Journal of Machine Learning Research 11 (2010) 2241-2259 Submitted 3/10; Revised 8/10; Published 8/10

Restricted Eigenvalue Properties for Correlated Gaussian Designs

Garvesh Raskutti GARVESHR@STAT.BERKELEY.EDU
Martin J. Wainwright* WAINWRIG@STAT.BERKELEY.EDU
Bin Yu* BINYU@STAT.BERKELEY.EDU
Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Editor: John Lafferty

Abstract

Methods based on ℓ_1-relaxation, such as basis pursuit and the Lasso, are very popular for sparse regression in high dimensions. The conditions for success of these methods are now well-understood: (1) exact recovery in the noiseless setting is possible if and only if the design matrix X satisfies the restricted nullspace property, and (2) the squared ℓ_2-error of a Lasso estimate decays at the minimax optimal rate $\frac{k \log p}{n}$, where k is the sparsity of the p-dimensional regression problem with additive Gaussian noise, whenever the design satisfies a restricted eigenvalue condition. The key issue is thus to determine when the design matrix X satisfies these desirable properties. Thus far, there have been numerous results showing that the restricted isometry property, which implies both the restricted nullspace and eigenvalue conditions, is satisfied when all entries of X are independent and identically distributed (i.i.d.), or the rows are unitary. This paper proves directly that the restricted nullspace and eigenvalue conditions hold with high probability for quite general classes of Gaussian matrices for which the predictors may be highly dependent, and hence restricted isometry conditions can be violated with high probability. In this way, our results extend the attractive theoretical guarantees on ℓ_1-relaxations to a much broader class of problems than the case of completely independent or unitary designs.

Keywords: Lasso, basis pursuit, random matrix theory, Gaussian comparison inequality, concentration of measure

1. Introduction

Many fields in modern science and engineering—among them computational biology, astrophysics, medical imaging, natural language processing, and remote sensing—involve collecting data sets in which the dimension of the data p exceeds the sample size n. Problems of statistical inference in this high-dimensional setting have attracted a great deal of attention in recent years. One concrete instance of a high-dimensional inference problem concerns the standard linear regression model, in which the goal is to estimate a vector $\beta^* \in \mathbb{R}^p$ that connects a real-valued response y to a vector of covariates $X = (X_1, \ldots, X_p)$. In the setting $p \gg n$, the classical linear regression model is unidentifiable, so that it is not meaningful to estimate the parameter vector $\beta^* \in \mathbb{R}^p$. However, many high-dimensional regression problems exhibit special structure that can lead to an identifiable model. In particular, sparsity in the regression vector β^* is an archetypal example of such struc-

*. Also in the Department of Electrical Engineering & Computer Science.

ture, and there is now a substantial and rapidly growing body of work on high-dimensional linear regression with sparsity constraints.

Using the ℓ_1-norm to enforce sparsity has been very successful, as evidenced by the widespread use of methods such as basis pursuit (Chen et al., 1998), the Lasso (Tibshirani, 1996) and the Dantzig selector (Candes and Tao, 2007). There is now a well-developed theory on what conditions are required on the design matrix $X \in \mathbb{R}^{n \times p}$ for such ℓ_1-based relaxations to reliably estimate β^*. In the case of noiseless observation models, it is known that imposing a *restricted nullspace property* on the design matrix $X \in \mathbb{R}^{n \times p}$ is both necessary and sufficient for the basis pursuit linear program to recover β^* exactly. The nullspace property and its link to the basis pursuit linear program has been discussed in various papers (Cohen et al., 2009; Donoho and Huo, 2001; Feuer and Nemirovski, 2003). In the case of noisy observations, exact recovery of β^* is no longer possible, and one goal is to obtain an estimate $\widehat{\beta}$ such that the ℓ_2-error $\|\widehat{\beta} - \beta^*\|_2$ is well-controlled. To this end, various sufficient conditions for the success of ℓ_1-relaxations have been proposed, including restricted eigenvalue conditions (Bickel et al., 2009; Meinshausen and Yu, 2009) and the restricted Riesz property (Zhang and Huang, 2008). Of the conditions mentioned, one of weakest known sufficient conditions for bounding ℓ_2-error are the restricted eigenvalue (RE) conditions due to Bickel et al. (2009) and van de Geer (2007). In this paper, we consider a restricted eigenvalue condition that is essentially equivalent to the RE condition in Bickel et al. (2009). As shown by Raskutti et al. (2009), a related restriction is actually necessary for obtaining good control on the ℓ_2-error in the minimax setting.

Thus, in the setting of linear regression with random design, the interesting question is the following: for what ensembles of design matrices do the restricted nullspace and eigenvalue conditions hold with high probability? To date, the main routes to establishing these properties have been via either incoherence conditions (Donoho and Huo, 2001; Feuer and Nemirovski, 2003) or via the restricted isometry property (Candes and Tao, 2005), both of which are sufficient but not necessary conditions (Cohen et al., 2009; van de Geer and Buhlmann, 2009). The restricted isometry property (RIP) holds with high probability for various classes of random matrices with i.i.d. entries, including sub-Gaussian matrices (Mendelson et al., 2008) with sample size $n = \Omega(k \log(p/k))$, and for i.i.d. subexponential random matrices (Adamczak et al., 2009) provided that $n = \Omega(k \log^2(p/k))$. It has also been demonstrated that RIP is satisfied for matrices from unitary ensembles (e.g., Guédon et al., 2007, 2008; Romberg, 2009; Rudelson and Vershynin, 2008), for which the rows are generated based on independent draws from a set of uncorrelated basis functions.

Design matrices based on i.i.d. or unitary ensembles are well-suited to the task of compressed sensing (Candes and Tao, 2005; Donoho, 2006), where the matrix X can be chosen by the user. However, in most of machine learning and statistics, the design matrix is not under control of the statistician, but rather is specified by nature. As a concrete example, suppose that we are fitting a linear regression model to predict heart disease on the basis of a set of p covariates (e.g., diet, exercise, smoking). In this setting, it is not reasonable to assume that the different covariates are i.i.d. or unitary—for instance, one would expect a strong positive correlation between amount of exercise and healthiness of diet. Nonetheless, at least in practice, ℓ_1-methods still work very well in settings where the covariates are correlated and non-unitary, but currently lacking is the corresponding theory that guarantees the performance of ℓ_1-relaxations for dependent designs.

The main contribution of this paper is a direct proof that both the restricted nullspace and eigenvalue conditions hold with high probability for a broad class of dependent Gaussian design matrices. In conjunction with known results on ℓ_1-relaxation, our main result implies as corollaries that the

basis pursuit algorithm reliably recovers β^* exactly in the noiseless setting, and that in the case of observations contaminated by Gaussian noise, the Lasso and Dantzig selectors produces a solution $\widehat{\beta}$ such that $\|\widehat{\beta} - \beta^*\|_2 = O(\sqrt{\frac{k \log p}{n}})$. Our theory requires that the sample size n scale as $n = \Omega(k \log p)$, where k is the sparsity index of the regression vector β^* and p is its dimensions. For sub-linear sparsity ($k/p \to 0$), this scaling matches known optimal rates in a minimax sense for the sparse regression problem (Raskutti et al., 2009), and hence cannot be improved upon by any algorithm. The class of matrices covered by our result allows for correlation among different covariates, and hence covers many matrices for which restricted isometry or incoherence conditions fail to hold but the restricted eigenvalue condition holds. Interestingly, one can even sample the rows of the design matrix X from a multivariate Gaussian with a degenerate covariance matrix Σ, and nonetheless, our results still guarantee that the restricted nullspace and eigenvalue conditions will hold with high probability (see Section 3.3). Consequently, our results extend theoretical guarantees on ℓ_1-relaxations with optimal rates of convergence to a much broader class of random designs.

The remainder of this paper is organized as follows. We begin in Section 2 with background on sparse linear models, the basis pursuit and Lasso ℓ_1-relaxations, and sufficient conditions for their success. In Section 3, we state our main result, discuss its consequences for ℓ_1-relaxations, and illustrate it with some examples. Section 4 contains the proof of our main result, which exploits Gaussian comparison inequalities and concentration of measure for Lipschitz functions.

2. Background

We begin with background on sparse linear models and sufficient conditions for the success of ℓ_1-relaxations.

2.1 High-dimensional Sparse Models and ℓ_1-relaxation

In the classical linear model, a scalar output $y_i \in \mathbb{R}$ is linked to a p-dimensional vector $X_i \in \mathbb{R}^p$ of covariates via the relation $y_i = X_i^T \beta^* + w_i$, where w_i is a scalar observation noise. If we make a set of n such observations, then they can be written in the matrix-vector form

$$y \quad = \quad X\beta^* + w, \tag{1}$$

where $y \in \mathbb{R}^n$ is the vector of outputs, the matrix $X \in \mathbb{R}^{n \times p}$ is the set of covariates (in which row $X_i \in \mathbb{R}^p$ represents the covariates for i^{th} observation), and $w \in \mathbb{R}^n$ is a noise vector where $w \sim \mathcal{N}(0, \sigma^2 I_{n \times n})$. Given the pair (y, X), the goal is to estimate the unknown regression vector $\beta^* \in \mathbb{R}^p$.

In many applications, the linear regression model is high-dimensional in nature, meaning that the number of observations n may be substantially smaller than the number of covariates p. In this $p \gg n$ regime, it is easy to see that without further constraints on β^*, the statistical model (1) is not identifiable, since (even when $w = 0$), there are many vectors β^* that are consistent[1] with the observations y and X. This identifiability concern may be eliminated by imposing some type of sparsity assumption on the regression vector $\beta^* \in \mathbb{R}^p$. The simplest assumption is that of *exact sparsity*: in particular, we say that $\beta^* \in \mathbb{R}^p$ is s-sparse if its support set

$$S(\beta^*) := \{ j \in \{1, \ldots, p\} \mid \beta_j^* \neq 0 \}$$

1. Indeed, any vector β^* in the nullspace of X, which has dimension at least $p - n$, leads to $y = 0$ when $w = 0$.

has cardinality at most s.

Disregarding computational cost, the most direct approach to estimating an s-sparse β^* in the linear regression model would be solving a quadratic optimization problem with an ℓ_0-constraint, say

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \qquad \text{such that } \|\beta\|_0 \leq s,$$

where $\|\beta\|_0$ simply counts the number of non-zero entries in β. Of course, this problem is non-convex and combinatorial in nature, since it involves searching over all $\binom{p}{s}$ subsets of size s. A natural relaxation is to replace the non-convex ℓ_0 constraint with the ℓ_1-norm, which leads to the *constrained form of the Lasso* (Chen et al., 1998; Tibshirani, 1996), given by

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \qquad \text{such that } \|\beta\|_1 \leq R,$$

where R is a radius to be chosen by the user. Equivalently, by Lagrangian duality, this program can also be written in the penalized form

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

where $\lambda > 0$ is a regularization parameter. In the case of noiseless observations, obtained by setting $w = 0$ in the observation model (1), a closely related convex program is the *basis pursuit linear program* (Chen et al., 1998), given by

$$\widehat{\beta} \in \arg\min_{\widehat{\beta} \in \mathbb{R}^p} \|\beta\|_1 \qquad \text{such that } X\beta = y. \tag{2}$$

Chen et al. (1998) also study the constrained Lasso (2.1), which they refer to as relaxed basis pursuit. Another closely related estimator based on ℓ_1-relaxation is the Dantzig selector (Candes and Tao, 2007).

2.2 Sufficient Conditions for Success

The high-dimensional linear model under the exact sparsity constraint has been extensively analyzed. Accordingly, as we discuss here, there is a good understanding of the necessary and sufficient conditions for the success of ℓ_1-based relaxations such as basis pursuit and the Lasso.

2.2.1 RESTRICTED NULLSPACE IN NOISELESS SETTING

In the noiseless setting ($w = 0$), it is known that the basis pursuit linear program (LP) (2) recovers β^* exactly if and only if the design matrix X satisfies a restricted nullspace condition. In particular, for a given subset $S \subset \{1, \ldots, p\}$ and constant $\alpha \geq 1$, let us define the set

$$\mathcal{C}(S; \alpha) := \left\{ \theta \in \mathbb{R}^p \mid \|\theta_{S^c}\|_1 \leq \alpha \|\theta_S\|_1 \right\}.$$

For a given sparsity index $k \leq p$, we say that the matrix X satisfies the *restricted nullspace (RN) condition* of order k if $\text{null}(X) \cap \mathcal{C}(S; 1) = \{0\}$ for all subsets S of cardinality k. Although this definition appeared in earlier work (Donoho and Huo, 2001; Feuer and Nemirovski, 2003), the terminology of restricted nullspace is due to Cohen et al. (2009).

This restricted nullspace property is important, because the basis pursuit LP recovers any vector k-sparse vector β^* exactly if and only if X satisfies the restricted nullspace property of order k. See the papers (Cohen et al., 2009; Donoho and Huo, 2001; Elad and Bruckstein, 2002; Feuer and Nemirovski, 2003) for more discussion of restricted nullspaces and equivalence to exact recovery of basis pursuit.

2.2.2 RESTRICTED EIGENVALUE CONDITION FOR ℓ_2 ERROR

In the noisy setting, it is impossible to recover β^* exactly, and a more natural criterion is to bound the ℓ_2-error between β^* and an estimate $\widehat{\beta}$. Various conditions have been used to analyze the ℓ_2-norm convergence rate of ℓ_1-based methods, including the restricted isometry property (Candes and Tao, 2007), various types of restricted eigenvalue conditions (van de Geer, 2007; Bickel et al., 2009; Meinshausen and Yu, 2009), and a partial Riesz condition (Zhang and Huang, 2008). Of all these conditions, the least restrictive are the restricted eigenvalue conditions due to Bickel et al. (2009) and van de Geer (2007). As shown by Bickel et al. (2009), their restricted eigenvalue (RE) condition is less severe than both the RIP condition (Candes and Tao, 2007) and an earlier set of restricted eigenvalue conditions due to Meinshausen and Yu (2009). All of these conditions involve lower bounds on $\|X\theta\|_2$ that hold uniformly over the previously defined set $\mathcal{C}(S;\alpha)$,

Here we state a condition that is essentially equivalent to the restricted eigenvalue condition due to Bickel et al. (2009). In particular, we say that the $p \times p$ sample covariance matrix $X^T X/n$ satisfies the *restricted eigenvalue (RE) condition* over S with parameters $(\alpha,\gamma) \in [1,\infty) \times (0,\infty)$ if

$$\frac{1}{n}\theta^T X^T X\theta \;=\; \frac{1}{n}\|X\theta\|_2^2 \;\geq\; \gamma^2\,\|\theta\|_2^2 \qquad \text{for all } \theta \in \mathcal{C}(S;\alpha).$$

If this condition holds uniformly for all subsets S with cardinality k, we say that $X^T X/n$ *satisfies a restricted eigenvalue condition of order k with parameters* (α,γ). On occasion, we will also say that a deterministic $p \times p$ covariance matrix Σ satisfies an RE condition, by which we mean that $\|\Sigma^{1/2}\theta\|_2 \geq \gamma\|\theta\|_2$ for all $\theta \in \mathcal{C}(S;\alpha)$. It is straightforward to show that the RE condition for some α implies the restricted nullspace condition for the same α, so that the RE condition is slightly stronger than the RN property.

Again, the RE condition is important because it yields guarantees on the ℓ_2-error of any Lasso estimate $\widehat{\beta}$. For instance, if X satisfies the RE condition of order k with parameters $\alpha \geq 3$ and $\gamma > 0$, then it can be shown that (with appropriate choice of the regularization parameter) any Lasso estimate $\widehat{\beta}$ satisfies the error bound $\|\widehat{\beta} - \beta^*\|_2 = O(\sqrt{\frac{k\log p}{n}})$ with high probability over the Gaussian noise vector w. A similar result holds for the Dantzig selector provided the RE condition is satisfied for $\alpha \geq 1$. Bounds with this scaling have appeared in various papers on sparse linear models (Bunea et al., 2007; Bickel et al., 2009; Candes and Tao, 2007; Meinshausen and Yu, 2009; van de Geer, 2007; van de Geer and Buhlmann, 2009). Moreover, this ℓ_2-convergence rate is known to be minimax optimal (Raskutti et al., 2009) in the regime $k/p \to 0$.

3. Main Result and Its Consequences

Thus, in order to provide performance guarantees (either exact recovery or ℓ_2-error bounds) for ℓ_1-relaxations applied to sparse linear models, it is sufficient to show that the RE or RN conditions hold. Given that our interest is in providing sufficient conditions for these properties, the remainder

of the paper focuses on providing conditions for the RE condition to hold for random designs, which implies that the RN condition is satisfied.

3.1 Statement of Main Result

Our main result guarantees that the restricted eigenvalue (and hence restricted nullspace) conditions hold for a broad class of random Gaussian designs. In particular, we consider the linear model $y_i = X_i^T \beta^* + w_i$, in which each row $X_i \sim \mathcal{N}(0, \Sigma)$. We define $\rho^2(\Sigma) = \max_{j=1,\ldots,p} \Sigma_{jj}$ to be the maximal variance, and let $\Sigma^{1/2}$ denote the square root of Σ.

Theorem 1 *For any Gaussian random design $X \in \mathbb{R}^{n \times p}$ with i.i.d. $\mathcal{N}(0, \Sigma)$ rows, there are universal positive constants c, c' such that*

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Sigma^{1/2} v\|_2 - 9\rho(\Sigma)\sqrt{\frac{\log p}{n}}\,\|v\|_1 \qquad \text{for all } v \in \mathbb{R}^p, \tag{3}$$

with probability at least $1 - c' \exp(-cn)$.

The proof of this claim is given later in Section 4. Note that we have not tried to obtain sharpest possible leading constants (i.e., the factors of $1/4$ and 9 can easily be improved).

In intuitive terms, Theorem 1 provides some insight into the eigenstructure of the sample covariance matrix $\widehat{\Sigma} = X^T X/n$. One implication of the lower bound (3) is that the nullspace of X cannot contain any vectors that are "overly" sparse. In particular, for any vector $v \in \mathbb{R}^p$ such that $\|v\|_1/\|\Sigma^{1/2} v\|_2 = o(\sqrt{\frac{n}{\log p}})$, the right-hand side of the lower bound (3) will be strictly positive, showing that v cannot belong to the nullspace of X. In the following corollary, we formalize this intuition by showing how Theorem 1 guarantees that whenever the population covariance Σ satisfies the RE condition of order k, then the sample covariance $\widehat{\Sigma} = X^T X/n$ satisfies the same property as long as the sample size is sufficiently large.

Corollary 1 (Restricted eigenvalue property) *Suppose that Σ satisfies the RE condition of order k with parameters (α, γ). Then for universal positive constants c, c', c'', if the sample size satisfies*

$$n > c'' \frac{\rho^2(\Sigma)\,(1+\alpha)^2}{\gamma^2}\, k \log p, \tag{4}$$

then the matrix $\widehat{\Sigma} = X^T X/n$ satisfies the RE condition with parameters $(\alpha, \frac{\gamma}{8})$ with probability at least $1 - c' \exp(-cn)$.

Proof Let S be an arbitrary subset of cardinality k, and suppose that $v \in \mathcal{C}(S; \alpha)$. By definition, we have

$$\|v\|_1 = \|v_S\|_1 + \|v_{S^c}\|_1 \leq (1+\alpha)\|v_S\|_1,$$

and consequently $\|v\|_1 \leq (1+\alpha)\sqrt{k}\|v\|_2$. By assumption, we also have $\|\Sigma^{1/2} v\|_2 \geq \gamma \|v\|_2$ for all $v \in \mathcal{C}(S; \alpha)$. Substituting these two inequalities into the bound (3) yields

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \left\{ \frac{\gamma}{4} - 9\,(1+\alpha)\,\rho(\Sigma)\sqrt{\frac{k \log p}{n}} \right\} \|v\|_2.$$

Under the assumed scaling (4) of the sample size, we have

$$9\,(1+\alpha)\,\rho(\Sigma)\,\sqrt{\frac{k\log p}{n}} \leq \gamma/8,$$

which shows that the RE condition holds for X^TX/n with parameter $(\alpha,\gamma/8)$ as claimed. ■

Remarks:

(a) From the definitions, it is easy to see that if the RE condition holds with $\alpha = 1$ and any $\gamma > 0$ (even arbitrarily small), then the RN condition also holds. Indeed, if the matrix X^TX/n satisfies the $(1,\gamma)$-RE condition, then for any $v \in \mathcal{C}(S;1)\backslash\{0\}$, we have $\frac{\|Xv\|_2}{\sqrt{n}} \geq \gamma\|v\|_2 > 0$, which implies that $\mathcal{C}(S,1) \cap (X) = \{0\}$.

(b) As previously discussed, it is known (Bickel et al., 2009; van de Geer, 2000; van de Geer and Buhlmann, 2009) that if X^TX/n satisfies the RE condition, then the ℓ_2 error of the Lasso under the sparse linear model with Gaussian noise satisfies the bound

$$\|\widehat{\beta} - \beta^*\|_2 = O(\sqrt{\frac{k\log p}{n}}) \qquad \text{with high probability.}$$

Consequently, in order to ensure that the ℓ_2-error is bounded, the sample size must scale as $n = \Omega(k\log p)$, which matches the scaling (4) required in Corollary 1, as long as the sequence of covariance matrices Σ have diagonal entries that stay bounded.

(c) Finally, we note that Theorem 1 guarantees that the sample covariance X^TX/n satisfies a property that is slightly stronger than the RE condition. As shown by Negahban et al. (2009), this strengthening also leads to error bounds for the Lasso when β^* is not exactly k-sparse, but belongs to an ℓ_q-ball. The resulting rates are known to be minimax-optimal for these ℓ_q-balls (Raskutti et al., 2009).

3.2 Comparison to Related Work

At this point, we provide a brief comparison of our results with some related results in the literature beyond the papers discussed in the introduction. Haupt et al. (2010) showed that a certain class of random Toeplitz matrices, where the entries in the first row and first column are Bernoulli random variables and the rest fill out the Toeplitz structure satisfy RIP (and hence the weaker RE condition) provided that $n = \Omega(k^3\log(p/k))$. In Section 3.3, we demonstrate that Gausssian design matrices where the covariance matrix is a Toeplitz matrix satisfies the RE condition under the milder scaling requirement $n = \Omega(k\log(p))$. It would be of interest to determine such scaling can be established for the random Toeplitz matrices considered by Haupt et al. (2010).

It is worth comparing the scaling (4) to a related result due to van de Geer and Buhlmann (2009). In particular, their Lemma 10.1 also provides sufficient conditions for a restricted eigenvalue condition to hold for design matrices with dependent columns. They show that if the true covariance matrix satisfies an RE condition, and if the elementwise maximum $\|\widehat{\Sigma} - \Sigma\|_\infty$ between the sample covariance $\widehat{\Sigma} = X^TX/n$ and true covariance Σ is suitably bounded, then the sample covariance also satisfies the RE condition. Their result applied to the case of Gaussian random matrices guarantees

that $\widehat{\Sigma}$ satisfies the RE property as long as $n = \Omega(k^2 \log p)$ and Σ satisfies the RE condition. By contrast, Corollary 1 guarantees the RE condition with the less restrictive scaling $n = \Omega(k \log p)$. Note that if $k = O(\sqrt{n})$, our scaling condition is satisfied while their condition fails. This quadratic-linear gap in sparsity between k^2 and k arises from the difference between a local analysis (looking at individual entries of $\widehat{\Sigma}$) versus the global analysis of this paper, which studies the full random matrix. On the other hand, the result of van de Geer and Buhlmann (2009) applies more generally, including the case of sub-Gaussian random matrices (e.g., those with bounded entries) in addition to the Gaussian matrices considered in Theorem 1.

Finally, in work that followed up on the initial posting of this work (Raskutti et al., 2009), a paper by Zhou (2009) provides an extension of Theorem 1 to the case of correlated random matrices with sub-Gaussian entries. Theorem 1.6 in her paper establishes that certain families of sub-Gaussian matrices satisfy the RE condition w.h.p. with sample size $n = \Omega(s \log(p/s))$. This extension is based on techniques developed by Mendelson et al. (2008), while we use Gaussian comparison inequalities and simple concentration results for the case of Gaussian design.

3.3 Some Illustrative Examples

Let us illustrate some classes of matrices to which our theory applies. We will see that Corollary 1 applies to many sequences of covariance matrices $\Sigma = \Sigma_{p \times p}$ that have much more structure than the identity matrix. Our theory allows for the maximal eigenvalue of Σ to be arbitrarily large, or for Σ to be rank-degenerate, or for both of these degeneracies to occur at the same time. In all cases, we consider sequences of matrices for which the maximum variance $\rho^2(\Sigma) = \max_{j=1,\ldots,p} \Sigma_{jj}$ stays bounded. Under this mild restriction, we provide several examples where the RE condition is satisfied with high probability. For suitable choices, these same examples show that the RE condition can hold with high probability, even when the restricted isometry property (RIP) of Candes and Tao (2005) is violated with probability converging to one.

Example 1 (Toeplitz matrices) *Consider a covariance matrix with the Toeplitz structure $\Sigma_{ij} = a^{|i-j|}$ for some parameter $a \in [0,1)$. This type of covariance structure arises naturally from autoregressive processes, where the parameter a allows for tuning of the memory in the process. The minimum eigenvalue of Σ is $1-a > 0$, independent of the dimension p, so that the population matrix Σ clearly satisfies the RE condition. Since $\rho^2(\Sigma) = 1$, Theorem 1 implies that the sample covariance matrix $\widehat{\Sigma} = X^T X/n$ obtained by sampling from this distribution will also satisfy the RE condition with high probability as long as $n = \Omega(k \log p)$. This provides an example of a matrix family with substantial correlation between covariates for which the RE property still holds.*

However, regardless of the sample size, the submatrices of the sample covariance $\widehat{\Sigma}$ will not satisfy restricted isometry properties (RIP) if the parameter a is sufficiently large. For instance, defining $S = \{1,2,\ldots,k\}$, consider the sub-block $\widehat{\Sigma}_{SS}$ of the sample covariance matrix. Satisfying RIP requires that that the condition number $\lambda_{\max}(\widehat{\Sigma}_{SS})/\lambda_{\min}(\widehat{\Sigma}_{SS})$ be very close to one. As long as $n = \Omega(k \log p)$, known results in random matrix theory (Davidson and Szarek, 2001) guarantee that the eigenvalues of $\widehat{\Sigma}_{SS}$ will be very close to the population versions Σ_{SS}; see also the concrete calculation in Example 2 to follow. Consequently, imposing RIP amounts to requiring that the population condition number $\lambda_{\max}(\Sigma_{SS})/\lambda_{\min}(\Sigma_{SS})$ be very close to one. This condition number grows as the parameter $a \in [0,1)$ increases towards one (Gray, 1990), so RIP will be violated once $a < 1$ is sufficiently large.

We now consider a matrix family with an even higher amount of dependency among the covariates, where the RIP constants are actually unbounded as the sparsity k increases, but the RE condition is still satisfied.

Example 2 (Spiked identity model) *For a parameter $a \in [0,1)$, the spiked identity model is given by the family of covariance matrices*

$$\Sigma := (1-a) I_{p \times p} + a \vec{1}\,\vec{1}^T,$$

where $\vec{1} \in \mathbb{R}^p$ is the vector of all ones. The minimum eigenvalue of Σ is $1-a$, so that the population covariance clearly satisfies the RE condition for any fixed $a \in [0,1)$. Since this covariance matrix has maximum variance $\rho^2(\Sigma) = 1$, Corollary 1 implies that a sample covariance matrix $\widehat{\Sigma} = X^T X/n$ will satisfy the RE property with high probability with sample size $n = \Omega(k \log p)$.

On the other hand, the spiked identity matrix Σ has very poorly conditioned sub-matrices, which implies that a sample covariance matrix $\widehat{\Sigma} = X^T X/n$ will violate the restricted isometry property (RIP) with high probability as n grows. To see this fact, for an arbitrary subset S of size k, consider the associated $k \times k$ submatrix Σ_{SS}. An easy calculation shows that $\lambda_{\min}(\Sigma_{SS}) = 1-a > 0$ and $\lambda_{\max}(\Sigma_{SS}) = 1 + a(k-1)$, so that the population condition number of this sub-matrix is

$$\frac{\lambda_{\max}(\Sigma_{SS})}{\lambda_{\min}(\Sigma_{SS})} = \frac{1+a(k-1)}{1-a}.$$

For any fixed $a \in (0,1)$, this condition number diverges as k increases. We now show that the same statement applies to the sample covariance with high probability, showing that the RIP is violated. Let $u \in \mathbb{R}^k$ and $v \in \mathbb{R}^k$ denote (respectively) unit-norm eigenvectors corresponding to the minimum and maximum eigenvalues of Σ_{SS}, and define the random variables $Z_u = \|Xu\|_2^2/n$ and $Z_v = \|Xv\|_2^2/n$. Since $\langle X_i, v\rangle \sim N(0, \lambda_{\max}(\Sigma_{SS}))$ by construction, we have

$$Z_v = \frac{1}{n}\sum_{i=1}^n \langle X_i, v\rangle^2 \stackrel{d}{=} \lambda_{\max}(\Sigma_{SS}) \Big\{ \frac{1}{n}\sum_{i=1}^n y_i^2 \Big\},$$

where $y_i \sim N(0,1)$ are i.i.d. standard Gaussians, and $\stackrel{d}{=}$ denotes equality in distribution. By χ^2 tail bounds, we have $\mathbb{P}[\frac{1}{n}\sum_{i=1}^n y_i^2 \geq \frac{1}{2}] \leq c_1 \exp(-c_2 n)$, so that $Z_v \geq \lambda_{\max}(\Sigma_{SS})/2$ with high probability. A similar argument shows that $Z_u \leq 2\lambda_{\min}(\Sigma_{SS})$ with high probability, and putting together the pieces shows that w.h.p.

$$\frac{\lambda_{\max}(\widehat{\Sigma}_{SS})}{\lambda_{\min}(\widehat{\Sigma}_{SS})} \geq \frac{1}{4}\frac{\lambda_{\max}(\Sigma_{SS})}{\lambda_{\min}(\Sigma_{SS})} \geq \frac{1}{4}\,\frac{1+a(k-1)}{1-a},$$

which diverges as k increases.

In both of the preceding examples, the minimum eigenvalue of Σ was bounded from below and the diagonal entries of Σ were bounded from above, which allowed us to assert immediately that the RE condition was satisfied for the population covariance matrix. As a final example, we now consider sampling from population covariance matrices that are actually rank degenerate, but for which our theory still provides guarantees.

Example 3 (Highly degenerate covariance matrices) *Let Σ be any matrix with bounded diagonal that satisfies the RE property of some order k. Suppose that we sample n times from a $N(0,\Sigma)$ distribution, and then form the empirical covariance matrix $\widehat{\Sigma} = X^T X/n$. If $n < p$, then $\widehat{\Sigma}$ must be rank degenerate, but Corollary 1 guarantees that $\widehat{\Sigma}$ will satisfy the RE property of order k with high probability as long as $n = \Omega(k \log p)$. Moreover, by χ^2-tail bounds, the maximal diagonal element $\rho^2(\widehat{\Sigma})$ will be bounded with high probability under this same scaling.*

Now if we condition on the original design matrix X in the high probability set, we may view $\widehat{\Sigma}$ as a fixed but highly rank-degenerate matrix. Suppose that we draw a new set of n i.i.d. vectors $\widetilde{X}_i \sim N(0,\widehat{\Sigma})$ using this degenerate covariance matrix. Such a resampling procedure could be relevant for a bootstrap-type calculation for assessing errors of the Lasso. We may then form a second empirical covariance matrix $\widetilde{\Sigma} = \frac{1}{n}\widetilde{X}^T\widetilde{X}$. Conditionally on $\widehat{\Sigma}$ having the RE property of order k and a bounded diagonal, Corollary 1 shows that the resampled empirical covariance $\widetilde{\Sigma}$ will also have the RE property of order k with high probability, again for $n = \Omega(k \log p)$.

This simple example shows that in the high-dimensional setting $p \gg n$, it is possible for the RE condition to hold with high probability even when the original population covariance matrix ($\widehat{\Sigma}$ in this example) has a $p-n$-dimensional nullspace. Note moreover that this is not an isolated phenomenon—rather, it will hold for almost every sample covariance matrix $\widehat{\Sigma}$ constructed in the way that we have described.

4. Proof of Theorem 1

We now turn to the proof of Theorem 1. The main ingredients are the Gordon-Slepian comparison inequalities (Gordon, 1985) for Gaussian processes, concentration of measure for Lipschitz functions (Ledoux, 2001), and a peeling argument. The first two ingredients underlie classical proofs on the ordinary eigenvalues of Gaussian random matrices (Davidson and Szarek, 2001), whereas the latter tool is used in empirical process theory (van de Geer, 2000).

4.1 Proof Outline

Recall that Theorem 1 states that the condition

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Sigma^{1/2}v\|_2 - 9\rho(\Sigma)\sqrt{\frac{\log p}{n}}\,\|v\|_1 \qquad \text{for all } v \in \mathbb{R}^p, \tag{5}$$

holds with probability at least $1 - c'\exp(-cn)$, where c, c' are universal positive constants. Hence, we are bounding the random quantity $\|Xv\|_2$ in terms of $\|\Sigma^{1/2}v\|_2$ and $\|v\|_1$ for all v with high probability. It suffices to prove Theorem 1 for $\|\Sigma^{1/2}v\|_2 = 1$. Indeed, for any vector $v \in \mathbb{R}^p$ such that $\Sigma^{1/2}v = 0$, the claim holds holds trivially. Otherwise, we may consider the re-scaled vector $\check{v} = v/\|\Sigma^{1/2}v\|_2$, and note that $\|\Sigma^{1/2}\check{v}\|_2 = 1$ by construction. By scale invariance of the condition (5), if it holds for the re-scaled vector $\check{v}$, it also holds for v.

Therefore, in the remainder of the proof, our goal is to lower bound the quantity $\|Xv\|_2$ over the set of v such that $\|\Sigma^{1/2}v\|_2 = 1$ in terms of $\|v\|_1$. At a high level, there are three main steps to the proof:

(1) We begin by considering the set $V(r) := \{v \in \mathbb{R}^p \mid \|\Sigma^{1/2}v\|_2 = 1, \|v\|_1 \leq r\}$, for a fixed radius r. Although this set may be empty for certain choices of $r > 0$, our analysis only concerns

those choices for which it is non-empty. Define the random variable

$$M(r,X) := 1 - \inf_{v \in V(r)} \frac{\|Xv\|_2}{\sqrt{n}} = \sup_{v \in V(r)} \left\{ 1 - \frac{\|Xv\|_2}{\sqrt{n}} \right\}.$$

Our first step is to upper bound the expectation $\mathbb{E}[M(r,X)]$, where the expectation is taken over the random Gaussian matrix X.

(2) Second, we establish that $M(r,X)$ is a Lipschitz function of its Gaussian arguments, and then use concentration inequalities to assert that for each fixed $r > 0$, the random variable $M(r,X)$ is sharply concentrated around its expectation with high probability.

(3) Third, we use a peeling argument to show that our analysis holds with high probability and uniformly over all possible choice of the ℓ_1-radius r, which then implies that the condition (5) holds with high probability as claimed.

In the remainder of this section, we provide the details of each of these steps.

4.2 Bounding the Expectation $\mathbb{E}[M(r,X)]$

This subsection is devoted to a proof of the following lemma:

Lemma 1 *For any radius $r > 0$ such that $V(r)$ is non-empty, we have*

$$\mathbb{E}[M(r,X)] \leq \frac{1}{4} + 3\rho(\Sigma)\sqrt{\frac{\log p}{n}}\, r.$$

Proof : Let $S^{n-1} = \{u \in \mathbb{R}^n \mid \|u\|_2 = 1\}$ be the Euclidean sphere of radius 1, and recall the previously defined set $V(r) := \{v \in \mathbb{R}^p \mid \|\Sigma^{1/2} v\|_2 = 1, \|v\|_1 \leq r\}$. For each pair $(u,v) \in S^{n-1} \times V(r)$, we may define an associated zero-mean Gaussian random variable $Y_{u,v} := u^T X v$. This representation is useful, because it allows us to write the quantity of interest as a min-max problem in terms of this Gaussian process. In particular, we have

$$-\inf_{v \in V(r)} \|Xv\|_2 = -\inf_{v \in V(r)} \sup_{u \in S^{n-1}} u^T X v = \sup_{v \in V(r)} \inf_{u \in S^{n-1}} u^T X v.$$

We may now upper bound the expected value of the above quantity via a Gaussian comparison inequality; here we state a form of Gordon's inequality used in past work on Gaussian random matrices (Davidson and Szarek, 2001). Suppose that $\{Y_{u,v}, (u,v) \in U \times V\}$ and $\{Z_{u,v}, (u,v) \in U \times V\}$ are two zero-mean Gaussian processes on $U \times V$. Using $\sigma(\cdot)$ to denote the standard deviation of its argument, suppose that these two processes satisfy the inequality

$$\sigma(Y_{u,v} - Y_{u',v'}) \leq \sigma(Z_{u,v} - Z_{u',v'}) \qquad \text{for all pairs } (u,v) \text{ and } (u',v') \text{ in } U \times V,$$

and this inequality holds with equality when $v = v'$. Then we are guaranteed that

$$\mathbb{E}[\sup_{v \in V} \inf_{u \in U} Y_{u,v}] \quad \leq \quad \mathbb{E}[\sup_{v \in V} \inf_{u \in U} Z_{u,v}].$$

We use Gordon's inequality to show that

$$\mathbb{E}[M(r,X)] = 1 + \mathbb{E}[\sup_{v \in V(r)} \inf_{u \in S^{n-1}} Y_{u,v}] \leq 1 + \mathbb{E}[\sup_{v \in V(r)} \inf_{u \in S^{n-1}} Z_{u,v}],$$

where we recall that $Y_{u,v} = u^T X v$ and $Z_{u,v}$ is a different Gaussian process to be defined shortly.

We begin by computing $\sigma^2(Y_{u,v} - Y_{u',v'})$. To simplify notation, we note that the $X \in \mathbb{R}^{n\times p}$ can be written as $W\Sigma^{1/2}$, where $W \in \mathbb{R}^{n\times p}$ is a matrix with i.i.d. $\mathcal{N}(0,1)$ entries, and $\Sigma^{1/2}$ is the symmetric matrix square root. In terms of W, we can write

$$Y_{u,v} = u^T W \Sigma^{1/2} v \;=\; u^T W \tilde{v},$$

where $\tilde{v} = \Sigma^{1/2} v$. It follows that

$$\sigma^2(Y_{u,v} - Y_{u',v'}) := \mathbb{E}\big(\sum_{i=1}^n \sum_{j=1}^p W_{i,j}(u_i\tilde{v}_j - u'_i\tilde{v}'_j)\big)^2 \;=\; |||u\tilde{v}^T - (u')(\tilde{v}')^T|||_F^2,$$

where $|||\cdot|||_F$ is the Frobenius norm (ℓ_2-norm applied elementwise to the matrix). This equality follows immediately since the W_{ij} variables are i.i.d $\mathcal{N}(0,1)$.

Now consider a second zero-mean Gaussian process $Z_{u,v}$ indexed by $S^{n-1} \times V(r)$, and given by

$$Z_{u,v} \;\;=\;\; \vec{g}^T u + \vec{h}^T \Sigma^{1/2} v,$$

where $\vec{g} \sim N(0, I_{n\times n})$ and $\vec{h} \sim N(0, I_{p\times p})$ are standard Gaussian random vectors. With $\tilde{v} = \Sigma^{1/2} v$, we see immediately that

$$\sigma^2(Z_{u,v} - Z_{u',v'}) = \|u - u'\|_2^2 + \|\tilde{v} - \tilde{v}'\|_2^2.$$

Consequently, in order to apply the Gaussian comparison principle to $\{Y_{u,v}\}$ and $\{Z_{u,v}\}$, we need to show that

$$|||u\tilde{v}^T - (u')(\tilde{v}')^T|||_F^2 \leq \|u - u'\|_2^2 + \|\tilde{v} - \tilde{v}'\|_2^2 \tag{6}$$

for all pairs $(u,\tilde{v})$ and $(u',\tilde{v}')$ in the set of interest. Since the Frobenius norm $|||\cdot|||_F$ is simply the ℓ_2-norm on the vectorized form of a matrix, we can compute

$$\begin{aligned}
|||u\tilde{v}^T - u'(\tilde{v}')^T|||_F^2 &= |||(u-u')\tilde{v}^T + u'(\tilde{v} - \tilde{v}')^T|||_F^2 \\
&= \sum_{i=1}^n \sum_{j=1}^p [(u_i - u'_i)\tilde{v}_j + u'_i(\tilde{v}_j - \tilde{v}'_j)]^2 \\
&= \|\tilde{v}\|_2^2 \|u - u'\|_2^2 + \|u'\|_2^2 \|\tilde{v} - \tilde{v}'\|_2^2 + 2(u^T u' - \|u'\|_2^2)(\|\tilde{v}\|_2^2 - \tilde{v}^T\tilde{v}') \\
&= \|u - u'\|_2^2 + \|\tilde{v} - \tilde{v}'\|_2^2 - 2(\|u'\|_2^2 - u^T u')(\|\tilde{v}\|_2^2 - \tilde{v}^T\tilde{v}'),
\end{aligned}$$

where we have used equalities $\|u\|_2 = \|u'\|_2 = 1$ and $\|\tilde{v}\|_2 = \|\tilde{v}'\|_2 = 1$. By the Cauchy-Schwarz inequality, we have $\|u\|_2^2 - u^T u' \geq 0$, and $\|\tilde{v}\|_2^2 - \tilde{v}^T\tilde{v}' \geq 0$, from which the claimed inequality (6) follows. When $v = v'$, we also have $\tilde{v} = \Sigma^{1/2} v = \Sigma^{1/2} v' = \tilde{v}'$, so that equality holds in the condition (6) when $\tilde{v} = \tilde{v}'$.

Consequently, we may apply Gordon's inequality to conclude that

$$\begin{aligned}
\mathbb{E}\big[\sup_{v\in V(r)} \inf_{u\in S^{n-1}} u^T X v\big] &\leq \mathbb{E}\big[\sup_{v\in V(r)} \inf_{u\in S^{n-1}} Z_{u,v}\big] \\
&= \mathbb{E}[\inf_{u\in S^{n-1}} \vec{g}^T u] + \mathbb{E}[\sup_{v\in V(r)} \vec{h}^T \Sigma^{1/2} v] \\
&= -\mathbb{E}[\|\vec{g}\|_2] + \mathbb{E}[\sup_{v\in V(r)} \vec{h}^T \Sigma^{1/2} v].
\end{aligned}$$

We now observe that by definition of $V(r)$, we have

$$\sup_{v \in V(r)} |\vec{h}^T \Sigma^{1/2} v| \leq \sup_{v \in V(r)} \|v\|_1 \, \|\Sigma^{1/2}\vec{h}\|_\infty \leq r \|\Sigma^{1/2}\vec{h}\|_\infty.$$

Each element $(\Sigma^{1/2}\vec{h})_j$ is zero-mean Gaussian with variance Σ_{jj}. Consequently, known results on Gaussian maxima (cf. Ledoux and Talagrand, 1991, Equation (3.13)) imply that $\mathbb{E}[\|\Sigma^{1/2}h\|_\infty] \leq 3\sqrt{\rho^2(\Sigma)\log p}$, where $\rho^2(\Sigma) = \max_j \Sigma_{jj}$. Noting[2] that $\mathbb{E}[\|\vec{g}\|_2] \geq \frac{3}{4}\sqrt{n}$ for all $n \geq 10$ by standard χ^2 tail bounds and putting together the pieces, we obtain the bound

$$\mathbb{E}[-\inf_{v \in V(r)} \|Xv\|_2] \quad \leq \quad -\frac{3}{4}\sqrt{n} + 3\big[\rho^2(\Sigma)\log p\big]^{1/2} r.$$

Dividing by $\sqrt{n}$ and adding 1 to both sides yields

$$\mathbb{E}[M(r,X)] = \mathbb{E}[1 - \inf_{v \in V(r)} \|Xv\|_2/\sqrt{n}] \leq 1/4 + 3\rho(\Sigma)\sqrt{\frac{\log p}{n}}\, r,$$

as claimed.

■

4.3 Concentration Around the Mean for $M(r,X)$

Having controlled the expectation, the next step is to establish concentration of $M(r,X)$ around its mean. Note that Lemma 1 shows that $\mathbb{E}[M(r,X)] \leq t(r)$, where

$$t(r) := \frac{1}{4} + 3r\rho(\Sigma)\sqrt{\frac{\log p}{n}}. \tag{7}$$

Now we prove the following claim:

Lemma 2 *For any r such that $V(r)$ is non-empty, we have*

$$\mathbb{P}\left[M(r,X) \geq \frac{3t(r)}{2}\right] \leq 2\exp(-nt^2(r)/8).$$

Proof In order to prove this lemma, it suffices to show that

$$\mathbb{P}\big[|M(r,X) - \mathbb{E}[M(r,X)]| \geq t(r)/2\big] \leq 2\exp(-nt^2(r)/8),$$

and use the upper bound on $\mathbb{E}[M(r,X)]$ derived in Lemma 1.

By concentration of measure for Lipschitz functions of Gaussians (see Appendix B), this tail bound will follow if we show that the Lipschitz constant of $M(r,X)$ as a function of the Gaussian random matrix is less than $1/\sqrt{n}$. To make this functional dependence explicit, let us write $M(r,X)$ as the function $h(W) = \sup_{v \in V(r)} \big(1 - \|W\Sigma^{1/2}v\|_2/\sqrt{n}\big)$. We find that

$$\sqrt{n}\big[h(W) - h(W')\big] = \sup_{v \in V(r)} -\|W\Sigma^{1/2}v\|_2 - \sup_{v \in V(r)} -\|W'\Sigma^{1/2}v\|_2.$$

2. In fact, $|\mathbb{E}[\|\vec{g}\|_2] - \sqrt{n}| = o(\sqrt{n})$, but this simple bound is sufficient for our purposes.

Since $V(r)$ is closed and bounded and the objective function is continuous, there exists $\hat{v} \in V(r)$ such that $\hat{v} = \arg\max_{v \in V(r)} -\|W\Sigma^{1/2}v\|_2$. Therefore

$$\begin{aligned}\sup_{v\in V(r)} \big(-\|W\Sigma^{1/2}v\|_2\big) - \sup_{v\in V(r)} \big(-\|W'\Sigma^{1/2}v\|_2\big) &= -\|W\Sigma^{1/2}\hat{v}\|_2 - \sup_{v\in V(r)} \big(-\|W'\Sigma^{1/2}v\|_2\big) \\ &\leq \|W'\Sigma^{1/2}\hat{v}\|_2 - \|W\Sigma^{1/2}\hat{v}\|_2 \\ &\leq \sup_{v\in V(r)} \big(\|(W'-W)\Sigma^{1/2}v\|_2\big).\end{aligned}$$

For a matrix A, we define its spectral norm $|||A|||_2 = \sup_{\|u\|_2=1} \|Au\|_2$. With this notation, we can bound the Lipschitz constant of h as

$$\begin{aligned}\sqrt{n}\big[h(W) - h(W')\big] &\leq \sup_{v\in V(r)} \big(\|(W-W')\Sigma^{1/2}v\|_2\big) \\ &\overset{(a)}{\leq} \big\{\sup_{v\in V(r)} \big(\|\Sigma^{1/2}v\|_2\big)\big\} |||(W-W')|||_2 \\ &\overset{(b)}{\leq} \big\{\sup_{v\in V(r)} \big(\|\Sigma^{1/2}v\|_2\big)\big\}\, |||(W-W')|||_F \\ &\overset{(c)}{=} |||W-W'|||_F.\end{aligned}$$

In this argument, inequality (a) follows by definition of the matrix spectral norm $|||\cdot|||_2$; inequality (b) follows from the bound $|||(W-W')|||_2 \leq |||(W-W')|||_F$ between the spectral and Frobenius matrix norms (Horn and Johnson, 1985); and equality (c) follows since $\|\Sigma^{1/2}v\|_2 = 1$ for all $v \in V(r)$. Thus, we have shown that h has Lipschitz constant $L \leq 1/\sqrt{n}$ with respect to the Euclidean norm on W (viewed as a vector with np entries). Finally we use a standard result on the concentration for Lipschitz functions of Gaussian random variables (Ledoux, 2001; Massart, 2003)—see Appendix B for one statement. Applying the concentration result (9) with $m = np$, $\tilde{g} = W$, and $t = t(r)/2$ completes the proof. ■

4.4 Extension to All Vectors Via Peeling

Thus far, we have shown that

$$M(r,X) = 1 - \inf_{v\in V(r)} \frac{\|Xv\|_2}{\sqrt{n}} = \sup_{v\in V(r)} \left\{1 - \frac{\|Xv\|_2}{\sqrt{n}}\right\} \geq 3t(r)/2, \tag{8}$$

with probability no larger than $2\exp(-nt^2(r)/8)$ where $t(r) = \frac{1}{4} + 3r\,\rho(\Sigma)\sqrt{\frac{\log p}{n}}$. The set $V(r)$ requires that $\|v\|_1 \leq r$ for some *fixed* radius r, whereas the claim of Theorem 1 applies to all vectors v. Consequently, we need to extend the bound (8) to an arbitrary ℓ_1 radius.

In order do so, we define the event

$$\mathcal{T} := \big\{\exists\, v \in \mathbb{R}^p \text{ s.t. } \|\Sigma^{1/2}v\|_2 = 1 \text{ and } \big(1 - \|Xv\|_2/\sqrt{n}\big) \geq 3t(\|v\|_1)\big\}.$$

Note that there is no r in the definition of $\mathcal{T}$, because we are setting $\|v\|_1$ to be the argument of the function t. We claim that there are constants positive constants c, c' such that $\mathbb{P}[\mathcal{T}] \leq c\exp(-c'n)$,

from which Theorem 1 will follow. We establish this claim by using a device known as peeling (Alexander, 1985; van de Geer, 2000); for the version used here, see Lemma 3 proved in the Appendix. In particular, we apply Lemma 3 with the functions

$$f(v,X) = 1 - \|Xv\|_2/\sqrt{n}, \qquad h(v) = \|v\|_1, \quad \text{and} \quad g(r) = 3t(r)/2,$$

the sequence $a_n = n$, and the set $A = \{v \in \mathbb{R}^p \mid \|\Sigma^{1/2}v\|_2 = 1\}$. Recall that the quantity t, as previously defined (7), satisfies $t(r) \geq 1/4$ for all $r > 0$ and is strictly increasing. Therefore, the function $g(r) = 3t(r)/2$ is non-negative and strictly increasing as a function of r, and moreover satisfies $g(r) \geq 3/8$, so that Lemma 3 is applicable with $\mu = 3/8$. We can thus conclude that $\mathbb{P}[\mathcal{T}^c] \geq 1 - c\exp(-c'n)$ for some numerical constants c and c'.

Finally, conditioned on the event $\mathcal{T}^c$, for all $v \in \mathbb{R}^p$ with $\|\Sigma^{1/2}v\|_2 = 1$, we have

$$1 - \|Xv\|_2/\sqrt{n} \;\leq\; 3t(\|v\|_1) \;=\; \frac{3}{4} + 9\,\|v\|_1\,\rho(\Sigma)\sqrt{\frac{\log p}{n}},$$

which implies that

$$\|Xv\|_2/\sqrt{n} \geq \frac{1}{4} - 9\,\|v\|_1\,\rho(\Sigma)\,\sqrt{\frac{\log p}{n}}.$$

As noted in the proof outline, this suffices to establish the general claim.

5. Conclusion

Methods based on ℓ_1-relaxations are very popular, and the weakest possible conditions on the design matrix X required to provide performance guarantees—namely, the restricted nullspace and eigenvalue conditions—are well-understood. In this paper, we have proved that these conditions hold with high probability for a broad class of Gaussian design matrices allowing for quite general dependency among the columns, as captured by a covariance matrix Σ representing the dependence among the different covariates. As a corollary, our result guarantees that known performance guarantees for ℓ_1-relaxations such as basis pursuit and Lasso hold with high probability for such problems, provided the population matrix Σ satisfies the RE condition. Interestingly, our theory shows that ℓ_1-methods can perform well when the covariates are sampled from a Gaussian distribution with a degenerate covariance matrix. Some follow-up work (Zhou, 2009) has extended these results to random matrices with sub-Gaussian rows. In addition, there are a number of other ways in which this work could be extended. One is to incorporate additional dependence across the rows of the design matrix, as would arise in modeling time series data for example. It would also be interesting to relate the allowable degeneracy structures of Σ to applications involving real data. Finally, although this paper provides various conditions under which the RE condition holds with high probability, it does not address the issue of how to determine whether a given sample covariance matrix matrix $\widehat{\Sigma} = X^T X/n$ satisfies the RE condition. It would be interesting to study if there are computationally efficient methods for verifying the RE condition.

Acknowledgments

We thank Arash Amini for useful discussion, particularly regarding the proofs of Theorem 1 and Lemma 3, and Rob Nowak for helpful comments on an earlier draft. This work was partially

supported by NSF grants DMS-0605165 and DMS-0907632 to MJW and BY. In addition, BY was partially supported by the NSF grant SES-0835531 (CDI), and a grant from the MSRA. MJW was supported by an Sloan Foundation Fellowship and AFOSR Grant FA9550-09-1-0466. During this work, GR was financially supported by a Berkeley Graduate Fellowship.

Appendix A. Peeling Argument

In this appendix, we state a result on large deviations of the constrained optimum of random objective functions of the form $f(v;X)$, where $v \in \mathbb{R}^p$ is the vector to be optimized over, and X is some random vector. Of interest is the problem $\sup_{h(v)\leq r,\, v\in A} f(v;X)$, where $h : \mathbb{R}^p \to \mathbb{R}_+$ is some non-negative and increasing constraint function, and A is a non-empty set. With this set-up, our goal is to bound the probability of the event defined by

$$\mathcal{E} \quad := \quad \big\{\exists\, v \in A \text{ such that } f(v;X) \geq 2g(h(v)))\big\},$$

where $g : \mathbb{R} \to \mathbb{R}$ is non-negative and strictly increasing.

Lemma 3 *Suppose that $g(r) \geq \mu$ for all $r \geq 0$, and that there exists some constant $c > 0$ such that for all $r > 0$, we have the tail bound*

$$\mathbb{P}\big[\sup_{v\in A,\, h(v)\leq r} f(v;X) \geq g(r)\big] \quad \leq \quad 2\exp(-c\, a_n\, g^2(r)),$$

for some $a_n > 0$. Then we have

$$\mathbb{P}[\mathcal{E}] \quad \leq \quad \frac{2\exp(-4ca_n\mu^2)}{1-\exp(-4ca_n\mu^2)}.$$

Proof : Our proof is based on a standard peeling technique (e.g., see van de Geer, 2000, p. 82). By assumption, as v varies over A, we have $g(r) \in [\mu,\infty)$. Accordingly, for $m = 1,2,\ldots$, defining the sets

$$A_m \quad := \quad \big\{v \in A \mid 2^{m-1}\mu \leq g(h(v)) \leq 2^m\mu\big\},$$

we may conclude that if there exists $v \in A$ such that $f(v,X) \geq 2g(h(v))$, then this must occur for some m and $v \in A_m$. By union bound, we have

$$\mathbb{P}[\mathcal{E}] \quad \leq \quad \sum_{m=1}^{\infty} \mathbb{P}\big[\exists\, v \in A_m \text{ such that } f(v,X) \geq 2g(h(v))\big].$$

If $v \in A_m$ and $f(v,X) \geq 2g(h(v))$, then by definition of A_m, we have $f(v,X) \geq 2\,(2^{m-1})\,\mu = 2^m\mu$. Since for any $v \in A_m$, we have $g(h(v)) \leq 2^m\mu$, we combine these inequalities to obtain

$$\begin{aligned}
\mathbb{P}[\mathcal{E}] \quad &\leq \quad \sum_{m=1}^{\infty} \mathbb{P}\big[\sup_{h(v)\leq g^{-1}(2^m\mu)} f(v,X) \geq 2^m\mu\big] \\
&\leq \quad \sum_{m=1}^{\infty} 2\exp\big(-ca_n\,[g(g^{-1}(2^m\mu))]^2\big) \\
&= \quad 2\sum_{m=1}^{\infty} \exp\big(-ca_n\,2^{2m}\mu^2\big),
\end{aligned}$$

from which the stated claim follows by upper bounding this geometric sum. ■

Appendix B. Concentration for Gaussian Lipschitz Functions

We say that a function $F : \mathbb{R}^m \to \mathbb{R}$ is Lipschitz with constant L if $|F(x) - F(y)| \leq L\|x-y\|_2$ for all $x, y \in \mathbb{R}^m$. It is a classical fact that Lipschitz functions of standard Gaussian vectors exhibit Gaussian concentration. We summarize one version of this fact in the following:

Theorem 2 (Theorem 3.8 from Massart 2003) *Let $w \sim \mathcal{N}(0, I_{m\times m})$ be an m-dimensional Gaussian random variable. Then for any L-Lipschitz function F, we have*

$$\mathbb{P}\left[|F(w) - \mathbb{E}[F(w)]| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2L^2}\right), \text{ for all } t \geq 0. \tag{9}$$

This result can be interpreted as saying that in terms of tail behavior, the random variable $F(w) - \mathbb{E}[F(w)]$ behaves like a zero-mean Gaussian with variance L^2.

References

R. Adamczak, A. Litvak, N. Tomczak-Jaegermann, and A. Pajor. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. Technical report, University of Alberta, 2009.

K. S. Alexander. Rates of growth for weighted empirical processes. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pages 475–493. UC Press, Berkeley, 1985.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

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

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

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.

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

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. of. American Mathematical Society*, 22(1):211–231, January 2009.

K. R. Davidson and S. J. Szarek. Local operator theory, random matrices, and Banach spaces. In *Handbook of Banach Spaces*, volume 1, pages 317–336. Elsevier, Amsterdan, NL, 2001.

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

D. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Info Theory*, 47(7):2845–2862, 2001.

M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Info Theory*, 48(9):2558–2567, September 2002.

A. Feuer and A. Nemirovski. On sparse representation in pairs of bases. *IEEE Trans. Info Theory*, 49(6):1579–1581, 2003.

Y. Gordon. Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

R. M. Gray. Toeplitz and Circulant Matrices: A Review. Technical report, Stanford University, Information Systems Laboratory, 1990.

O. Guédon, S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Subspaces and orthogonal decompositions generated by bounded orthogonal systems. *Journal of Positivity*, 11(2):269–283, 2007.

O. Guédon, S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Majorizing measures and proportional subsets of bounded orthonormal systems. *Journal of Rev. Mat. Iberoam*, 24(3):1075–1095, 2008.

J. Haupt, W. U. Bajwa, G. Raz, and R. Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. Technical report, University of Wisconsin-Madison, 2010.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

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

P. Massart. *Concentration Inequalties and Model Selection*. Ecole d'Eté de Probabilités, Saint-Flour. Springer, New York, 2003.

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

S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Journal of Constr. Approx.*, 28(3):277–289, 2008.

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

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. Technical report, U. C. Berkeley, October 2009. Posted as http://arxiv.org/abs/0910.2042.

Justin Romberg. Compressive sensing by random convolution. *SIAM Journal of Imaging Science*, 2(4):1098–1128, 2009.

M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Comm. Pure and Appl. Math.*, 61(8):1025–1045, 2008.

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

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

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

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.

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.

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