^p N(\theta,\epsilon)$ may be found among the values $\epsilon = |\theta|_{(k)}$. Hence we obtain

$$\|\theta\|_{w\ell_p}^p = \sup_{\epsilon>0} \epsilon^p N(\theta,\epsilon). \tag{9.8}$$

This representation makes it easy to establish the quasinorm property. Indeed, since

$$N(\theta+\theta',\epsilon) \le N(\theta,\epsilon/2) + N(\theta',\epsilon/2),$$

we obtain (9.7) immediately. Another immediate consequence of (9.8) is the implication

$$\epsilon^p N(\Theta,\epsilon) \le C^p \text{ for all } \epsilon \implies \Theta \subset w\ell_p(C). \tag{9.9}$$

9.2 Quasi-norm equivalences

In preceding subsections, we have defined three quantitative measures of the sparseness of a coefficient vector θ.

(a) $|\theta|_{c,\alpha}$ as a measure of the rate α of non-linear ℓ_2 approximation of θ using a given number of coefficients,
(b) $|\theta|_{IR,r}$ as a measure of the rate r of mean squared error decrease in ideal statistical estimation of θ in the presence of noise of scale ϵ, and
(c) $|\theta|_{w\ell_p}$ as a measure of the rate $1/p$ of decay of the individual coefficients $|\theta|_{(l)}$.

We now show that these measures are actually equivalent, if one makes the calibrations

$$r = 2\alpha/(2\alpha+1), \qquad p = 2/(2\alpha+1), \qquad \Longrightarrow \qquad p = 2(1-r). \tag{9.10}$$

Proposition 9.1 *Let $\alpha > 0$, and suppose that $r = r(\alpha)$ and $p = p(\alpha)$ are given by* (9.10). *Then, with $c_p = [2/(2-p)]^{1/p}$,*

$$3^{-1/p}|\theta|_{w\ell_p} \le |\theta|_{c,\alpha} \le |\theta|_{IR,r}^{2/p} \le c_p|\theta|_{w\ell_p}. \tag{9.11}$$

Proof We establish the inequalities proceeding from right to left in (9.11). Since all the measures depend only on the absolute values of (θ_i), by rearrangement we may suppose without loss of generality that θ is positive and decreasing, so that $\theta_k = |\theta|_{(k)}$.

1°. Suppose first that $C = |\theta|_{w\ell_p}$, so that $\theta_k \le Ck^{-1/p}$. Hence

$$\begin{aligned}\sum \theta_k^2 \wedge t^2 &\le \sum_1^\infty C^2k^{-2/p} \wedge t^2 \le \int_0^\infty (Cu^{-1/p})^2 \wedge t^2\, du \\ &= u_* t^2 + \frac{p}{2-p} C^2 u_*^{1-2/p} = \left(1+\frac{p}{2-p}\right) C^p t^{2r}.\end{aligned}$$

Here $u_* = C^p t^{-p}$ is the point of balance in the pairwise minimum. Hence $|\theta|_{IR}^2 = \sup_{t\ge 0} t^{-2r} \sum \theta_k^2 \wedge t^2 \le [2/(2-p)]|\theta|_{w\ell_p}^p$.

2°. Now let $C = |\theta|_{IR,r}$, so that for all positive t, $t^{-2r}\sum \theta_k^2 \wedge t^2 \le C^2$. In particular, when $t = \theta_k$, we obtain, for all $k \ge 1$,

$$\theta_k^{-2r}[k\theta_k^2 + c_k^2(\theta)] \le C^2.$$

Hence $\theta_k^p \le k^{-1}C^2$ and so

$$c_k^2(\theta) \le \theta_k^{2r} C^2 \le k^{-2r/p}(C^2)^{1+2r/p}.$$

Since $2r/p = 2\alpha$, we conclude for every $k \ge 1$, that $k^{2\alpha}c_k^2(\theta) \le C^{2(1+2\alpha)} = |\theta|_{IR}^{4/p}$. It remains to consider the exceptional case $k = 0$: putting $t = \theta_1$ in the definition of $|\theta|_{IR,r}$, we find $c_0^2(\theta) \le C^2\theta_1^{2r}$ and also that $\theta_1^2 \le C^2\theta_1^{2r}$. Hence $\theta_1^p \le C^2$ and so $c_0^2(\theta) \le C^{4/p}$, which completes the verification.

3°. Let $C = |\theta|_{c,\alpha}$, so that $c_k^2(\theta) \le C^2k^{-2\alpha}$ for $k \ge 1$ and $c_0^2(\theta) \le C^2$. This implies that $\theta_1^2 \le C^2$, and for $k \ge 2$ and $1 \le r < k$ that

$$\theta_k^2 \le (1/r)\sum_{k-r+1}^{k} \theta_j^2 \le C^2/r(k-r)^{2\alpha} \le C^2(3/k)^{1+2\alpha},$$

where for the last inequality we set $r = [k/2] \ge k/3$. Consequently, for all $k \ge 1$,

$$|\theta|_{w\ell_p}^2 = \sup_k k^{2/p}\theta_k^2 \le 3^{2/p}C^2. \qquad \square$$

9.3 A Risk Lower Bound via Embedding of hypercubes.

We have just seen that $N(\theta,\epsilon)$, the number of coefficients with modulus larger than ϵ, is a useful measure of sparsity. In combination with earlier minimax estimation results for hyperrectangles, it also leads to a simple, but important lower bound for minimax risk for solid, orthosymmetric Θ under squared error loss.

Suppose Θ is solid and orthosymmetric. For each $\theta \in \Theta$ and $\epsilon > 0$, the very definition shows that Θ contains a hypercube $\Theta(\epsilon)$ with center 0, side length 2ϵ and dimension $N(\theta,\epsilon)$. The *ϵ-hypercube dimension*

$$N(\Theta,\epsilon) := \sup_{\theta\in\Theta} N(\theta,\epsilon) \tag{9.12}$$

denote the maximal dimension of a zero-centered ϵ-hypercube embedded in Θ.

In the white Gaussian sequence model at noise level ϵ, the minimax risk for a p-dimensional ϵ-hypercube $[-\epsilon,\epsilon]^p$ is given, from (4.47) and (4.35) by

$$R_N([-\epsilon,\epsilon]^p,\epsilon) = p\epsilon^2\rho_N(1,1),$$

where $c_0 = \rho_N(1,1)$ is the minimax risk in the unit noise univariate bounded normal mean problem on $[-1,1]$. Since Θ contains the hypercube $\Theta(\epsilon)$, we arrive at a lower bound for the minimax risk

$$R_N(\Theta,\epsilon) \ge c_0\epsilon^2 N(\Theta,\epsilon). \tag{9.13}$$

Examples. 1. ℓ_p balls. In Chapters 11 and 13, we study at length estimation over

$$\ell_{n,p}(C) = \Theta_{n,p}(C) = \{\theta \in \mathbb{R} : \sum_1^n |\theta_i|^p \le C^p\}. \tag{9.14}$$

We clearly have $\#\{i : |\theta_i| \ge \epsilon\} \le \sum_1^n |\theta_i|^p/\epsilon^p$, and so the ϵ-hypercube dimension

$$N(\ell_{n,p}(C),\epsilon) = \min(n, [C^p/\epsilon^p]). \tag{9.15}$$

Hence, if $C > \epsilon$, we find from this and (9.13) that

$$R_N(\ell_{n,p}(C), \epsilon) \geq c_1 \min(n\epsilon^2, C^p \epsilon^{2-p}), \tag{9.16}$$

where $c_1 = c_0/2$. Since $\ell_{n,p}(C) \subset w\ell_{n,p(C)}$, the same lower bound applies also to the weak ℓ_p ball. In Section 11.5, we will see that for $p \geq 2$ this bound is sharp at the level of rates, while for $p < 2$ an extra log term is present.

2. Products. Since $N((\theta_1, \theta_2), \epsilon) = N(\theta_1, \epsilon) + N(\theta_2, \epsilon)$, we have

$$N(\Theta_1 \times \Theta_2, \epsilon) = N(\Theta_1, \epsilon) + N(\Theta_2, \epsilon). \tag{9.17}$$

9.4 Near Adaptive Minimaxity for (weak) ℓ_p balls

We are now ready to combine upper and lower bounds to arrive at an adaptive minimaxity result, up to logarithmic terms, for $\sqrt{2\log n}$ thresholding on ℓ_p balls, both strong and weak. More precise results will be given in later chapters, but the charm of the present statement lies in the relatively simple proof given the tools we have developed.

Consider the n-dimensional Gaussian white noise model $y_i = \theta_i + \epsilon z_i$ for $i = 1, \ldots, n$. Let $\hat{\theta}^U$ denote soft thresholding at $\epsilon\sqrt{2\log n}$. From the soft thresholding oracle inequality, Proposition 8.6, we have

$$r_\epsilon(\hat{\theta}^U, \theta) \leq (2\log n + 1)[\epsilon^2 + \mathcal{R}(\theta, \epsilon)].$$

If we have a bound on the weak ℓ_p norm of θ, then we can use Proposition 9.1 to bound

$$\mathcal{R}(\theta, \epsilon) \leq |\theta|^2_{IR,r}\epsilon^{2r} \leq [2/(2-p)]|\theta|^p_{w\ell_p}\epsilon^{2-p}. \tag{9.18}$$

In addition, $\mathcal{R}(\theta, \epsilon) = \sum_1^n \theta_i^2 \wedge \epsilon^2 \leq n\epsilon^2$, and so, with $c_p = 2/(2-p)$,

$$\sup_{\theta \in w\ell_{n,p}(C)} r_\epsilon(\hat{\theta}^U, \theta) \leq (2\log n + 1)[\epsilon^2 + c_p \min(n\epsilon^2, C^p\epsilon^{2-p})]. \tag{9.19}$$

To summarize, let $r^\circ_{n,p}(C, \epsilon) = \min(n\epsilon^2, C^p\epsilon^{2-p})$; this is also the main term in lower bound (9.16) for $R_N(\ell_{n,p}(C), \epsilon)$.

Theorem 9.2 *If $y \sim N_n(\theta, \epsilon^2 I)$ then for $0 < p < 2$ and $\epsilon < C$,*

$$\begin{aligned} c_1 r^\circ_{n,p}(C, \epsilon) &\leq R_N(\ell_{n,p}(C), \epsilon) \leq R_N(w\ell_{n,p}(C), \epsilon) \\ &\leq (2\log n + 1)[\epsilon^2 + c_p r^\circ_{n,p}(C, \epsilon)]. \end{aligned}$$

*The latter bound is attained for **all** ϵ by $\hat{\theta}^U$, soft thresholding at $\epsilon\sqrt{2\log n}$, compare* (9.19).

The minimax risks depend on parameters p, C and ϵ, whereas the threshold estimator $\hat{\theta}^U$ requires knowledge only of the noise level ϵ—which, if unknown, can be estimated as described in Chapter 7.5. Nevertheless, estimator $\hat{\theta}^U$ comes within a logarithmic factor of the minimax risk over a wide range of values for p and C. In the next section, we shall see how much of an improvement over linear estimators this represents.

The upper bound in Theorem 9.2 can be written, for $\epsilon < C$ and $n \geq 2$, as

$$c_2 \log n \cdot r^\circ_{n,p}(C, \epsilon)$$

if one is not too concerned about the explicit value for c_2. Theorem 11.7 gives upper and lower bounds that differ by constants rather than logarithmic terms.

Exercise 9.2 extends the weak ℓ_p risk bound (9.19) to general thresholds λ.

9.5 The woes of linear estimators.

We make some remarks about the maximum risk of linear estimators. While the techniques used are those of Section 4.8, the statistical implications are clearer now that we have established some properties of non-linear thresholding.

For any set $\Theta \subset \ell_2(I)$, we recall the notation for the "square" of Θ, namely $\Theta_+^2 = \{(\theta_i)^2 : \theta \in \Theta\}$. The *quadratically convex hull* of Θ is then defined as

$$\mathrm{QHull}(\Theta) = \{\theta : (\theta_i)^2 \in \mathrm{Hull}(\Theta_+^2)\}, \tag{9.20}$$

where $\mathrm{Hull}(S)$ denotes the closed convex hull of S. Of course, if Θ is closed and quadratically convex, then $\mathrm{QHull}(\Theta) = \Theta$. However, for ℓ_p- bodies with $p < 2$,

$$\mathrm{QHull}(\Theta_p(a)) = \{\theta : \sum a_i^2\theta_i^2 \le 1\}$$

is an *ellipsoid*. The key property of quadratic convexification is that it preserves the maximum risk of *linear* estimators.

Theorem 9.3 *Let Θ be solid orthosymmetric and compact. Then*

$$R_L(\Theta, \epsilon) = R_L(QHull(\Theta), \epsilon).$$

Proof Since Θ is orthosymmetric, (4.56) shows that linear minimax estimators may be found that are diagonal, with risk functions given by (4.50). Such risk functions are linear in $s = (\theta_i^2)$ and hence have the same maximum over $\mathrm{Hull}(\Theta_+^2)$ as over Θ_+^2. □

Remark. Combining Theorems 4.22 and 9.3 , we observe that the minimax linear risk of Θ is still determined by the hardest rectangular subproblem, but now of the *enlarged* set $\mathrm{QHull}(\Theta)$. Of course, $\mathrm{QHull}(\Theta)$ may be much larger that Θ, and so (in contrast to Corollary 4.23) it could certainly happen now that $R_L(\Theta) \gg R_N(\Theta)$: we will see examples in the later discussion of ℓ_p balls and Besov spaces.

For a key example, let $p < 2$ and consider $\Theta_{n,p}(C) = \{\theta : \sum_1^n |\theta_i|^p \le C^p\}$. Since $\mathrm{QHull}(\Theta_{n,p}(C)) = \Theta_{n,2}(C)$, we have using (4.58),

$$R_L(\Theta_{n,p}(C), \epsilon) = R_L(\Theta_{n,2}(C), \epsilon) = \frac{n\epsilon^2 C^2}{n\epsilon^2 + C^2} \in [\tfrac{1}{2}, 1] \min(n\epsilon^2, C^2). \tag{9.21}$$

Combining this with Theorem 9.2, which we may do by contrasting $r_{n,2}^\circ$ with $r_{n,p}^\circ$, we see that $C^p\epsilon^{2-p} \ll C^2$ exactly when $\epsilon \ll C$ and so for $p < 2$, the non-linear minimax risk is an algebraic order of magnitude smaller than the linear minimax risk. Furthermore, $\sqrt{2\log n}$ thresholding captures almost all of this gain, giving up only a factor logarithmic in n.

9.6 Function spaces and wavelet coefficients

To draw consequences of these results for function estimation, we need to relate sparsity ideas to smoothness classes of functions. We have seen when smoothness of functions is measured in, say, a mean-square sense—corresponding to L_2 integrals $\int (D^\alpha f)^2$—that linear estimators are close to optimal for mean-square error. On the other hand, it is apparent that non-linear estimators, for example using thresholding of wavelet coefficients, can greatly outperform linear estimators. In order to have a mathematical framework to describe this, we measure smoothness using other L_p measures, typically for $p < 2$, if estimation error is measured in mean-square. It might at first seem simplest, then, to consider L_p integrals of derivatives $\int |D^\alpha f|^p$, the Sobolev (semi-)norms. However, when working with wavelet bases $\{\psi_{jk}\}$, it turns out to be helpful to have the flexibility to sum separately over location k with an ℓ_p index and over scale j with an ℓ_q index. For this purpose it has proved helpful to formulate the notion of smoothness using Besov spaces.

This section gives some motivation for the definition of Besov measures of smoothness of functions. More systematic discussion can be found in the books by Meyer (1990), Frazier et al. (1991) and Triebel (1983). Instead the approach here is

- first, to give some heuristic remarks on L_p measures of smoothness and the tradeoff between worst-case, $p = \infty$, and average case, $p = 1$, measures,
- then, to explore the use of magnitudes of wavelet coefficients to describe smoothness of functions in examples with $p = 1, 2$ and ∞, and
- finally to give a definition of Besov norms on sequences of wavelet coefficients that encompasses the three examples.

This approach is somewhat roundabout, in that we do not begin with directly with Besov smoothness measures on functions. There are two reasons for this: the first is pragmatic: it is the sequence form that is most heavily used for the statistical theory. The second is to simplify exposition—while the rich theory of Besov spaces $B^\alpha_{p,q}(\Omega)$ on domains and $B^\alpha_{p,q}(\mathbb{R}^n)$ on Euclidean space can be approached in various, largely equivalent, ways, it does take some work to establish equivalence with the sequence form in terms of wavelet coefficients. To keep the treatment relatively self-contained, Appendix B gives the definition of $B^\alpha_{p,q}([0,1])$ in terms of moduli of smoothness and shows the equivalence with the sequence form using classical ideas from approximation theory.

Some Heuristics

Some of the traditional measures of smoothness are based on using L_p norms to measure the size of derivatives of the function: through the *seminorms* $|f|_{W^k_p} = (\int |D^k f|^p)^{1/p}$, where $1 \le p \le \infty$. When $p = \infty$, the integral is replaced by a supremum $\sup_x |D^k f(x)|$.

These seminorms vanish on polynomials of degree less than k, and so it is customary to add the L_p norm of the function in order to obtain an actual norm. Thus the (p^{th} power) of the Sobolev norm is defined by

$$\|f\|^p_{W^k_p} = \int |f|^p + \int |D^k f|^p.$$

The Sobolev space W_p^k of functions with k derivatives existing and integrable in L_p is then the (Banach) space of functions for which the norm is finite. Again, in the case $p = \infty$, the norm is modified to yield the *Hölder* norms

$$\|f\|_{C^k} = \|f\|_\infty + \|D^k f\|_\infty.$$

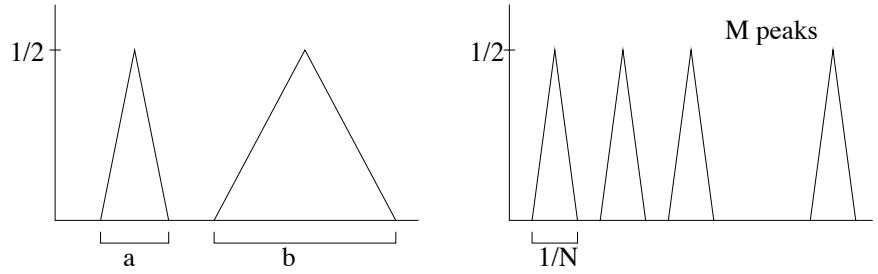

Figure 9.1

Figure 9.1 contains some examples to illustrate how smaller p corresponds to a more averaged and less worst-case measure of smoothness. For the function in the first panel,

$$\|f'\|_1 = 2, \qquad \|f'\|_2 = \sqrt{1/a + 1/b}, \qquad \|f'\|_\infty = 1/a.$$

In the $1-$norm the peaks have equal weight, while in the $2-$norm the narrower peak dominates, and finally in the $\infty-$norm, the wider peak has no influence at all. The second panel compares the norms of a function with M peaks each of width $1/N$:

$$\|f'\|_1 = M, \qquad \|f'\|_2 = \sqrt{MN}, \qquad \|f'\|_\infty = N.$$

The $1-$norm is proportional to the number of peaks, while the $\infty-$norm measures the slope of the narrowest peak (and so is unaffected by the number of spikes), while the $2-$norm is a compromise between the two. Thus, again smaller values of p are more forgiving of inhomegeneity.

Decay of wavelet coefficients–some examples

A basic idea is to use the magnitude of wavelet coefficients to describe the smoothness of functions. We explore this in three examples, $p = 1$, $p = 2$ and $p = \infty$, before showing how Besov sequence norms provide a unifying framework. To avoid boundary issues, we work with an orthonormal wavelet basis for $L_2(\mathbb{R})$, and so assume that a square integrable function f has expansion

$$f(x) = \sum_k \beta_{Lk}\varphi_{Lk}(x) + \sum_{j\geq L}\sum_k \theta_{jk}\psi_{jk}(x). \tag{9.22}$$

In the following proof, we see for the first time a pattern that recurs often with multiresolution models: a count or error that is a function of level j increases geometrically up to some critical level j_0 and decreases geometrically above j_0. The total count or error is then determined up to a constant by the critical level. While it is often easier to compute the

bound in each case as needed, we give a illustrative statement here. If $\beta, \gamma > 0$, then on setting $r = \gamma/(\beta+\gamma)$ and $c_\beta = (1-2^{-\beta})^{-1}$, we have

$$\sum_{j\in\mathbb{Z}} \delta 2^{\beta j} \wedge C2^{-\gamma j} \le (c_\beta + c_\gamma) C^{1-r}\delta^r. \tag{9.23}$$

The critical level may be taken as $j_0 = [j_*]$, where j_* is the solution to $\delta 2^{\beta j_*} = C2^{-\gamma j_*}$.

Hölder smoothness, $p = \infty$. We consider only $0 < \alpha < 1$, for which $|f(x) - f(y)| \le C|x-y|^\alpha$ for all x, y. Reflecting the uniformity in x, the conditions on the wavelet coefficients are uniform in k, with the decay condition applying to the scales j.

Theorem 9.4 *Suppose that $0 < \alpha < 1$ and that (φ, ψ) are C^1 and have compact support. Then $f \in C^\alpha(\mathbb{R})$ if and only if there exists $C > 0$ such that*

$$|\beta_{Lk}| \le C, \qquad |\theta_{jk}| \le C2^{-(\alpha+1/2)j}, \quad j \ge L. \tag{9.24}$$

Proof Assume first that $f \in C^\alpha$. Although this is a special case of Lemma 7.2, we give the detail here. What we rely on is that $\int \psi = 0$—this follows from Proposition 7.3 since ψ is C^1—and allows the wavelet coefficient to be rewritten as

$$|\langle f, \psi_{jk}\rangle| = 2^{-j/2}\int [f(x_k + 2^{-j}v) - f(x_k)]\psi(v)dv \tag{9.25}$$

for $x_k = k2^{-j}$. The Hölder smoothness now provides the claimed bound

$$\langle f, \psi_{jk}\rangle \le 2^{-j/2}|f|_\alpha 2^{-j\alpha}\int |v|^\alpha \psi(v)dv = c_{\psi,\alpha}|f|_\alpha 2^{-j(\alpha+1/2)}. \tag{9.26}$$

In the reverse direction, from (9.22), we can decompose the difference $f(x) - f(x')$ into terms $\Delta_\beta(f) + \Delta_\theta(f)$, where, for example,

$$\Delta_\theta(f) = \sum_{jk} \theta_{jk}[\psi_{jk}(x) - \psi_{jk}(x')].$$

We focus on $\Delta_\theta(f)$ here, since the argument for $\Delta_\beta(f)$ is similar and easier. Using the decay (9.24) of the coefficients θ_{jk},

$$|\Delta_\theta(f)| \le C\sum_{j\ge L} 2^{-(\alpha+1/2)j}\sum_k 2^{j/2}|\psi(2^j x - k) - \psi(2^j x' - k)|.$$

If the length of the support of ψ is S, then at most $2S$ terms in the sum over k are non-zero. In addition, the difference can be bounded using $\|\psi'\|_\infty$ when $|2^j x - 2^j x'| \le 1$, and using simply $2\|\psi\|_\infty$ otherwise. Hence

$$|\Delta_\theta(f)| \le c_\psi C\sum_{j\ge L} 2^{-\alpha j}\min\{2^j|x-x'|, 1\},$$

where $c_\psi = 2S\max\{2\|\psi\|_\infty, \|\psi'\|\}$. Let $j_* \in \mathbb{R}$ satisfy $2^{-j_*} = |x - x'|$. The summands above increase geometrically for $j < j_*$ (using the assumption that $\alpha < 1$!), and decrease geometrically for $j > j_*$. Consequently

$$|\Delta_\theta(f)| \le c_\alpha c_\psi C 2^{-\alpha j_*} \le C'|x-x'|^\alpha,$$

which, together with the bound for $\Delta_\beta(f)$ gives the Hölder bound we seek. □

Remark 9.5 We mention the extension of this result to $\alpha > 0$. Let $r = \lceil \alpha \rceil$. Assume that φ and ψ are C^r with compact support, and that ψ has at least r vanishing moments. If $f \in C^\alpha(\mathbb{R})$, then there exists positive C such that inequalities (9.24) hold. Conversely, if $\alpha > 0$ is not an integer, these inequalities imply that $f \in C^\alpha(\mathbb{R})$. The proof of these statements are a fairly straightforward extension the arguments given above (Exercise 9.4.)

When α is an integer, to achieve a characterization, a slight extension of C^α is needed, see Section B.3 for some extra detail.

Mean square smoothness, $p = 2$. Already in Chapter 3 we studied smoothness in the mean square sense, with norms $\|f\|^2_{W_2^r} = \int f^2 + \int (D^r f)^2$. Mean square smoothness also has a very natural expression in terms of wavelet coefficients. Suppose that (φ, ψ) are C^r. Then we may formally differentiate the homogeneous wavelet expansion $f = \sum_{jk} \theta_{jk} \psi_{jk}$ to obtain

$$D^r f(x) = \sum_{jk} 2^{rj} \theta_{jk} \psi^{(r)}_{jk}(x).$$

The system $\{\psi^{(r)}_{jk}\}$ is no longer orthonormal, but it turns out that it is the next best thing, namely a *frame*, meaning that there exist constants C_1, C_2 such that for all $f \in W_2^r$,

$$C_1 \sum_{jk} 2^{2rj} \theta^2_{jk} \le \Big\| \sum_{jk} 2^{rj} \theta_{jk} \psi^{(r)}_{jk}(x) \Big\|^2_2 \le C_2 \sum_{jk} 2^{2rj} \theta^2_{jk}. \tag{9.27}$$

These remarks render plausible the following result, proved in Appendix B.4.

Theorem 9.6 *If (ϕ, ψ) are C^r with compact support and ψ has $r+1$ vanishing moments, then there exist constants C_1, C_2 such that*

$$C_1 \|f\|^2_{W_2^r} \le \sum_k \beta^2_k + \sum_{j \ge 0, k} 2^{2rj} \theta^2_{jk} \le C_2 \|f\|^2_{W_2^r}. \tag{9.28}$$

Average smoothness, $p = 1$. Observing that $|\psi_{jk}|_{W_1^1} = 2^{j/2} \|\psi'\|_1$ and applying the triangle inequality to the wavelet expansion (9.22), we get

$$|f|_{W_1^1} \le 2^{L/2} \|\varphi'\|_1 \sum_k \beta_{Lk} + \|\psi'\|_1 \sum_j 2^{j/2} \sum_k |\theta_{jk}|$$

with a similar expression for $\|f\|_1$, with $2^{j/2}$ replaced by $2^{-j/2}$ and $\|\psi'\|_1$ by $\|\psi\|_1$.

We adopt the notation $\theta_{j\cdot}$ for the coefficients (θ_{jk}) at the jth level (and similarly for $\beta_{L\cdot}$). We have established the right hand half of

Theorem 9.7 *Suppose that (φ, ψ) are C^1 with compact support. Then there exist constants C_1 and C_2 such that*

$$C_1 \Big[\|\beta_{L\cdot}\|_1 + \sup_{j \ge L} 2^{j/2} \|\theta_{j\cdot}\|_1 \Big] \le \|f\|_{W_1^1} \le C_2 \Big[\|\beta_{L\cdot}\|_1 + \sum_{j \ge L} 2^{j/2} \|\theta_{j\cdot}\|_1 \Big].$$

For the left hand inequality, we suppose that $f \in W_1^1$. Since ψ is C^1, we conclude as before that $\int \psi = 0$, and it follows from integration by parts that if $\text{supp}\psi \subset I$, then

$$\int_I |f\psi| \le \tfrac{1}{2}\|\psi\|_1 \int_I |Df|$$

Suppose that ψ has support contained in $[-S+1, S]$. Applying the previous bound to $\theta_{jk} = \int f\psi_{jk}$ yields a bound $|\theta_{jk}| \le c_\psi 2^{-j/2}\int_{I_{jk}} |Df|$, where I_{jk} is the interval $2^{-j}[k-S+1, k+S]$. For j fixed, as k varies, any given point x falls in at most $2S$ intervals I_{jk}, and so adding over k yields, for each $j \ge L$,

$$2^{j/2}\sum_k |\theta_{jk}| \le 2S \cdot c_\psi \cdot |f|_{W_1^1}.$$

A similar but easier argument shows that we also have $\|\beta_{L\cdot}\|_1 \le 2^{L/2} \cdot 2S\|\varphi\|_\infty \cdot \|f\|_1$. Adding this to the last display yields the left bound.

Besov sequence norms

Comparing the three cases, we may contrast how the coefficients at a given level j are weighted and combined over k:

$$\begin{aligned} \text{Hölder, } p = \infty, &\qquad 2^{(\alpha+1/2)j}\|\theta_{j\cdot}\|_\infty, \\ \text{Mean square, } p = 2, &\qquad 2^{\alpha j}\|\theta_{j\cdot}\|_2, \\ \text{Average, } p = 1, &\qquad 2^{j(\alpha-1/2)}\|\theta_{j\cdot}\|_1. \end{aligned}$$

Introducing the index $a = \alpha + 1/2 - 1/p$, we can see each case as an example of $c_j = 2^{aj}\|\theta_{j\cdot}\|_p$. To combine the information in c_j across levels j, we use ℓ_q norms $(\sum_{j\ge L}|c_j|^q)^{1/q}$, which spans a range of measures from worst case, $q = \infty$, to average case, $q = 1$.

We use θ as an abbreviation for $\{\beta_{Lk}\} \cup \{\theta_{jk}, j \ge L, k \in \mathbb{Z}\}$, and define

$$\|\theta\|_{b^\alpha_{p,q}} = \|\beta_{L\cdot}\|_p + \Big(\sum_{j\ge L} 2^{ajq}\|\theta_{j\cdot}\|_p^q\Big)^{1/q}, \tag{9.29}$$

where again, $a = \alpha + 1/2 - 1/p$. In the case $q = \infty$, this is interpreted as

$$\|\theta\|_{b^\alpha_{p,\infty}} = \|\beta_{L\cdot}\|_p + \sup_{j\ge L} 2^{aj}\|\theta_{j\cdot}\|_p.$$

In full indicial glory, (9.29) becomes

$$\|\theta\|_{b^\alpha_{p,q}} = \Big(\sum_k |\beta_{Lk}|^p\Big)^{1/p} + \Big(\sum_{j\ge L} 2^{ajq}\big(\sum_k |\theta_{jk}|^p\big)^{q/p}\Big)^{1/q}.$$

Thus, the three parameters may be interpreted as follows:

$\alpha > 0$	smoothness
$p \in (0, \infty]$	averaging norm over locations k
$q \in (0, \infty]$	averaging norm over scales j.

With the Besov index notation, we may summarize the inequalities described in the three function class examples considered earlier as follows:

(i) Hölder smoothness, $p = \infty$. Then $\|\theta\|_{b^\alpha_{\infty,\infty}} \asymp \|f\|_{C^\alpha}$ for $\alpha > 0$, with the Zygmund class interpretation of C^α when $\alpha \in \mathbb{N}$, cf. B.11.

(ii) Mean-square smoothness, $p = 2$. Then $\|\theta\|^2_{b^\alpha_{2,2}} \asymp \int |f|^2 + |D^\alpha f|^2$.

(iii) Total variation, $p = 1$. Then $C_1\|\theta\|_{b^1_{1,\infty}} \le \int |f| + |Df| \le C_2\|\theta\|_{b^1_{1,\infty}}$, (i.e. $\alpha = 1$ here).

Example 9.8 Consider $f(x) = A|x|^\beta g(x)$. Here g is just a window function included to make f integrable; for example suppose that g is equal to 1 for $|x| \le 1/2$ and vanishes for $|x| \ge 1$ and is C^∞ overall. Assume that $\beta > -1/p$ so that $f \in L_p$. Suppose that the wavelet ψ has compact support, and $r > \beta + 1$ vanishing moments. Then it can be shown (Exercise 9.5) that $\|\theta\|_{b^\alpha_{p,\infty}} \le c_{\alpha\beta p} A < \infty$ whenever $\alpha \le \beta + 1/p$. Thus one can say that f has smoothness of order $\beta + 1/p$ when measured in L_p. Again, smaller p is more forgiving of a local singularity.

Besov function space norms

Our discussion here is brief; see Appendix B for more detail and references. To test if a function $f(x)$ belongs to function space $B^\alpha_{p,q}$, one starts with an integer $r > \alpha$ and the r-th order differences of $\Delta^r_h(f,x)$ of step length h, averaged over x in L_p. The largest such average for $h \le t$ defines the integral modulus of smoothness $\omega_r(f,t)_p$. The function $f \in B^\alpha_{p,\infty}$ if the ratio $\omega_r(f,t)_p/t^\alpha$ is uniformly bounded in $t > 0$. If instead the ratio belongs to $L_q((0,\infty), dt/t)$ then $f \in B^\alpha_{p,q}$. In each case the L_q norm of $\omega_r(f,t)_p/t^\alpha$ defines the seminorm $|f|_{B^\alpha_{p,q}}$ and then the norm $\|f\|_{B^\alpha_{p,q}} = \|f\|_p + |f|_{B^\alpha_{p,q}}$.

The discussion in Appendix B is tailored to Besov spaces on a finite interval, say $[0,1]$. It is shown there, Theorem B.22, that if (φ, ψ) are a C^r scaling function and wavelet of compact support giving rise to an orthonormal basis for $L_2[0,1]$ by the CDJV construction, then the sequence norm (9.29) and the function norm are equivalent

$$C_1\|f\|_{b^\alpha_{p,q}} \le \|f\|_{B^\alpha_{p,q}} \le C_2\|f\|_{b^\alpha_{p,q}}. \tag{9.30}$$

The constants C_i may depend on $(\varphi, \psi, \alpha, p, q, L)$ but *not* on f. The proof is given for $1 \le p, q \le \infty$ and $0 < \alpha < r$.

Relations among Besov spaces. The parameter q in the Besov definitions for averaging across scale plays a relatively minor role. It is easy to see, for example from (9.29), that

$$B^\alpha_{p,q_1} \subset B^\alpha_{p,q_2}, \qquad \text{for } q_1 < q_2$$

so that $B^\alpha_{p,q} \subset B^\alpha_{p,\infty}$ for all q, [1] and so we mainly focus on the $B^\alpha_{p,\infty}$ or more precisely the $b^\alpha_{p,\infty}$ norm in our discussion.

The relation between smoothness measured in different L_p norms as p varies is expressed by embedding theorems (see e.g. Peetre (1975, p. 63)

[1] If $(B_1, \|\cdot\|_1)$ and $(B_2, \|\cdot\|_2)$ are normed linear spaces, $B_1 \subset B_2$ means that for some constant C, we have $\|f\|_1 \le C\|f\|_2$ for all $f \in B_1$.

Proposition 9.9 *If $\alpha < \alpha'$ and $p > p'$ are related by $\alpha - 1/p = \alpha' - 1/p'$, then*

$$B^{\alpha'}_{p',q} \subset B^{\alpha}_{p,q}.$$

In fact, the proof becomes trivial using the sequence space form (9.29).

The situation can be summarized in Figure 9.2, which represents smoothness α in the vertical direction, and $1/p$ in the horizontal, for a fixed value of q. Thus the $y-$axis corresponds to uniform smoothness, and increasing spatial inhomogeneity to $1/p$. The imbeddings proceed down the lines of unit slope: for example, inhomogeneous smoothness $(\alpha', 1/p')$ with $\alpha' > 1/p'$ implies uniform smoothness of *lower* degree $\alpha = \alpha' - 1/p'$.

The line $\alpha = 1/p$ represents the boundary of continuity. If $\alpha > 1/p$, then functions in $B^{\alpha}_{p,q}$ are continuous by the embedding theorem just cited. However in general, the spaces with $\alpha = 1/p$ may contain discontinuous functions – one example is given by the containment $B^1_{1,1} \subset TV \subset B^1_{1,\infty}$.

Finally, for $B^{\alpha}_{p,q}([0,1])$, the line $\alpha = 1/p - 1/2$ represents the boundary of L_2 compactness - if $\alpha > 1/p - 1/2$, then $B^{\alpha}_{p,q}$ norm balls are compact in L_2: this observation is basic to estimation in the L_2 norm.

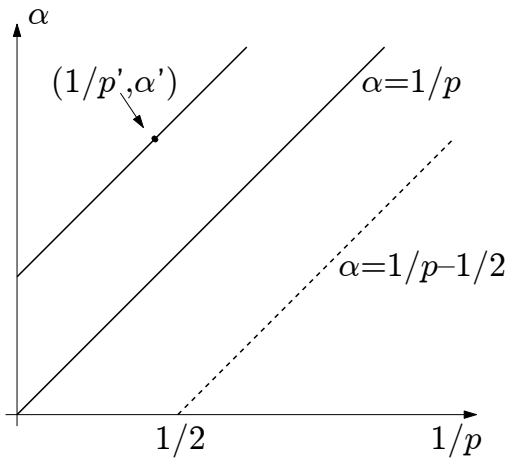

Figure 9.2 Summarizes the relation between function spaces through the primary parameters α (smoothness) and $1/p$ (integration in L_p).

Besov and Sobolev norms. While the Besov family does not match the Sobolev family precisely, we do have the containment, for $r \in \mathbb{N}$,

$$W^r_p \subset B^r_{p,\infty}.$$

In addition, when $p \leq 2$ we have

$$B^r_{p,p} \subset W^r_p.$$

We can write these embedding statements more explicitly. For $r \in \mathbb{N}$, there exists a constant C such that

$$\|f\|^p_{B^r_{p,\infty}} \leq C \int_0^1 |f|^p + |D^r f|^p. \tag{9.31}$$

In the other direction, for $0 < p \le 2$ and $r \in \mathbb{N}$, there exists a constant C such that

$$\int_0^1 |D^r f|^p \le C \|f\|^p_{b^r_{p,p}}. \tag{9.32}$$

A proof of (9.31) appears in Appendix B after (B.26), while for (9.32), see Johnstone and Silverman (2005b), though the case $p \le 1$ is elementary.

More generally, $W^r_p = F^r_{p,2}$ belongs to the Triebel class of spaces, in which the order of averaging over scale and space is reversed relative to the Besov class, see e.g. Frazier et al. (1991) or Triebel (1983). In particular, this approach reveals an exceptional case in which $W^r_2 = B^r_{2,2}$, cf Theorem 9.6.

Simplified notation

Consider a multiresolution analysis of $L_2[0,1]$ of one of the forms discussed in Section 7.1. For a fixed coarse scale L, we have the decomposition $L_2([0,1]) = V_L \oplus W_L \oplus W_{L+1} \oplus \cdots$, and associated expansion

$$f(x) = \sum_{k=0}^{2^L-1} \beta_k \varphi_{Lk}(x) + \sum_{j \ge L} \sum_{k=0}^{2^j-1} \theta_{jk} \psi_{jk}(x). \tag{9.33}$$

For the statistical results to follow, we adopt a simplified notation for the Besov sequence norms, abusing notation slightly. To this end, for $j < L$, define coefficients θ_{jk} to 'collect' all the entries of (β_k):

$$\begin{aligned} \theta_{jk} &= \beta_{2^j+k}, \qquad 0 \le j < L, 0 \le k < 2^j, \\ \theta_{-1,0} &= \beta_0. \end{aligned} \tag{9.34}$$

If we now write

$$\|\theta\|^q_{b^\alpha_{p,q}} = \sum_j 2^{ajq} \|\theta_{j\cdot}\|^q_p,$$

then we have an equivalent norm to that defined at (9.29). Indeed, since L is fixed and all norms on a fixed finite dimensional space, here $\mathbb{R}^{2^L}$ are equivalent, we have

$$\|\beta_\cdot\|_p \asymp \left(\sum_{j=-1}^{k-1} 2^{ajq} \|\theta_{j\cdot}\|^q_p \right)^{1/q}.$$

In the case of Besov spaces on $[0,1]$, we will therefore often write $\Theta^\alpha_{p,q}$ instead of $b^\alpha_{p,q}$.

Notation for norm balls. For $C > 0$, let

$$\Theta^\alpha_{p,q}(C) = \left\{ \theta : \sum_j 2^{ajq} \|\theta_{j\cdot}\|^q_p \le C^q \right\}.$$

Note that $\Theta^\alpha_{p,q}(C) \subset \Theta^\alpha_{p,\infty}(C)$, where

$$\Theta^\alpha_{p,\infty}(C) = \{\theta : \|\theta_{j\cdot}\|_p \le C 2^{-aj}, \text{for all } j \ge -1\}. \tag{9.35}$$

9.7 Besov Bodies and weak ℓ_p Balls

We have seen that the weak ℓ_p quasi-norm measures the sparsity of a coefficient sequence θ, with smaller p corresponding to greater sparsity. If a parameter set Θ is contained within $w\ell_p$, then all elements $\theta \in \Theta$ satisfy the same decay estimate. We now describe some relationships between the Besov and weak ℓ_p norms for the Besov spaces on $[0, 1]$. [As a matter of notation, we note that $c_{p\alpha}$ will denote a constant depending only on α and p, and not necessarily the same at each appearance.]

Proposition 9.10 *Suppose that $\alpha > 1/p - 1/2$, or equivalently that $p > p_\alpha = 2/(2\alpha+1)$. Then*

$$\Theta^\alpha_{p,q} \subset w\ell_{p_\alpha},$$

but $\Theta^\alpha_{p,q} \not\subset w\ell_s$ for any $s < p_\alpha$.

Recall that the notation $B_1 \subset B_2$ for (quasi-)normed linear spaces means that there exists a constant c such that $\|x\|_{B_2} \le c\|x\|_{B_1}$ for all x.

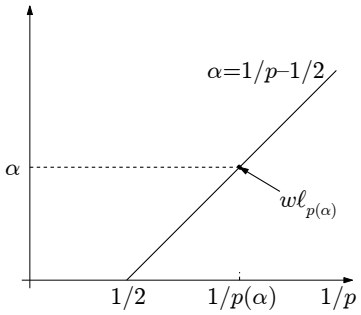

Figure 9.3 Besov spaces $\Theta^\alpha_{p,q}$ on the dotted line are included in $w\ell_p(\alpha)$

Proof Using the simplified notation for Besov norm balls, we need to show that, for some constant c_1 allowed to depend on α and p,

$$\Theta^\alpha_{p,q}(C) \subset w\ell_{p_\alpha}(c_1 C) \tag{9.36}$$

for $p > p_\alpha$, but that no such constant exists for $w\ell_s$ for $s < p_\alpha$.

Since $\Theta^\alpha_{p,q} \subset \Theta^\alpha_{p,\infty}$, it suffices to establish (9.36) for $\Theta^\alpha_{p,\infty}(C)$, which in view of (9.35) is just a product of ℓ_p balls $\ell_{2^j,p}(C2^{-aj})$. Hence, using (9.17) and (9.15) to calculate dimension bounds for products of ℓ_p balls, and abbreviating $\Theta = \Theta^\alpha_{p,q}(C)$, we arrive at

$$N(\Theta, \epsilon) \le 1 + \sum_j \min\{2^j, (C\epsilon^{-1}2^{-aj})^p\}.$$

The terms in the sum have geometric growth and decay away from the maximum j_* defined by equality between the two terms: thus $2^{j_*(\alpha+1/2)} = C/\epsilon$, independent of $p > p_\alpha$. Hence $N(\Theta, \epsilon) \le c_{\alpha p} 2^{j_*}$ where we may take $c_{\alpha p} = 3 + (1 - 2^{-ap})^{-1} < \infty$ for $ap > 0$, which is equivalent to $p > p_\alpha$. Now, from the definition of j_*, we have $\epsilon^{p_\alpha} 2^{j_*} = C^{p_\alpha}$, and so

$$\epsilon^{p_\alpha} N(\Theta, \epsilon) \le c_{p\alpha} C^{p_\alpha} \tag{9.37}$$

and so, using the criterion (9.9), we obtain (9.36) with $c_1 = c_{p\alpha}^{1/p_\alpha}$.

For the second part, consider the Besov shells $\Theta^{(j_0)} = \{\theta \in \Theta^\alpha_{p,q}(C) : \theta_{jk} = 0 \text{ unless } j = j_0.\} \equiv \ell_{2^j,p}(C2^{-ja})$. Consider the shell corresponding to level $j = [j_*]$ with j_* determined above: since this shell belongs to $\Theta = \Theta^\alpha_{p,q}(C)$ for all q, we have, from (9.15)

$$N(\Theta, \epsilon) \geq \min\{2^j, [(C2^{-ja}/\epsilon)^p]\} \geq \tfrac{1}{2}2^{j_*} = \tfrac{1}{2}(C/\epsilon)^{p_\alpha}, \tag{9.38}$$

and hence that $\epsilon^s N(\Theta, \epsilon) \geq \frac{1}{2}C^{p_\alpha}\epsilon^{s-p_\alpha}$ is unbounded in ϵ if $s < p_\alpha$. □

Remarks.

1.$(1/p, \alpha)$ `Diagram showing spaces embedding in` $w\ell_{p^*}$. See Figure

2. Note that in the case $\alpha = 1/p - 1/2$, we have $a = 0$, and so

$$\Theta^\alpha_{p,p}(C) = \{\theta : \sum_j \sum_k |\theta_{jk}|^p \leq C^p\} = \ell_p(C).$$

Note that there is no compactness here!

3. What happens to the embedding results when $p = p_\alpha$? For $q \leq p_\alpha$ we have

$$\Theta^\alpha_{p_\alpha,q}(C) \subset \Theta^\alpha_{p_\alpha,p_\alpha}(C) = \ell_{p_\alpha}(C) \subset w\ell_{p_\alpha}(C)$$

It can also be seen that $\ell_{p_\alpha}(C) \subset \Theta^\alpha_{p_\alpha,\infty}(C)$.

4. However, there is no containment relation between $w\ell_{p_\alpha}(C)$ and $\Theta^\alpha_{p_\alpha,\infty}(C)$:

(i) The vector θ defined by $\theta_{jk} = C\delta_{k0} \in \Theta^\alpha_{p_\alpha,\infty}(C)$ but is not in $w\ell_{p_\alpha}(C')$ for any C'.

(ii) The vectors θ^{j_0} defined by $\theta^{j_0}_{jk} = \delta_{jj_0}Ck^{-1/p_\alpha}$ for $k = 1, \ldots 2^j$ are each in $w\ell_{p_\alpha}(C)$, but $\|\theta^{j_0}\|_{b^\alpha_{p_\alpha,\infty}} \sim Cj_0^{1/p}$.

9.8 A framework for wavelet shrinkage results

As always our setting is the continuous Gaussian white noise model (1.18). This can be converted into a sequence model by taking coefficients in any orthonormal basis, as described in (1.21) - (1.23). Let us repeat this now explicitly in the context of an orthonormal wavelet basis adapted to $L_2[0, 1]$.

Given a fixed coarse scale L, suppose that we have an orthonormal basis $\{\varphi_{Lk}, k = 0, \ldots, 2^L - 1\} \cup \{\psi_{jk}, k = 0, \ldots, 2^j - 1, j \geq L\}$ leading to expansion (9.33) for any $f \in L_2[0, 1]$. Asking the reader's forbearance for an arrant abuse in service of compact notation, we propose to use the symbols ψ_{jk} for $\varphi_{L,2^j+k}$ when $0 \leq j < L, 0 \leq k < 2^j$ (and $\psi_{-,10}$ for $\varphi_{L,0}$). We may then, consistent with the convention (9.34), define an index set $\mathcal{I} = \{(jk) : j \geq 0, k = 0, \ldots, 2^j - 1\} \cup \{(-1, 0)\}$ and write $\theta_{jk} = \langle f, \psi_{jk}\rangle$ for $(jk) \in \mathcal{I}$. With this understanding, our wavelet sequence model becomes

$$y_{jk} = \theta_{jk} + \epsilon z_{jk}, \qquad (jk) \in \mathcal{I}, \tag{9.39}$$

with $y_{jk} = \langle \psi_{jk}, dY\rangle$ and $z_{jk} = \langle \psi_{jk}, dW\rangle$.

Every function $f \in L_2[0, 1]$ has the expansion $f = \sum \theta_I \psi_I$, and the Parseval relation $\int f^2 = \sum_I \theta_I^2$ shows that the mapping from f to θ is an isometry, which we sometimes write $\theta[f]$. Thus $\theta[f]_I = \langle f, \psi_I\rangle$ for $I \in \mathcal{I}$. For the inverse mapping, we write $f[\theta]$ for the function defined by $f[\theta](t) = \sum \theta_I \psi_I(t)$.

In the continuous white noise model, we estimate the function f using mean integrated squared error $\int(\hat{f} - f)^2$, and of course

$$\|\hat{f} - f\|_2^2 = \sum_{\mathcal{I}}(\hat{\theta}_I - \theta_I)^2 = \|\hat{\theta} - \theta\|_{\ell_2}^2. \tag{9.40}$$

We can now use the Besov bodies to define function classes

$$\mathcal{F} = \mathcal{F}_{p,q}^{\alpha}(C) = \{f : \theta[f] \in \Theta_{p,q}^{\alpha}(C)\}. \tag{9.41}$$

secure in the knowledge that under appropriate conditions on the multiresolution analysis, these function classes will be equivalent to norm balls in $B_{p,q}^{\alpha}[0,1]$.

Our choice of definitions has made the continuous white noise estimation problem exactly equivalent to the seqence model. Using the natural definition of minimax risks, we therefore have the identity

$$\begin{aligned} R_{\mathcal{E}}(\mathcal{F},\epsilon) &= \inf_{\hat{f}\in\mathcal{E}} \sup_{\mathcal{F}} E_f\|\hat{f} - f\|^2 \\ &= \inf_{\hat{\theta}\in\mathcal{E}} \sup_{\Theta} E_\theta\|\hat{\theta} - \theta\|^2 = R_{\mathcal{E}}(\Theta,\epsilon). \end{aligned} \tag{9.42}$$

Here $\mathcal{E}$ might denote the class of all estimators. We will also be particularly interested in certain classes of coordinatewise estimators applied to the wavelet coefficients. In the sequence model, this means that the estimator has the form $\hat{\theta}_I(y) = \hat{\delta}_I$, where $\hat{\delta}$ belongs to one of the four families in the following table.

Family	Description	Form of $\hat{\delta}_I(y)$
$\mathcal{E}_L$	Diagonal linear procedures in the wavelet domain	$\hat{\delta}_I^L(y) = c_I \cdot y$
$\mathcal{E}_S$	Soft thresholding of wavelet coefficients	$\hat{\delta}_I^S(y) = (\|y\| - \lambda_I)_+\mathrm{sgn}(y)$
$\mathcal{E}_H$	Hard thresholding of wavelet coefficients	$\hat{\delta}_I^H(y) = y1_{\{\|y\|\geq\lambda_I\}}$
$\mathcal{E}_N$	Scalar nonlinearities of wavelet coefficients	Arbitrary $\hat{\delta}_I^N(y)$

The corresponding estimators in classes $\mathcal{E}$ in (9.42) in the continous white noise model are defined by $\hat{f} = f[\hat{\theta}] = \sum_I \hat{\delta}_I(\langle\psi_I, dY\rangle)\psi_I$, where $\hat{\theta} \in \mathcal{E}_S, \mathcal{E}_L$ and so on.

9.9 Adaptive minimaxity for $\sqrt{2\log n}$ thresholding

We combine the preceding results about Besov bodies and weak ℓ_p with properties of thresholding established in Chapter 8 to derive adaptive near minimaxity results for $\sqrt{2\log n}$ thresholding over Besov bodies $\Theta_{p,q}^{\alpha}(C)$. Consider the dyadic sequence model (9.39) and apply soft thresholding to the first $n = \epsilon^{-2} = 2^J$ coefficients, using threshold $\lambda_\epsilon =$

$\sqrt{2\log\epsilon^{-2}} = \sqrt{2\log n}$:

$$\hat{\theta}^U_{jk} = \begin{cases} \eta_S(y_{jk}, \lambda_\epsilon \epsilon) & j < J \\ 0 & j \geq J. \end{cases} \tag{9.43}$$

Let $n = \epsilon^{-2} = 2^J$ and $\mathcal{I}_n = \{(jk) : j < J\}$ denote the collection of indices of the first n wavelet (and scaling) coefficients. The corresponding function estimate

$$\hat{f}_n(t) = \sum_{(jk)\in\mathcal{I}_n} \hat{\theta}^U_{jk}\psi_{jk}(t). \tag{9.44}$$

Remarks. 1. A variant that more closely reflects practice would spare the coarse scale coefficients from thresholding: $\hat{\theta}_{jk}(y) = y_{jk}$ for $j < L$. In this case, we have

$$\hat{f}_n(t) = \sum_{k=0}^{2^L-1} y_{Lk}\varphi_{Lk}(t) + \sum_{j=L}^{J-1}\sum_{k=0}^{2^j-1} \hat{\theta}^U_{jk}\psi_{jk}(t) \tag{9.45}$$

where $y_{Lk} = \langle \varphi_{Lk}, dY\rangle$. Since L remains fixed (and small), this will not affect the asymptotic results below.

2. Although not strictly necessary for the discussion that follows, we have in mind the situation of fixed equi-spaced regression: $y_i = f(i/n) + \sigma e_i$ – compare (2.64). After a discrete orthogonal wavelet transform, we would arrive at (9.39), restricted to $j < J = \log_2 n$, and with calibration $\epsilon = \sigma n^{-1/2}$. The restriction of thresholding in (9.43) to levels $j < J$ corresponds to what we might do with real data: namely threshold the n empirical discrete orthogonal wavelet transform coefficients.

The next theorem gives an indication of the broad adaptation properties enjoyed by wavelet thresholding.

Theorem 9.11 *Assume that $\alpha > (1/p - 1/2)_+$, $0 < p, q \leq \infty$, $0 < C < \infty$. If $p < 2$, then assume also that $\alpha \geq 1/p$. Let $\hat{\theta}^U$ denote soft thresholding at $\epsilon\sqrt{2\log n}$, defined at* (9.43) *Then for any Besov body $\Theta = \Theta^\alpha_{p,q}(C)$ and as $\epsilon \to 0$,*

$$\begin{aligned} \sup_\Theta r_\epsilon(\hat{\theta}^U, \theta) &\leq c_{\alpha p}(2\log\epsilon^{-2})C^{2(1-r)}\epsilon^{2r}(1 + o(1)) \\ &\leq c_{\alpha p}(2\log\epsilon^{-2})R_N(\Theta, \epsilon)(1 + o(1)). \end{aligned} \tag{9.46}$$

A key aspect of this theorem is that thresholding "learns" the rate of convergence appropriate to the parameter space Θ. The definition (9.43) of $\hat{\theta}^U$ does not depend at all on the parameters of $\Theta^\alpha_{p,q}(C)$, and yet, when restricted to such a set, the MSE attains the rate of convergence appropriate to that set, subject only to extra logarithmic terms.

Proof Let $\theta^{(n)}$ and $\hat{\theta}^{(n)}$ denote the first n coordinates – i.e. (j,k) with $j < J$ – of θ and $\hat{\theta}$ respectively. To compute a bound on the risk (mean squared error) of $\hat{\theta}$, we apply the soft thresholding risk bound (8.23) of Proposition 8.6 to $\hat{\theta}^{(n)}$. Since $\hat{\theta}_{jk} \equiv 0$ except in these first n coordinates, what remains is a "tail bias" term:

$$\begin{aligned} r(\hat{\theta}^U, \theta) &= E_\theta\|\hat{\theta}^{(n)} - \theta^{(n)}\|^2 + \|\theta^{(n)} - \theta\|^2 \\ &\leq (2\log\epsilon^{-2} + 1)[\epsilon^2 + \mathcal{R}(\theta^{(n)}, \epsilon)] + \sum_{j\geq J}\|\theta_j\|^2. \end{aligned} \tag{9.47}$$

Bound (9.47) is a pointwise estimate – valid for each coefficient vector θ. We now investigate its consequences for the worst case MSE of thresholding over Besov bodies $\Theta = \Theta^\alpha_{p,q}(C)$. Given α, we set as before,

$$r = 2\alpha/(2\alpha+1), \qquad p_\alpha = 2/(2\alpha+1) = 2(1-r).$$

Then, using the definition of the ideal risk seminorm, followed by the third bound of (9.11), we have for any $\theta \in \Theta^\alpha_{p,q}(C)$:

$$\mathcal{R}(\theta^{(n)}, \epsilon) \le |\theta|^2_{IR,r}\epsilon^{2r} \le c_\alpha |\theta|^{p(\alpha)}_{w\ell_{p(\alpha)}}\epsilon^{2r} \le c_{\alpha p} C^{2(1-r)}\epsilon^{2r}, \tag{9.48}$$

where the final inequality uses the Besov space embedding result of Proposition 9.10, compare (9.36).

Tail bias. First, note the simple bound

$$\sup\{\|\theta\|_2 : \|\theta\|_p \le C, \theta \in \mathbb{R}^n\} = C n^{(1/2-1/p)_+}. \tag{9.49}$$

which follows from a picture: when $p < 2$, the vectors having largest ℓ_2 norm in an ℓ_p ball are sparse, being signed permutations of the "spike" $C(1,0,\dots,0)$. When $p \ge 2$, the extremal vectors are *dense*, being sign flips of $Cn^{-1/p}(1,\dots,1)$.

Now we combine across levels to obtain a tail bias bound. Suppose that $\theta \in \Theta^\alpha_{p,q} \subset \Theta^\alpha_{p,\infty}$: we have $\|\theta_j\|_p \le C2^{-aj}$, and after using (9.49), also $\|\theta_j\|_2 \le C2^{-\alpha' j}$, after we set $\alpha' = \alpha - (1/p - 1/2)_+$. Clearly then $\sum_{j\ge J}\|\theta_j\|_2^2$ is bounded by summing the geometric series and we arrive at the tail bias bound

$$\sup_{\theta\in\Theta^\alpha_{p,q}(C)} \|\theta^{(n)} - \theta\|^2 \le c_{\alpha'} C^2 2^{-2\alpha' J}. \tag{9.50}$$

Inserting the ideal risk and tail bias bounds (9.48) and (9.50) into (9.47), we get the non-asymptotic bound

$$r(\hat\theta^U, \theta) \le (2\log\epsilon^{-2}+1)[\epsilon^2 + c_{\alpha p}C^{2(1-r)}\epsilon^{2r}] + c_{\alpha p}C^2\epsilon^{4\alpha'}. \tag{9.51}$$

Now suppose that C is fixed and $\epsilon \to 0$. We verify that $\epsilon^{2\alpha'} = o(\epsilon^r)$. This is trivial when $p \ge 2$, since $2\alpha > r$. When $p < 2$, the condition $\alpha \ge 1/p$ implies $2\alpha' = 2a \ge 1 > r$. This completes the proof of (9.47).

Lower Bounds. We saw in the proof of Proposition 9.10 that $\Theta^\alpha_{p,q}(C)$ contains $\epsilon-$ hypercubes of dimension $N(\Theta,\epsilon) \ge c_0(C/\epsilon)^{p(\alpha)}$. Hence the general hypercube lower bound (9.13) implies that

$$R_N(\Theta,\epsilon) \ge c_1(C/\epsilon)^{p(\alpha)}\epsilon^2 = c_1 C^{2(1-r)}\epsilon^{2r}. \tag{9.52}$$

□

Remark. The condition $\alpha \ge 1/p$ in the $p < 2$ case could be weakened to $\alpha > 1/p - 1/2$ by choosing to threshold, say $(\log_2\epsilon^{-2})^2$ levels rather than $\log_2\epsilon^{-2}$. However, we retain the latter choice in order to stay closer to what one does with data in practice. The condition $\alpha > 1/p$ implies, by embedding results mentioned in Section 9.6, that the functions $f[\theta]$ are continuous, which seems a reasonable condition in order to speak sensibly of point evaluation in model (2.64).

9.10 Estimation at a point.

In this section, we change point of view and consider the estimation of the value $f(t_0)$ of a function at a point $t_0 \in (0,1)$ on the basis of observations from dyadic sequence model (9.39). We again consider the wavelet threshold estimator with threshold $\delta_n = \epsilon_n \sqrt{2 \log n}$, this time without shrinkage of coarse scale coefficients, so that the estimator $\hat{f}_n(t_0)$ is given by (9.45).

In global estimation, we have seen that results are naturally obtained for average ($p < \infty$) as well as uniform ($p = \infty$) measures of smoothness. For estimation at a point, we need smoothness information locally, near that point, which would not be directly guaranteed by an average measure. For that reason, we adopt a hypothesis of Hölder smoothness here. Recall from (9.41) that $\mathcal{F}^{\alpha}_{\infty,\infty}(C) = \{f : \theta[f] \in \Theta^{\alpha}_{\infty,\infty}(C)\}$.

Theorem 9.12 *Suppose that the wavelet ψ is C^{α}, has compact support and has at least $\lceil \alpha \rceil$ vanishing moments. Let $r = 2\alpha/2\alpha + 1$. Then*

$$\sup_{f \in \mathcal{F}^{\alpha}_{\infty,\infty}(C)} E[\hat{f}_n(t_0) - f(t_0)]^2 \le c_{\psi,\alpha} C^{2(1-r)} (\frac{\log n}{n})^r (1 + o(1)). \tag{9.53}$$

Proof Decompose the estimation error over 'coarse', 'mid' and 'tail' scales:

$$\hat{f}_n(t_0) - f(t_0) = \sum_{I \in c} a_I + \sum_{I \in m} a_I + \sum_{I \in t} a_I. \tag{9.54}$$

The main term runs over the mid scales,

$$\sum_{I \in m} a_I = \sum_{j=L}^{J-1} \sum_k (\hat{\theta}_{jk} - \theta_{jk}) \psi_{jk}(t_0),$$

and points to the new point in the proof. In global estimation, the error $\|\hat{f} - f\|^2$ is expressed in terms of that of the coefficients, $\sum(\hat{\theta}_I - \theta_I)^2$, by Parseval's equality, using the orthonormality of the basis functions ψ_{jk}. In estimation at a point t_0, there is no orthogonality in t, and instead we bound the root mean squared (RMS) error of a sum by the sum of the RMS errors:

$$E\Big(\sum_I a_I\Big)^2 = \sum_{I,J} E a_I a_J \le \sum_{I,J} \sqrt{E a_I^2} \sqrt{E a_J^2} = \Big(\sum_I \sqrt{E a_I^2}\Big)^2. \tag{9.55}$$

We can use previous results to bound the individual terms $E a_I^2$. Indeed, recall from (8.9) the mean squared error bound for a soft threshold estimator with threshold λ, here given for noise level ϵ and $\bar{\lambda}^2 = 1 + \lambda^2$:

$$r_S(\lambda\epsilon, \theta; \epsilon) \le \epsilon^2 r(\lambda, 0) + \theta^2 \wedge \bar{\lambda}^2 \epsilon^2 \tag{9.56}$$

Since $\lambda = \sqrt{2 \log n}$, we have from (8.7) that $r(\lambda, 0) \le n^{-1}$. We use the Hölder continuity assumption and Lemma 7.2 to bound $|\theta_{jk}| \le c C 2^{-(\alpha+1/2)j}$. In conjunction with $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we obtain

$$\begin{aligned} \sqrt{E a_I^2} &\le |\psi_I(t_0)| \, [\epsilon \sqrt{r(\lambda, 0)} + |\theta_I| \wedge \bar{\lambda}\epsilon] \\ &\le c_{\psi} 2^{j/2} \;\; [1/n + C 2^{-(\alpha+1/2)j} \wedge \delta_n] \end{aligned}$$

where $\delta_n = \bar{\lambda}\epsilon$ can be taken as $\sqrt{2\log n/n}$ by increasing c_ψ slightly.

In the sum over I, to control the number of terms we use the compact support assumption on ψ: suppose that it has length S. Then for a given level j, at most S terms $\psi_{jk}(t_0)$ are non-zero. Hence

$$\begin{aligned}\sum_{I\in m}\sqrt{Ea_I^2} &\le cS2^{J/2}/n + cS\sum_{j<J}2^{j/2}(C2^{-(\alpha+1/2)j}\wedge\delta_n)\\ &\le c/\sqrt{n} + c_{\alpha,\psi}C^{1-r}\delta_n^r, \end{aligned} \tag{9.57}$$

where we have used geometric decay bound (9.23).

To organize the rest of the proof, combine (9.54) and (9.55); we obtain

$$E[\hat{f}_n(t_0) - f(t_0)]^2 = E\Big(\sum_{I\in c\cup m\cup t} a_I\Big)^2 \le \Big(\sum_{I\in c\cup m\cup t}\sqrt{Ea_I^2}\Big)^2.$$

In the coarse scale sum over $I \in c$, the terms $a_I = (y_{Lk} - \beta_{Lk})\varphi_{Lk}(t_0)$ for $k = 0,\dots,2^L-1$. We have $Ea_I^2 \le c_\varphi^2 n^{-1}$ and so

$$\sum_{I\in c}\sqrt{Ea_I^2} \le 2^L c_\varphi n^{-1/2}. \tag{9.58}$$

In the tail sum over $I \in t$, we have $a_I = \theta_I\psi_I(t_0)$ for $I = (jk)$ and $j \ge J$. Using again the Hölder coefficient decay bound and the compact support of ψ,

$$\sum_{I\in t}|a_I| \le c_\psi S\sum_{j\ge J} C2^{-(\alpha+1/2)j}\cdot 2^{j/2} \le cC2^{-\alpha J} = cCn^{-\alpha}. \tag{9.59}$$

Combining the coarse, mid and tail scale bounds (9.58), (9.57) and (9.59), we complete the proof:

$$E[\hat{f}_n(t_0) - f(t_0)]^2 \le (c_1 n^{-1/2} + c_2 C^{1-r}\delta_n^r + c_3 Cn^{-\alpha})^2 \le c_2^2 C^{2(1-r)}\delta_n^{2r}(1+o(1)).$$

□

Remark. If we knew α and C, then we could construct a linear minimax estimator $\hat{f}_n^{\alpha,C} = \sum_I c_I y_I$ where the (c_I) are the solution of a quadratic programming problem depending on C, α, n (Ibragimov and Khas'minskii (1982); Donoho and Liu (1991); Donoho (1994)). This estimator has worst case risk over $\Lambda^\alpha(C)$ asymptotic to $c_\alpha C^{2(1-r)}n^{-r}$. However, if the Hölder class is incorrectly specified, then this linear estimator will have a suboptimal rate of convergence over the true Hölder class (cf. also the discussion ending Section 10.6).

In contrast, the wavelet threshold estimator (9.44) does not depend on the parameters (C,α), and yet achieves nearly the optimal rate of convergence – up to a factor $\log^r n$ – over all the Hölder classes.

Lepskii (1991) and Brown and Low (1996b) have shown that this rate penalty $\log^r n$ is in fact optimal: even if the correct Hölder class is one of two, specified by pairs (α_0, C_0) and $(\alpha_1, C_1$ with $\alpha_0 < \alpha_1$, then

$$\inf_{\hat{f}_n}\max_{i=0,1}(C_i^{2(r_i-1)}n^{r_i})\sup_{\Lambda^{\alpha_i}}(C_i)E[\hat{f}_n(t_0) - f(t_0)]^2 \ge c_2\log^{r_0} n.$$

Remark. It is evident both intuitively from Lemma 7.2 that the full global constraint of

Hölder regularity on $[0, 1]$ is not needed: a notion of local Hölder smoothness near t_0 is all that is used. Indeed Lemma 7.2 is only needed for indices I with $\psi_I(t_0) \neq 0$.

9.11 *Outlook:* Overview of remaining chapters.

The statistical results which concluded the first part, Theorems 9.2 and 9.11, make quite informative statements about co-ordinatewise 'universal' thresholding. For example, the class of parameter spaces is broad enough to decisively distinguish thresholding from any linear estimator. The results do however raise or leave open a number of related questions, some of which are explored in more detail in the second part of the book, and are outlined here.

One basic theme, already apparent in the structure of this chapter, recurs in each setting. A result or technique is first formulated in a 'single sequence' model, as for example in Theorem 9.2. The same technique can then be carried over to function estimation by regarding each level j in the wavelet transform as an instance of the sequence model, and then combining over levels, as for example in Theorem 9.11

Other loss functions (Chapter 10). In Theorems 9.2 and 9.11, as in most of the rest of this book, the focus has been on the squared error loss function. We give an analog of the near-minimaxity result Theorem 9.11 for loss functions $\|\hat{\theta} - \theta\|_{b^{\alpha'}_{p',q'}}$ from the class of Besov norms. Wavelet thresholding, at threshold $\epsilon\sqrt{2\log n}$, is simultaneously near asymptotic minimax (up to at most a logarithmic factor) for all these loss functions. The technique is borrowed from the deterministic optimal recovery model of numerical analysis. The early sections do the preparatory work in the single sequence model.

Losing the log term: optimal rates (Chapters 11, 12). It is of both theoretical and practical interest to understand whether it is possible to remove the logarithmic gap ($\log n$ in Theorem 9.2 and $\log \epsilon^{-2}$ in Theorem 9.11) between upper and lower bounds, while still using adaptive estimators of threshold type. (Recall, for example, Figure 7.5, in which the threshold $\sqrt{2\log n}$ was too large).

This question is intimately linked with the use of data-dependent thresholds. We sketch a heuristic argument that suggests that an estimator using a constant threshold $\lambda\epsilon$ (even if λ depends on n) cannot be simultaneously minimax over ℓ_p balls $\ell_{n,p}(C)$ as p and C vary.

Suppose $y \sim N_n(\theta, \epsilon^2 I)$ and $\hat{\theta}_{\delta,i}(y) = \eta_S(y_i, \epsilon\lambda_\delta)$ where $\lambda = \sqrt{2\log \delta^{-1}}$. Using Corollary 8.4 and adding over co-ordinates yields[2]

$$r(\hat{\theta}_\delta, \theta) \leq 2\delta n\epsilon^2 + (1 + 2\log\delta^{-1}) \sum_{i=1}^{n} \theta_i^2 \wedge \epsilon^2.$$

Now maximize over $\lambda \in \ell_{n,p}(C)$–it can be shown, e.g. Lemma 13.19–that for $1 \leq (C/\epsilon)^p \leq n$, we have $\sum \theta_i^2 \wedge \epsilon^2 \leq C^p \epsilon^{2-p}$, and so

$$\sup_{\theta \in \ell_{n,p}(C)} r(\hat{\theta}_\delta, \theta) \leq 2\delta n\epsilon^2 + (1 + 2\log\delta^{-1}) C^p \epsilon^{2-p}.$$

We might select δ to minimize the right side bound: this immediately leads to a proposed choice $\delta = n^{-1}(C/\epsilon)^p$ and threshold $\lambda = \sqrt{2\log n(\epsilon/C)^p}$. Observe that as the signal to

[2] the factor $2\delta n\epsilon^2$, while a looser bound than given by (8.13), leads to cleaner heuristics here.

noise ratio C/ϵ increases from 1 to $n^{1/p}$, the nominally optimal threshold decreases from $\sqrt{2\log n}$ to 0, and no single threshold value appears optimal for anything other than a limited set of situations.

A number of approaches to choosing a data dependent threshold were reviewed in Section 7.6. In Chapter 11 we explore another alternative, based on complexity penalized model selection. Informally it may be described as imposing a penalty of order $2k\log(n/k)$ on models of size k. If we denote by $\hat{k}$ the size of the selected model, the associated threshold is often close to $\epsilon(2\log n/\hat{k})^{1/2}$, so that larger or 'denser' selected models correspond to smaller thresholds and 'sparser' models to larger ones. A virtue of the complexity penalized approach is the existence of oracle inequalities analogous to Proposition 8.6, but without the multiplicative log term—loosely, one may say that the logarithm was incorporated instead into the penalty. The corresponding estimator is defined adaptively, i.e. without reference to p and C, and yet satisfies *non-asymptotic* upper and lower bounds for MSE over the range of ℓ_p balls, that differ only at the level of constants.

The complexity penalized bounds have implications for wavelet shrinkage estimation of functions when applied separately at each level of a multiresolution analysis. In Chapter 12, we show that this leads to estimators that are rate-adaptive over a wide range of Besov spaces: essentially an analog of Theorem 9.11 without the $\log n$ multiplicative term. In this chapter we also return to the theme of linear inverse problems used as a class of examples in earlier chapters: the wavelet-vaguelette decomposition (WVD) allows one to construct adaptive rate-optimal wavelet shrinkage estimators for a class of inverse problems possessing a WVD.

Exact constants (Chapters 13, 14). In discussing adaptive minimaxity, we have emphasized the practical importance of estimators which do not not depend on the indices of parameter spaces such as $\ell_p(C)$ and $\Theta^{\alpha}_{p,q}(C)$. However, in order to calibrate the performance of these estimators, and to more fully understand the structure of these estimation settings, it is also of interest to evaluate exactly or asymptotically the minimax risk for specific parameter sets such as the ℓ_p balls or Besov bodies. Such an evaluation should be accompanied by a description of the (approximately) minimax estimator and their corresponding least favorable priors.

Thus, in Chapter 13, the optimality results for ℓ_p balls are summarized, and the thresholds $\lambda = \sqrt{2\log n(\epsilon/C)^p}$ derived heuristically above are shown in fact to be asymptotically minimax for $\ell_{n,p}(C)$. In particular, thresholding rules are found to be asymptotically optimal among *all* esimators in the limit $n^{-1}(C_n/\epsilon_n)^p \to 0$.

In Chapter 14 these considerations are extended to Besov bodies. A key structural result is that *separable* rules, one for which $\hat{\theta}_i(y)$ depends on y_i alone, can be found which are asymptotically minimax, and the corresponding least favorable priors make individual wavelet coefficients independent. Of course, these estimators and priors depend strongly on the indices α, p, q and C.

Epilogues.
A. Continuous versus discrete ...
B. Some related topics. ...

9.12 Notes

(Remark on $p/((2-p)$ as difference between weak and strong ℓ_p norm minimax risks. Also FDR connections?).

Meyer (1990, Section 6.4) explains that it is not possible to characterize the integer Hölder classes $C^m(\mathbb{R})$ in terms of moduli of wavelet coefficients.

Theorem 9.4 and Remark 9.5 extend to $C^\alpha([0,1])$ with the same proof, so long as the boundary wavelets satisfy the same conditions as ψ.

§6. Diagrams using the $(\alpha, 1/p)$ plane are used by Devore, for example in the survey article on nonlinear approximation DeVore (1998).

Exercises

9.1 *(Quasi-norm properties.)* (a) Give an example of θ and θ' for which

$$\|\theta + \theta'\|_{w\ell_p} > \|\theta\|_{w\ell_p} + \|\theta'\|_{w\ell_p}.$$

(b) Verify that for $a, b \in \mathbb{R}$ and $p > 0$,

$$2^{(1-p)_+}(a^p + b^p) \le (a+b)^p \le 2^{(p-1)_+}(a^p + b^p).$$

9.2 *(Fixed thresholds on weak ℓ_p balls).* Suppose that $y \sim N_n(\theta, \epsilon^2 I)$, and let $c_p = 2/(2-p)$.
(i) Let θ^λ denote soft thresholding at $\lambda\epsilon$. Show that

$$\bar{r}_\epsilon(\hat{\theta}^\lambda, w\ell_{n,p}(C)) = \sup_{\theta \in w\ell_{n,p}(C)} r_\epsilon(\hat{\theta}^\lambda, \theta) \le n\epsilon^2 r_S(\lambda, 0) + c_p(1+\lambda^2)^{1-p/2} C^p \epsilon^{2-p}.$$

This should be compared with bound (9.19) for $\lambda = \epsilon\sqrt{2\log n}$.
(ii) Let C_n, ϵ_n depend on n and define the normalized radius $\eta_n = n^{-1/p}(C_n/\epsilon_n)$. If $\eta_n \to 0$ as $n \to \infty$, set $\lambda_n = \sqrt{2\log \eta_n^{-p}}$ and show that

$$\bar{r}_\epsilon(\hat{\theta}^\lambda, w\ell_{n,p}(C)) \le c_p \cdot n\epsilon_n^2 \cdot \eta_n^p (2\log \eta_n^{-p})^{1-p/2}(1+o(1)).$$

[This turns out to be the minimax risk for weak ℓ_p; compare the corresponding result for strong ℓ_p in (13.43).]

9.3 *(James-Stein and thresholding on a sparse signal.)* Suppose that $X \sim N_n(\mu_n, I)$, let $\hat{\mu}^{JS}$ denote the James-Stein estimator (2.44), and $\hat{\mu}^\lambda$ soft thresholding at λ.
(i) Suppose that $\|\mu_n\|_2^2 \sim \gamma n$ as $n \to \infty$. Show that $r(\hat{\mu}^{JS}, \mu_n) \sim [\gamma/(\gamma+1)]n$.
(ii) Let $\mu_{n,k} = n^{1/2}k^{-1/p}, k = 1, \ldots, n$ be the weak ℓ_p extremal vector, with $0 < p < 2$. Show that with $\lambda_n = \sqrt{(2-p)\log n}$,

$$r(\hat{\mu}^{\lambda_n}, \mu_n) \le c_p n^{p/2}(\log n)^{1-p/2}, \qquad \text{while} \qquad r(\hat{\mu}^{JS}, \mu_n) \sim c_p' n.$$

9.4 *(Hölder smoothness and wavelet coefficients.)* Assume the hypotheses of Remark 9.5 and in particular that smoothness α satisfies $m < \alpha < m+1$ for $m \in \mathbb{N}$. Show that the bounds

$$|\beta_{Lk}| \le C, \qquad |\theta_{jk}| \le C 2^{-(\alpha+1/2)j},$$

imply that

$$|D^m f(x) - D^m f(y)| \le C'|x-y|^{\alpha-m}.$$

9.5 *(Besov norm of a singularity.)* Verify Example 9.8, for example as follows. Let $S(\psi_{jk})$ denote the support of wavelet ψ_{jk}. Establish the bounds

$$|\theta_{jk}| \le \begin{cases} C2^{-j(\beta+1/2)}|k|^{-(r-\beta)} & 0 \notin S(\psi_{jk}) \\ C2^{-j(\beta+1/2)} & 0 \in S(\psi_{jk}), \end{cases}$$

and hence show that $2^{ja}\|\theta_{j\cdot}\|_p \le c2^{j(\alpha-\beta-1/p)}$.

9.6 *(Thresholding at very fine scales.)* We wish to weaken the condition $\alpha \ge 1/p$ in Proposition 15.4 and Theorem 15.5 to $\alpha > 1/p - 1/2$. Instead of setting everything to zero at levels J and higher (compare (15.5)), one possibility for controlling tail bias better is to apply soft thresholding at very high scales at successively higher levels:

$$\hat{\theta}_{jk} = \begin{cases} \delta_S(y_{jk}, \lambda_j\epsilon), & j < J^2 \\ 0 & j \ge J^2 \end{cases}$$

where for $l = 0, 1, \ldots, J-1$,

$$\lambda_j = \sqrt{2(l+1)\log\epsilon^{-2}} \qquad \text{for} \quad lJ \le j < (l+1)J.$$

Show that if, now $\alpha > 1/p - 1/2$, then the upper risk bound in Theorem 15.5 continues to hold with $\log\epsilon^{-2}$ replaced by, say, $(\log\epsilon^{-2})^3$.

10

The optimal recovery approach to thresholding.

We have seen that the fact that the maximum of n independent standard normal variates is usually bounded by $\sqrt{2\log n}$ leads to some attractive properties for threshold estimators which use this relatively high threshold. In this chapter we will see how some quite general conclusions about $\sqrt{2\log n}$ thresholding may be drawn by analyzing a related *optimal recovery* problem with *deterministic* noise.

The plan is to consider a whole class of parameter spaces Θ and *loss functions* $\|\hat{\theta} - \theta\|$ (in contrast with our previous focus mainly on squared error loss). We again establish *near* optimality properties for a single estimator over many settings, rather than an exact optimality result for a single setting which may be dangerously misleading if that setting is not, in fact, the appropriate one.

The setting is the projected white noise model (15.12) with $n = 2^J$ observations, expressed in the wavelet domain as

$$y_I = \theta_I + \epsilon z_I, \qquad I \in \mathcal{I}^J, \tag{10.1}$$

with ϵ known and $z_I \overset{i.i.d}{\sim} N(0,1)$. We consider a soft threshold estimate, with threshold set at $\delta_n = \epsilon\sqrt{2\log n}$:

$$\hat{\theta}_{\delta_n,I}(y) = \begin{cases} \eta_S(y_I;\delta_n) & I \in \mathcal{I}^J \\ 0 & \text{otherwise.} \end{cases} \tag{10.2}$$

This estimator will be seen to have several useful features:

First, it can be easily used in practice. If data is observed that can be reasonably approximated by the finite sample regression model

$$Y_l = f(l/n) + \sigma\tilde{z}_l, \qquad l = 1,\dots n, \tag{10.3}$$

then the estimator (10.2) can be applied to the discrete wavelet transform of $\tilde{y}_i$. [Compare diagram (7.21).] Indeed, the algorithm runs in $O(n)$ time (if the wavelet filters have finite support) and is widely distributed in software. The results of the previous chapter show that properties of (10.2) proved in model (10.1) are at least asymptotically equally valid when the same estimator is applied to the wavelet transform of discrete data (10.3).

Secondly, estimator (10.2) possesses a number of pleasant properties. We have already discussed spatial adaptation, Section 7.5, and adaptation in estimation at a point, Section 9.10. In this chapter, we will focus on

(a) The function estimates $f[\hat{\theta}]$ corresponding to (10.2) are in a strong sense "as smooth

as" f, so that one has, with high probability, a guarantee of not "discovering" non-existent features. (Theorem 10.7)

(b) For a large class of parameter spaces Θ and global error measures $\|\cdot\|$ derived from Besov and Triebel norms, the estimator (10.2) is *simultaneously* near minimax (Theorem 10.10).

Finally, the proofs of the properties (a) and (c) exploit a useful connection with a deterministic problem of optimal recovery, and highlight the key role played by the concept of shrinkage in unconditional bases, of which wavelet bases are a prime example.

(Other remarks.) The modulus of continuity provides a convenient summary describing the rate of convergence corresponding to $\|\cdot\|_{b'}$ and Θ_b.

The device of 'Besov shells' consists in looking at signals θ whose only non-zero components lie in the j-th shell. Focusing on the j-th shell alone reduces the calculations to an ℓ_p ball. By studying the modulus as the shell index j varies, we see again the pattern of geometric decay away from a critical level $j_* = j_*(\mathbf{p})$.

With the expanded scale of loss functions, the shell calculations reveal a new phenomenon—a distinct, and slower, rate of convergence for parameter combinations $\mathbf{p}$ in a 'logarithmic' zone. (The reason for the name appears after the detailed statement of Theorem 10.10.)

While this structure emerges naturally here in the study of $\sqrt{2\log n}$ thresholding, it provides a basic point of reference for studying properties of other threshold selection schemes over the same range of $\mathbf{p}$. For example, this structure is used heavily in Johnstone and Silverman (2005b) to study wavelet shrinkage using an empirical Bayes choice of threshold, introduced in Section 7.6.

10.1 A Deterministic Optimal Recovery Model

Consider the following *deterministic* version of the sequence model. Data $x = (x_I)$ is observed that satisfies

$$x_I = \theta_I + \delta u_I \qquad |u_I| \le 1 \qquad I \in \mathcal{I}. \tag{10.4}$$

It is desired to recover the unknown vector θ, but it is assumed that the deterministic noise u might be chosen maliciously by an opponent, subject only to the uniform size bound. The noise level δ is assumed known. The worst case error suffered by an estimator $\hat{\theta}$ is then

$$e(\hat{\theta}, \theta; \delta) = \sup_{|u_I| \le 1} \|\hat{\theta}(x) - \theta\|. \tag{10.5}$$

We will see that a number of conclusions for the statistical (Gaussian) sequence model can be drawn, after appropriate calibration, from the deterministic model (10.4).

Assumptions on loss function and parameter space. Throughout this chapter we will assume:

(i) $\Theta \subset \ell_2(\mathcal{I})$ is solid and orthosymmetric, and

(ii) The error norm $\|\cdot\|$ is also solid and orthosymmetric, in the sense that

$$|\xi_I| \le |\theta_I| \ \ \forall I \quad \Rightarrow \quad \|\xi\| \le \|\theta\|.$$

The error norm can be convex, as usual, or at least ρ-convex, $0 < \rho \le 1$, in the sense that $\|\theta + \xi\|^\rho \le \|\theta\|^\rho + \|\xi\|^\rho$.

Remark. The literature on optimal recovery goes back to Golomb and Weinberger (1959) and a 1965 Moscow dissertation of Smolyak. See also Micchelli (1975); Micchelli and Rivlin (1977) and Donoho (1994), who makes the connection with statistical estimation. These latter references are concerned with estimation of a linear functional, while here we are concerned with the whole object θ.

The Uniform Shrinkage Property of Soft Thresholding. Soft thresholding at threshold λ can be used in the optimal recovery setting:

$$\hat{\theta}_{\lambda,I}(x_I) = \mathrm{sgn}(x_I)(|x_I| - \lambda)_+.$$

The shrinkage aspect of *soft* thresholding has the simple but important consequence that the estimate remains confined to the parameter space:

Lemma 10.1 *If Θ is solid orthosymmetric and $\lambda \geq \delta$, then $\theta \in \Theta$ implies $\hat{\theta}_\lambda \in \Theta$.*

Proof Since soft thresholding shrinks each data coordinate x_I towards 0 (but not past 0!) by an amount λ that is greater than the largest possible noise value δ that could be used to expand θ_I in generating x_I, it is clear that $|\hat{\theta}_{\lambda,I}| \leq |\theta_I|$. Since Θ is solid orthosymmetric, this implies $\hat{\theta}_\lambda \in \Theta$. □

Minimax Error. The minimax error of recovery in the determinstic model is

$$E(\Theta, \delta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} e(\hat{\theta}, \theta),$$

where $e(\hat{\theta}, \theta)$ is given by (10.5). Good bounds on this minimax error can be found in terms of a *modulus of continuity* defined by

$$\Omega(\delta) = \Omega(\delta; \Theta, \|\cdot\|) = \sup_{(\theta_0,\theta_1)\in\Theta\times\Theta} \{\|\theta_0 - \theta_1\| : \|\theta_0 - \theta_1\|_\infty \leq \delta\}. \tag{10.6}$$

Theorem 10.2 *Suppose that Θ and the error norm $\|\cdot\|$ are solid and orthosymmetric. Then*

$$(1/2)\Omega(\delta) \leq E(\Theta, \delta) \leq \Omega(2\delta).$$

In addition, soft thresholding $\hat{\theta}_\delta$ is near minimax simultaneously for all such parameter spaces and error norms.

Proof For each noise vector $u = (u_I)$ under model (10.4), and $\theta \in \Theta$, we have $\hat{\theta}_\delta \in \Theta$ by the uniform shrinkage property. In addition, for each u,

$$\|\hat{\theta}_\delta - \theta\|_\infty \leq \|\hat{\theta}_\delta - x\|_\infty + \|x - \theta\|_\infty \leq 2\delta.$$

Hence $(\hat{\theta}_\delta, \theta)$ is a feasible pair for the modulus, and so it follows from the definition that $e(\hat{\theta}_\delta, \theta) \leq \Omega(2\delta)$.

Turning now to a *lower bound*, suppose that the pair $(\theta_0, \theta_1) \in \Theta \times \Theta$ attains the value $\Omega(\delta)$ defining the modulus [1] The data sequence $x = \theta_1$ is potentially observable under (10.4) if either $\theta = \theta_0$ or $\theta = \theta_1$, and so for any estimator $\hat{\theta}$,

$$\sup_{\theta\in\Theta} e(\hat{\theta}, \theta) \geq \sup_{\theta\in\{\theta_0,\theta_1\}} \|\hat{\theta}(\theta_1) - \theta\| \geq \Omega(\delta)/2. \qquad \square$$

[1] If the supremum in (10.6) is not attained, the argument above can be repeated for an approximating sequence.

We now define a modified modulus of continuity which is more convenient for calculations with ℓ_p and Besov norm balls.

$$\Omega^\circ(\delta; \|\cdot\|, \Theta) = \sup\{\|\theta\| : \theta \in \Theta, \|\theta\|_\infty \le \delta\}.$$

If Θ is a norm ball $\Theta(C) = \{\theta : \|\theta\| \le C\}$ (so that $0 \in \Theta$), and if $\|\cdot\|$ is ρ-convex, then it follows easily that

$$\Omega^\circ(\delta) \le \Omega(\delta) \le 2^{1/\rho}\Omega^\circ(2^{-1/\rho}\delta). \tag{10.7}$$

10.2 Monoresolution model: upper bounds

In the deterministic model of optimal recovery, Theorem 10.2 is a strong statement of the near optimality of soft thresholding over a range of parameter spaces and error norms, phrased in terms of the modulus of continuity $\Omega(\delta)$.

Consider now a monoresolution Gaussian error model

$$y_i = \theta_i + \epsilon z_i \qquad z_i \overset{i.i.d.}{\sim} N(0,1), \;\; i = 1, \dots, n. \tag{10.8}$$

The connection with the optimal recovery model (with $\mathcal{I} = \{1, \dots, n\}$ is made by considering the event

$$A_n = \{\sup_{I \in \mathcal{I}} |z_I| \le \sqrt{2\log n}\}, \tag{10.9}$$

which because of the properties of maxima of i.i.d. Gaussians (c.f. Section 8.10) has probability approaching one:

$$P(A_n) = \pi_n \ge 1 - 1/\sqrt{\pi \log n} \nearrow 1 \qquad \text{as } n \to \infty.$$

The key idea is to apply results from the optimal recovery model with deterministic noise level $\delta_n = \epsilon\sqrt{2\log n}$ *on the set* A_n. Thus, in the statistical model we consider the soft thresholding estimator $\hat{\theta}_n$ of (10.2) at level $\epsilon\sqrt{2\log n}$. We therefore obtain immediately

Proposition 10.3 *Consider the Gaussian model* (10.8) *with n observations. If* $(\Theta, \|\cdot\|)$ *is solid, orthosymmetric and* Θ *is convex, then*

$$\sup_{\theta \in \Theta} P\{\|\hat{\theta}_n - \theta\| \le 2\Omega(\epsilon\sqrt{2\log n})\} \ge \pi_n \to 1.$$

In the next two sections, we explore the implications for estimation over ℓ_p-balls in $\mathbb{R}^n$ using error measured in $\ell_{p'}$ norms. We need first to evaluate the modulus Ω for this class of Θ and $\|\cdot\|$, and then to investigate lower bounds to match the upper bounds just proved.

10.3 Modulus of continuity for ℓ_p balls

In the definition of the modulus $\Omega(\delta)$, we take $\Theta = \Theta_{n,p}(C) = \{\theta \in \mathbb{R}^n : \sum_1^n |\theta_i|^p \le C^p\}$ and $\|\cdot\|$ equal to the norm of $\ell_{p',n}$ for $1 \le p' < \infty$. While the leading case is perhaps $p' = 2$, the method works equally well for more general p'. We introduce a new notation

$$\begin{aligned} W_{n;p',p}(\delta, C) &= \Omega^\circ(\delta;\ \Theta_{n,p}(C), \|\cdot\|_{p'}) \\ &= \sup\{\|\theta\|_{p'} \ :\ \|\theta\|_\infty \le \delta, \|\theta\|_p \le C\}. \end{aligned}$$

Usually we write more simply just $W_n(\delta, C)$. Equivalently,

$$W_n^{p'}(\delta, C) = \sup\{\sum_1^n \theta_i^{p'} \wedge \delta^{p'} \ : \ \sum_1^n |\theta_i|^p \le C^p\}.$$

We show that

$$W_n^{p'}(\delta, C) \doteq n_0 \delta_0^{p'},$$

with the *least favorable* configurations being given (up to permutations and sign changes) by

$$\theta^* = (\delta_0, \dots, \delta_0, 0, \dots, 0), \qquad \delta_0 \le \delta, \tag{10.10}$$

with n_0 non-zero coordinates and $1 \le n_0 \le n$. The explicit values of (n_0, δ_0) are shown in the diagrams below.

The verification is mostly by picture—compare Figure 10.1. First, however, set $x_i = |\theta_i|^p$, so that we may rewrite

$$W^{p'} = \sup\{\sum x_i^{p'/p} \ : \ \sum x_i \le C^p, \|x\|_\infty \le \delta^p\}.$$

The function $f(x) = \sum x_i^{p'/p}$ is concave for $p' \le p$ and strictly convex for $p' > p$, in both cases over a convex constraint set. We take the two cases in turn.

(i) $p \ge p'$. Let $\bar{x} = \text{ave } x_i$, and $\tilde{x} = (\bar{x}, \dots, \bar{x})$. By concavity, $f(\tilde{x}) \ge f(x)$, and so the maximum of f occurs at some vector $c(1, \dots, 1)$.

(ii) $p < p'$. Convexity implies that the maximum occurs at extreme points of the constraint set. For example, if $Cn^{-1/p} \le \delta \le C$, then

$$\theta^* = (\delta, \dots, \delta, \eta, 0, \dots, 0), \qquad \text{with } n_0\delta^p + \eta^p = C^p.$$

Hence $W_n^{p'}(\delta) = n_0\delta_0^{p'} + \eta^{p'}$ with $\delta_0 = \delta$, and we have $W_n^{p'}(\delta, C) \approx C^p\delta^{p'-p}$, or more precisely

$$\tfrac{1}{2}C^p\delta^{p'-p} \le W_n^{p'}(\delta, C) \le 2C^p\delta^{p'-p}.$$

Indeed, using the equation $n_0\delta^p + \eta^p = C^p$ with $n_0 \ge 1$, we find

$$\frac{W^{p'}}{C^p\delta^{p'-p}} = \frac{n_0\delta^{p'} + \eta^{p'}}{(n_0\delta^p + \eta^p)\delta^{p'-p}} \in \left[\frac{n_0}{n_0+1}, \frac{n_0+1}{n_0}\right] \in [\tfrac{1}{2}, 2].$$

Thus n_0 (or the ratio n_0/n) measures the *sparsity* of the least favorable configuration. When $p > 2$, the least favorable configurations are *always* dense, since the contours of the ℓ_2 loss touch those of the ℓ_p norm along the direction $(1, \dots, 1)$. On the other hand, when $p < 2$, the maximum value of ℓ_2 error over the intersection of the ℓ_p ball and $\delta-$cube is always attained on the boundary of the cube, which leads to sparser configurations when $C < \delta n^{1/p}$.

For later use, note the special case when there is no constraint on $\|\theta\|_\infty$:

$$W_{n;p',p}(\infty, C) = \sup\{\|\theta\|_{p'} : \|\theta\|_p \le C\} = n^{(1/p'-1/p)_+}C. \tag{10.11}$$

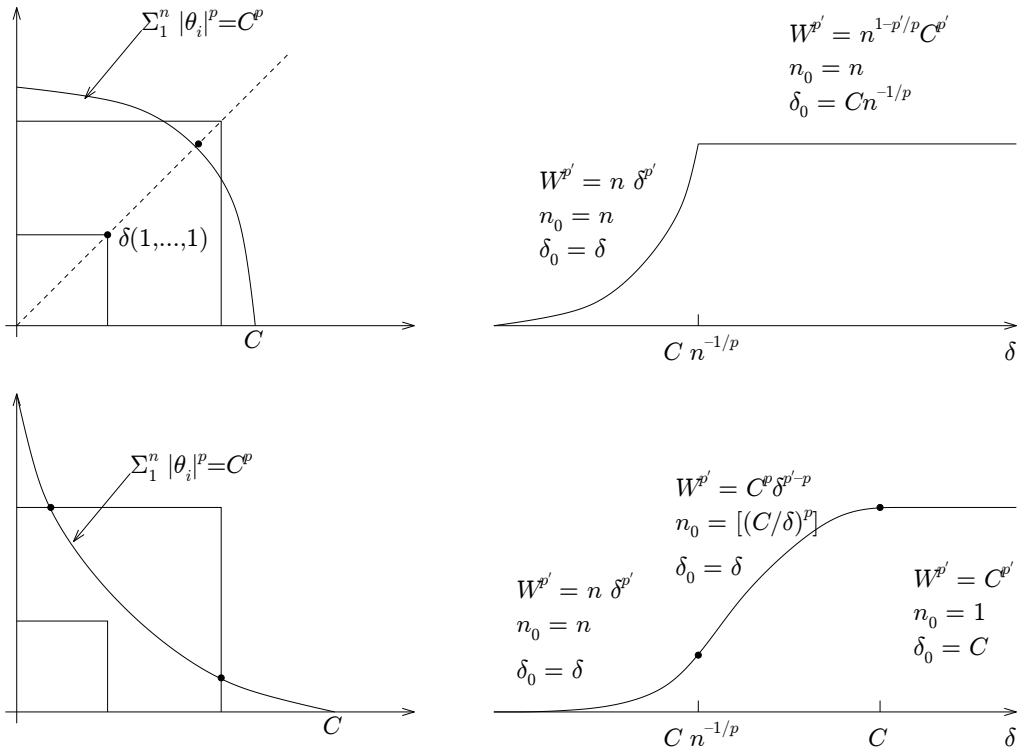

Figure 10.1 Top panel: Concave case $p \geq p'$, Bottom panel: Convex case $p < p'$

10.4 Lower Bounds for ℓ_p balls

In the statistical problem, one does not have an overtly malicious opponent choosing the noise, which suggests that statistical estimation might not be as hard as optimal recovery. However, a statistical lower bound argument, based on hypercubes, will show that in fact this is not true, and that in many cases, the modulus yields (up to logarithmic factors), a description of the difficulty of the statistical problem as well.

For now, we restrict to parameter spaces which are ℓ_p balls: $\Theta_{n,p}(C) = \{\theta \in \mathbb{R}^n : \sum_1^n |\theta_i|^p \leq C^p\}$. In stating lower bounds for the statistical model over ℓ_p balls, we need to recall the structure of extremal configurations for the modulus $\Omega(\epsilon) = \Omega(\epsilon; \Theta_{n,p}(C), \|\cdot\|_2)$. Indeed, let $n_0 = n_0(p, C, \epsilon, n)$ be the number of non-zero components in the extremal vectors θ_{n_0,δ_0} of (10.10). We develop two bounds, for the dense (n_0 large) and sparse ($n_0 = 1$) cases respectively.

Proposition 10.4 *Assume data is taken from model* (10.8).

(i) (Dense case). Let $n_0 = n_0(p, n, C, \epsilon)$ be the number of components of size δ_0 in the least favorable configuration for $\Theta_{n,p}(C)$. Let $\pi_0 = \Phi(-1)/2$. Then

$$\inf_{\hat{\theta}} \sup_{\Theta_{n,p}(C)} P\{\|\hat{\theta} - \theta\|_{p'} \geq (\pi_0/2)^{1/p'} W_n(\epsilon, C)\} \geq 1 - e^{-2n_0\pi_0^2}. \tag{10.12}$$

(ii) (Sparse case). Fix $\eta > 0$ small. There exist functions $\pi_\eta(n) \to 1$ as $n \to \infty$ such that

for any $\delta_n \le \epsilon\sqrt{(2-\eta)\log n}$, *then, as* $n \to \infty$,

$$\inf_{\hat\theta} \sup_{\Theta_{n,p}(\delta_n} P\{\|\hat\theta - \theta\| \ge \tfrac{1}{2}\delta_n\} \ge \pi_\eta(n). \tag{10.13}$$

Remarks. 1. Now imagine a sequence of problems indexed by n with $C = C_n$ and $\epsilon = \epsilon_n$. In the dense case, $p \ge p'$, we always have $n_0 = n$, compare Figure 10.1. Again from the figure, in the sparse case $p < p'$, now $n_0 \to \infty$ so long as $C_n/\epsilon_n \to \infty$. The improved lower bound of part (ii) applies so long as $C_n/\epsilon_n \le \sqrt{(2-\eta)\log n}$.

2. Thus, in the statistical model, an upper bound for estimation over $\Theta_{n,p}(C)$ is given, on a set of high probability, by $\Omega(\epsilon_n\sqrt{2\log n})$, whereas the lower bound in the dense case (10.12) is of order $\Omega(\epsilon_n)$. Thus there is a logarithmic gap between the two bounds. However, the near optimality of $\sqrt{2\log n}$ soft thresholding holds quite generally: the method of proof works for all $\ell_{p'}$ losses, and over all ℓ_p balls $\Theta_{n,p}(C)$.

3. In the sparse case, $p < p'$, the result is just a slight rephrasing of Proposition 8.12. One can rewrite the lower bound in terms of the modulus Ω by setting $c_\eta = (1-\eta/2)^{1/2}$. Then $\delta_n/2 = (c_\eta/2)\Omega(\epsilon_n\sqrt{2\log n})$. Thus in the sparse case the logarithmic term appears in the lower bound also, so that there are cases in which the optimal recovery method yields exact rate results in the statistical model.

Proof *Dense Case.* The argument uses a version of the hypercube method. Let (n_0, δ_0) be parameters of the worst case configuration for $W_n(\epsilon, C)$: from the figures

$$\delta_0 = \begin{cases} \min\{\epsilon, Cn^{-1/p}\} & \text{if } p \ge p' \\ \min\{\epsilon, C\} & \text{if } p < p'. \end{cases}$$

from which it is clear that $\delta_0 \le \epsilon$. Let π be the distribution on θ which makes θ_i independently equal to $\pm\delta_0$ with probability $\frac{1}{2}$ for $i = 1, \dots, n_0$, and all other co-ordinates 0. Since supp $\pi \subset \Theta$, we have for any $(\theta, y)-$measurable event A,

$$\sup_{\theta\in\Theta} P_\theta(A) \ge P_\pi(A). \tag{10.14}$$

Suppose now that $\hat\theta(y)$ is an arbitrary estimator and let $N(\hat\theta(y), \theta) = \sum_i I\{\hat\theta_i(y)\theta_i < 0\}$ be the number of sign errors made by $\hat\theta$, where the sum is over the first n_0 coordinates. Under P_π,

$$\|\hat\theta - \theta\|_{p'}^{p'} \ge \delta_0^{p'} N(\hat\theta(y), \theta). \tag{10.15}$$

Combining (10.14) and (10.15), we conclude that

$$\sup_{\theta\in\Theta} P_\theta\{\|\hat\theta - \theta\|_{p'}^{p'} \ge c\delta_0^{p'}\} \ge P_\pi\{N(\hat\theta, \theta) \ge c\}.$$

It was shown in Section 4.7 that the right side probability is minimized over $\hat\theta$ by the rule $\hat\theta_{\pi,i}(y) = \delta_0 \mathrm{sgn}(y_i)$ and that since $N(\hat\theta_\pi, \theta)$ counts sign errors in the data, the minimizing probability is a binomial tail probability. Hence

$$S(c) = \inf_{\hat\theta} \sup_{\Theta} P_\theta\{\|\hat\theta - \theta\| \ge c\delta_0^{p'}\} \ge P\{\mathrm{Bin}(n_0, \pi_1) \ge c\},$$

where $\pi_1 = P_\pi\{y_1\theta_1 < 0\} = P\{\delta_0 + \epsilon z < 0\} = \Phi(-\delta_0/\epsilon) \geq 2\pi_0$. We recall the Cramér-Chernoff large deviations principle[2] : if $\pi_1 > \pi_0$, then

$$P\{\text{Bin}\,(n_0, \pi_1) < n_0\pi_0\} \leq e^{-n_0 D(\pi_0,\pi_1)},$$

where $D(\pi_0, \pi_1) = K(\text{Be}(\pi_0), \text{Be}(\pi_1)) = \pi_0 \log(\pi_0/\pi_1) + \bar{\pi}_0 \log(\bar{\pi}_0/\bar{\pi}_1)$. Here $D(\pi_0, \pi_1)$ denotes the Kullback-Leibler divergence between two Bernoulli distributions, and $\bar{\pi}_i = 1 - \pi_i$. Noting also[3] that $D(\pi_0, \pi_1) \geq 2(\pi_1 - \pi_0)^2$, we conclude that

$$1 - S(n_0\pi_0) \leq e^{-n_0 D(\pi_0,\pi_1)} \leq e^{-2n_0\pi_0^2},$$

and since $n_0\delta_0^{p'} \geq (1/2)W_n^{p'}(\epsilon, C)$, this establishes (10.12). □

10.5 Multiresolution model: preservation of smoothness

An *unconditional basis* $\{\psi_I\}$ for a Banach space B can be defined by two properties: (i)*(Schauder basis)* $\forall v \in B, \exists$ unique sequence $\{\theta_I\} \subset \mathbb{C}$ such that $v = \sum_1^\infty \theta_I\psi_I$, and (ii) *(Multipliers)* $\exists C$ s.t. $\forall N$,and sequences $\{m_I\} \subset \mathbb{C}$ with $|m_I| \leq 1$

$$\|\sum_1^N m_I\theta_I\psi_I\| \leq C\|\sum_1^N \theta_I\psi_I\|. \tag{10.16}$$

Several equivalent forms and interpretations of the definition are given by Meyer (1990, I, Ch. VI). Here we note only that (10.16) says that shrinkage of coefficients can not grossly inflate the norm in unconditional bases. This suggests that traditional statistical shrinkage operations - usually introduced for smoothing or stabilization purposes - are best performed in unconditional bases.

A key consequence of the sequence norm characterisation results described in Section 9.6 is that *wavelets form unconditional bases for the Besov and Triebel scales of function spaces*. Indeed, when viewed in terms of the sequence norms

$$\|f\|^q_{\dot{B}^\alpha_{p,q}} \asymp \sum_j (2^{sj}\sum_k |\alpha_{jk}|^p)^{q/p},$$

the multiplier property is trivially satisfied, since $\|f\|$ depends on α_{jk} *only through* $|\alpha_{jk}|$. Donoho (1993, 1996) has shown that unconditional bases are in a certain sense optimally suited for compression and statistical estimation.

Example 10.5 Suppose that the orthonormal wavelet ψ is C^R and has D vanishing moments. Consider a *scale* of functional spaces

$$\mathcal{C}(R, D) = \{B^\alpha_{p,q}[0,1], F^\alpha_{p,q}[0,1] : 1/p < \alpha < \min(R, D)\}. \tag{10.17}$$

These spaces are (i) all embedded in $C[0,1]$ (since $\alpha > 1/p$), and (ii) the wavelet system $\{\psi_{jk}\}$ forms an unconditional basis for each of the spaces in the scale (since $\alpha < \min(R, D)$).

Example 10.6 *Preservation of Smoothness.* Suppose now that $\{\psi_I\}$ is an unconditional basis for a function space $\mathcal{F}$ with norm $\|\cdot\|_{\mathcal{F}}$. Data from model (10.4) can be used to construct an estimator of $f = \sum \theta_I \psi_I$ by setting $\hat{f} = \sum \hat{\theta}_{\lambda,I} \psi_I$. The uniform shrinkage property combined with the multiplier property (10.16) implies that whatever be the noise u,

$$\|\hat{f}\|_{\mathcal{F}} \le C \|f\|_{\mathcal{F}}.$$

This means that one can assert that $\hat{f}$ is *as smooth as* f. In particular, if f is identically 0, then so is $\hat{f}$! Furthermore, for a C^R wavelet ψ with D vanishing moments, this property holds *simultaneously* for all spaces $\mathcal{F}$ in the scale $\mathcal{C}(R, D)$ of (10.17).

Preservation of Smoothness

As a first illustration, consider the smoothness preservation property of Example 10.6. On the event A_n of (10.9), the uniform shrinkage property Lemma 10.1 implies that $\hat{\theta}_{\delta_n} \in \Theta$ whenever $\theta \in \Theta$. Hence, for function spaces in the scale $\mathcal{C}(R, D)$, we have on A_n that $\|\hat{f}_n\|_{\mathcal{F}} \le C(\mathcal{F})\|f\|_{\mathcal{F}}$. Hence

Theorem 10.7 *For each function space $\mathcal{F} \in \mathcal{C}(R, D)$ there exists a constant $C(\mathcal{F})$ such that*

$$P\{\|\hat{f}_n\|_{\mathcal{F}} \le C(\mathcal{F})\|f\|_{\mathcal{F}} \ \ \forall \mathcal{F} \in \mathcal{C}\} \ge \pi_n \to 1.$$

Thus, one can assert that with high probability, the estimator $\hat{f}_n$ is as smooth as the "truth" f simultaneously over many smoothness classes. In particular, if $f \equiv 0$, then $\hat{f}_n \equiv 0$ with probability at least π_n so that one can assert that $\hat{f}_n$ does not find "spurious structure".

10.6 Statistical Upper and Lower Bounds

For much the same reasons as in Section 15.2, we will also need to consider a *projected data* version of the optimal recovery model in which

$$x_I = \theta_I + \delta u_I \qquad I \in \mathcal{I}_{(n)}, \qquad |\mathcal{I}_{(n)}| = n.$$

Again, one still attempts to recover the entire object θ, and the corresponding minimax recovery error is

$$E(\Theta, \delta; n) = \inf_{\hat{\theta}(x^{(n)})} \sup_{\Theta} e(\hat{\theta}(x^{(n)}), \theta).$$

Projection onto the $n-$data model is defined by

$$(P_n\theta)_I = \begin{cases} \theta_I & I \in \mathcal{I}_{(n)} \\ 0 & \text{otherwise.} \end{cases}$$

Even when the noise level $\delta = 0$, there is still an error of recovery due to the attempt to infer the full vector θ from only n components. Hence we make the

Definition. The *tail $n-$width* of Θ in norm $\|\cdot\|$ is

$$\Delta(n, \Theta, \|\cdot\|) = \sup_{\theta \in \Theta}\{\|\theta\| : P_n\theta = 0\} = E(\Theta, 0; n). \tag{10.18}$$

It is then straightforward to establish the following finite data analog of Theorem 10.2.

Proposition 10.8 *Suppose that Θ is solid, orthosymmetric and convex, and that the error norm $\|\cdot\|$ is solid and orthosymmetric. Then*

$$\max\{\frac{\Omega(\delta)}{2}, \Delta(n)\} \le E(\Theta, \delta; n) \le 2\Omega(\delta) + \Delta(n).$$

In addition, soft thresholding $\hat{\theta}_\delta$ is near minimax simultaneously for all such parameter spaces and error norms.

Global Estimation Bounds

In a similar manner, we can immediately convert the upper-bound part of Proposition 10.8 to a statement in the projected Gaussian model with $\delta_n = \epsilon_n \sqrt{2 \log n}$: for the soft threshold estimator $\hat{\theta}_{\delta_n}$, we have for all solid, orthosymmetric (and convex) Θ that

$$\sup_\Theta P\{\|\hat{\theta}_{\delta_n} - \theta\| \le 2\Omega(\delta) + \Delta(n)\} \ge \pi_n \to 1.$$

Thus the statistical model is not harder than the optimal recovery model, up to factors involving $\sqrt{\log n}$. We may say, using the language of Stone (1980), that $2\Omega(\delta) + \Delta(n)$ is an achievable rate of convergence for all qualifying $(\Theta, \|\cdot\|)$.

Now specialize to the case of parameter space Θ and error norm $\|\cdot\|$ taken from the Besov scale.

We first summarize the results of calculation of the Besov modulus and bounds for the tail bias, the details being deferred to the next section.

An interesting feature is the appearance of distinct zones of parameters $\mathbf{p} = (\alpha, p, q, \alpha', p', q')$:

Regular	$\mathcal{R} = \{p' \le p\}$	$\cup$	$\{p' > p, (\alpha + 1/2)p > (\alpha' + 1/2)p'\}$
Logarithmic	$\mathcal{L} =$		$\{p' > p, (\alpha + 1/2)p < (\alpha' + 1/2)p'\}$

In the "critical case" $(\alpha + 1/2)p = (\alpha' + 1/2)p'$, the behavior is more complicated and is discussed in Donoho et al. (1997).

Theorem 10.9 *Let $\Theta = \Theta^\alpha_{p,q}(C)$ and $\|\cdot\| = \|\cdot\|_{b^{\alpha'}_{p',q'}}$. Assume that*

$$\tilde{\alpha} = \alpha - \alpha' - (1/p - 1/p')_+ > 0.$$

(a) Then the modulus $\Omega(\delta; \Theta, \|\cdot\|)$ given by (10.6) *satisfies*

$$\Omega(\delta) \asymp C^{1-r}\delta^r \qquad \text{as } \delta \to 0. \tag{10.19}$$

where the rate exponent is given by $r =$

$$r_R = \frac{(\alpha - \alpha')}{\alpha + 1/2}, \qquad \text{for } \mathbf{p} \in \mathcal{R}$$

$$r_L = \frac{\tilde{\alpha}}{\alpha + 1/2 - 1/p}, \qquad \text{for } \mathbf{p} \in \mathcal{L}$$

(b) the tail bias satisfies, with $c_2 = (1 - 2^{-\tilde{\alpha}q'})^{-1/q'}$,

$$\Delta(n) \le c_2 C n^{-\tilde{\alpha}}. \tag{10.20}$$

If in addition $\alpha > 1/p$, *then* $\Delta(n) = o(\Omega(n^{-1/2}))$.

Part (b) shows that the condition $\tilde{\alpha} > 0$ is needed for the tail bias to vanish with increasing n; we refer to it as a consistency condition. In particular, it forces $\alpha' < \alpha$. In the logarithmic zone, the rate of convergence is reduced, some simple algebra shows that for $\mathbf{p} \in \mathcal{L}$ we have $r_L < r_R$.

Some understanding of the regular and logarithmic zones comes from the smoothness parameter plots introduced in Chapter 9.6. For given values of the error norm parameters α' and p', Figure 10.2 shows corresponding regions in the $(1/p, \alpha)$ plane. The regular/logarithmic boundary is given by the solid line $\alpha = \omega/p - 1/2$ having slope $\omega = (\alpha' + 1/2)p'$. The consistency boundary corresponding to condition $\alpha > \alpha' + (1/p - 1/p')+$ is given by the broken line with inflection at $(1/p', \alpha')$. Note that the two lines in fact intersect exactly at $(1/p', \alpha')$.

If $\omega > 1$, or what is the same, if $a' = \alpha' + 1/2 - 1/p' > 0$, then there is a logarithmic zone. In this case, the consistency boundary lies wholly above the continuity boundary $\alpha = 1/p$, so the condition $\alpha > 1/p$ imposes no additional constraint.

On the other hand, if $\omega \leq 1$ or $a' \leq 0$, the zone boundary line is tangent to the consistency line and there is no logarithmic zone. This explains the why there is no logarithmic zone for traditional squared error loss, corresponding to $\alpha' = 0$, $p' = 2$. In this case the continuity boundary $\alpha = 1/p$ implies a further constraint to ensure negligibility of the tail bias.

As particular examples, on might contrast the error measure $\int |D^2 f|$, with $\alpha' = 2$, $p' = 1$ and $\omega = 5/2$, which has a logarithmic zone, with the measure $\int (Df)^2$, with $\alpha' = 1$, $p' = 2$ and $\omega = 3/4$, which does not.

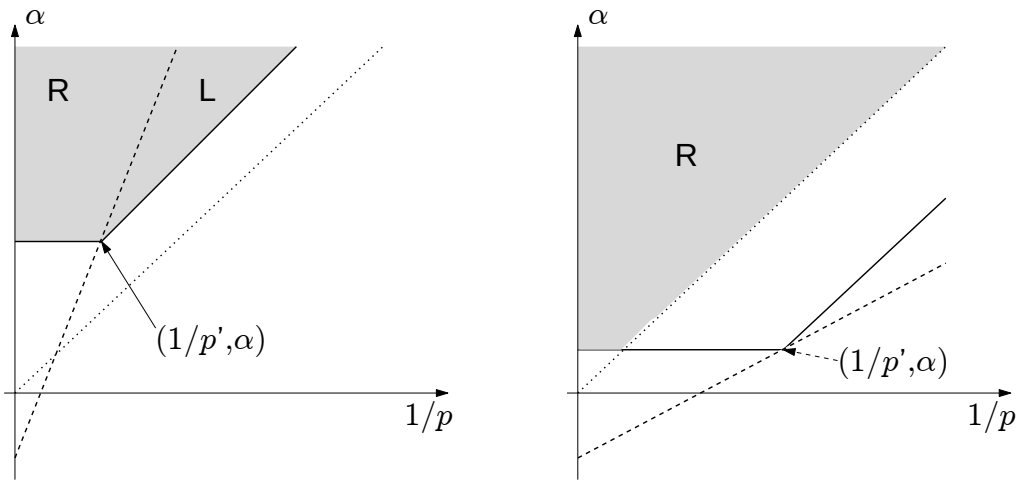

Figure 10.2 Schematic representation of regular $\mathcal{R}$ and logarithmic $\mathcal{L}$ zones in in two cases: left panel when $\omega = (\alpha' + 1/2)p' > 1$, and right panel with $\omega < 1$ and no logarithmic zone. In both cases, solid line is consistency boundary $\alpha = \alpha' + (1/p - 1/p')_+$, dashed line is the regular/logarithmic boundary $\alpha = \omega/p - 1/2$ and dotted line is the continuity boundary $\alpha = 1/p$.

Make the normalization $\epsilon = n^{-1/2}$. Using the bounds derived for the Besov modulus and for the tail bias in Theorem 10.9 we obtain

Theorem 10.10 *Let $\Theta = \Theta^{\alpha}_{p,q}(C)$ and $\|\cdot\| = \|\cdot\|_{b^{\alpha'}_{p',q'}}$. Assume that $\tilde{\alpha} = \alpha - \alpha' - (1/p - 1/p')_+ > 0$ and that $\alpha > 1/p$. Then*

$$\sup_{\theta \in \Theta(C)} P\{\|\hat{\theta}_{\delta_n} - \theta\| \le c\Omega(n^{-1/2}\sqrt{\log n})\} \ge \pi_n \to 1.$$

There exists a constant $c = c(\boldsymbol{p})$ such that

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| \ge c\Omega(n^{-1/2})\} \to 1. \tag{10.21}$$

In the logarithmic case, the lower bound can be strengthened to $\Omega(n^{-1/2}\sqrt{\log n})$.

Thus, soft thresholding at $\delta_n = \epsilon_n\sqrt{2\log n}$ is simultaneously nearly minimax (up to a logarithmic term) over all parameter spaces and loss functions in the (seven parameter) scale $\mathcal{C}(R, D)$, and indeed attains the optimal rate of convergence in the logarithmic case.

To appreciate the significance of adaptive estimation results such as this, note that an estimator that is exactly optimal for one pair $(\Theta, \|\cdot\|)$ may well have very poor properties for other pairs: one need only imagine taking a linear estimator (e.g. from Pinsker's theorem) that would be optimal for an ellipsoid $\Theta^{\alpha}_{2,2}$ and using it on another space $\Theta^{\alpha}_{p,q}$ with $p < 2$ in which linear estimators are known (e.g. Chapter 9.9) to have suboptimal rates of convergence.

10.7 Besov Modulus and Tail Bias

In this section we evaluate the asymptotic order of the modulus of continuity $\Omega(\delta)$ when both parameter space $\Theta^{\alpha}_{p,q}$ and error measure $\|\cdot\|_{b^{\alpha'}_{p',q'}}$ are taken from the Besov scale. The approach is to reduce the optimisation defining the modulus to a hardest resolution level j, where one is effectively dealing with scaled versions ℓ_p norms in both the error measure and in the reduced parameter space.

First define the *Besov shells*

$$\Theta^{(j)} = \{\theta \in \Theta : \theta_I = 0, I \notin \mathcal{I}_j\}.$$

If $\theta^{(j)}$ is derived from θ by setting to zero all components θ_I with $I \notin \mathcal{I}_j$, then

$$\|\theta^{(j)}\|_{b^{\alpha}_{p,q}} = 2^{aj}\|\theta_{j\cdot}\|_p. \tag{10.22}$$

This shows that $\Theta^{(j)}$ is isomorphic to a scaled ℓ_p-ball: $\Theta^{(j)} \cong \Theta_{2^j,p}(C2^{-aj})$. The modulus of continuity, when restricted to the jth shell, reduces in turn to a scaled form of the ℓ_p-modulus:

$$\begin{aligned}\Omega_j(\delta) &:= \Omega^{\circ}(\delta; \Theta^{(j)}, \|\cdot\|) \\ &= 2^{a'j} W_{2^j}(\delta, C2^{-aj}) = W_{2^j}(2^{a'j}\delta, C2^{-(a-a')j}),\end{aligned} \tag{10.23}$$

where we have used the invariance $bW_n(\delta, C) = W_n(b\delta, bC)$. It is easy to verify that nothing essential (at the level of rates of convergence) is lost by considering the shell moduli:

with $\rho = q' \wedge 1$ and $c_\rho = 2^{1/\rho}$,

$$\|\Omega_j(\delta)\|_{\ell_\infty} \le \Omega(\delta) \le c_\rho \|\Omega_j(\delta/c_\rho)\|_{\ell_{q'}}. \tag{10.24}$$

[*Proof of* (10.24). Using the scaling (10.23),

$$\begin{aligned}\Omega^{\circ q} &= \sup\Big\{\sum_j \|\theta^{(j)}\|_{b'}^{q'} : \sum_j \|\theta^{(j)}\|_b^q \le C^q, \|\theta^{(j)}\|_\infty \le \delta\Big\} \\ &\le \sum_j \sup\big\{\|\theta^{(j)}\|_{b'}^{q'} : \|\theta^{(j)}\|_b^q \le C^q, \|\theta^{(j)}\|_\infty \le \delta\big\}\end{aligned}$$

since doing the maximizations separately can only increase the supremum. The second expression is just $\sum_j \Omega_j^{q'}(\delta)$ and so the upper bound follows from (10.7). The lower bound is easier: simply restrict the supremum in the first line to the j-th shell: then $\Omega^\circ(\delta) \ge \Omega_j(\delta)$ for each j.]

In view of (10.23) we can use the ℓ_p-modulus results to compute $\Omega_j(\delta)$ by making the substitutions

$$n_j = 2^j, \qquad \delta_j = 2^{a'j}\delta, \qquad C_j = C2^{-(a-a')j}.$$

'Sparse' case $p < p'$. We use the lower panel of Figure 10.1: as $\delta = \delta_j$ increases, the three zones for W translate into three zones for $j \to \Omega_j$, illustrated in the top panel of Figure 10.3.

Zone (i): $\delta_j < C_j n_j^{-1/p}$. This corresponds to

$$2^{(a+1/p)j} = 2^{(\alpha+1/2)j} < C/\delta,$$

so that the zone (i)/(ii) boundary occurs at j_0 satisfying $2^{(\alpha+1/2)j_0} = C/\delta$. In zone (i),

$$\Omega_j^{p'} = n_j \delta_j^{p'} = \delta^{p'} 2^{(1+p'a')j},$$

and with $n_0 = 2^j$, the maximum possible, this is a 'dense' zone.

At the boundary j_0, on setting $r_0 = (\alpha - \alpha')/(\alpha + 1/2)$, we have

$$\Omega_{j_0} = \delta 2^{j_0(a'+1/p')} = \delta(C/\delta)^{(\alpha'+1/2)/(\alpha+1/2)} = C^{1-r_0}\delta^{r_0}.$$

Zone (ii): $C_j n_j^{-1/p} < \delta_j < C_j$. The right inequality corresponds to $\delta < C2^{-aj}$, so that the zone (ii)/(iii) boundary occurs at j_1 satisfying $2^{aj_1} = C/\delta$. In zone (ii),

$$\Omega_j^{p'} = C_j^p \delta_j^{p'-p} = C^p \delta^{p'-p} 2^{-(pa-p'a')j},$$

and observe using $a = \alpha + 1/2 - 1/p$ etc., that

$$pa - p'a' = p(\alpha + 1/2) - p'(\alpha' + 1/2)$$

is positive in the regular zone and negative in the logarithmic zone, so that Ω_j is (geometrically) decreasing in the regular zone and geometrically increasing in the logarithmic zone. The least favorable configuration has non-zero cardinality

$$n_0 = (C_j/\delta_j)^p = (C/\delta)^p 2^{-paj} = 2^{pa(j_1-j)},$$

decreasing from 2^{j_0} at $j = j_0$ to 1 at $j = j_1$, so this is a zone of increasing sparsity.

Zone (iii): $C_j < \delta_j$. In this sparse zone, $n_0 = 1$ and

$$\Omega_j^{p'} = C_j^{p'} = C^{p'} 2^{-p'(a-a')j},$$

where we note that for $p < p'$,

$$a - a' = \alpha - \alpha' - (1/p - 1/p') = \tilde{\alpha} > 0,$$

by our hypothesis. At the boundary j_1, we have, on setting $r_1 = 1 - a'/a = \tilde{\alpha}/(\alpha + 1/2 - 1/p)$,

$$\Omega_{j_1} = C 2^{-(a-a')j_1} = C(\delta/C)^{(a-a')/a} = C^{1-r_1}\delta^{r_1}.$$

The dense case, $p \geq p'$ is simpler. We refer to the bottom panel of Figure 10.3.

Zone (i) $\delta_j < C_j n_j^{-1/p}$. This zone is the same as in the sparse case, so for $j \leq j_0$ defined by $2^{(\alpha+1/2)j_0} = C/\delta$, we have

$$\Omega_j^{p'} = \delta^{p'} 2^{(1+p'a')j} = \delta^{p'} 2^{(\alpha+1/2)p'j}$$

and at the boundary level j_0, again $\Omega_{j_0} = C^{1-r_0}\delta^{r_0}$ with r_0 as before.

Zone (ii) $C_j n_j^{-1/p} < \delta_j$. We now have

$$\Omega_j^{p'} = n_j^{1-p'/p} C_j^{p'} = C^{p'} 2^{-(\alpha-\alpha')p'j}$$

and $\Omega_j = \Omega_{j_0} 2^{-(\alpha-\alpha')(j-j_0)}$.

Again we see that the geometric decay property (10.25) holds, with $j_* = j_0$ and $r = r_0$, and (as at all levels j) the least favorable configuration at level j_0 is dense, $n_0 = 2^{j_0}$.

To summarize, under the assumptions of the Theorem 10.10, and outside the critical case $(\alpha + 1/2)p = (\alpha' + 1/2)p'$, there exists $j_* \in \mathbb{R}$ and $\kappa = \kappa(\alpha, \alpha', p, p') > 0$ such that

$$\Omega_j(\delta) \leq \delta^r C^{1-r} 2^{-\kappa|j-j_*|}. \tag{10.25}$$

Thus we have geometric decay away from a single critical level. In the regular case, $j_* = j_0$ and $r = r_0$ and the least favorable configuration at level j_0 is dense, $n_0 = 2^{j_0}$. In the logarithmic case, $j_* = j_1$ and $r = r_1$, and the least favorable configuration at level j_1 is sparse, $n_0 = 1$.

The evaluation (10.19) follows from this and (10.24).

Evaluation of Besov tail widths These can be reduced to calculations on Besov shells by the same approach as used to prove (10.25). If we set

$$\Delta_j = \sup\{\|\theta^{(j)}\|_{b'} : \|\theta^{(j)}\|_b \leq C\},$$

then the full tail width is related to these shell widths by

$$\Delta_{J+1} \leq \Delta(2^J, \Theta) \leq \|(\Delta_j)_{j>J}\|_{\ell_{q'}}. \tag{10.26}$$

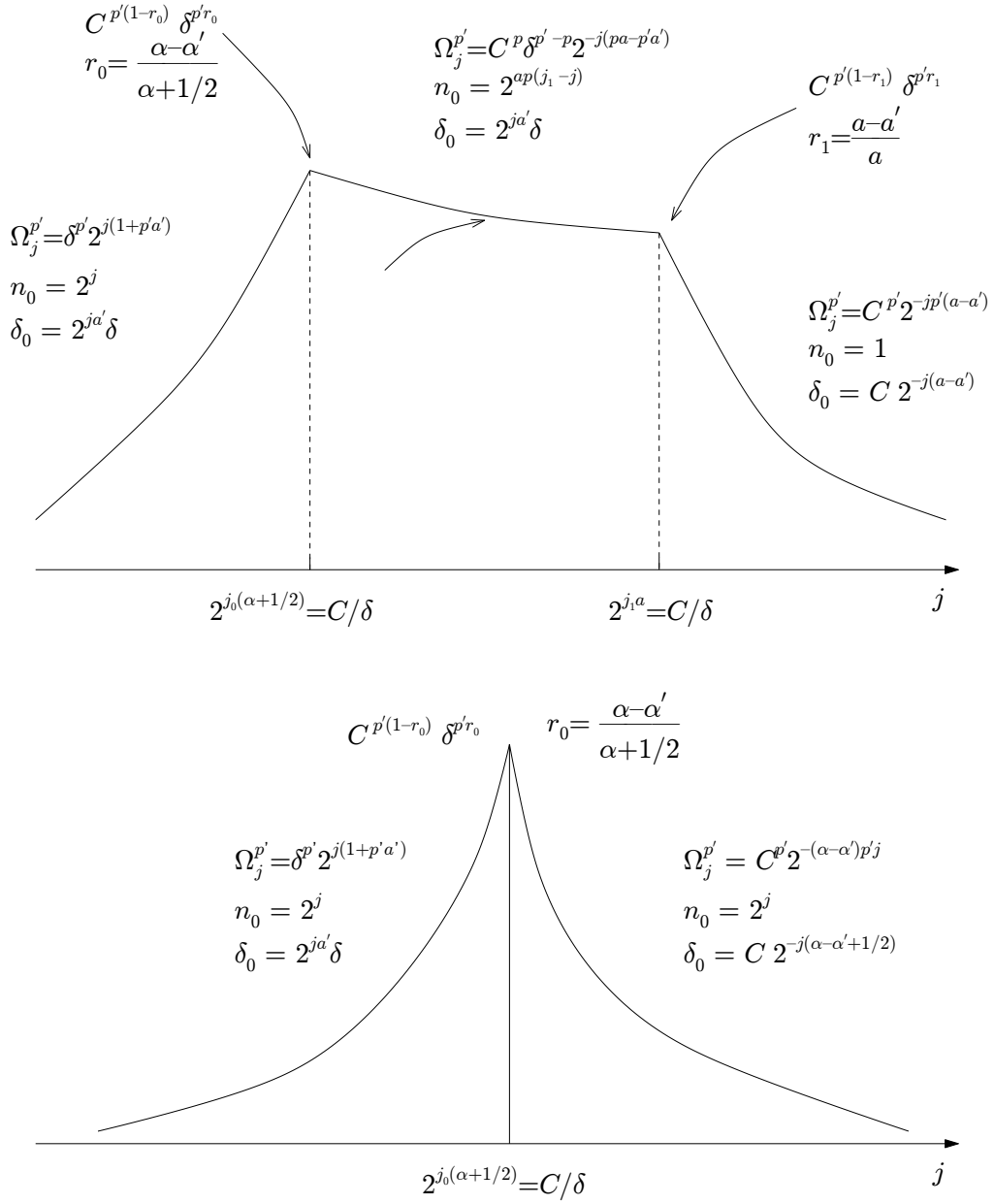

Figure 10.3 Schematic of the Besov modulus $\Omega_j(\delta)$, defined by (10.23), when viewed as a function of level j, with δ, C held fixed. Top panel is 'sparse' case, $p < p'$ (in the regular zone), bottom is 'dense' case $p \geq p'$

Using Besov shell identity (10.22),

$$\begin{aligned}\Delta_j &= 2^{ja'} \sup\{\|\theta_j\|_{p'} : \|\theta_j\|_p \leq C2^{-aj}\} \\ &= 2^{ja'} W_{2^j;p',p}(\infty, C2^{-aj}).\end{aligned}$$

Substituting the identity (10.11), $W_n(\infty, C) = n^{(1/p'-1/p)_+}C$, we find

$$\Delta_j = 2^{ja'}2^{j(1/p'-1/p)_+}C2^{-aj} = C2^{-j\tilde{\alpha}}.$$

In view of (10.26), the full tail bias $\Delta(2^J)$ is equivalent to $\Delta_J = C2^{J\tilde{\alpha}} = Cn^{-\tilde{\alpha}}$.

We now verify that the assumption $\alpha > 1/p$ (continuity) guarantees negligibility of the tail bias term: $\Delta(n) = o(\Omega(n^{-1/2}))$. From (10.20), $\Delta(n) = O(n^{-\tilde{\alpha}})$, while from (10.19), $\Omega(n^{-1/2}) \asymp n^{-r/2}$, so it is enough to verify that $\tilde{\alpha} > r/2$. If $\mathbf{p}$ is in the logarithmic zone, this is immediate when $\alpha > 1/p$.

If $\mathbf{p}$ is in the regular zone, the condition $\tilde{\alpha} > r/2$ becomes $a - \alpha' - (1/p - 1/p')_+ > (\alpha - \alpha')/(2\alpha + 1)$. If $p' \le p$ this is trivial, while for $p' > p$ it is the same as

$$\frac{2\alpha}{2\alpha + 1}(\alpha - \alpha') > (1/p - 1/p').$$

Now the condition for $\mathbf{p}$ to be regular, namely $(2\alpha' + 1)/(2\alpha + 1) < p/p'$, is equivalent to the previous display with the right side replaced by $\alpha p(1/p - 1/p')$. So, again using $\alpha > 1/p$, we are done.

10.8 Lower Bounds

We again use the device of Besov shells to reduce to previous results obtained for ℓ_p balls and their associated least favorable configurations.

For a shell at any level j, we have $\|\theta\|_{B'} \ge \|\theta^{(j)}\|_{B'}$ and also $\Theta^{(j)} \subset \Theta$, and so

$$\sup_{\Theta} P\{\|\hat{\theta} - \theta\|_{B'} \ge \gamma\} \ge \sup_{\Theta^{(j)}} P\{\|\hat{\theta}^{(j)} - \theta^{(j)}\|_{B'} \ge \gamma\}. \tag{10.27}$$

Now since $\|\theta^{(j)}\|_{B'} = 2^{a'j}\|\theta_{j\cdot}\|_{p'}$ and since $\theta^{(j)} \in \Theta^{(j)}$ if and only if $\|\theta_{j\cdot}\|_p \le C2^{-aj}$, the right hand side above equals

$$\sup_{\Theta_{2^j,p}(C2^{-aj})} P\{\|\hat{\theta}_{j\cdot} - \theta_{j\cdot}\|_{p'} \ge \gamma 2^{-a'j}\}. \tag{10.28}$$

Regular case. The Besov shell we use corresponds to the critical level $j_0 = (1/p(\alpha))\log_2(C/\delta)$, where $p(\alpha) = 2/(2\alpha + 1)$ and we set $\delta = \epsilon = n^{-1/2}$. The setting is 'dense' because (cf. top panel of Figure 10.3) there are $n_0 = 2^{j_0}$ non-zero components with size $\delta_0 = 2^{ja'}\epsilon$.

Hence, we apply the dense ℓ_p-ball modulus lower bound, Proposition 10.4, to $\Theta_{2^{j_0},p}(C2^{-aj_0})$. Hence, comparing (10.28) and (10.12), we are led to equate

$$\gamma 2^{-a'j_0} = c_{p'}W_{2^{j_0}}(\epsilon, C2^{-aj_0}),$$

after putting $c_{p'} = (\pi_0/2)^{1/p'}$. Recalling the definition of the shell modulus, (10.23), we get

$$\gamma = c_{p'}\Omega_{j_0}(\epsilon).$$

Because of the geometric decay of the shell modulus away from j_0, compare (10.25), there exists $c = c(\mathbf{p})$ for which

$$\Omega(\epsilon) \le c_1\Omega_{j_0}(\epsilon). \tag{10.29}$$

Combining the prior two displays, we can say that $\gamma \geq c_2\Omega(\epsilon)$ and hence

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\|_{B'} \geq c_2\Omega(\epsilon)\} \geq 1 - e^{-2n_0\pi_0^2}.$$

Here $n_0 = 2^{j_0} = (C/\epsilon)^{p(\alpha)} = (C\sqrt{n})^{p(\alpha)} \to \infty$ as $n \to \infty$, and so the regular case part of Theorem 10.10 is proven.

Logarithmic case. From the modulus calculation, we expect the least favorable configurations to be at shells near j_1 and to be highly sparse, perhaps a single spike. We therefore use the lower bounds derived for the 'bounded single spike' parameter spaces discussed in Section 8.9.

First we note that if $\delta_j \leq C2^{-aj}$, then $\Theta_{2^j}(\delta_j) \subset \Theta_{2^j,p}(C2^{-aj})$. If also $\delta_j \leq \epsilon\sqrt{(2-\eta)\log 2^j}$, then from Proposition (8.50), we can say that

$$\inf_{\hat{\theta}} \sup_{\Theta_{2^j,p}(C2^{-aj})} P\{\|\hat{\theta}_{j\cdot} - \theta_{j\cdot}\|_{p'} \geq \delta_j/2\} \geq \pi_\eta(2^j).$$

Bearing in mind the two conditions on δ_j, it is clear that the largest possible value for δ_j is

$$\bar{\delta}_j = \min\{\epsilon\sqrt{(2-\eta)\log 2^j}, C2^{-aj}\}.$$

The implied best bound in (10.27) that is obtainable using the j-th shell is then given by the solution to $\gamma_j 2^{-a'j} = \bar{\delta}_j/2$, namely

$$\gamma_j = \tfrac{1}{2}2^{a'j}\bar{\delta}_j.$$

Let $j_1 = \max\{j : \epsilon\sqrt{(2-\eta)\log 2^j} \leq C2^{-aj}\}$. It is clear that γ_j is increasing for $j \leq j_1$ and (since $a > a'$) decreasing for $j > j_1$, so our best shell bound will be derived from γ_{j_1}. Since we only observe data for levels $j < \log_2 n = \log_2 \epsilon^{-2}$, we also need to check that $j_1 < \log_2 \epsilon^{-2}$, and this is done below. To facilitate the bounding of γ_{j_1}, we first observe that from the definition of j_1, it follows that

$$2^{-a-1} \cdot C2^{-aj_1} \leq \bar{\delta}_{j_1} \leq C2^{-aj_1},$$

and, after inserting again $\bar{\delta}_{j_1} = c_\eta\epsilon\sqrt{j_1}$,

$$c_\eta\left(\frac{\epsilon\sqrt{j_1}}{C}\right)^{1/a} \leq 2^{-j_1} \leq c_\eta(a)\left(\frac{\epsilon\sqrt{j_1}}{C}\right)^{1/a},$$

and also, after taking logarithms, that for $\epsilon < \epsilon_1(a, C)$,

$$j_1 \geq (1/(2a))\log\epsilon^{-1}.$$

From the second display, we have $j_1 \leq (1/(2a))\log\epsilon^{-2} + \log(c(\eta, a)C^a) < \log_2\epsilon^{-2}$ for $\epsilon < \epsilon_2(a, C)$ since $2a > 1 > \log 2$. Hence, as claimed, $j_1 < \log_2 n$ for ϵ small.

Using the last three displays in turn, we find that

$$\gamma_{j_1} \geq 2^{-a-2}C2^{-j_1(a-a')} \geq cC\left(\frac{\epsilon\sqrt{j_1}}{C}\right)^{\frac{a-a'}{a}} \geq cC^{1-r}[\epsilon\sqrt{\log\epsilon^{-1}}]^r,$$

where the constant $c = c(a, a', \eta)$ may differ each time.

Returning to (10.27) and inserting γ_{j_1}, we have

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\|_{B'} \geq cC^{1-r}(\epsilon\sqrt{\log \epsilon^{-1}})^r\} \geq \pi_\eta(2^{j_1})$$

for $\epsilon < \epsilon(a, C)$; i.e. for $n > n(a, C)$. From the third display it is clear that $j_1 \to \infty$ as $n \to \infty$ so that $\pi_\eta(2^{j_1}) \to 1$.

10.9 Further Details

[2]. Here is a proof of the Cramér-Chernoff result, using a standard change of measure argument. Let P_π denote a binomial distribution Bin (n_0, π), and let B denote the corresponding random variable. The likelihood ratio

$$\frac{dP_{\pi_0}}{dP_{\pi_1}} = \big(\pi_0/\pi_1\big)^B \big(\bar{\pi}_0/\bar{\pi}_1\big)^{n_0 - B}.$$

Defining $\lambda = \log \pi_0/\pi_1$ and $\bar{\lambda} = \log \bar{\pi}_0/\bar{\pi}_1$, rewrite the loglikelihood ratio as

$$L = \log \frac{dP_{\pi_0}}{dP_{\pi_1}} = (\lambda - \bar{\lambda})B + n_0\bar{\lambda}.$$

Since $\pi_0 < \pi_1$ implies $\lambda < \bar{\lambda}$, it follows that $\{B \leq E_{\pi_0}B\} = \{L \geq E_{\pi_0}L\}$, while

$$E_{\pi_0}L = n_0 D(\pi_0, \pi_1) = n_0(\pi_0\lambda + \bar{\pi}_0\bar{\lambda}).$$

Consequently, using Markov's inequality along with $E_{\pi_1}e^L = 1$, we have

$$P_{\pi_1}\{B \leq n\pi_0\} = P_{\pi_1}\{e^L \geq e^{n_0 D}\} \leq e^{-n_0 D} E_{\pi_1} e^L = e^{-n_0 D}.$$

[3].

$$\begin{aligned} D(\pi_0, \pi_1) &= \pi_0 \log \frac{\pi_0}{\pi_1} + (1 - \pi_0) \log \frac{1 - \pi_0}{1 - \pi_1} \\ &= \pi_0 \int_{\pi_0}^{\pi_1} \frac{-du}{u} + (1 - \pi_0) \int_{\pi_0}^{\pi_1} \frac{du}{1 - u} \\ &= \int_{\pi_0}^{\pi_1} \frac{u - \pi_0}{u(1 - u)} du \geq 4 \int_{\pi_0}^{\pi_1} (u - \pi_0) du = 2(\pi_1 - \pi_0)^2. \end{aligned}$$

10.10 Exercises

1. Use (8.63) to show that if $n \geq 15$, then

$$\text{med} M_n \geq L_n - 1.$$

2. Use the concentration of measure bound (8.53) and the median bound (8.55) to show that

$$P\{M_{n-1} > Z + (1 - \eta)L_n\} \geq \pi_n(\eta) \to 1.$$

11

Model Selection, Penalization and Oracle Inequalities

Our ultimate goal is to obtain sharper bounds on rates of convergence - in fact exactly optimal rates, rather than with spurious log terms.

However, this is a situation where the tools introduced are perhaps of independent interest. These include model selection via penalized least squares, where the penalty function is not ℓ_2 or even ℓ_1 but instead a function of the *number* of terms in the model. We will call such things *complexity penalties*.

Many of the arguments work for general (i.e. non-orthogonal) linear models. While we will not ultimately use this extra generality in this book, there are important applications and the model is of such importance that it seems reasonable to present part of the theory in this setting.

While it is natural to start with penalties proportional to the number of terms in the model, it will turn out that for our later results on exact rates, it will be necessary to consider a larger class of "$2k\log(p/k)$" penalties, in which, roughly speaking, the penalty to enter the k^{th} variable is a function that decreases with k approximately like $2\log(p/k)$.

We will be looking essentially at "all subsets" versions of the model selection problem. If there are p variables, then there are $\binom{p}{k}$ distinct submodels with k variables, and this grows very quickly with k. In order to control the resulting model explosion, good exponential probability inequalities for the tails of chi-square distributions are needed. We will derive these as a consequence of a powerful *concentration* inequality for Gaussian measures in $\mathbb{R}^n$. We give a separate exposition of this result, as it is finding increasing application in statistics.

11.1 All subsets regression and complexity penalized least squares

We begin with the usual form of the general linear model with Gaussian errors:

$$y = X\beta + \epsilon z = \mu + \epsilon z, \qquad z \sim N_n(0, I). \tag{11.1}$$

There are n observations y and p unknown parameters β, connected by an $n \times p$ design matrix X with columns

$$X = [x_1, \cdots, x_p].$$

There is no restriction on p: indeed, we particularly wish to allow for situations in which $p \gg n$. We will assume that the noise level ϵ is known.

Example: Overcomplete dictionaries. Here is a brief indication of why one might wish to take $p \gg n$. Consider estimation of f in the continuous Gaussian white noise model (1.18), $dY(t) = f(t)dt + \epsilon dW(t)$,

and suppose that the observed data are inner products of Y with n orthonormal functions $\psi_1, \ldots, \psi_n$. Thus

$$y_i = \langle f, \psi_i \rangle + \epsilon z_i, \qquad i = 1, \ldots, n.$$

Now consider the possibility of approximating f by elements from a *dictionary* $\mathcal{D} = \{\phi_1, \phi_2, \ldots, \phi_p\}$. The hope is that by making $\mathcal{D}$ sufficiently rich, one might be able to represent f well by a linear combination of a very few elements of $\mathcal{D}$. This idea has been advanced by a number of authors. As a simple illustration, the ψ_i might be sinusoids at the first n frequencies, while the dictionary elements might allow a much finer sampling of frequencies

$$\phi_k(t) = \sin(2\pi k t/p), \qquad k = 1, \ldots, p = n^\beta \gg n.$$

with $p = n^\beta$ for some $\beta > 1$. If there is a single dominant frequency in the data, it is possible that it will be essentially captured by an element of the dictionary even if it does not complete an integer number of cycles in the sampling interval.

If we suppose that f has the form $f = \sum_{j=1}^p \beta_j \phi_j$, then these observation equation become an instance of the general linear model (11.1) with

$$X_{ij} = \langle \psi_i, \phi_j \rangle.$$

Again, the hope is that one can find an estimate $\hat{\beta}$ for which only a small number of components $\hat{\beta}_j \neq 0$.

All subsets regression. To each subset $J \subset \{1, \ldots, p\}$ of cardinality $n_J = |J|$ corresponds a regression model which fits only the variables x_j for $j \in J$. The possible fitted vectors μ that could arise from these variables lie in the model space

$$S_J = \text{span}\{x_j : j \in J\}.$$

The dimension of S_J is at most n_J, and could be less in the case of collinearity.

Let P_J denote orthogonal projection onto S_J: the least squares estimator $\hat{\mu}_J$ of μ is given by $\hat{\mu}_J = P_J y$. We include the case $J = \emptyset$, writing $n_\emptyset = 0$, $S_\emptyset = \{0\}$ and $\hat{\mu}_\emptyset(y) \equiv 0$. The issue in all subsets regression consists in deciding how to select a subset $\hat{J}$ on the basis of data y: the resulting estimate of μ is then $\hat{\mu} = P_{\hat{J}} y$.

Mean squared error properties can be used to motivate all subsets regression. We will use a predictive risk[1] criterion to judge an estimator $\hat{\beta}$ through the fit $\hat{\mu} = X\hat{\beta}$ that it generates:

$$E\|X\hat{\beta} - X\beta\|^2 = E\|\hat{\mu} - \mu\|^2.$$

The mean of a projection estimator $\hat{\mu}_J$ is just the projection of μ, namely $E\hat{\mu}_J = P_J\mu$, while its variance is $\epsilon^2 \text{tr} P_J = \epsilon^2 \dim S_J$. From the variance-bias decomposition of MSE,

$$E\|\hat{\mu}_J - \mu\|^2 = \|P_J\mu - \mu\|^2 + \epsilon^2 \dim S_J.$$

A *saturated* model arises from any subset with $\dim S_J = n$, so that $\hat{\mu}_J = y$ "interpolates the data". In this case the MSE is just the unrestricted minimax risk for $\mathbb{R}^n$:

$$E\|\hat{\mu} - \mu\|^2 = n\epsilon^2.$$

[1] Why the name "predictive risk"? Imagine that new data will be taken from the same design as used to generate the original observations y and estimator $\hat{\beta}$: $y^* = X\beta + \epsilon z^*$. A natural prediction of y^* is $X\hat{\beta}$, and its mean squared error, averaging over the distributions of both z and z^*, is

$$E\|y^* - X\hat{\beta}\|^2 = E\|X\beta - X\hat{\beta}\|^2 + n\epsilon^2,$$

so that the mean squared error of prediction equals $E\|\hat{\mu} - \mu\|^2$, up to an additive factor that doesn't depend on the model chosen.]

Comparing the last two displays, we see that if μ lies close to a low rank subspace — $\mu \doteq \sum_{j \in J} \beta_j x_j$ for $|J|$ small—then $\hat{\mu}_J$ offers substantial risk savings over a saturated model. Thus, it seems that one would wish to expand the dictionary $\mathcal{D}$ as much as possible to increase the possibilities for sparse representation. Against this must be set the dangers inherent in fitting over-parametrized models – principally overfitting of the data. Penalized least squares estimators are designed specifically to address this tradeoff.

This discussion also leads to a natural generalization of the notion of ideal risk introduced in Chapter 8.3. For each mean vector μ, there will be an optimal model subset $J = J(\mu)$ which attains the ideal risk

$$\mathcal{R}(\mu, \epsilon) = \min_J \ \|\mu - P_J \mu\|^2 + \epsilon^2 \dim S_J.$$

Of course, this choice $J(\mu)$ is not available to the statistician, since μ is unknown. The challenge, taken up below, is to see to what extent penalized least squares estimators can "mimick" ideal risk, in a fashion analagous to the mimicking achieved by threshold estimators in the orthogonal setting.

Complexity penalized least squares. The residual sum of squares (RSS) of model J is

$$\|y - \hat{\mu}_J\|^2 = \|y - P_J y\|^2,$$

and clearly decreases as the model J increases. To discourage simply using a saturated model, or more generally to discourage overfitting, we introduce a penalty on the size of the model, $\mathrm{pen}(n_J)$, that is increasing in n_J, and then define a complexity criterion

$$C(J, y) = \|y - \hat{\mu}_J\|^2 + \epsilon^2 \mathrm{pen}(n_J). \tag{11.2}$$

The complexity penalized RSS estimate $\hat{\mu}_{\mathrm{pen}}$ is then given by orthogonal projection onto the subset that minimizes the penalized criterion:

$$\hat{J}_{\mathrm{pen}} = \mathrm{argmin}_J \ C(J, y), \qquad \hat{\mu}_{\mathrm{pen}} = P_{\hat{J}_{\mathrm{pen}}} y. \tag{11.3}$$

The simplest penalty function grows linearly in the number of variables in the model:

$$\mathrm{pen}_0(k) = \lambda_p^2 \, k, \tag{11.4}$$

where we will take λ_p^2 to be roughly of order $2 \log p$. [The well known AIC criterion would set $\lambda_p^2 = 2$: this is effective for selection among a nested sequence of models, but is known to overfit in all-subsets settings.

For this particular case, we describe the kind of oracle inequality to be proved in this chapter. First, note that for $\mathrm{pen}_0(k)$, minimal complexity and ideal risk are related:

$$\begin{aligned} \min_J C(J, \mu) &= \min_J \left[\|\mu - P_J \mu\|^2 + \epsilon^2 \mathrm{pen}_0(n_J) \right] \\ &\le \lambda_p^2 \min \left[\|\mu - P_J \mu\|^2 + \epsilon^2 n_J \right] = \lambda_p^2 \, \mathcal{R}(\mu, \epsilon). \end{aligned}$$

Let $\lambda_p = \zeta(1 + \sqrt{2 \log p})$ for $\zeta > 1$ and $A(\zeta) = (1 - \zeta^{-1})^{-1}$. `[need to update.]` Then for penalty function (11.4) and arbitary μ,

$$E \|\hat{\mu}_{\mathrm{pen}} - \mu\|^2 \le A(\zeta) \lambda_p^2 [C\epsilon^2 + \mathcal{R}(\mu, \epsilon)].$$

Thus, the complexity penalized RSS estimator, for non-orthogonal and possibly over-complete dictionaries, comes within a factor of order $2 \log p$ of the ideal risk.

Remark. Another possibility is to use penalty functions monotone in the rank of the model, pen(dim J). However, when $k \to \text{pen}(k)$ is strictly monotone, this will yield the same models as minimizing (11.2), since a collinear model will always be rejected in favor of a sub-model with the same span.

11.2 Orthogonal Case

For this section we specialize to the n-dimensional white Gaussian sequence model:

$$y_i = \mu_i + \epsilon z_i, \qquad i = 1, \ldots, n, \qquad z_i \overset{i.i.d.}{\sim} N(0,1). \tag{11.5}$$

This is the canonical form of the more general orthogonal regression setting $Y = X\beta + \epsilon Z$, with N dimensional response and n dimensional parameter vector β linked by an orthogonal design matrix X satisfying $X^T X = I_n$, and with the noise $Z \sim N_N(0, I)$. This reduces to (11.5) after premultiplying by X^T and setting $y = X^T Y$, $\mu = \beta$ and $z = X^T Z$.

We will see in this section that, in the orthogonal regression setting, the penalized least squares estimator can be written in terms of a penalty on the number of non-zero elements (Lemma 11.1). There are also interesting connections to hard thresholding, in which the threshold is data dependent.

The columns of the design matrix implicit in (11.5) are the unit co-ordinate vectors e_i, consisting of zeros except for a 1 in the i^{th} position. The least squares estimator corresponding to a subset $J \subset \{1, \ldots, n\}$ is simply given by co-ordinate projection P_J:

$$(P_J y)_j = \begin{cases} y_j & j \in J \\ 0 & j \notin J. \end{cases}$$

The complexity criterion becomes

$$C(J, y) = \sum_{j \notin J} y_j^2 + \epsilon^2 \text{pen}(n_J),$$

where $n_J = |J|$ still. Using $|y|_{(j)}$ to denote the order statistics of $|y_j|$, in decreasing order, we can write

$$\min_J C(J, y) = \min_{0 \le k \le n} \sum_{j > k} |y|_{(j)}^2 + \epsilon^2 \text{pen}(k). \tag{11.6}$$

There is an equivalent form of the penalized least squares estimator in which the model selection aspect is less explicit. Let $N(\mu) = \#\{i : \mu_i \neq 0\}$ be the number of non-zero components of μ.

Lemma 11.1 *Suppose that $k \to pen(k)$ is monotone increasing. In orthogonal model* (11.5)*, the penalized least squares estimator can be written*

$$\hat{\mu}_{pen}(y) = \underset{\mu}{\text{argmin}} \, \|y - \mu\|^2 + \epsilon^2 \text{pen}(N(\mu)).$$

Proof The model space S_J corresponding to subset J consists of vectors μ whose components μ_j vanish for $j \notin J$. Let $S_J^+ \subset S_J$ be the subset on which the components $\mu_j \neq 0$

for every $j \in J$. The key point is that on S_J^+ we have $N(\mu) = n_J$. Since $\mathbb{R}^n$ is the disjoint union of all S_J^+—using $\{0\}$ in place of $S_\emptyset^+$—we get

$$\min_\mu \|y-\mu\|^2 + \epsilon^2 \mathrm{pen}(N(\mu)) = \min_J \min_{\mu \in S_J^+} \|y-\mu\|^2 + \epsilon^2 \mathrm{pen}(n_J).$$

The minimum over $\mu \in S_J^+$ can be replaced by a minimum over $\mu \in S_J$ without changing the value because if $\mu \in S_J \backslash S_J^+$ there is a smaller subset J' with $\mu \in S_{J'}^+$—here we use monotonicity of the penalty. So we have recovered precisely the model selection definition (11.3) of $\hat{\mu}_{\mathrm{pen}}$. □

When $\mathrm{pen}(k) = \lambda^2 k$, we recover the ℓ_0 penalty and the corresponding estimator is hard thresholding at $\epsilon\lambda$. To explore the connection with thresholding for more general penalties, consider the form $\mathrm{pen}(k) = \sum_{j=1}^k t_{n,j}^2$. Then

$$\hat{k} = \operatorname*{argmin}_k \sum_{j>k} |y|_{(j)}^2 + \epsilon^2 \sum_{j=1}^k t_{n,j}^2 \tag{11.7}$$

and $\hat{\mu}_{\mathrm{pen}}$ corresponds to hard thresholding at a *data-dependent* value $\hat{t}_{\mathrm{pen}} = t_{n,\hat{k}}$.

Proposition 11.2 *If $k \to t_{n,k}$ is strictly decreasing, then*

$$|y|_{(\hat{k}+1)} < \epsilon t_{n,\hat{k}} \le |y|_{(\hat{k})}, \tag{11.8}$$

and

$$\hat{\mu}_{pen,j}(y) = \begin{cases} y_j & |y_j| \ge \epsilon t_{n,\hat{k}} \\ 0 & \textit{otherwise.} \end{cases} \tag{11.9}$$

Proof Let $S_k = \epsilon^2 \sum_{j=1}^k t_k^2 + \sum_{j>k} |y|_{(j)}^2$, for notational simplicity, we write t_k instead of $t_{n,k}$. We have

$$S_k - S_{k-1} = \epsilon^2 t_k^2 - |y|_{(k)}^2.$$

Now $\hat{k}$ minimizes $k \to S_k$, so in particular we have both $S_{\hat{k}} \le S_{\hat{k}-1}$ and $S_{\hat{k}} \le S_{\hat{k}+1}$, which respectively imply that

$$|y|_{(\hat{k})} \ge \epsilon t_{\hat{k}}, \qquad \text{and} \qquad |y|_{(\hat{k}+1)} \le \epsilon t_{\hat{k}+1} < \epsilon t_{\hat{k}},$$

where at the last strict inequality we used the assumption on t_k. Together, these inequalities yield (11.8) and also the set identity

$$\{j : |y_j| \ge \epsilon t_{\hat{k}}\} = \{j : |y_j| \ge |y|_{\hat{k}}\}.$$

Since the set on the right side is $\hat{J}$, we have shown (11.9). □

Example. FDR estimation. In Chapter 7.6, (7.26) described a data dependent threshold choice that is closely related to penalized estimation as just described with $t_{n,k}^2 = z(kq/2n)$. Indeed, let $\hat{k}_F = \max\{k : |y|_{(k)} \ge \epsilon t_{n,k}\}$ denote the last crossing, and consider also the first crossing $\hat{k}_G + 1 = \min\{k : |y|_{(k)} < \epsilon t_{n,k}\}$. If $\hat{k}_{pen}$ denotes the penalized choice (11.7), then Section 11.6 shows that

$$\hat{k}_G \le \hat{k}_{pen} \le \hat{k}_F$$

and in simulations it is often found that all three agree.

In Exercise 11.1, it is verified that if k, possibly depending on n, is such that $k/n \to 0$ as $n \to \infty$, then

$$t_k^2 \le (1/k)\sum_1^k t_j^2 \le t_k^2 + c(k) \sim 2\log(n/k \cdot 2/q) \tag{11.10}$$

and hence that

$$\text{pen}(k) \sim 2k\log(n/k \cdot 2/q). \tag{11.11}$$

The idea to use penalties of the general form $2\epsilon^2 k\log(n/k)$ arose among several authors more or less simultaneously:

- Foster and Stine (1997) $\text{pen}(k) = \epsilon^2 \sum_1^k 2\log(n/j)$ via information theory.
- George and Foster (2000) Empirical Bayes approach. $[\mu_i \overset{i.i.d.}{\sim} (1-w)\delta_0 + wN(0,C)$ followed by estimation of $(w,C)]$. They argue that this approach penalizes the k^{th} variable by about $2\epsilon^2\log(((n+1)/k)-1)$.
- The covariance inflation criterion of Tibshirani and Knight (1999) in the orthogonal case leads to $\text{pen}(k) = 2\epsilon^2\sum_1^k 2\log(n/j)$.
- FDR - discussed above (?).
- Birgé and Massart (2001) contains a systematic study of complexity penalized model selection from the specific viewpoint of obtaining non-asymptotic bounds, using a penalty class similar to, but more general than that used here.

11.3 Oracle Inequalities

Consider a penalty of the form

$$\text{pen}(k) = \zeta k(1+\sqrt{2L_k})^2 \qquad (\zeta > 1, L_k \ge 0). \tag{11.12}$$

This form is chosen both to include the $2k\log(n/k)$ class introduced earlier and to be convenient for theoretical analysis. Thus, the penalty reduces to pen_0 if L_k is identically constant. Typically, however, $L_k = L_{p,k}$ is chosen so that $L_{p,k} \ge \log(p/k)$ and is decreasing in k. We will see in Section 11.5 and the next chapter that this property is critical for removing logarithmic terms in convergence rates.

As a concession to theoretical analysis, we will need $\zeta > 1$ and the extra "1" in (11.12), which are both needed for the technical arguments, but make the implied thresholds a bit larger than would otherwise be desirable in practice.

We abuse notation a little and write L_J for L_{n_J}. Associated with the penalty is a constant

$$M = \sum_J e^{-L_J n_J}. \tag{11.13}$$

Let us look at a couple of examples of penalty functions and the associated evaluation of M.

(i) Penalty (11.4), namely $\text{pen}_0(k) = \lambda_p^2 k$, takes the form (11.12) if λ_p is written in

the form $\lambda_p = \sqrt{\zeta}(1 + \sqrt{2\alpha \log p})$, and we set $L_k \equiv \alpha \log p$. Since there are at most $\binom{p}{k} \le p^k/k!$ subsets of $\{1, \ldots, p\}$ having cardinality $n_J = k$,

$$M = \sum_J e^{-n_J \alpha \log p} = \sum_{k=0}^{p} \binom{p}{k} e^{-k\alpha \log p} \le \sum_{k=0}^{\infty} \frac{(p \cdot p^{-\alpha})^k}{k!} \le \exp(p^{1-\alpha})$$

The last term is uniformly bounded in p so long as $\alpha \ge 1$. Thus, convergence of (11.13) and the theorem below require that $\lambda_p^2 \sim \zeta \cdot (2 \log p)$ or larger when p is large.

(ii) Now suppose that $L_k = \log(p/k) + \gamma$. Proceeding much as above,

$$M = \sum_{k=0}^{p} \binom{p}{k} e^{-kL_k} \le \sum_{0}^{\infty} \frac{p^k}{k!} \left(\frac{k}{p}\right)^k e^{-\gamma k} \le \sum_{k} \frac{1}{\sqrt{2\pi k}} e^{-(\gamma-1)k}. \tag{11.14}$$

using Stirling's formula, $k! = \sqrt{2\pi k} k^k e^{-k+\theta}$, with $(12k)^{-1} \le \theta \le (12k+1)^{-1}$. The last sum converges so long as $\gamma > 1$.

Theorem 11.3 *Let $\hat{\mu}$ be a penalized least squares estimator of (11.2)–(11.3) for a penalty and constant M as defined above. There exists a constant $K = K(\zeta)$ such that*

$$E\|\hat{\mu} - \mu\|^2 \le K[2M\epsilon^2 + \min_J C(J, \mu)]. \tag{11.15}$$

The constant K may be taken as $K(\zeta) = 2\zeta(\zeta+1)^3(\zeta-1)^{-3}$.

Proof 1°. Writing $y = \mu + \epsilon z$ and expanding (11.2), we have

$$C(\hat{J}, y) = \|\hat{\mu}_{\hat{J}} - \mu\|^2 + 2\epsilon\langle z, \mu - \hat{\mu}_{\hat{J}}\rangle + \epsilon^2\|z\|^2 + \epsilon^2 \mathrm{pen}(n_{\hat{J}}).$$

We aim to use the minimizing property, $C(\hat{J}, y) \le C(J, y)$, to get an upper bound for $\|\hat{\mu}_{\hat{J}} - \mu\|^2$. To this end, for an arbitrary index J, writing $P_J^\perp = I - P_J$ and $\mu_J = P_J\mu$, we have

$$\begin{aligned} \|P_J^\perp y\|^2 &= \|P_J^\perp \mu\|^2 + 2\epsilon\langle P_J^\perp z, P_J^\perp \mu\rangle + \epsilon^2\|P_J^\perp z\|^2 \\ &\le \|P_J^\perp \mu\|^2 + 2\epsilon\langle z, \mu - \mu_J\rangle + \epsilon^2\|z\|^2. \end{aligned}$$

Consequently

$$C(J, y) = \|P_J^\perp y\|^2 + \epsilon^2 \mathrm{pen}(n_J) \le C(J, \mu) + 2\epsilon\langle z, \mu - \mu_J\rangle + \epsilon^2\|z\|^2.$$

By definition, $C(\hat{J}, y) \le C(J, y)$, so combining the corresponding equations and cancelling terms yields a bound for $\hat{\mu}_{\hat{J}} - \mu$:

$$\|\hat{\mu}_{\hat{J}} - \mu\|^2 \le C(J, \mu) + 2\epsilon\langle z, \hat{\mu}_{\hat{J}} - \mu_J\rangle - \epsilon^2 \mathrm{pen}(n_{\hat{J}}). \tag{11.16}$$

The merit of this form is that we can hope to appropriately apply the Cauchy-Schwarz inequality, (11.20) below, to the linear term $\langle z, \hat{\mu}_{\hat{J}} - \mu_J\rangle$, and take a multiple of $\|\hat{\mu}_{\hat{J}} - \mu\|^2$ over to the left side to develop a final bound.

2°. We outline the strategy based on (11.16). We construct an increasing family of sets Ω_x for $x > 0$, with $P(\Omega_x^c) \le Me^{-x}$ and then show for each $\eta \in (0, 1)$ that there are constants $a(\eta), b(\eta)$ for which we can bound the last two terms of (11.16): when $\omega \in \Omega_x$,

$$2\epsilon\langle z, \hat{\mu}_{\hat{J}} - \mu_J\rangle - \epsilon^2 \mathrm{pen}(n_{\hat{J}}) \le (1-\eta^2)\|\hat{\mu}_{\hat{J}} - \mu\|^2 + a(\eta)C(J, \mu) + b(\eta)\epsilon^2 x. \tag{11.17}$$

Now we can move the first term on the right side to the left side of (11.16). We get

$$\|\hat{\mu}_{\hat{J}} - \mu\|^2 \le \eta^{-2}(1 + a(\eta))C(J, \mu) + \eta^{-2}b(\eta)\epsilon^2 X, \tag{11.18}$$

where $X(\omega) = \inf\{x : \omega \in \Omega_x\}$. Clearly $X(\omega) > x$ implies that $\omega \notin \Omega_x$, and so using the bound on $P(\Omega_x^c)$ gives $EX = \int_0^\infty P(X > x)dx \le M$. Hence, taking expectations, then minimizing over J, and setting $A(\eta) = \eta^{-2}(1 + a(\eta))$ and $B(\eta) = \eta^{-2}b(\eta)$, we get

$$E\|\hat{\mu}_{\hat{J}} - \mu\|^2 \le A(\eta) \min_J C(J, \mu) + B(\eta)\epsilon^2 M. \tag{11.19}$$

3°. We turn to the derivation of (11.17). Consider a pair of subsets J, J': we imagine J as fixed, and J' as being varied (it will later be set to $\hat{J}$.) To effectively bound the inner product term, introduce random variables

$$\chi_{J,J'} = \sup\{\langle z, u\rangle / \|u\|, u \in S_J \oplus S_{J'}\},$$

so that

$$\langle z, \hat{\mu}_{J'} - \mu_J\rangle \le \|\hat{\mu}_{J'} - \mu_J\| \cdot \chi_{J,J'}. \tag{11.20}$$

Clearly $\chi^2_{J,J'} \sim \chi^2_{(d)}$ with degrees of freedom $d = \dim(S_J \oplus S_{J'}) \le n_J + n_{J'}$.

We now use the Lipschitz concentration of measure bound (2.58), which says here that $P\{\chi_{(d)} \ge \sqrt{d} + t\} \le e^{-t^2/2}$ for all $t \ge 0$, and, crucially, for all non-negative integer d. (If $d = 0$, then $\chi_{(0)} = 0$.) For arbitrary $x > 0$, let $E_{J'}(x)$ be the event

$$\chi_{J,J'} \le \sqrt{n_J + n_{J'}} + \sqrt{2(L_{J'}n_{J'} + x)}, \tag{11.21}$$

and in the concentration bound set $t^2 = 2(L_{J'}n_{J'} + x)$. Let $\Omega_x = \cap_{J'} E_{J'}(x)$, so that

$$P(\Omega_x^c) \le e^{-x} \sum_{J'} e^{-L_{J'}n_{J'}} = Me^{-x}.$$

Using $\sqrt{a + b} \le \sqrt{a} + \sqrt{b}$ twice in (11.21) and then combining with (11.20), we conclude that on the set Ω_x,

$$\langle z, \hat{\mu}_{J'} - \mu_J\rangle \le \|\hat{\mu}_{J'} - \mu_J\| \cdot [\sqrt{n_{J'}}(1 + \sqrt{2L_{J'}}) + \sqrt{n_J} + \sqrt{2x}].$$

The key to extracting $\|\hat{\mu}_{J'} - \mu_J\|^2$ with a coefficient less than 1 is to use the inequality $2\alpha\beta \le c\alpha^2 + c^{-1}\beta^2$, valid for all $c > 0$. Thus, for $0 < \eta < 1$,

$$\begin{aligned} &2\epsilon\langle z, \hat{\mu}_{J'} - \mu_J\rangle \\ &\quad\le (1 - \eta)\|\hat{\mu}_{J'} - \mu_J\|^2 + \frac{\epsilon^2}{1 - \eta}\Big[\sqrt{n_{J'}}(1 + \sqrt{2L_{J'}}) + \sqrt{n_J} + \sqrt{2x}\Big]^2. \end{aligned} \tag{11.22}$$

Now use this trick again, now in the form $(\alpha + \beta)^2 \le (1 + \eta)\alpha^2 + (1 + \eta^{-1})\beta^2$, on each of the right side terms. In the first term, use $\|\hat{\mu}_{J'} - \mu_J\| \le \|\hat{\mu}_{J'} - \mu\| + \|\mu_J - \mu\|$ and get

$$(1 - \eta^2)\|\hat{\mu}_{J'} - \mu\|^2 + (\eta^{-1} - \eta)\|\mu_J - \mu\|^2.$$

In the second, use $\mathrm{pen}(n_{J'}) = \zeta n_{J'}(1 + \sqrt{2L_{J'}})^2$ and get

$$\frac{1 + \eta}{1 - \eta}\zeta^{-1}\epsilon^2\mathrm{pen}(n_{J'}) + \frac{1 + \eta^{-1}}{1 - \eta}\epsilon^2(2n_J + 4x).$$

Now, choose η so that $(1+\eta)/(1-\eta) = \zeta$, and then move the resulting $\epsilon^2 \mathrm{pen}(n_{J'})$ term to the left side. To bound the rightmost terms in the two previous displays, set

$$a(\eta) = \max\left\{\eta^{-1} - \eta, \frac{1+\eta^{-1}}{1-\eta}\frac{2}{\zeta}\right\}, \qquad b(\eta) = \frac{4(1+\eta^{-1})}{1-\eta}, \tag{11.23}$$

and note that $\zeta n_J \le \mathrm{pen}(n_J)$. Finally, setting $J' = \hat{J}$, we recover the desired inequality (11.17).

From the choice of η, we have $a(\eta) = 2\eta^{-1}$ and $b(\eta) = 4\eta^{-1}\zeta$. The constant

$$K(\zeta) = \max(A(\eta), B(\eta)/2) = \eta^{-3}\max(2+\eta, 2\zeta) = 2\eta^{-3}\zeta, \tag{11.24}$$

since $\zeta > 1 + \eta/2$ for all $0 < \eta < 1$ and the expression for $K(\zeta)$ follows. □

Remark. We have not sought to optimize the value of $K(\zeta)$ that might be attained with this method. There is room for such optimization: that this value of $K(\zeta)$ appears unchanged in the modification in Theorem 11.9 is a symptom of this.

11.4 Back to orthogonal case

In the orthogonal setting (11.5), we explore the links with thresholding for the penalties

$$\mathrm{pen}(k) = k\lambda_k^2, \qquad \lambda_k = \sqrt{\zeta}(1+\sqrt{2L_k})$$

introduced for the oracle inequalities of the last section.

Defining $t_k^2 = k\lambda_k^2 - (k-1)\lambda_{k-1}^2$, we can write $\mathrm{pen}(k) = \sum_1^k t_j^2$ in the form needed for the thresholding result Proposition 11.2, which interprets $\hat{\mu}_{\mathrm{pen}}$ as hard thresholding at $\hat{t} = t_{\hat{k}}$ where $\hat{k} = |\hat{J}|$ is the size of the selected model.

It is heuristically plausible that $t_k \approx \lambda_k$, but here is a more precise bound.

Lemma 11.4 *Suppose that the function $k \to L_k$ appearing in λ_k is decreasing, and for some constant $b \ge 0$ satisfies*

$$L_k \ge \max(\tfrac{1}{2}, 2b), \qquad k(L_k - L_{k-1}) \le b. \tag{11.25}$$

Then we have the bounds

$$\lambda_k - 4\zeta b/\lambda_k \le t_k \le \lambda_k.$$

Note in particular that if L_k is constant, then we can take $b = 0$ and $t_k = \lambda_k$. More generally, if $L_k = (1+2\beta)\log(n\gamma/k)$ for $\beta \ge 0$, then condition (11.25) holds with $b = 1 + 2\beta$ so long as $\gamma \ge e^2$.

In sparse cases, $k = o(n)$, we have $\lambda_k \asymp \sqrt{\log n}$ and t_k gets closer to λ_k as n grows.

Proof From the definition of t_k^2 and the monotonicity of λ_k^2 we have

$$t_k^2 - \lambda_k^2 = (k-1)(\lambda_k^2 - \lambda_{k-1}^2) \le 0,$$

so that $t_k = \lambda_k$. For the other bound, again use the definition of t_k^2, now in the form

$$\lambda_k - t_k \le \lambda_{k-1} - t_k = \frac{\lambda_{k-1}^2 - t_k^2}{\lambda_{k-1} + t_k} = k\frac{\lambda_{k-1} + \lambda_k}{\lambda_{k-1} + t_k}(\lambda_{k-1} - \lambda_k). \tag{11.26}$$

Setting $\delta = \lambda_k - t_k$ and $\Lambda = \lambda_{k-1} + \lambda_k$, this takes the form

$$\frac{\delta}{\Lambda} \le \frac{k(\lambda_{k-1} - \lambda_k)}{\lambda_{k-1} + t_k} \le \frac{k(\lambda_{k-1} - \lambda_k)}{\Lambda - \delta}. \tag{11.27}$$

Using now the definition of λ_k, then the bounds $L_{k-1} \ge \frac{1}{2}$ and $k(L_k - L_{k-1}) \le b$, we find

$$k(\lambda_{k-1} - \lambda_k) = \sqrt{2\zeta}\frac{k(L_{k-1} - L_k)}{\sqrt{L_{k-1}} + \sqrt{L_k}} \le \frac{2\zeta \cdot b}{\sqrt{\zeta}(1 + \sqrt{2L_k})} = \frac{2\zeta b}{\lambda_k}. \tag{11.28}$$

The bound $L_k \ge 2b$ implies $\lambda_k^2 \ge 4\zeta b$, and if we return to first inequality in (11.27) and simply use the crude bound $t_k \ge 0$ along with (11.28), we find that

$$\delta/\Lambda \le 2\zeta b/\lambda_k^2 \le 1/2.$$

Returning to the second inequality in (11.27), we now have $\delta/\Lambda \le 2k(\lambda_{k-1} - \lambda_k)/\Lambda$, and again using (11.28), we get $\delta \le 4\zeta b/\lambda_k$, which is the bound we claimed. □

An important simplification occurs in the theoretical complexity $C(J, \mu)$ in the orthogonal case. As in Section 11.2, but now using μ rather than y,

$$C(J, \mu) = \sum_{k \notin J} \mu_k^2 + \epsilon^2 \text{pen}(n_J)$$

The minimum theoretical complexity is denoted by

$$\mathcal{R}(\mu) = \min_J C(J, \mu).$$

Then, as at (11.6) we have

$$\mathcal{R}(\mu) = \min_{0 \le k \le n} \sum_{j>k} \mu_{(j)}^2 + k\lambda_k^2\epsilon^2. \tag{11.29}$$

There is a simple co-ordinatewise upper bound for theoretical complexity.

Lemma 11.5 *If pen$(k) = k\lambda_k^2$ with $k \to \lambda_k^2$ is non-increasing, then*

$$\mathcal{R}(\mu) \le \sum_{k=1}^{n} \mu_{(k)}^2 \wedge \lambda_k^2\epsilon^2.$$

Proof Without loss of generality, put $\epsilon = 1$. Let $\kappa = \max\{k \ge 1 : \lambda_k\epsilon \le |\mu|_{(k)}\}$ if such an index exists, otherwise set $\kappa = 0$. Let $M_k = \sum_{j>k} \mu_{(j)}^2$. Since both $k \to \lambda_k$ and $k \to |\mu|_{(k)}$ are non-increasing, we have

$$\begin{aligned} \sum_{k=1}^{n} \mu_{(k)}^2 \wedge \lambda_k^2 &= \sum_{1}^{\kappa} \mu_{(k)}^2 \wedge \lambda_k^2 + M_\kappa \ge \kappa(\mu_{(\kappa)}^2 \wedge \lambda_\kappa^2) + M_\kappa \\ &= \kappa\lambda_\kappa^2 + M_\kappa \ge \min_k M_k + k\lambda_k^2. \end{aligned} \tag{11.30}$$

□

We have been discussing several forms of minimization that turn out to be closely related. To describe this, we use a modified notation. Consider first

$$\mathcal{R}_S(s,\gamma) = \min_{0\le k\le n} \sum_1^k s_j + \sum_{k+1}^n \gamma_j. \tag{11.31}$$

With the identifications $s_k \leftrightarrow t^2_{n,k}$ and $\gamma_k \leftrightarrow |y|^2_k$, we recover the objective function in the thresholding formulation of penalization, (11.7). When using a penalty of the form $\mathrm{pen}(k) = k\lambda_k^2$, compare (11.29), we use a measure of the form

$$\mathcal{R}_C(s,\gamma) = \min_{0\le k\le n} ks_k + \sum_{k+1}^n \gamma_j. \tag{11.32}$$

Finally, the co-ordinatewise minimum

$$\mathcal{R}(s,\gamma) = \sum_1^n s_k \wedge \gamma_k. \tag{11.33}$$

Under mild conditions on the sequence $\{s_k\}$, these measures are equivalent up to constants. To state this, introduce a hypothesis:

(H) The values $s_k = \sigma(k/n)$ for $\sigma(u)$ a positive decreasing function on $[0,1]$ with

$$\lim_{u\to 0} u\sigma(u) = 0, \qquad \sup_{0\le u\le 1} |u\sigma'(u)| \le c_1.$$

For such a function, let $c_\sigma = 1 + c_1/\sigma(1)$.

A central example is given by $\sigma(u) = 2\log(\gamma/u)$, with $c_1 = 2$ and $c_\sigma = 1 + (\log\gamma)^{-1}$.

Proposition 11.6 *Let the sequence $\{s_k\}$ satisfy hypothesis (H). Let $\mathcal{R}_S, \mathcal{R}_C$ and $\mathcal{R}$ be the minima defined in* (11.31)–(11.33) *above. Then the measures are equivalent: for all non-negative, decreasing sequences $\gamma \in \mathbb{R}^n$,*

$$c_\sigma^{-1}\mathcal{R}_S(s,\gamma) \le \mathcal{R}_C(s,\gamma) \le \mathcal{R}(s,\gamma) \le \mathcal{R}_S(s,\gamma) \le c_\sigma \mathcal{R}_C(s,\gamma).$$

Remark. The central two inequalities, in which c_σ does not appear, are valid for any positive decreasing sequence $\{s_k\}$, without any need for hypothesis (H).

Proof Consider first the bounds not involving the constant c_σ. The bound $\mathcal{R}_C \le \mathcal{R}$ is precisely Lemma 11.5, while $\mathcal{R} \le \mathcal{R}_S$ is immediate since each sum appearing in (11.31) is bounded below by $\sum s_k \wedge \gamma_k$. The bounds with c_σ will follow if we show that (H) implies $\sum_1^k s_j \le c_\sigma k s_k$ for $k = 0,\dots,n$. But

$$\sum_1^k s_j = \sum_1^k \sigma(j/n) \le n\int_1^{k/n} \sigma(u)du,$$

and by partial integration

$$\int_0^v \sigma(u)du = v\sigma(v) + \int_0^v u|\sigma'(u)|du \le v[\sigma(v) + c_1] \le c_\sigma v\sigma(v).$$

Combining the previous two displays gives the bound we need. □

11.5 Non-asymptotic bounds for ℓ_p-balls

Suppose that we observe data from the n-dimensional Gaussian signal plus noise model (11.5), and that μ is constrained to lie in a ball of radius C defined by the ℓ_p norm:

$$\Theta = \Theta_{n,p}(C) = \{\mu \in \mathbb{R}^n : \sum_{i=1}^{n} |\mu_i|^p \le C^p\}. \tag{11.34}$$

We seek to evaluate the nonlinear minimax risk

$$R_N(\Theta) = \inf_{\hat{\mu}} \sup_{\mu \in \Theta} E\|\hat{\mu} - \mu\|_2^2.$$

In this section we will study *non-asymptotic* upper and lower bounds for the minimax risk – and will see that these lead to the optimal rates of convergence for these classes of parameter spaces.

The non-asymptotic bounds will have a number of consequences. We will see a sharp transition between the sparse case $p < 2$, in which non-linear methods clearly outperform linear ones, and the more requlars setting of $p \ge 2$.

The upper bounds will illustrate the use of the $2k\log(n/k)$ type oracle inequalities established in earlier sections. They will also be used in the next chapter to derive exactly optimal rates of convergence over Besov spaces for certain wavelet shrinkage estimators.

The lower bounds exemplify the use of minimax risk tools based on hyperrectangles.

While the non-asymptotic bounds have the virtue of being valid for finite $\epsilon > 0$, their disadvantage is that the upper and lower bounds may be too conservative. The optimal constants can be found from a separate asymptotic analysis as $\epsilon \to 0$ (see Chapter 13 below).

A control function. The non-asymptotic bounds will be expressed in terms of a control function $r_{n,p}(C)$ defined separately for $p \ge 2$ and $p < 2$. The control function captures key features of the minimax risk $R_N(\Theta_{n,p}(C,\epsilon)$ but is more concrete, and is simpler in form. As with the minimax risk, it can be reduced by rescaling to a unit noise version

$$r_{n,p}(C,\epsilon) = \epsilon^2 r_{n,p}(C/\epsilon). \tag{11.35}$$

For $p < 2$, the control function is given by

$$r_{n,p}(C) = \begin{cases} C^2 & \text{if } C \le \sqrt{1+\log n}, \\ C^p\big[1+\log(n/C^p)\big]^{1-p/2} & \text{if } \sqrt{1+\log n} \le C \le n^{1/p}, \\ n & \text{if } C \ge n^{1/p}. \end{cases} \tag{11.36}$$

See Figure 11.1. As will become evident from the proof, the three zones correspond to situations where the least favorable signals are 'near zero', 'sparse' and 'dense' respectively. A little calculus shows that $C \to r_{n,p}(C)$ is indeed monotone increasing in C for $0 < p < 2$, except at the discontinuity at $C = \sqrt{1+\log n}$. This discontinuity is not serious—even for $n = 2$ the ratio of right limit to left limit exceeds 0.82 for $0 < p < 2$, and the ratio approaches 1 for large n.

For $p \ge 2$, the control function is simpler:

$$r_{n,p}(C) = \begin{cases} n^{1-2/p}C^2 & \text{if } C \le n^{1/p}, \\ n & \text{if } C \ge n^{1/p}. \end{cases} \tag{11.37}$$

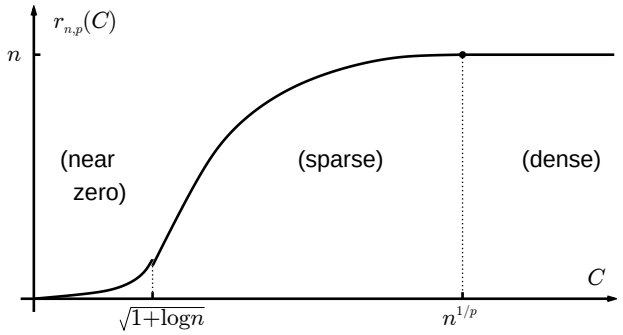

Figure 11.1 Schematic of the control function (11.36) for $p < 2$, showing the three zones for C, and the discontinuity at $C = \sqrt{1 + \log n}$.

To show that the bounds provided by the control function can be attained, we use a penalized least squares estimator $\hat{\mu}_P$ with a specific choice of penalty of the form (11.12), with

$$L_{n,k} = (1 + 2\beta) \log(n\gamma/k).$$

Thus $\text{pen}(k) = k\lambda_k^2$ with $\lambda_k = \sqrt{\zeta}(1 + \sqrt{2L_{n,k}})$.

The parameter β is included for applications to inverse problems in Chapter 12; for most other purposes we can take $\beta = 0$. The constant γ is included to obtain convergence of the sum defining the constant M: when $\beta = 0$ we need $\gamma > e$ (compare (11.14)).

Here is the main result of this section, saying that the minimax MSE for ℓ_p-balls is described, up to constants, by the control function $r_{n,p}(C)$, and that penalized least squares estimation can globally mimick the control function.

Theorem 11.7 *For $n \geq 1, 0 < p \leq \infty, 0 < C < \infty$, there exist constants a_1 and $c_1(\zeta, \beta, \gamma)$ so that*

$$a_1 r_{n,p}(C, \epsilon) \leq R_N(\Theta_{n,p}(C)) \tag{11.38}$$

$$\leq \sup_{\Theta_{n,p}(C)} E\|\hat{\mu}_P - \mu\|^2 \leq c_1[\epsilon^2 + r_{n,p}(C, \epsilon)]. \tag{11.39}$$

Note that a single estimator $\hat{\mu}_P$, defined without reference to either p or C, achieves the upper bound. We may thus speak of $\hat{\mu}_P$ as being adaptively optimal at the level of *rates* of convergence.

Constants convention. In the statement and proof, we use c_i to denote constants that depend on (ζ, β, γ) and a_j for to stand for absolute constants. While information is available about each such constant, we have not tried to assemble this into the final constants a_1 and c_1 above, as they would be far from sharp.

Proof 1°. *Upper Bounds.* We may assume, by scaling, that $\epsilon = 1$. As we are in the orthogonal setting, the oracle inequality of Theorem 11.3 takes the form

$$E\|\hat{\mu}_P - \mu\|^2 \leq K[2M + \mathcal{R}(\mu)],$$

where $K = K(\zeta), M = M(\beta, \gamma)$ and

$$\mathcal{R}(\mu) = \min_{0 \le k \le n} \sum_{j>k}^{n} \mu_{(j)}^2 + k\lambda_k^2. \tag{11.40}$$

For the upper bound, then, we need then to show that when $\mu \in \Theta_{n,p}(C)$,

$$\mathcal{R}(\mu) \le c_2 r_{n,p}(C). \tag{11.41}$$

We might guess that worst case bounds for (11.40) occur at gradually increasing values of k as C increases. In particular, the extreme zones for C will correspond to $k = 0$ and n. It turns out that these two extremes cover most cases, and then the main interest in the proof lies in the sparse zone for $p < 2$. Now to the details.

First put $k = n$ in (11.40). Since λ_n^2 is just a constant, c_3 say, we obtain

$$\mathcal{R}(\mu) \le c_3 n \tag{11.42}$$

valid for all C (and all p), but useful in the dense zone $C \ge n^{1/p}$.

For $p \ge 2$, simply by choosing $k = 0$ in (11.40), we also have

$$\mathcal{R}(\mu) \le n \cdot n^{-1} \sum \mu_j^2 \le n \left(n^{-1} \sum |\mu_j|^p \right)^{2/p} \le n^{1-2/p} C^2. \tag{11.43}$$

Combining the last two displays suffices to establish (11.41) in the $p \ge 2$ case.

For $p < 2$, note that $\sum |\mu_j|^p \le C^p$ implies that $|\mu|_{(j)} \le C j^{-1/p}$, and hence that

$$\sum_{j>k}^{n} \mu_{(j)}^2 \le C^{2-p}(k+1)^{1-2/p} \sum_{j>k} |\mu|_{(j)}^p \le C^2 (k+1)^{1-2/p}.$$

We can now dispose of the extreme cases. Putting $k = 0$, we get $\mathcal{R}(\mu) \le C^2$, as is needed for $C \le \sqrt{1 + \log n}$. For $C \ge n^{1/p}$, again use bound (11.42) corresponding to $k = n$.

We now work further on bounding $\mathcal{R}(\mu)$ for the range $C \in [\sqrt{1 + \log n}, n^{1/p}]$. Inserting the last display into (11.40) and ignoring the case $k = n$, we obtain

$$\mathcal{R}(\mu) \le \min_{0 \le k < n} C^2 (k+1)^{1-2/p} + k\lambda_k^2. \tag{11.44}$$

Now observe from the specific form of $L_{n,k}$ that we have $\lambda_{k-1}^2 \le c_4(1 + \log n/k)$ for $2 \le k \le n$. Putting this into (11.44), we arrive at

$$\mathcal{R}(\mu) \le c_4 \min_{1 \le k \le n} \{C^2 k^{1-2/p} + k(1 + \log n/k)\}. \tag{11.45}$$

We now pause to consider the lower bounds, as the structure turns out to be similar enough that we can finish the argument for both bounds at once in part 3° below.

2°. *Lower Bounds.* For $p \ge 2$, we use a hypercube lower bound. Since $\Theta_{n,p}(C)$ contains the cube $[-Cn^{-1/p}, Cn^{-1/p}]^n$, we have by (4.25) and (4.40), with $a_2 = 2/5$,

$$R_N(\Theta) \ge n\rho_N(Cn^{-1/p}, 1) \ge a_2 n \min(C^2 n^{-2/p}, 1).$$

For $p < 2$, we will use products of the single spike parameter sets $\Theta_m(\tau)$ consisting of a single non-zero component in $\mathbb{R}^m$ of magnitude at most τ, compare (8.46). Proposition 8.14 gave a lower bound for minimax mean squared error over such single spike sets.

Working in $\mathbb{R}^n$, for each fixed number k, one can decree that each block of $[n/k]$ successive coordinates should have a single spike belonging to $\Theta_{[n/k]}(\tau)$. Since minimax risk is additive on products, Proposition (4.25), we conclude from Proposition 8.14 that for each k

$$R_N(\Pi_1^k \Theta_{[n/k]}(\tau)) \geq a_3 k(\tau^2 \wedge (1 + \log[n/k]).$$

Now $\Theta_{n,p}(C)$ contains such a product of k copies of $\Theta_{[n/k]}(\tau)$ if and only if $k\tau^p \leq C^p$, so that we may take $\tau = Ck^{-1/p}$ in the previous display. Therefore

$$R_N(\Theta_{n,p}(C)) \geq a_4 \max_{1 \leq k \leq n} C^2 k^{1-2/p} \wedge (k + k \log(n/k)), \tag{11.46}$$

where we also used $1 + \log[x] \geq (1 + \log x)/(1 + \log 2)$ for $x \geq 1$.

Again we draw two quick conclusions: for $C \leq \sqrt{1 + \log n}$, the choice $k = 1$ yields the bound C^2, while for $C \geq n^{1/p}$, the choice $k = n$ gives the lower bound n.

3°. *Completion of proof.* Let us summarize the remaining task. Define two functions

$$g(x) = C^2 x^{1-2/p}, \qquad h(x) = x + x \log(n/x).$$

Then, for $\sqrt{1 + \log n} \leq C \leq n^{1/p}$, and $r(C) = C^p\big[1 + \log(n/C^p)\big]^{1-p/2}$, with $p < 2$, we seek absolute constants a_5 and a_6 so that

$$\begin{aligned} a_5 r(C) &\leq \max_{1 \leq k \leq n} g(k) \wedge h(k) \\ &\leq \min_{1 \leq k \leq n} g(k) + h(k) \leq a_6 r(C). \end{aligned} \tag{11.47}$$

Since g is decreasing and h is increasing for $0 \leq x \leq n$, it is natural to look for $x_\star = x_\star(C) \in \mathbb{R}$ at which $g(x_\star) = h(x_\star)$, compare Figure 11.2. At the point of intersection,

$$x_\star = C^p\big[1 + \log(n/x_\star)\big]^{-p/2}, \tag{11.48}$$

$$g(x_\star) = C^p\big[1 + \log(n/x_\star)\big]^{1-p/2}. \tag{11.49}$$

It is clear from Figure 11.2 that $C \to x_\star(C)$ is strictly increasing, with

$$x_\star(\sqrt{1 + \log n}) = 1, \quad \text{and} \quad x_\star(n^{1/p}) = n.$$

Hence $1 \leq x_\star \leq n$ if and only if $\sqrt{1 + \log n} \leq C \leq n^{1/p}$; this explains the choice of transition points for C in the definition of $r(C)$.

We now relate the intersection value $g(x_\star(C))$ to $r(C)$; we will show that

$$r(C) \leq g(x_\star(C)) \leq 2r(C). \tag{11.50}$$

One direction is easy: putting $x_\star \leq n$ into (11.48) shows that $x_\star \leq C^p$, and hence from (11.49) that $g(x_\star) \geq r(C)$. For the other direction, make the abbreviations

$$s = 1 + \log(n/x_\star), \qquad \text{and} \qquad t = 1 + \log(n/C^p).$$

Now taking logarithms in equation (11.48) shows that $s \leq t + \log s$. But $\log s \leq s/2$ (since $s \geq 1$ whenever $x_\star \leq n$), and so $s \leq 2t$. Plugging this into (11.49), we obtain (11.50).

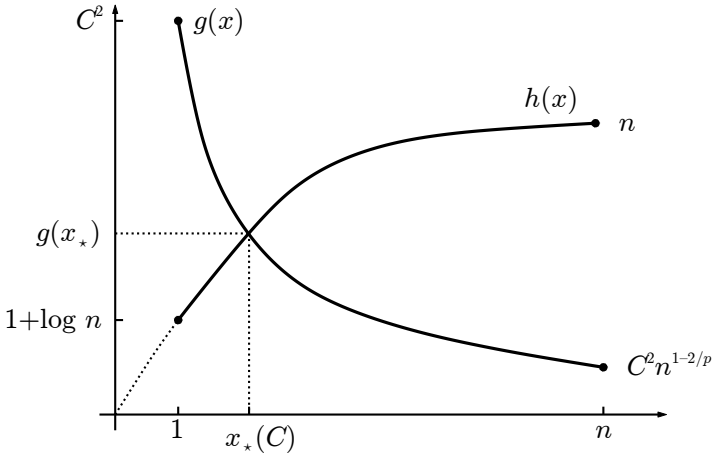

Figure 11.2 Diagram of functions g and h and their intersection, when $p < 2$ and $\sqrt{1+\log n} \le C \le n^{1/p}$.

A detail. We are not quite done since the extrema in the bounds (11.47) should be computed over *integers* $k, 1 \le k \le n$. The following remark is convenient: for $1 \le x \le n$, the function $h(x) = x + x\log(n/x)$ satisfies

$$\tfrac{1}{2}h(\lceil x \rceil) \le h(x) \le 2h(\lfloor x \rfloor). \tag{11.51}$$

Indeed, h is concave and $h(0) = 0$, and so for x positive, $h(x)/2 \le h(x/2)$. Since h is increasing for $0 \le y \le n$, it follows that if $x \le 2y$, then $h(x) \le 2h(y)$. Since $x \ge 1$ implies both $x \le 2\lfloor x \rfloor$ and $\lceil x \rceil \le 2x$, the bounds (11.51) follow.

For the upper bound in (11.47), take $k = \lceil x_\star \rceil$: since g is decreasing, and using (11.51) and then (11.50), we find

$$\min_{1\le k\le n} g + h \le (g+h)(\lceil x_\star \rceil) \le g(x_\star) + 2h(x_\star) = 3g(x_\star) \le 6r(C).$$

For the lower bound, take $k = \lfloor x_\star \rfloor$, and again from the same two displays,

$$\max_{1\le k\le n} g \wedge h \ge (g \wedge h)(\lfloor x_\star \rfloor) = h(\lfloor x_\star \rfloor) \ge \tfrac{1}{2}h(x_\star) = \tfrac{1}{2}g(x_\star) \ge \tfrac{1}{2}r(C). \qquad \square$$

11.6 Aside: Stepwise methods vs. complexity penalization.

Stepwise model selection methods have long been used as heuristic tools for model selection. In this aside, we explain a connection between such methods and a class of penalties for penalized least squares.

The basic idea with stepwise methods is to use a test statistic—in application, often an F-test—and a threshold to decide whether to add or delete a variable from the current fitted model. Let $\hat{J}_k$ denote the best submodel of size k:

$$\hat{J}_k = \text{argmax}_k\{\|P_J y\|^2 : n_J = k\},$$

and denote the resulting best k-variable estimator by $Q_k y = P_{\hat{J}_k} y$. The mapping $y \to Q_k(y)$ is non-linear since the optimal set $\hat{J}_k(y)$ will in general vary with y.

In the *forward stepwise* approach, the model size is progressively increased until a threshold criterion suggests that no further benefit will accrue by continuing. Thus, define

$$\hat{k}_G = \text{first } k \text{ s.t. } \|Q_{k+1} y\|^2 - \|Q_k y\|^2 \le \epsilon^2 t^2_{p,k+1}. \tag{11.52}$$

Note that we allow the threshold to depend on k: in practice it is often constant, but we wish to allow $k \to t^2_{p,k}$ to be decreasing.

In contrast, the *backward stepwise* approach starts with a saturated model and gradually decreases model size until there appears to be no further advantage in going on. So, define

$$\hat{k}_F = \text{last } k \text{ s.t. } \|Q_k y\|^2 - \|Q_{k-1} y\|^2 \ge \epsilon^2 t^2_{p,k}. \tag{11.53}$$

Remarks. 1. In the orthogonal case, $y_i = \mu_i + \epsilon z_i, i = 1, \ldots, n$ with order statistics $|y|_{(1)} \ge |y|_{(2)} \ge \ldots \ge |y|_{(n)}$, we find that

$$\|Q_k y\|^2 = \sum_{j=1}^{k} |y|^2_{(j)},$$

so that

$$\hat{k}_F = \max\{k : |y|_{(k)} \ge \epsilon t_{p,k}\}, \tag{11.54}$$

and that $\hat{k}_F$ agrees with the FDR definition (7.26) with $t_{p,k} = z(qk/2n)$. In this case, it is critical to the method that the thresholds $k \to t_{p,k}$ be (slowly) decreasing.

2. In practice, for reasons of computational simplicity, the forward and backward stepwise algorithms are often "greedy", i.e., they look for the best variable to add (or delete) without optimizing over all sets of size k.

The stepwise schemes are related to a penalized least squares estimator. Let

$$\begin{aligned} S(k) &= \|y - Q_k y\|^2 + \epsilon^2 \sum_{j=1}^{k} t^2_{p,j}, \\ \hat{k}_2 &= \text{argmin}_{0 \le k \le n}\, S(k). \end{aligned} \tag{11.55}$$

Thus the associated penalty function is $\text{pen}(k) = \sum_1^k t^2_{p,j}$ and the corresponding estimator is given by (11.2) and (11.3).

The optimal model size for $\text{pen}(k)$ is bracketed between the stepwise quantities.

Proposition 11.8 *Let $\hat{k}_G, \hat{k}_F$ be the forward and backward stepwise variable numbers defined at (11.52) and (11.53) respectively, and let $\hat{k}_2$ be the global optimum model size for pen(k) defined at (11.55). Then*

$$\hat{k}_G \le \hat{k}_2 \le \hat{k}_F.$$

Proof Since $\|y - Q_k y\|^2 = \|y\|^2 - \|Q_k y\|^2$,

$$S(k+1) - S(k) = \|Q_k y\|^2 - \|Q_{k+1} y\|^2 + \epsilon^2 t^2_{p,k+1}.$$

Thus

$$S(k+1) \begin{Bmatrix} < \\ = \\ > \end{Bmatrix} S(k) \quad \text{according as} \quad \|Q_{k+1}y\|^2 - \|Q_k y\|^2 \begin{Bmatrix} > \\ = \\ < \end{Bmatrix} \epsilon^2 t^2_{p,k+1}.$$

Thus, if it were the case that $\hat{k}_2 > \hat{k}_F$, then necessarily $S(\hat{k}_2) > S(\hat{k}_2 - 1)$, which would contradict the definition of $\hat{k}_2$ as a global minimum of $S(k)$. Likewise, $\hat{k}_2 < \hat{k}_G$ is not possible, since it would imply that $S(\hat{k}_2 + 1) < S(\hat{k}_2)$. □

11.7 A variant for use in inverse problems

Suppose that $y = \theta + \epsilon z$, now with z assumed to be zero-mean Gaussian, but *weakly correlated*: i.e.

$$\xi_0 I \preceq \mathrm{Cov}(z) \preceq \xi_1 I, \tag{11.56}$$

where $A \preceq B$ means that $B - A$ is non-negative definite. We modify the penalty to $\mathrm{pen}(k) = \zeta\xi_1 k(1 + \sqrt{2L_k})^2$. We want to replace the constant M in the variance term in (11.15) by one that excludes the zero model:

$$M' = \sum_{J \neq \{0\}} e^{-L_J n_J}.$$

Theorem 11.9 *Consider observations in the weakly correlated model* (11.56)*. Let $\hat{\mu}$ be a penalized least squares estimator of (11.2)- (11.3) for a penalty and constant M' as defined above. There exists a constant $K = K(\zeta)$ such that*

$$E\|\hat{\mu} - \mu\|^2 \leq K[2M'\xi_1\epsilon^2 + \inf_J C(J, \mu)]. \tag{11.57}$$

The value of K may be taken as in Theorem 11.3.

Proof 1°. We modify the proof of the previous theorem in two steps. First fix J and assume that $\mathrm{Cov}(z) = I$. Let $\Omega'_x = \cap_{J' \neq \{0\}} E_{J'}(x)$ and $X' = \inf\{x : \omega \notin \Omega'_x\}$. On the set $\hat{J} \neq \{0\}$, we have, as before,

$$\|\hat{\mu}_{\hat{J}} - \mu\|^2 \leq A(\eta)C(J, \mu) + B(\eta)\epsilon^2 X'.$$

Now consider the event $\hat{J} = \{0\}$. First, note that if $\|\mu\|^2 \leq \epsilon^2 \mathrm{pen}(1)$, we have on $\hat{J} = \{0\}$ that, for all J

$$\|\hat{\mu}_{\hat{J}} - \mu\|^2 = \|\mu\|^2 \leq C(J, \mu).$$

Suppose, instead, that $\|\mu\|^2 \geq \epsilon^2 \mathrm{pen}(1)$, so that $C(J, \mu) \geq \epsilon^2 \mathrm{pen}(1)$ for all J—here we use the monotonicity of $k \to \mathrm{pen}(k)$. Pick a J' with $n_{J'} = 1$; on Ω'_x we have

$$\langle z, -\mu_J \rangle \leq \|\mu_J\| \cdot \chi_{J,J'} \leq \|\mu_J\| \cdot [(1 + \sqrt{2L_1}) + \sqrt{n_J} + \sqrt{2x}].$$

We now proceed as in the argument from (11.22) to (11.23), except that we bound $\epsilon^2 \mathrm{pen}(1) \leq C(J, \mu)$, concluding that on Ω'_x and $\hat{J} = \{0\}$, we may use in place of (11.17)

$$\begin{aligned} 2\epsilon\langle z, -\mu_J \rangle \leq (1 - \eta^2)\|\mu\|^2 &+ C(J, \mu) \\ &+ a(\eta)C(J, \mu) + b(\eta)\epsilon^2 x. \end{aligned}$$

Consequently, combining all cases

$$\|\hat{\mu}_{\hat{J}} - \mu\|^2 \le \eta^{-2}(2 + a(\eta))C(J, \mu) + \eta^{-2}b(\eta)\epsilon^2 X',$$

which might be compared with e (11.18). Taking expectations, then minimizing over J, we obtain again (11.19), this time with $A(\eta) = \eta^{-2}(2 + a(\eta))$. In (11.24), now $K(\zeta) = \eta^{-3}\max(2 + 2\eta, 2\zeta)$, but since in fact $\zeta > 1 + \eta$ for all η, the value for $K(\zeta)$ is unchanged.

2°. The extension to weakly correlated z is straightforward. We write $y = \mu + \epsilon_1 z_1$, where $\epsilon_1 = \sqrt{\xi}_1\epsilon$ and $\Sigma_1 = \mathrm{Cov}(z_1) \prec I$. We apply the previous argument with ϵ, z replaced by ϵ_1 and z_1. The only point where the stochastic properties of z_1 are used is in the concentration inequality that is applied to $\chi_{J,J'}$. In the present case, if we put $z_1 = \Sigma_1^{1/2}Z$ for $Z \sim N(0, I)$, we can write

$$\chi_{J,J'} = \|P\,\Sigma_1^{1/2}Z\|,$$

where P denotes orthoprojection onto $S_J \oplus S_{J'}$. Since $\lambda_1(\Sigma_1^{1/2}) \le 1$, the map $Z \to \chi_{J,J'}(Z)$ is Lipschitz with constant at most 1, so that the concentration bound applies. □

In particular, we will in Chapter 12 make use of penalties for which

$$L_{n,k} = (1 + 2\beta)\log(\gamma_n n/k) \tag{11.58}$$

with $\gamma_n = \gamma\log^2 n$. For this choice, the constant M' in (11.57) satisfies (after using the Stirling formula bound $k! > \sqrt{2\pi k}\,k^k e^{-k}$),

$$\begin{aligned} M' \le \sum_{k=1}^{n} \frac{n^k}{k!}\Big(\frac{k}{n\gamma_n}\Big)^{k(1+2\beta)} &\le \sum_{k=1}^{\infty} \frac{1}{\sqrt{2\pi k}}\Big(\frac{k^{2\beta}}{n^{2\beta}}\frac{e}{\gamma_n^{1+2\beta}}\Big)^k \\ &\le \frac{1}{n^{2\beta}\gamma_n}\sum_{k\ge 1}\frac{k^{2\beta}e}{\sqrt{2\pi k}}\Big(\frac{e}{\gamma_n^{1+2\beta}}\Big)^{k-1} \le \frac{C_{\beta,\gamma}}{n^{2\beta}\gamma_n}, \end{aligned} \tag{11.59}$$

so long as either $\gamma_n \equiv \gamma > e^{1/(1+2\beta)}$ or $\gamma_n = \log^2 n$.

11.8 Notes

The formulation and proof of Theorem 11.3 is borrowed from Birgé and Massart (2001). Earlier versions in `[D-J, fill in.]`

2. The formulation and methods used for Theorem 11.7 are inspired by Birgé and Massart (2001).

Exercises

11.1 *(Gaussian quantiles and $2k\log(n/k)$ penalties.)* Define the Gaussian quantile $z(\eta)$ by the equation $\tilde{\Phi}(z(\eta)) = \eta$.
(a) Use (8.63) to show that

$$z^2(\eta) = 2\log\eta^{-1} - \log\log\eta^{-1} - r(\eta),$$

and that when $\eta \le 0.01$, we have $1.8 \le r(\eta) \le 3$ (Abramovich et al., 2006).

(b) Show that $z'(\eta) = -1/\phi(z(\eta))$ and hence that if $\eta_2 > \eta_1 > 0$, then

$$z(\eta_2) - z(\eta_1) \leq \frac{\eta_2 - \eta_1}{\eta_1 z(\eta_1)}.$$

(c) Verify (11.10) and (11.11).

12

Exact rates for estimation on Besov spaces

The claim made for wavelet shrinkage has been that it takes advantage of the local nature of wavelet basis functions to achieve spatial adaptation to inhomogeneous smoothness.

We have modelled inhomogeneous smoothness theoretically using Besov spaces $B^{\alpha}_{p,q}$ for $p < 2$, and their sequence space norm balls $\Theta^{\alpha}_{p,q}(C)$.

In studying $\sqrt{2\log n}$ thresholding, we showed that it was adaptively optimal up to a logarithmic factor of order $2\log\epsilon^{-1}$. That is, we showed under conditions give in Theorem 9.11 that for a wide range of Besov balls $\Theta = \Theta^{\alpha}_{p,q}$, as $\epsilon \to 0$,

$$\sup_{\Theta} E\|\hat{\theta}^U - \theta\|^2 \le c(\log\epsilon^{-1})R_N(\Theta,\epsilon)(1+o(1)).$$

While this already a quite strong adaptivity statement, the extra $\log\epsilon^{-1}$ is undesirable, and indeed reflects a practically important phenomenon: $\sqrt{2\log n}$ thresholds can be too high in some settings, for example in Figure 7.5, and lower choices of threshold can yield much improved reconstructions and MSE performance.

In the first section of this chapter, we apply the $2k\log n/k$ oracle inequality of Chapter 11 and its ℓ_p ball consequences to show that appropriate penalized least squares estimates (i.e. data dependent thresholding) adapt exactly to the correct rates of convergence over essentially all reasonable Besov bodies. Thus, we show that for an explicit $\hat{\theta}^P$,

$$\sup_{\Theta} E\|\hat{\theta}^P - \theta\|^2 \le cR_N(\Theta,\epsilon)(1+o(1))$$

simultaneously for all $\Theta = \Theta^{\alpha}_{p,q}(C)$ in a large set of values for (α, p, q, C).

Our approach is based on the inequalities of Chapter 11.5, which showed that the ℓ_p-ball minimax risk could, up to multiplicative constants, be described by the relatively simple control functions $r_{n_j,p}(C,\epsilon)$ defined there. The device of "Besov shells", consisting of vectors $\theta \in \Theta$ that vanish except on level j, and hence equivalent to ℓ_p-balls, allows the study of minimax risks on Θ to be reduced to the minimax risks and hence control functions $R_j = r_{n_j,p}(C_j,\epsilon_j)$ where the parameters $(n_j = 2^j, C_j, \epsilon_j)$ vary with j. Accordingly, a study of the shell bounds $j \to R_j$ yields our sharp rate results. Since this works for both direct and indirect estimation models, it is postponed to Section 12.5.

Also in this chapter, we finally return to the theme of linear inverse problems, introduced in Chapter 3 with the goal of broadening the class of examples to which the Gaussian sequence model applies. We now wish to see what advantages can accrue through using thresholding and wavelet bases, to parallel what we have studied at length for direct estimation in the white noise model.

We describe an alternative to the singular value decomposition, namely the *wavelet-vaguelette* decomposition, for a class of linear operators. The left and right singular function systems of the SVD are replaced by wavelet-like systems which still have multiresolution structure and yield sparse representations of functions with discontinuities. The function systems are not exactly orthogonal, but they are *nearly* orthogonal, in the sense of 'frames', and are in fact a sufficient substitute for analyzing the behavior of threshold estimators.

In Section 12.2, then, we indicate some drawbacks of the SVD for object functions with discontinuities and introduce the elements of the WVD.

Section 12.3 lists some examples of linear operators A having a WVD, including integration of integer and fractional orders, certain convolutions and the Radon transform. The common feature is that the stand-in for singular values, the *quasi-singular* values, decay at a rate algebraic in the number of coefficients, $\kappa_j \approx 2^{-\beta j}$ at level j.

Section 12.4 focuses on a particular idealisation, motivated by the WVD examples, that we call the "correlated levels model", cf (12.31). This generalizes the white noise model by allowing noise levels $\epsilon_j = 2^{\beta j}\epsilon$ that grow in magnitude with resolution level j, a key feature in inverting data in ill-posed inverse problems. In addition, the model allows for the kind of near-independence correlation structure of noise that appears in problems with a WVD.

Using co-ordinatewise thresholding–with larger thresholds chosen to handle the variance inflation with level–we easily recover the optimal rate of convergence up to a logarithmic factor. This analysis already makes it possible to show improvement in the rates of convergence, compared to use of the SVD, that are attainable by exploiting sparsity of representation in the WVD.

By returning to the theme of penalized least squares estimation with $2n \log n/k$ penalties, we are again able to dispense with the logarithmic terms in the rates of convergence in the correlated levels model. This is also done in Section 12.4.

12.1 Direct estimation

We consider the projected sequence model

$$y_{jk} = \theta_{jk} + \epsilon z_{jk} \qquad k = 1, \dots, 2^j;\ j = 0, \dots, J-1 \tag{12.1}$$

with $z_{jk} \sim N(0,1)$ independently and $J = \log_2 \epsilon^{-2}$.

Estimator: Use a penalized least squares estimator at each level j separately. We use a penalty of $2n \log(n/k)$ type used in Section 11.5, but with $n = n_j = 2^j$. Thus, introduce

$$\mathrm{pen}_j(k) = k\lambda_{j,k}^2, \qquad \lambda_{j,k} = \sqrt{\zeta}\big(1 + \sqrt{2\log(2^j\gamma/k)}\big),$$

and define

$$\hat{\theta}_P(y_j) = \operatorname*{argmin}_{\theta_j} \|y_j - \theta_j\|^2 + \epsilon^2 \mathrm{pen}_j(N(\theta_j)), \tag{12.2}$$

where $N(\theta_j)$ denotes the number of non-zero entries in θ_j. According to Lemma 11.1, this is an equivalent form of the complexity penalized projection estimator (11.2)–(11.3) to which the oracle inequality of Theorem 11.3 applies. From Proposition 11.2 and Lemma 11.4, it is

equivalent to hard thresholding at $\hat{t}_j = t_{\hat{k}_j} \approx \lambda_{j,\hat{k}_j}$ where $\hat{k}_j = N(\hat{\theta}_P(y_j))$ is the number of non-zero entries in $\hat{\theta}_P(y_j)$.

We put these levelwise estimates together to get a wavelet penalized least squares estimate $\hat{\theta}^P = (\hat{\theta}^P_j)$:

$$\hat{\theta}^P_j(y) = \begin{cases} \hat{\theta}_P(y_j) & j < J \\ 0 & j \geq J. \end{cases}$$

Theorem 12.1 *Let $\hat{\theta}^P$ be the wavelet penalized least squares estimate described above, and assume that $\gamma > e$. For $\alpha > (1/p - 1/2)_+$ along with $0 < p, q \leq \infty$ and $C > 0$, there exist constants c_i such that*

$$\begin{aligned} c_0 C^{2(1-r)}\epsilon^{2r} &\leq R_N(\Theta^\alpha_{p,q}(C), \epsilon) \\ &\leq \sup_{\Theta^\alpha_{p,q}(C)} E\|\hat{\theta}^P - \theta\|^2 \leq c_3 C^{2(1-r)}\epsilon^{2r} + c_2 C^2(\epsilon^2)^{2\alpha'} + c_1\epsilon^2 \log \epsilon^{-2}, \end{aligned}$$

where $r = 2\alpha/(2\alpha+1)$ while $\alpha' = \alpha$ if $p \geq 2$ and $a = \alpha + 1/2 - 1/p$ if $p < 2$.

Remarks. 1. The dependence of the constants on the parameters defining the estimator and Besov space is given by $c_1 = c_1(\zeta, \gamma), c_2 = c_2(\alpha, p)$ and $c_3 = c_3(\zeta, \gamma, \alpha, p)$.

2. Let us examine when the $C^{2(1-r)}\epsilon^{2r}$ term dominates as $\epsilon \to 0$. Since $r < 1$, the $\epsilon^2 \log^2 \epsilon^{-2}$ term is always negligible. If $p \geq 2$, then $2\alpha' = 2\alpha > r$ and so the tail bias term is also of smaller order. If $p < 2$, a convenient extra assumption is that $\alpha \geq 1/p$, for then $\alpha' = a \geq 1/2 > r/2$. Note that the condition $\alpha \geq 1/p$ is necessary for the Besov space $B^\alpha_{p,q}$ to embed in spaces of continuous functions.

3. One may ask more explicitly for what values of ϵ the tail bias $C^2(\epsilon^2)^{2\alpha'} < C^{2(1-r)}\epsilon^{2r}$. Simple algebra shows that this occurs when

$$\epsilon < C^{-r/(2\alpha'-r)},$$

showing the key role of the radius C.

Proof *Upper bound.* The levelwise structure of $\hat{\theta}^P$ yields the MSE decomposition

$$E\|\hat{\theta}^P - \theta\|^2 = \sum_{j<J} E\|\hat{\theta}_P(y_j) - \theta_j\|^2 + \Delta_J(\theta), \tag{12.3}$$

where $\Delta_J(\theta) = \sum_{j \geq J} \|\theta_j\|^2$ is the "tail bias" due to not estimating beyond level J. The maximum tail bias over $\Theta^\alpha_{p,q}(C)$ was evaluated at (9.50) and yields the bound $c_2(\alpha, p)C^2(\epsilon^2)^{2\alpha'}$.

Applying Theorem 11.3 on each level j, we obtain a bound

$$E\|\hat{\theta}_P(y_j) - \theta_j\|^2 \leq c_1\epsilon^2 + c_1\mathcal{R}_j(\theta_j, \epsilon), \tag{12.4}$$

where in accordance with (11.29), the level j theoretical complexity is given by

$$\mathcal{R}_j(\theta_j, \epsilon) = \min_{0 \leq k \leq n_j} \sum_{l>k} |\theta_j|^2_{(l)} + \epsilon^2 k \lambda^2_{j,k}, \tag{12.5}$$

where $|\theta_j|^2_{(l)}$ denotes the l-th largest value among $\{\theta^2_{jk}, j = 1, \ldots, 2^k\}$.

Summing over $j < J = \log_2 \epsilon^{-2}$, the first term on the right side of (12.4) yields the $c_1\epsilon^2 \log \epsilon^{-2}$ term in Theorem (12.3).

To bound $\sum_j \mathcal{R}_j(\theta_j, \epsilon)$ we use the Besov shells $\Theta^{(j)} = \{\theta \in \Theta : \theta_I = 0 \text{ for } I \notin \mathcal{I}_j\}$ introduced in Section 10.7, and their interpretation as ℓ_p–balls: $\Theta^{(j)} \equiv \Theta_{n_j,p}(C_j)$ for $n_j = 2^j$ and $C_j = C2^{-aj}$. The maximum of $\mathcal{R}_j(\theta_j, \epsilon)$ over Θ can therefore be obtained by maximizing over $\Theta^{(j)}$ alone, and so

$$\sup_{\Theta} \sum_j \mathcal{R}_j(\theta_j, \epsilon) \le \sum_j \sup_{\Theta^{(j)}} \mathcal{R}_j(\theta_j, \epsilon).$$

The maximization of theoretical complexity over ℓ_p-balls was studied in detail in Section 11.5. Let $r_{n,p}(C, \epsilon)$ be the control function for minimax mean squared error at noise level ϵ. The proof of Theorem 11.7 yields the bound

$$\mathcal{R}_j(\theta_j, \epsilon) \le c_2 r_{n_j,p}(C_j, \epsilon)$$

for $\theta_j \in \Theta_{n_j,p}(C_j)$, compare (11.41). Combining the two previous displays shows that we need to bound $\sum_j r_{n_j,p}(C_j, \epsilon)$.

In Section 12.5, we show that the shell bounds $R_j = r_{n_j,p}(C_j, \epsilon)$ peak at a critical level j_*, and decay geometrically away from the value R_* at this least favorable level, so that the series is indeed summable. So the final bound we need, namely

$$\sup_{\Theta} \sum_j \mathcal{R}_j(\theta_j, \epsilon) \le c_3 C^{2(1-r)} \epsilon^{2r}$$

follows from Propostion (12.37).

Lower bound. We saw already in Theorem 9.11 that $R_N(\Theta, \epsilon) \ge c C^{2(1-r)} \epsilon^{2r}$, but we can rewrite the argument using Besov shells and control functions for ℓ_p balls. Since each shell $\Theta^{(j)} \subset \Theta$, we have

$$R_N(\Theta, \epsilon) \ge R_N(\Theta^{(j)}, \epsilon) \ge R_N(\Theta_{n_j,p}(C_j), \epsilon) \ge a_1 r_{n_j,p}(C_j, \epsilon),$$

by the lower bound part of Theorem 11.7. Consequently $R_N(\Theta, \epsilon) \ge a_1 \max_j R_j$, and that this is bounded below by $c_1 C^{2(1-r)} \epsilon^{2r}$ is also shown in Proposition (12.37). □

12.2 Wavelet-Vaguelette Decomposition

Stochastic observation model. Let A be a linear operator from $\mathcal{D}(A) \subset L_2(T)$ to $\mathcal{R}(A) \subset L_2(U)$. We consider an idealized model in which Af is observed in additive Gaussian noise. We assume that we observe

$$Y = Af + \epsilon Z, \tag{12.6}$$

which is interpreted to mean that, for all $g \in L_2(U)$, we have

$$Y(g) = [Af, g] + \epsilon Z(g), \tag{12.7}$$

and the process $g \to Z(g)$ is Gaussian, with zero mean and covariance

$$\mathrm{Cov}(Z(g), Z(h)) = [g, h]. \tag{12.8}$$

A defect of the Singular Value Decomposition.

Suppose that $A : L_2(T) \to L_2(U)$ is a linear operator with singular value decomposition

$Ae_k = \lambda_k h_k$ in terms of orthogonal singular systems $\{e_k\}$ for $L_2(T)$ and $\{h_k\}$ for $L_2(U)$. In the examples of Chapter 3.8, and more generally, the singular functions are 'global' functions, supported on all of T and U respectively. Consequently, the representation of a smooth function with isolated singularities may not be sparse.

Consider a simple example in which $\{e_k\}$ is a trigonometric basis on $[0, 1]$ and f is a (periodic) step function, such as $I_{[1/4,3/4]}(t)$. If A is a convolution with a periodic kernel $a(t)$ with coefficients $\lambda_k = \langle a, e_k \rangle$, then in Chapter 3.8 we derived the sequence model $y_k = \theta_k + \epsilon_k z_k$ with $\epsilon_k = \epsilon/\lambda_k$. The coefficients $\theta_k = \langle f, e_k \rangle$ would typically have slow decay with frequency k, of order $|\theta_k| \asymp O(1/k)$. The (ideal) best linear estimator of form $(c_k y_k)$ for the given θ has the form

$$\inf_c r(\hat{\theta}_c, \theta) = \sum_k \frac{\theta_k^2 \epsilon_k^2}{\theta_k^2 + \epsilon_k^2} \asymp \sum_k \theta_k^2 \wedge \frac{\epsilon^2}{\lambda_k^2}. \tag{12.9}$$

For a typical convolution operator A, the singular values λ_k decrease quite quickly, while the coefficients θ_k do not. Hence even the ideal linear risk for a step function in the Fourier basis is apt to be uncomfortably large.

We might instead seek to replace the SVD bases by wavelet bases, in order to take advantage of wavelets' ability to achieve sparse representations of smooth functions with isolated singularities. As a running example for exposition, suppose that A is given by integration on $\mathbb{R}$:

$$(Af)(u) = f^{(-1)}(u) = \int_{-\infty}^{u} f(t)dt. \tag{12.10}$$

Let $\{\psi_\lambda\}$ be a nice orthonormal wavelet basis for $L_2(\mathbb{R})$: as usual we use λ for the double index (j, k), so that $\psi_\lambda(t) = 2^{j/2}\psi(2^j t - k)$. We may write

$$\begin{aligned} A\psi_\lambda(u) &= \int_{-\infty}^{u} 2^{j/2}\psi(2^j t - k)dt = 2^{-j} \cdot 2^{j/2}(\psi^{(-1)})(2^j u - k) \\ &= 2^{-j}(\psi^{(-1)})_\lambda(u). \end{aligned}$$

The initial difficulty is that $\{u_\lambda := (\psi^{(-1)})_\lambda\}$ is not orthonormal in the way that $\{\psi_\lambda\}$ is.

Suppose initially that we consider an arbitrary orthonormal basis $\{e_k\}$ for $L_2(T)$, so that $f = \sum \langle f, e_k \rangle e_k$. Suppose also that we can find *representers* $g_k \in L_2(U)$ for which

$$\langle f, e_k \rangle = [Af, g_k].$$

According to Proposition C.5, this occurs when each $e_k \in \mathcal{R}(A^*)$. The corresponding sequence of observations $Y_k = Y(g_k)$ has mean $[Af, g_k] = \langle f, e_k \rangle$ and covariance $\epsilon^2 \Sigma_{kl}$ where $\Sigma_{kl} = \mathrm{Cov}(Z(g_k), Z(g_l)) = [g_k, g_l]$. We might then consider using estimators of the form $\hat{f} = \sum_k c_k(Y_k)e_k$ for co-ordinatewise functions $c_k(Y_k)$, which might be linear or threshold functions. However, Proposition 4.26 shows that in the case of diagonal linear estimators and suitable parameter sets, the effect of the correlation of the Y_k on the efficiency of estimation is determined by $\lambda_{\min}(\rho(\Sigma))$, the minimum eigenvalue of the correlation matrix corresponding to covariance Σ. In order for this effect to remain bounded even as the noise level $\epsilon \to 0$, we need the representers g_k to be nearly orthogonal in an appropriate sense.

To see this, set $u_k = g_k/\|g_k\|_2$, and observe that

$$\begin{aligned}\lambda_{\min}(\rho(\Sigma)) &= \inf\{\alpha^T\rho(\Sigma)\alpha : \|\alpha\|_2 = 1\} \\ &= \inf\left\{\mathrm{Var}\left(\sum \frac{\alpha_k}{\|g_k\|} Y(g_k)\right) : \|\alpha\|_2 = 1\right\} \\ &= \inf\left\{\Big\|\sum \alpha_k u_k\Big\|^2 : \|\alpha\|_2 = 1\right\}.\end{aligned}$$

Hence, we obtain the necessary control if the normalized representers satisfy a bound

$$\Big\|\sum \alpha_k u_k\Big\|_2 \geq c\|\alpha\|_2 \qquad \text{for all } \alpha \in \ell_2. \tag{12.11}$$

We will see that this is indeed often possible if one starts with a wavelet basis $\{\psi_\lambda\}$ for $L_2(T)$.

Remark. In developing the WVD, it is convenient initially to take $T = \mathbb{R}$ to avoid boundary effects, and to exploit translation invariance properties of $\mathbb{R}$. In such cases, it may be that the operator A is only defined on a dense subset $\mathcal{D}(A)$ of $L_2(T)$. For example, with integration, (12.10), the Fourier transform formula $\widehat{Af}(\xi) = (i\xi)^{-1}\hat{f}(\xi)$ combined with the Parseval relation (C.8) shows that $\widehat{Af} \in L_2(\mathbb{R})$ if and only if f belongs to the subset of $L_2(\mathbb{R})$ defined by $\int |\xi|^{-2}|\hat{f}(\xi)|^2 d\xi < \infty$. Similarly, using $A^*g = \int_u^\infty g(t)dt$, it follows that $\mathcal{R}(A^*)$ is the subset of L_2 corresponding to $\int |\xi|^2|\hat{f}(\xi)|^2 d\xi < \infty$.

Let us turn again to wavelet bases. Suppose that $\{\psi_\lambda\}$ is an orthonormal wavelet basis for $L_2(T)$ such that $\psi_\lambda \in \mathcal{D}(A) \cap \mathcal{R}(A^*)$ for every λ. Proposition C.5 provides a representer g_λ such that

$$\langle f, \psi_\lambda\rangle = [Af, g_\lambda]. \tag{12.12}$$

Suppose, in addition, that $\|g_\lambda\| = c\kappa_j^{-1}$ is independent of k. Define two systems $\{u_\lambda\}, \{v_\lambda\} \in L_2(U)$ by the equations

$$u_\lambda = \kappa_j g_\lambda, \qquad v_\lambda = \kappa_j^{-1} A\psi_\lambda. \tag{12.13}$$

Since for every $f \in \mathcal{D}(A)$ we have $\langle f, A^*u_\lambda\rangle = [Af, \kappa_j g_\lambda] = \langle f, \kappa_j\psi_\lambda\rangle$, we may conclude that

$$A^*u_\lambda = \kappa_j\psi_\lambda, \qquad A\psi_\lambda = \kappa_j v_\lambda. \tag{12.14}$$

In addition, the $\{u_\lambda\}$ and $\{v_\lambda\}$ systems are *biorthogonal*:

$$[v_\lambda, u_\mu] = \kappa_j^{-1}\kappa_{j'}[A\psi_\lambda, g_\mu] = \kappa_j^{-1}\kappa_{j'}\langle\psi_\lambda, \psi_\mu\rangle = \delta_{\lambda\mu}. \tag{12.15}$$

Since $\langle f, \psi_\lambda\rangle = [Af, g_\lambda] = \kappa_j^{-1}[Af, u_\lambda]$, we have the formal reproducing formula

$$f = \sum\langle f, \psi_\lambda\rangle\psi_\lambda = \sum\kappa_j^{-1}[Af, u_\lambda]\psi_\lambda. \tag{12.16}$$

Example. Let A again correspond to integration. Suppose that the wavelet ψ is C^1, with compact support and $\int\psi = 0$, so that $\psi \in \mathcal{D}(A) \cap \mathcal{R}(A^*)$. Then formula (12.12) and integration by parts shows that the representer

$$g_\lambda = -(\psi_\lambda)' = -2^j(\psi')_\lambda.$$

Since $\|g_\lambda\|_2 = c_\psi 2^j$, with $c_\psi = \|\psi'\|_2$, we can set $\kappa_j = 2^{-j}$, and then from (12.13),

$$u_\lambda = -(\psi')_\lambda, \qquad v_\lambda = (\psi^{(-1)})_\lambda.$$

We now turn to showing that the (non-orthogonal) systems $\{u_\lambda\}$ and $\{v_\lambda\}$ satisfy (12.11).

To motivate the next definition, note that members of both systems $\{u_\lambda\}$ and $\{v_\lambda\}$ have, in our example, the form $w_\lambda(t) = 2^{j/2} w(2^j t - k)$. If we define a rescaling operator

$$(S_\lambda w)(x) = 2^{-j/2} w(2^{-j}(x + k)), \tag{12.17}$$

then in our example, but not in general, $(S_\lambda w_\lambda)(t) = w(t)$ is free of λ.

Definition 12.2 A collection $\{w_\lambda\} \subset L_2(\mathbb{R})$ is called a system of *vaguelettes* if there exist positive constants C_1, C_2 and exponents $0 < \eta < \eta' < 1$ such that for each λ, the rescaled function $\tilde{w} = S_\lambda w_\lambda$ satisfies

$$\tilde{w}(t) \le C_1 (1 + |t|)^{-1-\eta'}, \tag{12.18}$$

$$\int \tilde{w}(t) dt = 0 \tag{12.19}$$

$$|\tilde{w}(t) - \tilde{w}(s)| \le C_2 |t - s|^\eta \tag{12.20}$$

for $s, t \in \mathbb{R}$.

In some cases, the three vaguelette conditions can be verified directly. Exercise 12.2 gives a criterion in the Fourier domain that can be useful in some other settings.

The following is a key property of a vaguelette system, proved in Appendix B.4. We use the abbreviation $\|\alpha\|_2$ for $\|(\alpha_\lambda)\|_{\ell_2}$

Proposition 12.3 *(i) If $\{w_\lambda\}$ is a system of vaguelettes satisfying* (12.18)– (12.20)*, then there exists a constant C, depending on (C_1, C_2, η, η') such that*

$$\Big\| \sum_\lambda \alpha_\lambda w_\lambda \Big\|_2 \le C \|\alpha\|_2 \tag{12.21}$$

(ii) If $\{u_\lambda\}, \{v_\lambda\}$ are biorthogonal systems of vaguelettes, then there exist positive constants c, C such that

$$c\|\alpha\|_2 \le \Big\| \sum_\lambda \alpha_\lambda u_\lambda \Big\|_2, \Big\| \sum_\lambda \alpha_\lambda v_\lambda \Big\|_2 \le C \|\alpha\|_2. \tag{12.22}$$

The second part is a relatively straightforward consequence of the first key conclusion; it shows that having two vaguelette systems that are orthogonal allows extension of bound (12.21) to a bound in the opposite direction, which we have seen is needed in order to control $\lambda_{\min}(\rho(\Sigma))$.

Thus, if we have two biorthogonal systems of vaguelettes, then each forms a *frame*: up to multiplicative constants, we can compute norms of linear combinations using the coefficients alone.

Example continued. Suppose again that $Af(u) = \int_{-\infty}^u f(t)dt$ and that ψ is a C^2 orthonormal wavelet with compact support and two vanishing moments, so that $\int \psi = \int t\psi = 0$. We saw that $\{u_\lambda = (\psi^{(-1)})_\lambda\}$ and $\{v_\lambda = -(\psi')_\lambda\}$ satisfy (1) with $\kappa_j = 2^{-j}$, and (2).

In order to obtain the frame bounds for property (3), we verify conditions (12.18)–(12.20) for ψ^{-1} and ψ', and then appeal to Lemma 12.3. Indeed, ψ' and $\psi^{(-1)}$ have compact support, the latter because ψ does and $\int \psi = 0$. So (12.18) holds. Turning to (12.19), we have $\int \psi' = 0$ by compact support of ψ, and integration by parts shows (using compact support of ψ^{-1}) that $\int \psi^{(-1)} = -\int u\psi(u)du = 0$. Finally ψ' is C^1 and $\psi^{(-1)}$ is C^3 so the Hölder property (12.20) follows again from the compact support.

Definition 12.4 (Donoho (1995)) Let $\{\psi_\lambda\}$ be an orthonormal wavelet basis for $L_2(T)$ and $\{u_\lambda\}, \{v_\lambda\}$ be systems of vaguelettes for $L_2(U)$. Let A be a linear operator with domain $\mathcal{D}(A)$ dense in $L_2(T)$ and taking values in $L_2(U)$. The systems $\{\psi_\lambda\}, \{u_\lambda\}, \{v_\lambda\}$ form a *wavelet vaguelette decomposition* of A if they enjoy the following properties:
(1) quasi-singular values: (12.14)
(2) biorthogonality: (12.15)
(3) near-orthogonality: (12.22).

Note that the quasi-singular values κ_j depend on j, but not on k.

12.3 Examples of WVD

1. *r-fold integration.* If $(Af)(u) = \int_{-\infty}^{u} f(t)dt$ and r is a positive integer, we may define the r-fold iterated integral by $A_r f = A(A_{r-1} f)$. We also write $f^{(-r)}$ for $A_r f$. The WVD follows by extending the arguments used for $r = 1$. Suppose that ψ is a C^r orthonormal wavelet with compact support and $r + 1$ vanishing moments, then the WVD is given by

$$\kappa_j = 2^{rj}, \qquad u_\lambda = (\psi^{(-r)})_\lambda, \qquad v_\lambda = (\psi^{(r)})_\lambda.$$

In particular, for later use we note that $\{\psi_\lambda^{(r)}\}$ forms a system of vaguelettes and satisfies the frame bounds (12.22).

2. *Fractional Integration.* Suppose that A is the fractional integration operator

$$(Af)(u) = \frac{1}{\Gamma(\beta)} \int_{-\infty}^{u} \frac{f(t)}{(u-t)^{1-\beta}} dt = (\Psi_\beta \star f)(u) \tag{12.23}$$

for $0 < \beta < 1$ and $\Psi_\beta(u) = u_+^\beta / \Gamma(\beta)$. Define the order β fractional derivative and integral of ψ by $\psi^{(\beta)}$ and $\psi^{(-\beta)}$ respectively. The WVD of A is then obtained by setting

$$\kappa_j = 2^{-j\beta}, \qquad u_\lambda = (\psi^{(-\beta)})_\lambda, \qquad v_\lambda = (\psi^{(\beta)})_\lambda. \tag{12.24}$$

To justify these definitions, note that the Fourier transform of Ψ_β is given by (e.g. Gel'fand and Shilov (1964, p. 171)

$$\widehat{\Psi_\beta}(\xi) = \widehat{\Omega}(\xi)|\xi|^{-\beta},$$

where $\widehat{\Omega}(\xi)$ equals $c_\beta = ie^{i(\beta-1)\pi/2}$ for $\xi > 0$ and $c_\beta e^{-i\beta\pi}$ for $\xi < 0$. We use the Parseval formula (C.8) to express the representer equation (12.12) in the form $\int \widehat{f}\overline{\widehat{\psi_\lambda}} = \int \widehat{f}\overline{\widehat{\Psi_\beta}}\,\overline{\widehat{g_\lambda}}$ from which one formally obtains

$$\widehat{g_\lambda}(\xi) = \widehat{\psi_\lambda}/\widehat{\Psi_\beta}(\xi) = |\xi|^\beta \widehat{\psi_\lambda}(\xi)/\widehat{\Omega}(\xi).$$

It is easy to check that $\|\widehat{g_\lambda}\|^2 = \|\widehat{g_0}\|^2 2^{2j\beta}$ so that we may take $\kappa_j = 2^{-j\beta}$ and $u_\lambda = \kappa_j g_\lambda$, and, as in (12.13) set $v_\lambda = \kappa_j^{-1} A\psi_\lambda$.

Thus $\{u_\lambda\}$ and $\{v_\lambda\}$ are biorthogonal, and one checks that both systems are obtained by translation and dilation of in (12.24), with

$$\widehat{\psi^{(\beta)}} = |\xi|^\beta \widehat{\psi}(\xi)/\widehat{\Omega(\xi)}, \qquad \widehat{\psi^{(-\beta)}} = |\xi|^{-\beta}\widehat{\Omega(\xi)}\widehat{\psi}(\xi). \tag{12.25}$$

The biorthogonality relations for $\{u_\lambda\}$ and $\{v_\lambda\}$ will then follow if we verify that $\psi^{(\beta)}$ and $\psi^{(-\beta)}$ satisfy (12.18)–(12.20). The steps needed for this are set out in Exercise 12.2.

3. *Convolution.* The operator

$$(Af)(u) = \int_{-\infty}^{\infty} a(u-t)f(t)dt = (a \star f)(u)$$

is bounded on $L_2(\mathbb{R})$ if $\int |a| < \infty$, by (C.24), so we can take $\mathcal{D}(A) = L_2(\mathbb{R})$. The adjoint A^* is just convolution with $\tilde{a}(u) = a(-u)$, and so in the Fourier domain, the representer g_λ is given by

$$\hat{g}_\lambda = \hat{\psi}_\lambda / \hat{\tilde{a}}, \tag{12.26}$$

where $\hat{\tilde{a}}(\xi) = \hat{a}(-\xi)$.

As simple examples, we consider

$$a_1(x) = e^x I\{x < 0\}, \qquad a_2(x) = \tfrac{1}{2}e^{-|x|}. \tag{12.27}$$

It is easily checked that

$$\hat{a}_1(\xi) = (1 - i\xi)^{-1}, \qquad \hat{a}_2(\xi) = (1 + \xi^2)^{-1},$$

and hence that

$$g_\lambda = \psi_\lambda - (\psi_\lambda)', \qquad g_\lambda = \psi_\lambda - (\psi_\lambda)''. \tag{12.28}$$

Either from representation (12.26), or more directly from (12.28), one finds that with $\beta = 1$ and 2 in the two cases, that

$$\|g_\lambda\|_2 \sim \begin{cases} 2^{2j\beta} & \text{as } j \to \infty, \\ 1 & \text{as } j \to -\infty. \end{cases}$$

This is no longer homogeneous in j in the manner of fractional integration, but we can still set $\kappa_j = \min(1, 2^{-j\beta})$.

The biorthogonal systems $\{u_\lambda\}$ and $\{v_\lambda\}$ are given by (12.13). In the case of $u_\lambda = \kappa_j g_\lambda$, the rescaling $S_\lambda u_\lambda$ can be found directly from (12.28), yielding $2^{-j}\psi + \psi'$ in the case $j > 0$. The vaguelette properties (12.18)– (12.20) then follow from those of the wavelet ψ. For $v_\lambda = \kappa_j^{-1} A\psi_\lambda$, it is more convenient to work in the Fourier domain, see Exercise ??

4. *Radon transform.* For the Radon transform in $\mathbb{R}^2$—compare Section 3.8 for a version on the unit disk—Donoho (1995) develops a WVD with quasi-singular values $\kappa_j = 2^{j/2}$. The corresponding systems $\{u_\lambda\}$, $\{v_\lambda\}$ are localized to certain curves in the (s, ϕ) plane rather than to points, so they are not vaguelettes, but nevertheless they can be shown to have the near-orthogonality property.

Here is a formulation of the indirect estimation problem when a WVD of the operator is available, building on the examples presented above. Suppose that we observe A in the stochastic observation model (12.6)–(12.8), and that $\{\psi_\lambda, u_\lambda, v_\lambda\}$ form a wavelet-vaguelette decomposition of A. Consider the observations

$$Y(u_\lambda) = [Af, u_\lambda] + \epsilon Z(u_\lambda).$$

Writing $Y_\lambda = Y(u_\lambda), z_\lambda = Z(u_\lambda)$ and noting that $[Af, u_\lambda] = \kappa_j \langle f, \psi_\lambda \rangle = \kappa_j \theta_\lambda$, say, we arrive at

$$Y_\lambda = \kappa_j \theta_\lambda + \epsilon z_\lambda. \tag{12.29}$$

Let Σ be the covariance matrix of $z = (z_\lambda)$. Since

$$\beta \Sigma^T \beta = \mathrm{Var}\left(\sum \beta_\lambda z_\lambda\right) = \mathrm{Var} Z\left(\sum \beta_\lambda u_\lambda\right) = \| \sum \beta_\lambda u_\lambda \|_2^2,$$

the near orthogonality property guarantees that

$$cI \le \Sigma \le CI, \tag{12.30}$$

where the inequalities are in the sense of non-negative definite matrices. We say that the noise z is *nearly independent*.

We are now ready to consider estimation of f from observations on Y. The reproducing formula (12.16) suggests that we consider estimators of f of the form

$$\hat{f} = \sum_\lambda \eta_\lambda(\kappa_j^{-1} Y_\lambda)\psi_\lambda$$

for appropriate univariate estimators $\eta_\lambda(y)$. The near-independence property makes it plausible that restricting to estimators in this class will not lead to great losses in estimation efficiency; this is borne out by results to follow. Introduce $y_\lambda = \kappa_j^{-1} Y_\lambda \sim N(\theta_\lambda, \kappa_j^{-2}\epsilon^2)$. We have $\hat{f} - f = \sum_\lambda [\eta_\lambda(y_\lambda) - \theta_\lambda]\psi_\lambda$ and so, for the mean squared error,

$$E\| \hat{f} - f \|_2^2 = \sum_\lambda E[\eta_\lambda(y_\lambda) - \theta_\lambda]^2 = \sum_\lambda r(\eta_\lambda, \theta_\lambda; \kappa_j^{-1}\epsilon).$$

Notice that if $\kappa_j \sim 2^{-\beta j}$, then the noise level $\kappa_j^{-1}\epsilon \sim 2^{\beta j}\epsilon$ grows rapidly with level j. This is the noise amplification characteristic of linear inverse problems and seen also in Chapter 3.8. In the next section, we study in detail the consequences of using threshold estimators to deal with this amplification.

12.4 The correlated levels model

We assume that

$$\begin{aligned} &y_{jk} = \theta_{jk} + \epsilon_j z_{jk}, \qquad \epsilon_j = 2^{\beta j}\epsilon, \quad \beta \ge 0 \\ &\xi_0 I \le \Sigma \le \xi_1 I. \end{aligned} \tag{12.31}$$

Here the indices $j \ge 0, k = 1, \ldots, 2^j$ and the noise is assumed to be zero mean Gaussian, with the inequalities on the covariance matrix understood in the sense of non-negative definite matrices.

This is an extension of the Gaussian white noise model (12.1) in two significant ways: (i) level dependent noise $\epsilon_j = 2^{\beta j}\epsilon$, capturing the noise amplification inherent to inverting an operator A of smoothing type, and (ii) the presence of correlation among the noise components, although make the key assumption of near-independence.

Motivation for this model comes from the various examples of linear inverse problems in the previous section: when a wavelet-vaguelette decomposition exists, we have both properties (i) and (ii). The model is then recovered from (12.29)–(12.30) when $\lambda = (jk)$ has the standard index set, $\kappa_j = 2^{-\beta j}$ and $y_{jk} = \kappa_j^{-1} Y_{jk}$.

Let us first examine what happens in model (12.31) when on level j we use soft thresholding at a *fixed* value $\lambda_j \epsilon_j$. Thus $\hat{\theta}^S_{jk} = \eta_S(y_{jk}, \lambda_j \epsilon_j)$. Decomposing the mean squared error by levels, we have

$$r(\hat{\theta}^S, \theta) = \sum E\|\hat{\theta}^S_j - \theta_j\|^2,$$

and if $\lambda_j = \sqrt{2\log \delta_j^{-1}}$, we have from the soft thresholding risk bound (8.13) that

$$E\|\hat{\theta}^S_j - \theta_j\|^2 \le 2^j \delta_j \epsilon_j^2 + (\lambda_j^2 + 1)\sum_k \theta_{jk}^2 \wedge \epsilon_j^2.$$

The noise term $2^j \delta_j \epsilon_j^2 = \delta_j 2^{(1+2\beta)j}\epsilon^2$, showing the effect of the geometric inflation of the variances, $\epsilon_j^2 = 2^{2\beta j}\epsilon^2$. To control this term, we might take $\delta_j = 2^{-(1+2\beta)j} = n_j^{-(1+2\beta)}$. This corresponds to threshold

$$\lambda_j = \sqrt{2(1+2\beta)\log n_j},$$

which is higher than the 'universal' threshold $\lambda_j^U = \sqrt{2\log n_j}$ when $\beta > 0$. With this choice we arrive at

$$E\|\hat{\theta}^S_j - \theta_j\|^2 \le \epsilon^2 + c_\beta\, j \sum_k \theta_{jk}^2 \wedge 2^{2\beta j}\epsilon^2.$$

At this point we can do a heuristic calculation to indicate the benefits of using the sparse representation provided by the WVD. This will also set the stage for more precise results to follow.

Now suppose that the unknown function f is piecewise continuous with at most d discontinuities. Then the wavelet tranform of f is sparse, and in particular, if the suppport of ψ is compact, there are at most a bounded number of non-zero coefficients θ_{jk} at each level j, and those coefficients are bounded by $c2^{-j/2}$ by Lemma 7.1. Hence

$$\sum_k \theta_{jk}^2 \wedge 2^{2\beta j}\epsilon^2 \le c_{d\psi f}(2^{-j} \wedge 2^{2\beta j}\epsilon^2).$$

To find the worst level, we solve for $j = j_*$ in the equation $2^{-j} = 2^{2\beta j}\epsilon^2$, so that $2^{(1+2\beta)j_*} = \epsilon^{-2}$. On the worst level, this is bounded by $2^{-j_*} = (\epsilon^2)^{1/(1+2\beta)}$. The maxima on the other levels decay geometrically in $|j - j_*|$ away from the worst level, and so the sum converges and as a bound for the rate of convergence we obtain

$$j_* 2^{-j_*} \asymp (\log \epsilon^{-2})(\epsilon^2)^{1/(1+2\beta)}.$$

Comparison with SVD. For piecewise constant f, we can suppose that the coefficients in the singular function basis, $\theta_k = \langle f, e_k \rangle$ decay as $O(1/k)$. Suppose that the singular values $b_k \asymp k^{-\beta}$. Then from (12.9),

$$\sum_k \theta_k^2 \wedge \frac{\epsilon^2}{b_k^2} \asymp \sum k^{-2} \wedge k^{2\beta}\epsilon^2 \asymp k_*^{-1},$$

where k_* solves $k^{-2} = k^{2\beta}\epsilon^2$, so that $k_*^{-1} = (\epsilon^2)^{1/(2+2\beta)}$. Hence, the rate of convergence using linear estimators with the singular value decomposition is $O((\epsilon^2)^{1/(2+2\beta)})$, while we can achieve the faster rate $O(\log \epsilon^{-2}(\epsilon^2)^{1/(1+2\beta)})$ with thresholding and the WVD.

In fact, as the discussion of the direct estimation case (Section 12.1) showed, the $\log \epsilon^{-2}$ term can be removed by using data-dependent thresholding, and it will be the goal of the rest of this chapter to prove such a result.

We will see that the rate of convergence over $\Theta^\alpha_{p,q}(C)$ is $C^{2(1-r)}\epsilon^{2r}$, with $r = 2\alpha/(2\alpha + 2\beta + 1)$, up to constants depending only on (α, p, q) and β.

The proof has the same structure as in the direct case, Section 12.1. The lower bound, after bounding the effect of correlation, is found from the worst Besov shell. The upper bound uses a penalized least squares estimator, after a key modification to the oracle inequality, Section 11.7, to control the effect of noise inflation with level j. With these (not unimportant) changes, the argument is reduced to the analysis of the ℓ_p-ball control functions $r_{n_j,p}(C_j, \epsilon_j)$; this is deferred to the following section.

The penalized estimator is constructed levelwise, in a manner analogous to the direct case, Section 12.1, but allowing for the modified noise structure. Thus, at level j, we use a penalized least squares estimator $\hat\theta_P(y_j)$, (12.2), with $\mathrm{pen}_j(k) = k\lambda_{j,k}^2$. However, now

$$\lambda_{j,k} = \sqrt{\zeta\xi_1}(1 + \sqrt{2L_{n_j,k}}), \qquad L_{n_j,k} = (1+2\beta)\log(\gamma_{n_j} n_j/k), \tag{12.32}$$

where $n_j = 2^j$ and γ_n is specified below.

The penalized least squares estimator is equivalent to hard thresholding with level and data dependent threshold $\hat t_j = t_{n_j,\hat k_j}$ where $t_{n,k}^2 = k\lambda_k^2 - (k-1)\lambda_{k-1}^2 \approx \lambda_k^2$ and $\hat k_j = N(\hat\theta_P(y_j))$ is the number of non-zero entries in $\hat\theta_P(y_j)$.

The levelwise estimators are combined into an overall estimator $\hat\theta^P = (\hat\theta_j^P)$ with $\hat\theta_j^P(y) = \hat\theta_P(y_j)$ for $j \ge 0$. [Note that in this model there is no cutoff at a fixed level J.]

Theorem 12.5 *Assume the correlated blocks model* (12.31) *and that*

$$\alpha > (2\beta+1)(1/p - 1/2)_+. \tag{12.33}$$

Then for all such $\alpha, C > 0$ and $0 < p, q \le \infty$, for the penalized least squares estimator just described, there exist constants c_i such that if $\epsilon/C < c_0$,

$$\begin{aligned} c_0 C^{2(1-r)}\epsilon^{2r} &\le R_N(\Theta^\alpha_{p,q}(C), \epsilon) \\ &\le \sup_{\Theta^\alpha_{p,q}(C)} E\|\hat\theta_P - \theta\|^2 \le c_1\epsilon^2 + c_2 C^{2(1-r)}\epsilon^{2r}. \end{aligned} \tag{12.34}$$

with $r = 2\alpha/(2\alpha+2\beta+1)$. The constants $c_1 = c_1(\beta, \gamma, \zeta, \xi_1)$ and $c_2 = c_2(\alpha, \beta, \gamma, p, \zeta, \xi_1)$.

The key point is that the estimator $\hat{\theta}_P$ achieves the correct rate of convergence without having to specify any of (α, p, q, C) in advance, subject only to smoothness condition (12.33).

This is essentially a generalization of Theorem 12.1, to which it reduces if $\beta = 0$ and $\xi_0 = \xi_1 = 1$. We could modify $\hat{\theta}_P$ to cut off at a level $J = \log_2 \epsilon^{-2}$ as in that theorem; the result would be an additional tail bias term $cC^2(\epsilon^2)^{2\alpha'}$ in (12.34). In that case, we could also use $\gamma_n \equiv \gamma > e$ rather than $\gamma_n = \log^2 n$ in the definition (12.32), and the variance term $c_1\epsilon^2$ would change to $c_1\epsilon^2 \log \epsilon^{-2}$.

Proof We begin with the lower bound. It follows from the covariance comparison Lemma 4.24 that the minimax risk in correlated model (12.31) is bounded below by the risk in a corresponding independence model in which the z_{jk} are i.i.d. $N(0, \xi_0)$. We may then restrict attention to the Besov shell $\Theta^{(j)} \cong \Theta_{n_j,p}(C_j)$ and conclude that

$$R_N(\Theta^{\alpha}_{p,q}(C), \epsilon) \geq R_N(\Theta_{n_j,p}(C_j), \zeta_0\epsilon) \geq a_1 r_{n_j,p}(C_j, \xi_0\epsilon),$$

by the lower bound part of Theorem 11.7. It will be shown in the next section that this is bounded below by $c_1\xi_0^{2r}C^{2(1-r)}\epsilon^{2r}$.

Turning now to the upper bound, the levelwise structure of $\hat{\theta}_P$ implies that

$$E\|\hat{\theta}_P - \theta\|^2 = \sum_j E\|\hat{\theta}_{P,j} - \theta_j\|^2,$$

and we will apply at each level j the inverse problem variant, Theorem 11.9, of the oracle inequality for complexity penalized estimators. Indeed, from (11.15) with $n_j = 2^j$, and with the notation $\mathcal{R}_j(\theta_j, \epsilon_j) = \inf_J C_j(J, \theta_j)$, we can bound the overall risk

$$E\|\hat{\theta}_P - \theta\|^2 \leq 2K\xi_1 \sum_j M'_j\epsilon_j^2 + K\sum_j \mathcal{R}_j(\theta_j, \epsilon_j). \tag{12.35}$$

For the first term, note from (11.59) that $M'_j \leq C_{\beta\gamma}2^{-2\beta j}j^{-2}$—indeed, obtaining rapid decay with j was the primary reason for the variant in Theorem 11.9. Consequently,

$$\sum_j M'_j\epsilon_j^2 \leq C_{\beta\gamma}\sum_j 2^{-2\beta j}j^{-2}\cdot 2^{2\beta j}\epsilon^2 \leq c_{\beta\gamma}\epsilon^2 \tag{12.36}$$

so that the first term in the MSE bound for $\hat{\theta}_P$ is $O(\epsilon^2)$.

To bound $\sum_j \mathcal{R}_j(\theta_j, \epsilon_j)$ we using Besov shells as in the direct case, obtaining

$$\sup_{\Theta}\sum_j \mathcal{R}_j(\theta_j, \epsilon_j) \leq c_2\sum_j r_{n_j,p}(C_j, \epsilon_j)$$

which we show in Proposition 12.6 below is bounded by $c_3C^{2(1-r)}\epsilon^{2r}$. □

12.5 Taming the shell bounds

Proposition 12.6 *Suppose that for each $j \geq 0$, the shell and noise parameters are*

$$n_j = 2^j, \qquad C_j = C2^{-aj}, \qquad \epsilon_j = 2^{\beta j}\epsilon. \tag{12.37}$$

Here $0 < p \le \infty, a = \alpha + 1/2 - 1/p$ *and* $\beta \ge 0$. *Suppose that* $\alpha > (2\beta+1)(1/p - 1/2)_+$ *and put* $r = 2\alpha/(2\alpha + 2\beta + 1)$. *Then there exist constants* $c_i(\alpha, \beta, p)$ *such that if* $\epsilon/C \le c_0$, *then*

$$c_1 C^{2(1-r)}\epsilon^{2r} \le \max_{j\ge 0} r_{n_j,p}(C_j, \epsilon_j) \le \sum_{j\ge 0} r_{n_j,p}(C_j, \epsilon_j) \le c_2 C^{2(1-r)}\epsilon^{2r}. \tag{12.38}$$

Proof Let us make the abbreviations

$$R_j = r_{n_j,p}(C_j, \epsilon_j), \qquad R_* = C^{2(1-r)}\epsilon^{2r}. \tag{12.39}$$

The essence of the proof is to show that the shell bounds $j \to R_j$ peak at a critical level j_*, and decay geometrically away from the value R_* at this least favorable level, so that the series in (12.38) is summable. The behavior for $p < 2$ is indicated in Figure 12.1; the case $p \ge 2$ is similar and simpler.

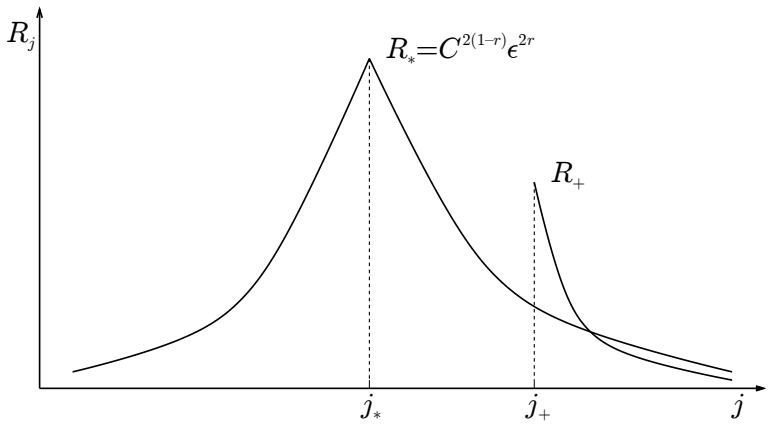

Figure 12.1 Schematic behavior of 'shell risks' R_j; with j treated as a real variable.

More specifically, in the case $p \ge 2$, we show that

$$R_j = \begin{cases} R_* 2^{(2\beta+1)(j-j_*)} & j \le j_* \\ R_* 2^{-2\alpha(j-j_*)} & j \ge j_* \end{cases} \tag{12.40}$$

with the critical level $j_* \in \mathbb{R}$ being defined by

$$2^{(\alpha+\beta+1/2)j_*} = C/\epsilon, \tag{12.41}$$

and the maximum shell bound being given by a multiple of R_* in (12.39).

In the case $p < 2$, by contrast, there are three zones to consider and we show that

$$R_j = \begin{cases} R_* 2^{(2\beta+1)(j-j_*)} & j \le j_* \\ R_* 2^{-p\rho(j-j_*)}[1 + \tau(j - j_*)]^{1-p/2} & j_* \le j < j_+ \\ R_+ 2^{-2a(j-j_+)} & j \ge j_+ \end{cases} \tag{12.42}$$

where R_* is as before and $\rho = \alpha - (2\beta+1)(1/p - 1/2)] > 0$ in view of smoothness assumption (12.33). The values of τ and R_+ are given below; we show that R_+ is of geometrically smaller order than R_*.

In either case, the geometric decay bounds are summable, leading to a bound that is just a multiple of R_*: $\sum R_j \le c_{\alpha\beta p} R_*$.

To complete the proof, we establish the geometric shell bounds in (12.40) and (12.42), starting with the simpler case $p \ge 2$. For convenience, we recall the scale–ϵ version of the control function

$$r_{n,p}(C,\epsilon) = \begin{cases} n^{1-2/p}C^2 & C \ge \epsilon n^{1/p} \\ n\epsilon^2 & C \le \epsilon n^{1/p}. \end{cases}$$

We apply this level by level, using the ℓ_p–ball interpretation of the Besov shell $\Theta^{(j)}$, so that

$$n_j = 2^j, \qquad C_j = C2^{-aj}, \qquad \epsilon_j = 2^{\beta j}\epsilon, \tag{12.43}$$

with $a = \alpha + 1/2 - 1/p$. Thus, on shell j, the boundary between samll C_j and large C_j zones in the control function is given by the equation $(C_j/\epsilon_j)n_j^{-1/p} = 1$. Inserting the definitions in the previous display, we recover formula (12.41) for the critical level j_*.

In the large signal zone, $j \le j_*$, the shell risks grow geometrically: $R_j = n_j\epsilon_j^2 = 2^{(2\beta+1)j}\epsilon^2$. The maximum is attained at $j = j_*$, and on substituting the definition of the critical level j_*, we obtain (12.39).

In the small signal zone, $j \ge j_*$, the shell bounds $R_j = C^2 2^{-2\alpha j}$ and it follows from (12.41) that $C^2 2^{-2\alpha j_*} = R_*$. We have established (12.40).

Turning to the case $p < 2$, we again recall the form of the control function, now given in scale-ϵ form by

$$r_{n,p}(C,\epsilon) = \begin{cases} C^2 & C \le \epsilon\sqrt{1+\log n} \\ C^p\epsilon^{2-p}[1+\log(n\epsilon^p/C^p)]^{1-p/2} & \epsilon\sqrt{1+\log n} \le C \le \epsilon n^{1/p} \\ n\epsilon^2 & C \le \epsilon n^{1/p}. \end{cases}$$

We may refer to these cases, from top to bottom, as the 'small signal', 'sparse' and 'dense' zones respectively, corresponding to the structure of the least favorable configurations in the lower bound proof of Theorem 11.7.

First, observe that the sparse/dense boundary, the definition of j_* and the behavior for $j \le j_*$ correspond to the small/large signal discussion for $p \ge 2$. However, in the sparse zone, $j \ge j_*$, the shell risks $R_j = C_j^p\epsilon_j^{2-p}[1+\log(n_j\epsilon_j^p C_j^{-p})]^{1-p/2}$. Using (12.37), the leading term

$$C_j^p\epsilon_j^{2-p} = C^p\epsilon^{2-p}2^{-p(a-2\beta/p+\beta)j}$$

decays geometrically for $j \ge j_*$, due to the smoothness assumption (12.33); indeed we have $a - 2\beta(1/p - 1/2) = \alpha - (2\beta+1)(1/p-1/2) > 0$. The logarithmic term can be rewritten using the boundary equation (12.41):

$$\log(n_j\epsilon_j^p C_j^{-p}) = p(\alpha+\beta+1/2)(j-j_*)\log 2.$$

Set $\tau = p(\alpha + \beta + 1/2)\log 2$ in (12.42), we have shown that

$$R_j = C^p \epsilon^{2-p} 2^{-p\rho j}[1 + \tau(j - j_*)]^{1-p/2}.$$

Putting $j = j_*$ gives $R_{j_*} = C^p \epsilon^{2-p} 2^{-p\rho j_*}$ and yields the middle formula in (12.42).

The boundary of the sparse and highly sparse zones is described by the equation

$$1 = (C_j/\epsilon_j)(1 + \log n_j)^{-1/2} = (C/\epsilon)2^{-(\alpha+\beta+1/2)j + j/p}(1 + j\log 2)^{-1/2}.$$

Using (12.41) as before, and taking base two logarithms, we find that the solution j_+ satisfies

$$(a + \beta)j_+ = (\alpha + \beta + \tfrac{1}{2})j_* - \ell(j_+), \tag{12.44}$$

where $\ell(j) = \frac{1}{2}[\log_2(1 + j\log 2)]$.

In the highly sparse zone, the shell risks $R_j = C_j^2 = C^2 2^{-2aj}$ decline geometrically from the maximum value $R_+ = C^2 2^{-2aj_+}$. It must be shown that this maximum is of smaller order than R_*. Indeed, using (12.39), $R_+/R_* = (C/\epsilon)^{2r} 2^{-2aj_+}$, and hence

$$\log_2(R_+/R_*) = 2(\alpha j_* - a j_+) = -[\rho j_+ - 2\alpha\ell(j_+)]/(\alpha + \beta + 1/2).$$

after some algebra using (12.44) and recalling the definition of ρ from below (12.42). For $j \geq 2$ we have $2\ell(j) \leq \log_2 j$, and setting $\rho_2 = \rho/(\alpha + \beta + 1/2)$ we arrive at

$$\log(R_+/R_*) \leq -\rho_2 j_+ + \log_2 j_+.$$

Since $j_* \leq j_+ \leq c_{\alpha\beta p} j_*$ from (12.44), we have

$$R_+/R_* \leq c_{\alpha\beta p} j_* 2^{-\rho_2 j_*} = c'(C/\epsilon)^{-\rho_3}\log(C/\epsilon).$$

Thus, if $\epsilon/C \leq c_0$, it is evident that $R_+ \leq R_*$.

For the lower bound, it is enough now to observe that

$$\max R_j = \max(R_{\lfloor j_* \rfloor}, R_{\lceil j_* \rceil}).$$

□

Note. We cannot argue that $R_+ \leq R_*$ directly from $j_+ > j_*$ because the control function $C \to r_{n,p}(C, \epsilon)$ is discontinuous at the highly sparse to sparse boundary, compare Figure 11.1.

Exercises

12.1 *(Simple Fourier facts)*
Recall or verify the following.
(a) Suppose that ψ is C^L with compact support. Then $\hat{\psi}(\xi)$ is infinitely differentiable and

$$|\hat{\psi}^{(r)}(\xi)| \leq C_r|\xi|^{-L} \qquad \text{for all } r.$$

(b) Suppose that ψ has K vanishing moments and compact support. Then for $r = 0, \ldots, K-1$, we have $\hat{\psi}^{(r)}(\xi) = O(|\xi|^{K-r})$ as $\xi \to 0$.
(c) For f such that the integral converges, $|f(t)| \leq (2\pi)^{-1}\int |\hat{f}(\xi)|d\xi$ and

$$|f(t) - f(s)| \leq (2\pi)^{-1}|t - s| \int |\xi||\hat{f}(\xi)|d\xi.$$

(d) If $\hat{f}(\xi)$ is C^2 for $0 < |\xi| < \infty$ and if $\hat{f}(\xi)$ and $\hat{f}'(\xi)$ vanish as $\xi \to 0, \infty$, then

$$|f(t)| \le (2\pi)^{-1} t^{-2} \int |\hat{f}''(\xi)| d\xi,$$

12.2 *(Vaguelette properties for convolution examples)* (a) Let S_λ be the rescaling operator (12.17), and suppose that the system of functions w_λ can be represented in the Fourier domain via $\widehat{s_\lambda} = \widehat{S_\lambda w_\lambda}$. Show that vaguelette conditions (12.18)–(12.20) are in turn implied by the existence of constants M_i, not depending on λ, such that

$$\text{(i)} \int |\hat{s}_\lambda(\xi)| d\xi \le M_0, \qquad \int |\hat{s}_\lambda''(\xi)| d\xi \le M_1,$$

$$\text{(ii)} \ \hat{s}_\lambda(0) = 0, \qquad \text{and} \qquad \text{(iii)} \int |\xi| \, |\hat{s}_\lambda(\xi)| d\xi \le M_2,$$

with $\widehat{s_\lambda}(\xi)$ and $\widehat{s_\lambda}'(\xi)$ vanishing at $0, \pm\infty$.
(b) Show that if $Af = a \star f$, then for the two systems

$$v_\lambda = \kappa_j^{-1} A \psi_\lambda, \qquad \widehat{s_\lambda}(\xi) = \kappa_j^{-1} \hat{a}(2^j \xi) \hat{\psi}(\xi)$$

$$u_\lambda = \kappa_j g_\lambda, \qquad \widehat{s_\lambda}(\xi) = [\kappa_j / \hat{a}(-2^j \xi)] \hat{\psi}(\xi).$$

(c) Suppose A is given by fractional integration, (12.23), for $0 < \beta < 1$. Suppose that ψ is C^3, of compact support and has $L = 2$ vanishing moments. Show that $\{u_\lambda\}$ and $\{v_\lambda\}$ are vaguelette systems.
(d) Suppose that A is given by convolution with either of the kernels in (12.27). Let $\beta = 1$ for a_1 and $\beta = 2$ for a_2. Suppose that ψ is $C^{2+\beta}$, of compact support and has $L = 2 + \beta$ vanishing moments. Show that $\{u_\lambda\}$ and $\{v_\lambda\}$ are vaguelette systems.

13

Sharp minimax estimation on ℓ_p balls

Suppose that we observe n-dimensional data

$$y_i = \theta_i + \epsilon z_i \qquad i = 1, \ldots, n \tag{13.1}$$

where θ is constrained to lie in a ball of radius C defined by the ℓ_p norm:

$$\Theta = \Theta_{n,p}(C) = \{\theta \in \mathbb{R}^n : \sum_{i=1}^{n} |\theta_i|^p \leq C^p\}. \tag{13.2}$$

We seek to estimate θ using squared error loss $\|\hat{\theta} - \theta\|^2 = \sum_i (\hat{\theta}_i - \theta_i)^2$, and in particular to evaluate the nonlinear minimax risk

$$R_N(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} E\|\hat{\theta} - \theta\|_2^2, \tag{13.3}$$

and make comparisons with the corresponding linear minimax risk $R_L(\Theta)$.

Although this model is finite dimensional, it is non-parametric in character since the dimension of the unknown parameter equals that of the data, and further we consider asymptotics as $n \to \infty$. The ℓ_p- constrained parameter space Θ is permutation symmetric and certainly solid, orthosymmetric and compact. It is thus relatively simple to study and yet yields a very sharp distinction between linear and non-linear estimators when $p < 2$. The setting also illustrates the Bayes minimax method discussed in Chapter 4.

Our object is to study the asymptotic behavior of the minimax risk R_N as n, the number of parameters, increases. We regard the noise level $\epsilon = \epsilon_n$ and ball radius $C = C_n$ as known functions of n. This framework accomodates a common feature of statistical practice: as the amount of data increases (here thought of as a decreasing noise level ϵ per parameter), so too does the number of parameters that one may contemplate estimating.

In previous chapters we have been content to describe the rates of convergence of $R_N(\Theta)$, or non-asymptotic bounds that differ by constant factors. In this chapter, and in the next for a multiresolution setting, we seek an exact, if often implicit, description of the asymptotics of $R_N(\Theta)$. Asymptotically, we will see that R_N depends on the size of $\Theta_{n,p}(C)$ through $n\epsilon^2$ times the dimension normalized radius $\eta_n = n^{-1/p}(C/\epsilon)$.

Although the asymptotic behavior depends on p and C (as well as on ϵ_n and n), the structure revealed in the least favorable distributions repays the effort expended. So long as the normalized radius η_n is not too small, the exact asymptotic behavior of $R_N(\Theta)$ is described by a Bayes-minimax problem in which the components θ_i of θ are drawn *independently* from an appropriate *univariate* near least favorable distribution $\pi_{1,n}$. When $p < 2$, this least

favorable distribution places most of its probability at 0, so that most components θ_i are zero and the corresponding vector θ is sparse.

The strategy for this chapter, then, is first to study a univariate problem $y = \theta + \epsilon z$, with $z \sim N(0,1)$ and θ having a prior distribution π, subject to a moment constraint $\int |\theta|^p d\pi \leq \tau^p$. In this univariate setting, we can compare linear, threshold and non-linear estimators and observe the distinction between $p \geq 2$, with "dense" least favorable distributions, and $p < 2$, with "sparse" least favorable distributions placing most of their mass at zero.

The second phase in this strategy is to "lift" the univariate results to the n-dimensional setting specified by (13.1)–(13.3). Here the independence of the co-ordinates of y_i in (13.1) and of the θ_i in the least favorable distribution is crucial. The details are accomplished using the Minimax Bayes approach sketched already in Chapter 4. In the sparse case ($p < 2$) with $\eta_n \to 0$, the least favorable distribution corresponds to vectors θ_i in which most co-ordinates are zero and a small fraction α_n at random positions have magnitude about $\epsilon_n \sqrt{2 \log \eta_n^{-p}}$.

The Minimax Bayes strategy is not, however, fully successful in extremely sparse cases when the expected number of spikes $n\alpha_n$, remains bounded as n grows. A different technique, using exchangeable rather than independent priors, is required for the lower bounds. The "random energy model" of statistical physics makes an appearance; its "phase transition" properties are needed to complete the argument.

We also study linear and threshold estimators as two simpler classes that might or might not come close in performance to the full class of non-linear estimators. In each case we also aim for exact asymptotics of the linear or threshold minimax risk.

13.1 Linear Estimators.

With linear estimators, exact calculations are relatively straightforward and serve as a point of reference for work with non-linear estimators in later sections.

The ℓ_p balls $\Theta_{n,p}(C)$ are solid and orthosymmetric and compact for all $0 < p \leq \infty$. However they are quadratically convex only if $p \geq 2$, while for $p < 2$,

$$\mathrm{QHull}[\Theta_{n,p}(C)] = \Theta_{n,2}(C).$$

Theorem 9.3 says that the linear minimax risk is determined by the quadratic hull, and so we may suppose that $p \geq 2$. Our first result evaluates the linear minimax risk, and displays the "corner" at $p = 2$.

Proposition 13.1 *Let $\bar{p} = p \vee 2$ and $\bar{\eta} = n^{-1/\bar{p}}(C/\epsilon)$. The minimax linear risk for squared error loss is*

$$R_L(\Theta_{n,p}(C), \epsilon) = n\epsilon^2 \bar{\eta}^2/(1 + \bar{\eta}^2),$$

with minimax linear estimator $\hat{\theta}_L$ given coordinatewise by

$$\hat{\theta}_{L,i}(y) = [\bar{\eta}^2/(1 + \bar{\eta}^2)]y_i.$$

Remark. For large C, and hence large $\bar{\eta}$, the minimax linear risk approaches the unconstrained minimax risk for $\mathbb{R}^n$, namely $n\epsilon^2$.

Proof Theorem 4.22 says that the linear minimax risk is found by looking for the hardest

rectangular subproblem:

$$R_L(\Theta) = \sup\left\{\sum \epsilon^2\tau_i^2/(\epsilon^2+\tau_i^2) : \sum \tau_i^p \le C^p\right\}.$$

In terms of new variables $u_i = \tau_i^p/C^p$, and a scalar function $\ell(t) = t^2/(1+t^2)$, this optimization can be rephrased as that of maximizing

$$f(u) = \epsilon^2 \sum_i \ell(C\epsilon^{-1}u_i^{1/p})$$

over the simplex $\sum_1^n u_i \le 1$ in the non-negative orthant of $\mathbb{R}^n$. Since f is symmetric and increasing in the co-ordinates u_i, and concave when $p \ge 2$, it follows that the maximum is attained at the centroid $u_i = n^{-1}(1,\dots,1)$. Introducing the normalized radius $\bar\eta = n^{-1/p\vee 2}(C/\epsilon)$, we may write the corresponding minimax risk as $n\epsilon^2\ell(\bar\eta)$. From (4.29), the corresponding linear minimax estimate is $\hat\theta_L = \ell(\bar\eta_n)y$. □

Example 13.2 The calibration $\epsilon = 1/\sqrt{n}$ arises frequently in studying sequence model versions of nonparametric problems (compare (1.24) in Chapter 3). Consider the ℓ_1 ball of radius $C = 1 : \Theta_{n,1} = \{\theta : \sum_1^n |\theta_i| \le 1\}$. We see that $\bar\eta = n^{-1/2}\cdot n^{1/2} = 1$ and that

$$R_L(\Theta_{n,1}) = 1/2, \qquad \hat\theta_L(y) = y/2.$$

Of course $\Theta_{n,1}$ has the same *linear* minimax risk as the solid sphere $\Theta_{n,2}$ which is much larger, for example in terms of volume. We have already seen, in Example 8.2, that nonlinear thresholding yields a much smaller maximum risk over $\Theta_{n,1}$ — the exact behavior of $R_N(\Theta_{n,1})$ is given at (13.27) below.

13.2 Univariate Bayes Minimax Problem

We consider a generalization of the bounded normal mean problem of Section 4.6 and the sparse normal mean setting of Section 8.7. Suppose that $y \sim N(\theta,\epsilon^2)$, and that θ is distributed according to a prior $\pi(d\theta)$ on $\mathbb{R}$. Assume that π belongs to a class satisfying the p-th moment constraint

$$\mathfrak{m}_p(\tau) = \{\pi(d\theta) : \int |\theta|^p\pi(d\theta) \le \tau^p\},$$

which is convex and weakly compact for all $p \le \infty$ and $\tau < \infty$. Such moment constraints are a population version of the "empirical" constraints on $(\theta_1,\dots\theta_n)$ defining an ℓ_p-ball—compare (13.2). We study the Bayes minimax risk

$$\beta_p(\tau,\epsilon) = \inf_{\hat\theta}\sup_{\pi\in\mathfrak{m}_p(\tau)} B(\hat\theta,\pi) = \sup\{B(\pi) : \pi \in \mathfrak{m}_p(\tau)\}. \tag{13.4}$$

where the second equality uses the minimax theorem (4.17) and (4.14) of Chapter 4.

Of course, $\mathfrak{m}_\infty(\tau)$ equals the set of priors supported on the bounded interval $[-\tau,\tau]$, and so $\beta_\infty(\tau,\epsilon) = \rho_N(\tau,\epsilon)$, compare (4.26). With an abuse of notation one can regard the sparse signal model of Section 8.7 as being the $p = 0$ limit of the p-th moment constraint. Since $\int |\theta|^p d\pi \to \pi\{\theta \ne 0\}$ as $p \to 0$, we can view $\mathfrak{m}_0(t) = \{\pi : \pi\{\theta \ne 0\} \le t\}$ as $\lim_{p\to 0}\mathfrak{m}_p(t^{1/p})$. In addition, the sparse Bayes minimax risk $\beta_0(\eta,\epsilon) = \lim_{p\to 0}\beta_p(\eta^{1/p},\epsilon)$.

Remark on Notation. We use the lower case letters β and ρ for Bayes and frequentist minimax risk in *univariate* problems, and the upper case letters B and R for the corresponding multivariate minimax risks.

We begin with some basic properties of $\beta_p(\tau, \epsilon)$, valid for all p and τ, and then turn to the interesting case of low noise, $\tau \to 0$, where the distinction between $p < 2$ and $p \geq 2$ emerges clearly.

Proposition 13.3 *The Bayes minimax risk $\beta_p(\tau, \epsilon)$, defined at* (13.4)*, is*
1. *decreasing in p,*
2. *increasing in ϵ,*
3. *strictly increasing, concave and continuous in $\tau^p > 0$,*
4. *and satisfies (i) $\beta_p(\tau, \epsilon) = \epsilon^2 \beta_p(\tau/\epsilon, 1)$, and*
   (ii) $\beta_p(a\tau, \epsilon) \leq a^2 \beta_p(\tau, \epsilon)$ for all $a \geq 1$.

Proof First, (1) and 4(i) are obvious, while (2) and 4(ii) are Exercises 4.1(a) and 4.6(a) respectively. Turning to (3), let $t = \tau^p$: the function $\tilde{\beta}(t) = \sup\{B(\pi) : \int |\theta|^p d\pi = t\}$ is concave in t because $\pi \to B(\pi)$ is concave and the constraint on π is linear. Monotonicity in τ^p is clear, and continuity follows from monotonicity and 4(ii). Strict monotonicity then follows from concavity. □

The scaling property 4(i) means that it suffices to study the unit noise situation. As in previous chapters, we use a special notation for this case: $x \sim N(\mu, 1)$, and write $\beta_p(\eta)$ for $\beta_p(\eta, 1)$ where $\eta = \tau/\epsilon$ denotes the signal to noise ratio.

Information about the least favorable distribution follows from an extension of our earlier results for $p = \infty$, Proposition 4.18, and $p = 0$, Proposition 8.9. (For the proof, see Exercise 13.3).

Proposition 13.4 *For p and τ in $(0, \infty)$, the Bayes minimax problem associated with $\mathfrak{m}_p(\tau)$ and $\beta_p(\tau)$ has a unique least favorable distribution π_τ. If $p = 2$, then π_τ is Gaussian, namely $N(0, \tau^2)$; while for $p \neq 2$ instead π_τ is proper, symmetric and has discrete support with $\pm\infty$ as the only possible accumulation points. When $p < 2$ the support must be countably infinite.*

Thus, the only case in which completely explicit solutions are available is $p = 2$, for which $\beta_2(\tau, \epsilon) = \tau^2\epsilon^2/(\tau^2 + \epsilon^2) = \rho_L(\tau, \epsilon)$ (Corollary 4.7). From now on, however, we will be especially interested in $p < 2$, and in general we will not have such explicit information about the value of $\beta_p(\tau, \epsilon)$, least favorable priors or corresponding estimators. We will therefore be interested in approximations, either by linear rules when $p \geq 2$, or more importantly, by threshold estimators for all $p > 0$.

$p \geq 2$ *versus* $p < 2$ *in low signal-to noise.*

When η is small and $p < 2$, appropriate choices of two point priors $\pi_{\alpha,\mu} = (1-\alpha)\delta_0 + \alpha\delta_\mu$ turn out to be approximately least favorable. We build on the discussion of sparse two point priors in Section 8.4. A one parameter family of priors $\pi_{\alpha,\mu(\alpha)}$ was defined there by requiring $\mu(\alpha)$ to satisfy the equation

$$\mu^2/2 + (2\log\alpha^{-1})^{1/4}\mu = \log(1-\alpha)/\alpha, \tag{13.5}$$

and the resulting sparse prior (defined for $\alpha < 1/2$) was said to have sparsity α and overshoot $a = (2\log\alpha^{-1})^{1/4}$.

Definition 13.5 The *sparse ℓ_p prior* $\pi_p[\eta]$ is the sparse prior $\pi_{\alpha,\mu(\alpha)}$ with $\alpha = \alpha_p(\eta)$ determined by the moment condition

$$\alpha\mu^p(\alpha) = \eta^p. \tag{13.6}$$

We write $\mu_p(\eta) = \mu(\alpha_p(\eta))$ for the non-zero support point.

Exercise 13.1 shows that this definition makes sense for η sufficiently small. Recalling from (8.31) that $\mu(\alpha) \sim \sqrt{2\log\alpha^{-1}}$ for α small, one can verify that as $\eta \to 0$,

$$\alpha_p(\eta) \sim \eta^p(2\log\eta^{-p})^{-p/2}$$
$$\mu_p(\eta) \sim (2\log\eta^{-p})^{1/2}.$$

We can now state the main result of this subsection.

Theorem 13.6 *As $\eta \to 0$,*

$$\beta_p(\eta) \sim \begin{cases} \eta^2 & 2 \le p \le \infty \\ \eta^p(2\log\eta^{-p})^{1-p/2} & 0 < p < 2. \end{cases} \tag{13.7}$$

If $p \ge 2$, then $\hat{\delta}_0 \equiv 0$ is asymptotically minimax and $\pi = (\delta_{-\eta} + \delta_\eta)/2$ is asymptotically least favorable.

If $p < 2$, then $\hat{\delta}_\lambda$, soft thresholding with threshold $\lambda = \sqrt{2\log\eta^{-p}}$, is asymptotically minimax. The sparse ℓ_p prior $\pi_p[\eta]$ of Definition 13.5 is asymptotically least favorable.

Remarks. 1. In the "nearly black" model of Section 8.7, corresponding to $p = 0$, we found that $\beta_0(\eta) \sim \eta \cdot (2\log\eta^{-1})$ with $\hat{\delta}_\lambda(x)$ being asymptotically minimax with $\lambda = \sqrt{2\log\eta^{-1}}$ and an asymptotically least favorable prior being $\pi_{\eta,\mu(\eta)}$. To see that this ℓ_p theorem is consistent with the $p = 0$ limit, observe that (13.7) implies $\beta_p(\eta^{1/p}) \sim \eta(2\log\eta^{-1})^{1-p/2}$ and recall that $\beta_0(\eta) = \lim_{p\to 0}\beta_p(\eta^{1/p})$.

2. Consider the special choice $\epsilon = n^{-1/2}$. Then $\eta_n^p = n^{-1}(C/\epsilon)^p = n^{-1+p/2}C^p$ and so $\lambda_n^2 = 2\log\eta_n^{-p} = (2-p)\log n - 2p\log C$. Hence larger signal strength, represented both in index p and in radius C, translates into a smaller choice of minimax threshold. Note that in a very small signal setting, $\eta_n^p = 1/n$, we recover the choice $\lambda_n = \sqrt{2\log n}$ discussed in earlier chapters.

3. The threshold estimator $\hat{\delta}_{\sqrt{2\log\eta^{-p}}}$ is also asymptotically minimax when $p \ge 2$.

Proof Consider first $p \ge 2$. For any prior $\pi \in \mathfrak{m}_p(\eta)$,

$$B(\hat{\delta}_0, \pi) = E_\pi\mu^2 \le (E_\pi|\mu|^p)^{2/p} \le \eta^2. \tag{13.8}$$

Consequently $B(\pi) \le \eta^2$, and so also $\beta_p(\eta) \le \eta^2$. In the other direction, consider the symmetric two point prior $\pi_\eta = (1/2)(\delta_\eta + \delta_{-\eta})$; formula (2.24) for the Bayes risk shows that $\beta_p(\eta) \ge B(\pi_\eta) \sim \eta^2$ as $\eta \to 0$.

Suppose now that $p < 2$. For the lower bound in (13.7), we use the priors $\pi_p[\eta]$ and the

asymptotics for their Bayes risks computed in Lemma 8.7. Note also that the p-th moment constraint $\alpha = \eta^p/\mu^p(\alpha)$ implies that

$$\mu^2(\alpha_\eta) \sim 2\log \alpha_\eta^{-1} = 2\log \eta^{-p} + p\log \mu^2(\alpha) \sim 2\log \eta^{-p}.$$

Putting these observations together, we obtain our desired lower bound as $\alpha \to 0$

$$\beta_p(\eta) \geq B(\pi_{\alpha(\eta)}) \sim \alpha\mu^2(\alpha) = \alpha\mu^p(\alpha)\cdot\mu^{2-p}(\alpha) \sim \eta^p(2\log\eta^{-p})^{1-p/2}.$$

For the upper bound, we use an inequality for the maximum integrated risk of soft thresholding:

$$\sup_{\pi\in\mathfrak{m}_p(\eta)} B(\hat{\delta}_\lambda,\pi) \leq r_S(\lambda,0) + \eta^p(1+\lambda^2)^{1-p/2}. \tag{13.9}$$

Assuming this for a moment, we note that $\beta_p(\eta)$ is bounded above by the left side, and in the right side we set $\lambda = \sqrt{2\log\eta^{-p}}$. Recalling from (8.7) that $r_S(\lambda,0) \sim 4\lambda^{-3}\phi(\lambda) = o(\eta^{-p})$ as $\eta \to 0$, we see that the second term is dominant and is asymptotically equivalent to $\eta^p(2\log\eta^{-p})^{1-p/2}$ as $\eta \to 0$.

It remains to prove (13.9). We use the risk bound for soft thresholding given at (8.12), and shown schematically in Figure 13.1. Now, define $\mu_\lambda = (1+\lambda^2)^{1/2}$, and then choose $c = c_\lambda$ so that

$$c\mu_\lambda^p = \mu_\lambda^2 = 1+\lambda^2,$$

that is, $c = (1+\lambda^2)^{1-p/2}$. Compare Figure 13.1. We conclude that

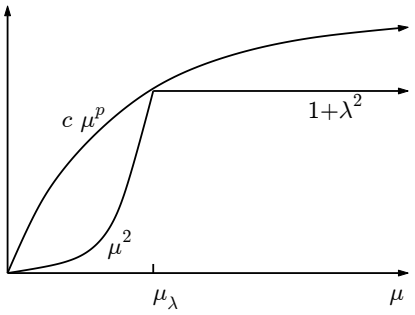

Figure 13.1 Schematic for risk bound: valid for $p < 2$

$$\begin{aligned} B(\hat{\delta}_\lambda,\pi) = \int r_S(\lambda,\mu)d\pi &\leq r_S(\lambda,0) + c\int \mu^p d\pi \\ &= r_S(\lambda,0) + \eta^p(1+\lambda^2)^{1-p/2}. \end{aligned}$$

As this holds for all $\pi \in \mathfrak{m}_p(\eta)$, we obtain (13.9). Here we used symmetry of $\mu \to r(\lambda,\mu)$ about 0 to focus on those π supported in $[0,\infty)$. □

Remark. There is an alternative approach to bounding $\sup_{\mathfrak{m}_p(\eta)} B_S(\lambda,\pi)$ which looks for the maximum of the linear function $\pi \to B_S(\lambda,\pi)$ among the extreme points of the convex $\mathfrak{m}_p(\eta)$ and shows that the maximum is actually of the two point form (8.26). This approach yields (see Exercise 13.6)

Proposition 13.7 *Let a threshold λ and moment space $\mathfrak{m}_p(\eta)$ be given. Then*

$$\begin{aligned}\sup\{B(\hat{\delta}_\lambda, \pi) : \pi \in \mathfrak{m}_p(\eta)\} &= \sup_{\mu \geq \eta} r(\lambda, 0) + (\eta/\mu)^p [r(\lambda, \mu) - r(\lambda, 0)] \\ &\leq r(\lambda, 0) + \eta^p \mu_\lambda^{2-p}\end{aligned} \tag{13.10}$$

where μ_λ is the unique solution of

$$r(\lambda, \mu_\lambda) - r(\lambda, 0) = (\mu_\lambda / p) r_\mu(\lambda, \mu_\lambda). \tag{13.11}$$

The least favorable prior for $\hat{\delta}_\lambda$ over $\mathfrak{m}_p(\eta)$ is of the two point prior form with α determined from η and $\mu = \mu_\lambda$ by (8.26)*. As $\lambda \to \infty$, we have*

$$\mu_\lambda \sim \lambda + \tilde{\Phi}^{-1}(p/2). \tag{13.12}$$

Hard thresholding. It is of some interest, and also explains some choices made in the analysis of Section 1.3, to consider when *hard* thresholding $\hat{\delta}_{H,\lambda}$ is asymptotically minimax.

Theorem 13.8 *If $p < 2$ and $\eta \to 0$, then the hard thresholding estimator $\hat{\delta}_{H,\lambda}$ is asymptotically minimax over $\mathfrak{m}_p(\eta)$ if*

$$\lambda^2 = \begin{cases} 2\log \eta^{-p} & \text{if } 0 < p < 1 \\ 2\log \eta^{-p} + \alpha \log(2\log \eta^{-p}) & \text{if } 1 \leq p < 2, \alpha > p - 1. \end{cases} \tag{13.13}$$

The introductory Section 1.3 considered an example with $p = 1$ and $\eta_n = n^{-1/2}$ so that $2\log \eta_n^{-1} = \log n$. In this case the threshold $\lambda = \sqrt{\log n}$ is not asymptotically minimax: the proof below reveals that the risk at 0 is too large. To achieve minimaxity for $p \geq 1$, a slightly larger threshold is needed, and in fact $\lambda_n = \sqrt{\log(n \log^\alpha n)}$ works for any $\alpha > 0$.

Proof We adopt a variant of the approach used for soft thresholding. It is left as Exercise 13.2 to use Lemma 8.5 to establish that if $c_\lambda = \lambda^{-p}(1 + \lambda^2)$ and $\lambda \geq \lambda_0(p)$, then

$$r_H(\lambda, \mu) \leq r_H(\lambda, 0) + c_\lambda \mu^p. \tag{13.14}$$

Consequently, integrating over any $\pi \in \mathfrak{m}_p(\eta)$, we obtain

$$B(\hat{\delta}_{H,\lambda}, \pi) \leq r_H(\lambda, 0) + c_\lambda \eta^p.$$

Since our choices $\lambda(\eta) \to \infty$ as $\eta \to 0$, we may use (8.15), namely $r_H(\lambda, 0) \sim 2\lambda\phi(\lambda)$, to conclude that

$$\sup_{\pi \in \mathfrak{m}_p(\eta)} B(\hat{\delta}_{H,\lambda}, \pi) \leq [2\lambda\phi(\lambda) + \lambda^{2-p}\eta^p](1 + o(1)).$$

Since $\lambda \sim 2\log \eta^{-p}$, we obtain minimaxity for hard thresholding so long as the term due to the risk at zero is negligible as $\eta \to 0$:

$$\lambda\phi(\lambda) = o(\lambda^{2-p}\eta^p).$$

It is easily checked that for $0 < p < 1$, this holds true for $\lambda^2 = 2\log \eta^{-p}$, whereas for $1 \leq p < 2$, we need the somewhat larger threshold choice in the second line of (13.13). □

For soft thresholding, the risk at zero $r_S(\lambda, 0) \sim 4\lambda^{-3}\phi(\lambda)$ is a factor λ^{-4} smaller than for hard thresholding with the same (large) λ; this explains why larger thresholds are only needed in the hard threshold case.

13.3 Univariate Thresholding

We have just seen that thresholding at an appropriate level is minimax in the low noise limit. In this section we look more systematically at the choice of threshold that minimizes mean squared error. We consider the optimal performance of the best threshold rule over the moment space $\mathfrak{m}_p(\tau)$ with the goal of comparing it to the minimax Bayes estimator, which although optimal, is not available explicitly. Define therefore

$$\beta_{S,p}(\tau, \epsilon) = \inf_{\lambda} \sup_{\pi \in \mathfrak{m}_p(\tau)} B(\hat{\delta}_\lambda, \pi), \tag{13.15}$$

where $\hat{\delta}_\lambda$ refers to a soft threshold estimator (8.4) with threshold λ. Throughout this section, we work with soft thresholding, sometimes emphasised by the subscript "S", though some analogous results are possible for hard thresholding (see Donoho and Johnstone (1994b).) A goal of this section is to establish an analogue of Theorem 4.16, which in the case of a bounded normal mean, bounds the worst case risk of linear estimators relative to all non-linear ones. Over the more general moment spaces $\mathfrak{m}_p(\tau)$, the preceding sections show that we have to replace linear by threshold estimators. To emphasize that the choice of estimator in (13.15) is restricted to thresholds, we write $B(\lambda, \pi)$ for $B(\hat{\delta}_\lambda, \pi)$.

Let $B_S(\pi) = \inf_\lambda B(\lambda, \mu)$ denote the best MSE attainable by choice of soft threshold. Our first task is to establish that a unique best $\lambda(\pi)$ exists, Proposition 13.10 below. Then follows a (special) minimax theorem for $B(\lambda, \pi)$. This is used to derive some properties of $\beta_{S,p}(\tau, \epsilon)$ which finally leads to the comparison result, Theorem 13.14.

To begin, we need some preliminary results about how the MSE varies with the threshold.

Dependence on threshold. Let $r_\lambda(\lambda, \mu) = (\partial/\partial\lambda) r(\lambda, \mu)$; from (8.59) and changes of variable one obtains

$$r_\lambda(\lambda, \mu) = 2 \int_{I(\mu)} w\phi(w - \lambda) dw,$$

where $I(\mu) = (-\infty, -\mu) \cup (-\infty, \mu)$. In particular, for all $\lambda \geq 0$ and μ

$$r_\lambda(0, \mu) = 4 \int_{-\infty}^{-|\mu|} w\phi(w) dw < 0, \quad \text{and} \tag{13.16}$$

$$r_\lambda(\lambda, 0) = 4 \int_{-\infty}^{0} w\phi(w - \lambda) dw < 0. \tag{13.17}$$

and by subtraction,

$$r_\lambda(\lambda, \mu) - r_\lambda(\lambda, 0) = 2 \int_{-|\mu|}^{|\mu|} |w|\phi(w - \lambda) dw. \tag{13.18}$$

After normalizing by $|r_\lambda(\lambda, 0)|$, the threshold risk derivative turns out to be monotone in

λ; a result reminiscent of the monotone likelihood ratio property. The proof is given at the end of the chapter.

Lemma 13.9 *For $\mu \neq 0$, the ratio*

$$V(\lambda, \mu) = \frac{r_\lambda(\lambda, \mu)}{|r_\lambda(\lambda, 0)|} \tag{13.19}$$

is strictly increasing in $\lambda \in [0, \infty)$, with $V(0, \mu) < 0$ and $V(\lambda, \mu) \nearrow \infty$ as $\lambda \to \infty$.

Integrated threshold risk. Define $B(\lambda, \pi)$ as above. Since $\mu \to r(\lambda, \mu)$ is a bounded (by $1 + \lambda^2$) and analytic function, $B(\lambda, \pi)$ is well defined and differentiable, with

$$(\partial/\partial\lambda)B(\lambda, \pi) = \int r_\lambda(\lambda, \mu)\pi(d\mu). \tag{13.20}$$

Now it can be shown that given π, there is always a unique best, i.e. risk minimizing, choice of threshold.

Proposition 13.10 *If $\pi = \delta_0$, then $\lambda \to B(\lambda, \pi)$ decreases to 0 as $\lambda \to \infty$. If $\pi \neq \delta_0$, then the function $\lambda \to B(\lambda, \pi)$ has a unique minimum $\lambda(\pi)$, $0 < \lambda(\pi) < \infty$, and is strictly decreasing for $\lambda < \lambda(\pi)$ and strictly increasing for $\lambda > \lambda(\pi)$.*

Proof First, $B(\lambda, \delta_0) = r(\lambda, 0)$ is strictly decreasing in λ by (13.17), and that it converges to 0 for large λ is clear from the risk function itself.

For $\pi \neq \delta_0$, it is convenient to normalize by $|r_\lambda(\lambda, 0)|$, and so to use (13.19) and (13.20) to define

$$W(\lambda) = \frac{(\partial/\partial\lambda)B(\lambda, \pi)}{|r_\lambda(\lambda, 0)|} = \int V(\lambda, \mu)\pi(d\mu).$$

From (13.16), it is clear that $W(0) < 0$, while Lemma 13.9 shows that $W(\lambda) \nearrow \infty$ as $\lambda \to \infty$. Hence there exists a zero, $W(\lambda_0) = 0$. Now for any λ

$$W(\lambda) - W(\lambda_0) = \int [V(\lambda, \mu) - V(\lambda_0, \mu)]\pi(d\mu),$$

and so strict monotonicity of $\lambda \to V(\lambda, \mu)$ for $\mu \neq 0$ guarantees that this difference is < 0 or > 0 according as $\lambda < \lambda_0$ or $\lambda > \lambda_0$. Consequently $(\partial/\partial\lambda)B(\lambda, \pi)$ has a single sign change from negative to positive, $\lambda(\pi) = \lambda_0$ is unique and the Proposition follows. □

The best threshold provided by the last proposition has a directional continuity property that will be needed for the minimax theorem below. (For proof, see Further Details).

Lemma 13.11 *If π_0 and π_1 are probability measures with $\pi_0 \neq \delta_0$, and $\pi_t = (1-t)\pi_0 + t\pi_1$, then $\lambda(\pi_t) \to \lambda(\pi_0)$ as $t \searrow 0$.*

A minimax theorem for thresholding. Just as in the full non-linear case, it is useful to think in terms of least favorable distributions for thresholding. Since the risk function $r(\lambda, \mu)$ is bounded and continuous in μ, the integrated threshold risk $B(\lambda, \pi)$ is linear and weakly continuous in π. Hence

$$B_S(\pi) = \inf_\lambda B(\lambda, \pi)$$

is concave and upper semicontinuous in π. Hence it attains its supremum on the weakly compact set $\mathfrak{m}_p(\tau)$, at a least favorable distribution π_0, say. Necessarily $\pi_0 \neq \delta_0$. Let $\lambda_0 = \lambda(\pi_0)$ be the best threshold for π_0, provided by Proposition 13.10.

The payoff function $B(\lambda, \pi)$ is *not* convex in λ, as is shown by consideration of, for example, the risk function $\lambda \to r_S(\lambda, 0)$ corresponding to $\pi = \delta_0$. On the other hand, $B(\lambda, \pi)$ is still linear in π, and this makes it possible to establish the following minimax theorem directly.

Theorem 13.12 *The pair (λ_0, π_0) is a saddlepoint: for all $\lambda \in [0, \infty)$ and $\pi \in \mathfrak{m}_p(\tau)$,*

$$B(\lambda_0, \pi) \leq B(\lambda_0, \pi_0) \leq B(\lambda, \pi_0), \tag{13.21}$$

and hence

$$\inf_{\lambda} \sup_{\mathfrak{m}_p(\tau)} B(\lambda, \pi) = \sup_{\mathfrak{m}_p(\tau)} \inf_{\lambda} B(\lambda, \pi)$$

and

$$\beta_{S,p}(\tau, \epsilon) = \sup\{B_S(\pi) : \pi \in \mathfrak{m}_p(\tau)\}. \tag{13.22}$$

Proof This is given as Theorem A.7, in which we take $\mathcal{P} = \mathfrak{m}_p(\tau)$. The hypotheses on $B(\lambda, \pi)$ are satisfied by virtue of Lemma 13.11 and Proposition 13.10. □

Remark. Proposition 13.10 shows that $B(\lambda, \pi)$ is quasi-convex in λ, and since it is also linear in π on a convex set, one could appeal to a general minimax theorem, e. g. Sion (1958). However, the general minimax theorems do not exhibit a saddlepoint, which emerges directly from the present more specialized approach.

With minimax threshold theorem in hand, we turn to understanding the threshold minimax risk $\beta_{S,p}(\tau, \epsilon)$ defined at (13.15).

Proposition 13.13 *The minimax Bayes threshold risk $\beta_{S,p}(\tau, \epsilon)$ also satisfies the properties (1) - (5) of $\beta_p(\tau, \epsilon)$ enumerated in Proposition 13.3.*

Proof Proposition 13.12 gives, in (13.22), a representation for $\beta_{S,p}(\tau, \epsilon)$ analogous to (13.4) for $\beta_p(\tau, \epsilon)$, and so we may just mimick the proof of Proposition 13.3. except in the case of monotonicity of in ϵ, compare Exercise 13.4. □

We have arrived at the destination for this section, a result showing that, regardless of the moment constraint, there is a threshold rule that comes quite close to the best non-linear minimax rule. It is an analog, for soft thresholding, of the Ibragimov-Has'minskii bound Theorem 4.16.

Theorem 13.14 *(i) For $0 < p \leq \infty$,*

$$\sup_{\tau, \epsilon} \frac{\beta_{S,p}(\tau, \epsilon)}{\beta_p(\tau, \epsilon)} = \Lambda(p) < \infty.$$

(ii) For $p \geq 2$, $\Lambda(p) \leq 2.22$.

Unpublished numerical work indicates that $\Lambda(1) = 1.6$, so that one may expect that even for $p < 2$, the inefficiency of the best threshold estimator is quite moderate. In addition, the proof below shows that the ratio

$$\mu_p(\eta) = \beta_{S,p}(\eta, 1)/\beta_p(\eta, 1) \to 1 \quad \text{as} \quad \eta \to 0, \infty. \tag{13.23}$$

Proof Most of the ingredients are present in Theorem 13.6 and Proposition 13.13, and we assemble them in a fashion parallel to the proof of Theorem 4.16. The scaling $\beta_{S,p}(\tau, \epsilon) = \epsilon^2 \beta_{S,p}(\tau/\epsilon, 1)$ reduces the proof to the case $\epsilon = 1$. The continuity of both numerator and denominator in $\eta = \tau/\epsilon$ shows that it suffices to establish (13.23).

For small η, we need only reexamine the proof of Theorem 13.6: the upper bounds for $\beta_p(\eta)$ given there are in fact provided by threshold estimators, with $\lambda = 0$ for $p \geq 2$ and $\lambda = \sqrt{2 \log \eta^{-p}}$ for $p < 2$.

For large η, use the trivial bound $\beta_{S,p}(\eta, 1) \leq 1$, along with the property (1) that $\beta_p(\eta)$ is decreasing in p to write

$$\mu_p(\eta) \leq 1/\beta_\infty(\eta) = 1/\rho_N(\eta, 1) \tag{13.24}$$

which decreases to 1 as $\eta \to \infty$, by (4.36) - (4.37). This completes the proof of (i).

For part (ii), use (13.24) to conclude that for any p and for $\eta \geq 1$, that $\mu_p(\eta) \leq 1/\rho_N(1, 1) \doteq 2.22$. For $\eta \leq 1$ and now using $p \geq 2$, we use $\beta_{S,p}(\eta) \leq \eta^2$ (compare (13.8)) to write $\mu_p(\eta) \leq \eta^2/\beta_\infty(\eta) = \eta^2/\rho_N(\eta, 1)$. The final part of the proof of Theorem 4.16 showed that the right side is bounded above by $1/\rho_N(1, 1) \doteq 2.22$. □

13.4 Minimax Bayes Risk for n−dimensional data.

We are at last able to return to the estimation of a n−dimensional parameter constrained to an ℓ_p ball and observed in white Gaussian noise of scale ϵ_n - compare model (13.1) and (13.2). The asymptotics of $R_N(\Theta_{n,p}(C_n))$ will be evaluated by the Bayes minimax approach of Section 4.10. This approach allows reduction to the basic one dimensional Bayes minimax problem studied in the previous section. We choose a collection of priors on $\mathbb{R}^n$

$$\mathcal{M}_n = \{\pi(d\theta) : E_\pi \sum_1^n |\theta_i|^p \leq C_n^p\}. \tag{13.25}$$

which relaxes the ℓ_p-ball constraint of $\Theta_{n,p}(C_n)$ to an in-mean constraint. The set $\mathcal{M}_n$ contains all point masses δ_θ for $\theta \in \Theta_n$, and is convex, so using (4.18), the minimax risk is bounded above by the Bayes minimax risk

$$\begin{aligned} R_N(\Theta_{n,p}(C_n)) &\leq B(\mathcal{M}_n) = \sup\{B(\pi), \pi \in \mathcal{M}_n\} \\ &:= B_{n,p}(C_n, \epsilon_n). \end{aligned}$$

We first show that that this upper bound is easy to evaluate in terms of a univariate quantity, and later investigate when the bound is asymptotically sharp.

Recall the dimension normalized radius $\eta_n = n^{-1/p}(C/\epsilon)$. This may be interpreted as the maximum scalar multiple in standard deviation units of the vector $(1, \ldots, 1)$ that is contained within $\Theta_{n,p}(C)$. Alternatively, it is the average signal to noise ratio measured in ℓ_p−norm: $(n^{-1} \sum |\theta_i/\epsilon|^p)^{1/p} \leq n^{-1/p}(C/\epsilon)$.

Proposition 13.15 *Let $\beta_p(\eta)$ denote the univariate Bayes minimax risk (13.4) for unit noise, and let $\eta_n = n^{-1/p}C_n/\epsilon_n$. Then*

$$B_{n,p}(C_n, \epsilon_n) = n\epsilon_n^2 \beta_p(\eta_n). \tag{13.26}$$

Proof We use the 'independence trick' to show that the maximisation in $B(\mathcal{M}_n)$ can be reduced to univariate priors. Indeed, for any $\pi \in \mathcal{M}_n$, construct a prior $\bar{\pi}$ from the product of the univariate marginals π_i of π. We have the chain of relations

$$B(\pi) \le B(\bar{\pi}) = \sum_i B(\pi_i) \le nB(\tilde{\pi}_1).$$

Indeed, Lemma 4.14 says that $\bar{\pi}$ is harder than π, yielding the first inequality. Bayes risk is additive for an independence prior: this gives the equality. For the second inequality, form the average $\tilde{\pi}_1 = n^{-1}\sum_i \pi_i$ and appeal to the concavity of Bayes risk.

The p-th moment of the univariate prior $\tilde{\pi}_1$ is easily bounded:

$$\int |\theta|^p d\tilde{\pi}_1 = n^{-1}\sum_1^n E_{\pi_i}|\theta_i|^p \le n^{-1}C_n^p,$$

because $\pi \in \mathcal{M}_n$, and so we can achieve the maximization of $B(\mathcal{M}_n)$ by restricting to univariate priors in $\mathfrak{m}_p(\tau)$ with $\tau = n^{-1/p}C_n$. In other words,

$$B_{n,p}(C_n, \epsilon_n) = n\beta_p(n^{-1/p}C_n, \epsilon_n)$$

and now the Proposition follows from the invariance relation 4(i) of Proposition 13.3. □

EXAMPLE 13.2 continued. Let us return to our original example in which $p = 1$, the noise $\epsilon_n = 1/\sqrt{n}$, and the radius $C_n = 1$. Thus $\eta_n = n^{-1} \cdot \sqrt{n} = n^{-1/2}$. It follows that

$$R_N(\Theta_{1,n}) \le B_{n,1}(C_n, \epsilon_n) = n \cdot (1/n) \cdot \beta_1(1/\sqrt{n}) \sim (\log n/n)^{1/2}, \tag{13.27}$$

and the next theorem will show that this rate and constant are optimal. Recall, for comparison, that $R_L(\Theta_{n,1}, \epsilon_n) = 1/2$.

The main result of this chapter describes the asymptotic behavior of the the non linear minimax risk $R_N(\Theta)$, and circumstances in which it is asymptotically equivalent to the Bayes minimax risk. In particular, except in the highly sparse settings to be discussed in the next section, the least favorable distribution for $R_N(\Theta)$ is essentially found by drawing n i.i.d rescaled observations from the least favorable distribution $\pi_p(\eta_n)$ for $\mathfrak{m}_p(\eta_n)$. We can thus incorporate the small η results from the previous section.

Theorem 13.16 *Let $R_N(C_n, \epsilon_n)$ denote the minimax risk* (13.3) *for estimation over the ℓ_p ball $\Theta_{n,p}(C_n)$ defined at* (13.2)*. Introduce normalized signal-to-noise ratios*

$$\eta_n = n^{-1/p}(C_n/\epsilon_n), \qquad \gamma_n = (2\log n)^{-1/2}(C_n/\epsilon_n).$$

For $2 \le p \le \infty$, if $\eta_n \to \eta \in [0, \infty]$, then

$$R_N(C_n, \epsilon_n) \sim n\epsilon_n^2 \beta_p(\eta_n). \tag{13.28}$$

For $0 < p < 2$, (a) if $\eta_n \to \eta \in [0, \infty]$ and $\gamma_n \to \infty$, then again (13.28) holds.

(b) If $\eta_n \to 0$ and $\gamma_n \to \gamma \in [0, \infty)$, *then*

$$R_N(C_n, \epsilon_n) \sim \begin{cases} \lambda_n^2 \epsilon_n^2 \big([\gamma^p] + \{\gamma^p\}^{2/p} \big), & \text{if } \gamma > 0 \\ \lambda_n \epsilon_n^2 \gamma_n^2 & \text{if } \gamma = 0, \end{cases} \tag{13.29}$$

where $[\cdot]$ *and* $\{\cdot\}$ *denote integer and fractional parts respectively.*

Before proving this result, we draw some implications.

Near minimaxity of thresholding in $\mathbb{R}^n$.

Let $\hat{\theta}_\lambda(y)$ denote soft thresholding at $\lambda\epsilon$ for data from n-dimensional model (13.1):

$$\hat{\theta}_{\lambda,i} = \hat{\delta}_S(y_i, \epsilon\lambda). \tag{13.30}$$

The minimax risk among soft thresholding estimators over the ℓ_p-ball $\Theta_{n,p}(C)$ is given by

$$R_S(C, \epsilon) = R_S(\Theta_{n,p}(C), \epsilon) = \inf_\lambda \sup_{\theta \in \Theta_{n,p}(C)} E_\theta \|\hat{\theta}_\lambda - \theta\|^2.$$

The next result is a fairly straightforward consequence of Theorems 13.14 and 13.16.

Theorem 13.17 *Adopt the assumptions of Theorem 13.16. If* $\eta_n \to \eta \in [0, \infty]$ *and, when* $p < 2$, *if also* $\gamma_n \to \infty$, *then there exists* $\Lambda(p) < \infty$ *such that*

$$R_S(C_n, \epsilon_n) \le \Lambda(p) R_N(C_n, \epsilon_n) \cdot (1 + o(1)). \tag{13.31}$$

If also $\eta_n \to 0$, *then*

$$R_S(C_n, \epsilon_n) \sim R_N(C_n, \epsilon_n).$$

The proof shows that $\Lambda(p)$ can be taken as the univariate quantity appearing in Theorem 13.14, as so from the remarks there, is likely to be not much larger than 1. Thus, in the high dimensional model (13.1), soft thresholding has bounded minimax efficiency among *all* estimators. In the case when $\eta_n \to 0$, the threshold choice $\lambda_n = \epsilon_n \sqrt{2 \log \eta_n^{-p}}$ is asymptotically minimax among all estimators.

Proof For a given vector $\theta = (\theta_i)$, define $\mu_i = \theta_i/\epsilon_n$ and let π_n denote the empirical measure $n^{-1} \sum_i \delta_{\mu_i}$. We can then rewrite the risk of soft thresholding at $\lambda\epsilon_n$, using our earlier notations, respectively as

$$E \sum_i (\hat{\theta}_{\lambda,i} - \theta_i)^2 = \epsilon_n^2 \sum_i r(\lambda, \mu_i) = n\epsilon_n^2 B(\lambda, \pi_n).$$

If $\theta \in \Theta_{n,p}(C_n)$, then the empirical measure satisfies a univariate moment constraint

$$\int |\mu|^p d\pi_n = n^{-1} \sum |\theta_i/\epsilon_n|^p \le n(C_n/\epsilon_n)^p = \eta_n^p. \tag{13.32}$$

Consequently $\pi_n \in \mathfrak{m}_p(\eta_n)$, and so

$$\inf_\lambda \sup_\theta E_\theta \|\hat{\theta}_\lambda - \theta\|^2 \le n\epsilon_n^2 \inf_\lambda \sup_{\pi \in \mathfrak{m}_p(\eta_n)} B(\lambda, \pi).$$

Now recalling definition (13.15) of $\beta_{S,p}(\eta)$ and then Theorem 13.14, the right side equals

$$n\epsilon_n^2\beta_{S,p}(\eta_n) \le \Lambda(p)\, n\epsilon_n^2\beta_p(\eta_n) = \Lambda(p)B_{n,p}(C_n,\epsilon_n),$$

where at the last equality we used the minimax Bayes structure Proposition 13.15. Putting this all together, we get

$$R_S(C_n,\epsilon_n) \le \Lambda(p)B_{n,p}(C_n,\epsilon_n)$$

and the conclusion (13.31) now follows directly from Theorem 13.16. If $\eta_n \to 0$, then $\beta_{S,p}(\eta_n) \sim \beta_p(\eta_n)$ by Theorem 13.6 and so we obtain the second statement. □

Remark 13.18 There is a fuller Bayes minimax theory for thresholding, which allows for a different choice of threshold in each co-ordinate. There is a notion of threshold Bayes minimax risk, $B_{S;n,p}(C,\epsilon)$ for priors satisfying (13.25), and a vector version of Theorem 13.14

$$B_{S;n,p}(C,\epsilon) \le \Lambda(p)B_{n,p}(C,\epsilon). \tag{13.33}$$

In this Bayes-minimax threshold theory, there is no advantage to allowing the thresholds to depend on the co-ordinate index: the minimax λ^* has all components the same. This provides some justification for the definition (13.30). Exercise 13.7 has details.

Proof of Asymptotic Equivalence Theorem 13.16

The approximation (13.28) follows from Proposition 13.15 once we establish the asymptotic equivalence of frequentist and Bayes minimax risks. The detailed behavior of R_N and the structure of the asymptotically least favorable priors and estimators follow from the results of the previous subsections on the univariate quantity $\beta_p(\eta,1)$ and will be described below. The asymptotic equivalence fails when γ_n remains bounded; the behavior described in part (b) of the theorem requires different tools, see Section 13.5.

Asymptotic equivalence of R_N and B. To show that the Bayes minimax bound is asymptotically sharp, we construct a series of asymptotically least favorable priors π_n that essentially concentrate on Θ_n. More precisely, following the recipe of Chapter 4.10, for each $\gamma < 1$ we construct priors π_n satisfying

$$B(\pi_n) \ge \gamma B_{n,p}(\gamma C_n,\epsilon_n) \tag{13.34}$$

$$\pi_n(\Theta_n) \to 1, \text{ and} \tag{13.35}$$

$$E_{\pi_n}\{\|\hat\theta_{v_n}\|^2 + \|\theta\|^2, \Theta_n^c\} = o(B_{n,p}(\gamma C_n,\epsilon_n)) \tag{13.36}$$

where $\hat\theta_{v_n}(y) = E_{\pi_n}(\theta|\theta\in\Theta_n, y)$.

In addition, we need the analog of (4.69), which in the present model becomes

$$\lim_{\gamma\nearrow 1}\lim_{n\to\infty}\frac{B_{n,p}(\gamma C_n,\epsilon_n)}{B_{n,p}(C_n,\epsilon_n)} = \lim_{\gamma\nearrow 1}\lim_{n\to\infty}\frac{\beta_p(\gamma\eta_n)}{\beta_p(\eta_n)} = 1. \tag{13.37}$$

As indicated at Lemma 4.28 and the following discussion, if we verify (13.34) - (13.37) we can conclude that $R_N(C_n,\epsilon_n) \sim B_{n,p}(C_n,\epsilon_n)$.

We will always define π_n by i.i.d rescaled draws from a univariate distribution $\pi_1(d\mu)$ on

$\mathbb{R}$ (in some cases $\pi_1 = \pi_{1n}$ depends on n): thus $\pi_n(d\theta) = \pi_{1n}^n(d\theta/\epsilon_n)$. Therefore, using (13.26), condition (13.34) can be reexpressed as

$$B(\pi_{1n}) \geq \gamma\beta_p(\gamma\eta_n), \tag{13.38}$$

and property (13.35) may be rewritten as

$$\pi_n(\Theta_n) = P_{\pi_{1n}}\{n^{-1}\sum |\mu_i|^p \leq \eta_n^p\}.$$

(1). Suppose first that $\eta_n \to \eta \in (0,\infty]$. Given $\gamma < 1$, there exists $M < \infty$ and a prior π_1 in $\mathfrak{m}_p(\gamma\eta)$ supported on $[-M, M]$ whose Bayes risk satisfies (13.38), compare Exercise 4.4. Noting $E_\pi|\mu|^p \leq \gamma^p\eta^p$ and that $|\mu_i| \leq M$, property (13.35) follows from the law of large numbers applied to the i.i.d. draws from π_1. Since $|\mu_i| \leq M$ under the prior π_n, both $\|\theta\|^2$ and $\|\hat{\theta}_v\|^2$ are bounded by $n\epsilon^2M^2$, the latter because $\|\hat{\theta}_v\|^2 \leq E_{\pi_n}\{\|\theta\|^2|\theta \in \Theta_n, y\}$. Hence the left side of (13.36) is bounded by $2n\epsilon_n^2M^2\pi_n(\Theta_n^c)$ while $B_{n,p}(\gamma C_n, \epsilon_n)$ is of exact order $n\epsilon_n^2$, and so (13.36) follows from (13.35). Property (13.37) follows from continuity of β_p, Proposition 13.3.

In summary, $R_N \sim n\epsilon_n^2\beta_p(\eta)$, and an asymptotically minimax estimator can be built from the Bayes estimator for a least favorable prior for $\mathfrak{m}_p(\eta)$.

Now suppose that $\eta_n \to 0$. First, observe from (13.7) that $\beta_p(\gamma\eta_n)/\beta_p(\eta_n) \to \gamma^{2\wedge p}$, so that (13.37) holds.

(2). Suppose first that $p \geq 2$. This case is straightforward: we know from the univariate case that the symmetric two point priors $\pi_n = \pi_{\eta_n} = (\delta_{\eta_n} + \delta_{-\eta_n})/2$ are asymptotically least favorable, so π_n satisfies (13.38) for large n. The corresponding measure π_n is already supported on Θ_n, so the remaining conditions are vacuous here.

In summary, $R_N \sim n\epsilon_n^2\eta_n^2$ and $\hat{\theta} = 0$ is asymptotically minimax.

(3). Suppose now that $p < 2$ (and still $\eta_n \to 0$). This case is more interesting. Given $\gamma < 1$, let π_{1n} be the sparse prior $\pi_p[\gamma\eta_n]$ of Definition 13.5 and set

$$\alpha_n = \alpha_p(\gamma\eta_n), \qquad \mu_n = \mu_p(\gamma\eta_n).$$

From the proof of Theorem 13.6 and Lemma 8.7, we have

$$\beta_p(\gamma\eta_n) \sim B(\pi_{1n}) \sim \alpha_n\mu_n^2. \tag{13.39}$$

Observe that the number N_n of non-zero components in a draw from $\pi_n = \pi_{1n}^n$ is a Binomial(n, α_n) variable, and that $\sum_i |\theta_i|^p = N_n\epsilon_n^p\mu_n^p$. The support requirement becomes

$$\{\theta \in \Theta_n\} = \{N_n \leq C_n^p/(\epsilon_n^p\mu_n^p)\}. \tag{13.40}$$

Rewriting the moment condition

$$\alpha_n\mu_n^p = (\gamma\eta_n)^p = \gamma^p n^{-1}C_n^p/\epsilon_n^p, \tag{13.41}$$

along with $EN_n = n\alpha_n$ and Chebychev's inequality leads to

$$\pi_n(\Theta^c) = P\{N_n > \gamma^{-p}n\alpha_n\} \leq c_{\gamma p}\mathrm{Var}\, N_n/(EN_n)^2. \tag{13.42}$$

We verify that $n\alpha_n \to \infty$ is equivalent to $\gamma_n \to \infty$. Indeed, from (13.41) and $\mu_n =$

$\mu(\alpha_n) \sim (2\log \alpha_n^{-1})^{1/2}$, we have

$$\gamma_n^p = \frac{(C_n/\epsilon_n)^p}{(2\log n)^{p/2}} = \frac{n\alpha_n}{\gamma^p}\left(\frac{\mu_n}{\sqrt{2\log n}}\right)^p \sim \frac{n\alpha_n}{\gamma^p}\left(1 - \frac{\log(n\alpha_n)}{\log n}\right)^{p/2}.$$

If $\gamma_n \to \infty$ then, for example arguing by contradiction, $n\alpha_n \to \infty$. Conversely, if $n\alpha_n \to \infty$, then argue separately the cases with $\log(n\alpha_n)/\log n \geq 1/2$ and $< 1/2$.

The right side of (13.42) converges to zero exactly when $EN_n = n\alpha_n \to \infty$ and so the previous paragraph shows that (13.35) follows from $\gamma_n \to \infty$. The proof of (13.36) also follows from the fact that $n\alpha_n \to \infty$, but is postponed to the appendix.

In summary,

$$R_N \sim n\epsilon_n^2 \eta_n^p (2\log \eta_n^{-p})^{1-p/2} \tag{13.43}$$

and soft thresholding with $\lambda_n = (2\log \eta_n^{-p})^{1/2}\epsilon_n$ provides an asymptotically minimax estimator. Hard thresholding is also asymptotically minimax so long as the thresholds are chosen in accordance with (13.13).

The role of the assumption that $\gamma_n \to \infty$ is to ensure that $EN_n \to \infty$. In other words, that Θ_n has large enough radius that the least favorable distribution in the Bayes minimax problem generates an asymptotically unbounded number of sparse spikes. Without this condition, asymptotic equivalence of Bayes and frequentist minimax risks can fail. For an example, return to the case $p = 1, \epsilon = n^{-1/2}$, but now with small radius $C_n = n^{-1/2}$. We have $\eta_n = n^{-1}$ and hence $B(C_n, \epsilon_n) \sim n^{-1}\sqrt{2\log n}$. However, the linear minimax risk is *smaller*: $R_L \sim n\epsilon_n^2\bar{\eta}^2 \sim n^{-1}$, and of course the non-linear minimax risk R_N is smaller still. In this case $EN_n = n\alpha_n = n\eta_n/\mu_n = 1/\mu_n \to 0$, since $\mu_n \sim \sqrt{2\log n}$.

The proof of equivalence thus demonstrates the existence of three different regimes for the least favorable distribution.

(i) *Dense:* $\eta_n \to \eta > 0$. The least favorable distribution $\pi_p(\eta) \in \mathfrak{m}_p(\eta)$ has high probability of yielding non zero values μ_i.

(ii) *Sparse:* $\eta_n \to 0, EN_n \to \infty$. Members of the least favorable sequence of two point distributions have an atom at 0 with probability increasing to 1, but still produce on average $EN_n = n\alpha_n \nearrow \infty$ non-zero "spikes" at $\mu_n = \sqrt{2\log \eta_n^{-p}}$ as $n \to \infty$.

(iii) *Highly sparse:* $\eta_n \to 0, \limsup_n EN_n < \infty$. In this case the signal to noise ratio n/ϵ_n is so small that only a finite number of non-zero spikes appear. The practical importance of this case has been highlighted by Mallat in a satellite image deconvolution/denoising application. Hence we devote the next section to its analysis.

13.5 Minimax Risk in the Highly Sparse Case

The moment constrained Bayesian approach to evaluating minimax risk $R_N(\Theta_n)$ fails in the highly sparse case because i.i.d. samples from the least favorable prior for the moment space $\mathfrak{m}_p(\eta_n)$ do not concentrate on Θ_n even asymptotically. In turn, this is a consequence of the small size of Θ_n, which entails that the expected number of non-zero 'spikes' is finite.

Thus in the highly sparse case, to obtain sharp lower bounds, we are forced to work with priors that concentrate entirely on Θ_n. We therefore abandon independence priors, and use instead exchangeable priors with a fixed number of non-zero components.

Remarks on Theorem 13.16, part (b) of $0 < p < 2$.

1. In this setting, $2\log \eta_n^{-p} = 2\log n - 2p\log(C_n/\epsilon_n) \sim 2\log n = \lambda_n^2$ and so the Minimax Bayes expression in the $p < 2$ sparse case, $n\epsilon_n^2\eta_n^p(2\log\eta_n^{-p})^{1-p/2} \sim C_n^p(\epsilon_n\lambda_n)^{2-p}$ agrees with the present result $\lambda_n^2\epsilon_n^2\gamma_n^2$ when when $\gamma_n \to \infty$, as expected. For finite limits γ however, it is strictly larger than (13.29) except when γ is an integer. Compare Figure 13.2.
2. In particular, note that when $\gamma < 1$ the bound $\epsilon_n^{2-p}C_n^p\lambda_n^{2-p}$ predicted by the Minimax Bayes method is too large and hence *incorrect* as the limiting value of R_N.
3. The parameter γ_n^p measures the number of 'spikes' of height about $\epsilon_n\sqrt{2\log n}$ that appear in the least favorable configuration for $\Theta_{n,p}(C_n)$.

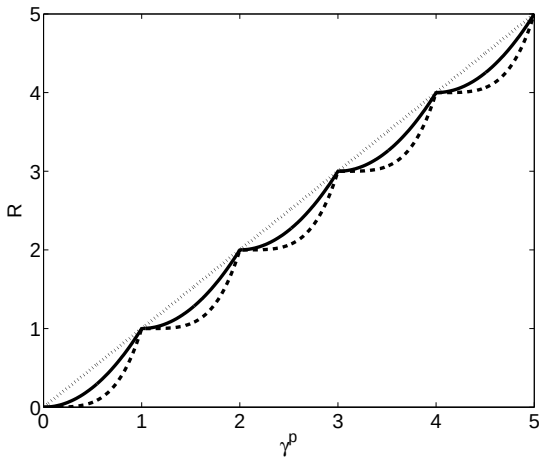

Figure 13.2 The function $R(\gamma) = [\gamma^p] + \{\gamma^p\}^{2/p}$ plotted against γ^p for $p = 1$ (solid) and $p = 1/2$ (dashed). The 45° line (dotted) shows the prediction of the Bayes minimax method.

Upper Bound. By scaling, it suffices to carry out the proof in the unit noise case $\epsilon_n = 1$. Use the risk bound (8.9) for soft thresholding with $\lambda_n = \sqrt{2\log n}$ and maximize over Θ:

$$R_N(\Theta) \le \sup_\Theta \sum_i r_S(\lambda_n, \theta_i) \le n r_S(\lambda_n, 0) + \sup_\Theta \sum_i \theta_i^2 \wedge (\lambda_n^2 + 1).$$

Using (8.7), $n r_S(\lambda_n, 0) \le c_1 n\phi(\lambda_n)/\lambda_n^3 \le c_2/(\log n)^{3/2} \to 0$. The upper bound now follows from the next lemma, on setting $\lambda^2 = \lambda_n^2 + 1 \sim \lambda_n^2$ as $n \to \infty$.

Lemma 13.19 *Let* $\gamma = C/\lambda$. *Then*

$$\sup_{\|\theta\|_p \le C} \sum_{i=1}^n \theta_i^2 \wedge \lambda^2 = \begin{cases} C^2 & \text{if } C \le \lambda \\ \lambda^2\big([\gamma^p] + \{\gamma^p\}^{2/p}\big) & \text{if } \lambda \le C \le n^{1/p}\lambda \\ n\lambda^2 & \text{if } C > n^{1/p}\lambda. \end{cases} \tag{13.44}$$

Proof If $C \le \lambda$, then the ℓ_p ball is entirely contained in the ℓ_∞ cube of side λ, and the maximum of $\sum \theta_i^2$ over the ℓ_p ball is attained at the spike $\theta^* = C(1, 0, \dots, 0)$ or permutations. This yields the first bound in (13.44). At the other extreme, if $C \ge n^{1/p}\lambda$, then the ℓ_∞ cube is contained entirely within the ℓ_p ball and the maximum of $\sum \theta_i^2$ is attained at the dense configuration $\theta^* = \lambda(1, \dots, 1)$.

If $\lambda < C < n^{1/p}\lambda$, the worst case vectors are subject to the ℓ_∞ constraint and are then permutations of the vector $\theta^* = (\lambda, \dots, \lambda, \mu, 0, \dots, 0)$ with n_0 components of size λ and the remainder $\mu = \{\lambda\}$ being determined by the ℓ_p condition:

$$n_0\lambda^p + \mu^p\lambda^p = C^p.$$

[To verify that this is indeed the worst case configuration, change variables to $u_i = \theta_i^p$: the problem is then to maximize the convex function $u \to \sum u_i^{2/p}$ subject to the convex constraints $\|u\|_1 \le C^p$ and $\|u\|_\infty \le \lambda^p$. This forces an extremal solution to occur on the boundary of the constraint set and to have the form described.] Thus $n_0 = [C^p/\lambda^p]$ and $\mu^p = \{C^p/\lambda^p\}$. Setting $\gamma^p = C^p/\lambda^p$, we obtain

$$\begin{aligned}\sum \theta_i^2 \wedge \lambda^2 &= n_0\lambda^2 + \mu^2\lambda^2 \\ &= \lambda^2[\gamma^p] + \lambda^2\{\gamma^p\}^{2/p}.\end{aligned}$$

□

Lower Bound. In the case $\gamma \le 1$, a single non-zero component occurs: this case was established in Proposition 8.13. In the general case, there will be $[\gamma]$ spikes of size approximately $\lambda_n = \sqrt{2\log n}$ and an additional spike of size approximately $\mu\lambda_n$, where $\mu = \{\gamma\}$. For notational simplicity, we consider in detail only the case $1 < \gamma \le 2$.

Perhaps the chief technical point of interest that appears, already in this case, is the need to bound sums of the form $S_n(\beta) = \sum_1^n e^{\beta z_k}$ for $z_1, \dots, z_n \overset{\text{i.i.d.}}{\sim} N(0,1)$ and β above the phase transition at $\lambda_n = \sqrt{2\log n}$. Indeed, suppose that $\gamma_n \to \gamma > 0$. In Lemma 8.15, we saw that when $\gamma \le 1$, the sums $S(\gamma_n\lambda_n)$ satisfy a weak law of large numbers, $S_n(\gamma_n\lambda_n) \sim e^{(1+\gamma^2)\log n}$. There is however a "phase change" at $\gamma = 1$. The following proposition, proved at the end of this section, is based on the discussion of the Random Energy Model in Talagrand (2003, Ch. 1.1, 2.2)

Proposition 13.20 *With the previous definitions,*

$$\frac{\log S_n(\gamma_n\lambda_n)}{\log n} \xrightarrow{\ p\ } \begin{cases} 1+\gamma^2 & \gamma \le 1 \\ 2\gamma & \gamma > 1. \end{cases} \tag{13.45}$$

We return to the lower bound in the case $1 < \gamma \le 2$. As in the proof of the single spike case, Proposition 8.13, the size of the primary spike needs to be somewhat smaller than λ_n. Thus, choose τ_n so that both $\tau_n \sim \lambda_n$ and $\lambda_n - \tau_n \to \infty$ and also fix $0 < \mu < 1$. The near least favorable prior π_n is defined by

$$\theta = \tau_n e_I + \mu\tau_n e_{I'},$$

where $I \ne I'$ are chosen at random from $\{1, \dots, n\}$. Clearly for large n, we have $\sum |\theta_i|^p = \tau_n^p + \mu^p\tau_n^p \le \gamma_n^p$ with probability one, and so supp $\pi_n \subset \Theta_{n,p}(C_n)$. We will show that

$$B(\pi_n) \ge \tau_n^2(1+\mu^2)(1+o(1)) \sim \lambda_n^2(1+\mu^2). \tag{13.46}$$

Since this works for all $\mu < \gamma - 1$, we conclude that $R_N(C_n, \epsilon_n) \geq \lambda_n^2[1+(\gamma-1)^2](1+o(1))$.

Let $\theta^{(k)}$ denote the $n(n-1)$ equiprobable support points of this prior. The risk function $r(\hat{\theta}_\pi, \theta^{(k)})$ is symmetric in k, and so the Bayes risk is given by any one of these values, $B(\pi_n) = r(\hat{\theta}_\pi, \theta^{(1)})$ say. Write the posterior probabilities $p_i(y) = P(I = i|y)$ and $p_i'(y) = P(I' = i|y)$. The components of the Bayes estimator are given by

$$\hat{\theta}_{\pi,i} = E(\theta_i|y) = \tau_n p_i(y) + \mu\tau_n p_i'(y).$$

Henceforth, write E_{1n} and P_{1n} for expectation and distribution conditional on $\theta = \theta^{(1)} = \tau_n e_1 + \mu\tau_n e_2$. We retain only the first two co-ordinates in the summed squared error loss, and then rely on the inequality

$$B(\pi_n) \geq \tau_n^2 E_{1n}[(p_1(y) + \mu p_1'(y) - 1)^2 + (p_2(y) + \mu p_2'(y) - \mu)^2].$$

It therefore suffices to show that $p_i(y)$ and $p_i'(y)$ converge in P_{1n}-probability to zero, for then we obtain (13.46).

Using the i.i.d. Gaussian structure, the joint posterior of (I, I') is seen to be

$$P(I = i, I' = i'|y) = \exp\{\tau_n y_i + \mu\tau_n y_{i'}\}/D_n \tag{13.47}$$

Writing $S_n'(\beta) = \sum_1^n e^{\beta y_k}$, the denominator

$$D_n = \sum_{j \neq j'} \exp\{\tau_n y_j + \mu\tau_n y_{j'}\} = S_n'(\tau_n)S_n'(\mu\tau_n) - S_n'((1+\mu)\tau_n) \tag{13.48}$$

Everything we need will follow from two convergence properties. Below, we write $X_n \sim Y_n$ to mean that the ratio converges in probability to 1. First, under P_{1n}, for $0 < \mu \leq 1$ we have for each fixed k

$$e^{\mu\tau_n y_k} = o_p(S_n(\mu\tau_n)), \qquad \text{and} \quad S_n'(\mu\tau_n) \sim S_n(\mu\tau_n) \sim e^{(\lambda_n^2 + \mu^2\tau_n^2)/2}. \tag{13.49}$$

Second, there is qualitatively different behavior above the phase transition: for $0 < \mu < 1$

$$e^{(1+\mu)\tau_n y_k} = o_p(e^{(3+\mu^2)\tau_n^2/2}), \qquad \text{and} \quad S_n'((1+\mu)\tau_n) = o_p(e^{(3+\mu^2)\tau_n^2/2}). \tag{13.50}$$

Indeed, assuming for now these properties, we can simplify D_n:

$$D_n \sim S_n'(\tau_n)S_n'(\mu\tau_n) \gg e^{(3+\mu^2)\tau_n^2/2},$$

where $a_n \gg b_n$ means that the ratio converges to ∞. Hence, using also (13.47),

$$p_i(y) = e^{\tau_n y_i}\Big(\sum_{i' \neq i} e^{\mu\tau_n y_{i'}}\Big)\Big/ D_n \leq e^{\tau_n y_i}/S_n'(\tau_n) \cdot (1 + o_p(1)) = o_p(1).$$

Interchanging i with i', and τ_n with $\mu\tau_n$, we get the same conclusion for $p_i'(y)$, and are done.

Finally we must verify (13.49)–(13.50). First, note that under P_{1n}, we have $y_1 = \tau_n + z_1$, $y_2 = \mu\tau_n + z_2$ and $y_k = z_k$ for $k \geq 3$. In the first equality of each of the displays, we need therefore only consider y_1, since under P_{1n} it is the stochastically largest of each y_k. The approximation for $S_n(\mu\tau_n)$ follows from Lemma 8.15 and $n = e^{\lambda_n^2}/2$. To see that $e^{\mu\tau_n y_1} = o_p(S_n(\mu\tau_n))$, use (8.52), after extracting an additional term $(1-\mu)^2\tau_n^2$. We

may then also conclude that $S_n'(\mu\tau_n) \sim S_n(\mu\tau_n)$ as the two differ in at most the first two summands.

Turning to (13.50) and behavior above the phase transition, note first that Proposition 13.20 implies that

$$\log S_n((1+\mu)\tau_n) \sim 2(1+\mu)\log n \sim (1+\mu)\tau_n^2,$$

and that for $0 < \mu < 1$, we have $(3+\mu^2) - 2(1+\mu) = (1-\mu)^2 > 0$. Hence the bound of (13.50) holds for $S_n((1+\mu)\tau_n)$. Again for $0 < \mu < 1$ one checks that $(3+\mu^2)\tau_n^2 - 2(1+\mu)\tau_n(\tau_n + z) \to \infty$ for each z. So the first part of (13.50) follows for $k = 1$, and hence for each k, and the second part is verified as well.

Remark. If one were falsely to approximate $S_n((1+\mu)\tau_n)$ by $e^{\lambda_n^2 + (1+\mu)^2\tau_n^2/2}$, then this term would seem to dominate $S_n(\tau_n)S_n(\mu\tau_n)$ for $\mu > 1/2$ – this shows how the phase transition 'rescues' the use of factorization (13.48).

Proof of Proposition 13.20 1°. Denote the function on the right side of (13.45) by $p(\gamma)$, and let $p_n(\gamma) = E\log S_n(\gamma\lambda_n)/\log n$. Concentration of measure shows that it suffices to establish (13.45) for the expectation $p_n(\gamma_n)$. Indeed, for the function $f(z) = \log\sum_1^n e^{\beta z_k}$, we find

$$\sum(\partial f/\partial z_k)^2 = \beta^2 \sum e^{2\beta z_k}/(\sum e^{\beta z_k})^2 \le \beta^2$$

and so f is Lipschitz with Lipschitz constant β. Hence, by Proposition 2.10,

$$P\{|\log S_n(\gamma_n\lambda_n) - E\log S_n(\gamma_n\lambda_n)| \ge \epsilon\lambda_n^2\} \le 2\exp\{-\epsilon^2\lambda_n^2/(2\gamma_n^2)\} \to 0.$$

2°. The upper bound $p_n(\gamma) \le p(\gamma)$ is given as Proposition C.9, setting there $\beta = \gamma\lambda_n$.

3°. We finally have to show that $\liminf p_n(\gamma_n) \ge p(\gamma)$. Let $N_n = \{j : z_j \ge s\lambda_n\}$; clearly $N_n \sim \mathrm{Bin}(n, q_{ns})$, where $q_{ns} = \tilde\Phi(s\lambda_n)$ is within a factor of two of $\phi(s\lambda_n)/s\lambda_n$, from the Mill's ratio inequalities (8.63) (for large n). Consider the event $A_n = \{N_n \le EN_n/2\}$, a short calculation using Chebychev's inequality shows that $P(A_n) \le 4/(nq_{ns}) \to 0$ if $s < 1$.

On A^c, we have $S_n \ge (n/2)q_{ns}e^{s\gamma_n\lambda_n^2}$, and so

$$E\log S_n \ge P(A_n^c)[\log(n/2) + \log q_{ns} + s\gamma_n\lambda_n^2] + EI_{A_n}\log S_n.$$

The second term on the right side is easily bounded: as $S_n \ge e^{\gamma_n\lambda_n z_1}$,

$$EI_{A_n}\log S_n \ge \gamma_n\lambda_n Ez_1 I_{A_n} \ge -\gamma_n\lambda_n E|z_1| = o(\log n).$$

For the first term, $P(A_n^c) \to 1$ and from the Mill's ratio bounds, $\log q_{ns} \sim -s^2\lambda_n^2/2$. Hence for $s < 1$,

$$\liminf p_n(\gamma_n) \ge 1 - s^2 + 2s\gamma.$$

For $\gamma < 1$, choose $s = \gamma$, while for $\gamma \ge 1$, let $s = 1 - \epsilon$ and then take ϵ small. □

13.6 Appendix: Further details

3°. *Proof of Lemma 13.9.* That $V(0,\mu) < 0$ follows from (13.16). From (13.18), we have

$$V(\lambda,\mu) = 2R(\lambda,\mu) - 1 \tag{13.51}$$

where, after writing ϕ_λ for $\phi(w-\lambda)$,

$$R(\lambda,\mu) = \int_N |w|\phi_\lambda \Big/ \int_D |w|\phi_\lambda = N(\lambda)/D(\lambda),$$

and the intervals $N = (-|\mu|, |\mu|)$ and $D = (-\infty, 0)$. One then checks that

$$\begin{aligned} D(\lambda)^2(\partial/\partial\lambda)R(\lambda,\mu) &= D(\lambda)N'(\lambda) - N(\lambda)D'(\lambda) \\ &= \int_D |w|\phi_\lambda \int_N w|w|\phi_\lambda - \int_D w|w|\phi_\lambda \int_N |w|\phi_\lambda, \end{aligned}$$

after cancellation, and each term on the right side is positive when $\mu \neq 0$ and $\lambda > 0$ since

$$\int_N w|w|\phi_\lambda = \int_0^{|\mu|} w^2[\phi(w-\lambda) - \phi(w+\lambda)]dw > 0.$$

This shows the monotonicity of $V(\lambda,\mu)$ in λ. We turn to the large λ limit: writing μ for $|\mu|$, a short calculation shows that as $\lambda \to \infty$

$$\begin{aligned} N(\lambda) &\geq \int_0^\mu w\phi(w-\lambda)dw = \lambda[\tilde\Phi(\lambda-\mu) - \tilde\Phi(\lambda)] + \phi(\lambda) - \phi(\mu-\lambda) \sim \frac{\lambda}{\mu}\phi(\lambda-\mu) \\ D(\lambda) &= -\lambda\tilde\Phi(\lambda) + \phi(\lambda) \sim \phi(\lambda)/\lambda^2, \end{aligned}$$

so that $R(\lambda,\mu) \geq \lambda\mu e^{\lambda\mu - \mu^2/2}(1+o(1)) \to \infty$ as $\lambda \to \infty$.

4°. *Proof of Lemma 13.11.* Let $D(\lambda,\pi) = \partial_\lambda B(\lambda,\pi)$; from Proposition 13.10 we know that $\lambda \to D(\lambda,\pi)$ has a single sign change from negative to positive at $\lambda(\pi)$. The linearity of $\pi \to D(\lambda,\pi)$ yields

$$D(\lambda,\pi_t) = D(\lambda,\pi_0) + tD(\lambda,\pi_1-\pi_0) = D(\lambda) + tE(\lambda),$$

say. Given $\epsilon > 0$, a sufficient condition for $\lambda_t = \lambda(\pi_t)$ to satisfy $|\lambda_t - \lambda_0| < \epsilon$ is that

$$D(\lambda_0+\epsilon) + tE(\lambda_0+\epsilon) > 0, \quad \text{and} \quad D(\lambda) + tE(\lambda) < 0$$

for all $\lambda \leq \lambda_0 - \epsilon$. Since $D(\lambda_0-\epsilon) < 0 < D(\lambda_0+\epsilon)$ and $\lambda \to E(\lambda)$ is continuous and bounded on $[0, \lambda_0+1]$, the condition clearly holds for all $t > 0$ sufficiently small.

5°. *Proof of* (13.36). From (13.42) we have that on Θ_n, necessarily $N_n \leq EN_n/\gamma^p$, and so

$$\|\hat\theta_{v_n}\|^2 \leq E\left\{\|\theta\|^2 \;\middle|\; \theta \in \Theta_n, y\right\} = \epsilon_n^2\mu_n^2 E\left\{N_n \;\middle|\; \theta \in \Theta_n, y\right\} \leq \epsilon_n^2\mu_n^2 EN_n/\gamma^p.$$

Thus

$$E_{\pi_n}\{\|\hat\theta_{v_n}\|^2 + \|\theta\|^2, \Theta_n^c\} \leq \gamma^{-p}\epsilon_n^2\mu_n^2 E_{\pi_n}\{EN_n + N_n, \Theta_n^c\},$$

whereas using Proposition 13.15 and (13.39),

$$B_{n,p}(\gamma C_n, \epsilon_n) \sim n\epsilon_n^2\alpha_n\mu_n^2 \sim \epsilon_n^2\mu_n^2 EN_n.$$

The ratio of the two preceding displays converges to zero because $n\alpha_n \to \infty$ implies both $P(\Theta_n^c) \to 0$ and $E|N_n - EN_n|/EN_n \le \sqrt{\mathrm{Var}\, N_n}/EN_n \le (n\alpha_n)^{-1/2} \to 0$. Since the order of magnitude of $B_{n,p}(\gamma C_n, \epsilon_n)$ does not depend on γ, we have shown (13.36).

13.7 Notes

`[To include:]` DJ 94 – ℓ_q losses and $p < q$ vs $p \ge q$.]

[Feldman, Mallat?, REM discussion/refs.]

Remark. `[later?]` A slightly artificial motivation for the ℓ_p balls model comes from the continuous Gaussian white noise model $dY_t = f(t)dt + n^{-1/2}dW$, $t \in [0,1]$ in which f has the form $f = \sum_1^n \theta_k \phi_{n,k}$, where $\phi_{n,k}(t) = n^{1/2}\phi(nt-k)$. If ϕ is the indicator of the unit interval $[0,1]$, then $\|\hat{f} - f\|_{L_2}^2 = \sum(\hat{\theta}_i - \theta_i)^2$ and since

$$\int |f|^p = n^{p/2-1} \sum_1^n |\theta_k|^p,$$

an L_p norm constraint on f corresponds to an ℓ_p constraint on θ. This connection becomes much more natural and useful in the context of sequence space characterizations of Besov and Triebel classes of functions to be discussed later (`ref`).

Remark. The discussion here will be confined to squared error loss, but the main results and phenomena remain valid for ℓ_q- loss, $0 < q < \infty$, with the non-linearity phenomena appearing in case $q < p$. Details are given in Donoho and Johnstone (1994b).

Exercises

13.1 Consider the sparse prior $\pi_{\alpha,\mu(\alpha)}$ specified by equation (8.28) with sparsity α and overshoot $a = (2\log\alpha^{-1})^{\gamma/2}$. Let $\eta > 0$ be small and consider the moment constraint equation $\alpha\mu(\alpha)^p = \eta^p$. Show that $m(\alpha) = \alpha\mu(\alpha)^p$ has $m(0+) = 0$ and is increasing for $\alpha > 0$ sufficiently small. Show also, for example numerically, that for some γ, $m(\alpha)$ ceases to be monotone for larger values of α.

13.2 Use Lemma 8.5 to establish (13.14) by considering in turn $\mu \in [0, \sqrt{5}]$, $\mu \in [\sqrt{5}, \lambda]$ and $\mu \ge \lambda$. Give an expression for $\lambda_0(p)$.

13.3 *(Structure of the p-th moment least favorable distributions.)* Establish Proposition 13.4 by mimicking the proof of Proposition 8.9, allowing for the fact that $\mathfrak{m}_p(\tau)$ is weakly compact.
(a) Let $\nu_\tau(d\theta) = \tau^{-p}|\theta|^p \pi_\tau(d\theta)$ and use strict monotonicity of $\beta_p(\tau)$ to show that ν_τ has total mass 1.
(b) Let $r(\theta) = \tau^p|\theta|^{-p}[r(\hat{\theta}_\tau, \theta) - r(\hat{\theta}_\tau, 0)]$ and verify that for $\theta \ne 0$,

$$r(\theta) \le \int r(\theta')\nu_\tau(d\theta').$$

(c) Complete the argument using Lemma 4.17 and Exercise 4.2.

13.4 *(Monotonicity of threshold minimax risk.)* Let $r(\lambda, \mu; \epsilon)$ denote the MSE of soft thresholding at λ when $x \sim N(\mu, \epsilon^2)$, and $r(\lambda, \mu) = r(\lambda, \mu; 1)$. Show that the proof of monotonicity of $\epsilon \to \beta_{S,p}(\tau, \epsilon)$ can be accomplished via the following steps:
(a) It suffices to show that if $\epsilon' < \epsilon$, then $r(\lambda\epsilon', \mu; \epsilon') \le r(\lambda\epsilon, \mu; \epsilon)$ for all μ and λ.
(b) Writing $\mu' = \mu/\rho$, verify that if $\rho \ge 1$,

$$(d/d\rho)r(\rho\lambda, \mu; \rho) = 2\rho r(\lambda, \mu') - \mu r_\mu(\lambda, \mu') \ge E_{\mu'}\{(\delta_\lambda(x) - \mu)^2; |x| \ge \lambda\}.$$

13.5 *(Motivation for minimax threshold value.)* Show that the value $\lambda_p(\eta)$ minimizing the right side of integrated risk bound (13.9) satisfies $\lambda_p(\eta) \sim \sqrt{2\log \eta^{-p}}$ as $\eta \to 0$.

13.6 *(Proof Outline for Proposition 13.7.)* (a) Let $\Phi_\mu(I_\lambda) = \int_{-\lambda}^{\lambda} \phi(x-\mu)dx$ and show that

$$p^2\kappa^2 D_\kappa^2 r(\lambda, \kappa^{1/p}) = \mu^2 \Phi_\mu(I_\lambda)\{(2-p)\mu^{-1} + D_\mu \log \Phi_\mu(I_\lambda)\}.$$

(b) For $0 < p \le 2$, there exists $\kappa_c > 0$ such that the function $\kappa \to r(\lambda, \kappa^{1/p})$, is convex for $\kappa \in (0, \kappa_c]$ and concave for $\kappa \in [\kappa_c, \infty)$. [Assume, from e.g. Prékopa (1980, Theorem 3 and Sec. 3), that $\mu \to \Phi_\mu(I_\lambda)$ is log-concave on $(0, \infty)$.]
(c) Show that the extreme points of $\mathfrak{m}_p(\tau)$ have the form $(1-\alpha)\delta_{\mu_0} + \alpha\delta_{\mu_1}$, but that it suffices to take $\mu_0 = 0$, and hence recover (13.10).
(d) Show that (13.11) is equivalent to solving for $\mu = \mu_\lambda$ in

$$R(\mu) := \frac{\int_0^\mu s\Phi_s(I_\lambda)ds}{\Phi_\mu(I_\lambda)\int_0^\mu sds} = \frac{2}{p}.$$

Show that $R(\lambda + u) \to 1/\tilde{\Phi}(u)$ uniformly on compact u-intervals and so conclude (13.12).

13.7 *(Bayes minimax theory for thresholding.)* Let $\lambda = (\lambda_i)$ be a vector of thresholds, and define now $\hat{\theta}_\lambda$ by $\hat{\theta}_{\lambda,i}(y) = \hat{\delta}_S(y_i, \lambda_i\epsilon)$. If π is a prior on $\theta \in \mathbb{R}^n$, set $B(\lambda, \pi) = E_\pi E_\theta \|\hat{\theta}_\lambda - \theta\|^2$. Define $\mathcal{M}_n$, the priors satisfying the $\Theta_{n,p}(C)$ constraint in mean, by (13.25) and then define the Bayes-minimax threshold risk by

$$B_{S;n,p}(C, \epsilon) = \inf_\lambda \sup_{\pi \in \mathcal{M}_n} B(\lambda, \pi).$$

(a) Let $B_S(\pi) = \inf_\lambda B(\lambda, \pi)$. Show that a minimax theorem holds

$$\inf_\lambda \sup_{\mathcal{M}_n} B(\lambda, \pi) = \sup_{\mathcal{M}_n} B_S(\pi),$$

and that a saddlepoint (λ^*, π^*) exists.
(b) Show that the least favorable π^* has i.i.d. co-ordinates and that the components λ_i^* of the minimax threshold do not depend on i.
(c) Conclude that the vector bound (13.33) holds.

14

Sharp minimax estimation on Besov spaces

14.1 Introduction

In previous chapters, we developed bounds for the behavior of minimax risk $R_N(\Theta(C), \epsilon)$ over Besov bodies $\Theta(C)$. In Chapters 9 and 10, we showed that thresholding at $\sqrt{2 \log \epsilon^{-1}}$ led to asymptotic minimaxity up to logarithmic factors $O(\log \epsilon^{-1})$, while in Chapter 12 we established that estimators derived from complexity penalties achieved asymptotic minimaxity up to constant factors.

In this chapter, we use the minimax Bayes method to study the exact asymptotic behavior of the minimax risk, at least in the case of squared error loss. The "price" for these sharper optimality results is that the resulting optimal estimators are less explicitly described and depend on the parameters of Θ.

In outline, we proceed as follows. In Section 14.3 we replace the minimax risk $R_N(\Theta(C), \epsilon)$ by an upper bound, the minimax Bayes problem with value $B(C, \epsilon)$, and state the main results of this chapter.

In Section 14.4, we begin study of the optimization over prior probability measures required for $B(C, \epsilon)$, and show that the least favorable distribution necessarily has independent co-ordinates, and hence the corresponding minimax rule is separable, i.e. acts coordinatewise. The $B(C, \epsilon)$ optimization is then expressed in terms of the univariate Bayes minimax risks $\beta_p(\tau, \epsilon)$ studied in Chapter 13.

In Section 14.5, a type of 'renormalization' argument is used to deduce the dependence of $B(C, \epsilon)$ on C and ϵ up to a periodic function of C/ϵ. At least in some cases, this function is almost constant.

In Section 14.6, we show that the upper bound $B(C, \epsilon)$ and minimax risk $R_N(\Theta(C), \epsilon)$ are in fact asymptotically equivalent as $\epsilon \to 0$, by showing that the asymptotically least favorable priors are asymptotically concentrated on $\Theta(C)$.

The minimax risk of *linear* estimators is evaluated in Section 14.7, using notions of quadratic convex hull from Chapter 4—revealing suboptimal rates of convergence when $p < 2$.

In contrast, *threshold* estimators, Section 14.4 can be found that come within a constant factor of $R_N(\Theta(C), \epsilon)$ over the full range of p; these results rely on the univariate Bayes minimax properties of thresholding established in Chapter 13.3.

14.2 The Dyadic Sequence Model

We consider the Gaussian sequence model (3.1) with countable index set, in the dyadic indexing regime

$$y_I = \theta_I + \epsilon z_I \tag{14.1}$$

where I denotes the pair (j,k), supposed to lie in the set $\mathcal{I} = \cup_{j\geq -1}\mathcal{I}_j$, where for $j \geq 0$, $\mathcal{I}_j = \{(j,k) : k = 1,\ldots,2^j\}$ and the exceptional $\mathcal{I}_{-1} = \{(-1,0)\}$.

Parameter Space. We restrict attention, for simplicity of exposition, to the Besov bodies

$$\Theta_p^\alpha(C) = \{\theta = (\theta_I) : \|\theta_{j\cdot}\|_p \leq C2^{-aj} \text{ for all} j\}, \qquad a = \alpha + 1/2 - 1/p.$$

This is the $q = \infty$ case of the Besov bodies $\Theta_{p,q}^\alpha$ considered in earlier chapters. They are supersets of the other cases, since $\Theta_{p,q}^\alpha(C) \subset \Theta_{p,\infty}^\alpha(C)$ for all q, and from the discussion of the general case in Donoho and Johnstone (1998), it is seen that the *rate* of convergence as $\epsilon \to 0$ is the same for all q.

We note that Θ is solid and orthosymmetric, and compact when $\alpha > (1/p - 1/2)_+$. (Exercise 14.1(a)).

We focuse on global ℓ_2 estimation: that is we evaluate estimators with the loss function $\|\hat\theta - \theta\|_2^2 = \sum(\hat\theta_I - \theta_I)^2$ and the minimax risk

$$R_N(\Theta,\epsilon) = \inf_{\hat\theta}\sup_{\Theta} E_\theta\|\hat\theta - \theta\|_2^2.$$

In principle, a similar development could be carried out for the ℓ_p loss $\|\hat\theta - \theta\|_p^p = \sum|\hat\theta_I - \theta_I|^p$, or weighted losses of the form $\sum_j 2^{jr}\sum_k|\hat\theta_{jk} - \theta_{jk}|^p$.

14.3 Bayes minimax problem

We relax the 'hard' constraint that $\|\theta\|_{b_{p\infty}^\alpha} \leq C$ by a constraint 'in mean' with respect to a prior π. We define a class of priors

$$\mathcal{M} = \mathcal{M}_p^\alpha(C) = \{\pi(d\theta) : E_\pi\sum_k|\theta_{jk}|^p \leq C^p2^{-ajp} \text{ for all } j\}.$$

As in earlier chapters, define the integrated risk $B(\hat\theta,\pi) = E_\pi E_\theta\|\hat\theta - \theta\|^2$ and the Bayes minimax risk

$$B(\mathcal{M},\epsilon) = \inf_{\hat\theta}\sup_{\pi\in\mathcal{M}} B(\hat\theta,\pi). \tag{14.2}$$

Since $\Theta \subset \mathcal{M}$, $R_N(\Theta,\epsilon) \leq B(\mathcal{M},\epsilon)$. We will again see that it is (relatively) easier to study and evaluate the Bayes minimax risk $B(\mathcal{M},\epsilon)$. To emphasize the dependence on C and ϵ, we sometimes write $B(C,\epsilon)$ for $B(\mathcal{M},\epsilon)$.

Our results build on the univariate Bayes minimax problem introduced in Section 13.2, with minimax risk $\beta_p(\tau,\epsilon)$ corresponding to observation $y = \theta + \epsilon z$ and moment constraint $E_\pi|\theta|^p \leq \tau^p$ for the prior π. We use the notation $\beta_p(\eta)$ for the normalized problem with noise $\epsilon = 1$. Let π_η denote the least favorable prior for $\beta_p(\eta)$ and $\delta_\eta = \delta(x;\eta)$ denote the corresponding Bayes-minimax estimator, so that $B(\delta_\eta,\pi_\eta) = B(\pi_\eta) = \beta_p(\eta)$.

A first key property of the Bayes minimax problem is that minimax estimators are *separable* into functions of each individual coordinate:

Theorem 14.1 *Suppose that* $0 < p \le \infty$ *and* $\alpha > (1/p - 1/2)_+$. *A minimax estimator for* $B(\mathcal{M}, \epsilon)$ *has the form*

$$\hat{\theta}_I^*(y) = \hat{\delta}_j^*(y_I), \qquad I \in \mathcal{I}, \tag{14.3}$$

where $\hat{\delta}_j^*(y)$ *is a scalar non-linear function of the scalar* y. *In fact there is a one parameter family of functions from which the minimax estimator is built: Let* $\hat{\delta}(x;\eta)$ *be the Bayes minimax estimator for the univariate Bayes minimax problem* $\beta_p(\eta)$ *recalled above. Then*

$$\hat{\delta}_j^*(y_I) = \epsilon\hat{\delta}(y_I/\epsilon; \eta_j), \tag{14.4}$$

where $\eta_j = (C/\epsilon)2^{-(\alpha+1/2)j}$.

For $p \neq 2$, the explicit form of $\hat{\delta}(\cdot;\eta)$ is not available, but we will see that useful approximations of $\hat{\delta}(\cdot;\eta)$ by threshold rules are possible.

Second, the exact asymptotic structure of the Bayes minimax risk can be determined.

Theorem 14.2 *Suppose that* $0 < p \le \infty$ *and* $\alpha > (1/p - 1/2)_+$. *Then* $B(C,\epsilon) < \infty$ *and*

$$B(C,\epsilon) \sim P(C/\epsilon) \cdot C^{2(1-r)}\epsilon^{2r}, \qquad \epsilon \to 0,$$

where $r = 2\alpha/(2\alpha+1)$ *and* $P(\cdot) = P(\cdot; \alpha + 1/2, p)$ *is a continuous, positive periodic function of* $\log_2(C/\epsilon)$.

This periodic function might be viewed as reflecting the arbitrary choice of the location of frequency octaves that is implicit in discrete dyadic wavelet bases.

Third, we establish asymptotic equivalence of frequentist and minimax Bayes risk.

Theorem 14.3 *For* $0 < p \le \infty$ *and* $\alpha > (1/p - 1/2)_+$,

$$R_N(\Theta,\epsilon) = B(C,\epsilon)(1 + o(1)), \qquad \epsilon \to 0. \tag{14.5}$$

Combining Theorems 14.2–14.3, we conclude that the estimator $\hat{\theta}^*$ is *asymptotically* minimax for R as $\epsilon \to 0$. In short: *a separable nonlinear rule is asymptotically minimax.*

14.4 Separable rules

We begin the proof of Theorem 14.1 by noting that $\mathcal{M}$ is convex—this follows immediately from the linearity in π of the expectation constraints. This allows use of the minimax theorem Theorem 4.11 to write that $B(\mathcal{M}) = \sup_{\mathcal{M}} B(\pi)$, so that we may look for a least favorable prior. The optimization is simplified by noting that $\mathcal{M}$ is closed under the operation of replacing π by the levelwise average of marginals. Given a prior $\pi \in \mathcal{M}$, form the univariate marginals π_{jk} and then levelwise averages $\bar{\pi}_j = \mathrm{ave}_k(\pi_{jk})$. Form a new prior $\bar{\pi}$ by making θ_{jk} independent, with $\theta_{jk} \sim \bar{\pi}_j$. By construction

$$\mathrm{ave}_k E_{\bar{\pi}}|\theta_I|^p = \mathrm{ave}_k E_\pi|\theta_I|^p,$$

so that $\bar{\pi} \in \mathcal{M}$. As we showed in earlier chapters, e.g. Lemma 4.14, the prior $\bar{\pi}$ is more difficult for Bayes estimation, so $B(\bar{\pi}) \geq B(\pi)$. Thus it suffices to maximise over priors $\bar{\pi} \in \mathcal{M}$.

The independence structure of $\bar{\pi}$ means that the Bayes estimator $\hat{\theta}_{\bar{\pi}}$ is separable - since prior and likelihood factorize, so does the posterior, and so

$$\hat{\theta}_{\bar{\pi},I} = E_{\bar{\pi}_I}(\theta_I | y_I).$$

In addition, the Bayes risk is additive: $B(\bar{\pi}) = \sum_I B(\bar{\pi}_I)$. The constraint for membership in $\mathcal{M}$ becomes, for $\bar{\pi}_j$,

$$E_{\bar{\pi}_I} |\theta_{j1}|^p \leq C^p 2^{-(ap+1)j} \qquad \text{for all } j.$$

Let $\omega = \alpha + 1/2$ and note that $ap + 1 = \omega p$. The optimization can now be carried out on each level separately, and, since $\bar{\pi}_j$ is a univariate prior, expressed in terms of the univariate Bayes minimax risk, so that

$$\begin{aligned} B(C, \epsilon) = \sup_{\pi \in \mathcal{M}} B(\pi) &= \sup\{\sum_{j \geq 0} 2^j B(\bar{\pi}_j) : E_{\bar{\pi}_j} |\theta_{j1}|^p \leq C^p 2^{-\omega p j}\} \\ &= \sum_{j \geq 0} 2^j \beta_p(C 2^{-\omega j}, \epsilon). \end{aligned} \tag{14.6}$$

Using the scale invariance of $\beta_p(\tau, \epsilon)$, Proposition 13.13, and introducing a parameter ζ through $2^{\omega\zeta} = C/\epsilon$, we have

$$B(C, \epsilon) = \epsilon^2 \sum_{j \geq 0} 2^j \beta_p(2^{-\omega(j-\zeta)}). \tag{14.7}$$

Hence the Bayes-minimax rule must be separable. Recalling the structure of minimax rules for $\beta_p(\eta)$, we have

$$\theta_I^*(y) = \epsilon \delta(y_I/\epsilon, \eta_j) \qquad \eta_j = (C/\epsilon) 2^{-\omega j}.$$

This completes the proof of Theorem 14.1.

14.5 Exact Bayes minimax asymptotics.

To start the proof of Theorem 14.2, we observe that, since $\beta_p(\eta) \leq 1$, we can extend the sum in (14.6) to all $j \in \mathbb{Z}$ at cost of at most ϵ^2:

$$Q(C, \epsilon) = \sum_{j \in \mathbb{Z}} 2^j \beta_p(C 2^{-\omega j}, \epsilon) = B(C, \epsilon) + O(\epsilon^2). \tag{14.8}$$

Since a discrepancy of order ϵ^2 is negligible in non-parametric problems as $\epsilon \to 0$, we may safely proceed to study $Q(C, \epsilon)$. Note that $Q(C, \epsilon)$ satisfies the invariances

$$Q(C, \epsilon) = \epsilon^2 Q(C/\epsilon, 1), \qquad Q(C 2^{\omega h}, \epsilon) = 2^h Q(C, \epsilon). \tag{14.9}$$

Starting now from (14.7) and writing $2^j = 2^{j-\zeta} \cdot 2^\zeta$, we have

$$Q(C, \epsilon) = \epsilon^2 \sum_{j \in \mathbb{Z}} 2^j \beta_p(2^{-\omega(j-\zeta)}) = \epsilon^2 2^\zeta P(\zeta),$$

where $P(\zeta)$ is the 1−periodic function

$$P(\zeta) = \sum_j 2^{j-\zeta}\beta_p(2^{-\omega(j-\zeta)}) = \sum_{v\in\mathbb{Z}-\zeta} 2^v \beta_p(2^{-\omega v}).$$

Since $2^\zeta = (C/\epsilon)^{1/\omega}$ with $1/\omega = 2/(2\alpha+1) = 2(1-r)$, we get

$$\epsilon^2 2^\zeta = C^{2(1-r)}\epsilon^{2r},$$

yielding the formula in the display in Theorem 14.2.

To check convergence of the sum defining $P(\zeta)$, observe that for large negative v, we have $F(v) = 2^v\beta_p(2^{-\omega v}) \asymp 2^v$, while for large positive v, referring to (13.7)

$$F(v) \asymp \begin{cases} 2^v \cdot 2^{-2\omega v} & \text{with } 2\omega - 1 = 2\alpha > 0 \qquad \text{if } \ p \geq 2 \\ 2^v \cdot 2^{-p\omega v} v^{1-p/2} & \text{with } p\omega - 1 = p(\alpha + 1/2) - 1 > 0 \ \text{ if } \ p < 2. \end{cases}$$

Continuity of $P(\zeta)$ follows from this convergence and the continuity of $\beta_p(\eta)$. This completes the proof of Theorem 14.2.

Remark. How does the location of the maximum j in $Q(C,\epsilon)$ depend on ϵ? Suppose that v_* is the location of the maximum of the function $v \to 2^v\beta_p(2^{-\omega v})$. Then the maximum in $Q(C,\epsilon)$ occurs at $u_* = v_* + \zeta = v_* + \omega^{-1}\log_2(C/\epsilon)$. Using the calibration $\epsilon = n^{-1/2}$ and $\omega = 1/(2\alpha+1)$, we can interpret this in terms of equivalent sample sizes as

$$u_* = \frac{\log_2 n}{1+2\alpha} + \frac{\log_2 C}{\alpha + 1/2} + v_*. \tag{14.10}$$

The "most difficult" resolution level for estimation is therefore at about $(\log_2 n)/(1+2\alpha)$. This is strictly smaller than $\log_2 n$ for $\alpha > 0$, meaning that so long as the sum (14.8) converges, the primary contributions to the risk $B(C,\epsilon)$ come from levels below the finest (with $\log_2 n$ corresponding to a sample of size n).

Example. When $p = 2$, explicit solutions are possible because $\beta_2(\eta) = \eta^2/(1+\eta^2)$ and $\hat\delta(x;\eta,2) = wx = [\eta^2/(1+\eta^2)]x$. Recall that $\eta_j = (C/\epsilon)2^{-\omega j} = 2^{-\omega(j-\zeta)}$ decreases rapidly with j above $\zeta = \omega^{-1}\log_2(C/\epsilon)$, so that $\hat\delta_j$ is essentially 0 for such j.

We have $P(\zeta) = \sum_j g(j-\zeta)$ for

$$g(v) = \frac{2^v}{1+2^{2\omega v}} = \frac{e^{av}}{1+e^{bv}}$$

for $a = \log 2$ and $b = (2\alpha+1)\log 2 > a$. An easy calculation shows that the maximum of g occurs at $v_* = \log_2(1/(2\alpha))/(1+2\alpha)$, compare also (14.10).

Figure 14.1 shows plots of the periodic function $P(\zeta)$ for several values of α. For small α, the function P is very close to constant, while for larger α it is close to a single sinusoidal cycle. This may be understood from the Poisson summation formula (C.10). Indeed, since g is smooth, its Fourier transform $\hat g(\xi)$ will decay rapidly, and so the primary contribution in the Poisson formula comes from $P_{0,\alpha} = \hat g(0) = \int_{-\infty}^{\infty} g(t)dt$. The integral may be expressed in terms of the beta function by a change of variables $w = (1+e^{bt})^{-1}$, yielding $b^{-1}B(c,1-c) = b^{-1}\Gamma(c)\Gamma(1-c)$ for $c = a/b$. From Euler's reflection formula $\Gamma(z)\Gamma(1-z) = \pi/\sin(\pi z)$, and using the normalized sinc function $\mathrm{sinc}(x) = \sin(\pi x)/(\pi x)$, we arrive at

$$P_{0,\alpha} = \big(\log 2 \cdot \mathrm{sinc}((2\alpha+1)^{-1})\big)^{-1}. \tag{14.11}$$

Figure 14.1 shows that $P_{0,\alpha}$ provides an adequate summary for $\alpha \leq 2$.

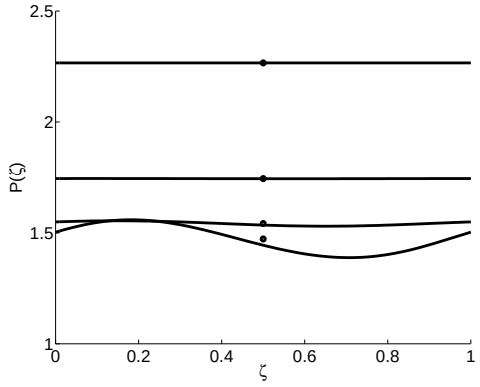

Figure 14.1 Periodic function $P(\zeta)$ appearing in Bayes-minimax risk, Theorem 14.2, for $p = 2$ and, from bottom to top, $\alpha = 4, 2, 1, 0.5$. Solid circles show the approximation by (14.11) in each case.

14.6 Asymptotic Efficiency

We again use the approach outlined in Chapter 4.10, which involves constructing near least favorable priors π_ϵ that asymptotically concentrate on Θ as $\epsilon \searrow 0$. More specifically, in line with the strategy (4.65) – (4.67), for each $\gamma < 1$, we construct $\pi_\epsilon \in \mathcal{M}_p$ such that $B(\pi_\epsilon) > \gamma B(\gamma C, \epsilon)$ and verify that $\pi_\epsilon(\Theta) \to 1$. as well as the technical step (4.67).

The idea is to use the renormalized problem $Q(1,1)$ and $Q(\gamma,1)$ to build approximately least favorable priors and then to "translate" them to the appropriate sets of resolution levels corresponding to noise level ϵ.

Thus, for each given value $\gamma < 1$, we choose $J = J(\gamma)$ and $M = M(\gamma)$ and then priors $\pi_j, j = -J, \ldots, J$ such that $\operatorname{supp} \pi_j \subset [-M, M]$ and $E_{\pi_j}|\mu|^p \leq \gamma^p 2^{-\omega j p}$ and together $\{\pi_j\}$ form a near maximizer of $Q(\gamma, 1)$:

$$\sum_{-J}^{J} 2^j B(\pi_j) \geq \gamma Q(\gamma, 1) = \gamma \sum_{-\infty}^{\infty} 2^j \beta_p(\gamma 2^{-\omega j}).$$

To construct the individual π_j, we may proceed as in case (1) of the proof of Theorem 13.16 for ℓ_p-balls. To obtain J, we rely on convergence of the sum established in the proof of Theorem 14.2.

To perform the "translation", we focus on a subsequence of noise levels ϵ_h defined by $C/\epsilon_h = 2^{\omega h}$, for $h \in \mathbb{N}$. The prior π_{ϵ_h} concentrates on the $2J + 1$ levels centered at $h = \omega^{-1} \log_2 C/\epsilon$. Let $\{\mu_{jk}, k \in \mathbb{N}\}$ be an i.i.d. sequence drawn from π_j. For $|j| \leq J$, set

$$\theta_{h+j,k} = \epsilon_h \mu_{j,k} \qquad k = 1, \ldots, 2^{h+j}. \tag{14.12}$$

Hence, as $\epsilon \to 0$, the near least favorable priors charge (a fixed number of) ever higher frequency bands.

We now verify conditions (4.65) – (4.67) for the sequence π_{ϵ_h}, noting that J and M are fixed. Working through the definitions and exploiting the invariances (14.9), we have

$$B(\pi_{\epsilon_h}) = \epsilon_h^2 \sum_{j=h-J}^{h+J} 2^j B(\pi_{j-h}) = \epsilon_h^2 2^h \sum_{j=-J}^{J} 2^j B(\pi_j)$$
$$\geq \gamma \epsilon_h^2 2^h Q(\gamma, 1) = \gamma Q(\gamma C, \epsilon_h) \geq \gamma B(\gamma C, \epsilon_h).$$

Recalling the definition of ϵ_h and that $a = \alpha + 1/2 - 1/p = \omega - 1/p$, we have with probability one under the prior π_{ϵ_h} that

$$\theta \in \Theta(C) \Leftrightarrow \sum_k |\theta_{h+j,k}|^p \leq C^p 2^{-a(h+j)p} \quad \text{for } |j| \leq J,$$
$$\Leftrightarrow n_{jh}^{-1} \sum_{k=1}^{n_{jh}} |\mu_{jk}|^p \leq 2^{-\omega j p} \quad \text{for } |j| \leq J,$$

where $n_{jh} = 2^{j+h}$.

Write $X_{jk} = |\mu_{jk}|^p - E|\mu_{jk}|^p$ and set $t_j = (1-\gamma)2^{-j\omega p}$. From the moment condition on π_j, it follows that $\{\theta \notin \Theta(C)\} \subset \cup_{j=-J}^{J} \Omega_{jh}$ where

$$\Omega_{jh} = \{n_{jh}^{-1} \sum_{k=1}^{n_{jh}} X_{jk} > t_j\}.$$

Since $P(\Omega_{jh}) \to 0$ as $h \to \infty$ by the law of large numbers, for each of a finite number of indices j, we conclude that $\pi_{\epsilon_h}(\Theta(C)) \to 1$.

Finally, to check (4.67), observe first that $\|\theta_{v_{\epsilon_h}}\|^2 \leq E_{\pi_{\epsilon_h}}[\|\theta\|^2 | \theta \in \Theta, y]$ and that for π_{ϵ_h} we have, with probability one,

$$\|\theta\|^2 = \epsilon_h^2 \sum_{j=-J}^{J} \sum_{k=1}^{2^{j+h}} |\mu_{jk}|^2 \leq M^2 2^{J+1} C^{2(1-r)} \epsilon_h^{2r}.$$

Consequently,

$$E\{\|\hat{\theta}_{v_\epsilon}\|^2 + \|\theta\|^2, \Theta^c\} \leq 2c(M, J) B(C, \epsilon_h) \pi_{\epsilon_h}(\Theta^c)$$

and the right side is $o(B(C, \epsilon_h))$ as required, again because $\pi_{\epsilon_h}(\Theta^c) \to 0$.

14.7 Linear Estimates

Using results from Chapter 4, it is relatively straightforward to show that over Besov bodies with $p < 2$, linear estimates are suboptimal, even at the level of rates of convergence.

First, we recall that the Besov bodies $\Theta = \Theta_p^\alpha(C)$ are solid and orthosymmetric, so that by Theorem 9.3 the linear minimax risk is determined by the quadratic hull of Θ. It follows from the definitions (Exercise 14.2) that

$$\text{QHull}(\Theta_p^\alpha) = \Theta_{p'}^{\alpha'} \qquad p' = p \vee 2, \;\; \alpha' = \alpha - 1/p + 1/p'. \tag{14.13}$$

In particular, Θ_p^α is quadratically convex only if p is at least 2. The Ibragimov-Hasminskii

theorem 4.16 shows that the linear minimax risk of a quadratically convex solid orthosymmetric set is between 1 and 5/4 times the non-linear minimax risk. Hence

$$R_L(\Theta_p^\alpha(C), \epsilon) \asymp R_N(\Theta_{p'}^{\alpha'}, \epsilon) \asymp C^{2(1-r')}\epsilon^{2r'} \qquad r' = 2\alpha'/(2\alpha' + 1). \tag{14.14}$$

In particular, when $p < 2$, we have $\alpha' = \alpha - (1/p - 1/2)$, so that the linear rate r' is strictly smaller than the minimax rate r. This property extends to all $q \leq \infty$ (Donoho and Johnstone, 1998). For example, on the Besov body $\Theta_{1,1}^1$ corresponding to the Bump Algebra, one finds that $\alpha' = 1/2$ and so the linear minimax rate is $O(\epsilon)$, whereas the non-linear rate is much faster, at $O(\epsilon^{4/3})$.

Let us conclude this section with some remarks about the structure of minimax linear estimators. Since the spaces $\Theta = \Theta_p^\alpha(C)$ are symmetric with respect to permutation of coordinates *within resolution levels*, it is intuitively clear that a minimax linear estimator will have the form $\hat{\theta} = (\hat{\theta}_{j,c_j})$, where for each j, $c_j \in [0, 1]$ is a scalar and

$$\hat{\theta}_{j,c_j} = c_j y_j, \tag{14.15}$$

and hence that

$$R_L(\Theta, \epsilon) = \inf_{(c_j)} \sup_{\Theta} \sum_j E\|\hat{\theta}_{j,c_j} - \theta_j\|^2. \tag{14.16}$$

A formal verification again uses the observation that $R_L(\Theta) = R_L(\bar{\Theta})$ where $\bar{\Theta} = \mathrm{QHull}(\Theta) = \Theta_{p'}^{\alpha'}$ as described earlier. Given $\tau \in \bar{\Theta}$, construct $\bar{\tau}$ by setting $\bar{\tau}_{jk}^2 \equiv \mathrm{ave}_k \tau_{jk}^2$: since $p' \geq 2$, one verifies that $\bar{\tau} \in \bar{\Theta}$ also. Formula (4.48) shows that $R(\Theta(\tau))$ is a concave function of (τ_i^2), and hence that $R(\Theta(\bar{\tau})) \geq R(\Theta(\tau))$. Consequently, the hardest rectangular subproblem lies among those hyperrectangles that are symmetric within levels j. Since the minimax linear estimator for rectangle $\bar{\tau}$ has the form $\hat{\theta}_{c(\bar{\tau}),I} = \bar{\tau}_I^2/(\bar{\tau}_I^2 + \epsilon^2)y_i$, it follows that the minimax linear estimator for $\bar{\Theta}$ has the form (14.15), which establishes (14.16).

14.8 Near Minimaxity of Threshold Estimators

Although described in terms of a two parameter family of co-ordinatewise Bayes estimators, the asymptotic minimax estimators derived at (14.4) are still not available in fully explicit form. In this section, we show that nearly minimax estimators exist within the family of soft threshold estimators.

Consider *level dependent* soft thresholding estimators, so that if $\lambda = (\lambda_j)$, we set

$$\hat{\theta}_{\lambda,jk}(y) = \hat{\delta}_S(y_{jk}, \lambda_j\epsilon),$$

where $\hat{\delta}_S(y, \lambda)$ is soft thresholding, cf (8.4). The minimax risk among such soft threshold estimators over Θ is defined by

$$R_S(\Theta, \epsilon) = \inf_{(\lambda_j)} \sup_{\Theta} E_\theta\|\hat{\theta}_\lambda - \theta\|^2.$$

Over the full range of p, and for a large range of α, thresholding is nearly minimax among all non-linear estimators.

Theorem 14.4 *For $0 < p \le \infty$ and $\alpha > (1/p - 1/2)_+$, with $\Theta = \Theta_p^\alpha(C)$, we have*

$$R_S(\Theta, \epsilon) \le \Lambda(p) R_N(\Theta, \epsilon)(1 + o(1)), \qquad \text{as } \epsilon \to 0.$$

Proof The argument is analogous to that for soft thresholding on ℓ_p-balls in $\mathbb{R}^n$, Theorem 13.17. We bound $R_S(\Theta, \epsilon)$ in terms of the Bayes minimax risk $B(C, \epsilon)$ given by (14.2) and (14.6), and then appeal to the equivalence theorem $R_N(\Theta, \epsilon) \sim B(C, \epsilon)$.

Given $\theta = (\theta_{jk})$, let $\mu_{jk} = \theta_{jk}/\epsilon$. Let π_j denote the empirical measure of $\{\mu_{jk}, k = 1, \dots, 2^j\}$, so that $\pi_j = 2^{-j} \sum_k \delta_{\mu_{jk}}$. Recalling the definitions of threshold risk $r(\lambda, \mu)$ and Bayes threshold risk $B(\lambda, \pi)$ for unit noise level from Chapter 13, we have

$$E_\theta \|\hat\theta_\lambda - \theta\|^2 = \sum_{jk} \epsilon^2 r(\lambda_j, \mu_{jk}) = \sum_j 2^j \epsilon^2 B(\lambda_j, \pi_j).$$

Let $\eta_j = (C/\epsilon) 2^{-\omega j}$; one verifies exactly as at (13.32) that $\theta \in \Theta_p^\alpha(C)$ implies $\pi_j \in \mathfrak{m}_p(\eta_j)$, so that

$$\inf_\lambda \sup_{\Theta_p^\alpha(C)} E_\theta \|\hat\theta_\lambda - \theta\|^2 \le \sum_j 2^j \epsilon^2 \beta_{S,p}(\eta_j),$$

since the minimization over thresholds λ_j can be carried out level by level. Now apply Theorem 13.14 to bound $\beta_{S,p}(\eta_j) \le \Lambda(p) \beta_p(\eta_j)$, and so the right side of the preceding display by $\Lambda(p) \sum_j 2^j \beta_p(C 2^{-\omega j}, \epsilon)$. Hence, using (14.6)

$$R_S(\Theta, \epsilon) \le \Lambda(p) B(C, \epsilon).$$

Our conclusion now follows from Theorem 14.3. □

Remark. In principle, one could allow the thresholds to depend on location k as well as scale j: $\lambda = (\lambda_{jk})$. Along the lines described in Remark 13.18 and Exercise 13.7, one can define a Bayes minimax threshold risk $B_S(\mathcal{M}, \epsilon)$, show that it is bounded by $\Lambda(p) B(\mathcal{M}, \epsilon)$, and that minimax choices of λ in fact depend only on j and not on k. Further details are in Donoho and Johnstone (1998, §5).

Since $\Lambda(p) \le 2.22$ for $p \ge 2$, and $\Lambda(1) \approx 1.6$, these results provide some assurance that threshold estimators achieve nearly optimal minimax performance. The particular choice of threshold still depends on the parameters (α, p, q, C), however. Special choices of threshold not depending on a prior specifications of these parameters will be discussed in later chapters.

Similar results may be established for hard thresholding.

14.9 Notes

General case. When $q < \infty$, the levels j do not decouple in the fashion that led to (14.8). We may obtain similar asymptotic behavior by using homogeneity properties of the $Q(C, \epsilon)$ problem with respect to scaling and level shifts – details may be found in Donoho and Johnstone (1998).

Remark. `[Retain?]` Here and in preceding chapters we have introduced various spaces of moment-constrained probability measures. These are all instances of a single method, as

is shown by the following slightly cumbersome notation. If π is a probability measure on $\ell_2(\mathcal{I})$, let $\tau_p(\pi)$ denote the sequence of marginal pth moments

$$\tau_p(\pi)_I = (E_\pi|\theta_I|^p)^{1/p}. \qquad I \in \mathcal{I}, \quad p \in (0, \infty].$$

If Θ is a parameter space contained in $\ell_2(\mathcal{I})$, then set

$$\mathcal{M}_p(\Theta) = \{\pi \in \mathcal{P}(\ell_2(\mathcal{I})) : \tau_p(\pi) \in \Theta\}.$$

In the following examples, the left side gives the notation used in the text, and the right side the notation according to the convention just introduced.

(i) Intervals $\Theta = [-\tau, \tau] \subset \mathbb{R}$:	$\mathcal{M}_p(\tau) = \mathcal{M}_p([-\tau, \tau])$.
(ii) ℓ_p balls:	$\mathcal{M}_n = \mathcal{M}_p(\Theta_{n,p}(C))$,
(iii) Ellipsoids in Pinsker's Theorem:	$\mathcal{M}(C) = \mathcal{M}_2(\Theta(C))$,
(iv) Besov bodies:	$\mathcal{M}^\alpha_{p,q}(C) = \mathcal{M}_{p\wedge q}(\Theta^\alpha_{p,q}(C))$.

`[Reference to Triebel case.]`

Exercises

14.1 *(Compactness criteria.)* (a) Show, using the total boundedness criterion REF, that $\Theta^\alpha_p(C)$ is ℓ_2-compact when $\alpha > (1/p - 1/2)_+$.
(b) Show, using the tightness criterion REF, that $\mathcal{M}^\alpha_p(C)$ is compact in the topology of weak convergence of probability measures on $\mathcal{P}(\ell_2)$ when $\alpha > (1/p - 1/2)_+$.

14.2 *(Quadratic hull of Besov bodies.)* Verify (14.13). [Hint: begin with finding the convex hull of sets of the form $\{(\sigma_{jk}) : \sum_j (\sum_k |\sigma_{jk}|^\pi)^{\mu/\pi} \le 1\}$]

14.3 *(Threshold minimax theorem.)* Formulate and prove a version of the threshold minimax theorem 13.12 in the Bayes minimax setting of this chapter.

15

Continuous v. Sampled Data

Our theory has been developed so far almost exclusively in the Gaussian sequence model (3.1). In this chapter, we indicate some implications of the theory for models that are more explicitly associated with function estimation. We first consider the *continuous white noise model*

$$Y_\epsilon(t) = \int_0^t f(s)ds + \epsilon W(t), \qquad t \in [0,1], \tag{15.1}$$

which we have seen is in fact an equivalent representation of (3.1).

Less trivial, but closer to many applications is the *sampled data model* in which one observes

$$\tilde{y}_l = f(t_l) + \sigma \tilde{z}_l, \qquad i = 1, \ldots n, \tag{15.2}$$

and it is desired to estimate the function $f \in L_2[0,1]$.

For many purposes, the models (15.1) and (15.2) are very similar, and methods and results developed in one should apply equally well in the other. A general equivalence result of Brown and Low (1996a) implies that for bounded loss function $\ell(.)$ and for collections $\mathcal{F}$ which are bounded subsets of Hölder classes $C^\alpha, \alpha > 1/2$, we have as $\epsilon \to 0$,

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} E\ell\Big(\|\hat{f}(Y) - f\|^2_{L^2[0,1]}\Big) \sim \inf_{\hat{f}} \sup_{f \in \mathcal{F}} E\ell\Big(\|\hat{f}(\tilde{y}) - f\|^2_{L^2[0,1]}\Big) \tag{15.3}$$

the expectation on the left-hand side being with respect to white noise observations Y in (15.1) and on the right hand-side being with respect to $\tilde{y}$ in (15.2). However, the general equivalence result fails for $\alpha \leq 1/2$ and we wish to establish results for the global estimation problem for the unbounded loss function $\|\hat{f} - f\|_2$ that are valid also for Besov (and Triebel) classes satisfying $\alpha > 1/p$, where p might be arbitrarily large.

In addition our development will address directly the common and valid complaint that theory is often developed for "theoretical" wavelet coefficients in model (15.1) while computer algorithms work with empirical wavelet coefficients derived from the sampled data model (15.2). We compare explicitly the sampling operators corresponding to pointwise evaluation and integration against a localized scaling function. The approach taken in this chapter is based on Donoho and Johnstone (1999) and Johnstone and Silverman (2004b).

15.1 The Sampled Data Model: A Wavelet Crime?

The simplest non-parametric regression model (15.2) posits an unknown function observed in homoscedastic Gaussian noise at equally spaced points $t_l = l/n$. We assume that the $\tilde{z}_l$

are i.i.d standard Gaussian variables and that the noise level σ is known. For convenience, suppose throughout that $n = 2^J$ for some integer J.

We have studied at length the white noise model (15.1) which after conversion to wavelet coefficients $y_I = \langle dY_\epsilon, \psi_I \rangle, \theta_I = \langle f, \psi_I \rangle, z_I = \langle dW, \psi_I \rangle$ takes the sequence model form

$$y_I = \theta_I + \epsilon z_I, \qquad I = (j,k),\ j \geq 0, k = 1, \ldots, 2^j. \tag{15.4}$$

This leads to a possibly troubling dichotomy. Much of the theory developed to study wavelet methods is carried out using functions of a continous variable, uses the multiresolution analysis and smoothness classes of functions on $\mathbb{R}$ or $[0,1]$, and the sequence model (15.4). Almost inevitably, most actual data processing is carried out on discrete, sampled data, which in simple cases might be modeled by (15.2).

There is therefore a need to make a connection between the continuous and sampled models, and to show that, under appropriate conditions, that conclusions in one model are valid for the other and vice versa. For this purpose, we compare minimax risks for estimation of f based on sequence data y from (15.4) with that based on sampled data $\tilde{y}$ from (15.2). Hence, set

$$\begin{aligned} R(\mathcal{F}, \epsilon) &= \inf_{\hat{f}(y)} \sup_{f \in \mathcal{F}} E \| \hat{f}(y) - f \|_2^2, \\ \tilde{R}(\mathcal{F}, n) &= \inf_{\hat{f}(\tilde{y})} \sup_{f \in \mathcal{F}} E \| \hat{f}(\tilde{y}) - f \|_2^2. \end{aligned} \tag{15.5}$$

Note that the error of estimation is measured in both cases in the norm of $L_2[0,1]$. The parameter space $\mathcal{F}$ is defined through the wavelet coefficients corresponding to f, as at (9.41):

$$\mathcal{F} = \{ f : \theta[f] \in \Theta^\alpha_{p,q}(C) \}.$$

Remark. One might also be interested in the error measured in the discrete norm

$$n^{-1} \| \hat{f} - f \|_n^2 = (1/n) \sum [\hat{f}(t_l) - f(t_l)]^2. \tag{15.6}$$

Section 15.5 shows that this norm is equivalent to $\int_0^1 (\hat{f} - f)^2$ under our present assumptions.

Assumption (A) on wavelet. In this chapter the choice of α, p and q is fixed at the outset, so that we focus on a fixed Besov space $B^\alpha_{p,q}[0,1]$. Given this selection, we choose a Daubechies pair (ϕ, ψ) and an orthonormal wavelet basis (ψ_I) for $L_2[0,1]$ consisting of wavelets of compact support, with elements having R continuous derivatives ($\psi_I \in C^R$) and $(D+1)$ vanishing moments. The basis is chosen so that $\min(R, D) \geq \alpha$, so that it is an unconditional basis of $B^\alpha_{p,q}[0,1]$, and the norm is equivalently given by the Besov sequence norm on wavelet coefficients. We also assume that the CDJV construction (cf. Section 7.1) is used for wavelets that intersect the boundary of $[0,1]$.

Theorem 15.1 *Let $\alpha > 1/p$ and $1 \leq p, q \leq \infty$; or else $\alpha = p = q = 1$. Then, with $\epsilon_n = \sigma/\sqrt{n}$ we have*

$$\tilde{R}(\mathcal{F}, n) \geq R(\mathcal{F}, \epsilon_n)(1 + o(1)), \quad n \to \infty. \tag{15.7}$$

In words, there is no estimator giving a worst-case performance in the sampled-data-problem (15.2) which is substantially better than what we can get for the worst-case performance of procedures in the white-noise-problem (15.4).

For *upper bounds*, we will specialize to estimators derived by applying certain coordinatewise mappings to the noisy wavelet coefficients.

For the white noise model, this means the estimate is of the form

$$\hat{f} = \sum_I \delta(y_I)\psi_I$$

where each function $\delta_I(y)$ either belongs to one of three specific families – *Linear, Soft Thresholding, or Hard Thresholding* – or else is a general scalar function of a scalar argument. The families are:

- $(\mathcal{E}_L)$ diagonal linear procedures in the wavelet domain, $\delta_I^L(y) = c_I \cdot y$,
- $(\mathcal{E}_S)$ soft thresholding of wavelet coefficients, $\delta_I^S(y) = (|y| - \lambda_I)_+ \mathrm{sgn}(y)$,
- $(\mathcal{E}_H)$ hard thresholding of wavelet coefficients, $\delta_I^H(y) = y1_{\{|y|\geq\lambda_I\}}$, and
- $(\mathcal{E}_N)$ scalar nonlinearities of wavelet coefficients, with arbitrary $\delta_I^N(y)$.

For the sampled-data problem, this means that the estimate is of the form

$$\hat{f} = \sum_I \delta_I(y_I^{(n)})\psi_I, \tag{15.8}$$

where $y_I^{(n)}$ is an empirical wavelet coefficient based on the sampled data $(\tilde{y}_i)$, see Section 15.4 below, and the δ_I belong to one of the families $\mathcal{E}$. Then define the $\mathcal{E}$-minimax risks in the two problems:

$$R_{\mathcal{E}}(\mathcal{F}, \epsilon) = \inf_{\hat{f}\in\mathcal{E}} \sup_{f\in\mathcal{F}} E_{Y_\epsilon} \| \hat{f} - f \|^2_{L^2[0,1]} \tag{15.9}$$

and

$$\tilde{R}_{\mathcal{E}}(\mathcal{F}, n) = \inf_{\hat{f}\in\mathcal{E}} \sup_{f\in\mathcal{F}} E_{\mathbf{y}_n} \| \hat{f} - f \|^2_{L^2[0,1]}. \tag{15.10}$$

With this notation established, we have

Theorem 15.2 *Let $\alpha > 1/p$ and $1 \leq p, q \leq \infty$ or $\alpha = p = q = 1$. Adopt assumption (A) on the wavelet basis. For each of the four classes $\mathcal{E}$ of coordinatewise estimators,*

$$\tilde{R}_{\mathcal{E}}(\mathcal{F}, n) \leq R_{\mathcal{E}}(\mathcal{F}, \epsilon_n)(1 + o(1)), \quad n \to \infty. \tag{15.11}$$

Our approach is to make an explicit construction transforming a sampled-data problem into a quasi-white-noise problem in which estimates from the white noise model can be employed. We then show that these estimates on the quasi-white-noise-model data behave nearly as well as on the truly-white-noise-model data. The observations in the quasi-white-noise problem have constant variance, but may be correlated. The restriction to coordinatewise estimators means that the correlation structure plays no role.

Furthermore, we saw in the last chapter in Theorems 14.1– 14.3 that co-ordinatewise non-linear rules were asymptotically minimax: $R(\mathcal{F}, \epsilon_n) \sim R_{\mathcal{E}_N}(\mathcal{F}\epsilon_n)$ for the $q = \infty$ cases considered there, and the same conclusion holds more generally for $p \leq q$ (Donoho and Johnstone, 1998).

Remark. The assumptions on (α, p, q) in Theorems 15.1 and 15.2 are needed for the bounds to be described in Section 15.4. Informally, they correspond to a requirement that point evaluation $f \to f(t_0)$ is well defined and continuous, as is needed for model (15.2) to be stably defined. For example, if $\alpha > 1/p$, then functions in $B^{\alpha}_{p,q}$ are uniformly continuous (by the embedding result Proposition 9.9), while if $\alpha = p = q = 1$, one can use the embedding $B^1_{1,1} \subset TV$ to make sense of point evaluation, by agreeing to use, say, the left continuous version of $f \in TV$. For further discussion, see (Donoho, 1992, Section 6.1).

15.2 The Projected White Noise Model

Finite dimensional submodels of (15.1) are of interest for a number of reasons. Firstly, when the noise level ϵ is of order $n^{-1/2}$, a model with n observed coefficients is a closer relative of the regression model (15.2). Secondly, for a given parameter space Θ, finite dimensional submodels can be found with dimension $m(\epsilon)$ depending on ϵ that are asymptotically as difficult as the full model. This proves to be a useful technical tool, for example in proving results for the sampling model.

Let ϕ be the scaling function corresponding to the orthonormal wavelet ψ used in the previous section. We consider only projections on to the increasing sequence of multiresolution spaces $V_j = \text{span}\,\{\phi_{ji}, i = 1, \dots 2^j\}$. Given ϵ, fix a level $J = J(\epsilon)$, set $m = m_\epsilon = 2^{J(\epsilon)}$ and define

$$y_i = \langle \phi_{Ji}, dY \rangle, \qquad z_i = \langle \phi_{Ji}, dW \rangle, \qquad i = 1, \dots m.$$

The *projected white noise model* refers to observations

$$y_i = \langle f, \phi_{Ji} \rangle + \epsilon z_i, \qquad i = 1, \dots m. \tag{15.12}$$

Write $y^{[m]}$ for the projected data $y_1, \dots, y_m$. When $\epsilon = n^{-1/2}$, the choice $J = \log_2 n$ yields an n-dimensional model which is an approximation to (15.2), in a sense to be explored below.

The projected white noise model can be expressed in terms of wavelet coefficients. Indeed, since $V_J = \oplus_{j<J} W_j$, it is equivalent to the 2^J-dimensional submodel of the sequence model given by

$$y_I = \theta_I + \epsilon z_I, \qquad I \in \mathcal{I}^J, \tag{15.13}$$

where we define $\mathcal{I}^J = \cup_{j<J} \mathcal{I}_j$.

Estimation of the unknown coefficients $\langle f, \phi_{Ji} \rangle$ is done in the wavelet basis. Recall that ϕ_{Ji} is an orthobasis for V_J and that $\{\psi_I, I \in \mathcal{I}^J\}$ is an orthobasis for the wavelet spaces $\{W_j, j < J\}$. The orthogonal change of basis transformation W on $\mathbb{R}^{2^J}$ that maps $\langle f, \phi_{Ji} \rangle$ to $\langle f, \psi_I \rangle = \theta_I$ is called the *discrete wavelet transform* W. Its matrix elements W_{Ii} are just the inner products $\langle \psi_I, \phi_{Ji} \rangle$.

The estimation procedure could then be summarized by the diagram

$$\begin{array}{ccc} (y_l) & \xrightarrow{W} & (y_I) \\ \downarrow & & \downarrow \\ (\hat{f}_{n,l}) & \xleftarrow{W^T} & (\hat{\delta}_I(y_I)) \end{array} \tag{15.14}$$

which is the same as (7.21), except that this diagram refers to observations on inner products, (15.12), whereas the earlier diagram used observations from the sampling model (7.20), here written in the form (15.2).

Consider now the minimax risk of estimation of $f \in \mathcal{F}$ using data from the projected model (15.12). Because of the Parseval relation (1.23), we may work in the sequence model and wavelet coefficient domain.

Suppose, as would be natural in the projected model, that $\hat{\theta}$ is an estimator which has non-zero co-ordinates only in $\mathcal{I}^J$. Set $\|\theta\|_{2,m}^2 = \sum_{I \in \mathcal{I}^J} \theta_I^2$ and $\|\theta\|_{2,m^\perp}^2 = \sum_{I \notin \mathcal{I}^J} \theta_I^2$ The following decomposition emphasises the "tail bias" term that results from estimating only up to level J :

$$\|\hat{\theta} - \theta\|^2 = \|\hat{\theta} - \theta\|_{2,m}^2 + \|\theta\|_{2,m^\perp}^2. \tag{15.15}$$

Of course, in terms of the equivalent $f = f[\theta]$, and with P_m denoting the orthogonal projection of $L_2[0,1]$ onto V_J, the tail bias $\|\theta\|_{2,m^\perp}^2 = \|f - P_m f\|^2$.

We write $y^{[m]}$ when needed to distinguish data in the projected model from data y in the full sequence model. In the projected model, we consider estimation with loss function

$$L(\hat{\theta}, \theta) = \|\hat{\theta}(y^{[m]}) - \theta\|_{2,m}^2. \tag{15.16}$$

and the projected parameter space

$$\Theta^{[m]}(C) = \{\theta \in \mathbb{R}^m : \|\theta\|_{b_{p,q}^\alpha} \le C\}$$

The minimax risk in this reduced problem is

$$R_N(\Theta^{[m]}(C); \epsilon) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta^{[m]}(C)} \mathbb{E}\|\hat{\theta}(y^{[m]}) - \theta\|_{2,m}^2.$$

We look for a condition on the dimension $m = 2^J$ so that the minimax risk in the projected model is asymptotically equivalent to (i.e. not easier than) the full model. For this it is helpful to recall, in current notation, a bound on the maximum tail bias over smoothness classes $\Theta_{p,q}^\alpha$ that was established at (9.50).

Lemma 15.3 *Let $\alpha' = \alpha - (1/p - 1/2)_+ > 0$. Then for a constant $K = K(\alpha')$,*

$$\Delta_m(\Theta) = \sup_{\mathcal{F}_{p,q}^\alpha(C)} \|f - P_{2^J} f\|^2 = \sup_{\Theta_{p,q}^\alpha(C)} \|\theta\|_{2,m^\perp}^2 \le K C^2 2^{-2J\alpha'}.$$

Suppose $J = J(\epsilon) = \gamma \log_2 \epsilon^{-2}$; one verifies that if $\gamma > (1/(2\alpha+1))(\alpha/\alpha')$, then the tail bias term becomes neglibible relative to the order $(\epsilon^2)^{2\alpha/(2\alpha+1)}$ of the minimax risk.

It will be helpful to use the minimax Theorem 4.11 to reexpress

$$R_N(\Theta^{[m]}(C); \epsilon) = \sup_{\pi \subset \Theta^{[m]}(C)} B(\pi; \epsilon) \doteq B(C, \epsilon), \tag{15.17}$$

where, as usual, $B(\pi, \epsilon)$ denotes the Bayes risk in (15.13) and (15.16) when the prior $\pi(d\theta)$ on $\Theta^{[m]}(C)$ is used.

We can now establish the equivalence of the projected white noise model with the full model.

Proposition 15.4 *Let $J(\epsilon) = \gamma \log_2 \epsilon^{-2}$. If $\gamma > (1/(2\alpha+1))(\alpha/\alpha')$, then*

$$R_N(\Theta^{[m_\epsilon]}(C), \epsilon) \sim R_N(\Theta(C), \epsilon) \qquad \epsilon \to 0.$$

Proof An arbitrary estimator $\hat{\theta}(y^{[m]})$ in the projected model can be extended to an estimator in the full sequence model by appending zeros–let $\mathcal{E}^{[m]}$ denote the class so obtained. From (15.15) we obtain

$$\inf_{\hat{\theta}\in\mathcal{E}^{[m]}} \sup_{\Theta(C)} E\|\hat{\theta}-\theta\|^2 \le R_N(\Theta^{[m]}(C),\epsilon) + \Delta_m(\Theta).$$

The left side exceeds $R_N(\Theta(C),\epsilon)$ while Lemma 15.3 shows that $\Delta_m(\Theta) = o(R_N(\Theta(C),\epsilon))$, so we conclude that the projected model is asymptotically no easier.

In the reverse direction, suppose that $\hat{\theta}_\pi$ is the Bayes estimator for the least favorable prior $\pi = \pi_\epsilon$ for $R_N(\Theta,\epsilon)$. Define $\pi_m = \pi_{m,\epsilon}$ as the marginal distribution of the first m coordinates of θ under π_ϵ : clearly the corresponding Bayes estimate $\hat{\theta}_{\pi_m} = E_{\pi_m}(\theta|y) = E(P_m\theta|y)$ depends only on $y^{(m)}$ and so is feasible in the projected problem. Since both π and π_m are supported on Θ, $\hat{\theta}_\pi - \hat{\theta}_{\pi_m} = E_\pi(\theta - P_m\theta|y)$ and so by Jensen's inequality and Lemma 15.3,

$$\|\hat{\theta}_\pi - \hat{\theta}_{\pi_m}\|_2 \le E(\|\theta - P_m\theta\|_2|y) \le KC2^{-J\alpha'} = o(\epsilon^r)$$

by the choice of $J(\epsilon)$ specified in the hypotheses. Using $E(X+Y)^2 \le [(EX^2)^{1/2} + (EY^2)^{1/2}]^2$, we have then, for all $\theta\in\Theta$,

$$E\|\hat{\theta}_{\pi_m} - \theta\|^2 \le (\{E_\theta\|\hat{\theta}_\pi - \theta\|^2\}^{1/2} + o(\epsilon^r))^2$$

and hence

$$R_N(\Theta^{[m]}(C),\epsilon) \le \sup_\theta E\|\hat{\theta}_{\pi_m} - \theta\|^2 \le R_N(\Theta,\epsilon)(1+o(1))$$

since $\hat{\theta}_\pi$ is minimax for Θ. □

Remark. The ratio $(1/(2\alpha+1))(\alpha/\alpha')$ is certainly less than 1 whenever (i) $p \ge 2$ and $\alpha > 0$, or (ii) $p < 2$ and $\alpha \ge 1/p$.

15.3 Sampling is not easier

It is perhaps intuitively clear that sampled data does not provide as much information as the continuous white noise model, but a formal argument is still necessary. Thus, in this section, we outline a proof of a lower bound to minimax risk in the sampling problem. The idea is to show that a prior distribution that is difficult in the continuous model sequence problem induces a difficult prior distribution in the sampled data setting.

Proposition 15.4 shows that the continous problem, in sequence space form, can be projected to a level $J_{0n} = \gamma \log_2 \epsilon_n^{-2}$ given by (15.25) without loss of difficulty. Let us formulate the sampling problem in a corresponding manner.

In the "sampling problem", we observe data in model (15.2) and seek to estimate f, in principle using loss function $\|\hat{f}(\tilde{y}) - f\|_2^2$. However, we only make our task easier by restricting attention to estimating $P_m f$, the projection of f onto V_{J_0}, and hence to estimation of $\theta = (\theta_I, I \in \mathcal{I}^{J_0})$. The loss function is then

$$L(\hat{\theta}(\tilde{y}),\theta) = \|\hat{\theta}(\tilde{y}) - \theta\|_{2,m}^2. \tag{15.18}$$

Since $f = \sum_{|I|<J_0} \theta_I \psi_I$, we may rewrite (15.2) as

$$\tilde{y}_l = (T\theta)_l + \sigma \tilde{z}_l, \tag{15.19}$$

where T is given in co-ordinates by

$$(T\theta)_l = \sum_I \theta_I \psi_I(t_l). \tag{15.20}$$

We regard this as a map from $(\mathbb{R}^m, \|\cdot\|_{2,m})$ to $(\mathbb{R}^n, \|\cdot\|_n)$, where $\|\cdot\|_n$ is the time domain norm (15.6). It is not a (partial) isometry since the vectors $(\psi_I(t_l) : l = 1, \ldots, n)$ are not orthogonal in the discrete norm. However, it comes close; at the end of the section we establish

Lemma 15.5 *Under assumption (A) on the wavelet system (ψ_I), if T is defined by* (15.20) *for $m = 2^{J_0} < n$, then*

$$\lambda_{\max}(T^t T) \le 1 + c J_0 2^{J_0}/n.$$

The minimax risk in the sampling problem is, setting $\epsilon_n = \sigma/\sqrt{n}$,

$$\begin{aligned} \tilde{R}_N(\Theta^{[m]}(C); \epsilon_n) &= \inf_{\hat{\theta}(\tilde{y})} \sup_{\theta \in \Theta^{[m]}(C)} \mathbb{E}\|\hat{\theta}(\tilde{y}) - \theta\|_{2,m}^2 \\ &= \sup_{\pi \subset \Theta^{[m]}(C)} \tilde{B}(\pi; \epsilon_n) \doteq \tilde{B}(C, \epsilon_n), \end{aligned} \tag{15.21}$$

where we have again used the minimax theorem and now $\tilde{B}(\pi, \epsilon_n)$ denotes the Bayes risk in (15.19) and (15.18) when the prior $\pi(d\theta)$ on $\Theta^{[m]}(C)$ is used.

As remarked earlier, estimation of $P_m f$ is easier than estimation of f, and so from (15.5) and (15.15) we have

$$\tilde{R}(\mathcal{F}, n) \ge \tilde{R}_N(\Theta^{[m]}(C); \epsilon_n)$$

With all this notational preparation, we have recast the "sampling is not easier" theorem as the statement

$$\tilde{B}(C, \epsilon_n) \ge B(C, \epsilon_n)(1 + o(1)). \tag{15.22}$$

Pushing the sequence model observations (at noise level ϵ_n) through T generates some heteroscedasticity which may be bounded using Lemma 15.5. To see this, we introduce e_l, a vector of zeros except for $\sqrt{n}$ in the ith slot, so that $\|e_l\|_n = 1$ and $(Ty)_l = \sqrt{n}\langle e_l, Ty\rangle_n$. Then

$$\mathrm{Var}(Ty)_l = n\epsilon_n^2 E\langle e_l, Tz\rangle_n^2 = \sigma^2 \|T^t e_l\|_{2,m}^2 \le \sigma^2 \lambda_n^2$$

where $\lambda_n^2 = \lambda_{\max}(TT^t) = \lambda_{\max}(T^t T)$ is bounded in the Lemma. Now let $\tilde{w}$ be a zero mean Gaussian vector, independent of y, with covariance chosen so that $\mathrm{Var}(Ty + \tilde{w}) = \lambda_n^2 \sigma^2 I_n$. By construction, then, $Ty + \tilde{w} \stackrel{\mathcal{D}}{=} \tilde{y} = T\theta + \lambda_n \sigma \tilde{z}$.

To implement the basic idea of the proof, let π be a least favorable prior in the sequence problem (15.17) so that $B(\pi, \epsilon_n) = B(C, \epsilon_n)$. Let $\tilde{\theta}_{\pi, \lambda_n \sigma}(\tilde{y})$ denote the Bayes estimator of θ in the sampling model (15.19) and (15.18) *with noise level* $\lambda_n \sigma$.

We construct a *randomized* estimator in the sequence model using the auxiliary variable $\tilde{w}$:

$$\hat{\theta}(y, \tilde{w}) = \tilde{\theta}_{\pi, \lambda_n \sigma}(Ty + \tilde{w}) \stackrel{\mathcal{D}}{=} \tilde{\theta}_{\pi, \lambda_n \sigma}(\tilde{y})$$

where the equality in distribution holds for the laws of $Ty + \tilde{w}$ and $\tilde{y}$ given θ. Consequently

$$B(\hat{\theta}, \pi; \epsilon_n) = \mathbb{E}_\pi \mathbb{E}_{\theta;\epsilon_n} \|\hat{\theta}(y, \tilde{w}) - \theta\|_{2,m}^2 = \mathbb{E}_\pi \mathbb{E}_{T\theta;\lambda_n\sigma} \|\tilde{\theta}_{\pi,\lambda_n\sigma}(\tilde{y}) - \theta\|_{2,m}^2 = \tilde{B}(\pi; \lambda_n\epsilon_n).$$

Use of randomized rules (with a convex loss function) does not change the Bayes risk $B(\pi)$–see e.g. (A.12) in Appendix A–and so

$$B(C, \epsilon_n) = B(\pi; \epsilon_n) \le B(\hat{\theta}, \pi; \epsilon_n) = \tilde{B}(\pi; \lambda_n\epsilon_n) \le \tilde{B}(C; \lambda_n\epsilon_n),$$

where the last inequality uses (15.21). Appealing to the scaling bounds for Bayes-minimax risks (e.g. Exercises 4.1 and 4.6) we conclude that

$$\tilde{B}(C; \lambda\epsilon) \le \begin{cases} \lambda^2 \tilde{B}(C/\lambda; \epsilon) \le \lambda^2 \tilde{B}(C; \epsilon) & \text{if } \lambda > 1 \\ \tilde{B}(C; \epsilon) & \text{if } \lambda \le 1. \end{cases}$$

In summary, using again Lemma 15.5,

$$B(C, \epsilon_n) \le (\lambda_n^2 \vee 1)\tilde{B}(C, \epsilon_n) \le \tilde{B}(C, \epsilon_n)(1 + o(1)).$$

This completes the proof of (15.22), and hence of Theorem 15.1.

Proof of Lemma 15.5 The matrix representation $(a_{II'})$ of $A = T^t T$ in the basis $(\psi_I, I \in \mathcal{I}^{J_0})$ is given by

$$a_{II'} = \langle \psi_I, \psi_{I'} \rangle_n = n^{-1} \sum_l \psi_I(t_l)\psi_{I'}(t_l).$$

Exercise 15.1 bounds on the distance of these inner products from exact orthogonality:

$$|\langle \psi_I, \psi_{I'} \rangle_n - \delta_{II'}| \le cn^{-1} 2^{j+j')/2} \chi(I, I'), \tag{15.23}$$

where $\chi(I, I') = 1$ if $\operatorname{supp} \psi$ intersects $\operatorname{supp} \psi'$ and $= 0$ otherwise.

We aim to apply Schur's lemma, Corollary C.19, to A with weights $x_I = 2^{-j/2}$, hence we consider

$$S_I = \sum_{I'} |a_{II'}| 2^{-j'/2} \le 2^{-j/2} + cn^{-1} \sum_{j'} 2^{(j+j')/2} \cdot 2^{(j'-j)_+} \cdot 2^{-j'/2}$$

where we used (15.23) and bounded $\sum_{k'} \chi(I, I')$, the number of $\psi_{j'k'}$ whose supports hits that of ψ_I, by $c2^{(j'-j)_+}$. The sum is over $j' \le J_0$ and hence

$$S_I \le 2^{-j/2}(1 + cn^{-1} J_0 2^{J_0})$$

and the result follows from Schur's lemma. □

15.4 Sampling is not harder

In this section, our goal is to show that, at least when using scaling functions and wavelets with adequate smoothness and vanishing moments, the standard algorithmic practice of using the cascade algorithm on discrete data does not significantly inflate minimax risk relative to its use on genuine wavelet coefficients.

To do this, we exploit a projected model sequence indexed by dyadic powers of n, using

less than $\log_2 n$ levels, but of full asymptotic difficulty. Indeed, Proposition 15.4 shows that given $\Theta^{\alpha}_{p,q}$, full asymptotic difficulty can be achieved by choosing $\eta > 0$ such that

$$\gamma = \frac{1}{2\alpha + 1}\frac{\alpha}{\alpha'} + \eta < 1, \tag{15.24}$$

and then setting

$$m_n = 2^{J_{0n}} \qquad J_{0n} = \gamma \log_2 n = \gamma J_n \tag{15.25}$$

Specifically, we prove

Theorem 15.6 *Suppose that $\alpha > 1/p, 1 \le p, q \le \infty$ and that (ϕ, ψ) satisfy Assumption A. Let $\mathcal{E}$ be any one of the four coordinatewise estimator classes of Section 15.1, and let m_n be chosen according to* (15.24) *and* (15.25)*. Then as $n \to \infty$,*

$$\tilde{R}_{\mathcal{E}}(\Theta^{[m]}(C), \epsilon_n) \le R_{\mathcal{E}}(\Theta^{[m]}(C), \epsilon_n)(1 + o(1)).$$

We outline the argument, referring to the literature for full details. A couple of approaches have been used; in each the strategy is to begin with the sampled data model (15.2) and construct from $(\tilde{y}_l)$ a related set of wavelet coefficients $(\tilde{y}_I)$ which satisfy a (possibly correlated) sequence model

$$\tilde{y}_I = \tilde{\theta}_I + \epsilon^{(n)}\tilde{z}_I. \tag{15.26}$$

We then take an estimator $\hat{\theta}(y)$ known to be good in the (projected) white noise model and apply it with the sample data wavelet coefficients $\tilde{y} = (\tilde{y}_I)$ in place of y. The aim then is to show that the performance of $\hat{\theta}(\tilde{y})$ for appropriate Θ and noise level $\epsilon^{(n)}$ is nearly as good as that for $\hat{\theta}(y)$ at original noise level ϵ_n.

(i) Deslauriers-Dubuc interpolation. Define a fundamental function $\tilde{\phi}$ satisfying the interpolation property $\tilde{\phi}(l) = \delta_{l,0}$ and other conditions, and then corresponding scaling functions $\tilde{\phi}_l(t) = \tilde{\phi}(nt - l), l = 1, \ldots, n$. Interpolate the sampled function and data values by

$$\tilde{P}_n f(t) = \sum_{l=1}^{n} f(l/n)\tilde{\phi}_l(t), \qquad \tilde{y}^{(n)}(t) = \sum_{l=1}^{n} \tilde{y}_l\tilde{\phi}_l(t). \tag{15.27}$$

Let $\{\psi_I\}$ be an orthonormal wavelet basis as specified in Assumption A and $\tilde{\theta}_I = \langle \tilde{P}_n f, \psi_I \rangle$ Let $\epsilon^{(n)}$ be the largest standard deviation among the variates $\langle \tilde{y}^{(n)}, \psi_I \rangle$ for $j \le J_0$: it can be shown, in a manner similar to Lemma 15.5 that $\epsilon^{(n)} \sim \epsilon_n$ for n large. Now let $\tilde{y}_I = \langle \tilde{y}^{(n)}, \psi_I \rangle + n_I$, where the n_I are noise inflating Gaussian variates independent of $\tilde{y}^{(n)}$ chosen so that $\mathrm{Var}(\tilde{y}_I) \equiv [\epsilon^{(n)}]^2$. We thus obtain (15.26) though here the variates $\tilde{z}_I$ are in general correlated. This approach is set out in Donoho and Johnstone (1999). Although somewhat more complicated in the processing of the observed data $\tilde{y}_l$ it has the advantage of working for general families of wavelets and scaling functions.

(ii) Coiflets. If the wavelet basis $\{\psi_I\}$ is chosen from a family with sufficient vanishing moments for the scaling function ϕ, then we may work directly with $\tilde{y}_I$ (and $\tilde{\theta}_I$) derived from the discrete wavelet transform of the observations $\tilde{y}_l$ (and $\tilde{\theta}_l$). This approach is set out in Johnstone and Silverman (2004b). While somewhat simpler in the handling of the sampled data $\tilde{y}_l$, it is restricted to scaling functions with sufficient vanishing moments. It has the advantage that, in decomposition (15.26), the interior noise variates $\tilde{z}_I$ are an orthogonal

transformation of the the original noise $\tilde{z}_I$ and hence are independent with $\epsilon^{(n)} = \epsilon_n$. The boundary noise variates $\tilde{z}_I$ may be correlated, but there are at most cJ_0 of these, with uniformly bounded variances $\mathrm{Var}\,\tilde{z}_I \le c\epsilon_n^2$. So in the coiflet case, we could actually take $\mathcal{E}$ to be the class of *all* estimators (scalar or not).

We will restrict attention to estimators vanishing for levels $j \ge J_{0n}$, where $2^{J_{0n}} = m = m_n$ is specified in (15.25). In view of the unbiasedness of $\tilde{y}$ for $\tilde{\theta}$ in (15.2), it is natural to decompose the error of estimation of θ in terms of $\tilde{\theta}$:

$$\|\hat{\theta}(\tilde{y}) - \theta\|_{2,m} \le \|\hat{\theta}(\tilde{y}) - \tilde{\theta}\|_{2,m} + \|\tilde{\theta} - \theta\|_{2,m}. \tag{15.28}$$

Concerning the second term on the right side, in either the Deslauriers-Dubuc or coiflet settings, one verifies that

$$\sup_{\Theta(C)} \|\tilde{\theta} - \theta\|_{2,m}^2 \le cC^2 2^{-2J\alpha'}, \tag{15.29}$$

where $m = 2^J$ and $\alpha' = \alpha - (1/p - 1/2)+$. [For Deslauriers-Dubuc, this is Lemma 4.1 in Donoho and Johnstone (1999), while for coiflets it follows from Proposition 5 as in the proof of Theorem 2 in Johnstone and Silverman (2004b)].

Turning to the first term on the right side of (15.28), the key remaining issue is to establish that if θ has bounded Besov norm, then the Besov norm of the interpolant coefficients $\tilde{\theta}$ *below level* J_{0n} is not much larger. To emphasise this, we write $P_m\tilde{\theta}$ for the vector whose $(j,k)-$th coefficient is $\tilde{\theta}_{jk}$ if $j < J_{0n}$ and 0 otherwise. The two references just cited show the existence of constants $\Delta_n = \Delta_n(\phi, \psi, \alpha, p, q)$ such that

$$\|P_m\tilde{\theta}\|_{b^\alpha_{p,q}} \le (1 + \Delta_n)\|\theta\|_{b^\alpha_{p,q}}. \tag{15.30}$$

Hence, if we set $C_n = (1 + \Delta_n)C$, then $\theta \in \Theta(C)$ implies that $P_m\tilde{\theta} \in \Theta(C_n)$. Suppose now that $\hat{\theta}^*$ is asymptotically $\mathcal{E}-$ minimax over $\Theta(C_n)$ – note that we have chosen J_{0n} expressly so that this can be achieved with an estimator that vanishes for $j \ge J_{0n}$. Thus, since we only attempt to estimate the first m components of $\tilde{\theta}$,

$$\begin{aligned}\sup_{\theta\in\Theta(C)} E\|\hat{\theta}^*(\tilde{y}) - \tilde{\theta}\|_{2,m}^2 &\le \sup_{\tilde{\theta}\in\Theta(C_n)} E\|\hat{\theta}^*(\tilde{y}) - \tilde{\theta}\|_{2,m}^2 \\ &\le R_{\mathcal{E}}(C_n, \epsilon^{(n)})(1 + o(1)).\end{aligned}$$

Lemma 15.7 *If $\epsilon_1 \ge \epsilon_0$ and $C_1 \ge C_0$, then for any of the four estimator classes $\mathcal{E}$*

$$R_{\mathcal{E}}(C_1, \epsilon_1) \le (\epsilon_1/\epsilon_0)^2(C_1/C_0)^2 R_{\mathcal{E}}(C_0, \epsilon_0). \tag{15.31}$$

For the proof, see Donoho and Johnstone (1999). Combining (15.28), (15.29) and (15.31), we obtain

$$\begin{aligned}\sup_{\theta\in\Theta(C)} E\|\hat{\theta}^*(\tilde{y}) - \theta\|^2 &\le (\epsilon^{(n)}/\epsilon_n)^2(C_n/C)^2 R_{\mathcal{E}}(C, \epsilon_n)(1 + o(1)) \\ &= R_{\mathcal{E}}(C, \epsilon_n)(1 + o(1)),\end{aligned}$$

which establishes Theorem 15.2.

Remark. One can rephrase the bound (15.29) in a form useful in the next section. Indeed,

let $\tilde{P}_n f$ be given in the Deslauriers-Dubuc case by (15.27) and in the Coiflet case by $\tilde{P}_n f = n^{-1/2} \sum f(t_l)\phi_{Jl}$. Then the arguments referred to following (15.29) also show that

$$\sup_{\mathcal{F}(C)} \| \tilde{P}_n f - f \|^2 \le c C^2 n^{-2\alpha'} = o(n^{-r}). \tag{15.32}$$

15.5 Estimation in discrete norms

We will now show that the condition (15.32) in fact implies that the quality of estimation in continuous and discrete norms is in fact equivalent:

$$\tilde{R}(\mathcal{F}, n; L_2) \sim \tilde{R}(\mathcal{F}, n; \ell_{2,n}) = \inf_{\hat{f}(\tilde{y})} \sup_{f \in \mathcal{F}} n^{-1} \sum_l E[\hat{f}(t_l) - f(t_l)]^2. \tag{15.33}$$

(and similarly for R.) We describe this in the Coiflet case, but a similar result would be possible in the Deslauriers-Dubuc setting.

Given a continuous function $f \in L_2[0,1]$, we may consider two notions of sampling operator:

$$(S_\phi f)_l = \sqrt{n}\langle f, \phi_{Jl}\rangle, \qquad (S_\delta f)_l = f(t_l).$$

Let P_n denote projection onto $V_J = \text{span}\,\{\phi_{Jl}\}$ and $\tilde{P}_n$ the "interpolation" operator, so that

$$P_n f = \sum_l \langle f, \phi_{Jl}\rangle \phi_{Jl}, \qquad \text{and} \qquad \tilde{P}_n g = \sum_l n^{-1/2} g(t_l)\phi_{Jl}.$$

From this we obtain Parseval identities like

$$\langle P_n f, \tilde{P}_n g\rangle_2 = n^{-1}\langle S_\phi f, S_\delta g\rangle_n$$

and

$$\| S_\phi \hat{f} - S_\delta f \|_n = \| P_n \hat{f} - \tilde{P}_n f \|_2. \tag{15.34}$$

First suppose that $\tilde{f} = (\tilde{f}(t_l))$ is a good estimator for $\ell_{2,n}$ loss. Construct the interpolation $\hat{f}(t) = n^{-1/2} \sum_1^n \tilde{f}(t_l)\phi_{Jl}(t)$. From the decomposition

$$\hat{f} - f = \hat{f} - \tilde{P}_n f + \tilde{P}_n f - f$$

and the identity $\| \hat{f} - \tilde{P}_n f \|_2 = \| \tilde{f} - S_\delta f \|_n$, we obtain from (15.32)

$$\| \hat{f} - f \|_2 \le \| \tilde{f} - f \|_n + o(n^{-r/2})$$

so that $\hat{f}$ has essentially as good performance for L_2 loss as does $\tilde{f}$ for loss $\ell_{2,n}$.

Now suppose on the other hand that $\hat{f}(t)$ is a good estimator for L_2 loss. Construct a discrete estimator $\tilde{f}$ using scaling function coefficients $\tilde{f}(t_l) = (S_\phi \hat{f})_l$. From the identity (15.34) and the decomposition

$$P_n \hat{f} - \tilde{P}_n f = P_n(\hat{f} - f) + P_n f - f + f - \tilde{P}_n f$$

we obtain first using (15.34), and then exploiting projection P_n, Lemma 15.3 and (15.32), that

$$n^{-1/2} \| \tilde{f} - S_\delta f \|_n \le \| \hat{f} - f \|_2 + o(n^{-r/2}).$$

Exercises

15.1 Show that

$$|\psi_I \psi_{I'}(s) - \psi_I \psi_{I'}(t)| \le c 2^{(j+j')/2} 2^{j \vee j'} |s - t|,$$

and that if $\|f\|_L = \sup |f(x) - f(y)|/|x - y|$, then

$$\left| n^{-1} f(t_l) - \int_{t_l}^{t_{l+1}} f \right| \le \tfrac{1}{2} n^{-2} \|f\|_L$$

and hence establish (15.23), and an improvement in which n^{-1} is replaced by $n^{-2} 2^{j \vee j'}$.

16

Epilogue

Brief mentions of topics of recent activity not discussed:

- Compressed sensing
- sparse non-orthogonal linear models
- covariance matrix estimation
- related non-Gaussian results

Appendix A

Appendix: The Minimax Theorem

The aim of this appendix is to give some justification for the minimax Theorem 4.11, restated below as Theorem A.5. Such *statistical* minimax theorems are a staple of statistical decision theory as initiated by Abraham Wald, who built upon the foundation of the two person zero-sum game theory of von Neumann and Morgenstern (1944). It is, however, difficult to find in the published literature a statement of a statistical minimax theorem which is readily seen to cover the situation of our nonparametric result Theorem A.5. In addition, published versions (e.g. Le Cam (1986, Theorem 2.1)) often do not pause to indicate the connections with game theoretic origins.

This appendix gives a brief account of von Neumann's theorem and one of its infinite-dimensional extensions (Kneser, 1952) which aptly indicates what compactness and continuity conditions are needed. Following Brown (1978), we then attempt an account of how statistical minimax theorems are derived, orienting the discussion towards the Gaussian sequence model. While the story does not in fact use much of the special structure of the sequence model, the Gaussian assumption is used at one point to assure the separability of L_1.

In later sections, a number of concepts and results from point set topology and functional analysis are needed, which for reasons of space we do not fully recall here. They may of course be found in standard texts such as Dugundji (1966) and Rudin (1973).

Finite two person zero sum games.

A finite two person, zero sum game can be described by an $m \times n$ *payoff matrix* $A = \{A(i, j)\}$, with the interpretation that if player I uses strategy $i \in \{1, \ldots, m\}$ and player II chooses strategy $j \in \{1, \ldots, n\}$, then player II receives a payoff $A(i, j)$ from player I.

If player I declares his strategy, i say, first, then naturally player II will choose the maximum payoff available in that row, namely $\max_j A(i, j)$. Expecting this, player I will therefore choose i to achieve $\min_i \max_j A(i, j)$. On the other hand, if player II declares his strategy j first, player I will certainly pay only $\min_i A(i, j)$, so that II will receive at most $\max_j \min_i A(i, j)$. Intuitively, II is better off if I has to declare first: indeed one may easily verify that

$$\max_j \min_i A(i, j) \le \min_i \max_j A(i, j). \tag{A.1}$$

When equality holds in (A.1), the game is said to have a *value*. This occurs, for example,

if the game has a *saddlepoint* (i_0, j_0), defined by the property

$$A(i_0, j) \le A(i_0, j_0) \le A(i, j_0) \qquad \text{for all } i, j.$$

However, saddlepoints do not exist in general, as is demonstrated already by the matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ The situation is rescued by allowing *mixed* or randomized strategies, which are probability distributions $x = (x(i))_1^m$ and $y = ((y(j))_1^n$ on the space of nonrandomized rules for each player. If the players use the mixed strategies x and y, then the *expected* payoff from I to II is given by

$$f(x, y) = x^T A y = \sum_{i,j} x(i) A(i, j) y(j). \tag{A.2}$$

Write S_m for the simplex of probability vectors $\{x \in \mathbb{R}^n : x_i \ge 0, \sum x_i = 1\}$. The classical minimax theorem of von Neumann states that for an arbitrary $m \times n$ matrix A in (A.2),

$$\min_{x \in S_m} \max_{y \in S_n} f(x, y) = \max_{y \in S_n} \min_{x \in S_m} f(x, y). \tag{A.3}$$

For the payoff matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, it is easily verified that the fair coin tossing strategies $x = y = (\frac{1}{2}\ \frac{1}{2})$ yield a saddlepoint.

We establish below a more general result that implies (A.3).

Bilinear semicontinuous payoffs

In (A.2) - (A.3), we observe that f is a bilinear function defined on compact, convex sets in Euclidean space. There have been numerous generalizations of this result, either relaxing bilinearity in the direction of convexity-convavity type assumptions on f, or in allowing more general convex spaces of strategies, or in relaxing the continuity assumptions on f. Frequently cited papers include those of Fan (1953) and Sion (1958), and a more recent survey is given by Simons (1995).

We give here a result for bilinear functions on general convex sets due to Kneser (1952) that has a particularly elegant and simple proof. In addition, Kuhn (1953) and Peck and Dulmage (1957) observed that the method extends directly to convex-concave f. First recall that a function $f : X \to \mathbb{R}$ on a topological space X is *lower semicontinuous* (lsc) iff $\{x : f(x) > t\}$ is open for all t, or equivalently if $\{x : f(x) \le t\}$ is closed for all t. [If X is 1st countable, then these conditions may be rewritten in terms of sequences as $f(x) \le \liminf f(x_n)$ whenever $x_n \to x$.] If X is also compact, then an lsc function f attains its infimum: $\inf_{x \in X} f = f(x_0)$ for some $x_0 \in X$.

Theorem A.1 (Kneser, Kuhn) *Let K, L be convex subsets of real vector spaces and $f : K \times L \to \mathbb{R}$ be convex in x for each $y \in L$, and concave in y for each $x \in K$. Suppose also that K is compact and that $x \to f(x, y)$ is lsc for all $y \in L$. Then*

$$\inf_{x \in K} \sup_{y \in L} f(x, y) = \sup_{y \in L} \inf_{x \in K} f(x, y). \tag{A.4}$$

A notable aspect of this extension of the von Neumann theorem is that there are no compactness conditions on L, nor continuity conditions on $y \to f(x, y)$: the topological conditions are confined to the x-slot.

Note that if $x \to f(x, y)$ is lower semi-continuous for all $y \in L$, then $x \to \sup_{y \in L} f(x, y)$ is also lower semi-continuous and so the infimumu on the left side of (A.4) is attained for some $x_0 \in K$.

Here is an example where f is *not* continuous, and only the semicontinuity condition of the theorem holds. Let $\mathbb{R}^\infty$ denote the space of sequences: a countable product of $\mathbb{R}$ with the product topology: $x^{(n)} \to x$ iff for each coordinate i, $x_i^{(n)} \to x_i$. Then the infinite simplex $K = \{x \in \mathbb{R}^\infty : x_i \geq 0, \sum_i x_i \leq 1\}$ is compact. Consider a simple extension of the payoff function (A.2), $f(x, y) = \sum x_i y_i$ for $y \in L = \{y : 0 \leq y_i \leq C \text{ for all } i\}$. Equality (A.4) can easily be checked directly. However, the function $x \to f(x, 1)$ is not continuous: the sequence $x^{(n)} = (1/n, \ldots, 1/n, 0, 0, \ldots.)$ converges to 0 but $f(x^{(n)}, 1) \equiv 1$. However, $f(x, y)$ is lsc in x, as is easily verified.

Kneser's proof nicely brings out the role of compactness and semicontinuity, so we present it here through a couple of lemmas.

Lemma A.2 *Let $f_1, \ldots f_n$ be convex, lsc real functions on a compact convex set K. Suppose for each $x \in K$ that $\max_i f_i(x) > 0$. Then there exists a convex combination that is positive on K: for some $\sigma \in S_n$,*

$$\sum_1^n \sigma_i f_i(x) > 0 \qquad \textit{for all } x \in K.$$

Remark. This lemma implies the standard separating hyperplane theorem in $\mathbb{R}^m$: if K is compact, convex with $0 \notin K$, then there exists a hyperplane separating 0 from K. Indeed, simply let $n = 2m$ and $f_i(x) = x_i$ and $f_{m+i}(x) = -x_i$.

Proof Once the case $n = 2$ is established ($n = 1$ is vacuous), an induction argument can be used. So, with a slight change of notation, assume for all x that $\max\{f(x), g(x)\} > 0$. By lower semicontinuity, the sets $M = \{x : f(x) \leq 0\}$ and $N = \{x : g(x) \leq 0\}$ are closed, and hence compact. The positivity condition implies that M and N are disjoint, and we may assume they are both non-empty, since otherwise the conclusion is trivial. Again by the positivity, on M, both $g > 0$ and the ratio $-f/g$ is defined and usc. Arguing similarly on N, we obtain

$$\max_M \frac{-f}{g} = \frac{-f}{g}(p) = \alpha \geq 0, \qquad \max_N \frac{-g}{f} = \frac{-g}{f}(q) = \beta \geq 0. \tag{A.5}$$

Since $f(p) \leq 0$ and $f(q) > 0$, there exists $\eta > 0$ such that $\eta f(p) + \bar{\eta} f(q) = 0$ [we have set $\bar{\eta} = 1 - \eta$.] Thus, writing $p_\eta = \eta p + \bar{\eta} q$, convexity of f implies $f(p_\eta) \leq 0$. By convexity of g and the positivity condition, $\eta g(p) + \bar{\eta} g(q) \geq g(p_\eta) > 0$. Combining these conclusions with (A.5),

$$\eta g(p) > -\bar{\eta} g(q) = \bar{\eta}\beta f(q) = -\eta\beta f(p) = \eta\alpha\beta g(p),$$

which implies that $\alpha\beta < 1$.

Thus, we may increase α to γ and β to δ in such a way that $\gamma\delta = 1$. Equalities (A.5) then become strict inequalities:

$$\text{On } M, \quad \frac{f + \gamma g}{1 + \gamma} > 0, \qquad\qquad \text{On } N, \quad \frac{\delta f + g}{1 + \delta} > 0.$$

Since $\gamma\delta = 1$, define $\sigma = 1/(1+\gamma) = \delta/(1+\delta)$. Thus on $M \cup N$, we get $\sigma f + \bar{\sigma} g > 0$, and on the rest of K this holds trivially, so the proof is done. □

Lemma A.3 *Either (I) for some* x, $\sup_y f(x,y) \leq 0$, *or*
(II) for some y, $\min_x f(x,y) > 0$.

Proof If (I) is false, then for every x, there exists some value of y, which we call $p(x)$, such that $f(x, p(x)) > 0$. Lower semicontinuity implies that each of the sets $A_y = \{x : f(x,y) > 0\}$ are open, and we have just shown that $x \in A_{p(x)}$. Hence K is covered by $\{A_{p(x)}\}$, so extract a finite subcover indexed by $y_i = p(x_i)$ for some $x_1, \ldots, x_n$. This means exactly that for each $x, \max_i f(x, y_i) > 0$. The previous lemma then gives a probability vector $\sigma \in S_n$ such that for each x, $\sum \sigma_i f(x, y_i) > 0$. By concavity, at $y^* = \sum_1^n \sigma_i y_i$, we have $f(x, y^*) > 0$ for each x. Again using compactness and lsc, $\min_{x \in K} f(x, y^*) > 0$, which implies alternative II. □

Proof of Theorem A.1 That the right side of (A.4) is less than or equal to the left side is elementary, just as in (A.1). Let us suppose, then, that the inequality is strict, so that for some c,

$$\sup_y \inf_x f \leq c < \inf_x \sup_y f. \tag{A.6}$$

Replacing f by $f - c$ does not harm any of the hypotheses, so we may assume that $c = 0$. The left inequality in (A.6) implies that Alternative II in the previous lemma fails, so Alternative I holds, and so $\inf_x \sup_y f \leq 0$, in contradiction with the right hand inequality of (A.6)! Hence there must be equality in (A.6). □

The following corollary is a trivial restatement of Theorem A.1 for the case when compactness and semicontinuity is known for the variable which is being *maximised*.

Corollary A.4 *Let* K, L *be convex subsets of real vector spaces and* $f : K \times L \to \mathbb{R}$ *be convex in* x *for each* $y \in L$, *and concave in* y *for each* $x \in K$. *Suppose also that* L *is compact and that* $y \to f(x,y)$ *is upper semicontinuous for each* $x \in K$. *Then*

$$\inf_{x \in K} \sup_{y \in L} f(x,y) = \sup_{y \in L} \inf_{x \in K} f(x,y). \tag{A.7}$$

Proof Apply Theorem A.1 to $\bar{f}(y,x) = -f(x,y)$. □

A statistical minimax theorem

First, we state the Gaussian sequence model in a little more detail. The sample space $\mathcal{X} = \mathbb{R}^\infty$, the space of sequences in the product topology of pointwise convergence, under which it is complete, separable and metrizable. [Terminology from point-set topology is recalled at Appendix C.11], It is endowed with the Borel $\sigma-$field, and as dominating measure, we take P_0, the centered Gaussian Radon measure (see Bogachev (1998, Example 2.3.5)) defined as the product of a countable number of copies of the standard $N(0,1)$ measure on $\mathbb{R}$. For each $\theta \in \Theta = \ell_2(\mathbb{N}, \lambda)$, the measure P_θ with mean θ is absolutely continuous (indeed equivalent) to P_0, and has density $f_\theta(x) = dP_\theta/dP_0 = \exp\{\langle \theta, x\rangle_\lambda - \|\theta\|_\lambda^2/2\}$. Because

P_0 is Gaussian, the space $L_2(\mathcal{X}, P_0)$ of square integrable functions is separable (Bogachev, 1998, Corollary 3.2.8), and hence so also is $L_1 = L_1(\mathcal{X}, P_0)$.

Let $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ denote the two point compactification of $\mathbb{R}$. As action space we take the countable product $\mathcal{A} = (\bar{\mathbb{R}})^\infty$ which with the product topology is compact, 2° countable and Hausdorff, and again equip it with the Borel $\sigma-$field.

We consider loss functions $L(a, \theta)$ that are non-negative, and perhaps extended-real valued: $L : \mathcal{A} \times \Theta \to [0, \infty]$.

Theorem A.5 *For the above Gaussian sequence model, we assume (i) that for each θ, the map $a \to L(a, \theta)$ is convex and lsc for the product topology on $\mathcal{A}$, and (ii) that $\mathcal{P}$ is a convex set of prior probability measures on $\ell_2(\mathbb{N}, \lambda)$. Then*

$$\inf_{\hat{\theta}} \sup_{\pi \in \mathcal{P}} B(\hat{\theta}, \pi) = \sup_{\pi \in \mathcal{P}} \inf_{\hat{\theta}} B(\hat{\theta}, \pi). \tag{A.8}$$

Our applications of this theorem will typically be to loss functions of the form $L(a, \theta) = w(\|a - \theta\|_p)$, with $w(\cdot)$ a continuous, increasing function. It is easy to verify that such loss functions are lsc in a in the topology of pointwise convergence. Indeed, if $a_i^{(n)} \to a_i^{(\infty)}$ for each i, then for each fixed m, one has

$$\sum_{i=1}^{m} |a_i^{(\infty)} - \theta_i|^p = \lim_n \sum_{i=1}^{m} |a_i^{(n)} - \theta_i|^p \le \liminf_n \|a^{(n)} - \theta\|_p^p.$$

Theorem A.5 and other statistical minimax theorems, while closely related to Theorem A.1, as will be seen below, do not seem to follow directly from it, using instead separating hyperplane results (compare Lemma A.2).

A general framework for statistical decision theory, including minimax and complete class results, has been developed by its chief exponents, including A. Wald, L. LeCam, C. Stein, and L. Brown, in published and unpublished works. A selection of references includes Wald (1950); LeCam (1955); Le Cam (1986); Diaconis and Stein (1983); Brown (1977, 1978).

The theory is general enough to handle abstract sample spaces and unbounded loss functions, but it is difficult to find a statement that immediately covers our Theorem A.5. We therefore give a summary description of the steps in the argument for Theorem A.5, using freely the version of the Wald-LeCam-Brown approach set out in Brown (1978). The theory of Brown (1978) was developed specifically to handle both parametric and nonparametric settings, but few nonparametric examples were then discussed explicitly. Proofs of results given there will be omitted, but we hope that this outline nevertheless has some pedagogic value in stepping through the general method in the concrete setting of the nonparametric Gaussian sequence model.

Remark. There is a special case (which includes the setting of a bounded normal mean, Section 4.6), in which our statistical minimax theorem can be derived directly from the Kneser-Kuhn theorem. Indeed, if $\Theta \subset \mathbb{R}^n$ is compact, and $\mathcal{P} = \mathcal{P}(\Theta)$, then $\mathcal{P}$ is compact for weak convergence of probability measures. Let K be the class of estimators $\hat{\theta}$ with finite risk functions on Θ, let $L = \mathcal{P}$ and for the payoff function f take $B(\hat{\theta}, \pi) = \int_\Theta r(\hat{\theta}, \theta)\pi(d\theta)$.

Observe that K is convex because $a \to L(a, \theta)$ is; that L is convex *and* compact; and that B is convex-linear. Finally $\pi \to B(\hat{\theta}, \pi)$ is continuous since in the Gaussian model

$y_i = \theta_i + \epsilon\lambda_i z_i$, the risk functions $\theta \to r(\hat{\theta}, \theta)$ are continuous and bounded on the compact set Θ. Hence the Kneser-Kuhn Corollary (A.7) applies to provide the minimax result.

Randomized decision rules.

The payoff function $B(\hat{\theta}, \pi)$ appearing in Theorem A.5 is linear in π, but not in $\hat{\theta}$. Just as in the two-person game case, the standard method in statistical decision theory for obtaining linearity is to introduce *randomized decision rules*. These are Markov kernels $\delta(da|x)$ with two properties: (i) for each $x \in \mathcal{X}$, $\delta(\cdot|x)$ is a probability measure on $\mathcal{A}$ which describes the distribution of the random action a given that x is observed, and (ii), for each measurable A, the map $x \to \delta(A|x)$ is measurable. The risk function of a randomized rule δ is

$$r(\delta, \theta) = \int\int L(a, \theta)\delta(da|x)P_\theta(dx), \tag{A.9}$$

and the payoff function we consider is the integrated risk against a probability measure π:

$$B(\delta, \pi) = \int r(\delta, \theta)\pi(d\theta).$$

A major reason for introducing $B(\delta, \pi)$ is that it is bilinear in δ and π. Further, writing $\mathcal{D}$ for the class of all randomized decision rules, we note that both it and $\mathcal{P}$ are convex. To establish a minimax statement

$$\inf_{\delta\in\mathcal{D}} \sup_{\pi\in\mathcal{P}} B(\delta, \pi) = \sup_{\pi\in\mathcal{P}} \inf_{\delta\in\mathcal{D}} B(\delta, \pi), \tag{A.10}$$

Kneser's theorem suggests that we need a topology on decision rules δ with two key properties:

(P1) $\mathcal{D}$ is compact, and
(P2) the risk functions $\delta \to B(\delta, \pi)$ are lower semicontinuous.

Before describing how this is done, we explain how (A.8) follows from (A.10) using the convexity assumption on the loss function. Indeed, given a randomized rule δ, the standard method is to construct a *non-randomized rule* by averaging: $\hat{\theta}_\delta(x) = \int a\delta(da|x)$. Convexity of $a \to L(a, \theta)$ and Jensen's inequality then imply that

$$L(\hat{\theta}_\delta(x), \theta) \le \int L(a, \theta)\delta(da|x).$$

Averaging over $X \sim P_\theta$, and recalling (A.9) shows that $\hat{\theta}_\delta$ is at least as good as δ:

$$r(\hat{\theta}_\delta, \theta) \le r(\delta, \theta) \qquad \text{for all } \theta \in \Theta. \tag{A.11}$$

Consequently, with convex loss functions, there is no reason ever to use a randomized decision rule, since there is always a better non-randomized one. In particular, integrating with respect to an arbitrary π yields

$$\sup_\pi B(\hat{\theta}_\delta, \pi) \le \sup_\pi B(\delta, \pi). \tag{A.12}$$

We then recover (A.8) from (A.10) via a simple chain of inequalities:

$$\inf_{\hat\theta}\sup_{\pi} B(\hat\theta,\pi) \le \inf_{\hat\theta_\delta}\sup_{\pi} B(\hat\theta_\delta,\pi) \le \inf_{\delta}\sup_{\pi} B(\delta,\pi)$$
$$= \sup_{\pi}\inf_{\delta} B(\delta,\pi) \le \sup_{\pi}\inf_{\hat\theta} B(\hat\theta,\pi) \le \inf_{\hat\theta}\sup_{\pi} B(\hat\theta,\pi),$$

and since the first and last terms are the same, all terms are equal.

A compact topology for $\mathcal{D}$

We return to establishing properties [P1] and [P2]. The approach is to identify decision rules δ with bilinear, bicontinuous functionals, and then use the Alaoglu theorem on weak compactness to induce a topology on $\mathcal{D}$.

For this section, we write $L_\theta(a)$ for the loss function to emphasise the dependence on a. The risk function of a rule δ may then be written

$$r(\delta,\theta) = \int\int L_\theta(a) f_\theta(x)\delta(da|x)P_0(dx) = b_\delta(f_\theta, L_\theta),$$

Here the probability density f_θ is regarded as a non-negative function in the Banach space $L_1 = L_1(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), P_0)$ which is separable as noted earlier. Since $\mathcal{A} = (\bar{\mathbb{R}})^\infty$ is compact, metrizable and second countable, the Banach space $C = C(\mathcal{A})$ of continuous functions on $\mathcal{A}$, equipped with the uniform norm, is also separable. The functional

$$b_\delta(g,c) = \int\int g(x)c(a)\delta(da|x)P_0(dx)$$

belongs to the Banach space B of bilinear, bicontinuous functionals on $L_1 \times C$ with the operator norm $\|b\|_B = \sup\{|b(g,c)| : \|g\|_{L_1} = \|c\|_C = 1\}$. Under assumptions satisfied here, Brown (1978) shows that the mapping $\iota : \delta \to b_\delta$ is a bijection of $\mathcal{D}$ onto

$$B_1^+ = \{b \in B : b \ge 0 \text{ and } b(g,1) = \|g\|_{L_1} \forall g \ge 0\}$$
$$\subset \{b : \|b\|_B \le 1\},$$

and the latter set, by Alaoglu's theorem, is compact in the weak topology, which by separability of L_1 and C is also metrizable on such norm bounded sets. Thus, B_1^+, being a closed subset, is also compact. The map ι is then used to induce a compact metrizable topology on $\mathcal{D} = \iota^{-1}(B_1^+)$ in which convergence may be described by sequences: thus $\delta_i \to \delta$ means that

$$b_{\delta_i}(g,c) \to b_\delta(g,c) \qquad \forall (g,c) \in L_1 \times C. \tag{A.13}$$

This topology also satisfies our second requirement: that the maps $\delta \to B(\delta,\pi)$ be lsc. Indeed, since $\mathcal{A}$ is second countable, the lsc loss functions can be approximated by an increasing sequence of continuous functions $c_i \in C$: $L_\theta(a) = \lim_i c_i(a)$. This implies that

$$r(\delta,\theta) = \sup_c\{b_\delta(f_\theta,c) : c \le L_\theta\}.$$

The definition (A.13) says that the maps $\delta \to b_\delta(f_\theta,c)$ are each continuous, and so $\delta \to r(\delta,\theta)$ appears as the upper envelope of a family of continuous functions, and is hence lsc. Finally Fatou's lemma implies that $\delta \to B(\delta,\pi) = \int r(\delta,\theta)\pi(d\theta)$ is lsc.

A separation theorem

We have now established $B(\delta, \pi)$ as a bilinear function on $\mathcal{D} \times \mathcal{P}$ which for each π fixed is lsc on the compact $\mathcal{D}$. What prevents us from applying Kneser's minimax theorem directly is that $B(\delta, \pi)$ can be infinite. The strategy used by Brown (1978) for handling this difficulty is to prove a separation theorem for *extended* real valued functions, and derive from this the minimax result.

Slightly modified for our context, this approach works as follows. Let $T = T(\mathcal{P}, [0, \infty])$ denote the collection of all functions $b : \mathcal{P} \to [0, \infty]$ – with the product topology, this space is compact by Tychonoff's theorem. Now define an upper envelope of the risk functions by setting $\Gamma = \rho(\mathcal{D})$ and then defining

$$\tilde{\Gamma} = \{b \in T : \text{there exists } b' \in \Gamma \text{ with } b' \leq b\}.$$

Brown uses the $\mathcal{D}$ topology constructed above, along with the compactness and lower semi-continuity properties [P1] and [P2] to show that $\tilde{\Gamma}$ is closed and hence compact in T.

Using the separating hyperplane theorem for Euclidean spaces – a consequence of Lemma A.2 – Brown shows

Theorem A.6 *Suppose that $\tilde{\Gamma}$ is convex and closed in T and that $b_0 \in T \backslash \tilde{\Gamma}$. Then there exists $c > 0$, a finite set $(\pi_i)_1^m \subset \mathcal{P}$ and a probability vector $(\xi_i)_1^m$ such that the convex combination $\pi_\xi = \sum \xi_i \pi_i \in \mathcal{P}$ satisfies*

$$b_0(\pi_\xi) < c < b(\pi_\xi) \qquad \text{for all } b \in \tilde{\Gamma}. \tag{A.14}$$

It is now easy to derive the minimax conclusion (A.10). Indeed, write $V = \inf_\delta \sup_\mathcal{P} B(\delta, \pi)$. If $V < \infty$, let $\epsilon > 0$ and choose $b_0 \equiv V - \epsilon$ – clearly $b_0 \notin \tilde{\Gamma}$. Convexity of $\mathcal{D}$ entails convexity of $\tilde{\Gamma}$, which is also closed in T as we saw earlier. Hence, the separation theorem produces $\pi_\xi \in \mathcal{P}$ such that

$$V - \epsilon = b_0(\pi_\xi) < \inf_\delta B(\delta, \pi_\xi).$$

In other words, $\sup_\pi \inf_\delta B(\delta, \pi) > V - \epsilon$ for each $\epsilon > 0$, and hence it must equal V. If $V = \infty$, a similar argument using $b_0 \equiv m$ for each finite m also yields (A.10).

A.1 A special minimax theorem for thresholding

It is sometimes of interest to restrict the estimator δ in $B(\delta, \pi)$ to a smaller class, for example threshold rules that depend on a single parameter, the threshold λ. We write $B(\lambda, \pi)$ for the payoff function in such cases (for details, see Section 14.4).

In such cases $\lambda \to B(\lambda, \pi)$ need not be convex and so our earlier minimax theorems do not directly apply. In addition, we would like to exhibit a saddle point. In this section, then, we formulate and prove a special minimax theorem tailored to this setting. First, a definition. We call a function $\lambda(\pi)$ defined for π in a convex set $\mathcal{P}$ *Gâteaux continuous* at π_0 if $\lambda\big((1-t)\pi_0 + t\pi_1\big) \to \lambda(\pi_0)$ as $t \to 0$ for each $\pi_1 \in \mathcal{P}$.

Theorem A.7 *Suppose $\Lambda \subset \mathbb{R}$ is an interval and that $\mathcal{P}$ is convex and compact. Suppose*

that $B : \Lambda \times \mathcal{P} \to \mathbb{R}$ is linear and continuous in π for each $\lambda \in \Lambda$. Then there exists a least favorable π_0.

Suppose also for each π that $B(\lambda, \pi)$ is continuous in λ, that there is a unique $\lambda(\pi)$ that minimizes B, and that $\lambda(\pi)$ is Gâteaux continuous at π_0. Set $\lambda_0 = \lambda(\pi_0)$.

Then the pair (λ_0, π_0) is a saddlepoint: for all $\lambda \in [0, \infty)$ and $\pi \in \mathcal{P}$,

$$B(\lambda_0, \pi) \le B(\lambda_0, \pi_0) \le B(\lambda, \pi_0), \tag{A.15}$$

and hence

$$\inf_{\lambda} \sup_{\mathcal{P}} B(\lambda, \pi) = \sup_{\mathcal{P}} \inf_{\lambda} B(\lambda, \pi) = \sup_{\mathcal{P}} B_S(\pi).$$

Proof The right side of (A.15) follows from the definition of $\lambda(\pi_0)$. For the left side, given an arbitrary $\pi_1 \in \mathcal{P}$, define $\pi_t = (1-t)\pi_0 + t\pi_1$ for $t \in [0, 1]$: by convexity, $\pi_t \in \mathcal{P}$. Let $\lambda_t = \lambda(\pi_t)$ be the best threshold for π_t, so that $B(\pi_t) = B(\lambda_t, \pi_t)$. Heuristically, since π_0 is least favorable, we have $(d/dt)B(\pi_t)|_{t=0} \le 0$, and we want to compute partial derivatives of $B(\lambda_t, \pi_t)$ and then exploit linearity in π.

More formally, for $t > 0$ we have

$$B(\lambda_t, \pi_t) - B(\lambda_0, \pi_0) = B(\lambda_t, \pi_0) - B(\lambda_0, \pi_0) + B(\lambda_0, \pi_t) - B(\lambda_0, \pi_0) + \Delta^2 B,$$

where the left side is ≤ 0 and

$$\Delta^2 B = B(\lambda_t, \pi_t) - B(\lambda_t, \pi_0) - B(\lambda_0, \pi_t) + B(\lambda_0, \pi_0).$$

Now also $B(\lambda_t, \pi_0) \ge B(\lambda_0, \pi_0)$ and by linearity $B(\lambda_0, \pi_t) - B(\lambda_0, \pi_0) = t[B(\lambda_0, \pi_1) - B(\lambda_0, \pi_0)]$ and so

$$0 \ge B(\lambda_0, \pi_1) - B(\lambda_0, \pi_0) + \Delta^2 B/t.$$

Again using the linearity in π,

$$\Delta^2 B/t = [B(\lambda_t, \pi_1) - B(\lambda_0, \pi_1] - [B(\lambda_t, \pi_0) - B(\lambda_0, \pi_0)] \to 0$$

as $t \to 0$, since $\lambda_t \to \lambda_0$ by Gâteaux continuity of $\lambda(\pi)$, and since $\lambda \to B(\lambda, \pi)$ is continuous. This shows that $B(\lambda_0, \pi_1) \le B(\lambda_0, \pi_0)$ for any $\pi_1 \in \mathcal{P}$ and completes the proof. □

Remark. Proposition 13.10 shows that $B(\lambda, \pi)$ is quasi-convex in λ, and since it is also linear in π on a convex set, one could appeal to a general minimax theorem, e. g. Sion (1958). However, the general minimax theorems do not exhibit a saddlepoint, which emerges directly from the present more specialized approach.

Exercise. Complete the induction step for the proof of Lemma A.2.

Appendix B

More on Wavelets and Function Spaces

B.1 Building scaling functions and wavelets

We sketch two common constructions of a scaling function φ (and then the corresponding wavelet ψ): (a) beginning from a Riesz basis, and (b) starting from discrete (especially finite) filters.

(a) Using a Riesz basis. A family $\{e_k\}_{k\in\mathbb{N}}$ is a *Riesz basis* for a Hilbert space H if (i) for all $h \in H$, there is a unique representation $h = \sum \alpha_k e_k$, and (ii) there exist positive absolute constants C_1, C_2 such that for all $h \in H$, $C_1\|h\|^2 \le \sum_k |\alpha_k|^2 \le C_2\|h\|^2$.

It is more common to replace the multiresolution analysis condition (iv) by the weaker condition

(iv') $\exists\, \theta \in V_0$ such that $\{\theta(x-k) : k \in \mathbb{Z}\}$ is a Riesz basis for V_0.[1]

That (iv') is equivalent to (iv) follows from the "orthonormalization trick" discussed below.

A key role in constructions and interpretations is played by the frequency domain and the Fourier transform (C.7). The Plancherel identity (C.8) leads to a frequency domain characterization of the orthnormality and Riesz basis conditions (iv) and (iv'):

Lemma B.1 *Suppose $\varphi \in L_2$. The set $\{\varphi(x-k), k \in \mathbb{Z}\}$ is (i) orthonormal iff*

$$\sum_k |\hat\varphi(\xi + 2k\pi)|^2 = 1 \qquad a.e., \tag{B.1}$$

and (ii) a Riesz basis iff there exist positive constants C_1, C_2 such that

$$C_1 \le \sum_k |\hat\varphi(\xi + 2k\pi)|^2 \le C_2 \qquad a.e. \tag{B.2}$$

Partial Proof We give the easy proof of (B.1) since it gives a hint of the role of frequency domain methods. The Fourier transform of $x \to \varphi(x-n)$ is $e^{-in\xi}\hat\varphi(\xi)$. Thus, orthonormality combined with the Plancherel identity gives

$$\delta_{0n} = \int_{-\infty}^{\infty} \varphi(x)\overline{\varphi(x-n)}dx = \int_{-\infty}^{\infty} e^{in\xi}|\hat\varphi(\xi)|^2 d\xi.$$

Now partition $\mathbb{R}$ into segments of length 2π, add the integrals, and exploit periodicity of $e^{in\xi}$ to rewrite the right hand side as

$$\int_0^{2\pi} e^{in\xi} \sum_k |\hat\varphi(\xi + 2k\pi)|^2 d\xi = \delta_{0n}.$$

[1] This use of the θ symbol, local to this appendix, should not be confused with the notation for wavelet coefficients in the main text.

The function in (B.1) has as Fourier coefficients the delta sequence δ_{0n} and so equals 1 a.e. □

The "orthonormalization trick" creates (B.1) by fiat:

Theorem B.2 *Suppose that $\{V_j\}$ is an MRA, and that $\{\theta(x-k), k \in \mathbb{Z}\}$ is a Riesz basis for V_0. Define*

$$\hat{\varphi}(\xi) = \hat{\theta}(\xi)/\{\sum_k |\hat{\theta}(\xi + 2k\pi)|^2\}^{1/2}. \tag{B.3}$$

Then φ is a scaling function for the MRA, and so for all $j \in \mathbb{Z}$, $\{\varphi_{jk} : k \in \mathbb{Z}\}$ is an orthonormal basis for V_j.

Example. Box spline MRA. (See also Chapter 7.1.) Given $r \in \mathbb{N}$, set $\chi = I_{[0,1]}$ and $\theta = \theta_r = \chi \star \cdots \star \chi = \chi^{\star(r+1)}$. Without any loss of generality, we may shift $\theta_r = \chi^{\star(r+1)}$ by an integer so that the center of the support is at 0 if r is odd, and at $1/2$ if r is even. Then it can be shown (Meyer, 1990, p61), (Mallat, 1999, Sec 7.1) that

$$\hat{\theta}_r(\xi) = \left(\frac{\sin \xi/2}{\xi/2}\right)^{r+1} e^{-i\epsilon\xi/2} \qquad \epsilon = \begin{cases} 1 & r \text{ even} \\ 0 & r \text{ odd} \end{cases},$$

$$\sum_k |\hat{\theta}_r(\xi + 2k\pi)|^2 = P_{2r}(\cos \xi/2),$$

where P_{2r} is a polynomial of degree $2r$. For example, in the piecewise linear case $r = 1$, $P_2(v) = (1/3)(1 + 2v^2)$. Using (B.2), this establishes the Riesz basis condition (iv') for this MRA. Thus (B.3) gives an explicit Fourier domain expression for φ which is amenable to numerical calculation. Mallat (1999, pp. 226-228) gives corresponding formulas and pictures for cubic splines.

(b) Using finite filters. The MRA conditions imply important structural constraints on $\hat{h}(\xi)$: using (B.1) and (7.2) it can be shown that

Proposition B.3 *If φ is an integrable scaling function for an MRA, then*

$$(CMF) \qquad |\hat{h}(\xi)|^2 + |\hat{h}(\xi + \pi)|^2 = 2 \qquad \forall \xi \in \mathbb{R} \tag{B.4}$$

$$(NORM) \qquad \hat{h}(0) = \sqrt{2}. \tag{B.5}$$

(B.4) is called the *conjugate mirror filter* (CMF) condition, while (B.5) is a normalization requirement. Conditions (B.5) and (B.4) respectively imply constraints on the discrete filters:

$$\sum h_k = \sqrt{2}, \qquad \sum h_k^2 = 1.$$

They are the starting point for a unified construction of many of the important wavelet families (Daubechies variants, Meyer, ...) that begins with the filter $\{h[k]\}$, or equivalently $\hat{h}(\xi)$. Here is a key result in this construction.

Theorem B.4 (Meyer, Mallat) *If $\hat{h}(\xi)$ is $2\pi-$periodic, C^1 near $\xi = 0$ and (a) satisfies* (B.4) *and* (B.5)*, and (b)* $\inf_{[-\pi/2,\pi/2]} |\hat{h}(\xi)| > 0$*, then*

$$\hat{\varphi}(\xi) = \prod_{l=1}^{\infty} \frac{\hat{h}(2^{-l}\xi)}{\sqrt{2}} \tag{B.6}$$

is the Fourier transform of a scaling function $\varphi \in L_2$ that generates an MRA.

Notes: 1. That $\hat{\varphi}$ is generated by an infinite product might be guessed by iteration of the two scale relation (7.2): the work lies in establishing that all MRA properties hold.

2. Condition (b) can be weakened to a necessary and sufficient condition due to Cohen (1990) (see also Cohen and Ryan (1995)).

Building wavelets. The next lemma gives the conditions on g in order that ψ be an orthonormal wavelet: it is an analog of Proposition B.3.

Lemma B.5 *[Mallat Lemma 7.1] $\{\psi_{jk}, k \in \mathbb{Z}\}$ is an orthonormal basis for W_j, the orthocomplement of V_j in V_{j+1} if and only if, for all $\xi \in \mathbb{R}$,*

$$|\hat{g}(\xi)|^2 + |\hat{g}(\xi + \pi)|^2 = 2 \tag{B.7}$$

$$\hat{g}(\xi)\hat{h}^*(\xi) + \hat{g}(\xi + \pi)\hat{h}^*(\xi + \pi) = 0. \tag{B.8}$$

Here $\hat{h}^*$ denotes the complex conjugate of $\hat{h}$.

One way to satisfy (B.7) and (B.8) is to set

$$\hat{g}(\xi) = e^{-i\xi}\hat{h}^*(\xi + \pi) \tag{B.9}$$

To understand this in the time domain, note that if $\hat{s}(\xi)$ has (real) coefficients s_k, then conjugation corresponds to time reversal: $\hat{s}^*(\xi) \leftrightarrow s_{-k}$, while modulation corresponds to time shift: $e^{i\xi}\hat{s}(\xi) \leftrightarrow s_{k+1}$, and the frequency shift by π goes over to time modulation: $\hat{s}(\xi + \pi) \leftrightarrow (-1)^k s_k$. To summarize, interpreting (B.9) this in terms of filter coefficients, one obtains the "mirror" relation

$$g_k = (-1)^{1-k} h_{1-k}. \tag{B.10}$$

Together, (7.4) and (B.9) provide a frequency domain recipe for constructing a candidate wavelet from φ:

$$\hat{\psi}(2\xi) = e^{-i\xi}\hat{h}^*(\xi + \pi)\hat{\varphi}(\xi). \tag{B.11}$$

Of course, there is still work to do to show that this does the job:

Theorem B.6 *[Mallat Th. 7.3] If g is defined by* (B.9)*, and ψ by* (7.4)*, then $\{\psi_{jk}, (j,k) \in \mathbb{Z}^2\}$ is an orthonormal basis for $L_2(\mathbb{R})$.*

Example. Box splines again. Given $\hat{\varphi}$, one constructs $\hat{h}$ from (7.2), $\hat{g}$ from (B.9) and $\hat{\psi}$ from (7.4). This leads to the *Battle-Lemarié spline wavelets* (see also Chui (1992)). The case $r = 0$ yields the Haar wavelet: $\psi(x) = I_{[1/2,1]}(x) - I_{[0,1/2]}(x)$ - verifying this via this construction is possibly a useful exercise in chasing definitions. However, the point of the construction is to yield wavelets with increasing regularity properties as r increases.

Example. The class of *Meyer wavelets* (Meyer, 1986) is built from a filter $\hat{h}(\xi)$ on $[-\pi, \pi]$ satisfying

$$\hat{h}(\xi) = \begin{cases} \sqrt{2} & |\xi| \le \pi/3 \\ 0 & |\xi| \ge 2\pi/3, \end{cases}$$

the CMF condition (B.4), and that is also required to be C^n at the join points $\pm\pi/3$ and

$\pm 2\pi/3$. In fact C^∞ functions exist with these properties, but for numerical implementation one is content with finite values of n, for which computable descriptions are available: for example $n = 3$ in the case given by Daubechies (1992, p137-8).

The scaling function $\hat{\varphi}(\xi) = \prod_1^\infty 2^{-1/2}\hat{h}(2^{-j}\xi)$ then has support in $[-4\pi/3, 4\pi/3]$, and the corresponding wavelet (defined from (7.4) and (B.9)) has support in the interval $\pm[2\pi/3, 8\pi/3]$. Since $\hat{\varphi}$ and $\hat{\psi}$ have compact support, both $\varphi(x)$ and $\psi(x)$ are C^∞ – unlike, say, Daubechies wavelets. However, they cannot have exponential decay in the time domain (which is impossible for C^∞ orthogonal wavelets, according to Daubechies (1992, Corollary 5.5.3)) – at least they are $O(|x|^{-n-1})$ if $\hat{h}$ is C^n. Finally, since $\hat{\psi}$ vanishes in a neighborhood of the origin, all its derivatives are zero at 0 and so ψ has an infinite number of vanishing moments.

Figure B.1 shows a schematic of the qualitative frequency domain properties of the squared modulus of $\hat{\varphi}, \hat{h}, \hat{g}$ and finally $\hat{\psi}$. It can be seen that the space V_0 generated by translates of φ corresponds roughly to frequencies around $\pm[0, \pi]$, while the space W_j contains frequencies around $\pm[2^j\pi, 2^{j+1}\pi]$. More precisely, it can be shown (Hernández and Weiss, 1996, p.332, and p.61) that φ and the dilations of ψ form a partition of frequency space in the sense that

$$|\hat{\varphi}(\xi)|^2 + \sum_{j=0}^{\infty} |\hat{\psi}(2^{-j}\xi)|^2 = 1 \qquad \text{a.e.} \tag{B.12}$$

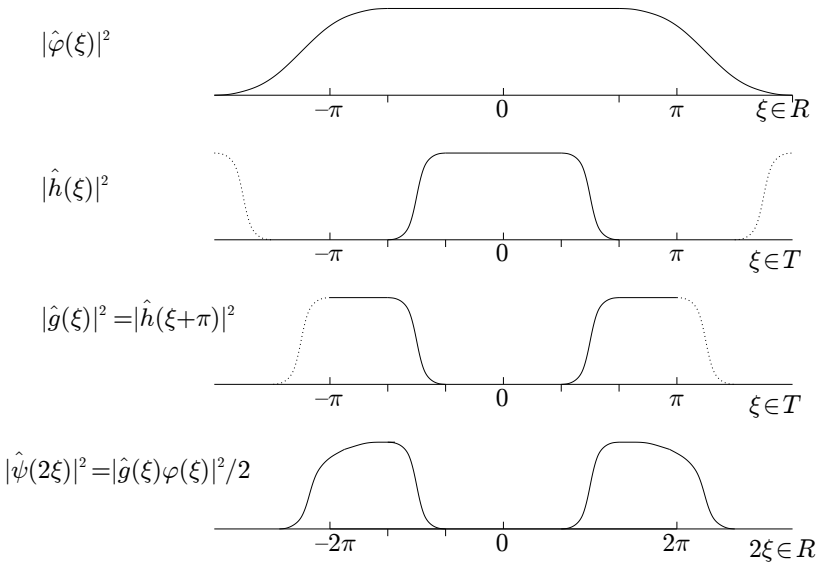

Figure B.1 qualitative frequency domain properties of scaling function $\hat{\varphi}$, transfer functions $\hat{h}, \hat{g}$ and wavelet $\hat{\psi}$ corresponding to the Meyer wavelet; dotted lines show extension by periodicity

Vanishing moments. The condition that ψ have r vanishing moments has equivalent formulations in terms of the Fourier transform of ψ and the filter h.

Lemma B.7 *Let ψ be an orthonormal wavelet. If $\hat{\psi}$ is C^p at $\xi = 0$, then the following are equivalent:*

$$\begin{aligned} &(i) && \int t^j \psi = 0, && j = 0, \dots, p-1. \\ &(ii) && D^j \hat{\psi}(0) = 0, && j = 0, \dots, p-1. \\ &(iii) && D^j \hat{h}(\pi) = 0 && j = 0, \dots, p-1. \quad (VM_p) \end{aligned} \tag{B.13}$$

See for example Mallat (1999)[Theorem 7.4] or Härdle et al. (1998)[Theorem 8.3].

Meyer (1990)[p38] shows that a wavelet deriving from an r-regular multresolution analysis necessarily has $r + 1$ vanishing moments.

Example. Daubechies wavelets. Here is a brief sketch, with a probabilistic twist, of some of the steps in Daubechies' construction of orthonormal wavelets of compact support. Of course, there is no substitute for reading the original accounts (see Daubechies (1988), Daubechies (1992, Ch. 6), and for example the descriptions by Mallat (1999, Ch. 7) and Meyer (1990, Vol I, Ch. 3)).

The approach is to build a filter $h = \{h_k\}_0^{N-1}$ with $h_k \in \mathbb{R}$ and transfer function $\hat{h}(\xi) = \sum_{k=0}^{N-1} h_k e^{-ik\xi}$ satisfying these conditions and then derive the conjugate filter g and the wavelet ψ from (B.9), (B.11) and Theorem B.6. The vanishing moment condition of order p (VM_p) implies that $\hat{h}(\xi)$ may be written

$$\hat{h}(\xi) = \Big(\frac{1 + e^{-i\xi}}{2}\Big)^p r(\xi), \qquad r(\xi) = \sum_0^m r_k e^{-ik\xi},$$

with $N = p + m + 1$ and $r_k \in \mathbb{R}$. Passing to squared moduli, one may write[2]

$$|\hat{h}(\xi)|^2 = 2(\cos^2 \tfrac{\xi}{2})^p \, P(\sin^2 \tfrac{\xi}{2})$$

for some real polynomial P of degree m. The conjugate mirror filter condition (B.4) then forces, on putting $y = \sin^2 \xi/2$,

$$(1-y)^p P(y) + y^p P(1-y) = 1 \qquad 0 \le y \le 1. \tag{B.14}$$

To have the support length N as small as possible, one seeks solutions of (B.14) of minimal degree m. One solution can be described probabilistically in terms of repeated independent tosses of a coin with $Pr(\text{Heads}) = y$. Either p tails occur before p heads or vice versa, so

$$\begin{aligned} P(y) &:= Pr\{p \; Ts \text{ occur before } p \; Hs\}/(1-y)^p \\ &= \sum_{k=0}^{p-1} \binom{p+k-1}{k} y^k \end{aligned}$$

certainly solves (B.14). Further, it is the *unique* solution of degree $p - 1$ or less[3] .

To return from the squared modulus scale, appeal to the F. Riesz lemma: if $s(\xi) = \sum_{-m}^{m} s_k e^{-ik\xi} \ge 0$, then there exists $r(\xi) = \sum_0^m r_k e^{-ik\xi}$ such that $s(\xi) = |r(\xi)|^2$, and if $\{s_k\}$ are real, then the $\{r_k\}$ can be chosen to be real also.

The lemma is applied to $s(\xi) = P(\sin^2 \frac{\xi}{2}) \ge 0$, and so one arrives at orthonormal wavelets with support length $N = 2p$ for $p = 1, 2, \dots$ The uniqueness argument shows

that $N < 2p$ is not possible. The choice $N = 2$ yields Haar wavelets and $N = 4$ gives the celebrated $D4$ wavelet of Daubechies. For $N \geq 6$ there are non-unique choices of solution to the construction of the "square root" $r(\xi)$ (a process called spectral factorization), and Daubechies (1992, Ch. 6) describes some families of solutions (for example, directed towards least asymmetry) along with explicit listings of coefficients.

Discussion. Table B.1 sets out some desiderata for a wavelet basis. The last three requirements are in a sense contradictory: it turns out that higher regularity of ψ can only be achieved with longer filters. One advantage of Daubechies' *family* of wavelets ψ_N, indexed by support size N, is to make this tradeoff directly apparent: the smoothness of ψ increases with N at approximate rate $0.2N$ (Daubechies, 1992, §7.12).

Table B.1 *Desirable properties of orthonormal wavelet family, together with corresponding conditions on the filter h*

1. Orthonormal wavelet ψ	$\leftrightarrow$ CMF (B.4) and NORM (B.5)
2. p vanishing moments	$\leftrightarrow$ VM_p (B.13)
3. (small) compact support	$\leftrightarrow$ N small
4. (high) regularity of ψ	

Proof of Lemma 7.2 We first recall that Hölder functions can be uniformly approximated by (Taylor) polynomials, cf. (C.20). So, let $p(y)$ be the approximating Taylor polynomial of degree $\lceil \alpha \rceil - 1$ at $x_k = k2^{-j}$. Using a change of variable and the vanishing moments property,

$$\int f(x)2^{j/2}\psi(2^j x - k)dx = 2^{-j/2}\int [f(x_k + 2^{-j}v) - p(2^{-j}v)]\psi(v)dv.$$

Hence, using the Hölder bound (C.20),

$$|\langle f, \psi_{jk}\rangle| \leq 2^{-j/2}C2^{-j\alpha}\int |v|^\alpha |\psi(v)|dv.$$

Setting c_ψ equal to the latter integral yields the result. □

Vanishing moments for the scaling function. The approximation of point values $f(t_i)$ of a function by scaling function coefficients $\langle f, 2^{j/2}\varphi_{jk}\rangle$ is similarly dependent on the smoothness of f and the number of vanishing moments of φ. Bearing in mind that the scaling function itself has $\int \varphi = 1$ (e.g. from (**??**)) we say that φ has r vanishing moments if

$$\int x^k \varphi(x)dx = 0 \qquad k = 1, \ldots r - 1.$$

Lemma B.8 *If f is C^α on $\mathbb{R}$ and φ has at least $r = \lceil \alpha \rceil$ vanishing moments,*

$$|\langle f, \varphi_{jk}\rangle - 2^{-j/2}f(k2^{-j})| \leq c_\psi C 2^{-j(\alpha+1/2)}.$$

Proof Modify the proof of Lemma 7.2 by writing the approximating polynomial at $x_k =$

$k2^{-j}$ in the form $p(y) = f(x_k) + p_1(y)$ where p_1 is also of degree $r-1$, but with no constant term, so that $\int p_1\varphi = 0$. Then

$$\int f\varphi_{jk} - 2^{-j/2}f(x_k) = 2^{-j/2}\int [f(x_k + 2^{-j}v) - f(x_k) - p_1(2^{-j}v)]\varphi(v)dv$$

and so

$$|\langle f, \varphi_{jk}\rangle - 2^{-j/2}f(x_k)| \le 2^{-j/2}C2^{-j\alpha}c_\varphi,$$

where again $c_\varphi = \int |v|^\alpha|\varphi(v)|dv$. □

Life on the interval.

Vanishing moments for wavelets on $[0,1]$. Let $\mathcal{P}_p$ denote the space of polynomials of degree p. The vanishing moments theorem (e.g. Mallat (1999, Theorem 7.4)) states that if φ and ψ have sufficiently rapid decay, then ψ has p vanishing moments if and only if the Strang-Fix condition is satisfied:

$$\theta_l(t) = \sum_{k=-\infty}^{\infty} k^l\varphi(t-k) \in \mathcal{P}_l \qquad l = 0, 1, \ldots, p-1. \tag{B.15}$$

The condition (B.15) says that $\mathcal{P}_{p-1} \subset V_j$ and further (see Cohen et al. (1993b)) that for $j \ge J_*$, $\mathcal{P}_{p-1} \subset V_j[0,1]$—the multiresolution spaces corresponding to the CDJV construction. Consequently $\mathcal{P}_{p-1} \perp W_j[0,1]$ and so for $j \ge J_*$, $k = 1, \ldots, 2^j$, we have

$$\int t^l\psi_{jk}^{\text{int}}(t)dt = 0, \qquad l = 0, 1, \ldots, p-1.$$

A key point is the existence of approximations by V_j with better smoothness and approximation properties than those of the Haar multiresolution. Following Meyer (1990, p22), we say that a multiresolution analysis $\{V_j\}$ of $L_2(\mathbb{R})$ is $r-$*regular* if $D^k\theta(x)$ is rapidly decreasing for $0 \le k \le r \in \mathbb{N}$. [A function f on $\mathbb{R}$ is *rapidly decreasing* if for all $m \in \mathbb{N}$, then $|f(x)| \le C_m(1+|x|)^{-m}$.]

B.2 Further remarks on function spaces and wavelet coefficients

In Section 9.6, we took an idiosyncratic route, exploring some function spaces on $\mathbb{R}$, then defining Besov sequence norms on $\mathbb{R}$ and finally focusing on Besov sequence and function norms on $[0,1]$. In this section, again without attempting to be comprehensive, we collect some complementary remarks on these topics, and prepare the way for a proof of equivalence of Besov function and sequence norms on $[0,1]$ in the next section.

Spaces on $\mathbb{R}$.

The Besov and Triebel scales of function spaces on $\mathbb{R}^n$ unify many of the classical spaces of analysis. They form the subject of several books, e.g. Frazier et al. (1991); Nikol'skii (1975); Peetre (1975); Triebel (1983, 1992).

Although it is not our main focus, for completeness we give one of the standard definitions. Let ψ be a "window" function of compact support in the frequency domain: assume, say, that supp $\hat{\psi} \subset \{1/2 < |\xi| < 2\}$ and that $|\hat{\psi}| > c$ on $\{3/5 < |\xi| < 5/3\}$.

Given a function f, define "filtered" versions f_j by $\hat{f}_j(\xi) = \hat{\psi}(2^{-j}\xi)\hat{f}(\xi)$: thus $\hat{f}_j(\xi)$ is concentrated on the double octave $|\xi| \in [2^{j-1}, 2^{j+1}]$. For $\alpha \in \mathbb{R}$ and $0 < p, q \leq \infty$, the Besov and Triebel seminorms are respectively defined by

$$|f|_{\dot{B}^{\alpha}_{p,q}} = \Big(\sum_j (2^{\alpha j} \|f_j\|_{L_p})^q \Big)^{1/q}, \qquad |f|_{\dot{F}^{\alpha}_{p,q}} = \Big\| \Big(\sum_j (2^{\alpha j} |f_j|)^q \Big)^{1/q} \Big\|_{L_p},$$

with the usual modifications if $p = \infty$ or $q = \infty$; thus $|f|_{\dot{B}^{\alpha}_{\infty,\infty}} = \sup_j 2^{\alpha j} \|f_j\|_{\infty}$. Thus the Besov norm integrates over location at each scale and then combines over scale, while the *Triebel* seminorm reverses this order. They merge if $p = q$: $\dot{B}^{\alpha}_{p,p} = \dot{F}^{\alpha}_{p,p}$. Despite the importance of the Triebel scale—$F^k_{p,2}$ equals the Sobolev space W^k_p, for example—we will not focus on them here.

These are the "homogeneous" definitions: if $f_t(x) = f(x/t)/t$, then the seminorms satisfy a scaling relation: $\|f_t\|_{\dot{B}} = t^{(1/p)-1-\alpha} \|f\|_{\dot{B}}$. There are only seminorms since they vanish on any polynomial. The "inhomogeneous" versions are defined by bringing in a "low frequency" function φ with the properties that Supp $\hat{\varphi} \subset [-2, 2]$, and $\hat{\varphi} > c$ on $[-5/3, 5/3]$. Then

$$\|f\|_{B^{\alpha}_{p,q}} = \|\varphi \star f\|_{L_p} + \Big(\sum_{j \geq 1} \big(2^{\alpha j} \|f_j\|_{L_p}\big)^q \Big)^{1/q},$$

with a corresponding definition for $\|f\|_{F^{\alpha}_{p,q}}$. These are norms for $1 \leq p, q \leq \infty$, otherwise they are still quasi-norms

Many of the traditional function spaces of analysis (and non-parametric statistics) can be identified as members of either or both of the Besov and Triebel scales. A remarkable table may be found in Frazier et al. (1991); here we mention the Hölder spaces $C^{\alpha} = B^{\alpha}_{\infty,\infty}, \alpha \notin \mathbb{N}$, the Hilbert-Sobolev spaces $W^{\alpha}_2 = B^{\alpha}_{2,2}$ and also the more general Sobolev spaces $W^{\alpha}_p = F^{\alpha}_{p,2}$ for $1 < p < \infty$, with in all cases $\alpha > 0$. If the window function ψ also satisfies the wavelet condition $\sum_j |\hat{\psi}(2^{-j}\xi)|^2 \equiv 1$ a.e., then it is straightforward to verify that $|f|_{\dot{B}^{\alpha}_{2,2}}$ as defined above satisfies

$$|f|_{\dot{B}^{\alpha}_{2,2}} \asymp \int |\xi|^{2\alpha} |\hat{f}(\xi)|^2 d\xi,$$

corresponding with the Fourier domain definition of $\int (D^{\alpha} f)^2$.

[TIDY UP:] Using the Meyer wavelet, Lemarié and Meyer (1986) established, among other things, the equivalence for $\alpha \in \mathbb{R}$ and $1 \leq p, q \leq \infty$. for homogeneous Besov norms. This result is extended to $0 < p, q \leq \infty$ and the Triebel scale by Frazier et al. (1991, Theorem 7.20) After a discussion of numerous particular spaces, the inhomogenous Besov case is written out in Meyer (1990, Volume 1, Chapter VI.10).

If (φ, ψ) have lower regularity (e.g. the Daubechies families of wavelets), then these characterisations hold for restricted ranges of (α, p, q). For example Meyer (1990, I, ch. VI.10), if φ generates an $r-$ regular MRA, then (9.30) holds for $p, q \geq 1, |\alpha| < r$.

B.3 Besov spaces and wavelet coefficients

Let (ϕ, ψ) be an orthonormal scaling and wavelet function pair, complemented with boundary scaling functions and wavelets to yield an orthonormal basis for $L_2[0,1]$:

$$f = \sum_k \beta_k \phi_{Lk} + \sum_{j \geq L} \sum_k \theta_{jk} \psi_{jk}.$$

We have made frequent use of Besov norms on the coefficients $\beta = (\beta_k)$ and $\theta = (\theta_{j\cdot}) = (\theta_{jk})$. To be specific, define

$$\|f\|_{b^\alpha_{p,q}} = \|\beta\|_p + |\theta|_{b^\alpha_{p,q}}, \tag{B.16}$$

where, setting $a = \alpha + 1/2 - 1/p$

$$|\theta|_b^q = |\theta|^q_{b^\alpha_{p,q}} = \sum_{j \geq L} [2^{aj} \|\theta_{j\cdot}\|_p]^q. \tag{B.17}$$

In these definitions, one can take $\alpha \in \mathbb{R}$ and $p, q \in (0, \infty]$ with the usual modification for p or $q = \infty$.

This appendix justifies the term 'Besov norm' by showing that these *sequence* norms are equivalent to standard definitions of Besov norms on *functions* on $L_p(I)$.

We use the term *CDJV multiresolution* to describe the multiresolution analysis of $L_2[0,1]$ resulting from the construction reviewed in Section 7.1. It is based on a Daubechies scaling function φ and wavelet ψ with compact support. If in addition, ψ is C^r—which is guaranteed for sufficiently large S, we say that the MRA is *r-regular.*

The main purpose of this section is to establish the following result.

Theorem B.9 *Let r be a positive integer and suppose that $\{V_j\}$ is a r-regular CDJV multresolution analysis of $L_2[0,1]$. Suppose that $1 \leq p, q \leq \infty$ and $0 < \alpha < r$. Let the Besov function space norm $\|f\|_{B^\alpha_{p,q}}$ be defined by (B.26), and the Besov sequence norm $\|f\|_{b^\alpha_{p,q}}$ by (B.16). Then the two norms are equivalent: there exist constants C_1, C_2 depending on (α, p, q) and the functions (ϕ, ψ), but not on f so that*

$$C_1 \|f\|_{b^\alpha_{p,q}} \leq \|f\|_{B^\alpha_{p,q}} \leq C_2 \|f\|_{b^\alpha_{p,q}}.$$

Equivalences of this type were first described by Lemarié and Meyer (1986) and developed in detail in Meyer (1992, Chapters 6 - 8). for $I = \mathbb{R}$. Their Calderón-Zygmund operator methods make extensive use of the Fourier transform and the translation invariance of $\mathbb{R}$.

The exposition here, however, focuses on a bounded interval, for convenience $[0,1]$, since this is needed for the white noise models of nonparametric regression. On bounded intervals, Fourier tools are less convenient, and our approach is an approximation theoretic one, inspired by Cohen et al. (2000) and DeVore and Lorentz (1993). The survey of nonlinear approximation, DeVore (1998), although more general in coverage than needed here, contains much helpful detail.

The conditions on α, p, q are not the most general. For example, Donoho (1992) develops a class of *interpolating* wavelet transforms using an analog of L_2 multiresolution analysis for continuous functions with coefficients obtained by sampling rather than integration. For

this transform, Besov (and Triebel) equivalence results are established for $0 < p, q \leq \infty$, but with α now in the range $(1/p, r)$.

An encyclopedic coverage of Besov and Triebel function spaces and their characterizations may be found in the books Triebel (1983, 1992, 2006, 2008).

Outline of approach. One classical definition of the Besov function norm uses a modulus of smoothness based on averaged finite differences. We review this first. The modulus of smoothness turns out to be equivalent to the $K-$functional

$$K(f,t) = \inf\{\|f - g\|_p + t\|f^{(r)}\|_p : g \in W_p^r(I)\}$$

which leads to the view of Besov spaces as being *interpolation spaces*, i.e. intermediate between $L_p(I)$ and $W_p(I)$.

The connection between multiresolution analyses $\{V_j\}$ and Besov spaces arises by comparing the $K-$functional at scale 2^{-rk}, namely $K(f, 2^{-rk})$, with the approximation error due to projection onto V_k,

$$e_k(f) = \|f - P_k f\|_p.$$

This comparison is a consequence of two key inequalities. The 'direct' or 'Jackson' inequality, Corollary B.17 below, bounds the approximation error in terms of the rth derivative

$$\|f - P_k f\|_p \leq C 2^{-rk} \|f^{(r)}\|_p.$$

Its proof uses bounds on kernel approximation, along with the key property that each V_j contains $\mathcal{P}_{r-1}$. The 'inverse' or 'Bernstein' inequality, Lemma B.19 below, bounds derivatives of $g \in V_k$:

$$\|g^{(r)}\|_p \leq C 2^{rk} \|g\|_p.$$

DeVore (1998) has more on the role of Jackson and Bernstein inequalities.

From this point, it is relatively straightforward to relate the approximation errors $e_k(f)$ with the wavelet coefficient norms (B.17). The steps are collected in the final equivalence result, Theorem B.22.

Moduli of smoothness and Besov spaces

This section sets out one of the classical definitions of Besov spaces, based on moduli of smoothness, and drawing on DeVore and Lorentz (1993), which contains a wealth of extra material. For more on the extensive literature on Besov spaces and the many equivalent definitions, see Peetre (1975); Triebel (1983, 1992). An expository account, limited to $\mathbb{R}$ and $0 < \alpha < 1$ is Wojtaszczyk (1997).

The definition does not explicitly use derivatives; instead it is built up from averages, in the L_p sense, of approximate derivatives given by finite differences. For L_p norms restricted to an interval A, write

$$\|f\|_p(A) = \Big(\int_A |f(x)|^p dx\Big)^{1/p},$$

and, as usual, $\|f\|_\infty(A) = \sup_{x \in A} |f(x)|$.

Let $T_h f(x) = f(x+h)$ denote translation by h. The first difference of a function is

$$\Delta_h(f,x) = f(x+h) - f(x) = (T_h - I)f(x).$$

Higher order differences, for $r \in \mathbb{N}$, are given by

$$\Delta_h^r(f,x) = (T_h - I)^r f(x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(x+kh). \tag{B.18}$$

To describe sets over which averages of differences can be computed, we need the (one sided) erosion of A: set $A_h = \{x \in A \,:\, x+h \in A\}$. The main example: if $A = [a,b]$, then $A_h = [a, b-h]$. The r^{th} integral modulus of smoothness of $f \in L_p(A)$ is then

$$\omega_r(f,t)_p = \sup_{0 \le h \le t} \|\Delta_h^r(f,\cdot)\|_p(A_{rh}).$$

For $p < \infty$, this is a measure of smoothness averaged over A; the supremum ensures monotonicity in t. If $p = \infty$, it is a uniform measure of smoothness, for example

$$\omega_1(f,t)_\infty = \sup\{|f(x) - f(y)|,\ x, y \in A, |x-y| \le r\}.$$

The differences $\Delta_h^r(f,x)$ are linear in f, and so for $p \ge 1$, there is a triangle inequality

$$\omega_r(f+g,t)_p \le \omega_r(f,t)_p + \omega_r(g,t)_p. \tag{B.19}$$

Again by linearity, $\|\Delta_h^r(f,\cdot)\|_p \le 2^r \|f\|_p$ and so also

$$\omega_r(f,t)_p \le 2^r \|f\|_p, \tag{B.20}$$

and more generally, for $0 \le k \le r$,

$$\omega_r(f,t)_p \le 2^{r-k} \omega_k(f,t)_p. \tag{B.21}$$

For $n \in \mathbb{N}$ and $1 \le p \le \infty$ it can be verified that

$$\omega_r(f,nt)_p \le n^r \omega_r(f,t)_p. \tag{B.22}$$

When derivatives exist, the finite difference can be expressed as a kernel smooth of bandwidth h of these derivatives:

Lemma B.10 *Let χ be the indicator of the unit interval $[0,1]$, and $\chi^{\star r}$ be its r^{th} convolution power. Then, for $f \in W_p^r$,*

$$\Delta_h^r(f,x) = h^r \int f^{(r)}(x+hu)\chi^{\star r}(u)du. \tag{B.23}$$

The easy proof uses induction and the fact that differentiation commutes with Δ_h^{r-1}. A simple consequence of (B.23) is the bound

$$\omega_r(f,t)_p \le t^r |f|_{W_p^r(I)}, \tag{B.24}$$

valid for all $t \ge 0$. Indeed, rewrite the right side of (B.23) as $h^r \int K(x,v) f^{(r)}(v)dv$, using the kernel

$$K(x,v) = h^{-1}\chi^{\star r}(h^{-1}(v-x))$$

for $x \in I_h$ and $v = x + hu \in I$. Now apply Young's inequality (C.22), which says that the

operator with kernel K is bounded on L_p. Note that both M_1 and $M_2 \leq 1$ since $\chi^{\star r}$ is a probability density, so that the norm of K is at most one. Hence

$$\|\Delta_h^r(f,\cdot)\|_p(I_{rh}) \leq h^r |f|_{W_p^r(I)},$$

and the result follows from the definition of ω_r.

B.11 *Uniform smoothness.* There are two ways to define uniform smoothness using moduli. Consider $0 < \alpha \leq 1$. The first is the usual Hölder/Lipschitz definition

$$|f|_{\mathrm{Lip}(\alpha)} = \sup_{t>0} t^{-\alpha}\omega_1(f,t)_\infty,$$

which is the same as (C.19). The second replaces the first-order difference by one of (possibly) higher order. Let $r = [\alpha] + 1$ denote the smallest integer larger than α and put

$$|f|_{\mathrm{Lip}^*(\alpha)} = \sup_{t>0} t^{-\alpha}\omega_r(f,t)_\infty.$$

Clearly these coincide when $0 < \alpha < 1$. When $\alpha = 1$, however, $\mathrm{Lip}^*(1) = Z$ is the Zygmund space, and

$$\|f\|_{\mathrm{Lip}^*(1)} = \|f\|_\infty + \sup_{x,x\pm h\in A} \frac{|f(x+h) - 2f(x) + f(x+h)|}{h}.$$

It can be shown (e.g. DeVore and Lorentz (1993, p. 52)) that $\mathrm{Lip}^*(1) \supset \mathrm{Lip}(1)$ and that the containment is proper, using the classical example $f(x) = x \log x$ on $[0,1]$.

Besov spaces. Let $\alpha > 0$ and $r = \lfloor \alpha \rfloor + 1$. Let $A = \mathbb{R}$, or an interval $[a,b]$. The Besov space $B_{p,q}^\alpha(A)$ is the collection of $f \in L_p(A)$ for which the seminorm

$$|f|_{B_{p,q}^\alpha} = \Big(\int_0^\infty \Big[\frac{\omega_r(f,t)_p}{t^\alpha} \Big]^q \frac{dt}{t} \Big)^{1/q} \tag{B.25}$$

is finite. If $q = \infty$, we use $|f|_{B_{p,\infty}^\alpha} = \sup_t \; t^{-\alpha}\omega_r(f,t)_p$. The seminorm vanishes if f is a polynomial of degree less than r. As norm on $B_{p,q}^\alpha(A)$, we take

$$\|f\|_{B_{p,q}^\alpha} = \|f\|_p + |f|_{B_{p,q}^\alpha}. \tag{B.26}$$

If $p = q = \infty$ and $\alpha < 1$, so that $r = 1$, we recover the $\mathrm{Lip}(\alpha)$ or Hölder-α seminorm. If $\alpha = 1$, then $r = 2$ and $B_{\infty,\infty}^1$ is the Zygmund space.

A simple inequality between Besov and Sobolev norms states that for $m \in \mathbb{N}$,

$$|f|_{B_{p,\infty}^m} \leq C \int_I |D^m f|^p.$$

Indeed, take $r = m + 1$ in the definition of the $B_{p,\infty}^m$ norm, then apply (B.21) and (B.24) to get

$$\omega_{m+1}(f,t)_p \leq 2\omega_m(f,t)_p \leq 2t^m |f|_{W_p^m}$$

so that $|f|_{B_{p,\infty}^m} \leq 2|f|_{W_p^m}$ as required.

Remarks. 1. The assumption that $r > \alpha$ is used in at least two places in the equivalence arguments to follow: first in the interpolation space identification of $B_{p,q}^\alpha$, Theorem B.12, and second in Theorem B.20 relating approximation error to the K-functional. This indicates

why it is the Zygmund space—and more generally $\mathrm{Lip}^*(\alpha)$—that appears in the wavelet characterizations of $B^\alpha_{\infty,\infty}$ for integer α, rather than the traditional C^α spaces.

2. The modulus based definition is equivalent (on $\mathbb{R}^n$) to the earlier Fourier form if $\alpha > n(p^{-1}-1)_+, 0 < p,q \le \infty$, (e.g. Triebel (1983, p. 110), [For $\alpha > 0, 1 \le p,q \le \infty$, see also Bergh and Löfström (1976, Th. 6.2.5))].

Besov spaces as interpolation spaces

This section shows that Besov spaces are *intermediate* spaces between $L_p(I)$ and $W^r_p(I)$. First we need the notion of *K-functional*, reminiscent of roughness penalized approximations in the theory of splines:

$$K(f,t) = K(f,t; L_p, W^r_p) = \inf\{\|f-g\|_p + t\|D^r g\|_p : g \in W^r_p\}.$$

The main fact about $K(f,t)$ for us is that it is equivalent to the r^{th} modulus of smoothness $\omega_r(f,t)_p$ – see Theorem B.13 below.

First some elementary remarks about $K(f,t)$. Since smooth functions are dense in L_p, it is clear that $K(f,0) = 0$. But $K(f,t)$ vanishes for all $t > 0$ if and only if f is a polynomial of degree at most $r-1$. Since K is the pointwise infimum of a collection of increasing linear functions, it is itself increasing and concave. Further, for any f

$$K(f,t) \ge \min(t,1)K(f,1), \tag{B.27}$$

while if $f \in W^r_p$ then by choosing g equal to f or 0 as $t \le 1$ or $t > 1$,

$$K(f,t) \le \min(t,1)\|f\|_{W^r_p}. \tag{B.28}$$

A sort of converse to (B.27) will be useful. We first state a result which it is convenient to prove later, after Proposition B.16. Given $g \in W^r_p$, let $\Pi_{r-1}g$ be the best (in $L_2(I)$) polynomial approximation of $r-1$ to g. Then for $C = C(I,r)$,

$$\|g - \Pi_{r-1}g\|_p \le C\|g^{(r)}\|_p. \tag{B.29}$$

Now, let $f \in L_p$ and $g \in W^r_p$ be given. From the definition of K and (B.29),

$$\begin{aligned} K(f,t) \le \|f - \Pi_{r-1}g\|_p &\le \|f-g\|_p + \|g - \Pi_{r-1}g\|_p \\ &\le \|f-g\|_p + C\|g^{(r)}\|_p, \end{aligned}$$

where $C = C(I,r)$. Hence, for all $t \ge a$,

$$K(f,t) \le \max(Ca^{-1}, 1)K(f,a). \tag{B.30}$$

The $K-$function $K(f,t)$ trades off between L_p and W^r_p at scale t. Information across scales can be combined via various weighting functions by defining, for $0 < \theta < 1$,

$$\rho(f)_{\theta,q} = \Big(\int_0^\infty \Big[\frac{K(f,t)}{t^\theta}\Big]^q \frac{dt}{t}\Big)^{1/q} \qquad 0 < q < \infty \tag{B.31}$$

and, when $q = \infty$, $\rho(f)_{\theta,\infty} = \sup_{0\le t\le\infty} t^{-\theta}K(f,t)$.

Replacing $K(f,t)$ by $\min(1,t)$ in the integral (B.31) leads to the sum of two integrals

$\int_0^1 t^{(1-\theta)q-1}dt$ and $\int_1^\infty t^{-\theta q-1}dt$, which both converge if and only if $0 < \theta < 1$. Hence property (B.27) shows that in order for $\rho(f)_{\theta,q}$ to be finite for any f other than polynomials, it is necessary that $0 < \theta < 1$.

On the other hand, property (B.28) shows that

$$\rho(f)_{\theta,q} \le c_{\theta q}\|f\|_{W_p^r}. \tag{B.32}$$

We therefore define intermediate, or *interpolation* spaces

$$X_{\theta,q} = (L_p, W_p^r)_{\theta,q} = \{f \in L_p : \rho(f)_{\theta,q} < \infty\}$$

for $0 < q \le \infty$ and $0 < \theta < 1$, and set $\|f\|_{X_{\theta,q}} = \|f\|_p + \rho(f)_{\theta,q}$.

From the definition and (B.32),

$$W_p^r \subset (L_p, W_p^r)_{\theta,q} \subset L_p.$$

The parameters (θ, q) yield a lexicographic ordering:

$$X_{\theta_1,q_1} \subset X_{\theta_2,q_2} \quad \text{if } \theta_1 > \theta_2, \text{ or if } \theta_1 = \theta_2 \text{ and } q_1 \le q_2.$$

The main reason for introducing interpolation spaces here is that they are in fact Besov spaces.

Theorem B.12 *For $r \in \mathbb{N}$, and $1 \le p \le \infty$, $0 < q \le \infty$, $0 < \alpha < r$,*

$$(L_p, W_p^r)_{\alpha/r,q} = B_{p,q}^\alpha.$$

This follows from the definitions and the next key theorem, which shows that the $K-$function is equivalent to the integral modulus of continuity.

Theorem B.13 (Johnen, ref) *Let $A = \mathbb{R}, \mathbb{R}_+, \mathbb{T}$ or $[0,1]$. For $1 \le p \le \infty$, and $r \in \mathbb{N}$, there exist $C_1, C_2 > 0$ depending only on r, such that for all $f \in L_p$,*

$$C_1\omega_r(f,t)_p \le K(f,t^r; L_p, W_p^r) \le C_2\omega_r(f,t)_p, \qquad t > 0. \tag{B.33}$$

Proof We work on the left inequality first: from the triangle inequality (B.19) followed by (B.20) and derivative bound (B.24), we have for arbitrary g,

$$\begin{aligned}\omega_r(f,t)_p &\le \omega_r(f-g,t)_p + \omega_r(g,t)_p \\ &\le 2^r\|f-g\|_p + t^r|g|_{W_p^r}.\end{aligned}$$

Minimizing over g, we obtain the left inequality in (B.33) with $C = 2^r$.

For the right inequality, we only give full details for $A = \mathbb{R}$. Given f, we choose

$$g(x) = f(x) + (-1)^{r+1}\int \Delta_{ut}^r(f,x)\chi^{\star r}(u)du. \tag{B.34}$$

By the Minkowski integral inequality (C.26),

$$\|g-f\|_p \le \int \|\Delta_{ut}^r(f,\cdot)\|_p\chi^{\star r}(u)du \le \omega_r(f,rt)_p \le r^r\omega_r(f,t)_p, \tag{B.35}$$

where the second inequality follows because $\chi^{\star r}$ is a probability density supported on $[0,r]$, and the third uses (B.22).

Now estimate $\|g^{(r)}\|_p$. Differentiate (B.34) with expansion (B.18) for $\Delta^r_{tu}(f,x)$ to get

$$g^{(r)}(x) = \sum_{k=1}^{r} \binom{r}{k} (-1)^{k+1} \int f^{(r)}(x + ktu)\chi^{\star r}(u)du$$
$$= \sum_{k=1}^{r} \binom{r}{k} (-1)^{k+1} (kt)^{-r} \Delta^r_{kt}(f,x),$$

where the second equality uses identity (B.23). Again using (B.22), we find

$$t^r \|g^{(r)}\|_p \le \sum_{k=1}^{r} \binom{r}{k} k^{-r} \omega_r(f,kt)_p \le 2^r \omega_r(f,t)_p.$$

Putting this last inequality and (B.35) into the definition of $K(f,t^r)$ yields the right hand bound.

If $A = [0,1]$, then g is defined in (B.34) for $x \in I_1 = [0,3/4]$ if $t \le 1/4r^2$. By symmetry, one can make an analogous definition and argument for $I_2 = [1/4,1]$. One patches together the two subinterval results, and takes care separately of $t > 1/4r^2$. For details see (DeVore and Lorentz, 1993, p. 176, 178). □

For work with wavelet coefficients, we need a discretized version of these measures.

Lemma B.14 *Let $L \in \mathbb{N}$ be fixed. With constants of proportionality depending on I, r, θ, q and L but not on f,*

$$\rho(f)^q_{\theta,q} \asymp \sum_{j=L}^{\infty} [2^{\theta r j} K(f, 2^{-rj})]^q. \tag{B.36}$$

Proof Since $K(f,t)$ is concave in t with $K(f,0) = 0$, we have $\epsilon K(f,t) \le K(f,\epsilon t)$, and since it is increasing in t, we have for $2^{-r(j+1)} \le t \le 2^{-rj}$,

$$2^{-r} K(f,2^{-rj}) \le K(f,2^{-r(j+1)}) \le K(f,t) \le K(f,2^{-rj}).$$

From this it is immediate that, with $a = 2^{-rL}$, the sum $S_L(f)$ in (B.36) satisfies

$$S_L(f) \asymp \int_0^a \Big[\frac{K(f,t)}{t^\theta}\Big]^q \frac{dt}{t}$$

with constants of proportionality depending only on (θ, q, r). From (B.30),

$$\int_a^\infty \left[\frac{K(f,t)}{t^\theta}\right]^q \frac{dt}{t} \le C[K(f,a)a^{-\theta}]^q$$

where C depends on (I, L, r, θ, q). With $a = 2^{-rL}$, this last term can be absorbed in the sum $S_L(f)$, completing the proof. □

MRAs on $[0,1]$

We use the term *CDJV multiresolution* to describe the multiresolution analysis of $L_2[0,1]$ resulting from the construction reviewed in Section 7.1. It is based on a scaling function φ and wavelet ψ with support in $[-S+1, S]$ and for which ψ has S vanishing moments.

The MRA of $L_2[0,1]$ is constructed using S left and S right boundary scaling functions $\varphi_k^L, \varphi_k^R, k = 0, \ldots S-1$.

Choose a coarse level L so that $2^L \geq 2S$. For $j \geq L$, we obtain scaling function spaces $V_j = \text{span}\{\varphi_{jk}\}$ of dimension 2^j. The orthogonal projection operators $P_j : L_2(I) \to V_j$ have associated kernels

$$E_j(x,y) = \sum_k \phi_{jk}(x)\phi_{jk}(y),$$

as may be seen by writing

$$P_j f(x) = \sum_k \langle f, \phi_{jk}\rangle \phi_{jk}(x) = \int \sum_k \phi_{jk}(x)\phi_{jk}(y) f(y) dy.$$

If in addition, ψ is C^r—which is guaranteed for sufficiently large S, we say that the MRA is *r-regular*. Since ψ is C^r it follows (e.g. by Daubechies (1992, Corollary 5.5.2)) that ψ has r vanishing moments. The CDJV construction then ensures that $\mathcal{P}_{r-1}$, the space of polynomials of degree $r-1$ on $[0,1]$ is contained in V_L. In fact, we abuse notation and write $V_{L-1} = \mathcal{P}_{r-1}$. The corresponding orthogonal projection operator $P_{L-1} : L_2(I) \to V_{L-1}$ has kernel

$$\Pi_{r-1}(x,y) = \sum_{k=0}^{r-1} p_k(x) p_k(y) \qquad x, y \in I. \tag{B.37}$$

Here $p_k(x) = \sqrt{2k+1} P_k(2x-1)$ are orthonormal on $L_2[0,1]$ with $P_k(x)$ being the classical Legendre polynomial of degree k on $[-1,1]$ (Szegö, 1967).

A simple fact for later use is that P_j have uniformly bounded norms on $L_p[0,1]$. To state it, define

$$a_q(\varphi) = \max\{\|\varphi\|_q, \|\varphi_k^L\|_q, \|\varphi_k^R\|_q, k = 0, \ldots, S-1\}.$$

Lemma B.15 *Suppose that $\{V_j\}$ is a CDJV multresolution analysis of $L_2[0,1]$. Then for $1 \leq p \leq \infty$,*

$$\|P_j\|_p \leq 2S a_1(\varphi) a_\infty(\varphi), \tag{B.38}$$

$$\|P_{L-1}\|_p \leq C(r). \tag{B.39}$$

Proof We simply apply Young's inequality (C.22). For $j \geq L$, we need the bounds

$$\sum_k |\varphi_{jk}(x)| \leq 2S 2^{j/2} a_\infty(\varphi), \qquad \int |\varphi_{jk}(y)| dy \leq 2^{-j/2} a_1(\varphi)$$

from which it follows that $\int |E_j(x,y)| dy \leq 2S a_1(\varphi) a_\infty(\varphi)$ and similarly for $\int |E_j(x,y)| dx$. We argue similarly for $j = L-1$ using the bounds

$$\sum_{k=0}^{r-1} |p_k(x)| \leq C r^{3/2}, \qquad \int |p_k(y)| dy \leq 1. \qquad \square$$

With the addition of boundary wavelets $\psi_k^L, \psi_k^R, k = 0, \ldots, S-1$, one obtains detail spaces $W_j = \text{span}\{\psi_{jk}, k = 0, \ldots, 2^j - 1\}$ and the decomposition

$$L_2[0,1] = V_L \oplus \oplus_{j \geq L} W_j.$$

Approximation Properties of Kernels and MRA's

We first look at the approximation power of a family of kernels $K_h(x, y)$. Let $I \subset \mathbb{R}$ be an interval – typically $I = [0, 1]$ or $\mathbb{R}$ itself. Define

$$K_h f(x) = \int_I K_h(x, y) f(y) dy \qquad x \in I.$$

In the proof to follow, $\|f\|_p = (\int_I |f|^p)^{1/p}$ is the L_p norm on I.

Proposition B.16 *Suppose that the kernel $K_h(x, y)$ satisfies*

$$\begin{aligned} &(i) \quad K_h \pi = \pi \quad \text{for } \pi \in \mathcal{P}_{r-1}, \\ &(ii) \quad K_h(x, y) = 0 \quad \text{if } |y - x| > Lh, \\ &(iii) \quad |K_h(x, y)| \le M h^{-1}. \end{aligned}$$

on an interval $I \subset \mathbb{R}$. For $p \ge 1$, there exists a constant $C = C(L, M, r)$ such that for $f \in W_p^r(I)$,

$$\|f - K_h f\|_p \le C h^r \|D^r f\|_p, \qquad h > 0.$$

The key requirement is that K_h preserve polynomials of degree at most $r-1$. Assumption (ii) could be weakened to require sufficient decay of K_h as $|x - y|$ grows.

Proof A function $f \in W_p^r(I)$ has continuous derivatives of order $k = 0, 1, \ldots, r-1$. If $x \in I$, we may therefore use the Taylor approximation to f at x by a polynomial π_x of degree $r-1$, so that $f(y) = \pi_x(y) + R_x(y)$ with the integral form of the remainder term

$$R_x(y) = c_{r-1} \int_x^y (D^r f)(u)(y - u)^{r-1} du, \qquad c_{r-1} = 1/(r-1)!$$

Since K_h leaves such polynomials invariant, $K_h f = \pi_x + K_h R_x$, and since $\pi_x(x) = f(x)$,

$$\begin{aligned} (K_h f)(x) - f(x) &= \int_I K_h(x, y) R_x(y) dy \\ &= c_{r-1} \int_I K_h(x, y) \int_x^y (y - u)^{r-1} f^{(r)}(u) du dy \\ &= \int_I \tilde{K}_h(x, u) f^{(r)}(u) du, \end{aligned}$$

where $\tilde{K}_h(x, u)$ is a new kernel on $I \times I$, about which we need only know a bound, easily derived from the above, along with conditions (ii) and (iii):

$$|\tilde{K}_h(x, u)| \le \begin{cases} c M h^{-1} (Lh)^r & \text{if } |x - u| \le Lh \\ 0 & \text{otherwise .} \end{cases}$$

Since $\int_I |\tilde{K}_h(x, u)| du \le 2cL^{r+1} M h^r$, with a similar bound for the corresponding integral over $x \in I$, our result follows from Young's inequality (C.22) with $M_1 = M_2 = 2cL^{r+1} M h^r$. □

A common special case occurs when $K_h(x, y) = h^{-1} K(h^{-1}(x - y))$ is a scaled translation invariant kernel on $\mathbb{R}$. Condition (i) is equivalent to the vanishing moment property

$\int t^k K(t)dt = \delta_{k0}$ for $k = 0, 1, \ldots, r-1$. If $K(y)$ is bounded and has compact support, then properties (ii) and (iii) are immediate.

As a second example, consider orthogonal polynomials on $I = [0, 1]$ and the associated kernel $\Pi_{r-1}(x, y)$ given in (B.37). Assumptions (i) - (ii) hold for $h = L = 1$. The bound $|P_n(x)| \le 1$ on $[-1, 1]$ shows that (iii) holds with $M = r^2$. Consequently, for $f \in W_p^r(I)$ and setting $C = 2r/(r-1)!$, we obtain $\|f - \Pi_{r-1} f\|_p \le C\|f^{(r)}\|_p$, which is just (B.29).

Our main application of Proposition B.16 is to multiresolution analyses.

Corollary B.17 *Suppose that $\{V_j\}$ is a CDJV multresolution analysis of $L_2[0, 1]$. Let P_j be the associated orthogonal projection onto V_j, and assume that $2^j \ge 2S$. Then there exists a constant $C = C(\varphi)$ such that for all $f \in W_p^r(I)$,*

$$\|f - P_j f\|_p \le C 2^{-rj} |f|_{W_p^r}.$$

Proof We claim that assumptions (i)-(iii) hold for the kernel E_j with h taken as 2^{-j}. The CDJV construction guarantees that $\mathcal{P}_{j-1} \subset V_j$ so that (i) holds. In addition the construction implies that (ii) holds with $L = 2S$ and that

$$\#\{k : \varphi_{jk}(x)\varphi_{jk}(y) \ne 0\} \le 2S.$$

It follows that (iii) holds with $M = 2pa_\infty^2(\varphi)$. □

Bernstein-type Inequalities

First a lemma, inspired by Meyer (1990, p.30), which explains the occurence of terms like $2^{j(1/2-1/p)}$ in sequence norms.

Lemma B.18 *Let $\{\gamma_{jk}, k \in K\}$ be an orthonormal sequence of functions satisfying*

$$\text{(i)} \qquad \sum_k |\gamma_{jk}(x)| \le b_\infty 2^{j/2}, \qquad \text{and}$$

$$\text{(ii)} \qquad \max_k \int |\gamma_{jk}| \le b_1 2^{-j/2}.$$

Then for all $1 \le p \le \infty$, and any sequence $\lambda = (\lambda_k, k \in K)$,

$$C_1 2^{j(1/2-1/p)} \|\lambda\|_p \le \Big\| \sum_k \lambda_k \gamma_{jk} \Big\|_p \le C_2 2^{j(1/2-1/p)} \|\lambda\|_p. \tag{B.40}$$

Here $C_1 = b_1^{-1}(b_1/b_\infty)^{1/p}$ and $C_2 = b_\infty(b_1/b_\infty)^{1/p}$.

Remarks. 1. If ϕ is an orthonormal scaling function and $\gamma_{jk}(x) = 2^{j/2}\phi(2^j x - k)$ for $k \in \mathbb{Z}$, and $|\operatorname{supp}\phi| \le B$, then (i) and (ii) are trivially satisfied with $b_\infty = B\|\phi\|_\infty$ and $b_1 = \|\phi\|_1$.

2. If $\{\gamma_{jk}\} = \{\phi_{jk}\}$ correspond to a CDJV boundary MRA for $[0, 1]$ derived from a scaling function ϕ with $\operatorname{supp}\phi \subset [-S+1, S]$, then (i) and (ii) hold with $b_\infty = 2Sa_\infty(\varphi)$ and $b_1 = a_1(\varphi)$. Analogous remarks apply with wavelets, when $\{\gamma_{jk}\} = \{\psi_{jk}\}$.

3. The right side in (B.40) does not require the assumption of orthonormality for $\{\gamma_{jk}\}$.

Proof This is just the extended Young inequality, Theorem C.17. Identify $\mu(dx)$ with Lebesgue measure on $\mathbb{R}$ and $\nu(dy)$ with counting measure on $k \in K$. Then match $K(x, y)$ with $\gamma_{jk}(x)$ and $f(y)$ with λ_k. Conditions (i) and (ii) imply that $M_1 = a_1 2^{-j/2}$ and $M_2 = a_\infty 2^{j/2}$ suffice for the conditions of the theorem. The right hand inequality above now follows from (C.22). Note that orthonormality of $\{\gamma_{jk}\}$ is not used.

For the left hand inequality, assume that $g(x) = \sum_k \lambda_k \gamma_{jk}$. Since the $\{\gamma_{jk}\}$ are orthonormal,

$$(K^* g)_k = \int \gamma_{jk}(x) g(x) dx = \lambda_k$$

and now the result follows from the adjoint form (C.23) of Young's inequality. □

Now to the variant of the Bernstein inequality that we need. We now require ψ to be C^r.

Lemma B.19 *Suppose that $\{V_j\}$ is a r-regular CDJV multresolution analysis of $L_2[0, 1]$. For $g \in V_j$ and $1 \le p \le \infty$,*

$$\|D^r g\|_p \le c 2^{jr} \|g\|_p.$$

Proof Since $g \in V_j$, it has an expansion $g = \sum \lambda_k \phi_{jk}$, and so

$$D^r g = \sum \lambda_k D^r \phi_{jk} = 2^{jr} \sum \lambda_k \gamma_{jk},$$

where the functions γ_{jk} are formed from the finite set $\{D^r\phi, D^r\phi_k^0, D^r\phi_k^1\}$ by exactly the same set of linear operations as used to form ϕ_{jk} from the set $\{\phi, \phi_k^0, \phi_k^1\}$.

Since the $\{\phi_{jk}\}$ system satisfy the conditions (i) and (ii) of Lemma B.18, the same is true of the $\{\gamma_{jk}\}$ system. From the right side of that Lemma,

$$\|D^r g\|_p = 2^{jr} \|\sum \lambda_k \gamma_{jk}\|_p \le c_2 2^{jr} 2^{j(1/2-1/p)} \|\lambda\|_p.$$

Now apply the left side of the same Lemma to the (orthogonal!) $\{\phi_{jk}\}$ system to get

$$\|D^r g\|_p \le c_2 c_1 2^{jr} \|\sum \lambda_k \phi_{jk}\|_p = c_2 c_1 2^{jr} \|g\|_p. \qquad \square$$

Approximation Spaces and Besov Spaces

This section relates the approximation properties of a multiresolution analysis to the behaviour of the $K-$functional near 0. Specifically, let the approximation error of a function $f \in W_p^r(I)$ by its orthogonal projection $P_k f$ onto the space V_k be given by

$$e_k(f) = \|f - P_k f\|_p.$$

We will show that the rate of decay of $e_k(f)$ is comparable to that of $K(f, 2^{-rk})$, using the Jackson and Bernstein inequalities, Corollary B.17 and Lemma B.19 respectively. In order to handle low frequency terms, we use the notation V_{L-1} to refer to the space of polynomials of degree at most $r-1$, and adjoin it to the spaces $V_k, k \ge L$ of the multiresolution analysis.

Theorem B.20 *Suppose that $\{V_j\}$ is a r-regular CDJV multresolution analysis of $L_2[0, 1]$.*

Let $r \in \mathbb{N}$ be given. For $1 \le p \le \infty, 0 < q < \infty$ and $0 < \alpha < r$. With constants depending on $(\alpha, p, q, r, \varphi, \psi)$, but not on f, we have

$$\sum_{L-1}^{\infty} [2^{\alpha k} e_k(f)]^q \asymp \sum_{L-1}^{\infty} [2^{\alpha k} K(f, 2^{-rk})]^q. \tag{B.41}$$

Proof 1°. The main work is to show that for $k \ge L-1$

$$C_1 e_k(f) \le K(f, 2^{-kr}) \le C_2 \sum_{j=L-1}^{k} 2^{-(k-j)r} e_j(f). \tag{B.42}$$

For the left hand inequality, let $f \in L_p$ and $g \in W_p^r$ be fixed. Write $f - P_k f$ as the sum of $(I - P_k)(f - g)$ and $g - P_k g$, so that

$$e_k(f) \le \|(I - P_k)(f - g)\|_p + e_k(g).$$

It follows from (B.38) that $\|I - P_k\|_p \le 1 + a_1 a_\infty$. Together with Jackson inequality Corollary B.17 for $k \ge L$ and (B.29) for $k = L-1$, this yields

$$e_k(f) \le C[\|f - g\|_p + 2^{-rk} |g|_{W_p^r}].$$

Minimizing now over g yields the left side of (B.42).

For the right inequality, set $\psi_j = P_j f - P_{j-1} f \in V_j$ and write $P_k f = \sum_{j=L}^{k} \psi_j + P_{L-1} f$. Now $P_{L-1} f$ is a polynomial of degree at most $r-1$, so $|P_{L-1} f|_{W_p^r} = 0$. For the other terms, apply the Bernstein inequality Lemma B.19 to obtain

$$|P_k f|_{W_p^r} \le \sum_{j=L}^{k} |\psi_j|_{W_p^r} \le c_2 \sum_{L}^{k} 2^{rj} \|\psi_j\|_p \le c_2 \sum_{L}^{k} 2^{rj} [e_{j-1}(f) + e_j(f)].$$

Finally, put this into the $K-$function definition:

$$\begin{aligned} K(f, 2^{-kr}) &\le \|f - P_k f\|_p + 2^{-kr} |P_k f|_{W_p^r} \\ &\le (1 + 2^{r+1} c_2) \sum_{j=L-1}^{k} 2^{-(k-j)r} e_j(f). \end{aligned}$$

2°. The left to right bound in (B.41) is immediate from (B.42). For the other inequality, let $b_k = 2^{\alpha k} e_k(f)$ and $c_k = 2^{\alpha k} K(f, 2^{-rk})$ for $k \ge L-1$ and 0 otherwise. Then bound (B.42) says that $c_k \le \sum_{j=L-1}^{\infty} a_{k-j} b_j$ for $k \ge L-1$, where $a_k = C_2 2^{-k(r-\alpha)} I\{k \ge 0\}$. Our bound $\|c\|_q \le c_{r\alpha} C_2 \|b\|_q$ now follows from Young's inequality (C.25). □

Wavelet coefficients, finally

The last step in this chain is now quite easy, namely to relate seminorms on wavelet coefficients to those on approximation errors. Let Q_j be orthogonal projection onto the details space W_j, thus $Q_j = P_{j+1} - P_j$. Suppose that for fixed j, $\{\psi_{jk}\}$ is the orthonormal basis for W_j so that

$$Q_j f = \sum_k \theta_{jk} \psi_{jk}, \qquad \theta_{jk} = \langle f, \psi_{jk} \rangle.$$

Let $\|\theta_{j\cdot}\|_p$ denote the ℓ_p-norm of (θ_{jk}), and $a = \alpha + 1/2 - 1/p$.

Lemma B.21 *For $\alpha > 0$ and $1 \le p \le \infty$, and an $r-$regular CDJV multiresolution analysis of $L_2[0,1]$,*

$$\sum_{j\ge L}[2^{\alpha j}\|Q_j f\|_p]^q \asymp \sum_{j\ge L}[2^{aj}\|\theta_{j\cdot}\|_p]^q \asymp \sum_{j\ge L}[2^{\alpha j}e_j(f)]^q$$

Proof The first equivalence follows from Lemma B.18 and the Remark 2 following it:

$$\|Q_j f\|_p \asymp 2^{j(1/2-1/p)}\|\theta_{j\cdot}\|_p, \tag{B.43}$$

For the second equivalence, let $\delta_k = \|Q_k f\|_p$ and $e_k = e_k(f) = \|f - P_k f\|_p$. Clearly $\delta_k \le e_k + e_{k+1}$, which suffices for one of the inequalities. On the other hand, $f - P_j f = \sum_{k\ge j} Q_k f$, and so $e_j \le \sum_{k\ge j}\delta_k$, or equivalently

$$2^{\alpha j}e_j \le \sum_{k\ge j} 2^{-\alpha(k-j)}2^{\alpha k}\delta_k.$$

The other inequality now follows from Young's inequality (C.25). □

Remark. The same argument as for (B.43) applies also to the projection onto V_L, given by $P_L f = \sum_k \beta_k \phi_{Lk}$ to show that, with $\beta = (\beta_k)$,

$$\|P_L f\| \asymp 2^{L(1/2-1/p)}\|\beta\|_p. \tag{B.44}$$

Summary: norm equivalence

We assemble the steps carried out in earlier sections.

Theorem B.22 *Let r be a positive integer and suppose that $\{V_j\}$ is a r-regular CDJV multresolution analysis of $L_2[0,1]$. Suppose that $1 \le p,q \le \infty$ and $0 < \alpha < r$. Let the Besov function space norm $\|f\|_{B^\alpha_{p,q}}$ be defined by (B.26), and the Besov sequence norm $\|f\|_{b^\alpha_{p,q}}$ by (B.16). Then the two norms are equivalent: there exist constants C_1, C_2 depending on (α, p, q) and the functions (ϕ, ψ), but not on f so that*

$$C_1\|f\|_{b^\alpha_{p,q}} \le \|f\|_{B^\alpha_{p,q}} \le C_2\|f\|_{b^\alpha_{p,q}}.$$

Proof We combine the definition of the Besov seminorm (B.25), the equivalence of modulus and $K-$functional (B.33) (with $s = t^r$ and $\theta = \alpha/r$), the dyadic discretization (B.36) and the $(\alpha, q)-$equivalence of $K-$functional and MRA-approximation errors (B.41) to find

$$\begin{aligned}
|f|^q_{B^\alpha_{p,q}} &= \int_0^\infty \left[\frac{\omega_r(f,t)_p}{t^\alpha}\right]^q \frac{dt}{t} \\
&\asymp \int_0^\infty \left[\frac{K(f,s)}{s^\theta}\right]^q \frac{ds}{s} \\
&\asymp \sum_{j\ge L-1}[2^{\alpha j}K(f,2^{-rj})]^q \\
&\asymp \sum_{j\ge L-1}[2^{\alpha j}e_j(f)]^q
\end{aligned}$$

Note that the sums here begin at $L-1$.

On the other hand, the previous section showed that for sums beginning at L, we may pass from the MRA approximation errors to the Besov seminorm on wavelet coefficients:

$$\sum_{j\ge L}[2^{\alpha j}e_j(f)]^q \asymp |\theta|_b^q. \tag{B.45}$$

Although the ranges of summation differ, this is taken care of by inclusion of L_p norm of f, as we now show. In one direction this is trivial since the sum from L is no larger than the sum from $L-1$. So, moving up the preceding chain, using also (B.44) with (B.38), we get

$$\|f\|_b = \|\beta\|_p + |\theta|_b \le C\|P_L f\|_p + C|f|_B \le C(\|f\|_p + |f|_B) = C\|f\|_B.$$

In the other direction, we connect the two chains by writing $|f|_B \le C[e_{L-1}(f) + |\theta|_b]$ and observing from (B.39) that $e_{L-1}(f) \le \|I - P_{L-1}\|_p \|f\|_p \le C\|f\|_p$. Consequently,

$$\|f\|_B = \|f\|_p + |f|_B \le C(\|f\|_p + |\theta|_b).$$

Now $\|f\|_p \le e_L(f) + \|P_L f\|_p$ which is in turn bounded by $C(|\theta|_b + \|\beta\|_p)$ by (B.45) and (B.44). Putting this into the last display finally yields $\|f\|_B \le C\|f\|_b$. □

B.4 Vaguelettes and frames

We rewrite Definition 12.2 without the rescaling operators. A collection $\{w_\lambda\}$ with $\lambda = (j,k)$ and $j \in \mathbb{Z}, k \in \Lambda_j \subset \mathbb{Z}$ is called a system of vaguelettes if there exist constants C_1, C_2 and exponents $0 < \eta < \eta' < 1$ such that

$$|w_\lambda(x)| \le C_1 2^{j/2}(1 + |2^j x - k|)^{-1-\eta'}, \tag{B.46}$$

$$\int w_\lambda(x)dx = 0, \tag{B.47}$$

$$|w_\lambda(x') - w_\lambda(x)| \le C_2 2^{j(1/2+\eta)}|x' - x|^\eta. \tag{B.48}$$

Proof of Proposition 12.3. (i) (Meyer and Coifman, 1997, Ch. 8.5) Let $K_{\lambda\lambda'} = \int w_\lambda \bar{w}_{\lambda'}$, our strategy is to use Schur's Lemma C.19 to show that K is bounded on ℓ_2. The ingredients are two bounds for $|K_{\lambda\lambda'}|$. To state the first, use (B.46) to bound $|K_{\lambda\lambda'}| \le C 2^{-|j'-j|/2} L_{\lambda\lambda'}$, where $L_{\lambda\lambda'}$ is the left side of the convolution bound

$$\int \frac{2^{j\wedge j'}dx}{(1+|2^j x - k|)^{1+\eta'}(1+|2^{j'}x - k'|)^{1+\eta'}} \le \frac{C}{(1+2^{j\wedge j'}|k'2^{-j'} - k2^{-j}|)^{1+\eta'}}, \tag{B.49}$$

verified in Exercise B.1. Denoting the right side by $CM_{\lambda\lambda'}^{1+\eta'}$, the first inequality states

$$|K_{\lambda\lambda'}| \le C_1 2^{-|j'-j|/2} M_{\lambda\lambda'}^{1+\eta'}. \tag{B.50}$$

For the next inequality, use the zero mean and Hölder hypotheses, (B.47) and (B.48), to argue, just as at (9.25) and (9.26), that for $j' \ge j$,

$$|K_{\lambda\lambda'}| \le C 2^{j(1/2+\eta)} \int |x - k'2^{-j'}|^\eta |w_{\lambda'}(x)|dx.$$

Using again (B.46) to bound $w_{\lambda'}$ and then $\eta < \eta'$ to assure convergence of the integral, we arrive at the second inequality

$$|K_{\lambda\lambda'}| \le C_2 2^{-|j'-j|(1/2+\eta)}. \tag{B.51}$$

The two bounds are combined by writing $|K_{\lambda\lambda'}|^{1-\theta}|K_{\lambda\lambda'}|^{\theta}$ and then using (B.50) in the first factor and (B.51) in the second to obtain

$$|K_{\lambda\lambda'}| \le C_3 2^{-|j'-j|(1/2+\delta)} M_{\lambda\lambda'}^{1+\delta} \tag{B.52}$$

by setting $\delta = \theta\eta$ for $\theta > 0$ sufficiently small that $1 + \delta < (1-\theta)(1+\eta')$.

We apply Schur's Lemm C.19 with weights $p_\lambda = q_\lambda = 2^{-j/2}$ so that, noting the symmetry of $K_{\lambda\lambda'}$, we need to show that $S_\lambda = 2^{j/2}\sum_{\lambda'} |K_{\lambda\lambda'}|2^{-j'/2}$ is uniformly bounded in $\lambda = (jk)$. From (B.52) we need to bound

$$\sum_{j'} 2^{-(j'-j)/2-|j'-j|(1/2+\delta)} \sum_{k'} M_{\lambda\lambda'}^{1+\delta}.$$

Consider the sum over k'. If $d = j' - j \ge 0$, then

$$2^{-d}\sum_{k'} M_{\lambda\lambda'}^{1+\delta} = \sum_{k'} \frac{2^{-d}}{(1+|k-2^{-d}k'|)^{1+\delta}} \le 2^{-d} + \int \frac{dt}{(1+|t|)^{1+\delta}} \le C_\delta,$$

while if $j' < j$ with $\varepsilon = 2^{j'-j}$, the terms $M_{\lambda\lambda'}^{1+\delta} \le C(1+|k'-k\varepsilon|)^{-1-\delta}$ have sum over k' uniformly bounded in k and $\varepsilon \le 1$. Hence in both cases, $\sum_{k'} M_{\lambda\lambda'}^{1+\delta}$ is bounded by $C_\delta 2^{(j'-j)_+}$. Since $u + |u| - 2u_+ = 0$, we have $S_\lambda \le C\sum_j 2^{-\delta|j'-j|} \le C$ uniformly in λ as required.

(ii). The biorthogonality means that $\sum |\alpha_\lambda|^2 = \langle \sum \alpha_\lambda u_\lambda, \sum \alpha_\mu v_\mu\rangle$, and hence by Cauchy-Schwarz that

$$\|\alpha\|^2 \le \|\sum \alpha_\lambda u_\lambda\| \|\sum \alpha_\mu v_\mu\|.$$

From part (i), we have $\|\sum \alpha_\mu v_\mu\| \le C\|\alpha\|$, so it follows that $\|\sum \alpha_\lambda u_\lambda\| \ge C^{-1}\|\alpha\|$. Reverse the roles of u and v to establish the same lower bound for $\|\sum \alpha_\mu v_\mu\|$. □

Proof of Theorem 9.6 We abbreviate $\|f\|_{W_2^r}$ by $\|f\|_r$ and the sequence norm in (9.28) by $|||f|||_r^2$. We establish $\|f\|_r \le C\,|||f|||_r$ for $f \in V_J$ and conclude by density. For $f \in V_J$ we can differentiate term by term to get

$$D^r f = \sum_k \beta_k \phi_{0k}^{(r)} + \sum_{j=0}^{J}\sum_k 2^{jr}\theta_{jk}\psi_{jk}^{(r)} = D^r f_0 + D^r f_1.$$

Under the hypotheses on ψ, it was shown in Section 12.3, example 1, that $\{(\psi^{(r)})_\lambda\}$ is a system of vaguelettes and hence by Proposition 12.3 satisfies the frame bounds (9.27). Apply the frame bound to conclude that $\|D^r f_1\|_2 \le C\,|||f|||_r$ and Lemma B.18 (for $p = 2$, $j = 0$ with orthogonality not required) to obtain $\|D^r f_0\|_2 \le C\sum \beta_k^2$. Putting these together, we get $\|f\|_r \le C\,|||f|||_r$ for $f \in V_J$. The density argument says that for $f \in W_2^r$, we have $P_J f \to f$ in L_2 and that $D^r P_J f$ is an L_2 Cauchy sequence (since $\|D^r(P_J f - P_K f)\|_2 \le C\,|||P_J f - P_K f|||_r$) so $P_J \to f$ in W_2^r.

In the other direction, for $f \in V_J$, we have $D^r f = \sum_{j\le J,k} 2^{jr}\psi_\lambda^{(r)}$, since the sum converges in L_2 at $J = -\infty$ from the frame bound. Hence

$$\sum_{j\ge 0,k} 2^{2rj}\theta_{jk}^2 \le \sum_{j\le J,k} (2^{rj}\theta_{jk})^2 \le C^2 \|D^r f\|_2^2,$$

while $\sum \beta_k^2 \le \|f\|_2^2$. Add the bounds to get $|||f|||_r^2 \le C^2\|f\|_r^2$ and extend by density. □

B.5 Notes

[2]. Footnotes such as this will be used to explain certain simple details. Thus, if $r(\xi) = \sum_0^m r_k e^{-ik\xi}$, with $r_k \in \mathbb{R}$, then $|r(\xi)|^2 = r(\xi)r^*(\xi) = r(\xi)r(-\xi) = \sum_{-m}^m s_k e^{-ik\xi}$ is both real and even, so $s_{-k} = s_k$ and hence it is a polynomial of degree m in $\cos\xi = 1 - 2\sin^2(\xi/2)$. In addition, $|(1 + e^{-i\xi})/2|^2 = \cos^2(\xi/2)$.

[3]. If P_1, P_2 are degree $p-1$ solutions of (B.14), then $Q = P_1 - P_2$ satisfies $(1-y)^p Q(y) + y^p Q(1-y) \equiv 0$, which implies that the degree $p-1$ polynomial Q has $Q^{(j)}(0) = 0$ for $0 \le j < p$ and so $Q \equiv 0$.

Exercises

B.1 *Verification of* (B.49). (a) Set $t = 2^{j'}x - k$, $\rho = 2^{j-j'}$ and $\lambda = k - \rho k'$ and show that the inequality reduces to

$$\int_{-\infty}^{\infty} \frac{dt}{(1+|\rho t - \lambda|)^\gamma (1+|t|)^\gamma} \le \frac{C(\gamma)}{(1+\lambda)^\gamma}$$

for $\gamma = 1 + \eta' > 1$ and $0 < \rho \le 1, \lambda \in \mathbb{R}$.
(b) Show that for $\lambda \le 1$ this bound is immediate and for $\lambda \ge 1$ set $g(t) = (1+|\lambda - \rho t|)(1+|t|)$ and obtain the inequality from the bounds

$$g(t) \ge \begin{cases} (1+\lambda)(1+|t|) & t \le 0, \\ (1+\lambda/2)(1+t) & 0 \le t < \lambda/(2\rho), \\ (\lambda/2)(1+|t-\lambda/\rho|) & \lambda/(2\rho) \le t \le \lambda/\rho, \\ \lambda(1+t-\lambda/\rho) & t \ge \lambda/\rho. \end{cases}$$

Appendix C

Background Material

> The reader ... should not be discouraged, if on first reading of §0, he finds that he does not have the prerequisites for reading the prerequisites. (Paul Halmos, *Measure Theory*).

Here we collect bits of mathematical background, with references, that are used in the main text, but are less central to the statistical development (and so, in that important sense, are not prerequisites). Not a systematic exposition, this collection has two aims: initially to save the reader a trip to an authoritative source, and later, if that trip is needed, to point to what is required. References in brackets, like [§1.4], indicate sections of the main text that refer here.

Hilbert spaces etc. [§1.4] `[add, also refs.]` If $\{\varphi_i, i \in I\}$ is a complete orthonormal basis for $L_2(T)$, then f can be expanded as $\sum_i c_i \varphi_i$ with coefficients $c_i = \int f \bar{\varphi}_i$ that satisfy the Parseval relation

$$\int_T |f(t)|^2 dt = \sum_{i \in I} |c_i|^2. \tag{C.1}$$

C.1 Norms and quasi-norms.

Compact operators, Hilbert-Schmidt and Mercer theorems. [§3.8]

We begin with some definitions and notation, relying for further detail on Reed and Simon (1980, Ch. VI.5,6) and Riesz and Sz.-Nagy (1955, Ch. VI, §97,98).

Let $\mathcal{H}$ and $\mathcal{K}$ be Hilbert spaces, with the inner product denoted by $\langle \cdot, \cdot \rangle$, with subscripts $\mathcal{H}$ and $\mathcal{K}$ shown as needed. A linear operator $A : \mathcal{H} \to \mathcal{K}$ is bounded if $\|A\| = \sup\{\|Ax\|_{\mathcal{K}} : \|x\|_{\mathcal{H}}\} < \infty$. The null space of A is $N(A) = \{x : Ax = 0\}$. The adjoint operator $A^* : \mathcal{K} \to \mathcal{H}$ is defined by the relations $\langle A^* y, x \rangle_{\mathcal{H}} = \langle y, Ax \rangle_{\mathcal{K}}$ for all $x \in \mathcal{H}$, $y \in \mathcal{K}$. Operator A is self-adjoint if $A^* = A$. We say A is compact if A takes bounded sets to sets with compact closure, or equivalently, if for every bounded sequence $\{x_n\} \subset \mathcal{H}$, the sequence $\{Ax_n\}$ has a convergent subsequence.

Theorem C.2 (Hilbert-Schmidt) *Let A be a compact self-adjoint linear operator on $\mathcal{H}$. There exists a complete orthonormal basis $\{\varphi_n\}$ for $\mathcal{H}$ such that*

$$A\varphi_n = \lambda_n \varphi_n, \qquad \text{with } \lambda_n \in \mathbb{R} \text{ and } \lambda_n \to 0 \text{ as } n \to \infty.$$

The Singular Value Decomposition. Suppose $A : \mathcal{H} \to \mathcal{K}$ is linear and compact. Then $A^*A : \mathcal{H} \to \mathcal{H}$ is self-adjoint and compact, and so the Hilbert-Schmidt theorem yields an orthonormal set $\{\varphi_n\}$ with *positive* eigenvalues

$$A^*A\varphi_n = b_n^2\varphi_n, \qquad b_n^2 > 0.$$

The set $\{\varphi_n\}$ need *not* be complete! However $A^*A = 0$ on the subspace $N(A) = N(A^*A)$ orthogonal to the closed linear span of $\{\varphi_n\}$. Define

$$\psi_n = \frac{A\varphi_n}{\|A\varphi_n\|} = b_n^{-1}A\varphi_n.$$

The set $\{\psi_n\}$ is orthnormal, and

$$A\varphi_n = b_n\psi_n, \qquad A^*\psi_n = b_n\varphi_n. \tag{C.2}$$

It can be verified that $\{\psi_n\}$ is a complete orthonormal basis for the closure of the range of A, and hence that for any $f \in \mathcal{H}$, using (C.2)

$$Af = \sum_n \langle Af, \psi_n\rangle\psi_n = \sum b_n\langle f, \varphi_n\rangle\psi_n. \tag{C.3}$$

Relations (C.2) and (C.3) describe the *singular value decomposition* of A, and $\{b_n\}$ are the singular values.

We have also

$$f = \sum b_n^{-1}\langle Af, \psi_n\rangle\varphi_n + u, \qquad u \in N(A). \tag{C.4}$$

In (C.3) and (C.4), the series converge in the Hilbert norms of $\mathcal{K}$ and $\mathcal{H}$ respectively.

C.3 *Kernels, Mercer's theorem.* [§3.9, §3.8] An operator $A \in \mathcal{L}(\mathcal{H})$ is *Hilbert-Schmidt* if for some orthobasis $\{e_i\}$

$$\|A\|_{HS}^2 = \sum_{i,j} |\langle e_i, Ae_j\rangle|^2 < \infty.$$

The value of $\|A\|_{HS}^2$ does not depend on the orthobasis chosen: regarding A as an infinite matrix, $\|A\|_{HS}^2 = \operatorname{tr} A^*A$. Hilbert-Schmidt operators are compact. An operator A is Hilbert-Schmidt if and only if its singular values are square summable.

Further, if $\mathcal{H} = L^2(T, d\mu)$, then A is Hilbert-Schmidt if and only if there is a square-integrable function $A(s,t)$ with

$$Af(s) = \int A(s,t)f(t)d\mu(t),$$

and in that case

$$\|A\|_{HS}^2 = \iint |A(s,t)|^2 d\mu(s)d\mu(t). \tag{C.5}$$

Suppose now that $T = [a,b] \subset \mathbb{R}$ and that $A : L^2(T, dt) \to L^2(T, dt)$ has kernel $A(s,t)$. The kernel $A(s,t)$ is called (i) *continuous* if $(s,t) \to A(s,t)$ is continuous on $T \times T$, (ii) *symmetric* if $A(s,t) = A(t,s)$, and (iii) *non-negative definite* if $(Af, f) \geq 0$ for all f.

These conditions imply that A is square-integrable, $\iint_{T\times T} A^2(s,t)dsdt < \infty$, and hence that A is self-adjoint, Hilbert-Schmidt and thus compact and so, by the Hilbert-Schmidt theorem, A has a complete orthonormal basis $\{\varphi_n\}$ of eigenfunctions with eigenvalues λ_n^2.

Theorem C.4 (Mercer) *If A is continuous, symmetric and non-negative definite, then the series*

$$A(s,t) = \sum_n \lambda_n^2 \varphi_n(s)\overline{\varphi_n(t)}$$

converges uniformly and in $L^2(T \times T)$.

[§12.2] In constructing the WVD in Chapter 12, in some cases it is necessary to consider possibly unbounded linear operators A defined on a dense subset $\mathcal{D}(A) \subset L_2(T)$. See, for example, Reed and Simon (1980, Ch. VIII). We give a useful criterion for the existence of *representers* g for linear functionals $\langle f, \psi\rangle$, in the sense that $[Af, g] = \langle f, \psi\rangle$. Let $\mathcal{R}(A)$ denote the range of A. The following formulation is from Donoho (1995) and Bertero (1989).

Proposition C.5 *Suppose that $A : \mathcal{D}(A) \subset L_2(T) \to L_2(U)$ with $\overline{\mathcal{D}(A)} = L_2(T)$ and that A is one to one. For a given $\psi \in L_2(T)$, the following are equivalent:*

(i) There exists $g \in L_2(U)$ such that

$$\langle f, \psi\rangle = [Af, g] \quad \text{for all } f \in \mathcal{D}(A).$$

(ii) There exists C such that $\langle f, \psi\rangle \le C\,\|Af\|_2$ for all $f \in \mathcal{D}(A)$.

(iii) $\psi \in \mathcal{R}(A^*)$.

Proof We prove (iii) $\Rightarrow$ (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii). If $\psi = A^*g$, then (i) follows from the definition of A^*. Then (i) $\Rightarrow$ (ii) follows from the Cauchy-Schwarz inequality with $C = \|g\|_2$.

(ii) $\Rightarrow$ (iii). The linear functional $Lh = \langle A^{-1}h, \psi\rangle$ is well defined on $\mathcal{R}(A)$ since A is one-to-one. From the hypothesis, for all $h = Af$, we have $|Lh| = |\langle f, \psi\rangle| \le C\,\|h\|_2$. Thus L is bounded on $\mathcal{R}(A)$ and so extends by continuity to a bounded linear functional on $\overline{\mathcal{R}(A)}$. The Riesz representation theorem gives a $g \in \overline{\mathcal{R}(A)}$ such that

$$[Af, g] = L(Af) = \langle f, \psi\rangle \qquad \text{for all } f \in \mathcal{D}(A).$$

Since $\langle f, A^*g\rangle = \langle f, \psi\rangle$ for all f on a dense subset of $L_2(T)$, we recover $\psi = A^*g$. □

[§4.2]. An extended form of the dominated convergence theorem, due to Young (1911) and rediscovered by Pratt (1960), has an easy proof, e.g. Bogachev (2007, Vol I, Theorem 2.8.8)

Theorem C.6 *If f_n, g_n and G_n are μ-integrable functions and*

1. *$f_n \to f$, $g_n \to g$ and $G_n \to G$ a.e. (μ), with g and G integrable,*
2. *$g_n \le f_n \le G_n$ for all n, and*
3. *$\int g_n \to \int g$ and $\int G_n \to \int G$,*

then f is integrable, and $\int f_n \to \int f$.

Covariance inequality. [Exer. 4.2]. Let Y be a real valued random variable and suppose that $f(y)$ is increasing and $g(y)$ is decreasing. Then, so long as the expectations exist,

$$E[f(Y)g(Y)] \le E[f(Y)]E[g(Y)]. \tag{C.6}$$

[§7.1, §12.2, §B.1]. The Fourier transform of an integrable function on $\mathbb{R}$ is defined by

$$\hat{f}(\xi) = \int_{\infty}^{\infty} f(x)e^{-i\xi x}dx. \tag{C.7}$$

If f is sufficiently nice (SEE REFS), it may be recovered from the inversion formula

$$f(x) = \frac{1}{2\pi}\int_{\infty}^{\infty} \hat{f}(\xi)e^{i\xi x}d\xi.$$

The function f has *p vanishing moments*, i.e. $\int x^k f(x)dx = 0$ for $k = 0, 1, \ldots, p-1$, exactly when the derivatives $\widehat{f}^{(k)}(0) = 0$ for $k = 0, 1, \ldots, p-1$.

The Parseval (or Plancherel) identity states that if $f, g \in L_1 \cap L_2$,

$$\int f(x)\overline{g(x)}dx = \frac{1}{2\pi}\int \widehat{f}(\xi)\overline{\widehat{g}(\xi)}d\xi. \tag{C.8}$$

[§3.5, §14.5]. The Poisson summation formula (Dym and McKean, 1972, for example) states that if $(1 + x^2)[|f(x)| + |f'(x)| + |f''(x)|]$ is bounded (or if the same condition holds for $\hat{f}$), then

$$\sum_{k\in\mathbb{Z}} f(k) = \sum_{k\in\mathbb{Z}} \hat{f}(2\pi k). \tag{C.9}$$

When applied to $f(x) = g(x + t)$, this yields a representation for the periodization of g (REF??):

$$\sum_k g(t + k) = \sum_k e^{2\pi ikt}\hat{g}(2\pi k), \qquad t \in \mathbb{R}. \tag{C.10}$$

Some further properties of the Gaussian distribution. [§2.8].

Lemma C.7 *. (a) If $X \sim N_n(\mu, \Sigma)$ and M is an $m \times n$ matrix, then $MX \sim N_m(M\mu, M\Sigma M^T)$.*

(b) If $X \sim N_n(0, \sigma^2 I)$ and U is an $n \times n$ orthogonal matrix, then $UX \sim N_n(0, \sigma^2 I)$ also.

C.8 *Brownian motion, Wiener integral.* [§1.4, §3.9]. A process $\{Z(t), t \in T\}$ is Gaussian if all finite-dimensional distributions $(Z(t_1), \ldots, Z(t_k))$ have Gaussian distributions for all $(t_1, t_2, \ldots, t_k) \in T^k$ and positive integer k. It is said to be continuous in quadratic mean if $E[Z(t+h) - Z(t)]^2 \to 0$ as $h \to 0$ at all t.

The following basic facts about Brownian motion and Wiener integrals may be found, for example, in Kuo (2006, Ch. 2). Standard Brownian motion on the interval $[0, 1]$ is defined as a Gaussian process $\{W(t)\}$ with mean zero and covariance function $\text{Cov}\,(W(s), W(t)) = s \wedge t$. It follows that $\{W(t)\}$ has independent increments: if $0 \le t_1 < t_2 < \cdots < t_n$, then the increments $W(t_j) - W(t_{j-1})$ are independent. In addition, the sample paths $t \to W(t, \omega)$ are continuous with probability one.

The *Wiener integral* $X = I(f) = \int_0^1 f(t)dW(t)$ of a deterministic function f is defined first for step functions and then for $f \in L_2[0, 1]$ by convergence of random variables in the Hilbert space $L_2(\Omega)$ with inner product $\langle X, Y \rangle = EXY$. In particular, the identity

$$\langle f, g \rangle_{L_2[0,1]} = EI(f)I(g)$$

holds, and $I(f) \sim N(0, \|f\|_2^2)$. If f is continuous and of bounded variation, then $I(f)$ can be interpreted as a Riemann-Stieltjes integral.

If $\{\varphi_i\}$ is an orthonormal basis for $L_2[0, 1]$, then $f = \sum \langle f, \varphi_i \rangle \varphi_i$ and

$$I(f) = \sum \langle f, \varphi_i \rangle I(\varphi_i),$$

where the variables $z_i = I(\varphi_i)$ are i.i.d. standard Gaussian, and the series converges almost surely. In particular,

$$W(t) \stackrel{\mathcal{D}}{=} \sum_{i=1}^{\infty} z_i \int_0^t \phi_i(s)ds$$

with the series converging almost surely (Shepp, 1966). Particular examples for which this representation was known earlier include the trigonmetric basis $\phi_k(t) = \sqrt{2}\cos(k - \frac{1}{2})\pi t$ (Wiener) and the Haar basis $\phi_{jk}(t) = 2^{j/2}h(2^j t - k)$ for $h(t)$ equal to 1 on $[0, \frac{1}{2}]$ and to -1 on $[\frac{1}{2}, 1]$ (Lévy).

[§8.10, §13.5]. The moment generating function of a standard Gaussian variable is

$$Ee^{\beta z} = e^{\beta^2/2}. \tag{C.11}$$

Proposition C.9 (Talagrand (2003), Proposition 1.1.4.) *Let $z_1, \ldots, z_n \sim N(0, 1)$. Then*

$$E \log \Big(\sum_1^n e^{\beta z_i} \Big) \le \begin{cases} \frac{1}{2}\beta^2 + \log n & \text{if } \beta \le \sqrt{2 \log n} \\ \beta\sqrt{2 \log n} & \text{if } \beta \ge \sqrt{2 \log n} \end{cases} \tag{C.12}$$

and, as a consequence,

$$E \max_{i \le n} z_i \le \sqrt{2 \log n}. \tag{C.13}$$

Note that the z_i need not be independent.

Proof Let $\phi_n(\beta) = E \log\left(\sum_1^n e^{\beta z_i}\right)$. From the moment generating function (C.11), we have $E \sum e^{\beta z_i} = n e^{\beta^2/2}$. The first bound in (C.12) follows from Jensen's inequality and concavity of the logarithm.

Since $\beta \max z_i \le \log\left(\sum_1^n e^{\beta z_i}\right)$, we conclude that $E \max z_i \le \frac{1}{2}\beta + \beta^{-1} \log n$, and substitute $\beta = \sqrt{2 \log n}$ to obtain (C.13).

Write now $\lambda_n = \sqrt{2 \log n}$ and observe that for all β.

$$\phi_n'(\beta) = E\Big(\sum z_i e^{\beta z_i} \Big/ \sum e^{\beta z_i}\Big) \le E \max z_i \le \lambda_n,$$

while from the first part of (C.12) we have $\phi_n(\lambda_n) \le \lambda_n^2$. Consequently $\phi_n(\beta) \le \lambda_n \beta$ for all $\beta \ge \lambda_n$, which is the second part of (C.12). □

[§8.9]. **Weak law of large numbers for triangular arrays.** Although designed for variables without finite second moment, the truncation method works well for the cases of rapidly growing variances that occur here. The following is taken from Durrett (2010, Thm 2.2.6).

Proposition C.10 *For each n let $X_{nk}, 1 \le k \le n$, be independent. Let $b_n > 0$ with $b_n \to \infty$, and let $\bar{X}_{nk} = X_{nk} I\{|X_{nk}| \le b_n\}$. Suppose that as $n \to \infty$,*
(i) $\sum_{k=1}^n P(|X_{nk}| > b_n) \to 0$, and
(ii) $b_n^{-2} \sum_{k=1}^n E X_{nk}^2 \to 0$ as $n \to \infty$.
Let $S_n = X_{n1} + \ldots + X_{nn}$ and put $a_n = \sum_{k=1}^n E \bar{X}_{nk}$. Then

$$S_n = a_n + o_p(b_n).$$

```
[Orthogonality of measures]
[Support of μ]
[Convex Set]
[l.s.c. and max on compact]
[metric space: seq cty = cty]
```

C.11 `[complete, separable, metrizable, Borel field, Radon measure, second countable, Hausdorff....]`

A subset K of a metric space is compact if every covering of K by open sets has a finite subcover.

A subset K of a metric space is totally bounded if it can be covered by finitely many balls of radius ϵ for every $\epsilon > 0$.

[Ref: Rudin FA p 369] If K is a closed subset of a complete metric space, then the following three properties are equivalent: (a) K is compact, (b) Every infinite subset of K has a limit point in K, (c) K is totally bounded.

[§4.2, §4.4]. First recall that a function $f : X \to \mathbb{R}$ on a topological space X is *lower*

semicontinuous (lsc) iff $\{x : f(x) > t\}$ is open for all t, or equivalently if $\{x : f(x) \le t\}$ is closed for all t.

If $\{f_\alpha : \alpha \in A\}$ is a set of lower semicontinous functions, then the pointwise supremum

$$f(x) = \sup_{\alpha \in A} f_\alpha(x)$$

is lower semicontinuous.

C.12 If X is compact, then an lsc function f attains its infimum: $\inf_{x \in X} f = f(x_0)$ for some $x_0 \in X$.

[If X is 1st countable, then these conditions may be rewritten in terms of sequences as $f(x) \le \liminf f(x_n)$ whenever $x_n \to x$.]

A function g is *upper semicontinuous* if $f = -g$ is lsc.

Weak convergence of probability measures. [§4.4]. Let Ω be a complete separable metric space–for us, usually a subset of $\mathbb{R}^n$ for some n. Let $\mathcal{P}(\Omega)$ denote the collection of probability measures on Ω with the Borel $\sigma-$algebra generated by the open sets. We say that $\pi_n \to \pi$ in the weak topology if

$$\int \psi d\pi_n \to \int \psi d\pi \tag{C.14}$$

for all bounded continuous $\psi : \Omega \to \mathbb{R}$.

A collection $\mathcal{P} \subset \mathcal{P}(\Omega)$ is called *tight* if for all $\epsilon > 0$, there exists a compact set $K \subset \Omega$ for which $\pi(K) > 1 - \epsilon$ for every $\pi \in \mathcal{P}$.

Prohorov's theorem (Billingsley, 1999, Ch. 1.5) provides a convenient description of compactness in $\mathcal{P}(\Omega)$: a set $\mathcal{P} \subset \mathcal{P}(\Omega)$ has compact closure if and only if $\mathcal{P}$ is tight.

Thus, if $\Omega = [-\tau, \tau]$ then $\mathcal{P}(\Omega)$ has compact closure. If $\Omega = \mathbb{R}$ and $\mathcal{P} = \{\pi : \int |\theta|^p \pi(d\theta) \le \eta^p\}$, then Markov's inequality shows that $\pi([-M, M]^c) \le \eta^p / M^p$ for any $\pi \in \mathcal{P}$, so that $\mathcal{P}$ is tight and hence weakly compact.

C.13 *Vague convergence.* [§4.4]. Let $\Omega = \mathbb{R}$ and $\mathcal{P}_+(\mathbb{R})$ be the collection of *sub-stochastic* measures on $\mathbb{R}$. Equivalently, $\mathcal{P}_+ = \mathcal{P}(\bar{\mathbb{R}})$ for $\bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, allowing mass at $\pm\infty$. We say that $\pi_n \to \pi$ in the *vague* topology if (C.14) holds for all continuous ψ with compact support, or (equivalently) for all continuous ψ that vanish at $\pm\infty$.

Clearly weak convergence implies vague convergence, and if $\mathcal{P} \subset \mathcal{P}(\mathbb{R})$ is weakly compact, then it is vaguely compact. However $\mathcal{P}(\mathbb{R})$ is not weakly compact (as mass can escape to $\pm\infty$) but $\mathcal{P}_+(\mathbb{R})$ *is* vaguely compact. [REF?]

[§4.2, §8.7]. The *Fisher information* for location of a distribution P on $\mathbb{R}$ is

$$I(P) = \sup_{\psi} \frac{\left(\int \psi' dP\right)^2}{\int \psi^2 dP}, \tag{C.15}$$

where the supremum is taken over the set C_0^1 of C^1 functions of compact support for which $\int \psi^2 dP > 0$. For this definition and the results quoted here, we refer to Huber and Ronchetti (2009, Chapter 4), [HR] below.

It follows from this definition that $I(P)$ is a convex function of P. The definition is however equivalent to the usual one: $I(P) < \infty$ if and only if P has an absolutely continuous density p, and $\int p'^2/p < \infty$. In either case, $I(P) = \int p'^2/p$.

Given P_0, P_1 with $I(P_0), I(P_1) < \infty$ and $0 \le t \le 1$, let $P_t = (1-t)P_0 + tP_1$. Differentiating $I(P_t) = \int p_t'^2/p_t$ under the integral sign (which is justified in HR), one obtains

$$\begin{aligned} \frac{d}{dt} I(P_t)|_{t=0} &= \int \frac{2p_0'}{p_0}(p_1' - p_0') - \frac{p_0'^2}{p_0^2}(p_1 - p_0) \\ &= \int [-2\psi_0 p_1' - \psi_0^2 p_1]dx - I(P_0), \end{aligned} \tag{C.16}$$

where we have set $\psi_0 = -p_0'/p_0$ for terms multiplying p_1' and p_1 and observed that the terms involving only p_0' and p_0 collapse to $-I(P_0)$.

Since $I(P)$ is the supremum of a set of vaguely (resp. weakly) continuous functions, it follows that $P \to I(P)$ is vaguely (resp. weakly) lower semicontinuous[1]. Consequently, from C.12, if $\mathcal{P} \subset \mathcal{P}_+(\mathbb{R})$ is vaguely compact, then there is an $P_0 \in \mathcal{P}$ minimizing $I(P)$.

Formula (C.16) yields a helpful variational criterion for characterizing a minimizing P_0. Let $\mathcal{P}_1 = \{P_1 \in \mathcal{P} : I(P_1) < \infty\}$ and for given P_0 and P_1, let $P_t = (1-t)P_0 + tP_1$. Since $I(P)$ is convex in P, a distribution $P_0 \in \mathcal{P}$ minimizes $I(P)$ if and only if $(d/dt)I(P_t) \ge 0$ at $t = 0$ for each $P_1 \in \mathcal{P}_1$.

A slight reformulation of this criterion is also useful. The first term on the right side of (C.16) is $\int -2\psi_0(p_1' - p_0') = \int 2\psi_0'(p_1 - p_0)$ `(justify!)` and so P_0 minimizes $I(P)$ over $\mathcal{P}$ if and only if

$$\int [2\psi_0' - \psi_0^2](p_1 - p_0) \ge 0. \tag{C.17}$$

C.14 (Uniqueness). Suppose (i) that $\mathcal{P}$ is convex and $P_0 \in \mathcal{P}$ minimizes $I(P)$ over $\mathcal{P}$ with $0 < I(P_0) < \infty$, and (ii) that the set on which p_0 is positive is an interval and contains the support of every $P \in \mathcal{P}$. Then P_0 is the unique minimizer of $I(P)$ in $\mathcal{P}$.

In our applications, P is typically the marginal distribution $\Phi \star \pi$ for a (substochastic) prior measure π. (For this reason, the notation uses $\mathcal{P}^\star$ for classes of distributions P, which in these applications correspond to classes $\mathcal{P}$ of priors through $\mathcal{P}^\star = \{P = \Phi \star \pi, \pi \in \mathcal{P}\}$.) In particular, in the uniqueness result, p_0 is then positive on all of $\mathbb{R}$ and so condition (ii) holds trivially.

C.15 Stein's Unbiased Estimate of Risk. [§2.6]. We provide some extra definitions and details of proof for the unbiased risk identity that comprises Proposition 2.5. As some important applications of the identity involve functions that are only "almost" differentiable, we begin with some remarks on weak differentiability, referring to standard sources, such as Gilbarg and Trudinger (1983, Chapter 7), for omitted details.

[1] indeed, if $V_\psi(P)$ denotes the ratio in (C.15), then $\{P : I(P) > t\}$ is the union of sets of the form $\{P : V_\psi(P) > t, \int \psi^2 dP > 0\}$ and hence is open.

A function $g : \mathbb{R}^d \to \mathbb{R}$ is said to be *weakly differentiable* if there exist functions $h_i : \mathbb{R}^d \to \mathbb{R}, i = 1, \ldots d$, such that

$$\int \psi h_i = -\int (D_i \psi) g \qquad \text{for all } \psi \in C_0^\infty,$$

where C_0^∞ denotes the class of C^∞ functions on $\mathbb{R}^d$ of compact support. We write $h_i = D_i g$.

To verify weak differentiability in particular cases, we note that it can be shown that g is weakly differentiable if and only if it is equivalent to a function $\bar{g}$ that is absolutely continuous on almost all line segments parallel to the co-ordinate axes and whose partial derivatives (which consequently exist almost everywhere) are locally integrable.

Proof of Proposition 2.5 First note that by a simple translation of parameter, it suffices to consider $\mu = 0$. Next, consider scalar C^∞ functions $\psi : \mathbb{R}^d \to \mathbb{R}$ of compact support, so that formula (2.41) becomes a simple integration by parts:

$$\begin{aligned} \int x_i \psi(x) \phi(x) dx &= \int \psi(x)[-D_i \phi(x)] dx \\ &= \int D_i \psi(x) \phi(x) dx. \end{aligned} \tag{C.18}$$

To verify (2.41) for general g we take limits in (C.18), and exploit a standard convergence criterion: suppose that h_i and g belong to $L_1(\Phi) = L_1(\mathbb{R}^d, \phi(x)dx)$. Then $h_i = D_i g$ if and only if there exists a sequence of C_0^∞ functions $\{\psi_n\}$ with $x_i \psi_m(x)$ converging to $x_i g(x)$ in $L_1(\Phi)$ such that $D_i \psi_m \to h_i$ in $L_1(\Phi)$.

Formula (2.42) follows immediately from (2.41) (since $E_\mu \|X - \mu\|^2 = d$). □

C.16 Hölder spaces. [§4.7, §7.1, §9.6, §B.3]. The Hölder spaces $C^\alpha(I)$ measure smoothness *uniformly* on an interval I, with smoothness parameter α. The norms have the form $\|f\|_{C^\alpha} = \|f\|_{\infty,I} + |f|_\alpha$, since the seminorm $|f|_\alpha$ reflecting the dependence on α will typically vanish on a finite dimensional space.

If α is a positive integer, then we require that f have α continuous derivatives, and set $|f|_\alpha = \|D^\alpha f\|_{\infty,I}$.

For $0 < \alpha < 1$, we require finiteness of

$$|f|_\alpha = \sup \left\{ \frac{|f(x) - f(y)|}{|x - y|^\alpha},\ x, y \in I \right\}. \tag{C.19}$$

If m is a positive integer and $m < \alpha < m + 1$, then we require that f have m uniformly continuous derivatives and also finiteness of

$$|f|_\alpha = |D^m f|_{\alpha - m}.$$

We note also that Hölder functions can be uniformly approximated by (Taylor) polynomials. Indeed, we can say that $f \in C^\alpha(I)$ implies that there exists a constant C such that for each $x \in I$, there exists a polynomial $p_x(y)$ of degree $\lceil \alpha \rceil - 1$ such that

$$|f(x + y) - p_x(y)| \le C |y|^\alpha, \qquad \text{if } x + y \in I. \tag{C.20}$$

The constant C can be taken as $c_\alpha |f|_\alpha$, where c_α equals 1 if $0 < \alpha < 1$ and equals $\prod_{j=1}^{[\alpha]}(\alpha + 1 - j)$ if $\alpha \geq 1$.

Total Variation. [§ 9.6] When $I = [a, b]$, this norm is defined by

$$\|f\|_{TV(I)} = \sup\{\sum_{i=1}^{n} |f(t_i) - f(t_{i-1})| \ : \ a = t_0 < t_1 < \cdots < t_n = b, n \in \mathbb{N}\}.$$

It represents a scientifically interesting enlargement of W_1^1, since when $f \in W_1^1$, we may write

$$\|f\|_{TV} = \int |Df|. \tag{C.21}$$

[explain equivalence of norms]

[§B.3]. **Background.** For convenience, we record a straightforward extension of Young's inequality for convolutions.

Theorem C.17 *Let $(X, \mathcal{B}_X, \mu)$ and $(Y, \mathcal{B}_Y, \nu)$ be σ–finite measure spaces, and let $K(x, y)$ be a jointly measurable function. Suppose that*

$$(i) \qquad \int |K(x, y)|\mu(dx) \leq M_1 \qquad \text{a.e. } (\nu), \qquad \text{and}$$

$$(ii) \qquad \int |K(x, y)|\nu(dy) \leq M_2 \qquad \text{a.e. } (\mu).$$

For $1 \leq p \leq \infty$, the operator

$$(Kf)(x) = \int K(x, y) f(y)\nu(dy)$$

maps $L_p(Y) \to L_p(X)$ with

$$\|Kf\|_p \leq M_1^{1/p} M_2^{1-1/p} \|f\|_p. \tag{C.22}$$

Proof For $p = \infty$ the result is immediate. For $1 < p < \infty$, let q be the conjugate exponent $1/q = 1 - 1/p$. Expand $|K(x, y)|$ as $|K(x, y)|^{1/q}|K(x, y)|^{1/p}$ and use Hölder's inequality:

$$|Kf(x)| \leq \Big[\int |K(x, y)|\nu(dy)\Big]^{1/q} \Big[\int |K(x, y)||f(y)|^p \nu(dy)\Big]^{1/p},$$

so that, using (ii),

$$|Kf(x)|^p \leq M_2^{p/q} \int |K(x, y)||f(y)|^p \nu(dy).$$

Now integrate over x, use Fubini's theorem and bound (i) to obtain (C.22). The proof for $p = 1$ is similar and easier. □

Remark. The adjoint $(K^* g)(y) = \int g(x)K(x, y)\mu(dx)$ maps $L_p(X) \to L_p(Y)$ with

$$\|K^* g\|_p \leq M_1^{1-1/p} M_2^{1/p} \|g\|_p. \tag{C.23}$$

Two traditional forms of Young's inequality are immediate consequences.

Corollary C.18 (§12.3, §B.3) *. Suppose that* $1 \le p \le \infty$.
(i) If $Kf(x) = \int_{\infty}^{\infty} K(x-y)f(y)dy$, *then*

$$\|Kf\|_p \le \|K\|_1 \|f\|_p. \tag{C.24}$$

(ii) If $c_k = \sum_{j\in\mathbb{Z}} a_{k-j}b_j$, *then*

$$\|c\|_p \le \|a\|_1 \|b\|_p. \tag{C.25}$$

Another consequence, in the L_2 setting, is a version with weights. Although true in the measure space setting of Theorem C.17, we need only the version for infinite matrices.

Corollary C.19 (Schur's Lemma) *[§15.3, §B.4]. Let* $K = (K(i,j))_{i,j\in\mathbb{N}}$ *be an infinite matrix and let* $(p(i))$ *and* $(q(j))$ *be sequences of positive numbers. Suppose that*

$$\text{(i)} \qquad \sum_i p(i)K(i,j) \le M_1 q(j) \qquad j \in \mathbb{N}, \qquad \text{and}$$

$$\text{(ii)} \qquad \sum_j K(i,j)q(j) \le M_2 p(i) \qquad i \in \mathbb{N},$$

Then the operator $(Kb)(i) = \sum_j K(i,j)b(j)$ *is bounded on* ℓ_2 *and*

$$\|Kb\|_2 \le \sqrt{M_1 M_2}\|b\|_2.$$

Proof Use the argument for Theorem C.17, this time expanding $|K(i,j)|$ as

$$|K(i,j)|^{1/2} q(j)^{1/2} \cdot |K(i,j)|^{1/2} q(j)^{-1/2}. \qquad \square$$

Theorem C.20 (Minkowski's integral inequality) *[§B.3]. Let* $(X, \mathcal{B}_X, \mu)$ *and* $(Y, \mathcal{B}_Y, \nu)$ *be* σ*–finite measure spaces, and let* $f(x,y)$ *be a jointly measurable function. Then for* $1 \le p \le \infty$,

$$\left(\int \left|\int f(x,y)\nu(dy)\right|^p \mu(dx)\right)^{1/p} \le \int \left(\int |f(x,y)|^p \mu(dx)\right)^{1/p} \nu(dy). \tag{C.26}$$

See, e.g. Okikiolu (1971, p. 159). [More canonical reference?]

Gauss' hypergeometric function [§3.8]. is defined for $|x| < 1$ by the series

$$F(\alpha,\beta,\gamma;x) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{x^n}{n!},$$

provided that $c \ne 0, -1, -2, \ldots$; and $(a)_n = a(a+1)(a+2)\cdots(a+n-1), (a)_0 = 1$ is the Pochhammer symbol. For $\operatorname{Re}\gamma > \operatorname{Re}\beta > 0$ and $|x| < 1$, Euler's integral representation says that

$$F(\alpha,\beta,\gamma;x) = B(\beta,\gamma-\beta)^{-1} \int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-tx)^{-\alpha}dt,$$

where $B(\beta,\gamma) = \Gamma(\beta)\Gamma(\gamma)/\Gamma(\beta+\gamma)$ is the beta integral. These and most identities given

here may be found in Abramowitz and Stegun (1964, Chs. 15, 22) See also Temme (1996, Chs. 5 and 6) for some derivations. Gel'fand and Shilov (1964, §5.5) show that this formula can be interpreted in terms of differentiation of fractional order

$$\frac{x^{\gamma-1}}{\Gamma(\gamma)}F(\alpha,\beta,\gamma;x) = D^{\beta-\gamma}\left(\frac{x_+^{\beta-1}(1-x)_+^{-\alpha}}{\Gamma(\beta)}\right). \tag{C.27}$$

They then show that the identity $D^{-\delta}D^{\beta-\gamma} = D^{\beta-\gamma-\delta}$ becomes, in integral form

$$x^{\gamma+\delta-1}F(\alpha,\beta,\gamma+\delta;x) = B(\gamma,\delta)^{-1}\int_0^x t^{\gamma-1}F(\alpha,\beta,\gamma;t)(x-t)^{\delta-1}dt. \tag{C.28}$$

Jacobi polynomials arise from the hypergeometric function when the series is finite

$$P_n^{a,b}(1-2x) = \binom{n+a}{n}F(-n,a+b+n+1,a+1;x),$$

where the generalized binomial coefficient is $\Gamma(n+a+1)/\Gamma(n+1)\Gamma(a+1)$. The polynomials $P_n^{a,b}(w), n \geq 0$ are orthogonal with respect to the weight function $(1-w)^a(1+w)^b$ on $[-1,1]$. Special cases include the *Legendre* polynomials $P_n(x)$, with $a = b = 0$, and the *Chebychev* polynomials $T_n(x)$ and $U_n(x)$ of first and second kinds, with $a = b = -1/2$ and $a = b = 1/2$ respectively.

The orthogonality relations, for the corresponding weight function on $[0,1]$, become

$$\int_0^1 P_m^{a,b}(1-2x)P_n^{a,b}(1-2x)\,x^a(1-x)^b dx = g_{a,b;n}^2\delta_{nm},$$

where

$$g_{a,b;n}^2 = \frac{n!}{2n+a+b+1}\frac{\Gamma(a+b+n+1)}{\Gamma(a+n+1)\Gamma(b+n+1)}. \tag{C.29}$$

Notation.
$\delta_{jk} = 1$ if $i = j$ and 0 otherwise.

Appendix D

To Do List

The general structure of the book is approaching final form, unless feedback now–which is welcome!–should lead to changes. Nevertheless, many smaller points still need clean-up. **Especially**, each chapter needs bibliographic notes discussing sources and references; this is by no means systematic at present.

Specific Sections:
§2.8 on more general linear models needs rewriting.
§6.6 Discussion (on adaptive minimaxity) needs rewriting
§7.6 (or elsewhere) discussion of block thresholding
A section/epilogue on topics not covered
Appendix C needs to be organized.
Some overview remarks about lower-bound techniques somewhere

Overall:
Each chapter clean up, attention to explaining flow. Also bibliographic notes, sources.
Table of Symbols/Acronyms,
Index.

References

Abel, N.H. 1826. Resolution dun probleme de mecanique. *J. Reine u. Angew. Math*, **1**, 153–157.

Abramovich, F., Sapatinas, T., and Silverman, B. W. 1998. Wavelet thresholding via a Bayesian approach. *J. Royal Statistical Society, Series B.*, **60**, 725–749.

Abramovich, Felix, Benjamini, Yoav, Donoho, David, and Johnstone, Iain. 2006. Adapting to Unknown Sparsity by controlling the False Discovery Rate. *Annals of Statistics*, **34**, 584–653.

Abramowitz, Milton, and Stegun, Irene A. 1964. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. National Bureau of Standards Applied Mathematics Series, vol. 55. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C.

Adler, Robert J., and Taylor, Jonathan E. 2007. *Random fields and geometry*. Springer Monographs in Mathematics. New York: Springer.

Anderson, Greg W., Guionnet, Alice, and Zeitouni, Ofer. 2010. *An Introduction to Random Matrices*. Cambridge University Press.

Ash, Robert B., and Gardner, Melvin F. 1975. *Topics in Stochastic Processes*. Academic Press.

Belitser, E.N., and Levit, B.Y. 1995. On Minimax Filtering over Ellipsoids. *Mathematical Methods of Statistics*, **4**, 259–273.

Bergh, J., and Löfström, J. 1976. *Interpolation spaces – An Introduction*. New York: Springer Verlag.

Berkhin, P.E., and Levit, B. Ya. 1980. Asymptotically minimax second order estimates of the mean of a normal population. *Problems of information transmission*, **16**, 60–79.

Bertero, M. 1989. Linear inverse and ill-posed problems. Pages 1–120 of: *Advances in Electronics and Electron Physics*, vol. 75. New York: Academic Press.

Bickel, Peter J. 1981. Minimax estimation of the mean of a normal distribution when the parametr space is restricted. *Annals of Statistics*, **9**, 1301–1309.

Bickel, Peter J. 1983. Minimax estimation of a normal mean subject to doing well at a point. Pages 511–528 of: Rizvi, M. H., Rustagi, J. S., and Siegmund, D. (eds), *Recent Advances in Statistics*. New York: Academic Press.

Billingsley, Patrick. 1999. *Convergence of probability measures*. Second edn. Wiley Series in Probability and Statistics: Probability and Statistics. New York: John Wiley & Sons Inc. A Wiley-Interscience Publication.

Birgé, L. 1983. Approximation dans les éspaces metriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **65**, 181–237.

Birgé, Lucien, and Massart, Pascal. 2001. Gaussian Model Selection. *Journal of European Mathematical Society*, **3**, 203–268.

Birkhoff, Garrett, and Rota, Gian-Carlo. 1969. *Ordinary Differential Equations*. Blaisdell.

Bogachev, V. I. 2007. *Measure theory. Vol. I, II*. Berlin: Springer-Verlag.

Bogachev, Vladimir I. 1998. *Gaussian Measures*. American Mathematical Society.

Borell, Christer. 1975. The Brunn-Minkowski inequality in Gauss space. *Invent. Math.*, **30**(2), 207–216.

Born, M., and Wolf, E. 1975. *Principles of Optics*. 5th edn. New York: Pergamon.

Breiman, Leo. 1968. *Probability*. Reading, Mass.: Addison-Wesley Publishing Company.

Breiman, Leo. 1995. Better subset selection using the non-negative garotte. *Technometrics*, **37**, 373–384.

Brown, L., DasGupta, A., Haff, L. R., and Strawderman, W. E. 2006. The heat equation and Stein's identity: connections, applications. *J. Statist. Plann. Inference*, **136**(7), 2254–2278.

Brown, L. D., and Low, M. G. 1996a. Asymptotic equivalence of nonparametric regression and white noise. *Annals of Statistics*, **3**, 2384–2398.

Brown, Lawrence D. 1971. Admissible estimators, recurrent diffusions and insoluble boundary value problems. *Annals of Mathematical Statistics*, **42**, 855–903. Correction: *Ann. Stat.* **1** 1973, pp 594–596.

Brown, Lawrence D. 1986. *Fundamentals of statistical exponential families with applications in statistical decision theory*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 9. Hayward, CA: Institute of Mathematical Statistics.

Brown, Lawrence D., and Gajek, Leslaw. 1990. Information Inequalities for the Bayes Risk. *Annals of Statistics*, **18**, 1578–1594.

Brown, Lawrence D., and Low, Mark G. 1996b. A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.*, **24**(6), 2524–2535.

Brown, Lawrence D., Low, Mark G., and Zhao, Linda H. 1997. Superefficiency in nonparametric function estimation. *Annals of Statistics*, **25**, 2607–2625.

Brown, Lawrence D., Carter, Andrew V., Low, Mark G., and Zhang, Cun-Hui. 2004. Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist.*, **32**(5), 2074–2097.

Brown, L.D. 1977. Closure theorems for sequential-design processes. In: Gupta, S.S., and Moore, D.S. (eds), *Statistical Decision Theory and Related Topics II*. Academic Press, New York.

Brown, L.D. 1978. *Notes on Statistical Decision Theory*. Unpublished Lecture Notes.

Büehlmann, Peter, and van de Geer, Sara. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Cai, T. Tony. 1999. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.*, **27**(3), 898–924.

Cai, T. Tony, and Zhou, Harrison H. 2009. Asymptotic equivalence and adaptive estimation for robust nonparametric regression. *Ann. Statist.*, **37**(6A), 3204–3235.

Candès, Emmanuel, and Romberg, Justin. 2007. Sparsity and incoherence in compressive sampling. *Inverse Problems*, **23**(3), 969–985.

Carter, Andrew V. 2011. *Asymptotic Equivalence of Nonparametric Experiments Bibliography*. webpage at University of California, Santa Barbara, Department of Statistics.

Carter, C.K., Eagleson, G.K., and Silverman, B.W. 1992. A comparison of the Reinsch and Speckman splines. *Biometrika*, **79**, 81–91.

Casella, George, and Strawderman, William E. 1981. Estimating a bounded normal mean. *Annals of Statistics*, **9**, 870–878.

Chatterjee, Sourav. 2009. Fluctuations of eigenvalues and second order Poincaré inequalities. *Probab. Theory Related Fields*, **143**(1-2), 1–40.

Chaumont, L., and Yor, M. 2003. *Exercises in probability*. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 13. Cambridge: Cambridge University Press. A guided tour from measure theory to random processes, via conditioning.

Chen, Scott Shaobing, Donoho, David L., and Saunders, Michael A. 1998. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, **20**(1), 33–61.

Chui, Charles K. 1992. *An Introduction to Wavelets*. San Diego: Academic Press.

Cirel'son, B.S., Ibragimov, I.A., and Sudakov, V.N. 1976. Norm of Gaussian sample function. Pages 20–41 of: *Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory*. Lecture Notes in Mathematics, 550.

Cleveland, William S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Cohen, A. 1966. All admissible linear estimates of the mean vector. *Annals of Mathematical Statistics*, **37**, 456–463.

Cohen, A., Daubechies, I., Jawerth, B., and Vial, P. 1993a. Multiresolution analysis, wavelets, and fast algorithms on an interval. *Comptes Rendus Acad. Sci. Paris (A)*, **316**, 417–421.

Cohen, A., Dahmen, W., and Devore, R. 2000. Multiscale Decompositions on Bounded Domains. *Transactions of American Mathematical Society*, **352**(8), 3651–3685.

Cohen, Albert. 1990. Ondelettes, analyses multirésolutions et filtres miroir en quadrature. *Annales Institut Henri Poincaré, Analyse Non Linéaire*, **7**, 439–459.

Cohen, Albert, and Ryan, Robert. 1995. *Wavelets and Multiscale Signal Processing*. Chapman and Hall.

Cohen, Albert, Daubechies, Ingrid, and Vial, Pierre. 1993b. Wavelets on the Interval and Fast Wavelet Transforms. *Applied Computational and Harmonic Analysis*, **1**, 54–81.

Coifman, R.R., and Donoho, D.L. 1995. Translation-Invariant De-Noising. In: Antoniadis, Anestis (ed), *Wavelets and Statistics*. Springer Verlag Lecture Notes.

Courant, R., and Hilbert, D. 1953. *Methods of Mathematical Physics, Volume 1*. Wiley-Interscience.

Cover, Thomas M., and Thomas, Joy A. 1991. *Elements of Information Theory*. Wiley.

Daubechies, I. 1988. Orthonormal Bases of Compactly Supported Wavelets. *Comm. Pure and Applied Math.*, **41**, 909–996.

Daubechies, I. 1992. *Ten Lectures on Wavelets*. CBMS-NSF Series in Applied Mathematics, no. 61. Philadelphia: SIAM.

DeVore, R., and Lorentz, G.G. 1993. *Constructive Approximation*. Springer Verlag.

DeVore, Ronald A. 1998. Nonlinear approximation. Pages 51–150 of: *Acta numerica, 1998*. Acta Numer., vol. 7. Cambridge: Cambridge Univ. Press.

Diaconis, Persi, and Stein, Charles. 1983. *Lectures on Statistical Decision Theory*. Unpublished Lecture Notes.

Diaconis, Persi, and Ylvisaker, Donald. 1979. Conjugate Priors for Exponential Families. *Annals of Statistics*, **7**, 269–281.

Diaconis, Persi, and Zabell, Sandy. 1991. Closed form summation for classical distributions: variations on a theme of de Moivre. *Statist. Sci.*, **6**(3), 284–302.

Donoho, D. L., and Johnstone, I. M. 1994a. Ideal Spatial Adaptation via Wavelet Shrinkage. *Biometrika*, **81**, 425–455.

Donoho, D. L., and Johnstone, I. M. 1994b. Minimax risk over ℓ_p-balls for ℓ_q-error. *Probability Theory and Related Fields*, **99**, 277–303.

Donoho, D. L., and Johnstone, I. M. 1995. Adapting to unknown smoothness via Wavelet shrinkage. *J. Amer. Statist. Assoc.*, **90**, 1200–1224.

Donoho, D. L., and Johnstone, I. M. 1998. Minimax Estimation via Wavelet shrinkage. *Annals of Statistics*, **26**, 879–921.

Donoho, D. L., and Johnstone, I. M. 1999. Asymptotic Minimaxity of Wavelet Estimators with Sampled Data. *Statistica Sinica*, **9**, 1–32.

Donoho, D. L., and Liu, R. C. 1991. Geometrizing Rates of Convergence, III. *Annals of Statistics*, **19**, 668–701.

Donoho, D. L., Liu, R. C., and MacGibbon, K. B. 1990. Minimax risk over hyperrectangles, and implications. *Annals of Statistics*, **18**, 1416–1437.

Donoho, D. L., Johnstone, I. M., Hoch, C.J., and Stern, A. 1992. Maximum Entropy and the nearly black object. *J. Royal Statistical Society, Ser. B.*, **54**, 41–81. With Discussion.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. 1995. Wavelet Shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B*, **57**, 301–369. With Discussion.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. 1997. Universal Near Minimaxity of Wavelet Shrinkage. Pages 183–218 of: D., Pollard, E., Torgersen, and G.L., Yang (eds), *Festschrift for L. Le Cam*. Springer Verlag.

Donoho, David L., and Huo, Xiaoming. 2001. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory*, **47**(7), 2845–2862.

Donoho, D.L. 1992. *Interpolating Wavelet Transforms*. Tech. rept. 408. Department of Statistics, Stanford University.

Donoho, D.L. 1993. Unconditional bases are optimal bases for data compression and statistical estimation. *Applied and Computational Harmonic Analysis*, **1**, 100–115.

Donoho, D.L. 1994. Statistical Estimation and Optimal recovery. *Annals of Statistics*, **22**, 238–270.

Donoho, D.L. 1995. Nonlinear solution of linear inverse problems by Wavelet-Vaguelette Decomposition. *Applied Computational and Harmonic Analysis*, **2**, 101–126.

Donoho, D.L. 1996. Unconditional Bases and Bit-Level Compression. *Applied Computational and Harmonic Analysis*, **3**, 388–392.

Dugundji, James. 1966. *Topology*. Allyn and Bacon, Boston.

Durrett, Rick. 2010. *Probability: theory and examples*. Fourth edn. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Dym, H., and McKean, H. P. 1972. *Fourier Series and Integrals*. Academic Press.

Efroimovich, S.Yu., and Pinsker, M.S. 1984. A learning algorithm for nonparametric filtering. *Automat. i Telemeh.*, **11**, 58–65. (in Russian), translated in *Automation and Remote Control*, 1985, p 1434-1440.

Efromovich, Sam. 1999. *Nonparametric curve estimation*. Springer Series in Statistics. New York: Springer-Verlag. Methods, theory, and applications.

Efromovich, Sam, and Samarov, Alex. 1996. Asymptotic equivalence of nonparametric regression and white noise model has its limits. *Statist. Probab. Lett.*, **28**(2), 143–145.

Efron, Bradley. 1993. Introduction to "James and Stein (1961) Estimation with Quadratic Loss". Pages 437–442 of: Kotz, Samuel, and Johnson, Norman (eds), *Breakthroughs in Statistics: Volume 1: Foundations and Basic Theory*. Springer.

Efron, Bradley. 2011. *Tweedie's formula and selection bias*. Tech. rept. Department of Statistics, Stanford University.

Efron, Bradley, and Morris, Carl. 1971. Limiting the Risk of Bayes and Empirical Bayes Estimators – Part I: The Bayes Case. *J. American Statistical Association*, **66**, 807–815.

Erdélyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. 1954. *Tables of Integral Transforms, Volume 1*. McGraw-Hill.

Eubank, Randall L. 1999. *Nonparametric regression and spline smoothing*. Second edn. Statistics: Textbooks and Monographs, vol. 157. New York: Marcel Dekker Inc.

Fan, K. 1953. Minimax theorems. *Prob. Nat. Acad. Sci. U.S.A.*, **39**, 42–47.

Feldman, Israel. 1991. Constrained minimax estimation of the mean of the normal distribution with known variance. *Ann. Statist.*, **19**(4), 2259–2265.

Foster, D.P., and Stine, R.A. 1997. *An information theoretic comparison of model selection criteria*. Tech. rept. Dept. of Statistics, University of Pennsylvania.

Frazier, M., Jawerth, B., and Weiss, G. 1991. *Littlewood-Paley Theory and the study of function spaces*. NSF-CBMS Regional Conf. Ser in Mathematics, **79**. Providence, RI: American Mathematical Society.

Freedman, David. 1999. On the Bernstein-von Mises Theorem with Infinite Dimensional Parameters. *Annals of Statistics*, **27**, 1119–1140.

Friedman, J.H., and Stuetzle, W. 1981. Projection Pursuit Regression. *J. Amer. Statist. Assoc.*, **76**, 817–823.

Galambos, Janos. 1978. *The asymptotic theory of extreme order statistics*. John Wiley & Sons, New York-Chichester-Brisbane. Wiley Series in Probability and Mathematical Statistics.

Gao, Hong-Ye. 1998. Wavelet Shrinkage DeNoising Using The Non-Negative Garrote. *J. Computational and Graphical Statistics*, **7**, 469–488.

Gao, Hong-Ye, and Bruce, Andrew G. 1997. Waveshrink with firm shrinkage. *Statistica Sinica*, **7**, 855–874.

Gasser, Theo, and Müller, Hans-Georg. 1984. Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.*, **11**(3), 171–185.

Gel'fand, I. M., and Shilov, G. E. 1964. *Generalized functions. Vol. I: Properties and operations*. Translated by Eugene Saletan. New York: Academic Press.

George, Edward I., and Foster, Dean P. 2000. Calibration and Empirical Bayes Variable Selection. *Biometrika*, **87**, 731–747.

Gilbarg, David, and Trudinger, Neil S. 1983. *Elliptic Partial Differential Equations of Second Order*. Second edition edn. Springer-Verlag.

Golomb, M., and Weinberger, H. F. 1959. Optimal approximation and error bounds. Pages 117–190 of: *On Numerical Approximation*. University of Wisconsin Press.

Golub, Gene H., and Van Loan, Charles F. 1996. *Matrix Computations*. 3rd edn. Johns Hopkins University Press.

Golubev, Georgi K., Nussbaum, Michael, and Zhou, Harrison H. 2010. Asymptotic equivalence of spectral density estimation and Gaussian white noise. *Ann. Statist.*, **38**(1), 181–214.

Gorenflo, Rudolf, and Vessella, Sergio. 1991. *Abel integral equations*. Lecture Notes in Mathematics, vol. 1461. Berlin: Springer-Verlag. Analysis and applications.

Gourdin, Eric, Jaumard, Brigitte, and MacGibbon, Brenda. 1994. Global Optimization Decomposition Methods for Bounded Parameter Minimax Risk Evaluation. *SIAM Journal of Scientific Computing*, **15**, 16–35.

Grama, Ion, and Nussbaum, Michael. 1998. Asymptotic Equivalence for Nonparametric Generalized Linear Models. *Probability Theory and Related Fields*, **111**, 167–214.

Gray, R. M. 2006. Toeplitz and Circulant Matrices: A review. *Foundations and Trends in Communications and Information Theory*, **2**, 155–239.

Green, P.J., and Silverman, B.W. 1994. *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Grenander, Ulf, and Rosenblatt, Murray. 1957. *Statistical Analysis of Stationary Time Series, Second Edition published 1984*. Chelsea.

Hall, P., and Patil, P. 1993. *Formulae for mean integrated squared error of nonlinear wavelet-based density estimators*. Tech. rept. CMA-SR15-93. Australian National University. To appear, Ann. Statist.

Hall, Peter. 1979. On the rate of convergence of normal extremes. *J. Appl. Probab.*, **16**(2), 433–439.

Härdle, Wolfgang, Hall, Peter, and Marron, Stephen. 1988. How far are automatically chosen regression smoothing parameters from their minimum? (with discussion). *J. American Statistical Association*, **83**, 86–101.

Härdle, Wolfgang, Kerkyacharian, Gerard, Picard, Dominique, and Tsybakov, Alexander. 1998. *Wavelets, approximation, and statistical applications*. Lecture Notes in Statistics, vol. 129. New York: Springer-Verlag.

Hardy, G. H., and Littlewood, J. E. 1928. Some properties of fractional integrals. I. *Math. Z.*, **27**(1), 565–606.

Hart, Jeffrey D. 1997. *Nonparametric smoothing and lack-of-fit tests*. Springer Series in Statistics. New York: Springer-Verlag.

Hastie, Trevor, Tibshirani, Robert, and Wainwright, Martin. 2012. L_1 *regression?* Chapman and Hall? forthcoming.

Hastie, Trevor J., and Tibshirani, Robert J. 1990. *Generalized Additive Models*. Chapman and Hall.

Hernández, E., and Weiss, G. 1996. *A First Course on Wavelets*. CRC Press.

Hida, T. 1980. *Brownian Motion*. Springer.

Huber, Peter J., and Ronchetti, Elvezio M. 2009. *Robust Statistics*. Wiley.

Hwang, Jiunn Tzon, and Casella, George. 1982. Minimax confidence sets for the mean of a multivariate normal distribution. *Ann. Statist.*, **10**(3), 868–881.

Ibragimov, I., and Khasminskii, R. 1997. Some estimation problems in infinite-dimensional Gaussian white noise. Pages 259–274 of: *Festschrift for Lucien Le Cam*. New York: Springer.

Ibragimov, I. A., and Has′minskiĭ, R. Z. 1977. Estimation of infinite-dimensional parameter in Gaussian white noise. *Dokl. Akad. Nauk SSSR*, **236**(5), 1053–1055.

Ibragimov, I. A., and Khas'minskii, R. Z. 1980. Asymptotic properties of some nonparametric estimates in Gaussian white nose. In: *Proceedings of Third International Summer School in Probability and Mathematical Statistics, (Varna 1978), Sofia*. in Russian.

Ibragimov, I. A., and Khas'minskii, R. Z. 1982. Bounds for the risks of non-parametric regression estimates. *Theory of Probability and its Applications*, **27**, 84–99.

Ibragimov, I. A., and Khas'minskii, R. Z. 1984. On nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory of Probability and its Applications*, **29**, 18–32.

Ibragimov, I.A., and Has'minskii, R.Z. 1981. *Statistical estimation : asymptotic theory*. New York: Springer.

Ingster, Yu. I., and Suslina, I. A. 2003. *Nonparametric goodness-of-fit testing under Gaussian models*. Lecture Notes in Statistics, vol. 169. New York: Springer-Verlag.

James, W., and Stein, C. 1961. Estimation with quadratic loss. Pages 361–380 of: *Proceedings of Fourth Berkeley Symposium on Mathematical Statistics and Probability Theory*. University of California Press.

Jansen, Maarten. 2001. *Noise reduction by wavelet thresholding*. Lecture Notes in Statistics, vol. 161. New York: Springer-Verlag.

Johnson, Norman L., and Kotz, Samuel. 1970. *Distributions in Statistics: Continuous Univariate Distributions - 2*. Wiley, New York.

Johnstone, I. M. 1994. Minimax Bayes, Asymptotic Minimax and Sparse Wavelet Priors. Pages 303–326 of: Gupta, S.S., and Berger, J.O. (eds), *Statistical Decision Theory and Related Topics, V*. Springer-Verlag.

Johnstone, I. M., and Silverman, B. W. 1997. Wavelet Threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B.*, **59**, 319–351.

Johnstone, I. M., and Silverman, B. W. 2004a. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, **32**, 1594–1649.

Johnstone, I. M., and Silverman, B.W. 1990. Speed of Estimation in Positron Emission Tomography and related inverse problems. *Annals of Statistics*, **18**, 251–280.

Johnstone, Iain M. 2001. Chi Square Oracle Inequalities. Pages 399–418 of: de Gunst, M., Klaassen, C., and van der Waart, A. (eds), *Festschrift for Willem R. van Zwet*. IMS Lecture Notes - Monographs, vol. 36. Institute of Mathematical Statistics.

Johnstone, Iain M. 2010. High dimensional Bernstein-von Mises: simple examples. *IMS Collections*, **6**, 87–98.

Johnstone, Iain M., and Silverman, Bernard W. 2004b. Boundary coiflets for wavelet shrinkage in function estimation. *J. Appl. Probab.*, **41A**, 81–98. Stochastic methods and their applications.

Johnstone, Iain M., and Silverman, Bernard W. 2005a. EbayesThresh: R Programs for Empirical Bayes Thresholding. *Journal of Statistical Software*, **12**(8), 1–38.

Johnstone, Iain M., and Silverman, Bernard W. 2005b. Empirical Bayes selection of wavelet thresholds. *Ann. Statist.*, **33**(4), 1700–1752.

Kakutani, S. 1948. On equivalence of infinite product measures. *Annals of Mathematics*, **49**, 214–224.

Katznelson, Yitzhak. 1968. *An Introduction to Harmonic Analysis*. Dover.

Keller, Joseph B. 1976. Inverse problems. *Amer. Math. Monthly*, **83**(2), 107–118.

Kempthorne, Peter J. 1987. Numerical specification of discrete least favorable prior distributions. *SIAM J. Sci. Statist. Comput.*, **8**(2), 171–184.

Kneser, H. 1952. Sur un théorème fondamental de la théorie des jeux. *C. R. Acad. Sci. Paris*, **234**, 2418–2420.

Kolaczyk, Eric D. 1997. Nonparametric Estimation of Gamma-Ray Burst Intensities Using Haar Wavelets. *The Astrophysical Journal*, **483**, 340–349.

Komlós, J., Major, P., and Tusnády, G. 1975. An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, **32**, 111–131.

Kotelnikov, V. 1959. *The Theory of Optimum Noise Immunity*. McGraw Hill, New York.

Kuhn, H.W. 1953. *Review of Kneser (1952)*.

Kuo, H-H. 1975. *Gaussian Measures in Banach Spaces*. Springer Verlag, Lecture Notes in Mathematics # 463.

Kuo, Hui-Hsiung. 2006. *Introduction to stochastic integration*. Universitext. New York: Springer.

Laurent, Béatric, and Massart, Pascal. 1998. *Adaptive estimation of a quadratic functional by model selection*. Tech. rept. Université de Paris-Sud, Mathématiques.

Le Cam, L. 1986. *Asymptotic Methods in Statistical Decision Theory*. Berlin: Springer.

Le Cam, Lucien, and Yang, Grace Lo. 2000. *Asymptotics in statistics*. Second edn. Springer Series in Statistics. New York: Springer-Verlag. Some basic concepts.

LeCam, Lucien. 1955. An extension of Wald's theory of statistical decision functions. *Annals of Mathematical Statistics*, **26**, 69–81.

Ledoux, M. 1996. Isoperimetry and Gaussian Analysis. In: Bernard, P. (ed), *Lectures on Probability Theory and Statistics, Ecole d'Eté de Probabilities de Saint Flour, 1994*. Springer Verlag.

Ledoux, Michel. 2001. *The concentration of measure phenomenon*. Mathematical Surveys and Monographs, vol. 89. Providence, RI: American Mathematical Society.

Lehmann, E. L., and Casella, George. 1998. *Theory of Point Estimation*. Second edn. Springer Texts in Statistics. New York: Springer-Verlag.

Lehmann, E. L., and Romano, Joseph P. 2005. *Testing statistical hypotheses*. Third edn. Springer Texts in Statistics. New York: Springer.

Lemarié, P.G., and Meyer, Y. 1986. Ondelettes et bases Hilbertiennes. *Revista Matematica Iberoamericana*, **2**, 1–18.

Lepskii, O.V. 1991. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and its Applications*, **35**, 454–466.

Levit, B. Ya. 1980. On asymptotic minimax estimates of second order. *Theory of Probability and its Applications*, **25**, 552–568.

Levit, B. Ya. 1982. Minimax estimation and positive solutions of elliptic equations. *Theory of Probability and its Applications*, **82**, 563–586.

Levit, B. Ya. 1985. Second order asymptotic optimality and positive solutions of Schrödinger's equation. *Theory of Probability and its Applications*, **30**, 333–363.

Loader, Clive R. 1999. Bandwidth selection: Classical or plug-in? *Annals of Statistics*, **27**, 415–438.

Mallat, Stéphane. 1998. *A Wavelet Tour of Signal Processing*. Academic Press.

Mallat, Stéphane. 1999. *A Wavelet Tour of Signal Processing*. Academic Press. 2nd, expanded, edition.

Mallows, C. 1973. Some comments on C_p. *Technometrics*, **15**, 661–675.

Mandelbaum, Avi. 1984. All admissible linear estimators of the mean of a Gaussian distribution on a Hilbert space. *Annals of Statistics*, **12**, 1448–1466.

Marr, R. B. 1974. On the reconstruction of a function on a circular domain from a sampling of its line integrals. *J. Math. Anal. Appl.*, **45**, 357–374.

Marron, J. S., and Wand, M. P. 1992. Exact mean integrated squared error. *Ann. Statist.*, **20**(2), 712–736.

Massart, Pascal. 2007. *Concentration inequalities and model selection*. Lecture Notes in Mathematics, vol. 1896. Berlin: Springer. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

McMurry, Timothy L., and Politis, Dimitris N. 2004. Nonparametric regression with infinite order flat-top kernels. *J. Nonparametr. Stat.*, **16**(3-4), 549–562.

Meyer, Y. 1986. Principe d'incertitude, bases hilbertiennes et algebres d'operatéurs. *Seminaire Bourbaki*, **662**.

Meyer, Y. 1990. *Ondelettes et Opérateurs, I: Ondelettes, II: Opérateurs de Calderón-Zygmund, III:* (with R. Coifman)*, Opérateurs multilinéaires*. Paris: Hermann. English translations of Vol I. and Vols II-III (combined) published by Cambridge University Press.

Meyer, Y. 1991. Ondelettes sur l'intervalle. *Revista Matematica Iberoamericana*, **7**, 115–133.

Meyer, Yves. 1992. *Wavelets and Operators*. Vol. 1. Cambridge University Press.

Meyer, Yves, and Coifman, Ronald. 1997. *Wavelets*. Cambridge Studies in Advanced Mathematics, vol. 48. Cambridge: Cambridge University Press. Calderón-Zygmund and multilinear operators, Translated from the 1990 and 1991 French originals by David Salinger.

Mézard, Marc, and Montanari, Andrea. 2009. *Information, physics, and computation*. Oxford Graduate Texts. Oxford: Oxford University Press.

Micchelli, C. A. 1975. *Optimal estimation of linear functionals*. Tech. rept. 5729. IBM.

Micchelli, C. A., and Rivlin, T. J. 1977. A survey of optimal recovery. Pages 1–54 of: Micchelli, C. A., and Rivlin, T. J. (eds), *Optimal Estimation in Approximation Theory*. New York: Plenum Press.

Miller, A. J. 1984. Selection of subsets of regression variables (with discussion). *J. Roy. Statist. Soc., Series A*, **147**, 389–425. with discussion.

Miller, A. J. 1990. *Subset Selection in Regression*. Chapman and Hall, London, New York.

Nason, G. P. 2008. *Wavelet methods in statistics with R*. Use R! New York: Springer.

Nemirovski, Arkadi. 2000. Topics in non-parametric statistics. Pages 85–277 of: *Lectures on probability theory and statistics (Saint-Flour, 1998)*. Lecture Notes in Math., vol. 1738. Berlin: Springer.

Nikol'skii, S.M. 1975. *Approximation of Functions of Several Variables and Imbedding Theorems*. Springer, New York.

Nussbaum, M. 1996. Asymptotic Equivalence of density estimation and white noise. *Annals of Statistics*, **24**, 2399–2430.

Nussbaum, M. N. 2004. Equivalence asymptotique des Expériences Statistiques. *Journal de la Société Francaise de Statistique*, **145**(1), 31–45. (In French).

Ogden, R. Todd. 1997. *Essential wavelets for statistical applications and data analysis*. Boston, MA: Birkhäuser Boston Inc.

Okikiolu, G. O. 1971. *Aspects of the Theory of Bounded Integral Operators in L^p Spaces*. Academic Press.

Peck, J.E.L., and Dulmage, A.L. 1957. Games on a compact set. *Canadian Journal of Mathematics*, **9**, 450–458.

Peetre, J. 1975. *New Thoughts on Besov Spaces, I*. Raleigh, Durham: Duke University Mathematics Series.

Pinsker, M.S. 1980. Optimal filtering of square integrable signals in Gaussian white noise. *Problems of Information Transmission*, **16**, 120–133. originally in Russian in *Problemy Peredatsii Informatsii* **16** 52-68.

Pratt, John W. 1960. On interchanging limits and integrals. *Annals of Mathematical Statistics*, **31**, 74–77.

Prékopa, András. 1980. Logarithmic concave measures and related topics. Pages 63–82 of: *Stochastic programming (Proc. Internat. Conf., Univ. Oxford, Oxford, 1974)*. London: Academic Press.

Reed, Michael, and Simon, Barry. 1980. *Functional Analysis, Volume 1, revised and enlarged edition*. Academic Press.

Rice, John, and Rosenblatt, Murray. 1981. Integrated mean squared error of a smoothing spline. *J. Approx. Theory*, **33**(4), 353–369.

Riesz, Frigyes, and Sz.-Nagy, Béla. 1955. *Functional Analysis*. Ungar, New York.

Robbins, Herbert. 1956. An empirical Bayes approach to statistics. Pages 157–163 of: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*. Berkeley and Los Angeles: University of California Press.

Rudin, Walter. 1973. *Functional Analysis*. McGraw Hill.

Schervish, Mark J. 1995. *Theory of statistics*. Springer Series in Statistics. New York: Springer-Verlag.

Shao, Peter Yi-Shi, and Strawderman, William E. 1994. Improving on the James-Stein positive-part estimator. *Ann. Statist.*, **22**(3), 1517–1538.

Shepp, L. A. 1966. Radon-Nikodym derivatives of Gaussian measures. *Annals of Mathematical Statistics*, **37**, 321–354.

Silverman, B. W. 1984. Spline smoothing: the equivalent variable kernel method. *Annals of Statistics*, **12**, 898–916.

Simons, Stephen. 1995. Minimax theorems and their proofs. Pages 1–23 of: Du, D.-Z., and Pardalos, P.M. (eds), *Minimax and Applications*. Kluwer Academic Publishers.

Sion, M. 1958. On general minimax theorems. *Pacific Journal of Mathematics*, **8**, 171–176.

Srinivasan, C. 1973. check. *Sankhya*, **check**, check.

Stein, C. 1956. Efficient nonparametric estimation and testing. Pages 187–195 of: *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*. University of California Press, Berkeley, CA.

Stein, Charles. 1981. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, **9**, 1135–1151.

Stoffer, D.S. 1991. Walsh-Fourier Analysis and Its Statistical Applications. *Journal of the American Statistical Association*, **86**, 461–479.

Stone, C.J. 1980. Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, **8**, 1348–1360.

Strawderman, William E. 1971. Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.*, **42**(1), 385–388.

Sudakov, V. N., and Cirel′son, B. S. 1974. Extremal properties of half-spaces for spherically invariant measures. *Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)*, **41**, 14–24, 165. Problems in the theory of probability distributions, II.

Szegö, Gabor. 1967. *Orthogonal Polynomials, 3rd edition*. American Mathematical Society.

Talagrand, Michel. 2003. *Spin glasses: a challenge for mathematicians*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics], vol. 46. Berlin: Springer-Verlag. Cavity and mean field models.

Tao, Terence. 2011. *Topics in Random Matrix Theory*. draft book mansucript.

Temme, Nico M. 1996. *Special functions*. A Wiley-Interscience Publication. New York: John Wiley & Sons Inc. An introduction to the classical functions of mathematical physics.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, **58**(1), 267–288.

Tibshirani, Robert, and Knight, Keith. 1999. The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society, Series B*, **61**, 529–546.

Triebel, H. 1983. *Theory of Function Spaces*. Basel: Birkhäuser Verlag.

Triebel, H. 1992. *Theory of Function Spaces II*. Basel: Birkhäuser Verlag.

Triebel, Hans. 2006. *Theory of function spaces. III*. Monographs in Mathematics, vol. 100. Basel: Birkhäuser Verlag.

Triebel, Hans. 2008. *Function spaces and wavelets on domains*. EMS Tracts in Mathematics, vol. 7. European Mathematical Society (EMS), Zürich.

Tsybakov, A. B. 2008. *Introduction to Nonparametric Estimation*. Springer.

Tsybakov, A.B. 1997. Asymptotically Efficient Signal Estimation in L_2 Under General Loss Functions. *Problems of Information Transmission*, **33**, 78–88. translated from Russian.

van der Vaart, A. W. 1997. Superefficiency. Pages 397–410 of: *Festschrift for Lucien Le Cam*. New York: Springer.

van der Vaart, Aad. 2002. The statistical work of Lucien Le Cam. *Ann. Statist.*, **30**(3), 631–682. Dedicated to the memory of Lucien Le Cam.

Van Trees, H. L. 1968. *Detection, Estimation and Modulation Theory, Part I*. New York: Wiley.

Vidakovic, Brani. 1999. *Statistical Modelling by Wavelets*. John Wiley and Sons.

Vidakovic, Brani, and Dasgupta, Anirban. 1996. Efficiency of linear rules for estimating a bounded normal mean. *Sankhyā Ser. A*, **58**(1), 81–100.

von Neumann, John, and Morgenstern, Oskar. 1944. *Theory of Games and Economic Behavior*. FILL IN.

Wahba, G. 1978. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B.*, **40**, 364–372.

Wahba, G. 1983. Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B.*, **45**, 133–150.

Wahba, G. 1990. *Spline Methods for Observational Data*. Philadelphia: SIAM.

Wald, Abraham. 1950. *Statistical Decision Functions*. Wiley.

Wasserman, Larry. 2006. *All of nonparametric statistics*. Springer Texts in Statistics. New York: Springer.

Watson, G. S. 1971. Estimating Functionals of Particle Size Distribution. *Biometrika*, **58**, 483–490.

Wicksell, S. D. 1925. The corpuscle problem. A mathematical study of a biometric problem. *Biometrika*, **17**, 84–99.

Williams, David. 1991. *Probability with Martingales*. Cambridge University Press, Cambridge.

Wojtaszczyk, P. 1997. *A Mathematical Introduction to Wavelets*. Cambridge University Press.

Woodroofe, Michael. 1970. On choosing a delta sequence. *Annals of Mathematical Statistics*, **41**, 1665–1671.

Young, W. H. 1911. On semi-integrals and oscillating successions of functions. *Proc. London Math. Soc. (2)*, **9**, 286–324.

Zygmund, A. 1959. *Trigonometric Series, Volume I*. Cambridge University Press, Cambridge.

Zygmund, A. 2002. *Trigonometric series. Vol. I, II*. Third edn. Cambridge Mathematical Library. Cambridge: Cambridge University Press. With a foreword by Robert A. Fefferman.