Risk of Penalized Least Squares, Greedy Selection and ℓ_1-Penalization for Flexible Function Libraries

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

Doctor of Philosophy

by

Cong Huang

Dissertation Director: Andrew R. Barron

May 2008

ABSTRACT

Risk of Penalized Least Squares, Greedy Selection and ℓ_1-Penalization for Flexible Function Libraries

Cong Huang

Supervisor: Andrew R. Barron

For function estimation using penalized squared error criteria, we derive generally applicable risk bounds, showing the balance of accuracy of approximation and penalty relative to the sample size. Attention is given to linear combinations of terms from a given class (such as used in neural network models, projection pursuit regression, function aggregation and multiple linear regression). The risk bounds apply to forward stepwise selection and other relaxed greedy algorithms with penalty on the number of terms, and to ℓ_1-penalized least squares, for which we develop a fast algorithm.

ACKNOWLEDGEMENT

I would like to address my most sincere appreciation to Professor Andrew R. Barron, my dissertation advisor, for the enormous amount of help and encouragement he has given to me during the past five years. Without his instruction, I cannot imagine to finish this dissertation. I would also like to express my ardent gratitude to David Pollard, Joseph T. Chang, John W. Emerson, Hannes Leeb, Mokshay Madiman and Harry Zhou for their helpful advice during my graduate study in Yale. And finally, thanks and love to my dear girlfriend Jie Chen who makes everyday a joy.

I would like to thank for the staff and students of the Statistics Department at Yale University for providing such a superb environment for learning and doing research.

Contents

Chapter 1

Introduction

Flexible regression models are built by combining simple functional forms. Fitting such models to data in a training sample, there is a role for empirical performance criteria such as penalized squared error in selecting components of the function from a given library of candidate terms. With suitable penalty, optimizing the criterion adapts the total weights of combination or the number of components as well as the subset of terms to include. The aim is to produce function estimates which accurately predict responses for new input values with the same distribution as the sample. This generalization capability is characterized by the mean squared error as the statistical risk. In this context, our paper has several interwoven objectives:

1. To analyze performance of penalized least squares estimators with theory of acceptable penalties, such that the estimator optimizing the empirical criterion has risk characterized by a corresponding population property of tradeoff of approximation and penalty relative to the sample size.

2. To allow for flexible function fitting using linear combinations of terms selected from various large or even infinite libraries of functions.

3. To establish that a greedy term selection solves the ℓ_1 penalized squared error problem with bounds on accuracy that compare favorably with competing convex optimization algorithms for large libraries.

4. To demonstrate that different estimators, one based on forward stepwise selection with penalty on the number of terms and another with penalty on the ℓ_1 norm of coefficients, both achieve approximately the same risk, for target functions that have control on the ℓ_1 norm of coefficients and for functions in the interpolation classes between these and all of L_2.

1.1 Problem Description and Setting

Suppose data $(X_i, Y_i)_{i=1}^n$ are independently drawn from the distribution of X, Y. To produce predictions of the real-valued response Y from its input X, the target regression function $f^*(x) = E[Y|X = x]$ is to be estimated. It is assumed the function f^* has magnitude bounded by a constant B. The domain $\mathcal{X}$ is an arbitrary measurable space. The error $\epsilon = Y - f^*(X)$ is assumed to satisfy moment conditions: namely, that $\text{var}(\epsilon|X)$ is bounded by a constant denoted σ^2 and higher order moments satisfy a Bernstein condition, as given in Chapter 2.

The empirical average squared error of a function f as a candidate fit to the observed data is $\|Y-f\|_n^2 = (1/n)\sum_{i=1}^n (Y_i - f(X_i))^2$. Given a collection of functions $\mathcal{F}$, a penalty $\text{pen}_n(f)$, $f \in \mathcal{F}$, and data, a penalized least squares estimator $\hat{f}$ arises by optimizing $\|Y-f\|_n^2 + \text{pen}_n(f)/n$. For parameterized functions f_β with penalty $\text{Pen}_n(\beta)$, it is a plug-in rule $\hat{f} = f_{\hat{\beta}}$, where $\hat{\beta}$ optimizes $\|Y - f_\beta\|_n^2 + \text{Pen}_n(\beta)/n$, accommodated by setting $\text{pen}_n(f) = \inf\{\text{Pen}_n(\beta) : f_\beta = f\}$.

Estimators $\hat{f}$ optimizing the criterion are then truncated to produce the final fit $T\hat{f}$,

where $Tf = \min\{B', |f|\}\text{sgn}(f)$ truncates the functions at a level B' chosen to be not less than B. Let $\|f\|^2 = \int f^2(x)P(dx)$ be the squared $L_2(P)$ norm, where P is the distribution of X. With the truncation, using the squared $L_2(P)$ loss, and taking the expectation with respect to the distribution of the data, the statistical risk of an estimator is $E\|T\hat{f} - f^*\|^2$, a function of f^* we wish to analyze.

1.2 Conditions on penalties

1.2.1 Motivation

Concerning objective (1) we determine in Chapter 2 a condition for a penalty, such that an estimator $\hat{f}$ approximately achieving the minimum of

$$\|Y - f\|_n^2 + \frac{\text{pen}_n(f)}{n}$$

will satisfy a corresponding risk inequality with positive δ,

$$\mathbb{E}\|T\hat{f} - f^*\|^2 \leq (1+\delta) \inf_{f \in \mathcal{F}} \left\{ \|f - f^*\|^2 + \mathbb{E}\frac{\text{pen}_n(f)}{n} \right\} + \frac{C_\delta}{n}. \tag{1.1}$$

This shows the accuracy of the estimator is controlled by the tradeoff between the squared L_2 approximation error and the penalty divided by n. When the target f^* is in $\mathcal{F}$ then $f = f^*$ yields a risk bound based on $\text{pen}_n(f^*)/n$. If $\text{pen}_n(f^*)$ is large compared to n, then the minimization will favor approximations f of smaller $\text{pen}_n(f)$ to achieve an appropriate balance in (1.1).

The penalty condition we develop has an information-theoretic flavor. Given $\mathcal{F}$, we require that there be a countable approximating set $\tilde{\mathcal{F}}$ of representors $\tilde{f}$, which we call

a variable-distortion, variable-complexity cover of $\mathcal{F}$, and a complexity function $L_n(\tilde{f})$, interpretable as a description length for $\tilde{f}$, with the property that for each f in $\mathcal{F}$ there is an $\tilde{f}$ in $\tilde{\mathcal{F}}$ such that $\text{pen}_n(f)$ is not less than $\gamma L_n(\tilde{f}) + \Delta_n(f, \tilde{f})$, where γ is a constant (depending on B, B' and δ) and $\Delta_n(f, \tilde{f})$ is given as a suitable empirical measure of distortion (based on sums of squared errors). Accordingly, accurate estimators are obtained when functions f near the target are close to functions of moderate complexity relative to n.

Associated with property (1.1), the quantity $\inf_{f\in\mathcal{F}}\{\|f-f^*\|^2 + \mathbb{E}\frac{\text{pen}_n(f)}{n}\}$ is an *index of resolvability* of f^* by functions in $\mathcal{F}$ with sample size n. This terminology is in accord with usage for minimum description length (MDL) procedures with countable $\mathcal{F}$ in [12], [6] and for Bayes predictors in [9]. Our penalty condition yields for uncountable $\mathcal{F}$ an extension of previous conditions based on information-theoretic complexity. Parallel to our penalized least squares work is the development of analogous conclusions for penalized log likelihood [13], extending risk analysis of MDL criteria to uncountable families of candidate functions.

1.2.2 Library $\mathcal{F}$ as a linear span

Suppose $\mathcal{F}$ is the linear span of a library $\mathcal{H}$ of candidate terms $h(x)$. These terms arise as candidate basis functions for approximating the target. Evaluated at $(X_i)_{i=1}^n$, the library yields a data-set of explanatory variables for regression, which may include transformations and interactions among original variables.

The library cardinality is denoted M or sometimes p and, for possibly infinite libraries of correlated variables, the effective cardinality M_n is the size of an empirical cover of $\mathcal{H}$ at a suitable precision. For libraries of metric dimension d, at precision $1/n$ the effective cardinality is of order n^d. Two examples, among several we discuss, are libraries of terms for product splines with variable knot locations and sigmoidal neural nets, with d the

number of original input variables.

For flexible function approximation, a large library of candidate terms is natural. Typically M_n is much larger than the sample size n; though $\log M_n$ is arranged to be small compared to n, as $(\log M_n)/n$ arises in the resolvability for the penalties we study. The dependence on library size only through the logarithm allows for very large libraries. Such large size increases the opportunity to find accurate linear combinations of moderate penalty. This tradeoff is facilitated by approximation properties of sparse linear combinations. Sparse combinations have a number of terms m small compared to both the library size and the sample size.

For subsets of size m, the log cardinality $\log \binom{M_n}{m}$ is near $m \log(M_n/m)$ plus a term of order m. This log cardinality plays a role in our analysis, both in directly giving the main term of a penalty based on the number of terms m as in Chapter 5 and, through a similar expression, as a key part of the demonstration in Chapter 4 of the validity of a penalty based on the ℓ_1 norm of coefficients.

One might think to favor the ℓ_1 penalization because of the convexity of optimization. However, we shall see that approximately the same accuracy is available by fast forward stepwise algorithms with either subset-size or ℓ_1 penalties. There are circumstances slightly favoring one of these penalties over the other, or even a combination of the two, as discussed after development of our main results.

1.3 Some commonly used penalty classes

Penalties for linear combinations are typically based on norms of the coefficients, and come in three varieties: (A) penalties on the ℓ_1 norm, or other ℓ_q norms with $0 < q < 2$, quantifying linear combinations that are close to sparse; (B) penalties based primarily

on the number terms m in linear combination, that is, the ℓ_0 norm of the coefficients, exactly capturing sparsity; and (C) penalties that capture traditional notions of roughness, through weighted ℓ_2 norms, which, in some cases, correspond to norms on derivatives or to reproducing kernels. Cases (A) and (B) make flexible use of possibly large libraries; whereas, for case (C) there is less flexibility as we shall discuss.

Functions in the linear span of $\mathcal{H}$ take the form $f(x) = f_\beta(x) = \sum_h \beta_h h(x)$ where each such $\beta = (\beta_h)_{h\in\mathcal{H}}$ has some subset of $\mathcal{H}$ within which the coefficients β_h are non-zero.

1.3.1 Penalty determined by the ℓ_1 norm of coefficients

The penalty is based on the weighted ℓ_1 norm $\|\beta\|_1 = \sum_h |\beta_h| a_h$ of the coefficients, for given positive weights a_h, usually taken to be a constant or a norm of h. For given a_h, the $\mathcal{L}_{1,\mathcal{H}}$ norm of functions f in the linear span of $\mathcal{H}$ is defined by $\|f\|_{1,\mathcal{H}} = \inf\{\|\beta\|_1 : f_\beta = f\}$, where the infimum is over all representations of f in $\mathcal{F}$. We consider $Pen_n(\beta)/n = \lambda\|\beta\|_1$ and determine acceptably small values for the multiplier $\lambda = \lambda_n$ for information-theoretic validity of the penalty. The ℓ_1 penalty is in agreement with Tibshirani's *LASSO* [79] and Chen and Donoho's *basis pursuit* [32, 33] as well as a precursor in Barron [5, 8] which involved optimization on a discrete net. With the continuum of parameters β, use of the ℓ_1 penalty is equivalent to solving the convex optimization

$$\min_\beta \left\{ \|Y - \sum_h \beta_h h\|_n^2 + \lambda\|\beta\|_1 \right\}. \tag{1.2}$$

The choice of λ should be typically of order $\sqrt{(\log M_n)/n}$, which allows for large libraries, though if the library size is of order $\sqrt{n}$ or smaller, then λ may be set to be of order $1/\sqrt{n}$, without the log factor. In Chapter 4 we show that such penalties are proper and hence (1.1) holds (adjusted by a negligible term of order $\frac{\log M_n}{n}$), yielding a risk perfor-

mance characterized by the corresponding tradeoff between squared approximation error plus the ℓ_1 penalty:

$$\inf_{f \in \mathcal{F}} \{\|f - f^*\|^2 + \lambda \|f\|_{1,\mathcal{H}}\}. \tag{1.3}$$

In verifying the requirement, the ℓ_1 penalty arises in a variable-complexity cover as a bound on a complexity-based term $\frac{m}{n} \log \frac{M}{m} + O(\frac{m}{n})$ plus a squared approximation error $\|f\|^2_{1,\mathcal{H}}/m$, at a near-optimal choice of $m = \|f\|_{1,\mathcal{H}} \sqrt{n/\log M}$. It yields risk comparable to what is achieved by stepwise selection algorithms with a subset-size penalty, in accordance with objective (4) above.

We show improvement of the rate based on covering properties of the library. Indeed, let ε_m be the radius of the cells in an empirical L_2 cover of $\mathcal{H}$ with m representors. We extend the validity of the criterion by showing all λ not less than $\varepsilon_{m_n} \sqrt{(\log M_n)/n}$ provide proper penalties. This expression arises as an optimized tradeoff of $\frac{m}{n} \log M_n$ and an improved squared approximation error bound $\varepsilon_m^2 \|f\|^2_{1,\mathcal{H}}/m$, achieved at an m_n now of order smaller than $\sqrt{n/\log M_n}$ for finite-dimensional libraries. This refinement makes a noticeable improvement when the dimension is low. For example, when $\mathcal{H}$ is the collection of indicators of half spaces in R^d, the rate matches (to within a log factor) what is best possible for functions of bounded variation on the line ($d = 1$). When the dimension is high the risk is of order close to what is achieved with the simpler form of λ of order $\sqrt{(\log M_n)/n}$. For large d, the rates are close to the minimax lower bounds of Yang and Barron [87] for variation balls.

This improved rate of ℓ_1 penalized least squares is also achieved by all-subset selection and by our new greedy implementation of ℓ_1 penalized least squares, while in contrast the bounds available for stepwise selection algorithms with subset-size penalty lock in at the slightly slower rate.

For a number of additional papers on ℓ_1 penalized least squares, see Bunea, Tsybakov,

Wegkamp [24, 25] and references cited therein. The result that the risk of ℓ_1 penalized least squares is bounded by the population tradeoff between approximation error and ℓ_1 norm is new. Along with our present manuscript, current work in this direction is a manuscript by Zhang [89] independent of our development, plus our risk bounds for ℓ_1 penalized log-likelihood in [13]. Fascinating prior results are in the cited [24, 25], from which we learned some detail of choice of λ in the small M case. They do not seek results of the type (1.1) in the form (1.3). Rather, imposing additional conditions (e.g., that pairs of candidate terms in the library are nearly orthogonal), they examine ℓ_1 penalized least squares as a subset selection rule for which, after some analysis, they apply conclusions of Birgé and Massart [21, 20] to obtain a bound that is not a minimum of approximation error plus ℓ_1 norm, but rather a minimum of approximation error plus a multiple of log subset size. In a roundabout way, bounds in the form (1.3) could follow from their bounds, by invoking approximation properties for suitable subsets, but that approach is limited to libraries that satisfy their additional conditions.

Estimators other than ℓ_1 penalized least squares have been shown also to achieve risk of order $\sqrt{(\log M)/n}$ governed by ℓ_1 properties. These include the aggregation method of Juditsky and Nemirovski [55], the exponentiated gradient on-line learning algorithm of Kivinen and Warmuth [56], and greedy algorithms with a line of development traced in Barron, Cohen, Dahmen, and Devore [11] and further developed here. Moreover, under specific assumptions on the noise distribution, one can use Cesàro averages of Bayes predictive density estimates to obtain general risk bounds analogous to (1.1) as in Barron [9] with applications to various regression settings, including ℓ_1 control on risk, as developed in the sequence of papers by Yang and Barron [87], Yang [84], Catoni [28], and Tsybakov [80]. In problems of nonparametric classification, Koltchinskii and Panchenko [57] have related results, including analogous improvements from covering properties with ℓ_1 control.

1.3.2 Penalty based on the ℓ_0 norm of the coefficients

We show the validity of penalties for which the main terms takes the form

$$C\left\{\log\binom{M_n}{m} + m\log n\right\}, \tag{1.4}$$

where C is a constant depending on B' and δ. This penalty has a more direct description-length interpretation, in accordance with the MDL principle [14, 13]. Representors are given by first describing a subset of size m out of M_n in the library cover, using code-length $\log\binom{M_n}{m}$ and then $m\log n$ represents additional description length required to represent truncated linear combinations to suitable precision. Setting aside for now secondary terms and the effect of greedy algorithms, our bound on the risk of the estimator takes the following form

$$(1+\delta)\inf_m \inf_{f\in\mathcal{F}_m}\left\{\|f-f^*\|^2 + C\frac{1}{n}\log\binom{M_n}{m} + C\frac{m}{n}\log n\right\}, \tag{1.5}$$

where $\mathcal{F}_m$ is the class of all m term linear combinations from the given library, with implications for the risk depending on the accuracy of approximation of f^* by members of $\mathcal{F}_m$. Details of conclusions of this type are in Chapter 5.

Some discussion puts the conclusion for subset size penalties in the context of past related work. Concerning library covering properties, as we have said, we allow parameterized candidate basis functions yielding a library of finite metric dimension (such as arise for neural nets or variable-knot splines) for which M_n is of order n^d. Typically, the dimension d corresponds to the parameter dimension of these basis functions. Accordingly, md is the total number of parameters used in representing an m term linear combination. Consequently, the main part of our penalty of order $m\log M_n$ is of order $md\log n$, equal to the total number of parameters times a $\log n$ factor, in accordance with typical MDL criteria.

For library covering using an L_∞ norm for smoothly parameterized basis functions, analogous risk properties for subset size penalties are in [10]. In contrast, we use empirical covering properties, as required for certain neural net or product spline expansions in which jumps are permitted (i.e. libraries consisting of indicators of half spaces or rectangles). Unlike [10] we do allow greedy algorithms.

The present work began with the thesis [29], and the conference reports [30, 31], building on the conclusions in [8] and [59]. Recent similar conclusions for subset size penalties and greedy algorithms are in [11], with a requirement that Y to be bounded. Here we allow unbounded noise, we have substantially improved constants, the risk theory applies more generally (not just to subset size penalty), we allow greedy optimization over the library (not just the cover), and we extend the greedy term selection theory to show that a variant of it also solves the ℓ_1-penalized least squares problem.

The use of empirical covering properties gives the indicated advantage of greater range of applicability, though to achieve it we do make a couple of concessions concerning the form of the penalty.

One concession is that we are content to arrange for the second term to be $m \log n$ rather than m times a constant. For our generalization, this enables us to avoid the more elaborate chaining argument and associated large constants in [10]. The need for the larger $\log \binom{M_n}{m}$ term somewhat mutes the debate between whether the other term should be of order $m \log n$ as in MDL and BIC criteria [71, 72, 14], [74] or of order m as in AIC and C_p criteria [1, 2],[64]. If one wants to have the whole penalty be a constant times m, risk analysis showing such to be acceptable relies on the log of the number of subsets considered of each size m to be of order not bigger than m, as developed in [76], [60], [83], [10, 21, 20], [3, 4]. As discussed also in [86], that is an undesirable restriction when addressing flexible high-dimensional modeling. In contrast, when either allowing all subsets of M_n terms of size m, or when building up the subsets by greedy algorithms,

addition of a constant times $\log \binom{M_n}{m}$, which is typically of larger order than m alone, does make for a valid penalty.

The other concession for infinite libraries of possibly correlated variables in managing the effect of the empirical cover is that we need some mild control on the size of the coefficients of linear combination. This is arranged through the inclusion of an additional ℓ_1 penalty, with a very small multiplier λ so that its effect is secondary to that of the subset size penalty.

Any ℓ_q norm with $0 \leq q < 2$ may be used to characterize linear combinations suitably approximated by sparse subsets. This property is most clear for the case of orthonormal basis functions for which the best subsets correspond to the sets of largest coefficients. For libraries with correlated variables, it is the ℓ_1 case that permits probabilistic arguments to cleanly demonstrate complexity and accuracy tradeoffs and to establish favorable computation time bounds for stepwise procedures. From risk conclusions in the ℓ_0 and ℓ_1 penalty cases, interpolation space properties then show appropriate order of risk for other function regularities.

1.3.3 Penalty determined by the weighted ℓ_2 norm of coefficients

Quadratic penalties such as L_2 norms on derivatives (Sobelev norms) and reproducing kernel norms are a third type. Early advocacy of such quadratic penalties is in Good and Gaskins [49, 50], de Montricher Tapia and Thompson [40], [77] and Wahba [82]. Functional analysis tools for analysis of quadratic penalties in a Hilbert space setting are developed in Cox and O'Sullivan [35, 36]. Metric entropy methods are developed for the case of smoothness constraints in Nemirovski, Polyak, and Tsybakov [66] and for minimum contrast estimators and sieves in Birgé and Massart [18, 19]. Shen [75] analyzes penalized criteria by an argument reducing consideration to functions with penalty not more than the value at the target, which is then addressed by the metric entropy methods

of the constrained case. That approach is limited to the case that the target function has a finite penalty value. Further developments in this direction are in Cucker and Smale [38]. In contrast our approach using variable-distortion variable-complexity covers avoids need for such reduction to the constrained case.

Quadratic penalties correspond to weighted ℓ_2 norms on coefficients in basis function expansions (e.g., for kernel methods these are the eigenfunctions) used to define classes of functions. Balls of functions determined by these norms correspond to ellipsoids in the coefficients.

A rigidity has been demonstrated for quadratic penalties that limits their performance potential and reduces the priority for their analysis compared to more flexible procedures. For a weighted ℓ_2 norm, consider the order of the basis functions induced the values of the weights. Rather than locking in one such class of functions, one can adapt to achieve the appropriate level of risk for all quadratic norms that preserve this order, simply by using least squares on the first m terms with a penalty (of order m) to select this number of terms, as shown in [10]. Indeed, the risk, bounded by $\min_m\{\|f_m^* - f^*\|^2 + Cm/n\}$ where f_m^* is the L_2 projection onto the first m terms, is within a constant factor of the minimax rate simultaneously for all these ellipsoids. This prioritization of the leading terms, with no rate advantage in ellipsoids for consideration of more general subsets, is the rigidity to which we refer.

This rigidity contributes to slow rates for function estimation in high-dimensional settings in traditional smoothness classes. Indeed, for multivariate formulations with domain in $\mathbb{R}^d$ suppose the library consists of products of one-dimensional basis functions in a specified order. Using all the products of the first k of each, the number of terms $m = k^d$ is exponential in d, requiring exponentially large sample size n. Though one may adapt the value of k by penalized least squares with penalty a constant multiple of k^d, to be minimax optimal in rate, e.g. for all Sobolev classes indexed by the order of smoothness s, the

minimax rates are disappointingly slow, of order $(1/n)^{2s/(2s+d)}$.

In contrast, if the target function has a moderate ℓ_1 norm of coefficients, for the library of $M = K^d$ possible terms consisting of products up to order K, then in accordance with the bounds from cases (A) or (B), with estimators based on ℓ_1 penalization or greedy subset selection picking a sparse subset of much smaller size m, the risk is bounded by order $\|f^*\|_{1,\mathcal{H}}\sqrt{(d \log K)/n}$, which does not require an exponentially large sample size to provide an accurate estimate. Fourier norm conditions that produce this favorable behavior are developed in [7, 8], [10], [11]. The reason for the extra flexibility with general subsets or the ℓ_1 penalty criteria is the attention these give to the basis functions that the data show most matter to the target, rather than to those prespecified to be important according to a weighted ℓ_2 control.

1.4 Risk on an evaluative sample

The analysis involves comparison of discrepancies between sample and population values of average squared error, and it is facilitated by consideration of both the training sample and a future sample at which the predicted responses are to be evaluated.

Let $\|f\|_{\underline{X}}^2 = \frac{1}{n}\sum_{i=1}^n f^2(X_i)$, also denoted $\|f\|_n^2$, be the squared $L_2(P_n)$ norm, where P_n is the empirical distribution for the input data $\underline{X} = (X_i)_{i=1}^n$, and likewise let $\|f\|_{\underline{X}'}^2 = \frac{1}{n}\sum_{i=1}^n f^2(X_i')$ be the squared $L_2(P_n')$ norm, where P_n', the empirical distribution for an independent copy $\underline{X}' = (X_i')_{i=1}^n$. The symmetrized empirical squared norm is $\|f\|_{\underline{X},\underline{X}'}^2 = [\|f\|_{\underline{X}}^2 + \|f\|_{\underline{X}'}^2]/2$, also denoted $\|f\|_{2n}^2$.

The statistical risk $\mathbb{E}\|T\hat{f} - f^*\|_{\underline{X}'}^2$ measures how well the estimator trained on $\underline{X}, \underline{Y}$ generalizes to an independent $\underline{X}'$ with the same distribution as $\underline{X}$. In a traditional setting, when forming the estimator from the training data, one does not have advance knowledge of the $\underline{X}'$ at which it will be evaluated and the risk matches $\mathbb{E}\|T\hat{f} - f^*\|^2$, using the squared

$L_2(P)$ loss. In addition to this traditional setting, we treat Vapnik's transductive inference setting [81], in which, when constructing $\hat{f}$, one makes use of advance knowledge of the random $\underline{X}'$ at which it is to be evaluated, and some slight advantages for reduced penalty are developed for this case. Then the risk $\mathbb{E}\|T\hat{f} - f^*\|^2_{\underline{X}'}$ is not expressible as the expected squared $L_2(P)$ norm. Nevertheless, the same techniques bound the risk in either setting working with the general risk expression $\mathbb{E}\|T\hat{f} - f^*\|^2_{\underline{X}'}$.

1.5 Summary of penalty analysis

Our task is to determine choices of penalties such that an estimator $\hat{f}$ approximately achieving the minimum of $\|Y - f\|_n^2 + \frac{\text{pen}_n(f)}{n}$ will satisfy

$$\mathbb{E}\|T\hat{f} - f^*\|^2_{\underline{X}'} \leq (1+\delta) \inf_{f \in \mathcal{F}} \left\{ \|f - f^*\|^2 + \mathbb{E}\frac{\text{pen}_n(f)}{n} \right\} + \frac{C_\delta}{n}. \tag{1.6}$$

Lower bounds on $\text{pen}_n(f)$ are established that permit this risk characterization.

Our underpinnings of penalized least squares begin with the case of a countable collection $\mathcal{F}$. Building on earlier work [6], for a specified constant γ, the penalty may be chosen to be $\text{pen}_n(f) = \gamma L(f)$ where $\sum_f e^{-L(f)} \leq 1$, so that $L(f)$ is interpretable as a complexity (in nats) or $e^{-L(f)}$ is interpretable as a prior probability of f. Then the counterpart of (1.6) holds showing that the risk is bounded by the index of resolvability $\inf_{f \in \mathcal{F}} \left\{ ||f - f^*||^2 + \frac{\gamma L(f)}{n} \right\}$ specifying the tradeoff between approximation and the complexity relative to the sample size.

Such countable collections can be effective in theory. For instance one may assign $L(f)$ to be the minimal log-cardinality (metric entropy) of covers of function classes, plus a description length of such classes, thereby simultaneously achieving the minimax optimal rates for each such class in accordance with [87].

In practice it is more customary to envision optimization over uncountable families $\mathcal{F}$ with continuous parameters (such as coefficients of linear combinations), optimized by penalized least squares. As we have said, we demonstrate that the desired risk behavior holds for such uncountable $\mathcal{F}$, provided good discrete approximations $\tilde{\mathcal{F}}_n$ can be formulated together with complexities $L_n(\tilde{f})$. Indeed, for satisfaction of the risk inequality (1.6), we show it suffices that the penalty $\text{pen}_n(f)$ be not less than

$$\min_{\tilde{f}\in\tilde{\mathcal{F}}_n} \left\{ \Delta_n(f,\tilde{f}) + \gamma L_n(\tilde{f}) \right\}. \tag{1.7}$$

expressing the distortion and complexity tradeoff. The distortion takes the form

$$\begin{aligned} \Delta_n(f,\tilde{f}) \;\; = \;\; & \left[\sum_{i=1}^n (Y_i - \tilde{f}(X_i))^2 - \frac{1}{c}\sum_{i=1}^n (f^*(X_i') - \tilde{f}(X_i'))^2 \right] \\ & - \left[\sum_{i=1}^n (Y_i - f(X_i))^2 - \frac{1}{c}\sum_{i=1}^n (f^*(X_i') - f(X_i'))^2 \right], \end{aligned} \tag{1.8}$$

with $c > 1$, where, to take advantage of boundedness properties, the $\tilde{f}$ may be replaced by $T\tilde{f}$ and the last occurrence of f may be replaced by Tf. This distortion between f and $\tilde{f}$ captures the essence of what is needed in the problem: there is a discrepancy between error on the training data and error on the independent copy and $\Delta_n(f,\tilde{f})$ is the difference in these discrepancies at $\tilde{f}$ and f. A slackening of the penalty requirement allows different values of c in the first and second lines of the definition of the distortion.

The tradeoff between distortion and complexity is analogous to what occurs in rate-distortion theory in Information Theory [17], [34]. The countable collection $\tilde{\mathcal{F}}_n$ of functions $\tilde{f}$, which we have called a variable-distance, variable-complexity cover of $\mathcal{F}$, would be called in Information Theory a variable-distortion, variable-rate code with codelength $L_n(\tilde{f})$. To verify that proposed choices of $\text{pen}_n(f)$ satisfy the property (1.7), it is equivalent to show for each $\underline{X}', \underline{X}, \underline{Y}$ that for each function f in $\mathcal{F}$ there is a representor $\tilde{f}$ in $\tilde{\mathcal{F}}_n$

for which the value of the distortion plus complexity $\Delta_n(f, \tilde{f}) + \gamma L_n(\tilde{f})$ is not more than $\text{pen}_n(f)$.

The sets $\tilde{\mathcal{F}}_n$ may depend on $\underline{X}$, $\underline{X}'$. We note also that $\Delta_n(f, \tilde{f})$ depends on f^*, and if one desires, the cover $\tilde{\mathcal{F}}_n$ may also depend on f^*. Key to our use of this analysis is that, even though the distortion has the indicated form, we arrange covers such that the requirement (1.7) is satisfied in such a way that the penalty that does not depend on f^*. Moreover, though for the transductive formulation dependence of the penalty on both $\underline{X}$ and $\underline{X}'$ is acceptable, for the traditional formulation we require also that the penalty not depend on $\underline{X}'$, which we may facilitate in some cases by replacing expression (1.7) by its expectation with respect to the distribution of the $\underline{X}'$.

In particular, by constructions of this type, when $\mathcal{F}$ is the linear span of a library, we demonstrate the existence of suitable $\tilde{\mathcal{F}}_n$ for which the familiar penalties based on $\log \binom{M}{m_f}$ or $\lambda \|f\|_{1,\mathcal{H}}$ do satisfy the requirements, where m_f is the number of non-zero terms in f. Armed with such we prove suitable risk properties for penalized least squares estimators.

For non-smooth classes (as arise with libraries $\mathcal{H}$ of functions with jumps, such as indicators of half spaces or rectangles), data-dependent covers and empirical norms are essential. In these cases the cover $\tilde{\mathcal{F}}_n$ we use depends on $\underline{X}$ and $\underline{X}'$. Nevertheless, for smooth function classes the collection $\tilde{\mathcal{F}}_n$ may be fixed and the difference in the empirical squared norms between $f^* - f$ and $f^* - \tilde{f}$ on $\underline{X}'$ in (1.7) may be replaced by its expectation inside the minimization.

In interpreting the distortion in the penalty condition, a convenient and slightly more general expression for the distortion is

$$\Delta_n(f, \tilde{f}) = \sum_{i=1}^{n} (Y_i - \tilde{f}(X_i))^2 - \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \frac{1}{c} \text{diff}_n(f, \tilde{f}) \qquad (1.9)$$

with

$$\text{diff}_n(f,\tilde{f}) = \frac{1}{1+\delta'}\sum_{i=1}^{n}(f^*(X_i')-f(X_i'))^2 - \sum_{i=1}^{n}(f^*(X_i')-\tilde{f}(X_i'))^2.$$

The form (1.8) above corresponds to $\delta' = 0$. The modification with $\delta' > 0$ slackens the penalty condition, but inflates the resulting risk bound by $(1+\delta')$. It allows a non-negative bound on $\text{diff}_n(f,\tilde{f})$ equal to $\frac{1}{\delta'}\sum_{i=1}^{n}(f(X_i')-\tilde{f}(X_i'))^2$ to be used in its place. This empirical squared distance is a more conventional distortion measure, it is independent of f^*, and it is helpful in dealing with the effect of truncation.

A natural question is whether the penalized squared error $\sum_{i=1}^{n}(Y_i-f(X_i))^2+\text{pen}_n(f)$ for f in the uncountable set $\mathcal{F}$ inherits the total description length interpretation of $\sum_{i=1}^{n}(Y_i-\tilde{f}(X_i))^2 + \gamma L_n(\tilde{f})$ for $\tilde{f}$ in the countable set $\tilde{\mathcal{F}}$ in accordance with [6]. Our penalty requirement is equivalent to requiring that the penalized squared error $\sum_{i=1}^{n}(Y_i-f(X_i))^2+\text{pen}_n(f)$ is greater than or equal to

$$\min_{\tilde{f}\in\tilde{\mathcal{F}}}\left\{\sum_{i=1}^{n}\left(Y_i-\tilde{f}(X_i)\right)^2+\gamma L_n(\tilde{f})+\frac{1}{c}\text{diff}_n(f,\tilde{f})\right\}. \tag{1.10}$$

With the formulation in which $\text{diff}_n(f,\tilde{f})$ is replaced by a non-negative quantity, we see that the penalized squared error in the uncountable case exceeds the penalized squared error in the countable case and hence inherits the description length interpretation for an information-theoretically valid criterion. The presence of the additional distortion term links the risk performance that f will have on future data $\underline{X}'$ to the performance more directly establishable for its representor $\tilde{f}$.

1.6 Example Libraries

Now we turn to discussion of Objective (2). We call attention to various flexible function models that are addressed by our analysis, depending on the choice of the library $\mathcal{H}$. In all

these cases, input vectors x are in $\mathbb{R}^d$. The simplest library $\mathcal{H}$ consists of the coordinate functions $\{x_1, \ldots, x_d\}$, in which case the library size M coincides the with number of variables d. Even in this simplest case there are challenges when $M \gg n$, as in the case of microarray gene data when the number of variables is hundreds of thousands. If there be only dozens of variables, very large libraries sizes M still arise by considering all interactions up to some order as in polynomial regression. Also, continuously parameterized libraries (of infinite cardinality) arise naturally in certain statistical models of interest to us. The framework of our paper can be used in the following models:

A. **Ordinary Linear Regression**: The library of candidate predictors is $\mathcal{H} = \{x_1, x_2, \ldots, x_d\}$ from which a data-driven subset is selected.

B. **Polynomial Regression**: The library $\mathcal{H}$ is the collection of all polynomial terms $x_{j_1}^{k_1} x_{j_2}^{k_2} \cdots x_{j_I}^{k_I}$ in subsets of the variables up to some maximum interactive order I and polynomial degree, often fit by forward stepwise selection. Strategies for building up a polynomial fit include those described in [78] and in [15].

C. **Projection Pursuit Regression**: $\mathcal{H}$ consists of all ridge functions $h(x) = \phi(a^T x)$ where $a \in \mathbb{R}^d$ provides the direction and ϕ is a scalar function constrained only in its smoothness [48], [53],[54].

D. **Neural Networks**: $\mathcal{H}$ consists of the functions $h_{a,b}(x) = \phi(a^T x - b)$ where $\phi(z)$ is a fixed function with distinct limits as $z \to \infty$ and $z \to -\infty$, usually taken to be an increasing function. Such functions ϕ are called sigmoids and linear combinations of them are called single hidden layer artificial neural networks [39], [7, 8], [59].

E. **Flexible Frequency Sinusoids**: $\mathcal{H}$ consists of functions $h_{a,b}(x) = \phi(a^T x - b)$ where $\phi(z) = \sin(z)$. Linear combinations of such and algorithms for fitting them were first considered (in the $d = 1$ case) in 1795 by Prony [70]. Combined L_1 and L_2

moment conditions on the Fourier transform of a function permit accurate estimates of it even if d is large [10].

F. **Multivariate Additive Regression Splines (MARS)**: $\mathcal{H}$ consists of functions $(x_{j_1} - t_1)^+ \cdot (x_{j_2} - t_2)^+ \cdots (x_{j_I} - t_I)^+$ with adaptation of the subset of variables $x_{j_1}, \ldots, x_{j_I}$ and the knot locations $t_1, \ldots, t_I$ and the interaction order I for each such term, linearly combined in the MARS algorithm [45]. Spline fitting by ℓ_1 penalized least squares is developed in [67, 68].

G. **Multiple Additive Regression Trees (MART)**: $\mathcal{H}$ consists of regression trees, which are linearly combined in the MART algorithm [46, 47, 43].

H. **Wavelet Basis Pursuit**: $\mathcal{H}$ consists of the union of several wavelet basis expansions (an over-determined system) as arises for instance by including all Daubechies wavelets up to some order [63], [32, 33]. Related libraries include ridgelets [26] and curvelets [27].

I. **Function Aggregation**: $\mathcal{H}$ consists of M functions $\hat{f}_1, \ldots, \hat{f}_M$ each already fit to an initial part of the data. A linear combination of these chosen to fit the rest of the data is used to aggregate into one improved estimator [65, 55], [85], [24, 25].

The risk analysis in this paper applies to these sorts of models with suitable search strategies for choosing h in $\mathcal{H}$. The presence of iterative algorithms to select terms from the library and to assign weights of linear combination is a common feature of flexible function methods associated with the above list of libraries.

It is critical to these applications that the theory is not only appropriate for fully optimized subsets, but also for approximate optimization by various iterative algorithms. This gives rise to our consideration of Objective (3).

1.7 Greedy Selection Summary

Iterative term selection such as *forward stepwise selection* is implemented for ordinary linear regression in nearly every statistical package. It is applicable to libraries of size that permit exhaustive consideration of every h in $\mathcal{H}$ for each step of forward selection. Moreover, iterative selection is also applicable to possibly infinite libraries provided an optimization strategy is identified to apply at each step of selection.

The essence of forward selection and other greedy algorithms is that a succession of terms $\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_m$ are selected in a data-driven fashion, as well as weights of linear combination $\hat{\beta}_{1,m}, \hat{\beta}_{2,m}, \ldots, \hat{\beta}_{m,m}$ leading to an estimator of the form

$$\hat{f}_m(x) = \sum_{j=1}^{m} \hat{\beta}_{j,m} \hat{h}_j(x)$$

where at step m, the first $\hat{h}_1, \ldots, \hat{h}_{m-1}$ are given, and a new term $\hat{h}_m$ in $\mathcal{H}$ is chosen to be linearly combined with the previous terms.

1.7.1 Pure greedy algorithms

A restrictive form of greedy algorithm is Jones' *pure greedy* [54], Mallat's *matching pursuit* [63] and Friedman's *L_2-boosting* [46] (also called stagewise regression) choosing β and a new term h from $\mathcal{H}$ to become $\hat{\beta}_m$ and $\hat{h}_m$, respectively, the new fit is restricted to be of the form

$$\hat{f}_m(x) = \hat{f}_{m-1}(x) + \beta h(x),$$

where h is chosen by least squares or to maximize the correlation with the residuals from $\hat{f}_{m-1}$. The projection pursuit algorithm of [48] is a pure greedy algorithm. Limitations of pure greedy algorithms are given in the succession of work of Temlyakov and his colleagues [41, 58, 61], who show that even if the target is in the convex hull of the library,

the pure greedy approximation converges to it, in squared L_2 norm, at a slow rate assured to only be as good as $(1/m)^{1/3}$, improved slightly to $(1/m)^{11/31}$, and indeed for some dictionaries there is such a target for which the rate is not better than $(1/m)^{.54}$, compared to the $1/m$ rate of the algorithms we discuss next. Further properties of pure greedy are explored in [88].

1.7.2 Relaxed greedy algorithm

Jones [54] provides a notion of a class of relaxed greedy algorithms with additional flexibility found to be essential for more desirable accuracy properties. The key idea is to endow the algorithm with the opportunity to adjust (typically downweigh) the weight of the previous fit via consideration of the form

$$(1-\alpha)\hat{f}_{m-1}(x) + \beta h(x). \tag{1.11}$$

For instance α, β and h may be optimized by least squares. Per his definition (and variants in [7] and [59]), given $\hat{f}_{m-1}$ and, if desired, given also $\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_{m-1}$, a procedure for choosing a new function $\hat{f}_m$, not necessarily of the form (1.11), is said to be a *relaxed greedy* algorithm if it produces sum of squared error not worse than at specific α_m, β_m and h_m as detailed further in Chapter 3.

Among algorithms that satisfy this requirement is standard *forward stepwise selection*, in which given $\hat{h}_1, \ldots, \hat{h}_{m-1}$, all the coefficients $\hat{\beta}_{1,m}, \ldots, \hat{\beta}_{m,m}$ and the new term $\hat{h}_m$ are chosen so that the linear combination is optimal by least squares. The relaxation requirement may be thought of as a minimal amount of back-fitting. There is freedom in the choice of the new term $\hat{h}_m$ for the updated relaxed fit (1.11). As we have said, it is permitted to be the choice in $\mathcal{H}$ of least angle with residuals from $\hat{f}_{m-1}$ or the choice which is best in linear combination with $\hat{f}_{m-1}$. Similar purposes are accomplished by the *LARS (least*

angle regression) algorithm [42], the term selection algorithms in [67, 68], the $B - Lasso$ algorithm in [90], and the coordinate optimization algorithms in [44]. In accordance with the computational theory we give here, we believe that some of these algorithms would be improved by inclusion of the relaxation parameter.

Another relaxed greedy variant valid for bounded libraries is to pick h to come within any constant factor (say $1/2$) of the maximum inner product with residuals (again, not to be purely added but rather with an 1-α relaxation factor). This freedom of optimization (only coarsely to within a constant factor) may be useful for certain high-dimensional libraries, where exact optimization is problematic.

1.7.3 ℓ_1-*Penalized greedy pursuit (LPGP)* algorithm

A relaxed greedy variant introduced in this paper is at each step to select the new term h along with α and β to best improve the ℓ_1 penalized sum of squares. We call it ℓ_1-*penalized greedy pursuit (LPGP)*. To be specific, at each step m, given the previous $\hat{f}_{m-1}$, and $v_{m-1} = \sum_{j=1}^{m-1} |\hat{\beta}_{j,m-1}| a_{\hat{h}_j}$, we choose h in $\mathcal{H}$ and coefficients α and β to form $\hat{f}_m = (1-\alpha)\hat{f}_{m-1} + \beta h$ and $v_m = (1-\alpha)v_{m-1} + |\beta| a_h$, iteratively to satisfy

$$\begin{aligned} & ||\hat{f}_m - Y||_n^2 + \lambda v_m \\ = & \min_{h \in \mathcal{H}, \beta \in \mathbb{R}, 0 \le \alpha \le 1} \left\{ ||(1-\alpha)\hat{f}_{m-1} + \beta h - Y||_n^2 + \lambda \left((1-\alpha) v_{m-1} + |\beta| a_h \right) \right\} \end{aligned}$$

where λ is a nonnegative constant and, as explained previously, a_h are weights which may be set to be the empirical L_2 norm $\|h\|_n$. This variant of relaxed greedy algorithms combines benefits of both term selection and ℓ_1 penalization. With this LPGP algorithm, for each h, one has two simple quadratics to solve for the scalars α and β, one for the possibility of positive β and one for the possibility of negative β, followed by a decision of which h and with which sign.

Advantageous properties of these algorithms are provided. In ordinary linear regression from a small size library, forward selection is a somewhat notorious method, criticized because it need not pick out the best subsets of each size m, leading in such settings to favoring of *backward selection* or *all-subset selection* procedures. However, the latter two procedures are not suitable computationally for large libraries needed for flexible fitting of functions of many variables. It is comforting then that the relaxed greedy algorithm theory establishes a sense in which forward selection procedures nearly optimize the squared error. Indeed, as shown in Chapter 3, for every data set $(X_i, Y_i)_{i=1}^n$, we have

$$\|Y - \hat{f}_m\|_n^2 \leq \inf_f \left\{ \|Y - f\|_n^2 + \frac{4\|f\|_{1,\mathcal{H}}^2}{m} \right\}, \tag{1.12}$$

and for the ℓ_1-penalized greedy pursuit estimates, we obtain

$$\|Y - \hat{f}_m\|_n^2 + \lambda \|\hat{f}_m\|_{1,\mathcal{H}} \leq \inf_f \left\{ \|Y - f\|_n^2 + \lambda \|f\|_{1,\mathcal{H}} + \frac{4\|f\|_{1,\mathcal{H}}^2}{m} \right\}. \tag{1.13}$$

The right sides of these bounds quantify a tradeoff between average squared error of fit and ℓ_1 norm of coefficients. It is a desirable bound when the target is close to functions f with finite $\|f\|_{1,\mathcal{H}}$. The validity of these bounds requires that, in forming the norm $\|\beta\|_1 = \sum_h |\beta_h| a_h$, the weights a_h are not less than $\|h\|_n$.

We use the bound (1.13) in two ways. In the first instance we keep λ very small (possibly 0) and emphasize the role of the number of terms m in selecting the subset.

In the second instance we choose $\lambda \geq \epsilon_{m_n} \sqrt{\frac{C \log M_n}{n}}$ emphasizing the role of the ℓ_1-penalty while picking m as large as we like. In this case (1.13) quantifies how *ℓ_1-penalized greedy pursuit* approximately achieves the ℓ_1-penalized least squares optimization. Thus it is an alternative algorithm for solving Tibshirani's LASSO [79] or equivalently Chen and Donoho's basis pursuit [32, 33]. An advantage of our analysis is that the term $4\|f\|_{1,\mathcal{H}}^2/m$ in (1.13) bounds the computational accuracy, quantifying how close to the minimum the

algorithm achieves.

Let $A_{f,m} = \|Y - \hat{f}_m\|_n^2 - \|Y - f\|_n^2$ be the empirical average difference between the squared error of the m term fit and the squared error of a particular comparator f. The inequality (1.12) corresponds to the result for relaxed greedy fits that $A_{f,m} \leq 4\|f\|_{1,\mathcal{H}}^2/m$ for all data sets and all m and all f.

1.7.4 Resolvability risk bound for estimators formed by greedy algorithms

Our risk conclusions allow for forward selection or other relaxed greedy computations as follows. We give conditions on the library $\mathcal{H}$ and on a penalty $\text{pen}_n(m)$, with the penalized criterion providing the stopping rule $\hat{m}$, such that the truncated estimator $T\hat{f} = T\hat{f}_{\hat{m}}$ has risk satisfying

$$\mathbb{E}\|T\hat{f} - f^*\|^2 \leq (1+\delta) \min_m \inf_{f \in \mathcal{F}_m^{\text{co}}} \left\{ \|f^* - f\|^2 + \frac{\text{pen}_n(m)}{n} + \mathbb{E}A_{f,m} \right\} + \frac{C_\delta}{n}, \quad (1.14)$$

where $\mathcal{F}_m^{\text{co}}$ is a comparison set for each m. For subset size control the primary term in the penalty takes the form $C \log \binom{M_n}{m}$ as previously discussed. In the case of all-subset regression, $\mathcal{F}_m^{\text{co}}$ is taken to be $\mathcal{F}_m$, the collection of all m-term linear combinations, yielding $\mathbb{E}A_{f,m} \leq 0$, and then the bound (1.14) becomes the bound (1.5) given in case (B) above expressing the trade-off between the approximation error of the best size subset and the penalty. For greedy algorithms, we use $\mathbb{E}A_{f,m} \leq 4\|f\|_{1,\mathcal{H}}^2/m$ (where now if empirical norms are used in defining the weights a_h, they are replaced on the right side by their expectations) and the comparison class is taken to be all of $\mathcal{F}$, that is, all functions which are linear combinations of terms in $\mathcal{H}$, leading to the risk bound

$$(1+\delta) \min_m \inf_f \left\{ \|f^* - f\|^2 + \frac{\text{pen}_n(m)}{n} + \frac{4\|f\|_{1,\mathcal{H}}^2}{m} \right\} + \frac{C_\delta}{n}. \quad (1.15)$$

In some cases (e.g., with near-orthogonality of the members of $\mathcal{H}$), forward stepwise algorithms should perform even better than indicated, at least for comparison functions in $\mathcal{F}_m$, which could produces a better performance than (1.15) in accordance with the extent to which $EA_{f,m}$ is less than $4\|f\|^2_{1,\mathcal{H}}/m$.

In the bound (1.15) for relaxed greedy selections with $pen_n(m) = Cm \log M_n$, for each f optimizing m yields $m(f) = \|f\|_{1,\mathcal{H}}\sqrt{n/(C \log M_n)}$, so equivalent to (1.15) we have

$$E\|T\hat{f} - f^*\|^2 \leq (1+\delta) \inf_f \left\{ \|f^* - f\|^2 + \|f\|_{1,\mathcal{H}} \sqrt{\frac{C \log M_n}{n}} \right\} + \frac{C_\delta}{n}. \qquad (1.16)$$

The penalization procedure using the data based selection $\hat{m}$ achieves this performance without advance knowledge of the best $m(f)$. One may try to fix $m = \sqrt{n/(C \log M_n)}$, though the bound would then have the wrong order of bevavior of $\|f\|_{1,\mathcal{H}}$ with a square instead of the first power. Moreover, if $A_{f,m}$ happens to be smaller than $4\|f\|^2_{1,\mathcal{H}}/m$ then the best m will be smaller and the risk bound expressed in (1.14) correspondingly better than the bound in (1.16). Thus it is better to use the data-based choice of $\hat{m}$ via penalized least squares instead of fixing $m = \sqrt{n/(C \log M_n)}$.

As we have said the bound (1.16) for these greedy algorithms is the same order of risk that can be achieved directly from certain types of ℓ_1 penalized least squares, as shown in Chapter 4. Refinements show the improvement by ε_{m_n} in the risk for both the ℓ_1 penalized least squares estimator and the all-subset selection estimator. This gives a slightly better level of performance for ℓ_1 penalized greedy pursuit than presently available for the greedy algorithms that don't incorporate the ℓ_1 penalty.

1.8 Additional Computational Concerns

We discuss issues regarding library search strategies. Direct use of forward stepwise selection or other relaxed greedy algorithms entails exhaustive consideration of every h in $\mathcal{H}$ for each iteration. With present computers, practicality of such computation restricts the size of the libraries to be not more than several million candidate terms. Though such cardinality strains computational resources, it is not an obstacle to our theory since the risk depends on the ratio of the logarithm of the cardinality to the sample size. Such libraries include those that arise from basis expansions of polynomial, spline, trigonometric, or wavelet variety, including interactions expressed through products of the one-dimensional basis functions, which leads to a manageable number of candidate terms when there are only a handful of original variables. But that number of candidate terms grows exponentially with the number of original variables d. Thus, when there are more than a handful of such variables, we have the ubiquitous but unwieldy situation in which the number of candidate basis functions is vastly greater than what can be considered by algorithms that seek to perform computations for every candidate for each step. Again while the computation is problematic, the statistical risk theory is not, provided $d\frac{\log n}{n}$ is small.

To address this computational difficulty, certain algorithms impose greater restriction on the search in the greedy algorithm. These include forward selection in polynomial and spline fitting in which a candidate new term is restricted to increment the form of an existing term. In polynomial fitting, such increments consist of increasing the degree of one of the variables by one in an existing term, as in the MAPS algorithm [15]; whereas, in multivariate linear spline fitting by the MARS algorithm [45], such increments consist of multiplying an existing term by a new factor of the form $(x_j - t)^+$ for some j and t. Likewise for regression trees there are restrictions in the CART [23] and MART [46, 47, 43] algorithms in which new partitions for piecewise constant regression are restricted to be

a recursive refinement of an existing partition. These algorithms are fast, but a limitation of existing theory (including ours) is that we lack understanding of the approximation capabilities to quantify their resolvability or risk. Certainly such estimates have favorable properties if, at each step, every member of the more complete library of (polynomial, spline, or piecewise constant) terms were considered. But that is not what existing algorithms are capable of doing in the case of exponentially large libraries.

Alternative computational tactics arise for libraries of functions $h_w(x)$ that are parameterized smoothly through a parameter vector w of moderate dimension $d_{\mathcal{H}}$. Such functions arise in the terms used in neural nets (in which w controls the orientation and gain of a sigmoid), in sinusoids (in which w is a frequency vector), in ridgelets (in which w determines the frequency and orientation), and in splines (in which w is the vector of knot locations). From this perspective the problem at each iterative m of a greedy algorithm is that of optimization of a function $l(w)$ which takes either the form $\|Y - (1-\alpha)\hat{f}_{m-1} - \beta h_w\|_n^2$ or, for bounded libraries, an empirical inner product between the residuals $Y - \hat{f}_{m-1}$ and h_w. For local search strategies, such as the gradient backpropagation algorithm [73] for least squares fitting of neural nets or other nonlinearly parameterized terms, even in the greedy term selection case, it is not known if the local optimum provides a theoretically satisfactory substitute to global optimization. Stochastic search strategies (such as Markov Chain Monte Carlo or simulated annealing) are designed to attempt to sample w from a density proportional to $\exp(-l(w)/\tau)$ for some temperature τ eventually small enough that the distribution is likely to produce nearly optimized w globally. It is a currently active topic of research to determine conditions on the form of $h_w(x)$ and the Markov chain steps such that the stochastic search sampling is provably accurate in a moderate number of computation steps, while retaining the flexibility of representation of a nonlinearly parameterized library. Some steps in this direction are in [16].

1.9 Layout of the Dissertation

This dissertation is organized as follows. In Chapter 2 we develop the general risk inequality for penalized least squares estimators. In Chapter 3, we develop establish the computational accuracy results for relaxed greedy computations, including the new ℓ_1-penalized greedy algorithm. In Chapter 4, we provide risk analysis for ℓ_1 penalized estimators. In Chapter 5, the results are applied to obtain risk bounds for all-subset selection, forward stepwise regression and other relaxed greedy computations. Concluding discussion and examples are give in Chapter 6 and 7 respectively. Some lemmas are relegated to the appendix.

Chapter 2

General Risk Bounds

The goal of this chapter is to obtain a resolvability bound on risk for general penalized least squares estimators. While the computation bounds hold for arbitrary data sets $(X_i, Y_i)_{i=1}^n$, the risk bounds are developed in the following context.

2.1 Assumption and setting

The following setting (B) is used throughout the dissertation. **Setting (B).** Data $(X_i, Y_i)_{i=1}^n$ are independently drawn from the distribution of (X, Y). The target function (or signal) is $f^*(x) = E[Y|X = x]$ and it is assumed to be bounded by a constant B. The error (or noise) is $\epsilon = Y - f^*(X)$ and it is assumed to satisfy the moment assumption (M).

Instead of restricting Y to be bounded, we allow the following.

Assumption (M) (Bernstein's moment condition). The error $\epsilon = Y - f^*(X)$ has a conditional distribution given X which satisfies the moment condition that for some positive constant h_{Bern} not depending on X,

$$\mathbb{E}[|\epsilon|^k|X] \leq \frac{\text{var}(\epsilon|X)}{2} k!\, h_{Bern}^{k-2},$$

for $k \geq 2$, with variance $\text{var}(\epsilon|X) \leq \sigma^2$ for all X for some finite σ^2.

Assumption (M) is satisfied in particular if Y is bounded or if the distribution of ϵ has tails that decay exponentially fast. Assumption (M) corresponds to the finiteness of certain moment generating functions of ϵ, that is, $D_1 = \mathbb{E}e^{|\epsilon|/\nu} < \infty$ for $\nu > h_{Bern}$. We also exhibit improvements in the conditions and conclusions that hold when ϵ is sub-Gaussian, that is, for some constant ν, a moment generating function of ϵ^2 is finite, i.e., $D_2 = \mathbb{E}e^{\epsilon^2/\nu} < \infty$; or when it is bounded, that is, $|\epsilon| \leq c_0$.

We work with a collection of functions $\mathcal{F}$. Our first result will assume a uniform bound B' on candidate functions. Extension to remove this boundedness constraint by using a truncation technique is provided later in this chapter.

Given a set $\mathcal{F}$, $\hat{f}$ is a penalized least squares estimator or approximate penalized least squares estimator if it satisfies the inequality

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(X_i))^2 + \frac{\text{pen}_n(\hat{f})}{n} \leq \inf_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \frac{\text{pen}_n(f)}{n} + A_f\right\}, \quad (2.1)$$

where A_f is a non-negative quantity. Here $\text{pen}_n(f)$ and A_f are permitted to depend on the data $\underline{X}$ and $\underline{Y}$.

The quantity A_f may be thought of as an index of the computational accuracy of approximate optimization. It is not to be neglected. The computational accuracy achievable by certain algorithms of interest is intertwined with the degree to which targets can be approximated from both approximation-theoretic and statistical risk standpoints. Building on the work of the present section, a similar formulation is investigated in Chapter 5, in which $A_{f,m}$ is indexed by the number of algorithm steps m, as for instance in the case of greedy algorithms, and the penalized criterion is used to adapt this number of steps.

We now give tools for development of the resolvability bound on risk. The case that $\mathcal{F}$ is countable is a starting point. Analysis for countable $\mathcal{F}$ and bounded Y is in [6]. From

both statistics and engineering standpoints, it is awkward to force a user to construct a discretization of his space of functions to which the optimization would be restricted. We overcome this difficulty to extend to uncountable $\mathcal{F}$. We also remove the boundedness condition of Y in our theorem.

2.2 Symmetrization Approach

The target f^* is not necessarily in $\mathcal{F}$. To each f in $\mathcal{F}$, there corresponds a function $\rho : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which assigns to (X, Y), the relative loss

$$\rho(X, Y) = \rho_f(X, Y) = (Y - f(X))^2 - (Y - f^*(X))^2. \tag{2.2}$$

To ease aspects of the analysis, we imagine a hypothetical independent copy $\underline{X}', \underline{Y}'$ of the data-set $\underline{X}, \underline{Y}$. Except in transductive analysis (where the penalty is allowed to depend on known $\underline{X}'$), we do not allow the penalty or the estimator $\hat{f}$ to depend on this copy data. The empirical loss with respect to the training data is denoted by $P_n(\rho) = \frac{1}{n}\sum_{i=1}^n \rho(X_i, Y_i)$ and that with respect to the independent copy is $P_n'(\rho) = \frac{1}{n}\sum_{i=1} \rho(X_i', Y_i')$. Using $Y = f^*(X) + \epsilon$, note that

$$\begin{aligned} \rho(X, Y) &= (f(X) - f^*(X))^2 - 2\epsilon(f(X) - f^*(X)) \\ &= g_1(X) - 2\epsilon g_2(X). \end{aligned}$$

where $g_1(X) = g_{1,f}(X) = (f^*(X) - f(X))^2$ and $g_2(X) = g_{2,f}(X) = f(X) - f^*(X)$. Let $\hat{g}_1(X) = g_{1,\hat{f}}(X)$ and $\hat{g}_2(X) = g_{2,\hat{f}}(X)$ and likewise $\hat{\rho} = \rho_{\hat{f}}$.

Because $\hat{f}$ is selected to (approximately) minimize the penalized empirical average squared error, the value $P_n(\hat{\rho})$ on the training sample tends to be smaller than the risk of $\hat{f}$, whereas its squared error $P_n'(\hat{\rho})$ on the independent sample is an unbiased estimate of its

risk. Indeed, since $\mathbb{E}(\epsilon_i'|X_i') = 0$ and ϵ_i' is independent of $\underline{X}, \underline{Y}$, we have $\mathbb{E}\epsilon_i'\hat{g}_2(X_i') = 0$. Hence the expected value of $P_n'(\hat{\rho})$ is

$$\mathbb{E}P_n'(\hat{\rho}) = \mathbb{E}P_n'(\hat{g}_1) = \mathbb{E}\|\hat{f} - f^*\|_{\underline{X}'}^2, \tag{2.3}$$

which measures the risk or generalization error of $\hat{f}$ that we study, reducing to $\mathbb{E}\|\hat{f} - f^*\|^2$ in the traditional setting. The heart of the idea is to control the empirical discrepancy $P_n'(\rho_f) - cP_n(\rho_f)$ between the loss on the hypothetical future data and the loss on the training data for a constant c near 1. Towards this end, one may seek a quantity $\mathcal{L}_n(f)$ to satisfy

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left\{P_n'(\rho_f) - cP_n(\rho_f) - \frac{c\mathcal{L}_n(f)}{n}\right\} \leq 0. \tag{2.4}$$

Working with a closely related empirical discrepancy $P_n'(g_{1,f}) - cP_n(\rho_f)$, which avoids need of further consideration of ϵ_i', we seek $\mathcal{L}_n(f)$ to satisfy

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left\{P_n'(g_{1,f}) - cP_n(\rho_f) - \frac{c\mathcal{L}_n(f)}{n}\right\} \leq 0. \tag{2.5}$$

Once either (2.4) or (2.5) holds, a similar inequality also holds for any data-based selection of $\hat{f}$ in $\mathcal{F}$, yielding an upper bound for the risk

$$\mathbb{E}P_n'(\rho_{\hat{f}}) = \mathbb{E}P_n'(g_{1,\hat{f}}) \leq c\,\mathbb{E}\left(P_n(\rho_{\hat{f}}) + \frac{\mathcal{L}_n(\hat{f})}{n}\right). \tag{2.6}$$

Then if $\hat{f}$ is the penalized least squares estimator, optimizing $P_n(\rho_{\hat{f}}) + \frac{\mathcal{L}_n(\hat{f})}{n}$ or approximately optimizing it within a specific accuracy, we obtain the desired risk bound

$$\mathbb{E}\|\hat{f} - f^*\|^2 \leq \inf_{f\in\mathcal{F}}\left\{c\,\mathbb{E}\left(P_n(\rho_f) + \frac{\mathcal{L}_n(f)}{n} + A_f\right)\right\}, \tag{2.7}$$

noting, for any fixed f, that $\mathbb{E}P_n(\rho_f) = \|f - f^*\|^2$.

We turn our attention to the determination of suitable $\mathcal{L}_n(f)$ to control the empirical discrepancy as in (2.5). First, as we will show in the next section, when $\mathcal{F}$ is countable, choices of the form $\mathcal{L}_n(f) = \gamma L_n(f)$ proportional to complexities satisfying the Kraft inequality $\sum_{f\in\mathcal{F}} e^{-L_n(f)} \leq 1$ are acceptable. Then for the general case, with $\mathcal{F}$ is not restricted to be countable, we define proper penalties to be those for which there is a suitable relationship between the penalized discrepancy for f in $\mathcal{F}$ and corresponding quantities for $\tilde{f}$ in some countable set $\tilde{\mathcal{F}}$. The essence for a penalty $\text{pen}_n(f)$, equivalent to (1.7), is that there exists $\tilde{\mathcal{F}}$ and $L_n(\tilde{f})$ satisfying the Kraft inequality with

$$\sup_{f\in\mathcal{F}} \left\{ \frac{1}{c} P'_n(g_{1,f}) - P_n(\rho_f) - \frac{\text{pen}_n(f)}{n} \right\} \leq \sup_{\tilde{f}\in\tilde{\mathcal{F}}} \left\{ \frac{1}{\tilde{c}} P'_n(g_{1,\tilde{f}}) - P_n(\rho_{\tilde{f}}) - \frac{\gamma L_n(\tilde{f})}{n} \right\}, \tag{2.8}$$

where $c \geq \tilde{c} > 1$. Then, in the uncountable case, the validity of the essential property (2.5) and hence its associated risk bound is immediately inherited from the validity in the countable case. In some cases, we will allow a slackening of this requirement on the penalty by allowing the inequality (2.8) to hold in expectation rather than point-wise.

2.3 Theorem for the countable $\mathcal{F}$

So we address the matter of showing for the countable case that complexity penalties do indeed satisfy the property (2.5). These complexity assignment $L(f) = L_n(f)$ may depend on n but for simplicity we often drop that index. First we think of $\tilde{\mathcal{F}}$ and $L(f)$ as fixed (not depending on any of the data). Subsequently, we will allow symmetric dependance of $\tilde{\mathcal{F}}$ and $L(f)$ on $\underline{X}$, $\underline{X}'$ as an aid in verification of (2.8) for penalties $\text{pen}_n(f)$ depending on $\underline{X}$.

2.3.1 Symmetric empirical process and two lemmas

A starting point is to examine the symmetric empirical process $P'_n(\rho) - P_n(\rho)$. If Y is bounded, then one can obtain bounds on $P'_n(\rho) - P_n(\rho)$ directly (as in [59] and [11]). For unbounded Y it is better to consider g_1 and g_2 separately. Define $\mathcal{G}_1 = \{g_{1,f}(\cdot) : f \in \tilde{\mathcal{F}}\}$ and $\mathcal{G}_2 = \{g_{2,f}(\cdot) : f \in \tilde{\mathcal{F}}\}$. There is a one-to-one correspondence of $\tilde{\mathcal{F}}$ with $\mathcal{G}_1$ and with $\mathcal{G}_2$. Therefore, we can define $\{L(g_1) : g_1 \in \mathcal{G}_1\}$ and $\{L(g_2) : g_2 \in \mathcal{G}_2\}$ according to $\{L(f) : f \in \tilde{\mathcal{F}}\}$. Since $\mathbb{E}P'_n(\hat{\rho}) = \mathbb{E}P'_n(\hat{g}_1)$, to bound the risk, the essence of the analysis is to demonstrate that $P'_n(\hat{g}_1)$ cannot be much larger than $P_n(\hat{g}_1)$, which is related to $P_n(\hat{\rho})$ provided $\frac{1}{n}\sum_{i=1}^n \epsilon_i \hat{g}_2(X_i)$ is not much greater than 0. Then we add bounds from these two sources of error together to give us a general risk bound for $\hat{f}$.

Two simple lemmas are tools in obtaining our risk bound. These differ from standard empirical process analysis primarily in the use of variable complexity. Also no chaining is invoked for the results we seek here. Lemma 2.1 provides a probability bound, uniformly over functions g in a countable class $\mathcal{G}$, on the differences between empirical means weighted by the complexity of g plus a multiple of the empirical variance. Lemma 2.2 provides a corresponding probability bound for weighted empirical averages of products of ϵ and functions g. In both lemmas, we make use of the inequality

$$2ab \leq \gamma a^2 + \frac{1}{\gamma} b^2 \tag{2.9}$$

for all $\gamma > 0$.

Lemma 2.1 *Let $(\underline{X}, \underline{X}') = (X_1, \ldots, X_n, X'_1, \ldots, X'_n)$ where $\underline{X}'$ is an independent copy of the data $\underline{X}$ and where $(X_1, \ldots, X_n)$ are component-wise independent but not necessarily identically distributed. A countable function class $\mathcal{G}$ and complexities $L(g)$ satisfying*

$\sum_{g\in\mathcal{G}} e^{-L(g)} \le 1$ *are given. Then for arbitrary positive* u *and* γ*, we have*

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}} \frac{P_n'(g) - P_n(g)}{u + \frac{\gamma}{n}L(g) + \frac{1}{2\gamma}s^2(g)} \ge 1\right\} \le \exp\left(-\frac{nu}{\gamma}\right), \tag{2.10}$$

where $s^2(g) = \frac{1}{n}\sum_{i=1}^n (g(X_i) - g(X_i'))^2$*. Moreover,*

$$\mathbb{E}\sup_{g\in\mathcal{G}}\left\{P_n'(g) - P_n(g) - \frac{\gamma L(g)}{n} - \frac{1}{2\gamma}s^2(g)\right\} \le 0. \tag{2.11}$$

Proof: The proof of the first inequality uses Hoeffding's inequality and a symmetry between $P_n(g)$ and $P_n'(g)$ together with the union of events bound. This inequality is equivalent to saying that $\sup_{g\in\mathcal{G}}\left\{P_n'(g) - P_n(g) - \frac{\gamma L(g)}{n} - \frac{1}{2\gamma}s^2(g)\right\}$ is stochastically less than an exponential random variable of mean γ/n. Accordingly its expectation is not more than γ/n. The second conclusion states that the expectation is actually not more than 0. See the appendix for details. ∎

Remark: For uncountable classes $\mathcal{G}$, if one has a countable $\tilde{\mathcal{G}}$ for which an analog of (2.8) holds

$$\begin{aligned} &\sup_{g\in\mathcal{G}}\left\{P_n'(g) - P_n(g) - \frac{\text{pen}_n(g)}{n} - \frac{1}{2\gamma}s^2(g)\right\} \\ \le\ &\sup_{\tilde{g}\in\tilde{\mathcal{G}}}\left\{P_n'(\tilde{g}) - P_n(\tilde{g}) - \frac{\gamma L_n(\tilde{g})}{n} - \frac{1}{2\gamma}s^2(\tilde{g})\right\}, \end{aligned} \tag{2.12}$$

then the same inequalities (2.10) and (2.11) are still valid for the uncountable $\mathcal{G}$ with $\text{pen}_n(g)$ in place of $\gamma L_n(g)$. As we have said, the condition (2.8) is just right for our purpose when the functions are bounded, so we do not make use of (2.12) here.

Lemma 2.2 *Assume* $\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ *are conditionally independent random variables given* $(X_i)_{i=1}^n$ *that have conditional mean zero and satisfy Bernstein's moment conditions. A countable class* $\mathcal{G}$ *and associated complexities* $L(g)$ *satisfying* $\sum_{g\in\mathcal{G}} e^{-L(g)} \le 1$ *are*

given. Assume a bound K, such that $|g(x)| \leq K$ for all g in $\mathcal{G}$. Then

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}} \frac{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i)}{u+\frac{\gamma}{n}L(g)+\frac{1}{An}\sum_{i=1}^{n}g^2(X_i)} \geq 1\right\} \leq \exp\left(-\frac{nu}{\gamma}\right) \tag{2.13}$$

where A and u are arbitrary positive constants, and $\gamma = A\sigma^2/2 + Kh_{Bern}$. Moreover,

$$\mathbb{E}\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i) - \frac{\gamma}{n}L(g) - \frac{1}{An}\sum_{i=1}^{n}g^2(X_i)\right\} \leq 0. \tag{2.14}$$

Proof: The proof of the first claim uses a Bernstein type inequality. The second claim is by the same device as for Lemma 2.1. See Lemma 8.2 in the appendix. ∎

2.3.2 Resolvability risk bound in the countable case

We now show how these Lemmas are used to obtain a risk bound in the countable case. First we apply Lemma 2.1 to the countable class $\mathcal{G}_1$ to bound the contribution to the risk from $\hat{g}_1$. By Lemma 2.1, the following random variable has expectation not more than 0 and it is stochastically less than an exponential random variable with mean γ_1/n,

$$P_n'(\hat{g}_1) - P_n(\hat{g}_1) - \frac{\gamma_1}{n}L(\hat{g}_1) - \frac{1}{2\gamma_1}s^2(\hat{g}_1). \tag{2.15}$$

Since $\hat{g}_1$ is nonnegative, we have $s^2(\hat{g}_1) \leq P_n'(\hat{g}_1^2) + P_n(\hat{g}_1^2)$. Also since $|\hat{f}| \leq B'$ and $|f^*| \leq B$, the $\hat{g}_1 = (f^* + \hat{f})^2$ is bounded by $(B+B')^2$. Hence $s^2(\hat{g}_1)$ is bounded by $(B+B')^2(P_n'(\hat{g}_1) + P_n(\hat{g}_1))$. Choosing $\gamma_1 = A_1(B+B')^2/2$ with A_1 to be specified later, we have that above random variable is greater than or equal to

$$P_n'(\hat{g}_1) - P_n(\hat{g}_1) - \frac{\gamma_1}{n}L(\hat{g}_1) - \frac{1}{A_1}\left(P_n'(\hat{g}_1) + P_n(\hat{g}_1)\right\}, \tag{2.16}$$

which is the same as

$$(1 - \frac{1}{A_1})P_n'(\hat{g}_1) - (1 + \frac{1}{A_1})P_n(\hat{g}_1) - \frac{\gamma_1}{n}L(\hat{f}), \tag{2.17}$$

since $L(\hat{g}_1) = L(\hat{f})$. Now we turn our attention to $\hat{g}_2 = g_{2,\hat{f}}(X)$. Note that $|g_2(X)|$ is bounded by $K = B + B'$ and $\hat{g}_1 = \hat{g}_2^2$. Using Lemma 2.2 for the class $\mathcal{G}_2$, we know for any positive A_2 that the following random variable also has expectation not more than 0 and it is stochastically less than an exponential random variable with mean γ_2/n,

$$P_n(\epsilon\hat{g}_2) - \frac{1}{A_2}P_n(\hat{g}_1) - \frac{\gamma_2}{n}L(\hat{f}). \tag{2.18}$$

where $\gamma_2 = A_2\sigma^2/2 + (B+B')h_{Bern}$ and $L(\hat{g}_2) = L(\hat{f})$, and where we denote $\frac{1}{n}\sum_{i=1}^n \epsilon_i\hat{g}_2(X_i)$ by $P_n(\epsilon\hat{g}_2)$.

With a constant a to be determined, we add (2.17) and $2a$ times (2.18) together to obtain

$$\left(1 - \frac{1}{A_1}\right)P_n'(\hat{g}_1) - \left(1 + \frac{1}{A_1} + \frac{2a}{A_2}\right)P_n(\hat{g}_1) + 2aP_n(\epsilon\hat{g}_2) - \frac{\gamma_1 + 2a\gamma_2}{n}L(\hat{f}). \tag{2.19}$$

To glue these terms together cleanly, observing the fact that $\hat{g}_1(X_i) - 2\epsilon_i\hat{g}_2(X_i) = \hat{\rho}(X_i, Y_i)$, we choose to set a to satisfy $1 + \frac{1}{A_1} + \frac{2a}{A_2} = a$ and then expression (2.19) becomes

$$\left(1 - \frac{1}{A_1}\right)P_n'(\hat{g}_1) - a\left[P_n(\hat{\rho}) + \frac{\gamma}{n}L(\hat{f})\right], \tag{2.20}$$

where $\gamma = (\gamma_1 + 2a\gamma_2)/a$. Alternatively, by choosing $A_1 = 1 + \frac{2}{\delta_1}$, $A_2 = 2 + \frac{2}{\delta_2}$, and $\tilde{c} = (1 + \delta_1)(1 + \delta_2)$, then $a = \tilde{c}(1 - 1/A_1)$ and dividing the expression by $1 - 1/A_1$, we have

$$P_n'(\hat{g}_1) - \tilde{c}\left[P_n(\hat{\rho}) + \frac{\gamma}{n}L(\hat{f})\right], \tag{2.21}$$

with γ equal to $\frac{(1+\delta_1/2)(1+2/\delta_1)}{2\tilde{c}}(B+B')^2+2(1+\frac{1}{\delta_2})\sigma^2+2(B+B')h_{Bern}$, where $\delta_1 > 0$ and $\delta_2 > 0$ are arbitrary positive constants. The conclusion is this expression has expectation not more than 0. Moreover, expression (2.21) concentrates to be not much more than 0, except with exponentially small probability. Indeed, for any positive u_1, u_2 the probability that it exceeds $[u_1 + 2au_2]/(1 - 1/A_1)$ is not more than $\exp(-nu_1/\gamma_1) + \exp(-nu_2/\gamma_2)$.

Taking the expectation and moving the part in brackets to the right side we have

$$\mathbb{E}P'_n(\hat{g}_1) \leq \tilde{c}\,\mathbb{E}\left[P_n(\hat{\rho}) + \frac{\gamma}{n}L(\hat{f})\right]. \tag{2.22}$$

This inequality (2.22) matches the desired risk bound (2.6) with a constant factor c slightly larger than 1. Indeed if the penalty $\text{pen}_n(f)$ were chosen as $\gamma L(f)$, then $\hat{f}$ minimizes $\left[P_n(\hat{\rho}) + \frac{\gamma}{n}L(\hat{f})\right]$. Bounding the expected infimum by the infimum of expectations, we may replace the right side by the resolvability expressing the approximation and complexity tradeoff.

We note also that the above analysis leading to (2.21) holds if in place of $\hat{f}$, one used any selection based on $\underline{Y}$, $\underline{X}$ and $\underline{X}'$ within the countable $\tilde{\mathcal{F}}$. Hence, we have the conclusion

$$\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\left\{P'_n(g_{1,f}) - \tilde{c}P_n(\rho_f) - \frac{\tilde{c}\mathcal{L}_n(f)}{n}\right\} \leq 0, \tag{2.23}$$

where $\tilde{c} = (1 + \delta_1)(1 + \delta_2)$ and $\mathcal{L}_n(f) = \gamma L(f)$. This verifies the desired form (2.5) in the countable case.

2.4 Data-dependent countable classes and resolvability risk bound

To extend the conclusion to general $\mathcal{F}$, we seek penalties for which (2.8) or equivalently (1.7) holds. We are to exhibit a countable $\tilde{\mathcal{F}}$, such that, for each $f \in \mathcal{F}$, the penalty $\text{pen}_n(f)$ exceeds the infimum over $\tilde{f} \in \tilde{\mathcal{F}}$ of an appropriate expression.

More freedom in choosing $\text{pen}_n(f)$ is made available by allowing the set $\tilde{\mathcal{F}} = \tilde{\mathcal{F}}_{\underline{X},\underline{X}'}$ and the complexities $L_n(\tilde{f}) = L_{\underline{X},\underline{X}'}(\tilde{f})$ we construct in the proofs to depend on the input data $\underline{X}$ and its independent copy $\underline{X}'$. For instance, if $\mathcal{F}$ were a bounded empirical metric entropy class, then we could work with an empirical L_2 cover on these $2n$ points. We use variable-complexity empirical covers to handle more general cases of interest including linear spans of libraries. With this freedom, we allow penalties $\text{pen}_n(f)$ to be at least

$$\inf_{\tilde{f} \in \tilde{\mathcal{F}}_{\underline{X},\underline{X}'}} \left\{ \gamma L_{\underline{X},\underline{X}'}(\tilde{f}) + \Delta_n(f, \tilde{f}) \right\}, \tag{2.24}$$

where $\Delta_n(f, \tilde{f})$ is the distortion as explained in the introduction. When we desire the penalty to not depend on $\underline{X}'$ the understanding is that it is to exceed the indicated expression (2.24) for all $\underline{X}'$. Alternatively, a less demanding requirement is that $\text{pen}_n(f)$ exceed the expectation of expression (2.24) with respect to $\underline{X}'$ conditional on $\underline{X}$ and $\underline{Y}$.

A useful devise in checking whether certain penalties of interest satisfy the requirement is to note that while $\text{pen}_n(f)$ as given might not exceed the required expression (2.24), the addition to it of some adjustment, denoted Adjust_n, that does not depend on f, may lead to $\big[\text{pen}_n(f) + \text{Adjust}_n\big]$ exceeding (2.24). If need be, this adjustment may depend on the data. Such adjustment does not change the penalized least squares estimator, but it will be reflected in the risk bound through the presence of the expected adjustment relative to the sample size, $\frac{1}{n}\mathbb{E}\big[\text{Adjust}_n\big]$. It is then preferable to have such adjustments be negligible in

size compared to the main $\text{pen}_n(f)$ term that adapts the choice of f.

For our analysis in the data-dependent penalty case, we note that key to the proof of Lemma 2.1 is the fact that the probabilities there are unchanged if one exchanges any coordinate pair (X_i, X_i'). We will need coordinate pair exchangeability to still hold for the classes $\tilde{\mathcal{F}}_{\underline{X},\underline{X}'}$. To allow this data-dependent freedom, we make use of the following definition and assumption.

Definition (Coordinate Pair Symmetry). We call a collection of classes $\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}$ indexed by $(\underline{X}, \underline{X}')$, each a subset of a given class $\mathcal{G}$, symmetric between $\underline{X}$ and $\underline{X}'$ if $\tilde{\mathcal{G}}_{\underline{X}_{(i)},\underline{X}'_{(i)}} = \tilde{\mathcal{G}}_{\underline{X},\underline{X}'}$, where $\underline{X}_{(i)} = (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$ and $\underline{X}'_{(i)} = (X_1', \ldots, X_{i-1}', X_i, X_{i+1}', \ldots, X_n')$, for each $i = 1, \ldots, n$. Likewise, for $\tilde{g}$ in $\tilde{\mathcal{G}}_{\underline{X}_{(i)},\underline{X}'_{(i)}} = \tilde{\mathcal{G}}_{\underline{X},\underline{X}'}$, we call $L_{\underline{X},\underline{X}'}(\tilde{g})$ symmetric between $\underline{X}$ and $\underline{X}'$ if $L_{\underline{X}_{(i)},\underline{X}'_{(i)}}(\tilde{g}) = L_{\underline{X},\underline{X}'}(\tilde{g})$ for each $i = 1, \ldots, n$.

Assumption (S) (Symmetry and Complexity Condition). The collection of classes $\tilde{\mathcal{F}}_{\underline{X},\underline{X}'}$ and associated complexities $L_{\underline{X},\underline{X}'}(\tilde{f})$ are coordinate pair symmetric between $\underline{X}$ and $\underline{X}'$ and the complexities $L_{\underline{X},\underline{X}'}(\tilde{f})$ satisfy the Kraft inequality

$$\sum_{\tilde{f} \in \tilde{\mathcal{F}}_{\underline{X},\underline{X}'}} e^{-L_{\underline{X},\underline{X}'}(\tilde{f})} \leq 1.$$

With this assumption, Lemma 2.1 and 2.2 are established in generalized forms in the appendix. In this setting, with symmetric dependence of $\tilde{\mathcal{F}}_{\underline{X},\underline{X}'}$ and $L_{\underline{X},\underline{X}'}$ on $\underline{X}, \underline{X}'$, the same argument we have used to derive the inequalities (2.21) and (2.23) holds with these generalized forms of the lemmas. Consequently, we have the following lemma and theorem.

Lemma 2.3 *For the regression setting (B), let $\tilde{\mathcal{F}}_{\underline{X},\underline{X}'}$ be a data-dependent countable set of functions with associated complexities $L_{\underline{X},\underline{X}'}$ satisfying Assumption (S). Also assume there exists a uniform bound B' for $\tilde{f} \in \tilde{\mathcal{F}}_{\underline{X},\underline{X}'}$. Given any positive δ_1 and δ_2, the following*

holds

$$\mathbb{E} \sup_{\tilde{f} \in \tilde{\mathcal{F}}_{\underline{X},\underline{X}'}} \left\{ \frac{1}{\tilde{c}} P'_n(g_{1,\tilde{f}}) - P_n(\rho_{\tilde{f}}) - \frac{\gamma L_{\underline{X},\underline{X}'}(\tilde{f})}{n} \right\} \leq 0, \tag{2.25}$$

where $\tilde{c} = (1+\delta_1)(1+\delta_2)$ *and* $\gamma = \frac{1+\delta_1/4+1/\delta_1}{\tilde{c}}(B+B')^2 + 2(1+\frac{1}{\delta_2})\sigma^2 + 2(B+B')h_{Bern}$.

2.5 Resolvability risk bound for uncountable $\mathcal{F}$

2.5.1 General theorem

Adapting (2.24), our general penalty requirement is that there is a collection $\tilde{\mathcal{F}}_{\underline{X},\underline{X}'}$ of functions $\tilde{f}$ bounded by B' and associated complexities $L_{\underline{X},\underline{X}'}(\tilde{f})$ satisfying Assumption (S) and an adjustment Adjust_n such that for every f in $\mathcal{F}$, the penalty has $\left[\text{pen}_n(f) + \text{Adjust}_n\right]$ at least

$$\inf_{\tilde{f} \in \tilde{\mathcal{F}}_{\underline{X},\underline{X}'}} \left\{ \Delta_n(f, \tilde{f}) + \gamma L_{\underline{X},\underline{X}'}(\tilde{f}) \right\}, \tag{2.26}$$

where, setting $c = (1+\delta) = (1+\delta_3)\tilde{c}$, the distortion is

$$\begin{aligned} \Delta_n(f, \tilde{f}) &= \sum_{i=1}^n (Y_i - \tilde{f}(X_i))^2 - \sum_{i=1}^n (Y_i - f(X_i))^2 \\ &+ \frac{1}{c} \sum_{i=1}^n (Tf(X'_i) - f^*(X'_i))^2 - \frac{1}{\tilde{c}} \sum_{i=1}^n (\tilde{f}(X'_i) - f^*(X'_i))^2. \end{aligned}$$

Theorem 2.4 establishes our general bound on the risk of the penalized least squares estimators for possibly uncountable $\mathcal{F}$. For such $\mathcal{F}$, this theorem and its refinements provide our main tool for obtaining risk bounds for ℓ_1 penalized estimators and those which are constructed from various term selection procedures.

Theorem 2.4 *For the same setting (B) as in Theorem 2.4, suppose we are given any positive* δ_1, δ_2 *and non-negative* δ_3 *and a proper penalty function* $\text{pen}_n(f)$ *which with an ad-*

justment Adjust$_n$ exceeds (2.26) or exceeds its expectation with respect to $\underline{X}'$. Then an approximate penalized least squares estimator $\hat{f} = \hat{f}_n$ (with optimization accuracy A_f), when truncated to the level B', has risk satisfying

$$\begin{aligned} & \mathbb{E}\|T\hat{f} - f^*\|^2_{\underline{X}'} \\ \leq \; & c \inf_{f \in \mathcal{F}} \left\{ \|f - f^*\|^2 + \mathbb{E}\left[\frac{pen_n(f)}{n} + A_f \right] + \frac{adjust}{n} \right\}, \end{aligned} \tag{2.27}$$

where adjust $= \mathbb{E}\big[Adjust_n\big]$ *and* $c = (1+\delta) = (1+\delta_1)(1+\delta_2)(1+\delta_3)$. *Here* γ *used in (2.26) is the same as given in Lemma 2.3.*

Remark: A simple choice of the constants is to set $\delta_1 = 1/2$, $\delta_2 = 1/3$ and $\delta_3 = 0$. In this setting, the coefficient $c = 2$ and the main term in γ is $\frac{25}{16}(B+B')^2$ with an additional term arising from unbounded noise as $\frac{8}{3}\sigma^2 + 2(B+B')h_{Bern}$.

Proof: We denote $\tilde{\mathcal{F}}_{\underline{X},\underline{X}'}$ and $L_{\underline{X},\underline{X}'}(\tilde{f})$ by $\tilde{\mathcal{F}}_n$ and $L_n(\tilde{f})$ for simplicity. Let $\text{pen}_n^+(f) = \text{pen}_n(f) + \text{Adjust}_n$. Rearrange the penalty condition (2.26) and take the expectation with respect to $\underline{X}'$ to obtain

$$\begin{aligned} & \mathbb{E}_{\underline{X}'} \sup_{f \in \mathcal{F}} \left\{ \frac{1}{c} P'_n(g_{1,Tf}) - P_n(\rho_f) - \frac{\text{pen}_n^+(f)}{n} \right\} \\ \leq \; & \mathbb{E}_{\underline{X}'} \sup_{\tilde{f} \in \tilde{\mathcal{F}}_n} \left\{ \frac{1}{\tilde{c}} P'_n(g_{1,\tilde{f}}) - P_n(\rho_{\tilde{f}}) - \frac{\gamma L_n(\tilde{f})}{n} \right\}, \end{aligned}$$

where $\tilde{c} = (1+\delta_1)(1+\delta_2)$. Applying Lemma 2.3, we know that the expectation with respect to $\underline{X}, \underline{Y}$ of the right side is less than or equal to 0. Consequently, the corresponding expectation of the left side is less than 0 as well, yielding risk for the penalized least squares estimator $\mathbb{E}\|T\hat{f} - f^*\|^2_{\underline{X}'} = \mathbb{E}P'_n(g_{1,T\hat{f}})$ bounded by,

$$c\mathbb{E}\left(P_n(\rho_{\hat{f}}) + \frac{\text{pen}_n^+(\hat{f})}{n} \right).$$

The above expression is an expected minimum which is bounded by the minimum expectation, from which the conclusion follows. ∎

The requirement that the class $\tilde{\mathcal{F}}_{\underline{X},\underline{X}'}$ of functions $\tilde{f}$ be bounded by B' forces us to either restrict attention to class $\mathcal{F}$ of candidate functions f which are also bounded by B' or to work with a truncated version $T\mathcal{F} = \{Tf : f \in \mathcal{F}\}$, where T is the truncation operator at level B' such that $Tf = \min\{B', |f|\}\text{sgn}(f)$, in which case our final fit is the truncated function $T\hat{f}$, where $\hat{f}$ is the penalized least squares estimator.

2.5.2 Rectifiable penalty requirement

Direct verification of the penalty requirement for the truncated function would be tricky in some cases (e.g., the ℓ_1-penalty case in Chapter 4). We find then that it is more natural to exhibit satisfaction of a suitable inequality for an unbounded $\mathcal{F}$ using an unbounded collection of covers $\tilde{\mathcal{F}}_{\underline{X},\underline{X}'}$. Then we would like satisfaction of a penalty requirement in the unbounded case to imply its satisfaction for the truncated functions, with $T\tilde{f}$ replacing $\tilde{f}$ and Tf replacing f. A penalty requirement that meets this aim is said to be *rectifiable*. A modification of our condition is shown to have this property.

First recall the form of our penalty requirement (1.10) expressed as a lower bound on the penalized sum of squared errors. As explained there, by introducing a positive $\delta_3 = \delta'$, we can bound $\text{diff}_n(f, \tilde{f})$ with $\frac{1}{\delta_3}\sum_{i=1}^n (f(X_i') - \tilde{f}(X_i'))^2$, which always is at least the same expression with f and $\tilde{f}$ replaced by Tf and $T\tilde{f}$, respectively. For the squared error of $\tilde{f}$ in the first term of (1.10), we shall see that it also is at least the corresponding squared errors of $T\tilde{f}$ with a small correction. Accordingly, we consider the following expression to use in controlling the penalized squared error,

$$\inf_{\tilde{f}\in\tilde{\mathcal{F}}_{\underline{X},\underline{X}'}} \left\{ \sum_{i=1}^n \Big(Y_i - \tilde{f}(X_i)\Big)^2 + \gamma L_{\underline{X},\underline{X}'}(\tilde{f}) + \frac{1}{\tilde{c}\delta_3}\sum_{i=1}^n (f(X_i') - \tilde{f}(X_i'))^2 \right\}. \qquad (2.28)$$

where functions in the countable set $\tilde{\mathcal{F}}_{\underline{X},\underline{X}'}$ may be unbounded. Indeed, the following lemma gives a sense in which a penalty requirement based on (2.28) with unbounded f, $\tilde{f}$ is rectified by truncation.

Lemma 2.5 *Expression (2.28) with the addition of an adjustment equal to Tail($\underline{Y}$) is greater than or equal to that expression with $\tilde{f}$ and f replaced by $T\tilde{f}$ and Tf, respectively, where the quantity Tail($\underline{Y}$)= $2\sum_{i=1}^{n}(|Y_i| - B')^2 1\{|Y_i| > B'\}$. Accordingly, if the penalized squared error exceeds (2.28), then $pen_n(f)$ is proper in the sense that with the indicated adjustment, the penalty exceeds (2.26) for the truncated $T\tilde{f}$ replacing the $\tilde{f}$.*

Remark: The quantity Tail(Y_i) defined using the square of the excess $(|Y_i| - B')$ for $|Y_i| > B'$, is also denoted $\text{Tail}_2(Y_i)$. Later we will also have similar use for $\text{Tail}_1(Y_i) = 4B'(|Y_i| - B')1\{|Y_i| > B'\}$ and $\text{Tail}_1(\underline{Y}) = 4B'\sum_{i=1}^{n}(|Y_i| - B')1\{|Y_i| > B'\}$ defined without the square.

Proof: From the discussion above, we only need to show the following inequality,

$$\sum_{i=1}^{n}(Y_i - \tilde{f}(X_i)^2 + \text{Tail}(\underline{Y}) \geq \sum_{i=1}^{n}(Y_i - T\tilde{f}(X_i))^2, \tag{2.29}$$

to be able to conclude that with the tail adjustment, the penalty exceeds (2.26). We show the above inequality is true term by term, that is,

$$(Y_i - T\tilde{f}(X_i))^2 \leq (Y_i - \tilde{f}(X_i))^2 + 2(|Y_i| - B')^2 1\{|Y_i| > B'\}. \tag{2.30}$$

We use truncation operators $T_i = T_{|Y_i| \vee B'}$ which are defined for $i = 1, 2, \ldots, n$ as $T_i f =$ $\min\{\max\{|Y_i|, B'\}, |f|\}\text{sgn(f)}$. By algebra for differences of squares $(Y_i - T\tilde{f}(X_i))^2$ is equal to,

$$(Y_i - T_i\tilde{f}(X_i))^2 + (T_i\tilde{f}(X_i) - T\tilde{f}(X_i))(2Y_i - T\tilde{f}(X_i) - T_i\tilde{f}(X_i)). \tag{2.31}$$

The first term is less than or equal to $(Y_i - \tilde{f}(X_i))^2$ from the definition of the operator $T_{|Y_i| \vee B'}$. If $|Y_i| \leq B'$ or if $|\tilde{f}(X_i)| \leq B'$, then $T_{|Y_i| \vee B'} \tilde{f}(X_i)$ and $T\tilde{f}(X_i)$ are equal to each other and the last term is zero, so (2.30) holds then. Also if Y_i and $\tilde{f}(X_i)$ are of opposite sizes then the first term on the right side of (2.30) already exceeds the left side, so (2.30) holds then as well. Otherwise, Y_i and $\tilde{f}(X_i)$ are of the same sign and both have magnitude at least B'. Then $2Y_i - T\tilde{f}(X_i) - T_i\tilde{f}(X_i)$ has magnitude not more than $2(|Y_i| - B')$. Also, since $T_i\tilde{f}(X_i)$ and $T\tilde{f}(X_i)$ have the same sign, the difference between them has magnitude less than or equal to $(|Y_i| - B')$, which completes the proof. ∎

2.5.3 Conclusion with respect to rectifiable penalties

A variant of Theorm 2.4 which removes need for consideration of boundedness is expressed in the following corollary.

Corollary 2.6 *For the same setting (B) as in Theorem 2.4, suppose we are given any positive δ_1, δ_2 and δ_3. If a penalty function $pen_n(f)$ is such that the penalized squared error exceeds expression (2.28) with a possibly unbounded $\tilde{\mathcal{F}}_{\underline{X},\underline{X}'}$, then with the adjustment by $Tail(Y)$ it exceeds (2.26) with the truncated representors. Accordingly, an approximate penalized least squares estimator $\hat{f} = \hat{f}_n$, when truncated to the level B', has risk satisfying (2.27), where $adjust = tail$. Here $tail = 4D_1\nu^2$ with $D_1 = \mathbb{E}e^{|\epsilon|/\nu}$ as defined before and $B' \geq B + \nu \log n$. If more strongly, ϵ is assumed to be sub-Gaussian with $D_2 = \mathbb{E}e^{\epsilon^2/\nu}$, then $tail = D_2\nu$ and $B' \geq B + \sqrt{\nu \log n}$; whereas if the error ϵ is bounded by a constant c_0, then $tail = 0$ and $B' \geq B + c_0$.*

Remark: In the risk bound, there are two terms effected by the value of ν. One is the penalty for which it is best to use as small a B' as allowed. It is also increasing w.r.t. ν, so we may be inclined to use as small as possible a value for ν. The other term is the $tail/n$ term, based on D_1 and D_2 which increases with decreasing ν. For some possible

distributions on ϵ, such as a two-sided exponential or a Gaussian, when ν goes to associated lower bounds, then D_1 and D_2, respectively, tend to infinity. Therefore, the ideal ν to apply is determined by the trade-off between these two terms. For the Gaussian, D_2 is finite for $\nu > 2\sigma^2$. A simple rule is to use a ν slightly larger than twice the variance.

Proof: From Lemma 2.5, the fact that the penalty $\text{pen}_n(f)$ is rectifiable provides the properness of the penalty function $\text{pen}_n(f)$ adjusted by adding $\text{Tail}(\underline{Y})$. Hence the risk bound for $T\hat{f}$ follows by using Theorem 2.4. It remains to bound the expectation of $\text{Tail}(\underline{Y})$ which is denoted $tail$. Lemma 8.3 in the appendix shows for $i = 1, 2, \ldots, n$ that, when ϵ_i has finite $\mathbb{E}e^{|\epsilon|/\nu}$, for $B' \geq B + \nu \log n$,

$$\mathbb{E}(|Y_i| - B')^2 1\{|Y_i| > B'\} \leq \frac{2D_1\nu^2}{n},$$

whereas, when ϵ_i has finite $\mathbb{E}e^{\epsilon^2/\nu}$, for $B' \geq B + \sqrt{\nu \log n}$,

$$\mathbb{E}(|Y_i| - B')^2 1\{|Y_i| > B'\} \leq \frac{D_2\nu}{2n}.$$

Finally, if ϵ is bounded by c_0, $|Y_i|$ is not more than $B + c_0$, which means $\text{Tail}(\underline{Y}) = 0$. ∎

Remark: The quantity $tail$ arises in bounding the expectation of the excess of $|Y_i|$ above B'. It is also denoted $tail_2$ in Chapter 4 and 5. Likewise $tail_1$ will denote corresponding bounds on $\mathbb{E}\text{Tail}_1(\underline{Y})$. According to Lemma 8.3, we may set $tail_1 = 4B'D_1v$ for $B' \geq B + v \log n$; if ϵ is assumed to be sub-Gaussian, then $tail_1 = 4B'D_2\sqrt{\pi v}$; whereas if the error ϵ is bounded, $tail_1 = 0$.

Chapter 3

Relaxed Greedy Computations and ℓ_1-Penalized Optimization

A general review of forward selection and other types of greedy algorithms is given in the introduction. As explained there, after having formed a linear combination of $m-1$ terms chosen from a library $\mathcal{H}$, one chooses the next term h_m from $\mathcal{H}$ such that a linear composition of it with proceeding terms provides a good improvement in the fit. In this section, we present two variants of our ℓ_1-penalized greedy pursuit (LPGP) algorithm and establish the claimed properties in Lemma 3.1. Traditional forward stepwise selection and other relaxed greedy algorithms and their properties correspond to a special case (with $\lambda = 0$).

3.1 Computation time of greedy algorithms

Concerning the computation time, suppose for a library of size M that each step of the search is conducted by trying every member of the library and evaluating required sums of size n for each. Then conducting m steps of the greedy algorithms entails a computation

time of order Mnm, with which statistically suitable accuracy is obtained with m not more than order $\sqrt{n}$. Of course, this is dramatically better than for all-subset regression for which the runtime is of order $\binom{M}{m} nm^2$, or, if the relevant inner products are precomputed, of order $\binom{M}{m} m^3 + M^2 n$.

For ℓ_1 penalized least squares, with fixed λ, the computation time of our LPGP algorithm is again of order Mnm, in which, though we are permitted to use larger m, one does not need to use a number of steps much larger than $\sqrt{n}$ to obtain a solution of statistical accuracy comparable to the exact penalized least squares limit. Other greedy ℓ_1 penalized least squares algorithms such as LARS [42] are said to provide solution for all λ in n steps for a computation time of order Mn^2. Treating ℓ_1 penalization as a convex optimization, and appealing to computation theory for interior point methods of Nemirovski and Nesterov as described in [22], would lead to computation time for which the dependence on the library size entails a somewhat higher power of M. So for ℓ_1 penalized least squares, greedy algorithms such as LPGP are to be preferred to general purpose interior point methods.

3.2 Function variation

For generality, and notational simplicity, we take our setting to be that of points in a Hilbert space with a norm $\| \cdot \|$ and an inner product $< \cdot, \cdot >$. The library $\mathcal{H}$ is taken to be a given set of points h. The special cases of interest are the spaces of functions in L_2 with respect to a probability measure with norm $\| \cdot \| = \| \cdot \|_{L_2(P)}$, and in particular the empirical L_2 space discussed above with $\| \cdot \|_n = \| \cdot \|_{L_2(P_n)}$. In the latter case the roles of f^* and f_m in the lemmas below are played by the point $\underline{Y}$ and the estimates $\hat{f}_m$ evaluated at the input data.

We first extend the definition of the $\mathcal{L}_{1,\mathcal{H}}$ norm.

Definition. The *variation* $V(f) = V_{\mathcal{H},a}(f)$ of f, with respect to a library $\mathcal{H}$ and positive weights $a = (a_h : h \in \mathcal{H})$, is

$$V(f) = \lim_{\varepsilon \to 0} \inf_{f_\varepsilon \in \mathcal{F}_\mathcal{H}} \left\{ \|\beta\|_1 : f_\varepsilon = \sum_h \beta_h h \text{ and } \|f_\varepsilon - f\| \leq \varepsilon, \beta_h \in \mathbb{R}, h \in \mathcal{H} \right\},$$

where $\mathcal{F}_\mathcal{H}$ is the linear span of $\mathcal{H}$ and $\|\beta\|_1 = \|\beta\|_{1,a} = \sum_h |\beta_h| a_h$.

Note that by the definition of $V(f)$, when it is finite, there will be finite linear combinations $f_\varepsilon = \sum \beta_{h,\varepsilon} h$ with $\|f - f_\varepsilon\|$ arbitrarily small and $\|\beta\|_1$ arbitrarily close to $V(f)$. We require the weights to satisfy $a_h \geq \|h\|$.

The variation $V(f)$ agrees with $\|f\|_{1,\mathcal{H}}$ for f in $\mathcal{F}_\mathcal{H}$ and extends the norm to the closure (so as to include all f that are limits of such linear combinations). With this extension, we denote $\mathcal{L}_{1,\mathcal{H}} = \mathcal{L}_{1,\mathcal{H},a}$ to be the set of functions with finite variation with respect to the library $\mathcal{H}$.

With the empirical distribution P_n on n points, we denote the empirical variation

$$V_n(f) = \lim_{\varepsilon \to 0} \inf_{f_\varepsilon \in \mathcal{F}_\mathcal{H}} \left\{ \|\beta\|_1 : f_\varepsilon = \sum_h \beta_h h \text{ and } \|f_\varepsilon - f\|_n \leq \varepsilon, \beta_h \in \mathbb{R}, h \in \mathcal{H} \right\},$$

where now $\|\beta\|_1 = \sum_h |\beta_h| a_h$ with a_h not less than $\|h\|_n = \|h\|_{\underline{X}}$.

The choice $a_h = \|h\|_n$ is most directly relevant for the bounds in this section. The choice $\max\{\|h\|_n, \eta\}$, for fixed $\eta > 0$, is used for risk bounds in Chapter 5. Symmetric forms such as $\sqrt{2}\|h\|_{\underline{X},\underline{X}'}$ or $\|h\|_\infty$ are used in Chapter 4.

3.3 Description of ℓ_1-*penalized greedy pursuit (LPGP)*

There are two variants in our ℓ_1-penalized greedy pursuit algorithm.

Definition (ℓ_1-Penalized Greedy Pursuit). Let $\mathcal{F}$ be a collection of points in the Hilbert space. Let f^* be a point or function we wish to fit. Initialize $f_0 = 0$. For $m =$

$1, 2, \ldots$, iteratively, given the terms of f_{m-1} as $h_1, \ldots, h_{m-1}$ and the coefficients of it as $\beta_{1,m-1}, \ldots, \beta_{m-1,m-1}$, with $v_{m-1} = \sum_{j=1}^{m-1} |\beta_{j,m-1}| a_{h_j}$, we proceed as follows in two cases with non-negative λ.

Variant 1

Let $f_m = \sum_{j=1}^{m} \beta_{j,m} h_j$ and $v_m = \sum_{j=1}^{m} |\beta_{j,m}| a_{h_j}$, with the term h_m in $\mathcal{H}$ and coefficients chosen such that

$$||f_m - f^*||^2 + \lambda v_m \leq \inf_{\alpha,\beta,h\in\mathcal{H}} \left\{ ||(1-\alpha) f_{m-1} + \beta h - f^*||^2 + \lambda \left((1-\alpha) v_{m-1} + |\beta| a_h \right) \right\} + \varepsilon_m^{\text{comp}}, \tag{3.1}$$

where the infimum is over $\beta \in \mathbb{R}$ and $\alpha \in [0, 1]$ and we require nonnegative $\varepsilon_m^{\text{comp}} \leq \frac{4\delta_0}{(m+1)^2}$ with $\delta_0 \geq 0$.

Variant 2

Choose the term h_m in $\mathcal{H}$ to come within a given constant factor $c \geq 1$ of the maximum normalized inner product (minimum angle) with the residual $f^* - f_{m-1}$, that is, $\langle \frac{h_m}{a_{h_m}}, f^* - f_{m-1} \rangle \geq \frac{1}{c} \sup_{h\in\mathcal{H}} \langle \frac{h}{a_h}, f^* - f_{m-1} \rangle$ and choose $f_m = (1-\alpha_m) f_{m-1} + \beta_m h_m$ with coefficients α_m and β_m in $\mathbb{R}$ such that

$$\begin{aligned} & ||f^* - f_m||^2 + \lambda v_m \\ \leq & \inf_{\beta\in\mathbb{R},\alpha\in[0,1]} \left\{ ||(1-\alpha) f_{m-1} + \beta h_m - f^*||^2 + \lambda \left((1-\alpha) v_{m-1} + |\beta| a_h \right) \right\}. \end{aligned} \tag{3.2}$$

Though optimization of α between 0 and 1 is desirable, it is acceptable to use $\alpha_m = \frac{2}{m+1}$ in (3.1) and (3.2) to yield the same bounds as in the following Lemma 3.1. As discussed in the introduction, the first variant with $\lambda = 0$ includes the forward stepwise regression, in

which case one optimizes the linear combination $\beta_{1,m}, \ldots, \beta_{m,m}$ at each step. Where we have multiplication by λ it can be replaced by any nonnegative convex function and the same conclusions will hold.

3.4 Computational accuracy of *LPGP*

The following lemma establishes the computational accuracy of ℓ_1-*penalized greedy pursuit*.

Lemma 3.1 *Let $\mathcal{H}$ be a collection of points in the Hilbert space. Let f^* be a target we wish to fit by linear combinations of elements of $\mathcal{H}$. Suppose the weights a_h which are associated with the variation $V(f) = V_{\mathcal{H},a}(f)$ are larger than or equal to the norm $\|h\|$ of the Hilbert space.*

Case 1

If f_m is chosen by using the first variant of the ℓ_1-penalized pursuit algorithm, then for every $m \geq 1$, the ℓ_1-penalized error satisfies

$$||f^* - f_m||^2 + \lambda \sum_{j=1}^{m} |\beta_{j,m}| a_{h_j} \leq \inf_f \left\{ ||f^* - f||^2 + \lambda V(f) + \frac{4(V^2(f) - \|f\|^2 + \delta_0)}{m+1} \right\}, \tag{3.3}$$

where the infimum is taken over all f in the Hilbert space.

Case 2

If f_m is chosen by using the second variant of the ℓ_1-penalized pursuit algorithm, then an analogous conclusion to Case 1 holds, but with a price for the slight suboptimality of each h_m. Indeed, for $m \geq 1$,

$$||f^* - f_m||^2 + \lambda \sum_{j=1}^{m} |\beta_{j,m}| a_{h_j} \leq \inf_f \left\{ ||f^* - f||^2 + c\lambda V(f) + \frac{4b_f}{m+1} \right\}. \tag{3.4}$$

where

$$b_f = \min\left\{[cV(f) + \|f^*\|]^2, [(1+c)V(f) + \|f - f^*\|]^2\right\} - \|f - f^*\|^2.$$

Remarks:

1. Thus after m steps, our algorithm is within order $1/m$ of the infimum.

2. There are possibly surprising aspects of this conclusion. Even though f_m is not the best m-term fit, the bound shows that its accuracy compares favorably with the infimum over all f. Also surprising is that on the left side, we have $\sum_{j=1}^m |\beta_{j,m}| a_{h_j}$ which may be greater than $V(f_m)$ when there are repeat visits to the same h, whereas on the right side we have the infimum over all f. Evidently this variation gap is also covered by the $4V^2(f)/(m+1)$ term.

3. For general algorithm weights, even those that don't satisfy $a_h \geq \|h\|$, inequalities (3.3) and (3.4) still hold with $V(f)$ replaced by $V_{\mathcal{H},a'}(f)$ on the right side, where $a'_h = \max\{a_h, \|h\|\}$.

4. We prove the counterpart to the inequalities (3.3) and (3.4) first for a fixed f on the right side and subsequently we take the infimum. These inequalities are trivial for f that have infinite norm $\|f\|$ or infinite variation $V(f)$, so suppose that f has finite norm $\|f\|$ and variation $V(f) = V_{\mathcal{H}}(f)$. There is no loss of generality if we assume $\mathcal{H}$ (replaced by $\mathcal{H} \cup -\mathcal{H}$ if need be) is closed under sign-change and that the coefficients of linear combination are kept non-negative. Then by the definition of $V(f)$, there will be a finite linear combination $f_\varepsilon = \sum_h \beta_{h,\varepsilon} h$ with $\|f - f_\varepsilon\|$ arbitrarily small and $\sum_h \beta_{h,\varepsilon} a_h$ arbitrarily close to $V(f)$. In that way it is enough to establish the inequalities for such finite linear combinations f in $\mathcal{F}_{\mathcal{H}}$. That is, we establish them for $f = f_\beta = \sum_h \beta_h h$ and $v = \|\beta\|_1 = \sum_h \beta_h a_h$, the variation associated with this particular expression of f.

5. The key step in the proof is a probabilistic sampling argument used to show that there exists an h_m yielding sufficient improvement at each step, following the idea first used in the approximation result of Jones [54]. It is of interest that this same idea is also used in our variable complexity argument in Chapter 4.

Proof: Our algorithm constructs a sequence of terms h_1, h_2, ..., h_m and a linear combination $f_m(x) = \sum_{j=1}^{m} \beta_{j,m} h_j$. The variation associated with this representation is $v_m = \sum_{j=1}^{m} \beta_{j,m} a_{h_j}$. Given the previous $\beta_{1,m-1}$, $\beta_{2,m-1}$, ..., $\beta_{m-1,m-1}$ and h_1, h_2, ..., h_{m-1}, this f_m with the new term h_m is chosen to compare favorably with the choice $(1-\alpha)f_{m-1}(x) + \beta h$ for all h in the library. Such a fit downweighs previous coefficients by the factor $(1-\alpha)$ and introduces a new term with coefficient β, and thus corresponds to a variation of $(1-\alpha)v_{m-1} + \beta a_h$. Let

$$e_m = \|f_m - f^*\|^2 - \|f - f^*\|^2 + \lambda v_m. \tag{3.5}$$

From (3.1) or (3.2), replacing the right side with optimized α and β with the not smaller value obtained with specific choices $\alpha = \frac{2}{m+1}$ and $\beta = \frac{\alpha v}{a_h}$, we have

$$e_m \leq \left\{ \|(1-\alpha)f_{m-1} + \alpha v h' - f^*\|^2 - \|f - f^*\|^2 \right\} + \lambda[(1-\alpha)v_{m-1} + \alpha v], \tag{3.6}$$

with $h'(x) = h(x)/a_h$ (where a small $\varepsilon_m^{\text{comp}}$ is permitted to be added to the right side of (3.6) and the corresponding expression below). Next use $\lambda[(1-\alpha)v_{m-1} + \alpha v] = (1-\alpha)\lambda v_{m-1} + \alpha\lambda v$, with the equality replaced by $\leq$ in the case that multiplication by λ is replaced (generalized) to be the use of a convex function per the remark above. Expanding

the square the inequality may be rearranged as

$$\begin{aligned} e_m \;\leq\; & (1-\alpha)e_{m-1} + \alpha^2 b(vh') + \alpha\lambda v \\ & -2\alpha(1-\alpha)\langle f^* - f_{m-1}, vh' - f\rangle \\ & -\alpha(1-\alpha)\|f_{m-1} - f\|^2, \end{aligned} \tag{3.7}$$

where $b(vh') = \|vh' - f^*\|^2 - \|f - f^*\|^2$.

Now in Case 1, h_m was chosen to perform at least as well as the infimum of the right side of (3.6) or equivalently (3.7) (to within the negligible $\varepsilon_m^{\text{comp}}$). Thus e_m is less than the average of the right side for any convenient distribution on the choices of h. For $f = \sum_h \beta_h h$ with $v = \sum_h \beta_h a_h$, we consider the average when h is chosen with probability $\beta_h \frac{a_h}{v}$ so that the expectation, the probability weighted average, of $\frac{vh(x)}{a_h}$ is $f(x)$. Then $\langle f^* - f_{m-1}, vh' - f\rangle$ has expectation 0 and $\|vh' - f^*\|^2 - \|f - f^*\|^2$ has expectation the same as that of $\|vh' - f\|^2$ which is less than or equal to $v^2 - \|f\|^2$ since a_h is greater than or equal to $\|h\|$. Throwing away the last term from (3.7), we thus have

$$e_m \leq (1-\alpha)e_{m-1} + \alpha^2 b_f + \lambda\alpha v + \varepsilon_m^{\text{comp}} \tag{3.8}$$

with $b_f = v^2 - \|f\|^2$.

Likewise $e_1 \leq b_f + \lambda v + \varepsilon_1^{\text{comp}}$. Then with $\varepsilon_m^{\text{comp}} \leq \frac{4\delta_0}{(m+1)^2}$ and assuming inductively that $e_{m-1} \leq \frac{4(b_f+\delta_0)}{m} + \lambda v$, we obtain from the inequality (3.8), with $\alpha = \frac{2}{m+1}$, that

$$e_m \leq \frac{4(b_f + \delta_0)}{m+1} + \lambda v$$

as desired. Taking the infimum over all f establishes the result for Case 1.

Now we turn our attention to Case 2. The argument is similar to Case 1 but differs in detail. With $\beta = \frac{c\alpha v}{a_h}$ in place of the minimizing β in (3.2), we obtain the same inequalities

as (3.6) and (3.7) but with cv in place of v and with the particular choice of h_m. Thus

$$\begin{aligned} e_m \leq\ & (1-\alpha)e_{m-1} + \alpha^2 b(cvh'_m) + c\alpha\lambda v \\ & -2\alpha(1-\alpha)\langle f^* - f_{m-1}, cvh'_m - f\rangle \\ & -\alpha(1-\alpha)\|f_{m-1} - f\|^2, \end{aligned} \tag{3.9}$$

where $h'_m = h_m/a_{h_m}$ and $b(cvh'_m) = \|cvh'_m - f^*\|^2 - \|f - f^*\|^2$. We bound $\|cvh'_m - f^*\|^2$ in two ways. One way is to simply use a triangle inequality to get an upper bound $(cv + \|f^*\|)^2$. The other is also to use the triangle inequality and $\|f\| \leq v$ to obtain the bound

$$(\|cvh' - f\| + \|f - f^*\|)^2 \leq [(1+c)v + \|f - f^*\|]^2.$$

Thus $b(cvh')$ is bounded by $b_f = \min\{[cv + \|f^*\|]^2, [(1+c)v + \|f - f^*\|]^2\} - \|f - f^*\|^2$.

The term $\langle f^* - f_{m-1}, cvh'_m - f\rangle$ is non-negative because of the selection rule of h_m using the fact that a maximum is bigger than the average. Therefore dropping this inner product term yields

$$e_m \leq (1-\alpha)e_{m-1} + \alpha^2 b_f + \lambda\alpha v + \varepsilon_m^{\text{comp}}. \tag{3.10}$$

Then from the same induction step, we prove the conclusion of Case 2. ∎

The heart of the proof is the demonstration that optimization of the improvement on each step, which is at least as good as the improvement one has on the average for certain distributions on h, is enough improvement for the claim to hold by induction. The distributions are constructed from the absolute values of the coefficients of functions approximating the target. This is the same strategy used in [54] and [7] in showing accuracy of approximation by greedy algorithms for targets in the convex hull of a multiple of a library and by [59] for targets possibly outside of such a convex hull. With the simple

modification to the squared error, adding the contribution to the ℓ_1 penalty at each step, the greedy algorithm is shown also to solve the ℓ_1 penalized least squares problem.

We end by considering the empirical situation to remind that the results of this section imply the validity of inequality (1.12) and (1.13) in the introduction. In the $\lambda = 0$ case, with $A_{f,m}$ defined as $\|Y - \hat{f}_m\|_n^2 - \|Y - f\|_n^2$, Lemma 3.1 demonstrates an upper bound of $A_{f,m}$ equal to $4V_n^2(f)/m$ for all m and f. Likewise in the general λ case, with $A_{f,m}$ defined as the difference of ℓ_1 penalized squared errors between the m-step estimator and any function f is bounded by the same quantity.

Chapter 4

Risk Bound for ℓ_1 Penalized Estimators

In this chapter, we apply tools developed in Chapter 2 to establish risk bounds for the ℓ_1 penalized least squares estimator. We start with special cases and extend the results to more general situations.

4.1 Setting and Goal

The class $\mathcal{F} = \mathcal{F}_{\mathcal{H}}$ is the linear span of a library $\mathcal{H}$ as in Chapters 1 and 3. Thus any f in $\mathcal{F}_{\mathcal{H}}$ is of the form $f(x) = f_\beta(x) = \sum_h \beta_h h(x)$ where the coefficient $\beta = (\beta_h : h \in \mathcal{H})$ has some finite subset of $\mathcal{H}$ within which β_h is non-zero. Without loss of generality, we assume that 0 is in $\mathcal{H}$ and, as before, we assume it is closed under sign changes in the sense that if a function h in $\mathcal{H}$, then $-h$ is also in this set. Otherwise replace $\mathcal{H}$ with $\mathcal{H} \cup -\mathcal{H} \cup \{0\}$. Accordingly, in this chapter, we assume all coefficient β_h in the linear combinations are non-negative. We recall that $\underline{X} = (X_i)_{i=1}^n$ and $\underline{Y} = (Y_i)_{i=1}^n$ are training data and $\underline{X}' = (X_i')_{i=1}^n$ is an independent copy of $\underline{X}$.

We want to show that a weighted ℓ_1 norm of the coefficient $\|\beta\|_1 = \|\beta\|_{1,a} = \sum_h |\beta_h| a_h$ can be used to formulate a proper penalty. Our first result in this chapter requires that a_h

exceeds $\|h\|_{\underline{X}}$ and $\|h\|_{\underline{X}'}$, which may be thought of as distances of h from 0. Later in this chapter we will allow smaller a_h that correspond to distances of h from certain sets that arise in approximating $\mathcal{H}$.

An estimator $\hat{f} = f_{\hat{\beta}} = \sum_{h \in \mathcal{H}} \hat{\beta}_h h$ in $\mathcal{F}_{\mathcal{H}}$ is an approximate ℓ_1 penalized least squares estimator with multiplier λ and weights a_h if it satisfies the following inequality,

$$\|Y - f_{\hat{\beta}}\|_n^2 + \lambda \|\hat{\beta}\|_{1,a} \leq \inf_{\beta} \left\{ \|Y - f_\beta\|_n^2 + \lambda \|\beta\|_{1,a} + A_\beta \right\}. \tag{4.1}$$

Exact ℓ_1 penalized least squares corresponds to $A_\beta = 0$, while computing a predetermined number of steps $m_{n,\text{comp}}$ of the ℓ_1 penalized greedy pursuit algorithm yields $A_\beta \leq 4\|\beta\|_{1,\|\cdot\|_n}^2 / m_{n,\text{comp}}$. (Data-based stopping rules are analyzed separately in Chapter 5). This definition of ℓ_1-penalized least squares with penalty given by $\lambda\|\beta\|_1$ matches the general concept, setting $\text{Pen}_n(\beta)/n = \lambda\|\beta\|_1$ or equivalently for $f \in \mathcal{F}$ setting $\text{pen}_n(f)/n$ to be the minimum of $\lambda\|\beta\|_1$ for coefficients β for which $f_\beta = f$. Hence if we prove this penalty function is indeed a proper penalty satisfying the requirements in Theorem 2.4 or Corollary 2.6, then the conclusion of those theorems may be applied to obtain a risk bound for $\hat{f}$.

4.2 Finite dictionary case

First, considering the case that $\mathcal{H}$ is finite with size $M = M_{\mathcal{H}}$, we show that λ exceeding $C\sqrt{(\log M)/n}$, with some constant C, makes the quantity $\lambda\|\beta\|_1$ a valid choice of $Pen_n(\beta)/n$, with adjustment by a smaller order $(\log M)/n$ term, satisfying the requirements of our theory. Then, subsequently, we will show reductions in λ taking advantage of possible covering properties of the library and allowing generalization to infinite size. The analysis in this first case displays the essence of the proof for the more general cases.

4.2.1 Constructing the countable set and complexities

Introduce the countable set $\tilde{\mathcal{F}}$ to be the set of all functions of the form

$$\tilde{f}(x) = \frac{v}{m} \sum_{k=1}^{m} h_k(x)/a_{h_k} \tag{4.2}$$

for terms h_k in $\mathcal{H}$ for any $m = 1, 2, \ldots$ and $v = m\eta$, with η to be specified later. We do not impose any upper bound on m in creating our cover.

For each $\tilde{f} \in \tilde{\mathcal{F}}$, the main part of the codelength $L(\tilde{f})$ is $m \log M$ nats to describe the choices of $h_1, \ldots, h_m$ for a specified m. Actually, because the order does not matter and because of the possibility of repeats, for specified m, a somewhat shorter description of $\tilde{f}$ is possible, as detailed in the appendix, using not more than $m \log(2eM/\min\{m, M\})$ nats.

The other part of the codelength is the description of m and it is negligible in comparison. Since the m are natural numbers, a crude codelength such as $m \log 2$ is enough. Thus adding these contributions together, we have the simple codelength, for $\tilde{f}$ of the form (4.2),

$$L(\tilde{f}) = m \log(2M),$$

and the refined codelength satisfying

$$L(\tilde{f}) \leq m \log_+(M/m) + m \log 4e.$$

If the quantities a_h are symmetric between $\underline{X}$ and $\underline{X}'$, which is true when $a_h = \|h\|_\infty$ or when $a_h = \sqrt{2}\|h\|_{\underline{X},\underline{X}'}$, then $\tilde{\mathcal{F}}$ and $L_n(\tilde{f})$ satisfy Assumption (S).

4.2.2 A preliminary analysis with boundedness restriction

Now we assume the functions $\|\tilde{f}\|_\infty$ are less than B'. If $a_h = \|h\|_\infty$, then noting that $\|\tilde{f}\|_\infty \leq v \leq \|\beta\|_1 + \eta$, that boundedness could be achieved by imposing the restriction that $\|\beta\|_1 \leq B' - \eta$.

For f in $\mathcal{F}$, let $f_\beta = f$ yield $\|f\|_{1,\mathcal{H}} = \|\beta\|_1$. The quantity which arises as a lower bound for a proper $\text{pen}_n(f)/n$ as in Theorem 2.4 is the minimum over $\tilde{\mathcal{F}}$ of the distortion plus complexity relative to sample size

$$\|Y - \tilde{f}\|_n^2 - \|Y - f\|_n^2 + \frac{1}{c}\left[\|f^* - f\|_{\underline{X}'}^2 - \|f^* - \tilde{f}\|_{\underline{X}'}^2\right] + \frac{\gamma}{n}L(\tilde{f}), \tag{4.3}$$

where c is a constant greater than 1. To verify the proper penalty condition we exhibit for each $f \in \mathcal{F}$, the existence of a representor $\tilde{f} = \tilde{f}_{m_f}$ in $\tilde{\mathcal{F}}$ with an $m = m_f$ depending on f, such that the inequality holds, namely that $\text{pen}_n(f)/n$ is not less than the expression (4.3). To establish the existence of such a representor, recalling that $\tilde{f}$ is built from choices of h_k, we consider a distribution, in which each h_k is selected independently, in which the probability of each h is specified based on the values of β_h. For instance, these probabilities may be proportional to $\beta_h a_h$. If the inequality we want holds on the average with respect to the chosen distribution, then there will exist a $\tilde{f}$ in $\tilde{\mathcal{F}}$ with the desired property.

Useful characteristics of the distribution are that, for each x, the expectation of $\tilde{f}(x)$ is equal to $f(x)$, and moreover, conditioning on $\underline{X}, \underline{X}'$, by independence the expectation of the squared norms $\|\tilde{f} - f\|_{\underline{X}}^2$ and $\|\tilde{f} - f\|_{\underline{X}'}^2$, respectively, are equal to $1/m$ times the corresponding expected squared norms we would have with a single term.

Such conclusions are given in Lemma 8.4 in the appendix. In particular, for values $v \geq \|\beta\|_{1,a}$ and $a_h \geq \|h\|_n$, i.i.d. sampling with probabilities proportional to $\beta_h a_h$ produces $\|\tilde{f} - f\|_n^2$ with expectation less than $(v/m)\|\beta\|_{1,a}$. Likewise for stratified sampling and other sampling designs given there, similar conclusions holds with improvements in some

cases in the values of v and a_h. We will take advantage of these improvements below, but first, for simplicity, we continue with the implications from the i.i.d. sampling bound.

Using the fact that associated cross-terms have mean zero, conditioning also on $\underline{Y}$, the difference $\|Y - \tilde{f}\|_n^2 - \|Y - f\|_n^2$ has the same expectation as $\|\tilde{f} - f\|_n$. Likewise, $\|f^* - f\|_{\underline{X}'}^2 - \|f^* - \tilde{f}\|_{\underline{X}'}^2$, which we have also called $\text{diff}_n(f, \tilde{f})/n$, has an expectation which is minus an expected squared norm, which is obviously non-positive and hence can be ignored in setting the penalty.

The minimum over all $\tilde{f}$ of the expression (4.3) is a value not more than the expectation over $\tilde{f}$. Thus, using $L(\tilde{f}) \leq m_f \log(2M)$, we obtain an upper bound of the minimum over $\tilde{\mathcal{F}}$ of the distortion plus complexity relative to the sample size

$$\frac{v}{m_f}\|\beta\|_{1,a} + \frac{\gamma m_f \log(2M)}{n}. \tag{4.4}$$

Here we arrange $v \geq \|\beta\|_{1,a}$, picking v as a function of m_f. Indeed, let $m_f = \lceil \|\beta\|_{1,a}/\eta \rceil$ which determines $v = m_f \eta$ equal to $\|\beta\|_{1,a}$ rounded up to the nearest value in a grid of spacings η. Consequently we have demonstrated there is an $\tilde{f}$ for which expression (4.4) is not more than

$$\eta\|\beta\|_1 + \left[\frac{\|\beta\|_1}{\eta} + 1\right] \frac{\gamma \log(2M)}{n}. \tag{4.5}$$

By choosing the optimal $\eta = \sqrt{\gamma(\log 2M)/n}$, the expression (4.5) is equal to $\lambda^*\|\beta\|_1 + \gamma(\log 2M)/n$, where $\lambda^* = 2\sqrt{\gamma(\log 2M)/n}$. Therefore, $\text{pen}_n(f_\beta)/n$ of the form

$$\lambda\|\beta\|_1 + \gamma\frac{\log(2M)}{n} \tag{4.6}$$

satisfies the requirement (2.24) in Theorem 2.4 for $\lambda \geq \lambda^*$. This penalty is equivalent to $\text{Pen}_n(\beta)/n = \lambda\|\beta\|_1$ with the adjustment by $\gamma(\log 2M)/n$ absorbed into the risk bound.

4.2.3 Removing the boundedness restriction

Per Chapter 2, to avoid the boundedness restriction, we use expression (2.28) with the bound on distortion available with positive δ_3. Accordingly, for each f we want an $\tilde{f}$ for which $\text{pen}_n(f)/n$ exceeds the distortion plus complexity relative to sample size

$$D_n(f, \tilde{f}) + \frac{\gamma L_n(\tilde{f})}{n}, \tag{4.7}$$

where

$$D_n(f, \tilde{f}) = \|Y - \tilde{f}\|_n^2 - \|Y - f\|_n^2 + c_3 \|f - \tilde{f}\|_{\underline{X}'}^2.$$

Here $c_3 = \frac{1}{\tilde{c}\delta_3}$ and $\tilde{c} = (1 + \delta_1)(1 + \delta_2)$. For concreteness we set $\delta_3 = 1/\tilde{c}$ so that $c_3 = 1$ and the constant c in Corollary 2.6 is $c = 1 + \tilde{c}$.

The analysis in this generality is the same except that it yields a larger expectation over choices of $\tilde{f}$, now invoking $a_h \geq \|h\|_{\underline{X}'}$ as well as $a_h \geq \|h\|_{\underline{X}}$. Namely, the expectation of the distortion part is now multiplied by a factor of $(1 + c_3) = 2$. Accordingly the expectation of expression (4.7) is not more than $2(v/m)\|\beta\|_1 + \gamma(m/n)\log(2M)$. Bounding it in the same manner as before, with $v = m\eta$ and $m = m_f = \lceil \|\beta\|_1/\eta \rceil$, for which the optimal η is $\sqrt{\gamma(\log 2M)/(2n)}$, leads to validity of the penalty $Pen_n(\beta)/n$ of the form $\lambda\|\beta\|_1 + \gamma(\log 2M)/n$ for

$$\lambda \geq 2\sqrt{\frac{2\gamma \log(2M)}{n}}. \tag{4.8}$$

Using the refined complexity bound $m_f \log\left(4e \max\{M/m_f, 1\}\right)$, as detailed in a Lemma in the appendix, establishes the validity of $Pen_n(\beta)/n$ equal to $\lambda^*\|\beta\|_1$ adjusted by $\frac{\gamma \log(4eM)}{n} + \sqrt{\frac{\gamma}{n}}\frac{1}{e}$, where now

$$\lambda^* = 2\sqrt{\frac{2\gamma \log(4e \max\{M\sqrt{\gamma/n}, 1\})}{n}}. \tag{4.9}$$

This allows for a smaller order λ in the case that M is not large compared to $\sqrt{n}$. Slight

refinements of this bound are possible in which there is a role for $\|\beta\|_1$ in the denominator inside the logarithm, but the improvements obtained thereby appear to only effect smaller order terms, and yield a penalty concave in $\|\beta\|_1$. We prefer to stick with what can be obtained for the valid penalties linear or convex in $\|\beta\|_1$.

We remind that in the present simplified development, in obtaining the v^2/m approximation bound with v near $\|\beta\|_1 = \|\beta\|_{1,a}$, we assumed the weights a_h exceed $\|h\|_{\underline{X}}$ and $\|h\|_{\underline{X}'}$. Accordingly we take here $a_h = \|h\|_\infty$ in the traditional setting and allow $a_h = \sqrt{2}\|h\|_{2n}$ in the transductive setting.

Summarizing the current conclusion, we have established the following.

Lemma 4.1 (Validity of the ℓ_1 penalty in the finite library size case) *An ℓ_1 penalized least squares estimator satisfying (4.1) with penalty $\lambda\|\beta\|_{1,a}$, with λ either at least $2\sqrt{2\gamma(\log 2M)/n}$ or at least λ^* as in (4.9), fulfills the requirement of Corollary 2.6 for γ as stated there, such that, with $a_h = \|h\|_\infty$, the risk satisfies*

$$\mathbb{E}\|T\hat{f} - f^*\|^2 \le c\left[\inf_\beta \left\{\|f_\beta - f^*\|^2 + \lambda\|\beta\|_{1,a} + \mathbb{E}A_\beta\right\} + adjust/n\right],$$

where $\frac{adjust}{n}$ consists of $\frac{tail}{n}$ plus either $\gamma\frac{\log(2M)}{n}$ or $\left[\frac{\gamma\log(4eM)}{n} + \sqrt{\frac{\gamma}{n}\frac{1}{e}}\right]$, respectively; whereas, for the estimator with ℓ_1 penalty with weights $a_h = \sqrt{2}\|h\|_{2n}$,

$$\mathbb{E}\|T\hat{f} - f^*\|^2_{\underline{X}'} \le c\left[\inf_\beta \left\{\|f_\beta - f^*\|^2 + \lambda\|\beta\|_{1,a^*} + \mathbb{E}A_\beta\right\} + adjust/n\right],$$

where $a^*_h = \sqrt{2}\|h\|$.

A simple choice of the constants is to set $\delta_1 = 1/2$ and $\delta_2 = 1/3$. Then $c = 3$ and $\gamma = \frac{25}{16}(B + B')^2 + \frac{8}{3}\sigma^2 + 2(B + B')h_{Bern}$.

4.2.4 Computational issues of *LPGP*

The adjustment terms are negligible compared to the main terms; likewise the computation accuracy term A_β is negligible if the optimization is performed with sufficiently many steps. The resolvability $\inf_\beta \{\|f_\beta - f^*\|^2 + \lambda_{n,M}\|\beta\|_{1,a}\}$ determines the behavior of the risk, with multiplier $\lambda_{n,M}$ of order $\sqrt{(\log M)/n}$. In particular, for functions f^* that are near such f_β with moderate size ℓ_1 norm of coefficients, the rate is controlled by λ_n near $1/\sqrt{n}$ times a log factor.

When the penalty uses weights $a_h = \sqrt{2}\|h\|_{2n}$ based on the empirical L_2 norm on $\underline{X}, \underline{X}'$, the resolvability bound on the risk bound involves the expected value of the weights which are not more than $a_h^* = \sqrt{2}\|h\|$. Accordingly, $\|\beta\|_{1,a^*}$ is used in the risk bound. This appearance of the $L_2(P)$ weights rather than the L_∞ weights is a risk advantage identified when one has knowledge of future input data.

The expectation of the computational accuracy term $\mathbb{E}A_\beta$ has a simplified bound $\mathbb{E}A_\beta \leq \|\beta\|^2_{1,\|\cdot\|}/m_{n,\text{comp}}$ when $A_\beta = 4\|\beta\|^2_{1,\|\cdot\|_n}/m_{n,\text{comp}}$. Indeed, $\|\beta\|_{1,\|\cdot\|_n} = \sum_h \beta_h \|h\|_n$, so its square is a sum over pairs h, h' involving $\|h\|_n\|h'\|_n$, each of which by the Cauchy–Schwartz inequality has expectation not more than $\|h\|\|h'\|$. Accordingly, the expected square is not more than $[\sum_h \beta_h \|h\|\,]^2$, where $\|h\|$ is the $L_2(P)$ norm.

Let's briefly discuss the choice of number of computation steps $m_{n,\text{comp}}$. If it is $\sqrt{n/\log M}$, then for functions $f^* = f_{\beta^*}$ with moderate ℓ_1 norm, the computation accuracy is sufficient to retain the order $\sqrt{(\log M)/n}$ risk. For each f^*, let $\beta_{n,M}$ be the coefficient vector that optimizes the resolvability with $\lambda_{n,M}$. If $\|\beta_{n,M}\|_1$ is large, then a minimal $m_{n,\text{comp}}$ needed to retain risk of the order of that resolvability, is somewhat larger than $\sqrt{n/\log M}$, though always of order smaller than n.

Recalling the heart of the analysis so far, we have obtained a risk bound for ℓ_1 penalized least squares dictated by the presence of various m term subsets of the library which, when

adapted to f_β, provide approximation at rate $\|\beta\|_1^2/m$ balanced with complexity when m is of size determined by $\|\beta\|_1\sqrt{n/\log M}$.

4.3 Refined risk bound of extension to the infinite dictionary case

We turn now to improvements in the approximation bound and to obtain finite complexity bounds in certain infinite library cases. To avoid complication we first give such improvements in a case in which the representors take a simple form, so we can retain the same tools for their descriptive complexity.

4.3.1 Two levels of cover

The improvements are based on covering properties of the library $\mathcal{H}$. We find usefulness of two levels of cover. For infinite size libraries, at a fine precision ε_1 typically of order approximately $1/\sqrt{n}$, we use finite empirical covers for the purpose of finding an effective library size M_1. This size M_1 serves as the surrogate for M in the expressions for complexity, with a small added price in distortion. This effective library size is permitted to be large compared to the sample size. As before it appears in the risk bounds through the ratio $(\log M_1)/n$.

At another precision ε_2, not nearly as small, we consider moderate improvement in the v^2/m approximation bound, and hence improvement in the distortion, by stratified approximation by partitioning of the library into a number of cells m_0 and maintaining $m \geq m_0$. Consequently, as we show, both the distortion and complexity terms in the penalty can be improved. As for the relationship between the two precisions, the best tradeoff will occur when ε_1 is of order $\varepsilon_2/\sqrt{n}$. Let's consider the improvement of the

distortion properties first.

Let $\tilde{\mathcal{H}}_2$ be a finite subset of functions from the library $\mathcal{H}$. This subset may depend on $\underline{X}, \underline{X}'$. Each $h \in \mathcal{H}$ has a distance denoted $\varepsilon_{h,2} = \min_{\tilde{h}\in\tilde{\mathcal{H}}_2} \|h - \tilde{h}\|_{2n}$, from the set $\tilde{\mathcal{H}}_2$, where for this part of the analysis, the appropriate norm is the empirical L_2 norm on the $2n$ points $\underline{X}, \underline{X}'$.

4.3.2 Refining risk bound using the L_2 covering property

Let $\varepsilon_2 \geq \sup_{h\in\mathcal{H}} \varepsilon_{h,2}$ bound the precision of $\tilde{\mathcal{H}}_2$ as a cover of the library $\mathcal{H}$. Let $m_0 \geq |\tilde{\mathcal{H}}_2|$ bound the cardinality of $\tilde{\mathcal{H}}_2$ and let $m_1 = m - m_0$. Appealing to the stratified sampling argument of Case 3 of Lemma 8.4, there is an equally weighted linear combination $f_m = (v/m)\sum_{i=1}^{m} h_k$ with terms h_k in $\mathcal{H}$ selected from those that form f, such that $\|Y - f_m\|_n^2 - \|Y - f\|_n^2 + \|f - f_m\|_{\underline{X}'}^2$ is not more than $2\varepsilon_2^2\|\beta\|_1 v/m$, so in the present case taking our representor $\tilde{f}$ to be this f_m the distortion satisfies

$$D_n(f, \tilde{f}) \leq 2\varepsilon_2^2 \frac{v}{m}\|\beta\|_1 = 2\eta\varepsilon_2\|\beta\|_1,$$

where $v/m \leq \|\beta\|_1/(m - m_0)$ and $\|\beta\|_1 = \|\beta\|_{1,1} = \sum_h |\beta_h|$. Moreover, for any specified $\eta > 0$ we arrange for $v/m = \eta/\epsilon_2$ and for our representor to use a total of $m = m_f$ terms, that is, the sum of the number of terms in each cell, where, due to integer rounding effects, m is between $\varepsilon_2 \sum_h \beta_h/\eta$ and $\varepsilon_2 \sum_h \beta_h/\eta + m_0$ (where η/ε_2 here plays the role of η in Lemma 8.4).

The complexity term $L_n(\tilde{f})$ is set to be $m \log(2M)$ as before, now interpreted as a sum of three parts: $m_0 \log 2$ for the description length of m_0, plus $m_1 \log 2$ for the description of m_1, and $m \log M$ for the description of the choices of h.

Using these bound on the distortion plus complexity, expression (4.7) is less than or

equal to

$$2\eta\varepsilon_2\|\beta\|_1 + \left[\varepsilon_2\frac{\|\beta\|_1}{\eta} + m_0\right]\frac{\gamma\log(2M)}{n}. \tag{4.10}$$

Note the similarity to the previous case, but with ε_2 multiplying $\|\beta\|_1$ and with the larger added term based on $m_0 \geq 1$. Again the optimal η^* is $\sqrt{\gamma(\log 2M)/(2n)}$. Accordingly, with the adjustment for $\gamma m_0(\log 2M)/n$, we have the validity of $\text{Pen}_n(\beta)/n$ of the form

$$\lambda\|\beta\|_1, \tag{4.11}$$

for $\lambda \geq 4\varepsilon_2\eta^* = 2\varepsilon_2\sqrt{2\gamma(\log 2M)/n}$. For the adjustment to remain small compared to this penalty requires that m_0 be of somewhat smaller order than $\sqrt{n}$. This restricts ε_2 to be not very small, potentially tending to zero at a slow polynomial rate, as will be discussed further for libraries with finite metric dimension properties.

The log factors here can be reduced. Indeed, the $m\log(2M)$ bound on the complexity may be reduced to $m\log(4e\max\{M/m, 1\})$ as in the previous Lemma. Now since $m \geq m_0$ and since we may assume that $m_0 \leq M$, this complexity may be replaced with the upper bound $m\log(4eM/m_0)$, retaining the linearity in m as needed in the above argument. Accordingly, each of the $\log(2M)$ expressions above may be replaced by $\log(4eM/m_0)$, which is an improvement for $m_0 > 6$.

4.3.3 Extension to the Infinite dictionary $\mathcal{H}$ using L_1 cover

Some of the examples involve continuously parameterized libraries, naturally infinite in cardinality. With finite empirical covering properties we can define an effective cardinality M_1 to use in place of M. Not only is this idea useful for infinite libraries, it can also apply for finite libraries to reduce the size of the multiplier for ℓ_1 penalties, if some of the functions in the library are highly correlated.

To determine the effective cardinality of $\mathcal{H}$, consider another empirical cover denoted $\tilde{\mathcal{H}}_1$, but with much finer precision and, accordingly, with cardinality typically much larger than the $\tilde{\mathcal{H}}_2$ considered above. Let M_1 denote an upper bound on the cardinality of this cover. Let $\varepsilon_1 \geq \sup_{h\in\mathcal{H}} \varepsilon_{h,1}$ bound its precision, with $\varepsilon_{h,1} = \min_{\tilde{h}\in\tilde{\mathcal{H}}_1} \|h - \tilde{h}\|_{2n,1}$, using the empirical L_1 norm defined by $\|h\|_{2n,1} = \frac{1}{2n}\sum_{i=1}^{n}(|h(X_i)| + |h(X_i')|)$, as this choice of norm is sufficient for analyzing the effect of this cover. One may arrange it to be an another empirical L_2 cover (indeed an L_2 cover is also an L_1 cover), but the best L_1 cover of a given precision may have somewhat smaller size.

Appropriate choices for the precision ϵ_1 are discussed following the proof of the Theorem below along with implications of the effective library size for libraries of finite metric dimension.

4.3.4 General ℓ_1 penalty conclusions

We give our general ℓ_1 penalty conclusions in the following Theorem. We remind that the covers $\tilde{\mathcal{H}}_1$ and $\tilde{\mathcal{H}}_2$ are permitted to depend on $\underline{X}, \underline{X}'$. We require that they be coordinate pair symmetric. Indeed, in accordance with the symmetry of the respective empirical norms, optimal size covers for specified precisions ε_1 and ε_2 have such symmetric. For the transductive setting we allow the cardinalities M_1 and m_0 and the precisions to depend on $\underline{X}, \underline{X}'$; whereas, for the traditional setting, we require that ε_1 and ε_2 be specified, not depending on the data, and that M_1 and m_0 denote constant bounds on the respective cardinalities, even though the covers are data-dependent.

The covers of the library are only used in setting a possibly smaller λ than before, otherwise the covers are not used in constructing the estimator. In implementation one only needs to know of bounds on the sizes and precisions of covers; we do not need explicit presentation of the covers we know to exist. Indeed, the estimator continues to be the optimizer of the penalized least squares over all $(\beta_h : h \in \mathcal{H})$ with the indicated λ.

Theorem 4.2 (Validity of the penalty $\lambda\|\beta\|_1$ with refinement of valid λ) *Given positive constants δ_1 and δ_2, the first two conclusions that follow are for $\|\beta\|_1 = \sum_{h\in\mathcal{H}} |\beta_h|$. First, if $M_{\mathcal{H}} = M$ is finite and has an empirical L_2 cover of precision ε_2 and cardinality not more than m_0, then with*

$$\lambda \geq 2\varepsilon_2\sqrt{\frac{2\gamma\log(2M)}{n}}$$

the penalty fulfills the requirement of Corollary 2.6, yielding risk in the traditional setting which satisfies

$$E\|T\hat{f} - f^*\|^2 \leq c\left[\min_{\beta}\left\{\|f_\beta - f^*\|^2 + \lambda\|\beta\|_1\right\} + \frac{adjust}{n}\right],$$

with $c = 1 + (1+\delta_1)(1+\delta_2)$ and adjust $= tail_2 + \gamma m_0 \log(2M)$.

Second, allowing M finite or infinite, if there is also an empirical L_1 cover of precision ε_1 and cardinality not more than M_1, then refining the allowed λ to be at least $2\varepsilon_2\sqrt{\frac{2\gamma\log(2M_1)}{n}} + 16B'\varepsilon_1$ yields the corresponding risk bound where now adjust $= tail_1 + tail_2 + 2\gamma m_0 \log(2M_1)$. Moreover, the statements above hold with the $\log(2M)$ and $\log(2M_1)$ factors replaced by $\log(4eM/m_0)$ and $\log(4eM_1/m_0)$, respectively, assuming M and M_1, respectively, are at least m_0.

Third, with M finite or infinite, we allow the penalty to be equal to $\|\beta\|_{1,a} = \sum_{h\in\mathcal{H}} |\beta_h| a_h$ with variable weights $a_h \geq \lambda_2\epsilon_{h,S,2} + \lambda_1\epsilon_{h,\tilde{\mathcal{H}}_1,1}$, where $\epsilon_{h,S,2} = \min_{g\in S}\|h - g\|_{2n}$ is the empirical L_2 distance of h to the linear span S of a given subset of $\tilde{\mathcal{H}}_2$ of size not more than m_0, and $\epsilon_{h,\tilde{\mathcal{H}}_1,1}$ is the empirical L_1 distance of h to the subset of cardinality M_1. Moreover, $\lambda_2 = 2\sqrt{\frac{2\gamma\log(eM_1c_n)}{n}}$ and $\lambda_1 = 8B'$, where $c_n = 4e^2\max\{1, nc'\}$ and $c' = 8(B')^2/[\gamma(m_0+1)]$. Then the risk bound holds with $\|\beta\|_{1,a}$ in place of $\lambda\|\beta\|_1$ and with adjust $= 2tail_1 + tail_2 + \gamma[2\log M_1 + (m_0+2)\log c_n e)]$.

In the transductive setting, the same conclusions hold where on the left side we put the risk $E\|T\hat{f} - f^*\|^2_{\underline{X}'}$ and on the right side we have also an expectation recognizing in that

case the possible data-dependence of M_1 and m_0 and accordingly also of λ and a_h.

Proof: The first claim is established in the development preceding the theorem statement and the third claim is proved in the appendix using a more elaborate approximation and covering argument. We give now the proof of the second claim.

We are to establish validity of a penalty of the form $\lambda\|\beta\|_1$ by showing that every $f = f_\beta$ has a representor such that $\lambda\|\beta\|_1$ exceeds expression (4.7). We begin by the same argument as above, noting that for λ at least $4\epsilon_2\eta$, this $\lambda\|\beta\|_1$ exceeds the expression

$$\|Y - f_m\|_n^2 - \|Y - f\|_n^2 + \|f - f_m\|_{\underline{X}'}^2 + \frac{\gamma L(\tilde{f}_m)}{n}, \tag{4.12}$$

where f_m is an m term approximation to f, with terms h_k in $\mathcal{H}$, where m depends on f. We form our representor $\tilde{f}_m$ by replacing each such h_k with the closest $\tilde{h}_k$ in $\tilde{\mathcal{H}}_1$. This yields $\tilde{f}_m = (v/m)\sum_{k=1}^m \tilde{h}_k$, in the present case of constant weights $a_k = 1$. The complexity expression $L(\tilde{f}_m)$ and the value of $\eta = \eta^*$ are taken to be the same as before, except now with M_1 in place of M.

We seek a lower bound on expression (4.12) using a corresponding expression with $\tilde{f}_m$ in place of f_m. Equivalently, adding $\|Y - f\|^2$ to $\lambda\|\beta\|_1$ and to expression (4.12), we seek such a bound on

$$\|Y - f_m\|_n^2 + \|f - f_m\|_{\underline{X}'}^2 + \frac{\gamma L(\tilde{f}_m)}{n}, \tag{4.13}$$

which we pursue by first replacing the functions in (4.13) by their truncations. Indeed, since (4.13) is of rectifiable form, we have the lower bound, as in Lemma 2.5,

$$\|Y - Tf_m\|_n^2 - \frac{1}{n}\text{Tail}_2 + \|Tf - Tf_m\|_{\underline{X}'}^2 + \frac{\gamma L(\tilde{f}_m)}{n}. \tag{4.14}$$

Next, when we replace the functions h_k by their representors $\tilde{h}_k$, we produce $\tilde{f}_m$ with $|Tf_m(x) - T\tilde{f}_m(x)| \le |f_m(x) - \tilde{f}_m(x)| \le (v/m)\sum_{k=1}^m |h_k(x) - \tilde{h}_k(x)|$. The average of

this bound across the $2n$ points of $\underline{X}$ and $\underline{X}'$ is denoted $V_{m,\varepsilon_1} = (v/m)\sum_{k=1}^{m} \varepsilon_{h_k,1}$, which is not more than $v\varepsilon_1$.

Using the rule for differences of squares, $[Tf(x) - Tf_m(x)]^2$ is the same as $[Tf(x) - T\tilde{f}_m(x)]^2 - [2Tf(x) - Tf_m(x) - T\tilde{f}_m(x)][Tf_m(x) - T\tilde{f}_m(x)]$, which is at least

$$[Tf(x) - T\tilde{f}_m(x)]^2 \; - \; 4B' \, |Tf_m(x) - T\tilde{f}_m(x)|.$$

Likewise, term by term, $[Y_i - Tf_m(X_i)]^2$ is at least $[Y_i - T\tilde{f}_m(X_i)]^2 - 4B'|Tf_m(X_i) - T\tilde{f}_m(X_i)| - \text{Tail}_1(Y_i)$, where $\text{Tail}_1(Y_i) = 4B'(|Y_i| - B')1\{|Y_i| > B'\}$. Combining these inequalities, using the bound on the average of $|Tf_m(x) - T\tilde{f}_m(x)|$ with respect to the $\underline{X}$ and $\underline{X}'$ points, we obtain that expression (4.14) is at least

$$\left[\|Y - T\tilde{f}_m\|_n^2 + \|Tf - T\tilde{f}_m\|_{\underline{X}'}^2 + \frac{\gamma L(\tilde{f}_m)}{n} \right] - \frac{\text{Tail}_1 + \text{Tail}_2}{n} - 8B'V_{m,\varepsilon_1} \qquad (4.15)$$

where Tail_1 and Tail_2 are the corresponding sums of $\text{Tail}_1(Y_i)$ and $\text{Tail}_2(Y_i)$, respectively. We recognize the expression in brackets is of the desired form.

Here V_{m,ε_1} is not more than $v\varepsilon_1$. We have two upper bounds on v, either of which we may put to use. On one hand $v \leq \|\beta\|_1 + m_0\eta/\varepsilon_2$, which yields $8B'V_{m,\varepsilon_1}$ not more than $8B'\varepsilon_1\|\beta\|_1 + 8B'm_0\eta\varepsilon_1/\varepsilon_2$. The term $8B'\varepsilon_1\|\beta\|_1$ is added to the penalty and the last term $8B'm_0\eta\varepsilon_1/\varepsilon_2$, which will be seen to be negligible, is added to the adjustment. Accordingly, with penalty at least $[4\eta\varepsilon_2 + 8B'\varepsilon_1]\|\beta\|_1$ and with adjustment by $\frac{\gamma m_0 \log(2M_1)}{n} + 8B'm_0\eta\varepsilon_1/\varepsilon_2$ along with the tail terms, we obtain a valid penalized squared error criterion exceeding the expression in brackets in (4.15) for satisfaction of the conditions of our theory.

Alternatively, we have $v \leq \|\beta\|_1 m/(m - m_0)$, with which we may arrange for the choice of m to be the maximum of the previous choice m_f and $2m_0$, so that $m/(m - m_0) \leq 2$. Hence $8B'V_{m,\varepsilon_1}$ has a second bound $16B'\varepsilon_1\|\beta\|_1$. Accordingly, the addition to the

penalty can be instead $16B'\varepsilon_1\|\beta\|_1$. In verifying the penalty condition, we use that the new $m \leq \epsilon_2\|\beta\|_1/\eta + 2m_0$, such that, with the new factor of 2 on the m_0, the adjustment term becomes $2\frac{\gamma m_0 \log(2M_1)}{n}$, verified in the same manner as at expressions (4.10), and (4.11), with M_1 in place of M.

Moreover, as before, assuming $m_0 \leq M_1$, each of the $\log(2M_1)$ factors may be replaced with $\log(4eM_1/m_0)$, as indicated in the Theorem statement. This completes our analysis for the second claim of the Theorem. As we said the proof of the third claim is in the appendix. ∎

Remarks: 1. In the special case of $m_0 = 1$, with $\tilde{\mathcal{H}}_2$ consisting of a single function equal to 0, the assumption concerning ε_2 corresponds to $\varepsilon_2 = b \geq \sup_h \|h\|_{2n}$, yielding conclusions for the equally weighted ℓ_1 norm $\|\beta\|_1 = \sum_h |\beta_h|$ that are analogous to the previous Lemma. Now allowing larger m_0, that result is improved by the ε_2 factor which bounds the empirical distance of functions in $\mathcal{H}$ from data dependent covers.

2. We note that the third conclusion is more flexible in what it allows of the penalty, regarding the use of the distance to the linear span S which is in smaller than the distance to $\tilde{\mathcal{H}}_2$. Also, like Lemma 4.1, this case allows for variable weights a_h, now possibly much smaller. With variable weights the conclusion is closely related to the Lemma 4.1 result when S is trivial, consisting only of the function equal to 0. Nevertheless, this third conclusion does not subsume the others, because the more elaborate covering argument it requires leads to the additional factor of n inside the logarithm in the definition of λ_2.

4.4 ℓ_1 Penalties for Libraries of Finite Metric Entropy

We now discuss the choices of ε_1 and ε_2 and the form of the penalty for libraries of finite metric dimension, using the results of the second claim in Theorem 4.2 above.

First focus attention on the choice of the precision ε_1 as a function of other character-

istics of the setting.

Recall that an infinite library $\mathcal{H}$ is said to have metric dimension $d_1 = d_{\mathcal{H},1}$ with respect to the empirical L_1 norm, if there are positive constants b_1 and c_1 such that for every positive $\varepsilon \leq b_1$, every $n \geq 1$, and every $(\underline{X}, \underline{X}')$, the best empirical L_1 cover of precision ε has cardinality $M_{\mathcal{H},\varepsilon}$ not more than $(c_1/\varepsilon)^{d_1}$, where $c_1 \geq b_1 \geq \sup_{h\in\mathcal{H}} \|h\|_{2n,1}$. Here the cover may be data-dependent, even though ε and the cardinality bound $(c_1/\varepsilon)^{d_1}$ are not.

An important class of examples, as we recall in the Chapter 7, are those for which the functions $h(x)$ are uniformly bounded by a constant b and the graph class of $\mathcal{H}$, that is, the class of sets $\{x : h(x) \leq \tau\}$, $\tau \in \mathbb{R}$, $h \in \mathcal{H}$, has VC-dimension d. By Theorem 13 of Chapter 10 in Pollard [69], which is a result based on Haussler [52], $\mathcal{H}$ has empirical dimension not more than d with respect to the empirical L_1 norm, with $M_{\mathcal{H},\varepsilon} \leq e(d+1)(4eb_1/\varepsilon)^d \leq e(4e^2b_1/\varepsilon)^d$ for all $\varepsilon \leq b_1$. In this case the associated c_1 is equal to $4eb_1[e(d+1)]^{1/d}$.

Let's explore consequences of $\mathcal{H}$ having finite metric dimension. Rounding down to an integer, we set

$$M_1 = (c_1/\varepsilon_1)^{d_1}.$$

Accordingly, for each ϵ_2, the associated best penalty multiplier $\lambda^*_{n,d}$ corresponds to optimization over choices of ε_1 of values of λ of order $\varepsilon_2\sqrt{\frac{d_1 \log(1/\varepsilon_1)}{n}} + \varepsilon_1$. The best such ε_1 is approximately $\varepsilon_2\sqrt{d_1/n}$ to within a log factor and produces a $\lambda^*_{n,d}$ of order

$$\varepsilon_2\sqrt{(d_1/n)\log(n/d_1)}. \tag{4.16}$$

This is a pleasing result. It shows, for sequences of libraries and sample sizes indexed by the metric dimension d_1 and the sample size n, that the multipliers $\lambda_{n,d}$ can be arranged to be small whenever d_1/n is small. Moreover, in view of the index of resolvability bound on

the risk of an ℓ_1 penalized least squares estimator, its risk tends to zero at a rate controlled by this $\varepsilon_2\sqrt{d_1/n}$ times a log factor, provided f^* is in the closure of the linear span of the library. In particular, if the target function f^* has finite $V(f^*)$ then the risk of the estimator tends to zero at rate $\lambda_{n,d}$ of order $\epsilon_2\sqrt{(d_1/n)\log(n/d_1)}$, assuming the adjustment terms have an order of behavior not larger than this.

What is pleasing is that this rate is at least as good as the power $1/2$ on the d_1/n term and this rate does not degrade to worse that this critical exponent $1/2$ as the library dimension gets large. This assurance of a rate that is at least as good as the dimension-independent rate $1/2$ is a type of avoidance of the curse of dimensionality for functions that have finite variation with respect to the library, for libraries of increasing dimension, so long as d is of smaller order than n. These properties hold even though the effective size of the library $M_1 = (c_1/\varepsilon_1)^{d_1}$ is much larger than the sample size n for $d_1 \geq 2$. Improvements that arise from the factor ε_2 are icing on the cake.

We pin down the specifics of a suitable multiplier $\lambda_{n,d}$ including the constants. We have $\lambda = 2\left[\varepsilon_2\sqrt{2\gamma\frac{d_1\log(c_1/\varepsilon_1)+\log 2}{n}} + 4B'\varepsilon_1\right]$. From the inequality $A + B \leq [2(A^2 + B^2)]^{1/2}$, it is not more than $4\left[\varepsilon_2^2\gamma\frac{d_1\log(c_1/\varepsilon_1)+\log 2}{n} + 8B'^2\varepsilon_1^2\right]^{1/2}$, which is exactly optimized at

$$\varepsilon_1 = \frac{\varepsilon_2}{4B'}\sqrt{\frac{\gamma d_1}{n}}, \tag{4.17}$$

with an assumption that $\frac{1}{4B'}\sqrt{\frac{\gamma d_1}{n}} \leq 1$, so that if ε_2 is not more than a constant needed for the metric dimension control then also ε_1 is not more than it.

Conveniently then we set this choice of ε_1, equal to a constant times $\varepsilon_2\sqrt{d_1/n}$ as expressed in (4.17). At this choice we have

$$\lambda_{n,d} = 2\varepsilon_2\left[\sqrt{2\gamma\frac{d_1\log(c_1/\varepsilon_1)+\log 2}{n}} + \sqrt{\gamma\frac{d_1}{n}}\right]. \tag{4.18}$$

This gives the advertised $\varepsilon_2\sqrt{(d_1/n)\log(n/d_1)}$ rate, including the trivial case with $m_0 = 1$ corresponding to the single function equal to 0, where then $\varepsilon_2 = b_{\mathcal{H}} \geq \sup_{h\in\mathcal{H}} \|h\|_{2n}$.

We proceed to take advantage of smaller ε_2 and corresponding larger m_0. Finite metric dimensionality with respect either the empirical L_1 norm or the empirical L_2 norm implies finite metric dimensionality with respect to the other, with $d_{\mathcal{H},1} \leq d_{\mathcal{H},2} \leq 2d_{\mathcal{H},1}$. With $d_2 = d_{\mathcal{H},2}$ there are constants b_2 and c_2 such that for every $\varepsilon_2 \leq b_2$ and every $\underline{X}, \underline{X}'$ there is an optimal empirical L_2 cover of $\mathcal{H}$ of precision ε_2 with cardinality not more than $(c_2/\varepsilon_2)^{d_2}$, where $c_2 \geq b_2 \geq \sup_{h\in\mathcal{H}} \|h\|_{2n}$. The following corollary demonstrates refinements for the libraries with finite metric dimension.

Corollary 4.3 *Assume the library $\mathcal{H}$ has finite metric dimension d_1 and d_2 with respect to L_1 norm and L_2 norm respectively.*

(1). The ℓ_1-penalized least squares estimator with $pen_n(f_\beta) = \lambda\|\beta\|_1$ with λ at least

$$\lambda_{n,d} = C_1(d_2)(1+\sqrt{d_1/\rho})\left(\frac{\gamma\rho}{n}\right)^{(d_2+2)/(2d_2+2)},$$

where $C_1(d_2) = d_2^{1/(d_2+1)}(2c_2)^{d_2/(d_2+1)}$ and $\rho = d_1\log(cn/d_1) + 2\log(4e)$ with $c = \frac{(4B'c_1)^2}{c_2^2\gamma}$, satisfies the requirement of the second conclusion in Theorem 4.2. In particular, if the target function f^* has finite variation $V(f^*)$, there exist such multipliers λ, such that the the risk tends to zero at rate of order

$$\left[\frac{\gamma d_1}{n}\log\frac{n}{d_1}\right]^{\frac{d_2+2}{2(d_2+1)}}.$$

Here if the noise ϵ in the regression is bounded, γ is a constant. Otherwise, according to Corollary 2.6, the quantity γ is of order $\log^2 n$; whereas if ϵ is sub-Gaussian, γ is of order $\log n$.

(2). Using penalty $pen_n(f_\beta) = \lambda \|\beta\|_1^{d_2/(d_2+1)}$ *with* λ *at least*

$$\lambda'_{n,d} = C(d_1, d_2) \left(\frac{\gamma\rho}{n}\right)^{(d_2+2)/(2d_2+2)},$$

where $C(d_1, d_2)$ *is defined in the proof and* ρ *is the same as in (1), the penalized least squares estimator* $\hat{f}$ *satisfies the index resolvability risk bound*

$$E\|T\hat{f} - f^*\|^2 \leq (1+\delta)\left[\min_\beta \left\{\|f_\beta - f^*\|^2 + \lambda\|\beta\|_1^{\frac{d_2}{d_2+1}}\right\} + \frac{adjust}{n}\right],$$

where $\frac{adjust}{n}$ *is of smaller order than* $\lambda'_{n,d}$.

Proof: Both conclusions are proven by similar arguments. First replace the $\log(2M_1)$, arising as $d\log(c_1/\varepsilon_1)+\log 2$ in the expression (4.18) for λ, with the alternative $\log(4eM_1/m_0)$. The choice of ϵ_1 set in equation (4.17) still has the indicated optimization property where now the $\log 2$ is replaced by $\log(4e/m_0)$.

Set

$$m_0 = \lceil (c_2/\varepsilon_2)^{d_2} \rceil. \tag{4.19}$$

Now since $d_2 \geq d_1$ and $c_2/\varepsilon_2 \geq 1$ we have that

$$\frac{M_1}{m_0} \leq \frac{(c_1/\varepsilon_1)^{d_1}}{(c_2/\varepsilon_2)^{d_2}} \leq \left[\frac{c_1/\varepsilon_1}{c_2/\varepsilon_2}\right]^{d_1} = \left[\frac{4B'c_1}{c_2}\sqrt{\frac{n}{\gamma d_1}}\right]^{d_1} = \left[c\frac{n}{d_1}\right]^{d_1/2},$$

where $c = \frac{(4B'c_1)^2}{c_2^2\gamma}$. Consequently, the multiplier for the ℓ_1 penalty may be set to be any value of λ at least

$$\lambda_{n,d} = 2\varepsilon_2\sqrt{\gamma}\left[\sqrt{\rho/n} + \sqrt{d_1/n}\right],$$

where $\rho = d_1 \log\left(c\frac{n}{d_1}\right) + 2\log 4e$. We note that the $\sqrt{d_1/n}$ is of smaller order than the $\sqrt{\rho/n}$ term by a log factor. Likewise, for the adjustment $\frac{\gamma m_0 \log(4eM_1/m_0)}{n} + 8B'm_0\eta\varepsilon_1/\varepsilon_2$, as suggested preceding the theorem statement, with $\eta = \sqrt{\frac{\gamma\rho}{4n}}$, we see that it can be upper

bounded by $\frac{\gamma m_0}{n}\left[\rho/2+\sqrt{\rho d_1}\right]$ since $2\log(4eM_1/m_0)$ is not more than ρ. Once again, the second part of the adjustment is negligible compared to the first by a log factor.

We now have the ingredients with which to address the choice of ε_2. Consider the contributions from $\lambda_{n,d}\|\beta\|_1$ and from the main part of the adjustment. Using the identity (4.19), we have that the adjusted penalty as appears in the index of resolvability is equal to

$$2\varepsilon_2\sqrt{\frac{\gamma\rho}{n}}\|\beta\|_1+\frac{1}{2}\left(\frac{c_2}{\varepsilon_2}\right)^{d_2}\frac{\gamma\rho}{n}. \tag{4.20}$$

(1) Picking a reference value v_0 for the variation $\|\beta\|_1$, a near optimal ε_2 for expression (4.20) is equal to $\varepsilon_2^*=\left(\frac{C_0(d_2)}{v_0}\right)^{1/(d_2+1)}\left(\frac{\gamma\rho}{n}\right)^{1/(2d_2+2)}$ where $C_0(d_2)=d_2c_2^{d_2}/4$. Plugging this value of ε_2 into the expression for $\lambda_{n,d}$, we allow multipliers λ that are at least

$$C_1(d_2)v_0^{-\frac{1}{d_2+1}}\left(1+\sqrt{\frac{d_1}{\rho}}\right)\left(\frac{\gamma\rho}{n}\right)^{\frac{d_2+2}{2(d_2+1)}},$$

where $C_1(d_2)=d_2^{1/(d_2+1)}(2c_2)^{d_2/(d_2+1)}$. The first statement of the Corollary is a special case with the reference v_0 to be 1. As we mentioned before, $\sqrt{d_1/n}$ is of smaller order than $\sqrt{\rho/n}$ by a log factor. Therefore, $\lambda_{n,d}$ is of order $\left(\frac{\gamma\rho}{n}\right)^{\frac{d_2+2}{2(d_2+1)}}$, which is the same as of order $\left[\frac{d_1}{n}\log\left(\frac{n}{d_1}\right)\right]^{\frac{d_2+2}{2(d_2+1)}}$. Consequently, if the target function f^* has finite $V(f^*)$, the index of resolvability tends to zero at the rate of $\lambda_{n,d}$. This proves the conclusion.

(2) Now we do not pick a fixed ε_2 for all β. Instead, we use the ε_2^* to optimize the expression (4.20), which is equal to

$$\left(\frac{C_0(d_2)}{\|\beta\|_1}\right)^{1/(d_2+1)}\left(\frac{\gamma\rho}{n}\right)^{1/(2d_2+2)}.$$

This ε_2^* and the corresponding $m_0=(c_2/\varepsilon_2^*)^{d_2}$ are functions of β, so we include the associated terms in our penalty. Hence, using the optimal ε_2^*, we obtain validity of penalties

to be at least,

$$C(d_1, d_2) \left(\frac{\gamma\rho}{n}\right)^{\frac{d_2+2}{2(d_2+1)}} \|\beta\|_1^{d_2/(d_2+1)},$$

where $C(d_1, d_2) = (d_2^{1/(d_2+1)}(1+\sqrt{d_1/\rho}) + d_2^{-d_2/(d_2+1)}(1+2\sqrt{d_1/\rho}))(2c_2)^{d_2/(d_2+1)}$. Consequently according the second conclusion of Theorem 4.2, the resolvability risk bound follows. ∎

Remarks:

1. When the target function f^* has finite variation $V(f^*)$, the risk of the penalized least squares estimators with both penalties $\lambda_{n,d}\|\beta\|_1$ and $\lambda'_{n,d}\|\beta\|_1^{d_2/(d_2+1)}$ tend to zero at the same rate. This rate is strictly better than the power $1/2$ on d_1/n. The smaller the empirical L_2 dimension d_2 is, the faster our risk converges to zero.

2. When the target function f^* has infinite variation, using the second penalty $\lambda_n\|\beta\|_1^r$, where $r < 1$, provides a faster rate. Indeed, consider the squared approximation error $App(f^*, v) = \inf_{f_\beta:\|\beta\|_1=v}\{\|f_\beta - f^*\|^2\}$, which is a decreasing function of v. Now the index of resolvability is $R_r^1(f^*, \lambda_n) = \inf_v\{App(f^*, v) + \lambda_n v^r\}$, with penalties to be $\lambda_n\|\beta\|_1^r$ for $r \leq 1$. If $App(f^*, v)$ is a polynomial function of v, the solution $R_r^1(f^*, \lambda_n)$ is also a polynomial function, which can be solved explicitly to reveal the power of the λ_n and to show the rate is decreasing with respect to r. Even if $R_r^1(f^*, \lambda_n)$ is not a polynomial function, one still gets a rate improvement for $r < 1$ compared to $r = 1$.

4.5 Comment on Variable Complexity Libraries

In Lemma 4.1, we used a constant complexity $\log M$ for members of a finite library $\mathcal{H}$. Variable complexities $L(h)$ for h in $\mathcal{H}$, satisfying $\sum_h e^{-L(h)} \leq 1$, may be used for countable libraries. Then the best a_h, via our technique without taking any advantage of metric covering properties of $\mathcal{H}$, is equal to $a_{L,h} = \|h\|_{2n}\sqrt{L(h)+\log 2}$ (for the traditional setting, we may use $\|h\|_\infty\sqrt{L(h)+\log 2}$). The analogous conclusion is achieved showing

$\lambda\|\beta\|_{1,a_L}$ to be a valid penalty for $\lambda \geq 2\sqrt{2\gamma/n}$ with corresponding risk bound. This extends Lemma 4.1 by using $\sqrt{L(h)+\log 2}$ inside the sum defining $\|\beta\|_{1,a_L}$ in place of the constant $\sqrt{\log M+\log 2}$ outside the sum. The proof is similar to that for Lemma 4.1 except that we express the complexity as $L(\tilde{f}) = \sum_{k=1}^{m}[L(h_k) + \log 2]$; the representors $\tilde{f} = \frac{C}{m}\sum_k h_k(x)/c_{h_k}$ are permitted to use different weights c_{h_k}; the distribution used on the h_k chooses each h in $\mathcal{H}$ with probability proportional to $\beta_h c_h$; and we compute the resulting expectation of the distortion plus complexity $D_n(f, \tilde{f}) + \gamma L(\tilde{f})/n$, leading to a penalty expression optimized at $c_h = \|h\|_{2n}/\sqrt{L(h) + \log 2}$.

Chapter 5

Risk Bound For Subset Selection

In this chapter, we first extend the result presented in Chapter 2 to allow penalty depending on indices m as well as the functions f. Then we apply these general risk conclusions along with the computation bounds for greedy algorithms provided in Chapter 3, to establish risk bound for these estimators based on subset selection. An improvement to generate an estimator better than both the ℓ_1 penalized estimator and the subset selection is given at the end of the chapter.

5.1 General resolvability risk bound allowing penalty depending on indices

Suppose $\mathcal{H}$ is a set of functions, each with finite $L_2(P)$ norm. Recall that $\underline{X} = (X_i)_{i=1}^n$ and $\underline{Y} = (Y_i)_{i=1}^n$ are training data and $\underline{X}' = (X_i')_{i=1}^n$ is an independent copy of $\underline{X}$. We now state a variant of the result in Chapter 2, taking advantage of properties of models $\{\mathcal{F}_m\}_{m\in\mathcal{M}}$ for which $\mathcal{F}$ is the union, where $\mathcal{M}$ is an index set. For each $m \in \mathcal{M}$, we allow a comparison class $\mathcal{F}_m^{\text{co}}$ for approximate optimization which might be larger than $\mathcal{F}_m$. Suppose $\hat{f}$, $\hat{m}$ approximately minimizes the penalized least squares criterion $\|Y -$

$f\|_n^2 + \frac{\text{pen}_n(f,m)}{n}$ relative to the comparison sets, in the sense that $\hat{f}$ is in $\mathcal{F}_{\hat{m}}$ and one has a non-negative quantity $A_{f,m}$ such that

$$\begin{aligned} & \|Y - \hat{f}\|_n^2 + \frac{\text{pen}_n(\hat{f}, \hat{m})}{n} \\ \leq & \inf_{m \in \mathcal{M}} \inf_{f \in \mathcal{F}_m^{\text{co}}} \left\{ \|Y - f\|_n^2 + \frac{\text{pen}_n(f,m)}{n} + A_{f,m} \right\}. \end{aligned}$$

Here $\text{pen}_n(f,m)$ and $A_{f,m}$ are permitted to depend on the data $\underline{X}$ and $\underline{Y}$ (and also depend on the evaluation inputs $\underline{X}'$ in the transductive case). We call this estimator $\hat{f}$ or its truncated counterpart $T\hat{f}$ a penalized least squared estimator with optimization accuracy $A_{f,m}$ with respect to $\{\mathcal{F}_m^{\text{co}}\}_{m \in \mathcal{M}}$.

Often, $\hat{f}$ and $\hat{m}$ are chosen by a two-step procedure. It is arranged for $\hat{f}_m = \hat{f}_{n,m}$ to approximately minimize $\|Y - f\|^2 + \frac{\text{pen}_n(f,m)}{n}$, for $\hat{m}$ to be the model index minimizing $\|Y - \hat{f}_m\|_n^2 + \frac{\text{pen}_n(\hat{f}_m,m)}{n}$ and for $\hat{f} = \hat{f}_{n,\hat{m}}$ to be the estimator obtained by plugging in the selected model. If $\hat{f}_m$ exactly minimizes the criterion among functions in $\mathcal{F}_m$, for $m \in \mathcal{M}$, then it is natural to set $\mathcal{F}_m^{\text{co}} = \mathcal{F}_m$ and $A_{f,m} = 0$. Nevertheless, we find that we can take advantage of larger comparison sets in some settings. For instance, using relaxed greedy fits in the setting of Chapter 3, we may have $\mathcal{F}_m^{\text{co}}$ equal to the whole $\mathcal{L}_{1,\mathcal{H}}$.

Corollary 5.1 below, similar to Theorem 2.4, gives a condition on the penalty such that an analogous risk conclusion holds for $\hat{f}_{\hat{m}}$. As for the cover it may be formed from a union of sets $\tilde{\mathcal{F}}_{\underline{X},\underline{X}',\tilde{m}}$, with $\tilde{m}$ in a subset $\tilde{\mathcal{M}}_{\underline{X},\underline{X}'}$ of the index set $\mathcal{M}$, with associated complexities $L_{\underline{X},\underline{X}'}(\tilde{f}, \tilde{m})$. In this setting the symmetry and complexity condition becomes the following.

Assumption (S′) The index set $\tilde{\mathcal{M}}_{\underline{X},\underline{X}'}$, the function sets $\tilde{\mathcal{F}}_{\underline{X},\underline{X}',\tilde{m}}$ and associated complexities $L_{\underline{X},\underline{X}'}(\tilde{f}, \tilde{m})$ are coordinate pair symmetric between $\underline{X}$ and $\underline{X}'$ and the complexities

satisfy the Kraft inequality

$$\sum_{\tilde{m}\in\tilde{\mathcal{M}}_{\underline{X},\underline{X}'}} \sum_{\tilde{f}\in\tilde{\mathcal{F}}_{\underline{X},\underline{X}',\tilde{m}}} e^{-L_{\underline{X},\underline{X}'}(\tilde{f},\tilde{m})} \leq 1.$$

Our general requirement on the penalty is that there exists countable sets $\tilde{\mathcal{F}}_{\underline{X},\underline{X}',\tilde{m}}$ of functions $\tilde{f}$ bounded by B' and associated complexities $L_{\underline{X},\underline{X}'}(\tilde{f},\tilde{m})$ satisfying Assumption (S′) and an adjustment Adjust_n, such that for every $m \in \mathcal{M}$ and $f \in \mathcal{F}_m$, the penalty has $[\text{pen}_n(f,m) + \text{Adjust}_n]$ at least

$$\inf_{\tilde{m}\in\tilde{\mathcal{M}}_{\underline{X},\underline{X}'}} \inf_{\tilde{f}\in\tilde{\mathcal{F}}_{\underline{X},\underline{X}',\tilde{m}}} \left\{\Delta_n(f,\tilde{f}) + \gamma L_{\underline{X},\underline{X}'}(\tilde{f},\tilde{m})\right\}, \tag{5.1}$$

where $\Delta_n(f,\tilde{f})$ is the distortion between f and $\tilde{f}$ defined at (2.27).

Corollary 5.1 *In the same setting as in Corollary 2.6, suppose the penalty $pen_n(f,m)$ not only depends on f, but also depends on an index m in an index set $\mathcal{M}$. Given positive constant δ_1, δ_2 and nonnegative δ_3, for $pen_n(f,m)$ with an adjustment $Adjust_n$ exceeding (5.1) or exceeding its expectation with respect to $\underline{X}'$, then a penalized least squares estimator (with optimization accuracy $A_{f,m}$) when truncated to the level B' satisfies the following risk bound*

$$\begin{aligned}
&\mathbb{E}\|T\hat{f} - f^*\|^2_{\underline{X}'} \\
&\leq (1+\delta) \inf_{m\in\mathcal{M}} \inf_{f\in\mathcal{F}_m^{CO}} \left\{\|f - f^*\|^2 + \mathbb{E}\left[\frac{pen_n(f,m)}{n} + A_{f,m}\right] + \frac{adjust}{n}\right\}.
\end{aligned}$$

where $(1+\delta)$, γ and $adjust$ are the same as in Theorem 2.4.

Proof: The proof is similar to that of Theorem 2.4. Denote $\tilde{\mathcal{M}}_{\underline{X},\underline{X}'}$, $\tilde{\mathcal{F}}_{\underline{X},\underline{X}',\tilde{m}}$ and $L_{\underline{X},\underline{X}'}(\tilde{f},\tilde{m})$ by $\tilde{\mathcal{M}}$, $\tilde{\mathcal{F}}_{\tilde{m}}$ and $L_n(f,m)$ for simplicity. Let $\text{pen}_n^+(f,m) = \text{pen}_n(f,m) + \text{Adjust}_n$. Rewrite

the condition (5.1) of the proper penalty as

$$\begin{aligned} & \mathbb{E}_{\underline{X}'} \sup_m \sup_{f \in \mathcal{F}_m} \left\{ \frac{1}{c} P_n'(g_{1,Tf}) - P_n(\rho_f) - \frac{\text{pen}_n^+(f,m)}{n} \right\} \\ \leq \; & \mathbb{E}_{\underline{X}'} \sup_{\tilde{m} \in \tilde{\mathcal{M}}} \sup_{\tilde{f} \in \tilde{\mathcal{F}}_m} \left\{ \frac{1}{\tilde{c}} P_n'(g_{1,\tilde{f}}) - P_n(\rho_{\tilde{f}}) - \frac{\gamma L_n(\tilde{f},\tilde{m})}{n} \right\}, \end{aligned} \tag{5.2}$$

where $\tilde{c} = (1+\delta_1)(1+\delta_2)$ and $c = \tilde{c}(1+\delta_3)$. Under the Assumption (S′), an analogous conclusion as Lemma 2.3 is achieved, namely, the expectation of the right side of (5.2) is non-positive. Consequently, the expectation of the left side is less than or equal to 0 as well, yielding a risk for the penalized least squares estimator $\hat{f}$ bounded by,

$$c\mathbb{E}\left(P_n(\rho_{\hat{f}}) + \frac{\text{pen}_n^+(\hat{f},\hat{m})}{n} \right),$$

which is an expected minimum both over the index set $\mathcal{M}$ and over the comparison set $\mathcal{F}_m^{\text{co}}$ and thus bounded by the minimum expectation. This provides our desired conclusion. ∎

5.2 $\mathcal{F}$ as the set of all finite linear combinations of functions

We now focus on the case that $\mathcal{F}$ may be the set of all finite linear combinations of functions from a dictionary $\mathcal{H}$, namely $\mathcal{F}_{\mathcal{H}}$. In all-subset regression, we use a penalty primarily determined by $\log \binom{M_{\mathcal{H}}}{m}$, the comparison class is $\mathcal{F}_m$, the set of all m-term linear combinations, and $A_{f,m} = 0$.

5.2.1 Performance of all-subset selection

The first theorem below is to demonstrate a risk bound with best tradeoff between approximation accuracy and the penalty function for the estimator $T\hat{f}_{\hat{m}}$ chosen by all-subset selection from a finite library.

Theorem 5.2 (Risk characterization for all-subset selection) *Assume the regression setting (B). Assume $\mathcal{H}$ is a finite dictionary with Cardinality $M = M_{\mathcal{H}}$ and $\mathcal{F}_m$ is the set of all m-term linear combinations of $\mathcal{H}$. Suppose $\hat{f}_m = f_{\hat{\beta}_m}$ is chosen to be the least squares estimators over $\mathcal{F}_{\mathcal{H},m}$. Let any positive δ_1 and δ_2 be given. Choose $\hat{m}$ among all $1 \le m \le \min\{n, M\}$ to minimize the penalized least squares*

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_m(X_i))^2 + \frac{pen_n(m)}{n}$$

where $pen_n(m)$ has the form $C\left(\log\binom{M}{m} + (m+1)\log n + \log\min\{n, M\}\right)$ with $C \ge \gamma$, where γ is the same as before.

Then for $n \ge 4e^2$, this selected estimator $\hat{f} = \hat{f}_{\hat{m}}$ when truncated to the level B' satisfies the following risk bound

$$\begin{aligned}&\mathbb{E}\|T\hat{f}_{\hat{m}} - f^*\|^2_{\underline{X}'}\\ &\le \ (1+\delta)\inf_m \inf_{f_m \in \mathcal{F}_{\mathcal{H},m}} \left\{\|f_m - f^*\|^2 + \frac{pen_n(m)}{n} + \frac{C_\delta}{n}\right\}\end{aligned}$$

where $1 + \delta = (1+\delta_1)(1+\delta_2)$ and $C_\delta \le \gamma' + \gamma + tail_1 + tail_2$ with $\gamma' = 32e^2B'^2$. Here $tail_1$, $tail_2$ and B' has the same properties as in Corollary 2.6 and its remark.

Remarks: This establishes for finite libraries the claim (1.5) in the introduction for Case (A) for all-subset regression. We point out that by allowing for a multiple of n inside the logarithm, we can ameliorate the effect of the otherwise large γ' and allow for arbitrary $n \ge 1$. These details are given in the proof.

A similar conclusion holds for infinite libraries with finite covers. For such infinite libraries our bounds are cleanest when we allow mild control (via penalties) on the size of the coefficients, which avoids numerical difficulties with correlated variables. Such control adjusts is obtained by the subset size penalty with a (possibly very small) multiple of $\|\beta\|_1$ or $\|\beta\|_2^2 = \sum_h \beta_h^2$, a familiar ridge regression modification. When that multiplier is appropriately small, the main part of the penalty remains based on subset size. Accordingly, we permit $\mathcal{H}$ to be an infinite library in Theorem 5.4 below using a small multiple of $\|\beta\|_1$. We omit infinity library details for all-subset selection conclusion here.

Proof: We prove the theorem in the following steps. Analogous steps are also used for the other results of this chapter.

1. Note that the estimator $\hat{f}_{\hat{m}}$ is a penalized least squares estimator;

2. Set $\tilde{\mathcal{M}}_{\underline{X},\underline{X}'} = \mathcal{M} = \{1, 2, \ldots, M\}$. Construct empirical covers $\tilde{\mathcal{F}}_{n,m} = \tilde{\mathcal{F}}_{\underline{X},\underline{X}',m}$ and associated complexities $\{L_n(\tilde{f}, m) : \tilde{f} \in \tilde{\mathcal{F}}_{n,m}\}$ to satisfy Assumption (S′);

3. Show that the penalty function $\text{pen}_n(f, m)$ with an adjustment satisfies condition (5.1).

Then the risk bound for the truncated estimator $T\hat{f}_{\hat{m}}$ follows from Corollary 5.1.

Step (1) is immediate for this case. The comparison class is taken to be $\mathcal{F}_m^{\text{CO}} = \mathcal{F}_m$. Our penalty is $\text{pen}_n(f, m) = \text{pen}_n(m)$ for f in $\mathcal{F}_m$. Using the definition, our estimator is a penalized least square estimator with $A_{f,m} = 0$.

Now we come to step (2). Given any set $\Lambda \subset \mathcal{H}$, we let $\mathcal{F}_\Lambda = \text{span}\{h : h \in \Lambda\}$ denote its linear span and $T\mathcal{F}_\Lambda$ denote the truncated version of this span. Using the results of Haussler [52] as in Lemma 8.7 in the appendix, we know that for any Λ with cardinality not more than m, the covering number $\mathcal{N}(t, T\mathcal{F}_\Lambda, \|.\|_{2n,1})$ is less than $e(\frac{4e^2B'}{t})^{m+1}$, where $\|.\|_{2n,1}$ is the empirical L_1 distance on $(\underline{X}, \underline{X}')$, and $t \leq B'$. Taking the union of

these covers over the $\binom{M}{m}$ choices of subsets, we bound the covering number of $T\mathcal{F}_m = \bigcup_{\Lambda:|\Lambda|=m} T\mathcal{F}_\Lambda$ as follows,

$$\mathcal{N}(t, T\mathcal{F}_{\mathcal{H},m}, \|.\|_{2n,1}) \leq e \left(\frac{4e^2 B'}{t}\right)^{m+1} \binom{M}{m}. \tag{5.3}$$

Then $\tilde{\mathcal{F}}_{n,m}$ is chosen to be a t-cover of $T\mathcal{F}_{\mathcal{H},m}$ with the minimum cardinality. In particular, for f in $\mathcal{F}_{\mathcal{H},m}$, there exists a function $\tilde{f}_m \in \tilde{\mathcal{F}}_{n,m}$, such that $\|\tilde{f}_m - Tf\|_{2n,1} \leq t$. We choose $t = \tau 4e^2 B'/n$, with $4e^2\tau/n \leq 1$, where as we shall see it may be advantageous to allow $\tau = \tau_m$ to depend on m; for simplicity, the claim of the theorem chooses $\tau = 1$. Then the log cardinality of $\tilde{\mathcal{F}}_{n,m}$ is not more than the logarithm of the right side of (5.3), that is,

$$1 + (m+1)\log\frac{n}{\tau_m} + \log\binom{M}{m}. \tag{5.4}$$

Now $\log M$ suffices for the description length of m in the set $\mathcal{M} = \{1, \ldots, M\}$. If one also imposes that models with not more than n terms be considered this part may be reduced to $\log\min\{n, M\}$ here and in what follows. Thus the complexities are set to be $L_n(\tilde{f}_m, m) = \log \text{Card}(\tilde{\mathcal{F}}_{n,m}) + \log M$ for which the Kraft inequality holds. Using the expression (5.4) to control the log cardinality of the covers, we have that

$$L_n(\tilde{f}_m, m) \leq \log\binom{M}{m} + (m+1)\log\frac{n}{\tau_m} + \log M + 1. \tag{5.5}$$

From the above construction, we recognize that both $\tilde{\mathcal{F}}_{n,m}$ and $L_n(\tilde{f}_m, m)$ depend on data only via the L_2 empirical norm on $(\underline{X}, \underline{X}')$, and so they are arranged to be coordinate pair symmetric.

For step (3), for any m and $f \in \mathcal{F}_m$, using Corollary 5.1, we are to show that the penalty is at least $\inf_{\tilde{f}\in\tilde{\mathcal{F}}_{n,m}} \left\{\Delta_n(f, \tilde{f}) + \gamma L_n(\tilde{f}, m)\right\}$ with our choice $\delta_3 = 0$. In the

proof of Lemma 2.5, we have the inequality that for any function f,

$$\|Y - f\|_n^2 \geq \|Y - Tf\|_n^2 - \frac{\text{Tail}_2(\underline{Y})}{n}$$

where $\text{Tail}_2(\underline{Y}) = 2\sum_{i=1}^n (|Y_i| - B')^2 1\{|Y_i| > B'\}$. Then we only need to show that for every $\underline{X}'$ that $\text{pen}_n(f, m)$ is not less than $\Delta_n(f, \tilde{f}_m) + \gamma L_n(\tilde{f}_m) + \text{Tail}_2(\underline{Y})$, where $\Delta_n(f, \tilde{f}_m)$ is given by

$$n\left(\|Y - \tilde{f}_m\|_n^2 - \|Y - Tf\|_n^2\right) + \frac{n}{c}\left(\|f^* - Tf\|_{\underline{X}'}^2 - \|f^* - \tilde{f}_m\|_{\underline{X}'}^2\right) \tag{5.6}$$

with $c = (1 + \delta_1)(1 + \delta_2)$ and $\tilde{f}_m$ is the function corresponding to f in the cover $\tilde{\mathcal{F}}_{n,m}$ as mentioned before.

Term by term, $[Y_i - \tilde{f}_m(X_i)]^2$ is at most $[Y_i - Tf(X_i)]^2 + 4B'|Tf(X_i) - \tilde{f}_m(X_i)| + \text{Tail}_1(Y_i)$, where $\text{Tail}_1(Y_i) = 4B'(|Y_i| - B')1\{|Y_i| > B'\}$. Then the first two terms in $\Delta_n(f, \tilde{f}_m)$ are upper bounded by $4B'\sum_{i=1}^n |Tf(X_i) - \tilde{f}_m(X_i)| + \text{Tail}_1$. Likewise, the last two terms are upper bounded by $\frac{2(B+B')}{c}\sum_{i=1}^n |Tf(X_i') - \tilde{f}_m(X_i')|$. Thus we obtain the inequality

$$\Delta_n(f, \tilde{f}_m) \leq 8nB'\|Tf - \tilde{f}_m\|_{2n,1} + \text{Tail}_1 \leq 32e^2 B'^2 \tau + \text{Tail}_1, \tag{5.7}$$

by substituting $t = \tau_m 4e^2 B'/n$. Now setting Adjust_n to be $\text{Tail}_1 + \text{Tail}_2 + \gamma$ and combining the upper bound on $\Delta_n(f, \tilde{f}_m)$ from (5.7) and the upper bound on the complexity from (5.5), we allow $\text{pen}_n(f, m)$ not less than

$$C\left[\log\binom{M}{m} + (m+1)\log\frac{n}{\tau_m} + \log M\right] + \gamma'\tau_m, \tag{5.8}$$

where $C \geq \gamma$ and $\gamma' = 32e^2B'^2$. Applying Corollary 5.1 with this penalty yields the risk bound for the penalized least squares estimator with all-subset selection

$$\begin{aligned} & \mathbb{E}\|T\hat{f}_{\hat{m}} - f^*\|^2 \\ \leq \; & c \inf_m \inf_{f \in \mathcal{F}_m} \left\{ \|f - f^*\| + \mathbb{E}\frac{\text{pen}_n(f,m)}{n} + \frac{adjust}{n} \right\}. \end{aligned} \tag{5.9}$$

where $adjust = \mathbb{E}[\text{Adjust}_n] = tail_1 + tail_2 + \gamma$.

The choice $\tau_m = 1$, valid provided $n \geq 4e^2$, corresponds to the claim of the theorem (with the adjustment that $\gamma'\tau_m$, now equal to the constant γ', is absorbed into the constant in the risk bound rather than kept in the penalty). Alternatively, we may choose $\tau_m = \tau_m^* = \min\{(C/\gamma')(m+1), n/4e^2\}$ to optimize expression (5.8), valid for all $n \geq 1$, for which the $\gamma'\tau_m$ term becomes not more than $C(m+1)$. Refined in this way we find acceptability of

$$\text{pen}_n(f,m) = C\left[\log\binom{M}{m} + (m+1)\log\frac{ne}{\tau_m^*} + \log M\right]$$

for the risk conclusion (5.9) to hold. This completes the proof of the theorem. ∎

The performance of the all-subset selection estimator is at least as good as that of the ℓ_1-penalized least squares estimator. Indeed, as a consequence of the argument yielding the third conclusion in Lemma 8.4 in the appendix, for any $m = m_1 + m_0$ and f^*, there is an m-term function f_m such that

$$\|f_m - f^*\|^2 \leq \inf_{f_\beta \in \mathcal{F}} \left\{ \|f_\beta - f^*\|^2 + \varepsilon_{m_0}^2 \|\beta\|_1/(m - m_0) \right\}, \tag{5.10}$$

where ε_{m_0} is the radius of the $L_2(P)$-cover of the library with cardinality not more than m_0. Thus using $\gamma(m+1)\log(Mn)$ to upper bound the penalty function $\text{pen}_n(m)$, and minimizing with respect to m first, the main term of the resolvability is $\inf_{f_\beta \in \mathcal{F}} \{\|f_\beta - f^*\|^2 + \lambda_n\|\beta\|_1\}$, where $\lambda_n = \varepsilon_{m_0}\sqrt{\frac{\gamma \log(Mn)}{n}}$, which has the same form as in the ℓ_1-penalty case within a log

factor. A difference here is that for ε_{m_0}, we use a $L_2(P)$-cover of $\mathcal{H}$ whereas in Chapter 4, the empirical L_2 cover is used.

5.2.2 Performance of relaxed greedy algorithms including forward stepwise selection

The argument in the proof of all-subset selection is readily applicable to other subset selection algorithms of interest to us, including forward stepwise selection, other relaxed greedy algorithms, and ℓ_1 penalized greedy pursuit. These allows a larger comparison class, all of $\mathcal{F}_{\mathcal{H}}$, and introduce an approximate computation term in the bounds equal to an expectation of $A_{f,m} \leq 4V_n^2(f)/m$.

If the library $\mathcal{H}$ is finite, then forward stepwise and other relaxed greedy algorithms may be used with penalty only on the number of terms and no need for control on the size of coefficients as shown in Theorem 5.3. The infinite library case is considered subsequently in Theorem 5.4.

We assume the greedy algorithm to run for a number of steps equal to $m_{n,\text{comp}}$. For forward stepwise, there is no reason to set $m_{n,\text{comp}}$ greater than $\min\{n, M_{\mathcal{H}}\}$. For other relaxed greedy algorithms, there is a continuous approximation improvement for larger m. However, $m_{n,\text{comp}} = n$ is large enough to let the data reveal the optimal $\hat{m}$ as we can see in the following theorem. The choice that optimizes the resolvability is of a smaller order than n.

Theorem 5.3 (Risk characterization for forward stepwise regression) *With the same setting as Theorem 5.2, suppose, for $m = 1, 2, \ldots, m_{n,comp}$, that $\hat{f}_m = f_{\hat{\beta}_m}$ is a sequence of m-term estimators obtained by a relaxed greedy algorithm (which includes forward stepwise regression). Let any positive δ_1 and δ_2 be given. Choose $\hat{m}$ among all $m \leq m_{n,comp}$ to*

minimize the penalized least squares

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_m(X_i))^2 + \frac{pen_n^0(m)}{n}$$

where $pen_n^0(m) = C\left(\log\binom{M}{\tilde{m}} + (\tilde{m}+1)\log n + \log M\right)$ *with* $\tilde{m} = \min\{m, M\}$ *and* $C \geq \gamma$.

Then this selected estimator $\hat{f}_{\hat{m}}$ *when truncated to the level* B' *satisfies the following risk bound,*

$$\begin{aligned}&\mathbb{E}\|T\hat{f}_{\hat{m}} - f^*\|^2_{\underline{X}'} \\ &\leq \ (1+\delta)\inf_m \inf_{f\in\mathcal{F}_m}\left\{\|f - f^*\|^2 + \mathbb{E}\frac{pen_n^0(m)}{n} + \frac{4V^2(f)}{m} + \frac{C_\delta}{n}\right\},\end{aligned}$$

where the infimum is over $m \leq m_{n,comp}$. *Here* $1+\delta$ *and* C_δ *are the same as in Theorem 5.2 and* $V(f) = V_{\|\cdot\|}(f)$ *is the variation of* f *with associated* a_h *equal to the* $L_2(P)$ *norm.*

Remarks: Thus we achieve the best tradeoff between the approximation error and the penalty as in the all-subset selection case except here we have an additional $4V^2(f)/m$ cost due to the forward stepwise procedure.

For large dictionaries with $M \gg n$, the main term of the penalty comes from the $\log\binom{M}{m}$. Using the bound $m\log M$, the value $m_f = 2V(f)\sqrt{n/(C\log(Mn))}$ optimizes the bound over m for each f. Then the risk bound becomes

$$(1+\delta)\inf_{f\in\mathcal{F}_m}\left\{\|f - f^*\|^2 + 4\sqrt{\frac{C\log(Mn)}{n}}V(f) + \frac{\log(Mn)}{n} + \frac{C_\delta}{n}\right\}.$$

We note that the order of this bound is in agreement with what was achieved in the first result in Chapter 4, when M is large compared to $\sqrt{n}$.

One may attempt similar refinement as achieved there for M or order $\sqrt{n}$ or smaller

(with the log factor removed), now optimizing using $m \log eM/m$ as an upper bound on $\log \binom{M}{m}$ for $m \leq M$. However, the other term of size $m \log n$ (or $m \log n/\tau_m$) in the present bound preserves the logarithm factor here. Other techniques such as chaining (as in [10]) might be able to avoid that log factor, but would need to be extended to allow for empirical covers and for greedy approximate least squares fits. As we are primarily interested in the large M case, we shall not concern ourselves with that in this paper.

Proof: We use the same argument as in Theorem 5.2. First, for f in $\mathcal{F}_{\mathcal{H}}$ and $m \leq m_{n,\text{comp}}$, we set $\text{pen}_n(f,m)$ to be $\text{pen}_n(m)$. Using the property of the relaxed greedy algorithm, our estimator is an approximate penalized least square estimator with $A_{f,m} \leq \frac{4V_n^2(f)}{m+1}$ and $\mathcal{F}_m^{\text{co}} = \mathcal{F}_{\mathcal{H}}$. Here our $\mathcal{F}_m$ is the set of m-term linear combinations but with repeated terms allowed.

Second, we set $\tilde{\mathcal{M}}_{\underline{X},\underline{X}'} = \{1, 2, \ldots, M\}$. For functions in $\mathcal{F}_m$, we construct our representors in $\tilde{\mathcal{F}}_{n,\tilde{m}}$ as in the step (2) of Theorem 5.2, where $\tilde{m} = \min\{m, M\}$. Then the complexities $L_n(\tilde{f}_m, \tilde{m})$ have the same form (5.5) with m replaced by $\tilde{m}$.

Then we follow the proof of Theorem 5.2 to get the conclusion. The only additional term appears in the risk bound is $\mathbb{E}A_{f,m} \leq \frac{4V^2(f)}{m}$ because $\mathbb{E}V_n^2(f) \leq V^2(f)$. ∎

5.2.3 Performance of ℓ_1 *penalized greedy pursuit (LPGP)*

Here we given the corresponding conclusion for ℓ_1 penalized greedy pursuit (LPGP) with a subset-size stopping criterion. In the theorem below, we denote $\|\beta\|_1 = \|\beta\|_{1,1} = \sum_h |\beta_h|$ as before.

Theorem 5.4 (Risk Characterization for LPGP) *Assume the regression setting (B). Assume $\mathcal{H}$ is a dictionary with empirical L_1 ε_1-covers $\tilde{\mathcal{H}}_1$ with cardinality $M_1 = M_{\varepsilon_1,\mathcal{H}}$. We take $\hat{f}_m$ to be the $m-$term fit from an LPGP algorithm (variant 1) as in Chapter 3 with library $\mathcal{H}$ and coefficient λ_0. Let any positive δ_1, δ_2 and ε_1 be given. Choose $\hat{m}$ among*

all $m \leq m_{n,comp}$ *to minimize the penalized least squares* $\|Y - \hat{f}_m\|_n^2 + \frac{pen_n(\hat{f}_m,m)}{n}$ *where* $pen_n(\hat{f}_m, m)$ *has the form* $n\lambda_0 v_m + \text{pen}_n^0(m)$ *with* $\lambda_0 \geq 8\varepsilon_1 B'$ *and*

$$pen_n^0(m) = C\left(\log\binom{M_1}{\tilde{m}} + (\tilde{m}+1)\log n + \log M_1\right)$$

with $\tilde{m} = \min\{m, M_1\}$ *and* $C \geq \gamma$. *Here* v_m *as described in the LPGP algorithm in Chapter 3 is* $\sum_{j=1}^m |\hat{\beta}_{j,m}|$, *where the* $\hat{\beta}_{j,m}$ *are coefficients of* $\hat{f}_m$.

Then for $n \geq 4e^2$, *the selected estimator* $\hat{f}_{\hat{m}}$ *when truncated to the level* B' *satisfies the following risk bound*

$$\begin{aligned}
&\mathbb{E}\|T\hat{f}_{\hat{m}} - f^*\|_{\underline{X}'}^2 \\
&\leq \ (1+\delta)\inf_m \inf_{f\in\mathcal{F}_{\mathcal{H}}}\left\{\|f - f^*\|^2 + \lambda_0 V_a(f) + \frac{4V_a^2(f)}{m} + \frac{\text{pen}_n^0(m)}{n} + \frac{C_\delta}{n}\right\}
\end{aligned}$$

where $1+\delta$ *and* C_δ *are the same as in Theorem 5.2 and* $V_a(f)$ *is the variation of* f *with associated weights* $a_h = \max\{1, \|h\|\}$. *Here* $\|h\|$ *is the* $L_2(P)$ *norm of* h.

Remarks: For any constant $\eta > 0$, a similar conclusion holds with using weights $a_h = \max\{\eta, \|h\|\}$. Now the risk bound becomes

$$(1+\delta)\inf_m \inf_f \left\{\|f - f^*\|^2 + \frac{\lambda_0}{\eta}V_a(f) + \frac{4V_a^2(f)}{m} + \frac{\text{pen}_n^0(m)}{n} + \frac{C_\delta}{n}\right\}.$$

The statement in the theorem is a special case with $\eta = 1$. We show the general statement in the proof.

Our primary interpretation of this theorem is the extension of greedy subset selection risk analysis to infinite libraries with a mild control on $\|\beta\|_1$. We accomplish this by using a small ε_1 and λ_0, for instance, of order $1/n$. Because λ_0 is negligible, there is room to use a small η to make the term λ_0/η negligible, which provides conclusion as an extension

of Theorem 5.3 to infinity libraries with the variation $V_a(f)$ close to $V_{\|\cdot\|}(f)$, where the associated weights equal the $L_2(P)$ norm.

One may also think of this theorem as producing a stopping rule for the LPGP algorithm. By the same argument as in the remark after Theorem 5.3, the risk bound is equal to

$$(1+\delta)\inf_f \left\{ \|f - f^*\|^2 + \left(\frac{\lambda_0}{\eta} + 4\sqrt{\frac{C\log(nM_1)}{n}} \right) V_a(f) + \frac{C_\delta}{n} \right\}.$$

We note that if λ_0 is chosen to have the order matching $\sqrt{\log M_1/n}$ or smaller, then the order of this bound is in agreement with what was achieved in Chapter 4 without the further refinement associated with empirical L_2 covers of $\mathcal{H}$. One advantage here is that even in the traditional non-transductive setting, the risk bound uses the $L_2(P)$ norm to form the weights used in the variation.

Here we assumed that the LPGP algorithm satisfies the iterative optimization requirement (3.1) with $\varepsilon_m^{\text{comp}} = 0$. If instead, we approximately achieve the optimization there with $\varepsilon_m^{\text{comp}} \le 4\delta_0/(m+1)^2$, then our risk bound in this theorem holds with $V_a^2(f) + \delta_0$ in place of $V_a^2(f)$. Actually we could use $V_a^2(f) - \|f\|^2 + \delta_0$.

As we have stated in the Remark of Lemma 3.1 in Chapter 3, we only need the infimum in the risk bound for f in $\mathcal{F}_\mathcal{H}$. So for the proof, we prove the conclusion for such $f = f_\beta = \sum_h \beta_h h$ in $\mathcal{F}_\mathcal{H}$, with the minimal $\|\beta\|_1$ among such representations of f providing its variation. Here, there are two different weights associated with $\|\beta\|_1$. We denote $\|\beta\|_{1,\eta} = \sum_h |\beta|\eta$ with positive η and $\|\beta\|_{1,a^{\text{emp}}}$ with $a_h^{\text{emp}} = \max\{\eta, \|h\|_{\underline{X}}\}$.

Proof: First, we know our estimator is an approximate penalized least square estimator with $A_{f,m} \le \frac{4\|\beta\|_{1,a^{\text{emp}}}^2}{m+1} - \frac{\lambda_0}{\eta}\|\beta\|_{1,\eta} + \frac{\lambda_0}{\eta}\|\beta\|_{1,a^{\text{emp}}}$ and $\mathcal{F}_m^{\text{co}} = \mathcal{F}_\mathcal{H}$ by using Lemma 3.1 and its Remark 3. Now with $\mathbb{E}\max\{\eta, \|h\|_{\underline{X}}\} \le \max\{\eta, \|h\|\}$, the expectation of $A_{f,m}$ is less than or equal to the same expression only with empirical weights a^{emp} replaced by $\max\{\eta, \|h\|\}$. Here our $\mathcal{F}_m$ is the m-term linear combinations with repeated terms allowed

as in Theorem 5.3.

Second, with finite empirical ε_1-cover $\tilde{\mathcal{H}}_1$ defined, we can use its cardinality to use in place of M as in Theorem 5.2. Now for each $f \in \mathcal{F}_m$, we form a corresponding $f_{\{\tilde{h}\}}$ by replacing each h_k with the closest $\tilde{h}_k$ in $\tilde{\mathcal{H}}_1$ and keeping the same coefficients. Then we know that $\|f - f_{\{\tilde{h}\}}\|_{2n,1} \leq \sum_{j=1}^{m} \beta_j \|h_j - \tilde{h}_j\|_{2n,1} \leq \varepsilon_1 \|\beta\|_{1,\eta}/\eta$. We denote the collection of such $f_{\{\tilde{h}\}}$ by $\mathcal{F}_{\tilde{\mathcal{H}}_1,\tilde{m}}$ with $\tilde{m} = \min\{m, M_1\}$. Our representor set $\tilde{\mathcal{F}}_{n,\tilde{m}}$ is chosen to be an empirical L_1 t-cover of $T\mathcal{F}_{\tilde{\mathcal{H}}_1,\tilde{m}}$ with the minimum cardinality. The same argument in Theorem 5.2 shows that the log cardinality of $\tilde{\mathcal{F}}_{n,\tilde{m}}$ is not more than $1 + (\tilde{m}+1)\log\frac{n}{\tau_m} + \log\binom{M_1}{\tilde{m}}$ with t set to be $\tau_m 4e^2 B'/n$. For each $f \in \mathcal{F}_m$, from the triangle inequality, there exists an $\tilde{f}_m$ closet to the corresponding $f_{\{\tilde{h}\}}$, such that $\|Tf - \tilde{f}_m\|_{2n,1} \leq \|Tf - Tf_{\{\tilde{h}\}}\|_{2n,1} + \|Tf_{\{\tilde{h}\}} - \tilde{f}_m\|_{2n,1} \leq t + \varepsilon_1\|\beta\|_{1,\eta}/\eta$.

Third, we are to show the penalty is at least $\inf_{\tilde{f}\in\tilde{\mathcal{F}}_{n,\tilde{m}}} \left\{\Delta_n(f,\tilde{f}) + \gamma L_n(\tilde{f},\tilde{m})\right\}$ for any m and $f \in \mathcal{F}_m$, with $\delta_3 = 0$. Using the same analysis as in Theorem 5.2 step (3), we note that $\Delta_n(f,\tilde{f}_m)$ is not more than $8nB'(t + \varepsilon_1\|\beta\|_{1,\eta}/\eta) + \text{Tail}_1$, which implies that with $\text{Adjust}_n = \text{Tail}_1 + \text{Tail}_2 + \gamma$, the penalty $\text{pen}_n(f,m)$ is allowed not less than

$$C\left[\log\binom{M_1}{\tilde{m}} + (\tilde{m}+1)\log\frac{n}{\tau_m} + \log M_1\right] + \frac{\lambda_0 n}{\eta}\|\beta\|_{1,\eta} + \gamma'\tau_m \tag{5.11}$$

by substituting $t = \tau_m 4e^2 B'/n$, where $C \geq \gamma$, $\lambda_0 \geq 8B'\varepsilon_1$ and $\gamma' = 32e^2 B'^2$. Applying Corollary 5.1 with this penalty yields the risk bound for the LPGP estimator. The choice of $\tau_m = 1$ corresponds to the claim of the theorem and the analysis of the optimal choice of τ_m is the same as in the final part of Theorem 5.2. ∎

5.3 Mixed penalty as a combination of both ℓ_0 and ℓ_1 norms of the coefficients

We end this chapter by combining results in Chapter 4 and 5 together to obtain an improvement. For simplicity, here we only discuss the case where functions in $\mathcal{H}$ have a uniform upper bound $b \leq 1$. Similar result as Theorem 5.5 below holds even if there is no uniform upper bound for $\mathcal{H}$ with more detailed analysis. First assume $\mathcal{H}$ is finite with cardinality not more than M. Let $\tilde{\mathcal{H}}_2$ be a finite empirical L_2 ϵ_2-cover of $\mathcal{H}$ with cardinality not more than m_0. Suppose $\hat{f}_1 = f_{\hat{\beta}_1}$ is the ℓ_1-penalized least squares estimator with $\text{Pen}_n(\beta) = n\lambda\|\beta\|_1$, where $\lambda = \lambda_1 \geq \lambda_1^* = 2\epsilon_2\sqrt{\frac{2\gamma\log(2M)}{n}}$. Also let $\hat{f}_{\hat{m}}$ be the estimator generated from a subset selection criterion such as forward stepwise regression, with proper penalty $\text{Pen}_n(\beta, m) = \text{pen}_n^0(m)$ defined in Theorem 5.3. The theorem holds with $A_{\beta,m} = \|Y - \hat{f}_m\|_n^2 - \|Y - f_\beta\|_n^2$.

Define our combined estimator $\hat{f}^{new}$ to be the one selected between $\hat{f}_1$ and $\hat{f}_{\hat{m}}$ to achieve the smaller penalized squared error. This $\hat{f}^{new}$ is an improvement compared to both $\hat{f}_1$ and $\hat{f}_{\hat{m}}$. Indeed, we have the following statement.

Theorem 5.5 *If $\mathcal{H}$ is finite, given positive constants δ_1 and δ_2, for $n \geq 4e^2$, the estimator $\hat{f}^{new}$ when truncated to the level B' satisfies the following risk bound*

$$\mathbb{E}\|T\hat{f}^{new} - f^*\|^2 \leq$$
$$c\inf_\beta \left\{\|f_\beta - f^*\|^2 + \min\left\{\lambda_1\|\beta\|_1, \min_m\left\{\frac{pen_n^0(m)}{n} + A_{\beta,m}\right\}\right\} + \frac{C_\delta^{new}}{n}\right\},$$

where $c = 1+(1+\delta_1)(1+\delta_2)$ and $C_\delta^{new} \leq tail_1+tail_2+\gamma\log 2+\max\{\gamma+\gamma', 2\gamma m_0\log(2M)\}$ with $\gamma' = 32e^2B'^2$.

Remark: The theorem shows the performance of the combined estimator is at least as

good as the better of the two procedures as revealed by which provides the smaller resolvability for the particular target f^*.

The same idea may also be used to combine the all-subset selection estimator and the ℓ_1-penalized least squares estimator. However, we believe that the all-subset selection estimator will win the minimization all the time.

Analogous conclusion holds for infinite libraries. Indeed, suppose $\mathcal{H}$ has an empirical L_1 ϵ_1-cover with cardinality M_1. Define $\hat{f}_1 = f_{\hat{\beta}_1}$ to be the ℓ_1-penalized least squares estimator with $\text{Pen}_n(\beta) = n\lambda\|\beta\|_1$, where $\lambda = \lambda_1 + 2\lambda_0$ with $\lambda_1 \geq \lambda_1^*$ and $\lambda_0 \geq \lambda_0^* = 8B'\epsilon_1$. Also define $\hat{f}_{\hat{m}}$ to be the estimator generated from a subset selection criterion with $\text{Pen}_n(\beta, m) = n\lambda_0\|\beta\|_1 + \text{pen}_n^0(m)$ with $\lambda_0 \geq \lambda_0^*$ and $\text{pen}_n^0(m)$ as in Theorem 5.4. Then a similar risk bound for $\hat{f}^{new}$ holds with the expression in the infimum over β in the theorem statement replaced by

$$\|f_\beta - f^*\|^2 + \min\left\{(\lambda_1 + \lambda_0)\|\beta\|_1, \min_m\left\{\frac{\text{pen}_n^0(m)}{n} + A_{\beta,m}\right\}\right\} + \lambda_0\|\beta\|_1 + \frac{C_\delta^{new}}{n},$$

where $A_{\beta,m} = \|Y - \hat{f}_m\|_n^2 + \lambda_0 v_m - \|Y - f_\beta\|_n^2 - \lambda_0\|\beta\|_1$. We show that the more general statement holds in the proof.

Proof: Set

$$\text{Pen}_n(\beta, m, k) = \begin{cases} n\lambda_0\|\beta\|_1 + \text{pen}_n^0(m) & \text{for } k = 0; \\ n\lambda\|\beta\|_1 & \text{for } k = 1. \end{cases} \tag{5.12}$$

The definition of $\hat{f}^{new}$ implies

$$\begin{aligned} &\|Y - \hat{f}^{new}\|_n^2 + \frac{Pen_n(\hat{\beta}, \hat{m}, \hat{k})}{n} \\ &\leq \min\left\{\|Y - \hat{f}_{\hat{m}}\|_n^2 + \lambda_0\|\hat{\beta}_0\|_1 + \frac{\text{pen}_n^0(\hat{m})}{n}, \|Y - \hat{f}_1\|_n^2 + \lambda\|\hat{\beta}_1\|_1\right\}. \end{aligned}$$

Here $\hat{k}$ is equal to 0 if the first term in the minimum is smaller and equal to 1 otherwise

and $\hat{\beta} = \hat{\beta}_{\hat{k}}$. The first term inside the minimum is not more than

$$\inf_{\beta} \min_{m} \left\{ \|Y - f_\beta\|_n^2 + \lambda_0 \|\beta\| + \frac{\text{pen}_n^0(m)}{n} + A_{\beta,m} \right\}$$

and the second term is equal to $\inf_\beta\{\|Y - f_\beta\|_n^2 + \lambda\|\beta\|_1\}$. Using $\min\{\inf_\beta A, \inf_\beta B\} = \inf_\beta\{\min\{A, B\}\}$, the right side of the above inequality is equal to

$$\inf_{\beta} \min_{k} \min_{m} \left\{ \|Y - f_\beta\|_n^2 + \frac{\text{Pen}_n(\beta, m, k)}{n} + A_{\beta,m,k} \right\},$$

where $A_{\beta,m,k} = A_{\beta,m}$ if $k = 0$ and zero if $k = 1$. Hence $\hat{f}^{new}$ is a penalized least squares estimator. Then since both $\text{Pen}_n(\beta, m, 0)$ and $\text{Pen}_n(\beta, m, 1)$ are proper penalties, our penalty $\text{Pen}_n(\beta, m, k)$ also satisfies the properness condition with $\log 2$ added to the complexity for the description of the choice of k, so that Theorem 2.4 yields the risk bound for $\hat{f}^{new}$. The additional description length appears in the term $\gamma \log 2$ in C_δ^{new}. ∎

Care is needed to interpret the resolvability of $\hat{f}^{new}$. If we insert $4\|\beta\|_1^2/m$ as the upper bound of $\mathbb{E}A_{\beta,m}$ into the resolvability, as we have stated in the remark of Theorem 5.3 and 5.4, the resolvability would have a contribution equal to $4\sqrt{C \log(nM_1)/n}$. This quatity tends to zero with a slower rate than λ_1^*, which has the rate $(\log M_1/n)^r$ with $r > 1/2$ as presented in Corollary 4.3. Therefore, when the empirical $A_{\beta,m}$ in each step matches the order $1/m$ bound, the ℓ_1-penalized least squares estimator achieves a smaller risk. It seems at first glance that there is no need for this $\hat{f}^{new}$ since the ℓ_1-penalized least squares estimator would appear superior. However, the accuracy of the fit of selected subsets can be much better than the bound. In other words, for some targets f^*, forward stepwise selection may outperform the order $1/m$ bound on $\mathbb{E}A_{\beta,m}$ for relevant f_β. When $A_{\beta,m}$ is smaller than order $1/m$, the criterion selects a smaller $\hat{m}$ and may produces a smaller risk than that given for the ℓ_1 penalized estimator. Armed with Theorem 5.5, we suggest to let the data decide the right choice and therefore achieve an improvement compared to both.

Chapter 6

Trade-off between the Approximation Error and the Complexity in the Resolvability

In this chapter, we discuss the trade-off between the approximation error and the complexity as expressed in the resolvability and its relationship to interpolation spaces between two classes of functions.

We have developed the resolvability risk bounds for ℓ_1-penalized least squares, all-subset selection, forward step-wise regression and the estimator formulated from the LPGP algorithm. We also constructed a combined estimator $\hat{f}^{new}$ to be an improvement compared to both the ℓ_1-penalization and the subset selection. In all these cases, the index of resolvability consists of two parts, the approximation error between the target f^* and the candidate f and the penalty $\text{pen}_n(f)/n$. It is natural to explore the trade-off between these two terms. This trade-off depends on the behavior of the unknown target function f^* and the accuracy with which it is approximated.

6.1 ℓ_1 Penalty case

For ℓ_1-penalization, as we mentioned in the remark of Corollary 4.3, we consider the squared approximation error $App(f^*, v) = \inf_{f_\beta : \|\beta\|_1 = v} \{\|f_\beta - f^*\|^2\}$. If 0 is contained in the library, $App(f^*, v)$ is a decreasing function of the ℓ_1 norm v. If the linear combination of the library is dense in the $L_2(P)$ space, the approximation error tends to 0 as v gets large. Hence $R^1(f^*, \lambda) = \inf_v \{App(f^*, v) + \lambda v\}$ goes to 0 as λ gets small. The index of resolvability is $R^1(f^*, \lambda_n)$, where λ_n is specified in Chapter 4. Thus, the resolvability rate is determined solely by the approximation rate and by how small is the permitted multiplier λ_n. One important case is when the target f^* has a finite variation, namely, $f^* \in \mathcal{L}_{1,\mathcal{H}}$. Then with $f = f^*$, the approximation error is zero and $R^1(f^*, \lambda) \leq \lambda V(f^*)$ goes to 0 with a linear rate. In general, if we only know f^* is in $L_2(P)$ space, the convergence rate can be arbitrarily slow.

We consider the function classes interpolating between $\mathcal{L}_{1,\mathcal{H}}$ and $L_2(P)$ indexed by $1 \leq p \leq 2$, denoted by $\mathcal{B}_{1,p}^{res}$ consisting of all functions f^* for which there is a constant c such that $R^1(f^*, \lambda) \leq c\lambda^{2-p}$ for all $\lambda > 0$. The infimum of such constants, denoted $C_{1,p}(f^*)$ is a measure of regularity of f^* in $\mathcal{B}_{1,p}^{res}$. As shown in Lemma 8.8, it is equivalent to a norm in a traditional interpolation space $\mathcal{B}_p$. The space $\mathcal{B}_p$, which is equivalent to $\mathcal{B}_{1,p}^{res}$, is developed in [11] where Fourier spectral norm conditions on a function f^* are shown to ensure membership in such interpolation spaces. Moreover, the space $\mathcal{B}_{1,p}^{res} = \mathcal{B}_p$ matches the weak space $w\mathcal{L}_p$ in the case where the library $\mathcal{H}$ is an orthonormal system. The interpolation space is a natural extension to non-orthonormal systems.

When $p=1$, we see $\mathcal{B}_{1,1}^{res}$ includes $\mathcal{L}_{1,\mathcal{H}}$. If $f^* \in \mathcal{B}_{1,p}^{res}$, the resolvability is of order λ_n^{2-p}. Results in Chapter 4 shows that λ_n is of order $\varepsilon_{m_0}\sqrt{\gamma(\log M)/n}$ for finite libraries, which provides a rate of $\varepsilon_{m_0}^{2-p}(\frac{\gamma \log M}{n})^{1-p/2}$ for the resolvability, where ε_{m_0} is the radius of the empirical L_2 cover of $\mathcal{H}$ with cardinality not more than m_0.

If $f^* \in \mathcal{B}_{1,p}^{res}$ and $\mathcal{H}$ has finite metric dimensions d_1 and d_2 w.r.t. the empirical L_1 and L_2 norms, respectively, the resolvability $R^1(f^*, \lambda_n)$ is of order $\left(\gamma \frac{d_1}{n} \log \frac{n}{d_1}\right)^{\xi}$, where $\xi = (1-p/2)(d_2+2)/(d_2+1)$, by using the first part of Corollary 4.3. As we mentioned in the second part of Corollary 4.3, a penalty with the form $\lambda_n \|\beta\|_1^{d_2/(d_2+1)}$ produces an estimator with smaller resolvability, namely, $R_r^1(f^*, \lambda_n)$, where $r = d_2/(d_2+1)$ and $R_r^1(f^*, \lambda) = \inf_v \{App(f^*, v) + \lambda v^r\}$ and λ_n is specified there. One may associate with this improved resolvability the class $\mathcal{B}_{1,r,p}^{res}$ of functions f^* for which $R_r^1(f^*, \lambda) \leq c\lambda^{\theta}$ for all positive λ with a constant c depending on f^*, where $\theta = (2-p)/(rp-p+2-r)$. Lemma 8.9 in the appendix demonstrates that $\mathcal{B}_{1,r,p}^{res}$ is the same space as $\mathcal{B}_{1,p}^{res}$. Consequently, if $f^* \in \mathcal{B}_{1,p}^{res}$, the resolvability $R_r^1(f^*, \lambda_n)$ yields a risk of order

$$\left(\gamma \frac{d_1}{n} \log \frac{n}{d_1}\right)^{\frac{(2-p)(d_2+2)}{2(d_2+2-p)}},$$

which is indeed smaller than $R^1(f^*, \lambda_n)$.

6.2 Subset selection case

Similarly, for all-subset selection, we define the squared approximation error $App^0(f^*, m) = \inf_{f_m \in \mathcal{F}_{\mathcal{H},m}} \{\|f_m - f^*\|^2\}$, which is a function of the ℓ_0 norm m. If the linear combination of the library is dense in the $L_2(P)$ space, $App^0(f^*, m)$ tends to 0 as m gets large. Hence $R^0(f^*, t) = \inf_m \{App^0(f^*, m) + tm\}$ goes to 0 as t gets small. From Theorem 5.2 in Chapter 5, the index of resolvability here is $R^0(f^*, t_n)$, where t_n is of order $\frac{\gamma \log(Mn)}{n}$, which is close to the square of λ_n in the previous case. For different targets f^*, the resolvability describes the performance of the all-subset selection estimator. $\mathcal{L}_{0,\mathcal{H}}$ is defined to be the set of linear combinations with finite number of terms. When the target f^* is in $\mathcal{L}_{0,\mathcal{H}}$ with m_{f^*} terms, the approximation error is zero and $R^0(f^*, t) \leq m_{f^*} t$ tends to 0

with a linear rate; whereas, for general $f^* \in L_2(P)$, as before, the convergence rate can be arbitrarily slow. We likewise consider the function classes between $\mathcal{L}_{0,\mathcal{H}}$ and $L_2(P)$ indexed by $0 \leq p \leq !2$, denoted by $\mathcal{B}^{res}_{0,p}$, consisting of all functions f^* for which there is a constant c such that $R^0(f^*, t) \leq ct^{1-p/2}$ for all $t > 0$. We likewise define $C_{0,p}(f^*)$ as the infimum of such constants c. When $p = 0$, the space $\mathcal{B}^{res}_{0,0}$ is indeed $\mathcal{L}_{0,\mathcal{H}}$. Then if $f^* \in \mathcal{B}^{res}_{0,p}$, the resolvability goes to 0 with rate $(\frac{\gamma \log(Mn)}{n})^{1-p/2}$.

$\mathcal{B}^{res}_{1,p}$ is a subset of $\mathcal{B}^{res}_{0,p}$, as follows by the argument used at inequality (5.10) after Theorem 5.2, with $m_0 = 1$ and either space yields risk of order bounded by $[(\log M)/n]^{1-p/2}$. Using larger m_0, if ε_{m_0} is polynormially small, that is, if $\mathcal{H}$ is of finite metric dimension w.r.t. $L_2(P)$, it follows that $\mathcal{B}^{res}_{1,p}$ is strictly smaller than $\mathcal{B}^{res}_{0,p}$, indeed, it is a subset of $\mathcal{B}^{res}_{0,p'}$, with p' smaller than p. Both provide resolvability of order $[\varepsilon_{m_0}(\log M)/n]^{1-p/2}$, close to rate $(1/n)^{1-p/2}$ if the dimension if large.

The issue arises as to whether these rates are best possible for these interpolation classes. Consider minimax risk of suitable subsets of $\mathcal{B}^{res}_{0,p}$. The balls $\{f^* \in \mathcal{B}^{res}_{0,p} : C_{0,p}(f^*) \leq C_0\}$ include the class of all f^* such that $App(f^*, m) \leq Cm^{-2r}$ where $r = 1/p - 1/2$ and C depends on C_0 and r. These approximation rates concern selection of arbitrary subsets of size m in $\mathcal{H}$. In the setting of Yang and Barron [86], Chapter 5 (which imposes additional structure), the minimax rate in such a sparse approximation class is m_n/n within logarithmic factors, where the subset size m_n is such that m_n/n matches the approximation bound m_n^{-2r}, which here yields rate $(1/n)^{2r/(2r+1)} = (1/n)^{1-p/2}$, and our results show it is achieved for all balls of $\mathcal{B}^{res}_{0,p}$. Thus not only for $p = 1$ (where the familiar rate $\sqrt{1/n}$ is known to be near optimal for functions of bounded variation w.r.t. $\mathcal{H}$), but also for the whole range of interpolation $\mathcal{B}^{res}_{0,p}$ with $0 \leq p \leq 2$, the approximate rates are identified.

Chapter 7

Examples

As we mentioned in the introduction, our main results in Chapter 4 and 5 are applicable to a bunch of flexible function fitting methods, depending on the choices of the library $\mathcal{H}$. Note that for our analysis, the covering property of our library is essential. Thus we concentrate in the following on the covering property of several libraries.

7.1 Smoothly Parameterized Libraries

Assume our target function $f^*(x)$ has domain $\mathcal{X}$. The library $\mathcal{H}$ consists of $\phi_\omega(x)$, continuously parameterized by a vector $\omega \in \Omega \subset \mathbb{R}^{d_\Omega}$, where Ω is a compact set and the functions ϕ_ω satisfy the Lipschitz condition

$$|\phi_\omega(x) - \phi_{\omega'}(x)| \leq \|\omega - \omega'\|_{l_1} \text{ for all } x \in \mathcal{X}.$$

Here $\|\cdot\|_{l_1}$ denote the l_1 norm in $\mathbb{R}^{d_\Omega}$. Also assume Ω is bounded by a constant R with respect to the norm $\|\cdot\|_{l_1}$. If $\mathcal{X} = [-1, 1]^d$, such models include trigonometric models with continuous parameters, certain multivariate wavelet models including ridglets, or single hidden layer sigmoidal networks with smooth sigmoids. Because of the Lips-

chitz condition of functions in $\mathcal{H}$, we know that $\|\phi_\omega - \phi_{\omega'}\|_\infty \leq |\omega - \omega'|_{l_1}$. Thus an l_1-covering of Ω, the domain of the parameters, yields an L_∞ covering of the library $\mathcal{H}$. Using Lemma 10 in [10], the cardinality of the ε-cover of the l_1-ball $\{\omega : \|\omega\|_{l_1} \leq R\}$ is bounded by $[2e(1 + R/\varepsilon)]^{d_\Omega}$, which implies $\mathcal{H}$ has covering number $\mathcal{N}(\varepsilon, \mathcal{H}, \|\cdot\|_{L_\infty}) \leq [2e\,(1 + R/\varepsilon)]^{d_\Omega}$. Since both the empirical L_1 norm and the empirical L_2 norm are not more than the L_∞ norm, an L_∞-cover is also an empirical L_1 and L_2 cover. Thus $d_1 \leq d_\Omega$ and $d_2 \leq d_\Omega$. Our current theory applies to such models. In particular, when the target function f^* has finite variation w.r.t. $\mathcal{H}$, the risk of the ℓ_1-penalized least squares estimator converges to zero at rate of order not more than $\left(\frac{d_\Omega}{n} \log \frac{n}{d_\Omega}\right)^{(d_\Omega+2)/(2d_\Omega+2)}$

7.2 Libraries of Indicator Functions

Suppose $\mathcal{D}$ is a class of sets with Vapnik-Červonenkis dimension $\mathbb{D}$ and $\mathcal{H}$ consists of all indicator functions of the sets in $\mathcal{D}$. One example here is single hidden layer sigmoidal networks with the sigmoid equal to a step function. This $\mathcal{H} = \mathcal{H}_{step}$ consists of functions $h_{a,b}(x) = \phi(a^T x - b)$ parametrized by (a, b) with internal weight vectors a in $\mathbb{R}^d$, internal location parameter b in $\mathbb{R}$ and $\phi(z) = 1_{\{z>0\}}$. Here $\mathcal{D}$ is the class of half-spaces in $\mathbb{R}^d$ and hence has VC-dimension $\mathbb{D} = d + 1$. Other examples includes the set of indicators of all rectangles or the set of indicators of all ellipsoids. we know that the empirical L_1 covering number of $\mathcal{H}$ is not more than $e(\mathbb{D} + 1)\left(\frac{4e}{\varepsilon_1}\right)^{\mathbb{D}}$. Thus $d_{1,\mathcal{H}} \leq \mathbb{D}$. Also for the indicator functions, the empirical L_2 norm is exactly equal to the square root of the empirical L_1 norm, which implies that $d_{2,\mathcal{H}} = 2d_{1,\mathcal{H}} \leq 2\mathbb{D}$. Therefore, given the target function f^* has finite variation, we know that the risk of the ℓ_1-penalized least squares estimator converges to zero at rate of order not more than

$$\left(\frac{\mathbb{D}}{n} \log \frac{n}{\mathbb{D}}\right)^{(\mathbb{D}+1)/(2\mathbb{D}+1)}$$

7.3 Tensor Product Models

Suppose Φ is a set of functions bounded by 1 and the library $\mathcal{H} = \Phi^d$ consists of all the functions of the form

$$h(x_1, \ldots, x_d) = \phi_1(x_1) \cdots \phi_d(x_d),$$

where $\phi_1, \ldots, \phi_d$ are functions in Φ. Typical examples are multivariate splines, multi-dimensional Fourier transformation with ϕ obtained from products of sines and cosines in the respective variables, and multivariate wavelets formed from products of univariate wavelet basis functions.

There are two variants of these models. One is to allow continuous knot locations with bounded range or continuous frequencies up to a maximum frequency. In this setting, the set Φ is infinite with covering properties essential to our analysis. If Φ has empirical L_1 and L_2 covers $\tilde{\Phi}_1$ and $\tilde{\Phi}_2$ of precision ε_1 and ε_2 respectively, then we can obtain a covering property of the library $\mathcal{H}$. Since $|\phi_1(x_1) \cdots \phi_d(x_d) - \phi'_1(x_1) \cdots \phi'_d(x_d)| \leq \sum_{i=1}^d |\phi_i(x_i) - \phi'_i(x_i)|$, we know that $\tilde{\Phi}_1^d$ is a $(d\varepsilon_1)$-cover of $\mathcal{H}$ with empirical L_1 norm. Also the fact $|\phi_1(x_1) \cdots \phi_d(x_d) - \phi'_1(x_1) \cdots \phi'_d(x_d)|^2 \leq d \sum_{i=1}^d |\phi_i(x_i) - \phi'_i(x_i)|^2$ implies that $\tilde{\Phi}_2^d$ is a $(d\varepsilon_2)$-cover of $\mathcal{H}$ with empirical L_2 norm. In particular, when Φ has finite metric dimensions $d_{1,\Phi}$ and $d_{2,\Phi}$ with respect to empirical L_1 and L_2 norms respectively, our $\mathcal{H}$ also has finite metric dimensions with $d_{1,\mathcal{H}} \leq d \cdot d_{1,\Phi}$ and $d_{2,\mathcal{H}} \leq d \cdot d_{2,\Phi}$.

The other case of interest is that the univariate library is a union of finite dictionaries, i.e., $\Phi = \bigcup_j \Phi_j$, where j is in an countable index set $\mathbb{J}$. The index j can specify, for example, the number of equal-spaced knots and the order of a spline or the number of levels and order of wavelets. Let K_j be the cardinality of Φ_j and $\mathcal{H}_{j,d} = \Phi_j^d$ be the product dictionary of terms $h(x_1, x_2, \ldots, x_d)$ as above as products of the functions in Φ_j with cardinality K_j^d. Suppose $L(j)$ is a complexity associated with the index j in $\mathbb{J}$ satisfying the Kraft inequality. By the same methods as in Chapter 5, for functions f that are m-term

linear combination of elements of $\mathcal{H}_{j,d}$, a valid penalty is

$$\text{pen}_n(m,j) = \gamma \left\{ \log \binom{K_j^d}{m} + (m+1)\log n + d \log K_j + L(j) \right\},$$

permitting optimization over choices of size of dictionaries as well as choices of subsets, and leading to corresponding risk conclusion in accordance with our theory.

7.4 Libraries with infinite metric dimension

The library $\mathcal{H}$ may be a much bigger library than those with finite metric dimension. For instance, the optimal covering number M_ε may be of order $\exp[(1/\varepsilon)^\theta]$ with $\theta > 0$. Typically, $\theta = d/s$ for functions of smoothness s in d variables. A fascinating result in this setting is that if the target is in $\mathcal{L}_{1,\mathcal{H}}$, the resolvability rate exponent is one-half of what it would be for a single term in the library. Indeed, recalling the result in [86], we know that $\min_{\hat{f}} \max_{f^* \in \mathcal{H}} \mathbb{E}\|\hat{f} - f^*\|^2$ is of order ε_n^2, where ε_n satisfies $\varepsilon_n^2 = (\log M_{\varepsilon_n})/n$. For both the ℓ_1 and the subset selection cases, for functions in $\mathcal{L}_{1,\mathcal{H}}$, the resolvability is of order $\sqrt{\frac{\log M_\varepsilon}{n}} + \varepsilon$, optimized w.r.t. ε, producing rate ε_n, which for $\mathcal{L}_{1,\mathcal{H}}$ is the squared root of the minimax rate for $\mathcal{H}$.

We caution that infinite-dimensional libraries $\mathcal{H}$ often have characteristics that make their consideration less pertinent for estimation of linear combinations of terms from $\mathcal{H}$. One such characteristic is that their size may lead to slow rates even in the case of a limited number of terms. Moreover, in smoothness class settings, such infinite-dimensional libraries may be closed under linear combination, which voids need for consideration of more than one term.

An interesting special case for use of linear combinations of an infinite-dimensional class is that of projection pursuit regression, where the library is a set of ridge functions of

x in $\mathcal{X}$, a bounded set of $\mathbb{R}^d$,

$$\mathcal{H} = \{\phi(a^T x) : a \in \Omega, \phi \in \Phi\},$$

with Φ a standard smoothness class of functions ϕ on $\mathbb{R}$ and Ω a compact subset of $\mathbb{R}^d$. Now $\mathcal{F}_{\mathcal{H}}$ consists of linear combinations of ridge functions which is a larger class than $\mathcal{H}$ for $d > 1$. For Φ, consider the class Φ_α of all functions satisfying a Lipschitz condition of order α with $0 < \alpha \leq 1$ and let $\mathcal{H}_{\alpha,d}$ be the associated set of ridge functions. The metric entropy of Φ_α is of order $(1/\varepsilon)^{1/\alpha}$ and of $\mathcal{H}_{\alpha,d}$ is of order $(1/\varepsilon)^{1/\alpha} + d\log(1/\varepsilon)$. A related library $\mathcal{H}_{step}$ (as in 7.2 above) uses Φ containing only a single step function. That may appear to be more restrictive. However, for $\alpha = 1$, the space $\mathcal{L}_{1,\mathcal{H}}$ of functions of finite variation w.r.t. $\mathcal{H}_{1,d}$ is included in the space of finite variation w.r.t. $\mathcal{H}_{step}$. In general, if $\mathcal{H}$ is constructable from a finite-dimensional set $\tilde{\mathcal{H}}$ via linear combination, the resolvability for functions in $\mathcal{L}_{1,\mathcal{H}}$ has rate of the same order as for the finite-dimensional case.

7.5 Concluding Comment

: One often hears emphasis on whether the number of terms M of a library $\mathcal{H}$ for linear combination is finite or infinite and, if it is finite, whether it is of larger or smaller order than n. We emphasize for subset selections and for ℓ_1-penalization that the key issue is not whether the library is finite or infinite but rather whether it has finite covering properties and, if it does, then for the effective cardinality M_n, the issue is not how it compares to n, but rather whether $\log M_n$ is small compared to n.

For functions in $\mathcal{L}_{1,\mathcal{H}}$ as well as the associated interpolation classes, forward stepwise and ℓ_1 penalized least squares produce risk of order equal to a power of $(\log M_n)/n$, where the power is between 0 and 1. With these procedures, one does not need to know which

class contains f^*. The risk is controlled by an index of resolvability showing the estimation simultaneously achieves desirable levels of performance of all such classes.

Chapter 8

Appendix

8.1 Lemmas for Chapter 2

Lemma 8.1 *Let* $(\underline{X}, \underline{X}') = (X_1, \ldots, X_n, X_1', \ldots, X_n')$ *where* $\underline{X}'$ *is an independent copy of the data* $\underline{X}$ *and* $(X_1, \ldots, X_n)$ *are component-wise independent but not necessarily identically distributed. Given a fixed function class* $\mathcal{G}$*, possibly uncountable, suppose* $\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}$ *is a countable subset of* $\mathcal{G}$ *with associated complexities* $L_{\underline{X},\underline{X}'}(\tilde{g})$ *satisfying Assumption (S). Then for arbitrary positive* u *and* γ*, we have*

$$\mathbb{P}\left\{\sup_{\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}} \frac{P_n'(\tilde{g}) - P_n(\tilde{g})}{u + \frac{\gamma}{n}L_{\underline{X},\underline{X}'}(\tilde{g}) + \frac{1}{2\gamma}s^2(\tilde{g})} \geq 1\right\} \leq \exp\left(-\frac{nu}{\gamma}\right), \tag{8.1}$$

where $s^2(\tilde{g}) = \frac{1}{n}\sum_{i=1}^n \left(\tilde{g}(X_i) - \tilde{g}(X_i')\right)^2$*. Moreover,*

$$\mathbb{E}\sup_{\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}} \left\{P_n'(\tilde{g}) - P_n(\tilde{g}) - \frac{\gamma L_{\underline{X},\underline{X}'}(\tilde{g})}{n} - \frac{1}{2\gamma}s^2(\tilde{g})\right\} \leq 0. \tag{8.2}$$

Proof: Set $u_1 = u + \frac{\gamma}{n}L_{\underline{X},\underline{X}'}(\tilde{g})$ for simplicity. Indeed to verify the claim, we bound the

probability of the event as follows

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}}\frac{P_n'(\tilde{g})-P_n(\tilde{g})}{u_1+\frac{1}{2\gamma}s^2(\tilde{g})}\geq 1\right\}\\
\leq\ &\mathbb{P}\left\{\exists\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}:P_n'(\tilde{g})-P_n(\tilde{g})\geq u_1+\frac{1}{2\gamma}s^2(\tilde{g})\right\}\\
\leq\ &\mathbb{P}\left\{\exists\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}:P_n'(\tilde{g})-P_n(\tilde{g})\geq\sqrt{\frac{2u_1}{\gamma}}s(\tilde{g})\right\}\\
=\ &\mathbb{P}\left\{\exists\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}:\frac{P_n'(\tilde{g})-P_n(\tilde{g})}{s(\tilde{g})}\geq\sqrt{\frac{2u_1}{\gamma}}\right\}
\end{aligned}\tag{8.3}$$

by using the inequality $\frac{a^2}{2}+\frac{b^2}{2}\geq ab$.

One lets $\underline{Z}=(Z_1,\ldots,Z_n)$ be independent ± 1 valued equiprobable random variables (so that $\mathbb{E}Z_i=0$). Since the distribution of $\underline{X},\underline{X}'$ is coordinate pair exchangable and Assumption (S) holds, we have, for any realization of $\underline{Z}$, that multiplying the differences $g(X_i)-g(X_i')$ by Z_i leaves the probability on the right side of inequality (8.3) unchanged and hence equal to the following probability with respect to the joint distribution of $\underline{X},\underline{X}',\underline{Z}$, which we then evaluate by conditioning on $\underline{X},\underline{X}'$ and invoking Hoeffding's inequality for the random $\underline{Z}$.

$$\begin{aligned}
&\mathbb{P}\left\{\exists\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}:\frac{1}{n}\sum_{i=1}^n\frac{Z_i(\tilde{g}(X_i')-\tilde{g}(X_i))}{\sqrt{\frac{1}{n}\sum_{i=1}^n(\tilde{g}(X_i')-\tilde{g}(X_i))^2}}\geq\sqrt{\frac{2u_1}{\gamma}}\right\}\\
\leq\ &\mathbb{E}\sum_{\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}}\mathbb{P}_{\underline{Z}|\underline{X},\underline{X}'}\left\{\frac{1}{n}\sum_{i=1}^n\frac{Z_i(\tilde{g}(X_i')-\tilde{g}(X_i))}{\sqrt{\frac{1}{n}\sum_{i=1}^n(\tilde{g}(X_i')-\tilde{g}(X_i))^2}}\geq\sqrt{\frac{2u_1}{\gamma}}\right\}\\
\leq\ &\mathbb{E}\sum_{\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}}\exp\left(-\frac{nu_1}{\gamma}\right)\\
=\ &\mathbb{E}\sum_{\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}}\exp\left(-\frac{nu}{\gamma}-L_{\underline{X},\underline{X}'}(\tilde{g})\right)\\
\leq\ &\exp\left(-\frac{nu}{\gamma}\right).
\end{aligned}$$

Thus the first claim is proven.

For the second conclusion, as before, by coordinate pair exchangability and symmetry, we know that the left side of (8.2) is equal to,

$$\mathbb{E}_{\underline{X},\underline{X}',\underline{Z}} \sup_{\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}} \left\{ \frac{1}{n}\sum_{i=1}^{n} Z_i(\tilde{g}(X_i') - \tilde{g}(X_i)) - \frac{\gamma L_{\underline{X},\underline{X}'}(\tilde{g})}{n} - \frac{1}{2\gamma}s^2(\tilde{g}) \right\}. \tag{8.4}$$

Using the identity $x = \lambda \log \exp(x/\lambda)$ inside the expectation with $\lambda = \gamma/n$, conditioning on $\underline{X}$ and $\underline{X}'$, and applying Jensen's inequality to move $\mathbb{E}_{\underline{Z}}$ inside the log function, the expression inside the expectation $\mathbb{E}_{\underline{X},\underline{X}'}$ is less than or equal to

$$\frac{\gamma}{n}\log \mathbb{E}_{\underline{Z}} \sup_{\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}} \exp\left\{ \frac{1}{\gamma}\left(\sum_{i=1}^{n} Z_i(\tilde{g}(X_i') - \tilde{g}(X_i)) \right) - L_{\underline{X},\underline{X}'}(\tilde{g}) - \frac{n}{2\gamma^2}s^2(\tilde{g}) \right\}. \tag{8.5}$$

Replacing the supremum with the sum and moving the expectation inside the sum, (8.5) is not more than

$$\frac{\gamma}{n}\log \sum_{\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}} \exp\left\{ -L_{\underline{X},\underline{X}'}(\tilde{g}) - \frac{n}{2\gamma^2}s^2(\tilde{g}) \right\} \mathbb{E}_{\underline{Z}} \exp\left\{ \frac{1}{\gamma}\sum_{i=1}^{n} Z_i(\tilde{g}(X_i') - \tilde{g}(X_i)) \right\}. \tag{8.6}$$

Since $\{Z_i\}_{i=1}^n$ are independent with each other, the expectation with respect to $\underline{Z}$ in the expression (8.6) is equal to the product of $\mathbb{E}_{Z_i} \exp\left\{ \frac{1}{\gamma} Z_i(\tilde{g}(X_i') - \tilde{g}(X_i)) \right\}$, which is less than or equal to $\exp\left\{ \frac{1}{2\gamma^2}(\tilde{g}(X_i') - \tilde{g}(X_i))^2 \right\}$ for each i by using the inequality $e^x + e^{-x} \leq 2e^{x^2/2}$. Hence, the expression (8.6) is upper bounded by

$$\frac{\gamma}{n}\log \sum_{\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}} \exp\left\{ -L_{\underline{X},\underline{X}'}(\tilde{g}) - \frac{n}{2\gamma^2}s^2(\tilde{g}) \right\} \exp\left\{ \frac{n}{2\gamma^2}s^2(\tilde{g}) \right\},$$

which is less than or equal to $\frac{\gamma}{n}\log \sum_{\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}} \exp\{-L_{\underline{X},\underline{X}'}(\tilde{g})\} \leq 0$. Then the conclusion follows. ∎

Lemma 8.2 *Assume* $\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ *given* $\underline{X} = (X_1, \ldots, X_n)$ *are conditionally independent with distributions for* ϵ_i *given* X_i *that satisfy Assumption (M). Also assume* $\underline{X}' = (X_1', \ldots, X_n')$ *is an independent copy of* $\underline{X}$. *Given a function class* $\mathcal{G}$, *suppose* $\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}$ *is a countable subset of* $\mathcal{G}$ *with associated complexities* $L_{\underline{X},\underline{X}'}(\tilde{g})$ *satisfying Assumption (S). Assume for all* $\tilde{g}$ *in* $\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}$ *that the absolute value of* $\tilde{g}(x)$ *is bounded by a constant* K. *Then*

$$\mathbb{P}\left\{\sup_{\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}} \frac{\frac{1}{n}\sum_{i=1}^n \epsilon_i\tilde{g}(X_i)}{u + \frac{\gamma}{n}L_{\underline{X},\underline{X}'}(\tilde{g}) + \frac{1}{An}\sum_{i=1}^n \tilde{g}^2(X_i)} \geq 1\right\} \leq \exp\left(-\frac{nu}{\gamma}\right) \tag{8.7}$$

where A *and* u *are arbitrary positive constants, and* $\gamma = A\sigma^2/2 + Kh_{Bern}$. *Moreover,*

$$\mathbb{E}\sup_{\tilde{g}\in\tilde{\mathcal{G}}_{\underline{X},\underline{X}'}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i\tilde{g}(X_i) - \frac{\gamma}{n}L_{\underline{X},\underline{X}'}(\tilde{g}) - \frac{1}{An}\sum_{i=1}^n \tilde{g}^2(X_i)\right\} \leq 0. \tag{8.8}$$

Proof: For simplicity, denote $\mathcal{G}_n = \mathcal{G}_{\underline{X},\underline{X}'}$, $L_n(\tilde{g}) = L_{\underline{X},\underline{X}'}(\tilde{g})$ and $h = h_{Bern}$.

By the union of events bound, the probability

$$\begin{aligned}
&\mathbb{P}\left\{\exists\tilde{g}\in\tilde{\mathcal{G}}_n : \frac{1}{n}\sum_{i=1}^n \epsilon_i\tilde{g}(X_i) \geq \frac{1}{A}\|\tilde{g}\|_n^2 + \frac{\gamma}{n}L_n(\tilde{g}) + u\right\} \\
\leq\ &\mathbb{E}_{\underline{X},\underline{X}'}\sum_{\tilde{g}\in\tilde{\mathcal{G}}_n}\mathbb{P}_{\epsilon|\underline{X}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i(\tilde{g}(X_i)) \geq \frac{1}{A}\|\tilde{g}\|_n^2 + \frac{\gamma}{n}L_n(\tilde{g}) + u\right\}.
\end{aligned} \tag{8.9}$$

Let $u_2 = \frac{\gamma}{n}L_n(\tilde{g}) + u$ for simplicity. Let $R_i = \epsilon_i\tilde{g}(X_i)$ and $\bar{R} = \frac{1}{n}\sum_{i=1}^n R_i$. Note also that $\mathbb{E}\bar{R} = 0$. Under Assumption (M), $var(\epsilon_i|X_i) \leq \sigma^2$, so that $\mathrm{var}(\bar{R}|\underline{X}) = \frac{1}{n^2}\sum_{i=1}^n(\tilde{g}(X_i))^2\mathrm{var}(\epsilon_i|X_i) \leq \frac{1}{n}\|g\|_n^2\sigma^2$ and R_i also satisfy the Bernstein's moment condition with $h' = B'h$. Then the right side of (8.9) satisfies

$$\begin{aligned}
&\mathbb{E}_{\underline{X},\underline{X}'}\sum_{\tilde{g}\in\tilde{\mathcal{G}}_n}\mathbb{P}_{\underline{\epsilon}|\underline{X}}\left\{\bar{R} \geq u_2 + \frac{1}{A}\|\tilde{g}\|_n^2\right\} \\
\leq\ &\mathbb{E}_{\underline{X},\underline{X}'}\sum_{\tilde{g}\in\tilde{\mathcal{G}}_n}\mathbb{P}_{\underline{\epsilon}|\underline{X}}\left\{\bar{R} \geq u_2 + \frac{n}{A\sigma^2}\mathrm{var}(\bar{R}|X)\right\}.
\end{aligned} \tag{8.10}$$

Here we do not use Bernstein's inequality directly. Instead we use the following inequality (8.11) that Craig [37] develops in his proof of Bernstein's inequality. If Z_i are independent random variables satisfying Bernstein's moment condition then

$$\mathbb{P}\left\{\bar{Z} - \mathbb{E}\bar{Z} \geq \frac{\tau}{nt} + \frac{nt\text{var}(\bar{Z})}{2(1-c)}\right\} \leq e^{-\tau} \tag{8.11}$$

for any $0 < th \leq c < 1$ and $\tau > 0$.

Now to apply to (8.10), we arrange that $u_2 = \frac{\tau}{nt}$ and $\frac{t}{2(1-c)} = \frac{1}{A\sigma^2}$ and $t = \frac{c}{Kh}$, which together yield $\tau = ntu_2 = \frac{nu_2}{A\sigma^2/2+Kh} = \frac{nt}{\gamma} + L(g)$, where $\gamma = A\sigma^2/2 + Kh$. Using Craig's inequality (8.11), the right side of (8.10) is less than or equal to

$$\mathbb{E}_{\underline{X},\underline{X}'} \sum_{g\in\mathcal{G}_n} \exp\left(-\frac{nt}{\gamma} - L_n(\tilde{g})\right) \leq \exp\left(-\frac{nt}{\gamma}\right). \tag{8.12}$$

The second claim uses a similar argument as in the proof of the second claim in Lemma 8.1. Applying Jensen's inequality to move the conditional expectation $\mathbb{E}_{\underline{\epsilon}|\underline{X}}$ inside the log function, the left side of (8.12) is less than or equal to

$$\frac{\gamma}{n}\mathbb{E}_{\underline{X},\underline{X}'} \log \mathbb{E}_{\underline{\epsilon}|\underline{X}} \sup_{\tilde{g}\in\tilde{\mathcal{G}}_n} \exp\left\{\frac{1}{\gamma}\left(\sum_{i=1}^{n} \epsilon_i \tilde{g}(X_i)\right) - L_n(\tilde{g}) - \frac{n}{\gamma A}\|\tilde{g}\|_n^2\right\}. \tag{8.13}$$

Replacing the supremum with the sum and moving the conditional expectation inside the sum, the expression inside the expectation $\mathbb{E}_{\underline{X},\underline{X}'}$ in (8.13) is not more than

$$\log \sum_{\tilde{g}\in\tilde{\mathcal{G}}_n} \exp\left\{-L_n(\tilde{g}) - \frac{n}{A\gamma}\|\tilde{g}\|_n^2\right\} \mathbb{E}_{\underline{\epsilon}|\underline{X}} \exp\left\{\frac{1}{\gamma}\sum_{i=1}^{n}\sum_{i=1}^{n} \epsilon_i \tilde{g}(X_i)\right\}. \tag{8.14}$$

Since $\{\epsilon_i\}_{i=1}^n$ are conditionally independent with each other, the conditional expectation in the expression (8.14) is equal to the product of $\mathbb{E}_{\epsilon_i|X_i} \exp\left\{\frac{1}{\gamma}\epsilon_i \tilde{g}(X_i)\right\}$, which is less than

or equal to

$$\exp\left\{\frac{\sigma^2\tilde{g}^2(X_i)}{2\gamma^2}\frac{1}{1-\tilde{g}(X_i)h/\gamma}\right\} \tag{8.15}$$

for each i by using the Bernstein's moment generating function inequality. Since $|\tilde{g}(X_i)| \leq K$, we know that $1-\tilde{g}(X_i)h/\gamma$ is not less than $1-Kh/\gamma$. Then (8.15) is not more than $\exp\left\{\frac{\sigma^2\tilde{g}^2(X_i)}{2\gamma^2}\frac{1}{1-Kh/\gamma}\right\}$, which is equal to $\exp\left\{\frac{1}{A\gamma}\tilde{g}^2(X_i)\right\}$ from the definition of γ. Hence, the expression (8.14) is upper bounded by

$$\log\sum_{\tilde{g}\in\tilde{\mathcal{G}}_n}\exp\left\{-L_n(\tilde{g})-\frac{n}{A\gamma}\|\tilde{g}\|_n^2\right\}\exp\left\{\frac{n}{A\gamma}\|\tilde{g}\|_n^2\right\},$$

which is less than or equal to $\log\sum_{\tilde{g}\in\tilde{\mathcal{G}}_n}\exp\{-L_{\underline{X},\underline{X}'}(\tilde{g})\} \leq 0$. Then the conclusion follows. ∎

Lemma 8.3 *Let $Y=\epsilon+f^*(X)$ with $|f^*(X)|\leq B$. (1) If ϵ is a random variable satisfying $\mathbb{E}\exp(|\epsilon|/\nu)=D_1<\infty$ for a positive constant ν, then for $B'\geq B+\nu\log n$,*

$$\mathbb{E}(|Y|-B')1\{|Y|>B'\}\leq\frac{D_1\nu}{n}$$

and

$$\mathbb{E}(|Y|-B')^2 1\{|Y|>B'\}\leq\frac{2D_1\nu^2}{n}.$$

(2) If ϵ satisfies $\mathbb{E}\exp(\epsilon^2/\nu)=D_2<\infty$ for some $\nu>0$, then for $B'\geq B+\sqrt{\nu\log n}$,

$$\mathbb{E}(|Y|-B')1\{|Y|>B'\}\leq\frac{D_2\sqrt{\pi\nu}}{n}$$

and

$$\mathbb{E}(|Y|-B')^2 1\{|Y|>B'\}\leq\frac{D_2\nu}{2n}.$$

Proof: In both cases we start with

$$\mathbb{P}\{|Y| > B' + t\} \leq \mathbb{P}\{|\epsilon| > B' - B + t\}.$$

Case (1): Using the definition of B', this is not more than $\mathbb{P}\{|\epsilon| > t + \nu \log n\}$. Inside we divide through by ν, exponentiate and apply Markov's inequality to obtain

$$\mathbb{P}\{|Y| > B' + t\} \leq D_1 \exp(-\frac{t}{\nu} - \log n) \tag{8.16}$$

Integrate with respect to t from 0 to ∞ to obtain the first claim of case (1). Multiply both side of (8.16) with t and then integrate to obtain the second claim of (1).

Case (2): Likewise, using the definition of B' yields,

$$\begin{aligned} & \mathbb{P}\{|Y| > B' + t\} \\ \leq\ & P\{|\epsilon| > t + \sqrt{\nu \log n}\} \\ \leq\ & D_2 \exp\left\{-(t + \sqrt{\nu \log n})^2/\nu\right\} \end{aligned} \tag{8.17}$$

Integrate with respect to t from 0 to ∞. The integral on the left side becomes $E(|Y| - B')1\{|Y| > B'\}$. By changing variable with $\tau = \sqrt{2}(t + \sqrt{\nu \log n})/\sqrt{\nu}$, the integral on the right side becomes $\sqrt{\nu/2} D_2 \int_{\sqrt{2 \log n}}^{\infty} \exp(-\tau^2/2) d\tau$, which not more than $\frac{D_2 \sqrt{\pi \nu}}{n}$. Finally, multiply both side of (8.17) with t and then integrate. The integral on the left becomes $E(|Y| - B')^2 1\{|Y| > B'\}$. By changing variable with $t' = (t + \sqrt{\nu \log n})/\sqrt{\nu}$, the integral on the right becomes $D_2 \nu \int_{\sqrt{\log n}}^{\infty} t' exp(-t'^2) dt'$, which is equal to $\frac{D_2 \nu}{2n}$. ∎

8.2 Lemmas and Proofs for Chapter 4

Next we collect some approximation properties for linear combinations of not more than m terms selected from a library. Assume the Hilbert space setting of Chapter 3, with $\mathcal{F}$ the linear span of a subset $\mathcal{H}$.

Lemma 8.4 *For $f = \sum_h \beta_h h$ in $\mathcal{F}$, there are choices of h_1, h_2, ..., h_m in $\mathcal{H}$, with repeats allowed, for which linear combinations f_m taking the following forms have the indicated approximation bounds.*

(1) Suppose $\|h\| \leq b$ in $\mathcal{H}$. For any $v \geq \sum_h |\beta_h|$ there is an $f_m = (v/m) \sum_{k=1}^m h_k$ such that

$$\|f - f_m\|^2 \leq \frac{(bv)^2}{m}. \tag{8.18}$$

(2) Suppose $a_h \geq \|h\|$ in $\mathcal{H}$. For any $v \geq \|\beta\|_{1,a}$ there is an $f_m = (v/m) \sum_{k=1}^m h_k/a_{h_k}$ such that $\|f - f_m\|^2 \leq (v/m)\|\beta\|_{1,a}$ and hence

$$\|f - f_m\|^2 \leq \frac{v^2}{m}. \tag{8.19}$$

(3) For $m = m_0 + m_1$, let a size m_0 subset $\tilde{\mathcal{H}} \subset \mathcal{H}$ be given and let $\varepsilon_{h,\tilde{\mathcal{H}}} = \inf_{\tilde{h} \in \tilde{\mathcal{H}}} \|h - \tilde{h}\|$ for $h \in \mathcal{H}$. There is a v between $\sum_h |\beta_h|$ and $\sum_h |\beta_h|(1 + m_0/m_1)$ and a choice of $f_m = (v/m) \sum_{k=1}^m h_k$ such that $\|f - f_m\|^2 \leq (v/m) \sum_h |\beta_h| \varepsilon_{h,\tilde{\mathcal{H}}}^2$ and hence

$$\|f - f_m\|^2 \leq \frac{(\sum_h |\beta_h|)(\sum_h |\beta_h| \varepsilon_{h,\tilde{\mathcal{H}}}^2)}{m_1}. \tag{8.20}$$

(4) For $m = m_0 + m_1$ and a size m_0 subset $\tilde{\mathcal{H}}$, let $S = \text{span}\tilde{\mathcal{H}}$ be its linear span and Π_S the operation of linear projection onto S, and let $a_h = \varepsilon_{h,S} = \|h - \Pi_S h\|$. Then for $v = \sum_h |\beta_h| \varepsilon_{h,S}$, there are choices of h_1, h_2, ..., h_{m_1} such that $f_m = \Pi_S f + (v/m) \sum_{k=1}^{m_1} (h_k -$

$\Pi_S h_k)/\varepsilon_{h_k,S}$ *satisfies*

$$\|f - f_m\|^2 \leq \frac{(\sum_h |\beta_h|\varepsilon_{h,S})^2}{m_1}. \tag{8.21}$$

Each of these four bounds on $\|f - f_m\|^2$ *is also a bound on the following difference,*

$$\|f^* - f_m\|^2 - \|f^* - f\|^2,$$

for any f^* *in the Hilbert space for some* f_m *of the indicated forms.*

Proof: All of these conclusions are consequences of reasonably familiar ideas of sampling. Without loss of generality assume $\mathcal{H}$ is closed under sign change. For $f = f_\beta$ in the linear span of $\mathcal{H}$ one has a finite set of h for which β_h is non-zero and we may take them to be positive. Case 1 relies on classical sample average facts, as in [7], and is a special case of the analysis for Case 2, which we now give. Consider the distribution in which each such h is assigned probability $\beta_h a_h/v$ and any left-over probability (due to v possibly strictly larger than $\sum_h \beta_h a_h$) is assigned to $h = 0$. Draw $h_1, h_2, \ldots, h_m$ independently from the indicated distribution. Treat h/a_h as equal to 0 if $h = 0$. Then the expectation of vh_k/a_{h_k} is equal to f as is the expectation of f_m. Furthermore, the inner product of $vh_k/a_{h_k} - f$ with $vh_j/a_{h_j} - f$ has expectation 0 for any $j \neq k$ and expectation $\sum_h \beta_h a_h \|vh/a_h - f\|^2/v = v\sum_h \beta_h \|h\|^2/a_h - \|f\|^2$ for $j = k$. Correspondingly, the expectation of $\|f_m - f\|^2$ is $(1/m)$ times the corresponding expectation in a single draw, so there exists such f_m with

$$\|f_m - f\|^2 \leq \frac{1}{m}\left(v\sum_h \beta_h \|h\|^2/a_h - \|f\|^2\right). \tag{8.22}$$

Using $\|h\| \leq a_h$ yields the stated conclusion for Case 2. Note then when $v = \sum_h |\beta_h| a_h$, using the Cauchy-Schwartz inequality, the choice $a_h = \|h\|$ is seen to optimize this bound, which may then be written $\frac{1}{m}\left[(\sum_h \beta_h \|h\|)^2 - \|f\|^2\right]$. Alternatively, if $a_h = 1$ so that the

coefficients of the terms in f_m are all equal then the bound (8.22) provides the conclusion for Case 1 in the strengthened form (here ignoring the subtraction of $\|f\|^2$), namely,

$$\|f_m - f\|^2 \leq \frac{v}{m} \sum_h \beta_h \|h\|^2. \tag{8.23}$$

The remaining conclusions improve upon these bounds by showing that the weight determined by the norm $\|h\|$, which is the distance of h from 0, can be replaced by the distance of h from a particular finite set of points (Case 3) or the distance from their linear span (Case 4).

Now we verify the claim for Case 3, which is related to a result in [62]. We show that f_m can take the simpler form shown here. The proof uses what we recognize to be a stratified sampling argument (though he did not use that terminology). Partition $\mathcal{H}$ into m_0 disjoint cells c where each cell consists of the points closest to a particular $\tilde{h}$ in $\tilde{H}$, breaking ties arbitrarily. Consider $v(c) \geq \sum_{h \in c} |\beta_h|$ with sum denoted $v = \sum_c v(c)$ and consider integers $m(c)$ with sum not more than m. For each cell c, draw $h_{c,k}$ for $k = 1, 2, \ldots, m(c)$ independently with choice h with probability β_h / v_c for h in c, and choice 0 with any left-over probability, due to $v(c)$ possibly strictly larger than $\sum_{h \in c} \beta_h$. Form the within-cell sample averages $f_{c,m} = \frac{1}{m(c)} \sum_{k=1}^{m(c)} h_{c,k}$, and, in general, the linear combination $f_m = \sum_c v(c) f_{c,m}$, for which the coefficients of the individual terms take the form $v(c)/m(c)$. In the special case of sizes $m(c)$ proportional to $v(c)$ these ratios are the same for all terms and f_m takes the indicated form. In this case, for some η we have $v(c) = \eta m(c)$ with sum $v = \eta m$, and so $v(c)/m(c) = \eta = v/m$. The within-cell sample averages have expectation f_c so that the overall average $f_m = \sum_c v(c) f_{c,m}$ has expectation $f = \sum_c v(c) f_c$. Then by independence the expectation of $\|f_m - f\|^2$ is $\sum_c v(c)^2 \frac{1}{m(c)} \sum_{h \in c} \|h - f_c\|^2 \beta_h / v(c)$. The inner sum is centered at the average f_c for members of $h \in c$ with the indicated weights. So that inner sum is less than what one

would have with f_c replaced by 0 or by any other point depending on c. We take in particular the representer of h, denoted $\tilde{h}$, shared by all h in c, at which $\|h - \tilde{h}\| = \varepsilon_{h,\tilde{\mathcal{H}}}$. Thus the expectation of $\|f_m - f\|^2$ is bounded by

$$\sum_c \frac{v(c)}{m(c)} \sum_{h \in c} \beta_h \min\{\|h\|^2, \varepsilon^2_{h,\tilde{\mathcal{H}}}\}, \tag{8.24}$$

and hence there exists such f_m for which $\|f_m - f\|^2$ has this bound. In particular, if we set $v(c) = \eta\lceil \sum_{h \in c} \beta_h / \eta \rceil$ (that is the value $\sum_{h \in c} \beta_h$ rounded up in a grid of spacings η) and set $m(c) = \lceil \sum_{h \in c} \beta_h / \eta \rceil$. Then $v(c)/m(c) = \eta$, the estimator takes the desired form, and the bound is

$$\eta \sum_h \beta_h \varepsilon^2_{h,\tilde{\mathcal{H}} \cup \{0\}}. \tag{8.25}$$

Here $\varepsilon_{h,\tilde{\mathcal{H}} \cup \{0\}} = \min\{\|h\|, \varepsilon_{h,\tilde{\mathcal{H}}}\}$. To complete the analysis, note that $v(c)$ is between $\sum_{h \in c} \beta_h$ and $\sum_{h \in c} \beta_h + \eta$, which when summed gives v between $\sum_h \beta_h$ and $\sum_h \beta_h + m_0 \eta$. Then from $\eta = v/m$ that yields $v \leq \sum_h \beta_h / (1 - m_0/m)$ or $\eta \leq \sum_h \beta_h / m_1$, which when plugged into (8.25) provides the desired bound

$$\|f_m - f\|^2 \leq \frac{(\sum_h |\beta_h|)(\sum_h |\beta_h| \varepsilon^2_{h,\tilde{\mathcal{H}}})}{m_1}. \tag{8.26}$$

With $\varepsilon_{\tilde{\mathcal{H}}} = \sup_{h \in \mathcal{H}} \varepsilon_{h,\tilde{\mathcal{H}}}$, that is, if $\tilde{\mathcal{H}}$ is a size m_0 cover with precision $\varepsilon_{\tilde{\mathcal{H}}}$, then

$$\|f_m - f\|^2 \leq \varepsilon^2_{\tilde{\mathcal{H}}} \frac{(\sum_h |\beta_h|)^2}{m_1}. \tag{8.27}$$

This conclusion is comparable to [62], with the improvements that f_m may take the simpler form and that the precision ρ is based on the radius of cells (distance from the representers), whereas his corresponding conclusion is for the diameter of the cells.

For Case 4 draw $h_1, h_2, \ldots, h_{m_1}$ independently with probability $\beta_h \varepsilon_{h,S}/v$. The other m_0 terms are those in $\Pi_S f$. By the definition of f_m we see that $f_m - \Pi_S f$ is the average of the m_1 terms $v(h_k - \Pi_S h_k)/\varepsilon_{h_k,S}$ for which the expectation is $\sum_h \beta_h (h - \Pi h) = f - \Pi_S f$, so one then follows the same argument as in Case 2, to obtain an expectation of $\|f - f_m\|^2$ and hence the existence of such an f_m for which

$$\|f - f_m\|^2 \leq \frac{(\sum_h |\beta_h| \varepsilon_{h,S})^2}{m_1}, \tag{8.28}$$

which is the desired conclusion for Case 4.

Likewise $\|f^* - f_m\|^2 - \|f^* - f\|^2$ has the same expectation as $\|f - f_m\|^2$ with f_m unbiased for f in all four cases. This completes the proof of Lemma 8.4. ∎

Case 4 provides an improved bound but with less explicit control on its coefficients. Note that the span S includes $\tilde{\mathcal{H}} \cup \{0\}$ so $\varepsilon_{h,S}$ is less than or equal to $\varepsilon_{h,\tilde{\mathcal{H}} \cup \{0\}}$, perhaps substantially less, so by that inequality and by Cauchy-Schwartz, the conclusion (8.28) is indeed superior to (8.26).

The next lemma considers the case that $\mathcal{H}$ is finite with cardinality M and provides log cardinality bounds on the number of f_m of the forms specified in the preceeding Lemma for Cases 1,2, and 3. As in Chapter 4, the choice of $v = \eta m$ makes v determined from m, so we only need to count the number of $h_1, h_2, \ldots, h_m$. This log cardinality will be less than $m \log M$ recognizing that repeats are allowed and the order does not matter.

Lemma 8.5 *The log cardinality of the set of terms $h_1, h_2, \ldots, h_m$, selected from a give library of size M with repeats permitted, is $\log \sum_{k=1}^{\min\{m,M\}} \binom{M}{k} \binom{m-1}{k-1}$, not more than $m \log(2e \max\{\frac{M}{m}, 1\})$.*

Adding $m \log 2$ for description of m, gives the variable complexity bounds used in Chapter 4.

Proof: The number of distinct terms k is between 1 and $\min\{m, M\}$. For each such k we have $\binom{M}{k}$ choices of subsets of size k. For each we have $\binom{m-1}{k-1}$, not more than 2^{m-1}, choices of how to assign counts of at least one of each. (This is in accordance with the standard stars and bars argument, for the placement of $k-1$ bars among $(m-k)+(k-1)$ positions to split $m-k$ extras among k distinct terms.) Consequently, there are $\sum_{k=1}^{\min\{m,M\}} \binom{M}{k}\binom{m-1}{k-1}$ such choices of $\tilde{f} = (R/m)\sum_{j-1}^{m} h_j/a_{h_j}$, for a specified m (our R is also determined by m). Thus, the codelength for $\tilde{f}$, for a specified m, may be set to be $\log \sum_{k=1}^{\min\{m,M\}} \binom{M}{k}\binom{m-1}{k-1}$. Using $\sum_{k=1}^{m}\binom{M}{k} \leq (eM/m)^m$, when $m \leq M$, this log cardinality is not more than $m\log(2eM/m)$. Using $\sum_{k=1}^{M}\binom{M}{k} = 2^M$, when $m \geq M$, it is not more than $(M+m-1)\log 2 \leq m \log 4$. This completes the proof. ∎

Lemma 8.6 *The minimum over integers $m \geq 1$ of $\frac{v^2}{m} + \frac{m}{n}\log(4e\max\{\frac{M}{m}, 1\})$ is not more than*

$$\lambda^* v + \frac{\log 4eM}{n} + \frac{1}{e}\min\{\frac{1}{\sqrt{n}}, \frac{M}{n}\}$$

where

$$\lambda^* = 2\sqrt{\frac{\log 4e\max\{\frac{M}{\sqrt{n}}, 1\}}{n}}.$$

Proof:

Set $A = 4e\max\{\frac{M}{\sqrt{n}}, 1\}$. If $v = 0$, taking $m = 1$ confirms the bound. Otherwise, for $v > 0$, consider $m = \lceil v/\eta \rceil$ and choose $\eta = \sqrt{\frac{\log A}{n}}$, at which we evaluate the expression of interest. If $M/m \geq 1$, that is, if $M \geq v/\eta$, then we bound $\frac{v^2}{m} + \frac{m}{n}\log(4e\frac{M}{m})$ by $\eta v + \frac{v}{\eta n}\log(4e\frac{M}{v/\eta}) + \frac{\log 4eM}{n}$. Multiply and divide by $\min\{\sqrt{n}, M\}$ inside the logarithm and note that $\eta v + \frac{v}{\eta n}\log A$ is optimized at the chosen η to obtain the bound $2\lambda^* v + \frac{\log 4eM}{n} + \frac{v}{\eta n}\log\min\{\sqrt{n}, M\}$. The last term in this expression may be written as $\frac{\min\{\sqrt{n},M\}}{n} r \log 1/r$ with $r = \frac{v\eta}{\min\{\sqrt{n},M\}}$. Since $r\log 1/r$ is never more than $1/e$, this establishes the desired bound for $M \geq v/\eta$. If instead, at this $m = \lceil v/\eta \rceil$, we have

$M/m \leq 1$, then the expression $\frac{v^2}{m} + \frac{m}{n}\log(4e)$ is bounded by $v\eta + \frac{v}{\eta n}\frac{\log 4e}{n} + \frac{\log 4e}{n}$, which is less than $2\lambda^* v + \frac{\log 4e}{n}$. This completes the proof. ∎

The penalty expression developed in Chapter 4 requires that the complexity term be multiplied by γ for which the resulting minimum is bounded as in Lemma 8.6 with n replaced by n/γ.

Lemma 8.7 *Given any set Λ of functions with Card(Λ)=m, we denote by $\mathcal{F}_\Lambda$ as the linear span of Λ and $T\mathcal{F}_\Lambda$ as its truncated version with truncation level B'. Then the empirical L_1-covering number satisfies,*

$$\mathcal{N}(t, T\mathcal{F}_\Lambda, \|\cdot\|_{n,1}) \leq e\left(\frac{4e^2 B'}{t}\right)^{m+1} \text{ for } 0 < t \leq B',$$

where $\|f\|_{n,1}^2 = \frac{1}{n}\sum_{i=1}^n |f(X_i)|$ for any f in $T\mathcal{F}_\Lambda$.

Proof: It is clear that any function in $T\mathcal{F}_\Lambda$ is bounded by the constant B'. We use Theorem 13 of Chapter 10 in [69], which is a result based on Haussler [52], to obtain an upper bound on the empirical L_1 packing number,

$$\mathcal{M}(tB', T\mathcal{F}_\Lambda, \|\cdot\|_{n,1}) \leq e(\mathbb{D}+1)\left(\frac{4e}{t}\right)^{\mathbb{D}} \text{ for } 0 < t \leq 1, \tag{8.29}$$

where $\mathbb{D}$ is the VC-dimension of the set of all subgraphs of $T\mathcal{F}_\Lambda$ Using the fact that a covering number is less than or equal to a packing number with the same diameter, we obtain an upper bound,

$$\mathcal{N}(t, T\mathcal{F}_\Lambda, \|\cdot\|_{n,1}) \leq e(\mathbb{D}+1)\left(\frac{4eB'}{t}\right)^{\mathbb{D}} \leq e\left(\frac{4e^2 B'}{t}\right)^{\mathbb{D}} \text{ for } 0 < t \leq B', \tag{8.30}$$

since $\mathbb{D} + 1 \leq e^{\mathbb{D}}$.

Also, it the VC-dimension of the set of all subgraphs of $T\mathcal{F}_\Lambda$ is not more than the VC-dimension of the set of subgraphs of $\mathcal{F}_\Lambda$, which is equal to $m+1$, e.g., by Theorem 9.5 in [51]. Thus $\mathbb{D} \leq m+1$. This provides the desired conclusion. ∎

Proof of Theorem 4.2, third conclusion. We are to show the validity of an ℓ_1 penalty $\|\beta\|_{1,a}$ with weights $a_h \geq \lambda_2 \varepsilon_{h,S,2} + \lambda_1 \varepsilon_{h,\tilde{\mathcal{H}}_1}$, where $\varepsilon_{h,\tilde{\mathcal{H}}_1}$ and $\varepsilon_{h,S,2}$ are the distances of h from a subset $\tilde{\mathcal{H}}_1$ of size M_1 and from the linear span S of a subset $\tilde{\mathcal{H}}_2$ of size m_0, using the empirical L_1 and L_2 norms, respectively, on the $2n$ points of $\underline{X}, \underline{X}'$. Here $\lambda_2 = 2\sqrt{\frac{2\gamma \log(eM_1 c_n)}{n}}$ and $\lambda = 8B'$ with c_n as specified. [Accordingly, M_1 is typically chosen to be much larger than m_0 so as to make $\varepsilon_{h,\tilde{\mathcal{H}}_1}$ small enough that the behavior of the allowed a_h is governed by the $\lambda_2 \varepsilon_{h,S,2}$ term.] A key tool in the proof is the refined approximation property in Lemma 8.4, case 4, above. It will be used in constructing our variable-complexity variable-distortion cover of $\mathcal{F}$, which is the linear span of all of $\mathcal{H}$.

The heart of the idea in building our cover of $\mathcal{F}$ is to consider the subclasses $\mathcal{F}_{m_0,m_1}$ which consist of union of linear spans of choices of $m = m_0 + m_1$ functions, where the union is over all subsets of size m_1 out of M_1 from $\tilde{\mathcal{H}}_1$ together with all m_0 functions from $\tilde{\mathcal{H}}_2$. Improved m term approximation properties hold in this subclass for approximating functions in $\mathcal{F}$ with error expressed through weighted ℓ_1 norms of their coefficients. However, the approximations we use have more general sorts of coefficients than before, so we need an additional step to get to suitable covers. For each such selection of terms, we appeal to the results of Lemma 8.7 concerning the optimal empirical L_1 cover of the set of all truncated linear combinations at a particular precision $t_{m,n}$. Taking the union of these covers for all subsets of size m_1 from $\tilde{\mathcal{H}}_1$ we have a set $\tilde{\mathcal{F}}_{m_0,m_1}$ and our $\tilde{\mathcal{F}}$ is their union for $1 \leq m_1 \leq M_1$.

We will take advantage of the fact that the $\mathcal{F}_{m_0,m_1}$ are nested in m_1. Indeed, if due to repeats, a linear combination constructed to use up to m_1 functions from $\tilde{\mathcal{H}}_1$ actually uses only $K < m_1$ then one may arbitrarily pick $m_1 - K$ other terms from $\tilde{\mathcal{H}}_1$ and assign them

0 coefficients. Accordingly $\tilde{\mathcal{F}}_{m_0,m_1}$ remains a cover at precision $t_{m,n}$ for such truncated linear combinations using less than or equal to m_1 terms from $\tilde{\mathcal{H}}_1$.

First we construct approximations using the m_0 terms from $\mathcal{H}_2$ and an arbitrary number m_1 of terms selected from the whole library $\mathcal{H}$. As in Lemma 8.4, case 4, let Π denote the operation of projection onto the linear span S of the given subset of $\mathcal{H}$ of size m_0, now using the empirical L_2 norm on $\underline{X}, \underline{X}'$. Let f be in $\mathcal{F}_{\mathcal{H}}$ and consider a representation $f = f_\beta$ of it with $f_\beta(x) = \sum_{h\in\mathcal{H}} \beta_h h(x)$. Consider the function $g = f - \Pi f$. It has a representation $g(x) = \sum_{h\in\mathcal{H}} \beta_h[h(x) - \Pi h(x)]$ as a linear combination of terms orthogonal to members of S. For $m \geq m_0$, we form an m term approximation of f_m to f. First when $m = m_0$ we let $f_{m_0} = \Pi f$ and then for $m = m_1 + m_0$ with $m_1 \geq 1$ we let $f_m = \Pi f + g_{m_1}$, where g_{m_1} is an m_1 term approximation to g of the form $g_{m_1}(x) = (v/m_1)\sum_{k=1}^{m_1}[h_k(x) - \Pi h_k(x)]/\varepsilon_{h_k,2}$, where $\varepsilon_{h,2} = \|h - \Pi h\|_{2n}$ is the distance of h from S, and $v = \sum_h \beta_h \varepsilon_{h,2}$. Here $h_1, h_2, \ldots, h_{m_1}$, with repeats allowed, is a selection of functions from $\mathcal{H}$ which will be selected depending on f.

While bounding the accuracy of f_m as an approximation of f, we will need at the same time to make sure that we have some control on its coefficients, or there will be difficulties when we switch from h_k in $\mathcal{H}$ to its representer $\tilde{h}_k$ in $\tilde{\mathcal{H}}_1$. With precision $\varepsilon_{h,1} = \min_{\tilde{h}\in\tilde{\mathcal{H}}_1}\{\|h - \tilde{h}\|_{2n,1}\}$, the relevant precision weighted variation of our approximation, which we have called V_{m,ε_1} now takes the form $V_{m,\varepsilon_1} = (v/m_1)\sum_{k=1}^{m_1} \varepsilon_{h_k,1}/\varepsilon_{h_k,2}$. This variation we need to control looks ugly, but conveniently it has a nice expectation, namely $\|\beta\|_{1,\varepsilon_1}$, with respect to a distribution on h_k in which h occurs with probability $\beta_h\varepsilon_{h,2}/v$.

We consider the following expression

$$\|Y - f_m\|_n^2 + \|f - f_m\|_{\underline{X}'}^2 + \lambda_1[V_{m,\varepsilon_1} + t_{m,n}] + \frac{\gamma L_n(\tilde{f}_m)}{n}, \tag{8.31}$$

analogous to expression (4.12) but now with the new form of approximation f_m and with the addition of a term $\lambda_1[V_{m,\varepsilon_1}+t_{m,n}]$ we need to be small in controlling the accuracy of the cover. Again to facilitate the approximation properties, the f_m is not yet restricted to be in our cover, as it is allowed to use arbitrary h in $\mathcal{H}$, and, for now, we allow arbitrarily large m_1. As in Lemma 8.4, Case 4, picking $h_1, \ldots, h_{m_1}$ independently, with each h having probability $\beta_h \epsilon_{h,2}/v$, we obtain that the expected value of expression (8.31) and hence its value for some such f_m is not more than

$$\|Y - f\|_n^2 + \frac{2\|\beta\|_{1,\varepsilon_2}^2}{m_1} + \lambda_1[\|\beta\|_{1,\varepsilon_1} + t_m] + \gamma \frac{L_{n,m_1}}{n}. \tag{8.32}$$

The L_{n,m_1} is used to bound the complexity $L_n(\tilde{f}_m)$ of functions $\tilde{f}_m$ in our cover which use not more than m_1 terms selected from $\tilde{\mathcal{H}}_1$. This complexity will depend on the precision $t_{m,n}$. As will soon be explained, the best $t_{m,n}$ for this bound satisfies $\lambda_1 t_{m,n} \leq \gamma(m+1)/n$, and at this best $t_{m,n}$,

$$L_{n,m_1} \leq (m_1 + 1) \log M_1 + (m_1 + m_0 + 1) \log c_n.$$

Now picking m_1 depending of f to equal $\lceil \|\beta\|_{1,\varepsilon_2}/\eta \rceil$ with $\eta = \sqrt{\frac{\gamma \log(eM_1 c_n)}{2n}}$ and setting λ_2 to be at least the specified value, expression (8.31) becomes not more than

$$\|Y - f\|_n^2 + \lambda_2 \|\beta\|_{1,\varepsilon_2} + \lambda_1 \|\beta\|_{1,\varepsilon_1} + \gamma \frac{2 \log M_1 + (m_0 + 2) \log(c_n e)}{n}, \tag{8.33}$$

which is our penalized least squares criterion, including the indicated adjustment.

We truncate the f_m and f in (8.31) and show that with certain replacements of Tf_m, ultimately leading to the representer of f, we have a suitable lower bound. We first replace f_m by a $\tilde{f}_{m,\text{temp}}$, replacing certain of the occurrences of h_k with their representers $\tilde{h}_k$. Indeed, in $f_m = \Pi f + (v/m_1) \sum_{k=1}^{m_1} [h_k - \Pi h_k]/\varepsilon_{h,2}$, the projected pieces Πh_k are already in

the span S of m_0 given functions, so we let $\tilde{f}_{m,\text{temp}} = \Pi f + (v/m_1) \sum_{k=1}^{m_1} [\tilde{h}_k - \Pi h_k]/\varepsilon_{h_k,2}$. So then their difference $\tilde{f}_{m,\text{temp}}(x) - f_m(x) = (v/m_1) \sum_{k=1}^{m_1} [\tilde{h}_k(x) - h_k(x)]/\varepsilon_{h_k,2}$ has empirical L_1 norm bounded by our updated expression V_{m,ϵ_1}. Hence $\|T\tilde{f}_{m,\text{temp}} - f_m\|_{2n,1} \leq V_{m,\epsilon_1}$. In forming $T\tilde{f}_{m,\text{temp}}$ not more than m_1 functions $\tilde{h}_k$ are selected from $\tilde{\mathcal{H}}_1$ and linearly combined with the m_0 functions from $\tilde{\mathcal{H}}_2$, and then thresholded to level B'. Accordingly, there is a representer $\tilde{f}_m$ in $\tilde{\mathcal{F}}_{m_0,\min\{m_1,M_1\}}$, which we may also arrange to be bounded by B', for which $\|T\tilde{f}_{m,\text{temp}} - \tilde{f}_m\|_{2n,1} \leq t_m$, where t_m is a precision to be specified. Consequently, the empirical L_1 distance between f_m and $\tilde{f}_m$ is not more than $V_{m,\varepsilon_1} + t_m$.

Then the argument proceeds as before in Chapter 4, in the development from expression (4.12) to expression (4.15), to obtain that expression (8.31) is at least

$$\|Y - T\tilde{f}_m\|_n^2 + \|Tf - T\tilde{f}_m\|_{\underline{X}'}^2 + \frac{\gamma L(\tilde{f}_m)}{n} - \frac{\text{Tail}_1 + \text{Tail}_2}{n} + (\lambda_1 - 8B')[V_{m,\varepsilon_1} + t_m], \tag{8.34}$$

where the $[V_{m,\varepsilon_1} + t_m]$ term may be dropped for $\lambda_1 \geq 8B'$.

Thus, for this case, reasoning as before, we have the ingredients for expression (8.33) to be a valid penalized squared error criterion, exceeding the corresponding expression here, for satisfaction of the conditions of our theory.

It remains to present the bound for the cardinality of $\tilde{\mathcal{F}}_{n,m_0,m_1}$ taken to be an empirical L_1 cover of $T\mathcal{F}_{m_0,m_1}$ with precision $t = t_{m,n}$, and to use it to verify our complexity expression and its bound. Here $T\mathcal{F}_{m_0,m_1}$ is the collection of truncated linear combinations with not more than m_1 terms from $\tilde{\mathcal{H}}_1$ and all of the m_0 terms from $\tilde{\mathcal{H}}_2$. Through padding linear combinations that use fewer terms with additional zero coefficient terms, as previously explained, this can be thought of as the collection of truncated linear combinations with exactly $\min\{m_1, M_1\}$ terms. In accordance with Lemma 8.7, for $t \leq B'$, its cardinality is

not more than

$$\binom{M_1}{m_1}\left(\frac{4B'e^2}{t}\right)^{m_0+m_1+1}, \tag{8.35}$$

with the understanding that if $m_1 \geq M_1$ then $\binom{M_1}{m_1} = 1$, and, if desired, in the exponent, m_1 may be replaced by $\min\{m_1, M_1\}$. Here $\binom{M_1}{m_1}$ is the number of subsets of size m_1 out of the M_1 functions in $\tilde{\mathcal{H}}_1$ and the other factor, as explained in the appendix, is the cardinality bound for the empirical L_1 cover of the set of thresholded linear combinations for any m specified functions. The logarithm of expression (8.35) plus $\log M_1$ provides a valid variable-complexity assignment for functions in $\tilde{\mathcal{F}} = \cup_{1\leq m_1\leq M_1}\tilde{\mathcal{F}}_{n,m_0,m_1}$ satisfying Condition (S). This complexity is the sum of three parts, $\log M_1$ corresponding to the description of $m_1 \leq M_1$, plus $\log\binom{M_1}{m_1}$ for the description of the subsets of terms, plus $(m+1)\log(4B'e^2/t_{m,n})$ the bound on the log-cardinality of the cover of thresholded linear combinations of such terms.

Taking γ/n times the log cardinality and adding the term $\lambda_1 t_{m,n}$, we have the part of the penalty expression that involves the choice of the precision $t_{m,n}$. Accordingly, the optimal $t_{m,n}$ is seen to be $\min\{\frac{\gamma(m+1)}{\lambda_1 n}, B'\}$, at which the term $\lambda_1 t_{m,n}$ is not more than $\gamma(m+1)/n$. Consequently, using $\lambda_1 \geq 8B'$, with this $t_{m,n}$ a valid complexity assignment is

$$L_{n,m} = \log\binom{M_1}{m_1} + (m+1)\log c_{n,m} + \log M_1, \tag{8.36}$$

where $c_{n,m} = 4e^2\max\{1, \frac{8(B')^2}{\gamma}\frac{n}{m+1}\}$, which is not more than $c_n = c_{n,m_0+1} = 4e^2\max\{1, nc'\}$ for $m_1 \geq 1$, where $c' = 8B'^2/[\gamma(m_0+2)]$. Interpret $\binom{M_1}{m_1} = 1$ for $m_1 > M_1$. Then replacing with $c_{n,m}$ with c_n, we have that $\log\binom{M_1}{m_1} + (m+1)\log c_n + \log M_1$ is an upper bound on $L_{n,\min\{m_1,M_1\}}$ for all $m_1 \geq 1$. Since $\log\binom{M_1}{m_1} \leq m_1 \log M_1$ this completes the verification of the form of the complexity bound used above, and, accordingly, it completes the proof of Theorem 4.2.

8.3 Lemmas for Chapter 6

Real interpolation spaces [11]

$$\mathcal{B}_p = [L_2(P), \mathcal{L}_{1,\mathcal{H}}]_{\theta,\infty},$$

with $0 \leq \theta \leq 1$ and p defined by $1/p = (1+\theta)/2$ consist of all functions f^* which satisfy

$$K(f^*, t) \leq Ct^\theta,$$

where $K(f^*, t) = \inf_{f \in \mathcal{L}_{1,\mathcal{H}}} \{\|f^* - f\|_{L_2(P)} + tV(f)\}$ is the so-called K-function. The smallest C such that the above holds is the norm of f^* in this interpolation space.

Lemma 8.8 *The interpolation space $\mathcal{B}_p$ is equivalent to the space $\mathcal{B}_{1,p}^{res}$ defined in Chapter 6 for $1 \leq p \leq 2$. Also*

$$0.6\left(C_{1,p}(f^*)\right)^{1/p} \leq \|f^*\|_{\mathcal{B}_p} \leq 2\left(C_{1,p}(f^*)\right)^{1/p}$$

and

$$0.38\|f^*\|_{\mathcal{B}_p}^p \leq C_{1,p}(f^*) \leq 2\|f^*\|_{\mathcal{B}_p}^p$$

Proof: First, if $f^* \in \mathcal{B}_{1,p}^{res}$, according to the definition, there exists a function $f \in \mathcal{L}_{1,\mathcal{H}}$, such that, $\|f^* - f\|^2 + \lambda V(f) \leq C_{1,p}(f^*)\lambda^{2-p}$ for any $\lambda > 0$. Hence, $\|f^* - f\|^2 \leq C_{1,p}(f^*)\lambda^{2-p}$ and $V(f) \leq C_{1,p}(f^*)\lambda^{1-p}$. Then the K-function satisfies

$$K(f^*, t) \leq \|f^* - f\| + tV(f) \leq \sqrt{C_{1,p}(f^*)}\lambda^{1-p/2} + C_{1,p}(f^*)\lambda^{1-p}t.$$

Since the above inequality is true for all $\lambda > 0$, it also holds with the infimum over λ on the right side, which yields

$$(C_{1,p}(f^*))^{1/p} \left(\frac{2p-2}{2-p}\right)^{2/p-2} \left(\frac{p}{2-p}\right) t^{2/p-1}.$$

Thus because $2/p - 1 = \theta$, this f^* is indeed in $\mathcal{B}_p$ and the norm $\|f^*\|_{\mathcal{B}_p}$ is less than or equal to $(C_{1,p}(f^*))^{1/p} \left(\frac{2p-2}{2-p}\right)^{2/p-2} \left(\frac{p}{2-p}\right)$.

Next, if $f^* \in \mathcal{B}_p$, using the same argument, we obtain that

$$R^1(f^*, \lambda) \leq \|f^*\|^2_{\mathcal{B}_p} t^{2\theta} + \|f^*\|_{\mathcal{B}_p} t^{\theta-1} \lambda$$

for all $t > 0$. Minimizing t on the right side and using $\theta = 2/p - 1$ yields

$$R^1(f^*, \lambda) \leq \|f^*\|^p_{\mathcal{B}_p} \left(\frac{p-1}{2-p}\right)^{1-p} \left(\frac{1}{2-p}\right) \lambda^{2-p},$$

which implies that f^* is in $\mathcal{B}^{res}_{1,p}$ and $C_{1,p}(f^*) \leq \|f^*\|^p_{\mathcal{B}_p} \left(\frac{p-1}{2-p}\right)^{1-p} \left(\frac{1}{2-p}\right)$.

Combining two bounds together yields

$$\frac{(2-p)^{2-p}}{(2p-2)^{2-2p} p^p} \|f^*\|^p_{\mathcal{B}_p} \leq C_{1,p}(f^*) \leq \frac{(p-1)^{1-p}}{(2-p)^{2-p}} \|f^*\|^p_{\mathcal{B}_p}.$$

Likewise, we can bound $\|f^*\|_{\mathcal{B}_p}$ using $(C_{1,p}(f^*))^{1/p}$. Extremizing the two coefficients in the upper and lower bounds produces the statement. ∎

Lemma 8.9 *The interpolation space $\mathcal{B}^{res}_{1,r,p}$ is equivalent to the space $\mathcal{B}^{res}_{1,p}$ for $1 \leq p \leq 2$ and any $r > 0$.*

Proof: First, if $f^* \in \mathcal{B}^{res}_{1,p}$, according to the definition, there exists a v, such that, $App(f^*, v) + \lambda v \leq C_{1,p}(f^*)\lambda^{2-p}$ for any $\lambda > 0$. Hence, $App(f^*, v) \leq C_{1,p}(f^*)\lambda^{2-p}$ and $v \leq C_{1,p}(f^*)\lambda^{1-p}$.

Then the function $R^{1,r}(f^*, t)$ is not more than

$$App(f^*, v) + tv^r \leq C_{1,p}(f^*)\lambda^{2-p} + C^r_{1,p}(f^*)\lambda^{(1-p)r}t.$$

Minimizing over λ on the right side yields

$$R^{1,r}(f^*, t) \leq c\, t^{\theta},$$

where c doesn't depend on t and $\theta = (2-p)/(rp-p+2-r)$. Therefore, the f^* is indeed in $\mathcal{B}^{res}_{1,r,p}$.

The other direction can be proved by the same argument. We omit it here. ∎

Bibliography

[1] H. Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247, 1969.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In P.N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, 1972.

[3] Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117:467–493, 2000.

[4] Y. Baraud. Model selection for regression on a random design. *ESAIM: Probability and Statistics*, 6:127–146, 2002.

[5] A. R. Barron. Approximation and estimation bounds for artificial neural networks. In L. Valiant, editor, *Computational Learning Theory: Proceedings of the Fourth Annual ACM Workshop*, pages 243–249. Morgan Kaufmann Publishers, 1991.

[6] A. R. Barron. Complexity regularization with applications to artificial neural networks. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 561–576, 1991.

[7] A. R. Barron. Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transactions on Information Theory*, 39:930–944, 1993.

[8] A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 114:113–143, 1994.

[9] A. R. Barron. Information-theoretic characterization of bayes performance and the choice of priors in parametric and nonparametric problems. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 6*, pages 27–52. Oxford University Press, 1998.

[10] A. R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

[11] A. R. Barron, Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Approximation and learning by greedy algorithms. *Annals of Statistics*, 36(1):64–94, 2008.

[12] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(1034-1054), 1991.

[13] A. R. Barron, J. Q. Li, C. Huang, and Xi Luo. MDL principle, enalized likelihood, and statistical risk. In *Feschrift for Jorma Rissanen*, 2008.

[14] A. R. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44:2734–2760, 1998.

[15] A. R. Barron and X. Xiao. Discussion of "multivariate adaptive regression splines" by J. H. Friedman. *Annals of Statistics*, 19:67–82, 1991.

[16] Andrew Barron and Xi Luo. Adaptive annealing. In *Proceedings of the Allerton Conference on Communications, Computation, and Control*, 2007.

[17] T. Berger. *Rate-distortion theory: A mathematical basis for data compression.* Prentice-Hall, Englewood Cliffs, NJ, 1971.

[18] L Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.

[19] L Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4:329–375, 1998.

[20] L Birgé and P. Massart. Gaussian model selection. *Journal of European Math. Society*, 3:203–268, 2001.

[21] L Birgé and P. Massart. A generalized C_p criterion for gaussian model selection. Technical report, Prépublication 647, Laboratoire de Probabilités et Modèles Aléatoires, Univ. Paris 6 and Paris 7, 2001. Available at www.proba.jussieu.fr/mathdoc/preprints/index.html.

[22] S. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[23] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Inc, Monterey, Calif., U.S.A., 1984.

[24] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35:1674–1697, 2007.

[25] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Sparse density estimation with ℓ_1 penalties. In N. Behouty and C. Gentile, editors, *Proceedings of 20th Annual Conference on Learning Theory, COLT 2007*, pages 530–543. Springer, New York, 2007.

[26] Emmanuel J. Candès and D. L. Donoho. Ridgelets: a key to higher-dimensional intermittency? *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 357(1760):2495–2509, 1999.

[27] Emmanuel J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise-C^2 singularities. *Comm. Pure Applied Math.*, 57:219–266, 2002.

[28] O Catoni. *Statistical Learning Theory and Stochastic Optimization.* Ecole d'Eté de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics. Springer, 2004.

[29] G. H. L. Cheang. *Approximation and Estimation Bounds for Two hidden-layer Sigmoidal Networks.* PhD thesis, Yale University, 1996.

[30] G. H. L. Cheang and A. R. Barron. Estimation with two hidden layer neural nets. In J. S. Boswell, editor, *Proceedings of the 1999 IJCNN*, 1999.

[31] G. H. L. Cheang and A. R. Barron. Penalized least squares, model selection, convex hull classes and penalized least squares, model selection, convex hull classes and neural nets. In M. Verleysen, editor, *Proceedings of the 9th ESANN*, pages 371–376, 2001.

[32] S. S. Chen. *Basis Pursuit.* PhD thesis, Stanford University, 1995.

[33] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *Society for Industrial and Applied Mathematics*, 20(1):33–61, 1998.

[34] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 2006.

[35] D. D. Cox and F. O'Sullivan. Asymptotic analysis of penalized likelihood-type estimators. *Annals of Statistics*, 18:1676–1695, 1990.

[36] D. D. Cox and F. O'Sullivan. Penalized likelihood-type estimators for generalized nonparametric regression. *Journal of Multivariate Analysis*, 56(2):185–206, 1996.

[37] Cecil C. Craig. On the tchebyshef inequality of bernstein. *Annals of Math. Statistics*, 4:94–102, 1933.

[38] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematics Society*, 39:1–49, 2001.

[39] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math Control Signal System*, 2:303–314, 1989.

[40] G. M. de Montricher, R. A. Tapia, and J. R. Thompson. Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *Annals of Statistics*, 3:1329–1348, 1975.

[41] R. A. DeVore and V. N. Temlyakov. Some remarks on greedy algorithms. *Advances in Computational Mathematics*, 5(1):173–187, 1996.

[42] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[43] J. Friedman. Tutorial: Getting started with mart in r, 2002.

[44] J. Friedman, T. Hastie, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

[45] J. H. Friedman. Multivariate additive regression splines. *Annals of Statistics*, 19:1–66, 1991.

[46] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[47] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38:367–378, 2002.

[48] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of American Statistical Association*, 76(376):817–823, 1981.

[49] I. J. Good and R. A. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58:255–277, 1971.

[50] I. J. Good and R. A. Gaskins. Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of American Statistical Association*, 75:42–73, 1980.

[51] L. Györfy, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression.* Springer-Verlag, 2002.

[52] David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69:217–232, 1995.

[53] P. J. Huber. Projection pursuit. *Annals of Statistics*, 13:435–525, 1985.

[54] L. K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics*, 20:608–613, 1992.

[55] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Annals of Statistics*, 28:681–712, 2000.

[56] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Journal of Information of Computation*, 132(1):1–64, 1997.

[57] V. Koltchinskii and D. Panchenko. Complexities of convex combinations and bounding the generalization error in classification. *Annals of Statistics*, 33(4):1455–1496, 2005.

[58] S. V. Konyagin and V. N. Temlyakov. Rates of convergence of pure greedy algorithms. *East Journal of Approximation*, 5:493–499, 1999.

[59] W. S. Lee, P. L. Bartlett, and R. C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42:2118–2132, 1996.

[60] K. C. Li. Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: discrete index set. *Annals of Statistics*, 15:958–975, 1987.

[61] E. D. Livshitz and V. N. Temlyakov. Two lower estimates in greedy approximation. *Construction Approximation*, 19:509–524, 2003.

[62] R. Makovoz. Random approximants and neural networks. *Journal of Approximation Theory*, 85:98–109, 1996.

[63] S. Mallat. Matching pursuits with time-frequency dictionary. *IEEE Transactions on Signal Processing*, 12(3):3397–3415, 1993.

[64] C. L. Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973.

[65] A. Nemirovski. *Topics in Non-parametric Statistics*. Ecole d'été de probabilités de Saint-Flour XXVIII-1998. Lecture Notes in Mathematics, no. 1738. Springer, New York, 2000.

[66] A. S. Nemirovski, B. T. Polyak, and A. B. Tsybakov. Rate of convergence of non-parametric estimates of maximum likelihood type. *Problems in Information Transmission*, 21:258–272, 1985.

[67] M. R. Osborne, B. Presnell, and B. A. Turlach. Knot selection for regression splines via the lasso. *Computing Science and Statistics*, 30:44–49, 1998.

[68] M. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

[69] David Pollard. Asymptopia, Chapter 10, 2007.

[70] R. Prony. Essai experimental et analytique. *Paris J. de l'Ecole, Polytechnique*, 1:24–76, 1795.

[71] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[72] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11(2):416–431, 1983.

[73] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representation by error propagation. In David E. Rumelhart and James A. McClelland, editors, *Parallel distributed processing: explorations in the microstructure of cognition*, volume 1. MIT Press, 1986.

[74] G. Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[75] X. Shen. On the method of penalization. *Statistica Sinca*, 8:337–357, 1998.

[76] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68:45–54, 1981.

[77] B Silverman. On the estimation of probability function by the maximum penalized likelihood method. *Annals of Statistics*, 10:795–810, 1982.

[78] Sandra E. Sinisi and Mark J. van der Laan. Deletion/substitution/addition algorithm in learning with applications in genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

[79] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society Series B*, 58:267–288, 1996.

[80] A. B. Tsybakov. Optimal rates of aggregation. In *Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines. Lecture Notes in Artificial Intel ligence 2777*, pages 303–313, Heidelberg, 2003. Springer-Verlag.

[81] V. N. Vapnik. *The nature of statistical learning theory.* Springer, New York, 1995.

[82] G Wahba. *Spline Models for Observational Data.* SIAM, Philadelphia, 1990.

[83] Y. Yang. Model selection for non-parametric regression. *Statistica Sinca*, 9:475–499, 1999.

[84] Y. Yang. Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, 74:135–161, 2000.

[85] Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10:25–47, 2004.

[86] Y. Yang and A. R. Barron. An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44:95–116, 1998.

[87] Y. Yang and A. R. Barron. Information theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27:1564–1599, 1999.

[88] B. Yu and T. Zhang. Boosting with early stopping: Covergence and consistency. *Annals of Statistics*, 33:1538–1579, 2005.

[89] Tong Zhang. Some sharp performance bounds for least squares regression with L_1 regularization. Technical report, Rutgers University, 2007.

[90] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.