Estimation with Norm Regularization

Arindam Banerjee         Sheng Chen         Farideh Fazayeli        Vidyashankar Sivakumar
{banerjee,shengc,farideh,sivakuma@cs.umn.edu}

Department of Computer Science & Engineering
University of Minnesota, Twin Cities

May 19, 2015

Abstract

Analysis of non-asymptotic estimation error and structured statistical recovery based on norm regularized regression, such as Lasso, needs to consider four aspects: the norm, the loss function, the design matrix, and the noise model. This paper presents generalizations of such estimation error analysis on all four aspects compared to the existing literature. We characterize the restricted error set where the estimation error vector lies, establish relations between error sets for the constrained and regularized problems, and present an estimation error bound applicable to any norm. Precise characterizations of the bound is presented for a variety of design matrices, including sub-Gaussian, anisotropic, and correlated designs, noise models, including both Gaussian and sub-Gaussian noise, and loss functions, including least squares and generalized linear models. A key result from the analysis is that the sample complexity of all such estimators depends on the Gaussian width of the spherical cap corresponding to the restricted error set. Further, once the number of samples \( n \) crosses the required sample complexity, the estimation error decreases as \( \frac{c}{\sqrt{n}} \), where \( c \) depends on the Gaussian width of the unit norm ball.

1 Introduction

Over the past decade, progress has been made in developing non-asymptotic bounds on the estimation error of structured parameters based on norm regularized regression. Such estimators are usually of the form \( \hat{\theta}_{\lambda_n} = \arg\min_{\theta \in \mathbb{R}^p} \mathcal{L}(\theta; Z^n) + \lambda_n R(\theta) \), where \( R(\theta) \) is a suitable norm, \( \mathcal{L}(\cdot) \) is a suitable loss function, \( Z^n = \{(y_i, X_i)\}_{i=1}^n \) where \( y_i \in \mathbb{R}, X_i \in \mathbb{R}^p \) is the training set, and \( \lambda_n > 0 \) is a regularization parameter. The optimal parameter \( \theta^* \) is often assumed to be ‘structured,’ usually characterized or approximated as a small value according to some norm \( R(\cdot) \). Recent work has viewed such characterizations in terms of atomic norms, which give the tightest convex relaxation of a structured set of atoms in which \( \theta^* \) belongs \[14\]. Since \( \hat{\theta}_{\lambda_n} \) is an estimate of the optimal structure \( \theta^* \), the focus has been on bounding a suitable measure of the error vector \( \Delta_n = (\hat{\theta}_{\lambda_n} - \theta^*) \), e.g., the \( L_2 \) norm \( \|\Delta_n\|_2 \).

To understand the state-of-the-art on non-asymptotic bounds on the estimation error for norm-regularized regression, four aspects of \[1\] need to be considered: (i) the norm \( R(\theta) \), (ii) properties of the design matrix \( X = [X_1, \ldots, X_n]^T \in \mathbb{R}^{n \times p} \), (iii) the loss function \( \mathcal{L}(\cdot) \), and (iv) the noise model, typically in terms of \( \omega_i = y_i - E[y_i|X_i] \). Most of the literature has focused on a linear model: \( y = X\theta + \omega \), and a squared-loss function: \( \mathcal{L}(\theta; Z^n) = \frac{1}{n} \|y - X\theta\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \theta, X_i \rangle)^2 \). Early work on such estimators focused on
the $L_1$ norm \cite{43,41,24}, and led to sufficient conditions on the design matrix $X$, including the restricted-isometry properties (RIP) \cite{13,12} and restricted eigenvalue (RE) conditions \cite{7,26,31}. While much of the development has focussed on isotropic Gaussian design matrices, recent work has extended the analysis for $L_1$ norm to correlated Gaussian designs \cite{31} as well as anisotropic sub-Gaussian design matrices \cite{32}.

Building on such development, \cite{26} presents a unified framework for the case of decomposable norms and also considers generalized linear models (GLMs) for certain norms such as $L_1$. Two key insights are offered in \cite{26}: first, the error vector $\hat{\Delta}_n$ lies in a restricted set, a cone or a star, for suitably large $\lambda_n$, and second, the loss function needs to satisfy restricted strong convexity (RSC), a generalization of the RE condition, on the restricted error set for the analysis to work out.

For isotropic Gaussian design matrices, additional progress has been made. \cite{14} considers a constrained estimation formulation for all atomic norms, where the gain condition, equivalent to the RE condition, uses Gordon’s inequality \cite{18,19,23} and is succinctly represented in terms of the Gaussian width of the intersection of the cone of the error set and a unit ball/sphere. \cite{29} considers three related formulations for generalized Lasso problems, establish recovery guarantees based on Gordon’s inequality, and quantities related to the Gaussian width. Sharper analysis for recovery has been considered in \cite{2}, yielding a precise characterization of phase transition behavior using quantities related to the Gaussian width. \cite{30} consider a linear programming estimator in a 1-bit compressed sensing setting and, interestingly, the concept of Gaussian width shows up in the analysis. In spite of the advances, with a few notable exceptions \cite{36,39}, most existing results are restricted to isotropic Gaussian design matrices. Further, while a suitable scale for $\lambda_n$ is known for special cases such as the $L_1$, a general analysis applicable to any norm $R(\cdot)$ has not been explored in the literature.

In this paper, we consider structured estimation problems with norm regularization of the form \cite{1}, and present a unified analysis which substantially generalizes existing results on all four pertinent aspects: the norm, the design matrix, the loss, and the noise model. The analysis we present applies to all norms, and the results can be divided into three groups: characterization of the error set and recovery guarantees, characterization of the regularization parameter $\lambda_n$, and characterization of the restricted eigenvalue conditions or restricted strong convexity. We provide a summary of the key results below.

**Restricted error set and recovery guarantee:** We start with a characterization of the error set $E_r$ in which the error vector $\hat{\Delta}_n$ belongs. For a suitably large $\lambda_n$, we show that $\hat{\Delta}_n$ belongs to the restricted error set

$$E_r = \left\{ \Delta \in \mathbb{R}^p \mid R(\theta^* + \Delta) \leq R(\theta^*) + \frac{1}{\beta} R(\Delta) \right\},$$

where $\beta > 1$ is a constant. The restricted error set has interesting structure, and forms the basis of subsequent analysis for bounds on $\|\hat{\Delta}_n\|_2$. As an alternative to regularized estimators, the literature has considered constrained estimators which directly focus on minimizing $R(\theta)$ under suitable constraints determined by the noise $(y - X\theta)$ and/or the design matrix $X$ \cite{13,7,14,15}. A recent example of such a constrained estimator is the generalized Dantzig selector (GDS) \cite{15}, which generalizes the Dantzig selector \cite{11} corresponding to the $L_1$ norm, and is given by:

$$\hat{\theta}_{\varphi_n} = \arg\min_{\theta \in \mathbb{R}^p} R(\theta) \quad \text{s.t.} \quad R^*(X^T(y - X\theta^*)) \leq \varphi_n,$$

where $R^*(\cdot)$ denotes the dual norm of $R(\cdot)$. One can show \cite{14,15} that the restricted error set for such constrained estimators are of the form:

$$E_c = \{ \Delta \in \mathbb{R}^p \mid R(\theta^* + \Delta) \leq R(\theta^*) \}.$$

One can readily see that $E_r$ is larger than $E_c$, i.e., $E_c \subseteq E_r$, and $E_r$ approaches $E_c$ as $\beta$ increases. We establish a geometric relationship between the two sets, which will possibly help in transforming analysis
done on regularized estimators as in [1] to corresponding constrained estimators as in [3] and vice versa. Let $\rho B^p_2$ denote a $L_2$ ball of any radius $\rho$ in $\mathbb{R}^p$. Then, with $A_r = E_r \cap \rho B^p_2$ and $A_c = \text{cone}(E_c) \cap \rho B^p_2$, we show that

$$w(A_r) \leq \left( 1 + \frac{2\|\theta^*\|_2}{\rho} \right) w(A_c),$$  

(5)

where $w(A) = \mathbb{E}_g[\sup_{\alpha \in A} \langle a, g \rangle]$, with $g$ being an isotropic Gaussian vector, denotes the Gaussian width\(^1\) of the set $A$. For example, when $\|\theta^*\|_2 = 1$ and one chooses $\rho = 1, \beta = 2$, we have $w(A_r) \leq 3w(A_c)$. Note that $A_r$ corresponds to the spherical cap of the error set $E_r$ at radius $\rho$, and $A_c$ corresponds to the spherical cap of the error cone $\text{cone}(E_c)$ at the same radius.

For the special case of $L_1$ norm, [7] considered a simultaneous analysis of the Lasso and the Dantzig selector, and characterized the structure of the error sets for regularized and constrained sets for the special case of $L_1$ norm. Further, while the characterization in [7] was also geometric, it was not based on Gaussian widths. In contrast, our results apply to any norm, not just $L_1$, and the geometric characterization is based on Gaussian widths. The utility of the Gaussian width based characterization becomes evident later when we establish sample complexity results for Gaussian and sub-Gaussian random matrices in terms of Gaussian widths of spherical caps.

We establish bounds on the estimation error $\hat{\Delta}_n$ under two assumptions, which are subsequently shown to hold with high probability for (sub)Gaussian designs and noise models. The first assumption is that the regularization parameter $\lambda_n$ is suitably large. In particular, for any $\beta > 1$, the regularization parameter $\lambda_n$ needs to satisfy

$$\lambda_n \geq \beta R^*(\nabla L(\theta^*; Z^n)),$$

(6)

where $R^*(\cdot)$ denotes the dual norm of $R(\cdot)$. The second assumption is that the design matrix $X \in \mathbb{R}^{n \times p}$ satisfies the restricted strong convexity (RSC) condition [7] in the error set $E_r$, so that for $\Delta \in E_r$, there exists a suitable constant $\kappa > 0$ so that

$$\delta \mathcal{L}(\Delta, \theta^*) \triangleq \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle \geq \kappa \|\Delta\|_2^2.$$

(7)

With such suitably large $\lambda_n$ and $X$ satisfying the RSC condition, we establish the following bound:

$$\|\hat{\Delta}_n\|_2 \leq \Psi(E_r) \frac{1 + \beta \lambda_n}{\beta \kappa},$$

(8)

where $\Psi(E_r) = \sup_{\omega \in E_r} \frac{R(\omega)}{\|\omega\|_2}$ is a norm compatibility constant [26]. Note that the above bound is deterministic, but relies on assumptions on $\lambda_n$ and $\kappa$. So, we focus on characterizations of $\lambda_n$ and $\kappa$ which hold with high probability for families of design matrices $X$ and noise models $\omega$.

**Bounds on the regularization parameter $\lambda_n$:** From (6) above, for the analysis to work, one needs to have $\lambda_n \geq R^*(\nabla L(\theta^*; Z^n))$. There are a few challenges in getting a suitable bound for $\lambda_n$. First, the bound depends on $\theta^*$, but $\theta^*$ is unknown and is the quantity one is interested in estimating. Second, the bound depends on $Z^n$, the samples, and is hence random. The goal will be to bound the expectation $\mathbb{E}[R^*(\nabla L(\theta^*; Z^n))]$ over all samples of size $n$, and obtain high-probability deviation bounds. Third, since the bound relies on the (dual) norm $R^*(\cdot)$ of a $p$-dimensional random vector, without proper care, the lower bound on $\lambda_n$ may end up having a large scaling dependency, say $\sqrt{p}$, on the ambient dimensionality. Since the error bound in (8) is directly proportional to $\lambda_n$, such dependencies will lead to weak bounds.

In Section 3, we characterize the expectation $\mathbb{E}[R^*(\nabla L(\theta^*; Z^n))]$ in terms of the geometry of the unit norm-ball of $R$, which leads to a sharp bound. Let $\Omega_R = \{u \in \mathbb{R}^p | R(u) \leq 1\}$ denote the unit norm-ball. Then,

\(^1\)A gentle exposition to Gaussian width and some of its properties is given in Appendix A.
for Gaussian design matrices and squared loss, we show that
\[
E[R^*(\nabla \mathcal{L}(\theta^*; Z^n))] \leq \frac{c}{\sqrt{n}} w(\Omega_R) ,
\] (9)
which scales as the Gaussian width of \( \Omega_R \). The result can be extended to the case of anisotropic Gaussian designs as well as correlated samples, where the constant \( c \) starts depending on the maximum eigenvalue (operator norm) of the corresponding covariance matrix. Further, the result also extends to the case of sub-Gaussian design matrices. In the sub-Gaussian case, the result is in terms of the ‘sub-Gaussian width’ of the unit norm-ball, which can be upper bounded by a constant times the Gaussian width using generic chaining \([33]\). For high-probability bounds, the Gaussian case follows from existing results in large deviation of Lipschitz functions of Gaussian vectors, or in this case, equivalently, supremum of Gaussian processes \([22, 8]\). While there is no equivalent result on supremum of general sub-Gaussian processes, large deviation bounds are possible for important special cases, e.g., bounded random variables \([22]\), random variables with sufficient smoothness, i.e., bounded Malčev derivatives \([40]\), etc. The results can also be extended to general convex losses from generalized linear models.

The above characterization allows one to choose \( \lambda_n \geq \frac{c}{\sqrt{n}} w(\Omega_R) \). For the special case of \( L_1 \) regularization, \( \Omega_R \) is the unit \( L_1 \) norm ball, and the corresponding Gaussian width \( w(\Omega_R) \leq c_1 \sqrt{\log p} \), which explains the \( \sqrt{\log p} \) term one finds in existing bounds for Lasso \([26, 10]\). When working with other norms, one simply needs to get an upper bound on the corresponding \( w(\Omega_R) \).

**Restricted eigenvalue conditions:** When the loss function under consideration is the squared loss, the RSC condition in \([7]\) reduces to the restricted eigenvalue (RE) condition on the design matrix. Our analysis focuses on establishing the RE condition on \( A = \text{cone}(E_r) \cap S^{p-1} \), the spherical cap obtained by intersecting the cone of the error set with the unit hypersphere, since it implies the RE condition on \( E_r \). We show that for isotropic design matrices, for both Gaussian and sub-Gaussian case, an inequality of the following form
\[
\inf_{u \in A} \| Xu \|_2 \geq c_1 \sqrt{n} - c_2 w(A) ,
\] (10)
where \( w(A) \) is the Gaussian width of the spherical cap \( A \), always holds with high probability. Thus, the sample complexity \( n_0 = c_2^2 w^2(A)/c_1^2 = O(w^2(A)) \), and for \( n > n_0 \), an RE condition of the form
\[
\inf_{u \in A} \| Xu \|_2 \geq c_1 \left( 1 - \frac{c_2 w(A)}{c_1 \sqrt{n}} \right) \sqrt{n} = c_1 \kappa_A(n, p) \sqrt{n} ,
\] (11)
is satisfied with high probability for \( \kappa_A(p, n) = \Theta(1) \). Thus, one does not need to treat the RE condition as an assumption for isotropic Gaussian or sub-Gaussian designs—it always holds with high probability, with the phase transition happening at \( O(w^2(A)) \) samples. Our analysis technique for this result, as well as all other RE/RSC type results, relies on a simple covering argument, in particular a union bound over an \( \epsilon \)-covering of the spherical cap \( A \). We use Sudakov’s inequality, also known as the weak converse of Dudley’s inequality, to go from covering numbers to Gaussian widths \([16, 23]\).

For anisotropic sub-Gaussian designs, with \( \Sigma \in \mathbb{R}^{p \times p} \) denoting the covariance matrix, we show an inequality of the following form
\[
\inf_{u \in A} \| Xu \|_2 \geq c_1 \sqrt{n} - c_2 \Lambda_{\text{max}}(\Sigma) w(A) ,
\] (12)
where \( \Lambda_{\text{max}}(\Sigma) \) denotes the largest eigenvalue of \( \Sigma \), always holds with high probability. We get a mildly sharper result for the case of anisotropic Gaussian designs. The constant \( c_1 \) has a dependency on the restricted minimum eigenvalue \( \nu = \inf_{u \in A} \| \Sigma^{1/2} u \|_2 \) of the true covariance \( \Sigma \), as expected. Thus, for the anisotropic case, the sample complexity of the RE condition is \( O(\Lambda_{\text{max}}^2(\Sigma) w^2(A)) \). If the largest eigenvalue \( \Lambda_{\text{max}}(\Sigma) \) is a constant or grows slowly with the dimensionality, the phase transition behavior is qualitatively similar.
to that for isotropic designs. However, for highly correlated covariates, the sample complexity may have a strong dependency on the dimensionality, and one may need to consider alternative estimators \cite{77,27}.

We also establish results for the setting where one has correlated isotropic samples. If $\Gamma \in \mathbb{R}^{n \times n}$ denotes the covariance matrix for the samples, then for both Gaussian and sub-Gaussian isotropic designs, we show that an inequality of the following form
\[
\inf_{u \in A} \|X u\|_2 \geq c_1 \sqrt{\text{Tr} (\Gamma)} - c_2 \lambda_{\max} (\Gamma) w (A),
\]
where $\lambda_{\max} (\Gamma)$ denotes the largest eigenvalue of $\Gamma$, always holds with high probability. Interestingly, the sample complexity of the RE condition here scales as the stable rank $\frac{\text{Tr} \Gamma}{\lambda_{\max} (\Gamma)}$ of the covariance matrix $\Gamma$. Further, for the special case of independent samples, we recover the results discussed above for isotropic designs.

**Generalized linear models and restricted strong convexity:** For convex loss functions coming from generalized linear models (GLMs), the sample complexity and associated phase transition behavior is determined by the Restricted Strong Convexity (RSC) condition \cite{26}. By generalizing our argument for RE conditions corresponding to square loss, we show that the RSC conditions are going to be satisfied for convex losses corresponding to GLMs for sub-Gaussian designs at the same order of sample complexity as that for squared loss. In particular, for $A = \text{cone} (E_r) \cap S^{p-1}$, we show a high probability lower bound of the form
\[
\inf_{w \in A} \sqrt{n} \| \delta \mathcal{L} (u, \theta^*) \|_2 \geq c_1 \sqrt{n} - c_2 w (A),
\]
where the constants $c_1, c_2$ depend on the probability mass on the tails of the design matrix distribution. Interestingly, the sample complexity still scales as $O(w^2 (A))$ for any spherical cap $A$. The result is thus a considerable generalization of earlier results on GLMs which had looked at specific norms and associated cones and/or did not express the results in terms of the Gaussian width of $A$ \cite{26}.

**Putting everything together:** With the above results in place, from (8), the main bound takes the form
\[
\| \tilde{\Delta}_n \|_2 \leq \frac{1 + \beta}{\beta} \Psi (E_r) \frac{c_2}{\max (0, 1 - c_1 w (A) / \sqrt{n})} \frac{w (\Omega_R)}{\sqrt{n}}
\]
with high probability, where $w (\Omega_R)$ is the Gaussian width of the unit norm ball, $w (A)$ is the Gaussian width of the spherical cap corresponding to the error set $E_r$, and the result is valid only when $n > n_A = O (w^2 (A))$ which corresponds to the sample complexity. For the special case of $L_1$ norm, i.e., Lasso, the sample complexity $n_A$ is of the order $w^2 (A) = O (s \log p)$. Further, $w (\Omega_R) = \sqrt{\log p}$ and $\Psi (E_r) = \sqrt{\pi}$. Plugging in these values, choosing $\beta = 2$, for $n > c_3 s \log p$, the bound $\| \tilde{\Delta}_n \|_2 \leq c_4 \sqrt{s \log p / n}$ holds with probability. For other norms, one can simply plug-in the widths to get the corresponding sample complexity and non-asymptotic error bounds.

The rest of the paper is organized as follows: Section 3 presents results on the restricted error set and deterministic error bounds under suitable bounds on the regularization parameter $\lambda_n$ and RSC assumptions. Section 4 presents a characterization of $\lambda_n$ in terms of the Gaussian width of the unit norm ball for Gaussian as well as sub-Gaussian designs and noise. Section 5 proves RE conditions and associated sample complexity results corresponding to squared loss functions. Results are presented for Gaussian and sub-Gaussian designs, including anisotropic and correlated cases, and always in terms of the Gaussian width of the spherical cap corresponding to the error set. Section 6 presents RSC conditions corresponding to general convex losses arising from generalized linear models, and the results are again in terms of the Gaussian width of the spherical cap corresponding to the error set. We conclude in Section 7. All technical arguments and proofs are in the appendix, along with a gentle exposition to Gaussian widths and related results.

A brief word on the notation used. We denote random matrices as $X$, and random vectors as $X_i$ where $i$ may be an index to a row or column of a random matrix. Vector norms are denoted as $\| \cdot \|$, e.g., $\| X_i \|_2$ for a (random) vector $X_i$, and norms of random variables are denoted as $\| \cdot \|$, e.g., $\| X \|_2 = E[\| X \|_2]$. 

$E_c = \{ \Delta \mid R(\theta^* + \Delta) \leq R(\theta^*) \}$

$A_c = \text{cone}(E_c) \cap \rho B^p_2$

$E_r = \{ \Delta \mid R(\theta^* + \Delta) \leq R(\theta^*) + \frac{1}{\beta} R(\Delta) \}$

$A_r = E_r \cap \rho B^p_2$

2 Restricted Error Set and Recovery Guarantees

In this section, we give a characterization of the restricted error set $E_r$ in which the error vector $\hat{\Delta}_n = (\hat{\theta}_{\lambda_n} - \theta^*)$ lies, establish clear relationships between the error sets for the regularized and constrained problems, and finally establish upper bounds on the estimation error. The error bound is deterministic, but has quantities which involve $\theta^*, X, \omega$, for which we develop high probability bounds in Sections 3, 4, and 5.

2.1 The Restricted Error Set and the Error Cone

We start with a characterization of the restricted error set $E_r$ where $\hat{\Delta}_n$ will belong.

Lemma 1 For any $\beta > 1$, assuming

$$\lambda_n \geq \beta R^*(\nabla L(\theta^*; Z^n)), \quad (16)$$

where $R^*(\cdot)$ is the dual norm of $R(\cdot)$. Then the error vector $\hat{\Delta}_n = \hat{\theta}_{\lambda_n} - \theta^*$ belongs to the set

$$E_r = E_r(\theta^*, \beta) = \left\{ \Delta \in \mathbb{R}^p \mid R(\theta^* + \Delta) \leq R(\theta^*) + \frac{1}{\beta} R(\Delta) \right\}. \quad (17)$$

The restricted error set $E_r$ need not be convex for general norms. Interestingly, for $\beta = 1$, the inequality in (17) is just the triangle inequality, and is satisfied by all $\Delta$. Note that $\beta > 1$ restricts the set of $\Delta$ which satisfy the inequality, yielding the restricted error set. In particular, $\Delta$ cannot go in the direction of $\theta^*$, i.e., $\Delta \neq \alpha \theta^*$ for any $\alpha > 0$. Further, note that the condition in (16) is similar to that in [26] for $\beta = 2$, but the above characterization holds for any norm, not just decomposable norms [26].
Figure 2: Schematic for norm regularized objective functions considered. The finite sample estimate $\hat{\theta}_n$ has lower empirical loss than the optimum $\theta^*$. Bounding the difference between the losses yields a bound on $\|\hat{\theta}_n - \theta^*\|$.

While $E_r$ need not be a convex set, we establish a relationship between $E_r$ and the error set $E_c$ corresponding to constrained estimators [13, 7, 14, 15]. A recent example of such a constrained estimator is the generalized Dantzig selector (GDS) [15] given by:

$$\hat{\theta}_{\varphi_n} = \arg\min_{\theta \in \mathbb{R}^p} R(\theta) \quad \text{s.t.} \quad R^*(X^T(y - X\theta^*)) \leq \varphi_n,$$

where $R^*(\cdot)$ denotes the dual norm of $R(\cdot)$. One can show [14, 15] that the restricted error set for such constrained estimators [14, 15, 36] are of the form:

$$E_c = \{ \Delta \in \mathbb{R}^p : R(\theta^* + \Delta) \leq R(\theta^*) \}.$$

By definition, it is easy to see that $E_c$ is always convex, and that $E_c \subseteq E_r$, as shown schematically in Figure 1.

The following results establishes a relationship between $E_r$ and $E_c$ in terms of their Gaussian widths.

**Theorem 1** Let $A_r = E_r \cap \rho B^p_2$ and $A_c = \text{cone}(E_c) \cap \rho B^p_2$, where $\rho B^p_2 = \{ u : \|u\|_2 \leq \rho \}$ is the $L_2$ ball of any radius $\rho > 0$. Then, for any $\beta > 1$ we have

$$w(A_r) \leq \left( 1 + \frac{2}{\beta - 1} \frac{\|\theta^*\|_2}{\rho} \right) w(A_c),$$

where $w(A)$ denotes the Gaussian width of any set $A$ given by: $w(A) = E_g \left[ \sup_{a \in A} \langle a, g \rangle \right]$, where $g$ is an isotropic Gaussian random vector, i.e., $g \sim N(0, \mathbb{I}_{p \times p})$.

Thus, the Gaussian width of the error sets of regularized and constrained problems are closely related. See Figure 1 for more details. In particular, for $\|\theta^*\|_2 = 1$, with $\rho = 1$, $\beta = 2$, we have $w(A_r) \leq 3w(A_c)$. Related observations have been made for the special case of the $L_1$ norm [7], although past work did not provide an explicit characterization in terms of Gaussian widths. The result also suggests that it is possible to move between the error analysis of the regularized and the constrained versions of the estimation problem.
2.2 Recovery Guarantees

In order to establish recovery guarantees, we start by assuming that restricted strong convexity (RSC) is satisfied by the loss function in $E_r$, the error set, so that for any $\Delta \in E_r$, there exists a suitable constant $\kappa$ so that
\[
\delta \mathcal{L}(\Delta, \theta^*) \triangleq \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle \geq \kappa \|\Delta\|^2 \tag{21}
\]
In Sections 4 and 5 we establish precise forms of the RSC condition for a wide variety of design matrices and loss functions. In order to establish recovery guarantees, we focus on the quantity
\[
\mathcal{F}(\Delta) = \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda_n (R(\theta^* + \Delta) - R(\theta^*)) \tag{22}
\]
Since $\hat{\theta}_{\lambda_n} = \theta^* + \hat{\Delta}_n$ is the estimated parameter, i.e., $\hat{\theta}_{\lambda_n}$ is the minimum of the objective, we clearly have $\mathcal{F}(\hat{\Delta}_n) \leq 0$, which implies a bound on $\|\hat{\Delta}_n\|_2$. Unlike previous analysis, the bound can be established without making any additional assumptions on the norm $R(\theta)$. We start with the following result, which expresses the upper bound on $\|\hat{\Delta}_n\|_2$ in terms of the gradient of the objective at $\theta^*$.

**Lemma 2** Assume that the RSC condition is satisfied in $E_r$ by the loss $\mathcal{L}(\cdot)$ with parameter $\kappa$. With $\hat{\Delta}_n = \hat{\theta}_{\lambda_n} - \theta^*$, for any norm $R(\cdot)$, we have
\[
\|\hat{\Delta}_n\|_2 \leq \frac{1}{\kappa} \|\nabla \mathcal{L}(\theta^*) + \lambda_n \nabla R(\theta^*)\|_2 \tag{23}
\]
where $\nabla R(\cdot)$ is any sub-gradient of the norm $R(\cdot)$.

Figure 3 illustrates the above results. Note that the right hand side is simply the $L_2$ norm of the gradient of the objective evaluated at $\theta^*$. For the special case when $\hat{\theta}_{\lambda_n} = \theta^*$, the gradient of the objective is zero, implying correctly that $\|\hat{\Delta}_n\|_2 = 0$. While the above result provides useful insights about the bound on $\|\hat{\Delta}_n\|_2$, the quantities on the right hand side depend on $\theta^*$, which is unknown. We present another form of the result in terms of quantities such as $\lambda_n$, $\kappa$, and the norm compatibility constant $\Psi(E_r) = \sup_{u \in E_r} \frac{R(u)}{\|u\|_2^2}$, which are often easier to compute or bound.

**Theorem 2** Assume that the RSC condition is satisfied in $E_r$ by the loss $\mathcal{L}(\cdot)$ with parameter $\kappa$. With $\hat{\Delta}_n = \hat{\theta}_{\lambda_n} - \theta^*$, for any norm $R(\cdot)$, we have
\[
\|\hat{\Delta}_n\|_2 \leq \Psi(E_r) \frac{1 + \beta \lambda_n}{\beta \kappa} \tag{24}
\]
The above result is deterministic, but contains $\lambda_n$ and $\kappa$. In Section 3 we give precise characterizations of $\lambda_n$, which needs to satisfy (16). In Sections 4 and 5 we characterize the RSC condition constant $\kappa$ for different losses and a variety of design matrices.

2.3 A Special Case: Decomposable Norms

In recent work, [26] considered regularized regression with the special case of decomposable norms, defined in terms of a pair of subspaces $\mathcal{M} \subseteq \bar{\mathcal{M}}$ of $\mathbb{R}^p$. The model is assumed to be in the subspace $\mathcal{M}$, and the definition considers the so-called perturbation subspace $\bar{\mathcal{M}} \perp$ which is the orthogonal complement of $\mathcal{M}$. A norm $R(\cdot)$ is considered decomposable with respect to subspaces $(\mathcal{M}, \bar{\mathcal{M}} \perp)$ if $R(\theta + \gamma) = R(\theta) + R(\gamma)$ for all $\theta \in \mathcal{M}$ and $\gamma \in \bar{\mathcal{M}} \perp$. 

8
Figure 3: Schematic illustrating the error bound in Lemma 2. Under restricted strong convexity (RSC) of the loss function in the error set $E_r$, the error $\|\hat{\Delta}_n\|_2$ can be bounded in terms of the gradient of the overall objective evaluated at $\theta^*$.

We show that for decomposable norms, the error set $E_r$ in our analysis is included in the error cone defined in [26]. In the current context, let $\beta = 2$, $\theta^* \in \mathcal{M}$, then for any $\Delta = \Delta_{\mathcal{M}^\perp} + \Delta_{\mathcal{M}} \in E_r$, we have

\begin{align}
R(\theta^* + \Delta) &\leq R(\theta^*) + \frac{1}{2} R(\Delta) \quad (25) \\
\Rightarrow R(\theta^* + \Delta_{\mathcal{M}^\perp} + \Delta_{\mathcal{M}}) &\leq R(\theta^*) + \frac{1}{2} R(\Delta_{\mathcal{M}^\perp} + \Delta_{\mathcal{M}}) \quad (26) \\
\Rightarrow R(\theta^*) + R(\Delta_{\mathcal{M}^\perp}) - R(\Delta_{\mathcal{M}}) &\leq R(\theta^*) + \frac{1}{2} R(\Delta_{\mathcal{M}^\perp}) + \frac{1}{2} R(\Delta_{\mathcal{M}}) \quad (27) \\
\Rightarrow R(\Delta_{\mathcal{M}^\perp}) &\leq 3R(\Delta_{\mathcal{M}}). \quad (29)
\end{align}

where inequality (a) follows from the triangle inequality and (b) follows from decomposability of the norm. The last inequality is precisely the error cone in [26] for $\theta^* \in \mathcal{M}$. As a result, for any $\Delta \in E_r$, for decomposable norms we have

\begin{align}
R(\Delta) = R(\Delta_{\mathcal{M}^\perp} + \Delta_{\mathcal{M}}) \leq R(\Delta_{\mathcal{M}^\perp}) + R(\Delta_{\mathcal{M}}) \leq 4R(\Delta_{\mathcal{M}}) \quad (30)
\end{align}

Hence, the norm compatibility constant can be bounded as

\begin{align}
\Psi(E_r) = \sup_{\Delta \in E_r} \frac{R(\Delta)}{\|\Delta\|_2} \leq 4 \sup_{\Delta \in E_r} \frac{R(\Delta_{\mathcal{M}^\perp})}{\|\Delta\|_2} \leq 4 \sup_{u \in \mathcal{M} \setminus \{0\}} \frac{R(u)}{\|u\|_2} = 4\Psi(\mathcal{M}). \quad (31)
\end{align}

where $\Psi(\mathcal{M})$ is the subspace compatibility in $\mathcal{M}$, as used in [26].

### 3 Bounds on the Regularization Parameter

Recall that the parameter $\lambda_n$ needs to satisfy the inequality

\begin{align}
\lambda_n \geq \beta R^\ast(\nabla L(\theta^*; Z_n)) . \quad (32)
\end{align}
Theorem 4

Elementwise independent Gaussian or sub-Gaussian noise. Where the expectation is taken over both \( L \) and also discuss the upper bounds. For ease of exposition, we present results for the case of squared loss, i.e., \( \omega \) is isotropic, independent anisotropic, or dependent isotropic, as considered for Gaussian designs.

Further for any \( \tau > 0 \), with probability at least \( 1 - 3 \exp(-\min(\frac{\tau^2}{2\Phi^2}, cn)) \), we have

\[
R^* (\nabla L(\theta^*; Z^n)) \leq \eta_2 (w(\Omega_R) + \tau)
\]

where \( \Phi^2 = \sup_{R(u) = 1} \|u\|^2 \), \( c \) is an absolute constant, \( \eta_2 = 2 \) when \( \omega \) is i.i.d. standard Gaussian, and \( \eta_2 = \sqrt{2K^2 + 1} \) when \( \omega \) is i.i.d. centered unit-variance sub-Gaussian with \( \|x\|_{\psi_2} \leq K \).

Sub-Gaussian Designs: We consider the setting of sub-Gaussian designs, where \( X \) can also be independent isotropic, independent anisotropic, or dependent isotropic, as considered for Gaussian designs. \( \omega \) is also elementwise independent Gaussian or sub-Gaussian noise.

Theorem 3

Let \( \Omega_R = \{ u : R(u) \leq 1 \} \). For Gaussian design \( X \) and Gaussian or sub-Gaussian noise \( \omega \), we have

\[
E[R^* (\nabla L(\theta^*; Z^n))] = E \left[ \frac{1}{n} X^T \omega \right] \leq \frac{\eta_1}{\sqrt{n}} w(\Omega_R)
\]

where the expectation is taken over both \( X \) and \( \omega \). The constant \( \eta_1 \) is given by

\[
\eta_1 = \begin{cases} 1 & \text{if } X \text{ is independent isotropic} \\ \sqrt{\Lambda_{\max}(\Sigma)} & \text{if } X \text{ is independent anisotropic} \\ \sqrt{\Lambda_{\max}(\Gamma)} & \text{if } X \text{ is dependent isotropic} \end{cases}
\]

Further for any \( \tau > 0 \), with probability at least \( 1 - 3 \exp(-\min(\frac{\tau^2}{2\Phi^2}, cn)) \), we have

\[
R^* (\nabla L(\theta^*; Z^n)) \leq \frac{\eta_1 \eta_2}{\sqrt{n}} (w(\Omega_R) + \tau)
\]

where \( \Phi^2 = \sup_{R(u) = 1} \|u\|^2 \), \( c \) is an absolute constant, \( \eta_2 = 2 \) when \( \omega \) is i.i.d. standard Gaussian, and \( \eta_2 = \sqrt{2K^2 + 1} \) when \( \omega \) is i.i.d. centered unit-variance sub-Gaussian with \( \|x\|_{\psi_2} \leq K \).
Interestingly, the analysis for the result above involves ‘sub-Gaussian width’ which can be upper bounded by a constant times the Gaussian width, using generic chaining [33]. Further, one can get Gaussian-like exponential concentration around the expectation for important classes of sub-Gaussian random variables, including bounded random variables [22], and when $X_u = \langle h, u \rangle$, where $u$ is any unit vector, are such that their Malliavin derivatives have almost surely bounded norm in $L^2[0,1]$, i.e., $\int_0^1 |D_r X_u|^2 dr \leq \eta$ [40]. For convenience, we denote $Z' = \sup_{R(u) \leq 1} \langle h, u \rangle$. The ideal scenario is to obtain dimension independent concentration inequalities of the form

$$P \left( \left| Z' - E[Z'] \right| \geq \tau \right) \leq \eta_3 \exp \left( -\eta_4 \tau^2 \right),$$  

where $\eta_3$ and $\eta_4$ are constants. We show that such inequalities are indeed satisfied by a wide variety of distributions.

**Theorem 5** Let $h = [h_1 \cdots h_p]$ be a vector of independent centered random variables, and let $Z' = R^*(h) = \sup_{R(u) \leq 1} \langle h, u \rangle$. Then, we have

$$P \left( \left| Z' - E[Z'] \right| \geq \tau \right) \leq \eta_3 \exp \left( -\eta_4 \tau^2 \right),$$

if $h$ satisfies any of the following conditions:

(i) $h$ is bounded, i.e., $\|h\|_\infty = L < \infty$,

(ii) for any unit vector $u$, $Y_u = \langle h, u \rangle$ are such that their Malliavin derivatives have almost surely bounded norm in $L^2[0,1]$, i.e., $\int_0^1 |D_r Y_u|^2 dr \leq L^2$.

Interestingly, both of the above conditions imply that $h$ is a sub-Gaussian vector. In particular, the condition in (i) implies finite sub-Gaussian norm [38]. The implication for the condition in (ii) is discussed in [40, Lemma 3.3] which in turn relies on [37, Theorem 9.1.1].

Based on Theorem 4 and Theorem 5, we can easily get the high-probability upper bound for $Z'$,

$$P \left( Z' \geq \eta_0 w(\Omega_R) + \tau \right) \leq \eta_3 \exp \left( -\eta_4 \tau^2 \right).$$

Hence, for any $\tau > 0$, with probability at least $1 - (\eta_3 + 2) \exp(-\min(\eta_4 \tau^2, cn))$, we have

$$R^* (\nabla L(\theta^*; Z^n)) \leq \frac{\eta_1 \eta_2}{\sqrt{n}} (\eta_0 w(\Omega_R) + \tau),$$

for sub-Gaussian design matrix $X$, where $\eta_1$ and $\eta_2$ are those defined in Theorem 4.

**Bounding the Gaussian width $w(\Omega_R)$**: In certain cases, one may be able to directly obtain a bound on the Gaussian width $w(\Omega_R)$. Here, we provide a mechanism for bounding the Gaussian width $w(\Omega_R)$ of the unit norm ball in terms of the Gaussian width of a suitable cone, obtained by shifting or translating the norm ball. In particular, the result involves taking any point on the boundary of the unit norm ball, considering that as the origin, and constructing a cone using the norm ball. Since such a construction can be done with any point on the boundary, the tightest bound is obtained by taking the infimum over all points on the boundary. The motivation behind getting an upper bound of the Gaussian width $w(\Omega_R)$ of the unit norm ball in terms of the Gaussian width of such a cone is because considerable advances have been made in recent years in upper bounding Gaussian widths of such cones [14, 2].

**Lemma 3** Let $\Omega_R = \{u : R(u) \leq 1\}$ be the unit norm ball and $\Theta_R = \{u : R(u) = 1\}$ be the boundary. For any $\hat{\theta} \in \Theta_R$, $\rho(\hat{\theta}) = \sup_{\theta \in \Theta_R} \|\theta - \hat{\theta}\|_2$ is the diameter of $\Omega_R$ measured with respect to $\hat{\theta}$. Let $G(\hat{\theta}) = \text{cone}(\Omega_R - \hat{\theta}) \cap \rho(\hat{\theta})B^2_2$, i.e., the cone of $(\Omega_R - \hat{\theta})$ intersecting the ball of radius $\rho(\hat{\theta})$. Then

$$w(\Omega_R) \leq \inf_{\hat{\theta} \in \Theta_R} w(G(\hat{\theta})).$$

(40)
Figure 4: Bounding the Gaussian width of a norm ball, e.g., corresponding to $L_1$ norm, by shifting the norm ball and using the width of the corresponding cone (Lemma 3). The approach allows one to directly use existing results on bounding Gaussian widths of certain cones. In some cases, it may be easier to directly bound the Gaussian width of the norm ball, rather than using the shifting argument.

The analysis and results for $\lambda_n$ presented above can be extended to general convex losses arising in the context of GLMs for sub-Gaussian designs and sub-Gaussian noise (see Section 5).

4 Least Squares Models: Restricted Eigenvalue Conditions

The error bound analysis in Theorem 2 depends on the restricted strong convexity (RSC) assumption. In this section, we establish RSC conditions for both Gaussian and sub-Gaussian random design matrices when the loss function is the squared loss. For squared loss, i.e., $L(\theta; Z^n) = \frac{1}{2n} \|y - X\theta\|^2$, the RSC condition (21) becomes equivalent to the Restricted Eigenvalue (RE) condition [7, 26], i.e., $\frac{1}{2n} \|X\Delta\|_2^2 \geq \kappa_E \|\Delta\|_2^2$ for all $\Delta \in E_r$, since

$$\delta L(\Delta, \theta^*) = \frac{1}{2n} \|X\Delta\|^2 = \frac{1}{2n} \sum_{i=1}^n \langle X_i, \Delta \rangle^2.$$  \hspace{1cm} (41)

We make two simplifications which lets us develop the RE results in terms of widths spherical caps rather than over the error set $E_r$. Let $n_{E_r}$ be the sample complexity for the RE condition over the set $E_r$, so that for $n > n_{E_r}$ samples, with high probability

$$\inf_{\Delta \in E_r} \frac{1}{2n} \|X\Delta\|_2^2 \geq \kappa_{E_r} \|\Delta\|_2^2,$$  \hspace{1cm} (42)

for some $\kappa_{E_r} > 0$. Let $C_r = \text{cone}(E_r)$ and let $n_{C_r}$ be the sample complexity for the RE condition over the cone $C_r$, so that for $n > n_{C_r}$ samples, with high probability

$$\inf_{\Delta \in C_r} \frac{1}{2n} \|X\Delta\|_2^2 \geq \kappa_{C_r} \|\Delta\|_2^2,$$  \hspace{1cm} (43)

for some $\kappa_{C_r} > 0$. Since $E_r \subseteq C_r$, we have $n_{E_r} \leq n_{C_r}$. Thus, it is sufficient to obtain (an upper bound to) the sample complexity $n_{C_r}$, since that will serve as an upper bound to $n_{E_r}$, the sample complexity over $E_r$. 

12
Further, since $C_r$ is a cone, the absolute magnitude $\|\Delta\|_2$ does not affect the sample complexity. As a result, it is sufficient to focus on a spherical cap $A = C_r \cap S^{p-1}$. In particular, if $n_A$ denotes the sample complexity for the RE condition over the spherical cap $A$, so that for $n > n_A$ samples, with high probability
\[
\inf_{u \in A} \frac{1}{2n} \| Xu \|_2^2 \geq \bar{\kappa}_A \| u \|_2^2 ,
\]
for some $\bar{\kappa}_A > 0$, then $n_A = n_{C_r} \geq n_{E_r}$. Noting that $\| u \|_2 = 1$ for $u \in C_r \cap S^{p-1}$, we present sample complexity results for RE conditions the form
\[
\inf_{u \in A} \| Xu \|_2 \geq \kappa_A(n, p) \sqrt{n}
\]
where $\kappa_A(n, p) > 0$ with high probability for $n > n_A$.

In this section, we characterize sample complexity $n_A$ over any given spherical cap $A$, and establish RE conditions for a variety of Gaussian and sub-Gaussian design matrices $X$, with isotropic, anisotropic, or dependent rows, i.e., when samples (rows of $X$) are not independent. Results for certain types of design matrices for certain types of norms, especially the $L_1$ norm, have appeared in the literature \cite{7,31,32}. Our analysis considers a wider variety of design matrices and establishes RE conditions for any $A \subseteq S^{p-1}$, thus corresponding to all norms. Interestingly, the Gaussian width $w(A)$ of the spherical cap $A$ shows up in all bounds, as a measure of the size of the set $A$, even for sub-Gaussian design matrices. In fact, all existing RE results do implicitly have the width term, but in a form specific to the chosen norm \cite{31,32}. The analysis on atomic norm in \cite{14} has the $w(A)$ term explicitly, but the analysis relies on Gordon’s inequality \cite{18,19,23}, which is applicable only for isotropic Gaussian design matrices.

The proof technique we use is simple, a standard covering argument, and is largely the same across all design matrices considered. A unique aspect of our analysis, used in all the proofs, is a way to go from covering numbers of $A$ to the Gaussian width of $A$ using Sudakov’s inequality, also known as the ‘weak converse’ of Dudley’s inequality, connecting Gaussian widths and covering numbers \cite{16,23}. Our simple techniques are in contrast to much of the existing literature on RE conditions, which use specialized tools such as Gaussian comparison principles \cite{31,26}, and/or specialized analysis geared to a particular norm such as $L_1$ \cite{32}.

### 4.1 Restricted Eigenvalue Conditions: Gaussian Designs

Here we focus on the case of Gaussian design matrices $X \in \mathbb{R}^{n \times p}$, and consider three settings: (i) independent-isotropic, where the entries are elementwise independent, (ii) independent-anisotropic, where rows $X_i$ are independent but each row has a covariance $E[X_i X_i^T] = \Sigma \in \mathbb{R}^{p \times p}$, and (iii) dependent-isotropic, where the rows are isotropic but the columns $X_j$ are correlated with $E[X_j X_j^T] = \Gamma \in \mathbb{R}^{n \times n}$. For convenience, we assume $E[x_{ij}^2] = 1$, noting that the analysis easily extends to the general case of $E[x_{ij}^2] = \sigma^2$.

**Independent Isotropic Gaussian (IIG) Designs:** The IIG setting has been extensively studied in the literature \cite{10,26}. As discussed in the recent work on atomic norms \cite{14}, one can use Gordon’s inequality \cite{19,23} to get RE conditions for the IIG setting. Our goal here is two-fold: first, we present the RE conditions obtained using our simple proof technique, and show that it is equivalent, up to constants, the RE condition obtained using Gordon’s inequality, an arguably heavy-duty technique only applicable to the IIG setting; and second, we go over some facets of how we present the results, which will apply to all subsequent RE-style results as well as give a way to plug-in $\kappa$ in the estimation error bound in \cite{24}.

**Theorem 6** Let the design matrix $X \in \mathbb{R}^{n \times p}$ be elementwise independent and normal, i.e., $x_{ij} \sim N(0, 1)$. Then, for any $A \subseteq S^{p-1}$, any $n \geq 2$, and any $\tau > 0$, with probability at least $(1 - \eta_1 \exp(-\eta_2 \tau^2))$, we have
\[
\inf_{u \in A} \| Xu \|_2 \geq \frac{1}{2} \sqrt{n} - \eta_0 w(A) - \tau,
\]
\[
(46)
\]
\(\eta_0, \eta_1, \eta_2 > 0\) are absolute constants.

As a result, the sample complexity \(n_A = O(w^2(A))\). We consider the equivalent result one could obtain by directly using Gordon’s inequality \([19, 23]\):

**Theorem 7 (Gordon’s inequality \([19]\))** Let the design matrix \(X\) be elementwise independent and normal, i.e., \(x_{ij} \sim N(0, 1)\). Then, for any \(A \subseteq S^{p-1}\) and any \(\tau > 0\), with probability at least \((1 - 2 \exp(-\tau^2/2))\), we have

\[
\inf_{u \in A} \|Xu\|_2 \geq \gamma_n - w(A) - \tau,
\]

where \(\gamma_n = E[\|g\|_2] > \frac{n}{\sqrt{n+1}}\) is the expected length of a Gaussian random vector in \(\mathbb{R}^n\).

Interestingly, the results are equivalent up to constants. However, unlike Gordon’s inequality, our proof technique generalizes to other design matrices, including sub-Gaussian designs.

We emphasize three additional aspects in the context of the above analysis, which will continue to hold for all the subsequent results but will not be discussed explicitly. First, to get a form of the result which can be technique generalizes to other design matrices, including sub-Gaussian designs.

We emphasize three additional aspects in the context of the above analysis, which will continue to hold for all the subsequent results but will not be discussed explicitly. First, to get a form of the result which can be used as \(\kappa\) and plugged in to the estimation error bound \([24]\), one can simply choose \(\tau = \frac{1}{2} (\frac{1}{2} \sqrt{n} - \eta_0 w(A))\) so as to get

\[
\inf_{u \in A} \|Xu\|_2 \geq \frac{1}{4} \sqrt{n} - \frac{\eta_0}{2} w(A),
\]

with high probability. Second, the result does not depend on the size of \(u\), i.e., \(\|u\|_2\). For the above analysis, \(u \in A \subseteq C_\rho \cap S^{p-1}\) so that \(\|u\|_2 = 1\). However, one can consider the cone \(C_\rho\) to be intersecting with a sphere \(\rho S^{p-1}\) of a different radius \(\rho\), to give \(A_\rho = C_\rho \cap \rho S^{p-1}\) so that \(u \in A_\rho\) has \(\|u\|_2 = \rho\). For simplicity, let \(A = A_1\), i.e., corresponding to \(\rho = 1\). Then, a straightforward extension yields \(\inf_{u \in A_\rho} \|Xu\|_2 \geq (\frac{1}{2} \sqrt{n} - \eta_0 w(A) - \tau)\|u\|_2\), with probability at least \((1 - \eta_1 \exp(-\eta_2 \tau^2))\), since \(\|Xu\|_2 = \|Xu\|_2 \|u\|_2\) and \(w(A\|u\|_2) = \|u\|_2 w(A)\) \([14]\). Finally, note that the leading constant \(\frac{1}{2}\) was a consequence of our choice of \(\epsilon = \frac{1}{4}\) for the \(\epsilon\)-net covering of \(A\) in the proof. One can get other constants, less than \(1\), with different choices of \(\epsilon\), and the constants \(\eta_0, \eta_1, \eta_2\) will change based on this choice.

**Independent Anisotropic Gaussian (IAG) Designs:** We consider a setting where the rows \(X_i\) of the design matrix are independent, and each row is sampled from an anisotropic Gaussian distribution, i.e., \(X_i \sim N(0, \Sigma_{p \times p})\) where \(X_i \in \mathbb{R}^p\). The setting has been considered in the literature \([31]\) for the special case of \(L_1\) norms, and results have been established using Gaussian comparison techniques \([23]\). We show that equivalent results can be obtained by our simple technique, which does not rely on Gaussian comparisons \([23, 26]\).

**Theorem 8** Let the design matrix \(X\) be row wise independent and each row \(X_i \sim N(0, \Sigma_{p \times p})\). Then, for any \(A \subseteq S^{p-1}\) and any \(\tau > 0\), with probability at least \(1 - \eta_1 \exp(-\eta_2 \tau^2)\), we have

\[
\inf_{u \in A} \|Xu\|_2 \geq \frac{1}{2} \sqrt{\nu} \sqrt{n} - \eta_0 \sqrt{\Lambda_{\max}(\Sigma)} w(A) - \tau,
\]

where \(\sqrt{\nu} = \inf_{u \in A} \|\Sigma^{1/2}u\|_2\), \(\Lambda_{\max}(\Sigma)\) denotes the largest eigenvalue of \(\Sigma\) and \(\eta_0, \eta_1, \eta_2 > 0\) are constants.

As a result, the sample complexity \(n_A = O(\Lambda_{\max}(\Sigma) w^2(A)/\nu)\). A comparison with the results of \([31]\) is instructive. The leading term \(\sqrt{\nu}\) appears in both results—the result in \([31]\) is for any \(u\) and has a \(\|\Sigma^{1/2}u\|_2\) term, whereas we have simply considered \(\inf_{u \in A}\) on both sides. The second term in \([31]\) depends on the largest entry in the diagonal of \(\Sigma\), \(\sqrt{\log p}\), and \(\|u\|_1\). These terms are a consequence of the special case analysis for \(L_1\) norm. In contrast, we consider the general case and simply get the scaled Gaussian width term \(\sqrt{\Lambda_{\max}(\Sigma)} w(A)\).
Dependent Isotropic Gaussian (DIG) Designs: We now consider a setting where the rows of the design matrix $\tilde{X}$ are isotropic Gaussians, but the columns $\tilde{X}_j$ are correlated with $E[\tilde{X}_j\tilde{X}_j^T] = \Gamma \in \mathbb{R}^{n \times n}$. Interestingly, correlation structure over the columns make the samples dependent, a scenario which has not yet been widely studied in the literature [44, 27]. We show that our simple technique continues to work in this scenario and gives a rather intuitive result.

**Theorem 9** Let $\tilde{X} \in \mathbb{R}^{n \times p}$ be a matrix whose rows $\tilde{X}_i$ are isotropic Gaussian random vectors in $\mathbb{R}^p$ and the columns $\tilde{X}_j$ are correlated with $E[\tilde{X}_j\tilde{X}_j^T] = \Gamma$. Then, for any set $A \subseteq S^{p-1}$ and any $\tau > 0$, with probability at least $(1 - \eta_1 \exp(-\eta_2 \tau^2))$, we have

$$\inf_{u \in A} \|\tilde{X} u\|_2 \geq \frac{3}{4} \sqrt{n} - \sqrt{\Lambda_{\max}(\Gamma)} \left( \eta_0 w(A) + \frac{5}{2} \right) - \tau$$

(50)

where $\eta_0, \eta_1, \eta_2 > 0$ are constants.

Note that with the assumption that $E[\tilde{X}_j^2] = 1$, $\Gamma$ will be a correlation matrix implying $\text{Tr}(\Gamma) = n$, and making the sample size dependence explicit. Intuitively, due to sample correlations, $n$ samples are effectively equivalent to $n \text{Tr}(\Gamma)/\Lambda_{\max}(\Gamma)$ samples. For the special case of $\Gamma = I_n$, $\text{Tr}(\Gamma) = n$, $\Lambda_{\max}(\Gamma) = 1$, and we recover the result for IIG designs.

4.2 Restricted Eigenvalue Conditions: Sub-Gaussian Designs

In this section, we focus on the case of sub-Gaussian design matrices $X \in \mathbb{R}^{n \times p}$, and consider three settings: (i) independent-isotropic, where the rows are independent and isotropic, (ii) independent-anisotropic, where the rows $X_i$ are independent but each row has a covariance $E[X_iX_i^T] = \Sigma_{p \times p}$, and (iii) dependent-isotropic, where the rows are isotropic and the columns $X_j$ are correlated with $E[X_jX_j^T] = \Gamma_{n \times n}$. For convenience, we assume $E[x_{ij}^2] = 1$ and the sub-Gaussian norm $\|x_{ij}\|_{\psi_2} \leq k$ [38]. In recent work, building on [1], [36] also considers generalizations of RE conditions to sub-Gaussian designs, although our proof techniques are different.

**Independent Isotropic Sub-Gaussian Designs:** We start with the setting where the sub-Gaussian design matrix $X \in \mathbb{R}^{n \times p}$ has independent rows $X_i$ and each row is isotropic.

**Theorem 10** Let $X \in \mathbb{R}^{n \times p}$ be a design matrix whose rows $X_i$ are independent isotropic sub-Gaussian random vectors in $R^p$. Then, for any set $A \subseteq S^{p-1}$ and any $\tau > 0$, with probability at least $(1 - 2 \exp(-\eta_1 \tau^2))$, we have

$$\inf_{u \in A} \|X u\|_2 \geq \sqrt{n} - \eta_0 w(A) - \tau,$$

(51)

where $\eta_0, \eta_1 > 0$ are constants which depend only on the sub-Gaussian norm $\|x_{ij}\|_{\psi_2} = k$.

As in the case of Gaussian designs, the sample complexity $n_A = O(w^2(A))$.

**Independent Anisotropic Sub-Gaussian Designs:** Consider a setting where the rows $X_i$ of the design matrix are independent, but each row is sampled from an anisotropic sub-Gaussian distribution, i.e., $\|x_{ij}\|_{\psi_2} = k$ and $E[X_iX_i^T] = \Sigma_{p \times p}$.

**Theorem 11** Let the sub-Gaussian design matrix $X$ be row wise independent, and each row $x_{ij}$ is sub-Gaussian, $X \in \mathbb{R}^{p \times p}$. Then, for any set $A \subseteq S^{p-1}$ and any $\tau > 0$, with probability at least $(1 - 2 \exp(-\eta_1 \tau^2))$, we have

$$\inf_{u \in A} \|X u\|_2 \geq \sqrt{\nu} - \eta_0 \Lambda_{\max}(\Sigma) w(A) - \tau,$$

(52)

where $\sqrt{\nu} = \inf_{u \in A} \|\Sigma^{1/2} u\|_2$, $\Lambda_{\max}(\Sigma)$ denotes the largest eigenvalue of $\Sigma$, and $\eta_0, \eta_1 > 0$ are constants which depend on the sub-Gaussian norm $\|x_{ij}\|_{\psi_2} = k$. 

15
As a result, the sample complexity \( n_A = O(\lambda^2_{\max}(\Sigma)w^2(A)/\nu) \). Note that [32] establish RE conditions for anisotropic sub-Gaussian designs for the special case of \( L_1 \) norm. In contrast, our results are general and in terms of the Gaussian width \( w(A) \).

**Dependent Isotropic Sub-Gaussian Designs:** We consider the setting where the sub-Gaussian design matrix \( \tilde{X} \) has isotropic sub-Gaussian rows, but the columns \( \tilde{X}_j \) are correlated with \( E[\tilde{X}_j\tilde{X}_j^T] = \Gamma \), implying dependent samples.

**Theorem 12** Let \( \tilde{X} \in \mathbb{R}^{n \times p} \) be a sub-Gaussian design matrix with isotropic rows and correlated columns with \( E[\tilde{X}_j\tilde{X}_j^T] = \Gamma \in \mathbb{R}^{n \times n} \). Then, for any \( A \subseteq S^{p-1} \) and any \( \tau > 0 \), with probability at least \((1 - 2\exp(-\eta_1\tau^2))\), we have

\[
\inf_{u \in A} \|\tilde{X}u\|_2 \geq \sqrt{\text{Tr}(\Gamma)} - \eta_0 \Lambda_{\max}(\Gamma)w(A) - \tau,
\]

where \( \eta_0, \eta_1 \) are constants which depend on the sub-Gaussian norm \( \|x_{ij}\|_{\psi_2} = k \).

### 4.3 Some Examples

In this section, we give examples of the analysis from previous sections for three norms: \( L_1 \) norm, group sparse norm, and \( L_2 \) norm. The summary of the results is given in Table 1. Other examples can be constructed for norms and error sets with known bounds on the Gaussian widths and norm compatibility constants [14, 26].

**L1 Norm:** Assume that the statistical parameter \( \theta^* \) is \( s \)-sparse, and note that \( \|\theta^*\|_1 \leq \sqrt{s}\|\theta^*\|_2 \). Since \( L_1 \) norm is a decomposable norm following the result in (29), we have \( \Psi(C_\tau) \leq 4\Psi(M) = 4\sqrt{s} \).

Applying Lemma 3, let \( \bar{\theta} \) be a 1-sparse vector, and \( \rho(\bar{\theta}) = 2 \), then \( w(\Omega_R) \) can be bounded by

\[
w(\Omega_R) \leq \inf_{\theta \in \Theta_R} w(G(\bar{\theta})) = w(G(\bar{\theta})) \overset{(a)}{=} O\left(\sqrt{\log p}\right),
\]

where (a) is obtained from the fact that Gaussian width of \( G(\bar{\theta}) \) with \( \bar{\theta} \) be a \( s \)-sparse vector is \( \sqrt{2s\log(\frac{p}{s}) + \frac{9}{4}s} \) [14]. See Figure 4 for more details. From Theorem 3 and (54), the bound on \( \lambda_n \) is

\[
\lambda_n \leq c\frac{w(\Omega_R)}{\sqrt{n}} = O\left(\sqrt{\frac{\log p}{n}}\right).
\]

Hence, the recovery error is bounded by

\[
\|\hat{\Delta}_n\|_2 \leq c_3 \frac{\Psi(C_\tau)\lambda_n}{\kappa} = O\left(\sqrt{\frac{s\log p}{n}}\right),
\]

which is similar to the results obtained in well known results [13, 26].

**Group Norm:** Suppose that the index set \( \{1, 2, \cdots, p\} \) can be partitioned into a set of \( T \) disjoint groups, say \( G = \{G_1, G_2, \cdots, G_T\} \). Define \((1, \nu)\)-group norm for a given vector \( \nu = (\nu_1, \cdots, \nu_T) \in [1, \infty]^T \) as

\[
\|\alpha\|_{G,\nu} = \sum_{t=1}^T \|\alpha_{G_t}\|_{\nu_t}
\]

As shown in [26] Group norm is a decomposable norm. For a given subset \( S_G \subset \{1, 2, \cdots, T\} \) with cardinality \(|S_G|\), define the subspace \( A(S_G) = \{\alpha \in \mathbb{R}^p | \alpha_{G_t} = 0, \ \forall t \notin S_G \} \). Let \( \nu_t \geq 2 \), then we have

\[
\|\Delta\|_{G,\nu} = \sum_{t \in S_G} \|\Delta_{G_t}\|_{\nu_t} \leq \sum_{t \in S_G} \|\Delta_{G_t}\|_2 \leq \sqrt{|S_G|} \|\Delta\|_2.
\]
Hence, from (31) and (58) we have
\[ \Psi(C_r) \leq 4 \sqrt{sG}. \]  
(59)

Applying Lemma 3, define \( \tilde{\theta} \) with 1-active group, and \( \rho(\tilde{\theta}) = 2 \), then \( w(\Omega_R) \) can be bounded by
\[ w(\Omega_R) \leq \inf_{\tilde{\theta} \in \Theta_R} w(G(\tilde{\theta})) = w(G(\tilde{\theta})) \overset{(a)}{=} O \left( \sqrt{m + \log T} \right), \]
where \( m = \max_t |G_t| \) and (a) is obtained from the fact that Gaussian width of \( G(\tilde{\theta}) \) where \( \tilde{\theta} \) has \( k \) active group is \( \sqrt{2k(m + \log(T - k)) + k} \) \[14\]. From Theorem 3 and (60), the bound on \( \lambda_n \) is
\[ \lambda_n \leq c \frac{w(\Omega_R)}{\sqrt{n}} = O \left( \sqrt{\frac{m + \log T}{n}} \right). \]
(61)

Hence, the recovery error is bounded by
\[ \| \hat{\Delta}_n \|_2 \leq c_3 \frac{\Psi(C_r) \lambda_n}{\kappa} = O \left( \sqrt{\frac{sG(m + \log T)}{n}} \right), \]  
(62)
which is similar to the results obtained in previous works \[14, 26\].

**L_2 Norm:** With L_2 norm as the regularizer, the norm constant is obtained as
\[ \Psi(C_r) = \sup_{\Delta \in C_r} \frac{\| \Delta \|_2}{\| \Delta \|_2} = 1. \]  
(63)

Applying Lemma 3, set \( \rho(\tilde{\theta}) = 1 \), then \( w(\Omega_R) \) can be bounded by
\[ w(\Omega_R) \leq \inf_{\tilde{\theta} \in \Theta_R} w(G(\tilde{\theta})) = O \left( \sqrt{p} \right). \]  
(64)

From Theorem 3 and (64), the bound on \( \lambda_n \) is
\[ \lambda_n \leq c \frac{w(\Omega_R)}{\sqrt{n}} = O \left( \sqrt{\frac{p}{n}} \right). \]
(65)

Hence, the recovery error is bounded by
\[ \| \hat{\Delta}_n \|_2 \leq c_3 \frac{\Psi(C_r) \lambda_n}{\kappa} = O \left( \frac{p}{\sqrt{n}} \right). \]
(66)

### 5 Generalized Linear Models: Restricted Strong Convexity

In this section, we consider the setting where the conditional probability distribution of \( y_i | X_i \) follows an exponential family distribution: \( p(y_i | X_i; \theta) = \exp\{y_i \langle \theta, X_i \rangle - \psi(\langle \theta, X_i \rangle)\} \), where \( \psi(\cdot) \) is the log-partition function \[9, 5, 42\]. Generalized linear models consider the negative likelihood of such conditional distributions as the loss function:
\[ \mathcal{L}(\theta; Z^n) = \frac{1}{n} \sum_{i=1}^{n} (\psi(\langle \theta, X_i \rangle) - \langle \theta, y_i X_i \rangle). \]  
(67)
Table 1: A summary of values for the regularization parameter $\lambda_n$, the RE condition constant $\kappa$, the norm constant $\Psi(C_r)$ and recovery bounds $\|\hat{\Delta}_n\|_2$ for $\ell_1$, $\ell_2$ and group norms in case of Gaussian Design matrix with Gaussian noise. All results are given up to constants with more emphasis on the scale of the results.

Least squares regression and logistic regression are popular special cases of GLMs. Since $\nabla \psi((\theta, X_i)) = E[y_i|X_i]$, we have $\nabla \mathcal{L}(\theta^*; Z^n) = \frac{1}{n} X^T \omega$, where $\omega_i = \nabla \psi((\theta, X_i)) - y_i = E[y_i|X_i] - y_i$ plays the role of noise. Hence, the analysis in Section 3 can be applied assuming $\omega_i$ is Gaussian or sub-Gaussian.

To obtain RSC conditions for GLMs, first note that

$$\delta \mathcal{L}(\theta^*, \Delta; Z^n) = \frac{1}{n} \sum_{i=1}^n \nabla^2 \psi((\theta^*, X_i) + \gamma_i(\Delta, X_i)) \langle \Delta, X_i \rangle^2,$$

where $\gamma_i \in [0, 1]$, by mean value theorem. Since $\psi$ is of Legendre type [4, 5, 9], the second derivative $\nabla^2 \psi(\cdot)$ is always positive. Since the RSC condition relies on a non-trivial lower bound for the above quantity, the analysis considers a suitable compact set where $\ell = \ell_{\psi}(T) = \min_{|a| \leq 2T} \nabla^2 \psi(a)$ is bounded away from zero. Outside this compact set, we will only use $\nabla^2 \psi(\cdot) > 0$. Then,

$$\delta \mathcal{L}(\theta^*, \Delta; Z^n) \geq \frac{\ell}{n} \sum_{i=1}^n \langle X_i, \Delta \rangle^2 \mathbb{P}[|\langle X_i, \theta^* \rangle| < T] \mathbb{P}[|\langle X_i, \Delta \rangle| < T].$$

We give a characterization of the RSC condition for independent isotropic sub-Gaussian design matrices $X \in \mathbb{R}^{n \times p}$. The analysis can be suitably generalized to the other design matrices considered in Section 4 by using the same techniques. As before, we denote $\Delta$ as $u$, and consider $u \in A \subseteq \mathbb{R}^{p-1}$ so that $\|u\|_2 = 1$. Further, we assume $\|\theta^*\|_2 \leq c_1$ for some constant $c_1$. Assuming $X$ has sub-Gaussian entries with $\|x_{ij}\|_{\psi_2} \leq k$, $\langle X_i, \theta^* \rangle$ and $\langle X_i, u \rangle$ are sub-Gaussian random variables with sub-Gaussian norm at most $Ck$ [39]. Let $\phi_1(\theta) = P\{\langle X_i, u \rangle > \tau\} \leq e \cdot \exp(-c_2 \tau^2/C_2^2 k^2)$, and $\phi_2(\theta^*; T, \tau) = P\{|\langle X_i, \theta^* \rangle| > \tau\} \leq e \cdot \exp(-c_2 \tau^2/C_2^2 k^2)$. The result we present is in terms of the constants $\ell = \ell_{\psi}(T)$, $\phi_1 = \phi(T; u)$ and $\phi_2 = \phi(T, \theta^*)$ for any suitably chosen $T$.

**Theorem 13** Let $X \in \mathbb{R}^{n \times p}$ be a design matrix with independent isotropic sub-Gaussian rows. Then, for any set $A \subseteq \mathbb{R}^{p-1}$, any $\alpha \in (0, 1)$, any $\tau > 0$, and any $n \geq \frac{2}{\alpha^2(1 - \phi_1 - \phi_2)}(c \|u\|_2^2(A) + c_2(1 - \phi_1 - \phi_2)^2(1 - \alpha)\tau^2)$ for suitable constants $c, c_2$, with probability at least $1 - 3 \exp\left(-\eta_1 \tau^2\right)$, we have

$$\inf_{u \in A} \mathbb{E}[\sqrt{n} \delta \mathcal{L}(\theta^*; u, X)] \geq \ell \sqrt{\pi} \left(\sqrt{n} - \eta_0 w(A) - \tau\right),$$

where $\pi = (1 - \alpha)(1 - \phi_1 - \phi_2)$, $\ell = \ell_{\psi}(T) = \min_{|a| \leq 2T + \beta} \nabla^2 \psi(a)$, and constants $(\eta_0, \eta_1)$ depend on the sub-Gaussian norm $\|x_{ij}\|_{\psi_2} = k$.
The form of the result is closely related to the corresponding result for the RE condition on $\inf_{u \in A} \|Xu\|_2$ in Section 4.2. Note that RSC analysis for GLMs was considered in [26] for specific norms, especially $L_1$, whereas our analysis applies to any set $A \subseteq S^{p-1}$, and hence to any norm. Further, following similar argument structure as in Section 4.2 the analysis for GLMs can be extended to anisotropic and dependent design matrices.

6 Conclusions

The paper presents a general set of results and tools for characterizing non-asymptotic estimation error in norm regularized regression problems. The analysis holds for any norm, and subsumes much of existing literature focused on structured sparsity and related themes. The work can be viewed as a direct generalization of results in [26], which presented related results for decomposable norms. Our analysis illustrates the important role Gaussian widths, as a measure of size of suitable sets, play in such results. Further, the error sets for regularized and constrained versions of such problems are shown to be closely related [7].

While the paper presents a unified geometric treatment of non-asymptotic structured estimation with regularized estimators, several technical questions need further investigation. The focus of the analysis has been on thin-tailed distributions, and the RE/RSC type analysis presented really gives two sided bounds, i.e., RIP, showing that thin-tailed distributions do satisfy the RIP condition. For heavy tailed measurements, the lower and upper tails of quadratic forms behave differently [28] [1], and it may be possible to establish geometric estimation error analysis for general norms, some special cases of which have been investigated in recent years [20, 21, 1]. Further, the sample complexity of the phase transitions in the RE/RSC conditions for anisotropic designs depend on the largest eigenvalue (operator norm) of the covariance matrix, making the estimator sample inefficient for highly correlated designs. Since real-world several problems, including spatial and temporal problems, do have correlated observations, it will be important to investigate estimators which perform well in such settings [17]. Finally, the focus of the work is on parametric estimation, and it will be interesting to explore generalizations of the analysis to non-parametric settings.

Acknowledgements: We thank the reviewers of the conference version [3] for helpful comments and suggestions on related work. We thank Sergey Bobkov, Snigdhansu Chatterjee, and Pradeep Ravikumar for helpful discussions related to the paper. The research was supported by NSF grants IIS-1447566, IIS-1422557, CCF-1451986, CNS-1314560, IIS-0953274, IIS-1029711, and by NASA grant NNX12AQ39A.

Appendix

A Background and Preliminaries

We start with a review of some definitions and well-known results which will be used for our proofs.

A.1 Gaussian Width

In several of our proofs, we use the concept of Gaussian width [19] [14], which is defined as follows.

Definition 1 (Gaussian width) For any set $A \in \mathbb{R}^p$, the Gaussian width of the set $A$ is defined as:

$$w(A) = E_g \sup_{u \in A} \langle g, u \rangle,$$  \hspace{1cm} (71)
where the expectation is over \( g \sim N(0, I_{p \times p}) \), a vector of independent zero-mean unit-variance Gaussian random variable.

The Gaussian width \( w(A) \) provides a geometric characterization of the size of the set \( A \). We consider three perspectives of the Gaussian width, and provide some properties which are used in our analysis. First, consider the Gaussian process \( \{Z_u\} \) where the constituent Gaussian random variables \( Z_u = \langle t, g \rangle \) are indexed by \( u \in A \), and \( g \sim N(0, I_{p \times p}) \). Then the Gaussian width \( w(A) \) can be viewed as the expectation of the supremum of the Gaussian process \( \{Z_u\} \). Bounds on the expectations of Gaussian and other empirical processes have been widely studied in the literature, and we will make use of generic chaining for some of our analysis \cite{33,34,8,23}. Second, \( \langle u, g \rangle \) can be viewed as a Gaussian random projection of each \( u \in A \) to one dimension, and the Gaussian width simply measures the expectation of largest value of such projections. Third, if \( A \) is the unit ball of any norm \( R(\cdot) \), i.e., \( A = \{x \in \mathbb{R}^p \mid R(x) \leq 1\} \), then \( w(A) = E_g[R^*(g)] \) by definition of the dual norm. Thus, the Gaussian width is the expected value of the dual norm of a standard Gaussian random vector. For instance, if \( A \) is unit ball of \( L_1 \) norm, \( w(A) = E||g||_\infty \).

Below we list some simple and useful properties of the Gaussian width of \( A \subseteq \mathbb{R}^p \):

**Property 1:** \( w(A) \leq w(B) \) for \( A \subseteq B \).

**Property 2:** \( w(A) = w(\text{conv}(A)) \), where \( \text{conv}(\cdot) \) denotes the convex hull of \( A \).

**Property 3:** \( w(cA) = cw(A) \) for any positive scalar \( c \), in which \( cA = \{cx \mid x \in A\} \).

**Property 4:** \( w(\Gamma A) = w(A) \) for any orthogonal matrix \( \Gamma \in \mathbb{R}^{p \times p} \).

**Property 5:** \( w(A + b) = w(A) \) for any \( A \subseteq \mathbb{R}^p \) and fixed \( b \in \mathbb{R}^p \).

The last two properties illustrate the Gaussian width is rotation and translation invariant.

### A.2 Dudley-Sudakov Inequality

The Dudley-Sudakov inequality connects Gaussian widths with covering numbers \cite{33, 23}. We begin with the definition of the covering number of a set \( A \).

**Definition 2** **Covering number:** For any set \( A \subseteq S_{p-1} \) and any \( \epsilon \), we say \( \mathcal{N}_\epsilon(A) \subseteq A \) is an \( \epsilon \)-net of \( A \), or \( \epsilon \)-cover of \( A \), if for every \( u \in A \) there exists a \( v \in \mathcal{N}_\epsilon(A) \) such that \( ||u - v||_2 \leq \epsilon \). Then the covering number \( N(A, \epsilon) \) is the smallest cardinality \( \epsilon \)-covering of \( A \), defined as

\[
N(A, \epsilon) = \min \{ |\mathcal{N}_\epsilon(A)| : \mathcal{N}_\epsilon(A) \text{ is an } \epsilon \text{ net of } A \} .
\]  

\( (72) \)

The covering number can be defined using distances \( d \) different from the \( L_2 \) norm, and such covering numbers are denoted as \( N(A, \epsilon, d) \). The Dudley-Sudakov inequality and our usage of covering numbers in this paper are based on the \( L_2 \) norm distance.

The Dudley-Sudakov inequality \cite{16, 23} gives upper and lower bound of the Gaussian width in terms of covering numbers. In particular, the inequality states that

\[
\sup_{\epsilon > 0} c\epsilon \sqrt{\log(N(A, \epsilon))} \leq w(A) \leq \int_0^\infty \sqrt{\log(N(A, \epsilon))} \, d\epsilon .
\]  

\( (73) \)
where $c > 0$ is an absolute constant.

For our analysis, we use the lower bound, which is often called weak converse of Dudley’s inequality or Sudakov’s inequality. Without loss of generality, our analysis will often use $\epsilon = \frac{1}{4}$. With such a specific constant choice of $\epsilon$, the weak converse of Dudley’s inequality gives an upper bound on the covering number in terms of the Gaussian width, i.e.,

$$N \left( A, \frac{1}{4} \right) \leq \exp \left( c_1 w^2 (A) \right). \quad (74)$$

The following lemmas give useful mechanisms to move from analysis done on a $\epsilon$-net of a set to the full set.

**Lemma 4** Let $\mathcal{N}_\epsilon(A)$ be a $\epsilon$-net of $A$ for some $\epsilon \in [0, 1)$. Then,

$$\sup_{x \in A} \langle x, s \rangle \leq \max_{v \in \mathcal{N}_\epsilon(A)} |\langle v, s \rangle| + \epsilon \|s\|_2. \quad (75)$$

**Lemma 5** Let $C$ be a symmetric $p \times p$ matrix, and let $\mathcal{N}_\epsilon(A)$ be an $\epsilon$-net of $A \subseteq S^{p-1}$ for some $\epsilon \in [0, \frac{1}{2})$. Then,

$$\sup_{u \in A} |\langle Cu, u \rangle| \leq \frac{1}{1 - 2\epsilon} \max_{v \in \mathcal{N}_\epsilon(A)} |\langle Cv, v \rangle|. \quad (76)$$

### A.3 Sub-Gaussian and Sub-exponential Random Variables (Vectors)

In the proof, we will also frequently use the properties of sub-Gaussian and sub-exponential random variables (vectors). In particular, we are interested in their definitions using moments.

**Definition 3** Sub-Gaussian (sub-exponential) random variable: We say that a random variable $x$ is sub-Gaussian (sub-exponential) if the moments satisfies

$$[E|x|^p]^{\frac{1}{p}} \leq K_2 \sqrt{p} \quad ([E|x|^p]^{\frac{1}{p}} \leq K_1p) \quad (77)$$

for any $p \geq 1$ with a constant $K_2$ ($K_1$). The minimum value of $K_2$ ($K_1$) is called sub-Gaussian (sub-exponential) norm of $x$, denoted by $\|x\|_{\psi_2}$ ($\|x\|_{\psi_1}$).

**Definition 4** Sub-Gaussian (sub-exponential) random vector: We say that a random vector $X$ in $\mathbb{R}^n$ is sub-Gaussian (sub-exponential) if the one-dimensional marginals $\langle X, x \rangle$ are sub-Gaussian (sub-exponential) random variables for all $x \in \mathbb{R}^p$. The sub-Gaussian (sub-exponential) norm of $X$ is defined as

$$\|X\|_{\psi_2} = \sup_{x \in S^{n-1}} \|\langle X, x \rangle\|_{\psi_2} \quad (\|X\|_{\psi_1} = \sup_{x \in S^{n-1}} \|\langle X, x \rangle\|_{\psi_1}) \quad (78)$$

The following definitions and lemmas are from [38].

**Lemma 6** Consider a finite number of independent centered sub-Gaussian random variables $X_i$. Then $\sum_i X_i$ is also a centered sub-Gaussian random variable. Moreover,

$$\left\| \sum_i X_i \right\|_{\psi_2}^2 \leq C \sum_i \|X_i\|_{\psi_2}^2 \quad (79)$$
Lemma 7 Let $X_1, \ldots, X_n$ be independent centered sub-Gaussian random variables. Then $X = (X_1, \ldots, X_n)$ is a centered sub-Gaussian random vector in $\mathbb{R}^n$, and

$$\|X\|_\psi^2 \leq C \max_{i \leq n} \|X_i\|_\psi^2$$

(80)

where $C$ is an absolute constant.

Lemma 8 Consider a sub-Gaussian random vector $X$ with sub-Gaussian norm $K = \max_i \|X_i\|_\psi$, then, $Z = \langle X, a \rangle$ is a sub-Gaussian random variable with sub-Gaussian norm $\|Z\|_\psi^2 \leq CK\|a\|_2$.

Lemma 9 A random variable $X$ is sub-Gaussian if and only if $X^2$ is sub-exponential. Moreover,

$$\|X\|_\psi^2 \leq \|X^2\|_\psi^1 \leq 2\|X\|_\psi^2$$

(81)

Lemma 10 If $X$ is sub-Gaussian (or sub-exponential), then so is $X - EX$. Moreover, the following holds,

$$\|X - EX\|_\psi^2 \leq 2\|X\|_\psi^2, \quad \|X - EX\|_\psi^1 \leq 2\|X\|_\psi^1$$

(82)

B Restricted Error Set and Recovery Guarantees

Section 2 is about the restricted error set. Lemma 1 characterizes the restricted error set. Theorem 1 establishes the relation between the constrained and restricted error sets. In particular, we prove that the Gaussian width of the regularized and constrained error sets (cone) are of the same order. Starting with the assumption that the RSC condition is satisfied Lemma 2 and Theorem 2 derive results on the upper bound on the $L_2$ norm of the error.

We collect the proofs of the different results in this section.

B.1 The Restricted Error Set

Lemma 1 in Section 2 characterizes the set to which the error vector belongs. We give the proof of Lemma 1 below:

**Lemma 1** For any $\beta > 1$, assuming

$$\lambda_n \geq \beta R^*(\nabla \mathcal{L}(\theta^*; Z^n))$$

(83)

where $R^*(\cdot)$ is the dual norm of $R(\cdot)$. Then the error vector $\hat{\Delta}_n = \hat{\theta}_{\lambda_n} - \theta^*$ belongs to the set:

$$E_r = E_r(\theta^*, \beta) = \left\{ \Delta \in \mathbb{R}^p \mid R(\theta^* + \Delta) \leq R(\theta^*) + \frac{1}{\beta} R(\Delta) \right\}.$$  

(84)

**Proof:** By the optimality of $\hat{\theta}_{\lambda_n} = \theta^* + \hat{\Delta}_n$, we have

$$\mathcal{L}(\theta^* + \hat{\Delta}_n) + \lambda_n R(\theta^* + \hat{\Delta}_n) - \{ \mathcal{L}(\theta^*) + \lambda_n R(\theta^*) \} \leq 0.$$  

(85)

Now, since $\mathcal{L}$ is convex,

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

(86)
Further, by generalized Holder’s inequality, we have
\[ |\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle| \leq R^* (\nabla \mathcal{L}(\theta^*)) R(\Delta) \leq \frac{\lambda_n}{\beta} R(\Delta), \tag{87} \]
where we have used \( \lambda_n \geq \beta R^* (\nabla \mathcal{L}(\theta^*; Z^n)) \). Hence, we have
\[ \mathcal{L}(\theta^* + \hat{\Delta}_n) - \mathcal{L}(\theta^*) \geq -\frac{\lambda_n}{\beta} R(\hat{\Delta}_n). \tag{88} \]
As a result,
\[ \lambda_n \left\{ R(\theta^* + \hat{\Delta}_n) - R(\theta^*) - \frac{1}{\beta} R(\hat{\Delta}_n) \right\} \leq 0. \tag{89} \]
Noting that \( \lambda_n > 0 \) and rearranging completes the proof. \( \blacksquare \)

### B.2 Relation between the Constrained and Regularized Error Cones

In this section we show that the sizes of the regularized and constrained error sets are of the same order. Recall from [14], that the error set for the constrained setting for atomic norms is a cone given by:
\[ C_c = C_c(\theta^*) = \text{cone} \{ \Delta \in \mathbb{R}^p \mid R(\theta^* + \Delta) \leq R(\theta^*) \} . \tag{90} \]
The error set \( E_r \) is given by:
\[ E_r = E_r(\theta^*, \beta) = \left\{ \Delta \in \mathbb{R}^p \mid R(\theta^* + \Delta) \leq R(\theta^*) + \frac{1}{\beta} R(\Delta) \right\} . \]

Below we provide the proof of Theorem [1]

**Theorem 1** Let \( A_r = E_r \cap \rho B_2^p \) and \( A_c = C_c \cap \rho B_2^p \), where \( B_2^p = \{ u \mid \|u\|_2 \leq 1 \} \) is the unit ball of \( L_2 \) norm and \( \rho \) is any suitable radius. Then, for any \( \beta > 1 \) we have
\[ w(A_r) \leq \left( 1 + \frac{2}{\beta - 1} \frac{\|\theta^*\|_2}{\rho} \right) w(A_c) , \tag{91} \]
where \( w(A) \) denotes the Gaussian width of any set \( A \) given by: \( w(A) = E_g[\sup_{a \in A} \langle a, g \rangle] \), where \( g \) is an isotropic Gaussian random vector.

**Proof:** From triangle inequality, we have
\[ R(\Delta) \leq R(\theta^* + \Delta) + R(\theta^*) . \tag{92} \]
Then,
\[ E_r(\theta^*, \beta) = \left\{ \Delta \in \mathbb{R}^p \bigg| R(\theta^* + \Delta) \leq R(\theta^*) + \frac{1}{\beta} R(\Delta) \right\} \]
\[ \subseteq \left\{ \Delta \in \mathbb{R}^p \bigg| R(\theta^* + \Delta) \leq R(\theta^*) + \frac{1}{\beta} R(\theta^* + \Delta) + \frac{1}{\beta} R(\theta^*) \right\} \]
\[ = \left\{ \Delta \in \mathbb{R}^p \bigg| \left( 1 - \frac{1}{\beta} \right) R(\theta^* + \Delta) \leq \left( 1 + \frac{1}{\beta} \right) R(\theta^*) \right\} \]
\[ = \left\{ \Delta \in \mathbb{R}^p \bigg| R(\theta^* + \Delta) \leq \frac{\beta + 1}{\beta - 1} R(\theta^*) \right\} = \bar{E}_r(\theta^*, \beta) . \]
Let $\bar{C}_r(\theta^*, \beta)$ denote the following set

$$\bar{C}_r = \bar{C}_r(\theta^*, \beta) = \text{cone}\left\{ \Delta - \frac{2}{\beta - 1} \theta^* \mid \Delta \in E_r \right\} + \frac{2}{\beta - 1} \theta^*. $$

(93)

It follows naturally from the construction that $E_r \subseteq \bar{C}_r$.

Let $\bar{A}_r = \bar{C}_r(\theta^*, \beta) \cap \rho B_2^p$. Since $E_r(\theta^*, \beta) \subseteq \bar{C}_r(\theta^*, \beta)$, we have $w(\bar{A}_r) \leq w(\bar{A}_r)$. We define two additional sets for our analysis:

$$\bar{B}_r = \bar{A}_r - \frac{2}{\beta - 1} \theta^* = \left\{ \Delta \in \mathbb{R}^p \mid \Delta + \frac{2}{\beta - 1} \theta^* \in \bar{A}_r \right\}, \quad (94)$$

$$\bar{D}_c = C_c(\theta^*, \beta) \cap \left( \rho + \frac{2}{\beta - 1} \| \theta^* \|_2 \right) \rho B_2^p = \left\{ \Delta \in \mathbb{R}^p \mid \Delta \in \bar{C}_c, \| \Delta \|_2 \leq \left( \rho + \frac{2}{\beta - 1} \| \theta^* \|_2 \right) \rho B_2^p \right\}. \quad (95)$$

For any set $S$ and any $t > 0$, following the properties of Gaussian width since $w(tS) = tw(S)$, we have

$$w(\bar{D}_c) = \left( 1 + \frac{2}{\beta - 1} \frac{\| \theta^* \|_2}{\rho} \right) w(A_c). \quad (96)$$

Further, since Gaussian width is translation invariant, we have

$$w(\bar{A}_r) = w(B_r). \quad (97)$$

From the construction it is clear that $\bar{B}_r \subseteq \bar{D}_c$. Hence we have

$$w(D_c) \geq w(\bar{B}_r) \quad (98)$$

Then, we have

$$w(\bar{A}_r) = w(\bar{B}_r) \leq w(\bar{D}_c) = \left( 1 + \frac{2}{\beta - 1} \frac{\| \theta^* \|_2}{\rho} \right) w(A_c)$$

Noting that $w(A_r) \leq w(\bar{A}_r)$ completes the proof. ■

### B.3 Recovery Guarantees

Lemma 2 and Theorem 2 in the paper are results which establish recovery guarantees. The result in Lemma 2 depends on $\theta^*$, which is unknown. On the other hand Theorem 2 gives the result in terms of quantities like $\lambda_n$ and the norm compatibility constant $\Psi(E_r) = \sup_{u \in E_r} \frac{R(u)}{\| u \|_2}$ which are easier to compute or bound. In this section we give proofs of Lemma 2 and Theorem 2.

**Lemma 2** Assume that the RSC condition is satisfied in $E_r$ by the loss $\mathcal{L}(\cdot)$ with parameter $\kappa$. With $\hat{\Delta}_n = \hat{\theta}_n - \theta^*$, for any norm $R(\cdot)$, we have

$$\| \hat{\Delta}_n \|_2 \leq \frac{1}{\kappa} \| \nabla \mathcal{L}(\theta^*) + \lambda_n \nabla R(\theta^*) \|_2, \quad (99)$$

where $\nabla R(\cdot)$ is any sub-gradient of the norm $R(\cdot)$.

**Proof:** By the RSC property in $E_r$, for any $\Delta \in E_r$ we have

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle + \kappa \| \Delta \|_2^2 \quad (100)$$

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle + \kappa \| \Delta \|_2^2$$

24
Figure 5: Error cone for $L_1$ norm in two dimensions: (a) The L1 norm ball in two dimensions; (b) The constrained error cone $A_c$; (c) The regularized error cone $\tilde{A}_r$ and the shifted cone $\tilde{B}_r$; (d) The constrained error cone $A_c$ and the shifted constrained error cone $\tilde{D}_c$; (e) All error cones.
Also, recall that any norm is convex, since by triangle inequality, for \( t \in [0, 1] \), we have
\[
R(t\theta_1 + (1-t)\theta_2) \leq R(t\theta_1) + R((1-t)\theta_2) = tR(\theta_1) + (1-t)R(\theta_2). 
\] (101)
As a result, for any sub-gradient \( \nabla R(\theta) \) of \( R(\theta) \), we have
\[
R(\theta^* + \Delta) - R(\theta^*) \geq \langle \Delta, \nabla R(\theta^*) \rangle. 
\] (102)
Adding (100) and (102), we get
\[
\lambda_n (\nabla L(\theta^*) + \lambda_n \nabla R(\theta^*)) \geq -\langle \lambda_n \nabla R(\theta^*), \Lambda \rangle + \kappa \|
\Delta \|^2. 
\] (103)
Now, by Cauchy-Schwartz inequality, we have
\[
|\langle \nabla L(\theta^*) + \lambda_n \nabla R(\theta^*), \Delta \rangle| \leq \|
\nabla L(\theta^*) + \lambda_n \nabla R(\theta^*)\|_2 \|
\Delta \|^2. 
\] (104)
Using (104) in (103), we have
\[
\mathcal{F}(\Delta) = \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda_n (R(\theta^* + \Delta) - R(\theta^*)) 
\geq -\|
\nabla \mathcal{L}(\theta^*) + \lambda_n \nabla R(\theta^*)\|_2 \|
\Delta \|^2 + \kappa \|
\Delta \|^2 
\geq \kappa \|
\Delta \|^2 \begin{cases} \|
\Delta \|^2 - \frac{\|
\nabla \mathcal{L}(\theta^*) + \lambda_n \nabla R(\theta^*)\|_2}{\kappa} \end{cases}. 
\] (105)
Now, since \( \mathcal{F}(\Delta_n) \leq 0 \), from (105), we have
\[
\|
\hat{\Delta}_n \|^2 \leq \frac{\|
\nabla \mathcal{L}(\theta^*) + \lambda_n \nabla R(\theta^*)\|_2}{\kappa}, 
\] (106)
which completes the proof.

**Theorem 2.** Assume that the RSC condition is satisfied in \( E_r \) by the loss \( \mathcal{L}(\cdot) \) with parameter \( \kappa \). With \( \hat{\Delta}_n = \hat{\theta}_{\lambda_n} - \theta^* \), for any norm \( R(\cdot) \), we have
\[
\|
\hat{\Delta}_n \|^2 \leq \frac{1 + \beta \lambda_n R(\theta^*)}{\kappa} \Psi(E_r). 
\] (107)

**Proof:** By the RSC property in \( E_r \), we have for any \( \Delta \in E_r \)
\[
\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle + \kappa \|
\Delta \|^2. 
\] (108)
By definition of a dual norm, we have
\[
|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle| \leq R^*(\nabla \mathcal{L}(\theta^*)) R(\Delta). 
\] (109)
Further, by construction, \( R^*(\nabla \mathcal{L}(\theta^*)) \leq \frac{\lambda_n}{\beta} \), implying
\[
|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle| \leq \frac{\lambda_n}{\beta} R(\Delta) 
\Rightarrow \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle \geq -\frac{\lambda_n}{\beta} R(\Delta). 
\] (110)
Further, from triangle inequality, we have
\[ R(\theta^* + \Delta) - R(\theta^*) \geq -R(\Delta) \quad (111) \]

Adding (110) and (111), we have
\[ F(\Delta) = L(\theta^* + \Delta) - L(\theta^*) + \lambda_n (R(\theta^* + \Delta) - R(\theta^*)) \geq -\frac{\lambda_n}{\beta} R(\Delta) + \kappa \|\Delta\|_2^2 - \lambda_n R(\Delta) \]
\[ = \kappa \|\Delta\|_2^2 - \lambda_n \frac{1 + \beta}{\beta} R(\Delta) . \quad (112) \]

By definition of the norm compatibility constant \( \Psi(E_r) \), we have \( R(\Delta) \leq \|\Delta\|_2 \Psi(E_r) \) implying \(-R(\Delta) \geq -\|\Delta\|_2 \Psi(E_r) \). Plugging the inequality back into (112), we have
\[ F(\Delta) \geq \kappa \|\Delta\|_2 \left\{ \|\Delta\|_2 - \frac{1 + \beta}{\beta} \frac{\lambda_n}{\kappa} \Psi(E_r) \right\} . \quad (113) \]

Since \( F(\hat{\Delta}_n) \leq 0 \), we have
\[ \|\hat{\Delta}_n\|_2 \leq \frac{1 + \beta}{\beta} \frac{\lambda_n}{\kappa} \Psi(E_r) , \quad (114) \]
which completes the proof.

C Bounds on the Regularization Parameter

In this section, we prove Theorem [3] and Theorem [4] in Section [3] of the paper. The regularization parameter should satisfy the condition \( \lambda_n \geq \beta R^*(\nabla L(\theta^*; Z^n)) \). In Theorem [3] we establish an upper bound on \( R^* (\nabla L(\theta^*; Z^n)) \) in terms of the Gaussian width of the error cone for least squares loss and Gaussian designs. In Theorem [4] we obtain similar bounds but for the case of sub-Gaussian design. Here we will give more details about the concentration result of \( R^* (\nabla L(\theta^*; Z^n)) \) for Theorem [4].

C.1 Proof of Theorem [3]

**Theorem [3]** Let \( \Omega_R = \{ u : R(u) \leq 1 \} \) and \( \Phi^2 = \sup_{R(u) = 1} \|u\|_2^2 \). For Gaussian design \( X \) and Gaussian or sub-Gaussian noise \( \omega \), we have
\[ E[R^* (\nabla L(\theta^*; Z^n)))] \leq \frac{\eta_1}{\sqrt{n}} w(\Omega_R) . \quad (115) \]

The constant \( \eta_1 \) is given by
\[ \eta_1 = \begin{cases} 1 & \text{if } X \text{ is independent isotropic} \\ \frac{1}{\sqrt{\Lambda_{\max}(\Sigma)}} & \text{if } X \text{ is independent anisotropic} \\ \frac{1}{\sqrt{\Lambda_{\max}(\Gamma)}} & \text{if } X \text{ is dependent isotropic} \end{cases} . \]

Further for any \( \tau > 0 \), with probability at least \( 1 - 3 \exp(- \min(\tau^2, cn)) \), we have
\[ R^* (\nabla L(\theta^*; Z^n)) \leq \frac{\eta_1 \eta_2}{\sqrt{n}} (w(\Omega_R) + \tau) , \quad (116) \]
where \( c \) is an absolute constant, \( \eta_2 = 2 \) when \( \omega \) is i.i.d. standard Gaussian, and \( \eta_2 = \sqrt{2K^2 + 1} \) when \( \omega \) is i.i.d. centered unit-variance sub-Gaussian with \( \|x\|_{\psi_2} \leq K \).
Proof: For least squares loss, we first note that
\[
R^*(\nabla \mathcal{L}(\theta^*; Z^n)) = R^*(\frac{1}{n} X^T \omega) = \sup_{R(u) \leq 1} \left\langle \frac{1}{n} X^T \omega, u \right\rangle = \frac{1}{n} \|\omega\|_2 \cdot \sup_{R(u) \leq 1} \left\langle X^T \frac{\omega}{\|\omega\|_2}, u \right\rangle.
\]
(117)

\|\omega\|_2 is the length of a Gaussian or sub-Gaussian random vector. Then we consider \( \sup_{R(u) \leq 1} \langle X^T \frac{\omega}{\|\omega\|_2}, u \rangle \) for different \( X \).

Case 1. If \( X \) is independent isotropic Gaussian, then
\[
\sup_{R(u) \leq 1} \left\langle X^T \frac{\omega}{\|\omega\|_2}, u \right\rangle = \sup_{R(u) \leq 1} \langle g, u \rangle,
\]
where \( g \) is a standard Gaussian random vector.

Case 2. If \( X \) is independent anisotropic Gaussian, then
\[
\sup_{R(u) \leq 1} \left\langle X^T \frac{\omega}{\|\omega\|_2}, u \right\rangle \leq \sup_{R(\Sigma^{\frac{1}{2}} u) \leq \sqrt{\Lambda_{\max}(\Sigma)}} \langle g, \Sigma^{\frac{1}{2}} u \rangle \leq \sup_{R(u) \leq \sqrt{\Lambda_{\max}(\Sigma)}} \langle g, u \rangle = \sqrt{\Lambda_{\max}(\Sigma)} \sup_{R(u) \leq 1} \langle g, u \rangle,
\]
where \( \sqrt{\Lambda_{\max}(\Sigma)} \) is the norm of the linear operator \( \Sigma^{\frac{1}{2}} \) mapping from the Banach space equipped with norm \( R(\cdot) \) to itself.

Case 3. If \( X \) is dependent isotropic Gaussian, then
\[
\sup_{R(u) \leq 1} \left\langle X^T \frac{\omega}{\|\omega\|_2}, u \right\rangle \leq \sqrt{\Lambda_{\max}(\Gamma)} \sup_{R(u) \leq 1} \left\langle X^T \Gamma^{\frac{1}{2}} \frac{\omega}{\|\Gamma^{\frac{1}{2}} \omega\|_2}, u \right\rangle \leq \sqrt{\Lambda_{\max}(\Gamma)} \sup_{R(u) \leq 1} \langle g, u \rangle,
\]
where \( \Lambda_{\max}(\Gamma) \) denotes the largest eigenvalue of \( \Gamma \).

For convenience, we denote \( Z = \sup_{R(u) \leq 1}(g, u) \). We also denote \( \eta_1 \) as the constant before \( \sup_{R(u) \leq 1}(g, u) \) for the three cases, which are 1, \( \sqrt{\Lambda_{\max}(\Sigma)} \), and \( \sqrt{\Lambda_{\max}(\Gamma)} \) respectively. According to the discussion above, we need upper bounds for \( \|\omega\|_2 \) and \( Z \) in order to bound \( R^*(\nabla \mathcal{L}(\theta^*; Z^n)) \).

Upper Bound for \( \|\omega\|_2 \): If \( \omega \) is standard Gaussian, i.e. \( \omega \sim N(0, I_{n \times n}) \), then \( E[\|\omega\|_2] \) is the expected length of a Gaussian random vector in \( \mathbb{R}^n \), which satisfies
\[
E[\|\omega\|_2] < \sqrt{E[\|\omega\|_2^2]} = \sqrt{n}
\]
(118)

Since \( L_2 \) norm \( \| \cdot \|_2 \) is a 1-Lipschitz function, using the concentration for Lipschitz function of Gaussian random vector, we have
\[
P(||\omega||_2 - E[||\omega||_2] \geq \tau) \leq 2 \exp(-c\tau^2).
\]
Setting $\tau = \sqrt{n}$ and using (118), we get the one-sided concentration

$$P(\|\omega\|_2 \geq 2\sqrt{n}) \leq 2 \exp(-cn).$$  \hfill (119)

If $\omega$ consists of i.i.d. centered unit-variance sub-Gaussian elements with $\|\omega_i\|_{\psi_2} < K$, $\omega_i^2$ is sub-exponential with $\|\omega_i\|_{\psi_1} < 2K^2$. Similarly we have

$$E[\|\omega\|_2] < \sqrt{E[\|\omega\|_2^2]} = \sqrt{n}. \hfill (120)$$

Applying Bernstein’s inequality to $\|\omega\|_2^2 = \sum_{i=1}^{n} \omega_i^2$, we obtain

$$P(\|\omega\|_2^2 - E[\|\omega\|_2^2] \geq \tau) \leq 2 \exp\left[-c \min\left(\frac{\tau^2}{4K^4n}, \frac{\tau}{2K^2}\right)\right]$$

Setting $\tau = 2K^2n$, we get

$$P(\|\omega\|_2 \geq \sqrt{2K^2 + 1}\sqrt{n}) \leq 2 \exp(-cn) \hfill (121)$$

For convenience, we denote $\eta_2$ as the constants before $\sqrt{n}$ term in both (119) and (121), which are 2 for Gaussian case, $\sqrt{2K^2 + 1}$ for sub-Gaussian case.

**Upper Bound for $Z$:** Next we need to bound $Z = \sup_{R(u) \leq 1} \langle g, u \rangle$, where $g$ is a standard Gaussian random vector. According to the definition of Gaussian width, the expectation of $Z$ is

$$\mathbb{E}_g[Z] = \mathbb{E}_g\left[\sup_{R(u) \leq 1} \langle g, u \rangle\right] = w(\Omega_R).$$

Combined with the bounds in (118) and (120), this immediately proves the upper bounds for $\mathbb{E}[R^*(\nabla\mathcal{L}(\theta^*; Z^n))]$ in (115). Then, by invoking the concentration inequality for supreme of Gaussian process, we have

$$P(Z - \mathbb{E}_g[Z] \geq \tau) \leq \exp\left(-\frac{\tau^2}{2\Phi^2}\right),$$

where $\Phi^2 = \sup_{R(u) \leq 1} \mathbb{E}_g[(g, u)^2] = \sup_{R(u) = 1} \|u\|_2^2$. Plugging $\mathbb{E}_g[Z] = w(\Omega_R)$ into the inequality, we have

$$P(Z \geq w(\Omega_R) + \tau) \leq \exp\left(-\frac{\tau^2}{2\Phi^2}\right). \hfill (122)$$

Last, we apply union bound to the inequality (119), (121), and (122), we get

$$R^* (\nabla\mathcal{L}(\theta^*; Z^n)) \leq \frac{\eta_1 \eta_2}{\sqrt{n}} (w(\Omega_R) + \tau),$$

with probability at least $1 - 3 \exp\left(-\min\left(\frac{\tau^2}{2\Phi^2}, cn\right)\right)$, where the choices of $\eta_1$ and $\eta_2$ are determined by the properties of $\omega$ and $X$. This completes the proof.
C.2 Proof of Theorem 4

Theorem 4. Let \( \Omega_R = \{ u : R(u) \leq 1 \} \). For sub-Gaussian design \( X \) and Gaussian or sub-Gaussian noise \( \omega \), we have

\[
E[R^*(\nabla L(\theta^*; Z_n))] \leq \frac{\eta_1 w(\Omega_R)}{\sqrt{n}}.
\]  

(123)

where \( \eta_0 \) is a constant. The constant \( \eta_1 \) is given by

\[
\eta_1 = \begin{cases} 
1 & \text{if } X \text{ is independent isotropic} \\
\sqrt{\Lambda_{\text{max}}(\Sigma)} & \text{if } X \text{ is independent anisotropic} \\
\sqrt{\Lambda_{\text{max}}(\Gamma)} & \text{if } X \text{ is dependent isotropic} 
\end{cases}
\]

In order to prove inequality (123), we first need the following theorem.

Theorem 14. Let \( \Omega_R = \{ u : R(u) \leq 1 \} \) be the unit norm ball. Assuming \( h \) is sub-Gaussian with \(||h||_{\psi_2} \leq k\), then we have

\[
E\left[ \sup_{R(u) \leq 1} \langle h, u \rangle \right] \leq \eta_0 w(\Omega_R),
\]

where \( \eta_0 \) is a constant independent of \( n, p \).

Proof: The quantity \( E[\sup_{R(u) \leq 1} \langle h, u \rangle] \) can be considered the “sub-Gaussian width” of \( \Omega_R \), the unit norm ball, since it has the exact same form as the Gaussian width, with \( h \) being an elementwise independent sub-Gaussian vector instead of a Gaussian vector. Next, we show that the sub-Gaussian width is always bounded by constant times the Gaussian width.

Consider the sub-Gaussian process \( Y = \{ Y_u \}, Y_u = \langle u, h \rangle \) indexed by \( u \in \Omega_R \), the unit norm ball. Consider the Gaussian process \( X = \{ X_u \}, X_u = \langle u, g \rangle \), where \( g \sim N(0, I) \), indexed by the same set, i.e., \( u \in \Omega_R \), the unit norm ball. First, note that \( |Y_u - Y_v| = |\langle h, u - v \rangle| \), so that by Hoeffding’s inequality [38, Proposition 5.10], we have

\[
P(|Y_u - Y_v| \geq \epsilon) = P \left( \left| \sum_{j=1}^{p} (u_j - v_j)h_j \right| \geq \epsilon \right) \leq \epsilon \cdot \exp \left( -\frac{\epsilon^2}{k^2 ||u - v||^2} \right),
\]

(124)

where \( k = \max_j ||h_j||_{\psi_2} \) and \( \epsilon > 0 \) is an absolute constant. As a result, a direct application of the generic chaining argument for upper bounds on such empirical processes [33, Theorem 2.1.5] gives

\[
E\left[ \sup_{u,v} |Y_u - Y_v| \right] \leq \eta_1 E\left[ \sup_u Y_u \right] = \eta_1 w(\Omega_R),
\]

(125)

where \( \eta_1 \) is an absolute constant. Further, since \( \{Y_u\} \) is a symmetric process, from [33, Lemma 1.2.8], we have

\[
E\left[ \sup_{u,v} |Y_u - Y_v| \right] = 2E\left[ \sup_u Y_u \right].
\]

(126)

As a result, with \( \eta_0 = \eta_1 / 2 \), we have

\[
E\left[ \sup_{R(u) \leq 1} \langle h, u \rangle \right] = E\left[ \sup_u Y_u \right] \leq \eta_0 w(\Omega_R).
\]  

(127)

30
That completes the proof.

To prove Theorem 4, we assume that for isotropic independent $X$, the sub-Gaussian norm of each $\langle X_i, u \rangle$ satisfies $|||\langle X_i, u \rangle|||_{\psi_2} \leq k$ for any $u \in S^{p-1}$.

**Proof of Theorem 4.** The equation (117) still holds for sub-Gaussian design matrix, and the analysis for $\|\omega\|_2$ remains the same as Gaussian designs. We also have three cases for $\sup_{R(u) \leq 1} \langle X^T \frac{\omega}{\|\omega\|_2}, u \rangle$.

**Case 4.** If $X$ is independent isotropic sub-Gaussian, then

$$\sup_{R(u) \leq 1} \langle X^T \frac{\omega}{\|\omega\|_2}, u \rangle = \sup_{R(u) \leq 1} \langle h, u \rangle,$$

where $h$ is an i.i.d. sub-Gaussian random vector with $|||h|||_{\psi_2} \leq k$, and its exact distribution depends on $\omega$.

**Case 5.** If $X$ is independent anisotropic sub-Gaussian, then

$$\sup_{R(u) \leq 1} \langle X^T \frac{\omega}{\|\omega\|_2}, u \rangle \leq \sqrt{\Lambda_{\max}(\Sigma)} \sup_{R(u) \leq 1} \langle h, u \rangle,$$

where $\Lambda_{\max}(\Sigma)$ is the maximum eigenvalue of $\Sigma$.

**Case 6.** If $X$ is dependent isotropic sub-Gaussian, then

$$\sup_{R(u) \leq 1} \langle X^T \frac{\omega}{\|\omega\|_2}, u \rangle \leq \sqrt{\Lambda_{\max}(\Gamma)} \sup_{R(u) \leq 1} \langle h, u \rangle,$$

Based on Theorem 14, it is easy to see that

$$E[R^*(\nabla L(\theta^*; Z^n))] \leq \frac{\eta_1}{\sqrt{n}} w(\Omega_R),$$

where $\eta_1$ remains the same as in Theorem 3. This completes the proof.

**C.3 Proof of Theorem 5**

**Theorem 5.** Let $h = [h_1 \cdots h_p]$ be a vector of independent centered random variables, and let $Z' = R^*(h) = \sup_{R(u) \leq 1} \langle h, u \rangle$. Then, we have

$$P\left( |Z' - E[Z']| \geq \tau \right) \leq \eta_3 \exp\left(-\eta_4 \tau^2\right),$$

(128)

if $h$ satisfies any of the following conditions:

(i) $h$ is bounded, i.e., $\|h\|_\infty = L < \infty$.

(ii) For any unit vector $u$, $Y_u = \langle h, u \rangle$ are such that their Malliavin derivatives have almost surely bounded norm in $L^2[0, 1]$, i.e., $\int_0^1 |D_r Y_u|^2 dr \leq L^2$.

**Proof:** The key aspect of the results is to show that the concentration happens in a dimensionality independent manner, i.e., the constants involved do not depend on $p$. We also recall that since $Z' = R^*(h) = \sup_{R(u) \leq 1} \langle h, u \rangle$, one can view $Z'$ as the 1-Lipschitz function of $h$, or the supremum of a sub-Gaussian process determined by $h$.

Part (i): When $h$ is bounded, i.e., $\|h\|_\infty = L < \infty$, the result follows from existing results on measure concentration of convex 1-Lipschitz functions on product spaces, e.g., see Chapter 4.2 in [22]. In particular, since the distribution over $h$ is a product measure, the elements being independent, and $R^*(\cdot)$ is a convex
1-Lipschitz function, following [22 Corollary 4.10] and subsequent discussions, it follows that for every \( \tau \geq 0 \),
\[
P(|Z' - m(Z')| \geq \tau) \leq 4 \exp(-\tau^2/L^2),
\]
where \( m(Z') \) is the median of \( Z' \) for \( P \). One can go from the median \( m(Z') \) to the mean \( E[Z'] \) with only changes in constants, as shown in [22 Proposition 1.8]; also see discussions leading to [22 Corollary 4.8].

Part (ii): In this setting, \( Y_u = (h, u) \) is assumed to such that their Malliavin derivatives have almost surely bounded norm in \( L^2[0, 1] \), i.e., \( \int_0^1 |D_r Y_u|^2 \, dr \leq L^2 \). As a result, it follows that \( \sup_u \int_0^1 |D_r Y_u|^2 \leq L^2 \).

Recall that \( Z' = \sup_{a : R(u) \leq 1} Y_u \) and \( E[Z'] = E[\sup_{a : R(u) \leq 1} Y_u] \). Then, following [40 Theorem 3.6], \( Z' - E[Z'] \) is sub-Gaussian relative to \( L^2 \), so that
\[
P(|Z' - E[Z']| > \tau) \leq 2 \exp(-\tau^2/L^2)
\]
That completes the proof.

\[ \blacksquare \]

C.4 Proof of Lemma 3

Lemma 3 Let \( \Omega_R = \{ u : R(u) \leq 1 \} \) be the unit norm ball and \( \Theta_R = \{ u : R(u) = 1 \} \) be the boundary. For any \( \theta \in \Theta_R \) define \( \rho(\hat{\theta}) = \sup_{\theta : R(\theta) \leq 1} \| \theta - \hat{\theta} \|_2 \) is the diameter of \( \Omega_R \) measured with respect to \( \hat{\theta} \). If \( G(\hat{\theta}) = \text{cone}(\Omega_R - \hat{\theta}) \cap \rho(\hat{\theta}) B_2^p \), i.e., the cone of \( \Omega_R - \hat{\theta} \) intersecting the ball of radius \( \rho(\hat{\theta}) \). Then
\[
w(\Omega_R) \leq \inf_{\hat{\theta} \in \Theta_R} w(G(\hat{\theta}))
\]

Proof: For any \( \hat{\theta} \in \Theta_R \), consider the set \( F_R(\hat{\theta}) = \Omega_R - \hat{\theta} = \{ u : R(u + \hat{\theta}) \leq 1 \} \). Since Gaussian width is translation invariant, the Gaussian width of \( \Omega_R \) and \( F_R \) are the same, i.e., \( w(\Omega_R) = w(F_R(\hat{\theta})) \). Since, \( \rho(\hat{\theta}) = \sup_{\theta : R(\theta) \leq 1} \| \theta - \hat{\theta} \|_2 \) is the diameter of \( \Omega_R \) as well as \( F_R(\hat{\theta}) \), a ball of radius \( \rho(\hat{\theta}) \) will include \( F_R(\hat{\theta}) \), so that \( F_R(\hat{\theta}) \subseteq \rho(\hat{\theta}) B_2^p \). Further, by definition, \( F_R(\hat{\theta}) \subseteq \text{cone}(F_R(\hat{\theta})) = \text{cone}(\Omega_R - \hat{\theta}) \). Let \( G(\hat{\theta}) = \text{cone}(\Omega_R - \hat{\theta}) \cap \rho(\hat{\theta}) B_2^p \). By construction, \( F_R(\hat{\theta}) \subseteq G(\hat{\theta}) \). Then,
\[
w(\Omega_R) = w(F_R(\hat{\theta})) \leq w(G(\hat{\theta})).
\]
Noting the analysis holds for any \( \hat{\theta} \in \Theta_R \), completes the proof.

\[ \blacksquare \]

D Restricted Eigenvalue Conditions: Gaussian Designs

We focus on results in Section 4.1 In particular we consider RE conditions for Gaussian design matrices. We give proofs of RE condition for three different cases: (i) The design matrix has i.i.d isotropic Gaussian rows (ii) The design matrix has rows which are independent but the columns are correlated and (iii) The columns are independent and the rows are correlated.

D.1 Independent Isotropic Gaussian (IIG) Designs

Our goal in this section is two-fold: first, we present the RE conditions obtained using our simple proof technique, and show that it is equivalent, up to constants, the RE condition obtained using Gordon’s inequality, an arguably heavy-duty technique only applicable to the IIG setting; and second, we go over some facets of how we present the results, which will apply to all subsequent RE-style results.
**Theorem 6** Let the design matrix \( X \in \mathbb{R}^{n \times p} \) be elementwise independent and normal, i.e., \( x_{ij} \sim N(0, 1) \). Then, for any \( A \subseteq S^{p-1} \), any \( n \geq 2 \), and any \( \tau > 0 \), with probability at least \( (1 - 2 \exp(-\tau^2/72)) \), we have
\[
\inf_{u \in A} \|Xu\|_2 \geq \sqrt{n} - \eta_0 w(A) - \tau ,
\] (132)
\( \eta_0 > 0 \) is an absolute constant.

**Proof:** Consider the function \( f(X) = \|Xu\|_2 \) where \( u \in A \subseteq S^{p-1} \). Then, from triangle inequality, with \( \|X\|_{op} \) denoting the operator norm, we have
\[
f(X_1) - f(X_2) = \|X_1u\|_2 - \|X_2u\|_2 \leq \|(X_1 - X_2)u\|_2
\]
\[
\leq \|X_1 - X_2\|_{op} \|u\|_2 = \|X_1 - X_2\|_{op} \|\text{vec}(X_1) - \text{vec}(X_2)\|_2 ,
\]
implying \( f(X) \) has a Lipschitz constant of 1. Then, since \( X \) is a Gaussian random matrix, following the large deviation bound for Lipschitz functions of Gaussian variables, we have
\[
P\{ \|Xu\|_2 - E[\|Xu\|_2] > \delta \} \leq 2 \exp\left(-\frac{\delta^2}{2}\right) .
\] (133)

Let \( Z_i = \langle X_i, u \rangle \), then \( Z_i \sim N(0, 1) \), and each \( Z_i \) is independent. Let \( Z \in \mathbb{R}^{n} \) be the vector with elements \( Z_i \). Then, \( E[\|Xu\|_2] = E[\|Z\|_2] = \gamma_n \). Then, we have
\[
P\{ \|Xu\|_2 - \gamma_n > \delta \} \leq 2 \exp\left(-\frac{\delta^2}{2}\right) .
\] (134)

Let \( \mathcal{N}_\epsilon(A) \) be an \( \epsilon \)-net covering of \( A \) based on \( L_2 \) norm, and let \( N(A, \epsilon) \) be the covering number. Then, for any \( v \in \mathcal{N}_\epsilon(A) \), we have
\[
P\{ \|Xv\|_2 - \gamma_n > \delta \} \leq 2 \exp\left(-\frac{\delta^2}{2}\right) .
\] (135)

Taking a union bound over all \( v \), and using the weak converse of Dudley’s inequality at \( \epsilon = \frac{1}{4} \), we have
\[
P\left\{ \max_{v \in \mathcal{N}_\epsilon(A)} \|Xv\|_2 - \gamma_n > \delta \right\} \leq 2 N(A, 1/4) \exp\left(-\frac{\delta^2}{2}\right) \leq 2 \exp\left(cw^2(a) - \frac{\delta^2}{2}\right) .
\] (136)

With \( \delta = \sqrt{2cw(A)} + \frac{\tau}{6} \), we have \( \frac{\delta^2}{2} \geq cw^2(A) + \frac{\tau^2}{72} \), so that
\[
P\left\{ \max_{v \in \mathcal{N}_\epsilon(A)} \|Xv\|_2 - \gamma_n > \eta_1 w(A) + \frac{\tau}{6} \right\} \leq 2 \exp\left(-\frac{\tau^2}{72}\right) .
\] (137)

Next, we focus on extending the result from \( \epsilon \)-net to the full set \( A \). We will use results from Lemmas 11 and 12 which are Lemmas 5.36 and 5.4 respectively from \( [38] \).

**Lemma 11** Let \( A \subseteq S^{p-1} \). Consider a matrix \( B \) which satisfies, for some \( \delta > 0 \)
\[
\|Bu\|_2^2 - 1 \leq \max(\delta, \delta^2) , \quad \forall u \in A .
\] (138)

Then,
\[
1 - \delta \leq \inf_{u \in A} \|Bu\|_2 \leq \sup_{u \in A} \|Bu\|_2 \leq 1 + \delta .
\] (139)

Conversely, if \( B \) satisfies (139) for some \( \delta > 0 \) then
\[
\|Bu\|_2^2 - 1 \leq 3 \max(\delta, \delta^2) , \quad \forall u \in A .
\] (140)
Lemma 12 Let $C$ be a symmetric $p \times p$ matrix, and let $\mathcal{N}_\epsilon(A)$ be an $\epsilon$-net of $A \subseteq S^{p-1}$ for some $\epsilon \in [0, 1)$. Then,

$$\sup_{u \in A} \left| \langle Cu, u \rangle \right| \leq \frac{1}{1 - 2\epsilon} \max_{v \in \mathcal{N}_\epsilon(A)} \left| \langle Cv, v \rangle \right|. \quad (141)$$

Note that previously, using the fact $\gamma_n \approx \sqrt{n}$ we have proved with probability $1 - 2 \exp \left( -\frac{\tau^2}{72} \right)$ and $\delta = \eta_1 \frac{w(A)}{\sqrt{n}} + \frac{\tau}{6\sqrt{n}}$

$$1 - \delta \leq \min_{v \in \mathcal{N}_\epsilon(A)} \frac{1}{\sqrt{n}} \|Xv\|_2 \leq \max_{v \in \mathcal{N}_\epsilon(A)} \frac{1}{\sqrt{n}} \|Xv\|_2 \leq 1 + \delta.$$ \quad (142)

Therefore, from the second result in Lemma 11, the following is true with probability $1 - 2 \exp \left( -\frac{\tau^2}{72} \right)$

$$\left| \frac{1}{n} \|Xv\|_2^2 - 1 \right| \leq 3 \max(\delta, \delta^2), \quad \forall v \in \mathcal{N}_\epsilon(A). \quad (143)$$

Now, using the result of Lemma 12 and choosing $\epsilon = \frac{1}{4}$ with probability $1 - 2 \exp \left( -\frac{\tau^2}{72} \right)$

$$\sup_{u \in A} \left| \frac{1}{n} \|Xu\|_2^2 - 1 \right| = \sup_{u \in A} \left| \frac{1}{n} \langle (XTX - I)u, u \rangle \right|$$

$$\leq \frac{1}{1 - 2\epsilon} \max_{v \in \mathcal{N}_\epsilon(A)} \left| \frac{1}{n} \langle (XTX - I)v, v \rangle \right|$$

$$\leq 2 \max_{v \in \mathcal{N}_\epsilon(A)} \left| \frac{1}{n} \|Xv\|_2^2 - 1 \right|$$

$$\leq 6 \max(\delta, \delta^2).$$

Using the above result along with (142), and using the first result in Lemma 11, the following holds with probability $1 - 2 \exp \left( -\frac{\tau^2}{72} \right)$

$$\inf_{u \in A} \frac{1}{\sqrt{n}} \|Xu\|_2 \geq 1 - 6\delta \geq 1 - \eta_1 \frac{w(A)}{\sqrt{n}} - \frac{\tau}{\sqrt{n}}.$$ \quad (144)

where in the last step we denote $\eta_0 = 6\eta_1$. That completes the proof.

We consider the equivalent result one could obtain by directly using Gordon’s inequality \[13, 19, 23\]:

Theorem 7 Let the design matrix $X$ be elementwise independent and normal, i.e., $x_{ij} \sim N(0, 1)$. Then, for any $A \subseteq S^{p-1}$ and any $\tau > 0$, with probability at least $\left( 1 - 2 \exp(-\tau^2/2) \right)$, we have

$$\inf_{u \in A} \|Xu\|_2 \geq \gamma_n - w(A) - \tau,$$ \quad (145)

where $\gamma_n = E[\|g\|_2] > \frac{n}{\sqrt{n+1}}$ is the expected length of a Gaussian random vector in $\mathbb{R}^n$.

Interestingly, the results are equivalent, up to constants. In fact, our technique provides an alternative proof to Gordon’s inequality using a covering argument, and not using Gaussian comparison techniques \[16, 23\].
Proof: We consider the function \( f(X) = \inf_{u \in A} \|Xu\|_2 \), which by the analysis for Theorem 6 has a Lipschitz constant of at most 1. Then, by the large deviation bound for functions of Gaussian variables, we have

\[
P \left\{ \left| \inf_{u \in A} \|Xu\|_2 - E \left[ \inf_{u \in A} \|Xu\|_2 \right] \right| > \delta \right\} \leq 2 \exp \left( -\frac{\delta^2}{2} \right).
\]  

(146)

Now, from an application of Gordon’s inequality \([18, 19, 23]\) (see Theorem 15), we have

\[
E \left[ \inf_{u \in A} \|Xu\|_2 \right] \leq \gamma_n - w(A),
\]

(147)

where \( \gamma_n \geq \frac{n}{\sqrt{n+1}} \) is the expected length of an isotropic Gaussian random vector in \( \mathbb{R}^n \). Then, we have

\[
P \left\{ \inf_{u \in A} \|Xu\|_2 < \gamma_n - w(A) - \delta \right\} \leq 2 \exp \left( -\frac{\delta^2}{2} \right).
\]

(148)

Changing the direction of the inequality and using \( \tau = \delta \) completes the proof.

The form of the result in Theorem 7 is the same as Theorem 6, but has an exact constant of 1 with the Gaussian width term instead of \( \eta_0 \) in Theorem 6. Further, the proof of Theorem 7 relies on Gordon’s inequality which is applicable only for Gaussian design matrices, and cannot be readily extended to other types of design matrices. On the other hand, our analysis technique for Theorem 6 is general, and can be extended with suitable modifications to other types of design matrices like sub-Gaussian design matrices as shown in later sections.

For the sake of completeness, we give a proof of the following key result used in the analysis of Theorem 7:

**Theorem 15** Let the design matrix \( X \) be elementwise independent and normal, i.e., \( x_{ij} \sim N(0, 1) \). Then, for any \( A \subseteq S^{p-1} \), we have

\[
E \left[ \inf_{u \in A} \|Xu\|_2 \right] \geq \gamma_n - w(A),
\]

(149)

where \( \gamma_n = E[\|h\|_2] \) is the expected length of a Gaussian random vector in \( \mathbb{R}^n \).

For the proof, we start by reviewing the Gaussian comparison form for Gordon’s inequality:

**Theorem 16 (Gordon’s inequality)** Let \( (X_{u,v}, u \in U, v \in V) \) and \( (Y_{u,v}, u \in U, v \in V) \) be centered Gaussian processes, i.e., \( E[X_{u,v}] = E[Y_{u,v}] = 0 \). Assume that

1. \( \|X_{u,v} - X_{u',v'}\|_2 \leq \|Y_{u,v} - Y_{u',v'}\|_2 \) if \( u \neq u' \),
2. \( \|X_{u,v} - X_{u,v'}\|_2 = \|Y_{u,v} - Y_{u,v'}\|_2 \).

Then,

\[
E \left[ \sup_{u \in U} \inf_{v \in V} X_{u,v} \right] \leq E \left[ \sup_{u \in U} \inf_{v \in V} Y_{u,v} \right].
\]

(150)

The proof of the result can be found in Chapter 3 of \([23]\). We will use the following form of the inequality, which can be obtained by applying the above form on \( -X_{u,v} \) and \( -Y_{u,v} \)

\[
E \left[ \inf_{u \in U} \sup_{v \in V} X_{u,v} \right] \leq E \left[ \inf_{u \in U} \sup_{v \in V} Y_{u,v} \right].
\]

(151)
Proof of Theorem 15: The proof follows by standard comparison techniques for Gaussian process. Consider the family of processes
\[ X_{u,v} = \langle v, Xu \rangle = \langle X, u \otimes v \rangle_{TV}, \] (152)
where \( v \in S^{n-1}, u \in S^{p-1} \) are unit vectors. For comparison, we consider the following family of processes
\[ Y_{u,v} = \langle h, v \rangle + \langle g, u \rangle \] (153)
where \( g \sim N(0, I) \) in \( \mathbb{R}^p \) and \( h \sim N(0, I) \) in \( \mathbb{R}^n \). Recall that for a canonical Gaussian process \( X_t = \langle g, t \rangle \), where \( g \sim N(0, I) \), we have
\[ ||X_s - X_t||_2 = ||t - s||_2. \] (154)
Hence, we have
\[ ||X_{u,v} - X_{u',v'}||_2 = ||u \otimes v - u' \otimes v'||_2, \] (155)
\[ ||Y_{u,v} - Y_{u',v'}||_2 = \left\| \begin{bmatrix} u \\ v \end{bmatrix} - \begin{bmatrix} u' \\ v' \end{bmatrix}\right\|. \] (156)
A direct calculation shows that, for unit vectors \( u, v, u', v' \), the conditions for Gordon’s inequality is satisfied by the two families of processes, i.e.,
\[ ||u \otimes v - u' \otimes v'||_2 \leq \sqrt{||u - u'||_2^2 + ||v - v'||_2^2}, \] (157)
\[ ||u \otimes v - u \otimes v'||_2^2 = ||v - v'||_2. \] (158)
Hence, for any \( A \subseteq S^{p-1} \), by Gordon’s inequality we have
\[ E \left[ \inf_{u \in A} ||X_u||_2 \right] = E \left[ \inf_{u \in A} \sup_{v \in S^{n-1}} \langle v, Xu \rangle \right] = E \left[ \inf_{u \in A} \sup_{v \in S^{n-1}} X_{u,v} \right] \leq E \left[ \inf_{u \in A} \sup_{v \in S^{n-1}} Y_{u,v} \right] = E \left[ \inf_{u \in A} \sup_{v \in S^{n-1}} \{ \langle h, v \rangle + \langle g, u \rangle \} \right] \] (159)
\[ = E \left[ \sup_{v \in S^{n-1}} \langle h, v \rangle \right] - E \left[ \sup_{u \in A} \langle g, u \rangle \right] \leq \gamma_n - w(A), \]
since \( E[\sup_{v \in S^{n-1}} \langle h, v \rangle] = E[||h||_2] = \gamma_n \), the expected length of a Gaussian random vector in \( \mathbb{R}^n \). That completes the proof.

D.2 Independent Anisotropic Gaussian (IAG) Designs

We consider a setting where the rows \( X_i \) of the design matrix are independent, but each row is sampled from an anisotropic Gaussian distribution, i.e., \( X_i \sim N(0, \Sigma) \) where \( X_i \in \mathbb{R}^p \). The setting has been considered in the literature [31] for the special case of \( L_1 \) norms, and sharp results have been established using Gaussian comparison techniques [18, 19, 23]. We show that equivalent results can be obtained by our simple technique, which does not rely on Gaussian comparisons [16, 23].

Theorem 8 Let the design matrix \( X \) be row wise independent and each row \( X_i \sim N(0, \Sigma_{p \times p}) \). Then, for any \( A \subseteq S^{p-1} \) and any \( \tau > 0 \), with probability at least \( 1 - 2 \exp(-\frac{\tau^2}{72}) \), we have
\[ \inf_{u \in A} ||X_u||_2 \geq \sqrt{n} - \eta_0 \sqrt{\Lambda_{\text{max}}(\Sigma)} w(A) - \tau, \] (160)
where $\sqrt{\nu} = \inf_{u \in A} \|\Sigma^{1/2} u\|_2$, $\sqrt{\Lambda_{\text{max}}(\Sigma)}$ denotes the largest eigenvalue of $\Sigma^{1/2}$ and $\eta_0 > 0$ is an absolute constant.

**Proof:** Note that $X \Sigma^{-1/2}$ is a matrix with independent rows and each row is isotropic Gaussian. For any $u \in A$, let $\xi = \Sigma^{1/2} u$, so that $u = \Sigma^{-1/2} \xi$. Then, $X u = X \Sigma^{-1/2} \xi$. Further, following the large deviation bound for Lipschitz functions of Gaussian variables, we have

$$P \left\{ \left| \frac{X \Sigma^{-1/2}}{\|\xi\|_2} \xi \right|_2 - \gamma_n \right| > \delta \right\} \leq 2 \exp \left( -\frac{\delta^2}{2} \right)$$

(161)

Note that, $\sqrt{\nu} = \inf_{u \in A} \|\Sigma^{1/2} u\|_2$ and $\|\Sigma^{1/2} u\|_2 \leq \sqrt{\Lambda_{\text{max}}(\Sigma)}$, so that $\sqrt{\nu} \leq \|\xi\|_2 = \|\Sigma^{1/2} u\|_2 \leq \sqrt{\Lambda_{\text{max}}(\Sigma)}$.

Let $\mathcal{N}_\epsilon(A)$ be an $\epsilon$-net covering of $A$ based on $L_2$ norm, and let $N(A, \epsilon)$ be the covering number. Then, for any $v \in \mathcal{N}_\epsilon(A)$, we have

$$P \left\{ \left| \frac{X v}{\|\Sigma^{1/2} v\|_2} - \gamma_n \right| > \delta \right\} \leq 2 \exp \left( -\frac{\delta^2}{2} \right).$$

(162)

Taking a union bound over all $v$, and using the weak converse of Dudley’s inequality, we have

$$P \left\{ \max_{v \in \mathcal{N}_\epsilon(A)} \left| \frac{X v}{\|\Sigma^{1/2} v\|_2} - \gamma_n \right| > \delta \right\} \leq 2 N(A, \epsilon) \exp \left( -\frac{\delta^2}{2} \right)$$

(163)

$$\leq 2 \exp \left( \frac{c w^2(A) - \delta^2}{2} \right).$$

With $\delta = \sqrt{2c w(A) + \frac{\tau}{6 \sqrt{\Lambda_{\text{max}}(\Sigma)}}}$, we have $\frac{\delta^2}{2} \geq c w^2(A) + \frac{\tau^2}{72 \Lambda_{\text{max}}(\Sigma)}$, so that

$$P \left\{ \max_{v \in \mathcal{N}_\epsilon(A)} \left| \frac{X v}{\|\Sigma^{1/2} v\|_2} - \gamma_n \right| > \eta_1 w(A) + \frac{\tau}{6 \sqrt{\Lambda_{\text{max}}(\Sigma)}} \right\} \leq 2 \exp \left( -\frac{\tau^2}{72 \Lambda_{\text{max}}(\Sigma)} \right).$$

(164)

Next, we focus on extending the result from $\epsilon$-net to the full set $A$. We will use results from Lemma [13] below, which is similar in spirit to the result of Lemma [11] and the proof of which is very similar to the proof of Lemma 14 in [38], and Lemma [12] which is stated in the previous subsection.

**Lemma 13** Let $A \subseteq S^{p-1}$. Consider a matrix $B$ which satisfies, for some $\delta > 0$

$$\left| \frac{\|B u\|_2}{\|\Sigma^{1/2} u\|_2} - 1 \right| \leq \max(\delta, \delta^2), \quad \forall u \in A.$$  

(165)

Then,

$$\sqrt{\nu} - \delta \sqrt{\Lambda_{\text{max}}(\Sigma)} \leq \inf_{u \in A} \|B u\|_2 \leq \sup_{u \in A} \|B u\|_2 \leq \sqrt{\Lambda_{\text{max}}(\Sigma)} + \delta \sqrt{\Lambda_{\text{max}}(\Sigma)}.$$  

(166)

Conversely, if $B$ satisfies (166) for some $\delta > 0$ then

$$\left| \frac{\|B u\|_2}{\|\Sigma^{1/2} u\|_2} - 1 \right| \leq 3 \max(\delta, \delta^2), \quad \forall u \in A.$$  

(167)
Note that previously, using the fact $\gamma_n \approx \sqrt{n}$ we have proved with probability $1 - 2 \exp \left( - \frac{\tau^2}{72\Lambda_{\max}(\Sigma)} \right)$ and 

$\delta = \eta_1 \frac{w(A)}{\sqrt{n}} + \frac{\tau}{6\sqrt{\Lambda_{\max}(\Sigma)n}}$ for all $v \in N'_\epsilon(A)$

\[
1 - \delta \leq \min_{v \in N'_\epsilon(A)} \frac{1}{\sqrt{n}} \frac{\|Xv\|_2^2}{\|\Sigma^{1/2}v\|_2^2} \leq \max_{v \in N'_\epsilon(A)} \frac{1}{\sqrt{n}} \frac{\|Xv\|_2^2}{\|\Sigma^{1/2}v\|_2^2} \leq 1 + \delta
\]

$\Rightarrow \sqrt{\nu} - \delta \sqrt{\Lambda_{\max}(\Sigma)} \leq \min_{v \in N'_\epsilon(A)} \frac{1}{\sqrt{n}} \frac{\|Xv\|_2^2}{\|\Sigma^{1/2}v\|_2^2} \leq \max_{v \in N'_\epsilon(A)} \frac{1}{\sqrt{n}} \frac{\|Xv\|_2^2}{\|\Sigma^{1/2}v\|_2^2} \leq \sqrt{\Lambda_{\max}(\Sigma)} + \delta \sqrt{\Lambda_{\max}(\Sigma)}.$

(168)

Therefore, from the second result in Lemma 13 the following is true with probability at least

$1 - 2 \exp \left( - \frac{\tau^2}{72\Lambda_{\max}(\Sigma)} \right)$

\[
\left| \frac{1}{n} \frac{\|Xv\|_2^2}{\|\Sigma^{1/2}v\|_2^2} - 1 \right| \leq 3 \max(\delta, \delta^2), \quad \forall v \in N'_\epsilon(A)
\]

(169)

Now, using the result of Lemma 12 and choosing $\epsilon = \frac{1}{4}$ with probability at least $1 - 2 \exp \left( - \frac{\tau^2}{72\Lambda_{\max}(\Sigma)} \right)$

\[
\sup_{u \in A} \left| \frac{1}{n} \frac{\|Xu\|_2^2}{\|\Sigma^{1/2}u\|_2^2} - \frac{1}{n} \langle (X^T X - n\Sigma)u, u \rangle \right| \leq \frac{1}{n} \max_{v \in N'_\epsilon(A)} \left| \frac{1}{n} \langle (X^T X - n\Sigma)v, v \rangle \right|
\]

\[
\leq 2 \max_{v \in N'_\epsilon(A)} \left| \frac{1}{n} \frac{\|Xv\|_2^2}{\|\Sigma^{1/2}v\|_2^2} - \frac{1}{n} \frac{\|\Sigma^{1/2}v\|_2^2}{\|Xv\|_2} \right| \leq 6\|\Sigma^{1/2}u\|_2^2 \max(\delta, \delta^2).
\]

Using the above result along with (168), and using the first result in Lemma 13 the following holds with probability at least $1 - 2 \exp \left( - \frac{\tau^2}{72\Lambda_{\max}(\Sigma)} \right)$

\[
\inf_{u \in A} \frac{1}{\sqrt{n}} \frac{\|Xu\|_2}{\|Xv\|_2} \geq \sqrt{\nu} - 6\delta \geq \sqrt{\nu} - \eta_0 \frac{w(A)}{\sqrt{n}} \frac{\sqrt{\Lambda_{\max}(\Sigma)}}{\sqrt{n}} - \frac{\tau}{\sqrt{n}}.
\]

(170)

where in the last step we denote $\eta_0 = 6\eta_1$. That completes the proof.

\[\Box\]

### D.3 Dependent Isotropic Gaussian Designs

We now consider a setting where the rows of the design matrix $\hat{X}$ are isotropic Gaussians, but the columns $\hat{X}_j$ are correlated with $E[\hat{X}_j \hat{X}_j^T] = \Gamma \in \mathbb{R}^{n \times n}$. Interestingly, correlation structure over the columns make the samples dependent, a scenario which has not yet been widely studied in the literature [27, 44]. We show that our simple technique continues to work in this scenario and gives a rather intuitive result.

**Theorem 9** Let $\hat{X} \in \mathbb{R}^{n \times p}$ be a matrix whose rows $\hat{X}_j$ are isotropic Gaussian random vectors in $\mathbb{R}^p$ and the columns $\hat{X}_j$ are correlated with $E[\hat{X}_j \hat{X}_j^T] = \Gamma$. Then, for any set $A \subseteq S^{p-1}$ and any $\tau > 0$, with
probability at least \((1 - 2 \exp \left(-\frac{\tau^2}{72}\right))\), we have

\[
\inf_{u \in A} \|\tilde{X}u\|_2 \geq \sqrt{\text{Tr}(\Gamma)} - \sqrt{\Lambda_{\text{max}}(\Gamma)} \left(\eta_0 w(A) + 12\right) - \tau
\]  

(171)

where \(\eta_0 > 0\) is an absolute constant.

The analysis will rely on the following two results, which respectively give a deviation bound for the norm of a correlated Gaussian random vector, and allows converting the continuous problem to a discrete problem over an \(\epsilon\)-net covering. Both results rely on the fact that \(f(g) = \|\Gamma^{1/2}g\|_2\) is Lipschitz with constant \(\Lambda_{\text{max}}(\Gamma^{1/2}) = \sqrt{\Lambda_{\text{max}}(\Gamma)}\).

**Lemma 14** Let \(\Gamma \in \mathbb{R}^{n \times n}\) be a positive (semi)definite matrix, and \(g \sim N(0, I_{n \times n})\) be isotropic Gaussian random vector. Then, for all \(\delta > 0\),

\[
P\left\{ \|\Gamma^{1/2}g\|_2 - \sqrt{\text{Tr}(\Gamma)} > \delta + 2\sqrt{\Lambda_{\text{max}}(\Gamma)} \right\} \leq 2 \exp \left(-\frac{\delta^2}{2\Lambda_{\text{max}}(\Gamma)}\right). \tag{172}
\]

**Proof:** Since \(\|\Gamma^{1/2}g\|_2\) is Lipschitz with constant \(\sqrt{\Lambda_{\text{max}}}\) by large deviation bounds for functions of Gaussian random variables we have

\[
P\left\{ \|\Gamma^{1/2}g\|_2 - E(\|\Gamma^{1/2}g\|_2) > \delta \right\} \leq 2 \exp \left(-\frac{\delta^2}{2\Lambda_{\text{max}}(\Gamma)}\right) \tag{173}
\]

Integrating this tail bound, we observe that \(\text{var}(\|\Gamma^{1/2}g\|_2) \leq 4\Lambda_{\text{max}}(\Gamma)\). Hence

\[
\left| \sqrt{E[\|\Gamma^{1/2}g\|_2^2]} - E[\|\Gamma^{1/2}g\|_2] \right| = \left| \sqrt{\text{Tr}(\Gamma)} - E[\|\Gamma^{1/2}g\|_2] \right| \leq 2\sqrt{\Lambda_{\text{max}}(\Gamma)} \tag{174}
\]

Combining (173) and (174) we get the required result.

**Lemma 15** Let \(f : \mathcal{X} \mapsto \mathbb{R}\), where \(\mathcal{X} \subseteq \mathbb{R}^n\) be \(L\)-Lipschitz, i.e., \(|f(x) - f(y)| \leq L\|x - y\|_2, \forall x, y \in \mathcal{X}\). Let \(\mathcal{N}_\epsilon\) be an \(\epsilon\)-net covering \(\mathcal{X}\). Then,

\[
\inf_{x \in \mathcal{X}} f(x) \geq \inf_{y \in \mathcal{N}_\epsilon} f(y) - L\epsilon. \tag{175}
\]

**Proof:** We have

\[
|f(x) - f(y)| \leq L\|x - y\|_2 \leq L\epsilon. \tag{176}
\]

Then by triangle inequality

\[
f(x) \geq f(y) - L\epsilon \tag{177}
\]

Taking \(\inf\) over \(x\) and \(y\) on the l.h.s and r.h.s gives the required result.

**Proof of Theorem 9** Let \(X \in \mathbb{R}^{n \times p}\) be a matrix whose rows are independent Gaussian isotropic random vectors. Then, \(\tilde{X} = \Gamma^{1/2}X\). Then, for any \(u \in S^{p-1}\), \(Xu = g \sim N(0, \mathbb{I}_{n \times n})\), so that

\[
\|\tilde{X}u\|_2 = \|\Gamma^{1/2}Xu\|_2 = \|\Gamma^{1/2}g\|_2. \tag{178}
\]
Then, following Lemma [14], we have
\[
P \left\{ \| \tilde{X} u \|_2 - \sqrt{\text{Tr}(\Gamma)} > \delta + 2\sqrt{\Lambda_{\max}(\Gamma)} \right\} \leq 2 \exp \left( -\frac{\delta^2}{2\Lambda_{\max}(\Gamma)} \right) .
\]  
(179)

Let \( \mathcal{N}_\epsilon(A) \) be an \( \epsilon \)-net covering of \( A \). For any \( v \in \mathcal{N}_\epsilon(A) \subseteq S^{p-1} \), we have
\[
P \left\{ \| \tilde{X} v \|_2 - \sqrt{\text{Tr}(\Gamma)} > \delta + 2\sqrt{\Lambda_{\max}(\Gamma)} \right\} \leq 2 \exp \left( -\frac{\delta^2}{2\Lambda_{\max}(\Gamma)} \right) .
\]  
(180)

Taking union bound over all \( v \in \mathcal{N}_\epsilon(A) \), we have
\[
P \left\{ \max_{v \in \mathcal{N}_\epsilon(A)} \left\| \tilde{X} v \|_2 - \sqrt{\text{Tr}(\Gamma)} \right\| > \delta + 2\sqrt{\Lambda_{\max}(\Gamma)} \right\} \leq 2N(A, \epsilon) \exp \left( -\frac{\delta^2}{2\Lambda_{\max}(\Gamma)} \right) .
\]  
(181)

where \( N(A, \epsilon) \) is the covering number. Choosing \( \epsilon = \frac{1}{7} \), and using the weak converse of Dudley’s inequality to convert the union bound in terms of the Gaussian width, we have
\[
P \left\{ \max_{v \in \mathcal{N}_\epsilon(A)} \left\| \tilde{X} v \|_2 - \sqrt{\text{Tr}(\Gamma)} \right\| > \delta + 2\sqrt{\Lambda_{\max}(\Gamma)} \right\} \leq 2 \exp(cw^2(\epsilon)) \exp \left( -\frac{\delta^2}{2\Lambda_{\max}(\Gamma)} \right) .
\]  
(182)

Let \( \delta = \sqrt{2c\Lambda_{\max}(\Gamma)}w(A) + \frac{\tau}{6} \), with \( \tau > 0 \), so that \( \delta^2 \geq 2c\Lambda_{\max}(\Gamma)w^2(A) + \frac{\tau^2}{36} \). Then, with \( \eta_0 = \sqrt{2e} \), we have
\[
P \left\{ \max_{v \in \mathcal{N}_\epsilon(A)} \left\| \tilde{X} v \|_2 - \sqrt{\text{Tr}(\Gamma)} \right\| > \sqrt{\Lambda_{\max}(\Gamma)(\eta_1 w(A) + 2) + \frac{\tau}{6}} \right\} \leq 2 \exp \left( -\frac{\tau^2}{72\Lambda_{\max}(\Gamma)} \right) .
\]  
(183)

Next, we focus on extending the result from \( \epsilon \)-net to the full set \( A \). We will use results from Lemma 16 below, which is similar in spirit to the result of Lemma 11 and the proof of which is very similar to the proof of Lemma 14 in [38], and Lemma 12

**Lemma 16** Let \( A \subseteq S^{p-1} \). Consider a matrix \( B \) which satisfies, for some \( \delta > 0 \)
\[
\left| \frac{\| Bu \|_2^2}{\text{Tr}(\Gamma)} - 1 \right| \leq \max(\delta, \delta^2) , \quad \forall u \in A .
\]  
(184)

Then,
\[
\sqrt{\text{Tr}(\Gamma)} - \delta \sqrt{\text{Tr}(\Gamma)} \leq \inf_{u \in A} \| Bu \|_2 \leq \sup_{u \in A} \| Bu \|_2 \leq \sqrt{\text{Tr}(\Gamma)} + \delta \sqrt{\text{Tr}(\Gamma)} .
\]  
(185)

Conversely, if \( B \) satisfies (183) for some \( \delta > 0 \) then
\[
\left| \frac{\| Bu \|_2^2}{\text{Tr}(\Gamma)} - 1 \right| \leq 3 \max(\delta, \delta^2) , \quad \forall u \in A .
\]  
(186)

Note that previously, we have proved with probability at least \( 1 - 2 \exp \left( -\frac{\tau^2}{72\Lambda_{\max}(\Gamma)} \right) \) and
\[
\delta = \sqrt{\frac{\Lambda_{\max}(\Gamma)}{\text{Tr}(\Gamma)}} (\eta_1 w(A) + 2) + \frac{\tau}{6\sqrt{\text{Tr}(\Gamma)}} \text{ for all } v \in \mathcal{N}_\epsilon(A)
\]
\[
\sqrt{\text{Tr}(\Gamma)} - \delta \sqrt{\text{Tr}(\Gamma)} \leq \min_{v \in \mathcal{N}_\epsilon(A)} \| \tilde{X} v \|_2 \leq \max_{v \in \mathcal{N}_\epsilon(A)} \| \tilde{X} v \|_2 \leq \sqrt{\text{Tr}(\Gamma)} + \delta \sqrt{\text{Tr}(\Gamma)} .
\]  
(187)
Therefore, from the second result in Lemma 16, the following is true with probability at least $1 - 2 \exp \left( -\frac{\tau^2}{72\Lambda_{\max}(\Sigma)} \right)$

$$\frac{\|\tilde{X}v\|^2}{\text{Tr}(\Gamma)} - 1 \leq 3 \max(\delta, \delta^2), \quad \forall v \in \mathcal{N}_c(A)$$

$$\Rightarrow \left| \frac{\|\tilde{X}v\|^2}{\text{Tr}(\Gamma)} - 3 \text{Tr}(\Gamma) \right| \leq 3 \text{Tr}(\Gamma) \max(\delta, \delta^2), \quad \forall v \in \mathcal{N}_c(A).$$

Now, using the result of Lemma 12 and choosing $\epsilon = \frac{1}{4}$ with probability at least $1 - 2 \exp \left( -\frac{\tau^2}{72\Lambda_{\max}(\Sigma)} \right)$

$$\sup_{u \in A} \left| \|\tilde{X}u\|^2 - \text{Tr}(\Gamma) \right| \leq 6 \text{Tr}(\Gamma) \max(\delta, \delta^2).$$

Using the above result along with (187), and using the first result in Lemma 16, the following holds with probability at least $1 - 2 \exp \left( -\frac{\tau^2}{72\Lambda_{\max}(\Sigma)} \right)$

$$\inf_{u \in A} \|\tilde{X}u\|_2 \geq \sqrt{n} - \eta_0 w(A) - \tau,$$

where in the last step we denote $\eta_0 = 6\eta_1$. That completes the proof.

### E Restricted Eigenvalue Conditions: Sub-Gaussian Designs

In this section we consider RE condition for sub-Gaussian design matrices with squared loss. As with the Gaussian case we prove results for three different cases: (i) Design matrices with isotropic i.i.d rows (ii) Design matrices with independent rows but dependent columns and (iii) Design matrices with independent columns but dependent rows.

#### E.1 Independent Isotropic Sub-Gaussian Designs

We start with the setting where the sub-Gaussian design matrix $X \in \mathbb{R}^{n \times p}$ has independent rows $X_i$ and each row is isotropic.

**Theorem 10** Let $X \in \mathbb{R}^{n \times p}$ be a design matrix whose rows $X_i$ are independent isotropic sub-Gaussian random vectors in $\mathbb{R}^p$. Then, for any set $A \subseteq S^{p-1}$ and any $\tau > 0$, with probability at least $(1 - 2 \exp(-\eta_1 \tau^2))$, we have

$$\inf_{u \in A} \|Xu\|_2 \geq \sqrt{n} - \eta_0 w(A) - \tau,$$

where $\eta_0, \eta_1 > 0$ are constants which depend only the sub-Gaussian norm $\|x_{ij}\|_\psi^2 = k$. 

41
The analysis will use results from Lemmas 11 and 12 which respectively help in converting the problem from a minimum singular value problem for $X$ to a minimum eigen-value problem, and from a continuous to a discrete problem.

With $B = \frac{1}{\sqrt{n}} X$, following Lemma 11, it suffices to show

$$\left| \frac{1}{n} \| Xu \|_2^2 - 1 \right| \leq \max(\delta, \delta^2) \delta \equiv t \ \forall u \in A, \text{ where } \delta = \eta_0 \frac{w(A)}{\sqrt{n}} + \frac{\tau}{\sqrt{n}}. \quad (191)$$

From Lemma 12 using an $\epsilon = \frac{1}{4}$-net $N$ on $A \subseteq S^{p-1}$, we have

$$\sup_{u \in A} \left| \frac{1}{n} \| Xu \|_2^2 - 1 \right| = \sup_{u \in A} \left| \frac{1}{n} \langle (X^T X - I)u, u \rangle \right| \leq 2 \max_{v \in N} \left| \frac{1}{n} \langle (X^T X - I)u, v \rangle \right| \leq 2 \max_{v \in N} \left| \frac{1}{n} \| Xu \|_2^2 - 1 \right|. \quad (192)$$

As a result, it suffices to prove that with high probability,

$$\max_{v \in N} \left| \frac{1}{n} \| Xu \|_2^2 - 1 \right| \leq \frac{t}{2}. \quad (193)$$

For any $v \in N \subseteq S^{p-1}$, and any row $X_i$ of $X$, $Z_i = \langle X_i, v \rangle$ are independent sub-Gaussian random variables with $E[Z_i^2] = 1$ and $\|Z_i\|_{\psi_2} \leq K$. Therefore, $Z_i^2 - 1$ are independent centered sub-exponential random variables with

$$\|Z_i^2 - 1\|_{\psi_1} \leq 2\|Z_i^2\|_{\psi_1} \leq 4\|Z_i\|_{\psi_2}^2 \leq 4K^2. \quad (194)$$

By large deviation inequality for sub-exponential variables, we have

$$P \left\{ \left| \frac{1}{n} \| Xu \|_2^2 - 1 \right| \geq \frac{t}{2} \right\} = P \left\{ \left| \frac{1}{n} \sum_{i=1}^n Z_i^2 - 1 \right| \geq \frac{t}{2} \right\} \leq 2 \exp \left( -\frac{c_2}{K^4} \min(t, t^2)n \right) = 2 \exp \left( -\frac{c_2}{K^4} \delta^2 n \right) \quad (195)$$

Taking union bound over all $v \in N$, and using (74), i.e., the upper bound on the cardinality of the covering in terms of the Gaussian width of $A$, we have

$$P \left\{ \max_{v \in N} \left| \frac{1}{n} \| Xu \|_2^2 - 1 \right| \geq \frac{t}{2} \right\} \leq 2N \left( A, \frac{1}{4} \right) \exp \left( -\frac{c_2}{K^4} (\eta_0^2 w^2(A) + \tau^2) \right) \leq 2 \exp(c_1 w^2(A)) \exp \left( -\frac{c_2}{K^4} \eta_0^2 w^2(A) \right) \exp \left( -\frac{c_2}{K^4} \tau^2 \right) = 2 \exp \left( -\left( \frac{c_2}{K^4} - c_1^2 \right) \eta_0^2 w^2(A) \right) \exp \left( -\frac{c_2}{K^4} \tau^2 \right) \leq 2 \exp \left( -\frac{c_2}{K^4} \tau^2 \right),$$

assuming $\eta_0 \geq \frac{K^2 c_1}{\sqrt{c_2}}$. Denoting $\eta_1 = \frac{c_2}{K^4}$ completes the proof. 

42
E.2 Independent Anisotropic Sub-Gaussian Designs

We consider a setting where the rows $X_i$ of the design matrix are independent, but each row is sampled from an anisotropic sub-Gaussian distribution, i.e., $\|x_{ij}\|_{\psi_2} = k$ and $E[X_i X_i^T] \sim N(0, \Sigma)$ where $X_i \in \mathbb{R}^p$. Further, to maintain the results at the same scale as previous results, we assume $\|x_{ij}\|_{2} = 1$ without loss of generality. The setting of anisotropic sub-Gaussian design matrices has been investigated in [32] for the special case of $L_1$ norm. In contrast, our analysis applies to any norm regularization with dependency captured by the Gaussian width of the corresponding error set. Further, the analysis can be viewed as an extension of our general proof technique used in earlier sections.

**Theorem 11** Let the sub-Gaussian design matrix $X$ be row wise independent, and each row has $E[X_i^T X_i] = \Sigma \in \mathbb{R}^{p \times p}$. Then, for any $A \subseteq S^{p-1}$ and any $\tau > 0$, with probability at least $(1 - 2 \exp(-\eta_1 \tau^2))$, we have

$$\inf_{u \in A} \|X u\|_2 \geq \sqrt{\nu} \sqrt{n} - \eta_0 \Lambda_{\max}(\Sigma) \ w(A) - \tau,$$

where $\sqrt{\nu} = \inf_{u \in A} \|\Sigma^{1/2} u\|_2$, $\sqrt{\Lambda_{\max}(\Sigma)}$ denotes the largest eigenvalue of $\Sigma^{1/2}$, and $\eta_0, \eta_1 > 0$ are constants which depend on the sub-Gaussian norm $\|x_{ij}\|_{\psi_2} = k$.

**Proof:** The proof will use the results from Lemmas 12 and 13.

With $B = \frac{1}{\sqrt{n}} X$ and following Lemma 13, it suffices to show that

$$\left| \frac{1}{n} \|X u\|_2^2 - \|\Sigma^{1/2} u\|_2^2 \right| \leq \max(\delta, \delta^2) \triangleq t \ \forall u \in A$$

(197)

where $\delta = \frac{\eta_0 \Lambda_{\max}(\Sigma) \ w(A)}{\sqrt{n}} + \frac{\tau}{\sqrt{n}}$. From Lemma 12, using a $\epsilon = \frac{1}{4}$-net $\mathcal{N}_\epsilon(A)$ on $A \subseteq S^{p-1}$, we have

$$\sup_{u \in A} \left| \frac{1}{n} \|X u\|_2^2 - \|\Sigma^{1/2} u\|_2^2 \right| = \sup_{u \in A} \left| \frac{1}{n} \left( X^T X - \Sigma \right) u, u \right| \leq 2 \max_{v \in \mathcal{N}_\epsilon(A)} \frac{1}{n} \left| \left( X^T X - \Sigma \right) v, v \right|$$

(198)

$$= 2 \max_{v \in \mathcal{N}_\epsilon(A)} \left| \frac{1}{n} \|X v\|_2^2 - \|\Sigma^{1/2} v\|_2^2 \right|$$

As a result, it suffices to prove with high probability

$$\max_{v \in \mathcal{N}_\epsilon(A)} \left| \frac{1}{n} \|X v\|_2^2 - \|\Sigma^{1/2} v\|_2^2 \right| \leq \frac{t}{2}$$

(199)

For any $v \in \mathcal{N}_\epsilon(A) \subseteq S^{p-1}$, and any row $X_i$ of $X$, $Z = (X_i, v)$ are independent sub-Gaussian random variables with $E[Z_i^2] = E[v^T X_i^T X_i v] = v^T E[X_i^T X_i] v = \|\Sigma^{1/2} v\|_2^2$ and $\|Z_i\|_{\psi_2} \leq K \sqrt{\Lambda_{\max}(\Sigma)}$. Therefore $Z_i^2 - \|\Sigma^{1/2} v\|_2^2$ are independent centered sub-exponential random variables with

$$\left\| Z_i^2 - \|\Sigma^{1/2} v\|_2^2 \right\|_{\psi_1} \leq 2 \left\| Z_i^2 \right\|_{\psi_1} \leq 4 \left\| Z_i \right\|_{\psi_2} \leq 4 K^2 \Lambda_{\max}(\Sigma)$$

(200)

Therefore, by large deviation inequality for sub-exponential variables, we have
\[
P \left\{ \frac{1}{n} \|Xv\|_2^2 - \|\Sigma^{1/2}v\|_2^2 \geq \frac{t}{2} \right\} = P \left\{ \frac{1}{n} \sum_{i=1}^n Z_i^2 - \|\Sigma^{1/2}v\|_2^2 \geq \frac{t}{2} \right\} \\
\leq 2 \exp \left( - \frac{c_2}{K^4 \Lambda_{\max}^2(\Sigma)} \min(t, t^2) n \right) \\
= 2 \exp \left( - \frac{c_2}{K^4 \Lambda_{\max}^2(\Sigma)} \delta^2 n \right) \\
\leq 2 \exp \left( - \frac{c_2}{K^4 \Lambda_{\max}^2(\Sigma)} (\eta_0^2 w^2(A) \Lambda_{\max}^2(\Sigma) + \tau^2) \right)
\]

Taking union bound over all \(v \in \mathcal{N}_c(A)\), and using (74), i.e., the upper bound on the cardinality of the covering number in terms of the Gaussian width of \(A\), we have

\[
P \left\{ \max_{v \in \mathcal{N}_c(A)} \frac{1}{n} \|Xv\|_2^2 - \|\Sigma^{1/2}v\|_2^2 \geq \frac{t}{2} \right\} \leq 2N \left( A, \frac{1}{4} \right) \exp \left( - \frac{c_2}{K^4 \Lambda_{\max}^2(\Sigma)} (\eta_0^2 w^2(A) \Lambda_{\max}^2(\Sigma) + \tau^2) \right) \\
\leq 2 \exp(c_1 w^2(A)) \exp \left( - \frac{c_2}{K^4 \Lambda_{\max}^2(\Sigma)} (\eta_0^2 w^2(A) \Lambda_{\max}^2(\Sigma) + \tau^2) \right) \\
= 2 \exp \left( - \left( \frac{c_2 \eta_0^2}{K^4} - c_1 \right) w^2(A) \right) \exp \left( - \frac{c_2}{K^4 \Lambda_{\max}^2(\Sigma)} \tau^2 \right) \\
\leq 2 \exp \left( - \frac{c_2}{K^4 \Lambda_{\max}^2(\Sigma)} \tau^2 \right)
\]

assuming \(\eta_0 \geq \frac{K^2 c_3}{\sqrt{c_2}}\). Denoting \(\eta_1 = \frac{c_2}{K^4 \Lambda_{\max}^2(\Sigma)}\) completes the proof.

**E.3 Dependent Isotropic Sub-Gaussian Designs**

In this section, we consider the setting where the design matrix \(\tilde{X}\) has isotropic sub-Gaussian rows, but the rows are dependent. In particular, we assume that \(\|\tilde{X}_{ij}\|_{\psi_2} = k\), and \(E[\tilde{X}_j \tilde{X}_j^T] = \Gamma \in \mathbb{R}^{n \times n}\). We show that the gain condition holds for such dependent sub-Gaussian designs:

**Theorem 12** Let \(\tilde{X} \in \mathbb{R}^{n \times p}\) be a sub-Gaussian design matrix with isotropic rows and correlated columns with \(E[\tilde{X}_j \tilde{X}_j^T] = \Gamma \in \mathbb{R}^{n \times n}\). Then, for any \(A \subseteq S^{p-1}\) and any \(\tau > 0\), with probability at least \((1 - 2\exp(-\eta_1 \tau^2))\), we have

\[
\inf_{u \in A} \|\tilde{X}u\|_2 \geq \sqrt{\text{Tr}(\Gamma)} - \Lambda_{\max}(\Gamma) \eta_0 w(A) - \tau ,
\]

where \(\eta_0, \eta_1\) are constants which depend on the sub-Gaussian norm \(\|x_{ij}\|_{\psi_2} = k\).

**Proof:** The analysis will use results from Lemma 12 and 16

With \(B = \tilde{X}\) and following Lemma 16 it suffices to show that

\[
\sup_{u \in A} \left\| \|\tilde{X}u\|_2^2 - \text{Tr}(\Gamma) \right\| \leq \max(\delta, \delta^2) \text{Tr}(\Gamma) \triangleq t, \quad \forall u \in A
\]

where \(\delta = \frac{\eta_0 w(A) \Lambda_{\max}(\Gamma)}{\sqrt{\text{Tr}(\Gamma)}} + \frac{\tau}{\sqrt{\text{Tr}(\Gamma)}}\). From Lemma 12 using a \(\epsilon = \frac{1}{4}\)-net \(\mathcal{N}_c(A)\) on \(A \subseteq S^{p-1}\), we have
\[
\sup_{u \in A} \left\| \tilde{X} u \right\|_2^2 - \text{Tr}(\Gamma) = \sup_{u \in A} \left| \langle (\tilde{X}^T \tilde{X} - \text{Tr}(\Gamma)) u, u \rangle \right| \leq 2 \max_{v \in N_i(A)} \left| \langle (\tilde{X}^T \tilde{X} - \text{Tr}(\Gamma)) v, v \rangle \right|
= 2 \max_{v \in N_i(A)} \left\| \tilde{X} v \right\|_2^2 - \text{Tr}(\Gamma)
\] (204)

As a result, it suffices to prove with high probability
\[
\max_{v \in N_i(A)} \left\| \tilde{X} v \right\|_2^2 - \text{Tr}(\Gamma) \leq \frac{t}{2}
\] (205)

Let \( \tilde{Z} = \tilde{X} v \). Also \( \tilde{X} = \Gamma^{1/2} X \) where \( E[\tilde{X}_j \tilde{X}_j^T] = \Gamma \) and \( X \in \mathbb{R}^{n \times p} \) has independent sub-Gaussian entries. Let \( Z = X v \) is an isotropic sub-Gaussian random vector, with each entry \( Z_i = \langle X_i, v \rangle \) such that \( \|Z_i\|_{\psi_2} \leq K \). Then
\[
\|\tilde{X} v\|_2^2 = \|\Gamma^{1/2} Z\|_2^2 = \| \sum_{j=1}^n Z_j \Gamma_j^{1/2} \|_2^2 = \sum_{j=1}^n Z_j^2 \|\Gamma_j^{1/2}\|_2^2 + \sum_{j,k \in [1, \ldots, n], j \neq k} Z_j Z_k \langle \Gamma_j^{1/2}, \Gamma_k^{1/2} \rangle
\] (206)

where \( \Gamma_j^{1/2}, \Gamma_k^{1/2} \) denote the jth and kth columns of \( \Gamma \) and \([1, \ldots, n]\) denotes a set containing the first n natural numbers. We assume that \( \| \sum_{j=1}^n Z_j \Gamma_j^{1/2} \|_2^2 = \text{Tr}(\Gamma) \) almost surely. Then,
\[
\|\tilde{X} v\|_2^2 = \sum_{j=1}^n \|\Gamma_j^{1/2}\|_2^2 + \sum_{j,k \in [1, \ldots, n], j \neq k} Z_j Z_k \langle \Gamma_j^{1/2}, \Gamma_k^{1/2} \rangle = \text{Tr}(\Gamma) + \sum_{j,k \in [1, \ldots, n], j \neq k} Z_j Z_k \langle \Gamma_j^{1/2}, \Gamma_k^{1/2} \rangle
\] (207)

Therefore, we get
\[
\|\tilde{X} v\|_2^2 - \text{Tr}(\Gamma) \leq \left| \sum_{j,k \in [1, \ldots, n], j \neq k} Z_j Z_k \langle \Gamma_j^{1/2}, \Gamma_k^{1/2} \rangle \right|
\] (208)

The sum on the r.h.s is \( \langle \Gamma_0 x, x \rangle \) where \( \Gamma_0 \) is the off-diagonal part of \( \Gamma \). This can also be written as follows:
\[
R_T(z) = \sum_{j,k \in T^c, z \in T^c} Z_j Z_k \langle \Gamma_j^{1/2}, \Gamma_k^{1/2} \rangle
\] (209)

We state the following decoupling Lemma, the proof of which is provided on Pg. 38 of [38] to bound the above quantity

**Lemma 17** Consider a double array of real numbers \((a_{ij})_{i,j=1}^n\), such that \( a_{ii} = 0 \) for all \( i \). Then
\[
\sum_{i,j \in [1, \ldots, n]} a_{ij} = 4 E \sum_{i \in T, j \in T^c} a_{i,j}
\] (210)

where \( T \) is a random subset of \([1, \ldots, n]\) with average size \( n/2 \). In particular
\[
4 \max_{T \subseteq [1, \ldots, n]} \sum_{i \in T \setminus T^c} a_{ij} \leq \sum_{i,j \in [1, \ldots, n]} a_{ij} \leq 4 \max_{T \subseteq [1, \ldots, n]} \sum_{i \in T \setminus T^c} a_{ij}
\] (211)
where the minimum and maximum are over all subsets of \([1, \ldots, n]\).

From the above Lemma,

\[
\|\tilde{X}v\|_2^2 - Tr(\Gamma) \leq 4 \max_{T \subseteq [1, \ldots, n]} |R_T(z)|
\]

Therefore we want to compute the probability of the following event:

\[
P \left\{ \max_{v \in N(A)} T \subseteq [1, \ldots, n] |R_T(z)| > \frac{t}{8} \right\} \leq N \left( A, \frac{1}{4} \right) \cdot 2^n \cdot \max_{v \in N(A), T \subseteq [1, \ldots, n]} P \left\{ |R_T(z)| > \frac{t}{8} \right\}
\]

The r.h.s follows from a union bounding argument.

To estimate the probability, we fix a vector \(v \in N(A)\) and a subset \(T \subseteq [1, \ldots, n]\) and we condition on a realization of random variables \((Z_k)_{k \in T^c}\). Therefore, we express

\[
R_T(z) = \sum_{j \in T} Z_j \langle \Gamma_j^{1/2}, b \rangle, \quad \text{where } b = \sum_{k \in T^c} Z_k \Gamma_k^{1/2}
\]

Under this conditioning \(b\) is a fixed vector, so \(R_T(z)\) is a sum of independent random variables. Moreover,

\[
\|Z\|_2 = \|Xv\|_2 \leq \sqrt{n} + \eta_0 w(A) + \tau \leq c \sqrt{n}
\]

The inequality above follows from our results for the isotropic sub-Gaussian design scenario. It can be proved using the same arguments,

\[
\|R_T(z)\|_{\psi_2} \leq K
\]

and

\[
\sum_{j \in T} \langle \Gamma_j^{1/2}, b \rangle \leq \|\Gamma_j\|_2 \|b\|_2 \leq \Lambda_{\max}(\Gamma) c \sqrt{n}
\]

Denoting the conditional probability by \(P_T = P(\cdot | (Z_k)_{k \in T^c})\) and the expectation with respect to \((A_k)_{k \in T^c}\) by \(E_{T^c}\), we obtain

\[
P \left\{ \left| R_T(z) \right| > \frac{t}{8} \right\} \leq E_{T^c} P_T \left\{ \left| R_T(z) \right| > \frac{t}{8} \right\}
\]

\[
\leq 2 \exp \left[ -c_4 \left( \frac{t/8}{K \Lambda_{\max}(\Gamma) \sqrt{n}} \right)^2 \right]
\]

\[
\leq 2 \exp \left[ -c_5 \left( \frac{\delta Tr(\Gamma)}{K \Lambda_{\max}(\Gamma) \sqrt{n}} \right)^2 \right]
\]

\[
\text{for a large enough } c \text{ the above result is almost always satisfied. Under these conditions } \sum_{j \in T} Z_j \langle \Gamma_j^{1/2}, b \rangle \text{ is a sum of independent sub-Gaussian random variables which in turn is a sub-Gaussian variable with sub-Gaussian norm}
\]

\[
\|R_T(z)\|_{\psi_2} \leq K
\]
Substituting $\delta = \frac{nw(A)\Lambda_{\text{max}}(\Gamma)}{\sqrt{\text{Tr}(\Gamma)}} + \frac{\tau}{\sqrt{\text{Tr}(\Gamma)}}$ and by the earlier argument we get

$$P\left\{ \max_{v \in \mathcal{N}(A), T \subseteq [1, n]} |R_T(z)| > \frac{t}{8} \right\} \leq \exp\left( c_1 w^2(A) \exp(n \ln 2) \exp\left[ -c_5 \left( \frac{\eta_0^2 w^2(A) \text{Tr}(\Gamma)}{K^2 n} + \frac{\tau^2 \text{Tr}(\Gamma)}{K^2 A_{\text{max}}^2(\Gamma) n} \right) \right] \right)$$

Choosing $\eta_0 > \sqrt{\frac{c_1 K^2 n}{c_5 \text{Tr}(\Gamma)}} + \frac{n^2 K^2 \ln n}{c_5 w^2(A) \text{Tr}(\Gamma)}$ and $\eta_1 = \frac{c_5 \text{Tr}(\Gamma)}{K^2 A_{\text{max}}^2(\Gamma) n}$ completes the proof.

## F Generalized Linear Models: Restricted Strong Convexity

We consider the bounds on the regularization parameter and the RE condition for GLM loss functions and any norm.

### F.1 Noise for GLM models

For any GLM the first derivative of the log-partition function $\psi(u)$ w.r.t $u$ is equal to the mean.

$$\psi^\prime(\langle \theta, x \rangle) = \frac{\int y|x \exp\{y(\langle \theta, x \rangle)\} y}{\int y|x \exp\{y(\langle \theta, x \rangle)\}} = \int y|x \mathbb{P}(y|x; \theta) y = \mathbb{E}(y|x)$$

Now the maximum likelihood loss function the negative of the log of the probability distribution.

$$\mathcal{L}(\theta; Z^n_1) = -\langle \theta, \frac{1}{n} \sum_{i=1}^n y_i X_i \rangle + \frac{1}{n} \sum_{i=1}^n \psi(\langle \theta, X_i \rangle)$$

Therefore

$$\nabla \mathcal{L}(\theta^*; Z^n_1) = -\frac{1}{n} \sum_{i=1}^n y_i X_i + \frac{1}{n} \sum_{i=1}^n X_i \psi'(\langle \theta^*, X_i \rangle) = \frac{1}{n} X^T (\mathbb{E}(y|x) - y) = \frac{1}{n} X^T w$$

If the noise is assumed to be gaussian or sub-Gaussian the analysis in Section 3 gives the analysis for the regularization parameter and the analysis for recovery bounds follows analogously.

### F.2 RSC condition for GLMs

The RSC condition is as follows:

$$\delta \mathcal{L}(\theta^*, u; Z^n_1) := \mathcal{L}(\theta^* + u; Z^n_1) - \mathcal{L}(\theta^*; Z^n_1) - \langle \nabla \mathcal{L}(\theta^*; Z^n_1), u \rangle \geq \kappa \|u\|^2_2 \quad (219)$$
where $v$.

For the general formulation of GLM’s given earlier

$$\delta \mathcal{L}(\theta^*, u; Z^n_i) = -\langle \theta^* + u, \frac{1}{n} \sum_{i=1}^{n} y_i X_i \rangle + \frac{1}{n} \sum_{i=1}^{n} \psi(\langle \theta^* + u, X_i \rangle) + \langle \theta^*, \frac{1}{n} \sum_{i=1}^{n} y_i X_i \rangle - \frac{1}{n} \sum_{i=1}^{n} \psi(\langle \theta^*, X_i \rangle) - \langle -\frac{1}{n} \sum_{i=1}^{n} y_i X_i + \frac{1}{n} \sum_{i=1}^{n} X_i \psi'(\langle \theta^*, x_i \rangle), u \rangle$$

Simplifying the expression and applying mean value theorem twice we get the following

$$\delta \mathcal{L}(\theta^*, u; Z^n_i) = \frac{1}{n} \sum_{i=1}^{n} \psi''(\langle \theta^*, X_i \rangle + \gamma_i \langle u, X_i \rangle) \langle u, X_i \rangle^2$$  \hspace{1cm} (220)

where $\gamma_i \in [0, 1]$.

The RSC condition for GLMs then needs to consider lower bounds for

$$\delta \mathcal{L}(\theta^*, u; Z^n_i) = \frac{1}{n} \sum_{i=1}^{n} \psi''(\langle \theta^*, X_i \rangle + \gamma_i \langle u, X_i \rangle) \langle u, X_i \rangle^2$$  \hspace{1cm} (221)

where $\gamma_i \in [0, 1]$. The second derivative of the log-partition function is always positive. Since the RSC condition relies on a non-trivial lower bound for the above quantity, the analysis will suitably consider a compact set where $\ell = \ell_\psi(T) = \min_{|a| \leq 2T} \psi''(a)$ is bounded away from zero. The only assumption outside this compact set $\{a : |a| \leq 2T \}$ is that the second derivative is greater than 0. Further, we assume $\|\theta^*\|_2 \leq c_1$ for some constant $c_1$. With these assumptions

$$\delta \mathcal{L}(\theta^*, u; Z^n_i) \geq \frac{\ell}{n} \sum_{i=1}^{n} \langle X_i, u \rangle^2 \mathbb{I}[\langle X_i, \theta^* \rangle < T] \mathbb{I}[\|\langle X_i, u \rangle \| < T]$$  \hspace{1cm} (222)

We give a characterization of the RSC condition for independent isotropic sub-Gaussian design matrices $X \in \mathbb{R}^{n \times p}$. The analysis can be suitably generalized to the other design matrices considered in earlier sections by using the same techniques. We consider $u \in A \subseteq \mathbb{S}^{p-1}$ so that $\|u\|_2 = 1$. Further, we assume $\|\theta^*\|_2 \leq c_1$ for some constant $c_1$. Assuming $X$ has sub-Gaussian entries with $\|x_{ij}\|_{\psi_2} \leq k$, $\langle X_i, \theta^* \rangle$ and $\langle X_i, u \rangle$ are sub-Gaussian random variables with sub-Gaussian norm at most $Ck$. Let $\phi_1 = \phi_1(T; u) = P\{|\langle X_i, u \rangle | > T\} \leq e \cdot \exp(-c_2 T^2 / C^2 k^2)$, and $\phi_2 = \phi_2(T; \theta^*) = P\{|\langle X_i, \theta^* \rangle | > T\} \leq e \cdot \exp(-c_2 T^2 / C^2 k^2)$. The result we present is in terms of the constants $\ell = \ell_\psi(T)$, $\phi_1 = \phi_1(T; u)$ and $\phi_2 = \phi_2(T, \theta^*)$ for any suitably chosen $T$.

**Theorem 13** Let $X \in \mathbb{R}^{n \times p}$ be a design matrix with independent isotropic sub-Gaussian rows. Then, for any set $A \subseteq \mathbb{S}^{p-1}$, any $\alpha \in (0, 1)$, any $\tau > 0$, and any $n \geq \frac{2}{\alpha^2(1-\phi_1 - \phi_2)}(cuv(A) + c_2(1-\phi_1 - \phi_2)^2(1-\alpha)\tau^2)$ for suitable constants $c, c_2$, with probability at least $1 - 3 \exp(-\eta_1 \tau^2)$, we have

$$\inf_{u \in A} \sqrt{n} \delta \mathcal{L}(\theta^*; u, X) \geq \ell \sqrt{\pi} \left( \sqrt{n} - \eta_0 w(A) - \tau \right),$$  \hspace{1cm} (223)

where $\pi = (1-\alpha)(1-\phi_1 - \phi_2)$, $\ell = \ell_\psi(T) = \min_{|a| \leq 2T+\beta} \nabla^2 \psi(a)$, and constants $(\eta_0, \eta_1)$ depend on the sub-Gaussian norm $\|x_{ij}\|_{\psi_2} = k$.

**Proof:** As in earlier sections we work with vectors $v \in \mathcal{N}_r(A)$. We make the observation that, for $\forall u \in A$ and the corresponding $v \in \mathcal{N}_r(A)$ and some fixed constant $\beta$, let $M_u(X), M_v(X)$ denote sets such that
For any fixed $v \in \mathcal{N}(A)$, we require that the submatrix $M_v(X) \subseteq M_u(X)$. This observation helps when we extend the argument from $v \in \mathcal{N}(A)$ to $u \in A$, as to apply Lemma 12 we require that the submatrix $M_v(X) \subseteq M_u(X)$.

For $v \in \mathcal{N}(A)$, let $Z_i = \langle X_i, v \rangle$. Then $Z_i$ is sub-Gaussian with $\|Z_i\|_2 = 1$ and $\|Z_i\|_{\psi_2} = CK = K$. Now for any fixed $T$, let $\hat{Z}_i = \langle X_i, v \rangle (\langle X_i, v \rangle \leq T) \mathbb{P}(\langle X_i, \theta^* \rangle \leq T)$. Then, the probability distribution over $\hat{Z}_i$ can be written as

$$P(\hat{Z}_i = z) = \frac{P(\langle X_i, v \rangle = z)\mathbb{P}(\langle X_i, v \rangle \leq T)\mathbb{P}(\langle X_i, \theta^* \rangle \leq T)}{P(\langle X_i, v \rangle \leq T, \langle X_i, \theta^* \rangle \leq T)} \leq \frac{1}{1 - \phi_1 - \phi_2} P(\langle X_i, v \rangle = z).$$

(224)

As a result, $\|\hat{Z}_i\|_{\psi_2} \leq \frac{K}{1 - \phi_1 - \phi_2}$. Let $\hat{Z} \in \mathbb{R}^n$ be the vector whose elements are $\hat{Z}_i$, implying that some elements can be zero. Note that

$$\sqrt{n}\partial \mathcal{L}(\theta^*; u, X) \geq \ell \|\hat{Z}\|_2,$$

(225)

where $\ell = \ell(T)$. Further, if $\hat{Z}_m \in \mathbb{R}^m$, for some $m \leq n$, be the vector whose elements are of the form $\hat{Z}_i$ corresponding to $m$ independent samples $X_i$, we follow the same analysis style as for Theorem 10.

Let $\max(\delta, \delta^2) = t$, $\delta = \eta_0 \frac{K}{\sqrt{n}} + \frac{T}{\sqrt{n}}$ and given (195), following the same analysis as Section E.1 we have

$$P \left\{ \left( \frac{1}{m} \right) \|Z_m\|_2^2 - 1 \geq \frac{t}{2} \right\} \leq 2 \exp \left( -\frac{c_2(1 - \phi_1 - \phi_2)^4}{K^4} \delta^2 m \right).$$

(226)

In particular, for any $\alpha \in (0, 1)$, for $m \geq (1 - \alpha)(1 - \phi_1 - \phi_2)n$, we have

$$P \left\{ \left( \frac{1}{1 - \alpha}(1 - \phi_1 - \phi_2)n \right) \|Z_m\|_2^2 - 1 \geq \frac{t}{2} \right\} \leq 2 \exp \left( -\frac{c_2(1 - \phi_1 - \phi_2)^5}{K^4} \delta^2 (1 - \alpha)n \right).$$

(227)

For convenience, let $\pi = (1 - \alpha)(1 - \phi_1 - \phi_2)$. Now, based on the design matrix $X$ consists of $n$ samples $X_i$, consider the discrete random variable $M_{v, \theta^*}(X) = \{ \langle X_i, v \rangle \leq T, \langle X_i, \theta^* \rangle \leq T \}$. Clearly, $M_{v, \theta^*}(X) \subseteq \{0, 1, \ldots, n\}$ follows a binomial distribution with success probability greater than $(1 - \phi_1 - \phi_2)$ and $\mathbb{E}[M_{v, \theta^*}(X)] \geq (1 - \phi_1 - \phi_2)n$. Then, for any $v \in \mathcal{N}(A)$ and noting that $\|Z\|_2 = \|Z_m\|_2$

$$P \left\{ \left( \frac{1}{\pi n} \right) \|Z\|_2^2 - 1 \geq \frac{t}{2} \right\} \leq P \left\{ \left( \frac{1}{\pi n} \right) \|Z\|_2^2 - 1 \geq \frac{t}{2} \middle| M_{v, \theta^*}(X) \geq (1 - \alpha)(1 - \phi_1 - \phi_2)n \right\} + P \left\{ M_{v, \theta^*}(X) \geq (1 - \alpha)(1 - \phi_1 - \phi_2)n \right\} + P \left\{ M_{v, \theta^*}(X) \geq (1 - \alpha)(1 - \phi_1 - \phi_2)n \right\} + P \left\{ M_{v, \theta^*}(X) \geq (1 - \alpha)(1 - \phi_1 - \phi_2)n \right\}

(228)

By application of Chernoff bounds for binomial distributions

$$P \{ M_{v, \theta^*}(X) \geq (1 - \alpha)(1 - \phi_1 - \phi_2)n \} \leq \exp \left( -\frac{\alpha^2(1 - \phi_1 - \phi_2)}{2} n \right) \tag{228}$$

\[\text{With abuse of notation, we treat the distribution over } \hat{Z}_i \text{ as discrete for ease of notation. A similar argument applies for the true continuous distribution, but more notation is needed.}\]
Noting that we are looking for an upper bound

\[
P \left\{ \left| \frac{1}{\pi n} \| Z \|_2^2 - 1 \right| \geq \frac{t}{2} \right\} = P \left\{ \left| \frac{1}{\pi n} \| Z \|_2^2 - 1 \right| \geq \frac{t}{2} \right\} \leq 2 \exp \left( -\frac{\alpha^2(1-\phi_1 - \phi_2)}{2} n \right)
\]

\[
\leq 2 \exp \left( -\frac{\alpha^2(1-\phi_1 - \phi_2)}{2} n \right) + \exp \left( \frac{\alpha^2(1-\phi_1 - \phi_2)}{2} n \right)
\]

where \( \delta = \eta_0 \frac{w(A)}{\sqrt{n}} + \frac{\tau}{\sqrt{n}} \) and \( \eta_1 = \frac{\alpha^2(1-\phi_1 - \phi_2)}{2} \).

Taking union bound over all \( v \in \mathcal{N}_t(A) \), and using (74), i.e., the upper bound on the cardinality of the covering number in terms of the Gaussian width of \( A \), we have

\[
P \left\{ \max_{v \in \mathcal{N}} \left| \frac{1}{\pi n} \| Z \|_2^2 - 1 \right| \geq \frac{t}{2} \right\} \leq P \left\{ \left| \frac{1}{\pi n} \| Z \|_2^2 - 1 \right| \geq \frac{t}{2} \right\} \leq 2 \exp \left( -\eta_1 \left[ \eta_0^2 w^2(A) + \tau^2 \right] \right) + \exp \left( \frac{\alpha^2(1-\phi_1 - \phi_2)}{2} n \right)
\]

\[
= 2 \exp \left( -(\eta_1 \eta_0^2 - c)w^2(A) \right) \exp \left( -\eta_1 \tau^2 \right) + \exp \left( cw^2(A) - \frac{\alpha^2(1-\phi_1 - \phi_2)}{2} n \right)
\]

\[
\leq 2 \exp \left( -\eta_1 \tau^2 \right) + \exp \left( cw^2(A) - \frac{\alpha^2(1-\phi_1 - \phi_2)}{2} n \right)
\]

\[
\leq 3 \exp \left( -\eta_1 \tau^2 \right),
\]

assuming \( \eta_0 \geq \sqrt{\frac{c}{n}} \) and for \( n \geq \frac{1}{\eta_2} (cw^2(A) + \eta_1 \tau^2) \) where \( \eta_2 = \frac{\alpha^2(1-\phi_1 - \phi_2)}{2} \).

As a result, from Lemma [II] and (225) we have

\[
P \left( \inf_{u \in A} \sqrt{n} \delta \mathcal{L}(\theta^0; u, X) \right) \geq \ell \sqrt{\pi} \left( \sqrt{n} - \eta_0 w(A) - \tau \right) \geq 1 - 3 \exp \left( -\eta_1 \tau^2 \right) .
\]

References

[1] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with norm regularization. In Advances in Neural Information Processing Systems (NIPS), 2014.

[2] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. Inform. Inference, 3(3):224–294, 2014.

[3] A. Banerjee, S. Merugu, I Dhillon, and J Ghosh. Clustering with \{B\}regman Divergences. Journal of Machine Learning Research, 6:1705–1749, 2005.

[4] O Barndorff-Nielsen. Information and Exponential Families in Statistical Theory. John Wiley and Sons, 1978.
[6] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

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

[8] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[9] I. Brown. *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, 1986.

[10] P. Buhlmann and S. van de Geer. *Statistics for High Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, 2011.

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

[12] E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52:489–509, 2006.

[13] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.

[14] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[15] S. Chatterjee, S. Chen, and A. Banerjee. Generalized dantzig selector: Application to the k-support norm. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[16] R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1:290–330.

[17] M. A. T. Figueiredo and R. D. Nowak. Sparse estimation with strongly correlated variables using ordered weighted l1 regularization. *arXiv:1409.4005*, 2014.

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

[19] Y. Gordon. On milmans inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric Aspects of Functional Analysis*, volume 1317 of Lecture Notes in Mathematics, pages 84–106. Springer, 1988.

[20] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *arXiv:1312.3580*, 2013.

[21] G. Lecué and S. Mendelson. Sparse recovery under weak moment assumptions. *arXiv:1401.2188*, 2014.

[22] M. Ledoux. *The concentration of measure phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society.

[23] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 2013.

51
[24] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246—270, 2009.

[25] A. Montanari and J. A. Pereira. Which graphical models are difficult to learn? In *Neural Information Processing Systems (NIPS)*, pages 1303–1311, 2009.

[26] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for the analysis of regularized M-estimators. *Statistical Science*, 27(4):538–557, 2012.

[27] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011.

[28] R. I. Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. *arXiv:1312.2903*, 2013.

[29] S. Oymak, C. Thrampoulidis, and B. Hassibi. The Squared-Error of Generalized Lasso: A Precise Analysis. *arXiv:1311.0830*, 2013.

[30] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.

[31] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted Eigenvalue Properties for Correlated Gaussian Designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

[32] Z. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, 2013.

[33] M. Talagrand. *The Generic Chaining*. Springer, 2005.

[34] M. Talagrand. *Upper and Lower Bounds for Stochastic Processes*. Springer, 2014.

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

[36] J. A. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*. (To Appear), 2015.

[37] A.-S. Ustunel. *An introduction to analysis on Weiner Space*. Springer-Verlag, 1995.

[38] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing*, chapter 5, pages 210–268. Cambridge University Press, 2012.

[39] R. Vershynin. Estimation in high dimensions: A geometric perspective. *Submitted*, 2014.

[40] A. B. Vizcarra and F. G. Viens. Some applications of the malliavin calculus to sub-gaussian and non-sub-gaussian random fields. In *Seminar on Stochastic Analysis, Random Fields and Applications, Progress in Probability*, volume 59, pages 363–396. Birkhauser, 2008.

[41] M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using \( \ell_1 \)-constrained quadratic programming(Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

[42] M J Wainwright and M I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.
[43] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, November 2006.

[44] S. Zhou. Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562, 2014.