Abstract

Adaptation in Estimation and Annealing

Sabyasachi Chatterjee

2014

We study general penalized log likelihood procedures and propose Information Theoretic conditions required of the penalty to obtain adaptive risk bounds. We demonstrate our conditions are natural and are satisfied in some canonical problems in statistics. We then investigate whether penalties are always required to obtain adaptation in the rates of estimation for M estimators. We show that the plain least squares estimators, without any penalization, in certain canonical shape-constrained regression problems indeed adapt to certain parametric complexities in the parameter space. We attempt to give a geometric characterization of this adaptation behaviour. We then move on to the important issue of computation. Motivated by a classical function estimation problem in non-parametric statistics, we study the possibility of a randomized algorithm, inspired by Simulated Annealing, being able to optimize multimodal functions in high dimensions. We explore the performance of this algorithm in low dimensions and explain the challenges faced in high dimensions. We provide myriads of possible routes for solving the statistically relevant optimization problem with the hope of encouraging further research in this direction.

Adaptation in Estimation and Annealing

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Sabyasachi Chatterjee

Dissertation Director: Andrew Barron

December 2014

Acknowledgements

I am deeply indebted to my advisor Professor Andrew Barron for my intellectual and professional growth in the field of Statistics and other related fields. I have benefited greatly from having spent numerous discussion sessions with him. He has been kind enough to share his vast knowledge of the fields of Statistics and Information Theory and I would like to think I have inculcated in myself some of his philosophies and style of advancing research in our field. I have also been influenced greatly by Professor David Pollard. He, again, has been kind enough to listen to some of my research efforts and have given me very valuable inputs and suggestions. I have also learnt a lot from the courses he has taught me in Probability and Statistics and his constant efforts to simplify exposition of complicated mathematical arguments is something I will attempt to practice in the future. I also want to thank Prof. Harrison Zhou, Prof. Joseph Chang, Prof. Sekhar Tatikonda, Prof. Sahand Negahban and Prof. Mokshay Madiman for encouraging me and lending a helpful ear to me whenever I wanted to talk to them.

I have had some very good friends here at Yale without whom life would not have been the same here. I would not be able to mention all of them here. However, I would like to mention some of the people with whom I have perhaps spent most of my time. I would like to thank Adityanand Guntuboyina for

being an exemplary role model and not the least, for introducing the problem of adaptation in monotone regression to me. I also thank Antony Joseph and Anup Rao for being great friends and tolerating me throughout my endless rants and complaints. I also thank all these three people for spending numerous hours with me on the squash and tennis courts and the various restaurants around the great town of New Haven. I thank Rajarshi Mukherjee and Wan Chen Lee for always being ready to give me company whenever I travelled to Boston and for all the wonderful game nights and the travels we have had together. Finally, I thank Dolon Bhattacharyya for being with me throughout and travelling this journey together.

Dedicated to my parents

Preface

Statistical methodologies have to be judged on two parameters. Firstly, their risk and predictive properties and secondly the potential to actually implement these methodologies using reasonable computational resources and time. In my dissertation I have focussed on both these issues, with more emphasis on the risk properties of estimators achieved by minimizing certain objective functions arising from data. I have worked on three topics and I have included them as the next three chapters in this dissertation.

Chapter 1 is on the Information Theoretic treatment of Penalized Likelihoods and derivation of general adaptive risk bounds for the penalized likelihood procedure. This is joint work with my advisor Andrew Barron. We develop a general framework for studying the penalized likelihood estimator and show that traditional penalties such as the ℓ_0 and ℓ_1 penalties in canonical statistical problems fit our framework.

Chapter 2 is on the study of the least squares estimator in certain shape-constrained regression problems. This work is done jointly with Adityanand Guntuboyina and Bodhisattva Sen. We develop improved risk bounds in these shape constrained problems and show adaptivity of the least squares estimator for certain parametric simplicities in the parameter space.

Chapter 3 deals with computational issues and considers a multimodal op-

timization problem in moderate to high dimensions motivated by a classical function estimation problem. This is again, jointly done with my advisor Andrew Barron. The optimization problem considered is very hard to solve because of multimodality in high dimensions. We explore an algorithm here which tries to improve the traditional Simulated Annealing idea. The main difference with Simulated Annealing and our algorithm is that while the Simulated Annealing algorithm makes transition steps based on the idea of an invariant or stationary distribution, we are inspired from the theory of Diffusion processes and attempt to solve a partial differential equation known as the Fokker Plank equation for our transition moves. While we do not claim we have been successful completely in our efforts in this direction, nevertheless it is an interesting idea and deserved to be pursued. It may look like these three topics are distinct entities but as I reveal now, there have been two main themes or threads which has driven the research in these three topics.

The first connecting thread concerns problems where we have a dictionary of functions or candidates to linearly combine and form estimates in a density estimation or regression setting. The work in Chapter 1 reveals the risk properties of choosing such estimators by penalizing for the complexity of the linear combination, which could be the number of non-zero coefficients needed to describe the parameter or the sum of absolute values of all the coefficients of the dictionary functions. It is nice to know the good risk properties of such estimators but we also need to compute them. If the number of dictionary elements is large but manageable such as a couple of million or so, as can happen in high-dimensional linear regression for example, then these estimators can be computed by looping over all elements of the dictionary by certain greedy algorithms. In the case when the dictionary elements are indeed too large to be looped over, as it happens in genuine non-parametric function estimation

problems in high dimensions, one needs flexible algorithms with some theoretical guarantees. The greedy algorithm that can be used in this case requires us to solve a non-convex multimodal optimization problem in each step which is precisely the motivation for our work described in Chapter 3. In this way we see that Chapter 1 and 3 are indeed two parts of one single puzzle.

The second theme which runs through my dissertation is the notion of constructing adaptive risk bounds. Chapter 1 reveals what are the conditions needed for a penalty function so that the resulting penalized likelihood estimator has risk upper bounded by a Information Theoretic quantity which is the minimum expected coding redundancy per symbol. This also permits us to prove that the risk bounds are adaptive, that is, the estimator adapts to the complexity of the parameter. The risk will be better for simpler elements of the parameter space. In Chapter 2 we show a very interesting fact. For parameter spaces which are polyhedral cones in euclidean space, the least squares estimator, without any need for penalization, adapts to certain parametric complexities of the parameter space. This applies to problems such as Monotone and Convex regression. This fact seems unique to shape-constrained problems and according to my opinion, is not very well understood yet. It seems the geometry of these particular cones plays a role and the truth being in a low dimensional face of these cones is advantageous as far as risk bounds for the least squares estimator is concerned. The fundamental question of what type of penalties are needed for adaptive estimation and when they may not be needed, drives the research that I describe in Chapters 1 and 2.

Contents

List of Figures

Chapter 1

Information Theory of Penalized Likelihoods

We extend the correspondence between two-stage coding procedures in data compression and penalized likelihood procedures in statistical estimation. Traditionally, this had required restriction to countable parameter spaces. We show how to extend this correspondence in the uncountable parameter case. Leveraging the description length interpretations of penalized likelihood procedures we devise new techniques to derive adaptive risk bounds of such procedures. We show that the existence of certain countable subsets of the parameter space implies adaptive risk bounds and thus our theory is quite general. We apply our techniques to illustrate risk bounds for ℓ_1 type penalized procedures in canonical high dimensional statistical problems such as linear regression and Gaussian graphical Models. In the linear regression problem, we also demonstrate how the traditional l_0 penalty plus lower order terms has a two stage description length interpretation and present risk bounds for this penalized likelihood procedure.

1.1 Introduction

There are close connections between good data compression and good estimation in statistical settings. Shannon's recipe for finding the minimum expected codelength when we know the data generating distribution shows the correspondence between probability distributions on data and optimal codelengths on the sample space. Also, Kraft's inequality stating that for every probability mass function there exists a prefix free code with lengths expressible as minus log probability gives an operational meaning to probability. The Kraft inequality allows one to think of prefix free codes and probabilities interchangeably. The MDL principle has further developed this connection by considering the case where we do not necessarily know the data generating distribution. In the MDL framework codes are always meant to be prefix free. In this framework one considers a family of codes or equivalently a set of probability distributions, possibly indexed by a parameter space Θ. The idea in one shot data compression is to compress the observed data sequence well. But for statistical purposes, we also want to devise a coding or estimation strategy based on the observed data that should compress or predict well for future data assumed to be arising from the same generating distribution.

A fundamental concept in the MDL philosophy is that of universal coding or modelling. The aim of universal coding or modelling is to find a single code that allows us to compress data almost as well as the best code in our class of codes Θ either in expectation or high probability with respect to the generation of the data X. This universal distribution can be constructed mainly in four different ways as described in [4]. These four ways can be categorized as Two-stage codes, Bayes mixture codes, Predictive codes and Normalized Maximum Likelihood codes. Penalized minus log likelihood on uncountable spaces Θ

provides another category of universal codes as we develop here, by relating it to Two-stage codes on appropriately defined countable subsets of Θ.

One of the earliest ways to build a universal code is to build what is called a two stage code [1]. The basic idea is to first devise a code or description of all the possible codes in Θ. Also for each possible code one encodes or describes the data using that code. Then one chooses the code which minimizes the sum of the two descriptions, one describing the code and the other describing data given the code. Now one can play the same game in the learning setup where now the codes are replaced by a family of probability distributions and the estimated probability distribution is the one which minimizes the sum of the two descriptions. This is indeed the penalized likelihood estimator where the penalty corresponds to the description lengths of probability distributions in our model. Traditionally, there have been various kinds of penalties that have been proposed. Firstly, penalties which penalize roughness or irregularity of the density as in [15],[22] and [21] have been considered. Secondly, penalties could be generally of the ℓ_2 type. Reproducing kernel Hilbert space penalties are championed in [23]. Statistical risk rate results for general quadratic penalties in Hilbert space settings which correspond to weighted ℓ_2 norms on coefficients in function expansions, including, in particular, Sobolev-type penalties (squared L_2 norms of derivatives) are developed in [14] and [13] based on functional analysis tools. Empirical process techniques for penalized likelihood built around metric entropy calculations are used to yield rate results for penalties designed for a wide variety of function classes in [20]. Theory related to penalized likelihood is developed for constrained maximum likelihood in nonparametric settings [19] and for minimum contrast estimators and sieves as in [11] and [12]. A general treatment of penalized likelihoods has been given in [26]. The authors there too give sufficient conditions, namely the decom-

posabiltiy and the restricted strong convexity condition on the penalty and the term corresponding to the log likelihood respectively, to obtain sharp risk bounds. For log likelihood penalization, it would be interesting to examine if there are connections between their conditions and the conditions proposed in this manuscript. General oracle type bounds in problems like high dimensional linear regression for the ℓ_1 penalty can be obtained by the method of aggregation as developed by Tsybakov and others [27]. These results have the leading constant 1 in the oracle inequalities. Our risk bounds do not achieve the leading constant 1 but can achieve a constant arbitrarily close to 1. Nevertheless, our work demonstrates the information theoretic side of penalized likelihood estimators and provides another framework to study penalized likelihood procedures. There has also been a lot of activity recently on investigating the risk rate results for ℓ_1 type penalties, including the Lasso estimator and it is indeed a daunting task to write down all the required references. A nice survey article in this topic where the major advancements and references can be found is [7].

Our treatment of penalized likelihoods extend the pattern of past MDL work. Traditionally, the statistical properties of this penalized likelihood procedure has been studied in the countable parameter space setting. Past work as in [2] and [3] shows how the expected pointwise redundancy controls the statistical risk in countable parameter spaces. For suitable penalties the performance is captured by an index of resolvability which is the minimum sum of relative entropy approximation error and the penalty relative to the sample size. Such results have been developed previously for the case that the function fits are restricted to a countable set which discretizes the parameter space, with a complexity penalty equal to an information-theoretic codelength for the discretized set as in [1], [9], [18], [16] and [17]. These estimators can also be interpreted as a maximum posterior probability estimator with the penalty

equal to a log reciprocal prior probability for the countable set. Resolvability bounds on risk have also been developed for the case that the function fits are optimized over a list of finite-dimensional families with penalty typically proportional to the dimension as in [25] and [10].

One of the main contributions of this present manuscript is to extend such risk bounds when the parameter space is uncountable maintaining the description length interpretation. The main idea here is to construct countable subsets of the parameter space Θ and leverage the results from the countable case. These subsets are constructed according to the interaction of the minus log likelihood and the penalty as is made clear in section (1.2.2). We show that the loss function we consider, is not much more than the pointwise redundancy, both in expectation and with high probability. As we would see, these risk bounds that we get also reveal the adaptation properties of these penalized likelihood procedures.

The main idea is to propose conditions on the penalty and the negative log likelihood in a general setting to derive adaptive risk bounds as long as the penalized likelihood estimator mirrors the construction of a two stage code. In a preliminary form, this idea has appeared in [2] and [3]. This manuscript lays out this general theory in more detail and then shows that our conditions are satisfied and our risk bounds are valid in canonical high-dimensional statistical problems such as linear regression and inverse covariance estimation in Gaussian models.

In section (1.2) we describe the general technique of how to relate penalized negative log likelihoods in the uncountable parameter space case to two stage description lengths on a appropriate countable subset. We also lay out the general strategy for deriving adaptive risk bounds whenever the codelength re-

lation holds. In this section, we describe the conditions needed on the penalty and the negative log likelihood which allows us to prove risk bounds. In Section (1.3) we apply our theory to the ℓ_1 penalty in linear regression case and fully illustrate different ways of verifying the conditions we need. We then present a new result on inverse covariance matrix estimation in a multivariate Normal setting which shows that our theory can handle not just location type problems but scale problems as well. In Section (1.4) we then turn our attention to the ℓ_0 penalty in the linear regression case. We devise a new way to interpret the ℓ_0 penalty times a $\log(n)/2$ factor as Kraft satisfying codelengths and leverage this interpretation to recover adaptive risk bounds. The penalties we consider in this manuscript are traditionally two of the most commonly used in statistics, namely the ℓ_0 penalty or the number of parameters times a suitable multiplier and ℓ_1 type penalties with suitable multipliers.

1.1.1 Notational Conventions

We denote a general sample space by $\mathbb{U}$ and its elements by u. In cases when the data is generated in an i.i.d fashion we set $\mathbb{U} = \mathcal{X}^n$ for some positive integer n which usually denotes the sample size. We generically take our model to be the class of densities $\{p_\theta : \theta \in \Theta\}$ with respect to some dominating measure ν. The class of densities is parametrized by a set Θ which is our parameter space.

We will also distinguish between countable and uncountable parameter spaces. Generically we denote a countable parameter space by $\tilde{\Theta}$ and an uncountable parameter space by Θ. We will also generically denote the elements of $\tilde{\Theta}$ by $\tilde{\theta}$ and elements in Θ by θ. We also consistently denote a penalty function on Θ by *pen* and a penalty function on $\tilde{\Theta}$ by V. We also measure codelengths in

nats instead of bits in this manuscript. So all the logarithms are with base e. Nevertheless, base 2 counterparts to log and exp can work as well and have bit interpretation.

1.2 General Technique

In this section, we first propose a way to relate the penalized log likelihood expression in the case when the parameter space is uncountable as akin to a two-stage codelength. Then we show how our proposed extension also helps us derive adaptive risk bounds for the penalized likelihood procedures. We introduce a terminology here which plays a key role in our discussions. Let $\tilde{\Theta}$ be a countable set and $V : \tilde{\Theta} \to \mathbb{R}_+$ be an associated complexity function on Θ. For any probability distribution with density q we denote the *sample resolvability* of q at the data point u with respect to the class of probability distributions indexed by $\tilde{\theta}$ to be the following expression

$$\min_{\tilde{\theta} \in \tilde{\Theta}} \left(\log \frac{p_\theta(u)}{p_{\tilde{\theta}}(u)} + V(\tilde{\theta}) \right).$$

The sample resolvability is the minimum of a sum of two terms. The first term is an approximation term of the log likelihood ratio between q and a member of the class. The second term is the complexity of a member of the class.

1.2.1 Codelength validity

First, let us describe the two-stage code in the case when we have a countable parameter space. Let the parameter space $\tilde{\Theta}$ be countable, and $V : \tilde{\Theta} \to \mathbb{R}_+$ be a penalty function on $\tilde{\Theta}$ satisfying Kraft's inequality $\sum_{\tilde{\theta} \in \tilde{\Theta}} exp(-V(\tilde{\theta})) \leq 1$.

Then the total two stage description length l is as follows

$$l(u) = \min_{\tilde{\theta} \in \tilde{\Theta}} \left(-\log p_{\tilde{\theta}}(u) + V(\tilde{\theta}) \right). \tag{1.1}$$

As one can notice, codelengths l are a sum of two description lengths; description of the parameter space by V and the description of the data u given the parameter by $-\log p_\theta(u)$. When the sample space $\mathbb{U}$ is also countable, we have

$$\sum_{u \in \mathbb{U}} \exp(-l(u)) = \sum_{u \in \mathbb{U}} \max_{\tilde{\theta} \in \tilde{\Theta}} \{ p_{\tilde{\theta}}(u) \exp(-V(\tilde{\theta})) \}.$$

In the right side of the above equation, the maximum can be upper bounded by the sum over all $\theta \in \tilde{\Theta}$. For each fixed θ the sum over u of $p_\theta(u)$ is 1 because $p_\theta(u)$ are a family of probability mass functions on $\mathbb{U}$ and the order of summation is interchanged. Then since V satisfies Kraft's inequality on $\tilde{\Theta}$ we have $\sum_{u \in \mathbb{U}} \exp(-l(u)) \leq 1$. Hence the two-stage codelengths l satisfy Kraft's inequality. When the sample space $\mathbb{U}$ is uncountable and $\{p_\theta(u) : \theta \in \Theta\}$ are probability densities with respect to a dominating measure ν on $\mathbb{U}$, the correspondence between probability distributions and codes can be extended as discussed in [8]. Hence we may think of negative log densities as Kraft satisfying codelengths. The summation in Kraft's inequality is now replaced by an integral and the two stage codelengths l in this case can be shown to satisfy $\int_{\mathbb{U}} \exp(-l(u)) \leq 1$.

In the case when the parameter space Θ is uncountable, one of the ways in which a penalized log likelihood expression could still be Kraft satisfying codelengths on the sample space is as follows. Let $pen : \Theta \to \mathbb{R}_+$ be a penalty function on Θ. Assume there exists a countable subset $\tilde{\Theta} \subset \Theta$ and any Kraft

summable penalty $V(\tilde{\theta})$ on $\tilde{\Theta}$ such that the following holds

$$\begin{aligned} &\min_{\theta\in\Theta}\{-\log p_\theta(u) + pen(\theta)\} \geq \\ &\min_{\tilde{\theta}\in\tilde{\Theta}}\{-\log p_{\tilde{\theta}}(u) + V(\tilde{\theta})\}. \end{aligned} \tag{1.2}$$

In this case the right side of the above display will satisfy Kraft's inequality by virtue of being a two-stage codelength on the countable set $\tilde{\Theta}$. Then the left side of the last display being not less than the right side also satisfies Kraft's inequality. So the upshot is, that for an uncountable parameter space Θ and a penalty function *pen*, as long as one verifies (1.2), one can assert that the following codelengths on Ω^n

$$l(u) = \min_{\theta\in\Theta}\{-\log p_\theta(u) + pen(\theta)\} \tag{1.3}$$

satisfy Kraft's inequality and hence again correspond to a prefix free code. In this way we link the countable and the uncountable cases. For a penalty function *pen* on Θ if there exists a countable F and Kraft satisfying V defined on $\tilde{\Theta}$ satisfying (1.2) then we say *pen* is a *codelength valid* penalty.

Remark 1.2.1. *The condition* (1.2) *can also be equivalently restated as the following: There exists a countable subset* $\tilde{\Theta} \subset \Theta$ *and any Kraft summable penalty* $V(\tilde{\theta})$ *on* $\tilde{\Theta}$ *such that for every* $\theta \in \Theta$ *and data* u *the following is satisfied:*

$$\min_{\tilde{\theta}\in\tilde{\Theta}} \left(\log \frac{p_\theta(u)}{p_{\tilde{\theta}}(u)} + V(\tilde{\theta}) \right) \leq pen(\theta). \tag{1.4}$$

Here V *is indeed a theoretical construct but in our applications would be very closely related to the pen we want to show is codelength valid. So for a penalty pen to be codelength valid we have to come up with a choice of a countable set* $\tilde{\Theta}$ *and a penalty function* V *such that* V *satisfies Kraft inequality on* $\tilde{\Theta}$

and (1.4) *holds.*

Remark 1.2.2. *The requirement on the penalty* (1.4) *says that for a penalty on an uncountable parameter space* Θ *to be codelength valid, one needs to find a countable subset* $\tilde{\theta}$ *and complexities* V *such that the penalty exceeds the sample resolvability of the probability distribution* p_θ *for all data* u *and all* $\theta \in \Theta$*. It is clear that defining* $pen(\theta) = V(\theta)$ *in case* $\theta \in \tilde{\Theta}$ *does not violate* (1.4)*.*

1.2.2 Risk Validity

Now we demonstrate how to derive adaptive risk bounds for penalized likelihood procedures.

Countable parameter Space

First we consider the countable parameter space case. The essentials of this argument can be found in [18] but we include it here for sake of completeness. Let $\tilde{\Theta}$ be a countable parameter space and $\{p_{\tilde{\theta}} : \tilde{\theta} \in \tilde{\Theta}\}$ denote probability mass functions or densities on $\mathbb{U}$ with respect to a dominating measure ν. Let V be a penalty function on $\tilde{\Theta}$. We want to investigate the statistical risk properties of the following penalized log likelihood estimator

$$\hat{\theta}(u) = \underset{\theta \in \tilde{\Theta}}{\operatorname{argmin}} \left(-\log p_\theta(u) + V(\theta) \right) \tag{1.5}$$

For any $0 < \alpha \leq 1$, we define a family, indexed by α, of loss functions between two probability measures with densities p and q on $\mathbb{U}$ by

$$L_\alpha(p, q) = -\frac{1}{\alpha} \log \mathbb{E}_p \left(\frac{q(u)}{p(u)} \right)^\alpha . \tag{1.6}$$

Remark 1.2.3. *We note that in the case* $\mathbb{U} = \mathcal{X}^n$ *is a* n *fold Cartesian product of* $\mathcal{X}$ *the for probability distributions* P_u *and* Q_u *on* $\mathbb{U}$ *having densities of the product form such as* $\Pi_{i=1}^n p(x_i)$ *and* $\Pi_{i=1}^n q(x_i)$ *respectively, then we have*

$$L_\alpha(P_u, Q_u) = nL_\alpha(P_x, Q_x). \tag{1.7}$$

where P_x, Q_x *are distributions on* $\mathcal{X}$ *with densities* p *and* q *respectively. In the literature, these are sometimes known as the Chernoff-Renyi divergences between probability measures.*

Remark 1.2.4. *It can be checked that*

$$\lim_{\alpha \to 0} L_\alpha(p, q) = D(p, q).$$

So, roughly speaking, a risk bound on L_α *for* α *near* 0 *would be nearly a risk bound for the Kulback Divergence loss.*

Remark 1.2.5. L_α *is not symmetric in general. However, it is symmetric when* $\alpha = \frac{1}{2}$*. In that case* $L_{\frac{1}{2}}$ *turns out to be the familiar Bhattacharya distance between two probability measures.*

Remark 1.2.6. *The Hellinger loss between two probability distributions with densities* p *and* q *on* $\mathbb{U}$ *is given by*

$$H^2(p, q) = 1 - \mathbb{E}_p \left(\sqrt{\frac{q(u)}{p(u)}} \right).$$

One can check that $L_{1/2}(p, q) = -2 \log \left(1 - \frac{1}{2} H^2(p, q)\right)$*. In particular we have that the Bhattacharya distance is a monotonic transformation of the Hellinger distance. Also, by properties of logarithms, we do have* $L_{1/2}(p, q) \geq H^2(p, q)$*. Hence risk bounds for the Bhattacharyya divergence immediately implies risk*

bounds for the Hellinger distance. The familiar Kulback Leibler divergence D between p and q is defined to be

$$D(p,q) = \mathbb{E}_p \log\left(\frac{p(u)}{q(u)}\right).$$

By Jensen's inequality one can check that $L_{1/2}(p,q) \leq D(p,q)$. In fact when the log likelihood ratios of p and q are bounded by constants then $L_{1/2}$ is within a constant factor of $D(p,q)$.

Remark 1.2.7. *The function $g(\alpha) = \alpha L_\alpha(p,q)$ as a function of α is concave on $[0,1]$. This can be checked by taking second derivative of g which can be interpreted as minus the variance of $\log(\frac{q(u)}{p(u)})$ with respect to some density and hence is non positive. It can also be checked that $g(0) = g(1) = 0$. Then by using the definition of concavity one obtains the following*

$$B(p,q) \leq \begin{cases} L_\alpha(p,q) & \text{if } 0 < \alpha \leq \frac{1}{2} \\ \frac{\alpha}{1-\alpha} L_\alpha(p,q) & \text{if } \frac{1}{2} \leq \alpha < 1. \end{cases}$$

A consequence of the above is that a bound on L_α is also a bound on the Bhattacharya divergence for all $0 < \alpha \leq \frac{1}{2}$ and a bound on the Bhattacharya divergence upto a constant factor for $\frac{1}{2} \leq \alpha < 1$.

Remark 1.2.8. *In case p and q are multivariate normals with mean vectors μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2 respectively, our loss function evaluates to the following expression*

$$\begin{aligned} L_\alpha(p,q) = \frac{1}{2\alpha} \log \frac{det(\alpha\Sigma_1 + (1-\alpha)\Sigma_2)}{det(\Sigma_1)^\alpha det(\Sigma_2)^{1-\alpha}} + \\ \frac{1-\alpha}{2}(\mu_1 - \mu_2)^T(\alpha\Sigma_1 + (1-\alpha)\Sigma_2)(\mu_1 - \mu_2). \end{aligned} \tag{1.8}$$

In case the covariance matrices are the same and identity then it is proportional

to the ℓ_2 squared norm between the mean vectors.

For any $\tilde{\theta} \in \tilde{\Theta}$ we now introduce a new notation. We define

$$\mathbb{D}_\alpha(\tilde{\theta}, u) = \log \frac{p^\star(u)}{p_{\tilde{\theta}}(u)} - L_\alpha(p^\star, p_{\tilde{\theta}}) \tag{1.9}$$

where the distribution generating the data u is denoted by $p^\star$. We note that $\mathbb{D}$ is almost of the form of a centered random variable. It is not quite because of L_α being subtracted off and not the Kulback Leibler divergence. That is why we call $\mathbb{D}_\alpha(\tilde{\theta}, u)$ the *discrepancy* of $p_{\tilde{\theta}}$ at the sample point u. Now we state a lemma:

Lemma 1.2.1. *Let the distribution generating the data u be denoted by $p^\star$. For the model $\{p_{\tilde{\theta}} : \tilde{\theta} \in \tilde{\Theta}\}$ and the penalized likelihood estimator defined as in (1.5), if the penalty function satisfies a slightly stronger Kraft type inequality as follows,*

$$\sum_{\theta \in \tilde{\Theta}} \exp(-\alpha\, V(\tilde{\theta})) \leq 1 \tag{1.10}$$

where $0 < \alpha \leq 1$ is any fixed number, we have the following moment generating inequality:

$$\mathbb{E} \exp\left(\alpha \max_{\tilde{\theta} \in \tilde{\Theta}} \{ -\mathbb{D}_\alpha(\tilde{\theta}, u) - V(\tilde{\theta}) \} \right) \leq 1. \tag{1.11}$$

Proof. By positivity of the exponential function and then by monotonicity and linearity of expectation we have

$$\mathbb{E} \exp\left(\alpha \max_{\tilde{\theta} \in \tilde{\Theta}} \{ -\mathbb{D}_\alpha(\tilde{\theta}, u) - V(\tilde{\theta}) \} \right) \leq$$
$$\sum_{\tilde{\theta} \in \tilde{\Theta}} \mathbb{E} \exp\left(\alpha(-\mathbb{D}_\alpha(\tilde{\theta}, u) - V(\tilde{\theta}) \right).$$

The right side of the above inequality can be expanded as

$$\sum_{\tilde{\theta}\in\tilde{\Theta}} \exp(\alpha L_\alpha(p^\star, p_{\tilde{\theta}}))\mathbb{E}(\frac{p_{\tilde{\theta}}(u)}{p^\star(u)})^\alpha \exp(-\alpha\, V(\tilde{\theta})). \tag{1.12}$$

By the definition of the loss function (1.6) the above simplifies to

$$\sum_{\tilde{\theta}\in\tilde{\Theta}} \exp(-\alpha\, V(\tilde{\theta})). \tag{1.13}$$

Now the summability condition (1.10) implies that the above display is not greater than 1. This completes the proof of lemma (1.2.1). □

Theorem 1.2.2. *Under the same conditions as in lemma (1.2.1) we have the following risk bound:*

$$\mathbb{E}L_\alpha(p^\star, p_{\hat{\theta}}) \leq \mathbb{E}\inf_{\tilde{\theta}\in\tilde{\Theta}} \left(\log\frac{p^\star(u)}{p_{\tilde{\theta}}(u)} + V(\tilde{\theta})\right). \tag{1.14}$$

Proof. Interchanging $\mathbb{E}_p$ and the exponential cannot increase the left side of equation (1.11) so we have the inequality

$$\exp(\alpha\, \mathbb{E}\max_{\tilde{\theta}\in\tilde{\Theta}}\{-\mathbb{D}_\alpha(\tilde{\theta}, u) - V(\tilde{\theta})\}) \leq 1.$$

Monotonicity of the exponential function and α being positive now implies

$$\mathbb{E}\max_{\tilde{\theta}\in\tilde{\Theta}}\{L_\alpha(p^\star, p_{\tilde{\theta}}) - \log(\frac{p^\star(u)}{p_{\tilde{\theta}}(u)}) - V(\tilde{\theta})\} \leq 0.$$

where we have expanded $\mathbb{D}_\alpha(\tilde{\theta}, u)$. Setting $\theta = \hat{\theta}$ in the left side of the above equation cannot increase it and hence we have

$$\mathbb{E}\{L_\alpha(p^\star, p_{\hat{\theta}}) - \log(\frac{p^\star(u)}{p_{\hat{\theta}}(u)}) - pen(\hat{\theta})\} \leq 0.$$

Taking the loss term on the other side and multiplying by -1, we get the desired risk bound by recalling the definition of $\hat{\theta}$. This completes the proof of Theorem (1.2.2). □

Remark 1.2.9. *The above theorem says that the expected loss $L_\alpha(p^\star, p_{\hat{\theta}})$ is upper bounded by the expected sample resolvability of the data generating distribution $p^\star$. By interchanging the expectation and the minimum we see that the upper bound is a tradeoff between Kulback approximation and complexity divided by the sample size. This gives us adaptive risk bounds.*

Now we extend Theorem (1.2.2) in the uncountable parameter space case.

Extension to Uncountable Parameter Spaces

The previous argument only works for countable parameter spaces. This is because we cannot take a sum over uncountable possibilities as in the first step of the proof of Lemma (1.2.1). In statistical applications, the estimators are optimized over continuous spaces and it is awkward to force a user to construct countable discretizations of the parameter space. In this section we show how to extend the idea of the previous section to obtain risk bounds for estimators minimizing negative log likelihood plus a penalty term over uncountable choices. We identify conditions on the penalty *pen* and the log likelihood in order to be able to mimic the countable case and derive risk bounds. Let Θ now denote the parameter space which is uncountable. Let *pen* be a penalty function defined on Θ. The penalized likelihood estimator is now defined as

$$\hat{\theta}(u) = \underset{\theta \in \Theta}{\operatorname{argmin}} \left(-\log p_\theta(u) + pen(\theta) \right). \tag{1.15}$$

For any $\theta \in \Theta$ we again denote

$$\mathbb{D}_\alpha(\theta, u) = \log \frac{p^\star(u)}{p_\theta(u)} - L_\alpha(p^\star, p_\theta).$$

Analogous to (1.2) let us assume the existence a countable subset $\tilde{\Theta} \subset \Theta$ and a penalty function V on $\tilde{\Theta}$ such that the following holds for any fixed $0 < \alpha < 1$ and data u,

$$\min_{\tilde{\theta} \in \tilde{\Theta}} \left(\mathbb{D}_\alpha(\theta, u) + V(\tilde{\theta}) \right) \leq \min_{\theta \in \Theta} \left(\mathbb{D}_\alpha(\theta, u) + pen(\theta) \right). \tag{1.16}$$

Also analogous to (1.10) let us assume V satisfies a similar inequality on F

$$\sum_{\tilde{\theta} \in \tilde{\Theta}} \exp(-\alpha\, V(\tilde{\theta})) \leq 1. \tag{1.17}$$

We now state the following theorem for the uncountable parameter case.

Theorem 1.2.3. *We again denote the distribution generating the data u by $p^\star$. For the model $\{p_\theta : \theta \in \Theta\}$, if the assumptions (1.16) and (1.17) are met then we have the desired risk bound for the estimator (1.15) as follows*

$$\mathbb{E} L_\alpha(p^\star, p_{\hat{\theta}}) \leq \mathbb{E} \inf_{\theta \in \Theta} \left(\log \frac{p^\star(u)}{p_{\hat{\theta}}(u)} + pen(\theta) \right). \tag{1.18}$$

Proof. Since V satisfies (1.10) on F which is countable, by lemma (1.2.1) we obtain

$$\mathbb{E} \exp(\alpha \max_{\tilde{\theta} \in \tilde{\Theta}} \{ -\mathbb{D}_\alpha(\tilde{\theta}, u) - V(\tilde{\theta}) \}) \leq 1.$$

Note that assumption (1.16) could be rewritten as

$$\max_{\tilde{\theta}\in\tilde{\Theta}}\left(-\mathbb{D}_\alpha(\theta,u)-V(\tilde{\theta})\right)\leq\max_{\theta\in\Theta}\left(-\mathbb{D}_\alpha(\theta,u)-pen(\theta)\right).$$

The last two displays now imply the moment generating inequality

$$\mathbb{E}\exp(\alpha\max_{\theta\in\Theta}\{-\mathbb{D}_\alpha(\theta,u)-pen(\theta)\})\leq 1. \tag{1.19}$$

Again by interchanging exponential and expectation and then by the monotonicity of the exponential function we have the following

$$E\max_{\theta\in\Theta}(-\mathbb{D}_\alpha(\theta,u)-pen(\theta))\leq 0.$$

By setting $\theta=\hat{\theta}$ we cannot increase the expectation and hence we have

$$E_p\left(-\mathbb{D}_\alpha(\hat{\theta},u)-pen(\hat{\theta})\right)\leq 0.$$

Expanding $\mathbb{D}$ and then taking the log term and the penalty term on the right side and recalling the definition of $\hat{\theta}$ we obtain the desired risk bound. This completes the proof of theorem (1.2.3). □

For a penalty function *pen* on Θ if there exists a countable F and a penalty function V defined on $\tilde{\Theta}$ satisfying (1.16) and (1.17) then we say *pen* is a *risk valid penalty.*

Remark 1.2.10. *Here again, the expected loss $L_\alpha(p^\star,p_{\hat{\theta}})$ is upper bounded by the sample resolvability of the data generating distribution $p^\star$ with respect to the uncountable class $\{p_\theta:\theta\in\Theta\}$. Also $p^\star$ denotes the data generating probability measure which need not be in the model we consider for theorem (1.2.3) to be valid.*

Remark 1.2.11. *The condition* (1.16) *is very similar to* (1.2) *with the loss terms added. Condition* (1.16) *can be interpreted in another way which is going to be sometimes more convenient for us. For a penalty pen defined on* Θ *to be valid for risk bounds such as* (1.18)*, condition* (1.16) *behooves us to find a countable* $\tilde{\Theta} \subset \Theta$ *and a penalty* V *defined on* $\tilde{\Theta}$ *satisfying Kraft* (1.10) *such that for any given* $\theta \in \Theta$ *and any given data point* u*, we have the following inequality*

$$\min_{\tilde{\theta} \in \tilde{\Theta}} (\mathbb{D}_\alpha(\tilde{\theta}, u) - \mathbb{D}_\alpha(\theta, u) + V(\tilde{\theta})) \leq pen(\theta). \tag{1.20}$$

Consequently, it is also enough to show for every $\theta \in \Theta$ *and fixed data* u *there must exist a representer* $\tilde{\theta} \in \tilde{\Theta}$ *such that*

$$L_\alpha(p^\star, p_{\tilde{\theta}}) - L_\alpha(p^\star, p_{\tilde{\theta}}) + \log \frac{p_\theta(u)}{p_{\tilde{\theta}}(u)} + V(\tilde{\theta}) \leq pen(\theta).$$

This representer may depend on the data u*. In applications we will show that for every* θ *its representer, perhaps dependent on* u*, can be found locally by searching over nearby lattice points. This will be made clear in the examples. This allows us to mimic the countable parameter space situation and lets us prove desired risk bounds.*

Remark 1.2.12. *We introduce some terminology which we use throughout this manuscript. For any* θ *we call the term* $\mathbb{D}_\alpha$ *or* $\log(\frac{p^\star_{\tilde{\theta}}(u)}{p_{\tilde{\theta}}(u)}) - L_\alpha(p^\star, p_{\tilde{\theta}})$ *as the discrepancy term because it is of the form of negative log likelihood ratio minus its population counterpart. We also call* $V(\theta)$ *the complexity term for* θ*. Then the condition* (1.20) *says that for all* $\theta \in \Theta$*, there should exist its representer* $\tilde{\theta}$*, such that the penalty at* θ *should be at least the difference in discrepancies at* $\tilde{\theta}$ *and* θ *respectively plus the complexity at* $\tilde{\theta}$ *in order to be risk valid.*

Remark 1.2.13. *Note that we have a variety of risk bounds with loss functions*

L_α parametrized by $0 < \alpha \leq 1$. If we want to get risk bounds for the Kulback Divergence we might want to take α near 0. The problem with that is our requirement on the penalty becomes more stringent as α decreases.

I.I.D case

Let the sample space $\mathbb{U} = \mathcal{X}^n$ for some space $\mathcal{X}$. We write a generic element $u = (x_1, \ldots, x_n)$. Let the model consist of densities of the product form $\{\Pi_{i=1}^n p_\theta(x_i) : \theta \in \Theta\}$ where p_θ are a family of densities on $\mathcal{X}$. Also let $p^\star$ denote the density of the data generating distribution on $\mathcal{X}$. In this setting, we write our risk bound in the following corollary.

Corollary 1.2.4. *Under the same assumptions as in theorem (1.2.2) we have the risk bound for all $0 < \alpha \leq 1$,*

$$\mathbb{E}L_\alpha(p^\star, p_{\hat{\theta}}) \leq \mathbb{E} \inf_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^{n} \log \frac{p^\star(x_i)}{p_\theta(x_i)} + \frac{pen(\theta)}{n}\right). \tag{1.21}$$

Proof. The proof follows by dividing throughout by n in equation (1.18) and because we are in the i.i.d setting. □

We note that by interchanging expectation and infimum in the right side of the risk bound in the last display we have

$$\mathbb{E}L_\alpha(p^\star, p_{\hat{\theta}}) \leq \inf_{\theta \in \Theta} \left(D(p^\star, p_\theta) + \frac{pen(\theta)}{n} \right). \tag{1.22}$$

The right side in the last display is called the index of resolvability as in [1]. As it can be seen, the index of resolvability is an ideal tradeoff between the KL approximation and the penalty or the complexity relative to the sample size. The index of resolvability bound shows adaptation for these penalized

likelihood estimators for parameter spaces with varying levels of complexity. One of the ways to see this is if the true data generating distribution lies in the model then the index of resolvability bound implies an upper bound of penalty of the true parameter divided by n. So we have better bounds for simpler truths. The index of resolvability bound also helps to show these estimators are simultaneously minimax optimal for all the complexity classes in many problems as shown in [1] and [25].

So far we have provided finite sample upper bounds for the expected loss. In case of i.i.d data finite sample high probability upper bounds are also readily available for the loss.

Corollary 1.2.5. *In case of i.i.d data we have the probability of the event that the loss exceeds the expected redundancy per symbol by a positive number $\tau > 0$ is exponentially small in n. We have the following inequality*

$$\begin{aligned} &\mathbb{P}\left(L_\alpha(p, p_{\hat{\theta}}) > \frac{1}{n} \inf_{\theta \in \Theta} \{ \sum_{i=1}^{n} \log(\frac{p^\star(x_i)}{p_{\hat{\theta}}(x_i)}) + \frac{pen(\hat{\theta})}{\alpha} \} + \tau \right) \\ &< e^{-n\alpha\tau}. \end{aligned} \tag{1.23}$$

Proof. We take equation (1.19) as our starting point. In the i.i.d setting we can rewrite it as

$$\begin{aligned} &\mathbb{E} \exp(n\alpha \max_{\theta \in \Theta} \{ L_\alpha(p^\star, p_\theta) - \frac{1}{n} \sum_{i=1}^{n} \log(\frac{p^\star(x_i)}{p_{\hat{\theta}}(x_i)}) - \frac{pen(\theta)}{n} \}) \\ &\leq 1. \end{aligned}$$

By setting $\theta = \hat{\theta}$ the above equation implies

$$\mathbb{E}\exp(n\alpha\{L_\alpha(p^\star, p_{\hat{\theta}}) - \frac{1}{n}\sum_{i=1}^{n}\log(\frac{p^\star(x_i)}{p_{\hat{\theta}}(x_i)}) - \frac{pen(\hat{\theta})}{n}\})$$

$$\leq 1.$$

Let τ be any positive number. By applying Markov's inequality and the previous equation we complete the proof of this corollary. □

Remark 1.2.14. *$p^\star$ denotes the data generating probability distribution which need not be in the model we consider for our risk bounds to be valid.*

Remark 1.2.15. *In order to apply theorem (1.2.2) to derive bounds in expectation or in probability as in corollary (1.2.5) for particular models, we need to be able to check condition (1.16) which means we have to come up with a choice of a countable subset $\tilde{\Theta} \subset \Theta$ and a penalty function V defined on $\tilde{\Theta}$ satisfying (1.17). We will show in the coming sections how to demonstrate that these conditions hold in canonical high-dimensional parametric problems such as Linear Models and Gaussian Graphical Models with the penalty being a suitable multiple of the l_1 penalty. We will also show how to use Theorem (1.2.2) to obtain adaptive risk bounds for a suitable multiplier times the l_0 penalty in the Linear model case. Our aim is to demonstrate that the existence condition of countable covers of the parameter space that we have proposed are natural and are satisfied for the canonical problems we consider in high-dimensional statistics.*

1.3 Validity of the l_1 penalty

In this section we show that a certain weighted ℓ_1 type penalty with a suitable multiplier is codelength valid and risk valid in the linear regression problem.

We also show that the ℓ_1 penalty is risk valid in the setting of Gaussian graphical models. We essentially verify conditions (1.16) and (1.17) in both these models. Our point is to convince the reader that our conditions are indeed satisfied in these canonical problems.

1.3.1 Linear Regression

To illustrate our techniques of obtaining adaptive risk bounds we first choose the setting of linear regression which is one of the canonical location problems in statistics. We have a real valued response variable y and a vector valued predictor vector x. We assume y conditional on x is Gaussian with conditional mean function $f^\star(x)$ and known variance σ^2. We are given n realizations $\{(y_i, x_i)\}_{i=1}^n$. The goal in this setting might be to estimate this unknown $f^\star$ as that completely specifies the conditional density of y given x under the Gaussian assumption.

In the fixed design case, the loss function measures how close our estimates are to the truth at the same design points that we had seen in the data and used to compute the estimate. In the random design case, one assumes that the pairs $\{(y_i, x_i)\}_{i=1}^n$ are i.i.d from some joint distribution. We use the design points we have seen in the data to compute our estimates but our loss evaluates how good we predict the response when we observe a new design point arising i.i.d from the marginal distribution of the covariates. Our theory can handle the random design case as well with appropriate assumptions on the marginal distribution of the covariates. For simplicity of exposition we treat the fixed design case here. So we now treat the predictor values $\{(x_i)\}_{i=1}^n$ as given. The random data here is the vector y and takes the role of u in the preceding section.

We assume that we have a dictionary $\mathcal{D}$ of fixed functions $\{f_j\}_{j=1}^p$ where p could be very large compared to n. The dictionary could have been obtained from a previous training sample or otherwise. We restrict attention to estimators of the conditional mean function, which take the form of a data dependent linear combination of the functions $f \in \mathcal{D}$. In other words, our estimators would be a member of the set

$$\{f : f = \sum_{j=1}^{p} \theta_j f_j\}$$

where $\theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$. Hence our parameter space Θ could be identified with $\mathbb{R}^p$. For any $\theta \in \mathbb{R}^p$ we denote the function $f = \sum_{j=1}^p \theta_j f_j$ by f_θ. Now we proceed to show risk validity of a certain weighted ℓ_1 penalty. We would need to define a countable set $\tilde{\Theta} \in \Theta$ and a penalty function V satisfying (1.17) defined on $\tilde{\Theta}$ such that equation (1.20) holds. In order to define $\tilde{\Theta}$ let us fix some notations. If we denote the design matrix by Ψ, where $\Psi_{ij} = f_i(x_j)$, then we define weights $\{w_j\}_{j=1}^p$ as follows

$$w_j = \frac{1}{n} \sum_{i=1}^{n} f_i(x_j)^2 (\Psi^T \Psi)_{jj} \tag{1.24}$$

which coincides with the j the diagonal entry of $\frac{1}{n}(\Psi^T \Psi)_{jj}$. Thus the weight vector w is nothing but the empirical ℓ_2 norms of the columns of the design matrix Ψ. For any vector $v \in \mathbb{R}^p$ we denote its weighted ℓ_1 norm as

$$|v|_{1,w} = \sum_{j=1}^{p} w_j |v_j|.$$

We now define our countable set $\tilde{\Theta}$. We define the set $\tilde{\Theta}$ as follows

$$\tilde{\Theta} = \{(\frac{\delta z_1}{w_1}, \ldots, \frac{\delta z_p}{w_p}) : z \in Z^p\} \tag{1.25}$$

where the value of $\delta > 0$ will be specified later. Clearly $\tilde{\Theta}$ is countable since Z^p is so. We now define a penalty function V on $\tilde{\Theta}$ derived from C. So we define V for all $\tilde{\theta} \in \tilde{\Theta}$ in the following manner

$$V(\tilde{\theta}) = \frac{\log(p+1)}{\delta} |\tilde{\theta}|_{1,w} + 2. \tag{1.26}$$

The fact that V, as defined above, satisfies the Kraft inequality (1.17) follows from the following lemma.

Lemma 1.3.1. *With Z^p being the integer lattice we have*

$$\sum_{z \in Z^p} \exp(-C(z)) \leq 1. \tag{1.27}$$

where

$$C(z) = |z|_1 \log(p+1) + 2.$$

The proof of this lemma is given in the appendix.

We now proceed to define a risk valid penalty, by upper bounding the difference in discrepancies plus complexity term as in equation (1.20). Our loss functions between conditional densities of y given the predictors with means f_θ and $f_{\theta'}$, as can be checked from (1.8) turn out to be

$$L_\alpha(\theta, \theta') = \frac{(1-\alpha)\sigma^2}{2} \| f_\theta(\underline{x}) - f_{\theta'}(\underline{x}) \|_2^2 \tag{1.28}$$

where $f_\theta(\underline{x}) = (f_\theta(x_1), \ldots, f_\theta(x_n))$ and $\|.\|_2^2$ denotes the square of the ℓ_2 norm. Hence the difference in discrepancies $\mathbb{D}_\alpha(\tilde{\theta}, y) - \mathbb{D}_\alpha(\tilde{\theta}, y)$ can be written as

$$\begin{aligned} &\frac{(1-\alpha)}{2\sigma^2} \left(\| f_\theta(\underline{x}) - f^\star(\underline{x}) \|_2^2 - \| f_{\tilde{\theta}}(\underline{x}) - f^\star(\underline{x}) \|_2^2 \right) \\ &+ \frac{1}{2\sigma^2} \left(\| y - f_{\tilde{\theta}}(\underline{x}) \|_2^2 - \| y - f_\theta(\underline{x}) \|_2^2 \right). \end{aligned}$$

The difference in discrepancies can be further simplified to

$$\frac{\alpha}{2\sigma^2}\|f_{\tilde{\theta}}(\underline{x}) - f_\theta(\underline{x})\|_2^2 + \frac{1}{\sigma^2}\langle y - f_\theta(\underline{x}), f_\theta(\underline{x}) - f_{\tilde{\theta}}(\underline{x})\rangle - \frac{1-\alpha}{\sigma^2}\langle f_{\tilde{\theta}}(\underline{x}) - f_\theta(\underline{x}), f_\theta(\underline{x}) - f^\star(\underline{x})\rangle. \tag{1.29}$$

where $\langle v_1, v_2\rangle$ denotes the inner product between two vectors v_1, v_2. To show risk validity of the ℓ_1 penalty, as in (1.20) we will need to upper bound the following expression

$$\min_{\tilde{\theta}\in\tilde{\Theta}}(\mathbb{D}_\alpha(\tilde{\theta}, u) - \mathbb{D}_\alpha(\theta, u) + V(\tilde{\theta})).$$

We can upper bound the minimum by an expectation over any distribution μ on $\tilde{\Theta}$. So it is enough to upper bound the following quantity for any distribution μ on $\tilde{\Theta}$.

$$E_\mu(\mathbb{D}_\alpha(\tilde{\theta}, y) - \mathbb{D}_\alpha(\theta, y) + V(\tilde{\theta})). \tag{1.30}$$

The trick is to choose this distribution carefully depending on θ. For any fixed θ, our general strategy will be to choose μ such that $E_\mu\tilde{\theta} = \theta$. That is, the random $\tilde{\theta}$ under the distributionμ is unbiased for θ.

Now we illustrate how to define a distribution μ on $\tilde{\Theta}$ for purposes elicited above. We will actually show how to do the above in another way in subsection (1.6.1) in the appendix. This other way was outlined in [3] and though quite interesting, is slightly suboptimal, compared to the distribution μ we describe now in the following subsection.

Sampling method 1

We now show a way of devising a probability distribution μ on the countable set $\tilde{\Theta}$ so that the average of the difference in discrepancy plus complexity with

respect to this distribution upper bounds the minimum of it over $\tilde{\Theta}$ and helps us set a risk valid penalty. Let $\theta \in \mathbb{R}^p$ be given and $\delta > 0$ be a given number. We can always write θ in the following way

$$\theta = \delta(\frac{m_1}{w_1}, \ldots, \frac{m_p}{w_p})$$

for some vector $(m_1, \ldots, m_p)$. We now describe our sampling strategy. For any integer $1 \leq l \leq p$ we define a random variable $\mathbf{h}_l$ in the following way.

$$\begin{aligned} \mathbf{h}_l &= \frac{\delta}{w_l}\lceil m_l \rceil \text{ with probability } (\lceil m_l \rceil - m_l) \\ &= \frac{\delta}{w_l}\lfloor m_l \rfloor \text{ with probability } (m_l - \lfloor m_l \rfloor) \\ &= \frac{\delta}{w_l} m_l \text{ with probability } 1 - (\lceil m_l \rceil - \lfloor m_l \rfloor) \end{aligned} \tag{1.31}$$

Basically the above definition says that for each coordinate l, in case m_l is an integer, $h_l = \frac{\delta}{w_l} m_l$ with probability 1. In case m_l is not an integer, h_l is a two valued random variable taking values $\frac{\delta}{w_l} a$ and $\frac{\delta}{w_l}(a+1)$ where a is the unique integer such that $a < m_l < a + 1$. The following facts about the random variable $\mathbf{h}_l$ can be easily checked. If θ_l is an integer multiple of δ then $\mathbf{h}_l = \theta_l$ with probability 1. Secondly, $\mathbf{h}_l$ by definition, takes values in $\tilde{\Theta}$ with probability 1. Thirdly and crucially, $\mathbf{h}_l$ is unbiased for θ_l. Now we define the random vector $\mathbf{h} = (\mathbf{h}_1, \ldots, \mathbf{h}_p)$ where the coordinate random variables $\{\mathbf{h}_l\}_{l=1}^p$ are jointly independent. We denote the distribution of $\mathbf{h}$ by μ.

Now, we are going to upper bound the expression in (1.30). We first consider $\mathbb{E}_\mu \left(\mathbb{D}(\tilde{\theta}, y) - \mathbb{D}(\theta, y) \right)$. Since $\mathbb{E}_\mu \tilde{\theta} = \theta$ we have $\mathbb{E}_\mu f_{\tilde{\theta}}(\underline{x}) = f_\theta(\underline{x})$. So, as can be seen from (1.29), the inner product terms are zero on an average. So we have

$$\mathbb{E}_\mu \mathbb{D}(\tilde{\theta}, y) - \mathbb{D}(\theta, y) = \mathbb{E}_\mu \frac{\alpha}{2\sigma^2} \| f_{\tilde{\theta}}(\underline{x}) - f_\theta(\underline{x}) \|_2^2$$

Hence we now control the expected quadratic term which we can write as follows by expanding as linear combination of the dictionary functions

$$\frac{\alpha}{2\sigma^2}\sum_{i=1}^{n}\mathbb{E}_{\mu}(\sum_{j=1}^{p}(\mathbf{h}_j-\theta_j)f_j(x_i))^2. \tag{1.32}$$

By unbiasedness of $\mathbf{h}$ and independence of each of its coordinates the expected crossproduct terms in the inner sum are zero. Hence after interchanging the order of summation and by recalling the definition of the weight vector w the last display can be written as the following

$$n\sum_{j=1}^{p}w_j^2\mathbb{E}_{\mu}(\mathbf{h}_j-\theta_j)^2.$$

Now after some calculations similar to the calculation of the variance of a Bernoulli random variable, it can be shown that for each $1\leq l\leq p$,

$$\mathbb{E}_{\mu}(\mathbf{h}_l-\theta_l)^2=(\frac{\delta}{w_l})^2(m_l-\lfloor m_l\rfloor)(\lceil m_l\rceil-m_l).$$

Also it can be checked that for all numbers m_l we have the following inequality

$$(m_l-\lfloor m_l\rfloor)(\lceil m_l\rceil-m_l)\leq|m_l|.$$

The above inequality is rather crude for large $|m_l|$ as we also have

$$(m_l-\lfloor m_l\rfloor)(\lceil m_l\rceil-m_l)\leq\min(|m_l|,\frac{1}{4}).$$

For this particular argument we really have the high-dimensional situation in mind, that is when p is large. In this situation it is okay to use the crude upper bound. So from the arguments above, we obtain an upper bound $\frac{\alpha}{2\sigma^2}n\delta^2\sum_{l=1}^{p}|m_l|$ for the expected quadratic term. Now by dividing and mul-

tiplying by w_l within every term in the sum and recalling the definition of θ we get the final upper bound for every y and θ,

$$\mathbb{E}_\mu \Big(\mathbb{D}(\tilde{\theta}, y) - \mathbb{D}(\theta, y) \Big) \leq \frac{\alpha}{2\sigma^2} n\delta |\theta|_{w,1}. \tag{1.33}$$

In order to control the complexity term, by (1.26) we have

$$\mathbb{E}_\mu V(\tilde{\theta}) = \frac{\log(p+1)}{\delta} \sum_{l=1}^{p} w_l \mathbb{E}_\mu |\tilde{\theta}_l| + 2.$$

For each l, we now claim that

$$\mathbb{E}_\mu |\tilde{\theta}_l| \leq |\theta_l|.$$

In fact there is equality in the above display but for us the inequality is enough. This can be seen as follows:

$$\mathbb{E}_\mu |\tilde{\theta}_l| = \mathbb{E}_\mu \tilde{\theta}_l \{\tilde{\theta}_l > 0\} - \mathbb{E}_\mu \tilde{\theta}_l \{\tilde{\theta}_l < 0\}. \tag{1.34}$$

Now we observe that $\{\tilde{\theta}_l > 0\} \leq \{\theta_l > 0\}$ and $\{\tilde{\theta}_l < 0\} \leq \{\theta_l < 0\}$ with probability 1 under μ. Hence substituting these inequalities in the last display we have the upper bound

$$\mathbb{E}_\mu |\tilde{\theta}_l| = \mathbb{E}_\mu \tilde{\theta}_l \{\theta_l > 0\} - \mathbb{E}_\mu \tilde{\theta}_l \{\theta_l < 0\}. \tag{1.35}$$

Now using the fact that $\mathbb{E}_\mu \tilde{\theta}_l = \theta_l$ we have the right side of the above display is just $\mathbb{E}_\mu |\theta_l|$ and hence we prove our claim. This implies

$$\mathbb{E}_\mu V(\mathbf{h}) = \frac{|\theta|_{w,1}}{\alpha\delta} \log(p+1) + \frac{2}{\alpha}. \tag{1.36}$$

Hence, from (1.33) and (1.36) we have the upper bound for the expectation of the sum of difference in discrepancy plus complexity to be

$$\frac{\alpha}{2\sigma^2} n\delta |\theta|_{w,1} + \frac{|\theta|_{w,1}}{\alpha\delta} \log(p+1) + \frac{2}{\alpha}.$$

Setting $\delta^2 = \frac{2\sigma^2 \log(p+1)}{\alpha^2 n}$ we obtain the following upper bound to the sum of difference in discrepancies and complexity

$$\frac{1}{\sigma}\sqrt{2n\log(p+1)}\,|\theta|_{w,1} + \frac{2}{\alpha}.$$

It follows that by defining the penalty function on Θ defined as follows

$$pen(\theta) = \frac{1}{\sigma}\sqrt{2n\log(p+1)}\,|\theta|_{w,1} + \frac{2}{\alpha}. \tag{1.37}$$

we have the risk validity of a weighted ℓ_1 penalty given by *pen*. Since *pen* is a risk valid penalty, by a direct application of theorem (1.2.2) and some minor rearranging of terms we obtain for all $0 < \alpha < 1$,

$$\begin{aligned} &\frac{1}{2n\sigma^2}\mathbb{E}\|f_{\hat{\theta}}(\underline{x}) - f^\star(\underline{x})\|_2^2 \leq \\ &(\frac{1}{1-\alpha})\mathbb{E}\inf_{\theta\in\mathbb{R}^p}\left(\frac{1}{2n\sigma^2}\|y - f_\theta(\underline{x})\|_2^2 - \|y - f^\star(\underline{x})\|_2^2\right. \\ &\left. + \frac{1}{\sigma}\sqrt{\frac{2\log(p+1)}{n}}\,|\theta|_{w,1} + \frac{2}{\alpha n}\right). \end{aligned} \tag{1.38}$$

By taking the expectation inside the infimum on the right side of the above display we present a theorem in this linear regression setting.

Theorem 1.3.2. *For the penalized likelihood estimator $\hat{\theta}$ defined as in (1.5) and the penalty given by (1.37) we have the following oracle inequality type*

result

$$\mathbb{E}\frac{1}{2n\sigma^2}\sum_{i=1}^{n}(f_{\hat{\theta}}(x_i) - f^\star(x_i))^2 \leq$$
$$(\frac{1}{1-\alpha}) \inf_{\theta\in\mathbb{R}^p} \left(\frac{1}{2n\sigma^2}\|f_{\hat{\theta}}(\underline{x}) - f^\star(\underline{x})\|_2^2 \right. \tag{1.39}$$
$$\left. + \sqrt{\frac{2\log(p+1)}{n}}\,|\theta|_{w,1} + \frac{2}{\alpha n} \right).$$

Remark 1.3.1. *The leading constant on the right side can be made to be arbitrarily close to* 1 *by choosing* α *arbitrarily near* 0 *but then we pay for it as we have to divide the penalty term by* α *in the risk bound.*

Remark 1.3.2. *We do not need any conditions on the design matrix* Ψ *in order for our risk bound to hold.*

Remark 1.3.3. *We have shown the risk validity of the penalty as defined in* (1.37)*. The codelength validity of the same penalty can be shown by exactly similar methods and is omitted here.*

1.3.2 Gaussian Graphical Models

A canonical scale problem in statistics is the problem of estimating the inverse covariance matrix of a multivariate Gaussian random vector. We observe $X = \{x_i\}_{i=1}^n$, each of which is drawn i.i.d from $N_p(0, \theta^{-1})$. Here $\theta_{p\times p}$ denotes the inverse covariance matrix of the random gaussian vectors. We denote the corresponding covariance matrices by $\Sigma = \theta^{-1}$. We assume that the model is well specified and we denote the true inverse covariance matrix to be $\theta^\star$. In this section we denote the $-\log det$ function on matrices by ϕ. We follow the convention that ϕ takes value $+\infty$ on any matrix that is not positive definite. Then it follows that ϕ is a convex function on the space of all $p \times p$ matrices.

Inspecting the log likelihood of this model we have

$$-\frac{1}{n}\log p_\theta(X) = \frac{p}{2}\log(2\pi) + \frac{1}{2}Tr(S\theta) + \frac{\phi(\theta)}{2}$$

Here, $Tr(S\theta)$ is the sum of diagonals of the matrix $S\theta$ and $S = \frac{1}{n}\sum_{i=1}^{n} \tilde{x_i}^T \tilde{x_i}$. In this setting $\theta_{ij} = 0$ means that the ith and jth variables are conditionally independent given the others. We outline the proof of the fact that the penalty $|\theta|_1$, which is just the sum of absolute values of all the entries of the inverse covariance matrix, is a risk valid penalty. We show our risk bounds in the case when the truth $\theta^\star$ is sufficiently positive definite in the following way. We assume that for any matrix in the set $\{\Delta : \|\Delta\|_\infty \leq \delta\}$ we have

$$(\theta^\star + \Delta) \succ 0. \tag{1.40}$$

Here $\|\Delta\|_\infty$ means the maximum absolute entry of the matrix Δ and a matrix being $\succ 0$ means it is positive definite. We remark that this is our only assumption on the true inverse covariance and the value of the δ in the assumption is specified later. Now we proceed with our scheme of things. Let us denote the space of $p \times p$ positive definite symmetric matrices by S_+^p. In this setting the parameter space could be identified with a convex cone of $\mathbb{R}^{p^2}$, the convex cone being the cone of positive definite symmetric matrices. We define $\tilde{\Theta}$ to be the δ integer lattice intersected with S_+^p. So we have

$$\tilde{\Theta} = \{\delta z \in \mathbb{R}^{p\times p} : vec(z) \in Z^{p^2}, z \in S_+^p\}. \tag{1.41}$$

Here $vec(z)$ is the vectorized form of the matrix $z \in \mathbb{R}^{p\times p}$ arranged to be a $p^2 \times 1$ column vector. Clearly, $\tilde{\Theta}$ is a countable set. We also define the penalty

function V on $\tilde{\Theta}$ in the following way

$$V(\delta z) = \frac{C(z)}{\alpha}. \tag{1.42}$$

By lemma (1.3.1) it is clear that V defined as above on $\tilde{\Theta}$ satisfies the Kraft type inequality (1.17). For this i.i.d model, our loss function for the joint distributions turns out to be

$$\begin{aligned} L_\alpha(\theta_1, \theta_2) = \frac{n}{2\alpha}[\alpha\phi(\theta_2) \\ + (1-\alpha)\phi(\theta_1) - \phi(\alpha\theta_2 + (1-\alpha)\theta_1)]. \end{aligned} \tag{1.43}$$

Since ϕ is a convex function, by Jensen' inequality one can see that $L_\alpha \geq 0$. In this setting, we will present our risk bounds for $0 < \alpha \leq \frac{1}{2}$ Now we need to verify (1.20) in order to set a risk valid penalty. Expanding and simplifying (1.20) we have

$$\begin{aligned} \frac{n}{2\alpha}[\phi(\alpha\tilde{\theta} + (1-\alpha)\theta^\star) \\ - \phi(\alpha\theta + (1-\alpha)\theta^\star)] + \frac{n}{2}Tr(S(\tilde{\theta} - \theta)) + V(\tilde{\theta}). \end{aligned} \tag{1.44}$$

One can check that by treating ϕ as a function of p^2 variables, one has for a given positive definite matrix $M_{p\times p}$ and for any pair of indices i, j we have $\frac{\partial}{\partial M_{i,j}}\phi(M) = -(M^{-1})_{i,j}$. Also for any other pair of indices k, l the second derivatives are given by $\frac{\partial^2}{\partial M_{k,l}\partial M_{i,j}}\phi(M) = (M^{-1}E_{k,l}M^{-1})_{i,j} = (M^{-1})_{i,k}(M^{-1})_{j,l}$. Here $E_{k,l}$ is a $p \times p$ matrix with all zero entries except a 1 at the k, l position. So the associated $p^2 \times p^2$ Hessian matrix is $M^{-1} \otimes M^{-1}$ where $\otimes$ denotes the Kronecker product between matrices. By Taylor expanding ϕ about A upto the second order term we have the following equality for all positive definite

symmetric matrices A and $A+B$ where t is some number between 0 and 1

$$\phi(A+B)-\phi(A)= \\ -Tr(BA^{-1})+vec(B)^T H\left((A+tB)^{-1}\right)vec(B). \tag{1.45}$$

where H evaluated at a positive definite matrix $M_{p\times p}$ is a $p^2\times p^2$ matrix and is given by

$$H_{(i,j),(k,l)}(M)=M_{i,k}M_{j,l}.$$

Let us now set $A=(1-\alpha)\theta^\star+\alpha\theta$ and $B=\alpha(\tilde{\theta}-\theta)$ in the above Taylor expansion. Then we can write (1.44) as

$$-\frac{n}{2\alpha}Tr(BA^{-1})+\frac{n}{2}Tr(SB)+V(\tilde{\theta})+ \\ \frac{n}{2\alpha}vec(B)^T H\left((A+tB)^{-1}\right)vec(B).$$

We again upper bound the minimum of the last expression over $\tilde{\theta}\in\tilde{\Theta}$ by an expectation over a chosen distribution μ on $\tilde{\Theta}$. This distribution μ is similar to the distribution μ used in the first sampling method in the linear regression setting. The only difference s that it is in dimension p^2 intead of p. So our random choice of $\tilde{\theta}$ is unbiased for θ and hence the average of B is zero. Consequently the trace terms are zero on an average. Then we have to control the difference in discrepancy which is a quadratic form and the penalty term. Since the coordinates of the random choice of $\tilde{\theta}$ are independent, the cross terms in the quadratic form are zero on an average. We note that an important property of our sampling strategy is that the ℓ_∞ distance between the random choice $\tilde{\theta}$ and θ is not greater than δ. Hence it follows that $|B|_\infty\leq\alpha\delta$. Now by assumption (1.40) one can check for all $0<t<1$ and all $0<\alpha\leq\frac{1}{2}$ it follows

that $\frac{(1-\alpha)}{2}\theta^\star + tB \succ 0$. Also we have by definition of A and B here,

$$A + tB - \frac{(1-\alpha)}{2}\theta^\star - \alpha\theta = \frac{(1-\alpha)}{2}\theta^\star + tB. \tag{1.46}$$

The above two equations imply that for all $0 < t < 1$ and all $0 < \alpha \leq \frac{1}{2}$ we have

$$A + tB \succ \frac{(1-\alpha)}{2}\theta^\star \succ 0. \tag{1.47}$$

In particular we are always inside the region of differentiability of ϕ and hence our Taylor expansion is valid. We first consider the following expected quadratic form for any $0 \leq t \leq 1$

$$\mathbb{E}_\mu(vec(B)^T H(A+tB)^{-1}vec(B)).$$

Since the cross terms are zero on an average due to independence of the coordinates and the fact that $\mathbb{E}vec(B) = 0$ we have the last display equalling

$$\mathbb{E}_\mu \sum_{l=1}^{p^2}(vec(B)_l)^2(H(A+tB)^{-1})_{ll}.$$

Now by definition of H any of the diagonals of $(H(A+tB)^{-1})$ is not greater than the maximum diagonal of $(A+tB)^{-1}$ squared. Now (1.47) implies that the maximum diagonal of $(A+tB)^{-1}$ is not greater than the maximum diagonal of $\frac{2}{1-\alpha}\Sigma^\star$. Let us denote the maximum diagonal of Σ by σ_{max}. Then we have the following inequality for all $1 \leq l \leq p^2$,

$$((A+tB)^{-1} \otimes (A+tB)^{-1})_{ll} \leq \frac{4(\sigma_{max})^2}{(1-\alpha)^2}. \tag{1.48}$$

Now, as in the linear regression case, it can be shown that for each coordinate l, the variance of $vec(B)_l$ is upper bounded by $\delta|vec(\theta)_l|$. Hence we can write

$$\mathbb{E}_\mu(vec(B)^T H\left((A+tB)^{-1}vec(B)\right) \leq \frac{4(\sigma_{max})^2}{(1-\alpha)^2}$$

$$\delta|vec(\theta)|_1.$$

As for the penalty term, the sampling method ensures that the signs of each of the coordinates of the random choice $\tilde{\theta}$ does not change. Hence the expected penalty term is just the penalty evaluated at θ. So then we have the expected difference in discrepancy plus complexity upper bounded by

$$\frac{4n(\sigma_{max})^2}{2\alpha(1-\alpha)^2}\delta|\theta|_1 + \frac{|\theta|_1}{\alpha\delta}\log(4p^2) + \frac{\log 2}{\alpha}.$$

Again by setting

$$\delta^2 = \frac{\log(4p^2)(1-\alpha)^2}{2n(\sigma_{max})^2}$$

it follows that by defining the penalty function on Θ as follows

$$pen(\theta) = \frac{\sqrt{\sigma_{max}\log(4p^2)2n}}{\alpha(1-\alpha)}|\theta|_1 + \frac{\log 2}{\alpha} \tag{1.49}$$

we construct a risk valid penalty. So with the definition of pen above, the estimator defined as follows

$$\hat{\theta} = \underset{\theta\in S_+^p}{\operatorname{argmin}}\left(\frac{1}{2}Tr(S\theta) + \frac{\phi(\theta)}{2} + \frac{pen(\theta)}{n}\right). \tag{1.50}$$

enjoys the adaptive risk properties we desire. Under the assumption (1.40) where now δ has been specified, we have the following risk bound for all $0 <$

$\alpha \leq \frac{1}{2}$

$$\mathbb{E}L_\alpha(\theta^\star, \hat{\theta}) \leq \mathbb{E} \inf_{\theta \in S_+^p} \left(\frac{1}{2} Tr(S(\theta - \theta^\star)) + \frac{\phi(\theta) - \phi(\theta^\star)}{2} + \frac{pen(\theta)}{n} \right).$$

By taking the expectation inside the infimum we now present our theorem.

Theorem 1.3.3. *For the estimator $\hat{\theta}$ as in (1.50) with $\hat{\Sigma}^{-1} = \hat{\theta}^{-1}$ and the penalty (1.49) we have the risk bound*

$$\mathbb{E}L_\alpha(\theta^\star, \hat{\theta}) \leq \inf_{\theta \in S_+^p} \left(\frac{1}{2}[Tr(\theta\Sigma^\star) - p] + \frac{1}{2}[\phi(\theta) - \phi(\theta^\star))] + \frac{pen(\theta)}{n} \right). \tag{1.51}$$

Remark 1.3.4. *By setting $\theta = \theta^\star$ in the right side of the bound, as long as $\theta^\star$ has finite l_1 norm, one has the standard risk bound $\sqrt{\frac{\log(4p^2)}{n}} \|\theta^\star\|_1$. The main purpose of the risk bound is to demonstrate the adaptation properties of the l_1 penalized estimator and to demonstrate redundancy, a coding notion, as the upper bound to the statistical risk which has been championed in [6].*

Remark 1.3.5. *The assumption (1.40) says that the true inverse covariance matrix $\theta^\star$ should be in the interior of the cone of positive definite matrix by a little margin. This assumption may be acceptable even in high dimensions as it does not prohibit collinearity.*

1.4 Validity of l_0 penalty in Linear Regression

In this section we return to the linear regression setup to show the codelength and risk validity of the l_0 penalty. We again consider the known variance σ^2

and the fixed design setup. Our model is

$$y_{n\times 1} = \Psi_{n\times p}\theta_{p\times 1} + \epsilon_{n\times 1}$$

where $\epsilon \sim N(0, \sigma^2 I_{n\times n})$ and Ψ is the design matrix. The log likelihood of the model is

$$-\log p_\theta(y) = \frac{1}{2\sigma^2}\|y - \Psi\theta\|_2^2 + \frac{n}{2}\log 2\pi\sigma^2.$$

We assume our model is well specified and there is a true vector of coefficients $\theta^\star$. Our results would be in the regime when the sample size n is larger than the number of explanatory variables p. We divide the data y into $y_{in} = (y_{in}, \Psi_{in})$ consisting of p samples and $y_f = (y_f, \Psi_f)$ consisting of $(n-p)$ samples. Here in is intended to suggest initial and f is intended to mean final. It does not really matter which p samples are chosen to represent the initial sample as long as it is done once and then remains frozen. We assume that the matrix Ψ_{in} is non singular. The purpose of such division of data is to use the initial p samples y_{in} to create a Kraft summable penalty on the countable cover we will choose and then this penalty together with the cover is used to derive codelength interpretation for the ℓ_0 penalized log likelihood or risk bounds for the estimator minimizing the ℓ_0 penalized log likelihood.

1.4.1 Codelength Validity

Ideally one would like a penalty, which is constant on all vectors with a specified number of non zeroes, to be codelength valid. This cannot be achieved and the codelength valid penalties have to diverge off to infinity as we go further out in the parameter space. In an attempt to construct a codelength valid penalty with the leading term proportional to the ℓ_0 norm only, our strategy

here is to condition on an initial segment of data. Accordingly let $pen(\theta|y_{in})$ be a penalty function defined on $\Theta = \mathbb{R}^p$ which is a function of y_{in} also. So it is infact a random penalty. The notation is deliberately designed to make the reader think of $pen(\theta|y_{in})$ as a penalty conditional on the initial data y_{in}. Analogous to (1.2) we intend to show the existence of a countable set $\tilde{\Theta} \subset \Theta$ and a Kraft valid codelength $V(\tilde{\theta}|y_{in})$ on $\tilde{\Theta}$ such that the following inequality holds

$$\begin{aligned} &\min_{\theta\in\Theta}\{-\log p_\theta(y) + pen(\theta|y_{in})\} \geq \\ &\min_{\tilde{\theta}\in\tilde{\Theta}}\{-\log p_{\tilde{\theta}}(y_f) + V(\tilde{\theta}|y_{in})\} \end{aligned} \tag{1.52}$$

where now the right side of (1.52) gives a two stage codelength interpretation provided we treat it as codelengths on y_f conditional on y_{in} and hence the left side as a function on y_f, being not less than the right side, also has a two stage conditional codelength interpretation. We now proceed to find out a suitable conditional penalty $pen(\theta|y_{in})$ which would satisfy (1.52).

We introduce some notations and tools which would be used to show both the codelength validity and the risk validity of a penalty with the main term of the order $k(\theta)\log n$. We now make some relevant definitions and set up some notations. Let $\theta \in \mathbb{R}^p$ be a given vector. We define $k(\theta) = \sum_{i=1}^p I\{\theta_i \neq 0\}$. In other words $k(\theta)$ is the number of non zeros of the vector θ. We denote the support of θ or the set of indices where θ is non zero by $S(\theta)$. Clearly $|S(\theta)| = k(\theta)$. Let $S^\star$ be the support of the true vector of coefficients $\theta^\star$. For any subset $S \subset [1:p]$, let $\Psi_{in,S}$ denote the initial part of the design matrix with column indices in S in natural order. Hence $\Psi_{in,S}$ is a p by $|S|$ matrix. Let us denote the matrix $(\Psi_{in,S}^T\Psi_{in,S})^{-1/2}$ by M_S. We also denote the quantity $\frac{1}{|S|}Tr\left((\Psi_{in,S}^T\Psi_{in,S})^{-1}(\Psi_{f,S}^T\Psi_{f,S})\right)$ by Υ_S where Tr refers to the trace or sum

of diagonals of a matrix.

Let $\mathcal{Z}$ denote the set of integers as before. Also fix some $\delta > 0$. For any given subset S we define a countable set

$$C_S = \{M_S v - w : v \in \delta(\mathcal{Z} - \{0\})^{|S|}\} \tag{1.53}$$

where w is the unique solution to the equation $\Psi_{in,S} w = O_{\Psi_{in},S} y_{in}$. Here $O_{\Psi_{in},S}$ is the orthogonal projection matrix onto the column space of the matrix $O_{\Psi_{in},S}$. As we have defined, C_S is a subset of $\mathbb{R}^{|S|}$ but by appending the coordinates in the complement of S as zeroes, we treat C_S as a subset of $\mathbb{R}^p$. We want to construct Kraft satisfying codelengths and hence subprobabilities on C_S which are proportional to $\left(\frac{p_\phi(y_{in})}{p_{\theta^\star}(y_{in})}\right)^\eta$ for any fixed but arbitrary $0 < \eta \leq 1$. For that purpose we want to estimate the normalizer which is the quantity $\sum_{\phi \in C_S} \left(\frac{p_\phi(y_{in})}{p_{\theta^\star}(y_{in})}\right)^\eta$. The following lemma helps us do exactly that.

Lemma 1.4.1. *For all $0 < \eta \leq 1$ we have*

$$\sum_{\phi \in C_S} \left(\frac{p_\phi(y_{in})}{p_{\theta^\star}(y_{in})}\right)^\eta \delta^{|S|} \leq U_\eta(y_{in}, S) \tag{1.54}$$

where

$$\begin{aligned} &U_\eta(y_{in}, S) = \exp\left(\frac{\eta}{2}\|O_{\Psi_{in,S\cup S^\star}} y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2\right) \\ &(\frac{2\pi}{\eta})^{|S|/2} \end{aligned} \tag{1.55}$$

and $O_{\Psi_{in,S\cup S^\star}}$ denotes the orthogonal projection matrix onto the column space of the matrix $\Psi_{in,S\cup S^\star}$.

The proof of this lemma is given in the appendix.

We now define the countable set $\mathcal{C} \subset \mathbb{R}^p$ as follows

$$\mathcal{C} = \cup_{k=0}^{p} \cup_{\{S:|S|=k\}} C_S. \tag{1.56}$$

$\mathcal{C}$ is the union of the countable sets $C_{S,\eta}$ over all subsets $S \subset [1:p]$. Hence $\mathcal{C}$ itself is a countable subset of $\mathbb{R}^p$. By definition, $\mathcal{C}$ varies with δ and in applications we will set δ to be something specific. We now define penalty functions satisfying Kraft type inequalities on the countable set $\mathcal{C}$. First we define a family of subprobabilities h_η on $\mathcal{C}$ as follows

$$h_\eta(\tilde{\theta}, y_{in}) = (\frac{1}{2})^{k(\tilde{\theta})+1} \frac{1}{\binom{p}{k(\tilde{\theta})}} (\frac{p_{\tilde{\theta}}(y_{in})}{p_{\theta^\star}(y_{in})})^\eta \ \delta^{k(\tilde{\theta})} \frac{1}{U_\eta(y_{in}, S(\tilde{\theta}))}. \tag{1.57}$$

We claim that $h_\eta(\tilde{\theta})$ is a subprobability on $\tilde{C}_\eta$ for every y_{in}. This can be seen by first summing $h_\eta(\tilde{\theta})$ over non negative integers k from 0 to p, then summing over all subsets of $[1:p]$ with cardinality k and then summing over $C_{S,\eta}$. The inner sum over $C_{S,\eta}$ of $(\frac{p_{\tilde{\theta}}(y_{in})}{p_{\theta^\star}(y_{in})})\delta^{|S|}\frac{1}{U_\eta(y_{in},S)}$ is no more than 1 by lemma (1.4.1). Then for each k we sum over $\binom{p}{k(\tilde{\theta})}$ subsets and the factor $\frac{1}{\binom{p}{k(\tilde{\theta})}}$ keeps the overall sum still no more than 1. Similarly, the factor $(\frac{1}{2})^{k(\tilde{\theta})+1}$ makes the whole sum less than or equal to 1 when we sum over k from 0 to p, which can be seen by summing up the geometric series. Hence, we prove our claim.

We can now define Kraft satisfying codelengths $l_\eta(\tilde{\theta}, y_{in})$ on $\mathcal{C}$ by defining

$$l_\eta(\tilde{\theta}, y_{in}) = -\frac{1}{\eta} \log h_\eta(\tilde{\theta}) \tag{1.58}$$

Then because of h_η being a subprobability, it is clear that l_η satisfies the

following inequality for all y_{in}

$$\sum_{\tilde{\theta}\in\mathcal{C}} \exp(-\eta l_\eta(\tilde{\theta}, y_{in})) \leq 1. \tag{1.59}$$

Now we are ready to state our theorem.

Theorem 1.4.2. *The penalty $pen(\theta|y_{in})$, defined as below, is conditionally codelength valid in the sense of (1.52).*

$$\begin{aligned} pen(\theta|y_{in}) =& \frac{k(\theta)}{2}\log(\frac{4n}{p}) + \log\binom{p}{k(\theta)} + \\ & k(\theta)\left(\frac{3\log(2)}{2} + \frac{\log(2\pi)}{2}\right) \\ & + \frac{1}{2}\|O_{\Psi_{in,S(\theta)\cup S^\star}} y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2. \end{aligned}$$

Proof. We declare our countable set $\tilde{\Theta} = \mathcal{C}$ as defined in (1.56). We also define $V = l_\eta$ with $\eta = 1$ as defined in (1.58). Then we have

$$\begin{aligned} V(\tilde{\theta}) =& (k(\tilde{\theta})+1)\log(2) + \log\binom{p}{k(\tilde{\theta})} + \\ & k(\tilde{\theta})\log(\frac{1}{\delta}) + \log(U(y_{in}, S(\theta)) - \log\frac{p_{\tilde{\theta}}(y_{in})}{p_{\theta^\star}(y_{in})}. \end{aligned}$$

The task now is to verify (1.52). An equivalent way to verify (1.52) is to verify the following for any given $\theta \in \Theta$ and data y,

$$\begin{aligned} & \min_{\tilde{\theta}\in\tilde{\Theta}}\{-\log\frac{p_{\tilde{\theta}}(y_f)}{p_{\theta^\star}(y_f)} + \log\frac{p_\theta(y)}{p_{\theta^\star}(y)} + V(\tilde{\theta}|y_{in})\} \\ & \leq pen(\theta|y_{in}). \end{aligned} \tag{1.60}$$

In the case when y_{in} and y_f are independent, the log likelihood of the full data y is the sum of log likelihoods of y_{in} and y_f and so we can write the left side

of the above equation as

$$\min_{\tilde{\theta}\in\tilde{\Theta}}\{-\log\frac{p_{\tilde{\theta}}(y)}{p_{\theta^\star}(y)}+\log\frac{p_{\theta}(y)}{p_{\theta^\star}(y)}+ \left(V(\tilde{\theta}|y_{in})+\log\frac{p_{\tilde{\theta}}(y_{in})}{p_{\theta^\star(y_{in})}}\right)\}. \tag{1.61}$$

Now our strategy to upper bound the minimum of the above expression is to restrict the minimum over $\tilde{\theta}\in C_{S(\theta)}$ where $C_{S(\theta)}$ is as defined in (1.53). Doing this cannot decrease the overall minimum because $C_{S(\theta)}\subset\tilde{\Theta}$ by definition of $\tilde{\Theta}$. Restricted to $\tilde{\theta}\in C_{S(\theta)}$ one can check that the term $V(\tilde{\theta}|y_{in})+\log\frac{p_{\tilde{\theta}}(y_{in})}{p_{\theta^\star(y_{in})}}$ remains a constant. Now we state a lemma which helps us in upper bounding (1.61).

Lemma 1.4.3.

$$\min_{\tilde{\theta}\in C_{S(\theta)}}\{-\log\frac{p_{\tilde{\theta}}(y)}{p_{\theta^\star}(y)}+\log\frac{p_{\theta}(y)}{p_{\theta^\star}(y)}\}\leq 2(1+\Upsilon_{S(\theta)})\,k(\theta)\delta^2. \tag{1.62}$$

The proof of the above lemma is given in the appendix.

By the above lemma and the fact that $V(\tilde{\theta}|y_{in})+\log\frac{p_{\tilde{\theta}}(y_{in})}{p_{\theta^\star(y_{in})}}$ is constant on $C_{S(\theta),1}$ we write down the upper bound we get for the left side of (1.61) which is as follows

$$2(1+\Upsilon_{S(\theta)})\,k(\theta)\delta^2+(k(\theta)+1)\log(2)+\log\binom{p}{k(\theta)}+ k(\theta)\log(\frac{1}{\delta})+\log(U(y_{in},S(\theta)).$$

Setting $\delta^2=\frac{1}{4(1+\Upsilon_{S(\theta)})}$ we see that a valid penalty satisfying (1.52) would be

$$pen(\theta|y_{in})=\frac{k(\theta)}{2}+(k(\theta)+1)\log(2)+ \log\binom{p}{k(\theta)}+\frac{k(\theta)}{2}\log(4(1+\Upsilon_{S(\theta)}))+\log(U(y_{in},S(\theta)). \tag{1.63}$$

Rearranging and expanding $U(y_{in}, S(\theta))$ we have

$$pen(\theta|y_{in}) = \frac{k(\theta)}{2}\log(4(1+\Upsilon_{S(\theta)})) + \log\binom{p}{k(\theta)} + k(\theta)\left(\frac{3\log(2)}{2} + \frac{\log(2\pi)}{2}\right) + \frac{1}{2}\|O_{\Psi_{in,S(\theta)\cup S^\star}} y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2.$$

With a fixed design matrix there is only one term in the above expression which is random. It can be checked that the term $\frac{1}{2}\|O_{\Psi_{in,S\cup S^\star}} y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2$ is distributed as a χ^2 random variable with degree of freedom at most $k(\theta)+k^\star$. So its expected value is going to be at most $k(\theta)+k^\star$. In the case when the design matrices Ψ_{in} and Ψ_f have orthogonal columns and the ℓ_2 norms of each of the columns of Ψ_{in} and Ψ_f are atmost p and $n-p$ respectively we then have for any subset S, $\Psi_{in,S}^T\Psi_{in,S} = pI_{|S|\times|S|}$ and $\Psi_{f,S}^T\Psi_{f,S} = (n-p)I_{|S|\times|S|}$. In that case it can be checked that $\gamma_S = \frac{n-p}{p}$. Hence in this situation, our codelength valid penalty conditional on y_{in} becomes exactly as defined in Theorem (1.68). This completes the proof of Theorem (1.68). □

Remark 1.4.1. *Note that the leading term of the expected penalty $pen(\theta|y_{in})$ is indeed going to be the traditional $\frac{\log(n))}{2}k(\theta)$ in case p does not grow with n. In case p grows as n^β for some $0<\beta<1$ then the leading term of of the expected penalty $pen(\theta|y_{in})$ is still some constant times $k(\theta)\log(n)$. We remind the reader that $k(\theta)\log(p/k(\theta)) \leq \log(\binom{p}{k(\theta)}) \leq k(\theta)\log(ep/k(\theta))$. So the term $\log(\binom{p}{k(\theta)})$ again contributes a constant times $k(\theta)\log(n)$ term in case p is growing as some power of n.*

1.4.2 Risk validity

In this section we show the risk validity of the l_0 penalty by leveraging its codelength interpretation as shown in the last subsection. To prove risk bounds by the same reasoning as in section (1.2.2) we need to adapt the arguments in section (1.2.2) to the case when we have data split into two parts. We define our family of loss functions between two probability distributions p and q on $\mathcal{X}^n$ in the same way as before except that it only depends on the final part of the data X_f. Let $0 < \alpha \leq 1$ be a fixed, arbitrary number. We define our loss function as follows

$$L_\alpha(p, q) = -\frac{1}{\alpha} \log(\mathbb{E}(\frac{q(y_f)}{p(y_f)})^\alpha). \tag{1.64}$$

Also for a penalty $pen(\theta|y_{in})$ depending on y_{in} we define our penalized likelihood estimator to be

$$\hat{\theta}(y) = \underset{\theta \in \Theta}{\operatorname{argmin}} \{-\log p_\theta(y) + pen(\theta|y_{in})\}. \tag{1.65}$$

We now present the theorem which will help us in proving risk bounds for the ℓ_0 penalized likelihood estimator. Fix $0 < \alpha < 1$. For our countable set $\tilde{\Theta} = \mathcal{C}$ and codelengths $V = l_\alpha$ as defined in (1.58), clearly the following is true by (1.59).

$$\sum_{\tilde{\theta} \in \tilde{\Theta}} \exp(-\alpha V(\tilde{\theta}|y_{in})) \leq 1. \tag{1.66}$$

We expand V to get

$$V(\tilde{\theta}|y_{in}) = \frac{1}{\alpha} \left((k(\tilde{\theta}) + 1) \log(2) + \log \binom{p}{k(\tilde{\theta})} + k(\tilde{\theta}) \log(\frac{1}{\delta}) + \log(U_\alpha(y_{in}, S(\tilde{\theta})) \right) - \log \frac{p_{\tilde{\theta}}(y_{in})}{p_{\theta^\star(y_{in})}}.$$

We would like to now verify the following

$$\begin{aligned}&\min_{\tilde{\theta}\in\tilde{\Theta}}\{-\log\frac{p_{\tilde{\theta}}(y_f)}{p_{\theta^\star}(y_f)}+\log\frac{p_\theta(y)}{p_{\theta^\star}(y)}+\\&L_\alpha(p_{\theta^\star},p_\theta)-L_\alpha(p_{\theta^\star},p_{\tilde{\theta}})+V(\tilde{\theta}|y_{in})\}\leq pen(\theta).\end{aligned}\tag{1.67}$$

Verifying the above gives us risk bounds as is shown in the following lemma.

Lemma 1.4.4. *Assuming the existence of a countable subset $\tilde{\Theta}\subset\Theta$ and a penalty function $V(.|y_{in})$ defined on F satisfying (1.66) and (1.67), we have the following risk bound*

$$\mathbb{E}L_\alpha(p_{\theta^\star},p_{\hat{\theta}})\leq\mathbb{E}\min_{\theta\in\Theta}\left(\log\frac{p_{\theta^\star}(y)}{p_\theta(y)}+pen(\theta|y_{in})\right).$$

The proof of this lemma parallels the proof of Theorem (1.2.3) and is given in the appendix.

Now we proceed to verify (1.66) in order to derive the risk bound in theorem (1.4.4) for the l_0 penalized estimator in the linear regression setting. We can write the left side in (1.67) as

$$\begin{aligned}&\min_{\tilde{\theta}\in\tilde{\Theta}}\{-\log\frac{p_{\tilde{\theta}}(y)}{p_{\theta^\star}(y)}+\log\frac{p_\theta(y)}{p_{\theta^\star}(y)}+L_\alpha(p_{\theta^\star},p_\theta)-\\&L_\alpha(p_{\theta^\star},p_{\tilde{\theta}})+\left(V_\alpha(\tilde{\theta}|y_{in})+\log\frac{p_{\tilde{\theta}}(y_{in})}{p_{\theta^\star(y_{in})}}\right)\}.\end{aligned}$$

Again we upper bound the minimum of the above expression by restricting to $\tilde{\theta}\in C_{S(\theta)}$ which cannot decrease the overall minimum. Restricted to $\tilde{\theta}\in C_{S(\theta)}$ it turns out that the term $V_\alpha(\tilde{\theta}|y_{in})+\log\frac{p_{\tilde{\theta}}(y_{in})}{p_{\theta^\star(y_{in})}}$ remains a constant. The following lemma now helps us.

Lemma 1.4.5. *Given any θ and data y we have the following inequality*

$$\min_{\tilde{\theta}\in C_{S(\theta)}} \left(\log(\frac{p_\theta(y)}{p_{\tilde{\theta}}(y)}) + L_\alpha(p_{\tilde{\theta}}, p_{\theta^\star}) - L_\alpha(p_\theta, p_{\theta^\star}) \right)$$
$$\leq 2k(\theta)\delta^2(1+\alpha\Upsilon_{S(\theta)})$$

Hence we get an upper bound for the left side of (1.67) which is as follows

$$2k(\theta)\delta^2(1+\alpha\Upsilon_{S(\theta)}) + \frac{(k(\theta)+1)\log(2)}{\alpha} + \frac{\log\binom{p}{k(\theta)}}{\alpha}$$
$$+ \frac{k(\theta)}{\alpha}\log(\frac{1}{\delta}) + \frac{\log(U(y_{in}, S(\theta))}{\alpha}.$$

Setting

$$\delta^2 = \frac{1}{4\alpha(1+\alpha\Upsilon_{S(\theta)})}$$

we see that a risk valid penalty would be

$$pen_\alpha(\theta|y_{in}) =$$
$$\frac{k(\theta)}{2\alpha}\log(4\alpha(1+\alpha\Upsilon_{S(\theta)})) + \frac{k(\theta)}{2\alpha} +$$
$$\frac{(k(\theta)+1)\log(2)}{\alpha} + \frac{\log\binom{p}{k(\theta)}}{\alpha} + \frac{\log(U_\alpha(y_{in}, S(\theta))}{\alpha}.$$

Rearranging and expanding $\log U_\alpha(y_{in}, S(\theta))$ we have

$$pen_\alpha(\theta|y_{in}) =$$
$$\frac{k(\theta)}{2\alpha}\log(4\alpha(1+\alpha\Upsilon_{S(\theta)})) + \alpha\Psi_f^T\Psi_f)) +$$
$$\frac{(k(\theta)+1)\log(2)}{\alpha} + \frac{\log\binom{p}{k(\theta)}}{\alpha} + \frac{k(\theta)}{2}\log(\frac{2\pi}{\alpha}) +$$
$$\frac{1}{2}\|O_{\Psi_{in,S\cup S^\star}} y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2. \tag{1.68}$$

By taking the expectation inside the minimum in the right side and then doing some algebraic manipulations of the statement of Theorem (1.4.4), we get the

resolvability risk bound which we write down below as a theorem.

Theorem 1.4.6. *With the estimator being defined as in* (1.65) *and* $pen_\alpha(\theta|y_{in})$ *as defined in* (1.68) *we have the risk bound for all* $0 < \alpha \leq 1$,

$$\mathbb{E}\frac{1}{2n}\|y_f(\hat{\theta}-\theta^\star)\|_2^2 \leq \frac{1-\alpha}{\sigma^2}\inf_{\theta\in\mathbb{R}^p}\left(\frac{1}{2n}\|y_f(\hat{\theta}-\theta^\star)\|_2^2+\frac{1}{n}\mathbb{E}_{in}\, pen(\theta|y_{in})\right).$$

Remark 1.4.2. *As we can see, as* α *is taken to be near zero, the constant outside the right side in theorem* (1.4.6) *approaches the desired value* 1. *But then we have to pay for the fact that the penalty contains terms divided by* α *which blow up when* α *is brought near zero.*

By setting $\theta = \theta^\star$ inside the infimum in the above theorem we obtain

$$\mathbb{E}\frac{1}{2n}\|y_f(\hat{\theta}-\theta^\star)\|_2^2 \leq \frac{1-\alpha}{\sigma^2}\mathbb{E}_{in}\,\frac{pen(\theta^\star|y_{in})}{n}. \tag{1.69}$$

Remark 1.4.3. *The random part depending on* y_{in} *in* $pen(\theta^\star|y_{in})$ *is*

$$\|O_{\Psi_{in,S\cup S^\star}}y_{in}-\Psi_{in,S^\star}\theta^\star\|_2^2.$$

The above term is distributed as a χ^2 *random variable with degrees of freedom at most* $k(\theta)+k(\theta^\star)$. *In the case when the design matrices* Ψ_{in} *and* Ψ_f *are orthogonal and the* ℓ_2 *norms of the columns of* Ψ_{in} *and* Ψ_f *are atmost* p *and* $n-p$ *respectively we then have* $\Upsilon_S = \frac{n-p}{p}$. *Then the leading term of the expected penalty is of the order* $k(\theta)\log(n)$ *and hence we atleast have a* $k(\theta^\star)\log(n)/n$ *rate of convergence of the left side in* (1.69).

Remark 1.4.4. *Our risk bounds are useful even when* p *grows like a constant fraction of* n.

1.5 Conclusion

We have contributed to a general theory of penalized likelihood estimation by building a connection to two stage coding procedures in MDL, even in uncountable parameter spaces. For a given penalty, we have proposed that existence of appropriate countable covers of the parameter space imply adaptive statistical risk bounds for the penalized likelihood estimator on any model as long as data is being generated in an i.i.d fashion. We then have exhibited multiple ways as to how to construct these countable covers and verify the needed properties in certain canonical problems in statistics.

1.6 Appendix

1.6.1 Proof of Lemma (1.3.1)

We are to show $\sum_{z\in\mathcal{Z}^p} \exp(-C(z)) \leq 1$ where

$$C(z) = |z|_1 \log(p+1) + 2.$$

Proof. It can be easily checked by summing up the geometric series that

$$\sum_{z\in\mathcal{Z}} (\frac{1}{p+1})^{|z|} = 1 + \frac{2}{p}.$$

Hence we have

$$\sum_{z\in\mathcal{Z}^p} (\frac{1}{p+1})^{|z|_1} = (1 + \frac{2}{p})^p.$$

We can divide by $(1+\frac{2}{p})^p$ both sides and exponentiate to get

$$\sum_{z\in\mathcal{Z}^p} exp\left(-|z|_1\log(p+1)-p\log(1+\frac{2}{p})\right)\leq 1.$$

Using the fact that $p\log(1+\frac{2}{p})\leq 2$ we have

$$\sum_{z\in\mathcal{Z}^p} exp\left(-|z|_1\log(p+1)-2\right)\leq 1.$$

Recalling the definition of $C(z)$ this completes the proof of Lemma (1.3.1). □

Sampling method 2 for linear regression

Here we show an alternative method for demonstrating an upper bound to the sum of discrepancies plus complexity term in the linear regression case. $\tilde{\Theta}$ and V are defined as in (1.25) and (1.26) respectively. Let θ be any given vector in $\mathbb{R}^p$ and let us first consider the quadratic term which is the following

$$\frac{\alpha}{2\sigma^2}\sum_{i=1}^{n}(f_\theta(x_i)-f_{\tilde{\theta}}(x_i))^2.$$

By expanding out f_θ and $f_{\tilde{\theta}}$ in terms of the dictionary functions the last display becomes

$$\frac{\alpha}{2\sigma^2}\sum_{i=1}^{n}(\sum_{j=1}^{p}(\tilde{\theta}_j-\theta_j)f_j(x_i))^2. \tag{1.70}$$

Let δ be a positive real number. Let $K(\theta)=\lceil\frac{|\theta|_{w,1}}{\delta}\rceil$. $K(\theta)$ is the least integer larger than or equal to $|\theta|_{w,1}$ divided by δ. We will write $K=K(\theta)$ to minimize notational clutter. In order to explain our sampling strategy, we first define a random variable h. Let $\{\tilde{e_j}\}_{j=1}^p$ denote the canonical basis of $\mathbb{R}^p$. The random vector h takes value $K\delta sign(\theta_j)\frac{\tilde{e_j}}{w_j}$ with probability $\frac{w_j\theta_j}{K\delta}$ for all $j=1,\ldots,p$. With the remaining probability, h takes the form of the zero vector. One

can check that h defined this way is unbiased for θ, that is $\mathbb{E}(h_1) = \theta$. Say we sample K i.i.d copies $h_1, \ldots, h_k$. of h. We now consider the mean of these random vectors $\overline{h} = \frac{1}{K}\sum_{i=1}^{K} h_i$. We denote the distribution of $\overline{h}$ by μ. Clearly, $\overline{h}$ is also unbiased for θ. We first note that

$$\mathbb{E}_{\mu}(\sum_{j=1}^{p}((\overline{h}_j - \theta_j)f_j(x_i))^2 = \frac{1}{K}\mathbb{E}_h(\sum_{j=1}^{p}((h_j - \theta_j)f_j(x_i))^2.$$

This is because $\overline{h}$ is the sum of K i.i.d copies of h and h is unbiased for θ. Now we can upper bound the expectation over $\{h_l\}_{l=1}^{K}$ of the above term as follows

$$\mathbb{E}_h(\sum_{j=1}^{p}((h_j - \theta_j)f_j(x_i))^2 \leq \mathbb{E}_h(\sum_{j=1}^{p} h_j f_j(x_i))^2.$$

The above inequality follows due to unbiasedness of h_1 and by the simple fact that the variance of any random variable is at most the expected square of that random variable. We note the fact that for any $j \neq l$ the cross product terms $h_j h_l = 0$ pointwise by definition of h. Now summing over i and combining the previous two inequalities we obtain the following result

$$\mathbb{E}_{\mu}\sum_{i=1}^{n}(\sum_{j=1}^{p}(\overline{h}_j - \theta_j)f_j(x_i))^2 \leq \frac{n}{K}\sum_{j=1}^{p} w_j^2 E_h(h_j)^2. \tag{1.71}$$

For any j we also have by definition of the random variable h

$$\mathbb{E}_h h_j^2 = \frac{K\delta\theta_j}{w_j}. \tag{1.72}$$

So combining the last two equations we have

$$\mathbb{E}_{\mu}\sum_{i=1}^{n}(\sum_{j=1}^{p}(\overline{h}_j - \theta_j)f_j(x_i))^2 \leq n\delta|\theta|_{w,1}. \tag{1.73}$$

Now we consider the penalty or the complexity term. We note that each coordinate of h has a fixed sign depending on the signs of the coordinates of θ. Therefore, linearity of expectation extends to absolute values also, in other words it can be checked that the following holds for any $j \in [1:p]$

$$\mathbb{E}_\mu |\overline{h}_j| = \mathbb{E}|h_j|$$

The above equation and the definition of V implies the following fact

$$\mathbb{E}_\mu V(\overline{h}) = \mathbb{E}_h V(h).$$

It is clear now from the definition of h and the definition of V the following holds

$$\mathbb{E}_\mu V(\overline{h}) = \frac{K \log(p+1) + 2}{\alpha}.$$

So by the above arguments we can conclude that H_θ on an average is upper bounded by the following expression

$$\frac{\alpha}{2\sigma^2} n\delta |\theta|_{w,1} + \frac{K \log(p+1) + 2}{\alpha}.$$

Using the fact that $K \leq \frac{|\theta|_{w,1}}{\delta} + 1$ we have the expression in the last display can be further upper bounded by the following expression

$$\frac{\alpha}{2\sigma^2} n\delta |\theta|_{w,1} + \frac{|\theta|_{w,1}}{\delta} \frac{\log(p+1)}{\alpha} + \frac{\log(p+1) + 2}{\alpha}.$$

Setting $\delta^2 = \frac{2\sigma^2 \log(p+1)}{\alpha^2 n}$ we obtain the following that by defining the penalty function on Θ as follows

$$pen(\theta) = \frac{1}{\sigma} \sqrt{2n \log(4p)} |\theta|_{w,1} + \frac{\log(p+1) + 2}{\alpha}. \tag{1.74}$$

we define a risk valid penalty.

Remark 1.6.1. *Comparing the penalties in* (1.37) *and* (1.74) *we see that* (1.74) *contains an extra* $\frac{\log(p+1)}{\alpha}$ *term. In this sense* (1.37) *is an improvement over* (1.37) *which was obtained in the linear regression problem in [3].*

This completes this subsection.

1.6.2 Proof of Lemma (1.4.1)

Proof. We fix any subset $S \subset [1:p]$. We first relate the sum over $\phi \in C_S$ of the following expression

$$(\frac{P_\phi(y_{in})}{P_{\theta^\star}(y_{in})})^\eta \delta^{|S|}$$

to the integral of the same expression over all of $\mathbb{R}^{|S|}$. By Pythagorus theorem, we have

$$\|y_{in} - \Psi_{in,S}\phi\|_2^2 =$$
$$\|y_{in} - O_{\Psi_{in},S}y_{in}\|_2^2 + \|O_{\Psi_{in},S}y_{in} - \Psi_{in,S}\phi\|_2^2$$

where $O_{\Psi_{in},S}$ denotes the orthogonal projection matrix to the column space of the matrix $\Psi_{in,S}$. Hence one can check

$$(\frac{P_\phi(y_{in})}{P_{\theta^\star}(y_{in})})^\eta = A \sum_{\phi \in C_S} \exp\left(-\frac{\eta}{2}\|O_{\Psi_{in},S}y_{in} - \Psi_{in,S}\phi\|_2^2\right) \tag{1.75}$$

where

$$A = \exp\left(-\frac{\eta}{2}\{\|y_{in} - O_{\Psi_{in},S}y_{in}\|_2^2 + \|y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2\}\right). \tag{1.76}$$

□

Note that by properties of orthogonal projections,

$$\begin{aligned}
&\|y_{in} - y_{in,S^\star}\theta^\star\|_2^2 - \|y_{in} - O_{y_{in,S}}y_{in}\|_2^2 \leq \\
&\|y_{in} - y_{in,S^\star}\theta^\star\|_2^2 - \|y_{in} - O_{y_{in,S\cup S^\star}}y_{in}\|_2^2 = \\
&\|O_{y_{in,S\cup S^\star}}y_{in} - y_{in,S^\star}\theta^\star\|_2^2.
\end{aligned}$$

Hence we have

$$A \leq \exp(-\frac{\eta}{2}\|O_{y_{in,S\cup S^\star}}y_{in} - y_{in,S^\star}\theta^\star\|_2^2). \tag{1.77}$$

Now we can always write $O_{\Psi_{in},S}y_{in} = \Psi_{in,S}w$ for some vector w because $O_{\Psi_{in},S}y_{in}$ lies in the column space of $O_{\Psi_{in},S}y_{in}$. So we have

$$\begin{aligned}
&\sum_{\phi\in C_S}\exp\left(-\frac{\eta}{2}\|O_{\Psi_{in},S}y_{in} - \Psi_{in,S}\phi\|_2^2\right) = \\
&\sum_{\phi\in C_S}\exp\left(-\frac{\eta}{2}(\phi - w)^T\Psi_{in,S}^T\Psi_{in,S}(\phi - w)\right)
\end{aligned} \tag{1.78}$$

Now recalling the definition of $C_S = \{M_S v - w : v \in \mathcal{G}^{|S|}\}$ where w is the unique solution to the equation $\Psi_{in,S}w = O_{\Psi_{in},S}y_{in}$ we can change variables and express the sum over C_S now as a sum over $\mathcal{G}^{|S|}$ as follows:

$$\begin{aligned}
&\sum_{\phi\in C_S}\exp(-\frac{\eta}{2}(\phi - w)^T\Psi_{in,S}^T\Psi_{in,S}(\phi - w) = \\
&\sum_{v\in\mathcal{G}^{|S|}}\exp(-\frac{\eta}{2}v^Tv).
\end{aligned} \tag{1.79}$$

Now the right side in the above equation appears while constructing an appropriate lower Reimann sum to an integral as we argue now. Imagine dividing up $\mathbb{R}^{|S|}$ into cubes of sidelength δ with the vertices of the cubes constiting the set $\delta\mathcal{Z}^p$. In order to minimize $\exp(-\frac{\eta}{2}v^Tv)$ for v in a cube, it is clear that one should choose v so that the absolute value of each of its coordinate is maxi-

mized. This implies that as we run over all the cubes, the minimizing points would be precisely the set $\mathcal{G}^{|S|}$. Since we are picking out the minimizing points and the volume of the cubes is $\delta^{|S|}$ we have

$$\sum_{v \in \mathcal{G}^{|S|}} \exp(-\frac{\eta}{2} v^T v) \delta^{|S|} \leq \int_{\mathbb{R}^{|S|}} \exp(-\frac{\eta}{2} v^T v) dv. \tag{1.80}$$

The above equation just states that the lower Reimann sum is upper bounded by the integral. Now it can be checked that $\int_{\mathbb{R}^{|S|}} \exp(-\frac{\eta}{2} v^T v) dv = (2\pi/\eta)^{|S|/2}$. Hence by (3.12) we have

$$A \int_{\mathbb{R}^{|S|}} \exp(-\frac{\eta}{2} v^T v) dv \leq U_\eta(y_{in,S}) \tag{1.81}$$

where we recall that

$$U_\eta(y_{in,S}) = \exp\left(\frac{\eta}{2} \|O_{\Psi_{in,S \cup S^\star}} y_{in} - \Psi_{in,S^\star} \theta^\star\|_2^2\right)$$
$$(\frac{2\pi}{\eta})^{|S|/2}.$$

So then by combining all the above equations (3.11), (3.12), (1.78), (1.79), (1.80), (1.81) we have the desired inequality

$$\sum_{\phi \in C_S} (\frac{p_\phi(y_{in})}{p_{\theta^\star}(y_{in})})^\eta \delta^{|S|} \leq U_\eta(y_{in}, S).$$

This completes the proof of Lemma (1.4.1).

1.6.3 Proof of Lemma (1.4.3)

Proof. For all $\tilde{\theta} \in C_{S(\theta)}$ clearly $k(\tilde{\theta}) \leq k(\theta)$ by definition of $C_{S(\theta)}$ as all the coordinates in the complement of S are set to zero. In this subsection when we write vectors $\tilde{\theta} \in C_{S(\theta),\alpha}$ and θ, we really mean $\tilde{\theta}_{S(\theta)}$ and $\theta_{S(\theta)}$ which are

$|S|$ dimensional vectors. We do not burden the notation here by adding extra subscripts. Now if we expand the log likelihoods we obtain

$$\log \frac{p_\theta(y)}{p_{\tilde{\theta}}(y)} = \frac{1}{2\sigma^2}\|y - \Psi_S\tilde{\theta}\|_2^2 - \frac{1}{2\sigma^2}\|y - \Psi_S\theta\|_2^2. \tag{1.82}$$

After simplifications and noting $\Psi^T\Psi = \Psi_{in}^T\Psi_{in} + \Psi_f^T\Psi_f$ we see that we have to upper bound the following expression

$$\min_{\tilde{\theta}\in C_{S(\theta)}} \{\frac{1}{2\sigma^2}\|\Psi_{in,S}\tilde{\theta} - \Psi_{in,S}\theta\|_2^2 + \frac{1}{2\sigma^2}\|\Psi_{f,S}\tilde{\theta} - \Psi_{f,S}\theta\|_2^2$$
$$+ l(\tilde{\theta} - \theta)\}$$

where l is an affine function. Setting $\tilde{\theta} = M_S\tilde{v} - w$ and $\theta = M_S v - w$ where $v = (M_S)^{-1}(\theta + w)$ we have that the expression in the last display equals

$$\min_{\tilde{v}\in\mathcal{G}^{|S|}} \{\frac{1}{2\sigma^2}(\tilde{v} - v)^T(\tilde{v} - v) + \frac{1}{2\sigma^2}(\tilde{v} - v)^T B(\tilde{v} - v)$$
$$+ l(M_S\tilde{v} - M_S v)\}$$

where the matrix B equals $M_S\Psi_{f,S}^T\Psi_{f,S}M_S$. Now our strategy is to upper bound the above minimum by an expectation with respect to a carefully chosen distribution. We now describe the choice of the distribution. Consider $v = (v_1, \ldots, v_{S(\theta)})$. For each coordinate l, one can devise a distribution taking only two values in $\mathcal{Z} - \{0\}$ such that the expectation of this distribution is v_l. The two values are namely the smallest non zero integer not lesser than v_l and the largest non zero integer not bigger than v_l. Hence one can devise a distribution on $\mathcal{G}^{|S|}$ with the property that each coordinate of the random vector drawn from two valued distributions as described is independent and the average of this distribution is the vector v. Now since we are only choosing points from closeby cubes any coordinate of $\tilde{v} - v$ is atmost 2δ in absolute value with

probability 1. For any quadratic form $(\tilde{v}-v)^T Q(\tilde{v}-v)$ where Q is some non negative definite matrix, by unbiasedness and independence of the coordinates, its expectation boils down to the expectation of the diagonal terms. Therefore we have

$$\mathbb{E}(\tilde{v}-v)^T Q(\tilde{v}-v) = (2\delta)^2 Tr(Q).$$

Also, l being an affine function of $\tilde{v}-v$ is zero on an average. So applying the above facts we have the following inequality

$$\begin{aligned} &\mathbb{E}\frac{1}{2}\|\Psi_{in}\tilde{\theta}-\Psi_{in}\theta\|_2^2 + \mathbb{E}\frac{1}{2}\|\Psi_f\tilde{\theta}-\Psi_f\theta\|_2^2 + \\ &\mathbb{E}l(\tilde{\theta}-\theta) \leq 2\delta^2(k(\theta)+Tr(B)). \end{aligned} \tag{1.83}$$

Now by definition of $\Upsilon_{S(\theta)}$, we have

$$2\delta^2(k(\theta)+Tr(B)) = 2\delta^2 k(\theta)(1+\Upsilon_{S(\theta)}).$$

Hence we get the desired bound

$$\min_{\tilde{\theta}\in C_{S(\theta)}} \log\frac{p_\theta(y)}{p_{\tilde{\theta}}(y)} \leq 2(1+\Upsilon_{S(\theta)})k(\theta)\delta^2. \tag{1.84}$$

This completes the proof of Lemma (1.4.3). The proof of Lemma (1.4.5) goes along very similar lines to this proof and hence is left to the reader. □

1.6.4 Proof of Lemma (1.4.4)

Proof. By the definition of $\hat{\theta}$ we have

$$L_\alpha(p_{\theta^\star}, p_{\hat{\theta}}) = \left(L_\alpha(p_{\theta^\star}, p_{\hat{\theta}}) - \log \frac{p_{\theta^\star}(y)}{p_{\hat{\theta}}(y)} - pen(\hat{\theta})\right) + \min_{\theta \in \Theta} \left(\log \frac{p_{\theta^\star}(y)}{p_\theta(y)} + pen(\theta)\right). \tag{1.85}$$

The exponential of the first term in the brackets in the above display can be upper bounded as follows.

$$\exp\left(\alpha\{L_\alpha(p_{\theta^\star}, p_{\hat{\theta}}) - \log \frac{p_{\theta^\star}(y)}{p_{\hat{\theta}}(y)} - pen(\hat{\theta}|y_{in})\}\right) \leq \sum_{\tilde{\theta} \in \tilde{\Theta}} \exp(\alpha L_\alpha(p_{\theta^\star}, p_{\tilde{\theta}}))(\frac{p_{\tilde{\theta}}(y_f)}{p(y_f)})^\alpha \exp(-\alpha V(\tilde{\theta}|y_{in})). \tag{1.86}$$

This follows by (1.66) and non negativity of the exponential function. We can now take expectation with respect to y_f conditional on y_{in}. The right side of the last display then becomes

$$\sum_{\tilde{\theta} \in \tilde{\Theta}} \exp(\alpha L_\alpha(p_{\theta^\star}, p_{\tilde{\theta}}))\mathbb{E}_f(\frac{p_{\tilde{\theta}}(y_f)}{p(y_f)})^\alpha \exp(-\alpha V(\tilde{\theta}|y_{in})).$$

By the definition of the loss function L_α and the summability condition on V for every y_{in} as in (1.67), the above expression is not greater than 1 and hence the expectation of the left side in (1.86) too is not greater than 1. Now by concavity of the logarithm and Jensen's inequality we obtain

$$\mathbb{E}_f \log \left(\exp(\alpha L_\alpha(p_{\theta^\star}, p_{\hat{\theta}}))(\frac{p_{\hat{\theta}}(y)}{p_{\theta^\star}(y)})^\alpha \exp(-\alpha\, pen(\hat{\theta}|y_{in}))\right) \leq 0. \tag{1.87}$$

The above is just α times the first term in brackets in (1.85). Hence from (1.85) and (1.87) we get an upper bound of the expected loss function conditional on y_{in} where expectation is taken over y_f,

$$\mathbb{E}_f L(p_{\theta^\star}, p_{\hat{\theta}}) \leq \mathbb{E}_f \min_{\theta \in \Theta} \left(\log \frac{p_{\theta^\star}(y)}{p_\theta(y)} + pen(\theta | y_{in}) \right).$$

Now by taking expectation with respect to y_{in} we obtain the following risk bound, where $\mathbb{E}$ refers to expectation taken over the whole data,

$$\mathbb{E} L_\alpha(p_{\theta^\star}, p_{\hat{\theta}}) \leq \mathbb{E} \min_{\theta \in \Theta} \left(\log \frac{p_{\theta^\star}(y)}{p_\theta(y)} + pen(\theta | y_{in}) \right).$$

This completes the proof of Lemma (1.4.4). □

Bibliography

[1] A.R. Barron and T.M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*. Vol.37, No.4, pp.1034–1054. 1991.

[2] A.R. Barron, C. Huang, J.Q. Li, X. Luo, "The MDL principle, penalized likelihoods, and statistical risk," In *Festschrift for Jorma Rissanen*. Tampere University Press, Tampere, Finland, 2008.

[3] A.R. Barron, C. Huang, J.Q. Li, X. Luo, "MDL, penalized likelihood and statistical risk," *IEEE Information Theory Workshop*. Porto Portugal, May 4-9, 2008.

[4] A.R. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inform. Theory*. Vol.44, No.6, pp.2743–2760. 1998. Special Commemorative Issue: Information Theory: 1948-1998.

[5] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by probability distributions," *Bull. Calcutta Math. Soc.* Vol.35, pp.99–109. 1943.

[6] P. Grünwald, *The Minimum Description Length Principle*. Cambridge, MA, MIT Press. 2007.

[7] P. Buhlmann S. Van De Geer, "Statistics for high-dimensional data.

Methods, theory and applications," Springer Series in Statistics. Springer, Heidelberg.

[8] J. Rissanen (2004), "An introduction to the MDL principle," Available at www.mdl-research.org.

[9] A. R. Barron, "Complexity Regularization with application to artificial neural networks," "Nonparametric Functional Estimation and Related Topics," G. Roussas (Ed.) pp. 561-576. Dordrecht, the Netherlands: Kluwer Academic Publishers.

[10] A. R. Barron, L. Birgé, P. Massart, "Risk bounds for model selection by penalization," Probability Theory and Related Fields, 1999 Volume 113, pp. 301-413.

[11] L. Birgé and P. Massart, "Rates of convergence for minimum contrast estimator," Probability Theory and Related Fields 1998 Volume 97, 113-150.

[12] L. Birgé and P. Massart, "Minimum contrast estimators on sieves: exponential bounds and rates of convergence," Bernoulli 1998 Volume 4, 329-375.

[13] F. Cucker, and S. Smale, On the mathematical foundations of learning. Bulletin of the American Mathematics Society 2001 Volume 39. pp. 1-49.

[14] D. D. Cox, F. O'Sullivan, "Asymptotic analysis of penalized likelihood and related estimators," Annals of Statistics, 1990 Volume 18, pp. 1676-1695.

[15] G. M. de Montricher, R. Tapia and J. R. Thompson, "Nonparametric maximum likelihood estimation of probability densities by penalty function methods," Annals of Statistics 1975 Volume 3, pp. 1329-1348

[16] E. D. Kolaczyk, R. D. Nowak, "Multiscale likelihood analysis and complexity penalized estimation," Annals of Statistics 2004 Volume 32, pp. 500-527.

[17] E. D. Kolaczyk, R. D. Nowak, "Multiscale generalized linear models for nonparametric function estimation," Biometrika 2005 Volume 92, No. 1, pp. 119-133.

[18] J. Q. Li, "Estimation of Mixture Models," Ph.D. Thesis, 1999, Department of Statistics, Yale University, New Haven, CT.

[19] A. S. Nemirovskii, B. T. Polyak, A. B. Tsybakov, "Rate of convergence of nonparametric estimates of maximum likelihood type," Problems in Information Transmission 1985 Volume 21, pp. 258-272.

[20] X. Shen, "On the method of Penalization," Statistica Sinica, 1998 Volume 8, pp. 337-357.

[21] B. Silverman, "On the estimation of probability function by the maximum penalized likelihood method," Annals of Statistics 1982 Volume 10, 795-810.

[22] R. A. Tapia, J. R. Thompson, "Nonparametric Probability Density Estimation," Baltimore, MD: The Johns Hopkins University Press 1978.

[23] G. Wahba, "Spline Models for Observational Data," Philadelphia, PA: SIAM 1990.

[24] Y. Yang and A. R. Barron, "An asymptotic property of model selection criteria," IEEE Transactions on Information Theory, 1998, Volume 44, pp. 117-133.

[25] Y. Yang and A. R. Barron, "Information-theoretic determination of minimax rates of convergence," Annals of Statistics, 1999 Volume 27, pp. 1564-1599.

[26] S. Negahban, P. Ravikumar, M. J. Wainwright, B. Yu, "A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers," Statistical Science 27 (2012), no. 4, 538–557. doi:10.1214/12-STS400. http://projecteuclid.org/euclid.ss/1356098555.

[27] F. Bunea, A. Tsybakov, M. Wegkamp, "Aggregation for Gaussian regression," The Annals of Statistics 35 (2007), no. 4, 1674–1697 http://projecteuclid.org/euclid.aos/1188405626.

Chapter 2

Improved Risk Bounds in Monotone Regression and other Shape Constraints

We consider the problem of estimating an unknown non-decreasing sequence $\theta \in \mathbb{R}^n$ from a noisy observation. We give an improved global risk upper bound for the isotonic least squares estimator (LSE) in this problem. The obtained risk bound behaves differently depending on the form of the true sequence θ – one gets a whole range of rates from $\log n/n$ (when θ is constant) to $n^{-2/3}$ (when θ is *uniformly increasing* in a certain sense). In particular, when θ has k constant pieces then the risk bound becomes $(k/n)\log(en/k)$. As a consequence, we illustrate the automatic adaptation properties of the LSE. We also derive local minimax lower bounds for this problem which show that the LSE is nearly optimal in a local non-asymptotic minimax sense. We prove an analogue of our risk bound for model misspecification where the true θ is not necessarily non-decreasing. We also derive global risk upper bounds

for the LSE of θ when θ belongs to a known but arbitrary convex polyhedral cone in $\mathbb{R}^n$.

2.1 Introduction

Consider the problem of estimating an unknown non-decreasing regression function f_0 from finitely many noisy observations. Specifically, only under the assumption that f_0 is non-decreasing, the goal is to estimate f_0 from data $(x_1, Y_1), \ldots, (x_n, Y_n)$ with

$$Y_i = f_0(x_i) + \epsilon_i, \qquad \text{for } i = 1, 2, \ldots, n, \tag{2.1}$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d mean zero errors with variance $\sigma^2 > 0$ and $x_1 < \cdots < x_n$ are fixed design points. This is one of the canonical problems in the area of shape-constrained nonparametric function estimation.

The most natural and commonly used estimator for this problem is the monotone least squares estimator (LSE), proposed in [6] and [2]; also see [16] for the related problem of estimating a non-increasing density. The LSE is defined as any minimizer of the LS criterion:

$$\hat{f}_{ls} \in \operatorname*{argmin}_{g \in \mathcal{C}} \sum_{i=1}^{n} (Y_i - g(x_i))^2 \tag{2.2}$$

where $\mathcal{C}$ denotes the set of all real-valued non-decreasing functions on $\mathbb{R}$. The values $\hat{f}_{ls}(x_1), \ldots, \hat{f}_{ls}(x_n)$ are unique and can be computed easily using the *pool adjacent violators algorithm*; see [26, Chapter 1].

The existing theoretical results on monotone function estimation can be grouped into two categories: (1) results on the behavior of the LSE at an interior point

(which is sometimes known as local behavior), and (2) results on the behavior of a global loss function measuring how far $\hat{f}_{ls}$ is from f_0. Before describing our results, let us briefly review the results on the local behavior of the LSE. The results on the behavior of the LSE under a global loss function will be detailed while describing our main results.

Results on the local behavior are proved, among others, in [7], [19], [17], [9], [10], and [22]. Under certain regularity conditions on the unknown function f_0 near the interior point x_0, [7] showed that $\hat{f}_{ls}(x_0)$ converges to $f_0(x_0)$ at the rate $n^{-1/3}$ and also characterized the limiting distribution of $n^{1/3}(\hat{f}_{ls}(x_0) - f_0(x_0))$. In the related (non-increasing) density estimation problem, [17], [9] and [22] showed that if the interior point x_0 lies on a flat stretch of the underlying function then the LSE (which is also the nonparametric maximum likelihood estimator, usually known as the Grenander estimator) converges to a non-degenerate limit at rate $n^{-1/2}$, and they characterized the limiting distribution. [10] showed that the rate of convergence of $\hat{f}_{ls}(x_0)$ to $f_0(x_0)$ depends on the local behavior of f_0 near x_0 and explicitly described this rate for each f_0. In this sense, the LSE $\hat{f}_{ls}$ adapts automatically to the unknown function f_0. [10] also proved optimality of the LSE for local behavior by proving a local asymptotic minimax lower bound.

Often in monotone regression, the interest is in the estimation of the entire function f_0 as opposed to just its value at one fixed point. In this sense, it is more appropriate to study the behavior of $\hat{f}_{ls}$ under a global loss function. The most natural and commonly studied global loss function in this setup is

$$\ell^2(f, g) := \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2$$

for real valued functions f and g. Note that under this loss function, the prob-

lem posited in (2.1) becomes a vector or sequence estimation problem where the goal is to estimate the vector $\theta := (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ from observations

$$Y_i = \theta_i + \epsilon_i, \qquad \text{for } i = 1, 2, \ldots, n, \tag{2.3}$$

under the constraint that the unknown sequence $\theta = (\theta_1, \ldots, \theta_n)$ satisfies $\theta_1 \leq \cdots \leq \theta_n$. In other words, we assume that θ lies in the closed convex polyhedral cone $\mathcal{M}$ defined as

$$\mathcal{M} := \{t \in \mathbb{R}^n : t_1 \leq t_2 \leq \cdots \leq t_n\}.$$

The connection between this sequence estimation problem and the earlier problem of monotone regression is apparent by the identification $\theta_i = f_0(x_i)$. As a result, it is obvious that the exact form of the design points $x_1, \ldots, x_n$ is irrelevant in the problem (2.1) as long as $x_1 < \cdots < x_n$.

Henceforth we would be stating our results for sequences although they could be thought of as being generated from a monotone function sampled at certain design points. The LSE, $\hat{\theta}$, of θ is defined as

$$\hat{\theta} := \operatorname*{argmin}_{t \in \mathcal{M}} \sum_{i=1}^{n} (Y_i - t_i)^2 \tag{2.4}$$

and it is easy to see that it exists and is unique. We study the risk of $\hat{\theta}$ under the loss function $\sum_{i=1}^n (t_i - s_i)^2/n$ which we denote by $\ell^2(t, s)$ by a slight abuse of notation.

The behavior of the LSE $\hat{\theta}$, under the loss ℓ^2, has been studied in a number of papers including [30, 31], [12], [4], [34], [23] and [35]. If one looks at the related (non-increasing) density estimation problem, [5] developed non-

asymptotic risk bounds for the Grenander estimator, measured with the L_1-loss, whereas [31] has results on the Hellinger distance.

The strongest results on the behavior of the LSE $\hat{\theta}$, under the loss ℓ^2, are obtained by [35] who proved non-asymptotic bounds on the risk $\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta)$ under a minimal set of assumptions on the errors $\epsilon_1, \ldots, \epsilon_n$. Among other things, [35, Theorem 2.2] showed that

$$\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta) \lesssim R_Z(n; \theta) \tag{2.5}$$

where

$$R_Z(n; \theta) := \left(\frac{\sigma^2 V(\theta)}{n} \right)^{2/3} + \frac{\sigma^2 \log n}{n}.$$

with

$$V(\theta) := \theta_n - \theta_1.$$

Here, by the symbol $\lesssim$ we mean $\leq$ up to a multiplicative constant. The quantity $V(\theta)$ is the variation of the sequence θ. When $V(\theta) = 0$, i.e., when θ is a constant sequence, this inequality implies that

$$\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta) \lesssim \frac{\sigma^2 \log n}{n}. \tag{2.6}$$

This result is an example of the adaptation behavior of the LSE in the sense that the risk bound when θ is constant is much smaller than the general upper bound (2.5). The rate $\log n/n$ also appears in the risk bound for the Grenander estimator when the underlying function is constant; see [20], [19] and [31]. In fact, it has been shown that in this case, the integrated squared error risk, properly normalized, converges to a normal distribution. Other results on the global behavior of the LSE can be found in [13, 14] who considered a different global loss function and characterized the limiting distribution of a suitably

normalized version.

An important shortcoming of the risk bound (2.5) is that it is not optimal in the sense that the risk $\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta)$ can, in general, be much smaller than $R_Z(n;\theta)$. The main result of this paper presents an improvement of (2.5). To describe our result, let us introduce some notation. We say an *interval partition* π of a positive integer n is a finite sequence of positive integers that sum to n. In combinatorics this is called a composition of n. Let the set of all interval partitions π of n be denoted by Π. Formally, Π can be written as

$$\Pi := \left\{ (n_1, n_2, .., n_k) : k \geq 1, n_i \in \mathbb{N} \text{ and } \sum_{i=1}^{k} n_i = n \right\}.$$

For each $\pi \in \Pi$, let $k(\pi)$ denote the length of the associated sequence $\{n_i\}$.

For each $\theta = (\theta_1, \ldots, \theta_n) \in \mathcal{M}$, there exist integers k and $n_1, \ldots, n_k$ with $n_i \geq 1$ and $n_1 + \cdots + n_k = n$ such that θ is constant on each set $\{j : s_{i-1} + 1 \leq j \leq s_i\}$ for $i = 1, \ldots, k$, where $s_0 := 0$ and $s_i = n_1 + \cdots + n_i$. We refer to the interval partition $\pi_\theta := (n_1, \ldots, n_k)$ as the interval partition *generated* by θ. By a slight abuse of notation, we denote $k(\pi_\theta)$ by simply $k(\theta)$. Note that $k(\theta)$ is just the number of distinct values among $\theta_1, \ldots, \theta_n$. $k(\theta)$ is a measure of sparsity of the differences $\theta_i - \theta_{i-1}$ for $i = 2, \ldots, n$.

For every $\theta \in \mathcal{M}$ and $\pi := (n_1, \ldots, n_k) \in \Pi$, we define

$$V_\pi(\theta) = \max_{1 \leq i \leq k} \left(\theta_{s_i} - \theta_{s_{i-1}+1} \right).$$

where $s_0 := 0$ and $s_i = n_1 + \cdots + n_i$ for $1 \leq i \leq k$. $V_\pi(\theta)$ can be treated as measure of variation of θ with respect to the partition π. An important property is that $V_{\pi_\theta}(\theta) = 0$ for every $\theta \in \mathcal{M}$. For the trivial partition $\pi = (n)$, it is easy to see that $V_\pi(\theta) = V(\theta)$.

We are now ready to state our main result. In Theorem 2.2.1, we prove the following non-asymptotic risk bound:

$$\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta) \leq R(n; \theta) \tag{2.7}$$

where

$$R(n; \theta) := 4 \inf_{\pi \in \Pi} \left(V_\pi^2(\theta) + \frac{4\sigma^2 k(\pi)}{n} \log \frac{en}{k(\pi)} \right). \tag{2.8}$$

Let us explain why our bound (2.7) is an improvement of (2.5) in certain cases. Suppose, for example, $\theta_j = I\{j > n/2\}$ (here I denotes the indicator function) and hence $V(\theta) = 1$. Then $R_Z(n; \theta)$ is essentially $(\sigma^2/n)^{2/3}$ while $R(n; \theta)$ is much smaller because it is at most $(32\sigma^2/n) \log(en/2)$ as can be seen by taking $\pi = \pi_\theta$ in (2.8).

More generally by taking $\pi = \pi_\theta$ in the infimum of the definition of $R(n; \theta)$, we obtain

$$\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta) \leq \frac{16 k(\theta) \sigma^2}{n} \log \frac{en}{k(\theta)}, \tag{2.9}$$

which is a stronger bound than (2.5) when $k(\theta)$ is small. The reader may observe that $k(\theta)$ is small precisely when the differences $\theta_i - \theta_{i-1}$ are sparse.

We prove some properties of our risk bound $R(n; \theta)$ in Section 2.3. In Theorem 2.3.1, we show that $R(n; \theta)$ is bounded from above by a multiple of $R_Z(n; \theta)$ that is at most logarithmic in n. Therefore, our inequality (2.7) is always only slightly worse off than (2.5) while being much better in the case of certain sequences θ. We also show in Section 2.3 that the risk bound $R(n; \theta)$ behaves differently depending on the form of the true sequence θ. This means that the bound (2.7) demonstrates adaptive behavior of the LSE. One gets a whole range of rates from $(\log n)/n$ (when θ is constant) to $n^{-2/3}$ up to logarithmic factors in the worst case (this worst case rate corresponds to the

situation where $\min_i(\theta_i - \theta_{i-1}) \gtrsim 1/n$). The bound (2.7) therefore presents a bridge between the two terms in the bound (2.5).

In addition to being an upper bound for the risk of the LSE, we believe that the quantity $R(n;\theta)$ also acts as a benchmark for the risk of any estimator in monotone regression. By this, we mean that, in a certain sense, no estimator can have risk that is significantly better than $R(n;\theta)$. We substantiate this claim in Section 2.4 by proving lower bounds for the *local minimax risk* near the "true" θ. For $\theta \in \mathcal{M}$, the quantity

$$\mathfrak{R}_n(\theta) := \inf_{\hat{t}} \sup_{t \in \mathfrak{N}(\theta)} \mathbb{E}_t \ell^2(t, \hat{t})$$

with

$$\mathfrak{N}(\theta) := \left\{ t \in \mathcal{M} : \ell_\infty^2(t, \theta) \lesssim R(n;\theta) \right\}$$

will be called the local minimax risk at θ (see Section 2.4 for the rigorous definition of the neighborhood $\mathfrak{N}(\theta)$ where the multiplicative constants hidden by the $\lesssim$ sign are explicitly given). In the above display ℓ_∞ is defined as $\ell_\infty(t, \theta) := \max_i |t_i - \theta_i|$. The infimum here is over all possible estimators $\hat{t}$. $\mathfrak{R}_n(\theta)$ represents the smallest possible (supremum) risk under the knowledge that the true sequence t lies in the neighborhood $\mathfrak{N}(\theta)$. It provides a measure of the difficulty of estimation of θ. Note that the size of the neighborhood $\mathfrak{N}(\theta)$ changes with θ (and with n) and also reflects the difficulty level of the problem.

Under each of two following setups for θ, and the assumption of normality of the errors, we show that $\mathfrak{R}_n(\theta)$ is bounded from below by $R(n;\theta)$ up to multiplicative logarithmic factors of n. Specifically,

1. when the increments of θ (defined as $\theta_i - \theta_{i-1}$, for $i = 2, \ldots, n$) grow like

$1/n$, we prove in Theorem 2.4.3 that

$$\mathfrak{R}_n(\theta) \gtrsim \left(\frac{\sigma^2 V(\theta)}{n}\right)^{2/3} \gtrsim \frac{R(n;\theta)}{\log(4n)}; \tag{2.10}$$

2. when $k(\theta) = k$ and the k values of θ are sufficiently well-separated, we show in Theorem 2.4.4 that

$$\mathfrak{R}_n(\theta) \gtrsim R(n;\theta)\left(\log\frac{en}{k}\right)^{-2/3}. \tag{2.11}$$

Because $R(n,\theta)$ is an upper bound for the risk of the LSE and also is a local minimax lower bound in the above sense, our results imply that the LSE is near-optimal in a local non-asymptotic minimax sense. Such local minimax bounds are in the spirit of [10] and [8] who worked with the problems of estimating monotone and convex functions respectively at a point. The difference between these works to ours is that we focus on the global estimation problem. In other words, [10] and [8] prove local minimax bounds for the local (pointwise) estimation problem while we prove local minimax bounds for the global estimation problem. On the other hand, global minimax bounds for the global estimation problem in isotonic shape-constrained problems can be found in [5].

We also study the performance of the LSE under model misspecification when the true sequence θ is not necessarily non-decreasing. Here we prove in Theorem 2.5.1 that $\mathbb{E}_\theta \ell^2(\hat{\theta}, \tilde{\theta}) \leq R(n;\tilde{\theta})$ where $\tilde{\theta}$ denotes the non-decreasing projection of θ (see Section 2.5 for its definition). This should be contrasted with the risk bound of [35] who proved that $\mathbb{E}_\theta \ell^2(\hat{\theta}, \tilde{\theta}) \lesssim R_Z(n;\tilde{\theta})$. As before our risk bound is at most slightly worse (by a multiplicative logarithmic factor in n) than R_Z but is much better when $k(\tilde{\theta})$ is small. We describe two situa-

tions where $k(\tilde{\theta})$ is small — when θ itself has few constant blocks (see (2.58) and Lemma 2.5.4) and when θ is non-increasing (in which case $k(\tilde{\theta}) = 1$; see Lemma 2.5.3).

Till now we have studied the risk behavior of the LSE $\hat{\theta}$, defined through (2.4), when $\mathcal{M}$ is the convex cone of all non-decreasing sequences. A natural question that arises is whether our results can be extended to general polyhedral cones. In Section 2.6 we give a generalization of our main result, Theorem 2.2.1, to more general convex cones. Our main result in this section, Theorem 2.6.1, shows that Theorem 2.2.1 can be extended to certain special kinds of polyhedral cones, which include as special cases isotonic regression and convex regression (see [18] and [21]). In fact, the results leading to Theorem 2.6.1 hold for any closed convex cone and are of independent interest. Our proof technique is completely new and very different from that of Theorem 2.2.1, where we crucially use the known analytic expression for $\hat{\theta}$. We use the characterizing properties of the projection operator on a closed convex cone to derive a general risk bound, under the additional assumption of the normality of the errors $\epsilon_1, \ldots, \epsilon_n$.

The paper is organized as follows: In Section 2.2 we state and prove our main upper bound for the risk $\mathbb{E}_\theta \ell^2(\theta, \hat{\theta})$. We investigate the behavior of $R(n; \theta)$ for different values of the true sequence θ and compare it with $R_Z(n; \theta)$ in Section 2.3. Local minimax lower bounds for $\mathfrak{R}_n(\theta)$ for two different scenarios for θ are proved in Section 2.4. We study the performance of the LSE under model misspecification in Section 2.5. In Section 2.6 we provide a generalization of Theorem 2.2.1 to more general polyhedral cones. Section 2.7 gives some auxiliary results needed in the proofs of our main results.

2.2 Our risk bound

In the following theorem, we present our risk bound (2.7); in fact, we prove a slightly stronger inequality than (2.7). Before stating our main result we introduce some notation. For any sequence $(a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$ and any $1 \le k \le l \le n$, let

$$\bar{a}_{k,l} := \frac{1}{l-k+1} \sum_{j=k}^{l} a_j. \tag{2.12}$$

We will use this notation mainly when a equals either the sequence Y or θ. Our proof uses similar ideas as in Section 2 of [35] and is based on the following explicit representation of the LSE $\hat{\theta}$ (see Chapter 1 of [26]):

$$\hat{\theta}_j = \min_{l \ge j} \max_{k \le j} \bar{Y}_{k,l}. \tag{2.13}$$

For $x \in \mathbb{R}$, we write $x_+ := \max\{0, x\}$ and $x_- := -\min\{0, x\}$. For $\theta \in \mathcal{M}$ and $\pi = (n_1, \ldots, n_k) \in \Pi$, let

$$D_\pi(\theta) = \left(\frac{1}{n} \sum_{i=1}^{k} \sum_{j=s_{i-1}+1}^{s_i} \left(\theta_j - \bar{\theta}_{s_{i-1}+1, s_i}\right)^2 \right)^{1/2}$$

where $s_0 = 0$ and $s_i = n_1 + \cdots + n_i$ for $1 \le i \le k$. Like $V_\pi(\theta)$, this quantity $D_\pi(\theta)$ can also be treated as a measure of the variation of θ with respect to π. This measure also satisfies $D_{\pi_\theta}(\theta) = 0$ for every $\theta \in \mathcal{M}$. Moreover

$$D_\pi(\theta) \le V_\pi(\theta) \qquad \text{for every } \theta \in \mathcal{M} \text{ and } \pi \in \Pi.$$

When $\pi = (n)$ is the trivial partition, $D_\pi(\theta)$ turns out to be just the standard deviation of θ. In general, $D_\pi^2(\theta)$ is analogous to the within group sum of squares term in ANOVA with the blocks of π being the groups. Below, we

prove a stronger version of (2.7) with $D_\pi(\theta)$ replacing $V_\pi(\theta)$ in (2.7).

Theorem 2.2.1. *For every $\theta \in \mathcal{M}$, the risk of the LSE satisfies the following inequality:*

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}) \leq 4 \inf_{\pi \in \Pi} \left(D_\pi^2(\theta) + \frac{4\sigma^2 k(\pi)}{n} \log \frac{en}{k(\pi)} \right). \tag{2.14}$$

Proof. Fix $1 \leq j \leq n$ and $0 \leq m \leq n - j$. By (2.13), we have

$$\hat{\theta}_j = \min_{l \geq j} \max_{k \leq j} \bar{Y}_{k,l} \leq \max_{k \leq j} \bar{Y}_{k,j+m} = \max_{k \leq j} \left(\bar{\theta}_{k,j+m} + \bar{\epsilon}_{k,j+m} \right)$$

where, in the last equality, we used $\bar{Y}_{k,l} = \bar{\theta}_{k,l} + \bar{\epsilon}_{k,l}$. By the monotonicity of θ, we have $\bar{\theta}_{k,j+m} \leq \bar{\theta}_{j,j+m}$ for all $k \leq j$. Therefore, for every $\theta \in \mathcal{M}$, we get

$$\hat{\theta}_j - \theta_j \leq (\bar{\theta}_{j,j+m} - \theta_j) + \max_{k \leq j} \bar{\epsilon}_{k,j+m}.$$

Taking positive parts, we have

$$\left(\hat{\theta}_j - \theta_j \right)_+ \leq (\bar{\theta}_{j,j+m} - \theta_j) + \max_{k \leq j} \left(\bar{\epsilon}_{k,j+m} \right)_+ .$$

Squaring and taking expectations on both sides, we obtain

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \theta_j \right)_+^2 \leq \mathbb{E}_\theta \left((\bar{\theta}_{j,j+m} - \theta_j) + \max_{k \leq j} \left(\bar{\epsilon}_{k,j+m} \right)_+ \right)^2 .$$

Using the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$ we get

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \theta_j \right)_+^2 \leq 2 \left(\bar{\theta}_{j,j+m} - \theta_j \right)^2 + 2\mathbb{E} \max_{k \leq j} \left(\bar{\epsilon}_{k,j+m} \right)_+^2 .$$

We observe now that, for fixed integers j and m, the process $\{\bar{\epsilon}_{k,j+m}, k = 1, \ldots, j\}$ is a martingale with respect to the filtration $\mathcal{F}_1, \ldots, \mathcal{F}_j$ where $\mathcal{F}_i$

is the sigma-field generated by the random variables $\epsilon_1, \ldots, \epsilon_{i-1}$ and $\bar{\epsilon}_{i,j+m}$. Therefore, by Doob's inequality for submartingales (see e.g., Theorem 5.4.3 of [15]), we have

$$\mathbb{E} \max_{k \le j} \left(\bar{\epsilon}_{k,j+m}\right)_+^2 \le 4\mathbb{E} \left(\bar{\epsilon}_{j,j+m}\right)_+^2 \le 4\mathbb{E} \left(\bar{\epsilon}_{j,j+m}\right)^2 \le \frac{4\sigma^2}{m+1}.$$

So using the above result we get a pointwise upper bound for the positive part of the risk

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \theta_j\right)_+^2 \le 2 \left(\bar{\theta}_{j,j+m} - \theta_j\right)^2 + \frac{8\sigma^2}{m+1}. \tag{2.15}$$

Note that the above upper bound holds for any arbitrary m, $0 \le m \le n - j$. By a similar argument we can get the following pointwise upper bound for the negative part of risk which now holds for any m, $0 \le m \le j$:

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \theta_j\right)_-^2 \le 2 \left(\theta_j - \bar{\theta}_{j-m,j}\right)^2 + \frac{8\sigma^2}{m+1}. \tag{2.16}$$

Let us now fix $\pi = (n_1, \ldots, n_k) \in \Pi$. Let $s_0 := 0$ and $s_i := n_1 + \cdots + n_i$ for $1 \le i \le k$. For each $j = 1, \ldots, n$, we define two integers $m_1(j)$ and $m_2(j)$ in the following way: $m_1(j) = s_i - j$ and $m_2(j) = j - 1 - s_{i-1}$ when $s_{i-1} + 1 \le j \le s_i$. We use this choice of $m_1(j)$ in (2.15) and $m_2(j)$ in (2.16) to obtain

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \theta_j\right)^2 \le A_j + B_j$$

where

$$A_j := 2 \left(\bar{\theta}_{j,j+m_1(j)} - \theta_j\right)^2 + \frac{8\sigma^2}{m_1(j)+1}$$

and

$$B_j := 2 \left(\theta_j - \bar{\theta}_{j-m_2(j),j}\right)^2 + \frac{8\sigma^2}{m_2(j)+1}.$$

This results in the risk bound

$$\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta) \leq \frac{1}{n} \sum_{j=1}^{n} A_j + \frac{1}{n} \sum_{j=1}^{n} B_j. \tag{2.17}$$

We shall now prove that

$$\frac{1}{n} \sum_{j=1}^{n} A_j \leq 2D_\pi^2(\theta) + \frac{8k\sigma^2}{n} \log \frac{en}{k} \tag{2.18}$$

and

$$\frac{1}{n} \sum_{j=1}^{n} B_j \leq 2D_\pi^2(\theta) + \frac{8k\sigma^2}{n} \log \frac{en}{k}. \tag{2.19}$$

We give below the proof of (2.18) and the proof of (2.19) is nearly identical. Using the form of A_j, we break up $\frac{1}{n}\sum_{j=1}^{n} A_j$ into two terms. For the first term, note that $j + m_1(j) = s_i$, for $s_{i-1} + 1 \leq j \leq s_i$ and therefore

$$\sum_{j=1}^{n} \left(\bar{\theta}_{j,j+m_1(j)} - \theta_j\right)^2 = \sum_{i=1}^{k} \sum_{j=s_{i-1}+1}^{s_i} \left(\bar{\theta}_{j,s_i} - \theta_j\right)^2.$$

By Lemma 2.7.2, we get

$$\sum_{j=s_{i-1}+1}^{s_i} \left(\bar{\theta}_{j,s_i} - \theta_j\right)^2 \leq \sum_{j=s_{i-1}+1}^{s_i} \left(\bar{\theta}_{s_{i-1}+1,s_i} - \theta_j\right)^2$$

for every $i = 1, \ldots, k$. Thus, summing over $i = 1, \ldots, k$, and multiplying by $2/n$ proves that the first term in $\frac{1}{n}\sum_{j=1}^{n} A_j$ is bounded from above by $2D_\pi^2(\theta)$. To bound the second term, we write

$$\sum_{j=1}^{n} \frac{1}{m_1(j)+1} = \sum_{i=1}^{k} \sum_{j=s_{i-1}+1}^{s_i} \frac{1}{s_i - j + 1} = \sum_{i=1}^{k} \left(1 + \frac{1}{2} + \cdots + \frac{1}{n_i}\right). \tag{2.20}$$

Since the harmonic series $\sum_{i=1}^{l} 1/l$ is at most $1 + \log l$ for $l \geq 1$, we obtain

$$\sum_{j=1}^{n} \frac{1}{m_1(j)+1} \leq k + \sum_{i=1}^{k} \log n_i \leq k + k \log \left(\frac{\sum_i n_i}{k} \right)$$

where the last inequality is a consequence of the concavity of the logarithm function. We thus obtain

$$\sum_{j=1}^{n} \frac{1}{m_1(j)+1} \leq k \log \frac{en}{k}.$$

This proves (2.18). Combining (2.18) and (2.19) proves the theorem. □

Remark 2.2.1. *Theorem 2.2.1 can be restated in the following way. For each $1 \leq k \leq n$, let $\mathcal{P}_k$ denote the set of all sequences $\alpha \in \mathcal{M}$ with $k(\alpha) \leq k$. With this notation, the statement of Theorem 2.2.1 can also be expressed as*

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}) \leq 4 \min_{1 \leq k \leq n} \left[\inf_{\alpha \in \mathcal{P}_k} \ell^2(\theta, \alpha) + \frac{4\sigma^2 k}{n} \log \frac{en}{k} \right]. \tag{2.21}$$

This follows from the fact that $\min_{\alpha \in \mathcal{P}_k} \ell^2(\theta, \alpha) = \min_{\pi \in \Pi_k} D_\pi^2(\theta)$ where Π_k is the set of all $\pi \in \Pi$ with $k(\pi) \leq k$.

The bound (2.21) reflects adaptation of the LSE with respect to the classes $\mathcal{P}_k$. Such risk bounds are usually provable for estimators based on empirical model selection criteria (see, for example, [3]) or aggregation (see, for example, [25]). Specializing to the present situation, in order to adapt over $\mathcal{P}_k$ as k varies, one constructs LSEs over each $\mathcal{P}_k$ and then either selects one estimator from this collection by an empirical model selection criterion or aggregates these estimators with data-dependent weights. In this particular situation, such estimators are very difficult to compute as minimizing the LS criterion over $\mathcal{P}_k$ is a non-convex optimization problem. In contrast, the LSE can be easily computed by

a convex optimization problem. It is remarkable that the LSE which is constructed with no explicit model selection criterion in mind achieves the adaptive risk bound $R(n;\theta)$. This adaptation property of the LSE in global estimation holds more generally; see Section 2.6 where we prove a more general version of Theorem 2.2.1 which includes other shape constrained problems such as convex regression.

Remark 2.2.2. *Note that $k(\theta)$ does not have to be small for (2.7) to be an improvement of (2.5). One only needs that $V_\pi(\theta)$ be small for some partition π with small $k(\pi)$. We demonstrate this with the following example. Suppose $\theta_i = -2^{-i+1}$ for $i = 1, \ldots, n$, so that $k(\theta) = n$ and $V(\theta) = 1 - 2^{-n+1}$. Because $V(\theta) \approx 1$ for large n, the upper bound (2.5) is of the order $(\sigma^2/n)^{2/3}$.*

On the other hand, the bound (2.7) will be much smaller as shown below. Assume for simplicity that $n = 2^k$ for some k. Let π_0 be the partition given by $k(\pi_0) = \log_2 n = k$, $n_1 = \ldots = n_{k-1} = 1$, and $n_k = n - (k-1)$. Thus, $s_i = i$, for $i = 0, \ldots, k-1$, and $s_k = n$. Further,

$$V_{\pi_0}(\theta) = \max_{1 \le i \le k} (\theta_{s_i} - \theta_{s_{i-1}+1}) = (\theta_n - \theta_k) < \frac{2}{2^k} = \frac{2}{n}.$$

Therefore, using (2.7),

$$\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta) \le 16 \left(\frac{1}{n^2} + \frac{\sigma^2 \log_2 n}{n} \log \frac{en}{\log_2 n} \right),$$

which is much smaller than the bound given by (2.5).

Example 2.2.2. *We prove in Theorem 2.3.1 in the next section that the bound given by Theorem 2.2.1 is always smaller than a logarithmic multiplicative factor of the usual cube root rate of convergence for every $\theta \in \mathcal{M}$ with $V(\theta) > 0$. Here, we shall demonstrate this in the special case of the sequence $\theta =$*

$(1/n, 2/n, \ldots, 1)$ *where the bound in* (2.14) *can be calculated exactly. Indeed, if* $\pi = (n_1, \ldots, n_k)$ *with* $n_i \geq 1$ *and* $\sum_{i=1}^k n_i = n$*, direct calculation gives*

$$D_\pi^2(\theta) = \frac{1}{12n^3}\left(\sum_{i=1}^k n_i^3 - n\right).$$

Now Holder's inequality gives $n = \sum_{i=1}^k n_i \leq (\sum_{i=1}^k n_i^3)^{1/3} k^{2/3}$ *which means that* $\sum_{i=1}^k n_i^3 \geq n^3/k^2$*. Therefore, for every fixed* $k \in \{1, \ldots, n\}$ *such that* n/k *is an integer,* $D_\pi^2(\theta)$ *is minimized over all partitions* π *with* $k(\pi) = k$ *when* $n_1 = n_2 = \cdots = n_k = n/k$*. This gives*

$$\inf_{\pi: k(\pi)=k} D^2(\pi) = \frac{1}{12}\left(\frac{1}{k^2} - \frac{1}{n^2}\right).$$

As a consequence, Theorem 2.2.1 yields the bound

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}) \leq \frac{1}{3} \inf_{k: n/k \in \mathbb{Z}} \left(\frac{1}{k^2} - \frac{1}{n^2} + \frac{48\sigma^2 k}{n} \log(en/k)\right).$$

Now with the choice $k \sim (n/\sigma^2)^{1/3}$*, we get the cube root rate for* $\hat{\theta}$ *up to logarithmic multiplicative factors in* n*. We generalize this to arbitrary* $\theta \in \mathcal{M}$ *with* $V(\theta) > 0$ *in Theorem 2.3.1.*

2.3 The quantity $R(n; \theta)$

In this section, we prove some results about our risk bound $R(n; \theta)$. In the first result below, we prove that $R(n; \theta)$ is always bounded from above by $R_Z(n; \theta)$ up to a logarithmic multiplicative factor in n. This implies that our risk bound (2.7) for the LSE is always only slightly worse off than (2.5) (by a logarithmic multiplicative factor) while being much better when θ is well-approximable by some $\alpha \in \mathcal{M}$ for which $k(\alpha)$ is small.

The proof of the next Theorem relies on Lemma 2.7.1 which is a result on the approximation of non-decreasing sequences by non-decreasing sequences with fewer pieces. Recall that $V(\theta) := \theta_n - \theta_1$.

Theorem 2.3.1. *For every $\theta \in \mathcal{M}$, we have*

$$R(n;\theta) \leq 16 \log(4n) \left(\frac{\sigma^2 V(\theta)}{n} \right)^{2/3} \tag{2.22}$$

whenever

$$n \geq \max \left(2, \frac{8\sigma^2}{V^2(\theta)}, \frac{V(\theta)}{\sigma} \right). \tag{2.23}$$

Proof. Fix $\theta \in \mathcal{M}$ and suppose (2.23) holds. Let $V = V(\theta)$. For every fixed $\delta \in [2V/n, V]$, we use Lemma 2.7.1 to assert the existence of π_δ such that $V_{\pi_\delta}(\theta) \leq \delta$ and

$$k(\pi_\delta) \leq \left\lceil \frac{V}{\delta} \right\rceil \leq \frac{V}{\delta} + 1 \leq \frac{2V}{\delta}.$$

By the definition of $R(n;\theta)$, we have

$$R(n;\theta) \leq 4 \left(V_{\pi_\delta}^2(\theta) + \frac{4\sigma^2 k(\pi_\delta)}{n} \log \frac{en}{k(\pi_\delta)} \right).$$

Note that the function $x \mapsto x \log(en/x)$ is non-decreasing for $x \in (0, n]$ and because $k(\pi_\delta) \leq 2V/\delta \leq n$, we deduce

$$R(n;\theta) \leq 4 \left(\delta^2 + \frac{8V\sigma^2}{n\delta} \log \frac{en\delta}{2V} \right) \leq 4 \left(\delta^2 + \frac{8V\sigma^2}{n\delta} \log \frac{en}{2} \right) \tag{2.24}$$

where in the last inequality above, we have used $\delta \leq V$. We now make the choice $\delta_0 = (8V\sigma^2/n)^{1/3}$. The constraint $2V/n \leq \delta_0 \leq V$ is guaranteed by the sample size condition in the statement of the theorem. The right hand side of (2.24) with $\delta = \delta_0$ gives the right hand side of (2.22), which completes the proof. □

In the next result, we characterize $R(n;\theta)$ for certain strictly increasing sequences θ where we show that it is essentially of the order $(\sigma^2 V(\theta)/n)^{2/3}$. In some sense, $R(n;\theta)$ is maximized for these strictly increasing sequences. The prototypical sequence we have in mind here is $\theta_i = i/n$ for $1 \le i \le n$.

Theorem 2.3.2. *Suppose* $\theta_1 < \theta_2 < \ldots < \theta_n$ *with*

$$\min_{2\le i\le n} (\theta_i - \theta_{i-1}) \ge \frac{c_1 V(\theta)}{n} \tag{2.25}$$

for a positive constant $c_1 \le 1$. *Then we have*

$$12\left(\frac{c_1\sigma^2 V(\theta)}{n}\right)^{2/3} \le R(n,\theta) \le 16\left(\frac{\sigma^2 V(\theta)}{n}\right)^{2/3}\log(4n) \tag{2.26}$$

provided

$$n \ge \max\left(2, \frac{8\sigma^2}{V^2(\theta)}, \frac{2V(\theta)}{\sigma}\right). \tag{2.27}$$

Proof. For notational convenience we write V for $V(\theta)$. The upper bound in (2.26) follows from Theorem 2.3.1. Note that the upper bound therefore does not need the assumption (2.25).

For the lower bound in (2.26), fix $\pi = (n_1, \ldots, n_k) \in \Pi$. Let $s_0 = 0$ and $s_i = n_1 + \cdots + n_i$ for $1 \le i \le k$. By (2.25), we have

$$(\theta_{s_i} - \theta_{s_{i-1}+1}) \ge \sum_{l=s_{i-1}+2}^{s_i} (\theta_l - \theta_{l-1}) \ge \frac{c_1 V}{n}(n_i - 1)$$

for each $1 \le i \le k$. Consequently,

$$V_\pi(\theta) \ge \frac{c_1 V}{n}\max_{1\le i\le k}(n_i - 1) \ge \frac{c_1 V}{nk}\sum_{i=1}^{k}(n_i - 1) = \frac{c_1 V(n-k)}{nk}.$$

As a result

$$R(n;\theta) \geq 4 \inf_{1\leq k\leq n} \left(\frac{c_1^2 V^2 (n-k)^2}{n^2 k^2} + \frac{4\sigma^2 k}{n} \log \frac{en}{k} \right).$$

Writing $x = k/n$ and using $\log(en/k) \geq 1$, we get $R(n;\theta) \geq 4 \inf_{0\leq x\leq 1} h(x)$ where

$$h(x) := \frac{c_1^2 V^2 (1-x)^2}{n^2 x^2} + 4\sigma^2 x.$$

Now $h(x) \geq 4\sigma^2 x \geq 2\sigma^2$ for $x \in [1/2, 1]$. On the other hand, for $x \in [0, 1/2]$, because $1 - x \geq 1/2$, we get

$$h(x) \geq \inf_{x\geq 0} \left(\frac{c_1^2 V^2}{4n^2 x^2} + 4\sigma^2 x \right),$$

which, by calculus, gives

$$h(x) \geq \ 3 \left(\frac{c_1 \sigma^2 V}{n} \right)^{2/3}.$$

Note that

$$3 \left(\frac{c_1 \sigma^2 V}{n} \right)^{2/3} \leq 2\sigma^2$$

whenever $n \geq 2V/\sigma \geq 3^{3/2} c_1 V/(2^{3/2}\sigma)$. Thus, under this condition, we obtain

$$R(n;\theta) \geq 4 \inf_{0\leq x\leq 1} h(x) \geq 12 \left(\frac{c_1 \sigma^2 V}{n} \right)^{2/3}$$

which proves the lower bound in (2.26). The proof is complete. □

Remark 2.3.1. *An important situation where (2.25) is satisfied is when θ arises from sampling a function on $[0, 1]$ at the points i/n for $i = 1, \ldots, n$, assuming that the derivative of the function is bounded from below by a positive constant.*

Next we describe sequences θ for which $R(n;\theta)$ is $(k(\theta)\sigma^2/n)\log(en/k(\theta))$, up to multiplicative factors. For these sequences our risk bound is potentially far superior to $R_Z(n;\theta)$.

Theorem 2.3.3. *Let $k = k(\theta)$ with*

$$\{y : y = \theta_j \text{ for some } j\} = \{\theta_{0,1}, \ldots, \theta_{0,k}\}$$

where $\theta_{0,1} < \cdots < \theta_{0,k}$. Then

$$\frac{\sigma^2 k}{n} \log \frac{en}{k} \le R(n;\theta) \le \frac{16\sigma^2 k}{n} \log \frac{en}{k} \tag{2.28}$$

provided

$$\min_{2 \le i \le k} (\theta_{0,i} - \theta_{0,i-1}) \ge \sqrt{\frac{k\sigma^2}{n} \log \frac{en}{k}}. \tag{2.29}$$

Proof. The upper bound in (2.28) is easily proved by taking $\pi = \pi_\theta$. For the lower bound, let us fix $\pi \in \Pi$. If $n \ge k(\pi) \ge k$, then

$$V_\pi^2(\theta) + \frac{4\sigma^2 k(\pi)}{n} \log \frac{en}{k(\pi)} \ge \frac{4\sigma^2 k(\pi)}{n} \log \frac{en}{k(\pi)} \ge \frac{4\sigma^2 k}{n} \log \frac{en}{k} \tag{2.30}$$

where we have used the fact that $m \mapsto (m/n)\log(en/m)$ is non-decreasing for $1 \le m \le n$. On the other hand, if $k(\pi) < k$, then it is easy to see that

$$V_\pi(\theta) \ge \min_{2 \le i \le k} (\theta_{0,i} - \theta_{0,i-1})$$

which implies, by (2.29), that

$$V_\pi^2(\theta) \ge \frac{k\sigma^2}{n} \log \frac{en}{k}.$$

Thus

$$V_\pi^2(\theta) + \frac{4\sigma^2 k(\pi)}{n} \log \frac{en}{k(\pi)} \geq V_\pi^2(\theta) \geq \frac{k\sigma^2}{n} \log \frac{en}{k} \tag{2.31}$$

for every $\pi \in \Pi$ with $k(\pi) < k$. The inequalities (2.30) and (2.31) therefore imply that

$$\inf_{\pi \in \Pi} \left(V_\pi^2(\theta) + \frac{4\sigma^2 k(\pi)}{n} \log \frac{en}{k(\pi)} \right) \geq \frac{k\sigma^2}{n} \log \frac{en}{k}$$

and this completes the proof. □

2.4 Local minimax optimality of the LSE

In this section, we establish an optimality property of the LSE. Specifically, we show that $\hat{\theta}$ is locally minimax optimal in a non-asymptotic sense. "Local" here refers to a ball $\{t : \ell_\infty^2(t, \theta) \leq cR(n; \theta)\}$ around the true parameter θ for a positive constant c. The reason we focus on local minimaxity as opposed to the more traditional notion of global minimaxity is that the rate $R(n; \theta)$ changes with θ. Note that, moreover, lower bounds on the global minimax risk follow from our local minimax lower bounds. Such an optimality theory based on local minimaxity has been pioneered by [8] and [10] for the problem of estimating a convex or monotone function at a point.

We start by proving an upper bound for the local supremum risk of $\hat{\theta}$. Recall that $\ell_\infty(t, \theta) := \max_{1 \leq i \leq n} |t_i - \theta_i|$.

Lemma 2.4.1. *The following inequality holds for every $\theta \in \mathcal{M}$ and $c > 0$*

$$\sup_{t \in \mathcal{M} : \ell_\infty^2(t,\theta) \leq cR(n;\theta)} \mathbb{E}_t \ell^2(t, \hat{\theta}) \leq 2(1 + 4c) R(n; \theta). \tag{2.32}$$

Proof. Inequality (2.7) gives $\mathbb{E}_t \ell^2(t, \hat{\theta}) \leq R(n; t)$ for every $t \in \mathcal{M}$. Fix $\pi \in \Pi$.

By the triangle inequality, we get

$$V_\pi(t) \le 2\ell_\infty(t,\theta) + V_\pi(\theta).$$

As a result, whenever $\ell_\infty(t,\theta) \le cR(n;\theta)$, we obtain

$$V_\pi^2(t) \le 2V_\pi^2(\theta) + 8\ell_\infty^2(t,\theta) \le 2V_\pi^2(\theta) + 8cR(n;\theta).$$

As a consequence,

$$\begin{aligned}\mathbb{E}_t\ell^2(t,\hat{\theta}) \le R(n;t) &\le \inf_{\pi\in\Pi}\left(2V_\pi^2(\theta) + \frac{4\sigma^2 k(\pi)}{n}\log\frac{n}{k(\pi)}\right) + 8cR(n;\theta)\\ &\le 2R(n;\theta) + 8cR(n;\theta).\end{aligned}$$

This proves (2.32). □

We now show that $R(n;\theta)$, up to logarithmic factors in n, is a lower bound for the local minimax risk at θ, defined as the infimum of the right hand side of (2.32) over all possible estimators $\hat{\theta}$. We prove this under each of the assumptions 1 and 2 (stated in the Introduction) on θ. Specifically, we prove the two inequalities (2.10) and (2.11). These results mean that, when θ satisfies either of the two assumptions 1 or 2, no estimator can have a supremum risk significantly better than $R(n;\theta)$ in the local neighborhood $\{t \in \mathcal{M} : \ell_\infty^2(t,\theta) \lesssim R(n;\theta)\}$. On the other hand, Lemma 2.4.1 states that the supremum risk of the LSE over the same local neighborhood is bounded from above by a constant multiple of $R(n;\theta)$. Putting these two results together, we deduce that the LSE is approximately locally non-asymptotically minimax for such sequences θ. We use the qualifier "approximately" here because of the presence of logarithmic factors on the right hand sides of (2.10) and (2.11).

We make here the assumption that the errors $\epsilon_1, \ldots, \epsilon_n$ are independent and normally distributed with mean zero and variance σ^2. For each $\theta \in \mathcal{M}$, let $\mathbb{P}_\theta$ denote the joint distribution of the data $Y_1, \ldots, Y_n$ when the true sequence equals θ. As a consequence of the normality of the errors, we have

$$D(\mathbb{P}_\theta \| \mathbb{P}_t) = \frac{n}{2\sigma^2} \ell^2(t, \theta)$$

where $D(P\|Q)$ denotes the Kullback-Leibler divergence between the probability measures P and Q. Our main tool for the proofs is Assouad's lemma, the following version of which is a consequence of Lemma 24.3 of [32, pp. 347].

Lemma 2.4.2 (Assouad)**.** *Let m be a positive integer and suppose that, for each $\tau \in \{0,1\}^m$, there is an associated non-decreasing sequence θ^τ in $N(\theta)$, where $N(\theta)$ is a neighborhood of θ. Then the following inequality holds:*

$$\inf_{\hat{t}} \sup_{t \in N(\theta)} \mathbb{E}_t \ell^2(t, \hat{t}) \geq \frac{m}{8} \min_{\tau \neq \tau'} \frac{\ell^2(\theta^\tau, \theta^{\tau'})}{\Upsilon(\tau, \tau')} \min_{\Upsilon(\tau,\tau')=1} \left(1 - \|\mathbb{P}_{\theta^\tau} - \mathbb{P}_{\theta^{\tau'}}\|_{TV}\right),$$

where $\Upsilon(\tau, \tau') := \sum_i I\{\tau_i \neq \tau_i'\}$ denotes the Hamming distance between τ and τ' and $\|\cdot\|_{TV}$ denotes the total variation distance between probability measures. The infimum here is over all possible estimators $\hat{t}$.

The two inequalities (2.10) and (2.11) are proved in the next two subsections.

2.4.1 Uniform increments

In this section, we assume that θ is a strictly increasing sequence with $V(\theta) = \theta_n - \theta_1 > 0$ and that

$$\frac{c_1 V(\theta)}{n} \leq \theta_i - \theta_{i-1} \leq \frac{c_2 V(\theta)}{n} \qquad \text{for } i = 2, \ldots, n \tag{2.33}$$

for some $c_1 \in (0,1]$ and $c_2 \geq 1$. Because $V(\theta) = \sum_{i=2}^{n}(\theta_i - \theta_{i-1})$, assumption (2.6.3) means that the increments of θ are in a sense uniform. An important example in which (2.6.3) is satisfied is when $\theta_i = f_0(i/n)$ for some function f_0 on $[0,1]$ whose derivative is uniformly bounded from above and below by positive constants.

In the next theorem, we prove that the local minimax risk at θ is bounded from below by $R(n;\theta)$ (up to logarithmic multiplicative factors) when θ satisfies (2.6.3).

Theorem 2.4.3. *Suppose θ satisfies (2.6.3) and let*

$$\mathfrak{N}(\theta) := \left\{ t \in \mathcal{M} : \ell_\infty^2(t,\theta) \leq \left(\frac{3c_2}{c_1}\right)^{2/3} \frac{R(n;\theta)}{12} \right\}.$$

Then the local minimax risk $\mathfrak{R}_n(\theta) := \inf_{\hat{t}} \sup_{t \in \mathfrak{N}(\theta)} \mathbb{E}_t \ell^2(t,\hat{t})$ satisfies the following inequality

$$\mathfrak{R}_n(\theta) \geq \frac{c_1^2 3^{2/3}}{256 c_2^{4/3}} \left(\frac{\sigma^2 V(\theta)}{n}\right)^{2/3} \geq \frac{c_1^2 3^{2/3}}{4096 c_2^{4/3}} \frac{R(n;\theta)}{\log(4n)} \tag{2.34}$$

provided

$$n \geq \max\left(2, \frac{24\sigma^2}{V^2(\theta)}, \frac{2c_2 V(\theta)}{\sigma}\right). \tag{2.35}$$

Proof. Let $V = V(\theta)$. Fix an integer $1 \leq k \leq n$ and let $m := \lfloor n/k \rfloor$ where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x. For each $\tau = (\tau_1, \ldots, \tau_m) \in \{0,1\}^m$, define $\theta^\tau \in \mathcal{M}$ by

$$\theta_j^\tau = \theta_j - \sum_{i=1}^{m} \tau_i \left(\theta_j - \theta_{(i-1)k+1}\right) I\left\{(i-1)k + 1 \leq j \leq ik\right\}$$

for $1 \leq j \leq n$. In other words, when j lies in the interval $[(i-1)k+1, ik]$, the value θ_j^τ equals θ_j if $\tau_i = 0$ and $\theta_{(i-1)k+1}$ if $\tau_i = 1$. The monotonicity of $\{\theta_i^\tau\}$

therefore follows easily from the monotonicity of $\{\theta_i\}$.

We apply Assouad's lemma to these functions θ^τ. Clearly for every $\tau, \tau' \in \{0,1\}^m$, we have

$$\ell^2(\theta^\tau, \theta^{\tau'}) = \frac{1}{n}\sum_{i=1}^{m} I\{\tau_i \neq \tau_i'\} \sum_{j=(i-1)k+1}^{ik} \left(\theta_j - \theta_{(i-1)k+1}\right)^2.$$

For a fixed $1 \leq i \leq m$ and $(i-1)k+1 \leq j \leq ik$, as

$$\theta_j - \theta_{(i-1)k+1} = \sum_{l=(i-1)k+2}^{j} (\theta_l - \theta_{l-1}),$$

assumption (2.6.3) gives

$$\theta_j - \theta_{(i-1)k+1} \geq \frac{c_1 V}{n}(j - (i-1)k - 1).$$

Therefore

$$\sum_{j=(i-1)k+1}^{ik} \left(\theta_j - \theta_{(i-1)k+1}\right)^2 \geq \frac{c_1^2 V^2}{n^2}\sum_{i=1}^{k-1} i^2 = \frac{c_1^2 V^2}{6n^2}k(k-1)(2k-1)$$

for every $1 \leq i \leq m$ and $(i-1)k+1 \leq j \leq ik$. Consequently, for every $\tau, \tau' \in \{0,1\}^m$, we have

$$\ell^2(\theta^\tau, \theta^{\tau'}) \geq \frac{c_1^2 V^2}{6n^3}k(k-1)(2k-1)\Upsilon(\tau, \tau'). \tag{2.36}$$

Similarly, by using the upper bound in (2.6.3), we obtain

$$\ell^2(\theta^\tau, \theta^{\tau'}) \leq \frac{c_2^2 V^2}{6n^3}k(k-1)(2k-1)\Upsilon(\tau, \tau'). \tag{2.37}$$

Also,

$$\max_{1\le j\le n}|\theta_j^\tau-\theta_j| = \max_{1\le i\le m}\max_{(i-1)k+1\le j\le ik}|\theta_j-\theta_{(i-1)k+1}|$$
$$\le \max_{1\le i\le m}\max_{(i-1)k+1\le j\le ik}\frac{c_2V}{n}(j-(i-1)k-1).$$

Consequently,

$$\ell_\infty(\theta^\tau,\theta)=\max_{1\le j\le n}|\theta_j^\tau-\theta_j|\le\frac{c_2Vk}{n}. \tag{2.38}$$

By Pinsker's inequality,

$$\|\mathbb{P}_{\theta^\tau}-\mathbb{P}_{\theta^{\tau'}}\|_{TV}^2\le\frac{1}{2}D(\mathbb{P}_{\theta^\tau}\|\mathbb{P}_{\theta^{\tau'}})=\frac{n}{4\sigma^2}\ell^2(\theta^\tau,\theta^{\tau'})$$

and when $\Upsilon(\tau,\tau')=1$, we have from (2.37) that

$$\|\mathbb{P}_{\theta^\tau}-\mathbb{P}_{\theta^{\tau'}}\|_{TV}^2\le\frac{c_2^2V^2}{24\sigma^2n^2}k(k-1)(2k-1)\le\frac{c_2^2V^2k^3}{12\sigma^2n^2}. \tag{2.39}$$

Thus, by Lemma 2.4.2, (2.36) and (2.39), we get

$$\mathfrak{R}_n(\theta)\ge\frac{c_1^2V^2m}{48n^3}k(k-1)(2k-1)\left[1-\frac{k^{3/2}Vc_2}{\sigma n\sqrt{12}}\right] \tag{2.40}$$

provided

$$\theta^\tau\in\mathfrak{N}(\theta)\qquad\text{for every }\tau\in\{0,1\}^m. \tag{2.41}$$

We now make the choice

$$k=\left(\frac{\sqrt{3}\sigma n}{c_2V}\right)^{2/3}.$$

It then follows from (2.38) that

$$\ell_\infty^2(\theta^\tau,\theta)\le\frac{c_2^2V^2k^2}{n^2}=(3c_2)^{2/3}\left(\frac{\sigma^2V}{n}\right)^{2/3}.$$

We now use Theorem 2.3.2 to show (2.41). Note first that (2.35) is a stronger condition than (2.27) because $c_2 \geq 1$. We can thus use the first inequality in (2.26) to get

$$\ell_\infty^2(\theta^\tau, \theta) \leq \left(\frac{3c_2}{c_1}\right)^{2/3} \frac{R(n;\theta)}{12}$$

which proves (2.41).

Also, with our choice of k, it is easy to check that $2 \leq k \leq n/2$ because (2.35) implies that

$$n \geq \max\left(\frac{2\sqrt{2}c_2 V}{\sqrt{3}\sigma}, \frac{24\sigma^2}{c_2^2 V^2}\right).$$

Inequality (2.40) with our choice of k gives (note that $k(k-1)(2k-1) \geq 3k^3/4$) that

$$\mathfrak{R}_n(\theta) \geq \frac{c_1^2 V^2 m k^3}{128 n^3} = \frac{3c_1^2\sigma^2 m}{128 c_2^2 n}.$$

Because $k \leq n/2$, we have $m = \lfloor n/k \rfloor \geq n/(2k)$. Thus

$$\mathfrak{R}_n(\theta) \geq \frac{c_1^2 V^2 m k^3}{128 n^3} \geq \frac{c_1^2 V^2 k^2}{256 n^2} = \frac{c_1^2 3^{2/3}}{256 c_2^{4/3}} \left(\frac{\sigma^2 V}{n}\right)^{2/3}.$$

This proves the first inequality in (2.34). The second inequality in (2.34) follows from Theorem 2.3.1. The proof is complete. □

2.4.2 Piecewise constant

Here, we again show that the local minimax risk at θ is bounded from below by $R(n;\theta)$ (up to logarithmic multiplicative factors). The difference from the previous section is that we work under a different assumption from (2.6.3). Specifically, we assume that $k(\theta) = k$ and that the k values of θ are sufficiently well-separated and prove inequality (2.11).

Let $k = k(\theta)$. There exist integers $n_1, \ldots, n_k$ with $n_i \geq 1$ and $n_1 + \cdots + n_k = n$ such that θ is constant on each set $\{j : s_{i-1} + 1 \leq j \leq s_i\}$ for $i = 1, \ldots, k$ where $s_0 := 0$ and $s_i := n_1 + \cdots + n_i$. Also, let the values of θ on the sets $\{j : s_{i-1} + 1 \leq j \leq s_i\}$ for $i = 1, \ldots, k$ be denoted by $\theta_{0,1} < \cdots < \theta_{0,k}$.

Theorem 2.4.4. *Suppose $c_1 n/k \leq n_i \leq c_2 n/k$ for all $1 \leq i \leq k$ for some $c_1 \in (0, 1]$ and $c_2 \geq 1$ and that*

$$\min_{2 \leq i \leq k} (\theta_{0,i} - \theta_{0,i-1}) \geq \sqrt{\frac{k\sigma^2}{n} \log \frac{en}{k}}. \tag{2.42}$$

Then, with $\mathfrak{N}(\theta)$ defined as $\{t \in \mathcal{M} : \ell_\infty^2(t, \theta) \leq R(n; \theta)\}$, the local minimax risk, $\mathfrak{R}_n(\theta) = \inf_{\hat{t}} \sup_{t \in \mathfrak{N}(\theta)} \mathbb{E}_t \ell^2(t, \hat{t})$, satisfies

$$\mathfrak{R}_n(\theta) \geq \frac{c_1^{7/3}}{2^{31/3} c_2^2} R(n; \theta) \left(\log \frac{en}{k} \right)^{-2/3} \tag{2.43}$$

provided

$$\frac{n}{k} \geq \max \left(\left(\frac{4}{c_1^2} \log \frac{en}{k} \right)^{1/3}, \exp \left(\frac{1 - 4c_1}{4c_1} \right) \right). \tag{2.44}$$

Proof. For notational convenience, we write

$$\beta_n^2 := \frac{k\sigma^2}{n} \log \frac{en}{k}.$$

First note that under assumption (2.42), Theorem 2.3.3 implies that $\beta_n^2 \leq R(n; \theta)$.

Let $1 \leq l \leq \min_{1 \leq i \leq k} n_i$ be a positive integer whose value will be specified later and let $m_i := \lfloor n_i / l \rfloor$ for $i = 1, \ldots, k$. We also write M for $\sum_{i=1}^k m_i$.

The elements of the finite set $\{0, 1\}^M$ will be represented as $\tau = (\tau_1, \ldots, \tau_k)$ where $\tau_i = (\tau_{i1}, \ldots, \tau_{im_i}) \in \{0, 1\}^{m_i}$. For each $\tau \in \{0, 1\}^M$, we specify $\theta^\tau \in \mathcal{M}$

in the following way. For $s_{i-1}+1 \leq u \leq s_i$, the quantity θ^τ_u is defined as

$$\theta_{0,i}+\frac{\beta_n}{m_i}\sum_{v=1}^{m_i}(v-\tau_{iv})I\{(v-1)l+1 \leq u-s_{i-1} \leq vl\}+\beta_n I\{s_{i-1}+m_i l+1 \leq u \leq s_i\}.$$

Because θ is constant on the set $\{u : s_{i-1}+1 \leq u \leq s_i\}$ where it takes the value $\theta_{0,i}$, it follows that $\ell_\infty(\theta^\tau, \theta) \leq \beta_n$. This implies that $\theta^\tau \in \mathfrak{N}(\theta)$ for every τ as $\beta_n^2 \leq R(n;\theta)$.

Also, because of the assumption $\min_{2\leq i\leq k}(\theta_{0,i}-\theta_{0,i-1}) \geq \beta_n$, it is evident that each θ^τ is non-decreasing. We will apply Assouad's lemma to $\theta^\tau, \tau \in \{0,1\}^M$. For $\tau, \tau' \in \{0,1\}^M$, we have

$$\ell^2(\theta^\tau, \theta^{\tau'}) = \frac{1}{n}\sum_{i=1}^{k}\sum_{v=1}^{m_i}\frac{l\beta_n^2}{m_i^2}I\left\{\tau_{iv} \neq \tau'_{iv}\right\} = \frac{l\beta_n^2}{n}\sum_{i=1}^{k}\frac{\Upsilon(\tau_i, \tau'_i)}{m_i^2}. \tag{2.45}$$

Because

$$m_i \leq \frac{n_i}{l} \leq \frac{c_2 n}{kl} \qquad \text{for each } 1 \leq i \leq k$$

we have

$$\ell^2(\theta^\tau, \theta^{\tau'}) \geq \frac{k^2 l^3 \beta_n^2}{c_2^2 n^3}\sum_{i=1}^{k}\Upsilon(\tau_i, \tau'_i) = \frac{k^2 l^3 \beta_n^2}{c_2^2 n^3}\Upsilon(\tau, \tau'). \tag{2.46}$$

Also, from (2.45), we get

$$\ell^2(\theta^\tau, \theta^{\tau'}) \leq \frac{l\beta_n^2}{n(\min_{1\leq i\leq k} m_i^2)} \qquad \text{when } \Upsilon(\tau, \tau') = 1. \tag{2.47}$$

The quantity $\min_i m_i^2$ can be easily bounded from below by noting that $n_i/l < m_i + 1 \leq 2m_i$ and that $n_i \geq c_1 n/k$. This gives

$$\min_{1\leq i\leq k} m_i \geq \frac{c_1 n}{2kl}. \tag{2.48}$$

Combining the above inequality with (2.47), we deduce

$$\ell^2(\theta^\tau, \theta^{\tau'}) \leq \frac{4k^2l^3\beta_n^2}{c_1^2n^3} \qquad \text{whenever } \Upsilon(\tau, \tau') = 1.$$

This and Pinsker's inequality give

$$\|\mathbb{P}_{\theta^\tau} - \mathbb{P}_{\theta^{\tau'}}\|_{TV}^2 \leq \frac{1}{2}D(\mathbb{P}_{\theta^\tau}\|\mathbb{P}_{\theta^{\tau'}}) = \frac{n}{4\sigma^2}\ell^2(\theta^\tau, \theta^{\tau'}) \leq \frac{k^2l^3\beta_n^2}{c_1^2n^2\sigma^2} \tag{2.49}$$

whenever $\Upsilon(\tau, \tau') = 1$.

The inequalities (2.46) and (2.49) in conjunction with Assouad's lemma give

$$\mathfrak{R}_n(\theta) \geq \frac{Mk^2l^3\beta_n^2}{8c_2^2n^3}\left(1 - \frac{k\beta_n l^{3/2}}{c_1n\sigma}\right).$$

Because of (2.48), we get $M = \sum_i m_i \geq k \min_i m_i \geq c_1 n/(2l)$ and thus

$$\mathfrak{R}_n(\theta) \geq \frac{c_1k^2l^2\beta_n^2}{16c_2^2n^2}\left(1 - \frac{k\beta_n l^{3/2}}{c_1n\sigma}\right). \tag{2.50}$$

The value of the integer l will now be specified. We take

$$l = \left(\frac{c_1n\sigma}{2k\beta_n}\right)^{2/3}. \tag{2.51}$$

Because $\min_i n_i \geq c_1 n/k$, we can ensure that $1 \leq l \leq \min_i n_i$ by requiring that

$$1 \leq \left(\frac{c_1n\sigma}{2k\beta_n}\right)^{2/3} \leq \frac{c_1n}{k}.$$

This gives rise to two lower bounds for n which are collected in (2.44).

As a consequence of (2.51), we get that $l^{3/2} \leq c_1n\sigma/(2k\beta_n)$, which ensures that the term inside the parantheses on the right hand side of (2.50) is atleast

1/2. This gives

$$\mathfrak{R}_n(\theta) \geq \frac{c_1 k^2 l^2 \beta_n^2}{32 c_2^2 n^2} \geq \frac{c_1^{7/3}}{2^{19/3} c_2^2} \frac{k\sigma^2}{n} \left(\log \frac{en}{k}\right)^{1/3}. \tag{2.52}$$

To complete the proof, we use Theorem 2.3.3. Specifically, the second inequality in (2.28) gives

$$\frac{k\sigma^2}{n} \geq \frac{R(n;\theta)}{16} \left(\log \frac{en}{k}\right)^{-1}.$$

The proof is complete by combining the above inequality with (2.52). □

2.5 Risk bound under model misspecification

We consider model (2.3) where now the true sequence θ is not necessarily assumed to be in $\mathcal{M}$. We study the behavior of the LSE $\hat{\theta}$ defined exactly as in (2.4). The goal of this section is to prove an inequality analogous to (2.7) for model misspecification. It turns out here that the LSE is really estimating the non-decreasing projection of θ on $\mathcal{M}$ defined as $\tilde{\theta} \in \mathcal{M}$ that minimizes $\ell^2(t, \theta)$ over $t \in \mathcal{M}$. From [26, Chapter 1], it follows that

$$\tilde{\theta}_j = \min_{l \geq j} \max_{k \leq j} \bar{\theta}_{k,l} \qquad \text{for } 1 \leq j \leq n, \tag{2.53}$$

where $\bar{\theta}_{k,l}$ is as defined in (2.12).

We define another measure of variation for $t \in \mathcal{M}$ with respect to an interval partition $\pi = (n_1, \ldots, n_k)$:

$$S_\pi(t) = \left(\frac{1}{n} \sum_{i=1}^{k} \sum_{j=s_{i-1}+1}^{s_i} (t_{s_i} - t_j)^2 \right)^{1/2}$$

where $s_0 = 0$ and $s_i = n_1 + \cdots + n_i$ for $1 \le i \le k$. It is easy to check that $S_\pi(t) \le V_\pi(t)$ for every $t \in \mathcal{M}$. The following is the main result of this section.

Theorem 2.5.1. *For every $\theta \in \mathbb{R}^n$, the LSE satisfies*

$$\mathbb{E}_\theta \ell^2(\tilde{\theta}, \hat{\theta}) \le 4 \inf_{\pi \in \Pi} \left(S_\pi^2(\tilde{\theta}) + \frac{4\sigma k(\pi)}{n} \log \frac{en}{k(\pi)} \right) \le R(n; \tilde{\theta}). \tag{2.54}$$

Proof. We only need to prove the first inequality in (2.54). The second inequality follows from the fact that $S_\pi(\tilde{\theta}) \le V_\pi(\tilde{\theta})$.

Fix $1 \le j \le n$ and $0 \le m \le n - j$. Recall the notation in (2.12). Let

$$l_j := \max \left\{ l : j + m \le l \le n \text{ and } \max_{k \le j+m} \bar{\theta}_{k,l} = \tilde{\theta}_{j+m} \right\}.$$

Then l_j is well defined by the definition of $\tilde{\theta}$ in (2.53). We now write

$$\hat{\theta}_j = \min_{l \ge j} \max_{k \le j} \bar{Y}_{k,l} \le \max_{k \le j} \bar{Y}_{k,l_j} = \max_{k \le j} \bar{\theta}_{k,l_j} + \max_{k \le j} \bar{\epsilon}_{k,l_j}.$$

As a result,

$$\hat{\theta}_j \le \max_{k \le j+m} \bar{\theta}_{k,l_j} + \max_{k \le j} \bar{\epsilon}_{k,l_j} = \tilde{\theta}_{j+m} + \max_{k \le j} \bar{\epsilon}_{k,l_j}.$$

This implies that

$$\left(\hat{\theta}_j - \tilde{\theta}_j \right)_+ \le (\tilde{\theta}_{j+m} - \tilde{\theta}_j) + \max_{k \le j} \left(\bar{\epsilon}_{k,l_j} \right)_+.$$

Now arguing as in the proof of Theorem 2.2.1 and noting that $l_j - j \ge m$, we get

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \tilde{\theta}_j \right)_+^2 \le 2 \left(\tilde{\theta}_{j+m} - \tilde{\theta}_j \right)^2 + \frac{8\sigma^2}{m+1}.$$

Also we have the corresponding inequality for the negative part:

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \tilde{\theta}_j\right)_-^2 \leq 2\left(\tilde{\theta}_{j+m} - \tilde{\theta}_j\right)^2 + \frac{8\sigma^2}{m+1}.$$

Now fix an interval partition π. For each $1 \leq j \leq n$, we define two integers $m_1(j)$ and $m_2(j)$ in the following way. For $s_{i-1}+1 \leq j \leq s_i$, let $m_1(j) = s_i - j$ and $m_2(j) = j - 1 - s_{i-1}$. This results in the risk bound

$$\mathbb{E}_\theta \ell^2(\hat{\theta}, \tilde{\theta}) \leq \frac{1}{n}\sum_{j=1}^n A_j + \frac{1}{n}\sum_{j=1}^n B_j.$$

where

$$A_j := 2\left(\tilde{\theta}_{j+m_1(j)} - \tilde{\theta}_j\right)^2 + \frac{8\sigma^2}{m_1(j)+1}. \tag{2.55}$$

and

$$B_j := 2\left(\tilde{\theta}_j - \tilde{\theta}_{j-m_2(j)}\right)^2 + \frac{8\sigma^2}{m_1(j)+1}.$$

We shall now prove that

$$\frac{1}{n}\sum_{j=1}^n A_j \leq 2S_\pi^2(\tilde{\theta}) + \frac{8k\sigma^2}{n}\log\frac{en}{k} \tag{2.56}$$

and

$$\frac{1}{n}\sum_{j=1}^n B_j \leq 2S_\pi^2(\tilde{\theta}) + \frac{8k\sigma^2}{n}\log\frac{en}{k}. \tag{2.57}$$

We give the argument for (2.56) below. The argument for (2.57) follows similarly. From (2.55), $\sum_{j=1}^n A_j/n$ can be broken into two terms. For the first term note that $j + m_1(j) = s_i$ for $s_{i-1} + 1 \leq j \leq s_i$, and therefore

$$2\sum_{j=1}^n \left(\tilde{\theta}_{j+m_1(j)} - \tilde{\theta}_j\right)^2 = 2\sum_{i=1}^k \sum_{j=s_{i-1}+1}^{s_i} \left(\tilde{\theta}_{s_i} - \tilde{\theta}_j\right)^2 = 2S_\pi^2(\tilde{\theta}).$$

For the second term, arguing exactly as in (2.20) we get the upper bound

$(8k\sigma^2/n)\log(en/k)$. This proves (2.56) which completes the proof. □

Remark 2.5.1. *By Theorem 2.3.1, the quantity $R(n;\tilde{\theta})$ is bounded from above by $(\sigma^2 V(\tilde{\theta})/n)^{2/3}$ up to a logarithmic multiplicative factor in n. Therefore, Theorem 2.5.1 implies that the LSE $\hat{\theta}$ converges to the projection of θ onto the space of monotone vectors at atleast the $n^{-2/3}$ rate, up to a logarithmic factor in n. The convergence rate will be much faster if $k(\tilde{\theta})$ is small or if $\tilde{\theta}$ is well-approximated by a monotone vector α with small $k(\alpha)$.*

By taking π in the infimum in the upper bound of (2.54) to be the interval partition generated by $\tilde{\theta}$, we obtain the following result which is the analogue of (2.9) for model misspecification.

Corollary 2.5.2. *For every arbitrary sequence θ of length n (not necessarily non-decreasing),*

$$\mathbb{E}_\theta \ell^2(\hat{\theta},\tilde{\theta}) \leq \frac{16\sigma^2 k(\tilde{\theta})}{n}\log\frac{en}{k(\tilde{\theta})}.$$

In the next pair of results, we prove two upper bounds on $k(\tilde{\theta})$. The first result shows that $k(\tilde{\theta}) = 1$ (i.e., $\tilde{\theta}$ is constant) when θ is non-increasing, i.e., $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_n$. This implies that the LSE converges to $\tilde{\theta}$ at the rate $\sigma^2\log(en)/n$ when θ is non-increasing.

Lemma 2.5.3. *$k(\tilde{\theta}) = 1$ if θ is non-increasing.*

Proof. By (2.53), $\tilde{\theta}_1 = \min_{l\geq 1}\bar{\theta}_{1,l}$. Because θ is non-increasing, we get therefore that $\tilde{\theta}_1 = \bar{\theta}_{1,n}$. Similarly $\tilde{\theta}_n = \max_{k\leq n}\bar{\theta}_{k,n} = \bar{\theta}_{1,n}$. Since $\tilde{\theta}$ has to be non-decreasing, it follows that $\tilde{\theta}_j = \bar{\theta}_{1,n}$ for all j. The proof is complete. □

To state our next result, let

$$b(t) := \sum_{i=1}^{n-1} I\{t_i \neq t_{i+1}\} + 1 \qquad \text{for } t \in \mathbb{R}^n. \tag{2.58}$$

$b(t)$ can be interpreted as the number of constant blocks of t. For example, when $n = 5$ and $t = (0, 0, 1, 1, 1, 0)$, $b(t) = 3$. Observe that $b(t) \geq k(t)$ for $t \in \mathbb{R}^n$ and $b(t) = k(t)$ for $t \in \mathcal{M}$.

Lemma 2.5.4. *For any sequence $\theta \in \mathbb{R}^n$ we have $k(\tilde{\theta}) \leq b(\theta)$.*

Proof. As $\tilde{\theta}$ is always non-decreasing it is enough to show that $b(\tilde{\theta}) \leq b(\theta)$. We will show that $I(\tilde{\theta}_i \neq \tilde{\theta}_{i+1}) \leq I(\theta_i \neq \theta_{i+1})$ for every $i \in \{1, \ldots, n-1\}$ using the method of contradiction. Fix $i \in \{1, \ldots, n-1\}$ and assume that $\theta_i = \theta_{i+1}$ and $\tilde{\theta}_i \neq \tilde{\theta}_{i+1}$. Let $\theta_i = \theta_{i+1} = c$. Define another sequence $t \in \mathbb{R}^n$ such that for $j = 1, \ldots, n$,

$$t_j = \left(\frac{\tilde{\theta}_i + \tilde{\theta}_{i+1}}{2} \right) \cdot I\left\{ i \leq j \leq i+1 \right\} + \tilde{\theta}_j \cdot I\left\{ i \leq j \leq i+1 \right\}^c .$$

We note that t is non-decreasing as $\tilde{\theta}$ is non-decreasing. Now,

$$\begin{aligned} n[\ell^2(\theta, \tilde{\theta}) - \ell^2(\theta, t)] &= \sum_{j=i}^{i+1} \left[(\theta_j - \tilde{\theta}_j)^2 - (\theta_j - t_j)^2 \right] \\ &= \sum_{j=i}^{i+1} \left[(\tilde{\theta}_j - c)^2 - (t_j - c)^2 \right] = \sum_{j=i}^{i+1} (\tilde{\theta}_j - t_j)^2 > 0. \end{aligned}$$

But this is a contradiction as $\tilde{\theta}$ is the non-decreasing projection of θ. □

As a consequence of the above lemma, we obtain that for every $\theta \in \mathbb{R}^n$, the quantity $\mathbb{E}_\theta \ell^2(\hat{\theta}, \tilde{\theta})$ is bounded from above by $(16 b(\theta) \sigma^2 / n) \log(en/b(\theta))$.

2.6 A general result

In this section, we prove a more general version of our main result, Theorem 2.2.1, that applies to other shape-constrained regression problems, such as

convex regression. A natural way of generalizing the isotonic regression problem (see e.g., [23]) is to consider the problem of estimating $\theta = (\theta_1, \ldots, \theta_n)$ from model (2.3) under the constraint that $\theta \in \mathcal{K}$ where $\mathcal{K}$ is an arbitrary known convex polyhedral cone in $\mathbb{R}^n$, i.e.,

$$\mathcal{K} := \{\theta \in \mathbb{R}^n : A\theta \gtrsim 0\} \tag{2.59}$$

where A is a matrix of order $m \times n$ and $\alpha = (\alpha_1, \ldots, \alpha_m) \gtrsim 0$ means that $\alpha_i \geq 0$ for each i (see, e.g., [28, Chapters 7 and 8] for basic facts about convex polyhedral cones). In this section, we shall assume normality of the errors $\epsilon_1, \ldots, \epsilon_n$.

This general setup includes the following as special cases:

1. Isotonic regression corresponds to $\mathcal{K} := \{\theta \in \mathbb{R}^n : \theta_1 \leq \cdots \leq \theta_n\}$.

2. Convex regression with uniformly spaced design points corresponds to $\mathcal{K} := \{\theta \in \mathbb{R}^n : 2\theta_i \leq \theta_{i-1} + \theta_{i+1} \text{ for } i = 2, \ldots, n-1\}$.

3. k-monotone regression corresponds to $\mathcal{K} := \left\{\theta \in \mathbb{R}^n : \nabla^k \theta \gtrsim 0\right\}$ where $\nabla : \mathbb{R}^n \to \mathbb{R}^n$ is defined by $\nabla(\theta) := (\theta_2 - \theta_1, \theta_3 - \theta_2, \ldots, \theta_n - \theta_{n-1}, 0)$ and ∇^k is the k-times composition of ∇ with itself.

4. Non-negative least squares regression corresponds to

$$\mathcal{K} := \{\beta_1 X_1 + \cdots + \beta_p X_p : \beta \gtrsim 0\}$$

for some explanatory variables $X_1, \ldots, X_p \in \mathbb{R}^n$ where $\beta := (\beta_1, \ldots, \beta_p)$.

Note that the cone $\mathcal{K}$ in the first three examples listed above is of the form

$$\mathcal{K}^n_{r,s} := \{\theta \in \mathbb{R}^n : \sum_{j=-r}^{s} w_j \theta_{t+j} \geq 0 \text{ for all } 1 + r \leq t \leq n - s\} \tag{2.60}$$

for some integers $r \geq 0$ and $s \geq 1$ and non-negative weights $w_j, -r \leq j \leq s$. When $n < 1 + r + s$, the condition in the definition of $\mathcal{K}^n_{r,s}$ is vacuous so that $\mathcal{K}^n_{r,s} = \mathbb{R}^n$. The integers r and s and the weights $w_j, -r \leq j \leq s$, do not depend on n. The dependence of the cone on the weights $\{w_j\}$ is suppressed in the notation $\mathcal{K}^n_{r,s}$. Isotonic regression corresponds to $r = 0, s = 1, w_0 = -1$ and $w_1 = 1$ while convex regression corresponds to $r = 1, s = 1, w_{-1} = w_1 = 1$ and $w_0 = -2$.

We denote the LSE for θ under the constraint $\theta \in \mathcal{K}$ by $\hat{\theta}(Y; \mathcal{K})$ where

$$\hat{\theta}(y; \mathcal{K}) := \operatorname*{argmin}_{\theta \in \mathcal{K}} \|\theta - y\|^2 \qquad \text{for } y \in \mathbb{R}^n, \tag{2.61}$$

and $Y = (Y_1, \ldots, Y_n)$ consists of the observations.

The function $y \mapsto \hat{\theta}(y; \mathcal{K})$ is well-defined (because for each y and $\mathcal{K}$, the quantity $\hat{\theta}(y; \mathcal{K})$ exists uniquely by the Hilbert projection theorem), non-linear in y (in general) and can be characterized by

$$\hat{\theta}(y; \mathcal{K}) \in \mathcal{K}, \ \left\langle y - \hat{\theta}(y; \mathcal{K}), \hat{\theta}(y; \mathcal{K}) \right\rangle = 0 \text{ and } \left\langle y - \hat{\theta}(y; \mathcal{K}), \omega \right\rangle \leq 0 \tag{2.62}$$

for all $\omega \in \mathcal{K}$.

We shall prove below bounds on the risk of $\hat{\theta}(y; \mathcal{K})$. Our bounds will involve the *statistical dimension* of $\mathcal{K}$ which is defined as

$$\delta(\mathcal{K}) := \mathbb{E} D(Z; \mathcal{K}) \qquad \text{where } \ D(y; \mathcal{K}) := \sum_{i=1}^{n} \frac{\partial}{\partial y_i} \hat{\theta}_i(y; \mathcal{K}) \tag{2.63}$$

and $Z = (Z_1, \ldots, Z_n)$ is a vector whose components are independent standard normal random variables. Note that the quantity $D(y; \mathcal{K})$ is well-defined because $\hat{\theta}(y; \mathcal{K})$ is a Lipschitz function of y (see [23]); in fact, $\hat{\theta}(y; \mathcal{K})$ is 1-Lipschitz. The statistical dimension is an important summary parameter for cones and it has been used in shape-constrained regression ([23]) and compressed sensing ([1, 24]). In fact, it is argued in [23] that $D(Y; \mathcal{K})$ provides a measure of the effective dimension of the model. To see how this conjecture generalizes simpler models, observe that if $\mathcal{K}$ is a linear space of dimension d, say, then $\hat{\theta}(y; \mathcal{K}) = QY$, where Q is the projection matrix onto $\mathcal{K}$, and $D(y; \mathcal{K}) = \text{trace}(Q) = d$ for all y. It was also shown in [23] that $D(Y; \mathcal{K})$ is the number of distinct values among $\hat{\theta}_1, \ldots, \hat{\theta}_n$ for isotonic regression.

An alternative definition of the statistical dimension involves the LSE and is given by

$$\delta(\mathcal{K}) = \mathbb{E}\|\hat{\theta}(Z; \mathcal{K})\|^2. \tag{2.64}$$

The equivalence of (2.63) and (2.64) was observed by [23, Proof of Proposition 2]. It is actually an easy consequence of Stein's lemma because the second identity in (2.62) implies $\mathbb{E}\|\hat{\theta}(Z; \mathcal{K})\|^2 = \mathbb{E}\left\langle Z, \hat{\theta}(Z; \mathcal{K})\right\rangle$ and, therefore, Stein's lemma on the right hand side gives the equivalence of (2.63) and (2.64).

The statistical dimension is closely related to the *Gaussian width* of $\mathcal{K}$ (see Section 10.3 of [1]) which is defined as

$$w(\mathcal{K}) := \mathbb{E} \sup_{x \in \mathcal{K} \cap S^{n-1}} \langle Z, x \rangle$$

where $S^{n-1} := \{u \in \mathbb{R}^n : u_1^2 + \cdots + u_n^2 = 1\}$ denotes the unit sphere. Gaussian width is an important quantity in geometric functional analysis (see e.g., Chapter 4 of [33]) and it has been used to prove recovery bounds in compressed sensing ([1, 11, 24, 27, 29]). The following inequality (see [1, Proposition 10.1])

relates the statistical dimension and the Gaussian width:

$$w^2(\mathcal{K}) \le \delta(\mathcal{K}) \le w^2(\mathcal{K}) + 1.$$

Below, we state and prove Theorem 2.6.1 which generalizes Theorem 2.2.1 and applies to any cone of the form (2.60). For the proof of Theorem 2.6.1, we prove certain auxiliary results which hold for any polyhedral cone (2.59).

Theorem 2.6.1. *Fix $n \ge 1$, $r \ge 0$ and $s \ge 1$. Consider the problem of estimating $\theta \in \mathcal{K}^n_{r,s}$ from (2.3) for independent $N(0,\sigma^2)$ errors $\epsilon_1, \dots, \epsilon_n$. Then for every $\theta \in \mathcal{K}^n_{r,s}$, the following bound holds for the risk of the LSE:*

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{K}^n_{r,s})) \le 6 \inf_{\alpha \in \mathcal{K}^n_{r,s}} \left(\ell^2(\theta, \alpha) + \frac{\sigma^2(1+k(\alpha))}{n} \delta(\mathcal{K}^n_{r,s}) \right) \tag{2.65}$$

where for each $\theta \in \mathcal{K}^n_{r,s}$, the integer $k(\theta)$ denotes the number of inequalities among $\sum_{j=-r}^{s} w_j \theta_{t+j} \ge 0$, for $1 + r \le t \le n - s$, that are strict.

Remark 2.6.1 (Stronger version)**.** *From the proof of Theorem 2.6.1, it will be clear that the risk of the LSE satisfies a stronger inequality than (2.65). For $\alpha \in \mathcal{K}^n_{r,s}$ with $k(\alpha) = k$, let $1 + r \le t_1 < \cdots < t_k \le n - s$ denote the values of t for which the inequalities $\sum_{j=-r}^{s} w_j \alpha_{t+j} \ge 0$ are strict. Let*

$$\tau(\alpha) := \delta(\mathcal{K}^{t_1 - 1 + s}_{r,s}) + \delta(\mathcal{K}^{t_2 - t_1}_{r,s}) + \cdots + \delta(\mathcal{K}^{t_k - t_{k-1}}_{r,s}) + \delta(\mathcal{K}^{n - t_k - s + 1}_{r,s}).$$

The proof of Theorem 2.6.1 will imply that

$$\mathbb{E}_\theta \ell^2 \left(\theta, \hat{\theta}(Y; \mathcal{K}^n_{r,s}) \right) \le 6 \inf_{\alpha \in \mathcal{K}^n_{r,s}} \left(\ell^2(\theta, \alpha) + \frac{\sigma^2}{n} \tau(\alpha) \right). \tag{2.66}$$

The trivial observation that $\delta(\mathcal{K}^n_{r,s})$ is increasing in n (note that the weights $w_j, -r \le j \le s$, do not depend on n) implies that $\tau(\alpha) \le (1 + k(\alpha))\delta(\mathcal{K}^n_{r,s})$ for

all $\alpha \in \mathcal{K}^n_{r,s}$ *and hence inequality* (2.66) *is stronger than* (2.65).

Remark 2.6.2 (Connection to the facial structure of $\mathcal{K}^n_{r,s}$)**.** *Every convex polyhedral cone* (2.59) *has a well-defined facial structure. Indeed, a standard result (see, for example, [28, Section 8.3]) states that a subset* F *of* $\mathcal{K}$ *is a face if and only if* F *is non-empty and* $F = \{\theta \in \mathcal{K} : \tilde{A}\theta = 0\}$ *for some* $\tilde{m} \times n$ *matrix* $\tilde{A}$ *whose rows are a subset of the rows of* A*. The dimension of* F *equals* $n - \rho(\tilde{A})$ *where* $\rho(\tilde{A})$ *denotes the rank of* $\tilde{A}$*. It is then clear that if* $\theta \in \mathcal{K}^n_{r,s}$ *is in a low-dimensional face of* $\mathcal{K}$*, then* $k(\theta)$ *must be small. Now if* $\delta(\mathcal{K}^n_{r,s})$ *is at most logarithmic in* n *(which is indeed the case for the case of isotonic and convex regression; see Examples 2.6.2 and 2.6.3), then the bound* (2.65) *implies that the risk of the LSE is bounded from above by the parametric rate* σ^2/n *(up to multiplicative logarithmic factors in* n*) provided* θ *is in a low-dimensional face of* $\mathcal{K}^n_{r,s}$*. Therefore, the LSE automatically adapts to vectors in low-dimensional faces of* $\mathcal{K}^n_{r,s}$*. For general* θ*, the risk is bounded from above by a combination of how close* θ *is to a* k*-dimensional face of* $\mathcal{K}^n_{r,s}$ *and* $\sigma^2\delta(\mathcal{K}^n_{r,s})(1+k)/n$ *as* k *varies.*

Example 2.6.2 (Isotonic regression)**.** *Isotonic regression corresponds to* $\mathcal{K}^n_{r,s}$ *with* $r = 0, s = 1, w_0 = -1$ *and* $w_1 = 1$*. It turns out that the statistical dimension of this cone satisfies*

$$\delta(\mathcal{K}^n_{0,1}) = 1 + \frac{1}{2} + \cdots + \frac{1}{n}, \qquad \text{for every } n \geq 1, \tag{2.67}$$

which immediately implies that $\delta(\mathcal{K}^n_{0,1}) \leq \log(en)$*. This can be proved using symmetry arguments formalized in the theory of finite reflection groups (see [1, Appendix C.4] where the proof of* (2.67) *is sketched). Theorem 2.6.1 (in its stronger form* (2.66)*) therefore gives an alternative proof of Theorem 2.2.1 with different multiplicative constants. It should be noted however that this*

proof only works for the case of normal $\epsilon_1, \ldots, \epsilon_n$ while Theorem 2.2.1 does not require normality.

Example 2.6.3 (Convex regression)**.** *Convex regression corresponds to $\mathcal{K}^n_{r,s}$ with $r = s = 1, w_{-1} = w_1 = 1$ and $w_0 = -2$. It turns out that the statistical dimension of this cone satisfies*

$$\delta(\mathcal{K}^n_{-1,1}) \leq C(\log n)^{5/4}, \qquad \text{for all } n \geq 1,$$

where C is a universal positive constant. This is proved in [21, Theorem 2.3] via metric entropy results for classes of convex functions. This gives the risk bound

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{K}^n_{-1,1})) \leq C \inf_{\alpha \in \mathcal{K}^n_{-1,1}} \left(\ell^2(\theta, \alpha) + \frac{\sigma^2(1 + k(\alpha))}{n} (\log n)^{5/4} \right).$$

The quantity $1 + k(\alpha)$ can be interpreted as the number of affine pieces of the convex sequence α. This risk bound is the analogue of Theorem 2.2.1 for convex regression.

2.6.1 Proof of Theorem 2.6.1

We now prove Theorem 2.6.1. We shall first prove some general results for the risk of $\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{K}))$ which hold for every $\mathcal{K}$ of the form (2.59). Theorem 2.6.1 will then be proved by specializing these results for $\mathcal{K} = \mathcal{K}^n_{r,s}$.

We begin by recalling a result of [23] who related the risk of $\hat{\theta}(Y; \mathcal{K})$ to the function $D(\cdot; \mathcal{K})$. Specifically, [23, Proposition 2] proved that

$$\mathbb{E}_0 \ell^2(0, \hat{\theta}(Y; \mathcal{K})) = \frac{\sigma^2 \delta(\mathcal{K})}{n} = \frac{\sigma^2}{n} \mathbb{E}_0 D(Y; \mathcal{K}) \tag{2.68}$$

and

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{K})) \leq \frac{\sigma^2}{n} \mathbb{E}_\theta D(Y; \mathcal{K}) \qquad \text{for every } \theta \in \mathcal{K}. \tag{2.69}$$

These can be proved via Stein's lemma (see [23, Proof of Proposition 2]). It might be helpful to observe here that the function $D(y; \mathcal{K})$ satisfies $D(ty; \mathcal{K}) = D(y; \mathcal{K})$ for every $t \in \mathbb{R}$ and this is a consequence of the fact that $\hat{\theta}(ty; \mathcal{K}) = t\hat{\theta}(y; \mathcal{K})$ and the characterization (2.62).

Our first lemma below says that the risk of the LSE is equal to $\sigma^2 \delta(\mathcal{K})/n$ for all θ belonging to the lineality space $\mathfrak{L} := \{\theta \in \mathbb{R}^n : A\theta = 0\}$ of $\mathcal{K}$. The lineality space $\mathfrak{L}$ will be crucial in the proof of Theorem 2.6.1. The lineality space of the cone for isotonic regression is the set of all constant sequences. The lineality space of the cone for convex regression is the set of all affine sequences.

Lemma 2.6.4. *For every $\theta \in \mathbb{R}^n$ with $\theta = \gamma_1 + \gamma_2$ for some $\gamma_1 \in \mathfrak{L}$ and $\gamma_2 \perp \mathcal{K}$ (i.e., $\langle \gamma_2, \omega \rangle = 0$ for all $\omega \in \mathcal{K}$), we have $\mathbb{E}_\theta D(Y; \mathcal{K}) = \delta(\mathcal{K})$.*

Proof. Let θ satisfy the statement of the lemma. By the characterization (2.62), it is clear that

$$\hat{\theta}(y - \theta; \mathcal{K}) = \hat{\theta}(y; \mathcal{K}) - \gamma_1.$$

Therefore,

$$D(y - \theta; \mathcal{K}) = D(y; \mathcal{K}).$$

The proof is now completed by taking expectation of both sides above with respect to $\mathbb{E}_\theta$. □

Our next lemma allows us to bound the risk of the LSE via the dimension for θ not necessarily in the lineality space.

Lemma 2.6.5. *Let $\mathcal{K}$ be an arbitrary convex polyhedral cone. Suppose $\mathcal{K}_1, \ldots, \mathcal{K}_l$ are orthogonal polyhedral cones with lineality spaces $\mathfrak{L}_1, \ldots, \mathfrak{L}_l$ such that $\mathcal{K} \subseteq \mathcal{K}_1 + \cdots + \mathcal{K}_l$. Then*

$$\mathbb{E}_\theta D(Y; \mathcal{K}) \leq 2(\delta(\mathcal{K}_1) + \cdots + \delta(\mathcal{K}_l)) \qquad \text{for every } \theta \in \mathcal{K} \cap (\mathfrak{L}_1 + \cdots + \mathfrak{L}_l).$$

Proof. Because $\mathcal{K} \subseteq \mathcal{K}_1 + \cdots + \mathcal{K}_l$, [23, Corollary 2] gives

$$\mathbb{E}_\theta D(Y; \mathcal{K}) \leq 2\mathbb{E}_\theta D(Y; \mathcal{K}_1 + \cdots + \mathcal{K}_l) \qquad \text{for every } \theta \in \mathcal{K}.$$

We shall show below, using the orthogonality of $\mathcal{K}_1, \ldots, \mathcal{K}_l$, that

$$\hat{\theta}(y; \mathcal{K}_1 + \cdots + \mathcal{K}_l) = \hat{\theta}(y; \mathcal{K}_1) + \cdots + \hat{\theta}(y; \mathcal{K}_l). \tag{2.70}$$

Assuming this, we have that

$$D(y; \mathcal{K}_1 + \cdots + \mathcal{K}_l) = D(y; \mathcal{K}_1) + \cdots + D(y; \mathcal{K}_l).$$

which implies that

$$\mathbb{E}_\theta D(Y; \mathcal{K}_1 + \cdots + \mathcal{K}_l) = \mathbb{E}_\theta D(Y; \mathcal{K}_1) + \cdots + \mathbb{E}_\theta D(Y; \mathcal{K}_l).$$

By the assumptions of the theorem, it is easy to see that for every $\theta \in \mathcal{K} \cap (\mathfrak{L}_1 + \cdots + \mathfrak{L}_l)$ and $1 \leq i \leq l$, we can write $\theta = \gamma_i + \tilde{\gamma}_i$ for some $\gamma_i \in \mathfrak{L}_i$ and $\tilde{\gamma}_i \perp \mathcal{K}_i$. Lemma 2.6.4 therefore gives that $\mathbb{E}_\theta D(Y; \mathcal{K}_i) = \delta(\mathcal{K}_i)$, which completes the proof.

It only remains to prove (2.70). Using the characterization (2.62) and the

orthogonality of $\mathcal{K}_i$, for $i = 1, \ldots, l$, it is easy to check that

$$\hat{\theta}(y; \sum_{i=1}^{l} \mathcal{K}_i) = \sum_{i=1}^{l} \hat{\theta}(y; \mathcal{K}_i),$$

which completes the proof of (2.70). □

The next lemma allows us to bound the risk of the LSE at θ by a combination of the risk at α and the distance between θ and α.

Lemma 2.6.6. *The risk of the LSE satisfies the following inequality*

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{K})) \leq 3 \inf_{\alpha \in \mathcal{K}} \left[2\ell^2(\theta, \alpha) + \mathbb{E}_\alpha \ell^2(\alpha, \hat{\theta}(Y; \mathcal{K})) \right] \qquad \textit{for every } \theta \in \mathcal{K}.$$

Proof. Let us denote $\hat{\theta}(y; \mathcal{K})$ by simply $\hat{\theta}(y)$ for the ease of notation. For every $\theta, \alpha \in \mathcal{K}$, the triangle inequality gives

$$\begin{aligned} \|\hat{\theta}(y) - \theta\|^2 &= \|\hat{\theta}(y) - \hat{\theta}(y - \theta + \alpha) + \hat{\theta}(y - \theta + \alpha) - \alpha + \alpha - \theta\|^2 \\ &\leq 3\|\hat{\theta}(y) - \hat{\theta}(y - \theta + \alpha)\|^2 + 3\|\hat{\theta}(y - \theta + \alpha) - \alpha\|^2 + 3\|\alpha - \theta\|^2. \end{aligned}$$

Because the map $y \mapsto \hat{\theta}(y)$ is 1-Lipschitz,

$$\|\hat{\theta}(y) - \hat{\theta}(y - \theta + \alpha)\| \leq \|\theta - \alpha\|.$$

Consequently, we get

$$\|\hat{\theta}(y) - \theta\|^2 \leq 3\|\hat{\theta}(y - \theta + \alpha) - \alpha\|^2 + 6\|\alpha - \theta\|^2.$$

The proof is now complete by taking expectations on both sides above with respect to $Y \sim N(\theta, \sigma^2 I)$. □

We are now ready to prove Theorem 2.6.1.

Proof of Theorem 2.6.1. By Lemma 2.6.6, it is enough to prove that

$$\mathbb{E}_\alpha \ell^2(\alpha, \hat{\theta}(Y; \mathcal{K}^n_{r,s})) \leq 2(1 + k(\alpha)) \frac{\sigma^2 \delta(\mathcal{K}^n_{r,s})}{n} \qquad \text{for every } \alpha \in \mathcal{K}^n_{r,s}.$$

Fix $\alpha \in \mathcal{K}^n_{r,s}$ and let $k = k(\alpha)$, which means that k of the inequalities $\sum_{j=-r}^{s} w_j \alpha_{t+j} \geq 0$ for $1 + r \leq t \leq n - s$ are strict. Let $1 + r \leq t_1 < \cdots < t_k \leq n - s$ denote the indices of the inequalities that are strict. We partition the set $\{1, \ldots, n\}$ into $k + 1$ disjoint sets $E_0, \ldots, E_k$ where

$$E_0 := \{1, \ldots, t_1 - 1 + s\}, \; E_k := \{t_k + s, \ldots, n\}$$

and

$$E_i := \{t_i + s, \ldots, t_{i+1} - 1 + s\} \qquad \text{for } 1 \leq i \leq k - 1.$$

Also for each $0 \leq i \leq k$, let

$$F_i := \{t \in \mathbb{Z} : t - r \in E_i \text{ and } t + s \in E_i\}.$$

We now apply Lemma 2.6.5 with

$$\mathcal{K}_i := \{\theta \in \mathbb{R}^n : \theta_j = 0 \text{ for } j \notin E_i \text{ and } \sum_{j=-r}^{s} w_j \theta_{t+j} \geq 0 \text{ for } t \in F_i\}$$

for $i = 0, \ldots, k$. The lineality space of $\mathcal{K}_i$ is, by definition,

$$\mathfrak{L}_i = \{\theta \in \mathbb{R}^n : \theta_j = 0 \text{ for } j \notin E_i \text{ and } \sum_{j=-r}^{s} w_j \theta_{t+j} = 0 \text{ for } t \in F_i\}.$$

$\mathcal{K}_0, \ldots, \mathcal{K}_k$ are orthogonal convex polyhedral cones because $E_0, \ldots, E_k$ are disjoint. Also $\mathcal{K} \subseteq \mathcal{K}_0 + \cdots + \mathcal{K}_k$ because every $\theta \in \mathcal{K}$ can be written as

$\theta = \sum_{i=0}^{k} \theta^{(i)}$ where $\theta_j^{(i)} := \theta_j I\{j \in E_i\}$ (it is easy to check that $\theta^{(i)} \in \mathcal{K}_i$ for each i). Further, note that $\alpha \in \mathfrak{L}_0 + \cdots + \mathfrak{L}_k$ since $\alpha^{(i)} \in \mathfrak{L}_i$ for every i. Lemma 2.6.5 thus gives $\mathbb{E}_\alpha D(Y; \mathcal{K}) \leq 2\sum_{i=0}^{k} \delta(\mathcal{K}_i)$. Inequality (2.69) then implies that

$$\mathbb{E}_\alpha \ell^2(\alpha, \hat{\theta}(Y; \mathcal{K})) \leq \frac{2\sigma^2}{n} \sum_{i=0}^{k} \delta(\mathcal{K}_i).$$

It is now easy to check that $\delta(\mathcal{K}_i) = \delta(\mathcal{K}_{r,s}^{|E_i|})$ for each i which proves (2.66). The proof of (2.65) is now complete by the observation $\delta(\mathcal{K}_{r,s}^{|E_i|}) \leq \delta(\mathcal{K}_{r,s}^{n})$ as $|E_i| \leq n$. □

2.7 Some auxiliary results

Lemma 2.7.1. *For every $\theta \in \mathcal{M}$ and $\delta > 0$, there exists an interval partition π with*

$$V_\pi(\theta) \leq \delta \text{ and } k(\pi) \leq \left\lceil \frac{V(\theta)}{\delta} \right\rceil \tag{2.71}$$

where $\lceil x \rceil$ denotes the smallest integer that is larger than or equal to x.

Proof. Fix $\theta \in \mathcal{M}$ and $\delta > 0$. Write V for $V(\theta)$. Let $s_0 := 0$ and we recursively define s_i by

$$s_i := \sup\left\{ s_{i-1} + 1 \leq s \leq n : \theta_s - \theta_{s_{i-1}+1} \leq \delta \right\}$$

while $s_{i-1} < n$. This construction will result in integers $s_0 = 0 < s_1 < \cdots < s_m = n$ having the property that $\theta_{s_i+1} - \theta_{s_{i-1}+1} > \delta$ for every $i = 1, \ldots, m$. As a result, we obtain

$$V \geq (\theta_{s_{m-1}+1} - \theta_1) > (m-1)\delta$$

which is the same as $m \leq \lceil V/\delta \rceil$. Let $\pi = (n_1, \ldots, n_m)$ where $n_i = s_i - s_{i-1}$.

Clearly $k(\pi) = m \leq \lceil V/\delta \rceil$. The definition of s_i ensures that $\theta_{s_i} - \theta_{s_{i-1}+1} \leq \delta$ for every i which implies that $V_\pi(\theta) \leq \delta$. The proof is complete. □

Lemma 2.7.2. *The following inequality holds for every finite sequence $a_1, \ldots, a_k$ of real numbers:*

$$\sum_{j=1}^{k} (a_j - \bar{a}_{j,k})^2 \leq \sum_{j=1}^{k} (a_j - \bar{a})^2 \tag{2.72}$$

where $\bar{a}_{j,k}$ denotes the mean of $a_j, \ldots, a_k$ and $\bar{a}$ is the mean of the entire sequence $a_1, \ldots, a_k$.

Proof. We shall prove this by induction on k. Note that the inequality is trivial for $k = 1$ and $k = 2$. Suppose we have proved it for all $k \leq m$ and for all sequences $a_1, \ldots, a_k$. We now consider $k = m + 1$ and a finite sequence $a_1, \ldots, a_{m+1}$ and prove that

$$\sum_{j=1}^{m+1} (a_j - \bar{a}_{j,m+1})^2 \leq \sum_{j=1}^{m+1} (a_j - \bar{a})^2 . \tag{2.73}$$

Note that the first term (for $j = 1$) on both sides of (2.73) equal each other because $\bar{a}_{1,m+1} = \bar{a}$ and can therefore be cancelled. After cancellation, note that the right hand side depends on a_1 while the left hand side does not. As a result, (2.73) holds for all sequences $a_1, \ldots, a_{m+1}$ if and only if

$$\sum_{j=2}^{m+1} (a_j - \bar{a}_{j,m+1})^2 \leq \inf_{a_1 \in \mathbb{R}} \sum_{j=2}^{m+1} (a_j - \bar{a})^2 .$$

Clearly the right hand side above is minimized when we choose a_1 so that $\bar{a} = \bar{a}_{2,m+1}$. Thus (2.73) holds if and only if

$$\sum_{j=2}^{m+1} (a_j - \bar{a}_{j,m+1})^2 \leq \sum_{j=2}^{m+1} (a_j - \bar{a}_{2,m+1})^2 .$$

But this is precisely the inequality (2.72) for $m = k$ and the sequence $a_2, \ldots, a_{m+1}$ which is true by the induction hypothesis. This completes the proof. $\square$

Bibliography

[1] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: a geometric theory of phase transitions in convex optimization. available at http://arxiv.org/abs/1303.6672, 2013.

[2] Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.*, 26:641–647, 1955.

[3] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalisation. *Probability Theory and Related Fields*, 113:301–413, 1999.

[4] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.

[5] Lucien Birgé. The Grenander estimator: a nonasymptotic approach. *Ann. Statist.*, 17(4):1532–1549, 1989.

[6] H. D. Brunk. Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, 26:607–616, 1955.

[7] H. D. Brunk. Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference (Proc. Sympos., Indiana Univ., Bloomington, Ind., 1969)*, pages 177–197. Cambridge Univ. Press, London, 1970.

[8] T. Cai and M. Low. A framework for estimation of convex functions. available at http://www-stat.wharton.upenn.edu/ tcai/, 2011.

[9] Chris Carolan and Richard Dykstra. Asymptotic behavior of the Grenander estimator at density flat regions. *Canad. J. Statist.*, 27(3):557–566, 1999.

[10] Eric Cator. Adaptivity and optimality of the monotone least-squares estimator. *Bernoulli*, 17:714–735, 2011.

[11] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12:805–849, 2012.

[12] D. Donoho. Gelfand n-widths and the method of least squares. Technical report, University of California, Berkeley, 1991. Department of Statistics, Technical report.

[13] C. Durot. Sharp asymptotics for isotonic regression. *Probability Theory and Related Fields*, 122:222–240, 2002.

[14] C. Durot. On the l_p error of monotonicity constrained estimators. *Annals of Statistics*, 35:1080–1104, 2007.

[15] Rick Durrett. *Probability: theory and examples.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, fourth edition, 2010.

[16] Ulf Grenander. On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, 39:125–153 (1957), 1956.

[17] P. Groeneboom. Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II*

(Berkeley, Calif., 1983), Wadsworth Statist./Probab. Ser., pages 539–555, Belmont, CA, 1985. Wadsworth.

[18] P. Groeneboom, G. Jongbloed, and J. A. Wellner. Estimation of convex functions: characterizations and asymptotic theory. *Annals of Statistics*, 29:1653–1698, 2001b.

[19] Piet Groeneboom. The concave majorant of Brownian motion. *Ann. Probab.*, 11(4):1016–1027, 1983.

[20] Piet Groeneboom and Ronald Pyke. Asymptotic normality of statistics based on the convex minorants of empirical distribution functions. *Ann. Probab.*, 11(2):328–345, 1983.

[21] Adityanand Guntuboyina and Bodhisattva Sen. Global risk bounds and adaptation in univariate convex regression. *Under review*, available at http://arxiv.org/abs/1305.1648, 2013.

[22] Hanna K. Jankowski. Convergence of linear functionals of the grenander estimator under misspecification. *Annals of Statistics*, 2014. to appear.

[23] Mary Meyer and Michael Woodroofe. On the degrees of freedom in shape-restricted regression. *Ann. Statist.*, 28(4):1083–1104, 2000.

[24] S. Omyak and B. Hassibi. Sharp MSE bounds for proximal denoising. available at http://arxiv.org/abs/1305.2714, 2013.

[25] Philippe Rigollet and Alexandre B. Tsybakov. Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4):558–575, 2012.

[26] Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics:

Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester, 1988.

[27] M. Rudelson and R. Vershynin. Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. In *Proceedings of the 40th Annual Conference on Information Sciences and Systems*, 2006.

[28] Alexander Schrijver. *Theory of linear and integer programming*. Wiley, Chichester, 1986.

[29] M. Stojnic. Various thresholds for ℓ_1-optimization in compressed sensing. available at http://arxiv.org/abs/0907.3666, 2009.

[30] S. Van de Geer. Estimating a regression function. *Annals of Statistics*, 18:907–924, 1990.

[31] Sara Van de Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, 21(1):14–44, 1993.

[32] Aad Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[33] Roman Vershynin. Lectures in geometric functional analysis. Available at www-personal.umich.edu/~romanv, 2014.

[34] Y. Wang. The l_2 risk of an isotonic estimate. *Comm. Statist. Theory Methods*, 25:281–294, 1996.

[35] Cun-Hui Zhang. Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555, 2002.

Chapter 3

Advances in Adaptive Annealing

Motivated by a classical non parametric function estimation problem, we explore a new annealing approach to randomized optimization of certain statistically relevant multimodal functions in high dimensions. Our hope is to bypass the problems of using algorithms based on Metropolis Hastings or traditional Simulated Annealing type ideas. We are not fixated on exact optimization but are content with maximization up to a constant factor. The objective functions we consider are bounded and have a bounded gradient in Euclidean space. We make an initial foray into this hard problem with the intent to determine a provably accurate approximate maximizer in manageable computational time.

3.1 Introduction and Motivation

3.1.1 Statistical Motivation

In statistics, a central problem is to estimate an unknown regression function f given data $(x_i, y_i)_{i=1}^n$ where x_i are d dimensional input variables and y_i is a single variable response. The responses could be modelled as $y_i = f(x_i) + \epsilon_i$ where ϵ_i are i.i.d zero mean errors. There is a rich history of trying to approximate this function f by a linear combination of ridge functions. A ridge function with parameter $\theta \in \mathbb{R}^d$, is of the form $\phi(\theta^T x)$ where $\phi : \mathbb{R} \to \mathbb{R}$, sometimes called the activation function, is a bounded smooth non polynomial function with bounded derivatives. In neural network terminology, such a ridge function is often called a neuron and then a linear combination of ridge functions is called a one layer feed forward neural network. The estimated regression function is chosen from the class of all finite linear combinations of ridge functions with a given activation function. The activation function ϕ could be any smooth non polynomial sigmoidal function. This is because the span of $\{\phi(x^T\theta) : \theta \in \mathbb{R}^d\}$ is dense in all of $C(K)$ where $C(K)$ is the space of all continuous functions on a compact set $K \in \mathbb{R}^d$ as shown in [9]. A typical m term member of our function class is of the form $\sum_{j=1}^m w_j \phi(x^T\theta_j)$ where w_j are the weights or coefficients of the members of the dictionary. For estimation purposes, a full least squares involving all the parameters seems out of reach computationally. A reasonable method to obtain statistically accurate fits of f is to use a greedy algorithm [1]. The estimates in this algorithm can obtained iteratively by setting

$$\hat{g}_k(x) = (1 - \alpha)\hat{g}_{k-1}(x) + \beta\varphi_k(x^T\theta).$$

One chooses the α and β by least squares and minimize over φ_k by choosing θ to satisfy

$$\frac{1}{n}\sum_{i=1}^{n} r_i \varphi(x_i^T \theta) \geq C \max_{\theta} \frac{1}{n}\sum_{i=1}^{n} r_i \varphi(x_i^T \theta)$$

where $r_i = r_{i,k-1} = y_i - \hat{f}_{k-1}(\theta)$, are the residuals and $0 < C \leq 1$. This is precisely the optimization problem we are concerned with here, with the objective function

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} r_i \varphi(x^T \theta) \tag{3.1}$$

Each step of this greedy algorithm requires us to globally maximize, up to a constant factor, an objective function of the above form as $J(\theta)$. This optimization problem is typically non convex and examples are known where there are exponentially many peaks in θ for certain choices of ϕ, r and x. Exact global optimization is known to be NP hard for ϕ equalling certain types of sigmoidal functions [2], [3]. Such problems have no deterministic provably good algorithms and are considered notoriously hard. A nice account of this problem is given in [8]. In this manuscript, we explore the possibility of determining a computationally manageable algorithm to maximize possibly multimodal objective functions such as (3.1) up to a constant factor.

3.1.2 Approximate Diffusion for Optimization

For the purposes of the greedy algorithm, our interest is in optimization of possibly multimodal functions in $\mathbb{R}^d$ with high dimensional settings in mind where even $d > 10$ can be typically thought of as high dimensional for us. Of particular interest to us are objective functions of the form (3.1) which are superpositions of ridge functions. It is also sufficient for our purposes to approximately maximize, that is maximize within a constant factor, say $C = \frac{1}{2}$,

of the maximum value. Here we explore what may be possible by stochastic optimization techniques. For optimization we need to sample from peaked distributions and the available methods consist of Markov chain methods with time homogenous transition probability rules like the Metropolis-Hastings algorithm and the time inhomogenous probability transition rules like the Simulated Annealing algorithm. The cornerstone of the Metropolis Hastings type methods is the fact that an aperiodic, irreducible Markov chain converges to its stationary distribution. The convergence can be extremely slow for multimodal target densities. The customary heuristic is that there is a reluctance for the chain to move from one local maxima to another because the transition moves have a preference towards higher density regions of the target.

In the 1980's, Simulated Annealing (see[2]) was developed to solve large optimization problems. The idea is to have a sequence of densities $\{p_t : t \geq 0\}$ which becomes more and more peaked at the maxima of the function. So presumably, we would be doing something good if we could sample from p_t where t is high. In this manuscript we sometimes call t as time. This is because we think of the probablity densities evolving with time. Now our task is to devise a stochastic process z_t whose distribution evolves with t towards the target density by closely matching the densities p_t. Ideally one would like z_t to have density exactly p_t. The basic transition step in traditional Simulated Annealing is such that if z_t already had density $p_{t+\delta}$ rather than p_t then $z_{t+\delta}$ too would have density $p_{t+\delta}$ where $\delta > 0$ is a small step size by which t increments. The transition steps are Markov chain steps invariant for $p_{t+\delta}$. However such transition rules lead to $z_{t+\delta}$ having marginal density different from $p_{t+\delta}$. If δ is very small then there exists results [6] showing that slow logarithmic growth of t is sufficient to approximate the target distribution in some sense. To reach a targeted t of order d as we will need, requires a number of steps

which is exponential in d. This makes us explore another idea, closely related in spirit to the annealing idea. We would like to devise transition probabilities so that the process at time t tracks the sequence of densities p_t more closely than in traditional Simulated Annealing. For this purpose, we now describe an algorithm which was laid out in a preliminary form in [10] and was referred to as adaptive annealing.

The task is to design a stochastic process which will maximize up to a constant factor a given function $J(\theta)$, possibly bounded and having bounded derivatives. We will start with a given distribution on $\mathbb{R}^d$, say p_0 which is easy to sample from. We define a sequence of densities $p_t(\theta)$ proportional to $e^{tf(\theta)}p_0(\theta)$. The aim is to track such a sequence of densities for $0 \leq t \leq T$ with sufficiently large T. With $t \to \infty$ the distributions given by the densities p_t converge to a distribution which has its support concentrated on the set of global maxima of the function $f(\theta)$. So the whole effort is to investigate whether we can sample from the target distribution p_T, with T high enough, starting from p_0 such that the computational effort needed grows polynomially with the dimension of the search space d. Generally there will exist multiple maps F such that $F \circ p_0 = p_T$. If we can find such a map which is easy to compute our task will be done. In general, it is hard to find such measure transforming maps. Taking a cue from the theory of diffusion processes we now describe a possible way to construct such a measure transforming map approximately. This would bring approximate diffusion into play for optimization as we now describe.

A diffusion process $\{z_t,\ 0 \leq t \leq T\}$ is denoted by the stochastic differential equation

$$dz_t = \mu(z_t, t)dt + \sigma(z_t, t)dB_t$$

where B_t is the standard d dimensional Brownian motion process. Under

suitable regularity conditions the time evolution of the probability density function is governed by a PDE named the Fokker-Planck equation and also called the Kolmogorov's forward equation:

$$\frac{\partial p(\theta,t)}{\partial t} = -\nabla^T(\mu(\theta,t)p(\theta,t)) + \frac{1}{2}\nabla^T\nabla(\sigma^2(\theta,t)p(\theta,t))$$

Here ∇ denotes the gradient operator with respect to θ and its inner product with a vector valued function is the divergence and its inner product with itself is the Laplacian operator. Let $(\overline{\mu}(t,\theta),\overline{\sigma}(t,\theta))$ be a reference solution for which the following holds:

$$0 = -\nabla^T(\overline{\mu}(\theta,t)p(\theta,t)) + \frac{1}{2}\nabla^T\nabla(\overline{\sigma}^2(\theta,t)p(\theta,t)).$$

A possible choice could be $(\overline{\mu},\overline{\sigma}) = (\frac{1}{2}\nabla\log p(\theta,t),1)$. Another possibility is $(\overline{\mu},\overline{\sigma}) = (0,\frac{1}{p(\theta,t)})$. We could even choose $(\overline{\mu},\overline{\sigma}) = (0,0)$ for that matter. In that case the only randomness is in sampling from the initial distribution and then each point would have a deterministic trajectory determined by the drift. It is easy to check that these choices are stationary distributions for the relevant density p. We can write a general drift function by adding a change function

$$\mu(\theta,t) = \overline{\mu}(\theta,t) + v(\theta,t)$$

and set the variance function to be $\sigma^2(\theta,t) = \overline{\sigma}(t)^2$ where $(\overline{\mu}(.,t),\overline{\sigma}(t))$ are some reference drift and variance functions for p_t The tactic now is to treat the sequence of densities $p(\theta,t) = p_t(\theta)$ as given and choose a suitable velocity field $v(.,.)$ in order to track them. It can then be checked that v needs to satisfy the following simplified Fokker Plank equation

$$\frac{\partial p(\theta,t)}{\partial t} = -\nabla^T(v(\theta,t)p(\theta,t)) \tag{3.2}$$

So to summarize, we advocate solving for the time dependent velocity field v in (3.2) and then using this v to devise a stochastic process z_t with drift function $\overline{\mu_t} + v(t, \theta)$ and a variance function $\overline{\sigma_t}^2$. This is in contrast with traditional use of the Fokker Plank equation which, given the velocity field asks for how the probability densities evolve. If the velocity field $v(t, \theta)$ satisfies regularity conditions, by Fokker Plank the process z_t will track the sequence of densities p_t and finally at $t = T$ we would sample from p_T which in turn would help in approximate maximization. We call this process Adaptive Annealing as it is very similar in spirit to Simulated Annealing but the transition steps are inspired from Fokker Plank equation and not from Markov Chain theory.

Once we have candidate drift functions satisfying Fokker Planck the next task is to numerically implement this continuous time stochastic diffusion. Numerically we need to take steps of the form $z_{t+\delta} = z_t + \delta v(z_t, t)$ where δ is the stepsize and t is the gain or inverse temperature. Hence the process z_t would be an approximate diffusion process. Unfortunately finding numerically stable velocity fields have been hard and implementing the approximate diffusion even in one or two dimenisons require a fair amount of computational resources. We are still experimenting with these solutions and though conducting and interpreting the results have been cumbersome and hard, we believe that our basic idea is interesting and deserves further attention. The above ideas were sketched in preliminary form in [10] and in this manuscript we record additional advancements and efforts.

3.2 Error from Discretization

The velocities we use to carry out our diffusion process are solutions of the Fokker Plank equation (3.2). To implement the diffusion in practice, we need to discretize and move our process in timesteps. This will introduce errors in tracking the sequence of densities we want and there will be accumulation of errors in each step. We now attempt to do the error analysis. Although in practice one can make Forward Euler type movements of the form

$$z_{t+\delta} = z_t + \delta v(t, z_t)$$

here we analyze,for reasons of ease, the backward Euler type movements of the form

$$z_{t+\delta} = z_t + \delta v(t, z_{t+\delta}). \tag{3.3}$$

The velocity field v is evaluated at the next point $z_{t+\delta}$ and so is analogous to backward Euler steps to solve an ODE. it is well known that Backward Euler methods can be much more stable than the forward Euler methods in many problems. The drawback of Backward Euler steps is that one needs to typically find the root of a non linear equation in every step. We analyze the backward Euler steps as it is more amenable to the usage of the change of variables Theorem as we will see. Let $\{p_t : 0 \leq t \leq T\}$ be a sequence of densities we would want to track. The goal is to have $z_t \sim p_t$ but there will be errors due to discretization of the variable t. We want to investigate the local error after taking a single step. Let us assume z_t is actually distributed as p_t for some t. When we take the step (3.3), let us denote the actual probability density of $z_{t+\delta}$ by $\tilde{p}_{t+\delta}$. Then we want to examine the discrepancy between $p_{t+\delta}$ and $\tilde{p}_{t+\delta}$. Let us denote $z_{t+\delta}$ by w to lessen the number of subscripts in the argument.

We assume the velocity field v satisfies the regularity conditions needed for the change of variables Theorem to hold. By the change of variables Theorem we have

$$\tilde{p}_{t+\delta}(w) = p_{t(w-\delta v(t,w))}|det(I + \delta J(t, w - \delta v(t, w)))|^{-1}$$

Now taking log of the above equation and doing Taylor's expansion up to second order terms we get

$$\log(\tilde{p}_{t+\delta}(w)) \simeq \log(p_t(w)) - \delta v(t,w)^T \nabla log p_t(w) + \frac{\delta^2}{2} v(t,w)^T H(\log p_t(w)) v(t,w)$$
$$- logdet(I + \delta J(t, w - \delta v(t, w))$$

where $H(\log p_t(w))$ is the Hessian matrix of $\log p_t$ evaluated at w and $\simeq$ means equality up to second order in δ. Now we can also Taylor expand the log determinant term to obtain

$$\log(\tilde{p}_{t+\delta}(w)) \simeq \log p_t(w) - \delta v(t,w)^T \nabla log p_t(w) + \frac{\delta^2}{2} v(t,w)^T H(\log p_t(w)) v(t,w)$$
$$- \delta Tr(J(t,w) + \delta^2 v(t,w)^T \nabla Tr(J(t,w)) + \frac{\delta^2}{2} \sum_{i=1}^{d} |\lambda_i(J(t,w))|^2$$

where $\lambda_i(J(t,w))$ is the square of the modulus of the eigenvalues of $J(t,w)$. The eigenvalues of $J(t,w)$ may be complex values as $J(t,w)$ need not be symmetric. So from the last display we get

$$\log(p_{t+\delta}(w)) - \log(\tilde{p}_{t+\delta}(w)) \simeq \log(p_{t+\delta}(w)) - \log(p_t(w)$$
$$- \delta \left(v(t,w)^T \nabla log p_t(w) + Tr(J(t,w)) \right) +$$
$$\frac{\delta^2}{2} \left(v(t,w)^T H(\log p_t(w)) v(t,w) - 2 v(t,w)^T \nabla Tr(J(t,w)) + \sum_{i=1}^{d} |\lambda_i(J(t,w))|^2 \right).$$

Now by Taylor, expanding $\log(p_{t+\delta}(w))$ about t in the right side of the above display, we obtain

$$\log(p_{t+\delta}(w)) - \log(\tilde{p}_{t+\delta}(w)) \simeq -\delta \left(\frac{\partial \log p_t(w)}{\partial t} + v(t,w)^T \nabla log p_t(w) + Tr(J(t,w)) \right)$$
$$+ \frac{\delta^2}{2} \left(\frac{\partial^2 \log p_t(w)}{\partial t^2} v(t,w)^T H(\log p_t(w)) v(t,w) - 2v(t,w)^T \nabla Tr(J(t,w)) \right.$$
$$\left. + \sum_{i=1}^{d} |\lambda_i(J(t,w))|^2 \right). \tag{3.4}$$

The coefficient of δ is exactly 0 when v satisfies the Fokker Plank PDE (3.2). Hence the local error is of order δ^2 with the coefficient as above. Unfortunately, for the velocity fields v we have been able to find by solving (3.2), we have not been able to show that the coefficient of δ^2 remains bounded so that we could claim a order δ cumulative or global error of pointwise approximation of log density.

3.3 Variable Augmentation Formulation

In this section we introduce a variable augmentation formulation of our optimization problem which appears to be potentially useful for us. We will denote the new variables by u which would be a $n$dimensional vector where n is the number of samples in our function estimation problem. Henceforth in this manuscript, we use θ in general sampling problems and u specifically when we are dealing with our variable augmentation problem. We describe the formulation as follows. We want to find an approximate global maximizer

of functions of the form

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} r_i \phi(x_i^T \theta) \tag{3.5}$$

where ϕ is some non-polynomial sigmoidal bounded function of one variable with bounded derivatives. We seek ϕ for which computation of the approximate maximizer might be manageable. We can assume that the r_i are given weights bounded in magnitude by 1 and n is a large integer bigger than d. We turn the optimization problem into a sampling problem and would like to sample from

$$p_t(\theta) \propto \exp(t\frac{1}{n}\sum_{i=1}^{n} r_i \phi(x_i^T \theta) - \frac{\|X\theta\|^2}{2\sigma^2}) \tag{3.6}$$

where t is high enough. We will discuss later how high this t needs to be. We prefer another formulation of the sampling problem in the last display. We introduce new variables $u = (u_1, u_2, ...u_n)$ such that the joint distribution for u and θ is

$$p_t(u,\theta) \propto \exp\,(\frac{t}{n}\sum_{i=1}^{n} r_i \phi(u_i))\frac{1}{(\sqrt{2\pi}\sigma)^n}\exp(-\frac{\|X\theta\|_2^2}{2n\sigma^2})\frac{1}{(\sqrt{2\pi}\alpha)^n}\exp(-\frac{\|u - X\theta\|^2}{2\alpha^2}). \tag{3.7}$$

The u_i is meant to be a surrogate for $x_i^T\theta$ and thus the term $\exp(-\frac{\|u-X\theta\|^2}{2\alpha^2})$ which keeps the u constrained to be near $X\theta$. We note that we can explicitly integrate out θ from the joint density to get marginal distributions of u and conditional distributions of θ given u. Hence it can be checked that the marginal density for u becomes

$$p_t(u) \propto \exp\,(\frac{t}{n}\sum_{i=1}^{n} r_i \phi(u_i) - \frac{u^T M u}{2}) \tag{3.8}$$

Here $M = \dfrac{I - \dfrac{\sigma^2}{\sigma^2+\alpha^2}P_X}{\alpha^2}$ and P_X denotes the orthogonal projection matrix onto the column space of X. Also the conditional density of θ given u becomes

$$\theta \mid u \sim N(\frac{(X^TX)^{-1}X^Tu}{1+\frac{\alpha^2}{\sigma^2}}, \frac{(X^TX)^{-1}}{\frac{1}{\sigma^2}+\frac{1}{\alpha^2}}) \tag{3.9}$$

This shows that provided we can sample from the marginal distribution of u it is easy to sample from the conditional distribution of θ given the u and hence easy to sample from the joint distribution of (u, θ). But we are really interested in the marginal distribution of θ. We now examine the form of the marginal distribution of θ. It is obtained by integrating out $u_1, u_2, ...u_n$ in the joint density. Since the u_i are conditionally independent given θ we can separately integrate out each u_i. So we get

$$p_t(\theta) \propto \exp -\frac{\|X\theta\|^2}{2n\sigma^2} \prod_{i=1}^{n} \int_{-\infty}^{\infty} \exp \frac{t}{n} r_i\phi(u_i) \frac{1}{\sqrt{2\pi}\alpha} \exp \frac{(u_i - {x_i}^T\theta)^2}{2\alpha^2} du_i \tag{3.10}$$

Now if $\frac{t}{n}$ is small, then we can write $\exp \frac{t}{n} r_i\phi(u_i) = 1 + \frac{t}{n} r_i\phi(u_i) + O(\frac{T^2}{n^2})$ since ϕ is bounded. Consider the product of the integrals

$$\Pi_{i=1}^{n} \int_{-\infty}^{\infty} \exp \frac{t}{n} r_i\phi(u_i) \frac{1}{\sqrt{2\pi}\alpha} \exp \frac{(u_i - {x_i}^T\theta)^2}{2\alpha^2} du_i \tag{3.11}$$

Expanding $\exp \frac{t}{n} r_i\phi(u_i)$, and since we are integrating against a Gaussian density we can write the integrand in (3.11) as

$$1 + \frac{t}{n} r_i \int_{-\infty}^{\infty} \phi(u_i) \frac{1}{\sqrt{2\pi}\alpha} \exp \frac{(u_i - {x_i}^T\theta)^2}{2\alpha^2} + O(\frac{t^2}{n^2}).$$

Define

$$\tilde{\phi}(t) = \int_{-\infty}^{\infty} \phi(u_i) \frac{1}{\sqrt{2\pi}\alpha} \exp \frac{(u_i - t)^2}{2\alpha^2} du_i.$$

Then we can simplify (3.11) and write the product of integrals as

$$\prod_{i=1}^{n} \left(1 + \frac{t}{n} r_i \tilde{\phi}(x_i{}^T\theta) + O(\frac{t^2}{n^2})\right).$$

Taking the log of the product of the integrals and writing $\log(1+x) = x+O(x^2)$ when x is small with $x = \frac{t}{n} r_i \phi(u_i) + O(\frac{t^2}{n^2})$ gives us

$$\sum_{i=1}^{n} \log \left(1 + \frac{t}{n} r_i \tilde{\phi}(x_i{}^T\theta) + O(\frac{t^2}{n^2})\right) = \sum_{i=1}^{n} \left(\frac{t}{n} r_i \tilde{\phi}(x_i{}^T\theta) + O(\frac{t^2}{n^2})\right)$$

Since summing over n terms of $O(\frac{t^2}{n^2})$ gives us a term of $O(\frac{t^2}{n})$ we obtain that the log of (3.11) is

$$\left(\sum_{i=1}^{n} \frac{t}{n} r_i \tilde{\phi}(x_i{}^T\theta)\right) + O(\frac{t^2}{n}).$$

So we can finally write the marginal density of θ as

$$p(\theta) \propto \exp -\frac{\|X\theta\|^2}{2n\sigma^2} \exp (\sum_{i=1}^{n} \frac{t}{n} r_i \tilde{\phi}(x_i{}^T\theta)) \exp \left(O(\frac{t^2}{n})\right). \tag{3.12}$$

We compare (3.6) and (3.12) and we notice that the marginal density of θ in (3.12) is exactly of the same form that we want with $\phi = \tilde{\phi}$ and there is an extra term $\exp \left(O(\frac{t^2}{n})\right)$ which will be negligible if $\frac{t^2}{n}$ is small. This shows that if we could sample from the marginal density of u with inverse temperature t and then sample θ given u which is Gaussian, we are succesful in sampling from the distribution on the θ space which we want. Now the question is how high should the multiplier t be? The following proposition shows that we need the final multiplier $T = O(d \log d)$. The proposition implies that we would

need the sample size n to be of a higher order than T^2. For example, having $n = d^3$ would suffice for our purposes.

Proposition 3.3.1. *Let p_t be as defined in (3.6). Let $E_{p_t} J$ denote the expectation of the function $J(\theta)$ with respect to the distribution whose density is p_t and $\theta \in \mathbb{R}^d$ be any given vector. Also let $T = O(d \log d)$. Then for all bounded J with bounded derivatives which has a global maxima, there exists a constant $0 < C < 1$ such that the following holds for all dimensions d*

$$E_{p_T} J > C \ max_\theta obj(\theta). \tag{3.13}$$

The proof of this proposition is outlined in the appendix. Since the expectation of our objective function can be shown to be bigger than a constant factor C of the global maximum value of the objective function, if we sample from p_T with $T = O(d \log(d))$ sufficiently many times, the probability of getting a point with the value of the objective function greater than the maximum value upto a constant factor is also high by Markov's inequality.

There are two reasons we like this variable augmentation formulation. One is that it makes the left side of the Fokker Plank equation sums of functions of single variables and hence lends itself to some natural solutions for the required velocity field as will be explained in the next section. Secondly, for any single variable u_1 say, the conditional distribution of it given all the other variables is very close to a Gaussian. We explain in section (3.5) as to how that might be useful.

3.4 Some Solutions of Fokker Plank and Simulations

In this section, we describe how we can solve for the velocity field in (3.2). We also present simulations showing the trajectories of the approximate diffusion running with these velocity fields.

3.4.1 Solution in 1 Dimension

Let us consider the simplest possible situation, which is the 1 dimensional case. In this setting, our problem is to track the sequence of densities of the form

$$p_t(\theta) \propto exp(tJ(\theta))exp(-\theta^2/2)$$

where the objective function J is of the form

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} r_i\varphi(x_i\theta).$$

In this case our probability density evolution equation becomes

$$\frac{\partial p_t(\theta)}{\partial t} = -\frac{\partial}{\partial\theta}(v_t(\theta)p_t(\theta))$$

Now $\frac{\partial p_t(\theta)}{\partial t}$ equals $p_t(\theta)(J(\theta) - E_tJ(\theta))$ where $E_tJ(\theta)$ is the expectation of J under the distribution whose density is p_t. Hence the equation we need to solve for becomes

$$p_t(\theta)(J(\theta) - E_tJ(\theta)) = -\frac{\partial}{\partial\theta}(v_t(\theta)p_t(\theta)) \tag{3.14}$$

The general solution for v in this case is

$$\frac{1}{p(t,\theta)}\int_c^{\theta} p(t,w)(J(w)-E_t(J))dw.$$

We prefer taking $c=-\infty$ because then, as we will see, one has a tapering behaviour of v at the tails. Then our solution v becomes

$$-\frac{1}{p(t,\theta)}\int_{-\infty}^{\theta} p(t,w)(J(w)-E_t(J))dw$$

We can reverse the limits and write our solution as

$$\frac{1}{p(t,\theta)}\int_{\theta}^{\infty} p(t,w)(J(w)-E_t(J))dw$$

The two ways of writing are exactly the same because $\int_{-\infty}^{\infty} p(t,w)(J(w)-E_t(J(w))dw = 0$. With J bounded and p_t proportional to $e^{(tJ(\theta)-\theta^2/2)}$ which has Gaussian tails the integral is seen to be controlled for θ close to $-\infty$ by the tail integral of the Gaussian, which is bounded by a constant times $\frac{1}{\theta}e^{-\theta^2/2}$. Consequently, despite the division by $p_t(\theta)$, the velcoity field $v(t,\theta)$ is seen to taper to 0 at a polynomial rate as θ goes to $-\infty$. Analogously, we can use the other representation of v and prove that v goes to 0 as θ goes to $+\infty$. at the same polynomial rate. We can also examine the derivative of v w.r.t to θ.

$$v'(t,\theta) = -(J(\theta)-E_tJ) + v(t,\theta)\frac{p_t'(\theta)}{p_t(\theta)}$$

Now the first term is bounded and in the second term we know v goes down like $\frac{1}{\theta}$ but derivative of p divided by p goes up like θ and hence the derivative of v remains bounded for each t. Now just for some qualititive understanding,

we can also write our solution as

$$v(t,\theta) = -\frac{\int_{-\infty}^{\theta} p(t,w)dw}{p(t,\theta)}(E_t(J|z \leq \theta) - E_t(J)).$$

The above equation can be understood in the following manner. At time t, wherever the process is, it compares the conditional mean given that it is less than the current value and the global mean with respect to p_t. If the conditional mean is greater than the overall mean v the solution is negative which means the process would go to a value less than the current value which intuitively makes sense. Also the jump sizes are inversely proportional to the conditional density at the current value given that the process is less than the current value. So in this one dimensional case it seems the solution v is manageable in terms of its derivatives being bounded and it not varying exponentially.

3.4.2 Extension to higher dimensions?

In dimension d we write the Fokker Plank pde as follows

$$-\frac{\partial \log p(\theta,t)}{\partial t} = \sum_{i=1}^{d}\left(\frac{\partial v_i(t,\theta)}{\partial \theta_i} + v_i(t,\theta)\frac{\partial \log p(\theta,t)}{\partial \theta_i}\right). \tag{3.15}$$

One way to satisfy the above equation is to set each term in the right side of (3.15) to be $1/d$ times the left side in (3.15) and then we are left with d one dimensional ode's to solve. For all $i \in [1:d]$ this results in solutions of the form

$$v_i(t,\theta) = -\frac{1}{p_t(\theta)}\int_{c(\theta_{-i})}^{\theta_i}(\frac{\partial p(\theta,t)}{\partial t})d\theta_i. \tag{3.16}$$

where $c(\theta_{-i})$ potentially depends on everything apart from θ_i. It is evident that instead of uniformly multiplying by $\frac{1}{d}$ one can also have fixed weights for each

coordinate which sum up to 1. If we consider the variable augmented form of our target densities we will see that the left side of (3.2) is of an additive form. This is one of the reasons we like the augmented formulation. We recall that our search space becomes n dimensional instead of d, where n is the sample size in our function estimation problem. In this setting, our Fokker Plank pde is

$$-\frac{\partial \log p(u,t)}{\partial t} = \sum_{i=1}^{n} \left(\frac{\partial v_i(t,u)}{\partial u_i} + v_i(t,u) \frac{\partial \log p(u,t)}{\partial u_i} \right).$$

Here $p_t(u)$ is as defined in (3.29). In this case it can be checked that

$$\frac{\partial \log p(u,t)}{\partial t} = \frac{1}{n} \sum_{i=1}^{n} r_i(\phi(u_i) - E_t\phi(u_i)) \tag{3.17}$$

where $E_t\phi(u_i)$ refers to expectation of $\phi(u_i)$ under the distribution whose density is p_t. A natural way to solve the Fokker Plank in this case is to equate terms for all $i \in [1:n]$ in the following way

$$\frac{1}{n} r_i(\phi(u_i) - E_t\phi(u_i)) = \frac{\partial v_i(t,u)}{\partial u_i} + v_i(t,u) \frac{\partial \log p(u,t)}{\partial u_i}$$

This results in the following solutions

$$v_i(t,\theta) = -\frac{1}{p_t(u_i|u_{-i})} \int_{c(u_{-i})}^{\theta_i} \frac{1}{n} r_i(\phi(u_i) - E_t\phi(u_i)) p_t(u_i|u_{-i}) du_i \tag{3.18}$$

where $p_t(u_i|u_{-i})$ is the conditional density of u_i given the others u_{-i}.

Unfortunately, these solutions exhibit instabilities when we implement them numerically. Let us explain the problem in the solution given by (3.18). We note that the solutions have $p_t(u_i|u_{-i})$ in the denominator. The denominator goes to 0 exponentially fast as u_i goes to $\pm\infty$ because f normal tails. Hence unless the numerator goes to 0 atleast equally fast the velocities blow up rather

quickly which we have seen experimentally. If we use $c(u_{-i}) = -\infty$ as in the case of 1 dimension, we can write the solution in (3.18) as a sum of two terms

$$-\frac{1}{p_t(u_i|u_{-i})}\left(\int_{-\infty}^{\theta_i}\frac{1}{n}r_i(\phi(u_i)-E_t[\phi(u_i|u_{-i})]p_t(u_i|u_{-i})du_i - \frac{E_t\phi(u_i)|u_{-i}-E_t\phi(u_i)}{n}\int_{-\infty}^{\theta_i}r_i p_t(u_i|u_{-i})du_i\right).$$

The first term again goes to 0 polynomially fast for the same reasons, as explained in the 1 dimensional case, as we approach infinities and hence is reasonable. The second term however behaves like a constant when plus infinity is approached so the denominator makes it blow up extremely fast. We have studied these solutions experimentally a lot and they do not seem to work. The trajectories blow up very quickly.

Another avenue one can consider is setting $v(t,\theta)p(t,u) = \nabla\psi(t,u)$ for some potential function ψ. In this case our Fokker Plank equation reduces to the familiar second order PDE known as the Poisson s equation.

$$-\frac{\partial p(u,t)}{\partial t} = \sum_{i=1}^{p}\left(\frac{\partial^2\psi(t,u)}{\partial u_i^2}\right). \tag{3.19}$$

One has analytical expressions for solutions of the Poisson's equation as an integral over the n dimensional space which cannot be computed in dimension more than 3. Another problem with this approach is that even in dimension 2 or 3, after solving for ψ one has to divide by p in order to obtain v. The trajectories then again quickly blow up.

3.4.3 Simulations in 1 dimension

In this subsection I present some of the simulations that we have done in order to investigate our method of using diffusion for approximate optimization for sampling 1 dimensional distributions. We take ϕ to be the arctan function scaled so that its limits are ± 1 at $\pm\infty$. Our objective function is of the form

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} r_i \varphi(x_i\theta).$$

For simplicity, we take all the r_i to be 1. We also take $n = 10$ and generate the x vector by sampling from a normal distribution. We end up with an objective function which looks like as shown in Fig 1. This is a challenging function to optimize because this has a broad local maxima around -0.2 and a very narrow global maxima at 0.3.We simulate our initial state θ_0 from a standard normal and then we follow the update rule $\theta_t = \theta_{t-1} + \delta v_t(\theta_{t-1})$ where v is a solution as given in (3.4.1) As mentioned before, v can be rewritten as

$$\frac{\int_{-\infty}^{\theta} p(t,w)dw}{p(t,\theta)}[condl - global]$$

where $condl = \frac{\int_{-\infty}^{\theta} p(t,w)J(w)dw}{\int_{-\infty}^{\theta} p(t,w)dw}$ and $global = E_t(J)$. Here, $condl$ denotes the conditional expectation of $J(w)$ given $w \leq \theta$ with respect to the density p_t and $global$ denotes the global mean of f with respect to the density p_t.

To compute $v_t(\theta)$ at some time t and state θ we need to evaluate 3 key quantities:

$$I(t,\theta) = \int_{-\infty}^{\theta} J(w)p(t,w)dw \tag{3.20}$$

$$F(t,\theta) = \int_{-\infty}^{\theta} p(t,w)dw \tag{3.21}$$

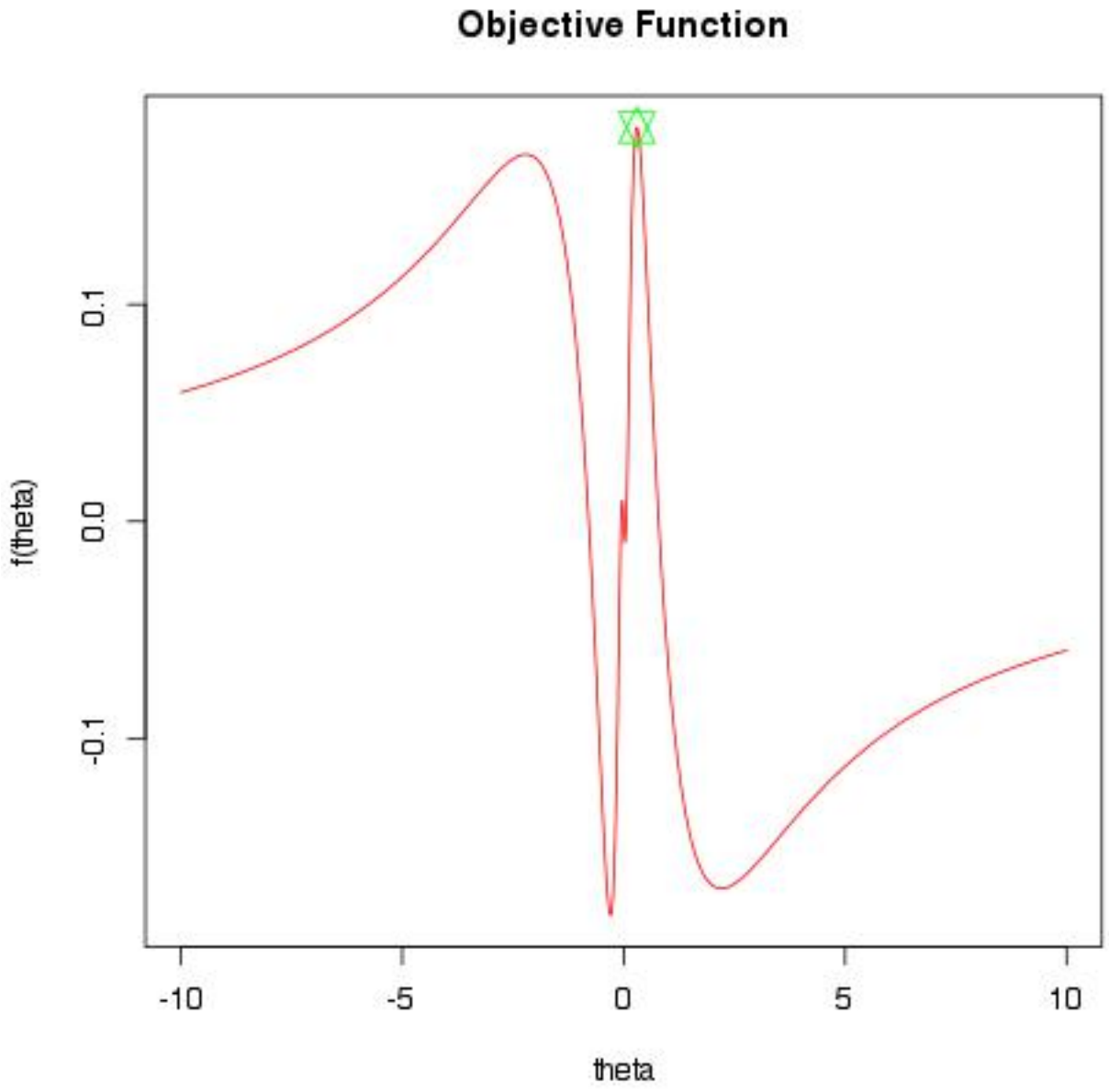

Figure 3.1: Objective Function

$$E_t = \int_{-\infty}^{\infty} J(w)p(t,w)dw \tag{3.22}$$

The quantities in (3.20) and (3.21) are used to compute *condl* and (3.22) is the overall mean of f with respect to the density p_t. To compute (3.20) and (3.21) we use the command integrate in R which performs numerical integration

once we specify the upper and lower limits and the number of subdivisions. Although we are now in the 1 dimensional case, even in higher dimensions we have to compute an analogous one dimensional integration which can be interpreted as a conditional mean and we can continue using the same R command. To compute (3.22) we could have used numerical integration in this one dimensional case. But in general, we would have to compute the overall mean anyways and in that case it is a d dimensional integration and is very expensive to compute and is in the first place the reason why we are investigating this whole approach. What we do is instead of computing the global mean with respect to the density p_t exactly, we estimate it by running several chains at once.

To be particular, using $\delta = 0.1$, in this case we run 100 chains and hence estimate the global mean at every time point t by taking the average across all the 100 chains. This will be a key feature of our proposed algorithm and is also the second source of departure from tracking p_t, the first being the discretization of the timesteps t. Since we have a 100 starting positions and a 100 chains we have a histogram of final positions after timesteps t =0, 25, 50, 100 which we show below: We can observe(fig 2) that at time $t = 0$ the positions were simulated from a standard Normal but as we move ahead in timesteps the picture starts to change. At $t = 25, 50$ the highest frequency position seems to be around 0. The global maxima and the second maxima(fig 1) are pretty much equal in frequency at these times. But at timestep $t = 100$ the global maxima starts to dominate and the highest frequency term is around 0.3 which is exactly where the global maxima is. So it seems that the process performs as how we would expect it to. Of course it is one particular objective function and one particular realization of 100 chains. In these simulations we have been using $\overline{\mu}, \overline{\sigma} = (0, 0)$ once the starting point has been determined then

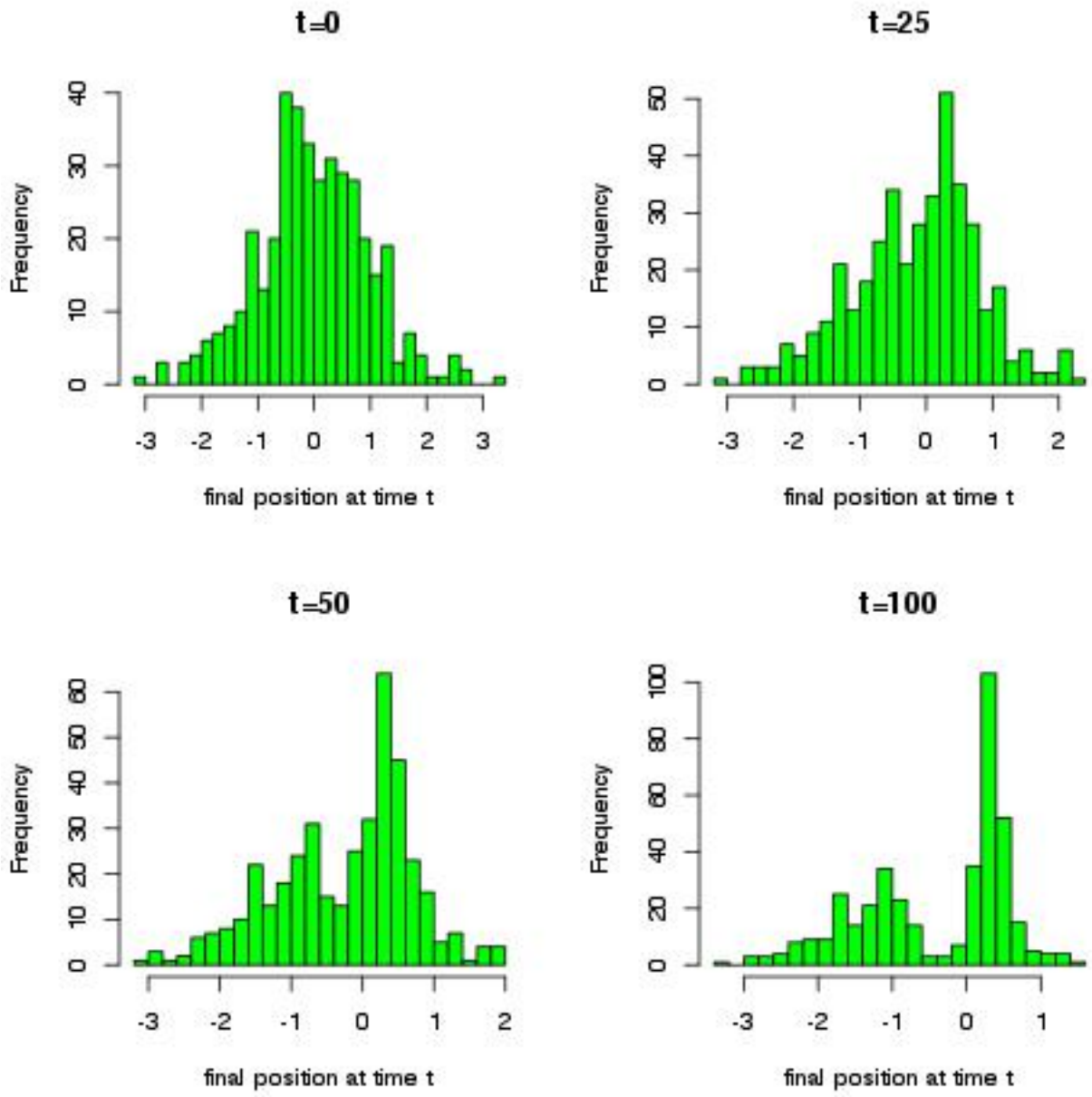

Figure 3.2: Histograms

there is a deterministic trajectory for a chain to follow. We now plot these trajectories for some chosen starting points, with timestep going from 0 to $t = 1000$. From fig 1 we can note that there is a narrow global maximum is at 0.3 and a broad local maxima around -0.2. Now fig 3 reveals some interesting facts. If we look at starting points $-0.2, 0.1$ and 1 they go straight to the

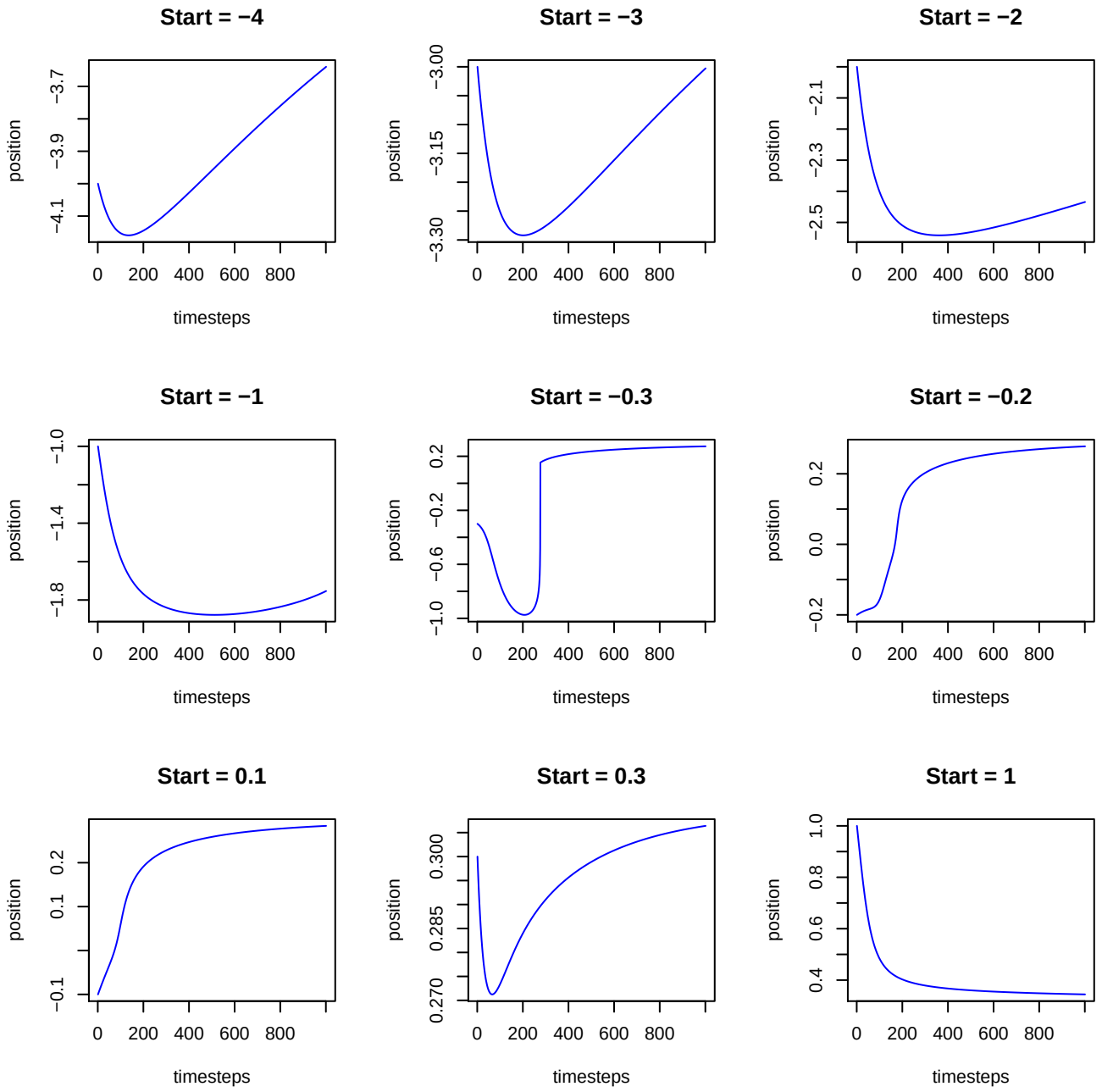

Figure 3.3: Trajectories with various starting points

global maximum around 0.3, all within 1000 timesteps. If we look at the starting point −0.3 which is a global minimum it first travels upward to the right which is the wrong direction. It keeps on going in the wrong direction till timestep about 200 and then zooms back and reaches the vicinity of 0.3.

This shows that our process is just not an ascent process. Similarly starting points like $-1, -2, -3$ all go to the left initially and then turn around at some point. The timesteps required till the change of direction varies. For example, starting point $= -1$ takes 600 timesteps to turn around and we can see in fig. 6 that even after timestep 1000 it has not really reached anywhere close to the global maxima. In fact with $\delta = 0.1$, for starting points between -0.6 and -1, which corresponds to the valley between the two maxima, we cant seem to run our chain for much longer than 1000 timesteps. The chain shoots off to infinity. We have to make δ smaller, say 0.01 to run the chain till $t = 100$ which means 10000 timesteps. This is the problem of discretizing a continuous time solution.

Remark 3.4.1. *This illustrates that the trajectory of some points may require smaller stepsizes than others.*

We did try some other objective functions by changing the x vector but keeping the functional form the same. We got similar results and sometimes we did see a chain shooting off to infinity. But the instances seem to be rare which backs our notion that the chain with high probability would stay in a region which would allow us to approximate the actual trajectory well. Unfortunately, the 1 dimensional case is not really representative of higher dimensions as our candidate velocity field has bounded derivative and tapers to 0 at $\pm\infty$ which is not the case in higher dimensions.

To conclude, we think that although Adaptive Annealing works reasonably well in 1 dimension one may see trajectories shooting off to infinity here and then. Of course there are more direct ways to sample univariate distributions. Adaptive annealing actually does more that just sample the final distribution p_T. It tracks the entire sequence of densities $\{p_t : 0 \leq t \leq T.\}$

Remark 3.4.2. *In 1 dimension there is of course a much easier thing to do. Let F and G be distribution functions of two probability distibutions absolutely continuous with respect to Lebesgue measure on the real line. Let F be our starting distribution and G be the target distribution we want to sample from. The map GF^{-1} will transport F from G. In fact the theory of optimal transport [11] says that this map is optimal in a certain sense.*

3.4.4 Sampling in 2 dimensions

In this section, we show the right way to use Adaptive Annealing to track a sequence of densities p_t defined on $\mathbb{R}^2$ with variables $\theta = (\theta_1, \theta_2)$. The Fokker Plank equation can be written as

$$\frac{\partial p_t(\theta)}{\partial t} = -\frac{\partial}{\partial \theta_1}(v_1(t,\theta)p_t(\theta)) - \frac{\partial}{\partial \theta_2}(v_2(t,\theta)p_t(\theta)) \tag{3.23}$$

In this case, resorting to breaking up the left side of the above and equating it to the two corresponding parts is a valid solution theoretically but we have observed them failing in practice as trajectories shoot off to infinity. A better solution in this case is to set $v = \nabla\psi$. That is, v itself is a gradient of some potential function ψ. It is a fact that among all possible solutions of (3.23) the v that is itself a gradient is a minimum norm solution minimizing $\int_0^T \int_{\mathbb{R}^2} |v(t,\theta)|_2^2 p(\theta,t)$ as can be seen in [11]. This is the key step in finding a well behaved solution for v. So by setting $v = \nabla\psi$ we are left with the following PDE to solve in ψ.

$$-\frac{\partial p_t(\theta)}{\partial t} = \frac{\partial^2}{\partial \theta_1^2}\psi(t,\theta) + \frac{\partial^2}{\partial \theta_2^2}\psi(t,\theta) + \frac{\partial}{\partial \theta_1}\psi(t,\theta)\frac{\partial}{\partial \theta_1}p(t,\theta) + \frac{\partial}{\partial \theta_2}\psi(t,\theta)\frac{\partial}{\partial \theta_2}p(t,\theta). \tag{3.24}$$

We numerically solve this PDE in a grid in 2 dimensions to to get a solution ψ. We then take numerical derivatives of ψ to obtain values for v in this grid. We then interpolate, correct upto first order, to obtain our velocity field v at any point inside the grid. While solving for (3.24) we represent the derivative operators by appropriate matrices. We do not go into the details of the constructions of such matrices here. We also have to choose a grid and specify a boundary value for ψ to solve the problem. For particular problems we have to choose this grid so that points are highly unlikely to go outside the grid during their entire trajectory. We generally set the boundary value to be 0 in our experiments unless stated otherwise. The boundary values have the potential to impact the behaviour of the solution near the boundary of the grid. Ideally we want to take the grid to encompass a wide range but we also need a fine enough grid for the interpolation of the solution to be close to correct. Also finer the grid, more the number of grid points and more computational resources are required to solve the extremely large but sparse linear systems that arise when solving such pde's. We now demonstrate one particular simulation that we did where taking the boundary values as 0 does not seem to affect the sampling of target densities.

3.4.5 Simulations in 2 dimensions

In this subsection we show simulations of sampling a mixture of Gaussian distributions with Adaptive Annealing. We start from a simple Gaussian distribution with mean 0 and covariance matrix a diagonal matrix with standard deviations 0.3 Let the density of this distribution be denoted by p_0. We choose the target density to be a bimodal density. We choose it to be an equally weighted mixture of Gaussians with mean vectors $c(0.6, 0.6)$ and $c(-0.6, -0.6)$

and let its density be denoted by p_1. We choose the log linear path of densities $p_t \propto p_0^{1-t} p_1^t$ to be tracked by our approximate diffusion. In order to solve the equation (3.24) we need to compute key quantities which are the space derivatives of log density and the time derivative of the log density. The space derivatives turn out to be

$$\nabla \log p_t(u) = t \left(\log p_1(u) - \log p_0(u) \right).$$

The time derivative turns out to be

$$\frac{\partial \log p_t(u)}{\partial t} = \left(\log p_1(u) - \log p_0(u) \right) - \mathbb{E}_t \left(\log p_1(u) - \log p_0(u) \right)$$

where $\mathbb{E}_t$ refers to expectation with respect to the distribution whose density is p_t. Again, we do not numerically compute $\mathbb{E}_t \left(\log p1(u) - \log p0(u) \right)$ because in higher dimensions we would be estimatng the analogous quantity by running several chains in parallel. Instead we run 500 chains in this particular case and take our time stepsize to be 0.1 In other words, we take 10 steps in all for each chain to increase t from 0 to 1. The following plot shows the scatter plot of the chains at all the time points. We see clearly that at the start there is a single cluster and slowly the two clusters start to emerge and at the final instant, we see approximately half of the points in each of the clusters which is exactly what our target distribution was like. The interesting thing to note is that none of the points shot off to infinities. This means that the solution velocity fields are extremely well behaved. This is in contrast to the one dimensional situation where we saw instabilities, albeit rarely. This is somewhat impressive because the target density is bimodal and there is a valley of low probability density between the two modes as was the case in our single variable example. So, in this sense being in 2 dimensions is bet-

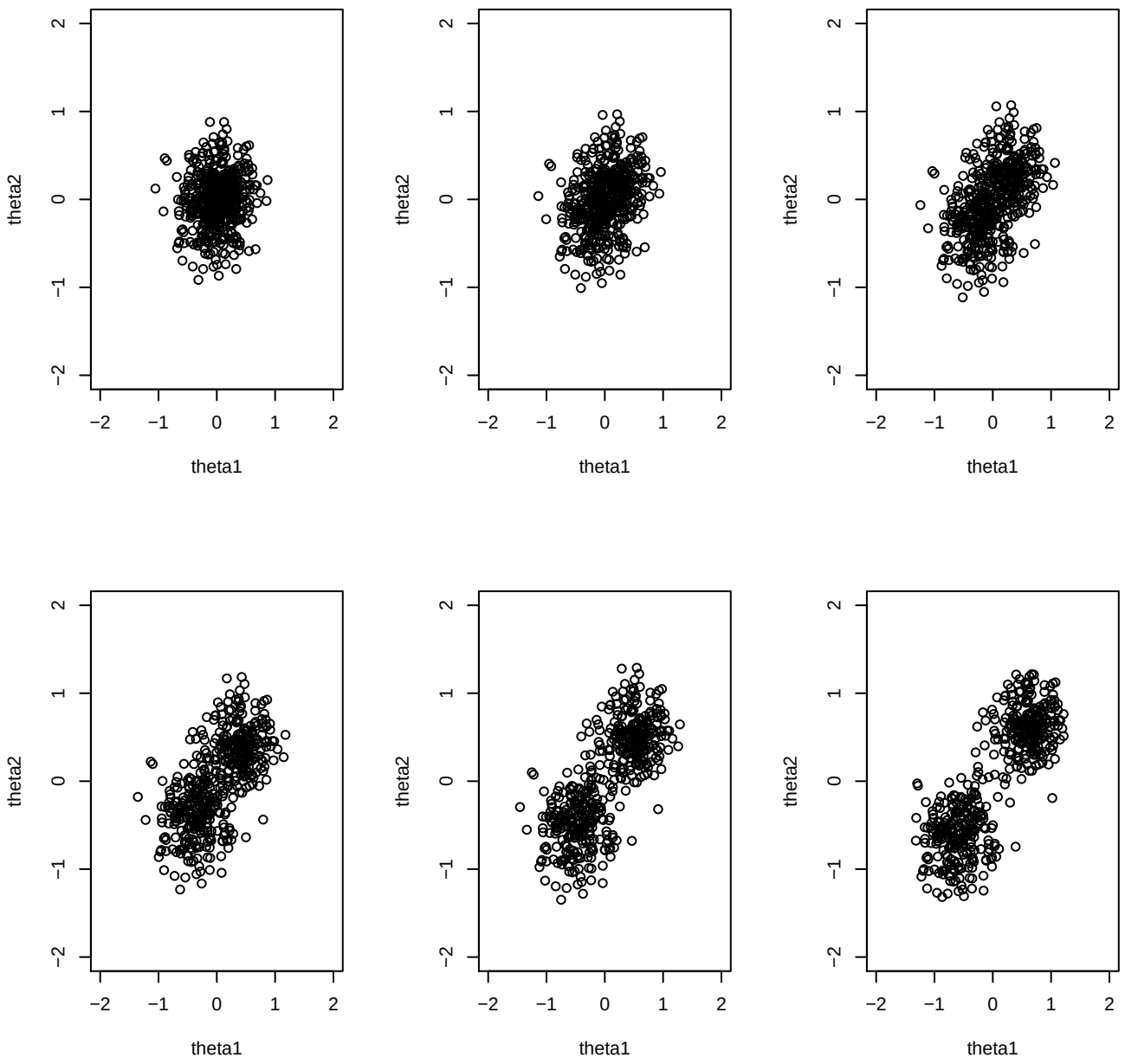

Figure 3.4: Sampling from a mixture of Gaussians

ter than 1. We can similarly carry out our approximate diffusion in order to sample from multimodal densities in 3 dimensions as well. Unfortunately, we cannot solve for the minimum norm velocity fields in higher dimensions. In that case, we need to rely on solutions which only need at most 3 dimensional

integrations to do or at most 3 dimensional partial differential equations to solve. There is one special case, where one indeed can sample by Adaptive Annealing from a high dimensional distribution with low dimensional differential equation solvers or low dimensional integration and that is sampling from an arbitrary multivariate Gaussian distribution as we describe in the following section.

3.5 Sampling Multivariate Gaussians and extensions

In this section we show how to sample from an arbitrary multivariate Gaussian distribution using Adaptive Annealing by just doing single variable integrations. Inspired from the solution to the multivariate Gaussian problem, we propose a way to solve for the velocity field in the Fokker Plank equation in the variable augmented setting. For simplicity, let the target multivariate Gaussian density have mean zero. The general mean situation can be handled easily. Let the covariance matrix of the target be Σ_1. We choose the starting distribution to be a zero mean Gaussian with a diagonal covariance matrix denoted by Σ_0. Let us choose any smoothly varying trajectory of covariance matrices $\{\Sigma_t : 0 \leq t \leq 1\}$ where Σ_1 is the target covariance matrix and Σ_0 is the initial covariance matrix. We denote $M_t = (\Sigma_t)^{-1}$. Some examples of such a trajectory would be $\Sigma_t = ((1-t)\Sigma_0 + t\Sigma_1)^2$ or the choice $\Sigma_t = ((1-t)M_0 + tM_1)^{-1}$. Then our sequence of distributions that we want to track are $\{N(0, \Sigma_t) : 0 \leq t \leq 1\}$. Let $p(t, u)$ denote the densities of $N(0, \Sigma_t)$. So we have

$$p(t, u) \propto exp(-\frac{1}{2} u^T M_t u). \tag{3.25}$$

The following Lemma says that for every t, a velocity function linear in u satisfies the Fokker Plank pde.

Lemma 3.5.1. *Let $\{p_t : 0 \leq t \leq 1\}$ denote densities as in (3.25). Let Σ_t be differentiable in t entrywise. Let $\frac{\partial M_t}{\partial t}$ be the matrix obtained by differentiating the matrix M_t as a function of t entrywise. Also let A_t be the matrix defined as $A_t = \frac{\partial M_t}{\partial t}\Sigma_t$. Then we have*

$$v(t,u) = -\frac{1}{2}A_t u \tag{3.26}$$

where $\frac{\partial M_t}{\partial t}$ is the $n \times n$ matrix of derivatives of each entry of M_t. Then v satisfies the Fokker Plank pde

$$\frac{\partial \log p(u,t)}{\partial t} = -\sum_{i=1}^{n}\left(\frac{\partial v_i(t,u)}{\partial u_i} + v_i(t,u)\frac{\partial \log p(u,t)}{\partial u_i}\right).$$

Proof. We have

$$\log p(u,t) = -\frac{1}{2}u^T M_t u - \log c_t$$

where $c_t = \int_{\mathbb{R}^p} \exp(-\frac{1}{2}u^T M_t u)$. Let us set $v(t,u) = A_t u$ for some matrix A_t and see check whether it solves the pde we want. Let us look at the left side of the pde. The left side of the pde turns out to be

$$-\frac{1}{2}u^T\frac{\partial M_t}{\partial t}u - \mathbb{E}_{p_t}(-\frac{1}{2}u^T\frac{\partial M_t}{\partial t}u).$$

We also have the space derivatives of log density to be

$$\frac{\partial \log p(u,t)}{\partial u_i} = -(M_t u)_i.$$

Hence the right side of the pde (3.5.1) becomes $-\sum_{i=1}^{n} A_{ti,i} + (u^T A_t)_i (M_t u)_i$ which equals $-Tr(A_t) - u^T A_t M_t u$. If the left side were to equal the right side

for all u then the quadratic forms have to match which implies $-\frac{1}{2}\frac{\partial M_t}{\partial t} = A_t M_t$. Hence we have to set

$$A_t = \frac{1}{2}\frac{\partial M_t}{\partial t}\Sigma_t. \tag{3.27}$$

Now by theory of quadratic forms of normal random variables, we have $\mathbb{E}_{p_t}(-\frac{1}{2}u^T\frac{\partial M_t}{\partial t}u) = Tr(\Sigma_t\frac{\partial M_t}{\partial t})$. Hence we see that (3.5.1) is indeed satisfied. □

The above proof shows something interesting. It tells us the right way to break up the left side in (3.5.1) if we want to recover the solution (3.26) by just doing one dimensional integrations. The above proof shows that the right way to try and solve for v is to solve the following for all $i = 1, \ldots, n$

$$u^T A_t[,i]M_t[i,]u - A_t[i,i] = \frac{\partial v_i(t,u)}{\partial u_i} + v_i(t,u)\frac{\partial \log p(u,t)}{\partial u_i}$$

where $A_t[,i]$ is the ith column of A_t and $M_t[i,]$ is the ith row of M_t. Solution to the above ode's coincide the solution (3.26). So by solving n one dimensional ode's which requires doing 1 dimensional integrations and estimating means of n functions by running several chains we can recover the linear velocity field we want (3.26). There is a price we have to pay however. We have to estimate the means of n functions at every step and each step requires very accurate estimation of the means otherwise the trajectories quickly become unstable. Hence we need to make sure we have enough chains for us to be able to estimate all the n means accurately.

This above way of sampling from a multivariate Gaussian motivates us to explore a similar idea in the variable augmentation formulation. The goal is to again obtain a well behaved velocity field as a solution to Fokker Plank by just doing one dimensional integrations. In the variable augmentation formulation

we want to sample from the density q_T as follows

$$q_t(u) \propto \exp\left(\frac{T}{n}\sum_{i=1}^{n} r_i\phi(u_i) - \frac{u^T M_t u}{2}\right). \tag{3.28}$$

The density q_t has a Gaussian term and the objective function term which is of a product form. In light of this, we propose a path or sequence of densities $\{q_t : 0 \leq t \leq 1\}$ defined the following way

$$q_t(u) \propto \exp\left(\frac{T}{n}\sum_{i=1}^{n} r_i\phi(u_i) - \frac{u^T M_t u}{2}\right) \tag{3.29}$$

where M_0 is a diagonal matrix and $M_1 = \dfrac{I - \dfrac{\sigma^2}{\sigma^2+\alpha^2}P_X}{\alpha^2}$ is the target positive definite symmetric matrix. This path q_t is different that what was defined in (3.29). There the multiplier in the exponent of the objective function term was growing from 0 to T. This path is a lot similar to the Gaussian path where only M_t varies smoothly with t. The difference here is that the extra factor $\exp(\frac{T}{n}\sum_{i=1}^{n} r_i\phi(u_i))$ sits throughout the path. It is essential to have M_0 a diagonal matrix because then q_0 is a product of one dimensional distributions and can be efficiently sampled. For solving the equations (3.5) in the Gaussian case we need to compute the space and time derivatives of the log densities. The space derivatives of $\log q_t$ are

$$\frac{\partial \log q(u,t)}{\partial u_i} = \frac{T}{n} r_i\phi(u_i) - (M_t u)_i.$$

This can be compared to the space derivatives of log density in the Gaussian case.

$$\frac{\partial \log p(u,t)}{\partial u_i} = -(M_t u)_i$$

Since $\frac{T}{n} r_i\phi(u_i)$ is very small when $n >> T = O(d\log d)$, the space derivatives

are almost the same as in the Gaussian case. Now we come to the time derivative of $\log q_t$ which is

$$\frac{\partial \log q(u,t)}{\partial t} = -\frac{1}{2}u^T \frac{\partial M_t}{\partial t}u - \mathbb{E}_{q_t}(-\frac{1}{2}u^T \frac{\partial M_t}{\partial t}u) \tag{3.30}$$

where $\mathbb{E}_{q_t}$ refers to expectation with respect to the density q_t. Although $\frac{\partial \log q(u,t)}{\partial t}$ looks exactly like $\frac{\partial \log q(u,t)}{\partial t}$, it differs from the Gaussian case in the expectation term because q_t is no longer a Gaussian density. Now in the Gaussian case, we broke up the time derivative of $\log p_t$ as follows

$$\frac{\partial \log p(u,t)}{\partial u_i} = \sum_{i=1}^{n} \left(u^T A_t[,i]M_t[i,]u - \mathbb{E}_{p_t} u^T A_t[,i]M_t[i,]u\right)$$

where A_t is defined as in (3.27). In the Gaussian case, $\mathbb{E}_{p_t} u^T A_t[,i]M_t[i,]u$ is known to be $A_t[i,i]$. Similarly, we can break up the time derivative of $\log q$ as follows

$$\frac{\partial \log q(u,t)}{\partial u_i} = \sum_{i=1}^{n} \left(u^T A_t[,i]M_t[i,]u - \mathbb{E}_{q_t} u^T A_t[,i]M_t[i,]u\right)$$

In this case $\mathbb{E}_{q_t} u^T A_t[,i]M_t[i,]$ is unknown and hence has to be estimated from multiple chains. So to summarize we propose solving the following bunch of differential equations

$$u^T A_t[,i]M_t[i,]u - \mathbb{E}_{q_t} u^T A_t[,i]M_t[i,]u = \frac{\partial v_i(t,u)}{\partial u_i} + v_i(t,u)\frac{T}{n} r_i \phi(u_i) - (M_t u)_i. \tag{3.31}$$

where at every timestep, we estimate the means of the quadratic forms $\mathbb{E}_t u^T A_t[,i]M_t[i,]u$ from multiple chains by taking the sample average. This gives us a velocity field which certainly satisfies the overall Fokker Plank and tries to mimic the way we solve for a well behaved velocity field in the Gaussian case by just

doing one dimensional integrations. The hope is that only a small fraction of the trajectories would be unstable and the remaining samples would be approximately valid for our target distribution.

3.6 Conclusion

We have proposed an optimization problem motivated by high dimensional function estimation. The problem is hard because the objective function may be highly multimodal. We have proposed a randomized algorithm to optimize such functions ala simulated annealing, only that the transition steps are motivated by the theory of diffusions and not Markov Chain theory. The transitions are made after solving a high dimensional pde, commonly known as the Fokker Plank equation. We have shown how to solve the pde in dimensions one, two and atmost three. In the special case of the objective function being the superposition of ridge functions we have also proposed a way of solving the required pde by just doing many one dimensional integrations. The hope is to encourage exploration solving multimodal sampling problems in high dimensions and to further understand the potential of solving such hard problems by our method of Adaptive Annealing.

3.7 Appendix

3.7.1 Proof of Lemma (3.3.1)

In the following we abuse notation by denoting any constants generically by C. Recall $f(\theta) = \frac{1}{n}\sum_{i=1}^{n} r_i \phi(x_i^T \theta)$ where ϕ is a smooth bounded function

taking values between ± 1 and also has bounded derivatives. We also have the sequence of densities defined by

$$p_t(\theta) = \exp(tf(\theta))exp(-\frac{1}{2}\theta^T M\theta)(\frac{1}{\sqrt{2\pi}})^d\sqrt{det(M)}$$

where $M = \frac{X^T X}{n}$ as can be seen from (**??**). The question is how large should t be so that $\frac{\mathbb{E}_t f(\theta)}{sup_\theta f(\theta)}$ is no less than a constant value C. Let $\theta^\star$ be any point in $\mathbb{R}^d$ maximizing f Since f is bounded in absolute value by 1, we note that

$$\frac{f(\theta^\star)+1}{f(\theta^\star)} \geq \frac{f(\theta^\star)-\mathbb{E}_t f(\theta)}{f(\theta^\star)}.$$

We want $\frac{f(\theta^\star)-\mathbb{E}_t f(\theta)}{f(\theta^\star)} \leq C$ where C is any constant strictly less than 1, say $\frac{1}{2}$. It can be checked that $\frac{\partial \log c_t}{\partial t} = \mathbb{E}_t f(\theta)$ and $\frac{\partial^2 \log c_t}{\partial t^2} = Var_t f(\theta) \geq 0$. So it follows that $\mathbb{E}_t f(\theta)$ is a non decreasing function of t. An application of mean value Theorem and the fact that $\log c_0 = 0$ now gives us

$$\frac{\log c_t}{t} \leq \mathbb{E}_t f(\theta). \tag{3.32}$$

The above result gives us a way to upper bound $f(\theta^\star) - \mathbb{E}_t f(\theta)$. Using (3.32) we have

$$f(\theta^\star)-\mathbb{E}_t f(\theta) \leq \frac{-1}{t}\log\int_{\mathbb{R}^d}\exp(t(f(\theta)-f(\theta^\star))exp(-\frac{1}{2}\theta^T M\theta)(\frac{1}{\sqrt{2\pi}})^d\sqrt{det(M)}d\theta.$$

We can further restrict the integral on the right side of the above equation in an ellipsoid around $\theta^\star$. This cannot decrease the right side because $\exp(t(f(\theta) - f(\theta^\star)) \leq 1$. Hence we have

$$f(\theta^\star)-\mathbb{E}_t f(\theta) \leq \frac{-1}{t}\log\int_{B_\epsilon}\exp(t(f(\theta)-f(\theta^\star))exp(-\frac{1}{2}\theta^T M\theta)(\frac{1}{\sqrt{2\pi}})^d\sqrt{det(M)}d\theta \tag{3.33}$$

where $B_\epsilon = \{\theta : (\theta-\theta^\star)^T M(\theta-\theta^\star) \leq \epsilon^2\}$. Now we want to say that $f(\theta)-f(\theta^\star)$ cannot be arbitrarily low in the set B_ϵ. By Taylor's expansion and the fact that the derivative of f at $\theta^\star$ is 0 we have

$$f(\theta) - f(\theta^\star) = (\theta - \theta^\star)^T H_f(\zeta)(\theta - \theta^\star)$$

where ζ is some point lying between θ and $\theta^\star$. The i, jth entry of the Hessian matrix H_f is given by

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\zeta) = \frac{1}{n}\sum_{l=1}^{n} r_l \phi''(x_l^T\theta) X_{li} X_{lj}.$$

So the maximum eigenvalue of H_f would be atmost the maximum eigenvalue of $\frac{1}{n}X^TX$ which would be atmost a constant C under most reasonable assumptions on the design matrix X. Then we have

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\zeta) \leq C(\theta - \theta^\star)^T(\theta - \theta^\star).$$

Multiplying by M and its inverse we have

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\zeta) \leq C(\theta - \theta^\star)^T M^{\frac{1}{2}} M^{-1} M^{\frac{1}{2}} (\theta - \theta^\star)$$

which further gives us

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\zeta) \leq C\lambda_{max}(M^{-1})(\theta - \theta^\star)^T M(\theta - \theta^\star)$$

. Hence we have for all $\theta \in B_\epsilon$ the following

$$f(\theta) - f(\theta^\star) \leq C\lambda_{max}(M^{-1})\epsilon^2.$$

The same argument could be repeated for $-(f(\theta) - f(\theta^\star))$ so that we can obtain for all $\theta \in B_\epsilon$ the following

$$f(\theta) - f(\theta^\star) \geq -C\lambda_{max}(M^{-1})\epsilon^2. \tag{3.34}$$

From (3.33) and (3.34) we now have

$$f(\theta^\star) - \mathbb{E}_t f(\theta) \leq C\lambda_{max}(M^{-1})\epsilon^2 - \frac{1}{t}\log \int_{B_\epsilon} exp(-\frac{1}{2}\theta^T M\theta)(\frac{1}{\sqrt{2\pi}})^d \sqrt{det(M)}d\theta$$

Now it can be checked that

$$\theta^T M\theta \leq 2(\theta - \theta^\star)^T M(\theta - \theta^\star) + 2\theta^{\star T} M\theta^\star. \tag{3.35}$$

Using the above inequality we then have

$$f(\theta^\star) - \mathbb{E}_t f(\theta) \leq C\lambda_{max}(M^{-1})\epsilon^2 - \frac{1}{t}\log \int_{B_\epsilon} exp(\epsilon^2 + \theta^{\star T} M\theta^\star)(\frac{1}{\sqrt{2\pi}})^d \sqrt{det(M)}d\theta$$

Pulling out the exponent and the constant terms out of the integral in the right side of the above equation we now have

$$f(\theta^\star) - \mathbb{E}_t f(\theta) \leq C\lambda_{max}(M^{-1})\epsilon^2 + \frac{\epsilon^2 + \theta^{\star T} M\theta^\star}{t} + \frac{d\log(2\pi)}{2t} - \frac{\log detM}{2t} - \frac{1}{t}\log \int_{B_\epsilon} d\theta.$$

The log volume of the ellipsoid B_ϵ is $\frac{\log detM}{2}$ plus log volume of the unit ball in $\mathbb{R}^d$ as can be seen from the change of variable theorem. Also it is well known that the log volume of the unit ball in $\mathbb{R}^d$ is $\frac{d}{2}\log\pi + d\log\epsilon - \log\Gamma(\frac{d}{2}+1)$ where Γ refers to the well known Gamma function satisfying the recurrence property $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. An important fact is $\Gamma(\frac{d}{2}+1)$ is

of the order $d \log d$. Hence combining the above facts we have the following

$$f(\theta^\star) - \mathbb{E}_t f(\theta) \leq C\lambda_{max}(M^{-1})\epsilon^2 + \frac{\epsilon^2 + \theta^{\star T} M \theta^\star}{t} + \frac{d \log(2\pi)}{2t} + \frac{1}{t}\log\Gamma(\frac{d}{2}+1) - \frac{d}{2t}\log\pi - \frac{d}{t}\log\epsilon.$$

Setting $\epsilon^2 = \frac{d}{2tC\lambda_{max}(M^{-1})+\frac{1}{t}}$ we have the final upper bound

$$f(\theta^\star) - \mathbb{E}_t f(\theta) \leq C\lambda_{max}(M^{-1})\frac{d}{2tC\lambda_{max}(M^{-1})+\frac{1}{t}} + \frac{\frac{d}{2tC\lambda_{max}(M^{-1})+\frac{1}{t}} + \theta^{\star T} M \theta^\star}{t} + \frac{d \log(2\pi)}{2t} + \frac{1}{t}\log\Gamma(\frac{d}{2}+1) - \frac{d}{2t}\log\pi - \frac{d}{t}\log\epsilon.$$

From the above we see that we need $t = O(d \log d)$ so that the difference $f(\theta^\star) - \mathbb{E}_t f(\theta)$ is upper bounded by a constant value. Hence the ratio $\frac{f(\theta^\star) - \mathbb{E}_t f(\theta)}{f(\theta^\star}$ is upper bounded by a constant C divided by $f(\theta^\star)$.

Bibliography

[1] A. Barron, A. Cohen, W. Dahmen, and R. DeVore, "Approximation and learning by greedy algorithms." Annals of Statistics, vol. 35, 2007.

[2] B. DasGupta, H. Siegelmann, and E. Sontag, "On the Intractability of Loading Neural Networks." Kluwer Academic Publishers, 1994, ch. X, pp. 357389.

[3] V. Vu, "On the infeasibility of training neural networks with small mean-squared error."
IEEE Trans. on Information Theory, vol. 44, no. 7, pp. 28921900, 1998.

[4] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Equations of state calculations by fast computing machines. Journal of Chemical Physics, vol. 21, no. 6, pp. 10871092, 1953.

[5] S. Kirkpatrick, C. D. G. Jr., and M. P. Vecchi, "Optimization by simulated annealing." Science, vol.220, no. 4598, pp. 671-680, 1983.

[6] J. Chang, "Stochastic processes."
http://pantheon.yale.edu/jtc5/251/stochasticprocesses.pdf.

[7] H. Risken, "The Fokker-Planck Equation: Methods of Solutions and Applications." Springer, 1996

[8] E. J. Candes, "New ties between computational harmonic analysis and approximation theory."
C.K. Chui, L.L. Schumaker, J. Stoeckler (Eds.), Approximation Theory X: Wavelets, Splin, Vanderbilt Univ. Press, Nashville, TN (2002), pp. 87153

[9] M. Leshno, V. Y. Lin, A. Pinkus, S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function." Neural networks 6 (6), 861-867

[10] A. R. Barron, X. Luo, "Adaptive Annealing." 45th Annual Allerton Conference, Allerton House, UIUC, Illinois, USA 2007.

[11] C. Villani, "Topics in Optimal Transportation." Graduate Studies in Mathematics 58, American Mathematical Society, Providence(2003).