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

Sahand Negahban
Department of EECS
UC Berkeley
sahand_n@eecs.berkeley.edu

Pradeep Ravikumar
Department of Computer Sciences
UT Austin
pradeepr@cs.utexas.edu

Martin J. Wainwright
Department of Statistics
Department of EECS
UC Berkeley
wainwrig@eecs.berkeley.edu

Bin Yu
Department of Statistics
Department of EECS
UC Berkeley
binyu@stat.berkeley.edu

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 $p/n \to 0$, a line of recent work has studied models with various types of 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 goodness-of-fit of the models to the data) with some regularization function that encourages the assumed structure. The goal of this paper is to provide a unified framework for establishing consistency and convergence rates for such regularized M-estimation procedures under high-dimensional scaling. We state one main theorem and show how it can be used to re-derive several existing results, and also to obtain several 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 the corresponding regularized M-estimators have fast convergence rates.

1 Introduction

In many fields of science and engineering such as genomics and natural language processing, it is of great interest to relate predictor variables (e.g. gene levels) to a response variable (e.g. cancer status). Due to the exploding size of problems, we often find ourselves in the "large p small n" regime—that is, the number of predictor variables p is comparable to or even larger than the number of observations n. For such high dimensional data, successful statistical modeling is possible only if the data follows models with restrictions. For instance, the data might be sparse in a suitably chosen basis, could lie on some manifold, or the dependencies among the variables might have Markov structure specifid by a graphical model.

In such settings, a common approach is to use *regularized M-estimators*, where some loss function (e.g., the negative log-likelihood of the data) is regularized by a function appropriate to the assumed structure. Such estimators may also be interpreted from a Bayesian perspective as the Maximum A Posterior (MAP) estimator, with the regularizer reflecting prior information. In this paper, we study such regularized M-estimation procedures, and attempt to provide a unifying framework that both recovers some existing results and provides new results on consistency and convergence rates under high-dimensional scaling.

As an illustration of the applications of our analysis, we work with three running examples of constrained parametric structures. The first are *sparse* models, both where the number of model parameters that are non-zero is small (*hard-sparse*) or more generally where the number of parameters above a certain threshold are limited (*weak-sparse*). The second are so called block-sparse models, where the parameters are matrix-structured, and entire rows are either zero or not. Our third class is the estimation of low-rank matrices, which arises in system identification, collaborative filtering, and other types of matrix completion problems.

To motivate the need for a unified analysis, let us provide a brief (and hence necessarily incomplete) overview of the broad range of work on high-dimensional models. For the case of sparse regression, a popular regularizer is the ℓ_1 norm of the parameter vector, which is the sum of the absolute values of the parameters. A number of researchers have studied the Lasso [15, 3] as well as the closely related Dantzig selector [2] and provided conditions on various aspects of its behavior, including ℓ_2-error bounds [6, 1, 20, 2] and model selection consistency [21, 19, 5, 16]. For generalized linear models (GLMs) and exponential family models, estimators based on ℓ_1-regularized maximum likelihood have also been studied, including results on risk consistency [18] and model selection consistency [11]. A body of work has focused on the case of estimating Gaussian graphical models, including convergence rates in Frobenius and operator norm [14], and results on operator norm and model selection consistency [12]. Motivated by inference problems involving block-sparse matrices, other researchers have proposed block-structured regularizers [17, 22], and more recently, high-dimensional consistency results have been obtained for model selection [7, 8] and parameter consistency [4]. In this paper, we derive a single main theorem, and show how we are able to rederive a wide range of known results on high-dimensional consistency, as well as some novel ones: such as estimation error rates for low-rank matrices, sparse matrices, and "weakly"-sparse vectors.

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. In particular, we define a notion of *decomposability* for regularizing functions r, and then prove that when it is satisfied, the error $\widehat{\Delta} = \widehat{\theta} - \theta^*$ of the regularized M-estimator must satisfy certain constraints. We use these constraints to define a notion of *restricted strong convexity* that the loss function must satisfy.

2.1 Problem set-up

Consider a random variable Z with distribution $\mathbb{P}$ taking values in a set $\mathcal{Z}$. Let $Z_1^n := \{Z_1, \ldots, Z_n\}$ denote n observations drawn in an i.i.d. manner from $\mathbb{P}$, and suppose $\theta^* \in \mathbb{R}^p$ is some parameter of this distribution. We consider the problem of estimating θ^* from the data Z_1^n. In order to do so, we consider the following class of regularized M-estimators. Let $\mathcal{L} : \mathbb{R}^p \times \mathcal{Z}^n \mapsto \mathbb{R}$ be some loss function that assigns a cost to any parameter $\theta \in \mathbb{R}^p$ given a set of observations. Let $r : \mathbb{R}^p \mapsto \mathbb{R}$ denote a regularization function. We then consider the regularized M-estimator given by

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

where $\lambda_n > 0$ is a regularization penalty. For ease of notation, in the sequel, we adopt the shorthand $\mathcal{L}(\theta)$ for $\mathcal{L}(\theta; Z_1^n)$. Throughout the paper, we assume that the loss function $\mathcal{L}$ is convex and differentiable, and that the regularizer r is a norm.

Our goal is to provide general techniques for deriving bounds on the error $\widehat{\theta} - \theta^*$ in some error metric d. A common example is the ℓ_2-norm $d(\widehat{\theta} - \theta^*) := \|\widehat{\theta} - \theta^*\|_2$. As discussed earlier, high-dimensional parameter estimation is made possible by structural constraints on θ^* such as sparsity, and we will see that the behavior of the error is determined by how well these constraints are captured by the regularization function $r(\cdot)$. We now turn to the properties of the regularizer r and the loss function $\mathcal{L}$ that underlie our analysis.

2.2 Decomposability

Our first condition requires that the regularization function r be decomposable, in a sense to be defined precisely, with respect to a family of subspaces. This notion is a formalization of the manner in which the regularization function imposes constraints on possible parameter vectors $\theta^* \in \mathbb{R}^p$. We begin with some abstract definitions, which we then illustrate with a number of concrete examples. Take some arbitrary inner product space $\mathcal{H}$, and let $\|\cdot\|_2$ denote the norm induced by the inner product. Consider a pair (A, B) of subspaces of $\mathcal{H}$ such that $A \subseteq B^\perp$. For a given subspace A and vector $u \in \mathcal{H}$, we let $\pi_A(u) := \operatorname{argmin}_{v \in A} \|u - v\|_2$ denote the orthogonal projection of u onto A. We let $\mathcal{V} = \{(A, B) \mid A \subseteq B^\perp\}$ be a collection of subspace pairs. For a given statistical model, our goal is to construct subspace collections $\mathcal{V}$ such that for any given θ^* from our model class, there exists a pair $(A, B) \in \mathcal{V}$ with $\|\pi_A(\theta^*)\|_2 \approx \|\theta^*\|_2$, and $\|\pi_B(\theta^*)\|_2 \approx 0$. Of most interest to us are subspace pairs (A, B) in which this property holds but the subspace A is relatively small and B is relatively large. Note that A represents the constraints underlying our model class, and imposed by our regularizer. In the remainder of this paper we assume that $\mathcal{H} = \mathbb{R}^p$ and use the standard Euclidean innerproduct, unless otherwise specified.

As a first concrete (but toy) example, consider the model class of all vectors $\theta^* \in \mathbb{R}^p$, and the subspace collection $\mathcal{T}$ that consists of a single subspace pair $(A, B) = (\mathbb{R}^p, 0)$. We refer to this choice ($\mathcal{V} = \mathcal{T}$) as the *trivial subspace collection*. In this case, for any $\theta^* \in \mathbb{R}^p$, we have $\pi_A(\theta^*) = \theta^*$ and $\pi_B(\theta^*) = 0$. Although this collection satisfies our desired property, it is not so useful since $A = \mathbb{R}^p$ is a very large subspace. As a second example consider the class of s-sparse parameter vectors $\theta^* \in \mathbb{R}^p$, meaning that $\theta^*_i \neq 0$ only if $i \in S$, where S is some s-sized subset of $\{1, 2, \ldots, p\}$. For any given subset S and its complement S^c, let us define the subspaces

$$A(S) = \{\theta \in \mathbb{R}^p \mid \theta_{S^c} = 0\}, \quad \text{and} \quad B(S) = \{\theta \in \mathbb{R}^p \mid \theta_S = 0\},$$

and the *s-sparse subspace collection* $\mathcal{S} = \{(A(S), B(S)) \mid S \subset \{1, \ldots, p\}, \ |S| = s\}$. With this set-up, for any s-sparse parameter vector θ^*, we are guaranteed that there exists some $(A, B) \in \mathcal{S}$ such that $\pi_A(\theta^*) = \theta^*$ and $\pi_B(\theta^*) = 0$. In this case, the property is more interesting, since the subspaces $A(S)$ are relatively small as long as $|S| = s \ll p$.

With this set-up, we say that the regularizer r is *decomposable* with respect to a given subspace pair (A, B) if

$$r(u + z) = r(u) + r(z) \quad \text{for all } u \in A \text{ and } z \in B. \tag{2}$$

In our subsequent analysis, we impose the following condition on the regularizer:

Definition 1. The regularizer r is decomposable with respect to a given subspace collection $\mathcal{V}$, meaning that it is decomposable for each subspace pair $(A, B) \in \mathcal{V}$.

Note that any regularizer is decomposable with respect to the trivial subspace collection $\mathcal{T} = \{(\mathbb{R}^p, 0)\}$. It will be of more interest to us when the regularizer decomposes with respect to a larger collection $\mathcal{V}$ that includes subspace pairs (A, B) in which A is relatively small and B is relatively large. Let us illustrate with some examples.

- *Sparse vectors and ℓ_1 norm regularization*. Consider a model involving s-sparse regression vectors $\theta^* \in \mathbb{R}^p$, and recall the definition of the s-sparse subspace collection $\mathcal{S}$ discussed above. We claim that the ℓ_1-norm regularizer $r(u) = \|u\|_1$ is decomposable with respect to $\mathcal{S}$. Indeed, for any s-sized subset S and vectors $u \in A(S)$ and $v \in B(S)$, we have $\|u + v\|_1 = \|u\|_1 + \|v\|_1$, as required.

- *Group-structured sparse matrices and $\ell_{1,q}$ matrix norms*. Various statistical problems involve matrix-valued parameters $\Theta \in \mathbb{R}^{k \times m}$; examples include multivariate regression problems or (inverse) covariance matrix estimation. We can define an inner product on such matrices via $\langle\langle \Theta, \ \Sigma \rangle\rangle = \operatorname{trace}(\Theta^T \Sigma)$ and the induced (Frobenius) norm $\sum_{i=1}^k \sum_{j=1}^m \Theta_{i,j}^2$. Let us suppose that Θ satisfies a group sparsity condition, meaning that the i^{th} row, denoted Θ_i, is non-zero only if $i \in S \subseteq \{1, \ldots, k\}$ and the cardinality of S is controlled. For a given subset S, we can define the subspace pair

$$B(S) = \big\{\Theta \in \mathbb{R}^{k \times m} \mid \Theta_i = 0 \quad \text{for all } i \in S^c\big\}, \qquad \text{and} \quad A(S) = (B(S))^\perp,$$

For some fixed $s \leq k$, we then consider the collection

$$\mathcal{V} = \{(A(S), B(S)) \mid S \subset \{1, \ldots, k\}, \, |S| = s\},$$

which is a group-structured analog of the s-sparse set $\mathcal{S}$ for vectors. For any $q \in [1, \infty]$, now suppose that the regularizer is the ℓ_1/ℓ_q matrix norm, given by $r(\Theta) = \sum_{i=1}^{k} [\sum_{j=1}^{m} |\Theta_{ij}|^q]^{1/q}$, corresponding to applying the ℓ_q norm to each row and then taking the ℓ_1-norm of the result. It can be seen that the regularizer $r(\Theta) = |||\Theta|||_{1,q}$ is decomposable with respect to the collection $\mathcal{V}$.

- *Low-rank matrices and nuclear norm.* The estimation of low-rank matrices arises in various contexts, including principal component analysis, spectral clustering, collaborative filtering, and matrix completion. In particular, consider the class of matrices $\Theta \in \mathbb{R}^{k \times m}$ that have rank $r \leq \min\{k, m\}$. For any given matrix Θ, we let $\mathrm{row}(\Theta) \subseteq \mathbb{R}^m$ and $\mathrm{col}(\Theta) \subseteq \mathbb{R}^k$ denote its row space and column space respectively. For a given pair of r-dimensional subspaces $U \subseteq \mathbb{R}^k$ and $V \subseteq \mathbb{R}^m$, we define a pair of subspaces $A(U, V)$ and $B(U, V)$ of $\mathbb{R}^{k \times m}$ as follows:

$$A(U, V) := \big\{\Theta \in \mathbb{R}^{k \times m} \mid \mathrm{row}(\Theta) \subseteq V, \ \mathrm{col}(\Theta) \subseteq U\big\}, \quad \text{and} \tag{3a}$$

$$B(U, V) := \big\{\Theta \in \mathbb{R}^{k \times m} \mid \mathrm{row}(\Theta) \subseteq V^{\perp}, \ \mathrm{col}(\Theta) \subseteq U^{\perp}\big\}. \tag{3b}$$

Note that $A(U, V) \subseteq B^{\perp}(U, V)$, as is required by our construction. We then consider the collection $\mathcal{V} = \{(A(U, V), B(U, V)) \mid U \subseteq \mathbb{R}^k, \ V \subseteq \mathbb{R}^m\}$, where (U, V) range over all pairs of r-dimensional subspaces. Now suppose that we regularize with the nuclear norm $r(\Theta) = |||\Theta|||_1$, corresponding to the sum of the singular values of the matrix Θ. It can be shown that the nuclear norm is decomposable with respect to $\mathcal{V}$. Indeed, since any pair of matrices $M \in A(U, V)$ and $M' \in B(U, V)$ have orthogonal row and column spaces, we have $|||M + M'|||_1 = |||M|||_1 + |||M'|||_1$ (e.g., see the paper [13]).

Thus, we have demonstrated various models and regularizers in which decomposability is satisfied with interesting subspace collections $\mathcal{V}$. We now show that decomposability has important consequences for the error $\widehat{\Delta} = \widehat{\theta} - \theta^*$, where $\widehat{\theta} \in \mathbb{R}^p$ is any optimal solution of the regularized M-estimation procedure (1). In order to state a lemma that captures this fact, we need to define the dual norm of the regularizer, given by $r^*(v) := \sup_{u \in \mathbb{R}^p} \frac{u^T v}{r(u)}$. For the regularizers of interest, the dual norm can be obtained via some easy calculations. For instance, given a vector $\theta \in \mathbb{R}^p$ and $r(\theta) = \|\theta\|_1$, we have $r^*(\theta) = \|\theta\|_\infty$. Similarly, given a matrix $\Theta \in \mathbb{R}^{k \times m}$ and the nuclear norm regularizer $r(\Theta) = |||\Theta|||_1$, we have $r^*(\Theta) = |||\Theta|||_2$, corresponding to the operator norm (or maximal singular value).

Lemma 1. *Suppose $\widehat{\theta}$ is an optimal solution of the regularized M-estimation procedure* (1)*, with associated error $\Delta = \widehat{\theta} - \theta^*$. Furthermore, suppose that the regularization penalty is strictly positive with $\lambda_n \geq 2\, r^*(\nabla \mathcal{L}(\theta^*))$. Then for any $(A, B) \in \mathcal{V}$*

$$r(\pi_B(\widehat{\Delta})) \ \leq \ 3r(\pi_{B^{\perp}}(\widehat{\Delta})) + 4r(\pi_{A^{\perp}}(\theta^*)).$$

This property plays an essential role in our definition of restricted strong convexity and subsequent analysis.

2.3 Restricted Strong Convexity

Next we state our assumption on the loss function $\mathcal{L}$. In general, guaranteeing that $\mathcal{L}(\widehat{\theta}) - \mathcal{L}(\theta^*)$ is small is *not sufficient* to show that $\widehat{\theta}$ and θ^* are close. (As a trivial example, consider a loss function that is identically zero.) 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 $\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \nabla \mathcal{L}(\theta^*)^T \Delta \geq \gamma d^2(\Delta)$ for all $\Delta \in \mathbb{R}^p$. In the high-dimensional setting, where the number of parameters p may be much larger than the sample size, the strong convexity assumption need not be satisfied. As a simple example, consider the usual linear regression model $y = X\theta^* + w$, where $y \in \mathbb{R}^n$ is the response vector, $\theta^* \in \mathbb{R}^p$ is the unknown parameter vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix, and $w \in \mathbb{R}^n$ is a noise vector, with i.i.d. zero mean elements. The least-squares loss is given by $\mathcal{L}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$, and has the Hessian $H(\theta) = \frac{1}{n}X^T X$. It is easy to check that the $p \times p$ matrix $H(\theta)$ will be rank-deficient whenever $p > n$, showing that the least-squares loss cannot be strongly convex (with respect to $d(\cdot) = \|\cdot\|_2$) when $p > n$.

Herein lies the utility of Lemma 1: it guarantees that the error $\widehat{\Delta}$ must lie within a restricted set, so that we only need the loss function to be strongly convex for a limited set of directions. More precisely, we have:

Definition 2. Given some subset $\mathcal{C} \subseteq \mathbb{R}^p$ and error norm $d(\cdot)$, we say that the loss function $\mathcal{L}$ satisfies *restricted strong convexity* (RSC) (with respect to $d(\cdot)$) with parameter $\gamma(\mathcal{L}) > 0$ over $\mathcal{C}$ if

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \nabla \mathcal{L}(\theta^*)^T \Delta \quad \geq \quad \gamma(\mathcal{L})\, d^2(\Delta) \qquad \text{for all } \Delta \in \mathcal{C}. \tag{4}$$

In the statement of our results, we will be interested in loss functions that satisfy RSC over sets $\mathcal{C}(A, B, \epsilon)$ that are indexed by a subspace pair (A, B) and a tolerance $\epsilon \geq 0$ as follows:

$$\mathcal{C}(A, B, \epsilon) := \big\{\Delta \in \mathbb{R}^p \mid r(\pi_B(\Delta)) \;\leq\; 3r(\pi_{B^\perp}(\Delta)) + 4r(\pi_{A^\perp}(\theta^*)), \quad d(\Delta) \geq \epsilon\big\}. \tag{5}$$

In the special case of least-squares regression with hard sparsity constraints, the RSC condition corresponds to a lower bound on the sparse eigenvalues of the Hessian matrix $X^T X$, and is essentially equivalent to a restricted eigenvalue condition introduced by Bickel et al. [1].

3 Convergence rates

We are now ready to state a general result that provides bounds and hence convergence rates for the error $d(\widehat{\theta} - \theta^*)$. Although it may appear somewhat abstract at first sight, we illustrate that this result has a number of concrete consequences for specific models. In particular, we recover some known results about estimation in s-sparse models [1], as well as a number of new results, including convergence rates for estimation under ℓ_q-sparsity constraints, estimation in sparse generalized linear models, estimation of block-structured sparse matrices and estimation of low-rank matrices.

In addition to the regularization parameter λ_n and RSC constant $\gamma(\mathcal{L})$ of the loss function, our general result involves a quantity that relates the error metric d to the regularizer r; in particular, for any set $A \subseteq \mathbb{R}^p$, we define

$$\Psi(A) := \sup_{\{u \in \mathbb{R}^p \mid d(u) = 1\}} r(u), \tag{6}$$

so that $r(u) \leq \Psi(A) d(u)$ for $u \in A$.

Theorem 1 (Bounds for general models)**.** *For a given subspace collection $\mathcal{V}$, suppose that the regularizer r is decomposable, and consider the regularized M-estimator* (1) *with $\lambda_n \geq 2\, r^*(\nabla \mathcal{L}(\theta^*))$. Then, for any pair of subspaces $(A, B) \in \mathcal{V}$ and tolerance $\epsilon \geq 0$ such that the loss function $\mathcal{L}$ satisfies restricted strong convexity over $\mathcal{C}(A, B, \epsilon)$, we have*

$$d(\widehat{\theta} - \theta^*) \quad \leq \quad \max\Big\{\epsilon,\, \frac{1}{\gamma(\mathcal{L})}\left[2\, \Psi(B^\perp)\, \lambda_n + \sqrt{2\, \lambda_n\, \gamma(\mathcal{L})\, r(\pi_{A^\perp}(\theta^*))}\right]\Big\}. \tag{7}$$

The proof is motivated by arguments used in past work on high-dimensional estimation (e.g., [9, 14]); we provide the details in the full-length version. In the remainder of this paper, we illustrate the consequences of Theorem 1 for specific models. The parameter λ_n will be selected as small as possible while satisfying the lower bound $2\, r^*(\nabla \mathcal{L}(\theta^*))$. For the sake of clarity, the error $d(\cdot)$ is taken to be $\|\cdot\|_2$. For all models $\epsilon = 0$, apart from the weak-sparse model in section 3.1.2.

3.1 Bounds for linear regression

Consider the standard linear regression $y = X\theta^* + w$ model, where $\theta^* \in \mathbb{R}^p$ is the regression vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix, and $w \in \mathbb{R}^n$ is a noise vector. Given the observations (y, X), our goal is to estimate the regression vector θ^*. Without any structural constraints on θ^*, we can apply Theorem 1 with the trivial subspace collection $\mathcal{T} = \{(\mathbb{R}^p, 0)\}$ to establish a rate $\|\widehat{\theta} - \theta^*\|_2 = \mathcal{O}(\sigma\sqrt{p/n})$ for ridge regression. Note that the RSC condition requires that X is full-rank so that $n > p$. Here we consider bounds for linear regression where θ^* is an s-sparse vector.

3.1.1 Lasso estimates of hard sparse models

More precisely, let us consider estimating an s-sparse regression vector θ^* by solving the Lasso program

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \Big\{\frac{1}{2n}\|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1\Big\}. \tag{8}$$

The Lasso is a special case of our M-estimator (1) with $r(\theta) = \|\theta\|_1$, and $\mathcal{L}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$.

Recall the definition of the s-sparse subspace collection $\mathcal{S}$ from Section 2.2. For this problem, let us set $\epsilon = 0$ so that the restricted strong convexity set (5) becomes $\mathcal{C}(A, B, 0) = \{\Delta \in \mathbb{R}^p \mid \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\}$. Establishing restricted strong convexity for the least-squares loss is equivalent to ensuring the following bound on the design matrix:

$$\|X\theta\|_2^2/n \;\geq\; \gamma(\mathcal{L})\, \|\theta\|_2^2 \qquad \text{for all } \theta \in \mathbb{R}^p \text{ s.t. } \|\theta_S\|_1 \leq 3\|\theta_S\|_1. \tag{9}$$

As mentioned previously, this condition is essentially the same as the restricted eigenvalue condition developed by Bickel et al. [1]. Moreover, we note that Raskutti et al. [10] have shown that condition (9) will hold with high probability for various random ensembles of Gaussian matrices. The i^{th} column of X, X_i, also satisfies the constraint $\|X_i\|_2 \leq \sqrt{n}$. Finally, we assume that the elements of w_i are zero-mean and have sub-Gaussian tails, meaning that there exists some constant $\sigma > 0$ such that $\mathbb{P}[|w_i| > t] \leq \exp(-t^2/2\sigma^2)$ for all $t > 0$. Under these conditions, we recover as a corollary of Theorem 1 the following known result [1, 6].

Corollary 1. *Suppose that the true vector $\theta^* \in \mathbb{R}^p$ is exactly s-sparse with support S, and that the design matrix X satisfies condition* (9)*. If we solve the the Lasso with $\lambda_n^2 = \frac{16\sigma^2 \log p}{n}$, then with probability at least $1 - c_1 \exp(-c_2 n\lambda_n^2)$, the solution satisfies*

$$\|\widehat{\theta} - \theta^*\|_2 \quad\leq\quad \frac{8\sigma}{\gamma(\mathcal{L})}\sqrt{\frac{s\log p}{n}}. \tag{10}$$

Proof. As noted previously, the ℓ_1-regularizer is decomposable for the sparse subspace collection $\mathcal{S}$, while condition (9) ensures that RSC holds for all sets $\mathcal{C}(A, B, 0)$ with $(A, B) \in \mathcal{S}$. We must verify that the given choice of regularization satisfies $\lambda_n \geq 2\, r^*(\nabla\mathcal{L}(\theta^*))$. Note that $r^*(\cdot) = \|\cdot\|_\infty$, and moreover that $\nabla\mathcal{L}(\theta^*) = X^T w/n$. Under the column normalization condition on the design matrix X and the sub-Gaussian nature of the noise, it follows that $\|X^T w/n\|_\infty \leq \sqrt{4\sigma^2 \frac{\log p}{n}}$ with high probability. The bound in Theorem 1 is thus applicable, and it remains to compute the form that its different terms take in this special case. For the ℓ_1-regularizer and the ℓ_2 error metric, we have $\Psi(A_S) = \sqrt{|S|}$. Given the hard sparsity assumption, $r(\theta^*_{S^c}) = 0$, so that Theorem 1 implies that $\|\widehat{\theta} - \theta^*\|_2 \leq \frac{2}{\gamma(\mathcal{L})}\sqrt{s}\lambda_n = \frac{8\sigma}{\gamma(\mathcal{L})}\sqrt{\frac{s\log p}{n}}$, as claimed. □

3.1.2 Lasso estimates of weak sparse models

We now consider models that satisfy a weak sparsity assumption. More concretely, suppose that θ^* lies in the ℓ_q-"ball" of radius R_q—namely, the set $\mathbb{B}_q(R_q) := \{\theta \in \mathbb{R}^p \mid \sum_{i=1}^p |\theta_i|^q \leq R_q\}$ for some $q \in (0, 1]$. Our analysis exploits the fact that any $\theta^* \in \mathbb{B}_q(R_q)$ can be well approximated by an s-sparse vector (for an appropriately chosen sparsity index s). It is natural to approximate θ^* by a vector supported on the set $S = \{i \mid |\theta_i^*| \geq \tau\}$. For any choice of threshold $\tau > 0$, it can be shown that $|S| \leq R_q\tau^{-q}$, and as shown in the full-length version, the optimal choice is to set $\tau = \lambda_n$, using the same regularization parameter as in Corollary 1. Accordingly, we consider the s-sparse subspace collection $\mathcal{S}$ with subsets of size $s = R_q\lambda_n^{-q}$. We assume that the noise vector $w \in \mathbb{R}^n$ is as defined above and that the columns are normalized as in the previous section. We also assume that the matrix X satisfies the condition

$$\|Xv\|_2 \;\geq\; \kappa_1\|v\|_2 - \kappa_2\Big(\frac{\log p}{n}\Big)^{\frac{1}{2}}\|v\|_1 \qquad \text{for constants } \kappa_1, \kappa_2 > 0. \tag{11}$$

Raskutti et al. [10] show that this property holds with high probablity for suitable Gaussian random matrices. Under this condition, it can be verified that RSC holds with $\gamma(\mathcal{L}) = \kappa_1/2$ over the set $\mathcal{C}\big(A(S), B(S), \epsilon_n\big)$, where $\epsilon_n = \big(4/\kappa_1 + \sqrt{4/\kappa_1}\big)R_q^{\frac{1}{2}}\big(\sqrt{\frac{16\,\sigma^2\log p}{n}}\big)^{1-q/2}$. The following result, which we obtain by applying Theorem 1 in this setting, is new to the best of our knowledge:

Corollary 2. *Suppose that the true vector $\theta^* \in \mathbb{B}_q(R_q)$, and the design matrix X satisfies condition* (11)*. If we solve the Lasso with $\lambda_n^2 = \frac{16\sigma^2\log p}{n}$, then with probability $1 - c_1\exp(-c_2 n\lambda_n^2)$, the solution satisfies*

$$\|\widehat{\theta} - \theta^*\|_2 \quad\leq\quad R_q^{\frac{1}{2}}\left(\sqrt{\frac{16\,\sigma^2\log p}{n}}\right)^{1-q/2}\left[\frac{2}{\gamma(\mathcal{L})} + \frac{\sqrt{2}}{\sqrt{\gamma(\mathcal{L})}}\right]. \tag{12}$$

We note that both of the rates—for hard-sparsity in Corollary 1 and weak-sparsity in Corollary 2—are known to be optimal in a minimax sense [10]. In [10], the authors also show that (12) is achievable by solving the computationally intractable problem of minimizing $\mathcal{L}(\theta)$ over the ℓ_q-ball.

3.2 Bounds for generalized linear models

Next, consider any generalized linear model with canonical link function, where the distribution of response $y \in \mathcal{Y}$, given predictor $X \in \mathbb{R}^p$, is given by $p(y|X;\theta^*) = \exp(y{\theta^*}^T X - a({\theta^*}^T X) + d(y))$, for some fixed functions $a : \mathbb{R} \mapsto \mathbb{R}$ and $d : \mathcal{Y} \mapsto \mathbb{R}$, where $|X| \leq A$, and $|y| \leq B$. We consider estimating θ^* from observations $\{(X_i, y_i)\}_{i=1}^n$ by ℓ_1-regularized maximum likelihood:

$$\widehat{\theta} \in \arg\min_{\theta\in\mathbb{R}^p} \big\{ -\frac{1}{n}\theta^T\big(\sum_{i=1}^n y_i X_i\big) + \frac{1}{n}\sum_{i=1}^n a(\theta^T X_i) + \|\theta\|_1 \big\}, \tag{13}$$

so that $\mathcal{L}(\theta) = -\theta^T\big(\frac{1}{n}\sum_{i=1}^n y_i X_i\big) + \frac{1}{n}\sum_{i=1}^n a(\theta^T X_i)$, and $r(\theta) = \|\theta\|_1$. Let $X \in \mathbb{R}^{n\times p}$ denote the matrix with X_i as row i. Again we use the s-sparse subspace collection $\mathcal{S}$ and $\epsilon = 0$, so that it can be verified that it suffices for the restricted strong convexity condition to hold if for some $c > 0$, $\ddot{a}(\theta^T x) > c, \text{ for } |x| \leq M, \; \theta \in \{\theta^* + \Delta : \|\Delta\|_2 \leq \frac{16AB}{\gamma(\mathcal{L})}\sqrt{\frac{s\log p}{n}}\}$, and that the design matrix X satisfies the restricted eigenvalue bound

$$\|X\theta\|_2^2/n \;\geq\; \frac{\gamma(\mathcal{L})}{c}\,\|\theta\|_2^2 \qquad \text{for all } \theta \in \mathbb{R}^p \text{ s.t. } \|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1. \tag{14}$$

Corollary 3. *Suppose that the true vector $\theta^* \in \mathbb{R}^p$ is exactly s-sparse with support S, and the design matrix X satisfies condition* (14)*. Suppose that we solve the ℓ_1-regularized M-estimator* (13) *with $\lambda_n^2 = \frac{32A^2B^2\log p}{n}$. Then with probability $1 - c_1\exp(-c_2 n\lambda_n^2)$, the solution satisfies*

$$\|\widehat{\theta} - \theta^*\|_2 \quad\leq\quad \frac{16AB}{\gamma(\mathcal{L})}\sqrt{\frac{s\log p}{n}}. \tag{15}$$

We defer the proof to the full-length version due to space constraints.

3.3 Bounds for sparse matrices

In this section, we consider some extensions of our results to estimation of regression matrices. Various authors have proposed extensions of the Lasso based on regularizers that have more structure than the ℓ_1 norm [17, 22]. Such regularizers allow one to impose various types of block-sparsity constraints, in which groups of parameters are assumed to be active (or inactive) simultaneously. We assume that the observation model takes on the form $Y = X\Theta^* + W$, where $\Theta^* \in \mathbb{R}^{k\times m}$ is the unknown fixed set of parameters, $X \in \mathbb{R}^{n\times k}$ is the design matrix, and $W \in \mathbb{R}^{n\times m}$ is the noise matrix. As a loss function, we use the Frobenius norm $\frac{1}{n}\mathcal{L}(\Theta) = |||Y - X\Theta|||_F^2$, and as a regularizer, we use the $\ell_{1,q}$-matrix norm for some $q \geq 1$, which takes the form $|||\Theta|||_{1,q} = \sum_{i=1}^k \|(\Theta_{i1}, \ldots, \Theta_{im})\|_q$. We refer to the resulting estimator as the q-group Lasso. We define the quantity $\eta(m;q) = 1$ if $q \in (1,2]$ and $\eta(m;q) = m^{1/2-1/q}$ if $q > 2$. We then set the regularization parameter as follows:

$$\lambda_n = \begin{cases} \frac{4\sigma}{\sqrt{n}}[\eta(m;q)\sqrt{\log k} + C_q m^{1-1/q}] & \text{if } q > 1 \\ 4\sigma\sqrt{\frac{\log(km)}{n}} & \text{for } q = 1. \end{cases}$$

Corollary 4. *Suppose that the true parameter matix Θ^* has non-zero rows only for indices $i \in S \subseteq \{1,\ldots,k\}$ where $|S| = s$, and that the design matrix $X \in \mathbb{R}^{n\times k}$ satisfies condition* (9)*. Then with probability at least $1 - c_1\exp(-c_2 n\lambda_n^2)$, the q-block Lasso solution satisfies*

$$|||\widehat{\Theta} - \Theta^*|||_F \quad\leq\quad \frac{2}{\gamma(\mathcal{L})}\Psi(S)\lambda_n. \tag{16}$$

Proof. We simply need to establish that the regularization parameter satisfies $\lambda_n \geq 2\,r^*(\nabla\mathcal{L}(\Theta^*))$. We note that for a matrix U, $r^*(U) = \max_{i=1,\ldots,k}\|U_{i\cdot}\|_{q'}$ for $1/q' = 1 - 1/q$. Moreover, we have $\nabla\mathcal{L}(\Theta^*) = \frac{1}{n}X^T W$. Concentration results on $\|\cdot\|_{q'}$ and the union bound yield that $r^*(\frac{1}{n}X^T W) \leq \frac{2\sigma}{\sqrt{n}}[\eta(m;q)\sqrt{\log k} + C_q m^{1-1/q}]$, as required. □

We will now consider three special cases of the above result. A simple argument shows that $\Psi(S) = \sqrt{s}$ if $q \geq 2$, and $\Psi(S) = m^{1/q-1/2}\sqrt{s}$ if $q \in [1,2]$. First, we consider $q = 1$, and note that solving the Group Lasso with $q = 1$ is identical solving a Lasso problem with sparsity sm and ambient dimension km. The resulting upper bound on the Frobenius norm reflects this fact: more specifically, for $q = 1$, the bound is $\frac{8\sigma}{\gamma(\mathcal{L})}\sqrt{\frac{s\,m\log(km)}{n}}$. For the case $q = 2$, Corollary 4 implies that the Frobenius error $|\!|\!|\widehat{\Theta} - \Theta^*|\!|\!|_F$ is upper bounded as $\frac{8\sigma}{\gamma(\mathcal{L})}\Big[\sqrt{\frac{s\log k}{n}} + \sqrt{\frac{sm}{n}}\Big]$. This is also a very natural result: the term $\frac{s\log k}{n}$ captures the difficulty of finding the s non-zero rows out of the total k, whereas the term $\frac{sm}{n}$ captures the difficulty of estimating the sm free parameters in the matrix (once the non-zero rows have been determined). We note that recent work by Lounici et al. [4] established the bound $\mathcal{O}(\frac{\sigma}{\gamma(\mathcal{L})}\sqrt{\frac{c\sqrt{m}\;s\log k}{n} + \frac{sm}{n}})$, which is equivalent apart from a term $\sqrt{m}$. Finally, for $q = \infty$, we obtain the upper bound $\frac{8\sigma}{\gamma(\mathcal{L})}\big[\sqrt{\frac{s\log k}{n}} + m\sqrt{\frac{s}{n}}\big]$, which is a novel result.

3.4 Bounds for estimating low rank matrices

Finally, we consider the implications of our main result for the problem of estimating low-rank matrices. This structural assumption is a natural generalization of sparsity, and has been studied by various authors (see the paper [13] and references therein). To illustrate our main theorem in this context, let us consider the following instance of low-rank matrix learning. Given a low-rank matrix $\Theta^* \in \mathbb{R}^{k\times m}$, suppose that we are given n noisy observations of the form $Y_i = \langle\!\langle X_i,\; \Theta^*\rangle\!\rangle + W_i$, where $W_i \sim N(0,1)$. Such an observation model arises, for instance, in system identification settings in control theory [13]. The following regularized M-estimator can be considered in order to estimate the desired low-rank matrix Θ^*:

$$\min_{\Theta\in\mathbb{R}^{m\times p}} \frac{1}{2n}\sum_{i=1}^{n} |Y_i - \langle\!\langle X_i,\; \Theta)\rangle\!\rangle|^2 + |\!|\!|\Theta|\!|\!|_1, \tag{17}$$

where the regularizer, $|\!|\!|\Theta|\!|\!|_1$, is the nuclear norm, or the sum of the singular values of Θ.

Recall the rank-r collection $\mathcal{V}$ defined for low-rank matrices in Section 2.2. Let $\Theta^* = U\Sigma W^T$ be the singular value decomposition (SVD) of Θ^*, so that $U \in \mathbb{R}^{k\times r}$ and $W \in \mathbb{R}^{m\times r}$ are orthogonal, and $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix. If we let $A = A(U,W)$ and $B = B(U,W)$, then, $\pi_B(\Theta^*) = 0$, so that by Lemma 1 we have that $|\!|\!|\pi_B(\Delta)|\!|\!|_1 \leq 3\,|\!|\!|\pi_{B^\perp}(\Delta)|\!|\!|_1$. Thus, for restricted strong convexity to hold it can be shown that the design matrices X_i must satisfy

$$\frac{1}{n}\sum_{i=1}^{n} |\langle\!\langle X_i,\; \Delta\rangle\!\rangle|^2 \geq \gamma(\mathcal{L})\,|\!|\!|\Delta|\!|\!|_F^2 \qquad \text{for all } \Delta \text{ such that } |\!|\!|\pi_B(\Delta)|\!|\!|_1 \leq 3\,|\!|\!|\pi_{B^\perp}(\Delta)|\!|\!|_1. \tag{18}$$

As with the analogous conditions for sparse vectors and linear regression, this condition can be shown to hold with high probability for Gaussian random matrices.

Corollary 5. *Suppose that the true matrix Θ^* has rank $r \ll \min(k,m)$, and that the design matrices $\{X_i\}$ satisfy condition* (18)*. If we solve the regularized M-estimator* (17) *with $\lambda_n = 4\frac{\sqrt{k}+\sqrt{m}}{\sqrt{n}}$, then with probability at least $1 - c_1\exp(-c_2(k+m))$, we have*

$$|\!|\!|\widehat{\Theta} - \Theta^*|\!|\!|_F \leq \frac{16}{\gamma(\mathcal{L})}\Big[\sqrt{\frac{rk}{n}} + \sqrt{\frac{rm}{n}}\Big]. \tag{19}$$

Proof. Note that if $\operatorname{rank}(\Theta^*) = r$, then $|\!|\!|\Theta^*|\!|\!|_1 \leq \sqrt{r}|\!|\!|\Theta^*|\!|\!|_F$ so that $\Psi(B^\perp) = \sqrt{2r}$, since the subspace $B(U,V)^\perp$ consists of matrices with rank at most $2r$. All that remains is to show that $\lambda_n \geq 2\,r^*(\nabla\mathcal{L}(\Theta^*))$. Standard analysis gives that the dual norm to $|\!|\!|\cdot|\!|\!|_1$ is the operator norm, $|\!|\!|\cdot|\!|\!|_2$. Applying this observation and the fact that $\nabla\mathcal{L}(\Theta^*) = -\frac{1}{n}\sum_{i=1}^{n} X_i W_i$ we can construct a bound on the operator norm of $\frac{1}{n}\sum_{i=1}^{n} X_i W_i$. We assume that the entries of X_i are i.i.d. $N(0,1)$. Then, conditioned on W, the entries of the matrix $\frac{1}{n}\sum_{i=1}^{n} X_i W_i$ are i.i.d. $N(0, \|W\|_2^2/n^2)$ from which it can be shown that with probability at least $1 - c_1\exp(-c_2 n)$, $\|W\|_2^2/n \leq 2$. Coupled with results on random matrix theory we have that $|\!|\!|\frac{1}{n}\sum_{i=1}^{n} X_i W_i|\!|\!|_2 \leq 2\frac{\sqrt{k}+\sqrt{m}}{\sqrt{n}}$ with probability at least $1 - c_1\exp(-c_2(k+m))$, verifying that $\lambda_n \geq 2\,r^*(\nabla\mathcal{L}(\theta^*))$. □

References

[1] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. Submitted to Annals of Statistics, 2008.

[2] 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.

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

[4] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. *Arxiv*, 2009.

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

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

[7] S. Negahban and M. J. Wainwright. Simultaneous support recovery in high-dimensional regression: Benefits and perils of $\ell_{1,\infty}$-regularization. Technical report, Department of Statistics, UC Berkeley, April 2009.

[8] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Union support recovery in high-dimensional multivariate regression. Technical report, Department of Statistics, UC Berkeley, August 2008.

[9] S. Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large: I. consistency. *Annals of Statistics*, 12(4):1296–1309, 1984.

[10] 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.

[11] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 2008. To appear.

[12] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Technical Report 767, Department of Statistics, UC Berkeley, September 2008.

[13] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *Allerton Conference 07, Allerton House, Illinois*, 2007.

[14] A.J. Rothman, P.J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Statist.*, 2:494–515, 2008.

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

[16] J. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Info Theory*, 52(3):1030–1051, March 2006.

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

[18] S. Van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.

[19] 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.

[20] C. Zhang and J. Huang. Model selection consistency of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36:1567–1594, 2008.

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

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

A Ridge-Regression

In this section, we apply Theorem 1 to ridge-regression. Consider solving the program

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \Big\{ \frac{1}{n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_2 \Big\}.$$

Assume that the underlying structure enforces $\|\theta^*\|_2 \leq M$ for some constant $M > 0$. As a result, the restricted strong convexity assumption reduces to $\lambda_{\min}(\frac{1}{n}X^T X) \geq \gamma(\mathcal{L}) > 0$. We may now present the following trivial corollary to Theorem 1. Note that the result is not new, and provides exactly the same bound as in the ordinary least-squares solution to the problem.

Corollary 6. *Suppose that the true vector $\theta^* \in \mathbb{R}^p$ and that the design matrix X has its smallest eigenvalue bounded below by $\gamma(\mathcal{L})$. Suppose that we solve the Ridge-regression program with $\lambda_n^2 = \frac{p}{n}$. Then, with probability $1 - c_1 \exp(-c_2 n \lambda_n^2)$, the solution satisfies*

$$\|\widehat{\theta} - \theta^*\|_2 \quad \leq \quad \frac{8\sigma}{\gamma(\mathcal{L})} \sqrt{\frac{p}{n}} \tag{20}$$

Proof. The restricted strong convexity condition clearly holds. Furthermore, let $\mathcal{V}$ be the space of all subspace-pairs. Therefore, we can apply the bound in Theorem 1. First note that $\Psi(A) = 1$ for any set A since $d(v) = r(v)\ \forall v \in \mathbb{R}^p$. The dual norm $r^*(\cdot)$ is $r(\cdot)$. Thus, we must establish the ℓ_2 norm of $\nabla\mathcal{L}(\theta^*) = X^T w/n$. However, the column normalization bounds yields that $\|X^T w/n\|_2 \leq 2\sigma\sqrt{p/n}$ with probability $1 - c_1 \exp(-c_2 p)$. Therefore, letting $\lambda_n = 2\|X^T w/n\|_2$ we have by Theorem 1 that $d(\widehat{\theta} - \theta^*) \leq \frac{1}{\gamma(\mathcal{L})}[8\sqrt{\frac{p}{n}} + \sqrt{8\lambda_n r(\pi_{A^\perp}(\theta^*))}]$. Thus, the bound is clealry minimized as long as $\bar{\theta}^*_A = 0$, which is the case if we let $A = \mathbb{R}^p$. Verifying the result. □

B Proof of Theorem 1

The argument is motivated by the methods of Rothman et al. [14], in their analysis of an ℓ_1-regularized log-determinant program. Consider the function

$$g(\Delta) \quad := \quad \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda_n \big\{ r(\theta^* + \Delta) - r(\theta^*) \big\}. \tag{21}$$

The convexity of $\mathcal{L}(\cdot)$ and $r(\cdot)$ implies that g is a convex function. Here, we have that $\Delta = \theta - \theta^*$ and $\widehat{\Delta} = \widehat{\theta} - \theta^*$. Observe that $g(0) = 0$ so that $g(\widehat{\Delta}) \leq 0$. From Lemma 1, we know that $\widehat{\Delta} \in \mathcal{C}$, where

$$\mathcal{C} \ := \ \{\Delta \in \mathbb{R}^p : r(\pi_B(\Delta)) \leq 3\, r(\overline{\pi}_B(\Delta)) + 4\, r(\pi_{A^\perp}(\theta^*))\}.$$

We also have that if $\Delta \in \mathcal{C}$, then $t\Delta \in \mathcal{C}$ for any $t \in [0, 1]$. Now suppose that $d(\widehat{\Delta}) > M$. Then there exists a $t \in (0, 1)$ such that $d(t\widehat{\Delta}) = M$ and $t\widehat{\Delta} \in \mathcal{C}$. Now suppose that $g(t\widehat{\Delta}) > 0$. Then, by the convexity of g

$$g((1-t)0 + t\widehat{\Delta}) \ \leq \ (1-t)g(0) + tg(\widehat{\Delta}).$$

We know $g(0) = 0$ and $t > 0$. Thus, $g(\widehat{\Delta}) > 0$, which is a contradiction. Therefore, $d(\widehat{\Delta}) \leq M$. Hence, it suffices to show that for any $\Delta \in \mathcal{C}$ such that $d(\Delta) = M$, $g(\Delta) > 0$, which we now prove.

Proof. Fix any arbitrary vector $\Delta \in \mathbb{R}^p$ such that $\Delta \in \mathcal{C}$ and $d(\Delta) = M$. We assume that restricted strong convexity holds for all such vectors Δ. Therefore,

$$\begin{aligned} g(\Delta) \quad &= \quad \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda_n \big\{ r(\theta^* + \Delta) - r(\theta^*) \big\} \\ &\geq \quad \nabla\mathcal{L}(\theta^*)^T \Delta + \gamma(\mathcal{L}) d(\Delta)^2 + \lambda_n \big\{ r(\theta^* + \Delta) - r(\theta^*) \big\}. \end{aligned} \tag{22}$$

Recall that $\lambda_n \geq 2r^*(\nabla\mathcal{L}(\theta^*))$, so that by Lemma 1

$$\begin{aligned} \nabla\mathcal{L}(\theta^*)^T \Delta + \lambda_n \big\{ r(\theta^* + \Delta) - r(\theta^*) \big\} \quad &\geq \quad \frac{\lambda_n}{2} \left\{ r(\pi_B(\Delta)) - 3r(\overline{\pi}_B(\Delta)) - 4r(\pi_{A^\perp}(\theta^*)) \right\} \\ &\geq \quad -\frac{\lambda_n}{2} \left\{ 3r(\overline{\pi}_B(\Delta)) + 4r(\pi_{A^\perp}(\theta^*)) \right\} \end{aligned}$$

Substituting the latter inequality into equaton (22) yields

$$g(\Delta) \;\geq\; \gamma(\mathcal{L})d(\Delta)^2 - \frac{\lambda_n}{2}\left\{3r(\overline{\pi}_B(\Delta)) + 4r(\pi_{A^\perp}(\theta^*))\right\}.$$

Noting that $r(\overline{\pi}_B(\Delta)) \;\leq\; \Psi(B^\perp)\, d(\overline{\pi}_B(\Delta)) \;\leq\; \Psi(B^\perp)\, d(\Delta)$, establishes that

$$g(\Delta) \;\geq\; \gamma(\mathcal{L})\, d(\Delta)^2 - \frac{\lambda_n}{2}\left\{3\Psi(B^\perp)d(\Delta) + 4r(\pi_{A^\perp}(\theta^*))\right\}.$$

Finally, substituting $M = \left\{\frac{1}{\gamma(\mathcal{L})}\;\left[2\,\Psi(B^\perp)\,\lambda_n + \sqrt{2\,\lambda_n\,\gamma(\mathcal{L})\, r(\pi_{A^\perp}(\theta^*))}\right]\right\}$ proves that $g(\Delta) > 0$. □

C Proofs and Auxiliary Results

Proof of Lemma 1. Recall the function

$$g(\Delta) \quad := \quad \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda_n\big\{r(\theta^* + \Delta) - r(\theta^*)\big\}. \tag{23}$$

We will start off by obtaining a lower bound for this function.

Loss Deviation: Using the convexity of the loss function $\mathcal{L}$, we have

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq \nabla\mathcal{L}(\theta^*)^T\Delta. \tag{24}$$

By the Cauchy-Schwartz inequality, we have

$$\begin{aligned} |\nabla\mathcal{L}(\theta^*)^T\Delta| \quad &\leq \quad r^*(\nabla\mathcal{L}(\theta^*))\, r(\Delta) \\ &\leq \quad \frac{\lambda_n}{2}\,\big[r(\pi_B(\Delta)) + r(\overline{\pi}_B(\Delta))\big], \end{aligned}$$

where we have used the assumption on $r^*(\nabla\mathcal{L}(\theta^*))$, and the triangle inequality. Substituting in (24)

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq -\frac{\lambda_n}{2}\,\big[r(\pi_B(\Delta)) + r(\overline{\pi}_B(\Delta))\big]. \tag{25}$$

Regularization Deviation: By the triangle inequality,

$$r(\theta^* + \Delta) \quad \geq \quad r(\pi_A(\theta^*) + \pi_B(\Delta)) - r(\pi_{A^\perp}(\theta^*)) - r(\overline{\pi}_B(\Delta)).$$

By the decomposition property,

$$r(\pi_A(\theta^*) + \pi_B(\Delta)) \quad = \quad r(\pi_A(\theta^*)) + r(\pi_B(\Delta)),$$

so that by another application of the triangle inequality,

$$r(\theta^* + \Delta) - r(\theta^*) \geq r(\pi_B(\Delta)) - r(\overline{\pi}_B(\Delta)) - 2r(\pi_{A^\perp}(\theta^*)). \tag{26}$$

Substituting the lower bounds for the loss and regularization function deviations (26) and (25) in (23),

$$g(\Delta) \quad \geq \quad \frac{\lambda_n}{2}\left[r(\pi_B(\Delta)) - 3r(\overline{\pi}_B(\Delta)) - 4r(\pi_{A^\perp}(\theta^*))\right]. \tag{27}$$

By construction $g(0) = 0$, and hence the deviation of the optimum Δ satisfies $g(\Delta) \leq 0$. Using in (27) and dividing by $\frac{\lambda_n}{2} > 0$ yields,

$$r(\pi_B(\Delta)) \leq 3\, r(\overline{\pi}_B(\Delta)) + 4\, r(\pi_{A^\perp}(\theta^*)),$$

as required. □

D Proof of Corollary 2

Proof. The subset $\mathcal{V}^*$ of the sparse-vectors decomposability-set collection we use in this corollary is the subset $\mathcal{V}^* = (A_S, A_{S^c})$ for sets $S \in \mathcal{S} = \{S \mid |S| \leq R_q(\log(p)/n)^{-q}\}$. As in the proof of Corollary 2, the assumptions of Theorem 1 are satisfied, so that we can use the bound in the theorem; its terms can be simplified as follows. Again, for the ℓ_1-regularizer and the ℓ_2 error metric, we have $\Psi(A_{S^*}) = \sqrt{|S^*|}$. Now $|S^*|$ can be bounded as follows:

$$\begin{aligned} R_q &\geq \sum_i |\theta_i^*|^q \geq \sum_{i \in S^*} |\theta_i^*|^q \\ &\geq \tau^q |S^*|, \end{aligned}$$

so that $|S^*| \leq \tau^{-q}\, R_q$. Further, given the soft sparsity assumption, $r(\theta^*_{S^{*c}})$ can be bound as follows:

$$\begin{aligned} \|\theta^*_{S^{*c}}\|_1 &= \sum_{i \in S^{*c}} |\theta_i^*| \\ &= \sum_{i \in S^{*c}} |\theta_i^*|^q |\theta_i^*|^{1-q} \leq R_q \tau^{1-q}. \end{aligned}$$

We thus obtain from Theorem 1 that

$$\begin{aligned} \|\widehat{\theta} - \theta^*\|_2 &\leq \frac{1}{\gamma(\mathcal{L})}\left[2\sqrt{|S^*|}\lambda_n + \sqrt{2\,\lambda_n\,\gamma(\mathcal{L})\,\|\theta^*_{S^{*c}}\|_1}\right] \\ &\leq \frac{1}{\gamma(\mathcal{L})}\left[2\sqrt{R_q}\tau^{-q/2}\lambda_n + \sqrt{2\,\lambda_n\,\gamma(\mathcal{L})\,R_q\tau^{1-q}}\right]. \end{aligned}$$

From the settings of τ and λ_n, it can be seen that $\lambda_n = \tau$, which when substituted in the previous expression yields,

$$\|\widehat{\theta} - \theta^*\|_2 \leq \sqrt{R_q \lambda_n^{2-q}}\left[\frac{2}{\gamma(\mathcal{L})} + \frac{\sqrt{2}}{\sqrt{\gamma(\mathcal{L})}}\right].$$

Substituting for the value of λ_n, we thus obtain the bound in the Corollary. □

D.1 Restricted Strong Convexity for Weak-Sparse Models

One sufficient condition for the restricted strong convexity condition to hold is that the design matrices $X \in \mathbb{R}^{n \times p}$ satisfy the conditioon

$$\|\frac{1}{\sqrt{n}}Xv\| \geq c_1\|v\|_2 - c_2\sqrt{\frac{\log p}{n}}\|v\|_1$$

for some constants $c_1 > 0$ and $c_2 > 0$.

In our setting, $\|v_{S^c}\|_1 \leq 3\|v_S\|_1 + 4\|\theta^*_{S^c}\|_1$ so that $\|v\|_1 \leq 4[\|v_S\|_1 + \|\theta^*_{S^c}\|_1]$, which further implies then that

$$\|v\|_1 \leq 4[\sqrt{|S|}\|v\|_1 + \|\theta^*_{S^c}\|_1.$$

Therefore, it immediately follows then that

$$\|\frac{1}{n}Xv\| \geq \left(c_1 - 4c_2\sqrt{\frac{|S|\log p}{n}}\right)\|v\|_2 - 4c_2\sqrt{\frac{\log p}{n}}\|\theta^*_{S^c}\|_1.$$

Recall from the arguments above that $\|\theta^*_{S^c}\|_1 \leq R_q\tau^{1-q}$ where we also set $\tau = \sqrt{\frac{\log p}{n}}$ and we are only concerned with sets such that $|S| \leq R_q\tau^{-q}$ so that

$$\|\frac{1}{\sqrt{n}}Xv\| \geq \left(c_1 - 4c_2\sqrt{R_q\tau^{2-q}}\right)\|v\|_2 - 4c_2R_q\tau^{2-q}.$$

For the applications of restricted strong convexity above, we only need it told hold for the vectors v such that $\|v\|_2 = \mathcal{O}(\frac{1}{c_1}\sqrt{R_q\tau^{2-q}})$ where we recall that $\tau = \lambda_n$, justifying the swap. Finally, applying the bound on $\|v\|_2$ yields that

$$\begin{aligned} \|\frac{1}{\sqrt{n}}Xv\|_2 &\geq \left(1 - 4c'\sqrt{R_q\tau^{2-q}}\right)\sqrt{R_q\tau^{2-q}} - 4c'R_q\tau^{2-q} \\ &\geq \left(1 - 8\,c'\sqrt{R_q\tau^{2-q}}\right)\sqrt{R_q\tau^{2-q}}, \end{aligned}$$

where $c' = c_2/c_1$. The constants c_1 and c_2 are independent of everything else and by the scaling of n, have that the term in the paranthesis can be made arbitrarily close to 1 by taking n sufficiently large. Therefore, have that

$$\|\frac{1}{\sqrt{n}}Xv\|_2 \geq \frac{c_1}{2}\|v\|_2,$$

which immediately implies then that $\gamma(\mathcal{L}) = \frac{c_1}{2}$ for $v \in \mathcal{G}$. Note, in fact that the bound holds for any v such that $\|v\|_2 \geq \frac{1}{c_1}\sqrt{R_q\tau^{2-q}}$, which implies then that the bound established in Corollary 2 is valid since $\frac{8+2\sqrt{8}}{2} \geq 1$.

E Restricted Strong Convexity for the Trace Observation Model

Recall the low-rank matrix observation model is

$$Y_i = \text{trace}(X_i^T\Theta^*) + W_i,$$

where $X_i, \Theta^* \in \mathbb{R}^{m\times p}$. Note that by we can convert each X_i and Θ^* to a vector to yield the usual linear regression observation model

$$Y = X\theta + W,$$

where $X \in \mathbb{R}^{n\times(pm)}$ and $\theta \in \mathbb{R}^{pm}$. We establish RSC for the simple case where the observation matrices X_i are drawn from the i.i.d. Gaussian ensemble. We will then appeal to the Gordon-Slepian Lemma to establish that

$$\inf_{\{\Delta:\|\Delta\|_2\|=1\}} \|\frac{1}{\sqrt{n}}X\Delta\|_2 \geq c_1\|\Delta\|_2 - c_2\frac{\sqrt{p}+\sqrt{m}}{\sqrt{n}}\|\Delta\|_1$$

where the norm $\|\Delta\|_1$ is the nuclear norm, and $\|\Delta\|_2$ is the Frobenius norm. Gordon-Slepain will lower bound the expected value of the random variable $\inf_\Delta \|\frac{1}{\sqrt{n}}X\Delta\|_2$, while we then apply concentration results to arrive at the above result with high probability, leaving that as an exercise. We know that

$$\inf_\Delta \|X\Delta\|_2 = \inf_\Delta \sup_U \text{trace}(U^TX\Delta).$$

Now, $\text{trace}(U^TX\Delta)$ is a centered Gaussian random process indexed by U and Δ. We may construct a second centered Gaussian random process index by U and Δ by defining $Y_{U,\Delta} = \text{trace}\,U^TW + \text{trace}\,\Delta^TZ$, where W, Z are independent normal i.i.d. Gaussian matrices. We thus have the following

$$\mathbb{E}[(X_{U,\Delta} - X_{U',\Delta'})^2] = \mathbb{E}[[\text{trace}\,(X(\Delta U^T - \Delta'(U')^T)]^2] = \vert\vert\vert\Delta U^T - \Delta'(U')^T\vert\vert\vert_F^2. \tag{28}$$

and

$$\begin{aligned} &\mathbb{E}[(\text{trace}((U-U')^TW) + \text{trace}((\Delta-\Delta')^TZ))^2] \\ =\;& \mathbb{E}[(\text{trace}((U-U')^TW))^2 + (\text{trace}((\Delta-\Delta')^TZ))^2] \\ =\;& \vert\vert\vert(U-U')\vert\vert\vert_F^2 + \vert\vert\vert(\Delta-\Delta')\vert\vert\vert_F^2 \end{aligned} \tag{29}$$

Recall that U and Δ are the vectorized versions of the corresponding matrices. Equation (28) is upper bounded by equation (29). On the other hand, if $\Delta = \Delta'$, then equation (28) equals equation (29), thus verifying the conditions of the Gordon-Slepain Lemma. Therefore, by the lemma, it immediately follows then that

$$\begin{aligned} \mathbb{E}\inf_\Delta\sup_U U^TX\Delta &\geq \mathbb{E}\inf_\Delta\sup_U U^TW + \Delta^TZ \\ &= \mathbb{E}\vert\vert\vert W\vert\vert\vert_F - \vert\vert\vert\Delta\vert\vert\vert_1\mathbb{E}\vert\vert\vert Z\vert\vert\vert_2 \\ &\geq \frac{\sqrt{n}}{2} - \vert\vert\vert\Delta\vert\vert\vert_1(\sqrt{p}+\sqrt{m}) \end{aligned}$$

as desired.