A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers

Sahand Negahban[1] Pradeep Ravikumar[2]
Martin J. Wainwright[1,3] Bin Yu[1,3]

Department of EECS[1] Department of CS[2] Department of Statistics[3]
UC Berkeley UT Austin UC Berkeley

October 2010

Abstract

High-dimensional statistical inference deals with models in which the the number of parameters p is comparable to or larger than the sample size n. Since it is usually impossible to obtain consistent procedures unless $p/n \rightarrow 0$, a line of recent work has studied models with various types of low-dimensional structure (e.g., sparse vectors; block-structured matrices; low-rank matrices; Markov assumptions). In such settings, a general approach to estimation is to solve a regularized convex program (known as a regularized M-estimator) which combines a loss function (measuring how well the model fits the data) with some regularization function that encourages the assumed structure. This paper provides a unified framework for establishing consistency and convergence rates for such regularized M-estimators under high-dimensional scaling. We state one main theorem and show how it can be used to re-derive some existing results, and also to obtain a number of new results on consistency and convergence rates. Our analysis also identifies two key properties of loss and regularization functions, referred to as restricted strong convexity and decomposability, that ensure corresponding regularized M-estimators have fast convergence rates, and which are optimal in many well-studied cases.

1 Introduction

High-dimensional statistics is concerned with models in which the ambient dimension of the problem p may be of the same order as—or substantially larger than—the sample size n. While its roots are quite old, dating back to classical random matrix theory and high-dimensional testing problems (e.g, [26, 46, 58, 82]), the past decade has witnessed a tremendous surge of research activity. Rapid development of data collection technology is a major driving force: it allows for more observations to be collected (larger n), and also for more variables to be measured (larger p). Examples are ubiquitous throughout science: astronomical projects such as the Large Synoptic Survey Telescope [1] produce terabytes of data in a single evening; each sample is a high-resolution image, with several hundred megapixels, so that $p \gg 10^8$. Financial data is also of a high-dimensional nature, with hundreds or thousands of financial instruments being measured and tracked over time, often at very fine time intervals for use in high frequency trading. Advances in biotechnology now allow for measurements of thousands of genes or proteins, and lead to numerous statistical challenges (e.g., see the paper [6] and references therein). Various types of imaging technology, among them magnetic resonance imaging in medicine [44] and hyper-spectral imaging in ecology [38], also lead to high-dimensional data sets.

In the regime $p \gg n$, it is well known that consistent estimators cannot be obtained unless additional constraints are imposed on the model. Accordingly, there are now several lines of work within high-dimensional statistics, all of which are based on imposing some type of low-dimensional constraint on the model space, and then studying the behavior of different estimators. Examples include linear regression with sparsity constraints, estimation of covariance or inverse covariance matrices, graphical model selection, sparse principal component analysis, low-rank matrix estimation, and sparse additive non-parametric models. The classical technique of regularization has proven fruitful in all of these contexts. Many well known estimators are based on solving a convex optimization problem formed by the sum of a loss function with a weighted regularizer; we refer to any such method as a *regularized M-estimator*. For instance, in application to linear models, the Lasso or basis pursuit approach [72, 22] is based on a combination of the least-squares loss with ℓ_1-regularization, and so involves solving a quadratic program. Similar approaches have been applied to generalized linear models, resulting in more general (non-quadratic) convex programs with ℓ_1-constraints. Several types of regularization have been used for estimating matrices, including standard ℓ_1-regularization, a wide range of sparse group-structured regularizers, as well as regularization based on the nuclear norm (sum of singular values).

Past work: Within the framework of high-dimensional statistics, the goal is to obtain bounds on the error—meaning the difference between the estimate and some nominal parameter setting—that hold with high probability for a finite sample size, and also allow the ambient dimension p, as well as other structural parameters (e.g., sparsity of a vector, degree of a graph, rank of matrix), to grow as a function of the sample size n. By now, for various types of M-estimators based on convex regularization, there are a large number of theoretical results in place that hold under high-dimensional scaling. It must be emphasized that our referencing is necessarily incomplete given the extraordinary number of papers that have appeared in recent years. Sparse linear regression has perhaps been the most active area, and multiple bodies of work can be differentiated by the error metric under consideration. They include work on exact recovery for noiseless observations (e.g., [24, 23, 17]), prediction error consistency (e.g., [27, 14, 76, 77]), consistency of the parameter estimates in ℓ_2 or some other norm (e.g., [14, 13, 77, 85, 50, 8, 18]), as well as variable selection consistency [49, 81, 87]. The information-theoretic limits of sparse linear regression are also well-understood, and ℓ_1-based methods are known to be optimal for ℓ_q-ball sparsity [59], and near-optimal for model selection [80]. For generalized linear models (GLMs), estimators based on ℓ_1-regularized maximum likelihood have also been studied, including results on risk consistency [78], consistency in ℓ_2 or ℓ_1-norm [4, 31, 47], and model selection consistency [63, 11]. Sparsity has also proven useful in application to different types of matrix estimation problems, among them sparse covariance matrices [7, 15, 32]. Another line of work has studied the problem of estimating Gaussian Markov random fields, or equivalently inverse covariance matrices with sparsity constraints. Here there are a range of results, including convergence rates in Frobenius, operator and other matrix norms [69, 64, 37, 89], as well as results on model selection consistency [64, 37, 49]. Motivated by applications in which sparsity arises in a structured manner, other researchers have proposed different types of block-structured regularizers (e.g., [74, 75, 86, 84, 2, 5, 29]), among them the group Lasso based on ℓ_1/ℓ_2 regularization. High-dimensional consistency results have been obtained for exact recovery based on noiseless observations [71, 5], convergence rates in ℓ_2-norm (e.g., [51, 28, 43, 5]) as well as model selection consistency (e.g., [57, 53, 51]). Problems of low-rank matrix estimation also arise in numerous applications. Techniques based on thresholding as well as nuclear norm regularization have been studied for different statistical models, including compressed sensing [66, 40, 67], matrix completion [19, 33, 34, 65, 52], multitask regression [83, 54, 68, 12, 3], and system identification [25, 54, 42]. Finally, although the primary emphasis of this paper is on high-dimensional parametric models, regularization methods have also proven effective for a class of high-dimensional non-parametric models that have a sparse additive decomposition (e.g., [62, 48, 35, 36]). Again, regularization methods have been been shown to achieve minimax-optimal rates in this setting [60].

Our contributions: As we have noted previously, almost all of these estimators can be seen as particular types of regularized M-estimators, with the choice of loss function, regularizer and statistical assumptions changing according to the model. This methodological similarity suggests an intriguing possibility: is there a *common set of theoretical principles* that underlies analysis of all these estimators? If so, it could be possible to gain a unified understanding of a large collection of techniques for high-dimensional estimation.

The main contribution of this paper is to provide an affirmative answer to this question. In particular, we isolate and highlight two key properties of a regularized M-estimator—namely, a *decomposability property* for the regularizer, and a notion of *restricted strong convexity* that depends on the interaction between the regularizer and the loss function. For loss functions and regularizers satisfying these two conditions, we prove a general result (Theorem 1) about consistency and convergence rates for the associated estimates. This result provides a family of bounds indexed by subspaces, and each bound consists of the sum of approximation error and estimation error. This general result, when specialized to different statistical models, yields in a direct manner a large number of corollaries, some of them known and others novel. In addition to our framework and main result (Theorem 1), other new results include minimax-optimal rates for sparse regression over ℓ_q-balls (Corollary 3), a general oracle-type result for group-structured norms with applications to generalized ℓ_q-ball sparsity (Corollary 4), as well as bounds for generalized linear models under minimal assumptions (Corollary 5), allowing for both unbounded and non-Lipschitz functions. In concurrent work, a subset of the current authors have also used this framework to prove several results on low-rank matrix estimation using the nuclear norm [54], as well as optimal rates for noisy matrix completion [52]. Finally, en route to establishing these corollaries, we also prove some new technical results that are of independent interest, including guarantees of restricted strong convexity for group-structured regularization (Proposition 1) and for generalized linear models (Proposition 2).

The remainder of this paper is organized as follows. We begin in Section 2 by formulating the class of regularized M-estimators that we consider, and then defining the notions of decomposability and restricted strong convexity. Section 3 is devoted to the statement of our main result (Theorem 1), and discussion of its consequences. Subsequent sections are devoted to corollaries of this main result for different statistical models, including sparse linear regression (Section 4), estimators based on group-structured regularizers (Section 5), estimation in generalized linear models (Section 6), and low-rank matrix estimation (Section 7).

2 Problem formulation and some key properties

In this section, we begin with a precise formulation of the problem, and then develop some key properties of the regularizer and loss function.

2.1 A family of M-estimators

Let $Z_1^n := \{Z_1, \ldots, Z_n\}$ denote n observations drawn i.i.d. according to some distribution $\mathbb{P}$, and suppose that we are interested in estimating some parameter θ of the distribution $\mathbb{P}$. Let $\mathcal{L} : \mathbb{R}^p \times \mathcal{Z}^n \rightarrow \mathbb{R}$ be some loss function that, for a given set of observations Z_1^n, assigns a cost $\mathcal{L}(\theta; Z_1^n)$ to any parameter $\theta \in \mathbb{R}^p$. We assume that that the population risk $R(\theta) = \mathbb{E}_{Z_1^n}[\mathcal{L}(\theta; Z_1^n)]$ is independent of n, and we let $\theta^* \in \arg\min\limits_{\theta \in \mathbb{R}^p} R(\theta)$ be a minimizer of the population risk. As is standard in statistics, in order to estimate the parameter vector θ^* from the data Z_1^n, we solve a convex program that combines the loss function with a regularizer. More specifically, let $r : \mathbb{R}^p \rightarrow \mathbb{R}$ denote a regularization function, and consider the regularized M-estimator given by

$$\widehat{\theta}_{\lambda_n} \quad \in \quad \arg\min_{\theta \in \mathbb{R}^p} \big\{ \mathcal{L}(\theta; Z_1^n) + \lambda_n r(\theta) \big\}, \tag{1}$$

where $\lambda_n > 0$ is a user-defined regularization penalty. Throughout the paper, we assume that the loss function $\mathcal{L}$ is convex and differentiable, and that the regularization function r is a norm.

Our goal is to provide general techniques for deriving bounds on the difference between any solution $\widehat{\theta}_{\lambda_n}$ to the convex program (1) and the unknown vector θ^*. In this paper, we derive bounds on the norm $\|\widehat{\theta}_{\lambda_n} - \theta^*\|_\star$, where the error norm is induced by an inner product $\langle \cdot, \cdot \rangle_\star$ on $\mathbb{R}^p$. Most often, this error norm will either be the Euclidean ℓ_2-norm on vectors, or the analogous Frobenius norm for matrices, but our theory also applies to certain types of weighted norms. In addition, we provide bounds on the quantity $r(\widehat{\theta}_{\lambda_n} - \theta^*)$, which measures the error in the regularizer norm. In the classical setting, the ambient dimension p stays fixed while the number of observations n tends to infinity. Under these conditions, there are standard techniques for proving consistency and asymptotic normality for the error $\widehat{\theta}_{\lambda_n} - \theta^*$. In contrast, the analysis of this paper is all within a high-dimensional framework, in which the tuple (n, p), as well as other problem parameters, such as vector sparsity or matrix rank etc., are all allowed to tend to infinity. In contrast to asymptotic normality, our goal is to obtain explicit finite sample error bounds that hold with high probability.

Throughout the paper, we make use of the following notation. For any given subspace $A \subseteq \mathbb{R}^p$, we let $A^\perp$ be its orthogonal complement, as defined by the inner product $\langle \cdot, \cdot \rangle_\star$—namely, the set $A^\perp := \big\{ v \in \mathbb{R}^p \mid \langle v, u \rangle_\star = 0 \quad \text{for all } u \in A \big\}$. In addition, we let $\Pi_A : \mathbb{R}^p \rightarrow A$ denote the projection operator, defined by

$$\Pi_A(u) := \arg\min_{v \in A} \|u - v\|_\star \,, \tag{2}$$

with the projection operator $\Pi_{A^\perp}$ defined in an analogous manner.

2.2 Decomposability

We now motivate and define the decomposability property of regularizers, and then show how it constrains the error in the M-estimation problem (1). This property is defined in terms of a pair of subspaces $A \subseteq B$ of $\mathbb{R}^p$. The role of the *model subspace* A is to capture the constraints specified by the model; for instance, it might be the subspace of vectors with a particular support (see Example 1), or a subspace of low-rank matrices (see Example 3). The orthogonal complement $B^\perp$ is the *perturbation subspace*, representing deviations away from the model subspace A. In the ideal case, we have $B^\perp = A^\perp$, but our definition allows for the possibility that B is strictly larger than A.

Definition 1. A norm-based regularizer r is ***decomposable*** with respect to the subspace pair $A \subseteq B$ if

$$r(\alpha + \beta) = r(\alpha) + r(\beta) \quad \text{for all } \alpha \in A \text{ and } \beta \in B^{\perp}. \tag{3}$$

In order to build some intuition, let us consider the ideal case $A = B$ for the time being, so that the decomposition (3) holds for all pairs $(\alpha, \beta) \in A \times A^{\perp}$. For any given pair (α, β) of this form, the vector $\alpha + \beta$ can be interpreted as perturbation of the model vector α away from the subspace A, and it is desirable that the regularizer penalize such deviations as much as possible. By the triangle inequality, we always have $r(\alpha + \beta) \leq r(\alpha) + r(\beta)$, so that the decomposability condition (3) holds if and only the triangle inequality is tight for all pairs $(\alpha, \beta) \in (A, A^{\perp})$. It is exactly in this setting that the regularizer penalizes deviations away from the model subspace A as much as possible.

In general, it is not difficult to find subspace pairs that satisfy the decomposability property. As a trivial example, any regularizer is decomposable with respect to $A = \mathbb{R}^p$ and its orthogonal complement $A^{\perp} = \{0\}$. As will be clear in our main theorem, it is of more interest to find subspace pairs in which the model subspace A is "small", so that the orthogonal complement $A^{\perp}$ is "large". Recalling the projection operator (2), of interest to us is its action on the unknown parameter $\theta^* \in \mathbb{R}^p$. In the most desirable setting, the model subspace A can be chosen such that $\Pi_A(\theta^*) \approx \theta^*$ and $\Pi_{A^{\perp}}(\theta^*) \approx 0$. If this can be achieved with the model subspace A remaining relatively small, then our main theorem guarantees that it is possible to estimate θ^* at a relatively fast rate. The following examples illustrate suitable choices of the spaces A and B in three concrete settings, beginning with the case of sparse vectors.

Example 1. *Sparse vectors and ℓ_1-norm regularization.* Suppose the error norm $\|\cdot\|_{\star}$ is the usual ℓ_2-norm, and that the model class of interest is the set of s-sparse vectors in p dimensions. For any particular subset $S \subseteq \{1, 2, \ldots, p\}$ with cardinality s, we define the model subspace

$$A(S) := \big\{\alpha \in \mathbb{R}^p \mid \alpha_j = 0 \quad \text{for all } j \notin S\big\}. \tag{4}$$

Here our notation reflects the fact that A depends explicitly on the chosen subset S. By construction, we have $\Pi_{A(S)}(\theta^*) = \theta^*$ for any vector that is supported on S.

Now define $B(S) = A(S)$, and note that the orthogonal complement with respect to the Euclidean inner product is given by

$$B^{\perp}(S) = A^{\perp}(S) = \big\{\beta \in \mathbb{R}^p \mid \beta_j = 0 \quad \text{for all } j \in S\big\}. \tag{5}$$

This set corresponds to the perturbation subspace, capturing deviations away from the set of vectors with support S. We claim that for any subset S, the ℓ_1-norm $r(\alpha) = \|\alpha\|_1$ is decomposable with respect to the pair $(A(S), B(S))$. Indeed, by construction of the subspaces, any $\alpha \in A(S)$ can be written in the partitioned form $\alpha = (\alpha_S, 0_{S^c})$, where $\alpha_S \in \mathbb{R}^s$ and $0_{S^c} \in \mathbb{R}^{p-s}$ is a vector of zeros. Similarly, any vector $\beta \in B^{\perp}(S)$ has the partitioned representation $(0_S, \beta_{S^c})$. Putting together the pieces, we obtain

$$\|\alpha + \beta\|_1 = \|(\alpha_S, 0) + (0, \beta_{S^c})\|_1 \; = \; \|\alpha\|_1 + \|\beta\|_1,$$

showing that the ℓ_1-norm is decomposable as claimed. ♢

As a follow-up to the previous example, it is also worth noting that the same argument shows that for a strictly positive weight vector ω, the *weighted* ℓ_1*-norm* $\|\alpha\|_\omega := \sum_{j=1}^p \omega_j |\alpha_j|$ is also decomposable with respect to the pair $(A(S), B(S))$. For another natural extension, we now turn to the case of sparsity models with more structure.

Example 2. *Group-structured norms.* In many applications, sparsity arises in a more structured fashion, with groups of coefficients likely to be zero (or non-zero) simultaneously. In order to model this behavior, suppose that the index set $\{1, 2, \ldots, p\}$ can be partitioned into a set of T disjoint groups, say $\mathcal{G} = \{G_1, G_2, \ldots, G_T\}$. With this set-up, for a given vector $\vec{\nu} = (\nu_1, \ldots, \nu_T) \in [1, \infty]^T$, the associated $(1, \vec{\nu})$*-group norm* takes the form

$$\|\alpha\|_{\mathcal{G},\vec{\nu}} \;:=\; \sum_{t=1}^{T} \|\alpha_{G_t}\|_{\nu_t}. \tag{6}$$

For instance, with the choice $\vec{\nu} = (2, 2, \ldots, 2)$, we obtain the group ℓ_1/ℓ_2-norm, corresponding to the regularizer that underlies the group Lasso [84]. The choice $\vec{\nu} = (\infty, \ldots, \infty)$ has also been studied in past work [75, 53].

We claim that the norm $\|\cdot\|_{\mathcal{G},\vec{\nu}}$ is again decomposable with respect to appropriately defined subspaces. Indeed, given any subset $S_\mathcal{G} \subseteq \{1, \ldots, T\}$ of group indices, say with cardinality $s_\mathcal{G} = |S_\mathcal{G}|$, we can define the subspace

$$A(S_\mathcal{G}) := \big\{\alpha \in \mathbb{R}^p \;\mid\; \alpha_{G_t} = 0 \quad \text{for all } t \notin S_\mathcal{G}\big\}, \tag{7}$$

as well as its orthogonal complement with respect to the usual Euclidean inner product

$$A^\perp(S_\mathcal{G}) \;=\; B^\perp(S_\mathcal{G}) := \big\{\alpha \in \mathbb{R}^p \;\mid\; \alpha_{G_t} = 0 \quad \text{for all } t \in S_\mathcal{G}\big\}. \tag{8}$$

With these definitions, for any pair of vectors $\alpha \in A(S_\mathcal{G})$ and $\beta \in B^\perp(S_\mathcal{G})$, we have

$$\|\alpha + \beta\|_{\mathcal{G},\vec{\nu}} = \sum_{t \in S_\mathcal{G}} \|\alpha_{G_t} + 0_{G_t}\|_{\nu_t} + \sum_{t \in S^c} \|0_{G_j} + \beta_{G_t}\|_{\nu_t} \;=\; \|\alpha\|_{\mathcal{G},\vec{\nu}} + \|\beta\|_{\mathcal{G},\vec{\nu}}, \tag{9}$$

thus verifying the decomposability condition. $\diamondsuit$

Example 3. *Low-rank matrices and nuclear norm.* Now suppose that each parameter $\Theta \in \mathbb{R}^{p_1 \times p_2}$ is a matrix; this corresponds to an instance of our general set-up with $p = p_1 p_2$, as long as we identify the space $\mathbb{R}^{p_1 \times p_2}$ with $\mathbb{R}^{p_1 p_2}$ in the usual way. We equip this space with the inner product $\langle\!\langle \Theta,\, \Gamma \rangle\!\rangle := \operatorname{trace}(\Theta \Gamma^T)$, a choice which yields the *Frobenius norm*

$$\|\Theta\|_\star := \sqrt{\langle\!\langle \Theta,\, \Theta \rangle\!\rangle} \;=\; \sqrt{\sum_{j=1}^{p_1} \sum_{k=1}^{p_2} \Theta_{jk}^2} \tag{10}$$

as the induced norm. In many settings, it is natural to consider estimating matrices that are low-rank; examples include principal component analysis, spectral clustering, collaborative filtering, and matrix completion. In particular, consider the class of matrices $\Theta \in \mathbb{R}^{p_1 \times p_2}$ that have rank $r \leq \min\{p_1, p_2\}$. For any given matrix Θ, we let $\text{row}(\Theta) \subseteq \mathbb{R}^{p_2}$ and $\text{col}(\Theta) \subseteq \mathbb{R}^{p_1}$ denote its row space and column space respectively. Let U and V be a given pair of r-dimensional subspaces $U \subseteq \mathbb{R}^{p_1}$ and $V \subseteq \mathbb{R}^{p_2}$; these subspaces will represent left and right singular vectors of the target matrix Θ^* to be estimated. For a given pair (U, V), we can define the subspaces $A(U, V)$ and $B^{\perp}(U, V)$ of $\mathbb{R}^{p_1 \times p_2}$ given by

$$A(U,V) := \big\{\Theta \in \mathbb{R}^{p_1 \times p_2} \mid \text{row}(\Theta) \subseteq V, \;\text{col}(\Theta) \subseteq U\big\}, \quad \text{and} \tag{11a}$$

$$B^{\perp}(U,V) := \big\{\Theta \in \mathbb{R}^{p_1 \times p_2} \mid \text{row}(\Theta) \subseteq V^{\perp}, \;\text{col}(\Theta) \subseteq U^{\perp}\big\}. \tag{11b}$$

Unlike the preceding examples, in this case, we have $A(U, V) \neq B(U, V)$. However, as is required by our theory, we do have the inclusion $A(U, V) \subseteq B(U, V)$. Indeed, given any $\Theta \in A(U, V)$ and $\Gamma \in B^{\perp}(U, V)$, we have $\Theta^T \Gamma = 0$ by definition, and hence $\langle\langle \Theta, \;\Gamma \rangle\rangle = \text{trace}(\Theta^T \Gamma) = 0$. Consequently, we have shown that Θ is orthogonal to the space $B^{\perp}(U, V)$, implying the claimed inclusion.

Finally, we claim that the *nuclear or trace norm* $r(\Theta) = |||\Theta|||_1$, corresponding to the sum of the singular values of the matrix Θ, satisfies the decomposability property with respect to the pair $(A(U, V), B^{\perp}(U, V))$. By construction, any pair of matrices $\Theta \in A(U, V)$ and $\Gamma \in B^{\perp}(U, V)$ have orthogonal row and column spaces, which implies the required decomposability condition—namely $|||\Theta + \Gamma|||_1 = |||\Theta|||_1 + |||\Gamma|||_1$. ♢

A line of recent work (e.g., [21, 20]) has studied matrix problems involving the sum of a low-rank matrix with a sparse matrix, along with the regularizer formed by a weighted sum of the nuclear norm and the elementwise ℓ_1-norm. By a combination of Examples 1 and Example 3, this regularizer also satisfies the decomposability property with respect to appropriately defined subspaces.

2.3 A key consequence of decomposability

Thus far, we have specified a class (1) of M-estimators based on regularization, defined the notion of decomposability for the regularizer and worked through several illustrative examples. We now turn to the statistical consequences of decomposability—more specifically, its implications for the error vector $\widehat{\Delta}_{\lambda_n} = \widehat{\theta}_{\lambda_n} - \theta^*$, where $\widehat{\theta} \in \mathbb{R}^p$ is any solution of the regularized M-estimation procedure (1). Letting $\langle \cdot, \cdot \rangle$ denote the usual Euclidean inner product, the dual norm of r is given by

$$r^*(v) := \sup_{u \in \mathbb{R}^p \backslash \{0\}} \frac{\langle u, v \rangle}{r(u)}. \tag{12}$$

It plays a key role in our specification of the regularization parameter λ_n in the convex program (1), in particular linking it to the data Z_1^n that defines the estimator.

Lemma 1. *Suppose that $\mathcal{L}$ is a convex function, and consider any optimal solution $\widehat{\theta}$ to the optimization problem* (1) *with a strictly positive regularization parameter satisfying*

$$\lambda_n \geq 2\, r^*(\nabla \mathcal{L}(\theta^*; Z_1^n)). \tag{13}$$

Then for any pair (A, B) over which r is decomposable, the error $\widehat{\Delta} = \widehat{\theta}_{\lambda_n} - \theta^*$ belongs to the set

$$\mathbb{C}(A, B; \theta^*) := \big\{\Delta \in \mathbb{R}^p \mid r(\Pi_{B^{\perp}}(\Delta)) \leq 3 r(\Pi_B(\Delta)) + 4 r(\Pi_{A^{\perp}}(\theta^*))\big\}. \tag{14}$$

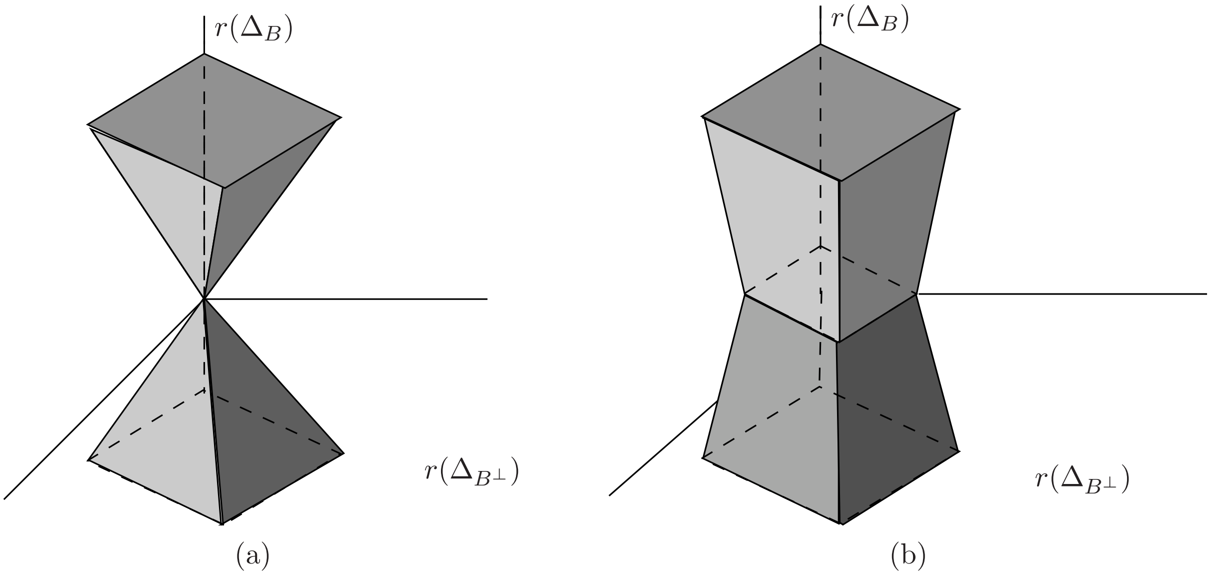

Figure 1. Illustration of the set $\mathbb{C}(A, B; \theta^*)$ in the special case $\Delta = (\Delta_1, \Delta_2, \Delta_3) \in \mathbb{R}^3$ and regularizer $r(\Delta) = \|\Delta\|_1$, relevant for sparse vectors (Example 1). This picture shows the case $S = \{3\}$, so that the model subspace is $A(S) = B(S) = \{\Delta \in \mathbb{R}^3 \mid \Delta_1 = \Delta_2 = 0\}$, and its orthogonal complement is given by $A^\perp(S) = B^\perp(S) = \{\Delta \in \mathbb{R}^3 \mid \Delta_3 = 0\}$. (a) In the special case when $\theta^* \in A(S)$ so that $r(\Pi_{A^\perp}(\theta^*)) = 0$, the set $\mathbb{C}(A, B; \theta^*)$ is a cone. (b) When θ^* does not belong to $A(S)$, the set $\mathbb{C}(A, B; \theta^*)$ is enlarged in the co-ordinates (Δ_1, Δ_2) that span $B^\perp(S)$. It is no longer a cone, but is still a star-shaped set.

We prove this result in Appendix A.1. It has the following important consequence: for any decomposable regularizer and an appropriate choice (13) of regularization parameter, we are guaranteed that the error vector $\widehat{\Delta}$ belongs to a very specific set, depending on the unknown vector θ^*. As illustrated in Figure 1, the geometry of the set $\mathbb{C}$ depends on the relation between θ^* and the model subspace A. When $\theta^* \in A$, then we are guaranteed that $r(\Pi_{A^\perp}(\theta^*)) = 0$. In this case, the constraint (14) reduces to $r(\Pi_{B^\perp}(\Delta)) \leq 3r(\Pi_B(\Delta))$, so that $\mathbb{C}$ is a cone, as illustrated in panel (a). In the more general case when $\theta^* \notin A$ and $r(\Pi_{A^\perp}(\theta^*)) \neq 0$, the set $\mathbb{C}$ is *not* a cone, but rather a star-shaped set excluding a small ball centered at the origin (panel (b)). As will be clarified in the sequel, this difference (between $\theta^* \in A$ and $\theta^* \notin A$) plays an important role in bounding the error.

2.4 Restricted strong convexity

We now turn to an important requirement of the loss function, and its interaction with the statistical model. Recall that $\widehat{\Delta} = \widehat{\theta}_{\lambda_n} - \theta^*$ is the difference between an optimal solution $\widehat{\theta}_{\lambda_n}$ and the true parameter, and consider the loss difference $\mathcal{L}(\widehat{\theta}_{\lambda_n}; Z_1^n) - \mathcal{L}(\theta^*; Z_1^n)$. To simplify notation, we frequently write $\mathcal{L}(\widehat{\theta}_{\lambda_n}) - \mathcal{L}(\theta^*)$ when the underlying data Z_1^n is clear from context. In the classical setting, under fairly mild conditions, one expects that that the loss difference should converge to zero as the sample size n increases. It is important to note, however, that such convergence on its own is *not sufficient* to guarantee that $\widehat{\theta}_{\lambda_n}$ and θ^* are close, or equivalently that $\widehat{\Delta}$ is small. Rather, the closeness depends on the curvature of the loss function, as illustrated in Figure 2. In a desirable setting (panel (a)), the loss function is sharply curved around its optimum $\widehat{\theta}_{\lambda_n}$, so that having a

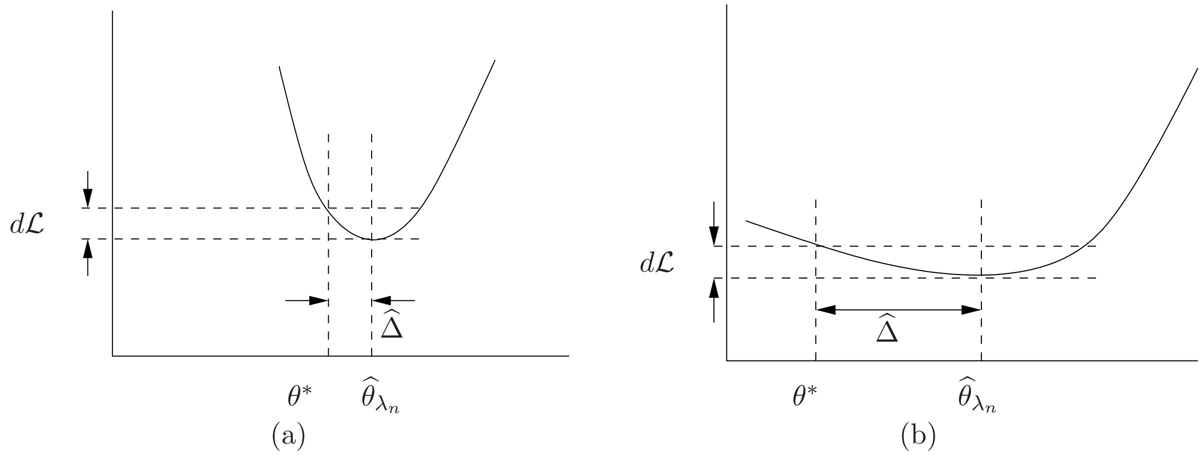

Figure 2. Role of curvature in distinguishing parameters. (a) Loss function has high curvature around $\widehat{\Delta}$. A small excess loss $d\mathcal{L} = |\mathcal{L}(\widehat{\theta}_{\lambda_n}) - \mathcal{L}(\theta^*)|$ guarantees that the parameter error $\widehat{\Delta} = \widehat{\theta}_{\lambda_n} - \theta^*$ is also small. (b) A less desirable setting, in which the loss function has relatively low curvature around the optimum.

small loss difference $|\mathcal{L}(\theta^*) - \mathcal{L}(\widehat{\theta}_{\lambda_n})|$ translates to a small error $\widehat{\Delta} = \widehat{\theta}_{\lambda_n} - \theta^*$. Panel (b) illustrates a less desirable setting, in which the loss function is relatively flat, so that the loss difference can be small while the error $\widehat{\Delta}$ is relatively large.

The standard way to ensure that a function is "not too flat" is via the notion of strong convexity—in particular, by requiring that there exist some constant $\gamma > 0$ such that

$$\delta\mathcal{L}(\Delta, \theta^*; Z_1^n) \;:=\; \mathcal{L}(\theta^* + \Delta; Z_1^n) - \mathcal{L}(\theta^*; Z_1^n) - \langle \nabla\mathcal{L}(\theta^*; Z_1^n), \Delta \rangle \geq \gamma \|\Delta\|_2^2 \tag{15}$$

for all $\Delta \in \mathbb{R}^p$ in a neighborhood of θ^*. When the loss function is twice differentiable, restricted strong convexity amounts to a uniform lower bound on the eigenvalues of the Hessian $\nabla^2\mathcal{L}$.

Under classical "fixed p, large n" scaling, the loss function will be strongly convex under mild conditions. For instance, suppose that population function $\mathcal{R}(\theta) := \mathbb{E}_{Z_1^n}[\mathcal{L}(\theta; Z_1^n)]$ is strongly convex, or equivalently, that the Hessian $\nabla^2\mathcal{R}(\theta)$ is strictly positive definite in a neighborhood of θ^*. As a concrete example, when the loss function $\mathcal{L}$ is defined based on negative log likelihood of a statistical model, then the Hessian $\nabla^2\mathcal{R}(\theta)$ corresponds to the Fisher information matrix, a quantity which arises naturally in asymptotic statistics. If the dimension p is fixed while the sample size n goes to infinity, standard arguments can be used to show that (under mild regularity conditions) the random Hessian $\nabla^2\mathcal{L}(\theta; Z_1^n)$ converges to $\nabla^2\mathcal{R}(\theta) = \mathbb{E}_{Z_1^n}[\nabla^2\mathcal{L}(\theta; Z_1^n)]$ uniformly for all θ in a neighborhood of θ^*.

Whenever the pair (n, p) both increase in such a way that $p > n$, the situation is drastically different: the $p \times p$ Hessian matrix $\nabla^2\mathcal{L}(\theta; Z_1^n)$ will always be rank-deficient. As a concrete example, consider linear regression based on samples $Z_i = (y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$, for $i = 1, 2, \ldots, n$. Using the least-squares loss $\mathcal{L}(\theta; Z_1^n) = \frac{1}{2n}\|y - X\theta\|_2^2$, the $p \times p$ Hessian matrix $\nabla^2\mathcal{L}(\theta; Z_1^n) = \frac{1}{n}X^T X$ has rank at most n, meaning that the loss cannot be strongly convex when $p > n$. In the high-dimensional setting, a typical loss function will take the form shown in Figure 3(a): while curved in certain directions, it will be completely flat in a large number of directions. Thus, it is impossible to guarantee global strong convexity.

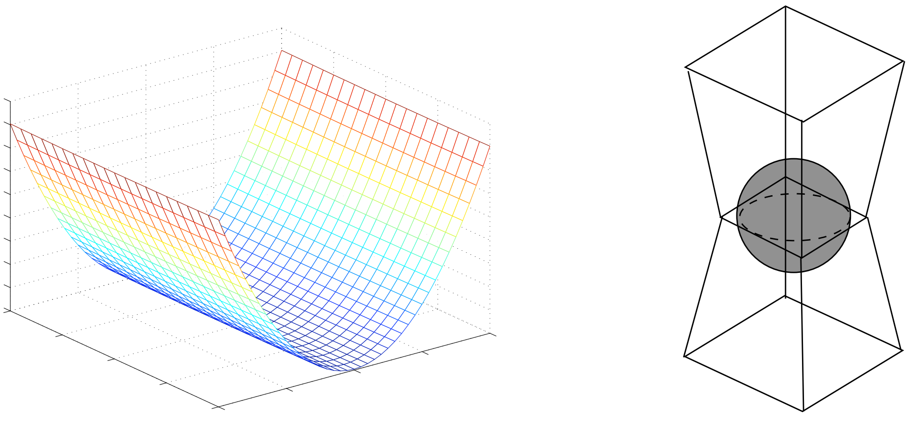
(a) (b)

Figure 3. (a) Illustration of a generic loss function in the high-dimensional $p > n$ setting: it is curved in certain directions, but completely flat in others. (b) Illustration of the set $\mathbb{K}(\delta; A, B, \theta^*) = \mathbb{C}(A, B; \theta^*) \cap \{\Delta \in \mathbb{R}^p \mid \|\Delta\|_\star = \delta\}$ over which restricted strong convexity must hold. It is the intersection of a ball of radius δ with the star-shaped set $\mathbb{C}(A, B, \theta^*)$.

In summary, we need to relax the notion of strong convexity by restricting the set of directions for which condition (15) holds. We formalize this idea as follows:

Definition 2. For a given set $\mathbb{S}$, the loss function satisfies ***restricted strong convexity*** (RSC) with parameter $\kappa_{\mathcal{L}} > 0$ if

$$\underbrace{\mathcal{L}(\theta^* + \Delta; Z_1^n) - \mathcal{L}(\theta^*; Z_1^n) - \langle \nabla \mathcal{L}(\theta^*; Z_1^n), \Delta \rangle}_{\delta \mathcal{L}(\Delta, \theta^*; Z_1^n)} \;\geq\; \kappa_{\mathcal{L}} \, \|\Delta\|_\star^2 \qquad \text{for all } \Delta \in \mathbb{S}. \tag{16}$$

It remains to specify suitable choices of sets $\mathbb{S}$ for this definition. Based on our discussion of regularizers and Lemma 1, one natural candidate is the set $\mathbb{C}(A, B; \theta^*)$. As will be seen, this choice is suitable only in the special case $\theta^* \in A$, so that $\mathbb{C}$ is a cone, as illustrated in Figure 1(a). In the more general setting ($\theta^* \notin A$), the set $\mathbb{C}$ is *not* a cone, but rather contains an $\| \cdot \|_\star$-ball of some positive radius centered at the origin. In this setting, even in the case of least-squares loss, it is impossible to certify RSC over the set $\mathbb{C}$, since it contains vectors in all possible directions.

For this reason, in order to obtain a generally applicable theory, it is essential to further restrict the set $\mathbb{C}$, in particular by introducing some tolerance $\delta > 0$, and defining the set

$$\mathbb{K}(\delta; A, B, \theta^*) := \mathbb{C}(A, B; \theta^*) \cap \big\{\theta \in \mathbb{R}^p \mid \|\Delta\|_\star = \delta\big\}. \tag{17}$$

As illustrated in Figure (3)(b), the set $\mathbb{K}(\delta)$ consists of the intersection of the sphere of radius δ in the error norm $\| \cdot \|_\star$ with the set $\mathbb{C}(A, B; \theta^*)$. As long as the tolerance parameter δ is suitably chosen, the additional sphere constraint, in conjunction with $\mathbb{C}$, excludes many directions in space. The necessity of this additional sphere case—essential for the inexact case $\theta^* \notin A$—does not appear to have been explicitly recognized in past work, even in the special case of sparse vector recovery.

3 Bounds for general M-estimators

We are now ready to state a general result that provides bounds and hence convergence rates for the error $\|\widehat{\theta}_{\lambda_n} - \theta^*\|_\star$, where $\widehat{\theta}_{\lambda_n}$ is any optimal solution of the convex program (1). Although it may appear somewhat abstract at first sight, we illustrate that this result has a number of concrete and useful consequences for specific models. In particular, we recover as an immediate corollary the best known results about estimation in sparse linear models with general designs [8, 50], as well as a number of new results, including minimax-optimal rates for estimation under ℓ_q-sparsity constraints, as well as results for sparse generalized linear models, estimation of block-structured sparse matrices and estimation of low-rank matrices.

Let us recall our running assumptions on the structure of the convex program (1).

(G1) The loss function $\mathcal{L}$ is convex and differentiable.

(G2) The regularizer is a norm, and is decomposable with respect to a subspace pair $A \subseteq B$.

The statement of our main result involves a quantity that relates the error norm and the regularizer:

Definition 3 (Subspace compatibility constant)**.** For any subspace B of $\mathbb{R}^p$, the ***subspace compatibility constant*** with respect to the pair $(r, \|\cdot\|_\star)$ is given by

$$\Psi(B) := \inf\big\{c > 0 \mid r(u) \leq c\,\|u\|_\star \quad \text{for all } u \in B\big\}. \tag{18}$$

This quantity reflects the degree of compatibility between the regularizer and the error norm over the subspace B. As a simple example, if B is a s-dimensional co-ordinate subspace, with regularizer $r(u) = \|u\|_1$ and error norm $\|u\|_\star = \|u\|_2$, then we have $\Psi(B) = \sqrt{s}$.

With this notation, we now come to the main result of this paper:

Theorem 1 (Bounds for general models)**.** *Under conditions (G1) and (G2), consider the convex program* (1) *based on a strictly positive regularization constant* $\lambda_n \geq 2r^*(\nabla\mathcal{L}(\theta^*; Z_1^n))$*. Define the critical tolerance*

$$\delta_n := \inf_{\delta > 0}\left\{\delta \;\mid\; \delta \geq \frac{2\lambda_n}{\kappa_{\mathcal{L}}}\Psi(B) + \sqrt{\frac{2\lambda_n r(\Pi_{A^\perp}(\theta^*))}{\kappa_{\mathcal{L}}}}, \quad \textit{and RSC holds over } \mathbb{K}(\delta; A, B, \theta^*). \right\}. \tag{19}$$

Then any optimal solution $\widehat{\theta}_{\lambda_n}$ *to the convex program* (1) *satisfies the bound* $\|\widehat{\theta}_{\lambda_n} - \theta^*\|_\star \leq \delta_n$*.*

Remarks: We can gain intuition by discussing in more detail some different features of this result.

(a) It should be noted that Theorem 1 is actually a *deterministic* statement about the set of optimizers of the convex program (1) for a fixed choice of λ_n. Since this program is not strictly convex in general, it may have multiple optimal solutions $\widehat{\theta}_{\lambda_n}$, and the stated bound holds for any of these optima. Probabilistic analysis is required when Theorem 1 is applied to particular statistical models, and we need to verify that the regularizer satisfies the condition

$$\lambda_n \geq 2r^*(\nabla\mathcal{L}(\theta^*; Z_1^n)), \tag{20}$$

and that the loss satisfies the RSC condition. A challenge here is that since θ^* is unknown, it is usually impossible to compute the quantity $r^*(\nabla \mathcal{L}(\theta^*; Z_1^n))$. Instead, when we derive consequences of Theorem 1 for different statistical models, we use large deviations techniques in order to provide bounds that hold with high probability over the data Z_1^n.

(b) Second, note that Theorem 1 actually provides a *family of bounds,* one for each pair (A, B) of subspaces for which the regularizer is decomposable. For any given pair, the error bound is the sum of two terms, corresponding to estimation error $\mathcal{E}_{\text{err}}$ and approximation error $\mathcal{E}_{\text{app}}$, given by (respectively)

$$\mathcal{E}_{\text{err}} := \frac{2\lambda_n}{\kappa_{\mathcal{L}}} \Psi(B), \quad \text{and} \quad \mathcal{E}_{\text{app}} := \sqrt{\frac{2\lambda_n r(\Pi_{A^\perp}(\theta^*))}{\kappa_{\mathcal{L}}}}. \tag{21}$$

As the dimension of the subspace A increases (so that the dimension of $A^\perp$ decreases), the approximation error tends to zero. But since $A \subseteq B$, the estimation error is increasing at the same time. Thus, in the usual way, optimal rates are obtained by choosing A and B so as to balance these two contributions to the error. We illustrate such choices for various specific models to follow.

A large body of past work on sparse linear regression has focused on the case of exactly sparse regression models for which the unknown regression vector θ^* is s-sparse. For this special case, recall from Example 1 in Section 2.2 that we can define an s-dimensional subspace A that contains θ^*. Consequently, the associated set $\mathbb{C}(A, B; \theta^*)$ is a cone (see Figure 1(a)), and it is thus possible to establish that restricted strong convexity (RSC) holds without any need for the additional tolerance parameter δ defining $\mathbb{K}$. This same reasoning applies to other statistical models, among them group-sparse regression, in which a small subset of groups are active, as well as low-rank matrix estimation. The following corollary provides a simply stated bound that covers all of these models:

Corollary 1. *If, in addition to the conditions of Theorem 1, the true parameter θ^* belongs to A and the RSC condition holds over $\mathbb{C}(A, B, \theta^*)$, then any optimal solution $\widehat{\theta}_{\lambda_n}$ to the convex program* (1) *satisfies the bounds*

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_\star \leq \frac{2\lambda_n}{\kappa_{\mathcal{L}}} \Psi(B), \qquad \textit{and} \tag{22a}$$

$$r(\widehat{\theta}_{\lambda_n} - \theta^*) \leq \frac{6\lambda_n}{\kappa_{\mathcal{L}}} \Psi^2(B). \tag{22b}$$

Focusing first on the bound (22a), it consists of three terms, each of which has a natural interpretation:

(a) The bound is inversely proportional to the RSC constant $\kappa_{\mathcal{L}}$, which measures the curvature of the loss function in a restricted set of directions.

(b) The bound is proportional the subspace compatibility constant $\Psi(B)$ from Definition 3, which measures the compatibility between the regularizer r and error norm $\| \cdot \|_\star$ over the subspace B.

(c) The bound also scales linearly with the regularization parameter λ_n, which must be strictly positive and satisfy the lower bound (20).

The bound (22b) on the error measured in the regularizer norm is similar, except that it scales quadratically with the subspace compatibility constant. As the proof clarifies, this additional dependence arises since the regularizer over the subspace B is larger than the norm $\| \cdot \|_\star$ by a factor of at most $\Psi(B)$ (see Definition 3).

Obtaining concrete rates using Corollary 1 requires some work in order to provide control on the three quantities in the bounds (22a) and (22b), as illustrated in the examples to follow.

4 Convergence rates for sparse linear regression

In order to illustrate the consequences of Theorem 1 and Corollary 1, let us begin with one of the simplest statistical models, namely the standard linear model. It is based on n observations $Z_i = (x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ of covariate-response pairs. Let $y \in \mathbb{R}^n$ denote a vector of the responses, and let $X \in \mathbb{R}^{n \times p}$ be the design matrix, where $x_i \in \mathbb{R}^p$ is the i^{th} row. This pair is linked via the linear model

$$y \;=\; X\theta^* + w, \tag{23}$$

where $\theta^* \in \mathbb{R}^p$ is the unknown regression vector, and $w \in \mathbb{R}^n$ is a noise vector. Given the observations $Z_1^n = (y, X) \in \mathbb{R}^n \times \mathbb{R}^{n \times p}$, our goal is to obtain a "good" estimate $\widehat{\theta}$ of the regression vector θ^*, assessed either in terms of its ℓ_2-error $\|\widehat{\theta} - \theta^*\|_2$ or its ℓ_1-error $\|\widehat{\theta} - \theta^*\|_1$.

It is worth noting that whenever $p > n$, the standard linear model (23) is unidentifiable, since the rectangular matrix $X \in \mathbb{R}^{n \times p}$ has a nullspace of dimension at least $p - n$. Consequently, in order to obtain an identifiable model—or at the very least, to bound the degree of non-identifiability—it is essential to impose additional constraints on the regression vector θ^*. One natural constraint is that of some type of sparsity in the regression vector; for instance, one might assume that θ^* has at most s non-zero coefficients, as discussed at more length in Section 4.2. More generally, one might assume that although θ^* is not exactly sparse, it can be well-approximated by a sparse vector, in which case one might say that θ^* is "weakly sparse", "sparsifiable" or "compressible". Section 4.3 is devoted to a more detailed discussion of this weakly sparse case.

A natural M-estimator for this problem is the constrained basis pursuit or Lasso [22, 72], obtained by solving the ℓ_1-penalized quadratic program

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\theta \in \mathbb{R}^p} \big\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \big\}, \tag{24}$$

for some choice $\lambda_n > 0$ of regularization parameter. Note that this Lasso estimator is a particular case of the general M-estimator (1), based on the loss function and regularization pair $\mathcal{L}(\theta; Z_1^n) = \frac{1}{2n}\|y - X\theta\|_2^2$ and $r(\theta) = \sum_{j=1}^p |\theta_j| \;=\; \|\theta\|_1$. We now show how Theorem 1 can be specialized to obtain bounds on the error $\widehat{\theta}_{\lambda_n} - \theta^*$ for the Lasso estimate.

4.1 Restricted eigenvalues for sparse linear regression

We begin by discussing the special form of restricted strong convexity for the least-squares loss function that underlies the Lasso. For the quadratic loss function $\mathcal{L}(\theta; Z_1^n) = \frac{1}{2n}\|y - X\theta\|_2^2$, the first-order Taylor series expansion from Definition 2 is exact, so that

$$\delta\mathcal{L}(\Delta, \theta^*; Z_1^n) \;=\; \langle \Delta, \frac{1}{n} X^T X \rangle \Delta \;=\; \frac{1}{n}\|X\Delta\|_2^2.$$

This exact relation allows for substantial theoretical simplification. In particular, in order to establish restricted strong convexity (see Definition 2) for quadratic losses, it suffices to establish a lower bound on $\|X\Delta\|_2^2/n$, one that holds uniformly for an appropriately restricted subset of p-dimensional vectors Δ.

As previously discussed in Example 1, for any subset $S \subseteq \{1, 2, \ldots, p\}$, the ℓ_1-norm is decomposable with respect to the subspace $A(S) = \{\alpha \in \mathbb{R}^p \mid \alpha_{S^c} = 0\}$ and its orthogonal complement. When the unknown regression vector $\theta^* \in \mathbb{R}^p$ is exactly sparse, it is natural to choose S equal to the support set of θ^*. By appropriately specializing the definition (14) of $\mathbb{C}$, we are led to consider the cone

$$\mathbb{C}(S) := \big\{\Delta \in \mathbb{R}^p \mid \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\big\}. \tag{25}$$

See Figure 1(a) for an illustration of this set in three dimensions. With this choice and the quadratic loss function, restricted strong convexity with respect to the ℓ_2-norm is equivalent to requiring that the design matrix X satisfy the condition

$$\frac{\|X\theta\|_2^2}{n} \geq \kappa_{\mathcal{L}} \, \|\theta\|_2^2 \qquad \text{for all } \theta \in \mathbb{C}(S). \tag{26}$$

This is a type of *restricted eigenvalue* (RE) condition, and has been studied in past work on basis pursuit and the Lasso (e.g., [8, 50, 59, 76]). It is a much milder condition than the restricted isometry property [18], since only a lower bound is required, and the strictly positive constant $\kappa_{\mathcal{L}}$ can be arbitrarily close to zero. Indeed, as shown by Bickel et al. [8], the restricted isometry property (RIP) implies the RE condition (26), but not vice versa. More strongly, Raskutti et al. [61] give examples of matrix families for which the RE condition (26) holds, but the RIP constants tend to infinity as $(n, |S|)$ grow. One could also enforce that a related condition hold with respect to the ℓ_1-norm—viz.

$$\frac{\|X\theta\|_2^2}{n} \geq \kappa'_{\mathcal{L}} \, \frac{\|\theta\|_1^2}{|S|} \qquad \text{for all } \theta \in \mathbb{C}(S). \tag{27}$$

This type of ℓ_1-based RE condition is less restrictive than the corresponding ℓ_2 version (26); see van de Geer and Bühlmann [77] for further discussion.

It is natural to ask whether there are many matrices that satisfy these types of RE conditions. This question was addressed by Raskutti et al. [59, 61], who showed that if the design matrix $X \in \mathbb{R}^{n \times p}$ is formed by independently sampling each row $X_i \sim N(0, \Sigma)$, referred to as the Σ-*Gaussian ensemble*, then there are strictly positive constants (κ_1, κ_2), depending only on the positive definite matrix Σ, such that

$$\frac{\|X\theta\|_2}{\sqrt{n}} \geq \kappa_1 \, \|\theta\|_2 - \kappa_2 \sqrt{\frac{\log p}{n}} \, \|\theta\|_1 \qquad \text{for all } \theta \in \mathbb{R}^p \tag{28}$$

with probability greater than $1 - c_1 \exp(-c_2 n)$. In particular, the results of Raskutti et al. [61] imply that this bound holds with $\kappa_1 = \frac{1}{4}\lambda_{min}(\sqrt{\Sigma})$ and $\kappa_2 = 9\sqrt{\max_{j=1,\ldots,p} \Sigma_{jj}}$, but sharper results are possible.

The bound (28) has an important consequence: it guarantees that the RE property (26) holds[1] with $\kappa_{\mathcal{L}} = \frac{\kappa_1}{2} > 0$ as long as $n > 64(\kappa_2/\kappa_1)^2 \, s \log p$. Therefore, not only do there exist matrices

[1] To see this fact, note that for any $\theta \in \mathbb{C}(S)$, we have $\|\theta\|_1 \leq 4\|\theta_S\|_1 \leq 4\sqrt{s}\|\theta_S\|_2$. Given the lower bound (28), for any $\theta \in \mathbb{C}(S)$, we have the lower bound $\frac{\|X\theta\|_2}{\sqrt{n}} \geq \big\{\kappa_1 - 4\kappa_2\sqrt{\frac{s \log p}{n}}\big\} \, \|\theta\|_2 \geq \frac{\kappa_1}{2}\|\theta\|_2$, where final inequality follows as long as $n > 64(\kappa_2/\kappa_1)^2 \, s \log p$.

satisfying the RE property (26), but it holds with high probability for any matrix random sampled from a Σ-Gaussian design. Related analysis by Zhou [88] extends these types of guarantees to the case of sub-Gaussian designs.

4.2 Lasso estimates with exact sparsity

We now show how Corollary 1 can be used to derive convergence rates for the error of the Lasso estimate when the unknown regression vector θ^* is s-sparse. In order to state these results, we require some additional notation. Using $X_j \in \mathbb{R}^n$ to denote the j^{th} column of X, we say that X is *column-normalized* if

$$\frac{\|X_j\|_2}{\sqrt{n}} \;\leq\; 1 \qquad \text{for all } j = 1, 2, \ldots, p. \tag{29}$$

Here we have set the upper bound to one in order to simplify notation as much as possible. This particular choice entails no loss of generality, since we can always rescale the linear model appropriately (including the observation noise variance) so that it holds.

In addition, we assume that the noise vector $w \in \mathbb{R}^n$ is zero-mean and has *sub-Gaussian tails*, meaning that there is a constant $\sigma > 0$ such that for any fixed $\|v\|_2 = 1$,

$$\mathbb{P}\big[|\langle v, w\rangle| \geq t\big] \leq 2 \exp\big(-\frac{\delta^2}{2\sigma^2}\big) \qquad \text{for all } \delta > 0. \tag{30}$$

For instance, this condition holds in the special case of i.i.d. $N(0, \sigma^2)$ Gaussian noise; it also holds whenever the noise vector w consists of independent, bounded random variables. Under these conditions, we recover as a corollary of Theorem 1 the following result:

Corollary 2. *Consider an s-sparse instance of the linear regression model* (23) *such that X satisfies the RE condition* (26)*, and the column normalization condition* (29)*. Given the Lasso program* (24) *with regularization parameter $\lambda_n = 4\,\sigma\sqrt{\frac{\log p}{n}}$, then with probability at least $1 - c_1 \exp(-c_2 n \lambda_n^2)$, any solution $\widehat{\theta}_{\lambda_n}$ satisfies the bounds*

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 \;\leq\; \frac{64\,\sigma^2}{\kappa_{\mathcal{L}}^2}\;\frac{s \log p}{n}, \quad \textit{and} \tag{31a}$$

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_1 \;\leq\; \frac{24\,\sigma}{\kappa_{\mathcal{L}}}\; s\,\sqrt{\frac{\log p}{n}}. \tag{31b}$$

Although these error bounds are known from past work [8, 50, 76], our proof illuminates the underlying structure that leads to the different terms in the bound—in particular, see equations (22a) and (22b) in the statement of Corollary 1.

Proof. We first note that the RE condition (27) implies that RSC holds with respect to the subspace $A(S)$. As discussed in Example 1, the ℓ_1-norm is decomposable with respect to $A(S)$ and its orthogonal complement, so that we may set $B(S) = A(S)$. Since any vector $\theta \in A(S)$ has at most s non-zero entries, the subspace compatibility constant is given by $\Psi(A(S)) = \sup\limits_{\theta \in A(S)\backslash\{0\}} \frac{\|\theta\|_1}{\|\theta\|_2} = \sqrt{s}$.

The final step is to compute an appropriate choice of the regularization parameter. The gradient of the quadratic loss is given by $\nabla \mathcal{L}(\theta; (y, X)) = \frac{1}{n} X^T w$, whereas the dual norm of the ℓ_1-norm is the ℓ_∞-norm. Consequently, we need to specify a choice of $\lambda_n > 0$ such that

$$\lambda_n \geq 2\, r^*(\nabla \mathcal{L}(\theta^*; Z_1^n)) = 2\big\|\frac{1}{n} X^T w\big\|_\infty$$

with high probability. Using the column normalization (29) and sub-Gaussian (30) conditions, for each $j = 1, \ldots, p$, we have the tail bound $\mathbb{P}\big[|\langle X_j, w\rangle / n| \geq t\big] \leq 2 \exp\big(-\frac{nt^2}{2\sigma^2}\big)$. Consequently, by union bound, we conclude that $\mathbb{P}\big[\|X^T w/n\|_\infty \geq t\big] \leq 2\exp\big(-\frac{nt^2}{2\sigma^2} + \log p\big)$. Setting $t^2 = \frac{4\sigma^2 \log p}{n}$, we see that the choice of λ_n given in the statement is valid with probability at least $1 - c_1 \exp(-c_2 n \lambda_n^2)$. Consequently, the claims (31a) and (31b) follow from the bounds (22a) and (22b) in Corollary 1. □

4.3 Lasso estimates with weakly sparse models

We now consider regression models for which θ^* is not exactly sparse, but rather can be approximated well by a sparse vector. One way in which to formalize this notion is by considering the ℓ_q "ball" of radius R_q, given by

$$\mathbb{B}_q(R_q) := \{\theta \in \mathbb{R}^p \mid \sum_{i=1}^p |\theta_i|^q \leq R_q\}, \qquad \text{where } q \in [0, 1] \text{ is parameter.}$$

In the special case $q = 0$, this set corresponds to an exact sparsity constraint—that is, $\theta^* \in \mathbb{B}_0(R_0)$ if and only if θ^* has at most R_0 non-zero entries. More generally, for $q \in (0, 1]$, the set $\mathbb{B}_q(R_q)$ enforces a certain decay rate on the ordered absolute values of θ^*.

In the case of weakly sparse vectors, the constraint set $\mathbb{C}$ takes the form

$$\mathbb{C}(A, B; \theta^*) = \big\{\Delta \in \mathbb{R}^p \mid \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1 + 4\|\theta^*_{S^c}\|_1\big\}. \tag{32}$$

It is essential to note that—in sharp contrast to the case of exact sparsity—the set $\mathbb{C}$ is no longer a cone, but rather contains a ball centered at the origin. To see the difference, compare panels (a) and (b) of Figure 1. As a consequence, it is *never* possible to ensure that $\|X\theta\|_2/\sqrt{n}$ is uniformly bounded from below for all vectors θ in the set (32). For this reason, it is essential that Theorem 1 require only that such a bound only hold over the set $\mathbb{K}(\delta)$, formed by intersecting $\mathbb{C}$ with a spherical shell $\{\Delta \in \mathbb{R}^p \mid \|\Delta\|_\star = \delta\}$.

The random matrix result (28), stated in the previous section, allows us to establish a form of RSC that is appropriate for the setting of ℓ_q-ball sparsity. More precisely, as shown in the proof of the following result, it guarantees that RSC holds with a suitably small $\delta > 0$. To the best of our knowledge, the following corollary of Theorem 1 is a novel result.

Corollary 3. *Suppose that X satisfies the condition (28) and the column normalization condition (29); the noise w is sub-Gaussian (30); and θ^* belongs to $\mathbb{B}_q(R_q)$ for a radius R_q such that $\sqrt{R_q}\,\big(\frac{\log p}{n}\big)^{\frac{1}{2} - \frac{q}{4}} = o(1)$. Then if we solve the Lasso with regularization parameter $\lambda_n = 4\sigma\sqrt{\frac{\log p}{n}}$, there are strictly positive constants (c_1, c_2) such that any optimal solution $\widehat{\theta}_{\lambda_n}$ satisfies*

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq 64\, R_q \left(\frac{\sigma^2}{\kappa_1^2}\, \frac{\log p}{n}\right)^{1 - \frac{q}{2}} \tag{33}$$

with probability at least $1 - c_1 \exp(-c_2 n \lambda_n^2)$.

Remarks: Note that this corollary is a strict generalization of Corollary 2, to which it reduces when $q = 0$. More generally, the parameter $q \in [0, 1]$ controls the relative "sparsifiability" of θ^*, with larger values corresponding to lesser sparsity. Naturally then, the rate slows down as q increases from 0 towards 1. In fact, Raskutti et al. [59] show that the rates (33) are minimax-optimal over the ℓ_q-balls—implying that not only are the consequences of Theorem 1 sharp for the Lasso, but more generally, no algorithm can achieve faster rates.

Proof. Since the loss function $\mathcal{L}$ is quadratic, the proof of Corollary 2 shows that the stated choice $\lambda_n = 4\sqrt{\frac{\sigma^2 \log p}{n}}$ is valid with probability at least $1 - c\exp(-c' n\lambda_n^2)$. Let us now show that the RSC condition holds. We do so via condition (28), in particular showing that it implies the RSC condition over an appropriate set as long as $\|\widehat{\theta} - \theta^*\|_2$ is sufficiently large. For a threshold $\tau > 0$ to be chosen, define the thresholded subset

$$S_\tau := \big\{ j \in \{1, 2, \ldots, p\} \;\mid\; |\theta^*_j| > \tau \big\}. \tag{34}$$

Now recall the subspaces $A(S_\tau)$ and $A^\perp(S_\tau)$ previously defined, as in equations (4) and (5) of Example 1 with $S = S_\tau$. The following lemma, proved in Appendix B, provides sufficient conditions for restricted strong convexity with respect to these subspace pairs:

Lemma 2. *Under the assumptions of Corollary 3 and the choice $\tau = \frac{\lambda_n}{\kappa_1}$, then the RSC condition holds with $\kappa_{\mathcal{L}} = \kappa_1/4$ over the set $\mathbb{K}(\delta; A(S_\tau), B(S_\tau), \theta^*)$ for all tolerance parameters*

$$\delta \geq \delta^* \;:=\; \frac{4\,\kappa_2}{\kappa_1}\sqrt{\frac{\log p}{n}}\, R_q\, \tau^{1-q} \tag{35}$$

This result guarantees that we may apply Theorem 1 with over $\mathbb{K}(\delta^*)$ with $\kappa_{\mathcal{L}} = \kappa_1/4$, from which we obtain

$$\|\widehat{\theta} - \theta^*\|_2 \leq \max\Bigg\{\delta^*,\; \frac{8\lambda_n}{\kappa_1}\Psi(A(S_\tau)) + \sqrt{\frac{8\lambda_n\, \|\theta^*_{S^c_\tau}\|_1}{\kappa_1}}\Bigg\}. \tag{36}$$

From the proof of Corollary 2, we have $\Psi(A(S_\tau)) = \sqrt{S_\tau}$. It remains to upper bound the cardinality of S_τ in terms of the threshold τ and ℓ_q-ball radius R_q. Note that we have

$$R_q \geq \sum_{j=1}^{p} |\theta^*_j|^q \;\geq\; \sum_{j \in S_\tau} |\theta^*_i|^q \;\geq\; \tau^q |S_\tau|, \tag{37}$$

whence $|S_\tau| \leq \tau^{-q}\, R_q$ for any $\tau > 0$. Next we upper bound the approximation error $\|\theta^*_{S^c_\tau}\|_1$, using the fact that $\theta^* \in \mathbb{B}_q(R_q)$. Letting S^c_τ denote the complementary set $S_\tau \backslash \{1, 2, \ldots, p\}$, we have

$$\|\theta^*_{S^c_\tau}\|_1 = \sum_{j \in S^c_\tau} |\theta^*_j| \;=\; \sum_{j \in S^c_\tau} |\theta^*_j|^q |\theta^*_j|^{1-q} \leq R_q\, \tau^{1-q}. \tag{38}$$

Setting $\tau = \lambda_n/\kappa_1$ and then substituting the bounds (37) and (38) into the bound (36) yields

$$\|\widehat{\theta} - \theta^*\|_2 \leq \max\Bigg\{\delta^*,\; 8\sqrt{R_q}\big(\frac{\lambda_n}{\kappa_1}\big)^{1-q/2} + \sqrt{8R_q\big(\frac{\lambda_n}{\kappa_1}\big)^{2-q}}\Bigg\} \;\leq\; \max\Bigg\{\delta^*,\; 16\sqrt{R_q}\big(\frac{\lambda_n}{\kappa_1}\big)^{1-q/2}\Bigg\}.$$

Under the assumption $\sqrt{R_q}\,\big(\frac{\log p}{n}\big)^{\frac{1}{2}-\frac{q}{4}} = o(1)$, the critical δ^* in the condition (35) is of lower order than the second term, so that the claim follows. □

5 Convergence rates for group-structured norms

The preceding two sections addressed M-estimators based on ℓ_1-regularization, the simplest type of decomposable regularizer. We now turn to some extensions of our results to more complex regularizers that are also decomposable. Various researchers have proposed extensions of the Lasso based on regularizers that have more structure than the ℓ_1 norm (e.g., [75, 84, 86, 47, 5]). Such regularizers allow one to impose different types of block-sparsity constraints, in which groups of parameters are assumed to be active (or inactive) simultaneously. These norms naturally arise in the context of multivariate regression, where the goal is to predict a multivariate output in $\mathbb{R}^m$ on the basis of a set of p covariates. Here it is natural to assume that groups of covariates are useful for predicting the different elements of the m-dimensional output vector. We refer the reader to the papers papers [75, 84, 86, 47, 5] for further discussion and motivation of such block-structured norms.

Given a collection $\mathcal{G} = \{G_1, \ldots, G_T\}$ of groups, recall from Example 2 in Section 2.2 the definition of the group norm $\| \cdot \|_{\mathcal{G}, \vec{\nu}}$. In full generality, this group norm is based on a weight vector $\vec{\nu} = (\nu_1, \ldots, \nu_T) \in [1, \infty]^T$, one for each group. For simplicity, here we consider the case when $\nu_t = \nu$ for all $t = 1, 2, \ldots, T$, and we use $\| \cdot \|_{\mathcal{G},\nu}$ to denote the associated group norm. As a natural generalization of the Lasso, we consider the *block Lasso* estimator

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \big\{ \frac{1}{n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_{\mathcal{G},\nu} \big\}, \tag{39}$$

where $\lambda_n > 0$ is a user-defined regularization parameter. Different choices of the parameter ν yield different estimators, and in this section, we consider the range $\nu \in [2, \infty]$. This range covers the two most commonly applied choices, $\nu = 2$, often referred to as the group Lasso, as well as the choice $\nu = +\infty$.

5.1 Restricted strong convexity for group sparsity

As a parallel to our analysis of ordinary sparse regression, our first step is to provide a condition sufficient to guarantee restricted strong convexity for the group-sparse setting. More specifically, we state the natural extension of condition (28) to the block-sparse setting, and prove that it holds with high probability for the class of Σ-Gaussian random designs. Recall from Theorem 1 that the dual norm of the regularizer plays a central role. For the block-$(1,\nu)$-regularizer, the associated dual norm is a block-(∞, ν^*) norm, where (ν, ν^*) are conjugate exponents satisfying $\frac{1}{\nu} + \frac{1}{\nu^*} = 1$.

Letting $\varepsilon \sim N(0, I_{p \times p})$ be a standard normal vector, we consider the following condition. Suppose that there are strictly positive constants (κ_1, κ_2) such that, for all $\Delta \in \mathbb{R}^p$, we have

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \kappa_1 \|\Delta\|_2 - \kappa_2 \, \frac{\rho_{\mathcal{G}}(\nu^*)}{\sqrt{n}} \, \|\Delta\|_{1,\nu} \qquad \text{where } \rho_{\mathcal{G}}(\nu^*) := \mathbb{E}\Big[\max_{t=1,2,\ldots,T} \|\varepsilon_{G_t}\|_{\nu^*} \Big]. \tag{40}$$

To understand this condition, first consider the special case of $T = p$ groups, each of size one, so that the group-sparse norm reduces to the ordinary ℓ_1-norm. (The choice of $\nu \in [2, \infty]$ is irrelevant when the group sizes are all one.) In this case, we have

$$\rho_{\mathcal{G}}(2) = \mathbb{E}[\|w\|_\infty] \; \leq \sqrt{3 \log p},$$

using standard bounds on Gaussian maxima. Therefore, condition (40) reduces to the earlier condition (28) in this special case.

Let us consider a more general setting, say with $\nu = 2$ and T groups each of size m, so that $p = Tm$. For this choice of groups and norm, we have $\rho_{\mathcal{G}}(2) = \mathbb{E}\big[\max_{t=1,\ldots,T} \|w_{G_t}\|_2\big]$ where each subvector w_{G_t} is a standard Gaussian vector with m elements. Since $\mathbb{E}[\|w_{G_t}\|_2] \leq \sqrt{m}$, tail bounds for χ^2-variates yield $\rho_{\mathcal{G}}(2) \leq \sqrt{m} + \sqrt{3 \log T}$, so that the condition (40) is equivalent to

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \kappa_1 \|\Delta\|_2 - \kappa_2 \bigg[\sqrt{\frac{m}{n}} + \sqrt{\frac{3 \log T}{n}}\bigg] \, \|\Delta\|_{\mathcal{G},2} \qquad \text{for all } \Delta \in \mathbb{R}^p.$$

Thus far, we have seen the form that condition (40) takes for different choices of the groups and parameter ν. It is natural to ask whether there are any matrices that satisfy the condition (40). As shown in the following result, the answer is affirmative—more strongly, almost every matrix satisfied from the Σ-Gaussian ensemble will satisfy this condition with high probability. (Here we recall that for a non-degenerate covariance matrix, a random design matrix $X \in \mathbb{R}^{n \times p}$ is drawn from the Σ-Gaussian ensemble if each row $x_i \sim N(0, \Sigma)$, i.i.d. for $i = 1, 2, \ldots, n$.)

Proposition 1. *For a design matrix $X \in \mathbb{R}^{n \times p}$ from the Σ-ensemble, there are constants (κ_1, κ_2) depending only Σ such that condition (40) holds with probability greater than $1 - c_1 \exp(-c_2 n)$.*

We provide the proof of this result in Appendix C.1. This condition can be used to show that appropriate forms of RSC hold, for both the cases of exactly group-sparse and weakly sparse vectors. As with ℓ_1-regularization, these RSC conditions are much milder than analogous RIP conditions (e.g., [28, 71, 5]), which require that all sub-matrices up to a certain size are close to isometries.

5.2 Convergence rates

Apart from RSC, we impose one additional condition on the design matrix. For a given group G of size m, let us view X_G as an operator from $\ell_\nu^m \to \ell_2^n$, and define the associated operator norm $|\!|\!|X_G|\!|\!|_{\nu \to 2} := \max_{\|\theta\|_\nu = 1} \|X_G \, \theta\|_2$. We then require that

$$\frac{|\!|\!|X_{G_t}|\!|\!|_{\nu \to 2}}{\sqrt{n}} \leq 1 \qquad \text{for all } t = 1, 2, \ldots, T. \tag{41}$$

Note that this is a natural generalization of the column normalization condition (29), to which it reduces when we have $T = p$ groups, each of size one. As before, we may assume without loss of generality, rescaling X and the noise as necessary, that condition (41) holds with constant one. Finally, we define the maximum group size $m = \max_{t=1,\ldots,T} |G_t|$. With this notation, we have the following novel result:

Corollary 4. *Suppose that the noise w is sub-Gaussian (30), and the design matrix X satisfies condition (40) and the block normalization condition (41). If we solve the group Lasso with*

$$\lambda_n \geq 2\sigma \bigg\{ \frac{m^{1-1/\nu}}{\sqrt{n}} + \sqrt{\frac{\log T}{n}} \bigg\}, \tag{42}$$

then with probability at least $1 - 2/T^2$, for any group subset $S_{\mathcal{G}} \subseteq \{1, 2, \ldots, T\}$ with cardinality $|S_{\mathcal{G}}| = s_{\mathcal{G}}$, any optimal solution $\widehat{\theta}_{\lambda_n}$ satisfies

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq \frac{4\lambda_n^2}{\kappa_{\mathcal{L}}^2} \, s_{\mathcal{G}} + \frac{4\lambda_n}{\kappa_{\mathcal{L}}} \sum_{t \notin S_{\mathcal{G}}} \|\theta^*_{G_t}\|_\nu. \tag{43}$$

Remarks: Since the result applies to any $\nu \in [2, \infty]$, we can observe how the choices of different group-sparse norms affect the convergence rates. So as to simplify this discussion, let us assume that the groups are all of equal size so that $p = mT$ is the ambient dimension of the problem.

Case $\nu = 2$: The case $\nu = 2$ corresponds to the block $(1, 2)$ norm, and the resulting estimator is frequently referred to as the group Lasso. For this case, we can set the regularization parameter as $\lambda_n = 2\sigma\big\{\sqrt{\frac{m}{n}} + \sqrt{\frac{\log T}{n}}\big\}$. If we assume moreover that θ^* is exactly group-sparse, say supported on a group subset $S_{\mathcal{G}} \subseteq \{1, 2, \ldots, T\}$ of cardinality $s_{\mathcal{G}}$, then the bound (43) takes the form

$$\|\widehat{\theta} - \theta^*\|_2^2 = \mathcal{O}\bigg(\frac{s_{\mathcal{G}}\, m}{n} + \frac{s_{\mathcal{G}}\, \log T}{n}\bigg). \tag{44}$$

Similar bounds were derived in independent work by Lounici et al. [43] and Huang and Zhang [28] for this special case of exact block sparsity. The analysis here shows how the different terms arise, in particular via the noise magnitude measured in the dual norm of the block regularizer.

In the more general setting of weak block sparsity, Corollary 4 yields a number of novel results. For instance, for a given set of groups $\mathcal{G}$, we can consider the block sparse analog of the ℓ_q-"ball"—namely the set

$$\mathbb{B}_q(R_q; \mathcal{G}, 2) := \bigg\{\theta \in \mathbb{R}^p \;\mid\; \sum_{t=1}^{T} \|\theta_{G_t}\|_2^q \leq R_q\bigg\}.$$

In this case, if we optimize the choice of S in the bound (43) so as to trade off the estimation and approximation errors, then we obtain

$$\|\widehat{\theta} - \theta^*\|_2^2 = \mathcal{O}\bigg(R_q\bigg[\frac{m}{n} + \frac{\log T}{n}\bigg]^{1-\frac{q}{2}}\bigg),$$

which is a novel result. This result is a generalization of our earlier Corollary 3, to which it reduces when we have $T = p$ groups each of size $m = 1$.

Case $\nu = +\infty$: Now consider the case of ℓ_1/ℓ_∞ regularization, as suggested in past work [75]. In this case, Corollary 4 implies that $\|\widehat{\theta} - \theta^*\|_2^2 = \mathcal{O}\big(\frac{s\, m^2}{n} + \frac{s \log T}{n}\big)$. Similar to the case $\nu = 2$, this bound consists of an estimation term, and a search term. The estimation term $\frac{sm^2}{n}$ is larger by a factor of m, which corresponds to amount by which an ℓ_∞ ball in m dimensions is larger than the corresponding ℓ_2 ball.

We provide the proof of Corollary 4 in Appendix C.2. It is based on verifying the conditions of Theorem 1: more precisely, we use Proposition 1 in order to establish RSC, and we provide a lemma that shows that the regularization choice (42) is valid in the context of Theorem 1.

6 Convergence rates for generalized linear models

The previous sections were devoted to the study of high-dimensional linear models, under either sparsity constraints (Section 4) or group-sparse constraints (Section 5). In these cases, the quadratic loss function allowed us to reduce the study of restricted strong convexity to the study of restricted eigenvalues. However, the assumption of an ordinary linear model is limiting, and may not be suitable in some settings. For instance, the response variables may be binary as in a classification problem, or follow another type of exponential family distribution (e.g., Poisson, Gamma etc.) It is natural then to consider high-dimensional models that involve non-quadratic losses. This section is devoted to an in-depth analysis of the setting in which the real-valued response $y \in \mathcal{Y}$ is linked to the p-dimensional covariate $x \in \mathbb{R}^p$ via a generalized linear model (GLM) with canonical link function. This extension, though straightforward from a conceptual viewpoint, introduces some technical challenges. Restricted strong convexity (RSC) is a more subtle condition in this setting, and one of the results in this section is Proposition 2, which guarantees that RSC holds with high probability for a broad class of GLMs.

6.1 Generalized linear models and M-estimation

We begin with background on generalized linear models. Suppose that conditioned on the covariate vector, the response variable has the distribution

$$\mathbb{P}(y \mid x; \theta^*, \sigma) \propto \exp\biggl\{\frac{y\,\langle\theta^*, x\rangle - \psi(\langle\theta^*, x\rangle)}{c(\sigma)}\biggr\} \tag{45}$$

Here the scalar $\sigma > 0$ is a fixed and known scale parameter, whereas the vector $\theta^* \in \mathbb{R}^p$ is fixed but unknown, and our goal is to estimate it. The function $\psi : \mathbb{R} \mapsto \mathbb{R}$ is known as the link function. We assume that ψ is defined on all of $\mathbb{R}$; otherwise, it would be necessary to impose constraints on the vector $\langle\theta^*, x\rangle$ so as to ensure that the model was meaningfully defined. From standard properties of exponential families [10], it follows that ψ is infinitely differentiable, and its second derivative ψ'' is strictly positive on the real line.

Let us illustrate this set-up with some standard examples:

Example 4(a) The model (45) includes the usual linear model as a special case. In particular, let the response y take values in $\mathcal{Y} = \mathbb{R}$, define the link function $\psi(u) = u^2/2$ and set $c(\sigma) = \sigma^2$, where σ is the standard deviation of the observation noise. The conditional distribution (45) takes the form

$$\mathbb{P}(y \mid x; \theta^*, \sigma) \propto \exp\biggl\{\frac{y\,\langle\theta^*, x\rangle - (\langle\theta^*, x\rangle)^2/2}{\sigma^2}\biggr\}.$$

In this way, we recognize that conditioned on x, the response y has a Gaussian distribution with mean $\langle\theta^*, x\rangle$ and variance σ^2.

(b) For the logistic regression model, the response y takes binary values ($\mathcal{Y} = \{0, 1\}$), and the link function is given by $\psi(u) = \log(1 + \exp(u))$. There is no scale parameter σ for this model, and we have

$$\mathbb{P}(y \mid x; \theta^*) = \exp\bigl\{y\langle\theta^*, x\rangle - \log(1 + \exp(\langle\theta^*, x\rangle))\bigr\},$$

so that conditioned on x, the response y is a Bernoulli variable with $\mathbb{P}(y = 1 \mid x; \theta^*) = \frac{\exp(\langle\theta^*, x\rangle)}{1+\exp(\langle\theta^*, x\rangle)}$.

(c) In the Poisson model, the response y takes a positive integer value in $\mathcal{Y} = \{0, 1, 2, \ldots\}$, and the conditional distribution of the response given the covariate takes the form

$$\mathbb{P}(y \mid x; \theta^*) = \exp\big\{y\langle\theta^*, x\rangle - \exp(\langle\theta^*, x\rangle)\big\},$$

corresponding to a GLM with link function $\psi(u) = \exp(u)$.

$\diamondsuit$

We now turn to the estimation problem of interest in the GLM setting. Suppose that we sample n i.i.d. covariate vectors $\{x_i\}_{i=1}^p$ from some distribution over $\mathbb{R}^p$. For each $i = 1, \ldots, n$, we then draw an independent response variable $y_i \in \mathcal{Y}$ according to the distribution $\mathbb{P}(y \mid x_i; \theta^*)$, Given the collection of observations $Z_1^n := \{(x_i, y_i)\}_{i=1}^n$, the goal is to estimate the parameter vector $\theta^* \in \mathbb{R}^p$. When θ^* is an s-sparse vector, it is natural to consider the estimator based on ℓ_1*-regularized maximum likelihood*—namely, the convex program

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\theta\in\mathbb{R}^p}\Big\{\underbrace{-\langle\theta, \frac{1}{n}\sum_{i=1}^n y_i x_i\rangle + \frac{1}{n}\sum_{i=1}^n \psi\big(\langle\theta, x_i\rangle\big)}_{\mathcal{L}(\theta; Z_1^n)} + \lambda_n\|\theta\|_1\Big\}. \tag{46}$$

This estimator is a special case of our M-estimator (1) with $r(\theta) = \|\theta\|_1$.

6.2 Restricted strong convexity for GLMs

In order to leverage Theorem 1, we need to establish that an appropriate form of RSC holds for generalized linear models. In the special case of linear models, we exploited a result in random matrix theory, as stated in the bound (28), in order to verify these properties, both for exactly sparse and weakly sparse models. In this section, we state an analogous result for generalized linear models. It applies to models with sub-Gaussian behavior (see equation (30)), as formalized by the following condition:

(GLM1) The rows x_i, $i = 1, 2, \ldots, n$ of the design matrix are i.i.d. samples from a zero-mean distribution with covariance matrix $\text{cov}(x_i) = \Sigma$ such that $\lambda_{min}(\Sigma) \geq \kappa_\ell > 0$, and for any $v \in \mathbb{R}^p$, the variable $\langle v, x_i\rangle$ is sub-Gaussian with parameter at most $\kappa_u\|v\|_2^2$.

We recall that RSC involves lower bounds on the error in the first-order Taylor series expansion—namely, the quantity

$$\delta\mathcal{L}(\Delta, \theta^*; Z_1^n) := \mathcal{L}(\theta^* + \Delta; Z_1^n) - \mathcal{L}(\theta^*; Z_1^n) - \langle\nabla\mathcal{L}(\theta^*; Z_1^n), \Delta\rangle.$$

When the loss function is more general than least-squares, then the quantity $\delta\mathcal{L}$ cannot be reduced to a fixed quadratic form, Rather, we need to establish a result that provides control uniformly over a neighborhood. The following result is of this flavor:

Proposition 2. *For any minimal GLM satisfying condition (GLM1) , there exist positive constants* κ_1 *and* κ_2*, depending only on* (κ_ℓ, κ_u) *and* ψ*, such that*

$$\delta\mathcal{L}(\Delta, \theta^*; Z_1^n) \;\geq\; \kappa_1\|\Delta\|_2\left\{\|\Delta\|_2 - \kappa_2\sqrt{\frac{\log p}{n}}\;\|\Delta\|_1\right\} \qquad \textit{for all } \|\Delta\|_2 \leq 1. \tag{47}$$

with probability at least $1 - c_1\exp(-c_2 n)$.

Note that Proposition 2 is a lower bound for an an empirical process; in particular, given the empirical process $\{\delta\mathcal{L}(\Delta, \theta^*; Z_1^n),\ \Delta \in \mathbb{R}^p\}$, it yields a lower bound on the infimum over $\|\Delta\|_2 \leq 1$. The proof of this result, provided in Appendix D.1, requires a number of techniques from empirical process theory, among them concentration bounds, the Ledoux-Talagrand contraction inequality, and a peeling argument. A subtle aspect is the use of an appropriate truncation argument to deal with the non-Lipschitz and unbounded functions that it covers.

In the current context, the most important consequence of Proposition 2 is that it guarantees that GLMs satisfy restricted strong convexity for s-sparse models as long as $n = \Omega(s \log p)$. This follows by the same argument as in the quadratic case (see Section 4.1).

6.3 Convergence rates

We now use Theorem 1 and Proposition 2 in order to derive convergence rates for the GLM Lasso (46). In addition to our sub-Gaussian assumptions on the covariates (Condition (GLM1)), we require the following mild regularity conditions on the link function:

(GLM2) For some constant $M_\psi > 0$, at least one of the following two conditions holds:

(i) The Hessian of the cumulant function is uniformly bounded: $\|\psi''\|_\infty \leq M_\psi$, or

(ii) The covariates are bounded ($|x_{ij}| \leq 1$), and

$$\mathbb{E}\big[\max_{|u|\leq 1}[\psi''(\langle\theta^*, x\rangle + u)]^\alpha\big] \leq M_\psi \qquad \text{for some } \alpha \geq 2. \tag{48}$$

Whereas condition (GLM1) applies purely to the random choice of the covariates $\{x_i\}_{i=1}^n$, condition (GLM2) also involves the structure of the GLM. Straightforward calculations show that the standard linear model and the logistic model both satisfy (GLM2) (i), whereas the Poisson model satisfies the moment condition when the covariates are bounded. The following result provides convergence rates for GLMs; we state it in terms of positive constants (c_0, c_1, c_2, c_3) that depend on the parameters in conditions (GLM1) and (GLM2) .

Corollary 5. *Consider a GLM satisfying conditions (GLM1) and (GLM2) , and suppose that $n > 9\kappa_2^2\, s \log p$. For regularization parameter $\lambda_n = c_0\sqrt{\frac{\log p}{n}}$, any optimal solution $\widehat{\theta}_{\lambda_n}$ to the GLM Lasso satisfies the bound*

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq c_1 \frac{s \log p}{n} \tag{49}$$

with probability greater than $1 - c_2/n^\alpha$.

The proof, provided in Appendix D.2, is based on a combination of Theorem 1 and Proposition 2, along with some tail bounds to establish that the specified choice of regularization parameter λ_n is suitable. Note that up to differences in constant factors, the rate (49) is the same as the ordinary linear model with an s-sparse vector.

7 Convergence rates for low rank matrices

Finally, we consider the implications of our main result for the problem of estimating a matrix $\Theta^* \in \mathbb{R}^{p_1 \times p_2}$ that is either exactly low rank, or well-approximated by low rank matrices. Such estimation problems arise in various guises, including principal component analysis, multivariate regression, matrix completion, and system identification.

7.1 Matrix-based regression and examples

In order to treat a range of instances in a unified manner, it is helpful to set up the problem of matrix regression. For a pair of $p_1 \times p_2$ matrices Θ and X, recall the usual trace inner product $\langle\langle X, \; \Theta \rangle\rangle := \operatorname{trace}(X^T \Theta)$. In terms of this notation, we then consider an observation model in which each observation $y_i \in \mathbb{R}$ is a noisy version of such an inner product. More specifically, for each $i = 1, \ldots, n$, we let $X_i \in \mathbb{R}^{p_1 \times p_2}$ denote a matrix, and define

$$y_i = \langle\langle X_i, \; \Theta^* \rangle\rangle + w_i, \qquad \text{for } i = 1, 2, \ldots, n, \tag{50}$$

where $w_i \in \mathbb{R}$ is an additive observation noise. When the unknown matrix Θ^* has low rank structure, it is natural to consider the estimator

$$\widehat{\Theta}_{\lambda_n} \in \arg \min_{\Theta \in \mathbb{R}^{p_1 \times p_2}} \Big\{ \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - \langle\langle X_i, \; \Theta \rangle\rangle \big)^2 + \lambda_n \, |||\Theta|||_1 \Big\}, \tag{51}$$

using the nuclear norm $|||\Theta|||_1 = \sum_{j=1}^{\min\{p_1, p_2\}} \sigma_j(\Theta)$ for regularization. Note that the estimator (51) is a special case of our general estimator (1). From the computational perspective, it is an instance of a semidefinite program [79], and there are a number of efficient algorithms for solving it (e.g., [9, 41, 45, 55, 56]).

Let us consider some examples to illustrate applications covered by the model (51).

Example 5(a) *Matrix completion* [19, 52, 65, 68, 70]: Suppose that one observes a relatively small subset of the entries of a matrix $\Theta^* \in \mathbb{R}^{p_1 \times p_2}$. These observations may be uncorrupted, or (as is relevant here) corrupted by additive noise. If n entries are observed, then one has access to the samples $y_i = \Theta^*_{a(i)b(i)} + w_i$ for $i = 1, 2, \ldots, n$, where $(a(i), b(i))$ corresponds to the matrix entry associated with observation i. This set-up is an instance of the model (50), based on the observation matrices $X_i = E_{a(i)b(i)}$, corresponding to the matrix with one in position $(a(i), b(i))$ and zeroes elsewhere.

(b) *Multivariate regression* [54, 68, 83]: Suppose that for $j = 1, 2, \ldots, p_2$, we are given a regression problem in p_1 dimensions, involving the unknown regression vector $\theta^*_j \in \mathbb{R}^{p_1}$. By stacking these regression vectors as columns, we can form a $p_1 \times p_2$ matrix

$$\Theta^* = \begin{bmatrix} \theta^*_1 & \theta^*_2 & \cdots & \theta^*_{p_2} \end{bmatrix}$$

of regression coefficients. A multivariate regression problem based on N observations then takes the form $Y = Z\Theta^* + W$, where $Z \in \mathbb{R}^{N \times p_1}$ is the matrix of covariates, and $Y \in \mathbb{R}^{N \times p_2}$ is the matrix of response vectors. In many applications, it is natural to assume that the regression vectors $\{\theta^*_j, j = 1, 2, \ldots, p_2\}$ are relatively similar, and so can be modeled as lying within or close to a lower dimensional subspace. Equivalently, the matrix Θ^* is either exactly low rank, and approximately low rank. As an example, in multi-view imaging, each regression problem is obtained from a different camera viewing the same scene, so that one expects substantial sharing among the different regression problems. This model can be reformulated as an instance of the observation model (50) based on $n = N \, p_2$ observations; see the paper [54] for details.

(c) *Compressed sensing* [16, 54, 66]: The goal of compressed sensing to estimate a parameter vector or matrix based on noisy random projections. In the matrix variant [66], one makes noisy observations of the form (50), where each X_i is a standard Gaussian random matrix (i.e., with i.i.d. $N(0,1)$ entries).

(d) *System identification in autoregressive processes* [25, 54, 42]: Consider a vector-valued stochastic process that is evolves according to the recursion $Z_{t+1} = \Theta^* Z_t + V_t$, where $W_t \sim N(0, \nu^2)$. Suppose that we observe a sequence $(Z_1, Z_2, \ldots, Z_N)$ generated by this vector autoregressive (VAR) process, and that our goal is to estimate the unknown system matrix $\Theta^* \in \mathbb{R}^{p \times p}$, where $p_1 = p_2 = p$ in this case. This problem of system identification can be tackled by solving a version of the M-estimator (51) with $n = N\,p$ observations, based on suitable definitions of the matrices X_i.

$\diamondsuit$

7.2 Restricted strong convexity for matrix estimation

Let us now consider the form of restricted strong convexity (RSC) relevant for the estimator (51). In order to do so, we first define an operator $\mathfrak{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$. This operator plays an analogous role to the design matrix in ordinary linear regression, and is defined elementwise by $[\mathfrak{X}(\Theta)]_i = \langle\langle X_i,\ \Theta \rangle\rangle$. Using this notation, the observation model can be re-written in a vector form as $y = \mathfrak{X}(\Theta^*) + w$, and the estimator (51) takes the form

$$\widehat{\Theta}_{\lambda_n} \in \arg \min_{\Theta \in \mathbb{R}^{p_1 \times p_2}} \Big\{ \underbrace{\frac{1}{2n} \|y - \mathfrak{X}(\Theta)\|_2^2}_{\mathcal{L}(\Theta; Z_1^n)} + \lambda_n \underbrace{|\!|\!|\Theta|\!|\!|_1}_{r(\Theta)} \Big\}.$$

Note that the Hessian of this quadratic loss function $\mathcal{L}$ is given by the operator $\frac{1}{n}(\mathfrak{X} \otimes \mathfrak{X})$, and the RSC condition amounts to establishing lower bounds of the form $\frac{1}{\sqrt{n}} \|\mathfrak{X}(\Delta)\|_2 \geq \kappa_{\mathcal{L}} \|\Delta\|_2$ that hold uniformly for a suitable subset of matrices $\Delta \in \mathbb{R}^{p_1 \times p_2}$.

Recall from Example 3 that the nuclear norm is decomposable with respect to subspaces of matrices defined in terms of their row and column spaces. We now show how suitable choices lead to a useful form of RSC for recovering near low-rank matrices. Given a rank k matrix $\Theta^* \in \mathbb{R}^{p_1 \times p_2}$, let $\Theta^* = U \Sigma V^T$ be its singular value decomposition (SVD), so that $U \in \mathbb{R}^{p_1 \times k}$ and $V \in \mathbb{R}^{p_2 \times k}$ are orthogonal, and $\Sigma \in \mathbb{R}^{k \times k}$ has the singular values of Θ^* along its diagonal, ordered from largest to smallest. For each integer $\ell = 1, 2, \ldots, k$, let $U^\ell \in \mathbb{R}^{p_1 \times \ell}$ denote the submatrix of left singular vectors associated with the top ℓ singular values, and define the submatrix $V^\ell \in \mathbb{R}^{p_2 \times \ell}$ similarly. As discussed in Example 3 of Section 2.2, the nuclear norm $r(\Theta) = |\!|\!|\Theta|\!|\!|_1$ is decomposable with respect to the subspaces $A^\ell = A(U^\ell, V^\ell)$ and $B = B(U^\ell, V^\ell)$. Recalling the definition (17) of the set $\mathbb{K}$, a matrix Δ belongs to $\mathbb{K}(\delta; A^\ell, B^\ell, \Theta^*)$ if and only if $|\!|\!|\Delta|\!|\!|_F = \delta$, and

$$|\!|\!|\Pi_{(B^\ell)^\perp}(\Delta)|\!|\!|_1 \leq 3 |\!|\!|\Pi_{B^\ell}(\Delta)|\!|\!|_1 + 4 \sum_{j=\ell+1}^{k} \sigma_j(\Theta^*),$$

and the RSC condition amounts to requiring

$$\frac{1}{\sqrt{n}} \|\mathfrak{X}(\Delta)\|_2 \geq \kappa_{\mathcal{L}} \|\Delta\|_2 \quad \text{for all } \Delta \in \mathbb{K}(\delta; A^\ell, B^\ell, \Theta^*). \tag{52}$$

This RSC condition can be shown to hold with high probability for different instantiations of the observation model (50). It is relatively straightforward to show [54] that it holds for multivariate regression and system identification of autoregressive models, as discussed in parts (b) and (d) of Example 5. For the case of compressed sensing, it has been shown that RIP conditions hold for certain types of sub-Gaussian matrices [66, 16]; other authors have imposed RIP conditions in the setting of multivariate regression [68]. As in the case of linear regression, RIP conditions are sufficient to guarantee the RSC property, but are much more restrictive in general. For instance, it is possible to construct sequences of models where the RIP constants tend to infinity while the RSC condition holds [54]. Even more strikingly, for the case of matrix completion, it is never possible for the RIP condition to hold; however, a fruitful version of restricted strong convexity does exist, and can be used to derive minimax-optimal bounds, up to logarithmic factors, for noisy matrix completion [52].

8 Discussion

In this paper, we have presented a unified framework for deriving convergence rates for a class of regularized M-estimators. The theory is high-dimensional and non-asymptotic in nature, meaning that it yields explicit bounds that hold with high probability for finite sample sizes, and reveals the dependence on dimension and other structural parameters of the model. Two properties of the M-estimator play a central role in our framework. We isolated the notion of a regularizer being *decomposable* with respect to a pair of subspaces, and showed how it constrains the error vector—meaning the difference between any solution and the nominal parameter—to lie within a very specific set. This fact is significant, because it allows for a fruitful notion of *restricted strong convexity* to be developed for the loss function. Since the usual form of strong convexity cannot hold under high-dimensional scaling, this interaction between the decomposable regularizer and the loss function is essential.

Our main result (Theorem 1) provides a deterministic bound on the error for a broad class of regularized M-estimators. By specializing this result to different statistical models, we derived various explicit convergence rates for different estimators, including some known results and a range of novel results. We derived convergence rates for sparse linear models, both under exact and approximate sparsity assumptions; our results for models with ℓ_q-ball are novel, and have been proven minimax-optimal elsewhere [59]. For generalized linear models with sparsity, we derived a convergence rate that covers a wide range of models, and also proved a novel technical result on restricted strong convexity that holds under relatively mild conditions. In the case of sparse group regularization, we established a novel upper bound of the oracle type, with a separation between the approximation and estimation error terms. For matrix estimation, the framework described here has been used to derive bounds on Frobenius error that are known to be minimax-optimal [52, 54, 68].

There are a variety of interesting open questions associated with our work. In this paper, for simplicity of exposition, we have specified the regularization parameter in terms of the dual norm r^* of the regularizer. In many cases, this choice leads to convergence rates that are known to be minimax-optimal, including linear regression over ℓ_q-balls (Corollary 3) for sufficiently small radii, and various instances of low-rank matrix regression. In other cases, some refinements of our convergence rates are possible; for instance, for the special case of linear sparsity regression (i.e., an exactly sparse vector, with a constant fraction of non-zero elements), our rates are sub-optimal by a logarithmic factor. However, optimal rates can be obtained by a more careful analysis of the noise term, which allows for a slightly smaller choice of the regularization parameter. Similarly, there are other non-parametric settings in which a more delicate choice of the regularization parameter is required [36, 60]. Last, we suspect that there are many other statistical models, not discussed in this paper, for which this framework can yield useful results. Some examples include different types of hierarchical regularizers and/or overlapping group regularizers [29, 30], as well as methods using combinations of decomposable regularizers, such as the fused Lasso [73], and work on matrix decomposition [21, 20].

Acknowledgements

All authors were partially supported by NSF grants DMS-0605165 and DMS-0907632. BY acknowledges additional support from NSF-grant SES-0835531 (CDI); MJW and SN acknowledge additional support from the NSF grant CDI-0941742, and AFOSR grant 09NL184; and PR acknowledges additional support from NSF grant IIS-101842. We thank a number of people, including Arash Amini, Francis Bach, Garvesh Raskutti, Alexandre Tsybakov, and Sara van de Geer, for helpful discussions.

A Proofs related to Theorem 1

In this section, we collect the proofs of Lemma 1 and our main result. All our arguments in this section are deterministic, and so we adopt the shorthand $\mathcal{L}(\theta^*) = \mathcal{L}(\theta^*; Z_1^n)$. For an arbitrary vector $\Delta \in \mathbb{R}^p$, we also use the simpler notation

$$\Delta_A := \Pi_A(\Delta), \quad \Delta_{A^\perp} := \Pi_{A^\perp}(\Delta), \quad \Delta_B := \Pi_B(\Delta), \quad \text{and} \quad \Delta_{B^\perp} := \Pi_{B^\perp}(\Delta).$$

Both proofs make use of the function $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$ given by

$$\mathcal{F}(\Delta) := \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda_n \big\{ r(\theta^* + \Delta) - r(\theta^*) \big\}. \tag{53}$$

In addition, we exploit the following fact: since $\mathcal{F}(0) = 0$, the optimal error $\widehat{\Delta} = \widehat{\theta} - \theta^*$ must satisfy $\mathcal{F}(\widehat{\Delta}) \leq 0$.

A.1 Proof of Lemma 1

Note that the function $\mathcal{F}$ consists of two parts: a difference of loss functions, and a difference of regularizers. In order to control $\mathcal{F}$, we require bounds on these two quantities:

Lemma 3 (Deviation inequalities). *For any decomposable regularizer and p-dimensional vectors θ^* and Δ, we have*

$$r(\theta^* + \Delta) - r(\theta^*) \geq r(\Delta_{B^\perp}) - r(\Delta_B) - 2r(\theta^*_{A^\perp}). \tag{54}$$

Moreover, as long as $\lambda_n \geq 2r^*(\nabla \mathcal{L}(\theta^*))$ and $\mathcal{L}$ is convex, we have

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq -\frac{\lambda_n}{2} \big[r\big(\Delta_B\big) + r\big(\Delta_{B^\perp}\big) \big]. \tag{55}$$

Proof. By triangle inequality, we have

$$\begin{aligned} r\big(\theta^* + \Delta\big) &= r\big(\theta^*_A + \theta^*_{A^\perp} + \Delta_B + \Delta_{B^\perp}\big) \\ &\geq r\big(\theta^*_A + \Delta_{B^\perp}\big) - r\big(\theta^*_{A^\perp} + \Delta_B\big) \\ &\geq r\big(\theta^*_A + \Delta_{B^\perp}\big) - r\big(\theta^*_{A^\perp}\big) - r\big(\Delta_B\big). \end{aligned}$$

By decomposability applied to θ^*_A and $\Delta_{B^\perp}$, we have $r\big(\theta^*_A + \Delta_{B^\perp}\big) = r\big(\theta^*_A\big) + r\big(\Delta_{B^\perp}\big)$, so that

$$r\big(\theta^* + \Delta\big) \geq r\big(\theta^*_A\big) + r\big(\Delta_{B^\perp}\big) - r\big(\theta^*_{A^\perp}\big) - r\big(\Delta_B\big). \tag{56}$$

Similarly, by triangle inequality, we have $r(\theta^*) \leq r\big(\theta^*_A\big) + r\big(\theta^*_{A^\perp}\big)$. Combining this inequality with the bound (56), we obtain

$$\begin{aligned} r\big(\theta^* + \Delta\big) - r(\theta^*) &\geq r\big(\theta^*_A\big) + r\big(\Delta_{B^\perp}\big) - r\big(\theta^*_{A^\perp}\big) - r\big(\Delta_B\big) - \big\{r\big(\theta^*_A\big) + r\big(\theta^*_{A^\perp}\big)\big\} \\ &= r\big(\Delta_{B^\perp}\big) - r\big(\Delta_B\big) - 2r\big(\theta^*_{A^\perp}\big), \end{aligned}$$

which yields the claim (54).

Turning to the loss difference, using the convexity of the loss function $\mathcal{L}$, we have

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle \;\geq\; -|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle|.$$

Applying the (generalized) Cauchy-Schwarz inequality with the regularizer and its dual, we obtain

$$|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle| \leq r^*(\nabla \mathcal{L}(\theta^*))\, r(\Delta) \;\leq\; \frac{\lambda_n}{2}\, \big[r\big(\Delta_B\big) + r\big(\Delta_{B^\perp}\big)\big],$$

where the final equality uses triangle inequality, and the assumed bound $\lambda_n \geq 2r^*(\nabla \mathcal{L}(\theta^*))$. Consequently, we conclude that $\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq -\frac{\lambda_n}{2}\, \big[r\big(\Delta_B\big) + r\big(\Delta_{B^\perp}\big)\big]$, as claimed. □

We can now complete the proof of Lemma 1. Combining the two lower bounds (54) and (55), we obtain

$$\begin{aligned} 0 \;\geq\; \mathcal{F}(\widehat{\Delta}) &\geq \lambda_n \left\{r(\Delta_{B^\perp}) - r(\Delta_B) - 2r(\theta^*_{A^\perp})\right\} - \frac{\lambda_n}{2}\, \big[r\big(\Delta_B\big) + r\big(\Delta_{B^\perp}\big)\big] \\ &= \frac{\lambda_n}{2} \big\{r(\Delta_{B^\perp}) - 3r(\Delta_B) - 4r(\theta^*_{A^\perp})\big\}, \end{aligned}$$

from which the claim follows.

A.2 Proof of Theorem 1

For an error tolerance $\delta > 0$ as given in the theorem, we are concerned with the behavior of the function $\mathcal{F}$ previously defined (53) over the set

$$\mathbb{K}(\delta; A, B, \theta^*) := \big\{\Delta \in \mathbb{R}^p \mid \|\Delta\|_\star = \delta\big\} \cap \mathbb{C}(A, B; \theta^*), \tag{57}$$

where

$$\mathbb{C}(A, B; \theta^*) := \big\{\Delta \in \mathbb{R}^p \mid r\big(\Delta_{B^\perp}\big) \leq 3r\big(\Delta_B\big) + 4r\big(\theta^*_{A^\perp}\big)\big\}. \tag{58}$$

Since the subspace pair (A, B) and true parameter θ^* remain fixed throughout this analysis, we adopt the shorthand notation $\mathbb{K}(\delta)$ and $\mathbb{C}$ for the remainder of this proof.

The following lemma shows that it suffices to control the sign of the function $\mathcal{F}$ over the set (57).

Lemma 4. *If $\mathcal{F}(\Delta) > 0$ for all vectors $\Delta \in \mathbb{K}(\delta)$, then $\|\widehat{\Delta}\|_\star \leq \delta$.*

Proof. We first claim that $\mathbb{C}$ is star-shaped, meaning that if $\widehat{\Delta} \in \mathbb{C}$, then the entire line $\{t\widehat{\Delta} \mid t \in (0,1)\}$ connecting $\widehat{\Delta}$ with the all-zeroes vector is contained with $\mathbb{C}$. This property is immediate whenever $\theta^* \in A$, since $\mathbb{C}$ is then a cone, as illustrated in Figure 1(a). Now consider the general case, when $\theta^* \notin A$. We first observe that for any $t \in (0,1)$,

$$\Pi_B(t\Delta) = \arg\min_{\beta \in B} \|t\Delta - \beta\|_\star \; = \; t \arg\min_{\beta \in B} \|\Delta - \frac{\beta}{t}\|_\star \; = \; t\,\Pi_B(\Delta),$$

using the fact that β/t also belongs to the subspace B. The equality $\Pi_{B^\perp}(t\Delta) = t\Pi_{B^\perp}(\Delta)$ follows similarly. Consequently, for all $\Delta \in \mathbb{C}$, we have

$$r(\Pi_{B^\perp}(t\Delta)) \; = \; r(t\Pi_{B^\perp}(\Delta)) \overset{(i)}{=} \; t\,r(\Pi_{B^\perp}(\Delta)) \; \overset{(ii)}{\leq} \; t\left\{3\,r(\Pi_B(\Delta)) + 4r(\Pi_{A^\perp}(\theta^*))\right\}$$

where step (i) uses the fact that any norm is positive homogeneous,[2] and step (ii) uses the inclusion $\Delta \in \mathbb{C}$. We now observe that $3\,t\,r(\Pi_B(\Delta)) = 3\,r(\Pi_B(t\Delta))$, and that $4t\,r(\Pi_{A^\perp}(\theta^*)) \leq 4r(\Pi_{A^\perp}(\theta^*))$, since $t \in (0,1)$, so that

$$r(\Pi_{B^\perp}(t\Delta)) \leq \; 3\,r(\Pi_B(t\Delta)) + t\,4\Pi_{A^\perp}(\theta^*) \; \leq \; 3\,r(\Pi_B(t\Delta)) + 4r(\Pi_{A^\perp}(\theta^*)).$$

We have thus shown that $t\Delta \in \mathbb{C}$, and hence that $\mathbb{C}$ is star-shaped.

Turning to the lemma itself, we prove the contrapositive statement: in particular, we show that if for some optimal solution $\widehat{\theta}$, the associated error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$ satisfies the inequality $\|\widehat{\Delta}\|_\star > \delta$, then there must be some vector $\widetilde{\Delta} \in \mathbb{K}(\delta)$ such that $\mathcal{F}(\widetilde{\Delta}) \leq 0$. If $\|\widehat{\Delta}\|_\star > \delta$, then the line joining $\widehat{\Delta}$ to 0 must intersect the set $\mathbb{K}(\delta)$ at some intermediate point $t^*\widehat{\Delta}$, for some $t^* \in (0,1)$. Since the loss function $\mathcal{L}$ and regularizer r are convex, the function $\mathcal{F}$ is also convex for any choice of the regularization parameter, so that by Jensen's inequality,

$$\mathcal{F}\big(t^*\Delta\big) \; = \; \mathcal{F}\bigg(t^*\Delta + (1-t^*)\,0\bigg) \leq t^*\,\mathcal{F}(\widehat{\Delta}) + (1-t^*)\mathcal{F}(0) \; \overset{(i)}{=} \; t^*\mathcal{F}(\widehat{\Delta}),$$

where equality (i) uses the fact that $\mathcal{F}(0) = 0$ by construction. But since $\widehat{\Delta}$ is optimal, we must have $\mathcal{F}(\widehat{\Delta}) \leq 0$, and hence $\mathcal{F}(t^*\Delta) \leq 0$ as well. Thus, we have constructed a vector $\widetilde{\Delta} = t^*\Delta$ with the claimed properties, thereby establishing Lemma 4. □

We now exploit Lemma 4 in order to prove Theorem 1. More specifically, it suffices to establish a lower bound on $\mathcal{F}(\Delta)$ over $\mathbb{K}(\delta)$ for the specified critical radius $\delta > 0$. For an arbitrary $\Delta \in \mathbb{K}(\delta)$, we have

$$\begin{aligned}
\mathcal{F}(\Delta) &= \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda_n\big\{r(\theta^* + \Delta) - r(\theta^*)\big\} \\
&\overset{(i)}{\geq} \; \langle \nabla\mathcal{L}(\theta^*), \Delta\rangle + \kappa_{\mathcal{L}}\|\Delta\|_\star^2 + \lambda_n\big\{r(\theta^* + \Delta) - r(\theta^*)\big\} \\
&\overset{(ii)}{\geq} \; \langle \nabla\mathcal{L}(\theta^*), \Delta\rangle + \kappa_{\mathcal{L}}\|\Delta\|_\star^2 + \lambda_n\big\{r(\Delta_{B^\perp}) - r(\Delta_B) - 2r(\theta^*_{A^\perp})\big\},
\end{aligned}$$

[2] Explicitly, for any norm and non-negative scalar t, we have $\|tx\| = t\|x\|$.

where inequality (i) holds because the RSC condition holds over $\mathbb{K}(\delta)$, and inequality (ii) follows from the bound (54).

By the Cauchy-Schwarz inequality applied to the regularizer r and its dual r^*, we have the bound $|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle| \leq r^*(\nabla \mathcal{L}(\theta^*))\, r(\Delta)$. Since $\lambda_n \geq 2r^*(\nabla \mathcal{L}(\theta^*))$ by assumption, we conclude that $|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle| \leq \frac{\lambda_n}{2} r(\Delta)$, and hence that

$$\mathcal{F}(\Delta) \geq \kappa_{\mathcal{L}} \|\Delta\|_\star^2 + \lambda_n \big\{ r(\Delta_{B^\perp}) - r(\Delta_B) - 2r(\theta^*_{A^\perp}) \big\} - \frac{\lambda_n}{2} r(\Delta)$$

By triangle inequality, we have

$$r(\Delta) \;=\; r(\Delta_{B^\perp} + \Delta_B) \;\leq\; r(\Delta_{B^\perp}) + r(\Delta_B),$$

and hence, following some algebra

$$\begin{aligned} \mathcal{F}(\Delta) &\geq \kappa_{\mathcal{L}} \|\Delta\|_\star^2 + \lambda_n \big\{ \frac{1}{2} r(\Delta_{B^\perp}) - \frac{3}{2} r(\Delta_B) - 2r(\theta^*_{A^\perp}) \big\} \\ &\geq \kappa_{\mathcal{L}} \|\Delta\|_\star^2 - \frac{\lambda_n}{2} \big\{ 3r(\Delta_B) + 4r(\theta^*_{A^\perp}) \big\}. \end{aligned} \tag{59}$$

Now by definition (18) of the subspace compatibility, we have the inequality $r(\Delta_B) \leq \Psi(B) \|\Delta_B\|_\star$. Since the projection $\Delta_B = \Pi_B(\Delta)$ is defined in terms of the norm $\|\cdot\|_\star$, it is non-expansive. Since $0 \in B$, we have

$$\|\Delta_B\|_\star \;=\; \|\Pi_B(\Delta) - \Pi_B(0)\|_\star \;\leq\; \|\Delta - 0\|_\star \;=\; \|\Delta\|_\star.$$

Combining with the earlier bound, we conclude that $r(\Delta_B) \leq \Psi(B)\|\Delta\|_\star$. Substituting into the lower bound (59), we obtain

$$\mathcal{F}(\Delta) \geq \kappa_{\mathcal{L}} \|\Delta\|_\star^2 - \frac{\lambda_n}{2} \big\{ 3\Psi(B)\, \|\Delta\|_\star + 4r(\theta^*_{A^\perp}) \big\}.$$

This is a strictly positive definite quadratic form in $\|\Delta\|_\star$, and so will be positive for $\|\Delta\|_\star$ sufficiently large; in particular, following some algebra, we find that it suffices to have $\|\Delta\|_\star \geq \frac{1}{\kappa_{\mathcal{L}}} \big\{ 2\lambda_n \Psi(B) + \sqrt{2\kappa_{\mathcal{L}} \lambda_n r(\theta^*_{A^\perp})} \big\}$. This completes the proof of Theorem 1.

B Proof of Lemma 2

For any Δ in the set $\mathbb{C}(S_\tau)$, we have

$$\|\Delta\|_1 \leq 4\|\Delta_{S_\tau}\|_1 + 4\|\theta^*_{S_\tau^c}\|_1 \leq \sqrt{|S_\tau|}\|\Delta\|_2 + 4R_q\, \tau^{1-q} \;\leq 4\sqrt{R_q}\, \tau^{-q/2}\, \|\Delta\|_2 + 4R_q\, \tau^{1-q},$$

where we have used the bounds (37) and (38). Therefore, for any vector $\Delta \in \mathbb{C}(S_\tau)$, the condition (28) implies that

$$\begin{aligned} \frac{\|X\Delta\|_2}{\sqrt{n}} &\geq\; \kappa_1\, \|\Delta\|_2 - \kappa_2 \sqrt{\frac{\log p}{n}} \big\{ \sqrt{R_q}\, \tau^{-q/2}\, \|\Delta\|_2 + R_q\, \tau^{1-q} \big\} \\ &= \|\Delta\|_2 \bigg\{ \kappa_1 - \kappa_2 \sqrt{\frac{R_q \log p}{n}} \tau^{-q/2} \bigg\} - \kappa_2 \sqrt{\frac{\log p}{n}}\, R_q\, \tau^{1-q}. \end{aligned}$$

By our choices $\tau = \frac{\lambda_n}{\kappa_1}$ and $\lambda_n = 4\,\sigma\sqrt{\frac{\log p}{n}}$, we have $\kappa_2\sqrt{\frac{R_q \log p}{n}}\,\tau^{-q/2} = \frac{\kappa_2}{(8\sigma)^{q/2}}\;\sqrt{R_q}\,\big(\frac{\log p}{n}\big)^{1-\frac{q}{2}}$, which is less than $\kappa_1/2$ under the stated assumptions. Thus, we obtain the lower bound

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \frac{\kappa_1}{2}\Big\{\|\Delta\|_2 - \frac{2\kappa_2}{\kappa_1}\,\sqrt{\frac{\log p}{n}}\,R_q\,\tau^{1-q}\Big\}.$$

Finally, recalling the choice $\delta^* = \frac{4\,\kappa_2}{\kappa_1}\,\sqrt{\frac{\log p}{n}}\,R_q\,\tau^{1-q}$ previously specified (35), then we have

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \frac{\kappa_1}{2}\Big\{\|\Delta\|_2 - \frac{\|\Delta\|_2}{2}\Big\} \;=\; \frac{\kappa_1}{4}\|\Delta\|_2$$

for any $\Delta \in \mathbb{C}(S_\tau)$ with $\|\Delta\|_2 \geq \delta^*$, as claimed.

C Proofs for group-sparse norms

In this section, we collect the proofs of results related to the group-sparse norms in Section 5.

C.1 Proof of Proposition 1

The proof of this result follows similar lines to the proof of condition (28) given by Raskutti et al. [61], hereafter RWY, who established this result in the special case of the ℓ_1-norm. Here we describe only those portions of the proof that require modification. For a radius $t > 0$, define the set

$$V(t) := \big\{\theta \in \mathbb{R}^p \;\mid\; \|\Sigma^{1/2}\theta\|_2 = 1, \|\theta\|_{\mathcal{G},\nu} \leq t\big\},$$

as well as the random variable $M(t;X) := 1 - \inf_{\theta \in V(t)} \frac{\|X\theta\|_2}{\sqrt{n}}$. The argument in Section 4.2 of RWY makes use of the Gordon-Slepian comparison inequality in order to upper bound this quantity. Following the same steps, we obtain the modified upper bound

$$\mathbb{E}[M(t;X)] \leq \frac{1}{4} + \frac{1}{\sqrt{n}}\,\mathbb{E}\big[\max_{j=1,\ldots,T}\|w_{G_j}\|_{\nu^*}\big]\;t,$$

where $w \sim N(0,\Sigma)$. The argument in Section 4.3 uses concentration of measure to show that this same bound will hold with high probability for $M(t;X)$ itself; the same reasoning applies here. Finally, the argument in Section 4.4 of RWY uses a peeling argument to make the bound suitably uniform over choices of the radius t. This argument allows us to conclude that

$$\inf_{\theta \in \mathbb{R}^p} \frac{\|X\theta\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Sigma^{1/2}\theta\|_2 - 9\,\mathbb{E}\big[\max_{j=1,\ldots,T}\|w_{G_j}\|_{\nu^*}\big]\;\|\theta\|_{\mathcal{G},\nu} \quad \text{for all } \theta \in \mathbb{R}^p$$

with probability greater than $1 - c_1\exp(-c_2 n)$. Recalling the definition of $\rho_{\mathcal{G}}(\nu^*)$, we see that in the case $\Sigma = I_{p\times p}$, the claim holds with constants $(\kappa_1, \kappa_2) = (\frac{1}{4}, 9)$. Turning to the case of general Σ, let us define the matrix norm $|||A|||_{\nu^*} := \max_{\|\beta\|_{\nu^*}=1}\|A\beta\|_{\nu^*}$. With this notation, some algebra shows that the claim holds with $\kappa_1 = \frac{1}{4}\lambda_{min}(\Sigma^{1/2})$ and $\kappa_2 = 9\max_{t=1,\ldots,T}|||(\Sigma^{1/2})_{G_t}|||_{\nu^*}$.

C.2 Proof of Corollary 4

In order to prove this claim, we need to verify that Theorem 1 may be applied. Doing so requires defining the appropriate model and perturbation subspaces, computing the compatibility constant, and checking that the specified choice (42) of regularization parameter λ_n is valid. For a given subset $S_{\mathcal{G}} \subseteq \{1, 2, \ldots, T\}$, define the subspaces

$$A(S_{\mathcal{G}}) := \big\{\theta \in \mathbb{R}^p \mid \theta_{G_t} = 0 \quad \text{for all } t \notin S_{\mathcal{G}}\big\}, \quad \text{and} \quad A^{\perp}(S_{\mathcal{G}}) := \big\{\theta \in \mathbb{R}^p \mid \theta_{G_t} = 0 \quad \text{for all } t \in S_{\mathcal{G}}\big\}.$$

As discussed in Example 2, the block norm $\|\cdot\|_{\mathcal{G},\nu}$ is decomposable with respect to these subspaces. Let us compute the regularizer-error compatibility function, as defined in equation (18), that relates the regularizer ($\|\cdot\|_{\mathcal{G},\nu}$ in this case) to the error norm (here the ℓ_2-norm). For any $\Delta \in A(S_{\mathcal{G}})$, we have

$$\|\Delta\|_{\mathcal{G},\nu} = \sum_{t \in S_{\mathcal{G}}} \|\Delta_{G_t}\|_\nu \overset{(a)}{\leq} \sum_{t \in S_{\mathcal{G}}} \|\Delta_{G_t}\|_2 \;\leq\; \sqrt{s}\,\|\Delta\|_2,$$

where inequality (a) uses the fact that $\nu \geq 2$.

Finally, let us check that the specified choice of λ_n satisfies the condition (20). As in the proof of Corollary 2, we have $\nabla\mathcal{L}(\theta^*; Z_1^n) = \frac{1}{n}X^T w$, so that the final step is to compute an upper bound on the quantity

$$r^*(\frac{1}{n}X^T w) = \frac{1}{n}\max_{t=1,\ldots,T} \|(X^T w)_{G_t}\|_{\nu^*}$$

that holds with high probability.

Lemma 5. *Suppose that X satisfies the block column normalization condition* (41)*, and the observation noise is sub-Gaussian* (30)*. Then we have*

$$\mathbb{P}\Big[\max_{t=1,\ldots,T} \|\frac{X_{G_t}^T w}{n}\|_{\nu^*} \geq 2\sigma\big\{\frac{m^{1-1/\nu}}{\sqrt{n}} + \sqrt{\frac{\log T}{n}}\big\}\Big] \leq 2\exp\big(-2\log T\big). \tag{60}$$

Proof. Throughout the proof, we assume without loss of generality that $\sigma = 1$, since the general result can be obtained by rescaling. For a fixed group G of size m, consider the submatrix $X_G \in \mathbb{R}^{n\times m}$. We begin by establishing a tail bound for the random variable $\|\frac{X_G^T w}{n}\|_{\nu^*}$.

Deviations above the mean: For any pair $w, w' \in \mathbb{R}^n$, we have

$$\Big|\|\frac{X_G^T w}{n}\|_{\nu^*} - \|\frac{X_G^T w'}{n}\|_{\nu^*}\Big| \leq \frac{1}{n}\|X_G^T(w - w')\|_{\nu^*} \;=\; \frac{1}{n}\max_{\|\theta\|_\nu = 1} \langle X_G\,\theta, (w - w')\rangle.$$

By definition of the $(\nu \to 2)$ operator norm, we have

$$\frac{1}{n}\|X_G^T(w - w')\|_{\nu^*} \leq \frac{1}{n}\,|\!|\!| X_G |\!|\!|_{\nu\to 2}\,\|w - w'\|_2 \overset{(i)}{\leq} \frac{1}{\sqrt{n}}\|w - w'\|_2,$$

where inequality (i) uses the block normalization condition (41). We conclude that the function $w \mapsto \|\frac{X_G^T w}{n}\|_{\nu^*}$ is a Lipschitz with constant $1/\sqrt{n}$, so that by Gaussian concentration of measure for Lipschitz functions [39], we have

$$\mathbb{P}\Big[\|\frac{X_G^T w}{n}\|_{\nu^*} \geq \mathbb{E}\Big[\|\frac{X_G^T w}{n}\|_{\nu^*}\Big] + \delta\Big] \leq 2 \exp\big(-\frac{n\delta^2}{2}\big) \qquad \text{for all } \delta > 0. \tag{61}$$

Upper bounding the mean: For any vector $\beta \in \mathbb{R}^m$, define the zero-mean Gaussian random variable $Z_\beta = \frac{1}{n}\langle \beta, X_G^T w\rangle$, and note the relation $\|\frac{X_G^T w}{n}\|_{\nu^*} = \max_{\|\beta\|_\nu = 1} Z_\beta$. Thus, the quantity of interest is the supremum of a Gaussian process, and can be upper bounded using Gaussian comparison principles. For any two vectors $\|\beta\|_\nu \leq 1$ and $\|\beta'\|_\nu \leq 1$, we have

$$\begin{aligned}\mathbb{E}\Big[(Z_\beta - Z_{\beta'})^2\Big] \;&=\; \frac{1}{n^2}\|X_G(\beta - \beta')\|_2^2 \;\overset{(a)}{\leq}\; \frac{2}{n}\,\frac{|\!|\!|X_G|\!|\!|_{\nu \to 2}^2}{n}\,\|\beta - \beta'\|_2^2 \\ &\overset{(b)}{\leq}\; \frac{2}{n}\|\beta - \beta'\|_2^2,\end{aligned}$$

where inequality (a) uses the fact that $\|\beta - \beta'\|_\nu \leq \sqrt{2}$, and inequality (b) uses the block normalization condition (41).

Now define a second Gaussian process $Y_\beta = \sqrt{\frac{2}{n}}\,\langle \beta, \varepsilon\rangle$, where $\varepsilon \sim N(0, I_{m \times m})$ is standard Gaussian. By construction, for any pair $\beta, \beta' \in \mathbb{R}^m$, we have $\mathbb{E}\big[(Y_\beta - Y_{\beta'})^2\big] = \frac{2}{n}\|\beta - \beta'\|_2^2 \geq \mathbb{E}[(Z_\beta - Z_{\beta'})^2]$, so that the Sudakov-Fernique comparison principle [39] implies that

$$\mathbb{E}\Big[\|\frac{X_G^T w}{n}\|_{\nu^*}\Big] = \mathbb{E}\Big[\max_{\|\beta\|_\nu = 1} Z_\beta\Big] \;\leq\; \mathbb{E}\Big[\max_{\|\beta\|_\nu = 1} Y_\beta\Big].$$

By definition of Y_β, we have

$$\begin{aligned}\mathbb{E}\Big[\max_{\|\beta\|_\nu = 1} Y_\beta\Big] \;&=\; \sqrt{\frac{2}{n}}\mathbb{E}\big[\|\varepsilon\|_{\nu^*}\big] = \sqrt{\frac{2}{n}}\mathbb{E}\Big[\big(\sum_{j=1}^m |\varepsilon_j|^{\nu^*}\big)^{1/\nu^*}\Big] \\ &\leq \sqrt{\frac{2}{n}}\; m^{1/\nu^*}\big(\mathbb{E}[|\varepsilon_1|^{\nu^*}]\big)^{1/\nu^*},\end{aligned}$$

using Jensen's inequality, and the concavity of the function $f(t) = t^{1/\nu^*}$ for $\nu^* \in [1, 2]$. Finally, we have $\big(\mathbb{E}[|\varepsilon_1|^{\nu^*}]\big)^{1/\nu^*} \leq \sqrt{\mathbb{E}[\varepsilon_1^2]} = 1$ and $1/\nu^* = 1 - 1/\nu$, so that we have shown that

$$\mathbb{E}\Big[\max_{\|\beta\|_\nu = 1} Y_\beta\Big] \leq 2\frac{m^{1-1/\nu}}{\sqrt{n}}. \tag{62}$$

Finally, combining the bound (62) with the concentration statement (61), we obtain

$$\mathbb{P}\Big[\|\frac{X_G^T w}{n}\|_{\nu^*} \geq 2\,\frac{m^{1-1/\nu}}{\sqrt{n}} + \delta\Big] \leq 2\exp\big(-\frac{n\delta^2}{2}\big).$$

We now apply the union bound over all groups, and set $\delta^2 = \frac{4 \log T}{n}$ to conclude that

$$\mathbb{P}\Big[\max_{t=1,\ldots,T} \|\frac{X_{G_t}^T w}{n}\|_{\nu^*} \geq 2\big\{\frac{m^{1-1/\nu}}{\sqrt{n}} + \sqrt{\frac{\log T}{n}}\big\}\Big] \leq 2\exp\big(-2\log T\big),$$

as claimed. □

D Proofs related to generalized linear models

In this section, we collect the proofs of several results concerning generalized linear models.

D.1 Proof of Proposition 2

Applying a second-order Taylor expansion to the negative log likelihood (46) between $\theta^* + \Delta$ and θ^*, we conclude that for some $v \in [0,1]$,

$$\delta\mathcal{L}(\Delta, \theta^*; Z_1^n) := \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \langle \nabla\mathcal{L}(\theta^*), \Delta\rangle = \frac{1}{n}\sum_{i=1}^n \psi''\big(\langle\theta^*, x_i\rangle + v\langle\Delta, x_i\rangle\big)\langle\Delta, x_i\rangle^2. \quad (63)$$

We now lower bound the right-hand side by a suitably truncated version. For a truncation level $\tau > 0$ to be chosen shortly (see equation (67)), define the functions

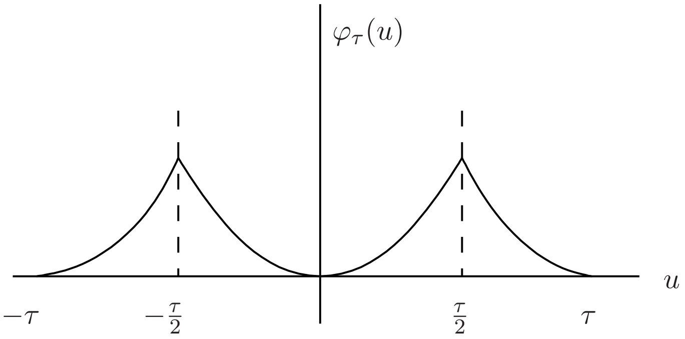

Figure 4. Illustration of truncation function φ_τ. It agrees with u^2 for all $|u| \leq \frac{\tau}{2}$, and then tapers off so that it is zero for all $|u| \geq \tau$. By construction, it is a Lipschitz function.

$$\varphi_\tau(u) = \begin{cases} u^2 & \text{if } |u| \leq \frac{\tau}{2}, \\ (\tau - u)^2 & \text{if } \frac{\tau}{2} \leq |u| \leq \tau\ , \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \alpha_\tau(u) = \begin{cases} u & \text{for } |u| \leq \tau, \\ 0 & \text{otherwise.} \end{cases} \quad (64)$$

Let us focus on a given term $A_i := \psi''\big(\langle\theta^*, x_i\rangle + v\langle\Delta, x_i\rangle\big)\langle\Delta, x_i\rangle^2$. Introducing a second truncation level $T \geq \tau$, consider the event:

$$|\langle\theta^*, x_i\rangle| > T \quad \text{or} \quad |\langle\Delta, x_i\rangle| > \tau. \quad (65)$$

Case 1: If the event (65) holds, then we use the lower bound $A_i \geq 0$.

Case 2: If the event (65) does not hold, then the definitions (64) of the truncation functions ensure that

$$\langle\theta^*, x_i\rangle = \alpha_T(\langle\theta^*, x_i\rangle), \qquad \langle\Delta, x_i\rangle = \alpha_\tau(\langle\Delta, x_i\rangle), \quad \text{and} \quad \langle\Delta, x_i\rangle^2 = \varphi_\tau(\langle\Delta, x_i\rangle),$$

and thus that

$$A_i = \psi''\big(\langle\theta^*, x\rangle + v\langle\Delta, x_i\rangle\big)\langle\Delta, x_i\rangle^2 = \psi''\big(\alpha_T(\langle\theta^*, x_i\rangle) + v\alpha_\tau(\langle\Delta, x_i\rangle)\big)\varphi_\tau\big(\langle\Delta, x_i\rangle\big).$$

In order to combine the two cases, let $\mathbb{I}\left[|\langle\theta^*, x\rangle| \leq T\right]$ be a $\{0, 1\}$-valued indicator function, and note that $\varphi_\tau\big(\langle\Delta, x_i\rangle\mathbb{I}\left[|\langle\theta^*, x_i\rangle| \leq T\right]\big) = 0$ whenever the event (65) holds. With this notation, we have established the lower bound

$$\psi''\big(\langle\theta^*, x\rangle + v\langle\Delta, x_i\rangle\big)\langle\Delta, x_i\rangle^2 \geq \psi''\Bigg(\alpha_T(\langle\theta^*, x_i\rangle) + v\alpha_\tau(\langle\Delta, x_i\rangle)\Bigg)\varphi_\tau\big(\langle\Delta, x_i\rangle\mathbb{I}\left[|\langle\theta^*, x_i\rangle| \leq T\right]\big).$$

Defining the constant $L_\psi(T) := \min\limits_{|u|\leq 2T} \psi''(u)$ and recalling that $T \geq \tau$ by choice, we have $\psi''\big(\alpha_T(\langle\theta^*, x_i\rangle) + v\alpha_\tau(\langle\Delta, x_i\rangle)\big) \geq L_\psi(T)$ for any sample. Consequently, summing over all n samples, we conclude that

$$\delta\mathcal{L}(\Delta, \theta^*; Z_1^n) \geq L_\psi(T) \underbrace{\frac{1}{n}\sum_{i=1}^n \varphi_\tau\big(\langle\Delta, x_i\rangle\mathbb{I}\left[|\langle\theta^*, x_i\rangle| \leq T\right]\big)}_{\widehat{\mathbb{E}}_n\left[\varphi_\tau\left(\langle\Delta, x\rangle\mathbb{I}\left[|\langle\theta^*, x\rangle| \leq T\right]\right)\right]} \tag{66}$$

where we have used $\widehat{\mathbb{E}}_n$ to denote empirical expectation over the n samples. We now define the constant $K_3 := \max\{1,\ 16\kappa_u^2 \log(\frac{24\kappa_u^2}{\kappa_\ell^2})\}$, and for a given ℓ_2-radius $\delta \in (0, 1]$, we choose the truncation parameters as

$$T^2 := K_3 \quad \text{and} \quad \tau^2(\delta) := K_3\delta^2. \tag{67}$$

Based on the lower bound (66), in order to prove the result, it suffices to show that there exist strictly positive constants κ_1 and κ_2, depending only on ψ and the covariance matrix Σ of the covariates, such that for any fixed $\delta \in (0, 1]$, we have

$$\widehat{\mathbb{E}}_n\left[\varphi_{\tau(\delta)}\big(\langle\Delta, x\rangle\mathbb{I}\left[|\langle\theta^*, x\rangle| \leq T\right]\big)\right] \geq \kappa_1\ \|\Delta\|_2\Bigg\{\|\Delta\|_2 - \kappa_2\sqrt{\frac{\log p}{n}}\ \|\Delta\|_1\Bigg\} \quad \forall\ \|\Delta\|_2 = \delta \tag{68}$$

with high probability. In fact, it is enough to prove the claim (68) for $\|\Delta\|_2 = \delta = 1$. To reduce from the general case, suppose that the claim (68) holds for $\delta = 1$. For any vector Δ with $\|\Delta\|_2 = \delta$, we can apply the claim to the rescaled vector $\widetilde{\Delta} = \Delta/\|\Delta\|_2$, thereby obtaining

$$\widehat{\mathbb{E}}_n\left[\varphi_{\tau(1)}\big(\langle\Delta/\|\Delta\|_2, x\rangle\mathbb{I}\left[|\langle\theta^*, x\rangle| \geq T\right]\big)\right] \geq \kappa_1\ \Bigg\{1 - \kappa_2\sqrt{\frac{\log p}{n}}\ \frac{\|\Delta\|_1}{\|\Delta\|_2}\Bigg\}. \tag{69}$$

Note that from the definition (64), for any $c > 0$ and $z \in \mathbb{R}$, we have the homogeneity property $\varphi_c(cz) = c^2\varphi_1(z)$. Multiplying both sides of the bound (69) by $\delta^2 = \|\Delta\|_2^2$ and using this homogeneity property yields the claim.

Consequently, for the remainder of the proof, we restrict our attention to $\delta = 1$, and and the truncation level $\tau = \tau(1) = K_3$. We use $\mathbb{S}_2(1)$ to denote the spherical shell with ℓ_2-radius one, and $\mathbb{S}_1(t) := \{\Delta \in \mathbb{R}^p \mid \|\Delta\|_1 = t\}$ to denote the ℓ_1-sphere of radius t to be chosen. For each $\Delta \in \mathbb{R}^p$, let us introduce the notation[3]

$$f_\Delta(x) := \langle \Delta, x \rangle \, \mathbb{I}\left[|\langle \theta^*, x\rangle| \leq T\right], \quad \text{and} \quad g_\Delta(x) := \varphi_\tau\big(f_\Delta(x)\big). \tag{70}$$

For each radius $t > 0$, define the event

$$\mathcal{E}(t) := \left\{ \widehat{\mathbb{E}}_n\big[g_\Delta(x)\big] < \kappa_1 \left[1 - \kappa_2 \sqrt{\frac{\log p}{n}}\, t\right], \quad \text{for some } \Delta \in \mathbb{S}_1(t) \cap \mathbb{S}_2(1) \right\}.$$

Our goal is to show that the probability of this event is very small, and we do so via the following two steps:

(a) First, we show that with the truncation parameters (67), for any fixed $\Delta \in S_2(1)$, we have

$$\mathbb{E}[g_\Delta(x)] \geq \frac{\kappa_\ell}{2}. \tag{71}$$

(b) Second, defining the random variable $Z(t) := \sup_{\Delta \in \mathbb{S}_2(1) \,\cap\, \mathbb{S}_1(t)} \big|\widehat{\mathbb{E}}_n\big[g_\Delta(x)\big] - \mathbb{E}[g_\Delta(x)]\big|$, we prove the tail bound

$$\mathbb{P}\left[Z(t) \geq \frac{\kappa_\ell}{4} + 50 K_3\, \kappa_u \sqrt{\frac{\log p}{n}}\, t\right] \leq c_1 \exp\big(-c_2 n - c_2' t \log p\big). \tag{72}$$

In conjunction, equations (71) and (72) imply that $\mathbb{P}[\mathcal{E}(t)] \leq c_1 \exp(-c_2 n - c_2' t \log p)$. This result implies the claim of Proposition 2 with slightly different constants c_i by a peeling argument (see Raskutti et al. [61] for details).

Let us first prove the lower bound (71). By assumption (GLM1), for any $\Delta \in \mathbb{S}_2(1)$, we have $\mathbb{E}[\langle \Delta, x\rangle^2] \geq \kappa_\ell \|\Delta\|_2^2 = \kappa_\ell$, so that it suffices to show that

$$\mathbb{E}_x\big[\langle \Delta, x\rangle^2 - g_\Delta(x))\big] \leq \frac{\kappa_\ell}{2}. \tag{73}$$

Note that $g_\Delta(x) \geq 0$ for all x, and $g_\Delta(x) = \langle \Delta, x\rangle^2$ for all x such $|\langle \theta^*, x\rangle| \leq T$ and $|\langle \Delta, x\rangle| \leq \frac{\tau}{2}$. Therefore, by union bound, we have

$$\mathbb{E}_x\big[\langle \Delta, x\rangle^2 - g_\Delta(x))\big] \leq \mathbb{E}_x\big[\langle \Delta, x\rangle^2 \mathbb{I}\left[|\langle \theta^*, x\rangle| \geq T\right]\big] + \mathbb{E}_x\big[\langle \Delta, x\rangle^2 \mathbb{I}\left[|\langle \Delta, x\rangle| \geq \frac{\tau}{2}\right]\big].$$

We bound each of the two terms on the right-hand side in turn. By the Cauchy-Schwarz inequality,

$$\mathbb{E}_x\big[\langle \Delta, x\rangle^2 \mathbb{I}\left[|\langle \theta^*, x\rangle| \geq T\right]\big] \leq \sqrt{\mathbb{E}_x\big[\langle \Delta, x\rangle^4\big]}\, \sqrt{\mathbb{P}[|\langle \theta^*, x\rangle| \geq T]}.$$

[3] Although f_Δ and g_Δ also depend on (θ^*, T) and (θ^*, T, τ), these quantities remain fixed throughout the argument, so that we omit the dependence.

By our assumptions, the variables $\langle \theta^*, x\rangle$ and $\langle \Delta, x\rangle$ are both zero-mean and sub-Gaussian with parameter κ_u^2. Consequently, we have $\mathbb{E}_x\big[\langle \Delta, x\rangle^4\big] \leq 3\kappa_u^4$ and $\mathbb{P}[|\langle \theta^*, x\rangle| \geq T] \leq 2\exp(-\frac{T^2}{2\kappa_u^2})$, and thus

$$\mathbb{E}_x\big[\langle \Delta, x\rangle^2 \mathbb{I}\,[|\langle \theta^*, x\rangle| \geq T]\big] \leq 6\kappa_u^2 \;\; \exp(-\frac{T^2}{4\kappa_u^2}). \tag{74}$$

A similar argument for the other term yields

$$\mathbb{E}_x\big[\langle \Delta, x\rangle^2 \mathbb{I}\,[|\langle \Delta, x\rangle| \geq \frac{\tau}{2}]\big] \leq 6\kappa_u^2 \;\exp(-\frac{\tau^2}{16\,\kappa_u^2}). \tag{75}$$

With $\delta = 1$, the specified choices (67) $\tau^2 = T^2 = K_3$ guarantee that each of these terms is at most $\frac{\kappa_\ell}{4}$, which completes the proof of the lower bound (73), and hence (71).

Now let us prove the tail bound (72). For any $\Delta \in \mathbb{S}_2(1)$, we have $\|g_\Delta\|_\infty \leq \tau^2 = K_3$. Therefore, by the Azuma-Hoeffding inequality, for any $z > 0$, we have $\mathbb{P}[Z(t) \geq \mathbb{E}[Z(t)] + z] \leq 2\exp\big(-\frac{nz^2}{4K_3^2}\big)$. Setting $z^*(t) = \frac{\kappa_\ell}{4} + 2K_3\,\kappa_u\sqrt{\frac{\log p}{n}}\;t$, we obtain

$$\mathbb{P}[Z(t) \geq \mathbb{E}[Z(t)] + z^*(t)]] \leq 2\exp\bigg\{-\frac{n\big(\frac{\kappa_\ell}{4} + 2K_3\,\kappa_u\sqrt{\frac{\log p}{n}}\;t\big)^2}{4K_3^2}\bigg\} \leq 2\exp\big(-\frac{n\,\kappa_\ell^2}{64K_3^2} - \kappa_u^2 t\log p\big). \tag{76}$$

Consequently, in order to establish the claim (72), it suffices to show that

$$\mathbb{E}[Z(t)] \leq 48K_3\;\kappa_u\;t\;\sqrt{\frac{\log p}{n}}. \tag{77}$$

Letting $\{\varepsilon_i\}_{i=1}^n$ be an i.i.d. sequence of Rademacher variables, a standard symmetrization argument [39] yields

$$\mathbb{E}[Z(t)] \leq 2\;\mathbb{E}_{x,\varepsilon}\big[\sup_{\Delta\in\mathbb{S}_2(1)\cap\mathbb{S}_1(t)} \big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g_\Delta(x_i)\big|\big] \;=\; 2\;\mathbb{E}_{x,\varepsilon}\big[\sup_{\Delta\in\mathbb{S}_2(1)\cap\mathbb{S}_1(t)} \big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \varphi_\tau\big(f_\Delta(x_i)\big)\big|\big],$$

where the final step uses the definition (70) of g_Δ. By the definition (64), the function φ_τ is Lipschitz with parameter at most $L = 2\tau = 2\sqrt{K_3} \leq 2K_3$, and $\varphi_\tau(0) = 0$. Therefore, by the Ledoux-Talagrand contraction inequality (see [39], p. 112), we have

$$\begin{aligned}
\mathbb{E}[Z(t)] &\leq 8\;K_3\;\mathbb{E}_{x,\varepsilon}\big[\sup_{\Delta\in\mathbb{S}_2(1)\cap\mathbb{S}_1(t)} \big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_\Delta(x_i)\big|\big] \\
&= 8\;K_3\;\mathbb{E}_{x,\varepsilon}\bigg[\sup_{\Delta\in\mathbb{S}_2(1)\cap\mathbb{S}_1(t)} \big|\big\langle\frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i \mathbb{I}\,[|\langle\theta^*, x_i\rangle| \leq T],\;\Delta\big\rangle\big|\bigg] \\
&\leq 8\;K_3\;t\;\mathbb{E}_{x,\varepsilon}\big\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i \mathbb{I}\,[|\langle\theta^*, x_i\rangle| \leq T]\big\|_\infty.
\end{aligned} \tag{78}$$

Finally, let us upper bound the remaining expectation. For each fixed index $j = 1, \ldots, p$, define the variable $u_{ij} := x_{ij} \mathbb{I}\left[|\langle \theta^*, x_i \rangle| \leq T\right]$, and note that by definition, we have $\frac{1}{n}\sum_{i=1}^n u_{ij}^2 \leq \frac{1}{n}\sum_{i=1}^n x_{ij}^2$. Since the variables $\{x_{ij}, i = 1, 2, \ldots, n\}$ are i.i.d zero-mean sub-Gaussian with parameter at most κ_u^2, standard bounds for the norms of sub-Gaussian vectors yields

$$\mathbb{P}\big[|\frac{1}{n}\sum_{i=1}^n u_{ij}^2| \geq 2\kappa_u^2\big] \leq \mathbb{P}\big[|\frac{1}{n}\sum_{i=1}^n x_{ij}^2| \geq 2\kappa_u^2\big] \;\leq\; 2\exp(-cn).$$

Thus, if we define the event $\mathcal{T} := \big\{|\frac{1}{n}\sum_{i=1}^n u_{ij}^2| \leq 2\kappa_u^2 \;\text{ for all } j = 1, 2, \ldots, p\big\}$, then the union bound implies that

$$\mathbb{P}[\mathcal{T}] \geq 1 - 2\exp(-cn + \log p) \;\overset{(i)}{\geq}\; 1 - 2\exp(-c'n),$$

where inequality (i) follows since $n \gg \log p$. Conditionally on $\mathcal{T}$, the variable $\frac{1}{n}\sum_{i=1}^n \varepsilon_i u_{ij}$ is sub-Gaussian with parameter at most $\frac{2\kappa_u^2}{n}$. Since $\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i u_i\|_\infty$ is the maximum of p such terms, known bounds on expectations of sub-Gaussian maxima (e.g., see Ledoux and Talagrand [39], p. 79) yield $\mathbb{E}\big\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i u_i\big\|_\infty \leq 6\kappa_u\sqrt{\frac{\log p}{n}}$. Combining with the earlier bound (78) yields the claim (77).

D.2 Proof of Corollary 5

For a given support set S, recall the subspaces $A(S)$ and $B(S)$ first defined in Example 1. Recall from our discussion following Proposition 2 that, as long as $n > 9\kappa_2^2\; |S|\log p$, restricted strong convexity with some $\kappa_{\mathcal{L}} > 0$ holds for the pair $A = B$ for all vectors $\Delta \in \mathbb{C}(A; B; \theta^*) \cap \{\|\Delta\|_2 \leq 1\}$. Since $\Pi_{A^\perp}(\theta^*) = 0$, Theorem 1 implies that, for a given valid choice of λ_n— meaning that it satisfies the bound (20)— any solution $\widehat{\theta}_{\lambda_n}$ satisfies the bound

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2 \leq \frac{2\Psi(A(S))}{\kappa_{\mathcal{L}}}\,\lambda_n. \tag{79}$$

As calculated previously, for the combination of ℓ_1-norm as regularizer and ℓ_2-norm as error metric, we have $\Psi(A(S)) = \sqrt{|S|}$.

The only remaining step is to verify that the specified choice of λ_n satisfies the condition (20). We compute the loss gradient $\nabla\mathcal{L}(\theta^*; Z_1^n) = -\frac{1}{n}\sum_{i=1}^n x_i\big(y_i - \psi'(\langle\theta^*, x_i\rangle)\big)$, as before, the dual norm of the ℓ_1-norm is the ℓ_∞-norm. The following result yields the requisite control on the ℓ_∞-norm of this loss gradient:

Lemma 6. *Under conditions (GLM1) and (GLM2), there are universal positive constants* (c_1, c_2) *such that*

$$\mathbb{P}\left[\|\nabla\mathcal{L}(\theta^*; Z_1^n)\|_\infty \geq c_1\sqrt{\frac{\log p}{n}}\right] \leq \frac{c_2}{n}.$$

Combining this claim with the earlier bound (79) yields the claim of Corollary 5.

In order to complete the proof, it remains to prove Lemma 6. For a fixed index $j \in \{1, 2, \ldots, p\}$, we begin by establishing an upper bound on the moment generating function of the sum $\frac{1}{n}\sum_{i=1}^n V_{ij}$, where $V_{ij} := x_{ij}\big(y_i - \psi'(\langle \theta^*, x_i\rangle)\big)$. Let us condition on $\{x_i\}_{i=1}^n$ to start, so that y_i is drawn from the exponential family with parameter $\langle \theta^*, x_i\rangle$. For any $t \in \mathbb{R}$, we compute the cumulant function

$$\begin{aligned}\log \mathbb{E}[\exp(tV_{ij}) \mid x_i] &= \log\big\{\mathbb{E}[\exp(tx_{ij}y_i)]\exp(-tx_{ij}\psi'(\langle\theta^*, x_i\rangle))\big\} \\ &= \psi(tx_{ij} + \langle\theta^*, x_i\rangle) - \psi(\langle\theta^*, x_i\rangle) - \psi'(\langle\theta^*, x_i\rangle)\,\big(tx_{ij}\big).\end{aligned}$$

Consequently, by second-order Taylor series expansion, we have

$$\log \mathbb{E}[\exp(tV_{ij}) \mid x_i] = \frac{t^2}{2}\, x_{ij}^2 \psi''(\langle\theta^*, x_i\rangle + v_i\, tx_{ij}) \quad \text{for some } v_i \in [0,1].$$

Since this upper bound holds for each $i = 1, 2, \ldots, n$, we have shown that

$$\frac{1}{n}\sum_{i=1}^n \log \mathbb{E}[\exp(tV_{ij}) \mid x_i] \le \frac{t^2}{2}\bigg\{\frac{1}{n}\sum_{i=1}^n x_{ij}^2 \psi''\big(\langle\theta^*, x_i\rangle + v_i\, tx_{ij}\big)\bigg\}. \tag{80}$$

At this point, we split the analysis into two cases, depending on whether condition (i) or (ii) holds in assumption (GLM2) .
Case (i): Using the uniform bound on the second derivative, we obtain

$$\frac{1}{n}\sum_{i=1}^n \log \mathbb{E}[\exp(tV_{ij}) \mid x_i] \le \frac{t^2\|\psi''\|_\infty}{2}\{\frac{1}{n}\sum_{i=1}^n x_{ij}^2\}.$$

For each index j, the variables $\{x_{ij}\}_{i=1}^n$ are i.i.d., zero-mean and sub-Gaussian with parameter at most κ_u (by assumption (GLM1)). Consequently, we have $\mathbb{E}[x_{ij}^2] \le \kappa_u^2$. Since the squared variables are sub-exponential, and we have the tail bound

$$\mathbb{P}[\frac{1}{n}\sum_{i=1}^n x_{ij}^2 \ge 2\kappa_u^2] \le 2\exp(-n/4). \tag{81}$$

Now define the event $\mathcal{E} = \big\{\max\limits_{j=1,\ldots,p} \frac{1}{n}\sum_{i=1}^n x_{ij}^2 \le 2\kappa_u^2\big\}$. By combining union bound with the tail bound (81), we have

$$\mathbb{P}[\mathcal{E}^c] \le 2\exp(-n/2 + \log p) \;\le 2\exp(-cn), \tag{82}$$

where we have used the fact that $n \gg \log p$. Conditioned on the event $\mathcal{E}$, we have

$$\frac{1}{n}\sum_{i=1}^n \log \mathbb{E}[\exp(tV_{ij}) \mid x_i] \le t^2\|\psi''\|_\infty \kappa_u^2 \qquad \text{for each } j = 1, 2, \ldots, p.$$

By the Chernoff bound, we obtain $\mathbb{P}\big[|\frac{1}{n}\sum_{i=1}^n V_{ij}| \ge t \mid \mathcal{E}\big] \le 2\exp\big(-n\frac{t^2}{\|\psi''\|_\infty\kappa_u^2}\big)$. Combining this bound with the union bound yields

$$\mathbb{P}[\max_{j=1,\ldots,p} |\frac{1}{n}\sum_{i=1}^n V_{ij}| \ge \delta \mid \mathcal{E}] \le 2\exp\big(-n\frac{\delta^2}{\|\psi''\|_\infty\kappa_u^2} + \log p\big). \tag{83}$$

Setting $\delta^* = \sqrt{2\|\psi''\|_\infty \kappa_u^2 \frac{\log p}{n}}$, and putting together the pieces yields

$$\mathbb{P}\Big[\max_{j=1,\ldots,p} |\frac{1}{n}\sum_{i=1}^n V_{ij}| \geq \delta^*] \leq \mathbb{P}[\mathcal{E}^c] + \mathbb{P}\Big[\max_{j=1,\ldots,p} |\frac{1}{n}\sum_{i=1}^n V_{ij}| \geq \delta^* \mid \mathcal{E}\Big]$$
$$\leq c_1 \exp(-c_2 \log p),$$

where the final inequality combines the bounds (82) and (83). Since $p \gg n$ in the regime of interest, the claim follows.

Case (ii): In this case, we assume that the model c satisfies conditions GLM (ii). Let us take $|t| \leq 1$ in our calculations above. Since $|x_{ij}| \leq 1$ by assumption, from the bound (80), we have

$$\frac{1}{n}\sum_{i=1}^n \log \mathbb{E}[\exp(tV_{ij}) \mid x_i] \leq \frac{t^2}{2}\Big\{\frac{1}{n}\sum_{i=1}^n \underbrace{\max_{|z|\leq 1} \psi''\big(\langle\theta^*, x_i\rangle + z\big)}_{U_i}\Big\}.$$

For a constant c_0, define the event $\mathcal{E}'(c_0) := \big\{|\frac{1}{n}\sum_{i=1}^n U_i| \leq c_0^2\big\}$. By the condition GLM (ii), the random variables U_i are non-negative, i.i.d., and have a bounded α^{th} moment. Therefore, by Chebyshev's inequality, as long the constant c_0 is chosen sufficiently large, there is a positive constant c_1 such that $\mathbb{P}[\mathcal{E}'(c_0)] \geq 1 - c_1/n^\alpha$. From the upper bound (80), we conclude that

$$\frac{1}{n}\sum_{i=1}^n \log \mathbb{E}\big[\exp(tV_{ij}) \mid \mathcal{E}'(c_0)\big] \leq \frac{c_0^2 t^2}{2} \qquad \text{for all } |t| \leq 1.$$

Consequently, for any $\delta > 0$, the Chernoff bound implies that

$$\mathbb{P}\big[|\frac{1}{n}\sum_{i=1}^n U_{ij}| \geq \delta \mid \mathcal{E}'(c_0)\big] \leq 2\exp\Big\{n\big(\frac{c_0^2 t^2}{2} - t\delta\big)\Big\} \quad \text{for all } |t| \leq 1.$$

For any $\delta \in [0, c_0^2]$, we may set $t = \frac{\delta}{c_0^2} \leq 1$ to obtain $\mathbb{P}\big[|\frac{1}{n}\sum_{i=1}^n U_{ij}| \geq \delta\, \mathcal{E}'(c_0)\big] \leq 2\exp\big(-\frac{n\delta^2}{2c_0^2}\big)$. Since this bound holds for each $j = 1, \ldots, p$, the union bound implies that

$$\mathbb{P}\big[\|\nabla\mathcal{L}(\theta^*; Z_1^n)\|_\infty \geq \delta \mid \mathcal{E}'(c_0)\big] \leq 2\exp\big(-\frac{n\delta^2}{2c_0^2} + \log p) \qquad \text{for all } \delta \in [0, c_0^2].$$

We now set $\delta = 2c_0\sqrt{\frac{\log p}{n}}$; this choice is less than c_0^2 since $n \gg \log p$. With this choice, we obtain

$$\mathbb{P}\Big[\|\nabla\mathcal{L}(\theta^*; Z_1^n)\|_\infty \geq 2c_0\sqrt{\frac{\log p}{n}} \mid \mathcal{E}'\Big] \leq 2\exp\big(-2\log p) \;\leq\; \frac{2}{n},$$

where the final inequality holds since $p \gg n$ in the regime of interest.

References

[1] Large Synoptic Survey Telescope, 2003. URL: `www.lsst.org`.

[2] F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[3] F. Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, June 2008.

[4] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[5] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. Technical report, Rice University, 2008. Available at arxiv:0808.3572.

[6] P. J. Bickel, J. B. Brown, H. Huang, and Q. Li. An overview of recent developments in genomics and associated statistical methods. *Phil. Trans. Royal Society A*, 367:4313–4337, 2009.

[7] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008.

[8] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[9] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

[10] L.D. Brown. *Fundamentals of statistical exponential families*. Institute of Mathematical Statistics, Hayward, CA, 1986.

[11] F. 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.

[12] F. Bunea, Y. She, and M. Wegkamp. Adaptive rank penalized estimators in multivariate regression. Technical report, Florida State, 2010. available at arXiv:1004.2995.

[13] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for gaussian regression. *Annals of Statistics*, 35(4):1674–1697, 2007.

[14] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, pages 169–194, 2007.

[15] T. Cai and H. Zhou. Optimal rates of convergence for sparse covariance matrix estimation. Technical report, Wharton School of Business, University of Pennsylvania, 2010. available at http://www-stat.wharton.upenn.edu/ tcai/paper/html/Sparse-Covariance-Matrix.html.

[16] E. Candès and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. Technical Report arXiv:1001.0339v1, Stanford, January 2010.

[17] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Info Theory*, 51(12):4203–4215, December 2005.

[18] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

[19] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[20] E. J. Candes, Y. Ma X. Li, and J. Wright. Stable principal component pursuit. In *International Symposium on Information Theory*, June 2010.

[21] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. Technical report, MIT, June 2009. Available at `arXiv:0906.2220v1`.

[22] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, 20(1):33–61, 1998.

[23] D. L. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, April 2006.

[24] D. L. Donoho and J. M. Tanner. Neighborliness of randomly-projected simplices in high dimensions. *Proceedings of the National Academy of Sciences*, 102(27):9452–9457, 2005.

[25] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford, 2002. Available online: http://faculty.washington.edu/mfazel/thesis-final.pdf.

[26] V. L. Girko. *Statistical analysis of observations of increasing dimension*. Kluwer Academic, New York, 1995.

[27] E. Greenshtein and Y. Ritov. Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli*, 10:971–988, 2004.

[28] J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

[29] L. Jacob, G. Obozinski, and J. P. Vert. Group Lasso with Overlap and Graph Lasso. In *International Conference on Machine Learning (ICML)*, pages 433–440, 2009.

[30] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. Technical report, HAL-Inria, 2010. available at inria-00516723.

[31] S. M. Kakade, O. Shamir, K. Sridharan, and A. Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *AISTATS*, 2010.

[32] N. El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Annals of Statistics*, 36(6):2717–2756, 2008.

[33] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from few entries. Technical report, Stanford, January 2009. Preprint available at http://arxiv.org/abs/0901.3150.

[34] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. Technical report, Stanford, June 2009. Preprint available at http://arxiv.org/abs/0906.2027v1.

[35] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Proceedings of COLT*, 2008.

[36] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. Technical report, Georgia Tech., April 2010.

[37] C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37:4254–4278, 2009.

[38] D. Landgrebe. Hyperspectral image data analsysis as a high-dimensional signal processing problem. *IEEE Signal Processing Magazine*, 19(1):17–28, January 2008.

[39] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

[40] K. Lee and Y. Bresler. Guaranteed minimum rank approximation from linear observations by nuclear norm minimization with an ellipsoidal constraint. Technical report, UIUC, 2009. Available at arXiv:0903.4742.

[41] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. Technical Report UILU-ENG-09-2214, Univ. Illinois, Urbana-Champaign, July 2009.

[42] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm optimization with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.

[43] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. Technical Report arXiv:0903.1468, ETH Zurich, March 2009.

[44] M. Lusting, D. Donoho, J. Santos, and J. Pauly. Compressed sensing MRI. *IEEE Signal Processing Magazine*, 27:72–82, March 2008.

[45] R. Mazumber, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. Technical report, Stanford, July 2009.

[46] M. L. Mehta. *Random matrices*. Academic Press, New York, NY, 1991.

[47] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70:53–71, 2008.

[48] L. Meier, S. van de Geer, and P. Buhlmann. High-dimensional additive modeling. *Annals of Statistics*, 37:3779–3821, 2009.

[49] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[50] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2009.

[51] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

[52] S. Negahban and M. J. Wainwright. Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. Technical report, UC Berkeley, August 2010.

[53] S. Negahban and M. J. Wainwright. Simultaneous support recovery in high-dimensional regression: Benefits and perils of $\ell_{1,\infty}$-regularization. *IEEE Transactions on Information Theory*, 2010. To appear.

[54] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, To appear. Originally posted as `http://arxiv.org/abs/0912.5100`.

[55] Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 76, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 2007.

[56] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming A*, 120(1):261–283, 2009.

[57] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Union support recovery in high-dimensional multivariate regression. *Annals of Statistics*, 2010. To appear.

[58] L. A. Pastur. On the spectrum of random matrices. *Theoretical and Mathematical Physics*, 10:67–74, 1972.

[59] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. Technical Report arXiv:0910.2042, UC Berkeley, Department of Statistics, 2009.

[60] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. Technical Report http://arxiv.org/abs/1008.3654, UC Berkeley, Department of Statistics, August 2010.

[61] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue conditions for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, August 2010.

[62] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SpAM: sparse additive models. *Journal of the Royal Statistical Society, Series B*, 2010. To appear.

[63] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

[64] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation: Convergence rates of ℓ_1-regularized log-determinant divergence. Technical report, Department of Statistics, UC Berkeley, September 2008.

[65] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 2010. Posted as arXiv:0910.0651v2.

[66] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, Vol 52(3):471–501, 2010.

[67] B. Recht, W. Xu, and B. Hassibi. Null space conditions and thresholds for rank minimization. Technical report, U. Madison, 2009. Available at http://pages.cs.wisc.edu/ brecht/papers/10.RecXuHas.Thresholds.pdf.

[68] A. Rohde and A. Tsybakov. Estimation of high-dimensional low-rank matrices. Technical Report arXiv:0912.5338v2, Universite de Paris, January 2010.

[69] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[70] N. Srebro, N. Alon, and T. S. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2005.

[71] M. Stojnic, F. Parvaresh, and B. Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Transactions on Signal Processing*, 57(8):3075–3085, 2009.

[72] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[73] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *J. R. Statistical Soc. B*, 67:91–108, 2005.

[74] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. *Signal Processing*, 86:572–602, April 2006. Special issue on ”Sparse approximations in signal and image processing”.

[75] B. Turlach, W.N. Venables, and S.J. Wright. Simultaneous variable selection. *Technometrics*, 27:349–363, 2005.

[76] S. van de Geer. The deterministic lasso. In *Proc. of Joint Statistical Meeting*, 2007.

[77] S. van de Geer and P. Buhlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[78] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36:614–645, 2008.

[79] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38:49–95, 1996.

[80] M. J. Wainwright. Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Info. Theory*, 55:5728–5741, December 2009.

[81] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55:2183–2202, May 2009.

[82] E. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62:548–564, 1955.

[83] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal Of The Royal Statistical Society Series B*, 69(3):329–346, 2007.

[84] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 1(68):49, 2006.

[85] C. H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

[86] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.

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

[88] S. Zhou. Restricted eigenvalue conditions on subgaussian random matrices. Technical report, Department of Mathematics, ETH Zürich, December 2009.

[89] S. Zhou, J. Lafferty, and L. Wasserman. Time-varying undirected graphs. In *21st Annual Conference on Learning Theory (COLT)*, Helsinki, Finland, July 2008.