2010

Learning Exponential Families in High-Dimensions: Strong Convexity and Sparsity

Sham M. Kakade
University of Pennsylvania

Ohad Shamir
Hebrew University

Karthik Sridharan

Ambuj Tewari

Learning Exponential Families in High-Dimensions: Strong Convexity and Sparsity

Abstract
The versatility of exponential families, along with their attendant convexity properties, make them a popular and effective statistical model. A central issue is learning these models in high-dimensions when the optimal parameter vector is sparse. This work characterizes a certain strong convexity property of *general* exponential families, which allows their generalization ability to be quantified. In particular, we show how this property can be used to analyze generic exponential families under L_1 regularization.

Disciplines
Statistics and Probability

Learning Exponential Families in High-Dimensions: Strong Convexity and Sparsity

Sham M. Kakade
University of Pennsylvania

Ohad Shamir
Hebrew University

Karthik Sridharan
TTI Chicago

Ambuj Tewari
TTI Chicago

Abstract

The versatility of exponential families, along with their attendant convexity properties, make them a popular and effective statistical model. A central issue is learning these models in high-dimensions when the optimal parameter vector is sparse. This work characterizes a certain strong convexity property of *general* exponential families, which allows their generalization ability to be quantified. In particular, we show how this property can be used to analyze generic exponential families under L_1 regularization.

1 INTRODUCTION

Exponential models are perhaps the most versatile and pragmatic statistical models for a variety of reasons: modelling flexibility (encompassing discrete variables, continuous variables, covariance matrices, time series, graphical models, etc); convexity properties allowing easy optimization; and robust generalization ability. For large scale problems, a key issue is estimating these models when the dimension p of parameters is much larger than the sample size n (the "$p \gg n$" regime).

Much recent work has focused on this problem in the special case of linear regression in high dimensions, where it is assumed that the optimal parameter vector is sparse (e.g. Zhao and Yu (2006); Candes and Tao (2007); Meinshausen and Yu (2009); Bickel et al. (2008)). This body of prior work focused on: sharply characterizing the convergence rates for the prediction loss; consistent model selection; and obtaining sparse models. As we tackle more challenging problems, there is a growing need for model selection in more general exponential families. Recent work here includes learning Gaussian graphs (Ravikumar et al., 2008b)) and Ising models (Ravikumar et al., 2008a).

Classical results established that consistent estimation in *general* exponential families is possible, in the asymptotic limit where the number of dimensions is held constant (though some work establishes rates under certain conditions as p is allowed to grow slowly with n (Portnoy, 1988; Ghosal, 2000)). However, in modern problems, we typically grow p rapidly with n. While we have a handle on this question for a variety of special cases, a pressing question here is understanding how fast p can scale as a function of n in *general* exponential families. We need to quantify the relevant aspects of the particular family at hand that govern its convergence rate. This is the focus of this work. We should emphasize that throughout this paper, while we are interested in *modelling* with an exponential family, we *do not* necessarily assume that the data generating process is from this family.

Our Contributions and Related Work The key issue in analyzing the convergence rates of exponential families in terms of their prediction loss (which we take to be the log loss) is in characterizing the nature in which they are strictly convex. Roughly speaking, in the large n regime (with p kept fixed), we have a central limit theorem effect where the log loss of any exponential family approaches the log loss of a Gaussian, with a covariance matrix corresponding to the Fisher information matrix. Our first main contribution is quantifying the rate at which this effect occurs in general exponential families.

In particular, we show that every exponential family satisfies a certain rather natural growth rate condition on their standardized moments and standardized cumulants (recall that the k-th standardized moment is the *unitless* ratio of the k-th central moment to the k-th power of the standard deviation, which for $k = 3, 4$ is the skew and kurtosis). This condition is rather mild: these moments can grow as fast as $k!$. We show that this growth rate characterizes the rate at which the prediction loss of the exponential family behaves as

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

a *strongly convex* loss function. In particular, our analysis draws many parallels to that of Newton's method, where there is a "burn in" phase in which a number of iterations must occur until the function behaves as a locally quadratic function. In our statistical setting, we show that beyond a (quantified) "burn-in" sample size, the prediction loss inherits the desired strong convexity properties.

Our second contribution is an analysis of L_1 regularization in generic families, in terms of both prediction loss and the sparsity level of the selected model. Under a particular condition on the design matrix (the Restricted Eigenvalue (RE) condition in Bickel et al. (2008)), we show how L_1 regularization in general exponential families enjoys a convergence rate of $O(\frac{s \log p}{n})$ (where s is the number of relevant features). The RE condition is one of the least stringent conditions which permit this optimal convergence rate for linear regression (see Bickel et al. (2008)). Stronger mutual incoherence/irrepresentable conditions considered in Zhao and Yu (2006) also provide this rate. We show that an essentially identical convergence rate can be achieved for *general* exponential families. Our results are *non-asymptotic* and precisely relate n and p. Also, our results hold no matter what the distribution generating the data is. For recent related work under distributional assumptions, see Bunea (2008); Bach (2009); Negahban et al. (2009). Distribution-free results are obtained in van de Geer (2008) but the strong convexity aspect of the problem is not investigated but assumed away.

Our final contribution is one of *approximate* sparse model selection, i.e. obtaining a sparse model with low prediction loss. A drawback of the RE condition in comparison to the more stringent mutual incoherence condition is that the latter permits perfect recovery of the true features. However, for the case of the linear regression, Zhao and Yu (2006); Bickel et al. (2008) show that, under a sparse eigenvalue or RE condition, the L_1 solution itself is sparse (with a multiplicative increase in the sparsity level, that depends on a certain condition number of the design matrix). So, while the L_1 solution may not precisely recover the true model, it still is sparse and does recover those features with large true weights.

For general exponential families, while we do not have a characterization of the sparsity level of the L_1 solution (an interesting open question), we do provide a simple 2-stage procedure (thresholding and refitting) that provides a sparse model, with support on no more than merely $2s$ features and that has nearly as good performance. This result is novel even for the square loss case. Thus, even under the rather mild RE condition, we can obtain both a favorable convergence rate and a sparse model for generic families.

2 THE SETTING

Our samples $t \in \mathbb{R}^p$ are distributed independently according to D, and we model the process with $P(t|\theta)$, where $\theta \in \Theta$. However, we *do not* necessarily assume that D lies in this model class. The class of interest is *exponential families*, which, in their natural form, we denote by $P(t|\theta) = h_t \exp\{\langle \theta, t\rangle - \log Z(\theta)\}$, where t is the natural sufficient statistic for θ, and $Z(\theta)$ is the partition function. Here, Θ is the natural parameter space: the (convex) set where Z is finite. While we work with an exponential family in this general form, it should be kept in mind that t can be the sufficient statistic for some prediction variable y, or, for a generalized linear model (such as for logistic/linear regression), t can be a function of both y and some covariate X (see Dobson (1990)). We return to this point later.

Our prediction loss is the log-loss and θ^* is the optimal parameter vector, i.e. $\mathcal{L}(\theta) = \mathbb{E}_{t\sim D}[-\log P(t|\theta)]$, and $\theta^\star = \operatorname{argmin}_{\theta\in\Theta} \mathcal{L}(\theta)$. We assume that θ^* is an interior point of Θ. Later we consider the case where $\theta^\star$ is sparse. We denote the Fisher information of $P(\cdot|\theta^\star)$ as $\mathcal{F}^\star = \mathbb{E}_{t\sim P(\cdot|\theta^\star)}\left[-\nabla^2 \log P(t|\theta^\star)\right]$, under the model of $\theta^\star$. The induced "Fisher risk" is $\|\theta - \theta^\star\|^2_{\mathcal{F}^\star} = (\theta - \theta^\star)^\top \mathcal{F}^\star(\theta - \theta^\star)$. We also consider the L_1 risk $\|\theta - \theta^\star\|_1$.

For a sufficiently large sample size, we expect that the Fisher risk of a reasonable estimator $\hat{\theta}$, $\|\hat{\theta} - \theta^\star\|^2_{\mathcal{F}^\star}$, will be close to $\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta^\star)$. One of our main contributions (Thm. 3.4) is quantifying when this occurs in general exponential families. This characterization is then used to quantify the convergence rate for L_1 methods in these families. We expect this strong convexity property to be useful for characterizing the performance of other regularization methods as well. All proofs are postponed till the appendix.

3 (Almost) STRONG CONVEXITY OF EXPONENTIAL FAMILIES

We first consider a certain bounded growth rate condition for standardized moments and standardized cumulants, satisfied by all exponential families. This growth rate is fundamental in establishing how fast the prediction loss behaves as a quadratic function.

3.1 ANALYTIC STANDARDIZED MOMENTS AND CUMULANTS

Moments: For a univariate random variable (r.v.) z distributed as ρ, denote its k-th central moment by $m_{k,\rho}(z) = \mathbb{E}_{z\sim\rho}\left[z - m_{1,\rho}(z)\right]^k$, where $m_{1,\rho}(z)$ is

the mean $\mathbb{E}_{z\sim\rho}[z]$. Recall that the k-th standardized moment is the ratio of the k-th central moment to the k-th power of the standard deviation, i.e. $m_{k,\rho}(z)/m_{2,\rho}(z)^{k/2}$. This normalization by the standard deviation makes the standard moments unitless quantities. For $k = 3, 4$, the standardized moments are the skew and kurtosis. We now define the *analytic* standardized moment for z. We use the term analytic to reflect that if the moment generating function of z is analytic[1] then z has an analytic moment.

Definition 3.1. *Let z be a univariate r.v. under ρ. Then z has an* analytic standardized moment *of α if the standardized moments exist and,*

$$\forall k \geq 3, \quad \left|\frac{m_{k,\rho}(z)}{m_{2,\rho}(z)^{k/2}}\right| \leq \tfrac{1}{2}k!\ \alpha^{k-2}$$

(where the above is assumed to hold if the denominator is 0). If $t \in \mathbb{R}^p$ is a multivariate r.v., we say that t has an analytic standardized moment of α with respect to a subspace $\mathcal{V} \subset \mathbb{R}^p$ if the above bound holds for all univariate $z = \langle v, t\rangle$ where $v \in \mathcal{V}$.

This condition is rather mild: the standardized moments can increase as fast as $k!\alpha^{k-2}$. This condition is closely related to those used in obtaining sharp exponential type tail bounds for the convergence of a r.v. to its mean. In fact, the Bernstein conditions (Bernstein, 1946) are almost identical, except that they use the k-th raw moments[2]. These moment conditions are weaker than requiring "sub-Gaussian" tails.

Cumulants: Recall that the cumulant-generating function f of z under ρ is the log of the moment-generating function, if it exists: $f(s) = \log \mathbb{E}[e^{sz}]$. The cumulants are given by the derivatives of f at 0: $c_{k,\rho}(z) = f^{(k)}(0)$. The 1st, 2nd and 3rd cumulants and central moments are identical. Higher cumulants are neither moments nor central moments, but rather more complicated polynomial functions of the moments. Analogously, the k-th standardized cumulant is $c_{k,\rho}(z)/c_{2,\rho}(z)^{k/2}$. Again, normalization by the standard deviation (the 2nd cumulant is the variance) makes these unitless quantities. Cumulants are viewed as equally fundamental as central moments, and we make use of their behavior as well. In certain settings, it is more natural to work with the cumulants. We define the *analytic* standardized cumulant analogously:

Definition 3.2. *Let z be a univariate r.v. under ρ. Then z has an* analytic standardized cumulant *of α if the standardized cumulants exist and,*

$$\forall k \geq 3, \quad \left|\frac{c_{k,\rho}(z)}{c_{2,\rho}(z)^{k/2}}\right| \leq \tfrac{1}{2}k!\ \alpha^{k-2}$$

(where the above is assumed to hold if the denominator is 0). If $t \in \mathbb{R}^p$ is a multivariate r.v., we say that t has an analytic standardized cumulant of α with respect to a subspace $\mathcal{V} \subset \mathbb{R}^p$ if the above bound holds for all univariate $z = \langle v, t\rangle$ where $v \in \mathcal{V}$.

Existence: While we do not expect analytic moments to exist for all distributions (e.g. heavy tailed ones), the next lemma shows that exponential families have (finite) analytic standardized moments and cumulants.

Lemma 3.3. *If t is the sufficient statistic of an exponential family with parameter θ and θ is an interior point of the natural parameter space, then t has both a finite analytic standardized moment and a finite analytic standardized cumulant, with respect to all of $\mathbb{R}^p$.*

We skip the technical proof that follows easily from the analyticity of the moment and cumulant generating functions. Reassured by this existence result, let us now consider some concrete examples.

3.2 EXAMPLES

Going through the examples, there are two issues to bear in mind. First, α is quantified only at a particular θ (later, $\theta^\star$ is the point we will be interested in). Note that we do not require any uniform conditions on any derivatives over all θ. Second, we are interested in how α could depend on the dimensionality. In some cases, α is dimension free and in other cases (like for generalized linear models), α depends on the dimension through spectral properties of $\mathcal{F}^\star$ (this dimension dependence can be relaxed in the sparse case that we consider, as discussed later).

3.2.1 One Dimensional Families

Bernoulli In the canonical form, the Bernoulli distribution is $P(y|\theta) = \exp\left(y\theta - \log(1 + e^\theta)\right)$ with $\theta \in \mathbb{R} = \Theta$. We have $m_1(\theta^\star) = e^{\theta^\star}/(1 + e^{\theta^\star})$. The central moments satisfy $m_2(\theta^\star) = m_1(\theta^\star)(1 - m_1(\theta^\star))$ and $m_k(\theta^\star) \leq m_2(\theta^\star)$ for $k \geq 3$. Thus, $\alpha = 1/\sqrt{m_2(\theta^\star)}$ is a standardized analytic moment at any $\theta^\star \in \Theta$. Further, $c_k(\theta^\star) \leq c_2(\theta^\star) = m_2(\theta^\star)$ for $k \geq 3$. Thus, α is also a standardized analytic cumulant at any $\theta^\star \in \Theta$.

Unit variance Gaussian In the canonical form, unit variance Gaussian is $P(y|\theta) \propto \exp(-y^2/2)\exp(y\theta - \theta^2/2)$ with $\theta \in \mathbb{R} = \Theta$ and $m_1(\theta^\star) = \theta^\star$, $m_2(\theta^\star) = 1$. Odd central moments are 0 and for even $k \geq 4$, we have $m_k(\theta^\star) = k!/2^{k/2}(k/2)!$. Thus, $\alpha = 1$ is a standardized analytic moment at any $\theta^\star \in \Theta$. However, the log-likelihood is already quadratic in this case (as

[1] Recall that a real valued function is analytic on some domain of $\mathbb{R}^p$ if the derivatives of all orders exist, and if for each interior point, the Taylor series converges in some sufficiently small neighborhood of that point.

[2] The Bernstein inequalities used in deriving tail bounds require that, for all $k \geq 2$, $\frac{\mathbb{E}[z^k]}{\mathbb{E}[z^2]} \leq \frac{1}{2}k!L^{k-2}$ for some constant L (which has units of z).

we shall see, there should be no "burn in" phase until it begins to look like a quadratic!). This becomes evident if we consider the cumulants instead. All cumulants $c_k(\theta^\star) = 0$ for $k \geq 3$ and hence $\alpha = 0$ is a standardized analytic cumulant at any $\theta^\star \in \Theta$.

3.2.2 Gaussian Covariance Estimation

Consider a mean zero high-dimensional ($p \gg 1$) multivariate Normal parameterized by the precision matrix Θ, $P(Y|\Theta) \propto \exp\left(-\frac{1}{2}\langle \Theta, YY^\top\rangle + \log\det(\Theta)\right)$. A "direction" here is a positive semi-definite (p.s.d.) matrix V, and we seek the cumulants of the r.v. $\langle V, YY^\top\rangle$. Note that $YY^\top$ has Wishart distribution $W_p(\Theta^{-1}, 1)$ with the moment generating function, $\det\left(\mathbf{I} - 2V\Theta^{-1}\right)^{-1/2}$. Let λ_i's be the eigenvalues of $V\Theta^{-1}$. Then, taking logs, the cumulant generating function is $f(s) = \log \mathbb{E}\left[\exp(s\langle V, YY^\top\rangle)\right] = \frac{-1}{2}\sum_{i=1}^p \log(1 - 2s\lambda_i)$. The derivatives are $f^{(k)}(s) = \frac{1}{2}\sum_{i=1}^p \frac{(k-1)!(2\lambda_i)^k}{(1-2s\lambda_i)^k}$. Thus, the cumulant $c_{k,\Theta}(V) = f^{(k)}(0) = 2^{k-1}(k-1)!\sum_i \lambda_i^k$. Hence, for $k \geq 3$,

$$\frac{c_{k,\Theta}(V)}{(c_{2,\Theta}(V))^{k/2}} = \frac{2^{k-1}(k-1)!\sum_i \lambda_i^k}{(2\sum_i \lambda_i^2)^{k/2}} \leq \frac{1}{2}2^{k/2-1}\cdot k!\ .$$

Thus, $\alpha = \sqrt{2}$ is a standardized analytic cumulant at Θ. It is harder to estimate the central moments in this case. This example is also interesting in connection to the analysis of Newton's method as $\log\det(\Theta)$ is self-concordant on the cone of p.s.d. matrices.

3.2.3 Generalized Linear Models

Suppose we have some covariate, response pair (X, Y) drawn from some distribution D. Consider a family of distributions $P(\cdot|\theta; X)$ such that, for each X, it is an exponential family with natural sufficient statistic $t_{y,X}$, $P(y|\theta; X) = h_y \exp\left(\langle \theta, t_{y,X}\rangle - \log Z_X(\theta)\right)$. The loss we consider is $\mathcal{L}(\theta) = \mathbb{E}_{X,Y\sim D}\left[-\log P(y|\theta; X)\right]$. A special case of this setup is as follows. Say we have a 1-dimensional exponential family $q_\nu(y) = h_y \exp(y\nu - \log Z(\nu))$, where $y, \nu \in \mathbb{R}$. The family $P(\cdot|\theta; X)$ can be simply $q_{\langle\theta,X\rangle}$ (i.e. take $\nu = \langle\theta, X\rangle$). Thus, $P(y|\theta; X) = h_y \exp\left(y\langle\theta, X\rangle - \log Z(\langle\theta, X\rangle)\right)$. We see that $t_{y,X} = yX$ and $Z_X(\theta) = Z(\langle\theta, X\rangle)$. For example, when q_ν is either the Bernoulli family or the Gaussian family, this corresponds to *logistic regression* or *least squares regression*, respectively. It turns out that the analog of having a standardized analytic moment of α at θ w.r.t. a direction v is to have $m_{k,\theta}(v)/(m_{2,\theta}(v))^{k/2} \leq \frac{1}{2}k!\alpha^{k-2}$, where $m_{k,\theta}(v) = \mathbb{E}_X\left[m_{k,P(\cdot|\theta;X)}(\langle t_{y,X}, v\rangle)\right]$. Here, the expectation is under $X \sim D_X$, the marginal of D on X. If $\|t_{y,X}\|_2 \leq B$ and the expected Fisher information matrix[3] has minimum eigenvalue $\lambda_{\min}$, then $\alpha = B/\lambda_{\min}$. Note that $\lambda_{\min}$ could be small but it arose only because we are considering arbitrary directions v. If the set of directions is smaller, one can often get less pessimistic bounds (see Sec. 5.0.3). Also note that similar bounds can be derived assuming subgaussian tails for $t_{y,X}$ rather than assuming it is bounded.

[3] i.e. $\mathbb{E}_X\left[\mathbb{E}_{y\sim P(\cdot|\theta;X)}\left[-\nabla^2 \log P(y|\theta; X)\right]\right]$.

3.3 ALMOST STRONG CONVEXITY

Recall that a strictly convex function F is strongly convex if the Hessian $\nabla^2 F$ has a (uniformly) lower bounded eigenvalue. In general, exponential families behave in a strongly convex manner only in a (sufficiently small) neighborhood of $\theta^\star$. Our first main result quantifies when this behavior is exhibited.

Theorem 3.4. *(Almost Strong Convexity) Let α be the analytic standardized moment/cumulant under $\theta^\star$ w.r.t. a subspace $\mathcal{V}$. For any θ s.t. $\theta - \theta^\star \in \mathcal{V}$, if*

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^\star) \leq \tfrac{1}{65\alpha^2} \quad \text{or} \quad \|\theta - \theta^\star\|_{\mathcal{F}^\star}^2 \leq \tfrac{1}{16\alpha^2}$$

then

$$\tfrac{1}{4}\|\theta - \theta^\star\|_{\mathcal{F}^\star}^2 \ \leq\ \mathcal{L}(\theta) - \mathcal{L}(\theta^\star)\ \leq\ \tfrac{3}{4}\|\theta - \theta^\star\|_{\mathcal{F}^\star}^2\ .$$

Both preconditions can be thought of as a "burn in" phase. The idea is that initially, a certain number of samples is needed until the loss of θ is somewhat close to the minimal loss; after which, the quadratic lower bound engages. This is analogous to the analysis of the Newton's method, which quantifies the number of steps needed to enter the quadratically convergent phase. The constants of 1/4 and 3/4 can be made arbitrarily close to 1/2 (with a longer "burn in" phase). A key idea in the proof is an expansion of the prediction regret in terms of the moments/cumulants. We use the shorthand notation of $c_{k,\theta}(\Delta)$ and $m_{k,\theta}(\Delta)$ to denote the cumulants and moments of the r.v. $\langle\Delta, t\rangle$ under the distribution $P(\cdot|\theta)$.

Lemma 3.5. *(Moment and Cumulant Expansion) Define $\Delta = \theta - \theta^\star$. For all $s \in [0, 1]$,*

$$\mathcal{L}(\theta^\star + s\Delta) - \mathcal{L}(\theta^\star) = \sum_{k=2}^{\infty} \tfrac{1}{k!} c_{k,\theta^\star}(\Delta) s^k$$

$$\mathcal{L}(\theta^\star + s\Delta) - \mathcal{L}(\theta^\star) = \log\left(1 + \sum_{k=2}^{\infty} \tfrac{1}{k!} m_{k,\theta^\star}(\Delta) s^k\right)$$

where the equalities hold if the r.h.s. converges.

The relatively straightforward proof of this lemma is skipped. The key technical step in the proof of Thm. 3.4 is characterizing when these expansions converge. Even if $\|\theta - \theta^\star\|_{\mathcal{F}^\star}^2 \leq \frac{1}{16\alpha^2}$ (one of our preconditions), a direct attempt at lower bounding $\mathcal{L}(\theta) - \mathcal{L}(\theta^\star)$

using the above expansions with the analytic moment condition would not imply these expansions converge; the proof requires a more delicate argument.

4 SPARSITY

We now consider the case where θ^* is sparse, with support S and sparsity level s, i.e.

$$S = \{i : [\theta^\star]_i \neq 0\}, \quad s = |S| \ .$$

To understand when L_1 regularized algorithms (for linear regression) converge at a rate comparable to that of L_0 algorithms (subset selection), Meinshausen and Yu (2009) considered a sparse eigenvalue condition on the design matrix, where the eigenvalues on any small (sparse) subset are bounded away from 0. Bickel et al. (2008) relaxed this condition by considering vectors most of whose support is on a small subset (see Bickel et al. (2008) for a discussion). We also consider this relaxed condition, but now on the Fisher matrix.

Assumption 4.1. *(Restricted Fisher Eigenvalues) For $\delta \in \mathbb{R}^p$, let $\delta_S \in \mathbb{R}^p$ be defined as $[\delta_S]_i = \delta_i \mathbf{1}_{(i \in S)}$ and let S^C be the complement of S. Assume that:*

$$\forall \delta \ s.t. \ \|\delta_{S^C}\|_1 \leq 3\|\delta_S\|_1, \quad \|\delta\|_{\mathcal{F}^\star} \geq \kappa^\star_{\min}\|\delta_S\|_2$$
$$\forall \delta \ s.t. \ \delta_{S^C} = 0, \quad \|\delta\|_{\mathcal{F}^\star} \leq \kappa^\star_{\max}\|\delta_S\|_2$$

The constant of 3 is for convenience. Note we only quantify on the support S: a substantially weaker condition than in Meinshausen and Yu (2009); Bickel et al. (2008), who quantify over *all* subsets (in fact, many previous algorithms/analysis actually use this condition on subsets different from S, e.g. Meinshausen and Yu (2009); Candes and Tao (2007); Zhang (2008)). Furthermore, with regards to our analyticity conditions, our proof shows that the subspace of directions we need to consider is now restricted to the set:

$$\mathcal{V} = \{v : \|v_{S^C}\|_1 \leq 3\|v_S\|_1\} \tag{1}$$

Under this Restricted Eigenvalue (RE) condition, we can replace the minimal eigenvalue used in Example 3.2.3 by $\kappa^\star_{\min}$ (section 5.0.3 in appendix), which could be significantly smaller.

4.1 FISHER RISK

Consider the L_1 regularized problem:

$$\hat{\theta} = \text{argmin}_{\theta \in \Theta} \ \widehat{\mathbb{E}}[-\log P(y|\theta)] + \lambda\|\theta\|_1 \tag{2}$$

where the empirical expectation is with respect to a sample. This reduces to the usual linear regression example (for Gaussian means) and involves the log-determinant in Gaussian graph setting (considered in Ravikumar et al. (2008b)) where θ is the precision matrix (see Example 3.2.2). Our next main result provides risk bounds for $\hat{\theta}$, under the RE condition. Typically, the regularization parameter λ is specified as a function of the noise level, under a particular noise model (e.g. for linear regression case, where $Y = \beta X + \eta$ with the noise model $\eta \sim \mathcal{N}(0, \sigma^2)$, λ is specified as $\sigma\sqrt{\frac{\log p}{n}}$ (Meinshausen and Yu, 2009; Bickel et al., 2008)). Here, our theorem is stated in a deterministic manner, to explicitly show that an appropriate value of λ is determined by the L_∞ norm of the measurement error $\|\mathbb{E}[t] - \widehat{\mathbb{E}}[t]\|_\infty$. We then easily quantify λ in a corollary under a mild distributional assumption. Also, we must have that this measurement error be (quantifiably) sufficiently small such that our "burn in" condition holds.

Theorem 4.2. *(Risk) Suppose that Assumption 4.1 holds and λ satisfies both*

$$\|\mathbb{E}[t] - \widehat{\mathbb{E}}[t]\|_\infty \ \leq \ \frac{\lambda}{2} \quad \text{and} \quad \lambda \leq \frac{1}{100\alpha^{\star 2}\|\theta^\star\|_1} \tag{3}$$

where $\alpha^\star$ is the analytic standardized moment or cumulant of $\theta^\star$ for the subspace $\mathcal{V}$ defined in (1)*. Then if $\hat{\theta}$ is a solution to* (2)*, the Fisher risk is bounded as:*

$$\tfrac{1}{4}\|\hat{\theta} - \theta^\star\|^2_{\mathcal{F}^\star} \leq \mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta^\star) \leq \frac{9s\lambda^2}{{\kappa^\star_{\min}}^2}$$

and the L_1 risk is bounded as:

$$\|\hat{\theta} - \theta^\star\|_1 \leq \frac{24s\lambda}{{\kappa^\star_{\min}}^2} \ .$$

We expect the measurement error $\|\mathbb{E}[t] - \widehat{\mathbb{E}}[t]\|_\infty$ to be $O(\sigma\sqrt{\log p/n})$, so we think of $\lambda = O(\sigma\sqrt{\log p/n})$. This would recover the usual (optimal) risk bound of $O(\sigma^2 \frac{s \log p}{n})$ Note that the mild dimension dependence enters through the measurement error. Hence, our theorem shows that *all* exponential families exhibit favorable convergence rates under the RE condition. The following proposition and corollary quantify this under a mild (and standard) distributional assumption.

Proposition 4.3. *If t is sub-Gaussian, ie. there exists $\sigma \geq 0$ such that $\forall i$ and $\forall s \in \mathbb{R}$, $\mathbb{E}\left[e^{s(t_i - \mathbb{E}t_i)}\right] \leq e^{\sigma^2 s^2/2}$, then for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\|\mathbb{E}[t] - \widehat{\mathbb{E}}[t]\|_\infty \leq \sigma\sqrt{\frac{\log\left(\frac{p}{\delta}\right)}{n}}$$

Bounded r.v.'s are sub-Gaussian (though boundedness is not necessary: Gaussian r.v.'s are obviously sub-Gaussian). The following corollary is immediate.

Corollary 4.4. *Suppose Assumption 4.1 and sub-Gaussian condition in Proposition 4.3 hold. For $n \geq K\alpha^{\star 4}\|\theta^\star\|_1^2\sigma^2 \log\left(\frac{p}{\delta}\right)$, (K is a universal constant), and $\lambda = 2\sigma\sqrt{\frac{\log(p/\delta)}{n}}$, with probability at least $1 - \delta$,*

$$\|\hat{\theta} - \theta^\star\|^2_{\mathcal{F}^\star} \leq \left(\frac{36}{{\kappa^\star_{\min}}^2}\right)\frac{\sigma^2 s \log\left(\frac{p}{\delta}\right)}{n} \ ,$$

$$\|\hat{\theta} - \theta^\star\|_1 \leq \frac{48\sigma s}{{\kappa^\star_{\min}}^2}\sqrt{\frac{\log(\frac{p}{\delta})}{n}} \ .$$

4.2 APPROXIMATE MODEL SELECTION

An important issue unaddressed by the previous result is the sparsity level of our estimate $\hat{\theta}$. For the linear regression case, Meinshausen and Yu (2009); Bickel et al. (2008) show that the L_1 solution is actually sparse, with a sparsity level of roughly $O((\frac{\kappa^\star_{\max}}{\kappa^\star_{\min}})^2 s)$, In the general setting, we do not have a characterization of the sparsity level of the L_1 solution. However, we now present a 2-stage procedure, which provides an estimate with support on merely $2s$ features, with nearly as good risk. Consider the procedure where we select the set of coordinates which have large weight under $\hat{\theta}$ (greater than some threshold τ). Then we refit to find an estimate with support only on these coordinates. That is, we restrict our estimate to the set $\Theta_\tau = \{\theta \in \Theta : \theta_i = 0 \text{ if } |\hat{\theta}_i| \leq \tau\}$. This algorithm is:

$$\tilde{\theta} = \text{argmin}_{\theta \in \Theta_\tau} \ \hat{\mathcal{L}}(\theta) + \lambda\|\theta\|_1 \tag{4}$$

Theorem 4.5. *(Sparsity) Suppose that 4.1 holds and the regularization parameter λ satisfies both*

$$\begin{aligned} \|\mathbb{E}[t] - \widehat{\mathbb{E}}[t]\|_\infty &\leq \tfrac{\lambda}{2} \\ \lambda \leq \min\{&\tfrac{1}{270{\alpha^\star}^2\|\theta^\star\|_1}, \tfrac{{\kappa^\star_{\min}}^2}{340\kappa^\star_{\max}\alpha^\star\sqrt{s}}\} \end{aligned} \tag{5}$$

where $\alpha^\star$ is the analytic standardized moment or cumulant of $\theta^\star$ for the subspace $\mathcal{V}$ defined in (1)*. If $\hat{\theta}, \tilde{\theta}$ are solutions of* (2), (4) *respectively with this λ and $\tau = \frac{18\lambda}{{\kappa^\star_{\min}}^2}$, then* **(i)** *$\tilde{\theta}$ has support on at most $2s$ coordinates, and* **(ii)** *the Fisher risk is bounded as:*

$$\tfrac{1}{4}\|\hat{\theta} - \theta^\star\|^2_{\mathcal{F}^\star} \leq \mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta^\star) \leq \left(12\frac{\kappa^\star_{\max}}{\kappa^\star_{\min}}\right)^2 \frac{9\ s\lambda^2}{{\kappa^\star_{\min}}^2}$$

Using Proposition 4.3, we have following corollary.

Corollary 4.6. *Suppose Assumption 4.1 and sub-Gaussian condition in Proposition 4.3 hold. For $n \geq K{\alpha^\star}^2\sigma^2\log(p/\delta)\max\{\frac{s{\kappa^\star_{\max}}^2}{{\kappa^\star_{\min}}^4}, {\alpha^\star}^2\|\theta^\star\|_1^2\}$ (K is a universal constant), $\lambda = 2\sqrt{\sigma^2\log(p/\delta)/n}$, and $\tau = 36\sqrt{\sigma^2\log(p/\delta)/(n{\kappa^\star_{\min}}^2)}$, with probability at least $1-\delta$,*

$$\|\tilde{\theta} - \theta^\star\|^2_{\mathcal{F}^\star} \leq \left(12\frac{\kappa^\star_{\max}}{\kappa^\star_{\min}}\right)^2 \left(\frac{36}{{\kappa^\star_{\min}}^2}\right)\frac{s\sigma^2\log\left(\frac{p}{\delta}\right)}{n} \ .$$

References

Francis Bach. Self-concordant analysis for logistic regression. *CoRR*, abs/0910.4627, 2009.

S. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2008.

Florentina Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1+\ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313, 2007.

A.J. Dobson. *An Introduction to Generalized Linear Models*. Chapman and Hall, 1990.

Subhashis Ghosal. Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivar. Anal.*, 74(1): 49–68, 2000. ISSN 0047-259X.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37:246, 2009.

S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *NIPS*, 2009.

S. Portnoy. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Annals of Statistics*, 16, 1988.

P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional ising model selection using l1-regularized logistic regression. Technical Report Technical Report 750, UC Berkeley, Department of Statistics., 2008a.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. Technical Report arXiv:0811.3628, Nov 2008b.

Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.

T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Advances in Neural Information Processing Systems 22*, 2008.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

5 APPENDIX

5.0.1 Proof of Theorem 3.4

We slightly abuse notation and let $m_k(\Delta)$ be the k-th central moment of the univariate r.v. $\langle\Delta, t\rangle$ distributed under $\theta^\star$. The following upper and lower bounds are useful in that they guarantee the sum converges for the choice of s specified.

Lemma 5.1. *Let α and θ be defined as in Thm. 3.4. Let $\Delta = \theta - \theta^\star$ and set $s = \min\{\frac{1}{4\alpha\sqrt{m_2(\Delta)}}, 1\}$. If α is an analytic moment, then*

$$\tfrac{1}{3}\frac{m_2(\Delta)}{\max\{16\alpha^2 m_2(\Delta), 1\}} \ \leq \ \sum_{k=2}^{\infty}\frac{m_k(\Delta)s^k}{k!} \ \leq \ \tfrac{2}{3}\frac{m_2(\Delta)}{\max\{16\alpha^2 m_2(\Delta), 1\}}$$

If α is an analytic cumulant, then

$$\tfrac{1}{3}\frac{c_2(\Delta)}{\max\{16\alpha^2 c_2(\Delta),1\}} \;\le\; \sum_{k=2}^{\infty}\frac{c_k(\Delta)s^k}{k!} \;\le\; \tfrac{2}{3}\frac{c_2(\Delta)}{\max\{16\alpha^2 c_2(\Delta),1\}}$$

Proof. See supplementary material. □

The core lemma below leads to the proof of Thm. 3.4.

Lemma 5.2. *Let α, θ be as in Thm. 3.4. We have:*

$$\tfrac{1}{4}\;\frac{\|\theta-\theta^\star\|^2_{\mathcal{F}^\star}}{\max\{16\alpha^2\|\theta-\theta^\star\|^2_{\mathcal{F}^\star},1\}} \;\le\; \mathcal{L}(\theta)-\mathcal{L}(\theta^\star) \qquad (6)$$

Furthermore, if $\|\theta-\theta^\star\|_{\mathcal{F}^\star}\le\frac{1}{16\alpha^2}$,

$$\tfrac{1}{4}\|\theta-\theta^\star\|^2_{\mathcal{F}^\star} \;\le\; \mathcal{L}(\theta)-\mathcal{L}(\theta^\star) \;\le\; \tfrac{2}{3}\|\theta-\theta^\star\|^2_{\mathcal{F}^\star}$$

Proof. See supplementary material. □

We are now ready to prove Theorem 3.4.

Proof of Theorem 3.4. If $\|\theta-\theta^\star\|^2_{\mathcal{F}^\star}\le\frac{1}{16\alpha^2}$, then the previous lemma implies the claim. Let us therefore assume $\mathcal{L}(\theta)-\mathcal{L}(\theta^\star)\le\frac{1}{65\alpha^2}$. If $\|\theta-\theta^\star\|^2_{\mathcal{F}^\star}\le\frac{1}{16\alpha^2}$, then we are done by the previous argument. So let $\|\theta-\theta^\star\|^2_{\mathcal{F}^\star} > \frac{1}{16\alpha^2}$. Hence, $\max\{16\alpha^2 m_2(\Delta),1\} = 16\alpha^2 m_2(\Delta)$. Using (6), we have that $\frac{1}{64\alpha^2}\le\mathcal{L}(\theta)-\mathcal{L}(\theta^\star)$, which is a contradiction. □

5.0.2 Proof of Theorem 4.2

Let $\widehat{\mathcal{L}}(\theta)=\widehat{\mathbb{E}}[-\log P(y|\theta)]$, $T=\mathbb{E}[t]$ and $\hat{T}=\widehat{\mathbb{E}}[t]$.

Lemma 5.3. *Suppose $\|T-\hat{T}\|_\infty\le\lambda/2$. Let $\hat{\theta}$ be a solution to* (2)*. Then, $\forall\theta\in\Theta$:*

$$\mathcal{L}(\hat{\theta})-\mathcal{L}(\theta)\le\tfrac{\lambda}{2}\|\hat{\theta}-\theta\|_1+\lambda\|\theta\|_1-\lambda\|\hat{\theta}\|_1\le\tfrac{3\lambda}{2}\|\theta\|_1 \quad (7)$$

Further, suppose that θ only has support on S, then:

$$\mathcal{L}(\hat{\theta})-\mathcal{L}(\theta)\le\tfrac{3\lambda}{2}\|\hat{\theta}_S-\theta\|_1 \qquad (8)$$

Proof. Since $\hat{\theta}$ solves (2), we have:

$$-\langle\hat{\theta},\hat{T}\rangle+\log Z(\hat{\theta})+\lambda\|\hat{\theta}\|_1\le-\langle\theta,\hat{T}\rangle+\log Z(\theta)+\lambda\|\theta\|_1$$

Hence,

$$\begin{aligned}&-\langle\hat{\theta},T\rangle+\log Z(\hat{\theta})+\lambda\|\hat{\theta}\|_1\\&\quad\le\langle\hat{\theta}-\theta,\hat{T}-T\rangle-\langle\theta,T\rangle+\log Z(\theta)+\lambda\|\theta\|_1\end{aligned}$$

The condition on λ gives the 1st inequality in (7):

$$\begin{aligned}\mathcal{L}(\hat{\theta})-\mathcal{L}(\theta)&\le\|\hat{\theta}-\theta\|_1\|\hat{T}-T\|_\infty+\lambda\|\theta\|_1-\lambda\|\hat{\theta}\|_1\\&\le\tfrac{\lambda}{2}\|\hat{\theta}-\theta\|_1+\lambda\|\theta\|_1-\lambda\|\hat{\theta}\|_1\end{aligned}$$

The 2nd inequality in (7) is by triangle inequality. For the final claim, using sparsity of θ, we have:

$$\begin{aligned}&\tfrac{\lambda}{2}\|\hat{\theta}-\theta\|_1+\lambda\|\theta\|_1-\lambda\|\hat{\theta}\|_1\\&=\tfrac{\lambda}{2}\|\hat{\theta}_S-\theta\|_1+\tfrac{\lambda}{2}\|\hat{\theta}_{S^C}\|_1+\lambda\left(\|\theta\|_1-\|\hat{\theta}_S\|_1\right)-\lambda\|\hat{\theta}_{S^C}\|_1\\&\le\tfrac{\lambda}{2}\|\hat{\theta}_S-\theta\|_1+\lambda\|\hat{\theta}_{S^C}\|_1+\lambda\|\hat{\theta}_S-\theta\|_1-\lambda\|\hat{\theta}_{S^C}\|_1\\&=\tfrac{3\lambda}{2}\|\hat{\theta}_S-\theta\|_1\;.\end{aligned}$$

□

Lemma 5.4. *Suppose that* (3) *holds. Let $\hat{\theta}$ be a solution the optimization problem in* (2)*. For any $\theta\in\Theta$, which only has support on S and such that $\mathcal{L}(\hat{\theta})\ge\mathcal{L}(\theta)$, then:*

$$\|\hat{\theta}_{S^C}\|_1\le 3\|\hat{\theta}_S-\theta\|_1 \qquad (9)$$

$$\|\hat{\theta}-\theta\|_1\le 4\|\hat{\theta}_S-\theta\|_1 \qquad (10)$$

Proof. By assumption on θ and (7),

$$0\le\mathcal{L}(\hat{\theta})-\mathcal{L}(\theta)\le\tfrac{\lambda}{2}\|\hat{\theta}-\theta\|_1+\lambda\|\theta\|_1-\lambda\|\hat{\theta}\|_1$$

Dividing by λ and adding $\frac{1}{2}\|\hat{\theta}-\theta\|_1$ to both sides,

$$\tfrac{1}{2}\|\hat{\theta}-\theta\|_1\le\|\hat{\theta}-\theta\|_1+\|\theta\|_1-\|\hat{\theta}\|_1$$

For $i\notin S$, $|\hat{\theta}_i-\theta_i|+|\theta_i|-|\hat{\theta}_i|=0$. Hence,

$$\tfrac{1}{2}\|\hat{\theta}-\theta\|_1\le\|\hat{\theta}_S-\theta\|_1+\|\theta\|_1-\|\hat{\theta}_S\|_1\le 2\|\hat{\theta}_S-\theta\|_1$$

This proves (10). From this, $\frac{1}{2}\|\hat{\theta}_S-\theta\|_1+\frac{1}{2}\|\hat{\theta}_{S^C}\|_1=\frac{1}{2}\|\hat{\theta}-\theta\|_1\le 2\|\hat{\theta}_S-\theta\|_1$ which proves (9), after rearranging. □

Proof of Theorem 4.2. First, by (3) and (7) we see that $\mathcal{L}(\hat{\theta})-\mathcal{L}(\theta^\star)\le\frac{1}{65\alpha^{\star 2}}$. Hence using Thm. 3.4 we see that

$$\tfrac{1}{4}\|\hat{\theta}-\theta^\star\|^2_{\mathcal{F}^\star}\le\mathcal{L}(\hat{\theta})-\mathcal{L}(\theta^\star)$$

On the other hand observe that:

$$\|\hat{\theta}_S-\theta^\star\|_1\le\sqrt{s}\|\hat{\theta}_S-\theta^\star\|_2\le\tfrac{\sqrt{s}}{\kappa^\star_{\min}}\|\hat{\theta}-\theta^\star\|_{\mathcal{F}^\star} \quad (11)$$

where the last step uses the RE condition, Assumption 4.1 (note that $\hat{\theta}$ satisfies the RE precondition, so $\hat{\theta}-\theta^\star\in\mathcal{V}$). Now using the above with (8) we have that $\frac{1}{4}\|\hat{\theta}-\theta^\star\|^2_{\mathcal{F}^\star}\le\mathcal{L}(\hat{\theta})-\mathcal{L}(\theta^\star)\le\frac{3\lambda\sqrt{s}}{2\kappa^\star_{\min}}\|\hat{\theta}-\theta^\star\|_{\mathcal{F}^\star}$. Hence,

$$\|\hat{\theta}-\theta^\star\|_{\mathcal{F}^\star}\le\tfrac{6\lambda\sqrt{s}}{\kappa^\star_{\min}} \qquad (12)$$

and so $\frac{1}{4}\|\hat{\theta}-\theta^\star\|^2_{\mathcal{F}^\star}\le\mathcal{L}(\hat{\theta})-\mathcal{L}(\theta^\star)\le\frac{9\lambda^2 s}{\kappa^{\star 2}_{\min}}$, which proves the first claim. Now plugging (12) into (11), we get $\|\hat{\theta}_S-\theta^\star\|_1\le\frac{6\lambda s}{\kappa^{\star 2}_{\min}}$. Hence by (10), $\|\hat{\theta}-\theta^\star\|_1$ is bounded by $\frac{24\lambda s}{\kappa^{\star 2}_{\min}}$. □

5.0.3 Analytic Standardized Moment for GLM and Sparsity

In the GLM example in Section 3.2.3, we showed that if the sufficient statistics are bounded by B and if $\mathcal{F}^\star$ has minimum eigenvalue $\lambda_{\min}$, then we can choose $\alpha = B/\lambda_{\min}$. However, when $\theta^\star$ is sparse we see that in both Thms. 4.2 and 4.5, we only care about $\alpha^\star$, the analytic standardized moment/cumulant of the set $\mathcal{V}$, specified in (1). Given this, it is clear from the exposition in the GLM example in Section 3.2.3 that $\alpha^\star$ can be bounded by $B/\kappa^\star_{\min}$, since all elements of the set $\mathcal{V}$ satisfy Assumption 4.1.

5.0.4 Proof of Theorem 4.5

Lemma 5.5. *(Sparsity or Restricted Set) If* $\tau = \frac{18\lambda}{{\kappa^\star_{\min}}^2}$, *then the size of the support of any* $\theta \in \Theta_\tau$ *is at most* $2s$

Proof. First notice that on the set S thresholding could potentially leave all the s coordinates. On the other hand notice that if we threshold using τ, then the number of coordinates that remain unclipped in the set S^C is bounded by $\|\hat{\theta}_{S^C}\|_1/\tau$. Hence $|\{i : |\hat{\theta}_i| > \tau\}| \leq s + \|\hat{\theta}_{S^C}\|_1/\tau$. By (9), (12) and the RE assumption, we have $\|\hat{\theta}_{S^C}\|_1 \leq 3\|\hat{\theta}_S - \theta^\star\|_1 \leq 3\sqrt{s}\|\hat{\theta}_S - \theta^\star\|_2 \leq \frac{18\lambda s}{{\kappa^\star_{\min}}^2}$. Using this we see that $|\{i : |\hat{\theta}_i| > \tau\}| \leq s + \frac{18\lambda s}{{\kappa^\star_{\min}}^2\tau}$. Plug in the value of τ to finish. □

Lemma 5.6. *(Bias) Choose* $\tau = 18\lambda/{\kappa^\star_{\min}}^2$. *Then,*

$$\mathcal{L}(\hat{\theta}^\tau_S) - \mathcal{L}(\theta^\star) \leq \frac{540{\kappa^\star_{\max}}^2 s\lambda^2}{{\kappa^\star_{\min}}^4} \ ,$$

where $\hat{\theta}^\tau$ *is defined as* $\hat{\theta}^\tau_i = \hat{\theta}_i \mathbf{1}_{(\hat{\theta}_i > \tau)}$.

Proof. Note that

$$\begin{aligned}
\|\hat{\theta}^\tau_S - \theta^\star\|^2_{\mathcal{F}^\star} &\leq {\kappa^\star_{\max}}^2 \|\hat{\theta}^\tau_S - \theta^\star\|_2^2 \\
&\leq 2{\kappa^\star_{\max}}^2 \left(\|\hat{\theta}^\tau_S - \hat{\theta}_S\|_2^2 + \|\hat{\theta}_S - \theta^\star\|_2^2 \right) \\
&\leq 2{\kappa^\star_{\max}}^2 \left(s\tau^2 + \|\hat{\theta}_S - \theta^\star\|_2^2 \right) \\
&\leq 2{\kappa^\star_{\max}}^2 \left(s\tau^2 + \tfrac{36 s\lambda^2}{{\kappa^\star_{\min}}^4} \right)
\end{aligned}$$

where the last step is obtained by applying Thm. 4.2. Substituting for τ,

$$\|\hat{\theta}^\tau_S - \theta^\star\|^2_{\mathcal{F}^\star} \leq \frac{720{\kappa^\star_{\max}}^2 s\lambda^2}{{\kappa^\star_{\min}}^4} \ . \tag{13}$$

Now the condition on λ in (5) implies that Thm. 3.4 is applicable, which completes the proof. □

Proof of Theorem 4.5. The first claim of the theorem follows from Lemma 5.5. We prove the second claim of the theorem by considering two cases. First, when $\mathcal{L}(\tilde{\theta}) \leq \mathcal{L}(\hat{\theta}^\tau_S)$. In this case, by Lemma 5.6, $\mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta^\star) \leq 540{\kappa^\star_{\max}}^2 s\lambda^2/{\kappa^\star_{\min}}^4$ Also by (5), applying Thm. 3.4, $\frac{1}{4}\|\tilde{\theta} - \theta^\star\|^2_{\mathcal{F}^\star} \leq \mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta^\star) \leq 540{\kappa^\star_{\max}}^2 s\lambda^2/{\kappa^\star_{\min}}^4$ which gives us the second claim of the theorem. In the second case, when $\mathcal{L}(\tilde{\theta}) > \mathcal{L}(\hat{\theta}^\tau_S)$, by applying Lemma 5.3 with $\theta = \hat{\theta}^\tau_S$, we see that

$$\begin{aligned}
\mathcal{L}(\tilde{\theta}) - \mathcal{L}(\hat{\theta}^\tau_S) \leq \tfrac{3\lambda}{2}\|\hat{\theta}^\tau_S\|_1 &\leq \tfrac{3\lambda}{2}\|\theta^\star - \hat{\theta}^\tau_S\|_1 + \tfrac{3\lambda}{2}\|\theta^\star\|_1 \\
&\leq \tfrac{3\lambda\sqrt{s}}{2}\|\theta^\star - \hat{\theta}^\tau_S\|_2 + \tfrac{3\lambda}{2}\|\theta^\star\|_1 \\
&\leq \tfrac{3\lambda\sqrt{s}}{2\kappa^\star_{\min}}\|\theta^\star - \hat{\theta}^\tau_S\|_{\mathcal{F}^\star} + \tfrac{3\lambda}{2}\|\theta^\star\|_1 \\
&\leq \tfrac{18\sqrt{5}\lambda^2 s\kappa^\star_{\max}}{{\kappa^\star_{\min}}^3} + \tfrac{3\lambda}{2}\|\theta^\star\|_1
\end{aligned}$$

where the last step is using (13). Hence we see that

$$\begin{aligned}
\mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta^\star) &\leq \mathcal{L}(\tilde{\theta}) - \mathcal{L}(\hat{\theta}^\tau_S) + \mathcal{L}(\hat{\theta}^\tau_S) - \mathcal{L}(\theta^\star) \\
&\leq \tfrac{581{\kappa^\star_{\max}}^2 s\lambda^2}{{\kappa^\star_{\min}}^4} + \tfrac{3\lambda}{2}\|\theta^\star\|_1
\end{aligned}$$

Thus, by condition (5) on λ, the pre-condition of the Thm. 3.4 is satisfied and hence,

$$\begin{aligned}
&\tfrac{1}{4}\|\tilde{\theta} - \theta^\star\|^2_{\mathcal{F}^\star} \leq \mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta^\star) && (14)\\
&\leq \mathcal{L}(\tilde{\theta}) - \mathcal{L}(\hat{\theta}^\tau_S) + \mathcal{L}(\hat{\theta}^\tau_S) - \mathcal{L}(\theta^\star) \\
&\leq \mathcal{L}(\tilde{\theta}) - \mathcal{L}(\hat{\theta}^\tau_S) + \tfrac{540{\kappa^\star_{\max}}^2 s\lambda^2}{{\kappa^\star_{\min}}^4} \\
&\leq \tfrac{6\lambda\sqrt{s}}{\kappa^\star_{\min}}\|\tilde{\theta} - \hat{\theta}^\tau_S\|_{\mathcal{F}^\star} + \tfrac{540{\kappa^\star_{\max}}^2 s\lambda^2}{{\kappa^\star_{\min}}^4} && (15)\\
&\leq \tfrac{6\lambda\sqrt{s}}{\kappa^\star_{\min}}\|\tilde{\theta} - \theta^\star\|_{\mathcal{F}^\star} + \tfrac{6\lambda\sqrt{s}}{\kappa^\star_{\min}}\|\theta^\star - \hat{\theta}^\tau_S\|_{\mathcal{F}^\star} + \tfrac{540{\kappa^\star_{\max}}^2 s\lambda^2}{{\kappa^\star_{\min}}^4} \\
&\leq \tfrac{6\lambda\sqrt{s}}{\kappa^\star_{\min}}\|\tilde{\theta} - \theta^\star\|_{\mathcal{F}^\star} + \tfrac{161\kappa^\star_{\max} s\lambda^2}{{\kappa^\star_{\min}}^2} + \tfrac{540{\kappa^\star_{\max}}^2 s\lambda^2}{{\kappa^\star_{\min}}^4} \ . && (16)
\end{aligned}$$

Here (15) is obtained by applying Lemmas 5.3 and 5.4 with $\Theta = \Theta_\tau$ and then using Assumption 4.1, and (16) is due to (13). Simplifying we conclude that

$$\begin{aligned}
\tfrac{1}{4}\|\tilde{\theta} - \theta^\star\|^2_{\mathcal{F}^\star} &\leq \mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta^\star) \\
&\leq \tfrac{6\lambda\sqrt{s}}{\kappa^\star_{\min}}\|\tilde{\theta} - \theta^\star\|_{\mathcal{F}^\star} + \tfrac{701{\kappa^\star_{\max}}^2 s\lambda^2}{{\kappa^\star_{\min}}^4} && (17)
\end{aligned}$$

From this, it can be shown that $\|\tilde{\theta} - \theta^\star\|_{\mathcal{F}^\star} \leq \frac{24\lambda\sqrt{s}}{\kappa^\star_{\min}} + \frac{75\kappa^\star_{\max}\lambda\sqrt{s}}{{\kappa^\star_{\min}}^2}$. Using this in (17)

$$\mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta^\star) \leq \tfrac{144\lambda^2 s}{{\kappa^\star_{\min}}^2} + \tfrac{450\kappa^\star_{\max}\lambda^2 s}{{\kappa^\star_{\min}}^3} + \tfrac{701{\kappa^\star_{\max}}^2 s\lambda^2}{{\kappa^\star_{\min}}^4}$$

Simplifying gives 2nd claim of the theorem for the 2nd case. □