Statistical and Computational Properties of Some "User-Friendly" Methods for High-Dimensional Estimation

Alnur Ali

May 2019
CMU-ML-19-105

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Sivaraman Balakrishnan
John C. Duchi (Stanford University)
J. Zico Kolter (Co-Chair)
Ameet Talwalkar
Ryan J. Tibshirani (Co-Chair)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was sponsored by the Department of Energy Computational Science Graduate Fellowship no. DE-FG02-97ER25308.

For Nayab

Abstract

Historically, the choice of method for a given statistical problem has been primarily driven by two criteria: a method's statistical properties, and its computational properties. But as tools for high-dimensional estimation become ubiquitous, it is clear that there are other considerations, going beyond pure accuracy and computational efficiency, that are equally (if not more) important. One such consideration is a method's "user-friendliness"— a term we use to encapsulate the various properties that make a method easy to work with in practice, exemplified by a method being (i) easy-to-implement, (ii) interpretable, and (iii) computationally cheap. In this thesis, we present new statistical and computational results for three different user-friendly methods in various high-dimensional estimation settings.

First, we give conditions for the existence and uniqueness of solutions to the *generalized lasso* problem, which is a generalization of the standard lasso problem that allows the user to easily impose domain-appropriate structure onto the fitted coefficients. The conditions are very weak, and essentially guarantee uniqueness in many settings of practical interest, even in high dimensions, which are useful results from the points-of-view of interpretability as well as prediction. Second, we consider early-stopped gradient descent (as an estimator), giving a number of results that tightly couple the risk profile of the iterates generated by gradient descent, when run on the fundamental problem of least squares regression, to that of ridge regression — these results are favorable for gradient descent, as it is relatively easy-to-implement as well as computationally cheap. We also discuss extending the analysis to give a similar coupling for (the arguably even more user-friendly) stochastic gradient descent. Finally, we present a new user-friendly, *pseudolikelihood*-based method for robust undirected graphical modeling that we call the Multiple Quantile Graphical Model (MQGM), showing that the MQGM recovers the population-level conditional independencies, with high probability — this is again a useful result, from an interpretability standpoint. We also give a highly efficient algorithm, based on the alternating direction method of multipliers, for fitting the MQGM to high-dimensional and potentially non-Gaussian data.

Acknowledgments

I have to begin by confessing something a little bit silly: before starting graduate school, I would occasionally leaf through other people's dissertations, which invariably contained an acknowledgements section (similar to this one) ... and I didn't really get why. Back then, the acknowledgements section seemed rather perfunctory to me. Were all those acknowledgements really all that necessary? But now that I am at the end of my Ph.D. journey, I get it. Doing a Ph.D. is indeed quite a journey, and no one gets by without a little help from their friends, advisors, etc.

There are a lot of people who I'd like to thank for helping me get here. But I'd like to start by thanking my advisors, Ryan Tibshirani and Zico Kolter. Zico took me on as a student, at the very beginning of my Ph.D. I was (and still am) impressed by how fearlessly he tackles many problems, and with the breadth of his understanding of machine learning and optimization. About halfway through my Ph.D., I decided I wanted to do more statistical theory, and I was (and still feel) quite lucky that Ryan agreed to co-advise me. In addition to being a fellow Canadian (instant points for him), Ryan has amazing mathematical instincts, as well as an admirable way of always pushing for simplicity (in proofs, writing, and other things). I hope these are qualities that I can emulate in my career.

I'd also like to thank my other committee members: Siva Balakrishnan, John Duchi, and Ameet Talwalkar. Everyone has been extremely gracious to me — not only in providing thoughtful feedback on my work, but also in taking the extra time out to give me helpful career advice (and reminding me to stay focused on problems that matter). For all of it, I am grateful.

Thanks also to my co-authors — Stephen Boyd, Steven Diamond, Kshitij Khare, Penporn Koanantakool, Sang-Yun Oh, Bala Rajaratnam, and Eric Wong — for making our projects together extra fun, and teaching me a lot about high-dimensional statistics and large-scale optimization. And thanks to Edgar Dobriban, Veeru Sadhanala, and Pratik Patil for fun and insightful discussions about new, ongoing projects.

Carnegie Mellon is an amazing place, in large part because of my fellow students. To Brandon Amos, Shaojie Bai, Filipe de Avila Belbute-Peres, Dallas Card, Jeremy Cohen, Christoph Dann, Jonathan Dinu, Priya Donti, Justin Hyun, Chun Kai Ling, Gaurav Manek, Vaishnavh Nagarajan, Pratik Patil, Leslie Rice, Mel Roderick, Veeru Sadhanala, Mariya Toneva, Yu-Xiang Wang, Po-Wei Wang, Josh Williams, Ezra Winston, Eric Wong, Matt Wytock, Jing Xiang, Yisong Yue, Manzil Zaheer, and Han Zhao: thank you all, for the fun discussions (even the ones about deep learning!), and for always keeping things entertaining. To anyone I (accidentally) forgot: thank you, too. Also, a big thanks goes to Diane Stidle for keeping the trains running on time, and always being a cheery face to see around Gates.

I'd like to reach back in time a little, and thank all of the mentors and collaborators I had at Microsoft and the University of Washington (many of whom I am now happy to bump into often at conferences), who always encouraged me to pursue graduate school. To Ahmad Abdulkader, Bodo von Billerbeck, Chris Burges, Rich Caruana, Max Chickering, Kevyn Collins-Thompson, Nick Craswell, Jianfeng Gao, Ashish Kapoor, Aparna Lakshmiratan, Chris Meek, and Marina Meilă: thank you all, for that push. It meant a lot. Going back even further, my parents deserve a huge amount of thanks for always telling me to never shy away from getting more education, and for always trying to put me in the best possible environment to succeed.

Lastly, but certainly not least, thanks to Nayab. You selflessly picked up from our home in Seattle, and moved across the country just to support me doing a Ph.D. There may have been a little homesickness and kicking (and screaming) here and there, but I really cannot believe how amazing you have been. I love you.

Contents

Chapter 1

Introduction

1.1 What is "User-Friendliness"?

In traditional applications of statistics and machine learning, the choice of method is usually driven by weighing two criteria: a method's statistical properties, and its computational properties. On the other hand, the recent explosion of interest in machine learning has shaped modern statistical practice in ways that challenge this traditional viewpoint. One prominent trend has been that of non-specialists increasingly being asked to deploy machine learning systems into the "wild"; as a consequence, these days, there is a growing preference for methods having properties that appear to lie along a "third axis", i.e., for methods that are (i) easy-to-implement, (ii) interpretable, and (iii) computationally cheap. In the rest of this thesis document, we will somewhat informally refer to a method possessing these properties as one that is "user-friendly".[1] Given the relevance of user-friendly methods to the modern practice of statistics and machine learning, it seems important to zero in on these user-friendly estimators in particular, and study their statistical and computational properties; characterizing the various properties of user-friendly methods can be helpful, as it gives practitioners (as well as statisticians) a more complete picture of the pros and cons of various methods. In this thesis, we present new statistical and computational results for three different user-friendly methods for statistical estimation. Before giving an overview of these results, we work through and discuss several examples of user-friendly estimators, in order to help make clear the concept for the reader.

1.2 Diving Deeper into User-Friendliness

1.2.1 A Semi-Synthetic Data Example

To clarify the notion of user-friendliness, it may help to see a concrete example. In what follows, we outline a semi-synthetic data example, describing an approach to estimation that is user-friendly, as well as several approaches that are evidently much less user-friendly. Here, we consider predicting a response variable, given functional magnetic resonance imaging (fMRI) data

[1] Giving a formal definition for user-friendliness seems to be a valuable and worthwhile pursuit; however, we defer doing this in a thorough and comprehensive way, to future work.

coming from several children diagnosed with attention deficient hyperactive disorder (ADHD) [10]; we elaborate on the construction of the repsonse in just a moment, but for now we focus on describing the predictors. Our specific data matrix $X \in \mathbb{R}^{n \times p}$ contains $p = 39$ columns, each one corresponding to a different physical location on the brain. The number of rows in X is given by $n = 703$, and this arises from the number of different subjects, as well as the time resolution, at which the measurements were taken.

There is evidence in the neuroscience literature that some physiological processes are reasonably modeled by a few spatially coherent (nearby) regions of the brain, having comparable effect sizes [8, 9, 30, 80, 147]. Therefore, thinking of the response $y \in \mathbb{R}^n$ as modeling attention span, it makes sense to express

$$y = a_1 \cdot (X_2 + X_{10} + X_{37}) + a_2 \cdot (X_3 + X_5 + X_{33}) + a_3 \cdot (X_1 + X_{16} + X_{35}) + \epsilon,$$

where $a_1 = a_2 = a_3 = 10$ are just fixed constants, and $\epsilon \in \mathbb{R}^n$ is standard Gaussian noise. Above, the groups of columns of X (e.g., $\{X_2,\ X_{10},\ X_{37}\}$) in the linear combination that drives y, were found to be spatially nearby after inspecting the data, reflecting the aforementioned principle of spatial coherence. Put slightly differently, in the above construction, we view the underlying signal $\beta_0 = (a_3, a_1, a_2, 0, 0, \ldots) \in \mathbb{R}^p$ as locally constant with respect to some underlying graph, where the vertices of the graph are put into a one-to-one correspondence with the dimensions. Figure 1.1 illustrates the idea, where we see that nearby coefficients (equivalently, vertices of the graph) are encouraged to take on similar values.

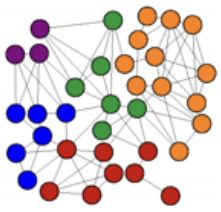

Figure 1.1: *Illustration of the graph associated with the fMRI data example. Here, nearby vertices in the graph (equivalently, coefficients) are modeled as taking similar values, emphasized by the different colorings.*

1.2.2 Potential Approaches

A first approach to estimation in the above example might be as follows. By analogy to some approaches for image deblurring [13, 16, 83], where denoising is done by averaging over patches constructed across an image, we could consider doing a lasso regression on an alternative design matrix, one that is constructed by forming averages of the columns in X corresponding to nearby dimensions (patches). This sort of approach might be reasonable when the underlying signal is suspected to be piecewise constant — but if it is not, then it is not at all clear that this approach is a suitable one, and remedies are not immediately apparent. However, even in the case of a piecewise constant signal, this approach requires that the user specify the size of the patches

(essentially introducing another tuning parameter); additionally, if many patches are employed, then this will in general drive up the overall computational cost.

Another more modern approach, emphasizing prediction, is to use a graph neural network [24, 34, 54, 81]. These methods are powerful, but apparently not so user-friendly — many of the usual problems with standard neural networks arise here, too: tuning the step size, optimization algorithm, network architecture, etc. can all be laborious, and training the network itself may be time-consuming. Arguably, ease-of-implementation is less of an issue these days, with the rise of widely available software packages for training deep neural networks; but if some customizations are required, then the implementation may pose its own set of challenges, as well. Along these lines, we point out that the graph neural network approach also suffers from the same extensibility issues raised with the first approach described above.

1.2.3 A User-Friendly Approach: The Generalized Lasso

By contrast, consider the *generalized lasso* estimator [136]: the generalized lasso is a generalization of the standard lasso, where the defining optimization problem now involves the usual squared error loss, plus a penalty given by the ℓ_1 norm composed with a linear transformation $D\beta$ of the coefficients (c.f. the pure ℓ_1 norm penalty, as in the standard lasso problem). The linear transformation $D \in \mathbb{R}^{m \times p}$ should be interpreted as a kind of penalty matrix, encoding essentially whatever sort of problem-specific structure the user has in mind. More concretely, the generalized lasso estimator is defined as a solution to the optimization problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \|D\beta\|_1, \tag{1.1}$$

where $y \in \mathbb{R}^n$ is the response, $X \in \mathbb{R}^{n \times p}$ is the data matrix, $D \in \mathbb{R}^{m \times p}$ is the penalty matrix, and $\lambda \geq 0$ is a tuning parameter. Much more will be said about the generalized lasso below, but for now we call out the *fused lasso* [134] as a simple and well-known special case, to convey the generality and user-friendliness of the broader generalized lasso framework. The fused lasso, employed when the underlying signal is suspected to be piecewise constant, is easily recovered as a special case from the generalized lasso framework, by simply specifying the appropriate penalty matrix. In this case, we set $X = I$ and

$$D = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(p-1) \times p},$$

which can be seen as the first-order discrete derivative operator, meaning that we expect the differences $D\beta$ to be sparse. (In general, X may be arbitrary, significantly expanding the scope of potential applications.) It is also straightforward to additionally encourage pure sparsity in the coefficients, by concatenating D together with the identity matrix. Finally, it is also worth repeating that alternative structural assumptions about a given problem may be enforced, by simply swapping out the current penalty matrix for another one.

The generalized lasso has been previously used as a modeling tool in many tasks [1, 67, 89, 133, 147], but we return now to the idea of using it in the context of the fMRI data example described above. To use the generalized lasso, we must specify a penalty matrix. There are several possible options, but we pursue here a relatively simple and natural one, known as the *fused lasso on the k-nearest neighbors (k-NN) graph* [136]. The construction is as follows: each row of D is formed by putting a -1 and a 1 in the column positions corresponding to the vertices that are adjacent in the k-NN graph (for some value of k) built from the physical locations of the dimensions. The same principle is at work here as with the usual fused lasso, i.e., now the fitted generalized lasso coefficients will generally obey a piecewise constant structure across many of the edges of the k-NN graph. Additionally, in what follows, we encourage pure sparsity in the coefficients in the same way that we mentioned before, i.e., by concatenating the identity matrix, and the penalty matrix that was just described.

It is worth pointing out that the generalized lasso is apparently much more user-friendly than either of the approaches described earlier. To be more specific, if we suspected the underlying signal to be, say, piecewise linear (instead of piecewise constant), then all we would need to do is change the penalty matrix to the analog of the second-order discrete derivative operator over a graph; see Wang et al. [142] for details. As mentioned previously, it is not clear how to (easily) modify either of the previous proposals to account for this sort of structural assumption.

1.2.4 Empirical Results

Having specified y, X, and D, we are now in a position to solve the generalized lasso problem. As the problem (1.1) is convex, an application of the alternating direction method of multipliers [15] admits an efficient and reasonably straightforward implementation; see Section 6.4.1 in Boyd et al. [15], for details. It is also easy to solve the generalized lasso problem with any number of the freely-available software packages for generic convex optimization (e.g., [46]).

The left panel of Figure 1.2, presents test error curves for both the generalized lasso and the standard lasso, plotted as a function of their (common) tuning parameter λ. In a little more detail, the curves were generated by solving the generalized and standard lasso problems over 600 out of 703 total samples (using the CVX software package [46]), then evaluating the fitted models on the remaining samples, for each value of λ. As can be seen from the plot, the generalized lasso test error is better than that of the standard lasso, across many values of λ — indicating that the presence of D is helpful. (To be clear, it is likely that the test error curve for the generalized lasso would have been even lower, had we allowed for different tuning parameter values governing sparsity in the differences $D\beta$ vs. sparsity in the coefficients β; we did not pursue this here, for simplicity.)

Inspecting the fitted coefficients is also instructive. Below, we list the nonzero fitted coefficients, grouping them together according to the underlying columns involved in constructing y (as described earlier), for the fitted generalized lasso model obtaining the best test error (i.e.,

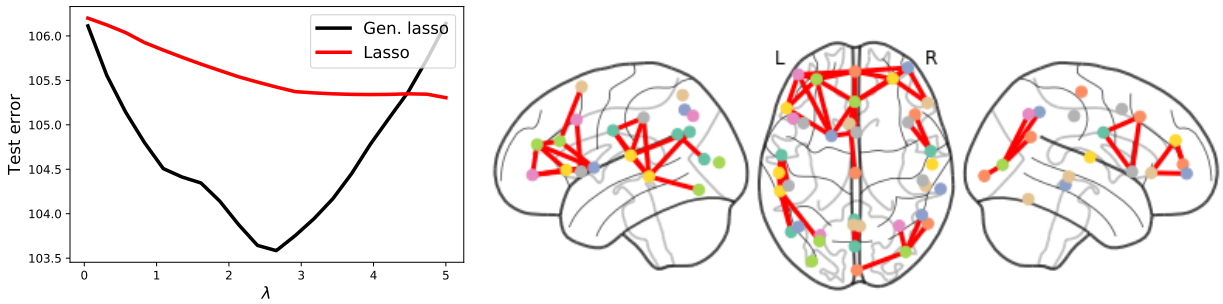

Figure 1.2: *The left panel shows test error curves for both the generalized lasso as well as the standard lasso, in the fMRI data example, obtained by varying the tuning parameter λ. The right panel shows a simplified k-nearest neighbors graph, produced by the generalized lasso coefficients yielding the best test error.*

stemming from the tuning parameter $\lambda = 2.6552$),

$$
\begin{aligned}
&\hat{\beta}_1 = 0.02, \quad \hat{\beta}_{16} = \hat{\beta}_{35} = 6.53,\\
&\hat{\beta}_2 = \hat{\beta}_{10} = \hat{\beta}_{37} = 5.10,\\
&\hat{\beta}_3 = \hat{\beta}_5 = \hat{\beta}_{33} = 4.74,\\
&\hat{\beta}_6 = -0.01,\\
&\hat{\beta}_{11} = -0.01,\\
&\hat{\beta}_{14} = -0.01,\\
&\hat{\beta}_{17} = -0.01,\\
&\hat{\beta}_{22} = -0.05,\\
&\hat{\beta}_{23} = -0.01,\\
&\hat{\beta}_{24} = -0.01,\\
&\hat{\beta}_{25} = -0.01,\\
&\hat{\beta}_{26} = 0.01,\\
&\hat{\beta}_{28} = 0.01,\\
&\hat{\beta}_{29} = -0.01,\\
&\hat{\beta}_{30} = -0.01,\\
&\hat{\beta}_{32} = 0.04.
\end{aligned}
$$

The above results reveal that the generalized lasso has recovered much of the underlying structure in the problem, which can be helpful from the point-of-view of interpretability (and no such structure was found in the fitted standard lasso models).

Continuing with this line of thought, another useful output of the generalized lasso can be seen as follows. Starting with the k-NN graph used to construct D, we may remove the edges connecting unfused vertices (corresponding to fitted coefficients that did not take on the same

value) in some solution $\hat{\beta}$. The result, presented in the right panel of Figure 1.2 (again, using the solution attaining the best test error), is a pruned and somewhat simplified graph that can be useful to practitioners.

1.2.5 Discussion

The takeaway message from the preceding discussion should be that the generalized lasso offers an appealing and user-friendly workflow for applied statistical modeling. Certainly, depending on the circumstances, there may be alternatives that are entirely appropriate (in the context of fMRI data, approaches to sparse inverse covariance estimation, as well as the group lasso, both come to mind, as in, e.g., Guo et al. [49], Hsieh et al. [60], Koanantakool et al. [70]). But the final message here is that the pros and cons of these alternative methods should be carefully weighed against the user-friendliness, statistical properties, and computational properties of the generalized lasso.

1.3 Another User-Friendly Approach: Gradient Descent

As a further illustration, we give a second example of a user-friendly estimator: the iterates generated by (early-stopping) gradient descent. It may be somewhat unusual to think of gradient descent as an estimator, as it is often viewed as an optimization algorithm. However, there has actually been a steady stream of both empirical and theoretical work over the years, suggesting a relationship between early-stopped gradient descent when applied to least squares regression, and ridge regression (see, e.g., Friedman and Popescu [40], Raskutti et al. [111], Yao et al. [150]). To be only slightly more specific for now (with many more details to follow, in subsequent chapters of this thesis), prior works have indicated a connection between the iterate $\beta^{(k)}$, obtained after running gradient descent for some number of iterations $k \geq 0$, and the ridge regression estimate $\hat{\beta}^{\text{ridge}}(\lambda) = (X^T X + \lambda I)^{-1} X^T y$, for a corresponding value of the ridge tuning parameter $\lambda \geq 0$ — in other words, we can think of each iterate of gradient descent as possessing a kind of implicit ℓ_2 regularity.

The reader may ask why we would ever pursue an (indirect) iterative solution, when the ridge solution can be computed in closed-form? The main reasons are arguably related to computation. In a high-dimensional setup, the ridge solution may be computed exactly in $O(n^2 p)$ time. The basic gradient descent iteration, by contrast, consists of simple matrix-vector products and is comparatively easy-to-implement as well as computationally cheap (i.e., user-friendly), costing $O(np)$ per iteration. It is worth mentioning that in large-scale applications, it can make sense to pursue an indirect method for computing the ridge solution $\hat{\beta}^{\text{ridge}}(\lambda)$ itself, with gradient descent and conjugate gradients being two natural approaches (i.e., for a fixed value of λ). Having said that, and keeping in mind the connection to implicit ℓ_2 regularization alluded to above, we may view running gradient descent for several iterations (on the least squares loss) as actually generating a suite of estimates, each one possessing a different level of regularization (c.f. the point estimate given by the ridge solution $\hat{\beta}^{\text{ridge}}(\lambda)$). Moreover, moving beyond the squared error loss and thinking of generalized linear models more broadly, the aforementioned computational

benefits belonging to a simple method such as gradient descent can be significant (consider, e.g., the case of logistic regression).

1.4 Outline

In this thesis document, we present statistical and computational results for three different user-friendly methods for high-dimensional estimation. To start with, in the second chapter of this thesis, we give new statistical results for the generalized lasso. An old question important to the modern practice of statistics, both from the standpoints of interpretability and prediction, is one related to identifiability: when can we say that the solution to a penalized loss minimization problem is unique? It turns out that answering this question for the generalized lasso problem is rather complicated, due to the presence of the (otherwise simple-looking) penalty matrix. Nevertheless, in Chapter 2, we present sufficient conditions characterizing both the existence and uniqueness of solutions to the generalized lasso problem. The conditions essentially guarantee uniqueness in many situations of practical interest, and are much weaker than those, e.g., required to establish estimation error rates and support recovery in sparse regression problems [79, 109, 141].

In the third chapter of this thesis, we consider early-stopped gradient descent, when applied to the fundamental problem of least squares regression. Whereas most prior works generally describe a somewhat coarse relationship between the two methods, we give a number of results that much more tightly couple the risk profile of the iterates generated by gradient descent to that of ridge regression.

Finally, in Chapter 4 of this thesis, we present a new user-friendly, *pseudolikelihood*-based method [12] for robust undirected graphical modeling, called the Multiple Quantile Graphical Model (MQGM). In this chapter, we also give statistical theory showing that the MQGM recovers the population-level conditional independence relationships with high probability. Finally, we present an algorithm for fitting the MQGM to high-dimensional heavy-tailed data, which is often an order of magnitude faster than alternatives.

In Chapter 5, we conclude with a brief discussion, but in doing so, we return to some of the points raised in Chapter 3, and outline how to extend the same ideas to characterize the risk profile of (the arguably even more user-friendly) stochastic gradient descent.

1.5 Notation

The notation we use in this thesis document is mostly standard. For a matrix $A \in \mathbb{R}^{m \times n}$, we write A^+ for its Moore-Penrose pseudoinverse and $\mathrm{col}(A), \mathrm{row}(A), \mathrm{null}(A), \mathrm{rank}(A)$ for its column space, row space, null space, and rank, respectively. We write A_J for the submatrix defined by the rows of A indexed by a subset $J \subseteq \{1, \ldots, m\}$, and use A_{-J} as shorthand for $A_{\{1,\ldots,m\} \setminus J}$. Similarly, for a vector $x \in \mathbb{R}^m$, we write x_J for the subvector defined by the components of x indexed by J, and use x_{-J} as shorthand for $x_{\{1,\ldots,m\} \setminus J}$.

For a set $S \subseteq \mathbb{R}^n$, we write $\mathrm{span}(S)$ for its linear span, and write $\mathrm{aff}(S)$ for its affine span. For a subspace $L \subseteq \mathbb{R}^n$, we write P_L for the (Euclidean) projection operator onto L, and write $P_{L^\perp}$

for the projection operator onto the orthogonal complement $L^{\perp}$. For a function $f : \mathbb{R}^m \to \mathbb{R}^n$, we write $\mathrm{dom}(f)$ for its domain, and $\mathrm{ran}(f)$ for its range.

Chapter 2

The Generalized Lasso

2.1 Introduction

In this chapter, we consider the *generalized lasso* problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \|D\beta\|_1, \tag{2.1}$$

where $y \in \mathbb{R}^n$ is a response vector, $X \in \mathbb{R}^{n \times p}$ is a predictor matrix, $D \in \mathbb{R}^{m \times p}$ is a penalty matrix, and $\lambda \geq 0$ is a tuning parameter. As explained in Tibshirani and Taylor [136], the generalized lasso problem (2.1) encompasses several well-studied problems as special cases, corresponding to different choices of D, e.g., the lasso [132], the fused lasso [119, 134], trend filtering [69, 127], the graph fused lasso [55], graph trend filtering [142], Kronecker trend filtering [121], among others. (For all problems except the lasso problem, the literature is mainly focused on the so-called "signal approximator" case, where $X = I$, and the responses have a certain underlying structure; but the "regression" case, where X is arbitrary, naturally arises whenever the predictor variables—rather than the responses—have an analogous structure.)

There has been an abundance of theoretical and computational work on the generalized lasso and its special cases. In the current chapter, we examine sufficient conditions under which the solution in (2.1) will be unique. While this is simple enough to state, it is a problem of fundamental importance. The generalized lasso has been used as a modeling tool in numerous application areas, such as copy number variation analysis [133], sMRI image classification [147], evolutionary shift detection on phylogenetic trees, [67], motion-capture tracking [89], and longitudinal prediction of disease progression [1]. In such applications, the structure of the solution $\hat{\beta}$ in hand (found by using one of many optimization methods applicable to (2.1), a convex quadratic program) usually carries meaning—this is because D has been carefully chosen so that sparsity in $D\hat{\beta}$ translates into some interesting and domain-appropriate structure for $\hat{\beta}$. Of course, nonuniqueness of the solution in (2.1) would cause complications in interpreting this structure. (The practitioner would be left wondering: are there other solutions providing compementary, or even contradictory structures?) Further, beyond interpretation, nonuniqueness of the generalized lasso solution would clearly cause complications if we are seeking to use this solution to make predictions (via $x^T\hat{\beta}$, for a new predictor vector $x \in \mathbb{R}^p$), as different solutions would lead to different predictions (potentially very different ones).

When $p \leq n$ and $\mathrm{rank}(X) = p$, there is always a unique solution in (2.1) due to strict convexity of the squared loss term. Our focus will thus be in deriving sufficient conditions for uniqueness in the high-dimensional case, where $\mathrm{rank}(X) < p$. It also worth noting that when $\mathrm{null}(X) \cap \mathrm{null}(D) \neq \{0\}$ problem (2.1) cannot have a unique solution. (If $\eta \neq 0$ lies in this intersection, and $\hat{\beta}$ is a solution in (2.1), then so will be $\hat{\beta} + \eta$.) Therefore, at the very least, any sufficient condition for uniqueness in (2.1) must include (or imply) the null space condition $\mathrm{null}(X) \cap \mathrm{null}(D) = \{0\}$.

In the lasso problem, defined by taking $D = I$ in (2.1), several authors have studied conditions for uniqueness, notably Tibshirani [135], who showed that when the entries of X are drawn from an arbitrary continuous distribution, the lasso solution is unique almost surely. One of the main results in this chapter yields this lasso result as a special case; see Theorem 2.1, and Remark 2.5 following the theorem. Moreover, our study of uniqueness leads us to develop intermediate properties of generalized lasso solutions that may be of interest in their own right—in particular, when we broaden our focus to a version of (2.1) in which the squared loss is replaced by a general loss function, we derive local stability properties of solutions that have potential applications beyond this work.

In the remainder of this introduction, we describe the implications of our uniqueness results for various special cases of the generalized lasso, discuss related work, and then cover notation and an outline of the rest of the chapter.

2.1.1 Uniqueness in Special Cases

The following is an application of Theorem 2.1 to various special cases for the penalty matrix D. The takeaway is that, for continuously distributed predictors and responses, uniqueness can be ensured almost surely in various interesting cases of the generalized lasso, provided that n is not "too small", meaning that the sample size n is at least the nullity (dimension of the null space) of D. (Some of the cases presented in the corollary can be folded into others, but we list them anyway for clarity.)

Corollary 2.1. *Fix any $\lambda > 0$. Assume the joint distribution of (X, y) is absolutely continuous with respect to $(np + n)$-dimensional Lebesgue measure. Then problem* (2.1) *admits a unique solution almost surely, in any one of the following cases:*

(i) $D = I \in \mathbb{R}^{p \times p}$ *is the identity matrix;*

(ii) $D \in \mathbb{R}^{(p-1) \times p}$ *is the first difference matrix, i.e., fused lasso penalty matrix (see Section 2.1.1 in Tibshirani and Taylor [136]);*

(iii) $D \in \mathbb{R}^{(p-k-1) \times p}$ *is the $(k+1)$st order difference matrix, i.e., kth order trend filtering penalty matrix (see Section 2.1.2 in Tibshirani and Taylor [136]), and $n \geq k + 1$;*

(iv) $D \in \mathbb{R}^{m \times p}$ *is the graph fused lasso penalty matrix, defined over a graph with m edges, n nodes, and r connected components (see Section 2.1.1 in Tibshirani and Taylor [136]), and $n \geq r$;*

(v) $D \in \mathbb{R}^{m \times p}$ *is the kth order graph trend filtering penalty matrix, defined over a graph with m edges, n nodes, and r connected components (see Wang et al. [142]), and $n \geq r$;*

(vi) $D \in \mathbb{R}^{(N-k-1)N^{d-1}d \times N^d}$ *is the kth order Kronecker trend filtering penalty matrix, defined over a d-dimensional grid graph with all equal side lengths $N = n^{1/d}$ (see Sadhanala et al.*

[121]), and $n \geq (k+1)^d$.

Two interesting special cases of the generalized lasso that fall outside the scope of our results here are *additive trend filtering* [120] and *varying-coefficient models* (which can be cast in a generalized lasso form, see Section 2.2 of Tibshirani and Taylor [136]). In either of these problems, the predictor matrix X has random elements but obeys a particular structure, thus it is not reasonable to assume that its entries overall follow a continuous distribution, so Theorem 2.1 cannot be immediately applied. Still, we believe that under weak conditions either problem should have a unique solution. Sadhanala and Tibshirani [120] give a uniqueness result for additive trend filtering by reducing this problem to lasso form; but, keeping this problem in generalized lasso form and carefully investigating an application of Lemma 2.6 (the deterministic result in this chapter leading to Theorem 2.1) may yield a result with simpler sufficient conditions. This is left to future work.

Furthermore, by applying Theorem 2.2 to various special cases for D, analogous results hold (for all cases in Corollary 2.1) when the squared loss is replaced by a generalized linear model (GLM) loss G as in (2.19). In this setting, the assumption that (X, y) is jointly absolutely continuous is replaced by the two assumptions that X is absolutely continous, and $y \notin \mathcal{N}$, where $\mathcal{N}$ is the set defined in (2.41). The set $\mathcal{N}$ has Lebesgue measure zero for some common choices of loss G (see Remark 2.12); but unless we somewhat artificially assume that the distribution of $y|X$ is continuous (this is artificial because in the two most fundamental GLMs outside of the Gaussian model, namely the Bernoulli and Poisson models, the entries of $y|X$ are discrete), the fact that $\mathcal{N}$ has Lebesgue measure zero set does not directly imply that the condition $y \notin \mathcal{N}$ holds almost surely. Still, it seems that $y \notin \mathcal{N}$ should be "likely"—and hence, uniqueness should be "likely"—in a typical GLM setup, and making this precise is left to future work.

2.1.2 Related Work

Several authors have examined uniqueness of solutions in statistical optimization problems en route to proving risk or recovery properties of these solutions; see Donoho [31], Dossal [32] for examples of this in the noiseless lasso problem (and the analogous noiseless ℓ_0 penalized problem); see Nam et al. [97] for an example in the noiseless generalized lasso problem; see Candes and Plan [19], Fuchs [44], Wainwright [141] for examples in the lasso problem; and lastly, see Lee et al. [79] for an example in the generalized lasso problem. These results have a different aim than ours, i.e., their main goal—a risk or recovery guarantee—is more ambitious than certifying uniqueness alone, and thus the conditions they require are more stringent. Our work in this chapter is more along the lines of direct uniqueness analysis in the lasso, as was carried out by Osborne et al. [103], Rosset et al. [117], Schneider and Ewald [122], Tibshirani [135].

2.1.3 Outline

Here is an outline for what follows. In Section 2.2, we review important preliminary facts about the generalized lasso. In Section 2.3, we derive sufficient conditions for uniqueness in (2.1), culminating in Theorem 2.1, our main result on uniqueness in the squared loss case. In Section 2.4, we consider a generalization of problem (2.1) where the squared loss is replaced by a smooth

and strictly convex function of $X\beta$; we derive analogs of the important preliminary facts used in the squared loss case, notably, we generalize a result on the local stability of generalized lasso solutions due to Tibshirani and Taylor [137]; and we give sufficient conditions for uniqueness, culminating in Theorem 2.2, our main result in the general loss case. In Section 2.5, we conclude with a brief discussion.

2.2 Preliminaries

2.2.1 Basic Facts, KKT Conditions, and the Dual

First, we establish some basic properties of the generalized lasso problem (2.1) relating to uniqueness.

Lemma 2.1. *For any y, X, D, and $\lambda \geq 0$, the following holds of the generalized lasso problem (2.1).*

(i) There is either a unique solution, or uncountably many solutions.

(ii) Every solution $\hat{\beta}$ gives rise to the same fitted value $X\hat{\beta}$.

(iii) If $\lambda > 0$, then every solution $\hat{\beta}$ gives rise to the same penalty value $\|D\hat{\beta}\|_1$.

Proof. The criterion function in the generalized lasso problem (2.1) is convex and proper, as well as closed (being continuous on $\mathbb{R}^p$). As both $g(\beta) = \|y - X\beta\|_2^2$ and $h(\beta) = \lambda\|D\beta\|_1$ are nonnegative, any directions of recession of the criterion $f = g + h$ are necessarily directions of recession of both g and h. Hence, we see that all directions of recession of the criterion f must lie in the common null space $\mathrm{null}(X) \cap \mathrm{null}(D)$; but these are directions in which the criterion is constant. Applying, e.g., Theorem 27.1 in Rockafellar [115] tells us that the criterion attains its infimum, so there is at least one solution in problem (2.1). Supposing there are two solutions $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}$, since the solution set to a convex optimization problem is itself a convex set, we get that $t\hat{\beta}^{(1)} + (1-t)\hat{\beta}^{(2)}$ is also a solution, for any $t \in [0,1]$. Thus if there is more than one solution, then there are uncountably many solutions. This proves part (i).

As for part (ii), let $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}$ be two solutions in (2.1), with $\hat{\beta}^{(1)} \neq \hat{\beta}^{(2)}$. Let $f^\star$ denote the optimal criterion value in (2.1). Proceeding by contradiction, suppose that these two solutions do not yield the same fit, i.e., $X\hat{\beta}^{(1)} \neq X\hat{\beta}^{(2)}$. Then for any $t \in (0,1)$, the criterion at $t\hat{\beta}^{(1)} + (1-t)\hat{\beta}^{(2)}$ is

$$
\begin{aligned}
&f\big(t\hat{\beta}^{(1)} + (1-t)\hat{\beta}^{(2)}\big) \\
&= \frac{1}{2}\big\|y - \big(tX\hat{\beta}^{(1)} + (1-t)X\hat{\beta}^{(2)}\big)\big\|_2^2 + \lambda\big\|D\big(t\hat{\beta}^{(1)} + (1-t)\hat{\beta}^{(2)}\big)\big\|_1 \\
&< t\frac{1}{2}\|y - X\hat{\beta}^{(1)}\|_2^2 + (1-t)\frac{1}{2}\|y - X\hat{\beta}^{(2)}\|_2^2 + \lambda t\|D\hat{\beta}^{(1)}\|_1 + (1-t)\lambda\|D\hat{\beta}^{(2)}\|_1 \\
&= tf(\hat{\beta}^{(1)}) + (1-t)f(\hat{\beta}^{(2)}) = f^\star,
\end{aligned}
$$

where in the second line we used the strict convexity of the function $G(z) = \|y - z\|_2^2$, along with the convexity of $h(z) = \|z\|_1$. That $t\hat{\beta}^{(1)} + (1-t)\hat{\beta}^{(2)}$ obtains a lower criterion than $f^\star$ is a contradiction, and this proves part (ii).

Lastly, for part (iii), every solution in the generalized lasso problem (2.1) yields the same fit by part (ii), leading to the same squared loss; and since every solution also obtains the same (optimal) criterion value, we conclude that every solution obtains the same penalty value, provided that $\lambda > 0$. □

Next, we consider the Karush-Kuhn-Tucker (or KKT) conditions to characterize optimality of a solution $\hat{\beta}$ in problem (2.1). Since there are no contraints, we simply take a subgradient of the criterion and set it equal to zero. Rearranging gives

$$X^T(y - X\hat{\beta}) = \lambda D^T \hat{\gamma}, \tag{2.2}$$

where $\hat{\gamma} \in \mathbb{R}^m$ is a subgradient of the ℓ_1 norm evaluated at $D\hat{\beta}$,

$$\hat{\gamma}_i \in \begin{cases} \{\mathrm{sign}((D\hat{\beta})_i)\} & \text{if } (D\hat{\beta})_i \neq 0 \\ [-1, 1] & \text{if } (D\hat{\beta})_i = 0 \end{cases}, \quad \text{for } i = 1, \ldots, m. \tag{2.3}$$

Since the optimal fit $X\hat{\beta}$ is unique by Lemma 2.1, the left-hand side in (2.2) is always unique. This immediately leads to the next result.

Lemma 2.2. *For any y, X, D, and $\lambda > 0$, every optimal subgradient $\hat{\gamma}$ in problem* (2.1) *gives rise to the same value of $D^T\hat{\gamma}$. Moreover, when D has full row rank, the optimal subgradient $\hat{\gamma}$ is itself unique.*

Remark 2.1. When D is row rank deficient, the optimal subgradient $\hat{\gamma}$ is not necessarily unique, and thus neither is its associated boundary set (to be defined in the next subsection). This complicates the study of uniqueness of the generalized lasso solution. In contrast, the optimal subgradient in the lasso problem is always unique, and its boundary set—called *equicorrelation set* in this case—is too, which makes the study of uniqueness of the lasso solution comparatively simpler [135].

Lastly, we turn to the dual of problem (2.1). Standard arguments in convex analysis, as given in Tibshirani and Taylor [136], show that the Lagrangian dual of (2.1) can be written as[1]

$$\underset{u \in \mathbb{R}^m,\, v \in \mathbb{R}^n}{\text{minimize}} \ \|y - v\|_2^2 \quad \text{subject to} \quad X^T v = D^T u, \ \|u\|_\infty \leq \lambda. \tag{2.4}$$

Any pair $(\hat{u}, \hat{v})$ optimal in the dual (2.4), and solution-subgradient pair $(\hat{\beta}, \hat{\gamma})$ optimal in the primal (2.1), i.e., satisfying (2.2), (2.3), must satisfy the primal-dual relationships

$$X\hat{\beta} = y - \hat{v}, \quad \text{and} \quad \hat{u} = \lambda\hat{\gamma}. \tag{2.5}$$

We see that $\hat{v}$, being a function of the fit $X\hat{\beta}$, is always unique; meanwhile, $\hat{u}$, being a function of the optimal subgradient $\hat{\gamma}$, is not. Moreover, the optimality of $\hat{v}$ in problem (2.4) can be expressed as

$$\hat{v} = P_C(y), \quad \text{where } C = (X^T)^{-1}\big(D^T B_\infty^m(\lambda)\big). \tag{2.6}$$

Here, $(X^T)^{-1}(S)$ denotes the preimage of a set S under the linear map X^T, $D^T S$ denotes the image of a set S under the linear map D^T, $B_\infty^m(\lambda) = \{u \in \mathbb{R}^m : \|u\|_\infty \leq \lambda\}$ is the ℓ_∞ ball of

[1]The form of the dual problem here may superficially appear different from that in Tibshirani and Taylor [136], but it is equivalent.

radius λ in $\mathbb{R}^m$, and $P_S(\cdot)$ is the Euclidean projection operator onto a set S. Note that C as defined in (2.6) is a convex polyhedron, because the image or preimage of any convex polyhedron under a linear map is a convex polyhedron. From (2.5) and (2.6), we may hence write the fit as

$$X\hat{\beta} = (I - P_C)(y), \tag{2.7}$$

the residual from projecting y onto the convex polyhedron C.

The conclusion in (2.7), it turns out, could have been reached via direction manipulation of the KKT conditions (2.2), (2.3), as shown in Tibshirani and Taylor [137]. In fact, much of what can be seen from the dual problem (2.4) can also be derived using appropriate manipulations of the primal problem (2.1) and its KKT conditions (2.2), (2.3). However, we feel that the dual perspective, specifically the dual projection in (2.6), offers a simple picture that can be used to intuitively explain several key results (which might otherwise seem technical and complicated in nature). We will therefore return to it periodically.

2.2.2 Implicit Form of Solutions

Fix an arbitrary $\lambda > 0$, and let $(\hat{\beta}, \hat{\gamma})$ denote an optimal solution-subgradient pair, i.e., satisfying (2.2), (2.3). Following Tibshirani and Taylor [136, 137], we define the *boundary set* to contain the indices of components of $\hat{\gamma}$ that achieve the maximum possible absolute value,

$$\mathcal{B} = \big\{i \in \{1, \ldots, m\} : |\hat{\gamma}_i| = 1\big\},$$

and the *boundary signs* to be the signs of $\hat{\gamma}$ over the boundary set,

$$s = \mathrm{sign}(\hat{\gamma}_{\mathcal{B}}).$$

Since $\hat{\gamma}$ is not necessarily unique, as discussed in the previous subsection, neither are its associated boundary set and signs $\mathcal{B}, s$. Note that the boundary set contains the *active set*

$$\mathcal{A} = \mathrm{supp}(D\hat{\beta}) = \big\{i \in \{1, \ldots, m\} : (D\hat{\beta})_i \neq 0\big\}$$

associated with $\hat{\beta}$; that $\mathcal{B} \supseteq \mathcal{A}$ follows directly from the property (2.3) (and strict inclusion is certainly possible). Restated, this inclusion tells us that $\hat{\beta}$ must lie in the null space of $D_{-\mathcal{B}}$, i.e.,

$$D_{-\mathcal{B}}\hat{\beta} = 0 \iff \hat{\beta} \in \mathrm{null}(D_{-\mathcal{B}}).$$

Though it seems very simple, the last display provides an avenue for expressing the generalized lasso fit and solutions in terms of $\mathcal{B}, s$, which will be quite useful for establishing sufficient conditions for uniqueness of the solution. Multiplying both sides of the stationarity condition (2.2) by $P_{\mathrm{null}(D_{-\mathcal{B}})}$, the projection matrix onto $\mathrm{null}(D_{-\mathcal{B}})$, we have

$$P_{\mathrm{null}(D_{-\mathcal{B}})}X^T(y - X\hat{\beta}) = \lambda P_{\mathrm{null}(D_{-\mathcal{B}})}D_{\mathcal{B}}^T s.$$

Using $\hat{\beta} = P_{\mathrm{null}(D_{-\mathcal{B}})}\hat{\beta}$, and solving for the fit $X\hat{\beta}$ (see 137 for details or the proof of Lemma 2.17 for the arguments in a more general case) gives

$$X\hat{\beta} = XP_{\mathrm{null}(D_{-\mathcal{B}})}(XP_{\mathrm{null}(D_{-\mathcal{B}})})^+\big(y - \lambda(P_{\mathrm{null}(D_{-\mathcal{B}})}X^T)^+ D_{\mathcal{B}}^T s\big). \tag{2.8}$$

Recalling that $X\hat{\beta}$ is unique from Lemma 2.1, we see that the right-hand side in (2.8) must agree for all instantiations of the boundary set and signs $\mathcal{B}, s$ associated with an optimal subgradient in problem (2.1). Tibshirani and Taylor [137] use this observation and other arguments to establish an important result that we leverage later, on the invariance of the space $X\text{null}(D_{-\mathcal{B}}) = \text{col}(XP_{\text{null}(D_{-\mathcal{B}})})$ over all boundary sets $\mathcal{B}$ of optimal subgradients, stated in Lemma 2.3 for completeness.

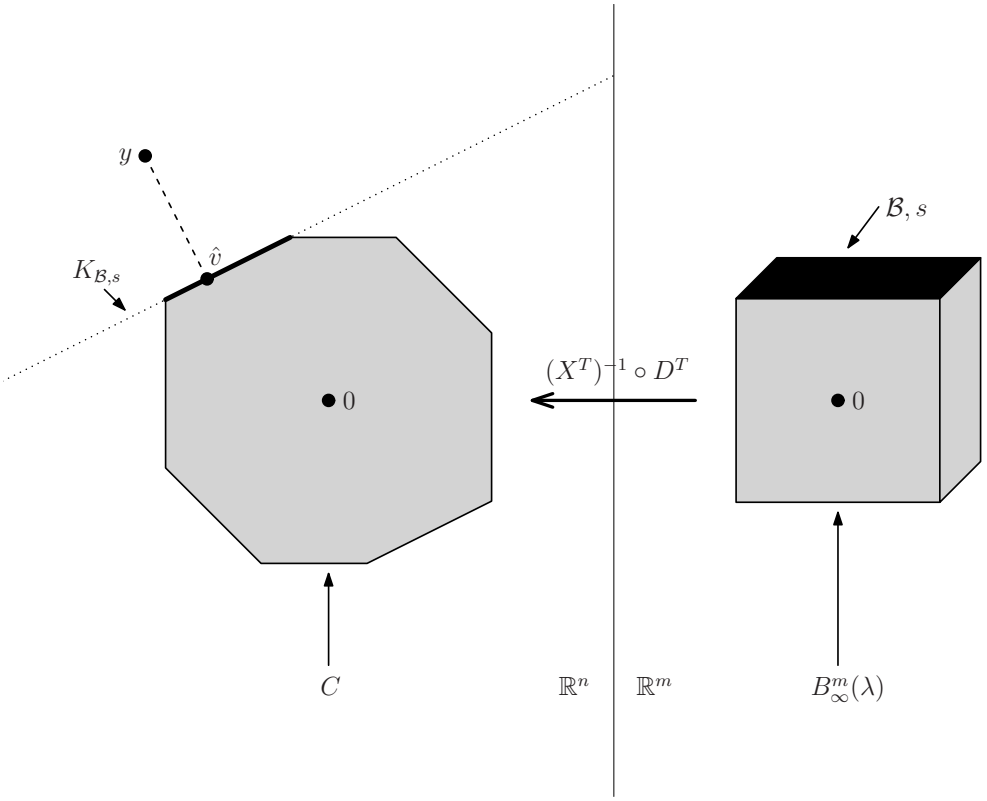

Figure 2.1: *Geometry of the generalized lasso dual problem* (2.4)*. As in* (2.6)*, the dual solution $\hat{v}$ may be seen as the projection of y onto a set C, and as in* (2.7)*, the primal fit $X\hat{\beta}$ may be seen as the residual from this projection. Here, $C = (X^T)^{-1}(D^T B_\infty^m(\lambda))$, and as $B_\infty^m(\lambda)$ is a polyhedron (and the image or inverse image of a polyhedron under a linear map is still a polyhedron), C is a polyhedron as well. This can be used to derive the implicit form* (2.8) *for $X\hat{\beta}$, based on the face of C on which $\hat{v}$ lies, as explained in Remark 2.2.*

Remark 2.2. As an alternative to the derivation based on the KKT conditions described above, the result (2.8) can be argued directly from the geometry surrounding the dual problem (2.4). See Figure 2.1 for an accompanying illustration. Given that $\hat{\gamma}$ has boundary set and signs $\mathcal{B}, s$, and $\hat{u} = \lambda\hat{\gamma}$ from (2.5), we see that $\hat{u}$ must lie on the face of $B_\infty^m(\lambda)$ whose affine span is $E_{\mathcal{B},s} = \{u \in \mathbb{R}^m : u_{\mathcal{B},s} = \lambda s\}$; this face is colored in black on the right-hand side of the figure. Since $X^T\hat{v} = D^T\hat{u}$, this means that $\hat{v}$ lies on the face of C whose affine span is $K_{\mathcal{B},s} = (X^T)^{-1}D^T E_{\mathcal{B},s}$; this face is colored in black on the left-hand side of the figure, and its affine span $K_{\mathcal{B},s}$ is drawn as a dotted line. Hence, we may refine our view of $\hat{v}$ in (2.6), and in turn, $X\hat{\beta}$ in (2.7): namely, we may view $\hat{v}$ as the projection of y onto the affine space

$K_{\mathcal{B},s}$ (instead of C), and the fit $X\hat{\beta}$ as the residual from this affine projection. A straightforward calculation shows that $K_{\mathcal{B},s} = \lambda(P_{\mathrm{null}(D_{-\mathcal{B}})}X^T)^+ D_{\mathcal{B}}^T s + \mathrm{null}(P_{\mathrm{null}(D_{-\mathcal{B}})}X^T)$, and another straightforward calculation shows that the residual from projecting y onto $K_{\mathcal{B},s}$ is (2.8).

From the expression in (2.8) for the fit $X\hat{\beta}$, we also see that the solution $\hat{\beta}$ corresponding to the optimal subgradient $\hat{\gamma}$ and its boundary set and signs $\mathcal{B}, s$ must take the form

$$\hat{\beta} = (XP_{\mathrm{null}(D_{-\mathcal{B}})})^+\big(y - \lambda(P_{\mathrm{null}(D_{-\mathcal{B}})}X^T)^+ D_{\mathcal{B}}^T s\big) + b, \tag{2.9}$$

for some $b \in \mathrm{null}(XP_{\mathrm{null}(D_{-\mathcal{B}})})$. Combining this with $b \in \mathrm{null}(D_{-\mathcal{B}})$ (following from $D_{-\mathcal{B}}\hat{\beta} = 0$), we moreover have that $b \in \mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}})$. In fact, *any* such point $b \in \mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}})$ yields a generalized lasso solution $\hat{\beta}$ in (2.9) provided that

$$s_i \cdot D_i\Big[(XP_{\mathrm{null}(D_{-\mathcal{B}})})^+\big(y - \lambda(P_{\mathrm{null}(D_{-\mathcal{B}})}X^T)^+ D_{\mathcal{B}}^T s\big) + b\Big] \geq 0, \quad \text{for } i \in \mathcal{B},$$

which says that $\hat{\gamma}$ appropriately matches the signs of the nonzero components of $D\hat{\beta}$, thus $\hat{\gamma}$ remains a proper subgradient.

We can now begin to inspect conditions for uniqueness of the generalized lasso solution. For a given boundary set $\mathcal{B}$ of an optimal subgradient $\hat{\gamma}$, if we know that $\mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}}) = \{0\}$, then there can only be one solution $\hat{\beta}$ corresponding to $\hat{\gamma}$ (i.e., such that $(\hat{\beta}, \hat{\gamma})$ jointly satisfy (2.2), (2.3)), and it is given by the expression in (2.9) with $b = 0$. Further, if we know that $\mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}}) = \{0\}$ for *all* boundary sets $\mathcal{B}$ of optimal subgradients, and the space $\mathrm{null}(D_{-\mathcal{B}})$ is invariant over all choices of boundary sets $\mathcal{B}$ of optimal subgradients, then the right-hand side in (2.9) with $b = 0$ must agree for all proper instantiations of $\mathcal{B}, s$ and it gives the unique generalized lasso solution. We elaborate on this in the next section.

2.2.3 Invariance of the Linear Space $X\mathrm{null}(D_{-\mathcal{B}})$

Before diving into the technical details on conditions for uniqueness in the next section, we recall a key result from Tibshirani and Taylor [137].

Lemma 2.3 (Lemma 10 in 137)**.** *Fix any X, D, and $\lambda > 0$. There is a set $\mathcal{N} \subseteq \mathbb{R}^n$ of Lebesgue measure zero (that depends on X, D, λ), such that for $y \notin \mathcal{N}$, all boundary sets $\mathcal{B}$ associated with optimal subgradients in the generalized lasso problem* (2.1) *give rise to the same subspace $X\mathrm{null}(D_{-\mathcal{B}})$, i.e., there is a single linear subspace $L \subseteq \mathbb{R}^n$ such that $L = X\mathrm{null}(D_{-\mathcal{B}})$ for all boundary sets $\mathcal{B}$ of optimal subgradients. Moreover, for $y \notin \mathcal{N}$, $L = X\mathrm{null}(D_{-\mathcal{A}})$ for all active sets $\mathcal{A}$ associated with generalized lasso solutions.*

2.3 Sufficient Conditions for Uniqueness

2.3.1 A Condition on Certain Linear Independencies

We start by formalizing the discussion on uniqueness in the paragraphs proceeding (2.9). As before, let $\lambda > 0$, and let $\mathcal{B}$ denote the boundary set associated with an optimal subgradient in (2.1).

Denote by $U(\mathcal{B}) \in \mathbb{R}^{p \times k(\mathcal{B})}$ a matrix with linearly independent columns that span $\mathrm{null}(D_{-\mathcal{B}})$. It is not hard to see that

$$\mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}}) = \{0\} \iff \mathrm{null}\big(XU(\mathcal{B})\big) = \{0\} \iff \mathrm{rank}\big(XU(\mathcal{B})\big) = k(\mathcal{B}).$$

Let us assign now such a basis matrix $U(\mathcal{B}) \in \mathbb{R}^{p \times k(\mathcal{B})}$ to each boundary set $\mathcal{B}$ corresponding to an optimal subgradient in (2.1). We claim that there is a unique generalized lasso solution, as given in (2.9) with $b = 0$, provided that the following two conditions holds:

$$\mathrm{rank}\big(XU(\mathcal{B})\big) = k(\mathcal{B}) \text{ for all such boundary sets } \mathcal{B}, \text{ and} \tag{2.10}$$

$$\mathrm{null}(D_{-\mathcal{B}}) \text{ is invariant across all such boundary sets } \mathcal{B}. \tag{2.11}$$

To see this, note that if the space $\mathrm{null}(D_{-\mathcal{B}})$ is invariant across all achieved boundary sets $\mathcal{B}$ then so is the matrix $P_{\mathrm{null}(D_{-\mathcal{B}})}$. This, and the fact that $P_{\mathrm{null}(D_{-\mathcal{B}})} D_B^T s = P_{\mathrm{null}(D_{-\mathcal{B}})} D^T \hat{\gamma}$ where $D^T \hat{\gamma}$ is unique from Lemma 2.2, ensures that the right-hand side in (2.9) with $b = 0$ agrees no matter the choice of boundary set and signs $\mathcal{B}, s$.

Remark 2.3. For any subset $\mathcal{B} \subseteq \{1, \ldots, m\}$, and any matrices $U(\mathcal{B}), \tilde{U}(\mathcal{B}) \in \mathbb{R}^{p \times k(\mathcal{B})}$ whose columns form a basis for $\mathrm{null}(D_{-\mathcal{B}})$, it is easy to check that $\mathrm{rank}(XU(\mathcal{B})) = k(\mathcal{B}) \iff \mathrm{rank}(X\tilde{U}(\mathcal{B})) = k(\mathcal{B})$. Therefore condition (2.10) is well-defined, i.e., it does not depend on the choice of basis matrix $U(\mathcal{B})$ associated with $\mathrm{null}(D_{-\mathcal{B}})$ for each boundary set $\mathcal{B}$.

We now show that, thanks to Lemma 2.3, condition (2.10) (almost everywhere) implies (2.11), so the former is alone sufficient for uniqueness.

Lemma 2.4. *Fix any X, D, and $\lambda > 0$. For $y \notin \mathcal{N}$, where $\mathcal{N} \subseteq \mathbb{R}^n$ has Lebesgue measure zero as in Lemma 2.3, condition* (2.10) *implies* (2.11)*. Hence, for almost every y, condition* (2.10) *is itself sufficient to imply uniqueness of the generalized lasso solution.*

Proof. Let $y \notin \mathcal{N}$, and let L be the linear subspace from Lemma 2.3, i.e., $L = X\mathrm{null}(D_{-\mathcal{B}})$ for any boundary set $\mathcal{B}$ associated with an optimal subgradient in the generalized lasso problem at y. Now fix a particular boundary set $\mathcal{B}$ associated with an optimal subgradient and define the linear map $\mathcal{X} : \mathrm{null}(D_{-\mathcal{B}}) \to L$ by $\mathcal{X}(u) = Xu$. By construction, this map is surjective. Moreover, assuming (2.10), it is injective, as

$$XU(\mathcal{B})a = XU(\mathcal{B})b \iff XU(\mathcal{B})(a - b) = 0,$$

and the right-hand side cannot be true unless $a = b$. Therefore, $\mathcal{X}$ is bijective and has a linear inverse, and we may write $\mathrm{null}(D_{-\mathcal{B}}) = \mathcal{X}^{-1}(L)$. As $\mathcal{B}$ was arbitrary, this shows the invariance of $\mathrm{null}(D_{-\mathcal{B}})$ over all proper choices of $\mathcal{B}$, whenever $y \notin \mathcal{N}$. □

From Lemma 2.4, we see that an (almost everywhere) sufficient condition for a unique solution in (2.1) is that the vectors $XU_i(\mathcal{B}) \in \mathbb{R}^n$, $i = 1, \ldots, k(\mathcal{B})$ are linearly independent, for all instantiations of boundary sets $\mathcal{B}$ of optimal subgradients. This may seem a little circular, to give a condition for uniqueness that itself is expressed in terms of the subgradients of solutions. But we will not stop at (2.10), and will derive more explicit conditions on y, X, D, and $\lambda > 0$ that imply (2.10) and therefore uniqueness of the solution in (2.1).

2.3.2 A Refined Condition on Linear Independencies

The next lemma shows that when condition (2.10) fails, there is a specific type of linear dependence among the columns of $XU(\mathcal{B})$, for a boundary set $\mathcal{B}$. The proof is not difficult, but involves careful manipulations of the KKT conditions (2.2), and we defer it until the supplementary material.

Lemma 2.5. *Fix any X, D, and $\lambda > 0$. Let $y \notin \mathcal{N}$, the set of zero Lebesgue measure as in Lemma 2.3. Assume that $\mathrm{null}(X) \cap \mathrm{null}(D) = \{0\}$, and that the generalized lasso solution is not unique. Then there is a pair of boundary set and signs $\mathcal{B}, s$ corresponding to an optimal subgradient in problem (2.1), such that for any matrix $U(\mathcal{B}) \in \mathbb{R}^{p \times k(\mathcal{B})}$ whose columns form a basis for $\mathrm{null}(D_{-\mathcal{B}})$, the following property holds of $Z = XU(\mathcal{B})$ and $\tilde{s} = U(\mathcal{B})^T D_{\mathcal{B}}^T s$: there exist indices $i_1, \ldots, i_k \in \{1, \ldots, k(\mathcal{B})\}$ with $k \leq n+1$ and $\tilde{s}_{i_1} \neq 0$, such that*

$$Z_{i_2} \in \mathrm{span}(\{Z_{i_3}, \ldots, Z_{i_k}\}), \tag{2.12}$$

when $\tilde{s}_{i_2} = \cdots = \tilde{s}_{i_k} = 0$, and

$$Z_{i_1}/\tilde{s}_{i_1} \in \mathrm{aff}(\{Z_{i_j}/\tilde{s}_{i_j} : \tilde{s}_{i_j} \neq 0,\, j \geq 2\}) + \mathrm{span}(\{Z_{i_j} : \tilde{s}_{i_j} = 0\}), \tag{2.13}$$

when at least one of $\tilde{s}_{i_2}, \ldots, \tilde{s}_{i_k}$ is nonzero.

The spaces on the right-hand sides of both (2.12), (2.13) are of dimension at most $n-1$. To see this, note that $\dim(\mathrm{span}(\{Z_{i_3}, \ldots, Z_{i_k}\})) \leq k-2 \leq n-1$, and also

$$\begin{aligned}\dim\big(\mathrm{aff}(\{Z_{i_j}/\tilde{s}_{i_j} : \tilde{s}_{i_j} \neq 0,\, j \geq 2\})\big) + \dim\big(\mathrm{span}(\{Z_{i_j} : \tilde{s}_{i_j} = 0\})\big) \\ \leq |\mathcal{J}| - 2 + |\mathcal{J}^c| = k - 2 \leq n - 1,\end{aligned}$$

where $\mathcal{J} = \{j \in \{1, \ldots, k\} : \tilde{s}_{i_j} \neq 0\}$. Hence, because these spaces are at most $(n-1)$-dimensional, neither condition (2.12) nor (2.13) should be "likely" under a continuous distribution for the predictor variables X. This is made precise in the next subsection.

Before this, we define a deterministic condition on X that ensures special linear dependencies between the (transformed) columns, as in (2.12), (2.13), never hold.

Definition 1. Fix $D \in \mathbb{R}^{m \times p}$. We say that a matrix $X \in \mathbb{R}^{n \times p}$ is in *D-general position* (or *D-GP*) if the following property holds. For each subset $\mathcal{B} \subseteq \{1, \ldots, m\}$ and sign vector $s \in \{-1, 1\}^{|\mathcal{B}|}$, there is a matrix $U(\mathcal{B}) \in \mathbb{R}^{p \times k(\mathcal{B})}$ whose columns form a basis for $\mathrm{null}(D_{-\mathcal{B}})$, such that for $Z = XU(\mathcal{B})$, $\tilde{s} = U(\mathcal{B})^T D_{\mathcal{B}}^T s$, and all $i_1, \ldots, i_k \in \{1, \ldots, k(\mathcal{B})\}$ with $\tilde{s}_{i_1} \neq 0$ and $k \leq n+1$, it holds that

(i) $Z_{i_2} \notin \mathrm{span}(\{Z_{i_3}, \ldots, Z_{i_k}\})$, when $\tilde{s}_{i_2} = \cdots = \tilde{s}_{i_k} = 0$;

(ii) $Z_{i_1}/\tilde{s}_{i_1} \notin \mathrm{aff}(\{Z_{i_j}/\tilde{s}_{i_j} : \tilde{s}_{i_j} \neq 0,\, j \geq 2\}) + \mathrm{span}(\{Z_{i_j} : \tilde{s}_{i_j} = 0\})$, when at least one of $\tilde{s}_{i_2}, \ldots, \tilde{s}_{i_k}$ is nonzero.

Remark 2.4. Though the definition may appear somewhat complicated, a matrix X being in D-GP is actually quite a weak condition, and can hold regardless of the (relative) sizes of n, p. We will show in the next subsection that it holds almost surely under an arbitrary continuous probability distribution for the entries of X. Further, when $X = I$, the above definition essentially reduces[2] to the usual notion of *general position* (refer to, e.g., 135 for this definition).

[2]We say "essentially" here, because our definition of D-GP with $D = I$ allows for a choice of basis matrix $U(\mathcal{B})$ for each subset $\mathcal{B}$, whereas the standard notion of generally position would mandate (in the notation of our definition) that $U(\mathcal{B})$ be given by the columns of I indexed by $\mathcal{B}$.

When X is in D-GP, we have (by definition) that (2.12), (2.13) cannot hold for *any* $\mathcal{B} \subseteq \{1, \ldots, m\}$ and $s \in \{-1, 1\}^{|\mathcal{B}|}$ (not just boundary sets and signs); therefore, by the contrapositive of Lemma 2.5, if we additionally have $y \notin \mathcal{N}$ and $\mathrm{null}(X) \cap \mathrm{null}(D) = \{0\}$, then the generalized lasso solution must be unique. To emphasize this, we state it as a lemma.

Lemma 2.6. *Fix any X, D, and $\lambda > 0$. If $y \notin \mathcal{N}$, the set of zero Lebesgue measure as in Lemma 2.3, $\mathrm{null}(X) \cap \mathrm{null}(D) = \{0\}$, and X is in D-GP, then the generalized lasso solution is unique.*

2.3.3 Absolutely Continuous Predictor Variables

We give an important result that shows the D-GP condition is met almost surely for continuously distributed predictors. There are no restrictions on the relative sizes of n, p. The proof of the next result uses elementary probability arguments and is deferred until the supplementary material.

Lemma 2.7. *Fix $D \in \mathbb{R}^{m \times p}$, and assume that the entries of $X \in \mathbb{R}^{n \times p}$ are drawn from a distribution that is absolutely continuous with respect to (np)-dimensional Lebesgue measure. Then X is in D-GP almost surely.*

We now present a result showing that the base condition $\mathrm{null}(X) \cap \mathrm{null}(D) = \{0\}$ is met almost surely for continuously distributed predictors, provided that $p \leq n$, or $p > n$ and the null space of D is not too large. Its proof is elementary and found in the supplementary material.

Lemma 2.8. *Fix $D \in \mathbb{R}^{m \times p}$, and assume that the entries of $X \in \mathbb{R}^{n \times p}$ are drawn from a distribution that is absolutely continuous with respect to (np)-dimensional Lebesgue measure. If either $p \leq n$, or $p > n$ and $\mathrm{nullity}(D) \leq n$, then $\mathrm{null}(X) \cap \mathrm{null}(D) = \{0\}$ almost surely.*

Putting together Lemmas 2.6, 2.7, 2.8 gives our main result on the uniqueness of the generalized lasso solution.

Theorem 2.1. *Fix any D and $\lambda > 0$. Assume the joint distribution of (X, y) is absolutely continuous with respect to $(np + n)$-dimensional Lebesgue measure. If $p \leq n$, or else $p > n$ and $\mathrm{nullity}(D) \leq n$, then the solution in the generalized lasso problem* (2.1) *is unique almost surely.*

Remark 2.5. If D has full row rank, then by Lemma 2.2 the optimal subgradient $\hat{\gamma}$ is unique and so the boundary set $\mathcal{B}$ is also unique. In this case, condition (2.11) is vacuous and condition (2.10) is sufficient for uniqueness of the generalized lasso solution for every y (i.e., we do not need to rely on Lemma 2.4, which in turn uses Lemma 2.3, to prove that (2.10) is sufficient for almost every y). Hence, in this case, the condition in Theorem 2.1 that $y|X$ has an absolutely continuous distribution is not needed, and (with the other conditions in place) uniqueness holds for every y, almost surely over X. Under this (slight) sharpening, Theorem 2.1 with $D = I$ reduces to the result in Lemma 4 of Tibshirani [135].

Remark 2.6. Generally speaking, the condition that $\mathrm{nullity}(D) \leq n$ in Theorem 2.1 (assumed in the case $p > n$) is not strong. In many applications of the generalized lasso, the dimension of the null space of D is small and fixed (i.e., it does not grow with n). For example, recall Corollary 2.1, where the lower bound n in each of the cases reflects the dimension of the null space.

2.3.4 Standardized Predictor Variables

A common preprocessing step, in many applications of penalized modeling such as the generalized lasso, is to *standardize* the predictors $X \in \mathbb{R}^{n \times p}$, meaning, center each column to have

mean 0, and then scale each column to have norm 1. Here we show that our main uniqueness results carry over, mutatis mutandis, to the case of standardized predictor variables. All proofs in this subsection are deferred until the supplementary material.

We begin by studying the case of centering alone. Let $M = I - \mathbb{1}\mathbb{1}^T/n \in \mathbb{R}^{n\times n}$ be the centering map, and consider the *centered generalized lasso* problem

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \frac{1}{2}\|y - MX\beta\|_2^2 + \lambda\|D\beta\|_1. \tag{2.14}$$

We have the following uniqueness result for centered predictors.

Corollary 2.2. *Fix any D and $\lambda > 0$. Assume the distribution of (X, y) is absolutely continuous with respect to $(np + n)$-dimensional Lebesgue measure. If $p \leq n - 1$, or $p > n - 1$ and $\text{nullity}(D) \leq n - 1$, then the solution in the centered generalized lasso problem* (2.14) *is unique almost surely.*

Remark 2.7. The exact same result as stated in Corollary 2.2 holds for the generalized lasso problem with intercept

$$\underset{\beta_0\in\mathbb{R},\,\beta\in\mathbb{R}^p}{\text{minimize}} \ \frac{1}{2}\|y - \beta_0\mathbb{1} - X\beta\|_2^2 + \lambda\|D\beta\|_1. \tag{2.15}$$

This is because, by minimizing over β_0 in problem (2.15), we find that this problem is equivalent to minimization of

$$\frac{1}{2}\|My - MX\beta\|_2^2 + \lambda\|D\beta\|_1$$

over β, which is just a generalized lasso problem with response $V_{-1}^T y$ and predictors $V_{-1}^T X$, where the notation here is as in the proof of Corollary 2.2.

Next we treat the case of scaling alone. Let $W_X = \text{diag}(\|X_1\|_2, \ldots, \|X_p\|_2) \in \mathbb{R}^{p\times p}$, and consider the *scaled generalized lasso* problem

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \frac{1}{2}\|y - XW_X^{-1}\beta\|_2^2 + \lambda\|D\beta\|_1. \tag{2.16}$$

We give a helper lemma, on the distribution of a continuous random vector, post scaling.

Lemma 2.9. *Let $Z \in \mathbb{R}^n$ be a random vector whose distribution is absolutely continuous with respect to n-dimensional Lebesgue measure. Then, the distribution of $Z/\|Z\|_2$ is absolutely continuous with respect to $(n-1)$-dimensional Hausdorff measure restricted to the $(n-1)$-dimensional unit sphere, $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$.*

We give a second helper lemma, on the $(n-1)$-dimensional Hausdorff measure of an affine space intersected with the unit sphere $\mathbb{S}^{n-1}$ (which is important for checking that the scaled predictor matrix is in D-GP, because here we must check that none of its columns lie in a finite union of affine spaces).

Lemma 2.10. *Let $A \subseteq \mathbb{R}^n$ be an arbitrary affine space, with $\dim(A) \leq n - 1$. Then $\mathbb{S}^{n-1} \cap A$ has $(n-1)$-dimensional Hausdorff measure zero.*

We present a third helper lemma, which establishes that for absolutely continuous X, the scaled predictor matrix XW_X^{-1} is in D-GP and satisfies the appropriate null space condition, almost surely.

Lemma 2.11. *Fix $D \in \mathbb{R}^{m\times p}$, and assume that $X \in \mathbb{R}^{n\times p}$ has entries drawn from a distribution that is absolutely continuous with respect to (np)-dimensional Lebesgue measure. Then XW_X^{-1} is in D-GP almost surely. Moreover, if $p \leq n$, or $p > n$ and $\mathrm{nullity}(D) \leq n$, then $\mathrm{null}(XW_X^{-1}) \cap \mathrm{null}(D) = \{0\}$ almost surely.*

Combining Lemmas 2.6, 2.11 gives the following uniqueness result for scaled predictors.

Corollary 2.3. *Fix any D and $\lambda > 0$. Assume the distribution of (X, y) is absolutely continuous with respect to $(np + n)$-dimensional Lebesgue measure. If $p \leq n$, or else $p > n$ and $\mathrm{nullity}(D) \leq n$, then the solution in the scaled generalized lasso problem* (2.16) *is unique almost surely.*

Finally, we consider the *standardized generalized lasso* problem,

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \frac{1}{2}\|y - MXW_{MX}^{-1}\beta\|_2^2 + \lambda\|D\beta\|_1, \tag{2.17}$$

where, note, the predictor matrix MXW_{MX}^{-1} has standardized columns, i.e., each column has been centered to have mean 0, then scaled to have norm 1. We have the following uniqueness result.

Corollary 2.4. *Fix any D and $\lambda > 0$. Assume the distribution of (X, y) is absolutely continuous with respect to $(np + n)$-dimensional Lebesgue measure. If $p \leq n - 1$, or $p > n - 1$ and $\mathrm{nullity}(D) \leq n - 1$, then the solution in the standardized generalized lasso problem* (2.17) *is unique almost surely.*

2.4 Smooth, Strictly Convex Loss Functions

2.4.1 Generalized Lasso with a General Loss

We now extend some of the preceding results beyond the case of squared error loss, as considered previously. In particular, we consider the problem

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ G(X\beta; y) + \lambda\|D\beta\|_1, \tag{2.18}$$

where we assume, for each $y \in \mathbb{R}^n$, that the function $G(\,\cdot\,; y)$ is *essentially smooth* and *essentially strictly convex* on $\mathbb{R}^n$. These two conditions together mean that $G(\,\cdot\,; y)$ is a closed proper convex function, differentiable and strictly convex on the interior of its domain (assumed to be nonempty), with the norm of its gradient approaching ∞ along any sequence approaching the boundary of its domain. A function that is essentially smooth and essentially strictly convex is also called, according to some authors, of *Legendre type*; see Chapter 26 of Rockafellar [115]. An important special case of a Legendre function is one that is differentiable and strictly convex, with full domain (all of $\mathbb{R}^n$).

For much of what follows, we will focus on loss functions of the form

$$G(z; y) = -y^T z + \psi(z), \tag{2.19}$$

for an essentially smooth and essentially strictly convex function ψ on $\mathbb{R}^n$ (not depending on y). This is a weak restriction on G and encompasses, e.g., the cases in which G is the negative log-likelihood function from a generalized linear model (GLM) for the entries of $y|X$ with

a canonical link function, where ψ is the cumulant generating function. In the case of, say, Bernoulli or Poisson models, this is

$$G(z;y) = -y^T z + \sum_{i=1}^n \log(1+e^{z_i}), \quad \text{or} \quad G(z;y) = -y^T z + \sum_{i=1}^n e^{z_i},$$

respectively. For brevity, we will often write the loss function as $G(X\beta)$, hiding the dependence on the response vector y.

2.4.2 Basic Facts, KKT Conditions, and the Dual

The next lemma follows from arguments identical to those for Lemma 2.1.

Lemma 2.12. *For any $y, X, D, \lambda \geq 0$, and for G essentially smooth and essentially strictly convex, the following holds of problem* (2.18)*.*

(i) There is either zero, one, or uncountably many solutions.

(ii) Every solution $\hat{\beta}$ gives rise to the same fitted value $X\hat{\beta}$.

(iii) If $\lambda > 0$, then every solution $\hat{\beta}$ gives rise to the same penalty value $\|D\hat{\beta}\|_1$.

Note the difference between Lemmas 2.12 and 2.1, part (i): for an arbitrary (essentially smooth and essentially strictly convex) G, the criterion in (2.18) need not attain its infimum, whereas the criterion in (2.1) always does. This happens because the criterion in (2.18) can have directions of strict recession (i.e., directions of recession in which the criterion is not constant), whereas the citerion in (2.1) cannot. Thus in general, problem (2.18) need not have a solution; this is true even in the most fundamental cases of interest beyond squared loss, e.g., the case of a Bernoulli negative log-likelihood G. Later in Lemma 2.14, we give a sufficient condition for the existence of solutions in (2.18).

The KKT conditions for problem (2.18) are

$$-X^T \nabla G(X\hat{\beta}) = \lambda D^T \hat{\gamma}, \tag{2.20}$$

where $\hat{\gamma} \in \mathbb{R}^m$ is (as before) a subgradient of the ℓ_1 norm evaluated at $D\hat{\beta}$,

$$\hat{\gamma}_i \in \begin{cases} \{\text{sign}((D\hat{\beta})_i)\} & \text{if } (D\hat{\beta})_i \neq 0 \\ [-1,1] & \text{if } (D\hat{\beta})_i = 0 \end{cases}, \quad \text{for } i = 1, \ldots, m. \tag{2.21}$$

As in the squared loss case, uniqueness of $X\hat{\beta}$ by Lemma 2.12, along with (2.20), imply the next result.

Lemma 2.13. *For any $y, X, D, \lambda > 0$, and G essentially smooth and essentially strictly convex, every optimal subgradient $\hat{\gamma}$ in problem* (2.18) *gives rise to the same value of $D^T\hat{\gamma}$. Furthermore, when D has full row rank, the optimal subgradient $\hat{\gamma}$ is unique, assuming that problem* (2.18) *has a solution in the first place.*

Denote by G^* the conjugate function of G. When G is essentially smooth and essentially strictly convex, the following facts hold (e.g., see Theorem 26.5 of Rockafellar [115]):

- its conjugate G^* is also essentially smooth and essentially strictly convex; and

- the map $\nabla G : \mathrm{int}(\mathrm{dom}(G)) \to \mathrm{int}(\mathrm{dom}(G^*))$ is a homeomorphism with inverse $(\nabla G)^{-1} = \nabla G^*$.

The conjugate function is intrinsically tied to duality, directions of recession, and the existence of solutions. Standard arguments in convex analysis, deferred to the supplementary material, give the next result.

Lemma 2.14. *Fix any y, X, D, and $\lambda \geq 0$. Assume G is essentially smooth and essentially strictly convex. The Lagrangian dual of problem* (2.18) *can be written as*

$$\underset{u \in \mathbb{R}^m,\, v \in \mathbb{R}^u}{\text{minimize}} \; G^*(-v) \quad \text{subject to} \quad X^T v = D^T u, \; \|u\|_\infty \leq \lambda, \tag{2.22}$$

where G^* is the conjugate of G. Any dual optimal pair $(\hat{u}, \hat{v})$ in (2.22)*, and primal optimal solution-subgradient pair $(\hat{\beta}, \hat{\gamma})$ in* (2.18)*, i.e., satisfying* (2.20)*,* (2.21)*, assuming they all exist, must satisfy the primal-dual relationships*

$$\nabla G(X\hat{\beta}) = -\hat{v}, \quad \textit{and} \quad \hat{u} = \lambda \hat{\gamma}. \tag{2.23}$$

Lastly, existence of primal and dual solutions is guaranteed under the conditions

$$0 \in \mathrm{int}(\mathrm{dom}(G)), \tag{2.24}$$

$$(-C) \cap \mathrm{int}(\mathrm{ran}(\nabla G)) \neq \emptyset, \tag{2.25}$$

where $C = (X^T)^{-1}(D^T B_\infty^m(\lambda))$. In particular, under (2.24) *and $C \neq \emptyset$, a solution exists in the dual problem* (2.22)*, and under* (2.24)*,* (2.25)*, a solution exists in the primal problem* (2.18)*.*

Assuming that primal and dual solutions exist, we see from (2.23) in the above lemma that $\hat{v}$ must be unique (by uniqueness of $X\hat{\beta}$, from Lemma 2.12), but $\hat{u}$ need not be (as $\hat{\gamma}$ is not necessarily unique). Moreover, under condition (2.24), we know that G is differentiable at 0, and $\nabla G^*(\nabla G(0)) = 0$, hence we may rewrite (2.22) as

$$\underset{u \in \mathbb{R}^m,\, v \in \mathbb{R}^n}{\text{minimize}} \; D_{G^*}\big(-v, \nabla G(0)\big) \quad \text{subject to} \quad X^T v = D^T u, \; \|u\|_\infty \leq \lambda, \tag{2.26}$$

where $D_f(x, z) = f(x) - f(z) - \langle \nabla f(z), x - z \rangle$ denotes the *Bregman divergence* between points x, z, with respect to a function f. Optimality of $\hat{v}$ in (2.26) may be expressed as

$$\hat{v} = -P_{-C}^{G^*}\big(\nabla G(0)\big), \quad \text{where } C = (X^T)^{-1}\big(D^T B_\infty^m(\lambda)\big). \tag{2.27}$$

Here, recall $(X^T)^{-1}(S)$ denotes the preimage of a set S under the linear map X^T, $D^T S$ denotes the image of a set S under the linear map D^T, $B_\infty^m(\lambda) = \{u \in \mathbb{R}^m : \|u\|_\infty \leq \lambda\}$ is the ℓ_∞ ball of radius λ in $\mathbb{R}^m$, and now $P_S^f(\cdot)$ is the projection operator onto a set S with respect to the Bregman divergence of a function f, i.e., $P_S^f(z) = \arg\min_{x \in S} D_f(x, z)$. From (2.27) and (2.23), we see that

$$X\hat{\beta} = \nabla G^*\Big(P_{-C}^{G^*}\big(\nabla G(0)\big)\Big). \tag{2.28}$$

We note the analogy between (2.27), (2.28) and (2.6), (2.7) in the squared loss case; for $G(z) = \frac{1}{2}\|y - z\|_2^2$, we have $\nabla G(0) = -y$, $G^*(z) = \frac{1}{2}\|y + z\|_2^2 - \frac{1}{2}\|y\|_2^2$, $\nabla G^*(z) = y + z$, $-P_{-C}^{G^*}(\nabla G(0)) = P_C(y)$, and so (2.27), (2.28) match (2.6), (2.7), respectively. But when G is non-quadratic, we

see that the dual solution $\hat{v}$ and primal fit $X\hat{\beta}$ are given in terms of a non-Euclidean projection operator, defined with respect to the Bregman divergence of G^*. See Figure 2.2 for an illustration. This complicates the study of the primal and dual problems, in comparison to the squared loss case; still, as we will show in the coming subsections, several key properties of primal and dual solutions carry over to the current general loss setting.

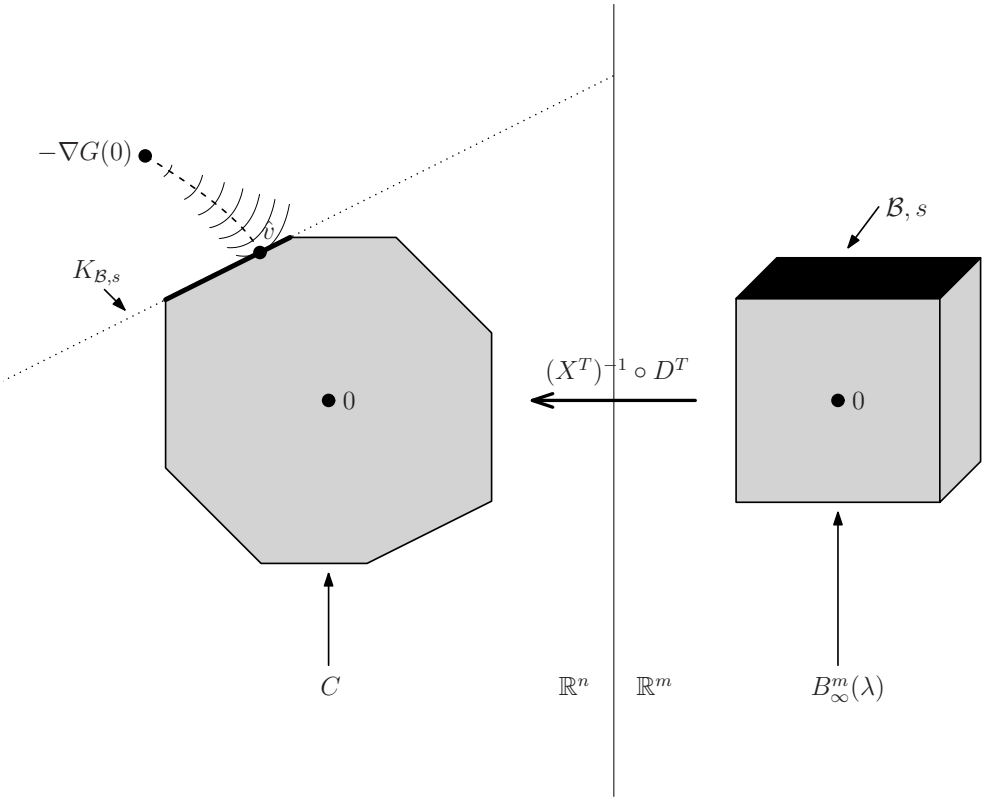

Figure 2.2: *Geometry of the dual problem* (2.26)*, for a general loss G. As in* (2.27)*, the dual solution $\hat{v}$ may be seen as the Bregman projection of $-\nabla G(0)$ onto a set C with respect to the map $x \mapsto G^*(-x)$ (where G^* is the conjugate of G). Shown in the figure are the contours of this map, around $-\nabla G(0)$; the point $\hat{v}$ lies at the intersection of the lowest-level contour and C. Here, as in the squared loss case, $C = (X^T)^{-1}(D^T B_\infty^m(\lambda))$, which is a polyhedron. This realization can be used to derive the implicit form* (2.38) *for $X\hat{\beta}$, based on* (2.28) *and the face of C on which $\hat{v}$ lies, as explained in Remark 2.10.*

2.4.3 Existence in (Regularized) GLMs

Henceforth, we focus on the case in which G takes the form (2.19). The stationarity condition (2.20) is

$$X^T\big(y - \nabla\psi(X\hat{\beta})\big) = \lambda D^T\hat{\gamma}, \tag{2.29}$$

and using the identities $G^*(x) = \psi^*(x+y)$, $P_S^{G^*}(x) = P_{S+y}^{\psi^*}(x+y) - y$, the dual and primal projections, (2.27) and (2.28), become

$$\hat{v} = y - P_{y-C}^{\psi^*}\big(\nabla\psi(0)\big), \quad \text{and} \quad X\hat{\beta} = \nabla\psi^*\Big(P_{y-C}^{\psi^*}\big(\nabla\psi(0)\big)\Big). \tag{2.30}$$

As a check, in the squared loss case, we have $\psi(z) = \frac{1}{2}\|z\|_2^2$, $\nabla\psi(0) = 0$, $\psi^*(z) = \frac{1}{2}\|z\|_2^2$, $\nabla\psi^*(z) = z$, $P_{y-C}^{\psi^*}(\nabla\psi(0)) = y - P_C(y)$, so (2.30) matches (2.6), (2.7). Finally, the conditions (2.24), (2.25) that guarantee the existence of primal and dual solutions become

$$0 \in \mathrm{int}(\mathrm{dom}(\psi)), \tag{2.31}$$

$$y \in \mathrm{int}(\mathrm{ran}(\nabla\psi)) + C, \tag{2.32}$$

where recall $C = (X^T)^{-1}(D^T B_\infty^m(\lambda))$.

We take somewhat of a detour from our main goal (establishing uniqueness in (2.18)), and study the existence conditions (2.31), (2.32). To gather insight, we examine them in detail for some cases of interest. We begin by looking at unregularized ($\lambda = 0$) logistic and Poisson regression. The proof of the next result is straightforward in all but the logistic regression case, and is given in the supplementary material.

Lemma 2.15. *Fix any y, X. Assume that G is of the form* (2.19)*, where ψ is essentially smooth and essentially strictly convex, satisfying $0 \in \mathrm{int}(\mathrm{dom}(\psi))$. Consider problem* (2.18)*, with $\lambda = 0$. Then the sufficient condition* (2.32) *for the existence of a solution is equivalent to*

$$y \in \mathrm{int}(\mathrm{ran}(\nabla\psi)) + \mathrm{null}(X^T). \tag{2.33}$$

For logistic regression, where $\psi(z) = \sum_{i=1}^n \log(1 + e^{z_i})$ and $y \in \{0,1\}^n$, if we write $Y_i = 2y_i - 1 \in \{-1, 1\}$, $i = 1, \ldots, n$, and we denote by $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ the rows of X, then condition (2.33) *is equivalent to*

$$\textit{there does not exist } b \neq 0 \textit{ such that } Y_i x_i^T b \geq 0,\ i = 1, \ldots, n. \tag{2.34}$$

For Poisson regression, where $\psi(z) = \sum_{i=1}^n e^{z_i}$ and $y \in \mathbb{N}^n$ (where $\mathbb{N} = \{0, 1, 2, \ldots\}$ denotes the set of natural numbers), condition (2.33) *is equivalent to*

$$\textit{there exists } \delta \in \mathrm{null}(X^T) \textit{ such that } y_i + \delta_i > 0,\ i = 1, \ldots, n. \tag{2.35}$$

Remark 2.8. For the cases of logistic and Poisson regression, the lemma shows that the sufficient condition (2.32) for the existence of a solution (note (2.31) is automatically satisfied, as $\mathrm{dom}(\psi) = \mathbb{R}^n$ in these cases) reduces to (2.34) and (2.35), respectively. Interestingly, in both cases, this recreates a well-known *necessary* and sufficient condition for the existence of the maximum likelihood estimate (MLE); see Albert and Anderson [2] for the logistic regression condition (2.34), and Haberman [51] for the Poisson regression condition (2.35). The former condition (2.34) is particularly intuitive, and says that the logistic MLE exists if and only if there is no hyperplane that "quasicompletely" separates the points x_i, $i = 1, \ldots, n$ into the positive and negative classes (using the terminology of Albert and Anderson [2]). For a modern take on this condition, see Candes and Sur [20].

Now we inspect the regularized case ($\lambda > 0$). The proof of the next result is straightforward and can be found in the supplementary material.

Lemma 2.16. *Fix any y, X, D, and $\lambda > 0$. Assume that G is of the form* (2.19)*, where we are either in the logistic case, $\psi(z) = \sum_{i=1}^n \log(1 + e^{z_i})$ and $y \in \{0, 1\}^n$, or in the Poisson case, $\psi(z) = \sum_{i=1}^n e^{z_i}$ and $y \in \mathbb{N}^n$ In either case, a sufficient (but not necessary) condition for* (2.32) *to hold, and hence for a solution to exist in problem* (2.18)*, is*

$$\mathrm{null}(D) \subseteq \mathrm{null}(X). \tag{2.36}$$

Remark 2.9. We note that, in particular, condition (2.36) always holds when $D = I$, which implies that lasso penalized logistic regression and lasso penalized Poisson regression always have solutions.

2.4.4 Implicit Form of Solutions

Fix an arbitrary $\lambda > 0$, and let $(\hat{\beta}, \hat{\gamma})$ denote an optimal solution-subgradient pair, i.e., satisfying (2.20), (2.21). As before, we define the boundary set and boundary signs in terms of $\hat{\gamma}$,

$$\mathcal{B} = \big\{i \in \{1, \ldots, m\} : |\hat{\gamma}_i| = 1\big\}, \quad \text{and} \quad s = \mathrm{sign}(\hat{\gamma}_{\mathcal{B}}).$$

and the active set and active signs in terms of $\hat{\beta}$,

$$\mathcal{A} = \mathrm{supp}(D\hat{\beta}) = \big\{i \in \{1, \ldots, m\} : (D\hat{\beta})_i \neq 0\big\}, \quad \text{and} \quad r = \mathrm{sign}(\hat{\gamma}_{\mathcal{A}}).$$

By (2.20), we have that $\mathcal{A} \subseteq \mathcal{B}$. In general, $\mathcal{A}, r, \mathcal{B}, s$ are not unique, as neither $\hat{\beta}$ nor $\hat{\gamma}$ are.

The next lemma gives an implicit form for the fit and solutions in (2.18), with G as in (2.19), akin to the results (2.8), (2.9) in the squared loss case. Its proof stems directly from the KKT conditions (2.29); it is somewhat technical and deferred until the supplementary material.

Lemma 2.17. *Fix any y, X, D, and $\lambda > 0$. Assume that G is of the form* (2.19)*, where ψ is essentially smooth and essentially strictly convex, and satisfies* (2.31)*,* (2.32)*. Let $\hat{\beta}$ be a solution in problem* (2.18)*, and let $\hat{\gamma}$ be a corresponding optimal subgradient, with boundary set and boundary signs $\mathcal{B}, s$. Define the affine subspace*

$$K_{\mathcal{B},s} = \lambda(P_{\mathrm{null}(D_{-\mathcal{B}})}X^T)^+ D_{\mathcal{B}}^T s + \mathrm{null}(P_{\mathrm{null}(D_{-\mathcal{B}})}X^T). \tag{2.37}$$

Then the unique fit can be expressed as

$$X\hat{\beta} = \nabla\psi^*\Big(P^{\psi^*}_{y-K_{\mathcal{B},s}}\big(\nabla\psi(0)\big)\Big), \tag{2.38}$$

and the solution can be expressed as

$$\hat{\beta} = (XP_{\mathrm{null}(D_{-\mathcal{B}})})^+\nabla\psi^*\Big(P^{\psi^*}_{y-K_{\mathcal{B},s}}\big(\nabla\psi(0)\big)\Big) + b, \tag{2.39}$$

for some $b \in \mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}})$. Similarly, letting $\mathcal{A}, r$ denote the active set and active signs of $\hat{\beta}$, the same expressions hold as in the last two displays with $\mathcal{B}, s$ replaced by $\mathcal{A}, r$ (i.e., with the affine subspace of interest now being $K_{\mathcal{A},r} = \lambda(P_{\mathrm{null}(D_{-\mathcal{A}})}X^T)^+ D_{\mathcal{A}}^T r + \mathrm{null}(P_{\mathrm{null}(D_{-\mathcal{A}})}X^T)$).

Remark 2.10. The proof of Lemma 2.17 derives the representation (2.38) using technical manipulation of the KKT conditions. But the same result can be derived using the geometry surrounding the dual problem (2.26). See Figure 2.2 for an accompanying illustration, and Remark 2.2 for a similar geometric argument in the squared loss case. As $\hat{\gamma}$ has boundary set and signs $\mathcal{B}, s$, and $\hat{u} = \lambda\hat{\gamma}$ from (2.23), we see that $\hat{u}$ must lie on the face of $B_\infty^m(\lambda)$ whose affine span is $E_{\mathcal{B},s} = \{u \in \mathbb{R}^m : u_{\mathcal{B},s} = \lambda s\}$; and as $X^T\hat{v} = D^T\hat{u}$, we see that $\hat{v}$ lies on the face of C whose affine span is $K_{\mathcal{B},s} = (X^T)^{-1}D^T E_{\mathcal{B},s}$, which, it can be checked, can be rewritten explicitly as the affine subspace in (2.37). Hence, the projection of $\nabla G(0)$ onto $-C$ lies on a face whose affine span is $-K_{\mathcal{B},s}$, and we can write

$$-\hat{v} = P^{G^*}_{-K_{\mathcal{B},s}}\big(\nabla G(0)\big),$$

i.e., we can simply replace the set $-C$ in (2.27) with $-K_{\mathcal{B},s}$. When G is of the form (2.19), repeating the same arguments as before therefore shows that the dual and primal projections in (2.30) hold with $-C$ replaced by $-K_{\mathcal{B},s}$, which yields the primal projection result in (2.38) in the lemma.

Though the form of solutions in (2.39) appears more complicated in form than the form (2.9) in the squared loss case, we see that one important property has carried over to the general loss setting, namely, the property that $b \in \mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}})$. As before, let us assign to each boundary set $\mathcal{B}$ associated with an optimal subgradient in (2.18) a basis matrix $U(\mathcal{B}) \in \mathbb{R}^{p \times k(\mathcal{B})}$, whose linearly independent columns that span $\mathrm{null}(D_{-\mathcal{B}})$. Then by the same logic as explained at the beginning of Section 2.3.1, we see that, under the conditions of Lemma 2.17, there is a unique solution in (2.18), given by (2.39) with $b = 0$, provided that conditions (2.10), (2.11) hold.

The arguments in the squared loss case, proceeding the observation of (2.10), (2.11) as a sufficient condition, relied on the invariance of the linear subspace $X\mathrm{null}(D_{-\mathcal{B}})$ over all boundary sets $\mathcal{B}$ of optimal subgradients in the generalized lasso problem (2.1). This key result was established, recall, in Lemma 10 of Tibshirani and Taylor [137], transcribed in our Lemma 2.3 for convenience. For the general loss setting, no such invariance result exists (as far as we know). Thus, with uniqueness in mind as the end goal, we take somewhat of a detour and study local properties of generalized lasso solutions, and invariance of the relevant linear subspaces, over the next two subsections.

2.4.5 Local Stability

We establish a result on the local stability of the boundary set and boundary signs $\mathcal{B}, s$ associated with an optimal solution-subgradient pair $(\hat{\beta}, \hat{\gamma})$, i.e., satisfying (2.20), (2.21). This is a generalization of Lemma 9 in Tibshirani and Taylor [137], which gives the analogous result for the case of squared loss. We must first introduce some notation. For arbitrary subsets $\mathcal{A} \subseteq \mathcal{B} \subseteq \{1, \ldots, m\}$, denote

$$M_{\mathcal{A},\mathcal{B}} = P_{[D_{\mathcal{B}\setminus\mathcal{A}}(\mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}}))]^\perp} D_{\mathcal{B}\setminus\mathcal{A}}(XP_{\mathrm{null}(D_{-\mathcal{B}})})^+. \tag{2.40}$$

(By convention, when $\mathcal{A} = \mathcal{B}$, we set $M_{\mathcal{A},\mathcal{B}} = 0$.) Define

$$\mathcal{N} = \bigcup_{\substack{\mathcal{A},\mathcal{B},s:\\ M_{\mathcal{A},\mathcal{B}} \neq 0}} \Big(K_{\mathcal{B},s} + \nabla\psi\big(\mathrm{col}(XP_{\mathrm{null}(D_{-\mathcal{B}})}) \cap \mathrm{null}(M_{\mathcal{A},\mathcal{B}})\big)\Big). \tag{2.41}$$

The union above is taken over all subsets $A \subseteq \mathcal{B} \subseteq \{1, \ldots, m\}$ and vectors $s \in \{-1, 1\}^{|\mathcal{B}|}$, such that $M_{\mathcal{A},\mathcal{B}} \neq 0$; and $K_{\mathcal{B},s}, M_{\mathcal{A},\mathcal{B}}$, are as defined in (2.37), (2.40), respectively. We use somewhat of an abuse in notation in writing $\nabla\psi(\mathrm{col}(XP_{\mathrm{null}(D_{-\mathcal{B}})}) \cap \mathrm{null}(M_{\mathcal{A},\mathcal{B}}))$; for an arbitrary triplet $(\mathcal{A}, \mathcal{B}, s)$, of course, $\mathrm{col}(XP_{\mathrm{null}(D_{-\mathcal{B}})}) \cap \mathrm{null}(M_{\mathcal{A},\mathcal{B}})$ need not be contained in $\mathrm{int}(\mathrm{dom}(\psi))$, and so really, each such term in the above union should be interpreted as $\nabla\psi(\mathrm{col}(XP_{\mathrm{null}(D_{-\mathcal{B}})}) \cap \mathrm{null}(M_{\mathcal{A},\mathcal{B}}) \cap \mathrm{int}(\mathrm{dom}(\psi)))$.

Next we present the local stability result. Its proof is lengthy and deferred until the supplementary material.

Lemma 2.18. *Fix any X, D, and $\lambda > 0$. Fix $y \notin \mathcal{N}$, where the set $\mathcal{N}$ is defined in* (2.41)*. Assume that G is of the form* (2.19)*, where ψ is essentially smooth and essentially strictly convex, satisfying* (2.31)*,* (2.32)*. That is, our assumptions on the response are succinctly: $y \in \mathcal{N}^c \cap (\mathrm{int}(\mathrm{ran}(\nabla\psi)) + C)$. Denote an optimal solution-subgradient pair in problem* (2.18) *by $(\hat{\beta}(y), \hat{\gamma}(y))$, our notation here emphasizing the dependence on y, and similarly, denote the associated boundary set, boundary signs, active set, and active signs by $\mathcal{B}(y), s(y), \mathcal{A}(y), r(y)$, respectively. There is a neighborhood U of y such that, for any $y' \in U$, problem* (2.18) *has a solution, and in particular, it has an optimal solution-subgradient pair $(\hat{\beta}(y'), \hat{\gamma}(y'))$ with the same boundary set $\mathcal{B}(y') = \mathcal{B}(y)$, boundary signs $s(y') = s(y)$, active set $\mathcal{A}(y') = \mathcal{A}(y)$, and active signs $r(y') = r(y)$.*

Remark 2.11. The set $\mathcal{N}$ defined in (2.41) is bigger than it needs to be; to be precise, the same result as in Lemma 2.18 actually holds with $\mathcal{N}$ replaced by the smaller set

$$\mathcal{N}^* = \bigcup_{\substack{\mathcal{A},\mathcal{B},s:\\ M_{\mathcal{A},\mathcal{B}} \neq 0}} \Big\{ z \in \mathbb{R}^n : \nabla\psi^*\Big(P^{\psi^*}_{z - K_{\mathcal{B},s}}\big(\nabla\psi(0)\big)\Big) \in \mathrm{null}(M_{\mathcal{A},\mathcal{B}}) \Big\}. \tag{2.42}$$

which can be seen from the proof of Lemma 2.18, as can be $\mathcal{N}^* \subseteq \mathcal{N}$. However, the definition of $\mathcal{N}$ in (2.41) is more explicit than that of $\mathcal{N}^*$ in (2.42), so we stick with the former set for simplicity.

Remark 2.12. For each triplet $\mathcal{A}, \mathcal{B}, s$ in the definition (2.41) over which the union is defined, the sets $K_{\mathcal{B},s}$ and $\mathrm{col}(XP_{\mathrm{null}(D_{-\mathcal{B}})}) \cap \mathrm{null}(M_{\mathcal{A},\mathcal{B}})$ both have Lebesgue measure zero, as they are affine spaces of dimension at most $n - 1$. When $\nabla\psi : \mathrm{int}(\mathrm{dom}(\psi)) \to \mathrm{int}(\mathrm{dom}(\psi^*))$ is a C^1 diffeomorphism—this is true when ψ is the cumulant generating function for the Bernoulli or Poisson cases—the image $\nabla\psi(\mathrm{col}(XP_{\mathrm{null}(D_{-\mathcal{B}})}) \cap \mathrm{null}(M_{\mathcal{A},\mathcal{B}}))$ also has Lebesgue measure zero, for each triplet $\mathcal{A}\,\mathcal{B}, s$, and thus $\mathcal{N}$ (being a finite union of measure zero sets) has measure zero.

2.4.6 Invariance of the Linear Space $X\mathrm{null}(D_{-\mathcal{B}})$

We leverage the local stability result from the last subsection to establish an invariance of the linear subspace $X\mathrm{null}(D_{-\mathcal{B}})$ over all choices of boundary sets $\mathcal{B}$ corresponding to an optimal

subgradient in (2.18). This is a generalization of Lemma 10 in problem Tibshirani and Taylor [137], which was transcribed in our Lemma 2.3. The proof is again deferred until the supplementary material.

Lemma 2.19. *Assume the conditions of Lemma 2.18. Then all boundary sets $\mathcal{B}$ associated with optimal subgradients in problem* (2.18) *give rise to the same subspace $X\mathrm{null}(D_{-\mathcal{B}})$, i.e., there is a single linear subspace $L \subseteq \mathbb{R}^n$ such that $L = X\mathrm{null}(D_{-\mathcal{B}})$ for all boundary sets $\mathcal{B}$ of optimal subgradients. Further, $L = X\mathrm{null}(D_{-\mathcal{A}})$ for all active sets $\mathcal{A}$ associated with solutions in* (2.18).

As already mentioned, Lemmas 2.18 and 2.19 extend Lemmas 9 and 10, respectively, of Tibshirani and Taylor [137] to the case of a general loss function G, taking the generalized linear model form in (2.19). This represents a significant advance in our understanding of the local nature of generalized lasso solutions outside of the squared loss case. For example, even for the special case $D = I$, that logistic lasso solutions have locally constant active sets, and that $\mathrm{col}(X_A)$ is invariant to all choices of active set A, provided y is not in an "exceptional set" $\mathcal{N}$, seem to be interesting and important findings. These results could be helpful, e.g., in characterizing the divergence, with respect to y, of the generalized lasso fit in (2.38), an idea that we leave to future work.

2.4.7 Sufficient Conditions for Uniqueness

We are now able to build on the invariance result in Lemma 2.19, just as we did in the squared loss case, to derive our main result on uniqueness in the current general loss setting.

Theorem 2.2. *Fix any X, D, and $\lambda > 0$. Assume that G is of the form* (2.19)*, where ψ is essentially smooth and essentially strictly convex, and satisfies* (2.31)*. Assume:*

(a) $\mathrm{null}(X) \cap \mathrm{null}(D) = \{0\}$*, and X is in D-GP; or*

(b) the entries of X are drawn from a distribution that is absolutely continuous on $\mathbb{R}^{np}$, and $p \leq n$; or

(c) the entries of X are drawn from a distribution that is absolutely continuous on $\mathbb{R}^{np}$, $p > n$, and $\mathrm{nullity}(D) \leq n$*.*

In case (a), the following holds deterministically, and in cases (b) or (c), it holds with almost surely with respect to the distribution of X: for any $y \in \mathcal{N}^c \cap (\mathrm{int}(\mathrm{ran}(\nabla\psi)) + C)$, where $\mathcal{N}$ is as defined in (2.41)*, problem* (2.18) *has a unique solution.*

Proof. Under the conditions of the theorem, Lemma 2.17 shows that any solution in (2.18) must take the form (2.39). As in the arguments in Section 2.3.1, in the squared loss case, we see that (2.10), (2.11) are together sufficient for implying uniqueness of the solution in (2.18). Moreover, Lemma 2.19 implies the linear subspace $L = X\mathrm{null}(D_{-\mathcal{B}})$ is invariant under all choices of boundary sets $\mathcal{B}$ corresponding to optimal subgradients in (2.18); as in the proof of Lemma 2.4 in the squared loss case, such invariance implies that (2.10) is by itself a sufficient condition. Finally, if (2.10) does not hold, then X cannot be in D-GP, which follows by the applying the arguments Lemma 2.5 in the squared loss case to the KKT conditions (2.29). This completes the proof under condition (a). Recall, conditions (b) or (c) simply imply (a) by Lemmas 2.7 and 2.8. □

As explained in Remark 2.12, the set $\mathcal{N}$ in (2.41) has Lebesgue measure zero for G as in (2.19), when $\nabla\psi$ is a C^1 diffeomorphism, which is true, e.g., for ψ the Bernoulli or Poisson cumulant generating function. However, in the case that ψ is the Bernoulli cumulant generating function, and G is the associated negative log-likelihood, it would of course be natural to assume that the entries of $y|X$ follow a Bernoulli distribution, and under this assumption it is not necessarily true that the event $y \in \mathcal{N}$ has zero probability. A similar statement holds for the Poisson case. Thus, it does not seem straightforward to bound the probability that $y \in \mathcal{N}$ in cases of fundamental interest, e.g., when the entries of $y|X$ follow a Bernoulli or Poisson model and G is the associated negative log-likehood, but intuitively $y \in \mathcal{N}$ seems "unlikely" in these cases. A careful analysis is left to future work.

2.5 Discussion

In this chapter, we derived sufficient conditions for the generalized lasso problem (2.1) to have a unique solution, which allow for $p > n$ (in fact, allow for p to be arbitrarily larger than n): as long as the predictors and response jointly follow a continuous distribution, and the null space of the penalty matrix has dimension at most n, our main result in Theorem 2.1 shows that the solution is unique. We have also extended our study to the problem (2.18), where the loss is of generalized linear model form (2.19), and established an analogous (and more general) uniqueness result in Theorem 2.2. Along the way, we have also shown some new results on the local stability of boundary sets and active sets, in Lemma 2.18, and on the invariance of key linear subspaces, in Lemma 2.19, in the generalized linear model case, which may be of interest in their own right.

An interesting direction for future work is to carefully bound the probability that $y \in \mathcal{N}$, where $\mathcal{N}$ is as in (2.41), in some typical generalized linear models like the Bernoulli and Poisson cases. This would give us a more concrete probabillistic statement about uniqueness in such cases, following from Theorem 2.2. Another interesting direction is to inspect the application of Theorems 2.1 and 2.2 to additive trend filtering and varying-coefficient models. Lastly, the local stability result in Lemma 2.18 seems to suggest that a nice expression for the divergence of the fit (2.38), as a function of y, may be possible (furthermore, Lemma 2.19 suggests that this expression should be invariant to the choice of boundary set). This may prove useful for various purposes, e.g., for constructing unbiased risk estimates in penalized generalized linear models.

2.6 Acknowledgements

The authors would like to thank Emmanuel Candes and Kevin Lin for several helpful conversations, that led to the more careful inspection of the existence conditions for logistic and Poisson regression, in Section 2.4.3.

Chapter 3

Early-Stopped Gradient Descent for Least Squares Regression

3.1 Introduction

There is mounting evidence that many simple and popular estimation methods perform a kind of *implicit regularization*, meaning that they appear to produce estimates exhibiting a kind of regularity, even though they do not employ an explicit regularizer. Research interest in implicit regularization is growing, but the foundations of the idea date back at least 30 years in machine learning, where early-stopped gradient descent was found to be effective in training neural networks [94], and at least 40 years in applied mathematics, where the same idea (here known as early-stopped Landweber iterations) was found ill-posed linear inverse problems [128]. After a wave of research on boosting with early stopping [17, 117, 150, 153], more recent work focuses on the regularity properties of particular algorithms for underdetermined problems in matrix factorization, regression, and classification [47, 48, 145]. More broadly, algorithmic regularization plays a key role in training deep neural networks, via batch normalization, dropout, and other techniques.

In this chapter, we focus on early stopping in gradient descent, when applied specifically to least squares regression. This is a basic problem and we are of course not the only authors to consider it; there is now a large literature on this topic (see references above, and more to come when we discuss related work shortly). However, our perspective differs from existing work in a few important ways: first, we study gradient descent in continuous-time (i.e., with infinitesimal step sizes), leading to a path of iterates known as *gradient flow*; second, we examine the regularity properties along *the entire path*, not just its convergence point (as is the focus in most of the work on implicit regularization); and third, we focus on analyzing and comparing the *risk* of gradient flow directly, which is arguably what we care about the most, in many applications.

A strength of the continuous-time perspective is that it facilitates the comparison between early stopping and ℓ_2 regularization. While the connection between these two mechanisms has been studied by many authors (and from many angles), our work provides some of the strongest evidence for this connection to date.

Summary of Contributions. Our contributions in this chapter are as follows.

- We prove that, in finite samples, under very weak assumptions on the data model (and with no assumptions on the feature matrix X), the estimation risk of gradient flow at time t is no more than 1.69 that of ridge regression at tuning parameter $\lambda = 1/t$, for all $t \geq 0$.
- We show that the same result holds for in-sample prediction risk.
- We show that the same result is also true for out-of-sample prediction risk, but now in an average (Bayes) sense, with respect to a spherical prior on the underlying signal β_0.
- For Bayes risk, under optimal tuning, our results on estimation, in-sample prediction, and out-of-sample prediction risks can all be tightened. We prove that the relative risk (measured in any of these three ways) of optimally-tuned gradient flow to optimally-tuned ridge is in between 1 and 1.22.
- We derive exact limiting formulae for the risk of gradient flow, in a Marchenko-Pastur asymptotic model where p/n (the ratio of the feature dimension to sample size) converges to a positive constant. We compare these to known limiting formulae for ridge regression.
- We support our theoretical results with numerical simulations that show the coupling between gradient flow and ridge can be extremely tight in practice (even tighter than suggested by theory).

Related Work. Various authors have made connections between ℓ_2 regularization and the iterates generated by gradient descent (when applied to different loss functions of interest): Friedman and Popescu [40] were among the first make this explicit, and gave supporting numerical experiments, followed by Ramsay [108], who adopted a continuous-time (gradient flow) view, as we do. Yao et al. [150] point out that early stopped gradient descent is a spectral filter, just like ℓ_2 regularization. Subsequent work in nonparametric data models (specifically, reproducing kernel Hilbert space models), studied early-stopped gradient descent from the perspective of risk bounds, where it is shown to perform comparably to explicit ℓ_2 regularization, when each method is optimally tuned [6, 88, 111, 144]. Other works have focused on the bias-variance trade-off in early-stopped gradient boosting [17, 153].

After completing this work, we became aware of the interesting recent paper by Suggala et al. [130], who gave deterministic bounds between gradient flow and ridge regularized estimates, for problems in which the loss function is strongly convex. Their results are very different from ours: they apply to a much wider variety of problem settings (not just least squares problems), and are driven entirely by properties associated with strong convexity; our analysis, specific to least squares regression, is much more precise, and covers the important high-dimensional case (in which the strong convexity assumption is violated).

There is also a lot of related work on theory for ridge regression. Recently, Dobriban and Wager [29] studied ridge regression (and regularized discriminant analysis) in a Marchenko-Pastur asymptotics model, deriving limiting risk expressions, and the precise form of the limiting optimal tuning parameter. Dicker [27] gave a similar asymptotic analysis for ridge, but under a somewhat different problem setup. Hsu et al. [61] established finite-sample concentration bounds for ridge risk. Low-dimensional theory for ridge dates back much further, see Goldenshluger and Tsybakov [45] and others. Lastly, we point out an interesting risk inflation result in that is vaguely

related to ours: Dhillon et al. [26] showed that risk of principal components regression is at most four times that of ridge, under a natural calibration between these two estimator paths (coupling the eigenvalue threshold for the sample covariance matrix with the ridge tuning parameter).

Outline. Here is an outline for the rest of the chapter. Section 3.2 covers preliminary material, on the problem and estimators to be considered. Section 3.3 gives basic results on gradient flow, and its relationship to ridge regression. Section 3.4 derives expressions for the estimation risk and prediction risk of gradient flow and ridge. Section 3.5 presents our main results on relative risk bounds (of gradient flow to ridge). Section 3.6 studies the limiting risk of gradient flow under standard Marchenko-Pastur asymptotics. Section 3.7 presents numerical examples that support our theoretical results, and Section 3.8 concludes with a short discussion.

3.2 Preliminaries

3.2.1 Least Squares, Gradient Flow, and Ridge

Let $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ be a response vector and a matrix of predictors or features, respectively. Consider the standard (linear) least squares problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{2n} \|y - X\beta\|_2^2. \tag{3.1}$$

Consider gradient descent applied to (3.1), with a constant step size $\epsilon > 0$, and initialized at $\beta^{(0)} = 0$, which repeats the iterations

$$\beta^{(k)} = \beta^{(k-1)} + \epsilon \cdot \frac{X^T}{n}(y - X\beta^{(k-1)}), \tag{3.2}$$

for $k = 1, 2, 3, \ldots$. Letting $\epsilon \to 0$, we get a continuous-time ordinary differential equation

$$\dot{\beta}(t) = \frac{X^T}{n}(y - X\beta(t)), \tag{3.3}$$

over time $t \geq 0$, subject to an initial condition $\beta(0) = 0$. We call (3.3) the *gradient flow* differential equation for the least squares problem (3.1).

To see the connection between (3.2) and (3.3), we simply rearrange (3.2) to find that

$$\frac{\beta^{(k)} - \beta^{(k-1)}}{\epsilon} = \frac{X^T}{n}(y - X\beta^{(k-1)}),$$

and setting $\beta(t) = \beta^{(k)}$ at time $t = k\epsilon$, we recognize the left-hand side above as the discrete derivative of $\beta(t)$ at time t, which approaches its continuous-time derivative as $\epsilon \to 0$.

In fact, starting from the differential equation (3.3), we can view gradient descent (3.2) as one of the most basic numerical analysis techniques—the *forward Euler method*—for discretely approximating the solution (3.3).

Now consider the ℓ_2 regularized version of (3.1), called ridge regression [56]:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2, \tag{3.4}$$

where $\lambda > 0$ is a tuning parameter. The explicit ridge solution is

$$\hat{\beta}^{\text{ridge}}(\lambda) = (X^T X + n\lambda I)^{-1} X^T y. \tag{3.5}$$

Though apparently unrelated, the ridge regression solution path and gradient flow path share striking similarities, and their relationship is our central focus.

3.2.2 The Exact Gradient Flow Solution Path

Thanks to our focus on least squares, the gradient flow differential equation in (3.3) is a rather special one: it is a continuous-time linear dynamical system, and has a well-known exact solution.

Lemma 3.1. *Fix a response y and predictor matrix X. Then the gradient flow problem* (3.3)*, subject to $\beta(0) = 0$, admits the exact solution*

$$\hat{\beta}^{\text{gf}}(t) = (X^T X)^+ (I - \exp(-tX^T X/n)) X^T y, \tag{3.6}$$

for all $t \geq 0$. Here A^+ is the Moore-Penrose generalized inverse of a matrix A, and $\exp(A) = I + A + A^2/2! + A^3/3! + \cdots$ is the matrix exponential.

Proof. This can be verified by differentiating (3.6) and using basic properties of the matrix exponential. □

In continuous-time, early stopping corresponds to taking the estimator $\hat{\beta}^{\text{gf}}(t)$ in (3.6) for any finite value of $t \geq 0$, with smaller t leading to greater regularization. We can already see that (3.6), like (3.5), applies a type of shrinkage to the least squares solution; their similarities will become more evident when we express both in spectral form, as we will do shortly in Section 3.3.1.

3.2.3 Discretization Error

In what follows, we will focus on (continuous-time) gradient flow rather than (discrete-time) gradient descent. Standard results from numerical analysis give uniform bounds between discretizations like the forward Euler method (gradient descent) and the differential equation path (gradient flow). In particular, the next result is a direct application of Theorem 212A in Butcher [18].

Lemma 3.2. *For least squares, consider gradient descent* (3.2) *initialized at $\beta^{(0)} = 0$, and gradient flow* (3.6)*, subject to $\beta(0) = 0$. For any step size $\epsilon < 1/s_{\max}$ where $s_{\max}$ is the largest eigenvalue of $X^T X/n$, and any $K \geq 1$,*

$$\max_{k=1,\ldots,k} |\beta^{(k)} - \hat{\beta}^{\text{gf}}(k\epsilon)| \leq \frac{\epsilon \|X^T y\|_2}{2n} (\exp(K\epsilon s_{\max}) - 1).$$

The results to come can therefore be translated to the discrete-time setting, by taking a small enough ϵ and invoking Lemma 3.2, but we omit details for brevity.

3.3 Basic Comparisons

3.3.1 Spectral Shrinkage Comparison

To compare the ridge (3.5) and gradient flow (3.6) paths, it helps to rewrite them in terms of the singular value decomposition of X. Let $X = \sqrt{n}US^{1/2}V^T$ be a singular value decomposition, so that $X^TX/n = VSV^T$ is an eigendecomposition. Then straightforward algebra brings (3.5), (3.6), on the scale of fitted values, to

$$X\hat{\beta}^{\text{ridge}}(\lambda) = US(S + \lambda I)^{-1}U^Ty, \tag{3.7}$$

$$X\hat{\beta}^{\text{gf}}(t) = U(I - \exp(-tS))U^Ty. \tag{3.8}$$

Letting s_i, $i = 1, \ldots, p$ denote the diagonal entries of S, and $u_i \in \mathbb{R}^n$, $i = 1, \ldots, p$ denote the columns of U, we see that (3.7), (3.8) are both linear smoothers (linear functions of y) of the form

$$\sum_{i=1}^{p} g(s_i, \kappa) \cdot u_i u_i^T y,$$

for a spectral shrinkage map $g(\cdot, \kappa) : [0, \infty) \to [0, \infty)$ and parameter κ. This map is $g^{\text{ridge}}(s, \lambda) = s/(s + \lambda)$ for ridge, and $g^{\text{gf}}(s, t) = 1 - \exp(-ts)$ for gradient flow. We see both apply more shrinkage for smaller values of s, i.e., lower-variance directions of X^TX/n, but do so in apparently different ways.

While these shrinkage maps agree at the extreme ends (i.e., set $\lambda = 0$ and $t = \infty$, or set $\lambda = \infty$ and $t = 0$), there is no single parametrization for λ as a function of t, say $\phi(t)$, that equates $g^{\text{ridge}}(\cdot, \phi(t))$ with $g^{\text{gf}}(\cdot, t)$, for all $t \geq 0$. But the parametrization $\phi(t) = 1/t$ gives the two shrinkage maps grossly similar behaviors: see Figure 3.1 for a visualization. Moreover, as we will show later in Sections 3.5–3.7, the two shrinkage maps (under the calibration $\phi(t) = 1/t$) lead to similar risk curves for ridge and gradient flow.

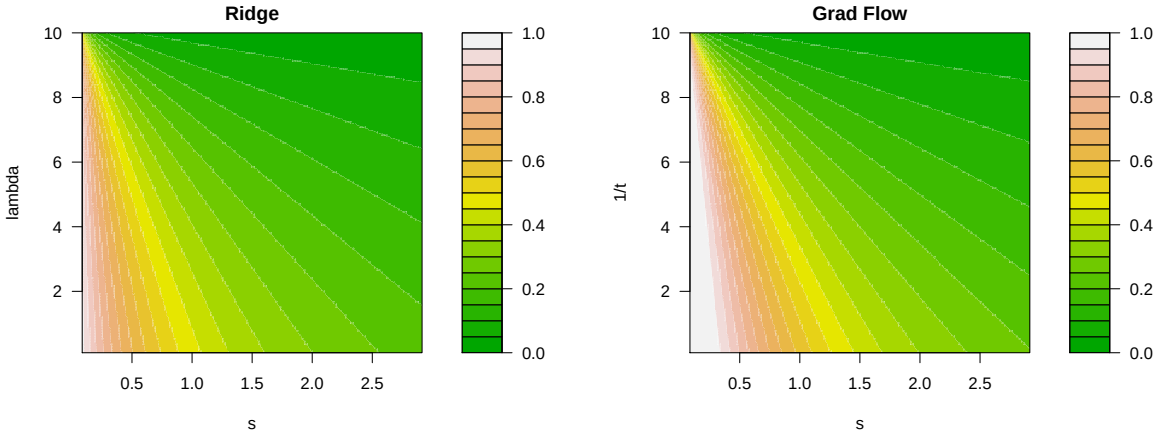

Figure 3.1: *Comparison of ridge and gradient flow spectral shrinkage maps, plotted as heatmaps over (s, λ) (ridge) and (s, t) (gradient flow) with the calibration $\lambda = 1/t$.*

3.3.2 Underlying Regularization Problems

Given our general interest in the connections between gradient descent and ridge regression, it is natural to wonder if gradient descent iterates can also be expressed as solutions to a sequence of regularized least squares problems. The following two simple lemmas certify that this is in fact the case, in both discrete- and continuous-time; their proofs may be found in the supplementary material.

Lemma 3.3. *Fix y, X, and let $X^TX/n = VSV^T$ be an eigendecomposition. Assume that we initialize $\beta^{(0)} = 0$, and we take the step size in gradient descent to satisfy $\epsilon < 1/s_{\max}$, with $s_{\max}$ denoting the largest eigenvalue of X^TX/n. Then, for each $k = 1, 2, 3, \ldots$, the iterate $\beta^{(k)}$ from step k in gradient descent* (3.2) *uniquely solves the optimization problem*

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{n}\|y - X\beta\|_2^2 + \beta^T Q_k \beta,$$

where $Q_k = VS((I - \epsilon S)^{-k} - I)^{-1}V^T$.

Lemma 3.4. *Fix y, X, and let $X^TX/n = VSV^T$ be an eigendecomposition. Under the initial condition $\beta(0) = 0$, for all $t > 0$, the solution $\beta(t)$ of the gradient flow problem* (3.3) *uniquely solves the optimization problem*

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{n}\|y - X\beta\|_2^2 + \beta^T Q_t \beta,$$

where $Q_t = VS(\exp(tS) - I)^{-1}V^T$.

Remark 3.1. The optimization problems that underlie gradient descent and gradient flow, in Lemmas 3.3 and 3.4, respectively, are both quadratically regularized least squares problems. In agreement with the intuition from the last subsection, we see that in both problems the regularizers penalize the lower-variance directions of X^TX/n more strongly, and this is relaxed as t or k grow. The proof of the continuous-time is nearly immediate from (3.8); the proof of the discrete-time result requires a bit more work. To see the link between the two results, set $t = k\epsilon$, and note that as $k \to \infty$:

$$((1 - ts/k)^{-k} - 1)^{-1} \to (\exp(ts) - 1)^{-1}.$$

3.4 Measures of Risk

3.4.1 Estimation Risk

We take the feature matrix $X \in \mathbb{R}^{n \times p}$ to be fixed and arbitrary, and consider a generic response model,

$$y|\beta_0 \sim (X\beta_0, \sigma^2 I), \tag{3.9}$$

which we write to mean $\mathbb{E}(y|\beta_0) = X\beta_0$, $\mathrm{Cov}(y|\beta_0) = \sigma^2 I$, for an underlying coefficient vector $\beta_0 \in \mathbb{R}^p$ and error variance $\sigma^2 > 0$. We consider a spherical prior,

$$\beta_0 \sim (0, (r^2/p)I) \tag{3.10}$$

for some signal strength $r^2 = \mathbb{E}\|\beta_0\|_2^2 > 0$.

For an estimator $\hat{\beta}$ (i.e., measurable function of X, y), we define its estimation risk (or simply, risk) as

$$\mathrm{Risk}(\hat{\beta}; \beta_0) = \mathbb{E}\big[\|\hat{\beta} - \beta_0\|_2^2 \,\big|\, \beta_0\big].$$

We also define its Bayes risk as $\mathrm{Risk}(\hat{\beta}) = \mathbb{E}\|\hat{\beta} - \beta_0\|_2^2$.

Next we give expressions for the risk and Bayes risk of gradient flow; the derivations are straightforward and found in the supplementary material. We denote by s_i, $i = 1, \dots, p$ and v_i, $i = 1, \dots, p$ the eigenvalues and eigenvectors, respectively, of $X^T X/n$.

Lemma 3.5. *Under the data model* (3.9)*, for any $t \geq 0$, the risk of the gradient flow estimator* (3.6) *is*

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{gf}}(t); \beta_0) = \sum_{i=1}^{p} \bigg(|v_i^T \beta_0|^2 \exp(-2ts_i) + \frac{\sigma^2}{n} \frac{(1 - \exp(-ts_i))^2}{s_i} \bigg), \tag{3.11}$$

and under the prior (3.10)*, the Bayes risk is*

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{gf}}(t)) = \frac{\sigma^2}{n} \sum_{i=1}^{p} \bigg(\alpha \exp(-2ts_i) + \frac{(1 - \exp(-ts_i))^2}{s_i} \bigg), \tag{3.12}$$

where $\alpha = r^2 n/(\sigma^2 p)$. Here and henceforth, we take by convention $(1 - e^{-x})^2/x = 0$ when $x = 0$.

Remark 3.2. Compare (3.11) to the risk of ridge regression,

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(\lambda); \beta_0) = \sum_{i=1}^{p} \bigg(|v_i^T \beta_0|^2 \frac{\lambda^2}{(s_i + \lambda)^2} + \frac{\sigma^2}{n} \frac{s_i}{(s_i + \lambda)^2} \bigg). \tag{3.13}$$

and compare (3.12) to the Bayes risk of ridge,

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(\lambda)) = \frac{\sigma^2}{n} \sum_{i=1}^{p} \frac{\alpha\lambda^2 + s_i}{(s_i + \lambda)^2}, \tag{3.14}$$

where $\alpha = r^2 n/(\sigma^2 p)$. These ridge results follow from standard calculations, found in many other papers; for completeness, we give details in the supplementary material.

Remark 3.3. For ridge regression, the Bayes risk (3.14) is minimized at $\lambda^* = 1/\alpha$. There are (at least) two easy proofs of this fact. For the first, we note the Bayes risk of ridge does not depend on the distributions of $y|\beta_0$ and β_0 in (3.9) and (3.10) (just on the first two moments); in the special case that both distributions are normal, we know that $\hat{\beta}^{\mathrm{ridge}}(\lambda^*)$ is the Bayes estimator, which achieves the optimal Bayes risk (hence certainly the lowest Bayes risk over the whole ridge family). For the second proof, following Dicker [27], we rewrite each summand in (3.14) as

$$\frac{\alpha\lambda^2 + s_i}{(s_i + \lambda)^2} = \frac{\alpha}{s_i + \alpha} + \frac{s(\lambda\alpha - 1)^2}{(s_i + \lambda)^2 (s_i + \alpha)},$$

and observe that this is clearly minimized at $\lambda^* = 1/\alpha$.

Remark 3.4. As far as we can tell, deriving the tuning parameter value t^* minimizing the gradient flow Bayes risk (3.12) is difficult. Nevertheless, as we will show in Section 3.5.3, we can still obtain interesting bounds on the optimal risk itself, $\mathrm{Risk}(\hat{\beta}^{\mathrm{gf}}(t^*))$.

3.4.2 Prediction Risk

We now define two predictive notions of risk. Let

$$x_0 \sim (0, \Sigma) \tag{3.15}$$

for a positive semidefinite matrix $\Sigma \in \mathbb{R}^{p\times p}$, and assume x_0 is independent of $y|\beta_0$. We define in-sample prediction risk and out-of-sample prediction risk (or simply, prediction risk) as, respectively,

$$\begin{aligned}
\mathrm{Risk}^{\mathrm{in}}(\hat{\beta};\beta_0) &= \frac{1}{n}\mathbb{E}\big[\|X\hat{\beta} - X\beta_0\|_2^2 \,\big|\, \beta_0\big], \\
\mathrm{Risk}^{\mathrm{out}}(\hat{\beta};\beta_0) &= \mathbb{E}\big[(x_0^T\hat{\beta} - x_0^T\beta_0)^2 \,\big|\, \beta_0\big],
\end{aligned}$$

and their Bayes versions as, respectively, $\mathrm{Risk}^{\mathrm{in}}(\hat{\beta}) = (1/n)\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2$, $\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}) = \mathbb{E}[(x_0^T\hat{\beta} - x_0^T\beta_0)^2]$.

For space reasons, in the remainder, we will focus on out-of-sample prediction risk, and defer detailed discussion of in-sample prediction risk to the supplementary material. The next lemma, proved in the supplement, gives expressions for the prediction risk and Bayes prediction risk of gradient flow. We denote $\hat{\Sigma} = X^TX/n$.

Lemma 3.6. *Under* (3.9)*,* (3.15)*, the prediction risk of the gradient flow estimator* (3.6) *is*

$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{gf}}(t);\beta_0) = \beta_0^T\exp(-t\hat{\Sigma})\Sigma\exp(-t\hat{\Sigma})\beta_0 + \frac{\sigma^2}{n}\mathrm{tr}\big[\hat{\Sigma}^+(I-\exp(-t\hat{\Sigma}))^2\Sigma\big], \tag{3.16}$$

and under (3.10)*, the Bayes prediction risk is*

$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{gf}}(t)) = \frac{\sigma^2}{n}\mathrm{tr}\big[\alpha\exp(-2t\hat{\Sigma})\Sigma + \hat{\Sigma}^+(I-\exp(-t\hat{\Sigma}))^2\Sigma\big]. \tag{3.17}$$

Remark 3.5. Compare (3.16) and (3.17) to the prediction risk and Bayes prediction risk of ridge, respectively,

$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{ridge}}(\lambda);\beta_0) = \lambda^2\beta_0^T(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\beta_0 + \frac{\sigma^2}{n}\mathrm{tr}\big[\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-2}\Sigma\big], \tag{3.18}$$

$$\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{ridge}}(\lambda)) = \frac{\sigma^2}{n}\mathrm{tr}\big[\lambda^2\alpha(\hat{\Sigma}+\lambda I)^{-2}\Sigma + \hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-2}\Sigma\big]. \tag{3.19}$$

These ridge results are standard, and details are given in the supplementary material.

Remark 3.6. The Bayes prediction risk of ridge (3.19) is again minimized at $\lambda^* = 1/\alpha$. This is not at all clear analytically, but it can be established by specializing to a normal-normal likelihood-prior pair, where (for fixed x_0) we know that $x_0^T\hat{\beta}^{\mathrm{ridge}}(\lambda^*)$ is the Bayes estimator for the parameter $x_0^T\beta_0$ (similar to the arguments in Remark 3.3 for the Bayes estimation risk).

3.5 Relative Risk Bounds

3.5.1 Relative Estimation Risk

We start with a simple but key lemma.

Lemma 3.7. *For all $x \geq 0$, we have (a) $e^{-x} \leq 1/(1+x)$ and (b) $1 - e^{-x} \leq 1.2985\, x/(1+x)$.*

Proof. Fact (a) can by shown via Taylor series and (b) by numerically maximizing $x \mapsto (1 - e^{-x})(1+x)/x$. □

A bound on the relative risk of gradient flow to ridge, under the calibration $\lambda = 1/t$, follows immediately.

Theorem 3.1. *Consider the data model* (3.9).

(a) For all $\beta_0 \in \mathbb{R}^p$, and all $t \geq 0$, $\mathrm{Risk}(\hat{\beta}^{\mathrm{gf}}(t); \beta_0) \leq 1.6862 \cdot \mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(1/t); \beta_0)$.

(b) The inequality in part (a) holds for the Bayes risk with respect to any prior on β_0.

(c) The results in parts (a), (b) also hold for in-sample prediction risk.

Proof. For part (a), set $\lambda = 1/t$ and compare the ith summand in (3.11), call it a_i, to that in (3.13), call it b_i. Then

$$\begin{aligned} a_i &= |v_i^T \beta_0|^2 \exp(-2ts_i) + \frac{\sigma^2}{n} \frac{(1 - \exp(-ts_i))^2}{s_i} \\ &\leq |v_i^T \beta_0|^2 \frac{1}{(1+ts_i)^2} + \frac{\sigma^2}{n} 1.2985^2 \frac{t^2 s_i}{(1+ts_i)^2} \\ &\leq 1.6862 \left(|v_i^T \beta_0|^2 \frac{(1/t)^2}{(1/t+s_i)^2} + \frac{\sigma^2}{n} \frac{s_i}{(1/t+s_i)^2} \right) \\ &= 1.6862\, b_i, \end{aligned}$$

where in the second line, we used Lemma 3.7. Summing over $i = 1, \ldots, p$ gives the desired result.

Part (b) follows by taking an expectation on each side of the inequality in part (a). Part (c) follows similarly, with details given in the supplementary material. □

Remark 3.7. For any $t > 0$, gradient flow is in fact a unique Bayes estimator, corresponding to a normal likelihood in (3.9) and normal prior $\beta_0 \sim N(0, (\sigma^2/n) Q_t^{-1})$, where Q_t is as in Lemma 3.4. It is therefore admissible. This means the result in part (a) in the theorem (and part (b), for the same reason) cannot be true for any universal constant strictly less than 1.

3.5.2 Relative Prediction Risk

We extend the two simple inequalities in Lemma 3.7 to matrix exponentials. We use $\preceq$ to denote the Loewner ordering on positive semidefinite matrices, i.e., we use $A \preceq B$ to mean that $B - A$ is positive semidefinite.

Lemma 3.8. *For all $X \succeq 0$, we have (a) $\exp(-2X) \preceq (I+X)^{-2}$ and (b) $X^+(I - \exp(-X))^2 \preceq 1.6862\, X(I+X)^{-2}$.*

Proof. All matrices in question are simultaneously diagonalizable, so the claims reduce to ones about eigenvalues, i.e., reduce to checking that $e^{-2x} \leq 1/(1+x)^2$ and $(1-e^{-x})^2/x \leq 1.6862\, x/(1+x)^2$, for $x \geq 0$, and these follow by manipulating the facts in Lemma 3.7. □

With just a bit more work, we can bound the relative Bayes prediction risk of gradient flow to ridge, again under the calibration $\lambda = 1/t$.

Theorem 3.2. *Consider the data model* (3.9)*, prior* (3.10)*, and (out-of-sample) feature distribution* (3.15)*. For all* $t \geq 0$*,* $\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{gf}}(t)) \leq 1.6862 \cdot \mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{ridge}}(1/t))$.

Proof. Consider the matrices inside the traces in (3.17) and (3.19). Applying Lemma 3.8, we have

$$\begin{aligned}
&\alpha \exp(-2t\hat{\Sigma}) + \hat{\Sigma}^+(I - \exp(-t\hat{\Sigma}))^2 \\
&\quad \preceq \alpha(I + t\hat{\Sigma})^{-2} + 1.6862\, t^2\hat{\Sigma}(I + t\hat{\Sigma})^{-2} \\
&\quad \preceq 1.6862\Big(\alpha(1/t)^2(I/t + \hat{\Sigma})^{-2} + \hat{\Sigma}(I/t + \hat{\Sigma})^{-2}\Big).
\end{aligned}$$

Let A, B be the matrices on the first and last lines in the above display, respectively. As $A \preceq B$ and $\Sigma \succeq 0$, we have $\mathrm{tr}(A\Sigma) \leq \mathrm{tr}(B\Sigma)$, completing the proof. □

Remark 3.8. The Bayes perspective here is critical; the proof breaks down for prediction risk, at an arbitrary fixed β_0, and it is not clear to us whether the result is true for prediction risk in general.

3.5.3 Relative Risks at Optima

We present one more helpful inequality, and defer its proof to the supplementary material (it is more technical than the proofs of Lemmas 3.7 and 3.8, but still straightforward).

Lemma 3.9. *For all* $X \succeq 0$*, it holds that* $\exp(-2X) + X^+(I - \exp(-X))^2 \preceq 1.2147\,(I + X)^{-1}$.

We now have the following result, on the relative Bayes risk (and Bayes prediction risk), of gradient descent to ridge regression, when both are optimally tuned.

Theorem 3.3. *Consider the data model* (3.9)*, prior* (3.10)*, and (out-of-sample) feature distribution* (3.15)*.*

(a) It holds that

$$1 \leq \frac{\inf_{t \geq 0} \mathrm{Risk}(\hat{\beta}^{\mathrm{gf}}(t))}{\inf_{\lambda \geq 0} \mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(\lambda))} \leq 1.2147.$$

(b) The same result as in part (a) holds for both in-sample and out-of-sample prediction risk.

Proof. For part (a), recall from Remark 3.3 that the optimal ridge tuning parameter is $\lambda^* = 1/\alpha$ and further, in the special case of a normal-normal likelihood-prior pair, we know that $\hat{\beta}^{\mathrm{ridge}}(\lambda^*)$ is the Bayes estimator so the Bayes risk of $\hat{\beta}^{\mathrm{gf}}(t)$, for any $t \geq 0$, must be at least that of $\hat{\beta}^{\mathrm{ridge}}(\lambda^*)$. But because these Bayes risks (3.12), (3.14) do not depend on the form of likelihood and prior (only on their first two moments), we know that the same must be true in general, which proves

the lower bound on the risk ratio. For the upper bound, we take $t = \alpha$, and compare the ith summand in (3.12), call it a_i, to that in (3.14), call it b_i. We have

$$\begin{aligned} a_i &= \alpha \exp(-2\alpha s_i) + \frac{(1 - \exp(-\alpha s_i))^2}{s_i} \\ &\leq 1.2147 \frac{\alpha}{1 + \alpha s_i} = 1.2147\, b_i, \end{aligned}$$

where in the second line, we applied Lemma 3.9 (to the case of scalar X). Summing over $i = 1, \ldots, p$ gives the desired result.

Parts (b) follows similarly, with details in the supplementary material. □

3.6 Asymptotic Risk Analysis

3.6.1 Marchenko-Pastur Asymptotics

Notice the Bayes risk for gradient flow (3.12) and ridge regression (3.14) depend only on the predictor matrix X via the eigenvalues of the (uncentered) sample covariance $\hat{\Sigma} = X^T X/n$. Random matrix theory gives us a precise understanding of the behavior of these eigenvalues, in large samples. The following assumptions are standard ones in random matrix theory (e.g., Bai and Silverstein [4]). Given a symmetric matrix $A \in \mathbb{R}^{p \times p}$, recall that its *spectral distribution* is defined as $F_A(x) = (1/p) \sum_{i=1}^p \mathbb{1}(\lambda_i(A) \leq x)$, where $\lambda_i(A)$, $i = 1, \ldots, p$ are the eigenvalues of A, and $\mathbb{1}(\cdot)$ denotes the 0-1 indicator function.

Assumption 1. A1 The predictor matrix satisfies $X = Z\Sigma^{1/2}$, for a random matrix $Z \in \mathbb{R}^{n \times p}$ of i.i.d. entries with zero mean and unit variance, and a deterministic positive semidefinite covariance $\Sigma \in \mathbb{R}^{p \times p}$.

Assumption 2. A2 The sample size n and dimension p both diverge, i.e., $n, p \to \infty$, with $p/n \to \gamma \in (0, \infty)$.

Assumption 3. A3 The spectral measure F_Σ of the predictor covariance Σ converges weakly as $n, p \to \infty$ to some limiting spectral measure H.

Under the above assumptions, the seminal Marchenko-Pastur theorem describes the weak limit of the spectral measure $F_{\hat{\Sigma}}$ of the sample covariance $\hat{\Sigma}$.

Theorem 3.4 (Bai and Silverstein [4], Marchenko and Pastur [90], Silverstein [124])**.** *Assuming 1–3, almost surely, the spectral measure $F_{\hat{\Sigma}}$ of $\hat{\Sigma}$ converges weakly to a law $F_{H,\gamma}$, called the* empirical spectral distribution*, that depends only on H, γ.*

Remark 3.9. In general, a closed form for the empirical spectral distribution $F_{H,\gamma}$ is not known, except in very special cases (e.g., when $\Sigma = I$ for all n, p). However, numerical methods for approximating $F_{H,\gamma}$ have been proposed (see Dobriban [28] and references therein).

3.6.2 Limiting Gradient Flow Risk

The limiting Bayes risk of gradient flow is now immediate from the representation in (3.12).

Theorem 3.5. *Assume 1–3, as well as a data model* (3.9) *and prior* (3.10)*. Then as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, for each $t \geq 0$, the Bayes risk* (3.12) *of gradient flow converges almost surely to*

$$\sigma^2 \gamma \int \left[\alpha_0 \exp(-2ts) + \frac{(1 - \exp(-ts))^2}{s} \right] dF_{H,\gamma}(s), \tag{3.20}$$

where $\alpha_0 = r^2/(\sigma^2\gamma)$, and $F_{H,\gamma}$ is the empirical spectral distribution from Theorem 3.4.

Proof. Note that we can rewrite the Bayes risk in (3.12) as $(\sigma^2 p)/n[\int \alpha h_1(s)\, dF_{\hat{\Sigma}}(s) + \int h_2(s)\, dF_{\hat{\Sigma}}(s)]$, where we let $h_1(s) = \exp(-2ts)$, $h_2(s) = (1 - \exp(-ts))^2/s$. Weak convergence of $F_{\hat{\Sigma}}$ to $F_{H,\gamma}$, from Theorem 3.4, implies $\int h(s)\, dF_{\hat{\Sigma}}(s) \to \int h(s)\, dF_{H,\gamma}(s)$ for all bounded, continuous functions h, which proves the result. □

A similar result is available for the limiting Bayes in-sample prediction risk. Studying the the limiting Bayes (out-of-sample) prediction risk is much more challenging, as (3.17) is not simply a function of eigenvalues of $\hat{\Sigma}$. The proof of the next result, deferred to the supplementary material, relies on a key fact on the Laplace transform of the map $x \mapsto \exp(xA)$, and the asymptotic limit of a certain trace functional involving $\hat{\Sigma}, \Sigma$, from Ledoit and Peche [77].

Theorem 3.6. *Under the conditions of Theorem 3.5, also assume $\mathbb{E}(Z_{ij}^{12}) \leq C_1$, $\|\Sigma\|_2 \leq C_2$, for all n, p and constants $C_1, C_2 > 0$. For each $t \geq 0$, the Bayes prediction risk* (3.17) *of gradient flow converges almost surely to*

$$\sigma^2 \gamma \left[\alpha_0 f(2t) + 2 \int_0^t (f(u) - f(2u))\, du \right], \tag{3.21}$$

where f is the inverse Laplace transform of the function

$$\theta(x) := \frac{1}{\gamma} \left(\frac{1}{1 - \gamma + \gamma x m(F_{H,\gamma})(-x)} - 1 \right),$$

and $m(F_{H,\gamma})$ is the Stieltjes transform of $F_{H,\gamma}$,

$$m(F_{H,\gamma})(z) = \int \frac{1}{u - z}\, dF_{H,\gamma}(u). \tag{3.22}$$

An interesting feature of the results (3.20), (3.21) is that they are asymptotically *exact* (no hidden constants).

3.6.3 Asymptotic Risk Comparisons

Under the conditions of Theorem 3.5, for each $\lambda \geq 0$, the Bayes risk (3.14) of ridge regression converges almost surely to

$$\sigma^2 \gamma \int \frac{\alpha_0 \lambda^2 + s}{(s + \lambda)^2}\, dF_{H,\gamma}. \tag{3.23}$$

This is simply an application of weak convergence of $F_{\hat{\Sigma}}$ to $F_{H,\gamma}$ (as argued the proof of Theorem 3.5), and can also be found in, e.g., Chapter 3 of Tulino and Verdu [138].

The limiting Bayes prediction risk is a more difficult calculation. Denote $\mathbb{C}_- = \{z \in \mathbb{C} : \mathrm{Im}(z) < 0\}$. By Lemma 2 in Ledoit and Peche [77], under the conditions stated in the theorem, for each $z \in \mathbb{C}_-$, we have

$$\lim_{n,p\to\infty} \frac{1}{p}\mathrm{tr}\big[(\hat{\Sigma} + zI)^{-1}\Sigma\big] \to \theta(z). \tag{3.24}$$

It is shown in Dobriban and Wager [29] that, under the conditions of Theorem 3.6, for each $\lambda \geq 0$, the Bayes prediction risk (3.19) of ridge regression converges almost surely to

$$\sigma^2\gamma\big[\theta(\lambda) + \lambda(1-\alpha_0\lambda)\theta'(\lambda)\big], \tag{3.25}$$

where $\theta(\lambda)$ is as defined in (3.24). The calculation (3.19) makes use of the Ledoit-Peche result (3.24), and Vitali's theorem (to assure the convergence of the derivative of the resolvent functional in (3.24)).

It is interesting to compare the limiting Bayes prediction risks (3.25) and (3.21). For concreteness, we can rewrite the latter as

$$\sigma^2\gamma\left[\alpha_0\mathcal{L}^{-1}(\theta)(2t) + 2\int_0^t (\mathcal{L}^{-1}(\theta)(u) - \mathcal{L}^{-1}(\theta)(2u))\,du\right]. \tag{3.26}$$

We see that (3.25) features θ and its derivative, while (3.26) features the inverse Laplace transform $\mathcal{L}^{-1}(\theta)$ and its antiderivative.

In fact, a similar structure can be observed by rewriting the limiting risks (3.23) and (3.20). By simply expanding $s = (s+\lambda) - \lambda$ in the numerator in (3.23), and using the definition of the Stieltjes transform (3.22), the limiting Bayes risk of ridge becomes

$$\sigma^2\gamma\big[m(F_{H,\gamma})(-\lambda) - \lambda(1-\alpha_0\lambda)m(F_{H,\gamma})'(-\lambda)\big]. \tag{3.27}$$

By following arguments similar to the treatment of the variance term in the proof of Theorem 3.6, in Section 6.2.10, the limiting Bayes risk of gradient flow becomes

$$\sigma^2\gamma\left[\alpha_0\mathcal{L}(f_{H,\gamma})(2t) + 2\int_0^t (\mathcal{L}(f_{H,\gamma})(u) - \mathcal{L}(f_{H,\gamma})(2u))\,du\right], \tag{3.28}$$

where $f_{H,\gamma} = dF_{H,\gamma}/ds$ denotes the density of the empirical spectral distribution $F_{H,\gamma}$, and $\mathcal{L}(f_{H,\gamma})$ its Laplace transform. We see (3.27) features $m(F_{H,\lambda})$ and its derivative, and (3.28) features $\mathcal{L}(f_{H,\gamma})$ and its antiderivative. But indeed $\mathcal{L}(\mathcal{L}(f_{H,\gamma}))(\lambda) = m(F_{H,\lambda})(-\lambda)$, since we can (in general) view the Stieltjes transform as an iterated Laplace transform. This creates a symmetric link between (3.27), (3.28) and (3.25), (3.26), where $m(F_{H,\gamma})(-\lambda)$ in the former plays the role of $\theta(\lambda)$ in the latter.

3.7 Numerical Examples

We give numerical evidence for our theoretical results: both our relative risk bounds in Section 3.5, and our asymptotic risk expressions in Section 3.6. We generated features via $X = \Sigma^{1/2}Z$,

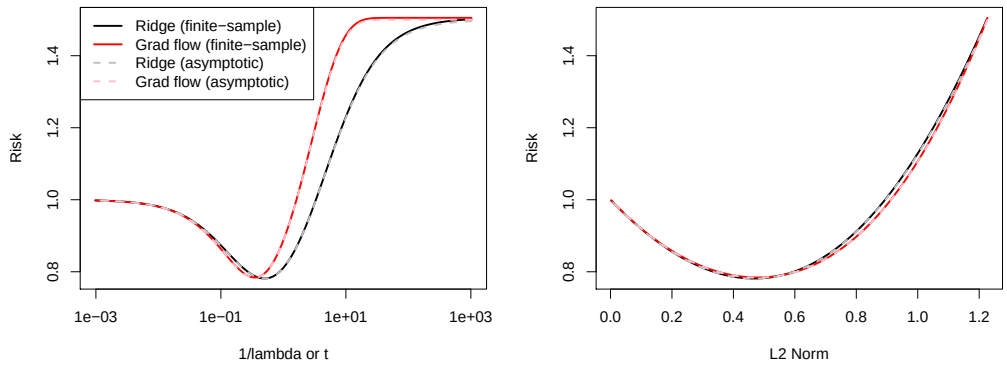

Figure 3.2: *Comparison of Bayes risks for gradient flow and ridge, with Gaussian features,* $\Sigma = I$*,* $n = 500$*,* $p = 1000$*.*

for a matrix Z with i.i.d. entries from a distribution G (with mean zero and unit variance), for three choices of G: standard Gaussian, Student t with 3 degrees of freedom, and Bernoulli with probability 0.5 (the last two distributions were standardized). We took Σ to have all diagonal entries equal to 1 and all off-diagonals equal to $\rho = 0$ (i.e., $\Sigma = I$), or $\rho = 0.5$. For the problem dimensions, we considered $n = 1000$, $p = 500$ and $n = 500$, $p = 1000$. For both gradient flow and ridge, we used a range of 200 tuning parameters equally spaced on the log scale from 2^{-10} to 2^{10}. Lastly, we set $\sigma^2 = r^2 = 1$, where σ^2 is the noise variance in (3.9) and r^2 is the prior radius in (3.10). For each configuration of G, Σ, n, p, we computed the Bayes risk and Bayes prediction risk gradient flow and ridge, as in (3.12), (3.14), (3.17), (3.19). For $\Sigma = I$, the empirical spectral distribution from Theorem 3.4 has a closed form, and so we computed the limiting Bayes risk for gradient flow (3.20) via numerical integration (and similarly for ridge, details in the supplementary material).

Figure 3.2 shows the results for Gaussian features, $\Sigma = I$, $n = 500$, and $p = 1000$; the supplementary material shows results for all other cases (the results are grossly similar). The top plot shows the risk curves when calibrated according to $\lambda = 1/t$ (as per our theory). Here we see fairly strong agreement between the two risk curves, especially around their minimums; the maximum ratio of gradient flow to ridge risks is 1.2164 over the entire path (cf. the upper bound of 1.6862 from Theorem 3.1), and the ratio of the minimums is 1.0036 (cf. the upper bound of 1.2147 from Theorem 3.3). The bottom plot shows the risks when parametrized by the ℓ_2 norms of the underlying estimators. We see remarkable agreement over the whole path, with a maximum ratio of 1.0050. Moreover, in both plots, we can see that the finite-sample (dotted lines) and asymptotic risk curves (solid lines) are identical, meaning that the convergence in Theorem 3.5 is very rapid (and similarly for ridge).

3.8 Discussion

In this work, we studied gradient flow (i.e., gradient descent with infinitesimal step sizes) for least squares, and pointed out a number of connections to ridge regression. We showed that, under minimal assumptions on the data model, and using a calibration $t = 1/\lambda$—where t denotes the time parameter in gradient flow, and λ the tuning parameter in ridge—the risk of gradient flow

is no more than 1.69 times that of ridge, for all $t \geq 0$. We also showed that the same holds for prediction risk, in an average (Bayes) sense, with respect to any spherical prior. Though we did not pursue this, it is clear that these risk couplings could be used to port risk results from the literature on ridge regression (e.g., Dicker [27], Dobriban and Wager [29], Hsu et al. [61], Raskutti et al. [111], etc.) to gradient flow.

Our numerical experiments revealed that calibrating the risk curves by the underlying ℓ_2 norms of the estimators results in a much tighter coupling; developing theory to explain this phenomenon is an important challenge left to future work. Other interesting directions are to analyze the risk of a continuum version of stochastic gradient descent, or to study gradient flow beyond least squares, e.g., for logistic regression.

3.9 Acknowledgements

We thank Veeranjaneyulu Sadhanala, whose insights led us to completely revamp the main results in this chapter.

Chapter 4

The Multiple Quantile Graphical Model

4.1 Introduction

In this chapter, we consider modeling the joint distribution $\mathbb{P}(y_1, \ldots, y_d)$ of d random variables, given n independent draws from this distribution $y^{(1)}, \ldots, y^{(n)} \in \mathbb{R}^d$, where possibly $d \gg n$. Later, we generalize this setup and consider modeling the distribution $\mathbb{P}(y_1, \ldots, y_d | x_1, \ldots, x_p)$, given n independent pairs $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^{p+d}$. Our starting point is the *neighborhood selection* method of [93], which is typically considered in the context of multivariate Gaussian data, and seen as a tool for *covariance selection* [25]: when $\mathbb{P}(y_1, \ldots, y_d)$ is a multivariate Gaussian distribution, it is a well-known fact that y_j and y_k are conditionally independent given the remaining variables if and only if the coefficent corresponding to y_k is zero in the (linear) regression of y_j on all other variables (e.g., [76]). Therefore, in neighborhood selection we compute, for each $k = 1, \ldots, d$, a lasso regression — in order to obtain a small set of conditional dependencies — of y_k on the remaining variables, i.e.,

$$\underset{\theta_k \in \mathbb{R}^d}{\text{minimize}} \sum_{i=1}^{n} \left(y_k^{(i)} - \sum_{j \neq k} \theta_{kj} y_j^{(i)} \right)^2 + \lambda \|\theta_k\|_1, \tag{4.1}$$

for a tuning parameter $\lambda > 0$. This strategy can be seen as a *pseudolikelihood* approximation [12],

$$\mathbb{P}(y_1, \ldots, y_d) \approx \prod_{k=1}^{d} \mathbb{P}(y_k | y_{\neg k}), \tag{4.2}$$

where $y_{\neg k}$ denotes all variables except y_k. Under the multivariate Gaussian model for $\mathbb{P}(y_1, \ldots, y_d)$, the conditional distributions $\mathbb{P}(y_k | y_{\neg k})$, $k = 1, \ldots, d$ here are (univariate) Gaussians, and maximizing the pseudolikelihood in (4.2) is equivalent to separately maximizing the conditionals, as is precisely done in (4.1) (with induced sparsity), for $k = 1, \ldots, d$.

Following the pseudolikelihood-based approach traditionally means carrying out three steps: (i) we write down a suitable family of joint distributions for $\mathbb{P}(y_1, \ldots, y_d)$, (ii) we derive the conditionals $\mathbb{P}(y_k | y_{\neg k})$, $k = 1, \ldots, d$, and then (iii) we maximize each conditional likelihood by (freely) fitting the parameters. Neighborhood selection, and a number of related approaches that came after it (see Section 4.2.1), can be all thought of in this workflow. In many ways,

step (ii) acts as the bottleneck here, and to derive the conditionals, we are usually limited to a homoskedastic and parameteric family for the joint distribution.

The approach we take in this work differs somewhat substantially, as we *begin* by directly modeling the conditionals in (4.2), without any preconceived model for the joint distribution — in this sense, it may be seen a type of *dependency network* [53] for continuous data. We also employ heteroskedastic, nonparametric models for the conditional distributions, which allows us great flexibility in learning these conditional relationships. Our method, called the Multiple Quantile Graphical Model (MQGM), is a marriage of ideas in high-dimensional, nonparametric, multiple quantile regression with those in the dependency network literature (the latter is typically focused on discrete, not continuous, data).

An outline for this chapter is as follows. Section 4.2 reviews background material, and Section 4.3 develops the MQGM estimator. Section 4.4 studies basic properties of the MQGM, and establishes a structure recovery result under appropriate regularity conditions, even for heteroskedastic, non-Gaussian data. Section 4.5 describes an efficient ADMM algorithm for estimation, and Section 4.6 presents empirical examples comparing the MQGM versus common alternatives. Section 4.7 concludes with a discussion.

4.2 Background

4.2.1 Neighborhood Selection and Related Methods

Neighborhood selection has motivated a number of methods for learning sparse graphical models. The literature here is vast; we do not claim to give a complete treatment, but just mention some relevant approaches. Many pseudolikelihood approaches have been proposed, see, e.g., [3, 42, 68, 85, 105, 114]. These works exploit the connection between estimating a sparse inverse covariance matrix and regression, and they vary in terms of the optimization algorithms they use and the theoretical guarantees they offer.

In a related but distinct line of research, [5, 41, 118, 151] proposed ℓ_1-penalized likelihood estimation in the Gaussian graphical model, a method now generally termed the *graphical lasso* (GLasso). Following this, several recent papers have extended the GLasso in various ways. [39] examined a modification based on the multivariate Student t-distribution, for robust graphical modeling. [126, 146, 152] considered conditional distributions of the form $\mathbb{P}(y_1, \ldots, y_d | x_1, \ldots, x_p)$. [78] proposed a model for mixed (both continuous and discrete) data types, generalizing both GLasso and pairwise Markov random fields. [86, 87] used copulas for learning non-Gaussian graphical models.

A strength of neighborhood-based (i.e., pseudolikelihood-based) approaches lies in their simplicity; because they essentially reduce to a collection of univariate probability models, they are in a sense much easier to study outside of the typical homoskedastic, Gaussian data setting. [58, 148, 149] elegantly studied the implications of using univariate exponential family models for the conditionals in (4.2). Closely related to pseudoliklihood approaches are dependency networks [53]. Both frameworks focus on the conditional distributions of one variable given all the rest; the difference lies in whether or not the model for conditionals stems from first specifying some family of joint distributions (pseudolikelihood methods), or not (dependency networks).

Dependency networks have been thoroughly studied for discrete data, e.g., [53, 100]. For continuous data, [140] proposed modeling the mean in a Gaussian neighborhood regression as a nonparametric, additive function of the remaining variables, yielding flexible relationships — this is a type of dependency network for continuous data (though it is not described by the authors in this way). Our method, the MQGM, also deals with continuous data, and is the first to our knowledge that allows for fully nonparametric conditional distributions, as well as nonparametric contributions of the neighborhood variables, in each local model.

4.2.2 Quantile Regression

In linear regression, we estimate the conditional mean of $y|x_1, \ldots, x_p$ based on data samples. Similarly, in α-*quantile regression* [73], we estimate the conditional α-quantile of $y|x_1, \ldots, x_p$, formally $Q_{y|x_1,\ldots,x_p}(\alpha) = \inf\{t : \mathbb{P}(y \leq t|x_1, \ldots, x_p) \geq \alpha\}$, for a given $\alpha \in [0, 1]$, by solving the convex optimization problem:

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{n} \psi_\alpha \left(y^{(i)} - \sum_{j=1}^{q} \theta_j x_j^{(i)} \right),$$

where $\psi_\alpha(z) = \max\{\alpha z, (\alpha - 1)z\}$ is the quantile loss (also referred to as the "pinball" or "tilted absolute" loss). Quantile regression can be useful when the conditional distribution in question is suspected to be heteroskedastic and/or non-Gaussian, e.g., heavy-tailed, or if we wish to understand properties of the distribution other than the mean, e.g., tail behavior. In multiple quantile regression, we solve several quantile regression problems simultaneously, each corresponding to a different quantile level; these problems can be coupled somehow to increase efficiency in estimation (see details in the next section). Again, the literature on quantile regression is quite vast (especially that from econometrics), and we only give a short review here. A standard text is [71]. Nonparametric modeling of quantiles is a natural extension from the (linear) quantile regression approach outlined above; in the univariate case (one conditioning variable), [74] suggested a method using smoothing splines, and [131] described an approach using kernels. More recently, [72] studied the multivariate nonparametric case (more than one conditioning variable), using additive models. In the high-dimensional setting, where p is large, [11, 37, 65] studied ℓ_1-penalized quantile regression and derived estimation and recovery theory for non-(sub-)Gaussian data. We extend results in [37] to prove structure recovery guarantees for the MQGM (in Section 4.4.3).

4.3 The Multiple Quantile Graphical Model

Many choices can be made with regards to the final form of the MQGM, and to help in understanding these options, we break down our presentation in parts. First fix some ordered set $\mathcal{A} = \{\alpha_1, \ldots, \alpha_r\}$ of quantile levels, e.g., $\mathcal{A} = \{0.05, 0.10, \ldots, 0.95\}$. For each variable y_k, and each level α_ℓ, we model the α_ℓ-conditional quantile given the other variables, using an additive

expansion of the form:

$$Q_{y_k|y_{\neg k}}(\alpha_\ell) = b^*_{\ell k} + \sum_{j \neq k}^{d} f^*_{\ell k j}(y_j), \tag{4.3}$$

where $b^*_{\ell k} \in \mathbb{R}$ is an intercept term, and $f^*_{\ell k j}$, $j = 1, \ldots, d$ are smooth, but not parametric in form.

Generic Functional Form of the MQGM. In its most general form, the MQGM estimator is defined as a collection of optimization problems, over $k = 1, \ldots, d$ and $\ell = 1, \ldots, r$:

$$\underset{b_{\ell k},\, f_{\ell k j} \in \mathcal{F}_{\ell k j},\, j=1,\ldots,d}{\text{minimize}} \sum_{i=1}^{n} \psi_{\alpha_\ell}\Big(y_k^{(i)} - b_{\ell k} - \sum_{j \neq k} f_{\ell k j}(y_j^{(i)})\Big) + \sum_{j \neq k} \Big(\lambda_1 P_1(f_{\ell k j}) + \lambda_2 P_2(f_{\ell k j})\Big)^{\omega}. \tag{4.4}$$

Here $\lambda_1, \lambda_2 \geq 0$ are tuning parameters. Also, $\mathcal{F}_{\ell k j}$, $j = 1, \ldots, d$ are spaces of univariate functions, $\omega > 0$ is a fixed exponent, and P_1, P_2 are sparsity and smoothness penalty functions, respectively, all to be decided as part of the modeling process. We give three examples below; several other variants are possible outside of what we describe.

Example 1: Basis Expansion Model. Consider taking $\mathcal{F}_{\ell k j} = \text{span}\{\phi_1^j, \ldots, \phi_m^j\}$, the span of m basis functions, e.g., radial basis functions (RBFs) with centers placed at appropriate locations across the domain of variable j, for each $j = 1, \ldots, d$. This means that each $f_{\ell k j} \in \mathcal{F}_{\ell k j}$ can be expressed as $f_{\ell k j}(x) = \theta_{\ell k j}^T \phi^j(x)$, for a coefficient vector $\theta_{\ell k j} \in \mathbb{R}^m$, where $\phi^j(x) = (\phi_1^j(x), \ldots, \phi_m^j(x))$. Also consider an exponent $\omega = 1$, and the sparsity and smoothness penalties

$$P_1(f_{\ell k j}) = \|\theta_{\ell k j}\|_2 \quad \text{and} \quad P_2(f_{\ell k j}) = \|\theta_{\ell k j}\|_2^2,$$

respectively, which are group lasso and ridge penalties, respectively. With these choices in place, the MQGM problem in (4.4) can be rewritten in finite-dimensional form:

$$\underset{b_{\ell k},\, \theta_{\ell k} = (\theta_{\ell k 1}, \ldots, \theta_{\ell k d})}{\text{minimize}} \psi_{\alpha_\ell}\Big(Y_k - b_{\ell k}\mathbb{1} - \Phi\theta_{\ell k}\Big) + \sum_{j \neq k} \Big(\lambda_1 \|\theta_{\ell k j}\|_2 + \lambda_2 \|\theta_{\ell k j}\|_2^2\Big). \tag{4.5}$$

Above, we used the abbreviation $\psi_{\alpha_\ell}(z) = \sum_{i=1}^n \psi_{\alpha_\ell}(z_i)$ for a vector $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$, and also $Y_k = (y_k^{(1)}, \ldots, y_k^{(n)}) \in \mathbb{R}^n$ for the observations along variable k, $\mathbb{1} = (1, \ldots, 1) \in \mathbb{R}^n$, and $\Phi \in \mathbb{R}^{n \times dm}$ for the basis matrix, with blocks of columns $\Phi_{ij} = \phi(y_j^{(i)})^T \in \mathbb{R}^m$.

The basis expansion model is simple and tends to works well in practice. For the majority of the chapter, we will focus on this model; in principle, everything that follows (methodologically, theoretically, algorithmically) extends to the next two models we describe, as well as many other variants.

Example 2: Smoothing Splines Model. Now consider taking $\mathcal{F}_{\ell k j} = \text{span}\{g_1^j, \ldots, g_n^j\}$, the span of $m = n$ natural cubic splines with knots at $y_j^{(1)}, \ldots, y_j^{(n)}$, for $j = 1, \ldots, d$. As before, we can then write $f_{\ell k j}(x) = \theta_{\ell k j}^T g^j(x)$ with coefficients $\theta_{\ell k j} \in \mathbb{R}^n$, for $f_{\ell k j} \in \mathcal{F}_{\ell k j}$. The work of [92], on high-dimensional additive smoothing splines, suggests a choice of exponent $\omega = 1/2$, and penalties

$$P_1(f_{\ell k j}) = \|G^j \theta_{\ell k j}\|_2^2 \quad \text{and} \quad P_2(f_{\ell k j}) = \theta_{\ell k j}^T \Omega^j \theta_{\ell k j},$$

for sparsity and smoothness, respectively, where $G^j \in \mathbb{R}^{n \times n}$ is a spline basis matrix with entries $G^j_{ii'} = g^j_{i'}(y^{(i)}_j)$, and Ω^j is the smoothing spline penalty matrix containing integrated products of pairs of twice differentiated basis functions. The MQGM problem in (4.4) can be translated into a finite-dimensional form, very similar to what we have done in (4.5), but we omit this for brevity.

Example 3: RKHS Model. Consider taking $\mathcal{F}_{\ell kj} = \mathcal{H}_j$, a univariate reproducing kernel Hilbert space (RKHS), with kernel function $\kappa^j(\cdot, \cdot)$. The representer theorem allows us to express each function $f_{\ell kj} \in \mathcal{H}_j$ in terms of the representers of evaluation, i.e., $f_{\ell kj}(x) = \sum_{i=1}^n (\theta_{\ell kj})_i \kappa^j(x, y^{(i)}_j)$, for a coefficient vector $\theta_{\ell kj} \in \mathbb{R}^n$. The work of [110], on high-dimensional additive RKHS modeling, suggests a choice of exponent $\omega = 1$, and sparsity and smoothness penalties

$$P_1(f_{\ell kj}) = \|K^j \theta_{\ell kj}\|_2 \quad \text{and} \quad P_2(f_{\ell kj}) = \sqrt{\theta^T_{\ell kj} K^j \theta_{\ell kj}},$$

respectively, where $K^j \in \mathbb{R}^{n \times n}$ is the kernel matrix with entries $K^j_{ii'} = \kappa^j(y^{(i)}_j, y^{(i')}_j)$. Again, the MQGM problem in (4.4) can be written in finite-dimensional form, now an SDP, omitted for brevity.

Structural Constraints. Different kinds of structural constraints can be placed on top of the MQGM optimization problem in order to guide the estimated component functions to meet particular shape requirements. An important example are *non-crossing constraints* (commonplace in nonparametric, multiple quantile regression [71, 131]): here, we optimize (4.4) jointly over $\ell = 1, \ldots, r$, subject to

$$b_{\ell k} + \sum_{j \neq k} f_{\ell kj}(y^{(i)}_j) \leq b_{\ell' k} + \sum_{j \neq k} f_{\ell' kj}(y^{(i)}_j), \quad \text{for all } \alpha_\ell < \alpha_{\ell'}, \text{ and } i = 1, \ldots, n. \tag{4.6}$$

This ensures that the estimated quantiles obey the proper ordering, at the observations. For concreteness, we consider the implications for the basis regression model, in Example 1 (similar statements hold for the other two models). For each $\ell = 1, \ldots, r$, denote by $F_{\ell k}(b_{\ell k}, \theta_{\ell k})$ the criterion in (4.5). Introducing the non-crossing constraints requires coupling (4.5) over $\ell = 1, \ldots, r$, so that we now have the following optimization problems, for each target variable $k = 1, \ldots, d$:

$$\underset{B_k, \Theta_k}{\text{minimize}} \sum_{\ell=1}^r F_{\ell k}(b_{\ell k}, \theta_{\ell k}) \quad \text{subject to} \quad (\mathbb{1} B_k^T + \Phi \Theta_k) D^T \geq 0, \tag{4.7}$$

where we denote $B_k = (b_{1k}, \ldots, b_{rk}) \in \mathbb{R}^r$, $\Phi \in \mathbb{R}^{n \times dm}$ the basis matrix as before, $\Theta_k \in \mathbb{R}^{dm \times r}$ given by column-stacking $\theta_{\ell k} \in \mathbb{R}^{dm}$, $\ell = 1, \ldots, r$, and $D \in \mathbb{R}^{(r-1) \times r}$ is the usual discrete difference operator, i.e.,

$$D = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & \ldots & 0 \\ \vdots & & \ddots & \ddots & \\ 0 & 0 & \ldots & -1 & 1 \end{bmatrix}.$$

(The inequality in (4.7) is to be interpreted componentwise.) Computationally, coupling the subproblems across $\ell = 1, \dots, r$ clearly adds to the overall difficulty of the MQGM, but statistically this coupling acts as a regularizer, by constraining the parameter space in a useful way, thus increasing our efficiency in fitting multiple quantile levels from the given data.

For a triplet ℓ, k, j, *monotonicity constraints* are easy to add, i.e., $f_{\ell kj}(y_j^{(i)}) \leq f_{\ell kj}(y_j^{(i')})$ for all $y_j^{(i)} < y_j^{(i')}$. *Convexity constraints*, where we require $f_{\ell kj}$ to be convex over the observations, for a particular ℓ, k, j, are also straightforward. Lastly, *strong non-crossing constraints*, where we enforce (4.6) but over all inputs $z \in \mathbb{R}^d$ (not just over the observations) are also possible with positive basis functions.

Exogenous Variables and Conditional Random Fields. So far, we have considered modeling the joint distribution $\mathbb{P}(y_1, \dots, y_d)$, corresponding to learning a Markov random field (MRF). It is not hard to extend our framework to model the conditional distribution $\mathbb{P}(y_1, \dots, y_d | x_1, \dots, x_p)$ given some exogenous variables $x_1, \dots, x_p$, corresponding to learning a conditional random field (CRF). To extend the basis regression model, we introduce the additional parameters $\theta_{\ell k}^x \in \mathbb{R}^p$ in (4.5), and the loss now becomes $\psi_{\alpha_\ell}(Y_k - b_{\ell k}\mathbb{1}^T - \Phi\theta_{\ell k} - X\theta_{\ell k}^x)$, where $X \in \mathbb{R}^{n \times q}$ is filled with the exogenous observations $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^q$; the other models are changed similarly.

4.4 Basic Properties and Theory

4.4.1 Quantiles and Conditional Independence

In the model (4.3), if a particular variable y_j has no contribution, i.e., satisfied $f_{\ell kj}^* = 0$ across all quantile levels α_ℓ, $\ell = 1, \dots, r$, what does this imply about the conditional independence between y_k and y_j, given the rest? Outside of the multivariate normal model (where the feature transformations need only be linear), nothing can be said in generality. But we argue that conditional independence can be understood in a certain *approximate sense* (i.e., in a projected approximation of the data generating model). We begin with a simple lemma. Its proof is elementary, and given in the supplementary material.

Lemma 4.1. *Let U, V, W be random variables, and suppose that all conditional quantiles of $U|V, W$ do not depend on V, i.e., $Q_{U|V,W}(\alpha) = Q_{U|W}(\alpha)$ for all $\alpha \in [0, 1]$. Then U and V are conditionally independent given W.*

By the lemma, if we knew that $Q_{U|V,W}(\alpha) = h(\alpha, U, W)$ for a function h, then it would follow that U, V are conditionally independent given W (n.b., the converse is true, as well). The MQGM problem in (4.4), with sparsity imposed on the coefficients, essentially aims to achieve such a representation for the conditional quantiles; of course we cannot use a *fully nonparametric* representation of the conditional distribution $y_k | y_{\neg k}$ and instead we use an *r-step approximation* to the conditional cumulative distribution function (CDF) of $y_k | y_{\neg k}$ (corresponding to estimating r conditional quantiles), and (say) in the basis regression model, limit the dependence on conditioning variables to be in terms of an additive function of RBFs in y_j, $j \neq k$. Thus, if at the solution in (4.5) we find that $\hat{\theta}_{\ell kj} = 0$, $\ell = 1, \dots, r$, we may interpret this to mean that y_k and y_j are conditionally independent given the remaining variables, but according to the distribution defined by the *projection* of $y_k | y_{\neg k}$ onto the space of models considered in (4.5) (r-step conditional

CDFs, which are additive expansions in y_j, $j \neq k$). This interpretation is no more tenuous (arguably, less so, as the model space here is much larger) than that needed when applying standard neighborhood selection to non-Gaussian data.

4.4.2 Gibbs Sampling and the "Joint" Distribution

When specifying a form for the conditional distributions in a pseudolikelihood approximation as in (4.2), it is natural to ask: what is the corresponding joint distribution? Unfortunately, for a general collection of conditional distributions, there need not exist a compatible joint distribution, even when all conditionals are continuous [143]. Still, pseudolikelihood approximations (a special case of composite likelihood approximations), possess solid theoretical backing, in that maximizing the pseudolikelihood relates closely to minimizing a certain (expected composite) Kullback-Leibler divergence, measured to the true conditionals [139]. Recently, [23, 149] made nice progress in describing specific conditions on conditional distributions that give rise to a valid joint distribution, though their work was specific to exponential families. A practical answer to the question of this subsection is to use Gibbs sampling, which attempts to draw samples consistent with the fitted conditionals; this is precisely the observation of [53], who show that Gibbs sampling from discrete conditionals converges to a unique stationary distribution, although this distribution may not actually be compatible with the conditionals. The following result establishes the analogous claim for continuous conditionals; its proof is in the supplementary material. We demonstrate the practical value of Gibbs sampling through various examples in Section 4.6.

Lemma 4.2. *Assume that the conditional distributions $\mathbb{P}(y_k|y_{\neg k})$, $k = 1, \ldots, d$ take only positive values on their domain. Then, for any given ordering of the variables, Gibbs sampling converges to a unique stationary distribution that can be reached from any initial point. (This stationary distribution depends on the ordering.)*

4.4.3 Graph Structure Recovery

When $\log d = O(n^{2/21})$, and we assume somewhat standard regularity conditions (listed as A1–A4 in the supplementary material), we will show that the MQGM estimate recovers the underlying conditional independencies with high probability (interpreted in the projected model space, as explained in Section 4.4.1). Importantly, we do not require a Gaussian, sub-Gaussian, or even parametric assumption on the data generating process; instead, we assume i.i.d. draws $y^{(1)}, \ldots, y^{(n)} \in \mathbb{R}^d$, where the conditional distributions $y_k|y_{\neg k}$ have quantiles that are specified by the model in (4.3) for $k = 1, \ldots, d$, $\ell = 1, \ldots, r$, and further, each $f^*_{\ell kj}(x) = \theta^T_{\ell kj}\phi^j(x)^*$ for coefficients $\theta^*_{\ell kj} \in \mathbb{R}^m$, $j = 1, \ldots, d$, as in the basis expansion model.

Let E^* denote the corresponding edge set of conditional dependencies from these neighborhood models, i.e., $\{k, j\} \in E^* \iff \max_{\ell=1,\ldots,r} \max\{\|\theta^*_{\ell kj}\|_2, |\theta^*_{\ell jk}\|_2\} > 0$. We define the estimated edge set $\hat{E}$ in the analogous way, based on the solution in (4.5). Without a loss of generality, we assume the features have been scaled to satisfy $\|\Phi_j\| \leq \sqrt{n}$ for $j = 1, \ldots, dm$. The following is our recovery result; its proof is provided in the supplementary material.

Theorem 4.1. *Assume $\log d = O(n^{2/21})$, and conditions A1–A4 in the supplementary material.*

Assume that the tuning parameters λ_1, λ_2 satisfy

$$\lambda_1 \asymp \sqrt{mn \log(d^2 mr/\delta) \log^3 n} \quad \textit{and} \quad \lambda_2 = o(n^{41/42}/\theta^*_{\max}),$$

where $\theta^*_{\max} = \max_{\ell,k,j} \|\theta^*_{\ell kj}\|_2$. Then for n large enough, the MQGM estimate in (4.5) *exactly recovers the underlying conditional dependencies, i.e., $\hat{E} = E^*$, with probability at least $1 - \delta$.*

The theorem shows that the nonzero pattern in the MQGM estimate identifies, with high probability, the underlying conditional independencies. But to be clear, we emphasize that the MQGM estimate is *not* an estimate of the inverse covariance matrix itself (this is also the case with neighborhood regression, SpaceJam of [140], and many other methods for learning graphical models).

4.5 Computational Approach

By design, the MQGM problem in (4.5) separates into d subproblems, across $k = 1, \ldots, d$ (it therefore suffices to consider only a single subproblem, so we omit notational dependence on k for auxiliary variables). While these subproblems are challenging for off-the-shelf solvers (even for only moderately-sized graphs), the key terms here all admit efficient *proximal operators* [104], which makes operator splitting methods like the alternating direction method of multipliers [15] a natural choice. As an illustration, we consider the non-crossing constraints in the basis regression model below. Reparameterizing so that we may apply ADMM:

$$\begin{array}{ll} \text{minimize}_{\Theta_k, B_k, V, W, Z} & \psi_{\mathcal{A}}(Z) + \lambda_1 \sum_{\ell=1}^r \sum_{j=1}^d \|W_{\ell j}\|_2 + \frac{\lambda_2}{2}\|W\|_F^2 + I_+(VD^T) \\ \text{subject to} & V = \mathbb{1}B_k^T + \Phi\Theta_k,\ W = \Theta_k,\ Z = Y_k\mathbb{1}^T - \mathbb{1}B_k^T - \Phi\Theta_k, \end{array} \tag{4.8}$$

where for brevity $\psi_{\mathcal{A}}(A) = \sum_{\ell=1}^r \sum_{j=1}^d \psi_{\alpha_\ell}(A_{\ell j})$, and $I_+(\cdot)$ is the indicator function of the space of elementwise nonnegative matrices. The augmented Lagrangian associated with (4.8) is:

$$\begin{aligned} L_\rho(\Theta_k, B_k, V, W, Z, U_V, U_W, U_Z) &= \psi_{\mathcal{A}}(Z) + \lambda_1 \sum_{\ell=1}^r \sum_{j=1}^d \|W_{\ell j}\|_2 + \frac{\lambda_2}{2}\|W\|_F^2 + I_+(VD^T) \\ &+ \frac{\rho}{2}\Big(\|\mathbb{1}B_k^T + \Phi\Theta_k - V + U_V\|_F^2 + \|\Theta_k - W + U_W\|_F^2 + \|Y_k\mathbb{1}^T - \mathbb{1}B_k^T - \Phi\Theta_k - Z + U_Z\|_F^2\Big), \end{aligned} \tag{4.9}$$

where $\rho > 0$ is the augmented Lagrangian parameter, and U_V, U_W, U_Z are dual variables corresponding to the equality constraints on V, W, Z, respectively. Minimizing (4.9) over V yields:

$$V \leftarrow P_{\text{iso}}\left(\mathbb{1}B_k^T + \Phi\Theta_k + U_V\right), \tag{4.10}$$

where $P_{\text{iso}}(\cdot)$ denotes the row-wise projection operator onto the isotonic cone (the space of componentwise nondecreasing vectors), an $O(nr)$ operation here [64]. Minimizing (4.9) over $W_{\ell j}$ yields the update:

$$W_{\ell j} \leftarrow \frac{(\Theta_k)_{\ell j} + (U_W)_{\ell j}}{1 + \lambda_2/\rho}\left(1 - \frac{\lambda_1/\rho}{\|(\Theta_k)_{\ell j} + (U_W)_{\ell j}\|_2}\right)_+, \tag{4.11}$$

where $(\cdot)_+$ is the positive part operator. This can be seen by deriving the proximal operator of the function $f(x) = \lambda_1\|x\|_2 + (\lambda_2/2)\|x\|_2^2$. Minimizing (4.9) over Z yields the update:

$$Z \leftarrow \mathbf{prox}_{(1/\rho)\psi_{\mathcal{A}}}(Y_k\mathbb{1}^T - \mathbb{1}b_k^T - \Phi\Theta_k + U_Z), \tag{4.12}$$

where $\mathbf{prox}_f(\cdot)$ denotes the proximal operator of a function f. For the multiple quantile loss function $\psi_{\mathcal{A}}$, this is a kind of generalized soft-thresholding. The proof is given in the supplementary material.

Lemma 4.3. *Let $P_+(\cdot)$ and $P_-(\cdot)$ be the elementwise positive and negative part operators, respectively, and let $a = (\alpha_1, \ldots, \alpha_r)$. Then $\mathbf{prox}_{t\psi_{\mathcal{A}}}(A) = P_+(A - t\mathbb{1}a^T) + P_-(A - t\mathbb{1}a^T)$.*

Finally, differentiation in (4.9) with respect to B_k and Θ_k yields the simultaneous updates:

$$\begin{bmatrix} \Theta_k \\ B_k^T \end{bmatrix} \leftarrow \frac{1}{2}\begin{bmatrix} \Phi^T\Phi + \frac{1}{2}I & \Phi^T\mathbb{1} \\ \mathbb{1}^T\Phi & \mathbb{1}^T\mathbb{1} \end{bmatrix}^{-1} \left([I\ 0]^T (W - U_W) + [\Phi\ \mathbb{1}]^T (Y_k\mathbb{1}^T - Z + U_Z + V - U_V) \right). \tag{4.13}$$

A complete description of our ADMM algorithm for solving the MQGM problem is given in Algorithm 4.1.

Algorithm 4.1 ADMM for the MQGM

Input: observations $y^{(1)}, \ldots, y^{(n)} \in \mathbb{R}^d$, feature matrix $\Phi \in \mathbb{R}^{n \times dm}$, quantile levels $\mathcal{A}$, constants $\lambda_1, \lambda_2 > 0$
Output: fitted coefficients $\hat{\Theta} = (\hat{\theta}_{\ell kj}, \hat{b}_{\ell k})$
for $k = 1, \ldots, d$ (in parallel, if possible) **do**
  initialize $\Theta_k, B_k, V, W, Z, U_V, U_W, U_Z$
  repeat
    update Θ_k using (4.13)
    update B_k using (4.13)
    update V using (4.10)
    update W using (4.11)
    update Z using (4.12) and Lemma 4.3
    update U_V, U_W, U_Z:

$$\begin{aligned} U_V &\leftarrow U_V + (\mathbf{1}B_k^T + \Phi_k\Theta - V) \\ U_W &\leftarrow U_W + (\Theta_k - W) \\ U_Z &\leftarrow U_Z + (Y_k\mathbf{1}^T - \mathbf{1}B_k^T - \Phi_k\Theta - Z) \end{aligned}$$

  until converged
end for

Gibbs Sampling. Having fit the conditionals $y_k|y_{\neg k}$, $k = 1, \ldots d$, we may want to make predictions or extract joint distributions over subsets of variables. As discussed in Section 4.4.2,

there is no general analytic form for these joint distributions, but the pseudolikelihood approximation underlying the MQGM suggests a natural Gibbs sampler. A careful implementation that respects the additive model in (4.3) yields a highly efficient Gibbs sampler, especially for CRFs; the supplementary material gives details.

4.6 Empirical Examples

4.6.1 Synthetic Data

We consider synthetic examples, comparing to neighborhood selection (MB), the graphical lasso (GLasso), SpaceJam [140], the nonparanormal skeptic [87], TIGER [85], and neighborhood selection using the absolute loss (Laplace).

Ring Example. As a simple but telling example, we drew $n = 400$ samples from a "ring" distribution in $d = 4$ dimensions. We used $m = 10$ expanded features and $r = 20$ quantile levels. Data were generated by first drawing a random angle $\nu \sim \text{Uniform}(0, 1)$, then a random radius $R \sim \mathcal{N}(0, 0.1)$, and finally computing the coordinates $y_1 = R\cos\nu$, $y_2 = R\sin\nu$ and $y_3, y_4 \sim \mathcal{N}(0, 1)$, i.e., y_1 and y_2 are the only dependent variables here. Figure 4.1 plots samples (blue) of the coordinates (y_1, y_2) as well as new samples (red) from the MQGM, MB, GLasso, and SpaceJam fitted to these same (blue) samples; the samples from the MQGM, obtained by using our Gibbs sampler (see the supplementary material for further details), appear to closely match the samples from the underlying ring.

Figure 4.2 shows the conditional independencies recovered by the MQGM, MB, GLasso, SpaceJam, the nonparanormal skeptic, TIGER, and Laplace, when run on the ring data. We visualize these independencies by forming a $d \times d$ matrix with the cell (j, k) set to white if j, k are conditionally independent given the others, and black otherwise. Across a range of tuning parameters for each method, the MQGM is the only one that successfully recovers the underlying conditional dependencies.

Table 4.1 presents an evaluation of the conditional CDFs given by the MQGM, MB, GLasso, SpaceJam, TIGER, and Laplace when run on the ring data. For each method, we averaged the total variation distances and Kolmogorov-Smirnoff statistics between the fitted and true conditional CDFs across all variables, and then reported the best values obtained across a range of tuning parameters (further details are given in the supplementary material); the MQGM outperforms all its competitors, in both metrics.

Larger Examples. To investigate performance at larger scales, we drew $n \in \{50, 100, 300\}$ samples from a multivariate normal and Student t-distribution (with 3 degrees of freedom), both in $d = 100$ dimensions, and parameterized by a random, sparse, diagonally dominant $d \times d$ inverse covariance matrix, following the procedure in [3, 68, 102, 105]. Over the same set of sample sizes, with $d = 100$, we also considered an autoregressive setup in which we drew samples of pairs of adjacent variables from the ring distribution. In all three data settings (normal, t, and autoregressive), we used $m = 10$ and $r = 20$ for the MQGM. To summarize the performances, we considered a range of tuning parameters for each method, computed corresponding

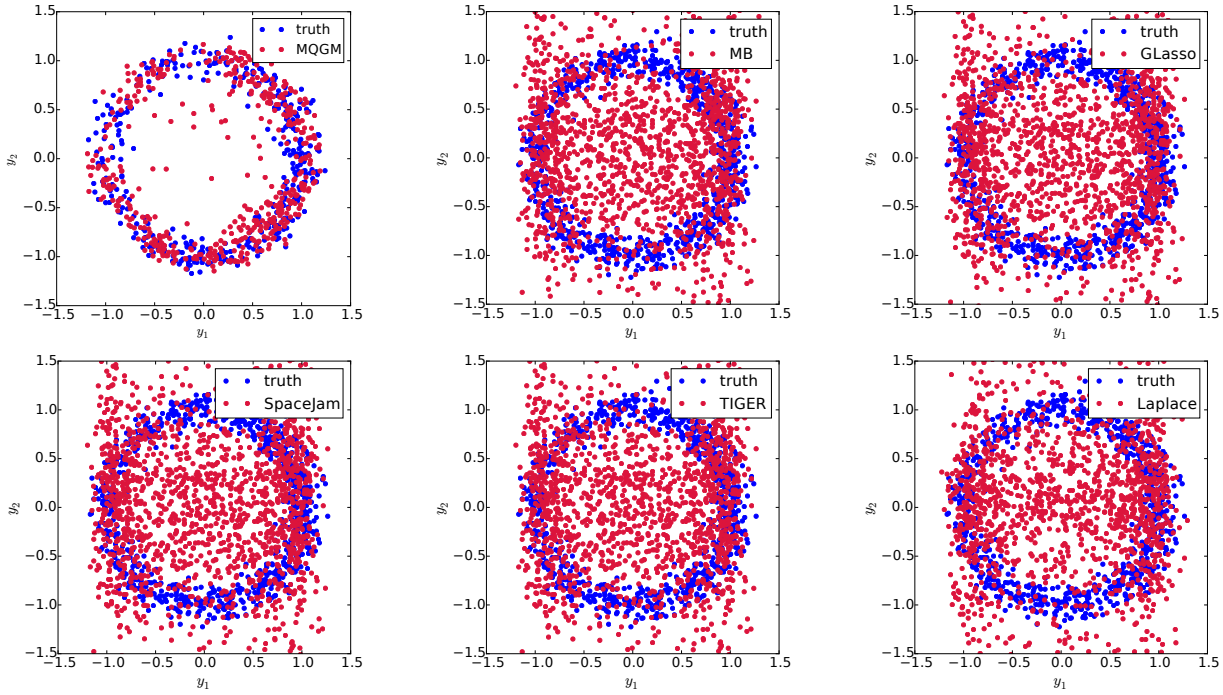

Figure 4.1: *Data from the ring distribution (blue) as well as new samples (red) from the MQGM, MB, GLasso, SpaceJam, TIGER, and Laplace fitted to the same (blue) data; the samples from the MQGM were obtained by using our Gibbs sampler.*

Table 4.1: *Total variation (TV) distance and Kolmogorov-Smirnoff (KS) statistic values for the MQGM, MB, GLasso, SpaceJam, TIGER, and Laplace on the ring data; lower is better, best in* **bold**.

	TV	KS
MQGM	**20.873**	**0.760**
MB	92.298	1.856
GLasso	92.479	1.768
SpaceJam	91.568	1.697
TIGER	88.284	1.450
Laplace	127.406	1.768

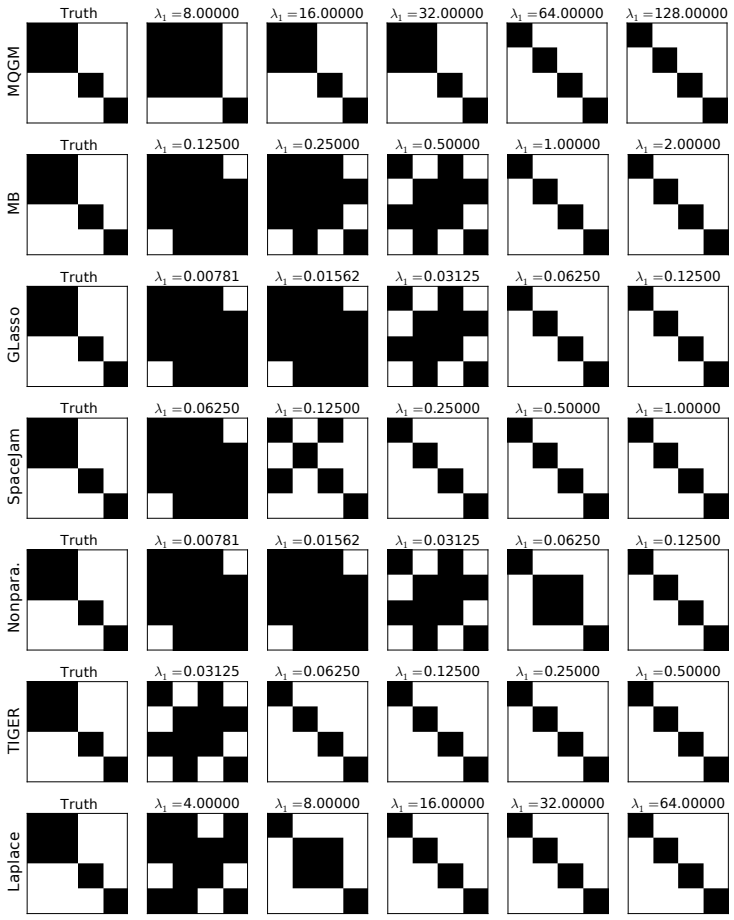

Figure 4.2: *Conditional independencies recovered by the MQGM, MB, GLasso, SpaceJam, the nonparanormal skeptic, TIGER, and Laplace on the ring data; black means conditional dependence. The MQGM is the only method that successfully recovers the underlying conditional dependencies.*

false and true positive rates (in detecting conditional dependencies), and then computed the corresponding area under the curve (AUC), following, e.g., [3, 68, 102, 105]. Table 4.2 reports the median AUCs (across 50 trials) for all three of these examples; the MQGM outperforms all other methods on the autoregressive example, as well as on the small-n normal and Student t examples.

4.6.2 Modeling Flu Epidemics

We study $n = 937$ weekly flu incidence reports from September 28, 1997 through August 30, 2015, across 10 regions in the United States (see the left panel of Figure 4.3), obtained from [21]. We considered $d = 20$ variables: the first 10 encode the current week's flu incidence (precisely, the percentage of doctor's visits in which flu-like symptoms are presented) in the 10 regions, and the last 10 encode the same but for the prior week. We set $m = 5$, $r = 99$, and also introduced exogenous variables to encode the week numbers, so $p = 1$. Thus, learning the MQGM here corresponds to learning the structure of a spatiotemporal graphical model, and reduces to solving

Table 4.2: *AUC values for the MQGM, MB, GLasso, SpaceJam, the nonparanormal skeptic, TIGER, and Laplace for the normal, t, and autoregressive data settings; higher is better, best in* **bold** *(standard errors are* $\approx 10^{-4}$ *or smaller).*

	Normal			Student t			Autoregressive		
	$n = 50$	$n = 100$	$n = 300$	$n = 50$	$n = 100$	$n = 300$	$n = 50$	$n = 100$	$n = 300$
MQGM	**0.953**	**0.976**	0.988	**0.928**	0.947	0.981	**0.726**	**0.754**	**0.955**
MB	0.850	0.959	0.994	0.844	0.923	0.988	0.532	0.563	0.725
GLasso	0.908	0.964	**0.998**	0.691	0.605	0.965	0.541	0.620	0.711
SpaceJam	0.889	0.968	0.997	0.893	**0.965**	0.993	0.624	0.708	0.854
Nonpara.	0.881	0.962	0.996	0.862	0.942	**0.998**	0.545	0.590	0.612
TIGER	0.732	0.921	0.996	0.420	0.873	0.989	0.503	0.518	0.718
Laplace	0.803	0.931	0.989	0.800	0.876	0.991	0.530	0.554	0.758

20 multiple quantile regression subproblems, each of dimension $(19 \times 5 + 1) \times 99 = 9504$. All subproblems took about 1 minute on a 6 core 3.3 Ghz Core i7 X980 processor.

The left panel of Figure 4.4 plots the wallclock time (seconds) for solving one subproblem using ADMM versus SCS [101], a cone solver that has been advocated as a reasonable choice for a class of problems encapsulating (4.4); ADMM outperforms SCS by roughly two orders of magnitude. The right panel of Figure 4.3 presents the conditional independencies recovered by the MQGM. Nonzero entries in the upper left 10×10 submatrix correspond to dependencies between the y_k variables for $k = 1, \ldots, 10$; e.g., the nonzero (0,2) entry suggests that region 1 and 3's flu reports are dependent. The lower right 10×10 submatrix corresponds to the y_k variables for $k = 11, \ldots, 20$, and the nonzero banded entries suggest that at any region the previous week's flu incidence (naturally) influences the next week's. The left panel of Figure 4.3 visualizes these relationships by drawing an edge between dependent regions; region 6 is highly connected, suggesting that it is a bellwether for other regions, which is a qualitative observation also made by the CDC. To draw samples from the fitted distributions, we ran our Gibbs sampler over the year, generating 1000 total samples, making 5 passes over all coordinates between each sample, and with a burn-in period of 100 iterations. The right panel of Figure 4.4 plots samples from the marginal distribution of the percentages of flu reports at region six throughout the year, revealing the heteroskedastic nature of the data; we also see that flu incidence (naturally) increases towards the end of the year. Similarly, Figure 4.5 presents samples from the marginal distributions at other regions (one, five, ten).

4.6.3 Sustainable Energy Application

We evaluate the ability of MQGM to recover the conditional independencies between several wind farms on the basis of large-scale, hourly wind power measurements; wind power is intermittent, and thus understanding the relationships between wind farms can help farm operators plan. We obtained hourly wind power measurements from July 1, 2009 through September 14, 2010 at seven wind farms ($n = 877$, see [3, 59, 146] for details). The primary variables here encode the hourly wind power at a farm over two days (i.e., 48 hours), thus $d = 7 \times 48 = 336$. Exogenous variables were used to encode forecasted wind power and direction as well as other historical measurements, for a total of $q = 3417$. We set $m = 5$ and $r = 20$. Fitting the

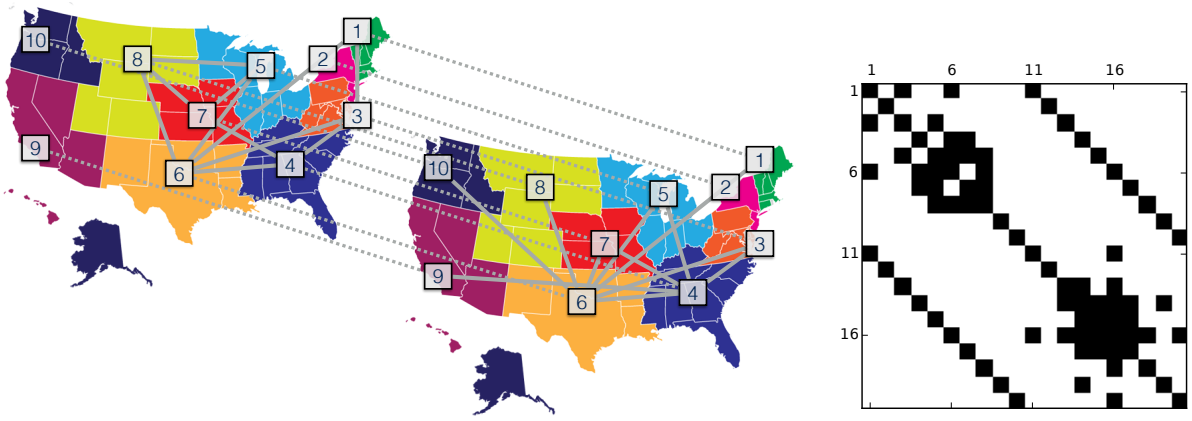

Figure 4.3: *Conditional dependencies recovered by the MQGM on the flu data; each of the first ten cells on the right corresponds to a region of the U.S. on the left, and black means dependence.*

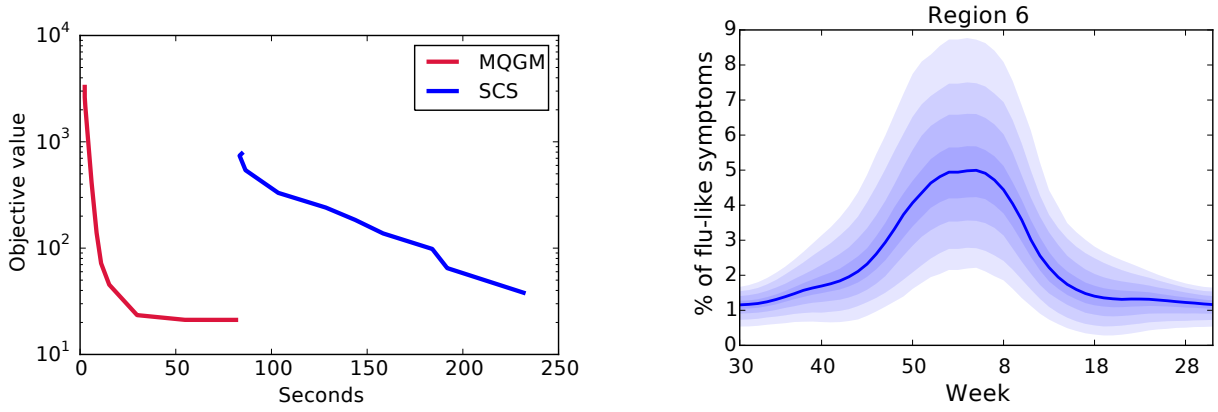

Figure 4.4: *Wallclock time (seconds) for solving one subproblem using ADMM versus SCS, on the left. Samples from the fitted marginal distribution of the weekly flu incidence rates at region six, on the right. Samples at larger quantiles are shaded lighter; the median is in darker blue.*

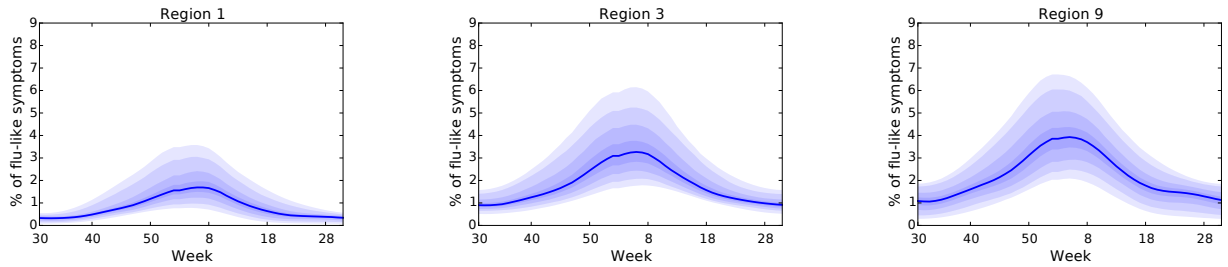

Figure 4.5: *Samples from the fitted marginal distributions of the weekly flu incidence rates at several regions of the U.S.; samples at larger quantile levels shaded lighter, median in darker blue.*

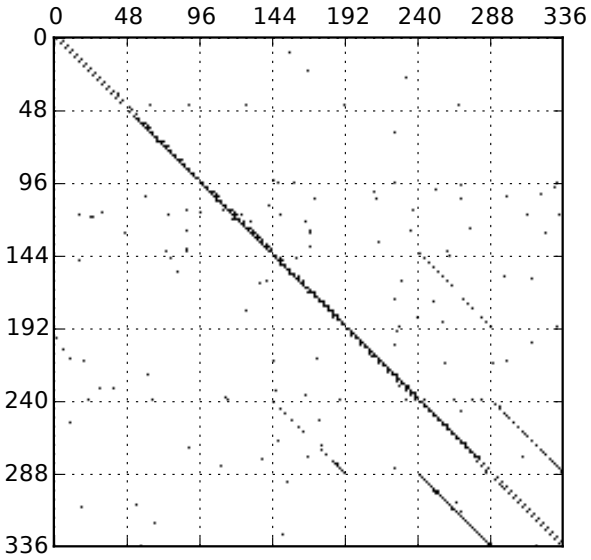

Figure 4.6: *Conditional independencies recovered by the MQGM on the wind farms data; each block corresponds to a wind farm, and black indicates dependence.*

MQGM here hence requires solving $48 \times 7 = 336$ multiple quantile regression subproblems each of dimension $((336-1) \times 5 + 3417) \times 20 = 101,840$. Each subproblem took roughly 87 minutes, comparable to the algorithm of [146].

Figure 4.6 presents the recovered conditional independencies; the nonzero super- and sub-diagonal entries suggest that at any wind farm, the previous hour's wind power (naturally) influences the next hour's, while the nonzero off-diagonal entries, e.g., in the (4,6) block, uncover farms that may influence one another. [146], whose method placed fifth in a Kaggle competition, as well as [3] report similar findings (see the left panels of Figures 7 and 3 in these papers, respectively).

4.7 Discussion

We proposed and studied the Multiple Quantile Graphical Model (MQGM). We established theoretical and empirical backing to the claim that the MQGM is capable of compactly representing relationships between heteroskedastic, non-Gaussian variables. We developed efficient algorithms for estimation and sampling in the MQGM. All in all, we believe that our work represents a step forward in the design of flexible yet tractable graphical models.

Chapter 5

Conclusion

5.1 Discussion

In this thesis, we presented new statistical and computational results for three different user-friendly estimators: the generalized lasso (Chapter 2), early-stopped gradient descent (Chapter 3), and the Multiple Quantile Graphical Model (Chapter 4). Taken together, we hope these results paint a more complete picture of the pros and cons of various methods, and therefore might be of use to practitioners as well as statisticians.

Each of the chapters in this thesis presented some ideas for follow-up work; some high-level directions for future investigation are as follows. First and foremost, as discussed in the introduction, it seems both valuable and worthwhile to give a formal definition of user-friendliness. A first step might be to make precise what it really means for a method to be "interpretable", as there is certainly not widespread agreement on that, at the moment. Perhaps one of the many definitions (see Lipton [84], Murdoch et al. [95], for a survey) that have been floated is more appropriate than the others, for the purpose of constructing a broader definition of user-friendliness. Or, perhaps an entirely new and different definition of interpretability should be proposed. Similarly, although what it means for a method to be "computationally cheap" is relatively unambiguous, the same is not exactly true for "easy-to-implement". Is it enough to characterize how "easy-to-implement" a method is, by simply counting lines of code? Or, should the "difficulty" of the code matter? (And if so, then who/what should assess difficulty?) Finally, one might consider expanding the criteria for what it means for a method to be considered user-friendly. In any event, with a more precise definition of user-friendliness in hand, it might be interesting to formally study the user-friendliness-statistical-computational trade-offs of various methods.

Another line of inquiry that seems interesting, is to study various user-friendly methods for inference (whereas most of the work in this thesis focused on user-friendly methods for estimation); there seem to be many possible avenues for investigation here.

Finally, at a high-level, this thesis is concerned with studying what might be referred to as the "unreasonable effectiveness" of (three) simple methods. Building off the work presented in Chapter 3, it may be that the continuous-time viewpoint is a broadly useful device, perhaps simplifying the statistical analysis of other user-friendly methods; therefore, in what follows, and before concluding the main body of the thesis, we show how to apply some of the continuous-

time tools developed in Chapter 3 to study another user-friendly method: mini-batch stochastic gradient descent.

5.2 Mini-Batch Stochastic Gradient Descent for Least Squares Regression

Consider once again the standard least squares regression problem,

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{2n}\|y - X\beta\|_2^2. \tag{5.1}$$

When the number of samples n is large, running (batch) gradient descent may no longer be feasible due to memory and/or computational constraints. In this case, running mini-batch stochastic gradient descent on (5.1) can be much more convenient. Initializing $\beta^{(0)} = 0$ and using a constant step size $\epsilon > 0$, mini-batch stochastic gradient descent is just given by the recursion

$$\beta^{(k)} = \beta^{(k-1)} + \frac{\epsilon}{m} \cdot X_{\mathcal{S}}^T (y_{\mathcal{S}} - X_{\mathcal{S}} \beta^{(k-1)}), \quad k = 1, 2, 3, \ldots. \tag{5.2}$$

Here, $\mathcal{S} \subseteq \{1, \ldots, n\}$ is an index set, called the *mini-batch*, and is often thought of as being sampled uniformly at random, either with or without replacement, from $\{1, \ldots, n\}$; the notation $X_{\mathcal{S}}, y_{\mathcal{S}}$ extracts the rows indexed by $\mathcal{S}$ from the relevant matrix/vector. Mini-batch stochastic gradient descent is arguably even more user-friendly than (batch) gradient descent: all we need to do, is sample one (or more) of the design points, and then take a step in the negative stochastic gradient direction. This simplicity, in part, helps explain stochastic gradient descent's popularity in practice.

As was the case in Chapter 3, our aim now is to briefly outline how one might go about characterizing the (exact) risk profile of stochastic gradient descent in a relatively simple way, just as before, and relating it to that of ridge regression. However, as we will hint at below, it turns out that the stochastic setup is fundamentally different than the non-stochastic setup. Therefore, this time, in order to make some progress, we will need to make two approximations before proceeding. The first approximation is that the stochastic gradient descent iteration (5.2) may be expressed as the batch gradient descent iteration (3.2), plus Gaussian noise, i.e.,

$$\beta^{(k)} = \beta^{(k-1)} + \epsilon \cdot \frac{X^T}{n}(y - X\beta^{(k-1)}) + \sqrt{\frac{\epsilon}{m}} \cdot \sqrt{\epsilon} \cdot \hat{\Sigma}^{1/2} Z.$$

Here, Z follows a multivariate standard normal distribution. The second approximation is that the covariance matrix of the Gaussian noise in the preceding display is set to the sample covariance matrix, $\hat{\Sigma} = (1/n)X^T X$.

Following these approximations, the last display above can now be seen as the Euler discretization of the following stochastic process,

$$d\beta(t) = -\hat{\Sigma}\beta(t)dt + \frac{1}{n}X^T y dt + \sqrt{\frac{\epsilon}{m}} \cdot \hat{\Sigma}^{1/2} dW(t),$$

under the intial condition that $\beta(0) = 0$, where $W(t)$ denotes the usual Brownian motion. This process has a closed-form solution (which is itself a well-known stochastic process, called the Ornstein-Uhlenbeck process), given by

$$\hat{\beta}^{\text{sgf}}(t) = \hat{\beta}^{\text{gf}}(t) + \sqrt{\frac{\epsilon}{m}} \cdot \exp(-t\hat{\Sigma}) \int_0^t \exp(\tau\hat{\Sigma})\hat{\Sigma}^{1/2} dW(\tau).$$

We call the process $\hat{\beta}^{\text{sgf}}(t)$ *stochastic gradient flow*. Interestingly, the step size ϵ appears in both the stochastic differential equation and its solution (c.f. the non-stochastic setup); it is worth pointing out that prior work [38] has shown it is possible to derive a stochastic differential equation that is free of the step size, by leveraging the theory of diffusion approximations [33, 35, 43, 129].

As a sanity check on the above approximations, in Figure 5.1, we present the out-of-sample predictive risk curves for stochastic gradient flow, as well as for a few other related methods. The plots reveal some interesting phenomena (some of which have been points of discussion in the literature).

- Surveying all the plots, we can see that the continuous-time stochastic gradient flow risk curves appear to reflect the underlying trends present in both the discrete-time stochastic gradient descent as well as the ridge regression risk curves; this seems to imply that there is an implicit regularization effect at work, even with stochastic gradient descent (along its entire optimization path).
- The step size and mini-batch size appear to be linked, in at least a couple of ways. First of all, the risk curves in the upper right panel, where we used a mini-batch size of $m = 50$, are almost identical to the risk curves generated by fixing the mini-batch size at $m = 1$ and decreasing the step size by a factor of 50 (these curves are not shown, to reduce clutter). Second, increasing both the step size as well as the mini-batch size by the same constant factor had virtually no effect on the risk curves (these figures are also not shown); this effect has been pointed out in the literature (e.g., Hoffer et al. [57], Nacson et al. [96], Smith et al. [125]).
- Comparing the first vs. second row of plots, where a step size of 0.0002 vs. 0.0001 was used, we see that when the step size shrinks (and the mini-batch size grows), the risk curves for the iterative algorithms tend towards that of (batch) gradient flow, which serves as a check on the continuous-time modeling approach.
- For all the plots, the mini-batches were formed by sampling with replacement. The risk curves for sampling without replacement were virtually identical (not shown). Formally characterizing the risk of stochastic gradient descent under various popular sampling schemes is an open problem, as far as we can tell (for some relevant work, see HaoChen and Sra [52], Jain et al. [62, 63], Recht and Re [112], Shamir [123]).

It seems important to understand when the previously mentioned approximations are valid. Additionally, it would be valuable to perform formal risk comparisons involving stochastic gradient flow and ridge regression, along the lines of what was done in Chapter 3. In terms of estimation risk, it seems likely that the risk of stochastic gradient flow could be bounded by that of ridge, plus some unavoidable error due to random sampling. It would be interesting to see if the same situation holds for prediction risk.

5.3 Related Work

Before concluding the main body of the thesis, we discuss some work related to the above extension. Stochastic gradient descent has been the object of rather intense study, in both the statistics and optimization communities, over the last few decades. See Nemirovski et al. [98], Polyak and Juditsky [107], Robbins and Monro [113], for examples of earlier work, and Chaturapruek et al. [22], Jain et al. [62], Lin and Rosasco [82], Neu and Rosasco [99], Pillaud-Vivien et al. [106] for a (necessarily abridged) sampling of some of the more recent work, related to our approach here. Therefore, a skeptical reader may (justifiably) wonder what can be gained by adopting the perspective that was just described? We feel there may be several benefits, which we walk through now.

The first broad point to be made, is regarding the decision to focus on least squares regression. Most recent works studying stochastic gradient descent seek to establish error rates for simple modifications (e.g., iterate averaging [62, 99, 106]) to the "vanilla" stochastic gradient descent scheme, showing that these rates are optimal in a certain sense, for a broad class of loss functions. Although valuable, this sort of generality often requires making strong assumptions that are not likely to hold in practice. On the other hand, our goal here is somewhat more direct: we aim to characterize the risk properties of the basic (mini-batch) stochastic gradient descent scheme that is most commonly used in practice, when applied to least squares regression. Focusing on the special case of least squares seems like it would allow us to conduct a more comprehensive study of stochastic gradient descent, without commiting to (many of) the strong assumptions that are currently found in the literature; we elaborate on this, in the next two points.

Looking at the proofs of the results presented in Chapter 3, the main benefits of the continuous-time approach seem to be that it (i) facilitates the risk comparison to ridge regression, and (ii) allows for a more precise characterization of the risk of gradient descent than has been found in previous papers. We might expect the same benefits to transfer over to our study of stochastic gradient descent. To make the point, observe that, by building on the work done in the proof of Lemma 3.3, we may express the estimation risk of the (discrete-time batch) gradient descent iteration as

$$\mathrm{Risk}(\beta^{(k)};\beta_0) = \sum_{i=1}^{p} \left(|v_i^T \beta_0|^2 (1-\epsilon s_i)^{2k} + \frac{\sigma^2}{n} \frac{(1-(1-\epsilon s_i)^k)^2}{s_i} \right).$$

Although the preceding expression is not completely unwieldy, relating the continuous-time risk found in (3.11) to that of ridge in (3.13) seems to be comparatively straightforward (at least, with Lemma 3.7 in hand).

Finally, it appears that the continuous-time perspective may help unify the risk analyses of several popular mini-batching strategies: with replacement, without replacement, and cyclic.

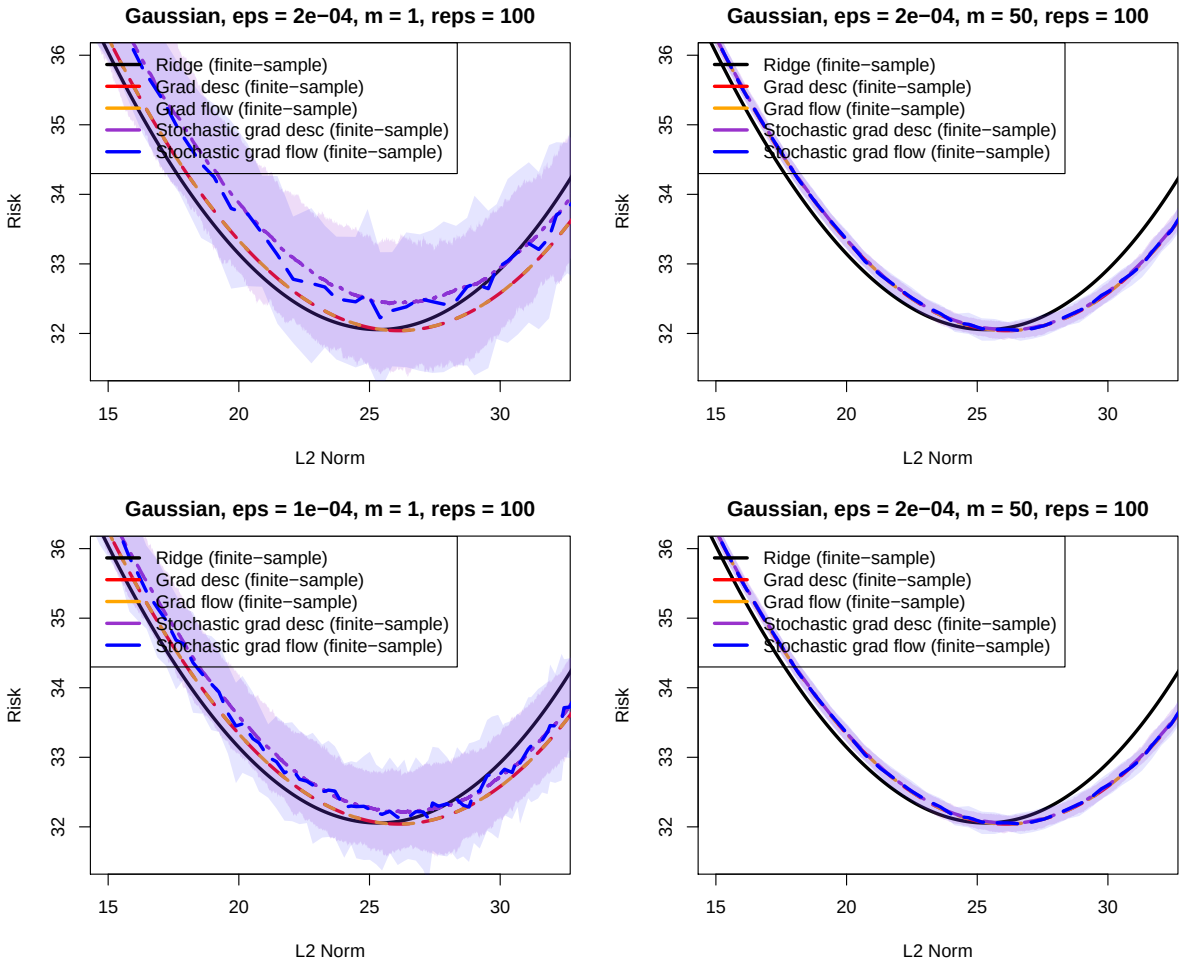

Figure 5.1: *Out-of-sample predictive risk curves for ridge regression, as well as several variants of gradient descent. The risk curves for discrete and continuous-time stochastic gradient descent actually show the average risk over 100 trials, while the shaded bands show the 95th and 5th risk percentiles. We set the step size ϵ to 0.0002 and 0.0001, respectively, in order to generate the plots found in the first and second rows. We set the mini-batch size m to 1 and 50 (the former setting representing pure stochastic gradient descent), respectively, in order to generate the plots found in the first and second columns.*

Chapter 6

Appendix

The supplementary material appearing here and below consists of additional proofs, experiments, and details that support the work presented in the main body of the thesis above.

6.1 Supplementary Material for The Generalized Lasso

6.1.1 Proof of Lemma 2.5

As the generalized lasso solution is not unique, we know that condition (2.10) cannot hold, and there exist $\mathcal{B}, s$ associated with an optimal subgradient in problem (2.1) for which $\mathrm{rank}(XU(\mathcal{B})) < k(\mathcal{B})$, for any $U(\mathcal{B}) \in \mathbb{R}^{p \times k(\mathcal{B})}$ whose linearly independent columns span $\mathrm{null}(D_{-\mathcal{B}})$. Thus, fix an arbitrary choice of basis matrix $U(\mathcal{B})$. Then by construction we have that $Z_i = XU_i(\mathcal{B}) \in \mathbb{R}^n$, $i = 1, \ldots, k(\mathcal{B})$ are linearly dependent.

Note that multiplying both sides of the KKT conditions (2.2) by $U(\mathcal{B})^T$ gives

$$U(\mathcal{B})^T X^T (y - X\hat{\beta}) = \tilde{s}, \tag{6.1}$$

by definition of $\tilde{s}$. We will first show that the assumptions in the lemma, $\tilde{s} \neq 0$. To see this, if $\tilde{s} = 0$, then at any solution $\hat{\beta}$ as in (2.9) associated with $\mathcal{B}, s$,

$$\|D\hat{\beta}\|_1 = \|D_{\mathcal{B}}\hat{\beta}\|_1 = s^T D_{\mathcal{B}}\hat{\beta} = 0,$$

since $\hat{\beta} \in \mathrm{col}(U(\mathcal{B}))$. Uniqueness of the penalty value as in Lemma 2.1 now implies that $\|D\hat{\beta}\|_1 = 0$ at *all* generalized lasso solutions (not only those stemming from $\mathcal{B}, s$). Nonuniqueness of the solution is therefore only possible if $\mathrm{null}(X) \cap \mathrm{null}(D) \neq \{0\}$, contradicting the setup in the lemma.

We may now choose $i_1 \in \{1, \ldots, k(\mathcal{B})\}$ such that $\tilde{s}_{i_1} \neq 0$, and $i_2, \ldots, i_k \in \{1, \ldots, k(\mathcal{B})\}$ such that $k \leq n + 1$ and

$$\sum_{j=1}^{k} c_j Z_{i_j} = 0. \tag{6.2}$$

for some $c \neq 0$. Taking an inner product on both sides with the residual $y - X\hat{\beta}$, and invoking the modified KKT conditions (6.1), gives

$$\sum_{j=1}^{k} c_j \tilde{s}_{i_j} = 0. \tag{6.3}$$

There are two cases to consider. If $\tilde{s}_{i_j} = 0$ for all $j = 2, \ldots, k$, then we must have $c_1 = 0$, so from (6.2),

$$\sum_{j=2}^{k} c_j Z_{i_j} = 0. \tag{6.4}$$

If instead $\tilde{s}_{i_j} \neq 0$ for some $j = 2, \ldots, k$, then define $\mathcal{J} = \{j \in \{1, \ldots, k\} : \tilde{s}_{i_j} \neq 0\}$ (which we know in the present case has cardinality $|\mathcal{J}| \geq 2$). Rewrite (6.3) as

$$c_1 \tilde{s}_{i_1} = - \sum_{j \in \mathcal{J} \setminus \{1\}} c_j \tilde{s}_{i_j},$$

and hence rewrite (6.2) as

$$\sum_{j \in \mathcal{J}} c_j \tilde{s}_{i_j} \frac{Z_{i_j}}{\tilde{s}_{i_j}} + \sum_{j \notin \mathcal{J}} c_j Z_{i_j} = 0,$$

or

$$\frac{Z_{i_1}}{\tilde{s}_{i_1}} = \frac{-1}{c_1 \tilde{s}_{i_1}} \sum_{j \in \mathcal{J} \setminus \{1\}} c_j \tilde{s}_{i_j} \frac{Z_{i_j}}{\tilde{s}_{i_j}} + \frac{-1}{c_1 \tilde{s}_{i_1}} \sum_{j \notin \mathcal{J}} c_j Z_{i_j}.$$

or letting $a_{i_j} = -c_j \tilde{s}_{i_j} / (c_1 \tilde{s}_{i_1})$ for $j \in \mathcal{J}$,

$$\frac{Z_{i_1}}{\tilde{s}_{i_1}} = \sum_{j \in \mathcal{J} \setminus \{1\}} a_{i_j} \frac{Z_{i_j}}{\tilde{s}_{i_j}} + \frac{-1}{c_1 \tilde{s}_{i_1}} \sum_{j \notin \mathcal{J}} c_j Z_{i_j}, \quad \text{where} \sum_{j \in \mathcal{J} \setminus \{1\}} a_{i_j} = 1. \tag{6.5}$$

Reflecting on the two conclusions (6.4), (6.5) from the two cases considered, we can reexpress these as (2.12), (2.13), respectively, completing the proof. □

6.1.2 Proof of Lemma 2.7

Fix an arbitrary $\mathcal{B} \subseteq \{1, \ldots, m\}$ and $s \in \{-1, 1\}^{|\mathcal{B}|}$. Define $U(\mathcal{B}) \in \mathbb{R}^{p \times k(\mathcal{B})}$ whose columns form a basis for $\mathrm{null}(D_{-\mathcal{B}})$ by running Gauss-Jordan elimination on $D_{-\mathcal{B}}$. We may assume without a loss of generality that this is of the form

$$U(\mathcal{B}) = \begin{bmatrix} I \\ F \end{bmatrix},$$

where $I \in \mathbb{R}^{k(\mathcal{B}) \times k(\mathcal{B})}$ is the identity matrix and $F \in \mathbb{R}^{(p-k(\mathcal{B})) \times k(\mathcal{B})}$ is a generic dense matrix. (If need be, then we can always permute the columns of X, i.e., relabel the predictor variables, in order to obtain such a form.) This allows us to express the columns of $Z = XU(\mathcal{B})$ as

$$Z_i = \sum_{\ell=1}^{p} X_\ell U_{\ell i}(\mathcal{B}) = X_i + \sum_{\ell=1}^{p-k(\mathcal{B})} X_{\ell + k(\mathcal{B})} F_{\ell i}, \quad \text{for } i = 1, \ldots, k(\mathcal{B}).$$

Importantly, for each $i = 1, \ldots, k(\mathcal{B})$, we see that only Z_i depends on X_i (i.e., no other Z_j, $j \neq i$ depends on X_i). Select any $i_1, \ldots, i_k \in \{1, \ldots, k(\mathcal{B})\}$ with $\tilde{s}_{i_1} \neq 0$ and $k \leq n+1$. Suppose first that $\tilde{s}_{i_2} = \cdots = \tilde{s}_{i_k} = 0$. Then

$$Z_{i_2} \in \mathrm{span}(\{Z_{i_3}, \ldots, Z_{i_k}\}) \iff X_{i_2} \in -\sum_{\ell=1}^{p-k(\mathcal{B})} X_{\ell+k(\mathcal{B})} F_{\ell i} + \mathrm{span}(\{Z_{i_3}, \ldots, Z_{i_k}\}).$$

Conditioning on X_j, $j \neq i_2$, the right-hand side above is just some fixed affine space of dimension at most $n-1$, and so

$$\mathbb{P}\bigg(X_{i_2} \in -\sum_{\ell=1}^{p-k(\mathcal{B})} X_{\ell+k(\mathcal{B})} F_{\ell i} + \mathrm{span}(\{Z_{i_3}, \ldots, Z_{i_k}\}) \,\bigg|\, X_j, j \neq i_2 \bigg) = 0,$$

owing to the fact that $X_{i_2} \,|\, X_j, j \neq i_2$ has a continuous distribution over $\mathbb{R}^n$. Integrating out over X_j, $j \neq i_2$ then gives

$$\mathbb{P}\bigg(X_{i_2} \in -\sum_{\ell=1}^{p-k(\mathcal{B})} X_{\ell+k(\mathcal{B})} F_{\ell i} + \mathrm{span}(\{Z_{i_3}, \ldots, Z_{i_k}\}) \bigg) = 0,$$

which proves a violation of case (i) in the definition of D-GP happens with probability zero. Similar arguments show that a violation of case (ii) in the definition of D-GP happens with probability zero. Taking a union bound over all possible $\mathcal{B}, s, i_1, \ldots, i_k$, and k shows that any violation of the defining properties of the D-GP condition happens with probability zero, completing the proof. □

6.1.3 Proof of Lemma 2.8

Checking that $\mathrm{null}(X) \cap \mathrm{null}(D) = \{0\}$ is equivalent to checking that the matrix

$$M = \begin{bmatrix} X \\ D \end{bmatrix}$$

has linearly independent columns. In the case $p \leq n$, the columns of X will be linearly independent almost surely (the argument for this is similar to the arguments in the proof of Lemma 2.7), so the columns of M will be linearly independent almost surely.

Thus assume $p > n$. Let $q = \mathrm{nullity}(D)$, so $r = \mathrm{rank}(D) = p - q$. Pick r columns of D that are linearly independent; then the corresponding columns of M are linearly independent. It now suffices to check linear independence of the remaining $p - r$ columns of M. But any n columns of X will be linearly independent almost surely (again, the argument for this is similar to the arguments from the proof of Lemma 2.7), so the result is given provided $p - r \leq n$, i.e., $q \leq n$. □

6.1.4 Proof of Corollary 2.2

Let $V = [\, V_1 \;\; V_{-1} \,] \in \mathbb{R}^{n \times n}$ be an orthogonal matrix, where $V_1 = \mathbb{1}/\sqrt{n} \in \mathbb{R}^{n \times 1}$ and $V_{-1} \in \mathbb{R}^{n \times (n-1)}$ has columns that span $\mathrm{col}(M)$. Note that the centered generalized lasso criterion in (2.14) can be written as

$$\frac{1}{2}\|y - MX\beta\|_2^2 + \lambda\|D\beta\|_1 = \frac{1}{2}\|V_1^T y\|_2^2 + \|V_{-1}^T y - V_{-1}^T X\beta\|_2^2 + \lambda\|D\beta\|_1,$$

hence problem (2.14) is equivalent to a regular (uncentered) generalized lasso problem with response $V_{-1}^T y \in \mathbb{R}^{n-1}$ and predictor matrix $V_{-1}^T X \in \mathbb{R}^{(n-1) \times p}$. By straightforward arguments (using integration and change of variables), (X, y) having a density on $\mathbb{R}^{np+n}$ implies that $(V_{-1}^T X, V_{-1}^T y)$ has a density on $\mathbb{R}^{(n-1)p+(n-1)}$. Thus, we can apply Theorem 2.1 to the generalized lasso problem with response $V_{-1}^T y$ and predictor matrix $V_{-1}^T X$ to give the desired result. □

6.1.5 Proof of Lemma 2.9

Let σ^{n-1} denote the $(n-1)$-dimensional spherical measure, which is just a normalized version of the $(n-1)$-dimensional Hausdorff measure $\mathcal{H}^{n-1}$ on the unit sphere $\mathbb{S}^{n-1}$, i.e., defined by

$$\sigma^{n-1}(S) = \frac{\mathcal{H}^{n-1}(S)}{\mathcal{H}^{n-1}(\mathbb{S}^{n-1})}, \quad \text{for } S \subseteq \mathbb{S}^{n-1}. \tag{6.6}$$

Thus, it is sufficient to prove that the distribution of $Z/\|Z\|_2$ is absolutely continuous with respect to σ^{n-1}. For this, it is helpful to recall that an alternative definition of the $(n-1)$-dimensional spherical measure, for an arbitrary $\alpha > 0$, is

$$\sigma^{n-1}(S) = \frac{\mathcal{L}^n(\mathrm{cone}_\alpha(S))}{\mathcal{L}(\mathbb{B}_\alpha^n)}, \quad \text{for } S \subseteq \mathbb{S}^{n-1}. \tag{6.7}$$

where $\mathcal{L}^n$ denotes n-dimensional Lebesgue measure, $\mathbb{B}_\alpha^n = \{x \in \mathbb{R}^n : \|x\|_2 \leq \alpha\}$ is the n-dimensional ball of radius α, and $\mathrm{cone}_\alpha(S) = \{tx : x \in S,\ t \in [0, \alpha]\}$. That (6.7) and (6.6) coincide is due to the fact that any two measures that are uniformly distributed over a separable metric space must be equal up to a positive constant (see Theorem 3.4 in Mattila [91]), and as both (6.7) and (6.6) are probability measures on $\mathbb{S}^{n-1}$, this positive constant must be 1.

Now let $S \subseteq \mathbb{S}^{n-1}$ be a set of null spherical measure, $\sigma^{n-1}(S) = 0$. From the representation for spherical measure in (6.7), we see that $\mathcal{L}^n(\mathrm{cone}_\alpha(S)) = 0$ for any $\alpha > 0$. Denoting $\mathrm{cone}(S) = \{tx : x \in S,\ t \geq 0\}$, we have

$$\mathcal{L}^n(\mathrm{cone}(S)) = \mathcal{L}^n\bigg(\bigcup_{k=1}^\infty \mathrm{cone}_k(S)\bigg) \leq \sum_{k=1}^\infty \mathcal{L}^n(\mathrm{cone}_k(S)) = 0.$$

This means that $\mathbb{P}(Z \in \mathrm{cone}(S)) = 0$, as the distribution of Z is absolutely continuous with respect to $\mathcal{L}^n$, and moreover $\mathbb{P}(Z/\|Z\|_2 \in S) = 0$, since $Z \in \mathrm{cone}(S) \iff Z \in Z/\|Z\|_2 \in S$. This completes the proof. □

6.1.6 Proof of Lemma 2.10

Denote the n-dimensional unit ball by $\mathbb{B}^n = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$. Note that the relative boundary of $\mathbb{B}^n \cap A$ is precisely

$$\mathrm{relbd}(\mathbb{B}^n \cap A) = \mathbb{S}^{n-1} \cap A.$$

The boundary of a convex set has Lebesgue measure zero (see Theorem 1 in Lang [75]), and so we claim $\mathbb{S}^{n-1} \cap A$ has $(n-1)$-dimensional Hausdorff measure zero. To see this, note first that we can assume without a loss of generality that $\dim(A) = n-1$, else the claim follows immediately. We can now interpret $\mathbb{B}^n \cap A$ as a set in the ambient space A, which is diffeomorphic—via a change of basis—to $\mathbb{R}^{n-1}$. To be more precise, if $V \in \mathbb{R}^{n \times (n-1)}$ is a matrix whose columns are orthonormal and span the linear part of A, and $a \in A$ is arbitrary, then $V^T(\mathbb{B}^n \cap A - a) \subseteq \mathbb{R}^{n-1}$ is a convex set, and by the fact cited above its boundary must have $(n-1)$-dimensional Lebesgue measure zero. It can be directly checked that

$$\mathrm{bd}(V^T(\mathbb{B}^n \cap A - a)) = V^T(\mathrm{relbd}(\mathbb{B}^n \cap A) - a) = V^T(\mathbb{S}^{n-1} \cap A - a).$$

As the $(n-1)$-dimensional Lebesgue measure and $(n-1)$-dimensional Hausdorff measure coincide on $\mathbb{R}^{n-1}$, we see that $V^T(\mathbb{S}^{n-1} \cap A - a)$ has $(n-1)$-dimensional Hausdorff measure zero. Lifting this set back to $\mathbb{R}^n$, via the transformation

$$VV^T(\mathbb{S}^{n-1} \cap A - a) + a = \mathbb{S}^{n-1} \cap A,$$

we see that $\mathbb{S}^{n-1} \cap A$ too must have Hausdorff measure zero, the desired result, because the map $x \mapsto Vx + a$ is Lipschitz (then apply, e.g., Theorem 1 in Section 2.4.1 of Evans and Gariepy [36]). □

6.1.7 Proof of Lemma 2.11

Let us abbreviate $\tilde{X} = XW_X^{-1}$ for the scaled predictor matrix, whose columns are $\tilde{X}_i = X_i/\|X_i\|_2$, $i = 1, \ldots, p$. By similar arguments to those given in the proof of Lemma 2.7, to show $\tilde{X}$ is in D-GP almost surely, it suffices to show that for each $i = 1, \ldots, p$,

$$\mathbb{P}\big(\tilde{X}_i \in A \,\big|\, \tilde{X}_j, j \neq i\big) = 0,$$

where $A \subseteq \mathbb{R}^n$ is an affine space depending on $\tilde{X}_j$, $j \neq i$. This follows by applying our previous two lemmas: the distribution of $\tilde{X}_i$ is absolutely continuous with respect $(n-1)$-dimensional Hausdorff measure on $\mathbb{S}^{n-1}$, by Lemma 2.9, and $\mathbb{S}^{n-1} \cap A$ has $(n-1)$-dimensional Hausdorff measure zero, by Lemma 2.10.

To establish that the null space condition $\mathrm{null}(\tilde{X}) \cap \mathrm{null}(D) = \{0\}$ holds almost surely, note that the proof of Lemma 2.8 really only depends on the fact that any collection of k columns of X, for $k \leq n$, are linearly independent almost surely. It can be directly checked that the scaled columns of $\tilde{X}$ share this same property, and thus we can repeat the same arguments as in Lemma 2.8 to give the result. □

6.1.8 Proof of Corollary 2.4

Let $V = [\, V_1 \; V_{-1} \,] \in \mathbb{R}^{n \times n}$ be as in the proof of Corollary 2.2, and rewrite the criterion in (2.17) as

$$\frac{1}{2}\|y - MXW_{MX}^{-1}\beta\|_2^2 + \lambda\|D\beta\|_1 = \frac{1}{2}\|V_1^T y\|_2^2 + \|V_{-1}^T y - V_{-1}^T XW_{MX}^{-1}\beta\|_2^2 + \lambda\|D\beta\|_1.$$

Now for each $i = 1, \ldots, p$, note that $\|V_{-1}^T X_i\|_2^2 = X_i^T V_{-1} V_{-1}^T X_i = \|MX_i\|_2^2$, which means that

$$V_{-1}^T XW_{MX} = V_{-1}^T XW_{V_{-1}^T X}^{-1},$$

precisely the scaled version of $V_{-1}^T X$. From the second to last display, we see that the standardized generalized lasso problem (2.17) is the same as a scaled generalized lasso problem with response $V_{-1}^T y$ and scaled predictor matrix $V_{-1}^T XW_{V_{-1}^T X}^{-1}$. Under the conditions placed on y, X, as explained in the proof of Corollary 2.2, the distribution of $(V_{-1}^T X, V_{-1}^T y)$ is absolutely continuous. Therefore we can apply Corollary 2.3 to give the result. □

6.1.9 Proof of Lemma 2.14

Write $h(\beta) = \lambda\|D\beta\|_1$. We may rewrite problem (2.18) as thus

$$\underset{\beta \in \mathbb{R}^p,\, z \in \mathbb{R}^n}{\text{minimize}} \; G(z) + h(\beta) \quad \text{subject to} \quad z = X\beta. \tag{6.8}$$

The Lagrangian of the above problem is

$$L(\beta, z, v) = G(z) + h(\beta) + v^T(z - X\beta), \tag{6.9}$$

and minimizing the Lagrangian over β, z gives the dual problem

$$\underset{v \in \mathbb{R}^n}{\text{maximize}} \; -G^*(-v) - h^*(X^T v), \tag{6.10}$$

where G^* is the conjugate of G, and h^* is the conjugate of h. Noting that $h(\beta) = \max_{\eta \in D^T B_\infty^m(\lambda)} \eta^T \beta$, we have

$$h^*(\alpha) = I_{D^T B_\infty^m(\lambda)}(\alpha) = \begin{cases} 0 & \alpha \in D^T B_\infty^m(\lambda) \\ \infty & \text{otherwise} \end{cases},$$

and hence the dual problem (6.10) is equivalent to the claimed one (2.22).

As G is essentially smooth and essentially strictly convex, the interior of its domain is nonempty. Since the domain of h is all of $\mathbb{R}^p$, this is enough to ensure that strong duality holds between (6.8) and (6.10) (see, e.g., Theorem 28.2 of Rockafellar [115]). Moreover, if a solution $\hat{\beta}, \hat{z}$ is attained in (6.8), and a solution $\hat{v}$ is attained in (6.10), then by minimizing the Lagrangian $L(\beta, z, \hat{v})$ in (6.9) over z and β, we have the relationships

$$\nabla G(\hat{z}) = -\hat{v}, \quad \text{and} \quad X^T \hat{v} \in \partial h(\hat{\beta}), \tag{6.11}$$

respectively, where $\partial h(\cdot)$ is the subdifferential operator of h. The first relationship in (6.11) can be rewritten as $\nabla G(X\hat{\beta}) = -\hat{v}$, matching the first relationship in (2.23). The second relationship

in (6.11) can be rewritten as $D^T\hat{u} \in \partial h(\hat{\beta})$, where $\hat{u} \in B_\infty^m(\lambda)$ is such that $X^T\hat{v} = D^T\hat{u}$, and thus we can see that $\hat{u}/\lambda$ is simply a relabeling of the subgradient $\hat{\gamma}$ of the ℓ_1 norm evaluated at $D\hat{\beta}$, matching the second relationship in (2.23).

Finally, we address the constraint qualification conditions (2.24), (2.25). When (2.24) holds, we know that G^* has no directions of recession, and so if $C \neq \emptyset$, then the dual problem (2.22) has a solution (see, e.g., Theorems 27.1 and 27.3 in Rockafellar [115]), equivalently, problem (6.10) has a solution. Suppose (2.25) also holds, or equivalently,

$$(-C) \cap \mathrm{int}(\mathrm{dom}(G^*)) \neq \emptyset,$$

which follows as $\mathrm{int}(\mathrm{dom}(G^*)) = \mathrm{int}(\mathrm{ran}(\nabla G))$, due to the fact that the map $\nabla G : \mathrm{int}(\mathrm{dom}(G)) \to \mathrm{int}(\mathrm{dom}(G^*))$ is a homeomorphism. Then we have know further that $-\hat{v} \in \mathrm{int}(\mathrm{dom}(G^*))$ by essential smoothness and essential strict convexity of G^* (in particular, by the property that $\|\nabla G^*\|_2$ diverges along any sequence convering to a boundary point of $\mathrm{dom}(G^*)$; see, e.g., Theorem 3.12 in Bauschke and Borwein [7]), so $\hat{z} = \nabla G^*(-\hat{v})$ is well-defined; by construction it satisfies the first relationship in (6.11), and minimizes the Lagrangian $L(\beta, z, \hat{v})$ over z. The second relationship in (6.11), recall, can be rewritten as $D^T\hat{u} \in \partial h(\hat{\beta})$; that the Lagrangian $L(\beta, z, \hat{v})$ attains its infimum over β follows from the fact that the map $\beta \mapsto h(\beta) - \hat{u}^T D\beta$ has no strict directions of recession (directions of recession in which this map is not constant). We have shown that the Lagrangian $L(\beta, z, \hat{v})$ attains its infimum over β, z. By strong duality, this is enough to ensure that problem (6.8) has a solution, equivalently, that problem (2.18) has a solution, completing the proof.

6.1.10 Proof of Lemma 2.15

When $\lambda = 0$, note that $C = \mathrm{null}(X^T)$, so (2.32) becomes (2.33). For Poisson regression, the condition (2.35) is an immediate rewriting of (2.33), because $\mathrm{int}(\mathrm{ran}(\nabla\psi)) = \mathbb{R}^n_{++}$, where $\mathbb{R}_{++} = (0, \infty)$ denotes the positive real numbers. For logistic regression, the argument leading to (2.34) is a little more tricky, and is given below.

Observe that in the logistic case, $\mathrm{int}(\mathrm{ran}(\nabla\psi)) = (0, 1)^n$, hence condition (2.33) holds if and only if there exists $a \in (0, 1)^n$ such that $X^T(y - a) = 0$, i.e., there exists $a' \in (0, 1)^n$ such that $X^T D_Y a' = 0$, where $D_Y = \mathrm{diag}(Y_1, \ldots, Y_n)$. The latter statement is equivalent to

$$\mathrm{null}(X^T D_Y) \cap \mathbb{R}^n_{++} \neq \emptyset. \tag{6.12}$$

We claim that this is actually in turn equivalent to

$$\mathrm{col}(D_Y X) \cap \mathbb{R}^n_+ = \{0\}. \tag{6.13}$$

where $\mathbb{R}_+ = [0, \infty)$ denotes the nonnegative real numbers, which would complete the proof, as the claimed condition (6.13) is a direct rewriting of (2.34).

Intuitively, to see the equivalence of (6.12) and (6.13), it helps to draw a picture: the two subspaces $\mathrm{col}(D_Y X)$ and $\mathrm{null}(X^T D_Y)$ are orthocomplements, and if the former only intersects the nonnegative orthant at 0, then the latter must pass through the negative orthant. This intuition is formalized by Stiemke's lemma. This is a theorem of alternatives, and a close relative of Farkas' lemma (see, e.g., Theorem 2 in Chapter 1 of Kemp and Kimura [66]); we state it below for reference.

Lemma 6.1. *Given $A \in \mathbb{R}^{n\times p}$, exactly one of the following systems has a solution:*

- $Ax = 0$, $x < 0$ *for some* $x \in \mathbb{R}^p$;
- $A^T y \geq 0$ *for some* $y \in \mathbb{R}^n$, $y \neq 0$.

Applying this lemma to $A = X^T D_Y$ gives the equivalence of (6.12) and (6.13), as desired. □

6.1.11 Proof of Lemma 2.16

We prove the result for the logistic case; the result for the Poisson case follows similarly. Recall that in the logistic case, $\mathrm{int}(\mathrm{ran}(\nabla\psi)) = (0,1)^n$. Given $y \in \{0,1\}^n$, and arbitrarily small $\epsilon > 0$, note that we can always write $y = z + \delta$, where $z \in (0,1)^n$ and $\delta \in B_\infty^m(\epsilon)$. Thus (2.32) holds as long as

$$C = (X^T)^{-1}\big(D^T B_\infty^m(\lambda)\big) = \big\{u \in \mathbb{R}^n : X^T u = D^T v,\ v \in B_\infty^m(\lambda)\big\}$$

contains a ℓ_∞ ball of arbitrarily small radius centered at the origin. As $\lambda > 0$, this holds provided $\mathrm{row}(X) \subseteq \mathrm{row}(D)$, i.e., $\mathrm{null}(D) \subseteq \mathrm{null}(X)$, as claimed. □

6.1.12 Proof of Lemma 2.17

We first establish (2.38), (2.39). Multiplying both sides of stationarity condition (2.29) by $P_{\mathrm{null}(D_{-\mathcal{B}})}$ yields

$$P_{\mathrm{null}(D_{-\mathcal{B}})} X^T\big(y - \nabla\psi(X\hat\beta)\big) = \lambda P_{\mathrm{null}(D_{-\mathcal{B}})} D_{\mathcal{B}}^T s.$$

Let us abbreviate $M = P_{\mathrm{null}(D_{-\mathcal{B}})} X^T$. After rearranging, the above becomes

$$M\nabla\psi(X\hat\beta) = M(y - \lambda M^+ P_{\mathrm{null}(D_{-\mathcal{B}})} D_{\mathcal{B}}^T s).$$

where we have used $P_{\mathrm{null}(D_{-\mathcal{B}})} D_{\mathcal{B}}^T s = M M^+ P_{\mathrm{null}(D_{-\mathcal{B}})} D_{\mathcal{B}}^T s$, which holds as $P_{\mathrm{null}(D_{-\mathcal{B}})} D_{\mathcal{B}}^T s \in \mathrm{col}(M)$, from the second to last display. Moreover, we can simplify the above, using $M^+ P_{\mathrm{null}(D_{-\mathcal{B}})} = M^+$, to yield

$$M\nabla\psi(X\hat\beta) = M(y - \lambda M^+ D_{\mathcal{B}}^T s),$$

and multiplying both sides by M^+,

$$P_{\mathrm{row}(M)}\nabla\psi(X\hat\beta) = P_{\mathrm{row}(M)}(y - \lambda M^+ D_{\mathcal{B}}^T s). \tag{6.14}$$

Lastly, by virtue of the fact that $D_{-\mathcal{B}}\hat\beta = 0$, we have $X\hat\beta = X P_{\mathrm{null}(D_{-\mathcal{B}})}\hat\beta = M^T\hat\beta \in \mathrm{row}(M)$, so

$$P_{\mathrm{null}(M)} X\hat\beta = 0. \tag{6.15}$$

We will now show that (6.14), (6.15) together imply $\nabla\psi(X\hat\beta)$ can be expressed in terms of a certain Bregman projection onto an affine subspace, with respect to ψ^*. To this end, consider

$$\hat x = P_S^f(a) = \operatorname*{arg\,min}_{x\in S}\Big(f(x) - f(a) - \langle\nabla f(a), x - a\rangle\Big),$$

for a function f, point a, and set S. The first-order optimality conditions are

$$\big\langle\nabla f(\hat x) - \nabla f(a), z - \hat x\big\rangle \geq 0 \ \text{ for all } z \in S, \quad \text{and} \quad \hat x \in S.$$

When S is an affine subspace, i.e., $S = c + L$ for a point c and linear subspace L, this reduces to

$$\big\langle \nabla f(\hat{x}) - \nabla f(a), v \big\rangle = 0 \text{ for all } v \in L, \quad \text{and} \quad \hat{x} \in c + L.$$

i.e.,

$$P_L \nabla f(\hat{x}) = P_L \nabla f(a), \quad \text{and} \quad P_{L^\perp} \hat{x} = P_{L^\perp} c. \tag{6.16}$$

In other words, $\hat{x} = P_S^f(a)$, for $S = c + L$, if and only if (6.16) holds.

Set $\hat{x} = \nabla\psi(X\hat{\beta})$, $f = \psi^*$, $a = \nabla\psi(0)$, $c = y - \lambda M^+ D_{-\mathcal{B}}^T s$, and $L = \mathrm{null}(M)$. We see that (6.14) is equivalent to $P_{L^\perp}\hat{x} = P_{L^\perp} c$. Meanwhile, using $(\nabla\psi)^{-1} = \nabla\psi^*$ as guaranteed by essential smoothness and essential strict convexity of ψ, we see that (6.15) is equivalent to $P_{\mathrm{null}(M)} \nabla\psi^*(\nabla\psi(X\hat{\beta})) = 0$, in turn equivalent to $P_L \nabla f(\hat{x}) = P_L \nabla f(a)$. From the first-order optimality conditions (6.16), this shows that $\nabla\psi(X\hat{\beta}) = P_{c+L}^f(a) = P_{y - K_{\mathcal{B},s}}^{\psi^*}(\nabla\psi(0))$. Using $(\nabla\psi)^{-1} = \nabla\psi^*$, once again, establishes (2.38).

As for (2.39), this follows by simply writing (2.38) as

$$M^T\hat{\beta} = \nabla\psi^*\Big(P_{y - K_{\mathcal{B},s}}^{\psi^*}\big(\nabla\psi(0)\big)\Big),$$

where we have again used $X\hat{\beta} = XP_{\mathrm{null}(D_{-\mathcal{B}})}\hat{\beta} = M^T\hat{\beta}$. Solving the above linear system for $\hat{\beta}$ gives (2.39), where $b \in \mathrm{null}(M^T) = \mathrm{null}(XP_{\mathrm{null}(D_{-\mathcal{B}})})$. This constraint together with $b \in \mathrm{null}(D_{-\mathcal{B}})$ implies $b \in \mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}})$, as claimed.

Finally, the results with $\mathcal{A}, r$ in place of $\mathcal{B}, s$ follow similarly. We begin by multiplying both sides of (2.29) by $P_{\mathrm{null}(D_{-\mathcal{A}})}$, and then proceed with the same chain of arguments as above. □

6.1.13 Proof of Lemma 2.18

The proof follows a similar general strategy to that of Lemma 9 in Tibshirani and Taylor [137]. We will abbreviate $\mathcal{B} = \mathcal{B}(y)$, $s = s(y)$, $\mathcal{A} = \mathcal{A}(y)$, and $r = r(y)$. Consider the representation for $\hat{\beta}(y)$ in (2.39) of Lemma 2.17. As the active set is $\mathcal{A}$, we know that

$$D_{\mathcal{B}\setminus\mathcal{A}}(XP_{\mathrm{null}(D_{-\mathcal{B}})})^+ \nabla\psi^*\Big(P_{y - K_{\mathcal{B},s}}^{\psi^*}\big(\nabla\psi(0)\big)\Big) + D_{\mathcal{B}\setminus\mathcal{A}} b = 0,$$

i.e.,

$$D_{\mathcal{B}\setminus\mathcal{A}}(XP_{\mathrm{null}(D_{-\mathcal{B}})})^+ \nabla\psi^*\Big(P_{y - K_{\mathcal{B},s}}^{\psi^*}\big(\nabla\psi(0)\big)\Big) = -D_{\mathcal{B}\setminus\mathcal{A}} b \in D_{\mathcal{B}\setminus\mathcal{A}}\big(\mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}})\big),$$

and so

$$P_{[D_{\mathcal{B}\setminus\mathcal{A}}(\mathrm{null}(X)\cap\mathrm{null}(D_{-\mathcal{B}}))]^\perp} D_{\mathcal{B}\setminus\mathcal{A}}(XP_{\mathrm{null}(D_{-\mathcal{B}})})^+ \nabla\psi^*\Big(P_{y - K_{\mathcal{B},s}}^{\psi^*}\big(\nabla\psi(0)\big)\Big) = 0.$$

Recalling $M_{\mathcal{A},\mathcal{B}}$ as defined in (2.40), and abbreviating $\hat{x} = P_{y - K_{\mathcal{B},s}}^{\psi^*}(\nabla\psi(0))$, we may write this simply as

$$\nabla\psi^*(\hat{x}) \in \mathrm{null}(M_{\mathcal{A},\mathcal{B}}).$$

Since $\nabla\psi^*(\hat{x}) = X\hat{\beta}(y)$, we have $\nabla\psi^*(\hat{x}) \in \mathrm{col}(XP_{\mathrm{null}(D_{-\mathcal{B}})})$, so combining this with above display, and using $(\nabla\psi^*)^{-1} = \nabla\psi$, gives

$$\hat{x} \in \nabla\psi\big(\mathrm{col}(XP_{\mathrm{null}(D_{-\mathcal{B}})}) \cap \mathrm{null}(M_{\mathcal{A},\mathcal{B}})\big).$$

And since $\hat{x} \in y - K_{\mathcal{B},s}$, with $K_{\mathcal{B},s}$ an affine space, as defined in (2.37), we have $y \in \hat{x} + K_{\mathcal{B},s}$, which combined with the last display implies

$$y \in K_{\mathcal{B},s} + \nabla\psi\big(\mathrm{col}(XP_{\mathrm{null}(D_{-\mathcal{B}})}) \cap \mathrm{null}(M_{\mathcal{A},\mathcal{B}})\big).$$

But as $y \notin \mathcal{N}$, where the set $\mathcal{N}$ is defined in (2.41), we arrive at

$$M_{\mathcal{A},\mathcal{B}} = P_{[D_{\mathcal{B}\setminus\mathcal{A}}(\mathrm{null}(X)\cap\mathrm{null}(D_{-\mathcal{B}}))]^\perp} D_{\mathcal{B}\setminus\mathcal{A}}(XP_{\mathrm{null}(D_{-\mathcal{B}})})^+ = 0,$$

which means

$$\mathrm{col}\big(D_{\mathcal{B}\setminus\mathcal{A}}(XP_{\mathrm{null}(D_{-\mathcal{B}})})^+\big) \subseteq D_{\mathcal{B}\setminus\mathcal{A}}\big(\mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}})\big). \tag{6.17}$$

This is an important realization that we will return to shortly.

As for the optimal subgradient $\hat{\gamma}(y)$ corresponding to $\hat{\beta}(y)$, note that we can write

$$\begin{aligned}
\hat{\gamma}_{\mathcal{B}}(y) &= \lambda s, \\
\hat{\gamma}_{-\mathcal{B}}(y) &= \frac{1}{\lambda}(D_{-\mathcal{B}}^T)^+\Big[X^T\Big(y - P^{\psi^*}_{y-K_{\mathcal{B},s}}\big(\nabla\psi(0)\big)\Big) - \lambda D_{\mathcal{B}}^T s\Big] + c,
\end{aligned} \tag{6.18}$$

for some $c \in \mathrm{null}(D_{-\mathcal{B}}^T)$. The first expression holds by definition of $\mathcal{B}, s$, and the second is a result of solving for $\hat{\gamma}_{-\mathcal{B}}(y)$ in the stationarity condition (2.29), after plugging in for the form of the fit in (2.38).

Now, at a new response y', consider defining

$$\begin{aligned}
\hat{\beta}(y') &= (XP_{\mathrm{null}(D_{-\mathcal{B}})})^+\nabla\psi^*\Big(P^{\psi^*}_{y'-K_{\mathcal{B},s}}\big(\nabla\psi(0)\big)\Big) + b', \\
\hat{\gamma}_{\mathcal{B}}(y') &= \lambda s, \\
\hat{\gamma}_{-\mathcal{B}}(y') &= \frac{1}{\lambda}(D_{-\mathcal{B}}^T)^+\Big[X^T\Big(y' - P^{\psi^*}_{y'-K_{\mathcal{B},s}}\big(\nabla\psi(0)\big)\Big) - \lambda D_{\mathcal{B}}^T s\Big] + c,
\end{aligned}$$

for some $b' \in \mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}})$ to be specified later, and for the same value of $c \in \mathrm{null}(D_{-\mathcal{B}}^T)$ as in (6.18). By the same arguments as given at the end of the proof of Lemma 2.14, where we discussed the constraint qualification conditions (2.24), (2.25), the Bregman projection $P^{\psi^*}_{y'-K_{\mathcal{B},s}}(\nabla\psi(0))$ in the above expressions is well-defined, for any y', under (2.31). However, this Bregman projection need not lie in $\mathrm{int}(\mathrm{dom}(\psi^*))$—and therefore $\nabla\psi^*(P^{\psi^*}_{y'-K_{\mathcal{B},s}}(\nabla\psi(0)))$ need not be well-defined—unless we have the additional condition $y' \in \mathrm{int}(\mathrm{ran}(\nabla\psi)) + C$. Fortunately, under (2.32), the latter condition on y' is implied as long as y' is sufficiently close to y, i.e., there exists a neighborhood U_0 of y such that $y' \in \mathrm{int}(\mathrm{ran}(\nabla\psi)) + C$, provided $y' \in U_0$. By Lemma 2.14, we see that a solution in (2.18) exists at such a point y'. In what remains, we will show that this solution and its optimal subgradient obey the form in the above display.

Note that, by construction, the pair $(\hat{\beta}(y'), \hat{\gamma}(y'))$ defined above satisfy the stationarity condition (2.29) at y', and $\hat{\gamma}(y')$ has boundary set and boundary signs $\mathcal{B}, s$. It remains to show that $(\hat{\beta}(y'), \hat{\gamma}(y'))$ satisfy the subgradient condition (2.21), and that $\hat{\beta}(y')$ has active set and active signs $\mathcal{A}, r$; equivalently, it remains to verify the following three properties, for y' sufficiently close to y, and for an appropriate choice of b':

(i) $\|\hat{\gamma}_{-\mathcal{B}}(y')\|_\infty < 1$;

(ii) $\mathrm{supp}(D\hat{\beta}(y')) = \mathcal{A}$;

(iii) $\mathrm{sign}(D_{\mathcal{A}}\hat{\beta}(y')) = r$.

Because $\hat{\gamma}(y)$ is a subgradient corresponding to $\hat{\beta}(y)$, and has boundary set and boundary signs $\mathcal{B}, s$, we know that $\hat{\gamma}_{-\mathcal{B}}(y)$ in (6.18) has ℓ_∞ norm strictly less than 1. Thus, by continuity of

$$x \mapsto \left\| \frac{1}{\lambda}(D^T_{-\mathcal{B}})^+ \Big[X^T \Big(x - P^{\psi^*}_{x - K_{\mathcal{B},s}}\big(\nabla\psi(0)\big)\Big) - \lambda D^T_{\mathcal{B}} s \Big] + c \right\|_\infty$$

at y, which is implied by continuity of $x \mapsto P^{\psi^*}_{x-K_{\mathcal{B},s}}(\nabla\psi(0))$ at y, by Lemma 6.2, we know that there exists some neighborhood U_1 of y such that property (i) holds, provided $y' \in U_1$.

By the important fact established in (6.17), we see that there exists $b' \in \mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}})$ such that

$$D_{\mathcal{B}\setminus\mathcal{A}} b' = -D_{\mathcal{B}\setminus\mathcal{A}}(XP_{\mathrm{null}(D_{-\mathcal{B}})})^+ \nabla\psi^*\Big(P^{\psi^*}_{y'-K_{\mathcal{B},s}}\big(\nabla\psi(0)\big)\Big),$$

which implies that $D_{\mathcal{B}\setminus\mathcal{A}}\hat{\beta}(y') = 0$. To verify properties (ii) and (iii), we must show this choice of b' is such that $D_{\mathcal{A}}\hat{\beta}(y')$ is nonzero in every coordinate and has signs matching r. Define a map

$$T(x) = (XP_{\mathrm{null}(D_{-\mathcal{B}})})^+ \nabla\psi^*\Big(P^{\psi^*}_{x-K_{\mathcal{B},s}}\big(\nabla\psi(0)\big)\Big),$$

which is continuous at y, again by continuity of $x \mapsto P^{\psi^*}_{x-K_{\mathcal{B},s}}(\nabla\psi(0))$ at y, by Lemma 6.2. Observe that

$$D_{\mathcal{A}}\hat{\beta}(y') = D_{\mathcal{A}}T(y') + D_{\mathcal{A}}b' = D_{\mathcal{A}}T(y') + D_{\mathcal{A}}b + D_{\mathcal{A}}(b - b').$$

As $D_{\mathcal{A}}\hat{\beta}(y) = D_{\mathcal{A}}T(y) + D_{\mathcal{A}}b$ is nonzero in every coordinate and has signs equal to r, by definition of $\mathcal{A}, r$, and T is continuous at y, there exists a neighborhood U_2 of y such that $D_{\mathcal{A}}T(y') + D_{\mathcal{A}}b$ is nonzero in each coordinate with signs matching r, provided $y' \in U_2$. Furthermore, as

$$\|D_{\mathcal{A}}(b - b')\|_\infty \le \|D^T\|_{2,\infty}\|b - b'\|_2,$$

where $\|D^T\|_{2,\infty}$ denotes the maximum ℓ_2 norm of rows of D, we see that $D_{\mathcal{A}}T(y') + D_{\mathcal{A}}b'$ will be nonzero in each coordinate with the correct signs, provided b' can be chosen arbitrarily close to b, subject to the restrictions $b' \in \mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}})$ and $D_{\mathcal{B}\setminus\mathcal{A}}b' = -D_{\mathcal{B}\setminus\mathcal{A}}T(y')$.

Such a b' does indeed exist, by the bounded inverse theorem. Let $L = \mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}})$, and $N = \mathrm{null}(D_{\mathcal{B}\setminus\mathcal{A}}) \cap L$. Consider the linear map $D_{\mathcal{B}\setminus\mathcal{A}}$, viewed as a function from L/N (the quotient of L by N) to $D_{\mathcal{B}\setminus\mathcal{A}}(L)$: this is a bijection, and therefore it has a bounded inverse. This means that there exists some $R > 0$ such that

$$\|b - b'\|_2 \le R\big\|D_{\mathcal{B}\setminus\mathcal{A}}T(y') - D_{\mathcal{B}\setminus\mathcal{A}}T(y)\big\|_2,$$

for a choice of $b' \in \mathrm{null}(X) \cap \mathrm{null}(D_{-\mathcal{B}})$ with $D_{\mathcal{B}\setminus\mathcal{A}}b' = -D_{\mathcal{B}\setminus\mathcal{A}}T(y')$. By continuity of T at y, once again, there exists a neighborhood U_3 of y such that the right-hand side above is sufficiently small, i.e., such that $\|b - b'\|_2$ is sufficiently small, provided $y' \in U_3$.

Finally, letting $U = U_0 \cap U_1 \cap U_2 \cap U_3$, we see that we have established properties (i), (ii), and (iii), and hence the desired result, provided $y' \in U$. □

6.1.14 Continuity result for Bregman projections

Lemma 6.2. *Let f, f^* be a conjugate pair of Legendre (essentially smooth and essentially strictly convex) functions on $\mathbb{R}^n$, with $0 \in \mathrm{int}(\mathrm{dom}(f^*))$. Let $S \subseteq \mathbb{R}^n$ be a nonempty closed convex set. Then the Bregman projection map*

$$x \mapsto P^f_{x-S}\big(\nabla f^*(0)\big)$$

is continuous on all of $\mathbb{R}^n$. Moreover, $P^f_{x-S}(\nabla f^*(0)) \in \mathrm{int}(\mathrm{dom}(f))$ for any $x \in \mathrm{int}(\mathrm{dom}(f)) + S$.

Proof. As $0 \in \mathrm{int}(\mathrm{dom}(f^*))$, we know that f has no directions of recession (e.g., by Theorems 27.1 and 27.3 in Rockafellar [115]), thus the Bregman projection $P^f_{x-S}(\nabla f^*(0))$ is well-defined for any $x \in \mathbb{R}^n$. Further, for $x - S \in \mathrm{int}(\mathrm{dom}(f))$, we know that $P^f_{x-S}(\nabla f^*(0)) \in \mathrm{int}(\mathrm{dom}(f))$, by essential smoothness of f (by the property that $\|\nabla f\|_2$ approaches ∞ along any sequence that converges to boundary point of $\mathrm{dom}(f)$; e.g., see Theorem 3.12 in Bauschke and Borwein [7]).

It remains to verify continuity of $x \mapsto P^f_{x-S}(\nabla f^*(0))$. Write $P^f_{x-S}(\nabla f^*(0)) = \hat{v}$, where $\hat{v}$ is the unique solution of

$$\underset{v \in x-S}{\text{minimize}} \;\; f(v),$$

or equivalently, $P^f_{x-S}(\nabla f^*(0)) = \hat{w} + x$, where $\hat{w}$ is the unique solution of

$$\underset{w \in -S}{\text{minimize}} \;\; f(w + x).$$

It suffices to show continuity of the unique solution in the above problem, as a function of x. This can be established using results from variational analysis, provided some conditions are met on the bi-criterion function $f_0(w, x) = f(w + x)$. In particular, Corollary 7.43 in Rockafellar and Wets [116] implies that the unique minimizer in the above problem is continuous in x, provided f_0 is a closed proper convex function that is level-bounded in w locally uniformly in x. By assumption, f is a closed proper convex function (it is Legendre), and thus so is f_0. The level-boundedness condition can be checked as follows. Fix any $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^n$. The α-level set $\{w : f(w + x) \leq \alpha\}$ is bounded since $x \mapsto f(x + w)$ has no directions of recession (to see that this implies boundedness of all level sets, e.g., combine Theorem 27.1 and Corollary 8.7.1 of Rockafellar [115]). Meanwhile, for any $x' \in \mathbb{R}^n$,

$$\{w : f(w + x') \leq \alpha\} = \{w : f(w + x) \leq \alpha\} + x' - x.$$

Hence, the α-level set of $f_0(\cdot, x')$ is uniformly bounded for all x' in a neighborhood of x, as desired. This completes the proof. □

6.1.15 Proof of Lemma 2.19

The proof is similar to that of Lemma 10 in Tibshirani and Taylor [137]. Let $\mathcal{B}, s$ be the boundary set and signs of an arbitrary optimal subgradient in $\hat{\gamma}(y)$ in (2.18), and let $\mathcal{A}, r$ be the active set and active signs of an arbitrary solution in $\hat{\beta}(y)$ in (2.18). (Note that $\hat{\gamma}(y)$ need not correspond to $\hat{\beta}(y)$; it may be a subgradient corresponding to another solution in (2.18).)

By (two applications of) Lemma 2.18, there exist neighborhoods U_1, U_2 of y such that, over U_1, optimal subgradients exist with boundary set and boundary signs $\mathcal{B}, s$, and over U_2, solutions exist with active set and active signs $\mathcal{A}, r$. For any $y' \in U = U_1 \cap U_2$, by Lemma 2.17 and the uniqueness of the fit from Lemma 2.12, we have

$$X\hat{\beta}(y) = \nabla\psi^*\Big(P^{\psi^*}_{y-K_{\mathcal{B},s}}\big(\nabla\psi(0)\big)\Big) = \nabla\psi^*\Big(P^{\psi^*}_{y-K_{\mathcal{A},r}}\big(\nabla\psi(0)\big)\Big),$$

and as $\nabla\psi^*$ is a homeomorphism,

$$P^{\psi^*}_{y'-K_{\mathcal{B},s}}\big(\nabla\psi(0)\big) = P^{\psi^*}_{y'-K_{\mathcal{A},r}}\big(\nabla\psi(0)\big). \tag{6.19}$$

We claim that this implies $\mathrm{null}(P_{\mathrm{null}(D_{-\mathcal{B}})}X^T) = \mathrm{null}(P_{\mathrm{null}(D_{-\mathcal{A}})}X^T)$.

To see this, take any direction $z \in \mathrm{null}(P_{\mathrm{null}(D_{-\mathcal{B}})}X^T)$, and let $\epsilon > 0$ be sufficiently small so that $y' = y + \epsilon z \in U$. From (6.19), we have

$$P^{\psi^*}_{y'-K_{\mathcal{A},r}}\big(\nabla\psi(0)\big) = P^{\psi^*}_{y'-K_{\mathcal{B},s}}\big(\nabla\psi(0)\big) = P^{\psi^*}_{y-K_{\mathcal{B},s}}\big(\nabla\psi(0)\big) = P^{\psi^*}_{y-K_{\mathcal{A},r}}\big(\nabla\psi(0)\big),$$

where the second equality used $y' - K_{\mathcal{B},s} = y - K_{\mathcal{B},s}$, and the third used the fact that (6.19) indeed holds at y. Now consider the left-most and right-most expressions above. For these two projections to match, we must have $z \in \mathrm{null}(P_{\mathrm{null}(D_{-\mathcal{A}})}X^T)$; otherwise, the affine subspaces $y' - K_{\mathcal{A},r}$ and $y - K_{\mathcal{A},r}$ would be parallel, in which case clearly the projections cannot coincide. Hence, we have shown that $\mathrm{null}(P_{\mathrm{null}(D_{-\mathcal{B}})}X^T) \subseteq \mathrm{null}(P_{\mathrm{null}(D_{-\mathcal{A}})}X^T)$. The reverse inclusion follows similarly, establishing the desired claim.

Lastly, as $\mathcal{B}, \mathcal{A}$ were arbitrary, the linear subspace $L = \mathrm{null}(P_{\mathrm{null}(D_{-\mathcal{B}})}X^T) = \mathrm{null}(P_{\mathrm{null}(D_{-\mathcal{A}})}X^T)$ must be unchanged for any choice of boundary set $\mathcal{B}$ and active set $\mathcal{A}$ at y, completing the proof. □

6.2 Supplementary Material for Early-Stopped Gradient Descent for Least Squares Regression

6.2.1 Proof of Lemma 3.3

Let $X^TX/n = VSV^T$ be an eigendecomposition of X^TX/n. Then we can rewrite the gradient descent iteration (3.2) as

$$\beta^{(k)} = \beta^{(k-1)} + \frac{\epsilon}{n} \cdot X^T(y - X\beta^{(k-1)}) = (I - \epsilon VSV^T)\beta^{(k-1)} + \frac{\epsilon}{n} \cdot X^Ty.$$

Rotating by V^T, we get

$$\tilde{\beta}^{(k)} = (I - \epsilon S)\tilde{\beta}^{(k-1)} + \tilde{y},$$

where we let $\tilde{\beta}^{(j)} = V^T\beta^{(j)}$, $j = 1, 2, 3, \ldots$ and $\tilde{y} = (\epsilon/n)V^TX^Ty$. Unraveling the preceding display, we find that

$$\tilde{\beta}^{(k)} = (I - \epsilon S)^k\tilde{\beta}^{(0)} + \sum_{j=0}^{k-1}(I - \epsilon S)^j\tilde{y}.$$

Furthermore applying the assumption that the initial point $\beta^{(0)} = 0$ yields

$$\tilde{\beta}^{(k)} = \sum_{j=0}^{k-1} (I - \epsilon S)^j \tilde{y} = (\epsilon S)^{-1}(I - (I - \epsilon S)^k)\tilde{y},$$

with the second equality following after a short inductive argument.

Now notice that $\beta^{(k)} = V\tilde{\beta}^{(k)}$, since VV^T is the projection onto the row space of X, and $\beta^{(k)}$ lies in the row space. Rotating back to the original space then gives

$$\beta^{(k)} = V(\epsilon S)^{-1}(I - (I - \epsilon S)^k)\tilde{y} = \frac{1}{n} V S^{-1}(I - (I - \epsilon S)^k)V^T X^T y.$$

Compare this to the solution of the optimization problem in Lemma 3.3, which is

$$(X^T X + nQ_k)^{-1} X^T y = \frac{1}{n}(VSV^T + Q_k)^{-1} X^T y.$$

Equating the last two displays, we see that we must have

$$VS^{-1}(I - (I - \epsilon S)^k)V^T = (VSV^T + Q_k)^{-1}.$$

Inverting both sides and rearranging, we get

$$Q_k = VS(I - (I - \epsilon S)^k)^{-1}V^T - VSV^T,$$

and an application of the matrix inversion lemma shows that $(I - (I - \epsilon S)^k)^{-1} = I + ((I - \epsilon S)^{-k} - I)^{-1}$, so

$$Q_k = VS((I - \epsilon S)^{-k} - I)^{-1}V^T,$$

as claimed in the lemma.

6.2.2 Proof of Lemma 3.4

Recall that Lemma 3.1 gives the gradient flow solution at time t, in (3.6). Compare this to the solution of the optimization problem in Lemma 3.4, which is

$$(X^T X + nQ_t)^{-1} X^T y.$$

To equate these two, we see that we must have

$$(X^T X)^+(I - \exp(-tX^T X/n)) = (X^T X + nQ_t)^{-1},$$

i.e., writing $X^T X/n = VSV^T$ as an eigendecomposition of $X^T X/n$,

$$VS^+(I - \exp(-tS))V^T = (VSV^T + Q_t)^{-1}.$$

Inverting both sides and rearranging, we find that

$$Q_t = VS(I - \exp(-tS))^{-1}V^T - VSV^T,$$

which is as claimed in the lemma.

6.2.3 Proof of Lemma 3.5

For fixed β_0, and any estimator $\hat{\beta}$, recall the bias-variance decomposition

$$\mathrm{Risk}(\hat{\beta}; \beta_0) = \|\mathbb{E}(\hat{\beta}) - \beta_0\|_2^2 + \mathrm{tr}[\mathrm{Cov}(\hat{\beta})].$$

For the gradient flow estimator in (3.6), we have

$$\begin{aligned}
\mathbb{E}[\hat{\beta}^{\mathrm{gf}}(t)] &= (X^T X)^+ (I - \exp(-tX^T X/n)) X^T X \beta_0 \\
&= (X^T X)^+ X^T X (I - \exp(-tX^T X/n)) \beta_0 \\
&= (I - \exp(-tX^T X/n)) \beta_0. \qquad (6.20)
\end{aligned}$$

In the second line, we used the fact that $X^T X$ and $(I - \exp(-tX^T X/n))$ are simultaneously diagonalizable, and so they commute; in the third line, we used the fact that $(X^T X)^+ X^T X = X^+ X$ is the projection onto the row space of X, and the image of $I - \exp(-tX^T X/n)$ is already in the row space. Hence the bias is, abbreviating $\hat{\Sigma} = X^T X/n$,

$$\big\|\mathbb{E}[\hat{\beta}^{\mathrm{gf}}(t)] - \beta_0\big\|_2^2 = \|\exp(-t\hat{\Sigma})\beta_0\|_2^2 = \sum_{i=1}^p |v_i^T \beta_0|^2 \exp(-2ts_i). \qquad (6.21)$$

As for the variance, we have

$$\begin{aligned}
\mathrm{tr}\big(\mathrm{Cov}[\hat{\beta}^{\mathrm{gf}}(t)]\big) &= \sigma^2 \mathrm{tr}\big[(X^T X)^+ (I - \exp(-t\hat{\Sigma}))(X^T X)(I - \exp(-t\hat{\Sigma}))(X^T X)^+\big] \\
&= \frac{\sigma^2}{n} \mathrm{tr}\big[\hat{\Sigma}^+ (I - \exp(-t\hat{\Sigma}))^2\big] \\
&= \frac{\sigma^2}{n} \sum_{i=1}^p \frac{(1 - \exp(-ts_i))^2}{s_i}, \qquad (6.22)
\end{aligned}$$

where in the second line we used the fact that $\hat{\Sigma}^+$ and $(I - \exp(-t\hat{\Sigma}))$ are simultaneously diagonalizable, and hence commute, and also the fact that $\hat{\Sigma}^+ \hat{\Sigma} \hat{\Sigma}^+ = \hat{\Sigma}^+$. Putting together (6.21) and (6.22) proves the result in (3.11).

When β_0 follows the prior in (3.10), the variance (6.22) remains unchanged. The expectation of the bias (6.21) (over β_0) is

$$\mathbb{E}\big[\beta_0^T \exp(-2t\hat{\Sigma})\beta_0\big] = \mathrm{tr}\big[\mathbb{E}(\beta_0 \beta_0^T) \exp(-2t\hat{\Sigma})\big] = \frac{r^2}{p} \sum_{i=1}^p \exp(-2ts_i),$$

which leads to (3.12), after the appropriate definition of α.

6.2.4 Derivation of (3.13), (3.14)

As in the calculations in the last section, consider for the ridge estimator in (3.5),

$$\mathbb{E}[\hat{\beta}^{\mathrm{ridge}}(\lambda)] = (X^T X + n\lambda I)^{-1} X^T X \beta_0 = (\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma} \beta_0, \qquad (6.23)$$

where we have again abbreviated $\hat{\Sigma} = X^T X/n$. The bias is thus

$$\begin{aligned}\big\|\mathbb{E}[\hat{\beta}^{\text{ridge}}(\lambda)] - \beta_0\big\|_2^2 &= \big\|(\hat{\Sigma}+\lambda I)^{-1}(\hat{\Sigma}-I)\beta_0\big\|_2^2 \\ &= \big\|\lambda(\hat{\Sigma}+\lambda I)^{-1}\beta_0\big\|_2^2 \\ &= \sum_{i=1}^p |v_i^T\beta_0|^2 \frac{\lambda^2}{(s_i+\lambda)^2}, \end{aligned} \tag{6.24}$$

the second equality following after adding and subtracting λI to the second term in parentheses, and expanding. For the variance, we compute

$$\begin{aligned}\text{tr}\big(\text{Cov}[\beta^{\text{ridge}}(\lambda)]\big) &= \sigma^2 \text{tr}\big[(X^TX+n\lambda I)^{-1}X^TX(X^TX+n\lambda I)^{-1}\big] \\ &= \frac{\sigma^2}{n}\text{tr}\big[\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-2}\big] \\ &= \frac{\sigma^2}{n}\sum_{i=1}^p \frac{s_i}{(s_i+\lambda)^2}, \end{aligned} \tag{6.25}$$

the second equality following by noting that $\hat{\Sigma}$ and $(\hat{\Sigma}+\lambda I)^{-1}$ are simultaneously diagonalizable, and therefore commute. Putting together (6.24) and (6.25) proves the result in (3.13). The Bayes result (3.14) follows by taking an expectation of the bias (6.24) (over β_0), just as in the last section for gradient flow.

6.2.5 Proof of Lemma 3.6

First, observe that for fixed β_0, and any estimator $\hat{\beta}$,

$$\text{Risk}^{\text{out}}(\hat{\beta};\beta_0) = \mathbb{E}\|\hat{\beta}-\beta_0\|_\Sigma^2,$$

where $\|z\|_A^2 = z^TAz$. The bias-variance decomposition for out-of-sample prediction risk is hence

$$\text{Risk}^{\text{out}}(\hat{\beta};\beta_0) = \|\mathbb{E}(\hat{\beta})-\beta_0\|_\Sigma^2 + \text{tr}[\text{Cov}(\hat{\beta})\Sigma].$$

For gradient flow, we can compute the bias, from (6.20),

$$\big\|\mathbb{E}[\hat{\beta}^{\text{gf}}(t)] - \beta_0\big\|_\Sigma^2 = \|\exp(-t\hat{\Sigma})\beta_0\|_\Sigma^2 = \beta_0^T\exp(-t\hat{\Sigma})\Sigma\exp(-t\hat{\Sigma})\beta_0, \tag{6.26}$$

and likewise the variance,

$$\begin{aligned}\text{tr}\big(\text{Cov}[\beta^{\text{gf}}(t)]\big) &= \sigma^2\text{tr}\big[(X^TX)^+(I-\exp(-t\hat{\Sigma}))(X^TX)(I-\exp(-t\hat{\Sigma}))(X^TX)^+\Sigma\big] \\ &= \frac{\sigma^2}{n}\text{tr}\big[\hat{\Sigma}^+(I-\exp(-t\hat{\Sigma}))^2\Sigma\big]. \end{aligned} \tag{6.27}$$

Putting together (6.26) and (6.27) proves the result in (3.16). The Bayes result (3.17) follows by taking an expectation over the bias, as argued previously.

We note that the in-sample prediction risk is given by the same formulae except with Σ replaced by $\hat{\Sigma}$, which leads to

$$\begin{aligned}\mathrm{Risk}^{\mathrm{in}}(\hat{\beta}^{\mathrm{gf}}(t);\beta_0) &= \beta_0^T \exp(-t\hat{\Sigma})\hat{\Sigma}\exp(-t\hat{\Sigma})\beta_0 + \frac{\sigma^2}{n}\mathrm{tr}\big[\hat{\Sigma}^+(I-\exp(-t\hat{\Sigma}))^2\hat{\Sigma}\big] \\ &= \sum_{i=1}^{p}\bigg(|v_i^T\beta_0|^2 s_i \exp(-2ts_i) + \frac{\sigma^2}{n}(1-\exp(-ts_i))^2\bigg), \end{aligned} \tag{6.28}$$

and

$$\begin{aligned}\mathrm{Risk}^{\mathrm{out}}(\hat{\beta}^{\mathrm{gf}}(t)) &= \frac{\sigma^2}{n}\mathrm{tr}\big[\alpha\exp(-2t\hat{\Sigma})\hat{\Sigma} + \hat{\Sigma}^+(I-\exp(-t\hat{\Sigma}))^2\hat{\Sigma}\big] \\ &= \frac{\sigma^2}{n}\sum_{i=1}^{p}\big[\alpha s_i\exp(-2ts_i) + (1-\exp(-ts_i))^2\big]. \end{aligned} \tag{6.29}$$

6.2.6 Derivation of (3.18), (3.19)

For ridge, we can compute the bias, from (6.23),

$$\big\|\mathbb{E}[\hat{\beta}^{\mathrm{ridge}}(\lambda)] - \beta_0\big\|_\Sigma^2 = \big\|\lambda(\hat{\Sigma}+\lambda I)^{-1}\beta_0\big\|_\Sigma^2 = \lambda^2\beta_0^T(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\beta_0, \tag{6.30}$$

and also the variance,

$$\begin{aligned}\mathrm{tr}\big(\mathrm{Cov}[\beta^{\mathrm{ridge}}(\lambda)]\Sigma\big) &= \sigma^2\mathrm{tr}\big[(X^TX+n\lambda I)^{-1}X^TX(X^TX+n\lambda I)^{-1}X^T\Sigma\big] \\ &= \frac{\sigma^2}{n}\mathrm{tr}\big[\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-2}\Sigma\big]. \end{aligned} \tag{6.31}$$

Putting together (6.30) and (6.31) proves (3.18), and the Bayes result (3.19) follows by taking an expectation over the bias, as argued previously.

Again, we note that the in-sample prediction risk expressions is given by replacing Σ replaced by $\hat{\Sigma}$, yielding

$$\begin{aligned}\mathrm{Risk}^{\mathrm{in}}(\hat{\beta}^{\mathrm{ridge}}(\lambda);\beta_0) &= \lambda^2\beta_0^T(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\beta_0 + \frac{\sigma^2}{n}\mathrm{tr}\big[\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}\big] \\ &= \sum_{i=1}^{p}\bigg(|v_i^T\beta_0|^2\frac{\lambda^2 s_i}{(s_i+\lambda)^2} + \frac{\sigma^2}{n}\frac{s_i^2}{(s_i+\lambda)^2}\bigg), \end{aligned} \tag{6.32}$$

and

$$\begin{aligned}\mathrm{Risk}^{\mathrm{in}}(\hat{\beta}^{\mathrm{ridge}}(\lambda)) &= \frac{\sigma^2}{n}\mathrm{tr}\big[\lambda^2\alpha(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma} + \hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}\big] \\ &= \frac{\sigma^2}{n}\sum_{i=1}^{p}\frac{\alpha\lambda^2 s_i + s_i^2}{(s_i+\lambda)^2}. \end{aligned} \tag{6.33}$$

6.2.7 Proof of Theorem 3.1, Part (c)

As we can see from comparing (3.11), (3.13) to (6.28), (6.32), the only difference in the latter in-sample prediction risk expressions is that each summand has been multiplied by s_i. Therefore the exact same relative bounds apply termwise, i.e., the arguments for part (a) apply here. The Bayes result again follows just by taking expectations.

6.2.8 Proof of Lemma 3.9

As in the proof of Lemma 3.8, because all matrices here are simultaneously diagonalizable, the claim reduces to one about eigenvalues, and it suffices to check that $e^{-2x} + (1-e^{-x})^2/x \leq 1.2147/(1+x)$ for all $x \geq 0$. Completing the square and simplifying,

$$\begin{aligned} e^{-2x} + \frac{(1-e^{-x})^2}{x} &= \frac{(1+x)e^{-2x} - 2e^{-x} + 1}{x} \\ &= \frac{(\sqrt{1+x}e^{-x} - \frac{1}{\sqrt{1+x}})^2}{x} + \frac{x}{1+x}. \end{aligned}$$

Now observe that, for any constant $C > 0$,

$$\begin{aligned} \frac{(\sqrt{1+x}e^{-x} - \frac{1}{\sqrt{1+x}})^2}{x} + \frac{x}{1+x} \leq (1+C^2)\frac{1}{1+x} \\ \iff |(1+x)e^{-x} - 1| \leq C\sqrt{x} \\ \iff 1 - (1+x)e^{-x} \leq C\sqrt{x}, \end{aligned} \tag{6.34}$$

the last line holding because the basic inequality $e^x \geq 1+x$ implies that $e^{-x} \leq 1/(1+x)$, for $x > -1$. We see that for the above line to hold, we may take

$$C = \max_{x \geq 0} \big[1 - (1+x)e^{-x}\big]/\sqrt{x} = 0.4634,$$

which has been computed by numerical maximization, i.e., we find that the desired inequality (6.34) holds with $(1+C^2) = 1.2147$.

6.2.9 Proof of Theorem 3.3, Part (b)

The lower bounds for the in-sample and out-of-sample prediction risks follow by the same arguments as in the estimation risk case (the ridge estimator here is the Bayes estimator in the case of a normal-normal likelihood-prior pair, and the risks here do not depend on the specific form of the likelihood and prior).

For the upper bounds, for in-sample prediction risk, we can see from comparing (3.12), (3.14) to (6.29), (6.33), the only difference in the latter expressions is that each summand has been multiplied by s_i, and hence the same relative bounds apply termwise, i.e., the arguments for part (a) carry over directly here.

And for out-of-sample prediction risk, the matrix inside the trace in (3.17) when $t = \alpha$ is

$$\alpha \exp(-2\alpha\hat{\Sigma}) + \hat{\Sigma}^+(I - \exp(-\alpha\hat{\Sigma}))^2,$$

and the matrix inside the trace in (3.19) when $\lambda = 1/\alpha$ is

$$1/\alpha(\hat{\Sigma} + (1/\alpha)I)^{-2} + \hat{\Sigma}(\hat{\Sigma} + (1/\alpha)I)^{-2} = \alpha(\alpha\hat{\Sigma} + I)^{-1}.$$

By Lemma 3.9, we have

$$\alpha \exp(-2\alpha\hat{\Sigma}) + \hat{\Sigma}^{+}(I - \exp(-\alpha\hat{\Sigma}))^2 \preceq 1.2147\alpha(\alpha\hat{\Sigma} + I)^{-1}.$$

Letting A, B denote the matrices on the left- and right-hand sides above, since $A \preceq B$ and $\Sigma \succeq 0$, it holds that $\mathrm{tr}(A\Sigma) \leq \mathrm{tr}(B\Sigma)$, which gives the desired result.

6.2.10 Proof of Theorem 3.6

It is evident that (3.24) is helpful for understanding the Bayes prediction risk of ridge regression (3.19), where the resolvent functional $\mathrm{tr}[(\hat{\Sigma} + zI)^{-1}\Sigma]$ plays a prominent role.

For the Bayes prediction risk of gradient flow (3.17), the connection is less clear. However, the Laplace transform is the key link between (3.17) and (3.24). In particular, defining $g(t) = \exp(tA)$, it is a standard fact that its Laplace transform $\mathcal{L}(g)(z) = \int e^{-tz} g(t)\, dt$ (meaning elementwise integration) is in fact

$$\mathcal{L}(\exp(tA))(z) = (A - zI)^{-1}. \tag{6.35}$$

Using linearity (and invertibility) of the Laplace transform, this means

$$\exp(-2t\hat{\Sigma})\Sigma = \mathcal{L}^{-1}\big((\hat{\Sigma} + zI)^{-1}\Sigma\big)(2t), \tag{6.36}$$

Therefore, we have for the bias term in (3.17),

$$\begin{aligned}\frac{\sigma^2\alpha}{n}\mathrm{tr}\big[\exp(-2t\hat{\Sigma})\Sigma\big] &= \frac{\sigma^2\alpha}{n}\mathrm{tr}\big[\mathcal{L}^{-1}\big((\hat{\Sigma} + zI)^{-1}\Sigma\big)(2t)\big] \\ &= \frac{\sigma^2 p\alpha}{n}\mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma\big]\Big)(2t),\end{aligned} \tag{6.37}$$

where in the second line we again used linearity of the (inverse) Laplace transform. In what follows, we will show that we can commute the limit as $n, p \to \infty$ with the inverse Laplace transform in (6.37), allowing us to apply the Ledoit-Peche result (3.24), to derive an explicit form for the limiting bias. We first give a more explicit representation for the inverse Laplace transform in terms of a line integral in the complex plane

$$\frac{\sigma^2 p\alpha}{n}\mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma\big]\Big)(2t) = \frac{\sigma^2 p\alpha}{n}\frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} \mathrm{tr}\big[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma\big]\exp(2tz)\, dz,$$

where $i = \sqrt{-1}$, and $a \in \mathbb{R}$ is chosen so that the line $[a - i\infty, a + i\infty]$ lies to the right of all singularities of the map $z \mapsto \mathrm{tr}[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma]$. Thus, we may fix any $a > 0$, and reparametrize the integral above as

$$\begin{aligned}\frac{\sigma^2 p\alpha}{n}\mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma\big]\Big)(2t) &= \frac{\sigma^2 p\alpha}{n}\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{tr}\big[p^{-1}\big(\hat{\Sigma} + (a+ib)I\big)^{-1}\Sigma\big]\exp(2t(a+ib))\, db \\ &= \frac{\sigma^2 p\alpha}{n}\frac{1}{\pi}\int_{-\infty}^{0} \mathrm{Re}\Big(\mathrm{tr}\big[p^{-1}\big(\hat{\Sigma} + (a+ib)I\big)^{-1}\Sigma\big]\exp(2t(a+ib))\Big)\, db.\end{aligned} \tag{6.38}$$

The second line can be explained as follows. A straightforward calculation, given in Lemma 6.3, shows that the function $h_{n,p}(z) = \mathrm{tr}[p^{-1}\big(\hat{\Sigma} + zI\big)^{-1}\Sigma]\exp(2tz)$ satisfies $h_{n,p}(\overline{z}) = \overline{h_{n,p}(z)}$; another short calculation, deferred to Lemma 6.4, shows that for any function with such a property, its integral over a vertical line in the complex plane reduces to the integral of twice its real part, over the line segment below the real axis. Now, noting that the integrand above satisfies

$$|h_{n,p}(z)| \leq \|(\hat{\Sigma} + zI)^{-1}\|_2 \|\Sigma\|_2 \leq C_2/a,$$

for all $z \in [a - i\infty, a + i\infty]$, we can take limits in (6.38) and apply the dominated convergence theorem, to yield that almost surely,

$$\begin{aligned}
\lim_{n,p\to\infty} \frac{\sigma^2 p\alpha}{n}\mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma\big]\Big)(2t) \\
&= \sigma^2\gamma\alpha_0\frac{1}{\pi}\int_{-\infty}^{0} \lim_{n,p\to\infty} \mathrm{Re}\Big(\mathrm{tr}\big[p^{-1}\big(\hat{\Sigma} + (a+ib)I\big)^{-1}\Sigma\big]\exp(2t(a+ib))\Big)\,db \\
&= \sigma^2\gamma\alpha_0\frac{1}{\pi}\int_{-\infty}^{0} \mathrm{Re}\big(\theta(a+ib)\exp(2t(a+ib))\big)\,db \\
&= \sigma^2\gamma\alpha_0\frac{1}{2\pi}\int_{-\infty}^{\infty} \theta(a+ib)\exp(2t(a+ib))\,db \\
&= \sigma^2\gamma\alpha_0\mathcal{L}^{-1}(\theta)(2t). \qquad (6.39)
\end{aligned}$$

In the second equality, we used the Ledoit-Peche result (3.24), which applies because $a+ib \in \mathbb{C}_-$ for b in the range of integration. In the third and fourth equalities, we essentially reversed the arguments leading to (6.37), but with $h(z) = \theta(z)\exp(2tz)$ in place of $h_{n,p}$ (note that h must also satisfy $h(\overline{z}) = \overline{h(z)}$, as it is the pointwise limit of $h_{n,p}$, which has this same property).

As for the variance term in (3.17), consider differentiating with respect to t, to yield

$$\begin{aligned}
\frac{d}{dt}\frac{\sigma^2}{n}\mathrm{tr}\big[\hat{\Sigma}^+(I - \exp(-t\hat{\Sigma}))^2\Sigma\big] &= \frac{2\sigma^2}{n}\mathrm{tr}\big[\hat{\Sigma}^+\hat{\Sigma}(I - \exp(-t\hat{\Sigma}))\exp(-t\hat{\Sigma})\Sigma\big] \\
&= \frac{2\sigma^2}{n}\mathrm{tr}\big[(I - \exp(-t\hat{\Sigma}))\exp(-t\hat{\Sigma})\Sigma\big],
\end{aligned}$$

with the second line following because the column space of $I - \exp(-t\hat{\Sigma})$ matches that of $\hat{\Sigma}$. The fundamental theorem of calculus then implies that the variance equals

$$\begin{aligned}
&\frac{2\sigma^2}{n}\int_0^t \mathrm{tr}\big[(\exp(-u\hat{\Sigma}) - \exp(-2u\hat{\Sigma}))\Sigma\big]\,du = \\
&\qquad \frac{2\sigma^2 p}{n}\int_0^t \Big[\mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma\big]\Big)(u) - \mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma\big]\Big)(2u)\Big]\,du,
\end{aligned}$$

where the equality is due to inverting the Laplace transform fact (6.35), as done in (6.36) for the bias. The same arugments for the bias now carry over here, to imply

$$\begin{aligned}
\lim_{n,p\to\infty} \frac{2\sigma^2}{n}\int_0^t \Big[\mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma\big]\Big)(u) - \mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma\big]\Big)(2u)\Big]\,du = \\
2\sigma^2\gamma\int_0^t \big(\mathcal{L}^{-1}(\theta)(u) - \mathcal{L}^{-1}(\theta)(2u)\big)\,du. \qquad (6.40)
\end{aligned}$$

Putting together (6.39) and (6.40) completes the proof.

6.2.11 Supporting Lemmas

Lemma 6.3. *For any real matrices $A, B \succeq 0$ and $t \geq 0$, define*

$$f(z) = \operatorname{tr}\big[(A + zI)^{-1}B\big] \exp(2tz),$$

over $z \in \mathbb{C}_+ = \{z \in \mathbb{C} : \operatorname{Im}(z) > 0\}$. Then $f(\overline{z}) = \overline{f(z)}$.

Proof. First note that $\exp(2t\overline{z}) = \overline{\exp(2tz)}$ by Euler's formula. As the conjugate of a product is the product of conjugates, it suffices to show that $\operatorname{tr}[(A + \overline{z}I)^{-1}B] = \overline{\operatorname{tr}[(A + zI)^{-1}B]}$. To this end, denote $C_z = (A + zI)^{-1}$, and denote by C_z^* its adjoint (conjugate transpose). Note that $\overline{\operatorname{tr}(C_z B)} = \operatorname{tr}(C_z^* B)$; we will show that $C_z^* = C_{\overline{z}}$, which would then imply the desired result. Equivalent to $C_z^* = C_{\overline{z}}$ is $\langle C_z x, y\rangle = \langle x, C_{\overline{z}} y\rangle$ for all complex vectors x, y (where $\langle \cdot, \cdot \rangle$ denotes the standard inner product). Observe

$$\begin{aligned}
\langle C_z x, y\rangle &= \langle C_z x,\, (A + \overline{z}I) C_{\overline{z}} y\rangle \\
&= \langle (A + \overline{z}I)^* C_z x,\, C_{\overline{z}} y\rangle \\
&= \langle (A + zI) C_z x,\, C_{\overline{z}} y\rangle \\
&= \langle x, C_{\overline{z}} y\rangle,
\end{aligned}$$

which completes the proof. □

Lemma 6.4. *If $f : \mathbb{C} \to \mathbb{C}$ satisfies $f(\overline{z}) = \overline{f(z)}$, then for any $a \in \mathbb{R}$,*

$$\int_{-\infty}^{\infty} f(a + ib)\, db = 2 \int_{-\infty}^{0} \operatorname{Re}(f(a + ib))\, db.$$

Proof. The property $f(\overline{z}) = \overline{f(z)}$ means that $\operatorname{Re}(f(a - ib)) = \operatorname{Re}(f(a + ib))$, and $\operatorname{Im}(f(a - ib)) = -\operatorname{Im}(f(a + ib))$. Thus

$$\begin{aligned}
\int_{-\infty}^{\infty} f(a + ib)\, db &= \int_{-\infty}^{\infty} \operatorname{Re}(f(a + ib))\, db + i \int_{-\infty}^{\infty} \operatorname{Im}(f(a + ib))\, db \\
&= 2 \int_{-\infty}^{0} \operatorname{Re}(f(a + ib))\, db + 0,
\end{aligned}$$

which completes the proof. □

6.2.12 Additional Numerical Results

Here we show the complete set of numerical results comparing gradient flow and ridge regression. The setup is as described in Section 4.6. Figure 6.1 shows the results for Gaussian features in the low-dimensional case ($n = 1000$, $p = 500$). The first row shows the estimation risk when $\Sigma = I$, with the left plot using $\lambda = 1/t$ calibration, and the right plot using ℓ_2 norm calibration (details on this calibration explained below). The second row shows the estimation risk when Σ has all off-diagonals equal to $\rho = 0.5$. The third row shows the prediction risk for the same Σ (n.b., the prediction risk when $\Sigma = I$ is the same as the estimation risk, so it is redundant to

show both). The conclusions throughout are similar to that made in Section 4.6. Calibration by ℓ_2 norm gives extremely good agreement: the maximum ratio of gradient flow to ridge risk (over the entire path, in any of the three rows) is 1.0367. Calibration by $\lambda = 1/t$ is still quite good, but markedly worse: the maximum ratio of gradient flow to ridge risk (again over the entire path, in any of the three rows) is 1.4158.

Figures 6.2 shows analogous results for Gaussian features in the high-dimensional case ($n = 500$, $p = 1000$). Figures 6.3–6.6 show the results for Student t and Bernoulli features. The results are similar throughout: the maximum ratio of gradient flow to ridge risk, under ℓ_2 norm calibration (over the entire path, in any setting), is 1.0371; the maximum ratio, under $\lambda = 1/t$ calibration (over the entire path, in any setting), is 1.4154. (One noticeable, but unremarkable difference between the settings is that the finite-sample risks seem to be converging slower to their asymptotic analogs in the case of t features. This is likely due to the fact that the tails here are very fat—they are as fat as possible for the t family, subject to the second moment being finite.)

It helps to give further details for a few of the calculations. For ℓ_2 norm calibration, note that we can compute the expected squared ℓ_2 norm of the ridge and gradient flow estimators under the data model (3.9) and prior (3.10):

$$\begin{aligned}
\mathbb{E}\|\hat{\beta}^{\mathrm{ridge}}(\lambda)\|_2^2 &= \frac{1}{n}\Big(\mathrm{tr}\big[\alpha(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}^2\big] + \mathrm{tr}\big[(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}\big]\Big) \\
&= \frac{1}{n}\sum_{i=1}^{p} \frac{\alpha s_i^2 + s_i}{(s_i+\lambda)^2}, \\
\mathbb{E}\|\hat{\beta}^{\mathrm{gf}}(t)\|_2^2 &= \frac{1}{n}\Big(\mathrm{tr}\big[\alpha(I-\exp(-t\hat{\Sigma}))^2\big] + \mathrm{tr}\big[(I-\exp(-t\hat{\Sigma}))^2\hat{\Sigma}^+\big]\Big) \\
&= \frac{1}{n}\sum_{i=1}^{p}\left(\alpha(1-\exp(-ts_i))^2 + \frac{(1-\exp(-ts_i))^2}{s_i}\right).
\end{aligned}$$

We thus calibrate according to the square root of the quantities above (this is what is plotted on the x-axis in the left columns of all the figures). The above expressions have the following limits under the asymptotic model studied in Theorem 3.5:

$$\begin{aligned}
\mathbb{E}\|\hat{\beta}^{\mathrm{ridge}}(\lambda)\|_2^2 &\to \gamma \int \frac{\alpha_0 s^2 + s}{(s+\lambda)^2}\, dF_{H,\gamma}(s), \\
\mathbb{E}\|\hat{\beta}^{\mathrm{gf}}(t)\|_2^2 &\to \gamma \int \left(\alpha_0(1-\exp(-ts))^2 + \frac{(1-\exp(-ts))^2}{s}\right) dF_{H,\gamma}(s).
\end{aligned}$$

Furthermore, we note that when $\Sigma = I$, the empirical spectral distribution from Theorem 3.4 abbreviated as F_γ, sometimes called the *Marchenko-Pastur (MP) law* and has a closed form. For $\gamma \le 1$, its density is

$$\frac{dF_\gamma(s)}{ds} = \frac{1}{2\pi\gamma s}\sqrt{(b-s)(s-a)},$$

and is supported on $[a, b]$, where $a = (1-\sqrt{\gamma})^2$ and $b = (1+\sqrt{\gamma})^2$. For $\gamma > 1$, the MP law F_γ has an additional point mass at zero of probability $1 - 1/\gamma$. This allows us to evaluate the integrals in (3.20), (3.23) via numerical integration, to compute limiting risks for gradient flow

and ridge regression. (It also allows us to compute the integrals in the second to last display, to calibrate according to limiting ℓ_2 norms.)

6.3 Supplementary Material for The Multiple Quantile Graphical Model

6.3.1 Proof of Lemma 4.1

If the conditional quantiles satisfy $Q_{U|V,W}(\alpha) = Q_{U|W}(\alpha)$ for all $\alpha \in [0,1]$, then the conditional CDF must obey the same property, i.e., $F_{U|V,W}(t) = F_{U|W}(t)$ for all t in the support of U. This is simply because any CDF may be expressed in terms of its corresponding quantile function (i.e., inverse CDF), as in

$$F_{U|V,W}(t) = \sup\{\alpha \in [0,1] : Q_{U|V,W}(\alpha) \leq t\},$$

and the right-hand side does not depend on V, so neither can the left-hand side. But this precisely implies that the distribution of $U|V,W$ equals that of $U|W$, i.e., U and V are conditionally independent given W. We note that the converse of the statement in the lemma is true as well, by just reversing all the arguments here. □

6.3.2 Proof of Lemma 4.2

This result can be seen as a generalization of Theorem 3 in [53].

First, we define an *iteration* of Gibbs sampling to be a single pass through all the variables (without a loss of generality, we take this order to be $y_1, \ldots, y_d$). Now, consider a particular iteration of Gibbs sampling; let $\tilde{y}_1, \ldots, \tilde{y}_d$ be the values assigned to the variables on the previous iteration. Then the transition kernel for our Gibbs sampler is given by

$$\begin{aligned}
\mathbb{P}(y_1, \ldots, y_d | \tilde{y}_1, \ldots, \tilde{y}_d) &= \mathbb{P}(y_d | y_{d-1}, \ldots, y_1, \tilde{y}_1, \ldots, \tilde{y}_d) \mathbb{P}(y_{d-1}, \ldots, y_1 | \tilde{y}_1, \ldots, \tilde{y}_d) && (6.41)\\
&= \mathbb{P}(y_d | y_{d-1}, \ldots, y_1) \mathbb{P}(y_{d-1}, \ldots, y_1 | \tilde{y}_1, \ldots, \tilde{y}_d) && (6.42)\\
&= \mathbb{P}(y_d | y_{d-1}, \ldots, y_1) \mathbb{P}(y_{d-1} | y_{d-2}, \ldots y_1, \tilde{y}_d) \cdots \mathbb{P}(y_1 | \tilde{y}_2, \ldots, \tilde{y}_d), && (6.43)
\end{aligned}$$

where (6.41) follows by the definition of conditional probability, (6.42) by conditional independence, and (6.43) by repeated applications of these tools. Since each conditional distribution is assumed to be (strictly) positive, we have that the transition kernel is also positive, which in turn implies [14, page 544] that the induced Markov chain is ergodic with a unique stationary distribution that can be reached from any initial point. □

6.3.3 Statement and Discussion of Regularity Conditions for Theorem 4.1

For each $k = 1, \ldots, r$, $\ell = 1, \ldots, r$, let us define the "effective" (independent) error terms $\epsilon_{\ell k i} = y_k^{(i)} - b^*_{\ell k} - \sum_{j \neq k} \phi(y_j^{(i)})^T \theta^*_{\ell k j}$, over $i = 1, \ldots, n$. Denote by $F_{\epsilon_{\ell k}}$ the conditional CDF of

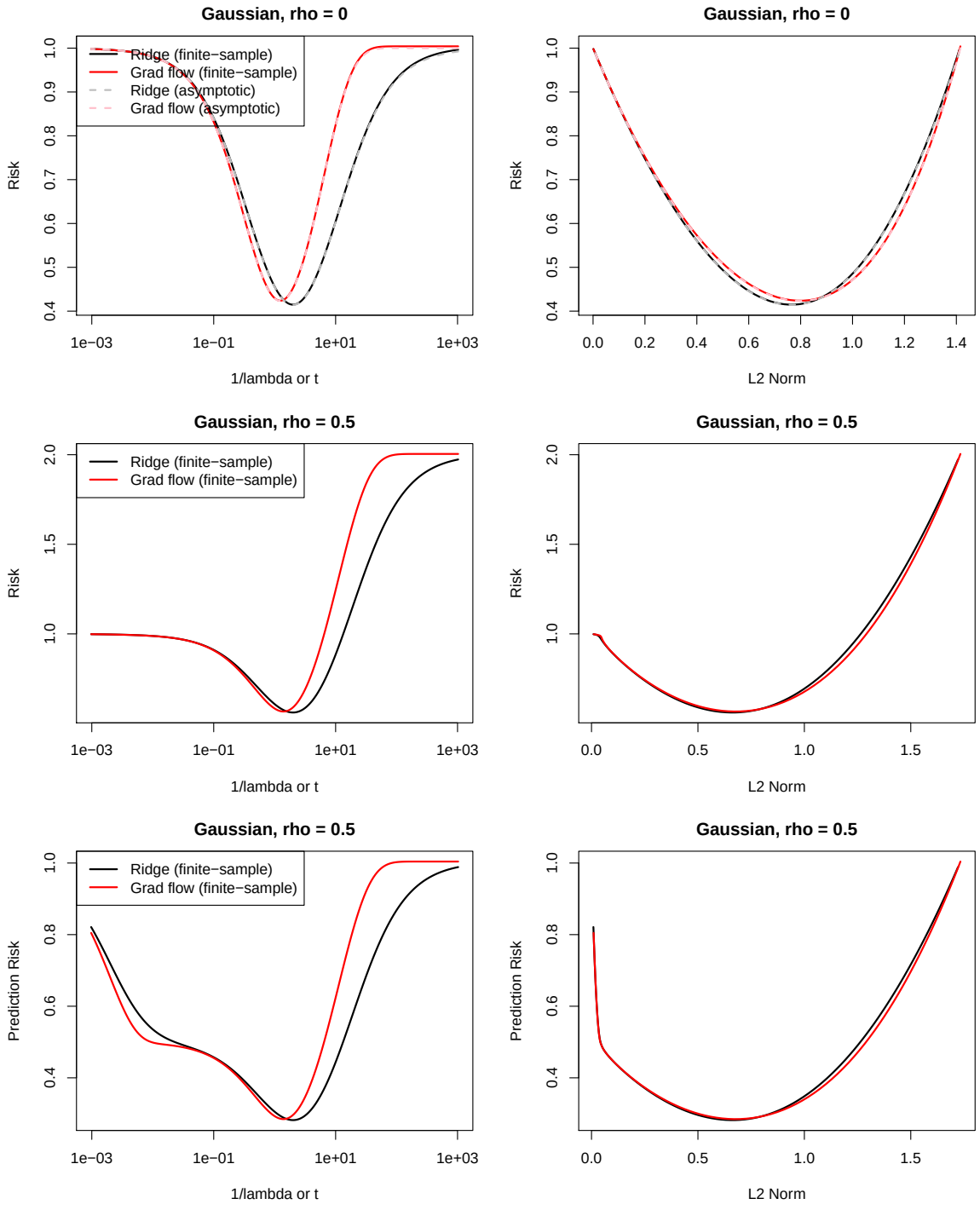

Figure 6.1: *Gaussian features, with* $n = 1000$ *and* $p = 500$.

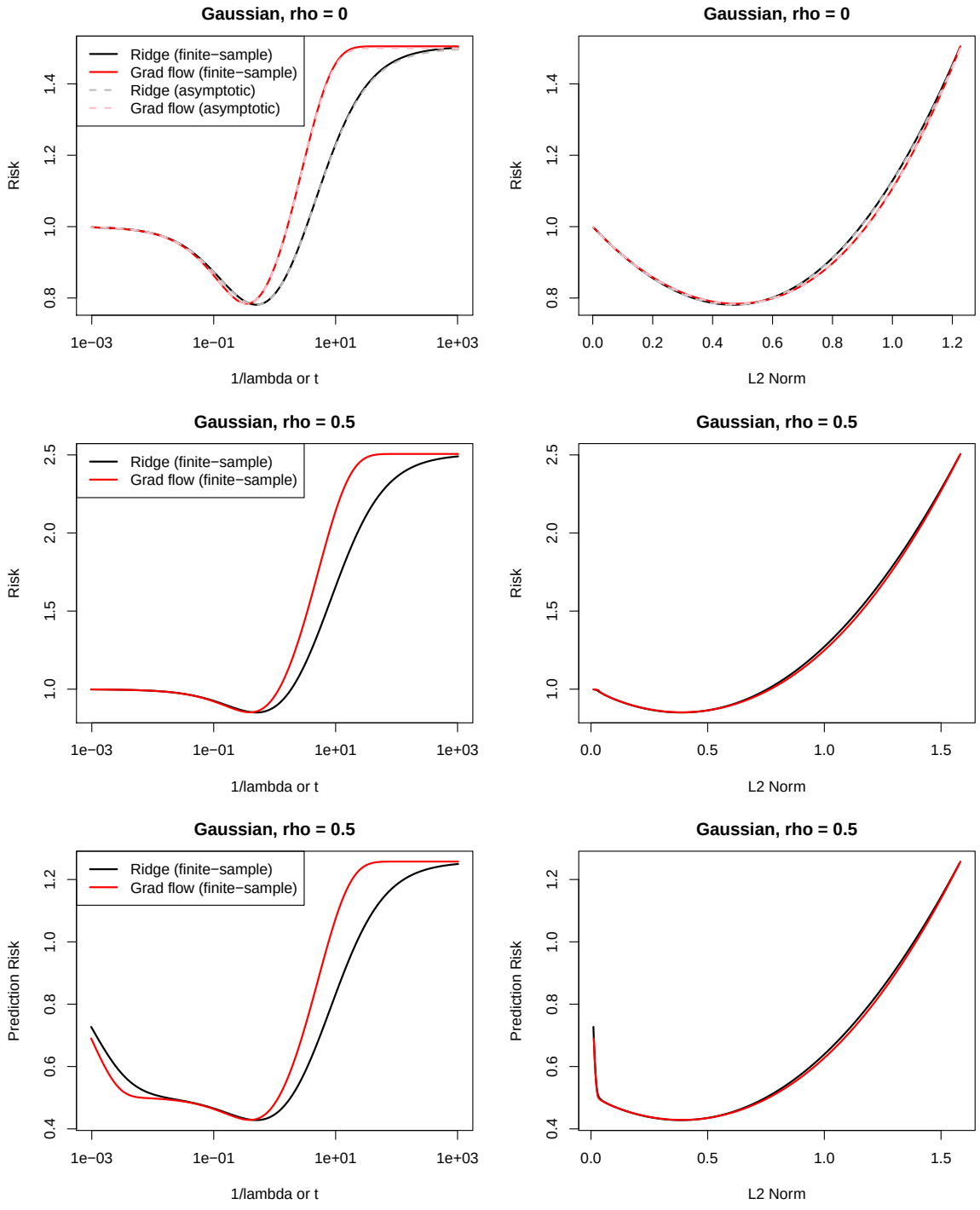

Figure 6.2: *Gaussian features, with* $n = 500$ *and* $p = 1000$.

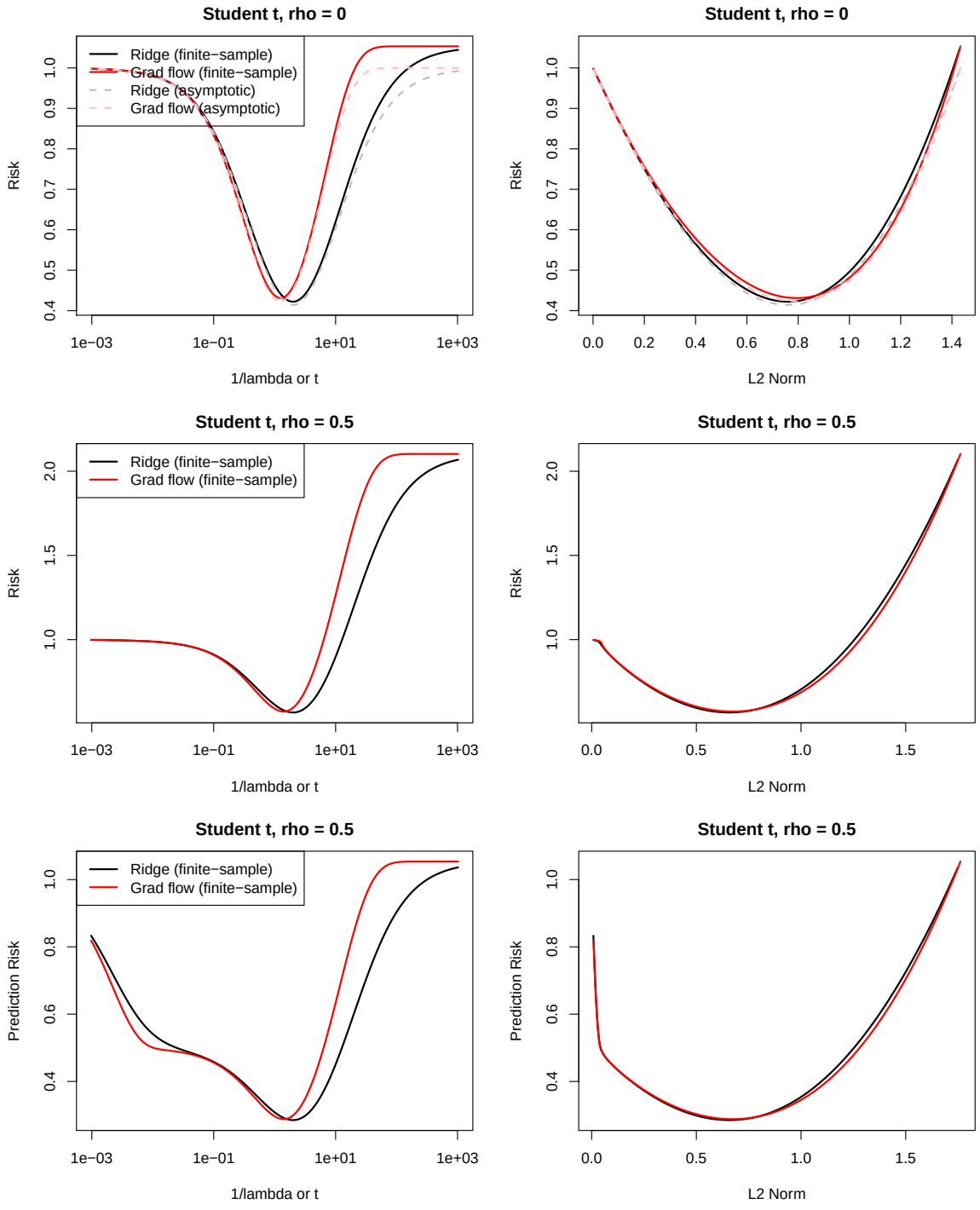

Figure 6.3: *Student t features, with* $n = 1000$ *and* $p = 500$.

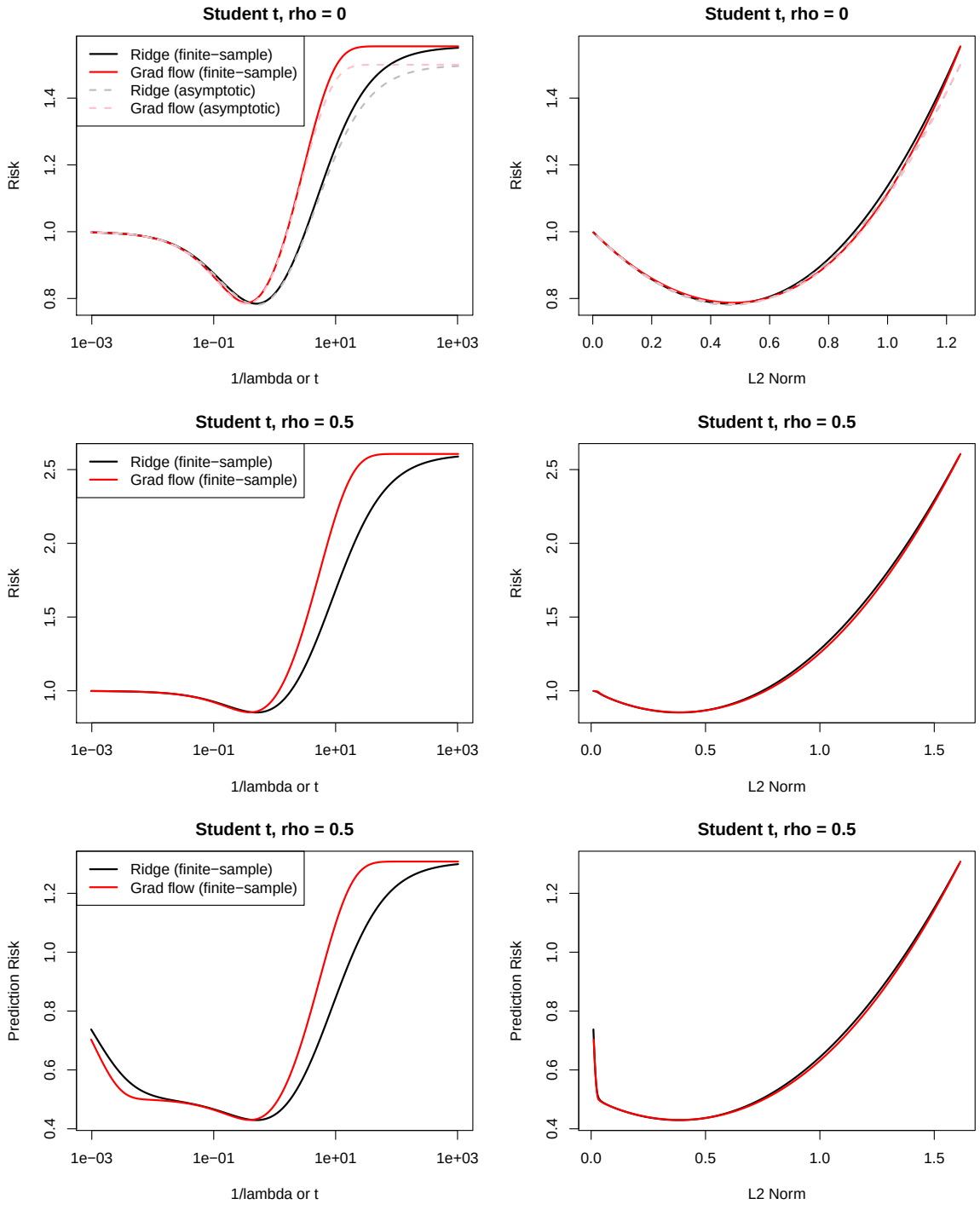

Figure 6.4: *Student t features, with* $n = 500$ *and* $p = 1000$.

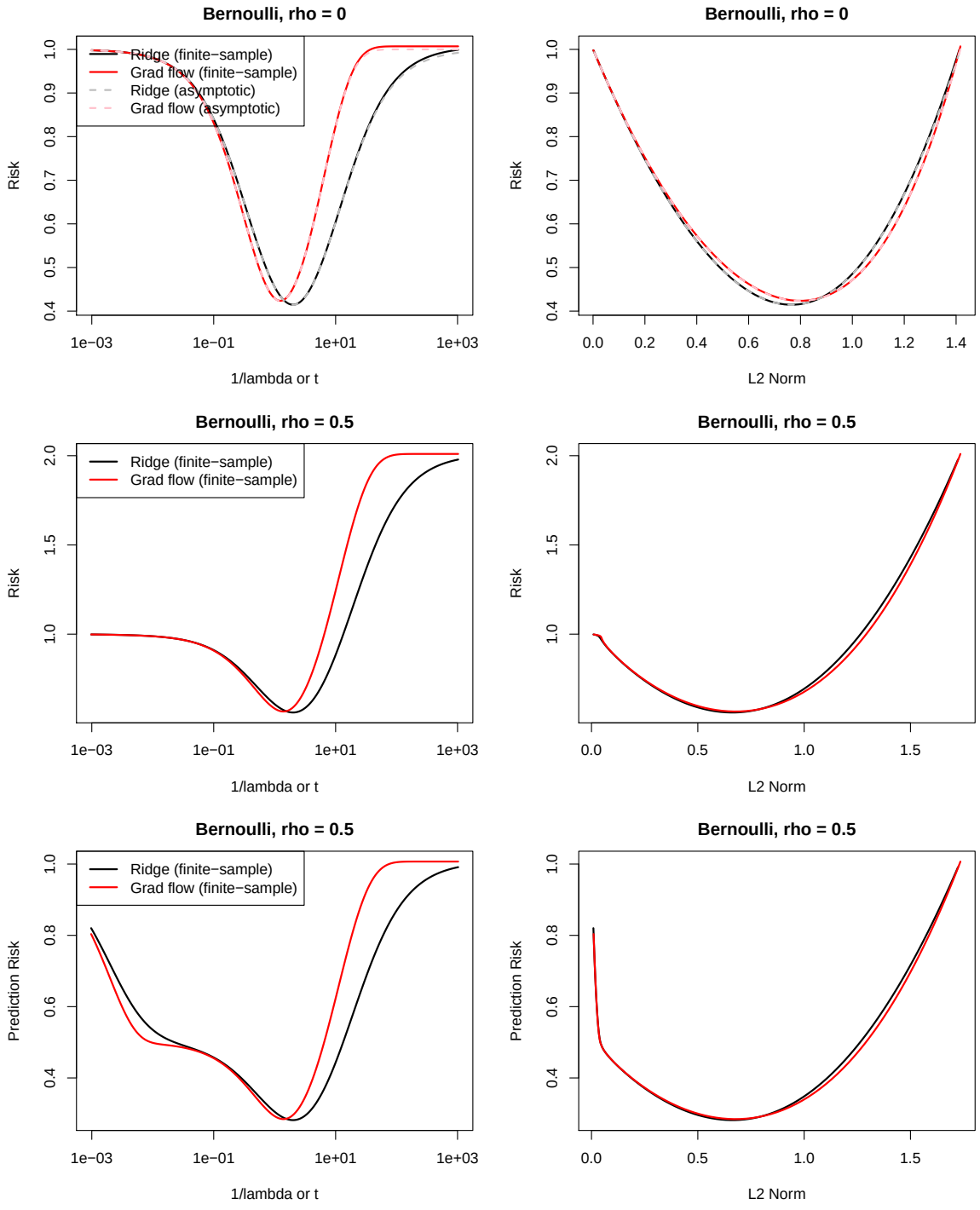

Figure 6.5: *Bernoulli features, with* $n = 1000$ *and* $p = 500$.

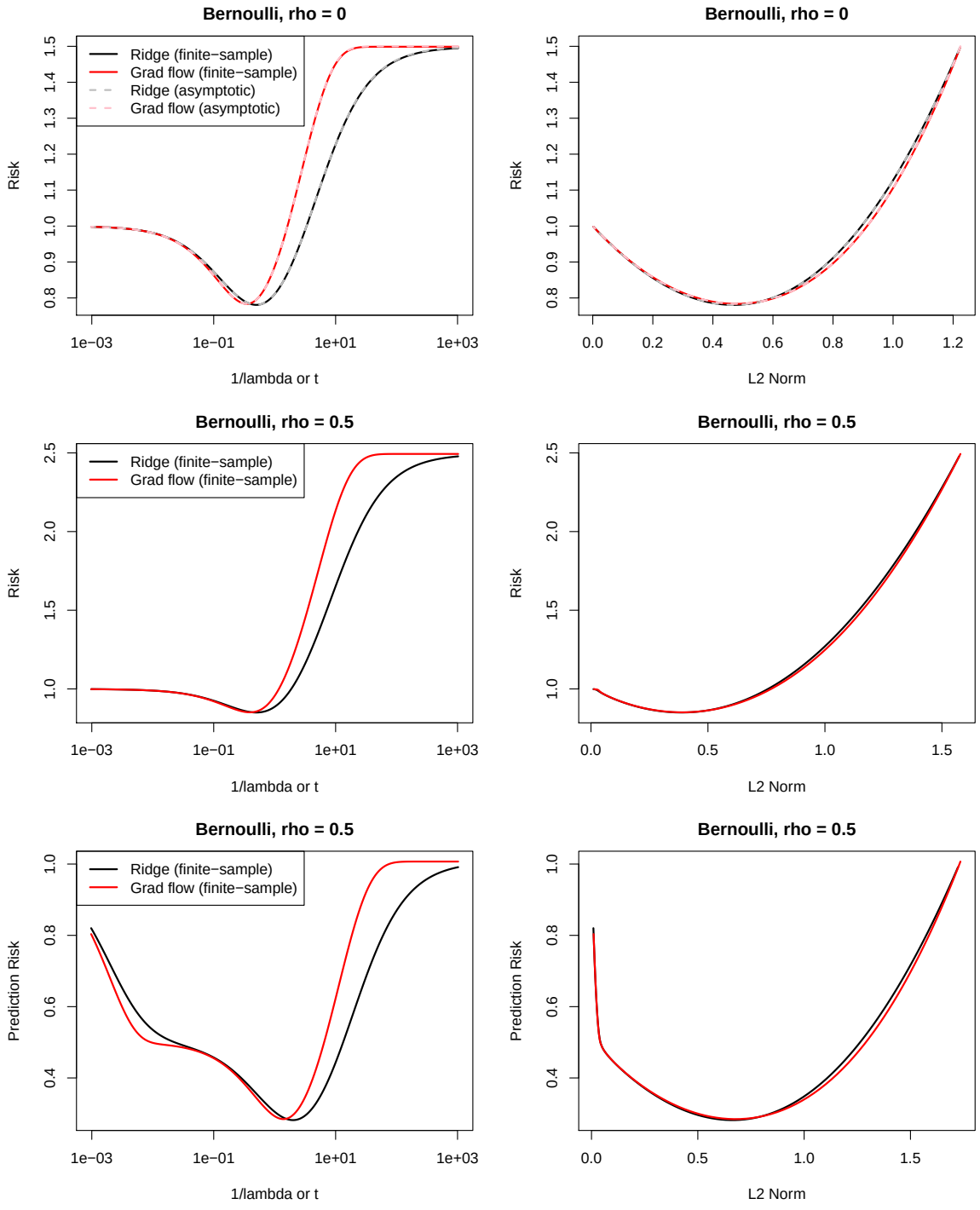

Figure 6.6: *Bernoulli features, with* $n = 500$ *and* $p = 1000$.

$\epsilon_{\ell ki}|y^{(i)}_{-k}$, $i = 1, \ldots, n$, which by construction satisfies $F_{\epsilon_{\ell k}}(0) = \alpha_\ell$. Also define the underlying support

$$S_{\ell k} = \big\{j \in \{1, \ldots, d\} : \theta^*_{\ell kj} \neq 0\big\}.$$

Here we take a moment to explain a somewhat subtle indexing issue with the columns of the feature matrix $\Phi \in \mathbb{R}^{n \times dm}$. For a single fixed index $j = 1, \ldots, d$, we will extract an appropriate block of columns of $\Phi \in \mathbb{R}^{n \times dm}$, corresponding to the basis expansion of variable j, by writing Φ_j. More precisely, we use Φ_j to denote the block of m columns

$$[\Phi_{(j-1)m+1}, \Phi_{(j-1)m+2}, \ldots, \Phi_{jm}]. \tag{6.44}$$

We do this because it simplifies notation considerably. (Occasionally, to be transparent, we will use the more exhaustive notation on the right-hand side in (6.44), but this is to be treated as an exception, and the default is to use the concise notation as in Φ_j.) The same rule will be used for subsets of indices among $1, \ldots, d$, so that $\Phi_{S_{\ell k}}$ denotes the appropriate block of $m|S_{\ell k}|$ columns corresponding to the basis expansions of the variables in $S_{\ell k}$.

For all $k = 1, \ldots, d$, $\ell = 1, \ldots, r$, we will assume the following regularity conditions.

A1. *Groupwise irrepresentability:* for $j \in S^c_{\ell k}$, we require that $\|\Phi_j^T \Phi_{S_{\ell k}}\|_F < \lambda_1/(6 f_{\epsilon_{\ell k}}(0)\gamma)$, where $S_{\ell k} = \{j \in \{1, \ldots, dm\} : \theta^*_{\ell kj} \neq 0\}$, $f_{\epsilon_{\ell k}}$ is the density of $F_{\epsilon_{\ell k}}$, and $\gamma > 0$ is a quantity prescribed by Lemma 6.3.5.

A2. *Distributional smoothness:* we assume that $|F_{\epsilon_{\ell k}}(x) - F_{\epsilon_{\ell k}}(0) - x f_{\epsilon_{\ell k}}(0)| \leq C_1 x^2$ for all $|x| \leq C_2$, where $C_1, C_2 > 0$ are constants.

A3. *Correlation restriction:* we assume that $C_3 \leq (f_{\epsilon_{\ell k}}(0)/n)\lambda_{\min}(\Phi^T_{S_{\ell k}} \Phi_{S_{\ell k}}) \leq C_4$ for constants $C_3, C_4 > 0$, where $\lambda_{\min}(A)$ denotes the minimum eigenvalue of A.

A4. *Basis and support size restrictions:* we assume that $m = O(n^{1/9})$ and $s = O(n^{1/21})$, where $s = |S_{\ell k}|$. We also assume, with probability tending to one, that $\Phi_{\max} = \Omega(1)$ and $\Phi_{\max} = o(n^{1/21}/\log^{1/2} n)$, where we write $\Phi_{\max}$ to denote the maximum absolute entry of the basis matrix Φ.

Next, we provide some intuition for these conditions.

Condition A1. Fix some $j \in S^c_{\ell k}$. For notational convenience, we let

$$A = \Phi_j^T \Phi_{S_{\ell k}} \in \mathbb{R}^{m \times sm}.$$

Observe that each entry of A can be expressed as

$$A_{ip} = n\rho_{i,p} \|\Phi_{(j-1)m+i}\|_2 \|\Phi_p\|_2, \tag{6.45}$$

for $i = 1, \ldots, m$, p denoting an index into the basis expansion of the columns $\Phi_{S_{\ell k}}$, and $\rho_{i,p}$ denoting the sample correlation coefficient for the columns Φ_i and Φ_p. Since $\|A_p\|_F \leq \sqrt{m}\|A_p\|_\infty$, we have that

$$\max_{i,p} \ \rho_{i,p} < \frac{\lambda_1}{6n^2 f_{\epsilon_{\ell k}}(0)\sqrt{m}}$$

is sufficient for condition A1; here, we have also used the column scaling assumption $\|\Phi_p\|_2 \leq \sqrt{n}$.

So, roughly speaking, bounded correlation between each pair of columns in the submatrices Φ_j and $\Phi_{S_{\ell k}}$ is enough for condition A1 to hold; note that this is trivially satisfied when $\Phi_i^T \Phi_p = 0$, for $i = 1, \ldots, m$, and p as defined above. Condition A1 is therefore similar to, e.g., the mutual incoherence condition of [141] for the lasso, which is given by

$$\left\| \Phi_{S^c}^T \Phi_S \left(\Phi_S^T \Phi_S \right)^{-1} \right\|_\infty \leq 1 - \tilde{\gamma} \iff \max_{j \in S^c} \left\| \left(\Phi_S^T \Phi_S \right)^{-1} \Phi_S^T \Phi_j \right\|_1 \leq 1 - \tilde{\gamma},$$

where again Φ_S extracts the appropriate block of columns of Φ, $\| \cdot \|_\infty$ here denotes the ℓ_∞ operator norm (maximum ℓ_1 norm of a row), $\| \cdot \|_1$ here denotes the elementwise ℓ_1 norm, and $\tilde{\gamma} \in (0, 1]$ is a constant. This condition can be seen as requiring bounded correlation between each column in the submatrix Φ_{S^c} and all columns in the submatrix Φ_S.

Condition A2. This condition is similar to requiring that $f_{\epsilon_{\ell k}}(x)$ be Lipschitz, over some x in a neighborhood of 0. We can show that the Laplace distribution, e.g., satisfies this condition.

The density and distribution functions for the Laplace distribution with location zero and unit scale are given by

$$f_{\epsilon_{\ell k}}(x) = (1/2) \exp(-|x|)$$

and

$$F_{\epsilon_{\ell k}}(x) = \begin{cases} 1 - (1/2) \exp(-x) & \text{if } x \geq 0 \\ (1/2) \exp(x) & \text{if } x < 0, \end{cases}$$

respectively.

Now, suppose $0 \leq x \leq C_2$. Then we can express condition A2 as

$$|f_{\epsilon_{\ell k}}(x) - f_{\epsilon_{\ell k}}(0) - x f_{\epsilon_{\ell k}}(0)| \leq C_1 x^2 \iff -2C_1 x^2 \leq \exp(-x) + x - 1 \leq 2C_1 x^2.$$

For the first inequality, since $1 - x \leq \exp(-x)$, it is sufficient to check that $0 \leq C_1 x^2$, which is true for $C_1 > 0$ and all x. For the second inequality, by differentiating and again using $1 - x \leq \exp(-x)$, we have that the function

$$2C_1 x^2 - \exp(-x) - x + 1 \tag{6.46}$$

is nondecreasing in $x \geq 0$; thus, it is sufficient to check that this function is nonnegative for $x = 0$, which is true.

Now, suppose $-C_2 \leq x < 0$. Then we can express condition A2 as

$$|f_{\epsilon_{\ell k}}(x) - f_{\epsilon_{\ell k}}(0) - x f_{\epsilon_{\ell k}}(0)| \leq C_1 x^2 \iff -2C_1 x^2 \leq \exp(x) - x - 1 \leq 2C_1 x^2.$$

By symmetry with the preceding case, the first inequality here holds. The second inequality here also holds, since $\exp(x) - 2C_1 x^2 - x - 1$ is continuous and increasing in $x < 0$; taking the limit as $x \uparrow 0$ gives that this function is nonpositive as required.

Condition A3. This condition is a generalization of the minimum eigenvalue condition of [141], i.e., $c_{\min} \leq \lambda_{\min} \left((1/n) \Phi_S^T \Phi_S \right)$, for some constant $c_{\min} > 0$, and where we write Φ_S to extract the appropriate block of columns of Φ.

Condition A4. This condition allows the number of basis functions m in the expansion to grow with n, at a polynomial rate (with fractional exponent). This is roughly in line with standard nonparametric regression; e.g., when estimating a continuous differentiable function via a spline expansion, one typically takes the number of basis functions m to scale as $n^{1/3}$, and the more derivatives that are assumed, the smaller the fractional exponent [50]. The condition also restricts, for any given variable, the number of variables s that contribute to its neighborhood model to be polynomial in n (with a smaller fractional exponent).

Finally, the condition assumes that the entries of the basis matrix Φ (i.e., the matrix of transformed variables) to be at least of constant order, and at most of polynomial order (with small fractional exponent), with n. We note that this implicitly places a restriction on the tails of distribution governing the data $y_j^{(i)}$, $i = 1, \ldots, n$, $j = 1, \ldots, d$. However, the restriction is not a strong one, because it allows the maximum to grow polynomially large with n (whereas a logarithmic growth would be expected, e.g., for normal data). Furthermore, it is possible to trade off the restrictions on m, s, $\Phi_{\max}$, and d (presented in the statement of the theorem), making each of these restrictions more or less stringent, if required.

6.3.4 Proof of Theorem 4.1

The general strategy that we use here for support recovery is inspired by that in [37], for ℓ_1-penalized quantile regression.

Fix some $k = 1, \ldots, d$ and $\ell = 1, \ldots, r$. We consider the conditional distribution $y_k|y_{\neg k}$, whose α_ℓ-quantile is assumed to satisfy (4.3). Hence, to be perfectly clear, all expectations and probability statements in what follows are to be interpreted with respect to the observations $y_k^{(i)}$, $i = 1, \ldots, n$ conditional on $y_j^{(i)}$, $i = 1, \ldots, n$, for $j \neq k$ (and thus we can treat the feature matrix Φ as fixed throughout). In the setting assumed by the theorem, the conditional quantile model in (4.3) is, more explicitly,

$$Q_{y_k|y_{\neg k}}(\alpha_\ell) = b^*_{\ell k} + \sum_{j \neq k}^{d} (\theta^*_{\ell kj})^T \phi^j(y_j),$$

for some unknown parameters $b^*_{\ell k}$ and $\theta^*_{\ell kj}$, $j = 1, \ldots, d$. For simplicity, in this proof, we will drop the intercept term completely both from the model (denoted $b^*_{\ell k}$) and the optimization problem in (4.4) (here denoted $b_{\ell k}$) that defines the estimator in question. Including the intercept is not at all difficult, and it just requires some extra bookkeeping at various places. Recall that we define

$$S_{\ell k} = \big\{ j \in \{1, \ldots, d\} : \theta^*_{\ell kj} \neq 0 \big\},$$

and analogously define

$$\hat{S}_{\ell k} = \big\{ j \in \{1, \ldots, d\} : \hat{\theta}_{\ell kj} \neq 0 \big\},$$

where $\hat{\theta}_{\ell k} = (\hat{\theta}_{\ell k1}, \ldots, \hat{\theta}_{\ell kd}) \in \mathbb{R}^{dm}$ is the solution in (4.5).

We will show that, with probability at least $1 - \delta/(dr)$, it holds that $S_{\ell k} = \hat{S}_{\ell k}$. A union bound (over all choices $k = 1, \ldots, d$ and $\ell = 1, \ldots, r$) will then tell us that $E^* = \hat{E}$ with probability at least $1 - \delta$, completing the proof.

To certify that $S_{\ell k} = \hat{S}_{\ell k}$, we will show that the unique solution in (4.5) is given by

$$\hat{\theta}_{\ell k(S_{\ell k})} = \tilde{\theta}_{\ell k(S_{\ell k})}, \quad \hat{\theta}_{\ell k(S_{\ell k}^c)} = 0, \tag{6.47}$$

where $\tilde{\theta}_{\ell k(S_{\ell k})}$ solves the "restricted" optimization problem:

$$\underset{\theta_{\ell k(S_{\ell k})}}{\text{minimize}} \ \psi_{\alpha_\ell}\Big(Y_k - \Phi_{S_{\ell k}}\theta_{\ell k(S_{\ell k})}\Big) + \lambda_1 \sum_{j \in S_{\ell k}} \|\theta_{\ell k j}\|_2 + \frac{\lambda_2}{2}\|\theta_{\ell k(S_{\ell k})}\|_2^2. \tag{6.48}$$

Now, to prove that $\hat{\theta}_{\ell k}$ as defined above in (6.47) indeed the solution in (4.5), we need to check that it satisfies the KKT conditions for (4.5), namely

$$\Phi_{S_{\ell k}}^T v_\ell\Big(Y_k - \Phi_{S_{\ell k}}\tilde{\theta}_{\ell k(S_{\ell k})}\Big) - \lambda_2\tilde{\theta}_{\ell k(S_{\ell k})} = \lambda_1 u_{\ell k(S_{\ell k})}, \tag{6.49}$$

$$\Phi_{S_{\ell k}^c}^T v_\ell\Big(Y_k - \Phi_{S_{\ell k}}\tilde{\theta}_{\ell k(S_{\ell k})}\Big) = \lambda_1 u_{\ell k(S_{\ell k}^c)}, \tag{6.50}$$

where $v_\ell(Y_k - \Phi_{S_{\ell k}}\tilde{\theta}_{\ell k(S_{\ell k})}) \in \mathbb{R}^n$ is a subgradient of $\psi_{\alpha_\ell}(\cdot)$ at $Y_k - \Phi_{S_{\ell k}}\tilde{\theta}_{\ell k(S_{\ell k})}$, i.e.,

$$\Big[v_\ell\Big(Y_k - \Phi_{S_{\ell k}}\tilde{\theta}_{\ell k(S_{\ell k})}\Big)\Big]_i = \alpha_\ell - I_-\Big(y_k^{(i)} - \Phi_{i(S_{\ell k})}\tilde{\theta}_{\ell k(S_{\ell k})}\Big), \quad i = 1, \ldots, n$$

where $I_-(\cdot)$ is the indicator function of the nonpositive real line, and where each $u_{\ell k j} \in \mathbb{R}^m$ is a subgradient of $\|\cdot\|_2$ at $\tilde{\theta}_{\ell k j}$, i.e.,

$$u_{\ell k j} \in \begin{cases} \{\tilde{\theta}_{\ell k j}/\|\tilde{\theta}_{\ell k j}\|_2\} & \text{if } \theta_{\ell k j} \neq 0 \\ \{x \in \mathbb{R}^m : \|x\|_2 \leq 1\} & \text{if } \theta_{\ell k j} = 0, \end{cases}$$

for $j = 1, \ldots, d$. Note that, since $\tilde{\theta}_{\ell k(S_{\ell k})}$ is optimal for the restricted problem (6.48), we know that there exists a collection of subgradients $u_{\ell k(S_{\ell k})}$ to satisfy (6.49), from the KKT conditions for (6.48) itself.

It remains to satisfy (6.50), and for this, we can use $u_{\ell k j} = \Phi_j^T v_\ell(Y_k - \Phi_{S_{\ell k}}\tilde{\theta}_{\ell k(S_{\ell k})})$ as a valid choice of subgradient, for each $j \in S_{\ell k}^c$, provided that

$$\Big\|\Phi_j^T v_\ell\Big(Y_k - \Phi_{S_{\ell k}}\tilde{\theta}_{\ell k(S_{\ell k})}\Big)\Big\|_2 < \lambda_1, \quad \text{for } j \in S_{\ell k}^c. \tag{6.51}$$

Define $z_j(\vartheta) = \Phi_j^T v_\ell(Y_k - \Phi_{S_{\ell k}}\vartheta)$, for $j \in S_{\ell k}^c$, and define a ball

$$B^* = \big\{\vartheta \in \mathbb{R}^{sm} : \|\vartheta - \theta^*_{\ell k(S_{\ell k})}\|_2 \leq \gamma\big\},$$

where we write $s = |S_{\ell k}|$. To show (6.51), then, it suffices to show that

$$\underbrace{\tilde{\theta}_{\ell k(S_{\ell k})} \in B^*}_{E_1}, \quad \text{and} \quad \underbrace{\max_{j \in S_{\ell k}^c} \sup_{\vartheta \in B^*} \|z_j(\vartheta)\|_2 < \lambda_1}_{E_2}. \tag{6.52}$$

In Lemma 6.5, given in Section 6.3.5, it is shown that the event E_1 defined above occurs with probability at least $1 - \delta/(2dr)$, with a choice of radius

$$\gamma = C\left(\frac{\lambda_1 s\sqrt{m}}{n} + \sqrt{\frac{s\log n}{n}}\right),$$

for a constant $C > 0$. Below we show that E_2 occurs with probability at least $1 - \delta/(2dr)$, as well.

For $j = 1, \ldots, d$, let us expand

$$z_j(\vartheta) = \underbrace{\Phi_j^T v_\ell(\epsilon_{\ell k})}_{\Delta_1^j} + \underbrace{\Phi_j^T \mathbb{E}\Big[v_\ell\Big(Y_k - \Phi_{S_{\ell k}}\vartheta\Big) - v_\ell(\epsilon_{\ell k})\Big]}_{\Delta_2^j} + \underbrace{\Phi_j^T\bigg(v_\ell\Big(Y_k - \Phi_{S_{\ell k}}\vartheta) - v_\ell(\epsilon_{\ell k}\Big) - \mathbb{E}\Big[v_\ell(Y_k - \Phi_{S_{\ell k}}\vartheta) - v_\ell(\epsilon_{\ell k})\Big]\bigg)}_{\Delta_3^j}, \quad (6.53)$$

where $\epsilon_{\ell k} = (\epsilon_{\ell k 1}, \ldots, \epsilon_{\ell k n}) \in \mathbb{R}^n$ is a vector of the effective error terms, which recall, is defined by $\epsilon_{\ell k} = Y_k - \Phi\theta^*_{\ell k}$. Therefore, to show that the event E_2 in (6.52) holds, we can show that for each $p = 1, 2, 3$,

$$\max_{j \in S^c_{\ell k}} \sup_{\vartheta \in B^*} \|\Delta_p^j\|_2 < \frac{\lambda_1}{3}.$$

Further, to show that E_2 holds with probability at least $1 - \delta/(2dr)$, we can show that the above holds for $p = 1, 3$ each with probability at least $1 - \delta/(4dr)$, as the statement for $p = 2$ is deterministic. We now bound the terms $\Delta_1^j, \Delta_2^j, \Delta_3^j$ one by one.

Bounding $\|\Delta_1^j\|_2$. Fix $j \in S^c_{\ell k}$, and write

$$\Phi_j^T v_\ell(\epsilon_{\ell k}) = \bigg(\sum_{i=1}^n \Phi_{i,(j-1)m+1} v_\ell(\epsilon_{\ell k i}), \ldots, \sum_{i=1}^n \Phi_{i,jm} v_\ell(\epsilon_{\ell k i})\bigg),$$

where, as a reminder that the above quantity is a vector, we have returned momentarily to the more exhaustive notation for indexing the columns of Φ, as in the right-hand side of (6.44).

Straightforward calculations reveal that, for each $i = 1, \ldots, n$, and $p = 1, \ldots, m$,

$$\mathbb{E}\Phi_{i,(j-1)m+p} v_\ell(\epsilon_{\ell k i}) = 0, \quad \text{and} \quad -|\Phi_{i,(j-1)m+p}| \le \Phi_{i,(j-1)m+p} v_\ell(\epsilon_{\ell k i}) \le |\Phi_{i,(j-1)m+p}|.$$

Hence,

$$\begin{aligned}
\mathbb{P}\Big(\|\Phi_j^T v_\ell(\epsilon_{\ell k i})\|_2 \ge \sqrt{m}t\Big) &\le \mathbb{P}\bigg(\Big|\sum_{i=1}^n \Phi_{i,(j-1)m+p} v_\ell(\epsilon_{\ell k i})\Big| \ge t, \ \text{ some } p = 1, \ldots, m\bigg) \\
&\le \sum_{p=1}^m 2\exp\bigg(-\frac{t^2}{2\sum_{i=1}^n \Phi^2_{i,(j-1)m+p}}\bigg) \\
&\le 2m\exp\bigg(-\frac{t^2}{2n}\bigg).
\end{aligned}$$

Above, the first inequality used the simple fact that $\|x\|_2 \le \sqrt{m}\|x\|_\infty$ for $x \in \mathbb{R}^m$; the second used Hoeffding's bound and the union bound; and the third used our assumption that the columns

of Φ have norm at most $\sqrt{n}$. Therefore, taking $t = \lambda_1/(3\sqrt{m})$, we see that, by the above and the union bound,

$$\mathbb{P}\bigg(\max_{j \in S_{\ell k}^c} \|\Delta_1^j\|_2 < \frac{\lambda_1}{3} \bigg) \geq 1 - 2dm \exp\bigg(- \frac{\lambda_1^2}{18mn} \bigg).$$

By choosing $\lambda_1 = C' \sqrt{18mn \log(8d^2mr/\delta)}$ for a constant $C' > 0$, we see that the probability in question is at least $1 - \delta/(4dr)$, as desired.

Bounding $\|\Delta_2^j\|_2$. Recall that $F_{\epsilon_{\ell k}}(\cdot)$ is used to denote the CDF of the effective error distribution, and $f_{\epsilon_{\ell k}}(\cdot)$ is used for its its density. By construction, $F_{\epsilon_{\ell k}}(0) = \alpha_\ell$. Direct calculation, using the definition of $v_\ell(\cdot)$, shows that, for any $\vartheta \in B^*$, and each $i = 1, \ldots, n$,

$$\mathbb{E}\Big[v_\ell(\epsilon_{\ell k}) - v_\ell\Big(Y_k - \Phi_{S_{\ell k}} \vartheta \Big) \Big] = F_{\epsilon_{\ell k}}\Big(\Phi_{S_{\ell k}} \big(\vartheta - \theta^*_{\ell k(S_{\ell k})} \big) \Big) - F_{\epsilon_{\ell k}}(\mathbf{0}),$$

where we apply $F_{\epsilon_{\ell k}}$ componentwise, and so

$$\Phi_j^T \mathbb{E}\Big[v_\ell(\epsilon_{\ell k}) - v_\ell\Big(Y_k - \Phi_{S_{\ell k}} \vartheta \Big) \Big] = f_{\epsilon_{\ell k}}(0) \Phi_j^T \Phi_{S_{\ell k}} \big(\vartheta - \theta^*_{\ell k(S_{\ell k})} \big) + \Delta_4^j$$

with $\Delta_4^j \in \mathbb{R}^m$ being the appropriate remainder term, i.e.,

$$\big[\Delta_4^j\big]_t = \sum_{i=1}^n \Phi_{it} \Big[F_{\epsilon_{\ell k}}\Big(\Phi_{i(S_{\ell k})} \big(\vartheta - \theta^*_{\ell k(S_{\ell k})} \big) \Big) - F_{\epsilon_{\ell k}}(0) - f_{\epsilon_{\ell k}}(0) \Phi_{i(S_{\ell k})} \big(\vartheta - \theta^*_{\ell k(S_{\ell k})} \big) \Big],$$

for $t = j(m-1) + 1, \ldots, jm$.

Now, we have that

$$\big\| f_{\epsilon_{\ell k}}(0) \Phi_j^T \Phi_{S_{\ell k}} \big(\vartheta - \theta^*_{\ell k(S_{\ell k})} \big) \big\|_2 \leq f_{\epsilon_{\ell k}}(0) \big\| \Phi_j^T \Phi_{S_{\ell k}} \big\|_F \big\| \vartheta - \theta^*_{\ell k(S_{\ell k})} \big\|_2 \leq \frac{\lambda_1}{6},$$

where we have used $\|\vartheta - \theta^*_{\ell k(S_{\ell k})}\|_2 \leq \gamma$ and the groupwise irrepresentability condition in A1.

We also have the following two facts, which we will use momentarily:

$$\Phi_{\max}^3 ns\gamma^2 = o(\lambda_1) \tag{6.54}$$

$$\sqrt{s} \Phi_{\max} \gamma \to 0. \tag{6.55}$$

Note that (6.54) can be obtained as follows. Since $(1/2)(x+y)^2 \leq x^2 + y^2$ for $x, y \in \mathbb{R}$, we can plug in

$$\gamma = C \left(\frac{\lambda_1 s \sqrt{m}}{n} + \sqrt{\frac{s \log n}{n}} \right),$$

and check that both terms on the right-hand side of

$$\frac{\Phi_{\max}^3 ns}{\lambda_1} \left(\frac{\lambda_1^2 s^2 m}{n^2} + \frac{s \log n}{n} \right) = \frac{\Phi_{\max}^3 s^3 \lambda_1 m}{n} + \frac{\Phi_{\max}^3 s^2 \log n}{\lambda_1}$$

tend to zero. For the first term on the right-hand side, it is enough to show that

$$\Phi_{\max}^6 s^6 m^3 \log(d^2 mr) (\log^3 n)/n \to 0,$$

where we have plugged in $\lambda_1 = C'\sqrt{mn\log(d^2mr/\delta)\log^3 n}$. Using the assumptions in condition A4, we get that $\log(d^2mr) = O(\log d + \log m) = O(n^{2/21})$, and furthermore that

$$\Phi_{\max}^6 s^6 m^3 \log(d^2mr)(\log^3 n)/n = o\left(\frac{n^{1/3}\cdot n^{2/21}\cdot n^{6/21}\cdot n^{6/21}}{\log^3 n}\right)\frac{\log^3 n}{n} \to 0,$$

as required. A similar calculation shows that the second term on the right-hand side also tends to zero, i.e., $\Phi_{\max}^3 s^2(\log n)/\lambda_1 \to 0$, which establishes (6.54). Lastly, (6.55) follows since its left-hand side is dominated by the left-hand side of (6.54).

So, we now compute

$$\begin{aligned}
\|\Delta_4^j\|_2 &\leq \sqrt{m}\max_t \sum_{i=1}^n \Big|\Phi_{it}\Big[F_{\epsilon_{\ell k}}\Big(\Phi_{i(S_{\ell k})}\big(\vartheta - \theta^*_{\ell k(S_{\ell k})}\big)\Big) - \\
&\qquad\qquad F_{\epsilon_{\ell k}}(0) - f_{\epsilon_{\ell k}}(0)\Phi_{i(S_{\ell k})}\big(\vartheta - \theta^*_{\ell k(S_{\ell k})}\big)\Big]\Big| \\
&\leq C_1\Phi_{\max}\sqrt{m}\sum_{i=1}^n \Big(\Phi_{i(S_{\ell k})}\big(\vartheta - \theta^*_{\ell k(S_{\ell k})}\big)\Big)^2 \\
&\leq C_1\Phi_{\max}\sqrt{m}\sum_{i=1}^n \|\Phi_{i(S_{\ell k})}\|_2^2\|\vartheta - \theta^*_{\ell k(S_{\ell k})}\|_2^2 \\
&\leq C_1\Phi_{\max}^3\sqrt{m}ns\gamma^2 \\
&= o(\lambda_1).
\end{aligned}$$

Here the first inequality follows from the fact that $\|x\|_2 \leq \sqrt{m}\|x\|_\infty$ for $x \in \mathbb{R}^m$, and the triangle inequality; the second follows from the distributional smoothness condition in A2, which is applicable since (6.55) holds; the third uses Cauchy-Schwarz; the fourth uses our column norm assumption on Φ, and $\|\vartheta - \theta^*_{\ell k(S_{\ell k})}\|_2 \leq \gamma$; the last uses (6.54). As $\|\Delta_4^j\|_2 = o(\lambda_1)$, it will certainly be strictly less than $\lambda_1/6$ for n large enough. We have hence shown, noting that none of our above arguments have depended on the particular choice of $j = 1, \ldots, d$ or $\vartheta \in B^*$,

$$\max_{j\in S_{\ell k}^c}\sup_{\vartheta\in B^*}\|\Delta_2^j\|_2 < \frac{\lambda_1}{3}.$$

Bounding $\|\Delta_3^j\|_2$. For this part, we can use the end of the proof of Lemma 2 in [37], which uses classic entropy-based techniques to establish a bound very similar to that which we are seeking. By carefully looking at the conditions required for this lemma, we see that under the distributional smoothness condition in A2, condition A3, and also

$$\begin{aligned}
\sqrt{n\log(dm)} &= o(\lambda_1) \\
n\Phi_{\max}\gamma^2 &= o(\lambda_1) \\
(1 + \gamma\Phi_{\max}^2 s^{3/2})\log^2 n &= o(\lambda_1^2/n),
\end{aligned}$$

all following directly from condition A4 by calculations similar to the ones we used when bounding $\|\Delta_2^j\|$, we have

$$\mathbb{P}\bigg(\max_{j \in S_{\ell k}^c} \sup_{\vartheta \in B^*} \|\Delta_3^j\|_2 \geq \frac{\lambda_1}{3}\bigg) \leq \mathbb{P}\bigg(\max_{j \in S_{\ell k}^c} \sup_{\vartheta \in B^*} \|\Delta_3^j\|_\infty \geq \frac{\lambda_1}{3\sqrt{m}}\bigg);$$

the probability on the right-hand side can be made arbitrarily small for large n, by the arguments at the end of Lemma 2 in [37], and hence clearly smaller than the desired $\delta/(4dr)$ level.

Putting it together. Returning to the logic in (6.51), (6.52), (6.53), we have shown that the subgradient condition in (6.51) holds with probability at least $1 - (\delta/(2dr) + \delta/(4dr) + \delta/(4dr)) = 1 - \delta/(dr)$. Taking a union bound over $k = 1, \ldots, d$ and $\ell = 1, \ldots, r$, which were considered fixed at the start of our analysis, gives the result stated in the theorem. □

6.3.5 Statement and Proof of Lemma 6.5

We show that with probability at least $1 - \delta/(2dr)$, it holds that $\tilde{\theta}_{\ell k(S_{\ell k})} \in B^*$, where $\tilde{\theta}_{\ell k(S_{\ell k})}$ is the solution to the restricted problem (6.48), for some fixed $k = 1, \ldots, d$ and $\ell = 1, \ldots, r$, and B^* is a ball defined in the proof of Theorem 4.1 in Section 6.3.4. This fact is used a few times in the proof of Theorem 4.1.

Lemma 6.5. *Fix some $k = 1, \ldots, d$ and $\ell = 1, \ldots, r$. Define the ball*

$$B^* = \{\vartheta \in \mathbb{R}^{sm} : \|\vartheta - \theta^*_{\ell k(S_{\ell k})}\|_2 \leq \gamma\}$$

centered at the underlying coefficients $\theta^*_{\ell k(S_{\ell k})}$ with radius

$$\gamma = C\left(\frac{\lambda_1 s\sqrt{m}}{n} + \sqrt{\frac{s \log n}{n}}\right),$$

for some constant $C > 0$. Then, with probability at least $1 - \delta/(2dr)$, it holds that $\tilde{\theta}_{\ell k(S_{\ell k})} \in B^*$, where $\tilde{\theta}_{\ell k(S_{\ell k})}$ is the solution to the restricted problem (6.48)*.*

Proof. We will follow the strategy for the proof of Theorem 1 in [37] closely. We begin by considering the ball

$$B = \{\vartheta \in \mathbb{R}^{sm} : \|\vartheta - \theta^*_{\ell k(S_{\ell k})}\|_2 \leq R\}$$

with center $\theta^*_{\ell k(S_{\ell k})}$ and radius R. We also introduce some useful notational shorthand, and write the quantile loss term in the restricted problem (6.48) as

$$L_{\ell k}(\vartheta) = \psi_{\alpha_\ell}\left(Y_k - \Phi_{S_{\ell k}}\vartheta\right).$$

Below, we show that a particular function of R serves as an upper bound for the quantity $\mathbb{E}[L_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) - L_{\ell k}(\theta^*_{\ell k(S_{\ell k})})]$, where the expectation here is taken over draws of the data, and $\tilde{\vartheta}_{\ell k(S_{\ell k})}$ is a particular point in B that we define in a moment. This in turn implies, with probability at least $1 - \delta/(2dr)$, that $\tilde{\theta}_{\ell k(S_{\ell k})} \in B^*$, as claimed.

First, we define $\tilde{\vartheta}_{\ell k(S_{\ell k})}$ more precisely: it is a point on the line segment between the solution to the restricted problem $\tilde{\theta}_{\ell k(S_{\ell k})}$ and the underlying coefficients $\theta^*_{\ell k(S_{\ell k})}$, i.e.,

$$\tilde{\vartheta}_{\ell k(S_{\ell k})} = \beta\tilde{\theta}_{\ell k(S_{\ell k})} + (1-\beta)\theta^*_{\ell k(S_{\ell k})},$$

for a particular choice

$$\beta = \frac{R}{R + \|\tilde{\theta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2},$$

which guarantees that $\tilde{\vartheta}_{\ell k(S_{\ell k})} \in B$ even if $\tilde{\theta}_{\ell k(S_{\ell k})} \notin B$, as we establish next. Observe that we always have

$$\begin{aligned}
& \|\tilde{\theta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2 \leq R + \|\tilde{\theta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2 \\
\iff & R\frac{\|\tilde{\theta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2}{R + \|\tilde{\theta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2} \leq R \\
\iff & \beta\|\tilde{\theta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2 \leq R \\
\iff & \|\beta\tilde{\theta}_{\ell k(S_{\ell k})} - \beta\theta^*_{\ell k(S_{\ell k})} + \theta^*_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2 \leq R \\
\iff & \|\tilde{\vartheta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2 \leq R,
\end{aligned}$$

as claimed. The second line here follows by rearranging and multiplying through by R; the third by using the definition of β above; the fourth by adding and subtracting the underlying coefficients; and the fifth by using the definition of $\tilde{\vartheta}_{\ell k(S_{\ell k})}$.

Now, the beginning of the proof of Theorem 1 in [37] establishes, for any $\tilde{\vartheta}_{\ell k(S_{\ell k})} \in B$, for some constant $C_5 > 0$, and using condition A3, that

$$\mathbb{E}\left[L_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) - L_{\ell k}(\theta^*_{\ell k(S_{\ell k})})\right] \geq C_5 n\|\tilde{\vartheta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2^2, \tag{6.56}$$

and so, by direct calculation, since

$$\|\tilde{\vartheta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2 \leq R \iff \beta\|\tilde{\theta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2 \leq R \iff \|\tilde{\theta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2 \leq R/2, \tag{6.57}$$

it suffices to obtain a suitable upper bound for $\mathbb{E}[L_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) - L_{\ell k}(\theta^*_{\ell k(S_{\ell k})})]$, in order to get the result in the statement of the lemma. To this end, we introduce one more piece of shorthand, and denote the objective for the restricted problem (6.48) as $J_{\ell k}(\vartheta)$.

We proceed with the following chain of (in)equalities:

$$\begin{aligned}
&\mathbb{E}\left[L_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) - L_{\ell k}(\theta^*_{\ell k(S_{\ell k})})\right] \\
&= \mathbb{E}\left[L_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) - L_{\ell k}(\theta^*_{\ell k(S_{\ell k})})\right] + J_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) - J_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) + \\
&\qquad\qquad J_{\ell k}(\theta^*_{\ell k(S_{\ell k})}) - J_{\ell k}(\theta^*_{\ell k(S_{\ell k})}) \qquad (6.58)\\
&= \underbrace{L_{\ell k}(\theta^*_{\ell k(S_{\ell k})}) - \mathbb{E}L_{\ell k}(\theta^*_{\ell k(S_{\ell k})}) - L_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) + \mathbb{E}L_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})})}_{\Delta(\theta^*_{\ell k(S_{\ell k})}, \tilde{\vartheta}_{\ell k(S_{\ell k})})} + \\
&\qquad J_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) - J_{\ell k}(\theta^*_{\ell k(S_{\ell k})}) + \lambda_1 \sum_{j \in S_{\ell k}} \|\theta^*_{\ell kj}\|_2 - \lambda_1 \sum_{j \in S_{\ell k}} \|\tilde{\vartheta}_{\ell kj}\|_2 \\
&\qquad\qquad - (\lambda_2/2)\|\tilde{\vartheta}_{\ell k(S_{\ell k})}\|_2^2 + (\lambda_2/2)\|\theta^*_{\ell k(S_{\ell k})}\|_2^2 \qquad (6.59)\\
&\leq \Delta(\theta^*_{\ell k(S_{\ell k})}, \tilde{\vartheta}_{\ell k(S_{\ell k})}) + J_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) - J_{\ell k}(\theta^*_{\ell k(S_{\ell k})}) + \lambda_1 \sum_{j \in S_{\ell k}} \|\theta^*_{\ell kj} - \tilde{\vartheta}_{\ell kj}\|_2 + o(1) \qquad (6.60)\\
&\leq \Delta(\theta^*_{\ell k(S_{\ell k})}, \tilde{\vartheta}_{\ell k(S_{\ell k})}) + J_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) - J_{\ell k}(\theta^*_{\ell k(S_{\ell k})}) + \lambda_1 sR\sqrt{m} + o(1) \qquad (6.61)\\
&\leq \Delta(\theta^*_{\ell k(S_{\ell k})}, \tilde{\vartheta}_{\ell k(S_{\ell k})}) + 2\lambda_1 sR\sqrt{m} \qquad (6.62)\\
&\leq \sup_{\tilde{\vartheta}_{\ell k(S_{\ell k})} \in B} |\Delta(\theta^*_{\ell k(S_{\ell k})}, \tilde{\vartheta}_{\ell k(S_{\ell k})})| + 2\lambda_1 sR\sqrt{m}. \qquad (6.63)
\end{aligned}$$

Here, (6.58) follows by adding and subtracting like terms, and (6.59) by rearranging (6.58). In (6.60) we use the triangle inequality and the following argument to show that the terms involving λ_2 are $o(1)$. Under the assumption that $\lambda_2 = o(n^{41/42}/\theta^*_{\max})$, combined with the restriction that $s = o(n^{1/21})$, we have $\lambda_2 = o(n/(\sqrt{s}\theta^*_{\max}))$. Therefore, under our choice of $R = 1/n$ (as specified below), we have

$$\lambda_2\sqrt{s}\theta^*_{\max}R \to 0.$$

This in turn is used to argue that

$$\begin{aligned}
-(\lambda_2/2)\|\tilde{\vartheta}_{\ell k(S_{\ell k})}\|_2^2 + (\lambda_2/2)\|\theta^*_{\ell k(S_{\ell k})}\|_2^2 &= (\lambda_2/2)\|\tilde{\vartheta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2^2 \\
&\qquad - \lambda_2\|\tilde{\vartheta}_{\ell k(S_{\ell k})}\|_2^2 + \lambda_2\tilde{\vartheta}^T_{\ell k(S_{\ell k})}\theta^*_{\ell k(S_{\ell k})} \\
&\leq (\lambda_2/2)R^2 - \lambda_2\|\tilde{\vartheta}_{\ell k(S_{\ell k})}\|_2(\|\tilde{\vartheta}_{\ell k(S_{\ell k})}\|_2 - \|\theta^*_{\ell k(S_{\ell k})}\|_2) \\
&\leq (\lambda_2/2)R^2 + \lambda_2\|\tilde{\vartheta}_{\ell k(S_{\ell k})}\|_2 R \\
&\leq (\lambda_2/2)R^2 + \lambda_2\|\theta^*_{\ell k(S_{\ell k})}\|_2 R \\
&\leq (\lambda_2/2)R^2 + \lambda_2\sqrt{s}\theta^*_{\max}R \to 0.
\end{aligned}$$

In the second to last line, we have applied $\|\tilde{\vartheta}_{\ell k(S_{\ell k})}\|_2 \leq \|\theta^*_{\ell k(S_{\ell k})}\|_2$, as, outside of this case, the term in question $-(\lambda_2/2)\|\tilde{\vartheta}_{\ell k(S_{\ell k})}\|_2^2 + (\lambda_2/2)\|\theta^*_{\ell k(S_{\ell k})}\|_2^2$ would be negative, anyway.

Continuing on, (6.61) holds because $\|\theta^*_{\ell k(S_{\ell k})} - \tilde{\vartheta}_{\ell k(S_{\ell k})}\|_2 \leq R$ implies $\|\theta^*_{\ell kj} - \tilde{\vartheta}_{\ell kj}\|_2 \leq R$. Finally, (6.62) follows because of the following argument. Since $J_{\ell k}$ is convex, we can use the definition of $\tilde{\vartheta}_{\ell k(S_{\ell k})}$ and get

$$J_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) \leq \beta J_{\ell k}(\tilde{\theta}_{\ell k(S_{\ell k})}) + (1-\beta)J_{\ell k}(\theta^*_{\ell k(S_{\ell k})}) = J_{\ell k}(\theta^*_{\ell k(S_{\ell k})}) + \beta(J_{\ell k}(\tilde{\theta}^*_{\ell k(S_{\ell k})}) - J_{\ell k}(\theta^*_{\ell k(S_{\ell k})}));$$

notice that the last term here is nonpositive, since $\tilde{\theta}_{\ell k(S_{\ell k})}$ is the solution to the restricted problem (6.48), and thus we have that

$$J_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) \leq J_{\ell k}(\theta^*_{\ell k(S_{\ell k})}),$$

which lets us move from (6.61) to (6.62).

Lemma 1 in [37] states, with probability at least $1-\delta$, where $\delta = \exp(-C_6 s \log n)$ and $C_6 > 0$ is some constant, that

$$\sup_{\tilde{\vartheta}_{\ell k(S_{\ell k})} \in B} |\Delta(\theta^*_{\ell k(S_{\ell k})}, \tilde{\vartheta}_{\ell k(S_{\ell k})})| \leq 6R\sqrt{sn \log n},$$

so from (6.63), with probability at least $1-\delta$, we see that

$$\mathbb{E}\left[L_{\ell k}(\tilde{\vartheta}_{\ell k(S_{\ell k})}) - L_{\ell k}(\theta^*_{\ell k(S_{\ell k})})\right] \leq 6R\sqrt{sn \log n} + 2\lambda_1 sR\sqrt{m}$$

and, using (6.56), that

$$n\|\tilde{\vartheta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2^2 \leq C'\left(R\sqrt{sn \log n} + \lambda_1 sR\sqrt{m}\right),$$

for some constant $C' > 0$.

Plugging in $R = 1/n$, dividing through by n, and using the fact that the square root function is subadditive, we get, with probability at least $1-\delta$, that

$$\begin{aligned}\|\tilde{\vartheta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2 &\leq C'\left(\frac{(s\log n)^{1/4}}{n^{3/4}} + \frac{(\lambda_1 s)^{1/2} m^{1/4}}{n}\right) \\ &\leq C'\left(\sqrt{\frac{s\log n}{n}} + \frac{\lambda_1 s\sqrt{m}}{n}\right).\end{aligned}$$

Finally, we complete the proof by applying (6.57), in order to get that

$$\|\tilde{\theta}_{\ell k(S_{\ell k})} - \theta^*_{\ell k(S_{\ell k})}\|_2 \leq \gamma,$$

where we have defined

$$\gamma = C\left(\frac{\lambda_1 s\sqrt{m}}{n} + \sqrt{\frac{s\log n}{n}}\right),$$

and $C > 0$ is some constant, with probability at least $1 - \delta/(2dr)$, for large enough n. □

6.3.6 Proof of Lemma 4.3

The prox operator $\mathbf{prox}_{\lambda\psi_{\mathcal{A}}}(A)$ is separable in the entries of its minimizer X, so we focus on minimizing over X_{ij} the expression

$$\begin{aligned}&\max\{\alpha_j X_{ij}, (\alpha_j - 1)X_{ij}\} + (1/(2\lambda))\,(X_{ij} - A_{ij})^2 \\ &\quad = \alpha_j \max\{0, X_{ij}\} + (1-\alpha_j)\max\{0, -X_{ij}\} + (1/(2\lambda))\,(X_{ij} - A_{ij})^2. \qquad (6.64)\end{aligned}$$

Suppose $X_{ij} > 0$. Then differentiating (6.64) gives $X_{ij} = A_{ij} - \lambda\alpha_j$ and the sufficient condition $A_{ij} > \lambda\alpha_j$. Similarly, assuming $X_{ij} < 0$ gives $X_{ij} = A_{ij} + \lambda(1-\alpha_j)$ when $A_{ij} < \lambda(\alpha_j - 1)$. Otherwise, we can take $X_{ij} = 0$. Putting these cases together gives the result. □

6.3.7 Additional Details on Gibbs Sampling

In the MQGM, there is no analytic solution for parameters like the mean, median, or quantiles of these marginal and conditional distributions, but the pseudolikelihood approximation makes for a very efficient Gibbs sampling procedure, which we highlight in this section. As it is relevant to the computational aspects of the approach, in this subsection we will make explicit the conditional random field, where y_k depends on both $y_{\neg k}$ and fixed input features x.

First, note that since we are representing the distribution of $y_k|y_{\neg k}, x$ via its inverse CDF, to sample from from this conditional distribution we can simply generate a random $\alpha \sim \text{Uniform}(0, 1)$. We then compute

$$\hat{Q}_{y_k|y_{\neg k}}(\alpha_\ell) = \phi(y)^T \theta_{\ell k} + x^T \theta^x_{\ell k}$$
$$\hat{Q}_{y_k|y_{\neg k}}(\alpha_{\ell+1}) = \phi(y)^T \theta_{(\ell+1) k} + x^T \theta^x_{(\ell+1) k}$$

for some pair $\alpha_\ell \leq \alpha \leq \alpha_{\ell+1}$ and set y_k to be a linear interpolation of the two values

$$y_k \leftarrow \hat{Q}_{y_k|y_{\neg k}}(\alpha_\ell) + \frac{\left(\hat{Q}_{y_k|y_{\neg k}}(\alpha_{\ell+1}) - \hat{Q}_{y_k|y_{\neg k}}(\alpha_\ell)\right)(\alpha - \alpha_\ell)}{\alpha_{\ell+1} - \alpha_\ell}.$$

This highlights the desirability of having a range of non-uniformly spaced α terms that reach values close to zero and one as otherwise we may not be able to find a pair of α's that lower and upper bound our random sample α. However, in the case that we model a sufficient quantity of α, a reasonable approximation (albeit one that will not sample from the extreme tails) is also simply to pick a random $\alpha_\ell \in \mathcal{A}$ and use just the corresponding column $\theta_{\ell k}$ to generate the random sample.

Computationally, there are a few simple but key points involved in making the sampling efficient. First, when sampling from a conditional distribution, we can precompute $x^T \Theta^x_k$ for each k, and use these terms as a constant offset. Second, we maintain a "running" feature vector $\phi(y) \in \mathbb{R}^{dm}$, i.e., the concatenation of features corresponding to each coordinate $\phi(y_k)$. Each time we sample a new coordinate y_k, we generate just the new features in the $\phi(y_k)$ block, leaving the remaining features untouched. Finally, since the Θ_k terms are sparse, the inner product $\phi(y)^T \theta_{\ell k}$ will only contain a few nonzeros terms in the sum, and will be computed more efficiently if the Θ_k are stored as a sparse matrices.

6.3.8 Additional Details on the Evaluation of Fitted Conditional CDFs

Here, we elaborate on the evaluation of each method's conditional CDFs that we first presented in Section 4.6.1. For simplicity, we describe everything that follows in terms of the conditional CDF $y_1|y_2$ only, with everything being extended in the obvious way to other conditionals. (We omit the nonparanormal skeptic from our evaluation as it is not clear how to sample from its conditionals, due to the nature of a particular transformation that it uses.)

First, we carried out the following steps in order to compute the true (empirical) conditional CDF.

1. We drew $n = 400$ samples from the ring distribution, by following the procedure described in Section 4.6.1; these observations are plotted across the top row of Figure 6.7.

2. We then partitioned the y_2 samples into five equally-sized bins, and computed the true empirical conditional CDF of y_1 given each bin of y_2 values.

Next, we carried out the following steps in order to compute the estimated (empirical) conditional CDFs, for each method.

3. We fitted each method to the samples obtained in step (1) above.
4. Then, for each method, we drew a sample of y_1 given *each* y_2 sample, using the method's conditional distribution; these conditionals are plotted across the second through fifth rows of Figure 6.7 (for representative values of λ_1).

   Operationally, we drew samples from each method's conditionals in the following ways.

   - MQGM: we used the Gibbs sampler described in Section 6.3.7.
   - MB: we drew $y_1 \sim \mathcal{N}(\hat{\theta}_1^T y_2^{(i)}, \hat{\sigma}_{1|2}^2)$, where $\hat{\theta}_1$ is the fitted lasso regression coefficient of y_1 on y_2; $y_2^{(i)}$ for $i = 1, \ldots, n$ is the ith observation of y_2 obtained in step (1) above; and $\hat{\sigma}_{1|2}^2 = \mathrm{Var}(Y_1 - Y_2\hat{\theta}_1)$ denotes the sample variance of the underlying error term $Y_1 - Y_2\hat{\theta}_1$ with $Y_i = (y_i^{(1)}, \ldots, y_i^{(n)}) \in \mathbb{R}^n$ collecting all observations along variable i.
   - SpaceJam: we drew $y_1 \sim \mathcal{N}(\hat{\theta}_1^T \phi(y_2^{(i)}), \hat{\sigma}_{1|2}^2)$, where ϕ is a suitable basis function, and $\hat{\theta}_1$ as well as $\hat{\sigma}_{1|2}^2$ are defined in ways analogous to the neighborhood selection setup.
   - GLasso: we drew $y_1 \sim \mathcal{N}(\hat{\mu}_{1|2}, \hat{\sigma}_{1|2}^2)$, where

     $$\hat{\mu}_{1|2} = \hat{\mu}_1 + \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}(y_2^{(i)} - \hat{\mu}_2)$$
     $$\hat{\sigma}_{1|2}^2 = \hat{\Sigma}_{11} - \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{21}$$

     with $\hat{\mu}_i$ denoting the sample mean of Y_i, and $\hat{\Sigma}$ denoting the estimate of the covariance matrix given by GLasso (subscripts select blocks of this matrix).
5. Finally, we partitioned the y_2 samples into five equally-sized bins (just as when computing the true conditional CDF), and computed the estimated empirical conditional CDF of y_1 given each bin of y_2 values.

Having computed the true as well as estimated conditional CDFs, we measured the goodness of fit of each method's conditional CDFs to the true conditional CDFs, by computing the total variation (TV) distance, i.e.,

$$(1/2)\sum_{i=1}^{q} \left| \hat{F}_{y_1|y_2}^{\text{method}_j}(z^{(i)}|\zeta) - \hat{F}_{y_1|y_2}^{\text{true}}(z^{(i)}|\zeta) \right|,$$

as well as the (scaled) Kolmogorov-Smirnoff (KS) statistic, i.e.,

$$\max_{i=1,\ldots,q} \left| \hat{F}_{y_1|y_2}^{\text{method}_j}(z^{(i)}|\zeta) - \hat{F}_{y_1|y_2}^{\text{true}}(z^{(i)}|\zeta) \right|.$$

Here, $\hat{F}^{\text{true}}_{y_1|y_2}(z^{(i)}|\zeta)$ is the true empirical conditional CDF of $y_1|y_2$, evaluated at $y_1 = z^{(i)}$ and given $y_2 = \zeta$, and $\hat{F}^{\text{method}_j}_{y_1|y_2}(z^{(i)}|\zeta)$ is a particular method's ("method$_j$" above) estimated empirical conditional CDF, evaluated at $y_1 = z^{(i)}$ and given $y_2 = \zeta$. For each method, we averaged these TV and KS values across the method's conditional CDFs. Table 4.1 reports the best (across a range of tuning parameters) of these averaged TV and KS values.

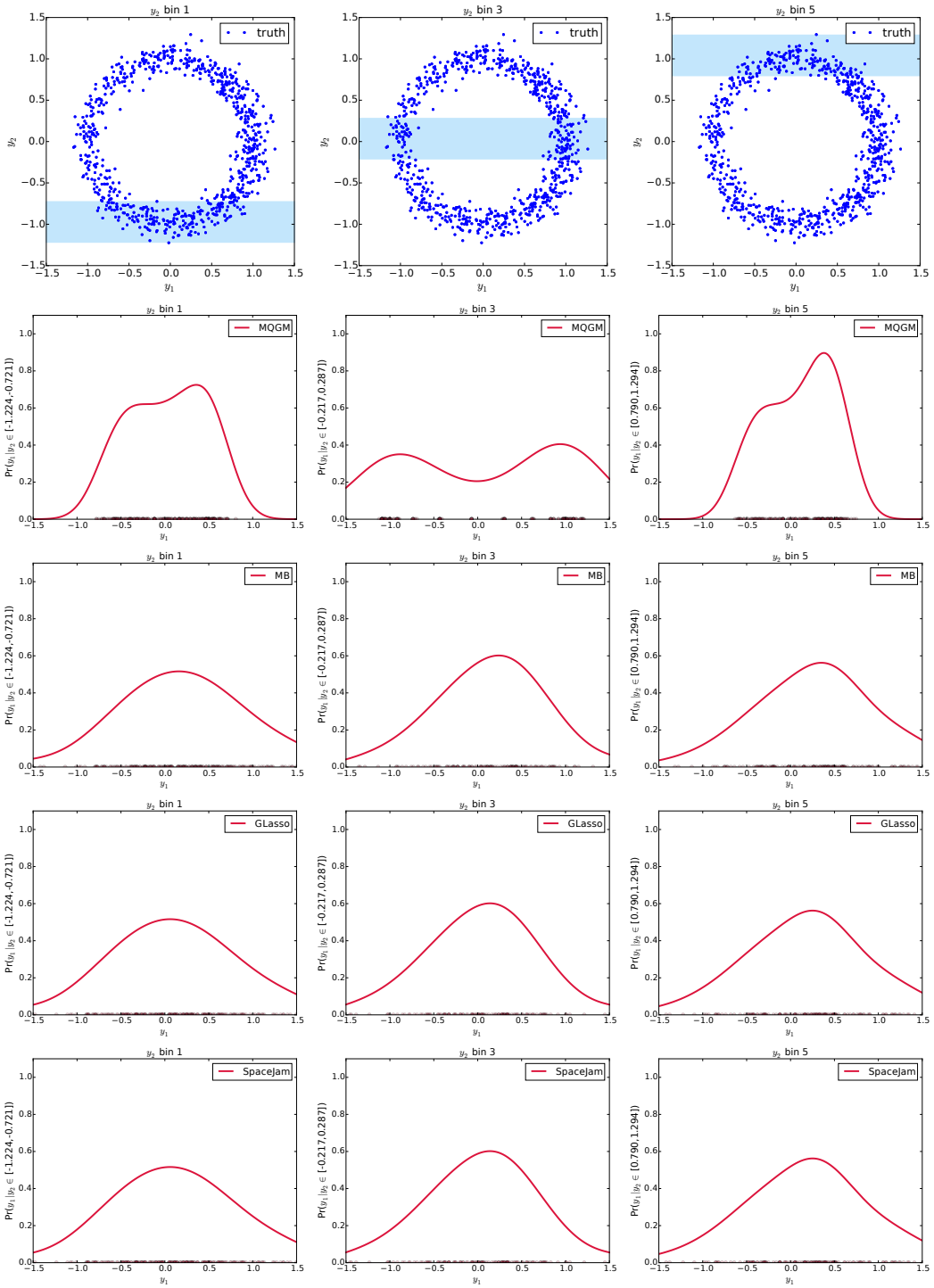

Figure 6.7: *Conditional distributions for MQGM, MB, GLasso, and SpaceJam, fitted to samples from the ring distribution (TIGER and Laplace's conditionals both look similar to MB's). First row: samples from the ring distribution, where each plot highlights the samples falling into a particular shaded bin on the y_2 axis. Second through fifth rows: conditional distributions of y_1 given y_2 for each method, where each plot conditions on the appropriate y_2 bin as highlighted in the first row. The MQGM's conditional distributions are intuitive, appearing bimodal for bin 3, and more peaked for bins 1 and 5. MB, GLasso, and SpaceJam's densities appear (roughly) Gaussian, as expected.*

Bibliography

[1] Samrachana Adhikari, Fabrizio Lecci, James T. Becker, Brian W. Junker, Lewis H. Kuller, Oscar L. Lopez, and Ryan J. Tibshirani. High-dimensional longitudinal classification with the multinomial fused lasso. *Statistics in Medicine*, 38(12):2184–2205, 2019. 1.2.3, 2.1

[2] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984. 2.8

[3] Alnur Ali, Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. Generalized pseudolikelihood methods for inverse covariance estimation. Technical report, 2016. Available at `http://arxiv.org/pdf/1606.00033.pdf`. 4.2.1, 4.6.1, 4.6.3

[4] Zhidong Bai and Jack Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010. 3.6.1, 3.4

[5] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008. 4.2.1

[6] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007. 3.1

[7] Heinz H. Bauschke and Jonathan M. Borwein. Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997. 6.1.9, 6.1.14

[8] Joanne C Beer, Howard J Aizenstein, Stewart J Anderson, and Robert T Krafty. Incorporating prior information with fused sparse group lasso: Application to prediction of clinical measures from neuroimages. *arXiv preprint arXiv:1801.06594*, 2018. 1.2.1

[9] Eugene Belilovsky, Gaël Varoquaux, and Matthew B Blaschko. Testing for differences in gaussian graphical models: applications to brain connectivity. In *Advances in Neural Information Processing Systems*, pages 595–603, 2016. 1.2.1

[10] Pierre Bellec, Carlton Chu, Francois Chouinard-Decorte, Yassine Benhajali, Daniel S Margulies, and R Cameron Craddock. The neuro bureau adhd-200 preprocessed repository. *Neuroimage*, 144:275–286, 2017. 1.2.1

[11] Alexandre Belloni and Victor Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*, 39(1):82–130, 2011. 4.2.2

[12] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B*, 36(2):192–236, 1974. 1.4, 4.1

[13] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical*

Society: Series B (Methodological), 48(3):259–279, 1986. 1.2.2

[14] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 6.3.2

[15] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 1.2.4, 4.5

[16] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005. 1.2.2

[17] Peter Buhlmann and Bin Yu. Boosting with the ℓ_2 loss. *Journal of the American Statistical Association*, 98(462):324–339, 2003. 3.1, 3.1

[18] John C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, 2016. 3.2.3

[19] Emmanuel J. Candes and Yaniv Plan. Near ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37(5):2145–2177, 2009. 2.1.2

[20] Emmanuel J. Candes and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. arXiv: 1804.09753, 2018. 2.8

[21] Centers for Disease Control and Prevention (CDC). Influenza national and regional level graphs and data, August 2015. URL `http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html`. 4.6.2

[22] Sorathan Chaturapruek, John C Duchi, and Christopher Ré. Asynchronous stochastic convex optimization: the noise is in the noise and sgd don't care. In *Advances in Neural Information Processing Systems*, pages 1531–1539, 2015. 5.3

[23] Shizhe Chen, Daniela Witten, and Ali Shojaie. Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64, 2015. 4.4.2

[24] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016. 1.2.2

[25] Arthur Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972. 4.1

[26] Paramveer Dhillon, Dean Foster, Sham Kakade, and Lyle Ungar. A risk comparison of ordinary least squares vs ridge regression. *The Journal of Machine Learning Research*, 14:1505–1511, 2013. 3.1

[27] Lee H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37, 2016. 3.1, 3.3, 3.8

[28] Edgar Dobriban. Efficient computation of limit spectra of sample covariance matrices. *Random Matrices: Theory and Applications*, 4(4):1550019, 2015. 3.9

[29] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: ridge regression and classification. *Annals of Statistics*, 46(1):247–279, 2018. 3.1, 3.6.3, 3.8

[30] Frank Dondelinger and Sach Mukherjee. The joint lasso: high-dimensional regression for

group structured data. *Biostatistics*, 2018. 1.2.1

[31] David L. Donoho. For most large underdetermined systems of linear equations, the minimal ℓ_1 solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006. 2.1.2

[32] Charles Dossal. A necessary and sufficient condition for exact sparse recovery by ℓ_1 minimization. *Comptes Rendus Mathematique*, 350(1–2):117–120, 2012. 2.1.2

[33] Richard Durrett. *Probability models for DNA sequence evolution*. Springer Science & Business Media, 2008. 5.2

[34] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015. 1.2.2

[35] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009. 5.2

[36] Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 1992. 6.1.6

[37] Jianqing Fan, Yingying Fan, and Emre Barut. Adaptive robust variable selection. *Annals of Statistics*, 42(1):324–351, 2014. 4.2.2, 6.3.4, 6.3.4, 6.3.5, 6.3.5

[38] Jianqing Fan, Wenyan Gong, Chris Junchi Li, and Qiang Sun. Statistical sparse online regression: A diffusion approximation perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 1017–1026, 2018. 5.2

[39] Michael Finegold and Mathias Drton. Robust graphical modeling of gene networks using classical and alternative t-distributions. *Annals of Applied Statistics*, 5(2A):1057–1080, 2011. 4.2.1

[40] Jerome Friedman and Bogdan Popescu. Gradient directed regularization. Working paper, 2004. URL `http://www-stat.stanford.edu/~jhf/ftp/pathlite.pdf`. 1.3, 3.1

[41] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. 4.2.1

[42] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Technical report, 2010. Available at `http://statweb.stanford.edu/~tibs/ftp/ggraph.pdf`. 4.2.1

[43] Christiane Fuchs. *Inference for Diffusion Processes: With Applications in Life Sciences*. Springer Science & Business Media, 2013. 5.2

[44] Jean Jacques Fuchs. Recovery of exact sparse representations in the presense of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608, 2005. 2.1.2

[45] Alexander Goldenshluger and Alexandre Tsybakov. Adaptive prediction and estimation in linear regression with infinitely many parameters. *Annals of Statistics*, 29(6):1601–1619, 2001. 3.1

[46] Michael Grant and Stephen Boyd. Cvx: Matlab software for disciplined convex programming, version 2.1, 2014. 1.2.4

[47] Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, 2017. 3.1

[48] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, 2018. 3.1

[49] Hao Guo, Yao Li, Yong Xu, Yanyi Jin, Jie Xiang, and Junjie Chen. Resting-state brain functional hyper-network construction based on elastic net and group lasso methods. *Frontiers in neuroinformatics*, 12, 2018. 1.2.5

[50] Laszlo Gyrofi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. 2002. 6.3.3

[51] Shelby J. Haberman. Log-linear models for frequency tables derived by indirect observation: Maximum likelihood equations. *Annals of Statistics*, 2(5):911–924, 1974. 2.8

[52] Jeffery Z HaoChen and Suvrit Sra. Random shuffling beats sgd after finite epochs. *arXiv preprint arXiv:1806.10077*, 2018. 5.2

[53] David Heckerman, David Maxwell Chickering, David Meek, Robert Rounthwaite, and Carl Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–75, 2000. 4.1, 4.2.1, 4.4.2, 6.3.2

[54] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015. 1.2.2

[55] Holger Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010. 2.1

[56] Arthur E. Hoerl and Robert W. Kennard. Ridge regression iterative estimation of the biasing parameter. *Communications in Statistics: Theory and Methods*, 5(1):77–88, 1976. 3.2.1

[57] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pages 1731–1741, 2017. 5.2

[58] Holger Höfling and Robert Tibshirani. Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*, 10:883–906, 2009. 4.2.1

[59] Tao Hong, Pierre Pinson, and Shu Fan. Global energy forecasting competition 2012. *International Journal of Forecasting*, 30:357–363, 2014. 4.6.3

[60] Cho-Jui Hsieh, Mátyás A Sustik, Inderjit S Dhillon, Pradeep K Ravikumar, and Russell Poldrack. Big & quic: Sparse inverse covariance estimation for a million variables. In *Advances in neural information processing systems*, pages 3165–3173, 2013. 1.2.5

[61] Daniel Hsu, Sham Kakade, and Tong Zhang. Random design analysis of ridge regression.

In *Annual Conference on Learning Theory*, pages 9.1–9.24, 2012. 3.1, 3.8

[62] Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223): 1–42, 2018. 5.2, 5.3

[63] Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. *arXiv preprint arXiv:1903.01463*, 2019. 5.2

[64] Nicholas Johnson. A dynamic programming algorithm for the fused lasso and ℓ_0-segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260, 2013. 4.5

[65] Kengo Kato. Group lasso for high dimensional sparse quantile regression models. Technical report, 2011. Available at `http://arxiv.org/pdf/1103.1458.pdf`. 4.2.2

[66] Murray C. Kemp and Yoshio Kimura. *Introduction to Mathematical Economics*. Springer, 1978. 6.1.10

[67] Mohammad Khabbazian, Ricardo Kriebel, Karl Rohe, and Cecile Ane. Fast and accurate detection of evolutionary shifts in Ornstein-Uhlenbeck models. *Evolutionary Quantitative Genetics*, 7:811–824, 2016. 1.2.3, 2.1

[68] Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B*, 77(4):803–825, 2014. 4.2.1, 4.6.1

[69] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009. 2.1

[70] Penporn Koanantakool, Alnur Ali, Ariful Azad, Aydin Buluc, Dmitriy Morozov, Leonid Oliker, Katherine Yelick, and Sang-Yun Oh. Communication-avoiding optimization methods for distributed massive-scale sparse inverse covariance estimation. *arXiv preprint arXiv:1710.10769*, 2017. 1.2.5

[71] Roger Koenker. *Quantile Regression*. Cambridge University Press, 2005. 4.2.2, 4.3

[72] Roger Koenker. Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3):239–262, 2011. 4.2.2

[73] Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. 4.2.2

[74] Roger Koenker, Pin Ng, and Stephen Portnoy. Quantile smoothing splines. *Biometrika*, 81(4):673–680, 1994. 4.2.2

[75] Robert Lang. A note on the measurability of convex sets. *Archiv der Mathematik*, 47(1): 90–92, 1986. 6.1.6

[76] Steffen Lauritzen. *Graphical models*. Oxford University Press, 1996. 4.1

[77] Olivier Ledoit and Sandrine Peche. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1–2):233–264, 2011. 3.6.2, 3.6.3

[78] Jason Lee and Trevor Hastie. Structure learning of mixed graphical models. In *Proceed-*

ings of the 16th International Conference on Artificial Intelligence and Statistics, pages 388–396, 2013. 4.2.1

[79] Jason Lee, Yuekai Sun, and Jonathan Taylor. On model selection consistency of M-estimators with geometrically decomposable penalties. *Electronic Journal of Statistics*, 9(1):608–642, 2015. 1.4, 2.1.2

[80] Sang H Lee, Donghyeon Yu, Alvin H Bachman, Johan Lim, and Babak A Ardekani. Application of fused lasso logistic regression to the study of corpus callosum thickness in early alzheimer's disease. *Journal of neuroscience methods*, 221:78–84, 2014. 1.2.1

[81] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015. 1.2.2

[82] Junhong Lin and Lorenzo Rosasco. Optimal learning for multi-pass stochastic gradient methods. In *Advances in Neural Information Processing Systems*, pages 4556–4564, 2016. 5.3

[83] Michael Lindenbaum, M Fischer, and A Bruckstein. On gabor's contribution to image enhancement. *Pattern Recognition*, 27(1):1–8, 1994. 1.2.2

[84] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016. 5.1

[85] Han Liu and Lie Wang. TIGER: A tuning-insensitive approach for optimally estimating Gaussian graphical models. Technical report, 2012. Available at `http://arxiv.org/pdf/1209.2437.pdf`. 4.2.1, 4.6.1

[86] Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009. 4.2.1

[87] Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, pages 2293–2326, 2012. 4.2.1, 4.6.1

[88] Laura Lo Gerfo, Lorenzo Rosasco, Francesca Odone, Ernesto De Vito, and Alessandro Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008. 3.1

[89] Oscar Hernan Madrid-Padilla and James Scott. Tensor decomposition with generalized lasso penalties. *Journal of Computational and Graphical Statistics*, 26(3):537–546, 2017. 1.2.3, 2.1

[90] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967. 3.4

[91] Pertti Mattila. *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge University Press, 1995. 6.1.5

[92] Lukas Meier, Sara van de Geer, and Peter Buhlmann. High-dimensional additive modeling. *Annals of Statistics*, 37(6):3779–3821, 2009. 4.3

[93] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006. 4.1

[94] Nelson Morgan and Herve Bourlard. Generalization and parameter estimation in feedforward nets: Some experiments. In *Advances in Neural Information Processing Systems*, 1989. 3.1

[95] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019. 5.1

[96] Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. *arXiv preprint arXiv:1806.01796*, 2018. 5.2

[97] Sangnam Nam, Mike E. Davies, Michael Elad, and Remi Gribonval. The cosparse analysis model and algorithms. *Applied and Computational Harmonic Analysis*, 34(1):30–56, 2013. 2.1.2

[98] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009. 5.3

[99] Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. *arXiv preprint arXiv:1802.08009*, 2018. 5.3

[100] Jennifer Neville and David Jensen. Dependency networks for relational data. In *Proceedings of Fourth IEEE International Conference on the Data Mining*, pages 170–177. IEEE, 2004. 4.2.1

[101] Brendan O'Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Operator splitting for conic optimization via homogeneous self-dual embedding. Technical report, 2013. Available at `https://stanford.edu/~boyd/papers/pdf/scs.pdf`. 4.6.2

[102] Sang-Yun Oh, Onkar Dalal, Kshitij Khare, and Bala Rajaratnam. Optimization methods for sparse pseudolikelihood graphical model selection. In *Advances in Neural Information Processing Systems 27*, pages 667–675, 2014. 4.6.1

[103] Michael Osborne, Brett Presnell, and Berwin Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000. 2.1.2

[104] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013. 4.5

[105] Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009. 4.2.1, 4.6.1

[106] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018. 5.3

[107] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. 5.3

[108] James Ramsay. Parameter flows. Working paper, 2005. 3.1

[109] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57 (10):6976–6994, 2011. 1.4

[110] Garvesh Raskutti, Martin Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13:389–427, 2012. 4.3

[111] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and nonparametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15:335–366, 2014. 1.3, 3.1, 3.8

[112] Benjamin Recht and Christopher Re. Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies. Technical report, and consequences. Technical report, University of Wisconsin-Madison, 2012. 5.2

[113] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 5.3

[114] Guilherme Rocha, Peng Zhao, and Bin Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (SPLICE). Technical report, 2008. Available at `https://www.stat.berkeley.edu/~binyu/ps/rocha.pseudo.pdf`. 4.2.1

[115] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970. 2.2.1, 2.4.1, 2.4.2, 6.1.9, 6.1.9, 6.1.14

[116] R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*. Springer, 2009. 6.1.14

[117] Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004. 2.1.2, 3.1

[118] Adam Rothman, Peter Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008. 4.2.1

[119] Leonid I. Rudin, Stanley Osher, and Emad Faterni. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. 2.1

[120] Veeranjaneyulu Sadhanala and Ryan J. Tibshirani. Additive models via trend filtering. arXiv: 1702.05037, 2017. 2.1.1

[121] Veeranjaneyulu Sadhanala, Yu-Xiang Wang, James Sharpnack, and Ryan J. Tibshirani. Higher-total variation classes on grids: Minimax theory and trend filtering methods. *Advances in Neural Information Processing Systems*, 30, 2017. 2.1, vi

[122] Ulrike Schneider and Karl Ewald. On the distribution, model selection properties and uniqueness of the lasso estimator in low and high dimensions. arXiv: 1708.09608, 2017. 2.1.2

[123] Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in neural information processing systems*, pages 46–54, 2016. 5.2

[124] Jack Silverstein. Strong convergence of the empirical distribution of eigenvalues of large

dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995. 3.4

[125] Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017. 5.2

[126] Kyung-Ah Sohn and Seyoung Kim. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse covariance regularization. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pages 1081–1089, 2012. 4.2.1

[127] Gabriel Steidl, Stephan Didas, and Julia Neumann. Splines in higher order TV regularization. *International Journal of Computer Vision*, 70(3):214–255, 2006. 2.1

[128] Otto Neall Strand. Theory and methods related to the singular-function expansion and Landweber's iteration for integral equations of the first kind. *SIAM Journal on Numerical Analysis*, 11(4):798–825, 1974. 3.1

[129] Daniel W Stroock and SR Srinivasa Varadhan. *Multidimensional diffusion processes*. Springer, 2007. 5.2

[130] Arun S. Suggala, Adarsh Prasad, and Pradeep Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, 2018. 3.1

[131] Ichiro Takeuchi, Quoc Le, Timothy Sears, and Alexander Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006. 4.2.2, 4.3

[132] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996. 2.1

[133] Robert Tibshirani and Pei Wang. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9(1):18–29, 2008. 1.2.3, 2.1

[134] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67 (1):91–108, 2005. 1.2.3, 2.1

[135] Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7: 1456–1490, 2013. 2.1, 2.1.2, 2.1, 2.4, 2.5

[136] Ryan J. Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371, 2011. 1.2.3, 1.2.3, 2.1, ii, iii, iv, 2.1.1, 2.2.1, 1, 2.2.2

[137] Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *Annals of Statistics*, 40(2):1198–1232, 2012. 2.1.3, 2.2.1, 2.2.2, 2.2.2, 2.2.3, 2.3, 2.4.4, 2.4.5, 2.4.6, 2.4.6, 6.1.13, 6.1.15

[138] Antonia M. Tulino and Sergio Verdu. Random matrix theory and wireless communications. *Foundations and Trends in Communications and Information Theory*, 1(1):1–182, 2004. 3.6.3

[139] Cristiano Varin and Paolo Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528, 2005. 4.4.2

[140] Arend Voorman, Ali Shojaie, and Daniela Witten. Graph estimation with joint additive

models. *Biometrika*, 101(1):85–101, 2014. 4.2.1, 4.4.3, 4.6.1

[141] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009. 1.4, 2.1.2, 6.3.3, 6.3.3

[142] Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan J. Tibshirani. Trend filtering on graphs. *Journal of Machine Learning Research*, 17(105):1–41, 2016. 1.2.3, 2.1, v

[143] Yuchung Wang and Edward Ip. Conditionally specified continuous distributions. *Biometrika*, 95(3):735–746, 2008. 4.4.2

[144] Yuting Wei, Fanny Yang, and Martin J. Wainwright. Early stopping for kernel boosting algorithms: a general analysis with localized complexities. In *Advances in Neural Information Processing Systems*, 2017. 3.1

[145] Ashia Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, 2017. 3.1

[146] Matt Wytock and Zico Kolter. Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1265–1273, 2013. 4.2.1, 4.6.3

[147] Bo Xin, Yoshinobu Kawahara, Yizhou Wang, and Wen Gao. Efficient generalized fused lasso and its application to the diagnosis of alzheimers disease. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014. 1.2.1, 1.2.3, 2.1

[148] Eunho Yang, Pradeep Ravikumar, Genevera Allen, and Zhandong Liu. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems 25*, pages 1358–1366, 2012. 4.2.1

[149] Eunho Yang, Pradeep Ravikumar, Genevera Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16:3813–3847, 2015. 4.2.1, 4.4.2

[150] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. 1.3, 3.1, 3.1

[151] Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007. 4.2.1

[152] Xiao-Tong Yuan and Tong Zhang. Partial Gaussian graphical model estimation. *IEEE Transactions on Information Theory*, 60(3):1673–1687, 2014. 4.2.1

[153] Tong Zhang and Bin Yu. Boosting with early stopping: convergence and consistency. *Annals of Statistics*, 33(4):1538–1579, 2005. 3.1, 3.1