A New Perspective on Debiasing Linear Regressions

Yufei Yi Matey Neykov
Department of Statistics & Data Science
Carnegie Mellon University

Abstract

In this paper, we propose an abstract procedure for debiasing constrained or regularized potentially high-dimensional linear models. It is elementary to show that the proposed procedure can produce $\frac{1}{\sqrt{n}}$-confidence intervals for individual coordinates (or even bounded contrasts) in models with unknown covariance, provided that the covariance has bounded spectrum. While the proof of the statistical guarantees of our procedure is simple, its implementation requires more care due to the complexity of the optimization programs we need to solve. We spend the bulk of this paper giving examples in which the proposed algorithm can be implemented in practice. One fairly general class of instances which are amenable to applications of our procedure include convex constrained least squares. We are able to translate the procedure to an abstract algorithm over this class of models, and we give concrete examples where efficient polynomial time methods for debiasing exist. Those include the constrained version of LASSO, regression under monotone constraints, regression with positive monotone constraints and non-negative least squares. In addition, we show that our abstract procedure can be applied to efficiently debias SLOPE and square-root SLOPE, among other popular regularized procedures under certain assumptions. We provide thorough simulation results in support of our theoretical findings.

1 Introduction

Linear regression is a pillar in statistics. Due to its simplicity and interpretability, it is possibly the most widely known and used statistical modeling and estimation technique both within and outside the field of statistics. The amount of literature on linear regression is vast, and ever-growing. In addition, with the big data boom, high-dimensional regression has steadily become an indispensable tool in practice, and has been in the focus of statisticians and practitioners for the past number of years. By far the most widely used estimator for the linear model is the ordinary least squares estimator (OLS). Unfortunately, OLS does not allow the practitioner to build in prior knowledge on the coefficients of interest. However, prior knowledge, e.g. sparsity, can be crucial for performing reasonable estimation especially in modern large datasets like genome-wide association studies where the number of samples can be smaller than the number of covariates. Incorporating prior knowledge (in a frequentist sense) may come at a price — it is not immediately obvious how to perform inference since the resulting estimator might not have a closed form, in contrast to the OLS, and in addition the estimated coefficients are likely biased. In this paper we tackle questions of this flavor: we suggest an abstract procedure which can perform inference
for certain estimators in linear models which are “non-OLS”, such as some convex constraint least squares estimators and some regularized estimators such as the Sorted L-One Penalized Estimator (SLOPE) and square-root SLOPE.

As we mentioned, parameter estimation in high-dimensional statistical models typically requires solving a regularized (or constrained) optimization problem. Regularization is necessitated in order to help fight the curse of dimensionality. Since the resulting estimators are non-linear, it is difficult to directly characterize their limiting distributions. A notable exception where asymptotic results have been obtained for regularized estimators, is the LASSO estimator Tibshirani (1996) (and more generally the so called Bridge estimators) see Knight and Fu (2000); however, importantly, these asymptotic results are valid in the fixed dimensional setting and not in the high-dimensional setting, and moreover, are difficult to apply to draw inference or construct confidence intervals since the limiting distribution is not pivotal. This underscores that performing statistical inference is non-trivial in the high-dimensional setting. In a low-dimensional setting (where the need for regularization is less apparent), one can use large sample theory on an unregularized estimator (such as the OLS) to get an asymptotic result (Van der Vaart, 2000). Even in low-dimensional settings however, if one chooses to use a constrained likelihood or more generally a constrained $M$-estimator, e.g., the asymptotic distribution may be highly non-trivial Chernoff (1954); Self and Liang (1987); Geyer et al. (1994). A high-dimensional setting only exacerbates this issue, since as we mentioned, it necessitates the regularization.

In high-dimensional models, one is often interested in one of three directions: oracle inequalities (Bunea et al., 2007; Van de Geer et al., 2008; Bickel et al., 2009), variable selection (Meinshausen et al., 2006; Zhao and Yu, 2006; Fan and Lv, 2008), and statistical inference (Van de Geer et al., 2014; Neykov et al., 2018; Feng and Ning, 2019). The latter reference list is far from complete and we refer the reader to the excellent books by Bühlmann and Van De Geer (2011) and Wainwright (2019) for a full introduction to high-dimensional statistics. Since in this paper we focus on the inference direction, below we review in depth only articles which are related to this direction.

At first, the efforts of statisticians were naturally devoted to enable performing inference in the high-dimensional linear model, as it has ubiquitous applications in a variety of fields such as statistical genetics, bioinformatics, econometrics, finance, among many others. For instance, high-dimensional problems have been recently recognized in signal processing (Lustig et al., 2008), genetics (Peng et al., 2010) and collaborative filtering (Koren et al., 2009). Early approaches of high-dimensional statistical inference were based on variable selection consistency (Wasserman and Roeder, 2009; Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013), which only works for sparse signal vectors. Specifically, the estimator is computed on the oracle set only, so the statistical inference is reduced to a low-dimensional setting. A limitation of this approach is that the variable selection consistency requires the magnitude of all non-zero coefficients to be greater than a threshold (Wainwright, 2009; Zhang et al., 2010), which may be unrealistic in many applications. The above reasoning motivated various approaches for deriving tractable and pivotal distributions for high-dimensional models which can be used to construct confidence intervals and draw inferences for individual coefficients. While there are approaches which consider a conditional hypothesis test of the coefficients from a LASSO (Lockhart et al., 2014; Lee and Taylor, 2014; Lee et al., 2016, among others), in this paper we follow a line of work initiated by Zhang and Zhang (2014); Van de Geer et al. (2014); Javanmard and Montanari (2014); Belloni et al. (2014, 2015) where it was proposed how to correct the LASSO estimate (often called debiasing) in order to achieve asymptotic normality on individual coefficients. These works spurred a lot of follow-ups including
(Ning et al., 2017; Jankova et al., 2015; Neykov et al., 2018; Javanmard et al., 2018; Jankova et al., 2018, among others). Until recently, the majority of debiasing methods focused exclusively on \( \ell_1 \) penalized (generalized) linear models. Of note there is a recent exception which can handle more general penalties than the \( \ell_1 \) (Bellec and Zhang, 2019b). A notable limitation of this work however is that this debiasing scheme works only in the asymptotic regime \( p/n \to \gamma \) for some constant \( \gamma \), and furthermore it requires the knowledge of the covariance matrix \( \Sigma \) of the predictors. Further, some other more recent works integrate a degrees-of-freedom adjustment to the debiasing procedure (Bellec and Zhang, 2019a; Celentano et al., 2020). This is something that we do not exploit in the current work, although we think there may be some promising connection between this idea and our algorithm. Finally we would like mention the work of Bradic et al. (2018); Zhu and Bradic (2018) which studies how one can perform inference in linear models where sparsity may be absent. This is related to our work in the sense that some models which we consider, like the monotone regression, are non-sparse. However, there is a big difference in the settings in that the algorithms given in Bradic et al. (2018); Zhu and Bradic (2018) work without having to respect the prior knowledge that the coefficients are monotone, e.g.

In this paper, we propose an abstract debiasing procedure for some regularized or constrained linear models. We illustrate that our procedure is applicable to convex constrained least squares with unknown covariance, in cases when the convex constraint set \( K \) has a simple geometric structure. In addition, we demonstrate that our approach can successfully debias SLOPE and square-root SLOPE under the assumption that we have a known upper bound on the sparsity of the signal. Our debiasing approach relies on solving a cascade of two optimization problems. The first optimization restricts the initial coefficient estimator to have a small tangent cone, which is used to facilitate the second optimization program. The second optimization is inspired by the work of Javanmard and Montanari (2014). Specifically, the constraint set of this convex program is designed in such a way so that any feasible solution can be used for debiasing. Next the objective function is selected to minimize the variance of the limiting distribution of debiased estimator. Our second optimization uses a newly-designed constraint set in comparison with the LASSO debiasing approach from (Javanmard and Montanari, 2014, Algorithm 1). In the case of convex constrained least squares for example, our debiasing constraint is designed to respect the geometry of the constraint set \( K \), which turns out to be the key for generalizing the debiasing from \( \ell_1 \)-regularized problems to general constraint problems.

1.1 Notation and Definitions

Here we introduce some notation and concepts which will be used throughout the paper. Given a set \( T \subset \mathbb{R}^p \), define its Gaussian complexity as

\[
w(T) = \mathbb{E} \sup_{x \in T} \langle g, x \rangle, \quad \text{where } g \sim \mathcal{N}(0, I_p).
\]

\( w(T) \) is the expectation of maximum magnitude of the canonical Gaussian process on \( T \). The Gaussian complexity is a basic geometric property of \( T \). It measures the size of \( T \) and is related to the metric entropy of \( T \) (Vershynin, 2018, Theorem 8.1.13). In addition to \( w(T) \), throughout the paper we denote with \( \overline{w}(T) \) any known (and ideally easily computable and as small as possible) upper bound of \( w(T) \), i.e., \( \overline{w}(T) \) satisfies:

\[
w(T) \leq \overline{w}(T) \tag{1.1}\]
Next we formalize the concept of a tangent cone which is frequently used in optimization. The tangent cone of a convex set \( K \subset \mathbb{R}^p \) at \( x \in K \) consists of all the possible directions from which a sequence in \( K \) can converge to \( x \). It is defined as

\[
T_K(x) = \{ t(v - x) : t \geq 0, v \in K \}.
\]

The projection of a vector \( v \in \mathbb{R}^p \) onto a convex set \( K \subset \mathbb{R}^p \) is defined as

\[
\Pi_K(v) = \arg\min_{x \in K} \|v - x\|,
\]

where here and throughout we will use \( \|\cdot\| \) as a shorthand for the Euclidean norm \( \|\cdot\|_2 \). Furthermore let \( \|\cdot\|_{op} \) denote the operator norm of a matrix. In addition we will also use \( \wedge \) and \( \vee \) as a shorthand for min and max of two numbers respectively, and \([n] = \{1, \ldots, n\}\) for an integer \( n \in \mathbb{N} \). We also make use of standard asymptotic notation: we write \( X_n = o_p(1) \) if \( \mathbb{P}(|X_n| > \epsilon) \to 0 \) for all \( \epsilon > 0 \), and \( X_n = O_p(1) \) if for any \( \epsilon > 0 \) there exists an \( M > 0 \) and a finite \( N > 0 \) such that \( \mathbb{P}(|X_n| > M) < \epsilon \) for all \( n > N \). We write \( X_n = o_p(a_n) \) if \( X_n/a_n = o_p(1) \), and \( X_n = O_p(a_n) \) if \( X_n/a_n = O_p(1) \) for some non-zero sequence \( \{a_n\} \). Furthermore, given two non-negative sequences \( \{a_n\}, \{b_n\} \) we write \( a_n = O(b_n) \) (or \( a_n \lesssim b_n \)) if there exists a constant \( C < \infty \) such that \( a_n \leq Cb_n \), \( a_n = o(b_n) \) if \( a_n/b_n \to 0 \), and \( a_n \asymp b_n \) if there exists positive constants \( c \) and \( C \) such that \( c < a_n/b_n < C \).

1.2 Problem Formulation

Suppose that we are given \( n \) i.i.d. observations from a linear model

\[
Y_i = X_i^\top \beta^* + \epsilon_i, i \in [n],
\]

where the predictors \( X_i \) are also considered i.i.d. and random. For simplicity we assume that every observation \( X_i \) is zero-mean (i.e. the covariates are centered). This can always be achieved at the price of splitting the data evenly and subtracting the \( Y_i \) and \( X_i \) values from the first half from those values of the second half (this not only ensures that \( X_i \) will be zero-mean but also preserves other subsequent assumptions that we make on the data). In addition we will require that \( X_i \) is either a Gaussian or a bounded random variable with covariance \( \Sigma \). Furthermore, for the most part of the manuscript we will assume that \( \epsilon_i \sim N(0, \sigma^2) \) in order to simplify the presentation. In Section 6 we elaborate on a slight modification of our procedure, inspired by Javanmard and Montanari (2014), that can handle general sub-Gaussian noise.

Suppose now that instead of fitting OLS to (1.2), a practitioner fits a regularized or constrained least squares estimator. An example where such a situation may arise is when the practitioner has prior knowledge that \( \beta^* \in K \) for some fixed and known convex set \( K \). In such a setting the practitioner may opt for outputting the following natural estimate of \( \beta^* \):

\[
\hat{\beta} = \arg\min_{\beta \in K} n^{-1} \sum_{i \in [n]} (Y_i - X_i^\top \beta)^2.
\]

In addition, especially in settings when \( p \gg n \) and an assumption on the sparsity of \( \beta^* \) is appropriate, the practitioner may opt for running a regularized procedure such as LASSO (Tibshirani, 1996), SLOPE (Bogdan et al., 2015) or square-root SLOPE (Stucky and Van De Geer, 2017). Unlike the OLS, constraints or regularizations incur bias on \( \hat{\beta} \), and make the limiting distribution of \( \hat{\beta} \) complicated. Thus performing statistical inference on \( \hat{\beta} \) becomes non-straightforward.
The goal of the present paper is to develop what became known as debiasing techniques for \( \hat{\beta} \) in such scenarios. In particular we would like to construct confidence intervals for any bounded contrast of \( \beta^* \) (i.e. \( \gamma^\top \beta^* \) with \( \|\gamma\| < B < \infty \)) — using a non-OLS pilot estimator \( \hat{\beta} \) of \( \beta^* \) in (1.2) — in a high-dimensional setting. It is worthy to mention that the majority of previous works on debiasing focus exclusively on debiasing \( \ell_1 \)-penalized regression. There are some exceptions such as Bellec and Zhang (2019b), but their setting is substantially different from the present work.

The algorithm proposed in this paper is capable of debiasing any estimator \( \hat{\beta} \) which can be used to produce the following quantities:

- An estimator \( v \) of a vector sufficiently close to \( \beta^* \) (or ideally \( \beta^* \) itself) in the \( \ell_2 \) sense.
- A convex set \( K \) such that \( v, \beta^* \in K \) (here \( K \) may be given or may be constructed from \( \hat{\beta} \)).
- \( v \) is a boundary point in \( K \) such that the tangent cone of \( K \) at \( v \) is sufficiently small.

We will make use of sample splitting to produce \( \hat{\beta} \) on one half of the sample, and estimate a projection direction used in the debiasing on the other half. For more detailed information on our abstract procedure refer to Section 2.

Finally we mention that our debiasing procedure does not require prior knowledge of the inverse population covariance matrix — \( \Sigma^{-1} \) — which is known to make inference easier (Javanmard et al., 2018; Bellec and Zhang, 2019b).

### 1.3 Paper Organization

The paper is structured as follows. Section 2 describes our abstract debiasing procedure and shows how the program from the second step can be solved with subgradient descent. Section 3 proves the main theorem of the paper and provides a confidence interval construction. Section 4 is dedicated to convex constrained least squares, where we formally describe how one can solve step 1 of our abstract debiasing procedure in such a setting. Section 5 discusses applications to SLOPE and square-root SLOPE. Section 6 contains an extension to non-Gaussian noise. Section 7 illustrates our results with some numerical studies and finally in Section 8 we give a brief discussion. All technical proofs are deferred to the supplement.

### 2 The Debiasing Algorithm

In this section we propose an optimization-based Algorithm 1 as a general procedure to debias an individual coordinate, as well as any contrast of \( \beta^* \) using a non-OLS estimator \( \hat{\beta} \). Then in Section 2.2 and 2.3 we provide details for how to solve the optimization problem in step 2 of the proposed Algorithm 1.

#### 2.1 The Debiasing Algorithm

For simplicity of the presentation, and without loss of generality we will assume that we are given \( 2n \) samples from model (1.2). If the actual number of samples is odd we can simply drop one sample. We randomly split the data set \( (X, Y) \) where \( X = (X_1, \ldots, X_{2n})^\top \), \( Y = (Y_1, \ldots, Y_{2n})^\top \) into two equally-sized partitions \( (\tilde{X}, \tilde{Y}) \) and \( (\hat{X}, \hat{Y}) \). The first half \( (\tilde{X}, \tilde{Y}) \) is used to obtain an estimator \( \hat{\beta} \) of the true coefficient \( \beta^* \), and then is used to obtain \( v \) and \( K \). The second half \( (\hat{X}, \hat{Y}) \) is used to construct the debiased \( \hat{\beta}_d \) based on \( v \) and \( K \).
Step 1 of Algorithm 1 uses the first half of the data to construct a vector $v$ which is close to $\beta^*$ in $\ell_2$-distance, and a convex set $K$ which has a small tangent cone at $v$. In all of our examples to follow, such a construction uses a pilot estimator $\hat{\beta}$ which can be a constrained or regularized estimator. We therefore view our procedure as a procedure for debiasing the pilot vector $\hat{\beta}$, but in principle one may bypass estimating $\beta$ and may use the first half of the data to directly find $v$ and $K$ obeying the desired properties.

Next we solve an optimization program (see step 2 of Algorithm 1) to get an auxiliary vector $\hat{\eta}$ which is used in the final debiasing formula as a proxy to the $j$-th row of $\Sigma^{-1}$. In fact, as implied by Theorem 3.1, any feasible point of the optimization program in step 2 would successfully produce an asymptotically normal debiased estimator. In other words, the limiting distribution of $\sqrt{n}(\hat{\beta}_d^{(j)} - \beta^{*(j)})$ would be a zero-mean Gaussian random variable, but its variance might be large. To achieve a small variance for the limiting distribution, we pick the objective function in the optimization of step 2 to minimize such a variance, which is inspired by (Javanmard and Montanari, 2014, Algorithm 1). The following Algorithm 1 summarizes our debiasing procedure.

**Algorithm 1** Debiasing the $j$th Coordinate of A Non-Ordinary Least Squares Estimator

**Input:** Two equal size partitions of the data $(\bar{X}, \bar{Y})$ and $(\tilde{X}, \tilde{Y})$.

**Initialize:** Empirical Gram matrix of the second partition $\tilde{\Sigma} = \frac{1}{n} \tilde{X}^\top \tilde{X}$.

1. Using the first data split find a convex set $K$ and a vector $v$, such that: $v, \beta^* \in K$ with high probability, and $\mathbb{P}(T_K(v) \cap S^{p-1})\|v - \beta^*\| = o_p(1)$.

2. The debiased $j$th coefficient $\hat{\beta}_d^{(j)} \leftarrow e^{(j)\top}v + n^{-1}\hat{\eta}^\top \tilde{X}^\top (\tilde{Y} - \tilde{X}v)$, where $\hat{\eta}$ is computed by

\[
\hat{\eta} \leftarrow \arg\min_{\eta} \|\tilde{\Sigma}^{\frac{1}{2}}\eta\| \quad \text{subject to} \quad \sup_{u \in T_K(v) \cap S^{p-1}} |(\eta^\top \tilde{\Sigma} - e^{(j)\top})u| \leq \frac{\mathbb{P}(T_K(v) \cap S^{p-1})}{\sqrt{n}},
\]

for some sufficiently large tuning parameter $\rho > 0$.

**Remark 2.1.** Several remarks regarding Algorithm 1 are in order. First we comment on step 1. One may wonder how to construct a set $K$ and vector $v$ with the desired properties, and if that is even possible. While it is hard to answer this without having a concrete example at hand, we will give a couple of comments. The set $K$ may be naturally given to the practitioner — for example it may be the constraint set if the practitioner is solving convex constrained least squares. On the other hand, a set $K$ could be constructed via the vector $\hat{\beta}$. If $\hat{\beta}$ for instance is known to satisfy $\|\hat{\beta} - \beta^*\| \leq b(n,p,\beta^*)$ for some explicitly quantifiable upper bound $b(n,p,\beta^*)$ one may start the construction of $K$ based on the Euclidean ball around $\hat{\beta}$ with radius $b(n,p,\beta^*)$ (for more details on approach this we refer to Section 5 where we build a convex set $K$ for the SLOPE and square-root SLOPE estimators). The vector $v$ on the other hand should be selected to respect the geometry of $K$ and will likely have to possess additional properties (e.g. sparsity or other adequate restrictions which make the tangent cone at it small). We provide a detailed process of finding $v$ for each type of estimator $\hat{\beta}$ in our examples; see Section 4 and Section 5.

We now comment on the condition $\mathbb{P}(T_K(v) \cap S^{p-1})\|v - \beta^*\| = o_p(1)$ required in step 1.
Intuitively, we need \( \|v - \beta^*\| \) to be small because in the final step the debiased estimator \( \hat{\beta}_d \) is constructed from \( v \). A small upper bound on the Gaussian complexity of the tangent cone \( \overline{w}(T_K(v) \cap \mathbb{S}^{p-1}) \) is needed to guarantee fast convergence rate of the debiased estimator \( \hat{\beta}_d \), and fast computation of the optimization in step 2.

Finally we comment on step 2. Step 2 of our abstract procedure is reminiscent of previous ideas on debiasing which attempt to estimate the inverse covariance (aka precision) matrix along a direction of interest. We stress on the fact that our proposal is distinct from previous works however, and even in the “classical” example of LASSO will produce a distinct projection direction \( \hat{\eta} \). In addition, we mention that if one is interested in performing inference on general bounded contrasts of \( \beta^* \), i.e., \( \gamma^T \beta^* \) for some \( \|\gamma\| \leq B \) with a finite \( B \), step 2 can be readily modified by changing \( \hat{\eta} \) to

\[
\hat{\eta} \leftarrow \arg\min_{\eta} \|\hat{\Sigma}^{1/2} \eta\| \text{ subject to } \sup_{u \in T_K(v) \cap \mathbb{S}^{p-1}} |(\eta^T \hat{\Sigma} - \gamma^T) u| \leq \frac{\rho \overline{w}(T_K(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}}.
\]

For simplicity of presentation we stick to our formulation with \( e^{(j)} \) but all of our proofs and results can be easily modified to the more general setting described above by changing \( e^{(j)} \) to \( \gamma \).

In the next two subsections, we address two questions regarding the optimization (2.1) of step 2 of Algorithm 1. The first question is whether the constraint in (2.1) is empty. In Section 2.2 we will show that (2.1) is guaranteed to have a feasible point with high probability, and furthermore the interior of such a constraint is not empty if \( \rho \) is sufficiently large.

In addition, the above optimization (2.1) can be solved by subgradient descent. An explicit formula of the subgradient is complicated by the unconventional constraint, which makes the program in step 2 a semi-infinite program. See Hettich and Kortanek (1993) for details about semi-infinite programming. Section 2.3 gives out the explicit formula of the subgradient, and proves the convergence of such a subgradient descent method.

### 2.2 Studying the Constraint Set of Step 2

We begin by showing that \( \eta = \Sigma^{-1} e^{(j)} \) is a feasible point of the optimization (2.1). In fact, the right hand side of the constraint in (2.1) — \( \frac{\rho \overline{w}(T_K(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}} \) — is inspired by analyzing the magnitude of \( \sup_{u \in T_K(v) \cap \mathbb{S}^{p-1}} \left| |(\eta^T \hat{\Sigma} - e^{(j)T}) u| \right| \) when evaluated at \( \eta = \Sigma^{-1} e^{(j)} \). The intuition is that \( \hat{\eta} \) is a proxy of \( \Sigma^{-1} e^{(j)} \). This idea is of course standard and central in all previous debiasing works, but the challenge in our setting is to analyze the empirical process \( \sup_{u \in T_K(v) \cap \mathbb{S}^{p-1}} \left| (\eta^T \hat{\Sigma} - e^{(j)T}) u \right| \) at \( \eta = \Sigma^{-1} e^{(j)} \).

**Theorem 2.2.** Suppose that \( X = (X_1, \ldots, X_n)^T \) where every observation \( X_i \) is a zero-mean bounded or a zero-mean Gaussian random variable with covariance matrix \( \Sigma \), and the eigenvalues of \( \Sigma \) are bounded from above and below. Let \( \hat{\Sigma} = \frac{1}{n} X^T X \) be the empirical Gram matrix. Suppose that the upper bound \( \overline{w}(T_K(v) \cap \mathbb{S}^{p-1}) \) is chosen so that \( \overline{w}(T_K(v) \cap \mathbb{S}^{p-1}) \to \infty \) as \( n \to \infty \), and \( \overline{w}(T_K(v) \cap \mathbb{S}^{p-1}) = O(\sqrt{n}) \). Then for \( \eta = \Sigma^{-1} e^{(j)} \), with probability converging to one we have

\[
\sup_{u \in T_K(v) \cap \mathbb{S}^{p-1}} \left| |(\eta^T \hat{\Sigma} - e^{(j)T}) u| \right| \leq \frac{\overline{w}(T_K(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}}.
\]
Remark 2.3. In Theorem 2.2, the condition on bounded $X_i$ can be slightly relaxed to the following condition — $Z^T := \eta^T X_i$ is a sub-Gaussian random vector. The latter means that there exists a constant $C \in \mathbb{R}_+$ such that for any unit vector $w \in S^{p-1}$ and any $\lambda \in \mathbb{R}$, $\mathbb{E}\exp(\lambda (Z - \mathbb{E}Z)^T w) \leq \exp(\lambda^2 C)$. This modification requires a different proof which relies on a result in Mendelson (2010) and we do not give the proof here.

Theorem 2.2 requires that $\mathbb{w}(T_K(v) \cap S^{p-1}) \to \infty$. Since $v$ is random, it is convenient to assume this holds for all $v$. If one knows an upper bound on $u(T_K(v) \cap S^{p-1}) \leq u(T_K(v) \cap S^{p-1})$ for all vectors $v$, obtaining a diverging upper bound is simple: just take $\mathbb{w}(T_K(v) \cap S^{p-1}) = u(T_K(v) \cap S^{p-1}) \vee a_n$ for any slowly diverging sequence $a_n$. For future reference we will always assume that $\mathbb{w}(T_K(v) \cap S^{p-1})$ is constructed in such a way, and we do not explicitly mention the term “$a_n$”.

Note that in the result of Theorem 2.2, $T_K(v) \cap S^{p-1}$ can be substituted by a general compact set in $\mathbb{R}^p$ since the proof of Theorem 2.2 does not rely on the the fact that $T_K(v)$ is a convex cone. Here we stated the theorem with $T_K(v) \cap S^{p-1}$ because this is the only set of interest for us. Also, the result of Theorem 2.2 still holds if $e^{(j)}$ is replaced by any other unit norm vector, which supports the generalization of Algorithm 1 to debias a linear combination of coordinates. See also Remark 3.2.

The following Corollary proves that the constraint of (2.1) has a non-empty interior. It is a sufficient condition for the convergence of the subgradient descent in the next section.

Corollary 2.4 (Non-empty Interior of the Constraint). Under the same assumptions of Theorem 2.2 the set

$$Q = \left\{ \eta : \sup_{u \in T_K(v) \cap S^{p-1}} \left| (\eta^T \tilde{\Sigma} - e^{(j)T})u \right| \leq \frac{\rho \mathbb{w}(T_K(v) \cap S^{p-1})}{\sqrt{n}} \right\},$$

has a non-empty interior with high probability for sufficiently large $\rho$.

2.3 Solving the Optimization Problem (2.1) by Subgradient Descent

We will now explain how to solve the optimization program (2.1) by subgradient descent for constrained optimization. We implicitly assume in this section that the projection $\Pi_{T_K(v)}$ can be computed in a reasonable time. This may not always hold in practice due to the fact that both the set $K$ and estimator $v$ are random variables and depend on the first sample split. However we note that in all of our examples to be considered (see Sections 4 & 5) this projection is indeed feasible and can be computed fast. In addition finding a projection on a convex set is always a convex optimization problem, which can be solved in principle. Define

$$\psi(\eta) = \sup_{u \in T_K(v) \cap S^{p-1}} \left| (\eta^T \tilde{\Sigma} - e^{(j)T})u \right| - \frac{\rho \mathbb{w}(T_K(v) \cap S^{p-1})}{\sqrt{n}}.$$ (2.2)

The constraint in (2.1) can be written as $Q = \{\eta : \psi(\eta) \leq 0\}$. According to (Boyd et al., 2003, Section 7), the subgradient descent moves towards the optima by generating a sequence $\{\eta_n\}$ as

$$\eta_{n+1} = \eta_n - h_n \mathbf{g}_n,$$ (2.3)
where $h_n$ is the step size, and $g_n$ is the gradient of the objective function $f(\eta) = \| \tilde{\Sigma}^{\frac{1}{2}} \eta \|$ if $\eta_n \in Q$; otherwise is a subgradient of the constraint function $\psi(\eta)$ if $\eta_n \notin Q$. Put

$$
\phi_0(\eta) = \frac{\Pi_{T_K(v)}(\tilde{\Sigma} \eta - e^{(j)})}{\| \Pi_{T_K(v)}(\tilde{\Sigma} \eta - e^{(j)}) \|}, \quad \phi_1(\eta) = \frac{\Pi_{-T_K(v)}(\tilde{\Sigma} \eta - e^{(j)})}{\| \Pi_{-T_K(v)}(\tilde{\Sigma} \eta - e^{(j)}) \|}.
$$

Lemma 2.5 below, shows that the explicit form of $g_n$ is given by:

$$
g_n = \begin{cases} 
\tilde{\Sigma} \eta_n / \| \tilde{\Sigma}^{\frac{1}{2}} \eta_n \| & , \text{if } \eta_n \in Q \\
\tilde{\Sigma} \phi_1(\eta_n) / \| \tilde{\Sigma}^{\frac{1}{2}} \eta_n \| & , \text{if } \eta_n \notin Q.
\end{cases}
$$

It is clear that the first expression in (2.4) for $\eta_n \in Q$ is the gradient of the objective function $f(\eta) = \| \tilde{\Sigma}^{\frac{1}{2}} \eta \|$ at $\eta_n$ when $\eta_n \neq 0$. If $\eta_n$ turns out to be 0, $g_n$ can be taken as $\tilde{\Sigma}^{1/2} w$ for any unit vector $w$. However, if $\eta_n = 0$ is a feasible point, it is necessarily an optimal value so that the algorithm should terminate. In Lemma 2.5 we show that the second expression in (2.4) is a subgradient of $\psi(\eta)$ at $\eta_n$ when $\eta_n \notin Q$.

**Lemma 2.5.** For $\eta_n \notin Q$, the expression of $g_n$ at (2.4) is a subgradient of $\psi(\eta)$ at $\eta_n$.

We observe that if one can compute $\Pi_{T_K(v)}$, one can clearly compute

$$
\Pi_{-T_K(v)}(x) = - \arg\min_{w \in T_K(v)} \| w - (-x) \| = - \Pi_{T_K(v)}(-x).
$$

We provide Algorithm 2 as a summary of solving (2.1), assuming $\Pi_{T_K(v)}$ is computable in a reasonable time. In Section 4 and Section 5 we will see that such a projection $\Pi_{T_K(v)}$ can be obtained efficiently for some specific convex cones with a simple structure.

**Algorithm 2** Solve the Optimization (2.1) in Step 2 of Algorithm 1

**Input:** The convex set $K$, the vector $v$ from step 2, empirical Gram matrix of the second partition $\tilde{\Sigma} = \frac{1}{n} \tilde{X}^{\top} \tilde{X}$.

**Initialize:** $\eta_1$

Run for sufficiently long time:

1. Compute $P_+ \leftarrow \Pi_{T_K(v)}(\tilde{\Sigma} \eta_n - e^{(j)})$, $P_- \leftarrow \Pi_{-T_K(v)}(\tilde{\Sigma} \eta_n - e^{(j)})$.
2. if $\max\{\|P_+\|, \|P_-\|\} \leq \frac{\rho_m(T_K(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}}$
   1. if $\| \tilde{\Sigma}^{\frac{1}{2}} \eta_n \| \leq \| \tilde{\Sigma}^{\frac{1}{2}} \eta_{out} \|$: $\eta_{out} \leftarrow \eta_n$
   2. $\eta_{n+1} \leftarrow \eta_n - h_n \tilde{\Sigma} \eta_n / \| \tilde{\Sigma}^{\frac{1}{2}} \eta_n \|$.
3. else:
   1. $\phi_0(\eta_n) \leftarrow P_+ / \|P_+\|
   2. $\phi_1(\eta_n) \leftarrow P_- / \|P_-\|.
   3. $\eta_{n+1} \leftarrow \eta_n - h_n \tilde{\Sigma} \phi_1(\eta_n - \phi_0(\eta_n)) \phi_1(\eta_{n-1}) < 0) \eta_n
\end{align*}

$\hat{\eta} \leftarrow \eta_{out}$.

We note that the condition

$$
\max\{\|P_+\|, \|P_-\|\} \leq \frac{\rho_m(T_K(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}},
$$

on page 9.
used in Algorithm 2 is equivalent to checking feasibility, i.e., checking
\[
\psi(\eta_n) \leq \frac{\rho \mu (T_K(v) \cap S^{p-1})}{\sqrt{n}},
\]
since \(\langle \Pi_{T_K(v)}(\Sigma \eta_n - e^{(j)}), \Sigma \eta_n - e^{(j)} \rangle = \|\Pi_{T_K(v)}(\Sigma \eta_n - e^{(j)})\|^2\) as can be seen from Lemma B.7 in the supplementary material. Picking an adequate \(\rho\) is not hard in practice: one can start with a small constant (for example \(\rho = 1\)). If a feasible point is not found within a reasonable number of iterations, this possibly implies that the current \(\rho\) is too small, so one can enlarge \(\rho\) by setting \(\rho = 2\rho\) and so on.

Let \(\eta^* = \text{argmin}_{\eta \in Q} \|\Sigma^{\frac{1}{2}} \eta\|\) be the constrained minimas of (2.1). It is proved in Lemma 2.6 that there exists a subsequence of \(\{\eta_n\}\) in (2.3) converging to \(\eta^*\), and it takes \(n = O(1/\epsilon^2)\) iterations to get an \(\epsilon\)-suboptimal solution, i.e. \(\|\Sigma^{\frac{1}{2}} \eta_n\| - \|\Sigma^{\frac{1}{2}} \eta^*\| \leq \epsilon\). Therefore the subgradient descent is an appropriate method for solving program (2.1). As we mentioned earlier the constraint in the optimization program (2.1) is unconventional since the sup can be regarded as infinite number of constraints. Such programs are called semi-infinite programs. The proof of Lemma 2.6 is inspired by (Boyd et al., 2003, Section 7) which is suitable for unconventional constraints. For completeness we also mention that Polyak (1967) was the first to prove the convergence of subgradient descent with rather general constraints.

**Lemma 2.6 (Convergence of subgradient descent).** For any bounded starting point \(\eta_1\), one can construct a sequence \(\{\eta_n\}\) by (2.3), (2.4). As detailed in Algorithm 2 at every step of the iteration, we record the best candidate found so far as
\[
\eta^{\text{best}}_n = \text{argmin}_{\eta_i} \{\|\Sigma^{\frac{1}{2}} \eta\| \mid \eta_i \in Q, i \in [n]\},
\]
Let \(\eta^*\) achieve the minima of (2.1) and \(h_n\) be the step size of the subgradient descent. Suppose we run Algorithm 2 for \(k\) iterations. Then for some absolute constants \(C_1, C_2\),
\[
\epsilon := \|\Sigma^{\frac{1}{2}} \eta^{\text{best}}_k\| - \|\Sigma^{\frac{1}{2}} \eta^*\| \leq \frac{C_1^2 + C_2^2 \sum_{n=1}^{k} h_n^2}{\sum_{n=1}^{k} h_n^2}.
\]

For \(h_n\) satisfying \(\sum_{n=0}^{\infty} h_n = +\infty\) and \(\sum_{n=0}^{\infty} h_n^2 = o(\sum_{n=0}^{\infty} h_n)\), we have \(\epsilon \to 0\) so that \(\lim_{n \to \infty} \|\Sigma^{\frac{1}{2}} \eta^{\text{best}}_n\| = \|\Sigma^{\frac{1}{2}} \eta^*\|\), which implies the convergence of the subgradient descent in asymptotic time. Moreover, different choices of the step size \(h_n\) give different convergence rates. For example, if \(h_n = 1/\sqrt{n}\), the convergence rate is nearly quadratic as \(k = O(\log^2 k/\epsilon^2)\); if \(h_n = h \asymp \epsilon\) is a fixed small constant, the exact quadratic convergence rate \(k = O(1/\epsilon^2)\) is achieved (although the algorithm does not converge in asymptotic time in this case).

### 3  Asymptotic Distribution and Confidence Interval of the Debiased Estimator

In this section we derive the limiting distribution of the debiased estimator obtained by Algorithm 1. We then construct a confidence interval using a consistent estimator of \(\sigma\) — the standard deviation of the noise \(\epsilon\). The following Theorem 3.1 shows that Algorithm 1 successfully debiases the \(j\)-th coordinate of an estimator of \(\beta^*\) given model (1.2), when the population covariance matrix \(\Sigma\) has bounded spectrum.
**Theorem 3.1.** Consider a linear model in (1.2) with Gaussian errors $\varepsilon_i \sim N(0, \sigma^2)$. Suppose the eigenvalues of $\Sigma$ are bounded from both above and below. Then the debiased $j^{th}$ coefficient $\hat{\beta}_d^{(j)}$ obtained by Algorithm 1 is conditionally asymptotically normal with mean equal to $\beta^{*(j)}$. In particular, if $Z_j = \frac{1}{\sqrt{n}} \bar{\eta}^\top \bar{X}^\top \varepsilon$, we have

$$\sqrt{n}(\hat{\beta}_d^{(j)} - \beta^{*(j)}) = Z_j + \Delta_j, \quad Z_j |\bar{X}, \bar{Y}, \bar{X} \sim N(0, \sigma^2 \bar{\eta}^\top \bar{\Sigma} \bar{\eta}), \quad \Delta_j = \sqrt{n}(\bar{\eta}^\top \bar{\Sigma} - \epsilon^{(j)^\top})(\beta^* - \nu),$$

and $\Delta_j = o_p(1)$ converges to zero with probability converging to one.

**Remark 3.2.** We will reiterate that our debiasing procedure works for a linear combination of coordinates (i.e. a contrast). It is not hard to see from the proof of Theorem 2.2 and Theorem 3.1 that if we replace $\epsilon^{(j)}$ by any bounded in Euclidean norm vector, the same results will also hold. In terms of implementation, to debias a contrast, one simply needs to replace $\epsilon^{(j)}$ by the relevant vector with bounded norm in step 2.

**Remark 3.3.** For simplicity of exposition the above theorem assumes that the errors are Gaussian. Our procedure also works with non-Gaussian errors using a modification similar in spirit to the one proposed in (Javanmard and Montanari, 2014, Section 4). Details will be given in Section 7.

### 3.1 Confidence Intervals

Based on Theorem 3.1, a $(1 - \alpha)$-level confidence interval of $\beta^{*(j)}$ can be constructed as

$$\left(\hat{\beta}_d^{(j)} - z_{\alpha/2} \sigma \frac{\|\bar{\Sigma} \bar{\eta}\|}{\sqrt{n}}, \hat{\beta}_d^{(j)} + z_{\alpha/2} \sigma \frac{\|\bar{\Sigma} \bar{\eta}\|}{\sqrt{n}}\right).$$

(3.1)

Usually the variance of the noise $\sigma$ is unknown. Thus the need for consistent estimation of $\sigma$ arises. In order to estimate $\sigma$ we assume there exists an estimator $\hat{\sigma}$ which does well in terms of mean squared prediction error (see Theorem 3.4 for the precise assumption on $\hat{\sigma}$). We use only the first half of the data to estimate $\sigma$ with $\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i \in [n]} (Y_i - X_i^\top \hat{\beta})^2}$. Alternatively, for this step one could estimate $\hat{\sigma}$ using the entire data set, since we do not need sample splitting when we estimate $\sigma$ (we only need a consistent estimator). The following Theorem 3.4 proves the consistency of such an estimator of $\sigma$. Theorem 3.4 does not require the noise to be Gaussian, and even sub-Gaussian. It only assumes the existence of a 6-th moment.

**Theorem 3.4.** Let $\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i \in [n]} (Y_i - X_i^\top \hat{\beta})^2}$. Suppose $\mathbb{E} \varepsilon^6 < +\infty$, and that the eigenvalues of $\Sigma$ are bounded from above and below. Let $\hat{\beta}$ be an estimator of $\beta^*$ such that with probability converging to 1 we have $\|X(\hat{\beta} - \beta^*)\| \lesssim \sigma \delta$ for some $\delta = o(\sqrt{n})$. Then with probability converging to $1 - e^{-\delta^2/2}$, we have

$$|\hat{\sigma}^2 - \sigma^2| \lesssim \frac{(\sqrt{Var(\varepsilon_i^2)} \lor \sigma^2) \delta}{\sqrt{n}}.$$

In the above since $\delta$ can be taken such that $\delta \to \infty$ as $n \to \infty$ (as long as $\delta = o(\sqrt{n})$), the result shows that $\hat{\sigma}$ is consistent. Note that the assumption $\|X(\hat{\beta} - \beta^*)\| \lesssim \sigma \delta$ is achieved by many estimators. For example, (Neykov, 2019, Lemma A.1) implies that convex constrained least squares estimators satisfy this condition; (Bellec et al., 2018b, Corollary 6.2) and (Derumigny et al., 2018,
Corollary 6.2) imply that it holds for SLOPE and square-root SLOPE. The explicit order of $\delta$ for those cases can be found in Lemma 4.3 and Lemma 5.5 when we consider applying our general procedure to some special cases. In the case when $\hat{\sigma}$ is consistent, it follows by Slutsky’s theorem that $\sigma$ in the confidence interval in (3.1) can be substituted with $\hat{\sigma}$:

$$
\left(\hat{\beta}_d^{(j)} - z_\alpha \frac{\|\hat{\Sigma}\|}{\sqrt{n}}, \hat{\beta}_d^{(j)} + z_\alpha \frac{\|\hat{\Sigma}\|}{\sqrt{n}}\right).
$$

(3.2)

In the following Section 4 and Section 5, we discuss in details how to implement the debiasing procedure Algorithm 1 for some commonly used estimators including onotone regression, positive monotone regression, non-negative least squares, LASSO, SLOPE and square-root SLOPE. More concretely, the next section, Section 4 is dedicated to convex constrained least squares, while Section 5 discusses an application to SLOPE and square-root SLOPE.

## 4 Convex Constrained Least Squares

In this section we are interested in the estimator (1.3) which we mentioned in the introduction section. Clearly this estimator is a form of constrained least squares, where the practitioner has knowledge that the true coefficient $\beta^*$ belongs to a convex set $K$. Assuming that least squares is a reasonable criteria to estimate $\beta^*$, the practitioner further imposes a restriction that $\hat{\beta} \in K$. Similarly to how LASSO biases the coefficients by shrinking them towards zero, imposing a constraint on $\hat{\beta}$ also biases the coefficients and standard inference methods do not work even in the low-dimensional setting. This motivates us to debias individual coordinates or contrasts of the estimator $\hat{\beta}$. In this section, we will assume that $X_i \sim N(0, \Sigma)$. The sole reason why we require this, is that there are known estimation and in-sample prediction guarantees for the performance of $\hat{\beta}$ given in Neykov (2019) which require the same condition. We do anticipate that at least some of those results may be generalized to broader distributional settings, as suggested by the works of Genzel and Kipp (2020); Li et al. (2015), but this is out of the scope of the present paper.

Since a set $K$ with the property $\beta^* \in K$ is given, it is natural to try and use that knowledge in our abstract debiasing procedure. In particular, we will use $K$ as the convex set required in step 1 and step 2 of Algorithm 1. It remains to construct a vector $v \in K$ which obeys the requirements of step 1. We now provide such a construction. We claim that the solution of the following optimization program

$$
v := \arg\min_{w \in K} \|\hat{\beta} - w\| + \frac{\bar{w}(T_K(w) \cap S^{p-1})}{\sqrt{n}},
$$

(4.1)

would satisfy the properties required of $v$. We now give a high level intuition why such $v$ is worth considering. Recall that the condition $\bar{w}(T_K(v) \cap S^{p-1})\|v - \beta^*\| = o_p(1)$ in step 1 of Algorithm 1. This condition will be met if both $\bar{w}(T_K(v) \cap S^{p-1})$ and $\|v - \beta^*\|$ are “small”. Suppose there exists a vector $v'$ such that $\|v' - \beta^*\|$ is small, and in addition $v'$ has a “small” tangent cone, in the sense that $\frac{\bar{w}(T_K(v') \cap S^{p-1})}{\sqrt{n}}$ is small. By the definition of $v$ it follows that

$$
\|\hat{\beta} - v\| + \frac{\bar{w}(T_K(v) \cap S^{p-1})}{\sqrt{n}} \leq \|\hat{\beta} - v'\| + \frac{\bar{w}(T_K(v') \cap S^{p-1})}{\sqrt{n}}.
$$
Therefore both terms $\|\hat{\beta} - v\|$ and $\frac{\|T_K(v)\cap S^{p-1}\|}{\sqrt{n}}$ are “small”. By the triangle inequality $\|v - \beta^*\| \leq \|\hat{\beta} - v\| + \|\hat{\beta} - \beta^*\|$. Finally we know by a result of (Neykov, 2019, see Corollary 2.7) that $\|\hat{\beta} - \beta^*\|$ is “small”. This implies that $\|v - \beta^*\|$ is “small”. Theorem 4.1 makes the above intuition precise and proves why the solution of program (4.1) satisfies the condition needed in step 1.

**Theorem 4.1.** Consider the same setting as Theorem 3.1, and further assume that $X_i \sim N(0, \Sigma)$. Suppose there exists $\nu' \in K$ such that $\|\nu' - \beta^*\|^2 = o(1/\sqrt{n})$, and the tangent cone of $K$ at $\nu'$ has a simple structure such that $\bar{w}^2(\mathcal{T}_K(\nu') \cap S^{p-1}) = o(\sqrt{n})$ and $\bar{w}^2(\mathcal{T}_K(\nu') \cap S^{p-1}) \to \infty$. Then for $\hat{\beta}$ being the constrained least squares estimator obtained via (1.3), the solution $v$ of (4.1) satisfies the condition needed in step 1 of Algorithm 1 with probability converging to 1 asymptotically.

**Remark 4.2.** Some comments are in order. The existence of a vector $\nu'$ which is close to $\beta^*$, with a sufficiently small tangent cone is natural. If $\nu' = \beta^*$, this condition requires that $\beta^*$ has a simple structure; otherwise when $\nu' \neq \beta^*$ it does not require that $\beta^*$ has a simple structure, as long as it is close enough to a vector $\nu'$ with a simple structure. This enables consistent estimation of $\beta^*$ in high-dimensional settings. As an example (for a case when $\nu' = \beta^*$) consider the set $K = \{\beta : \|\beta\|_1 \leq \|\beta^*\|_1\}$ which is the LASSO constraint. Requiring that $\beta^*$ has a cone with small Gaussian complexity is equivalent to imposing a sparsity assumption on $\beta^*$. This example is considered in more details in Subsection 4.5 below.

In addition, notice that the vector $\nu'$ in Theorem 4.1 is not necessarily the same as the vector $v$ found by (4.1). However, it may be useful to think that the vector $v$ is attempting to estimate $\nu'$ (although this intuition too is not necessarily precise). The existence of $\nu'$ guarantees that we can find a “useful” $v$ by (4.1) in step 1. After we find the desired $v$, one can compute the auxiliary vector $\hat{\eta}$ in step 2 based on $v$ and $K$, and then use $\hat{\eta}$ to construct the debiased estimator $\hat{\beta}_d$ and the confidence interval as (3.1) or (3.2).

Of course, in practice, in order to construct the confidence interval (3.2) we need to estimate $\sigma$. As discussed in Lemma 4.3 below, consistent estimation of $\sigma$ is possible in the convex constrained least squares case.

**Lemma 4.3.** Consider the same setting as Theorem 4.1 where $\hat{\beta}$ is a convex constrained least squares estimator. Then Theorem 3.4, applies with

$$
\delta \asymp \frac{\sqrt{n}}{\sigma} \|\nu' - \beta^*\| + \bar{w}(\mathcal{T}_K(\nu') \cap S^{p-1}),
$$

where $\delta = o(\sqrt{n})$ as required.

Our debiasing algorithm does not require the population covariance matrix $\Sigma$ to be known as long as it has bounded spectrum. Can one do better if one is given knowledge of $\Sigma$? It is known (Javanmard et al., 2018) that with prior knowledge of $\Sigma$, the LASSO estimator $\hat{\beta}$ can be debiased with the following formula:

$$
\hat{\beta}_d = \hat{\beta} + n^{-1} \Sigma^{-1} \bar{X}^\top (\bar{Y} - \bar{X}\hat{\beta}).
$$

(4.2)

What is more, Javanmard et al. (2018) show that when the design is Gaussian the requirement for the debiasing procedure to work with known $\Sigma$ is much weaker compared to the requirement with unknown $\Sigma$. See also Bellec and Zhang (2019a) for a sharpened version of this result. In
fact Javanmard et al. (2018) also show that the same debiased estimator works without sample splitting under more stringent assumptions, but this is out of the scope of the present paper. Lemma 4.4 will show that the debiasing formula in (4.2) also works for any convex constrained least squares estimator under proper conditions. Afterwards we will compare the conditions needed to successfully debias a convex constrained least squares estimator \( \hat{\beta} \) for the known and unknown \( \Sigma \) cases. Similarly to the LASSO case, without the knowledge of \( \Sigma^{-1} \), we impose more stringent assumptions on the structure of tangent cones of the parameter space \( K \).

**Lemma 4.4.** Consider a linear model as in (1.2) with Gaussian errors \( \varepsilon_i \sim N(0, \sigma^2) \). Further assume that \( X_i \sim N(0, \Sigma) \). Let \( \{a_n\}_{n=1}^\infty \) be any slowly diverging sequence with \( n \), and let \( \mathbf{v}' \in K \), be a vector such that \( \|\mathbf{v}' - \beta^*\|_2 = o(1) \), \( \overline{\mathbb{w}}(T_K(\mathbf{v}') \cap \mathbb{S}^{p-1})a_n = o(\sqrt{n}) \) and \( \overline{\mathbb{w}}(T_K(\mathbf{v}') \cap \mathbb{S}^{p-1}) \to \infty \). Let \( \hat{\beta} \) be a convex constrained least squares estimator obtained by (1.3) on the first half of the data. The debiased \( j \)-th coefficient \( \hat{\beta}_d^{(j)} \) obtained by (4.2) is conditionally asymptotically normal with mean equal to \( \beta^{*(j)} \). In particular, let \( Z = \frac{1}{\sqrt{n}} \Sigma^{-1} \hat{X}^\top \varepsilon \), and \( \hat{X} = \frac{1}{n} \hat{X}^\top \hat{X} \) be the empirical Gram matrix of the second half, we have

\[
\sqrt{n}(\hat{\beta}_d^{(j)} - \beta^{*(j)}) = Z^{(j)} + \Delta^{(j)}, \quad Z^{(j)}|\hat{X} \sim N(0, \sigma^2 e^{(j)\top} \Sigma^{-1} \hat{X} \Sigma^{-1} e^{(j)}),
\]

and \( \Delta^{(j)} = o_p(1) \) converges to zero with probability converging to one.

Suppose \( \beta^* = \mathbf{v}' \) is s-sparse and \( K = \{\beta : ||\beta||_1 \leq ||\beta^*||_1\} \). The condition \( \overline{\mathbb{w}}(T_K(\mathbf{v}') \cap \mathbb{S}^{p-1})a_n = o(\sqrt{n}) \) in Lemma 4.4 is in fact a condition on the sparsity \( s \). The Gaussian complexity of the tangent cone \( T_K(\beta^*) \) can be evaluated in terms of the sparsity \( s \) as (Chandrasekaran et al., 2012, Proposition 3.10)

\[
\overline{\mathbb{w}}(T_K(\beta^*) \cap \mathbb{S}^{p-1}) = O\left(\sqrt{s \log \frac{ep}{s}}\right).
\]

Thus if \( s \) doesn’t scale with \( n, p \) we have \( s = o(n/(a_n^2 \log p)) \). If one selects \( a_n = \sqrt{\log p} \), the condition in Lemma 4.4 becomes \( s = o(n/(\log p)^2) \) (assuming \( p \to \infty \) as \( n \to \infty \)), which matches the condition needed in debiasing the regularized LASSO for the known covariance case (Javanmard et al., 2018). Assuming \( a_n = \sqrt{\log p} \) is convenient since in this case by tracking the proof of Lemma 4.4 and applying the union bound one may claim that (4.3) holds for all \( j \in [p] \), which is precisely the setting of (Javanmard et al., 2018).

The condition \( \overline{\mathbb{w}}^2(T_K(\mathbf{v}') \cap \mathbb{S}^{p-1}) = o(\sqrt{n}) \) needed in Theorem 4.1 is more stringent than the condition \( \overline{\mathbb{w}}(T_K(\mathbf{v}') \cap \mathbb{S}^{p-1})a_n = o(\sqrt{n}) \) in Lemma 4.4, which can be viewed as the price we pay for having an unknown covariance. On an important note, presently we do not have corresponding lower bounds showing that these conditions are also necessary. We may say however that in the case when \( K = \{\beta : ||\beta||_1 \leq ||\beta^*||_1\} \) the condition \( \overline{\mathbb{w}}^2(T_K(\mathbf{v}') \cap \mathbb{S}^{p-1}) = o(\sqrt{n}) \) reduces to a condition on the sparsity parameter \( s \) by (4.4). The equivalent condition in terms of \( s \) is \( s \log ep/s = o(\sqrt{n}) \) which matches the assumption needed in debiasing the regularized LASSO for the unknown covariance case (Cai et al., 2017; Javanmard et al., 2018).

### 4.1 Lower Bounds on Confidence Interval Length

We will now show that under certain conditions the \( \frac{1}{\sqrt{n}} \)-rate of the confidence intervals that we provide cannot be improved in a worst case sense. Of course one should not expect this is always
Finally, define the worst case expected confidence interval length over $\Theta$:

$$ L(CI_{\alpha}(e^{(j)^T} \beta, X, Y), \Theta) := \sup_{\beta \in \Theta} E_\beta L(CI_{\alpha}(e^{(j)^T} \beta, X, Y)). $$

The above definitions are extracted from Cai et al. (2017) whose work forms the basis of our Lemma 4.6. We need one final definition before we state the result.

**Definition 4.5.** For a fixed upper bound $\bar{w}(T_K(\beta) \cap S^{p-1})$ of $w(T_K(\beta) \cap S^{p-1})$, and $\delta > 0$ let

$$ r_n := \inf_{\beta \in S} \frac{\bar{w}^2(T_K(\beta) \cap S^{p-1})}{\sqrt{n}}, \quad S := \left\{ \beta \in K : \beta(+ \text{ or } -) \frac{\delta \sigma \| \Sigma \|_{op}^{-1}}{\sqrt{n}} e^{(j)} \in K \right\}, $$

where $r_n = \infty$ if $S = \emptyset$.

We suppress the dependence of $r_n$ on $\delta$, $K$ and $\bar{w}(T_K(\beta) \cap S^{p-1})$ to ease the notation. In the above definition observe that the set $K$ and the dimension $p$ are also allowed to change with $n$. We have

**Lemma 4.6.** Let $K \subset \mathbb{R}^p$ be a convex set. For a fixed upper bound $\bar{w}(T_K(\mathbf{v}) \cap S^{p-1})$ of $w(T_K(\mathbf{v}) \cap S^{p-1})$ such that for all $\mathbf{v} \in K$, $\bar{w}(T_K(\mathbf{v}) \cap S^{p-1}) \to \infty$ and a fixed $\delta > 0$, suppose that $r_n = o(1)$. For any sequence $R_n \geq 2r_n$ such that $R_n = o(1)$, define the parameter space

$$ \mathcal{H} := \mathcal{H}(R_n) = \{ \beta \in K : \| \beta - \mathbf{v} \|^2 \leq R_n / \sqrt{n}, \text{ for } \mathbf{v} \in K \text{ and } \bar{w}^2(T_K(\mathbf{v}) \cap S^{p-1}) \leq R_n \sqrt{n} \}. \quad (4.5) $$

Then for any $\beta^* \in \mathcal{H}$ and sufficiently large $n$ we have

$$ \inf_{CI_{\alpha}(e^{(j)^T} \beta^*, X, Y) \in \mathcal{L}_{\alpha}(\mathcal{H})} L(CI_{\alpha}(e^{(j)^T} \beta^*, X, Y), \mathcal{H}) \geq \delta \left( 1 - 2\alpha - \sqrt{\exp(2\delta^2) - 1} \right) \frac{\sigma \| \Sigma^\frac{1}{2} \|_{op}^{-1}}{\sqrt{n}}. $$

**Remark 4.7.** Notice that given a convex set $K$, our Algorithm 1 is able to perform debiasing asymptotically over the parameter space $\mathcal{H}$ according to Theorem 4.1. The result of Lemma 4.6 shows that the length of our confidence interval (3.1) for a single coefficient $\beta^*(j)$ cannot be much
improved asymptotically in a worst case sense, since its length times $\sqrt{n}$ is at least of the order of a constant (assuming $\Sigma$ has bounded spectrum). As mentioned earlier, we cannot expect that the sequence $r_n = o(1)$ for all convex sets $K$. But in all examples we consider in this work, $r_n = o(1)$ holds. For instance, if $K$ is a monotone cone or positive monotone cone as we will study later in Section 4.2 and Section 4.3, a monotone vector comprised of two constant pieces whose jump from the $(j-1)$-th coordinate to the $j$-th coordinate is greater than $\delta_\sigma \| \Sigma_1/2 \|_{op}^{-1}/\sqrt{n}$ will produce $r_n \asymp \frac{2 \log(ep/2)}{\sqrt{n}}$ (Bellec et al., 2018a, see (1.19), (1.22), Proposition 3.1). Also, if $K = \{ \beta : \| \beta \|_1 \leq \| \beta^* \|_1 \}$, there exists a 1-sparse vector $\mathbf{v}$ (with $j$-th coefficient equal to $\| \beta^* \|_1$) which gives $r_n = o(1)$ whenever $1 = o(\sqrt{n}/\log p)$ and $\| \beta^* \|_1 \geq \sigma \| \Sigma_1/2 \|_{op}^{-1}/(2\sqrt{n})$. If $K$ is the non-negative orthant cone, a vector of zeros with exception of its $j$-th coordinate being equal to $\sigma \| \Sigma_{1/2} \|_{op}^{-1}/\sqrt{n}$ will yield $r_n \lesssim \frac{1}{\sqrt{n}}$ so when $p = o(\sqrt{n})$, $r_n = o(1)$.

We end up this section with a result slightly stronger than Lemma 4.6 for the special case when $K$ is a polyhedral cone (i.e. $K = \{ x \in \mathbb{R}^n : Ax \geq 0 \}$ for some matrix $A$) as is the case when $K$ is the monotone or positive monotone cone or the non-negative orthant cone. It is well known that polyhedral cones are finitely-generated, i.e., there exists a $k \in \mathbb{N}$ and unit norm vectors $w_1, \ldots, w_k$ such that $K = \{ \sum_{i \in [k]} \alpha_i w_i : \alpha_i \geq 0 \}$. We have the following

**Lemma 4.8.** Fix a number $j \in [k]$. Let $\mathcal{H}(R_n)$ be defined as in (4.5), and set $\nu_n := 2R_n + 2\delta^2 \| \Sigma_{1/2} \|_{op}^{-2}/\sqrt{n}$. Then for any $\beta^* \in \mathcal{H}(R_n)$ we have

$$\inf_{CI_\alpha(w_j^\top \beta^*, X, Y) \in IC_\alpha(\mathcal{H}(\nu_n))} L(CI_\alpha(w_j^\top \beta^*, X, Y)) \geq \delta \left( 1 - 2\alpha - \sqrt{\exp(2\delta^2) - 1} \right) \frac{\sigma \| \Sigma_{1/2} \|_{op}^{-1}}{\sqrt{n}}.$$

In other words, if one is interested in performing inference along a generating direction of the cone, the confidence interval length has to be at least $\frac{1}{\sqrt{n}}$ for any $\beta^* \in \mathcal{H}(R_n)$ for all algorithms which return valid $(1-\alpha)$-confidence intervals for all vectors in $\mathcal{H}(\nu_n)$. Note that since $\nu_n = o(1)$ our debiasing algorithm will produce $(1-\alpha)$-level confidence intervals on $\mathcal{H}(\nu_n)$ asymptotically, and therefore the length of the confidence intervals for contrasts equal to generating directions of the cone cannot be improved. Unlike Lemma 4.6, Lemma 4.8 is not a worst case result since we are not taking sup over all vectors in the parameter space. We now give concrete examples of sets $K$ for which our algorithm is fully implementable.

### 4.2 Monotone Cone Regression

Consider the case where the true coefficient $\beta^*$ is in a monotone cone parameter space $M^p$ in $\mathbb{R}^p$ defined as

$$M^p = \{ (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p : \beta_1 \leq \beta_2 \leq \ldots \leq \beta_p \}.$$  

Notice that $M^p$ is convex. Moreover, the set of monotone vectors with $l$ constant pieces is defined as (Gao et al., 2017)

$$M^p_l = \{ (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p : \text{there exist } \{ a_j \}_{j=0}^l \text{ and } \{ u_j \}_{j=0}^l \text{ such that } 0 = a_0 \leq a_1 \leq \ldots \leq a_l = p, u_1 \leq u_2 \leq \ldots \leq u_l, \text{ and } \beta_1 = u_j \text{ for all } i \in (a_{j-1}, a_j) \}.$$
Given the prior knowledge $\beta^* \in M^p$, the constrained least squares estimator $\hat{\beta}$ in (1.3) can be solved by incorporating isotonic regression in projected gradient descent.

To find the desired vector $v$ in step 1, we solve (4.1) with $\pi(T_{M^p}(v) \cap S^{p-1}) = \sqrt{l \log(ep/l)}$. The latter is a legitimate upper bound on the Gaussian complexity of the tangent cone, as the result in (Bellec et al., 2018a, (1.19), (1.22), Proposition 3.1) shows that for a monotone cone $M^p \in \mathbb{R}^p$, the complexity of the tangent cone at any vector $v$ comprised of $l$ constant pieces has an explicit upper bound $w(T_{M^p}(v) \cap S^{p-1}) \leq \sqrt{l \log(ep/l)}$. Thus the optimization problem (4.1) can be simplified to

$$\arg\min_{v \in M^p} \|\hat{\beta} - v\| + \sqrt{l \log(ep/l)}. \quad (4.6)$$

For a fixed $l$, the term $\sqrt{(l/n) \log(ep/l)}$ is constant for all $v \in M^p$. Thus in each $M^p_l$, the solution of $\arg\min_{v \in M^p_l} \|\hat{\beta} - v\| + \sqrt{(l/n) \log(ep/l)}$ should minimize $\|\hat{\beta} - v\|$, which is exactly the projection of $\hat{\beta}$ to $M^p_l$, denoted as $\Pi_{\hat{\beta}}(\hat{\beta})$. Let $p'$ be the number of constant pieces in $\hat{\beta}$, where $p' \leq p$. The optimization problem (4.6) can be converted to an optimization problem over finitely many candidates. Define

$$\hat{v} = \arg\min_{l \in [1, p]} \|\hat{\beta} - \Pi_{M^p_l}(\hat{\beta})\| + \sqrt{\frac{l \log(ep/l)}{n}}.$$

Since there is no point in looking for values of $l > p'$ as this will only increase the loss function (compared to when $l = p'$), the desired $v$ in (4.6) is exactly $\Pi_{M^p_l}(\hat{\beta})$. There is an efficient projection algorithm of $\hat{\beta}$ to $M^p_l$ as proposed by (Gao et al., 2017, Algorithm 1) which takes $O(p^3)$ time to compute all projections for $l \in [1, p']$.

Once $v$ is obtained, we solve the optimization program (2.1) using subgradient descent as in Algorithm 2. The final piece of the puzzle is to show how to calculate the projections $\Pi_{T_{M^p_v}(v)}(\cdot)$ and $\Pi_{-T_{M^p_v}(v)}(\cdot)$. We compute them by decomposing $T_{M^p_v}(v)$. Since $v$ is a piece-wise constant, the tangent cone of $M^p$ at $v$ can be decomposed as (Bellec et al., 2018a, Proposition 3.1)

$$T_{M^p_v}(v) = M^{p_1} \times M^{p_2} \times \ldots \times M^{p_t},$$

where each $p_i$ is the length of each constant piece of $v$, and $p_1 + \ldots + p_t = p$. Thus for any vector $u = (u_1, u_2, \ldots, u_p)^\top \in \mathbb{R}^p$, the projection of $u$ to $T_{M^p_v}(v)$ is (Amelunxen et al., 2014, Equation B.2)

$$\Pi_{T_{M^p_v}(v)}(u) = \left(\Pi_{M^{p_1}}((u_1, \ldots, u_{p_1}))^\top, \Pi_{M^{p_2}}((u_{p_1+1}, \ldots, u_{p_1+p_2}))^\top, \ldots, \Pi_{M^{p_t}}((u_{p_1+\ldots+p_{t-1}+1}, \ldots, u_p))^\top\right)^\top,$$

noting that projections into a monotone cone, as in (4.7) can be efficiently implemented via the PAVA algorithm for isotonic regression (Robertson, 1988, see e.g.). Once we have computed $\hat{\eta}$, we can debias $\hat{\beta}$ using the formula in step 2. The entire procedure to get a debiased estimation $\hat{\beta}_d^{(j)}$ for monotone cone regression is summarized in Algorithm 3.

**Remark 4.9.** We remark that thanks to Theorem 4.1, $\beta^*$ need not be piecewise constant. In fact, by Lemma 2 of Bellec and Tsybakov (2015) we know that any vector $\beta^* \in M^p$, can be
approximated within \( \| \beta^* - \nu' \| \leq \frac{\beta^*(p) - \beta^{*(1)}}{2k} \) by a vector \( \nu' \) consisting of at most \( k \) constant pieces. So long as \( \beta^*(p) - \beta^{*(1)} \) is bounded, it suffices that \( p \) is such that we can select \( k \gg n^{1/4} \) with \( k \log p / k = o(\sqrt{n}) \), and the regression with signal \( \beta^* \) can be debiased.

Algorithm 3 Debias the \( j^{th} \) Coefficient for Monotone Cone Regression

**Input:** Two equal size partitions \((\bar{X}, \bar{Y})\) and \((\tilde{X}, \tilde{Y})\), \( \hat{\beta} \) obtained by projected gradient descent with isotonic regression.

**Initialize:** Empirical Gram matrix of the second partition \( \Sigma = \frac{1}{n} \bar{X}^\top \bar{X} \).

1. Solve \( \hat{l} \leftarrow \arg\min_{l \in [1,p']} \| \hat{\beta} - \Pi_{M^p l}(\hat{\beta}) \| + \sqrt{\frac{4}{n} \log p} \).
   \( \nu \leftarrow \Pi_{M^p l}(\hat{\beta}) \).

2. Run Algorithm 2. Compute \( \Pi_{\mathcal{T}_{M^p l}(\nu)()} \) by isotonic regression (PAVA) with (4.7). For \( \Pi_{-\mathcal{T}_{M^p l}(\nu)()} \) use (2.5). The debiased \( j^{th} \) coefficient equals \( \hat{\beta}^{d(j)}_d \leftarrow v^{(j)} + n^{-1} \hat{\eta}^\top \bar{X}^\top (\tilde{Y} - \tilde{X}v) \).

4.3 Positive Monotone Cone Regression

Based on the analysis in Section 4.2 for the monotone cone \( M^p \), we can analogously develop the debiasing technique when the true coefficient is inside of a positive monotone cone defined as

\[ M^{p+} = \{ (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p : 0 \leq \beta_1 \leq \beta_2 \leq \ldots \leq \beta_p \} .\]

The algorithm to debias the \( j^{th} \) coefficient in positive monotone cone regression is the same as Algorithm 3 except for some minor modifications. Specifically, \( \hat{\beta} \) can also be obtained by projected gradient descent, but such a projection onto a positive monotone cone is done by fitting an isotonic regression followed by assigning zeros to all the negative coordinates (Németh and Németh, 2012). The procedure of finding \( \nu \) in step 1 is the same as the monotone cone case. This is so since \( \hat{\beta} \) is always positive and the algorithm in Gao et al. (2017) computes the projections of \( \hat{\beta} \) onto \( M^p_l \) by further averaging itself, all the projections automatically belong to the positive monotone cone. For step 2, we need to project a vector \( u = (u_1, \ldots, u_p)^\top \in \mathbb{R}^p \) onto \( \mathcal{T}_{M^{p+}}(v) \)—the tangent cone of the positive monotone cone \( M^{p+} \) at \( v \). By Proposition 4.10, \( \mathcal{T}_{M^{p+}}(v) \) can be decomposed into Cartesian products of a positive monotone cone and several other monotone cones. Thus the projection onto \( \mathcal{T}_{M^{p+}}(v) \) can be computed as a Cartesian product of the projection onto every component.

**Proposition 4.10.** Suppose \( v \in M^{p+} \) has \( l \) constant pieces, and the length of each constant piece is \( p_i \) for \( i \in [l] \). If the first constant piece consists of zeros, the tangent cone of \( M^{p+} \) at \( v \) can be decomposed as

\[ \mathcal{T}_{M^{p+}}(v) = M^{p_1+} \times M^{p_2} \times \ldots \times M^{p_l} , \]

otherwise it is

\[ \mathcal{T}_{M^{p+}}(v) = M^{p_1} \times M^{p_2} \times \ldots \times M^{p_l} . \]

**Remark 4.11.** Similarly to the monotone cone case, the \( \beta^* \) vector need not be piecewise constant. See Remark 4.9.
4.4 Non-negative Least Squares

In this section we suppose that \( K = \{ \beta : \beta^{(i)} \geq 0 \, \forall i \in [p] \} \) is the non-negative orthant cone. Clearly, implementing the non-negative least squares can be done via a quadratic program, or with a projected gradient descent, where the projection onto the non-negative orthant is given by setting to 0 any negative coefficients.

In order to implement \((4.1)\) and find \( v \) in step 1, we need to evaluate an upper bound on the Gaussian complexity of \( T_K(v) \cap \mathbb{S}^{p-1} \) for any \( v \in K \); see Lemma 4.12.

**Lemma 4.12.** If \( K = \{ \beta : \beta^{(i)} \geq 0 \, \forall i \in [p] \} \) is the non-negative orthant cone, for any \( v \in K \) the following bound holds
\[
w(T_K(v) \cap \mathbb{S}^{p-1}) \leq \sqrt{p - |\{ i : v^{(i)} = 0 \}|}/2.
\]

Then as in the monotone cone case, the problem \((4.1)\) boils down to an optimization over finitely many candidates. Let \( v_s \) be the projection of \( \hat{\beta} \) onto the set of non-negative vectors with exactly \( s \) zero coefficients. We then need to solve
\[
\hat{s} = \arg\min_{s \in [0,p]} \| \hat{\beta} - v_s \| + \sqrt{\frac{p-s/2}{n}},
\]
and our final solution is \( v = v_{\hat{s}} \). What is left to show is how to obtain a vector \( v_s \), which is discussed in Lemma 4.13.

**Lemma 4.13.** Let \( S \) denote the index set of the \( s \) smallest in magnitude coefficients of \( \hat{\beta} \). The vector \( v_s \) is given by
\[
v_s^{(i)} = \hat{\beta}^{(i)} \mathbf{1}(i \in S^c).
\]
In other words \( v_s \) greedily takes the largest \( p-s \) entries in \( \hat{\beta} \), where ties are broken arbitrarily.

After we obtain \( v \) in step 1, we also need to write down the explicit form of the projection \( \Pi_{T_K(v)} \) to solve step 2. Such a projection is provided in Lemma 4.14.

**Lemma 4.14.** We have that
\[
\Pi_{T_K(v)}(x) = (x^{(i)} \mathbf{1}(v^{(i)} \neq 0) + (x^{(i)})_+ \mathbf{1}(v^{(i)} = 0))_{i \in [p]}
\]

We summarize the procedure in Algorithm 4.

**Algorithm 4** Debias the \( j \)th Coefficient for Non-negative Least Squares

**Input:** Two equal size partitions \((X, Y)\) and \((\tilde{X}, \tilde{Y})\), \( \hat{\beta} \) obtained by projected gradient descent with isotonic regression.

**Initialize:** Empirical Gram matrix of the second partition \( \hat{\Sigma} = \frac{1}{n} \tilde{X}^\top \tilde{X} \).

1. Solve \( \hat{s} = \arg\min_{s \in [0,p]} \| \hat{\beta} - v_s \| + \sqrt{\frac{p-s/2}{n}}, \)
\( v \leftarrow v_{\hat{s}} \).

2. Run Algorithm 2. Compute \( \Pi_{T_K(v)}(\cdot) \) by the result of Lemma 4.14. For \( \Pi_{-T_K(v)}(\cdot) \) use (2.5). The debiased \( j \)th coefficient equals \( \hat{\beta}_d^{(j)} \leftarrow v^{(j)} + n^{-1} \eta^\top \tilde{X}^\top (\tilde{Y} - \tilde{X} v) \).
4.5 LASSO Constrained Version

The next example is a constrained LASSO problem. There are many methods for debiasing LASSO, but we remark that our algorithm below is distinct from all of these proposals as it creates a different projection direction \( \hat{y} \). We note however, that since the constrained LASSO has an \( \ell_1 \) guarantee (Wainwright, 2019, see Theorem 7.1), some previous debiasing methods are applicable according to (Javanmard et al., 2018, eq. 9). Hence our intent with this section is to create a proof of concept that our debiasing scheme is also applicable to constrained LASSO. More importantly however, this section will serve as a building block to our algorithm which debiases SLOPE and square-root SLOPE (see Section 5 below).

Suppose that the convex set is \( K = \{ \beta : \| \beta \|_1 \leq \| \beta^* \|_1 \} \). Here we assume a prior knowledge of \( \| \beta^* \|_1 \), which is a common assumption in works analyzing the constrained version of LASSO (Thrampoulidis et al., 2014; Wainwright, 2019, see e.g.). The constrained LASSO can be converted to an ordinary quadratic program with \( 2p \) variables and \( 2p + 1 \) constraints by rewriting \( \beta^{(j)} \) as \( \beta^{+(j)} - \beta^{-(j)} \) (Tibshirani, 1996, Section 6).

\[
\begin{align*}
\arg\min_{\beta^{+}, \beta^{-} \in \mathbb{R}^p} & \left\| \mathbf{Y} - \mathbf{X}(\beta^{+} - \beta^{-}) \right\|^2 \\
\text{subject to} & \beta^{+} \succeq 0; \beta^{-} \succeq 0; \sum_{i=1}^{p} \beta^{+(i)} + \sum_{i=1}^{p} \beta^{-(i)} \leq \| \beta^* \|_1. \tag{4.8}
\end{align*}
\]

We follow the outline of Algorithm 1 to debias the constrained LASSO estimator. In step 1, in order to solve (4.1), we use \( \sqrt{\eta(T_K(v) \cap \mathbb{S}^{p-1})} = \sqrt{s \log(ep/s)} \) (see Chandrasekaran et al., 2012, Proposition 3.10), where \( s \) is the number of non-zero coordinates in \( v \). Let \( v_s \) be the projection of \( \hat{\beta} \) onto the set of \( s \)-sparse vectors with \( \ell_1 \)-norm \( \| \beta^* \|_1 \). The optimization (4.1) reduces to an optimization with finite candidates

\[
\hat{s} \leftarrow \arg\min_{s \in [1, \| \beta^* \|_0]} \| \hat{\beta} - v_s \| + \sqrt{\frac{s \log(ep/s)}{n}},
\]

and we find the output of step 1 by choosing \( v = v_{\hat{s}} \). According to Lemma 4.15, the computation of the projection \( v_s \) has a complexity \( O(s) \) (after the entries of \( \hat{\beta} \) have been ordered by magnitude), by greedily taking the largest \( s \) coefficients of \( \hat{\beta} \) and distributing the remaining of the \( \ell_1 \)-norm equally across the \( s \)-coefficients.

**Lemma 4.15.** Let \( S \) be the set of indices of the \( s \) largest in magnitude coordinates of \( \hat{\beta} \), and \( \Lambda \geq 0 \) be a constant. Let \( v_s \) be the projection of \( \hat{\beta} \) onto the set \( T = \{ \beta : \| \beta \|_1 = \Lambda \text{ and } \| \beta \|_0 = s \} \). Then \( v_s \) and satisfies

\[
v_{s}^{(i)} = \begin{cases} 
0, & \text{if } i \notin S \\
\hat{\beta}_i + \text{sign}(\hat{\beta}_i) \frac{\Lambda - \sum_{j \in S} |\hat{\beta}_j|}{s}, & \text{if } i \in S.
\end{cases}
\]

In the above, ties in ordering the coefficients of \( \hat{\beta} \) in magnitude can be broken arbitrarily. Once we obtain the vector \( v \) in step 1, the projection onto the tangent cone \( T_K(v) \) needed in step 2 can be done efficiently by first finding the projection onto its polar cone — the normal cone at \( v \) with respect to the set \( K \) (see Chandrasekaran et al., 2012, eq (9)):

\[
\mathcal{N}_K(v) = \{ w : \langle w, v' - v \rangle \leq 0, v' \in K \}.
\]
Then the projection \( \Pi_{\mathcal{T}_K(v)}(z) = z - \Pi_{\mathcal{N}_K(v)}(z) \) by applying Moreau’s decomposition (Moreau, 1962). Let \( S \) be the set of non-zero coordinates of \( v \). For the set \( K \) equal to the \( \ell_1 \) ball with radius \( \| \beta^* \|_1 \), the normal cone has an explicit form (see Chandrasekaran et al., 2012, eq (60))

\[
\mathcal{N}_K(v) = \{ v' : v'_i = t \cdot \text{sign}(v_i) \text{ for } i \in S; \ |v'_i| \leq t \text{ for } i \notin S, \text{ where } t > 0 \text{ is a constant} \}.
\]

Based on the expression above, the projection of a vector \( z \) onto the normal cone \( \mathcal{N}_K(v) \) can be converted to a one-dimensional convex optimization program with an auxiliary parameter \( t \) (Chandrasekaran et al., 2012, eq (62)) which can be solved by golden section search, e.g. (Kiefer, 1953)

\[
\hat{t} = \arg\min_{t \in [0,\|z\|_{\infty}]} \sum_{i \in S} (z_i - t \cdot \text{sign}(v_i))^2 + \sum_{i \notin S} \text{sign}(z_i)(|z_i| - t)_+ ,
\]

where \((x)_+ = \max\{x, 0\}\). The search interval of \( t \) has an upper bound \( \|z\|_{\infty} \) since the objective function will have a larger value for all \( t > \|z\|_{\infty} \) compared with \( t = \|z\|_{\infty} \). Once \( \hat{t} \) is obtained the projection onto \( \mathcal{N}_K(v) \) is

\[
\Pi_{\mathcal{N}_K(v)}(z) = \begin{cases} \hat{t} \cdot \text{sign}(v_i), & \text{if } i \in S \\ \text{sign}(z_i)(\hat{t} \land |z_i|), & \text{if } i \notin S. \end{cases}
\]

We remark that golden section search can get arbitrarily close to the optimal value, which is good from a computational standpoint. If one would like to obtain the exact solution (which is desirable for theoretical purposes), one can order all \( |z_i| \) values and look for \( t \) in between them. Each problem is a constrained quadratic polynomial so it is easy to optimize. This approach will solve (4.9) precisely. A summary of the debiasing procedure specific for constrained LASSO estimator is given in Algorithm 5.

---

**Algorithm 5** Debias the \( j^{th} \) Coefficient for Constrained LASSO

**Input:** Two equal size partitions \((\bar{X}, \bar{Y})\) and \((\bar{X}, \bar{Y})\); \( \hat{\beta} \) obtained by solving (4.8). \( K = \{ \beta : \| \beta \|_1 \leq \| \beta^* \|_1 \} \).

**Initialize:** Empirical Gram matrix of the second partition \( \Sigma = \frac{1}{n} \bar{X}^\top \bar{X} \).

1. Solve \( \tilde{s} \leftarrow \arg\min_{s \in [1,\|\hat{\beta}\|_0]} \| \hat{\beta} - v_s \| + C\sqrt{\frac{s \log p/s}{n}} \). For each \( s \), the projection \( v_s \) is computed according to Lemma 4.15.

\[
v \leftarrow v_{\tilde{s}} .
\]

2. Run Algorithm 2. Compute \( \Pi_{\mathcal{N}_K(v)}(\cdot) \) by (4.9), (4.10), and apply Moreau’s decomposition to get \( \Pi_{\mathcal{T}_K(v)}(\cdot) \). For \( \Pi_{-\mathcal{T}_K(v)}(\cdot) \) use (2.5). The debiased \( j^{th} \) coefficient \( \hat{\beta}^{(j)}_d \leftarrow v^{(j)} + n^{-1} \bar{\eta}^\top \bar{X}^\top (\bar{Y} - \bar{X}v) \).

---

5 SLOPE and Square-Root SLOPE

In this section we show how our debiasing scheme can be used in SLOPE and square-root SLOPE estimators. It is worthwhile to also mention that even though this section is dedicated to SLOPE
and square-root SLOPE, the same debiasing procedure also works for some other types of estimators whose error rate $\|\hat{\beta} - \beta^*\|$ is tractable. Examples include the LASSO penalized version (Tibshirani, 1996), MCP (Zhang et al., 2010), SCAD penalized estimator (Fan and Li, 2001), elastic net (Zou and Hastie, 2005) etc.

SLOPE was first proposed by Bogdan et al. (2015) as

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \|Y - X\beta\|^2 + \lambda_1|\beta_{#1}| + \lambda_2|\beta_{#2}| + \ldots + \lambda_p|\beta_{#p}|,$$  (5.1)

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, and $|\beta_{#1}| \geq |\beta_{#2}| \geq \ldots \geq |\beta_{#p}|$ are the entries of $\beta$ sorted in a decreasing order in terms of their absolute value. Let $A \geq 2(4 + \sqrt{2})$ be a constant. According to (Bellec et al., 2018b, Corollary 6.2), if one picks $\lambda_i = A\sqrt{\log(2p/i)} / n, i \in [p]$, (5.2) the SLOPE estimator achieves the optimal error rate:

$$\|\hat{\beta} - \beta^*\| \leq C\sigma \sqrt{s \log(ep/s) / n},$$  (5.3)

where $C > 0$ is a constant and $s$ is the number of non-zero coordinates in $\beta^*$.

The square-root SLOPE (Stucky and Van De Geer, 2017) is introduced to alleviate the restriction of knowing $\sigma$ while still achieving the optimal rate (5.3). It estimates $\sigma$ and $\beta$ simultaneously:

$$(\hat{\beta}, \hat{\sigma}) \in \arg\min_{\beta \in \mathbb{R}^p, \sigma > 0} \frac{1}{n}\|Y - X\beta\|^2 + \lambda_1|\beta_{#1}| + \lambda_2|\beta_{#2}| + \ldots + \lambda_p|\beta_{#p}|.$$  (5.4)

Let $A' \geq 4(4 + \sqrt{2})$ be a constant. (Derumigny et al., 2018, Theorem 6.1) shows that if the constraint parameters are picked as

$$\lambda_i = A'\sqrt{\log(2p/i)} / n, i \in [p],$$  (5.5)

the square-root SLOPE will achieve the optimal rate (5.3).

We will now suggest two ways to debias both the SLOPE and square-root SLOPE estimator $\hat{\beta}$. The first assumes knowledge on $\|\beta^*\|_1$, while the second assumes knowledge of an upper bound on sparsity $\|\beta^*\|_0 \leq s^u$.

First, suppose that we know $\|\beta^*\|_1$ and $\beta^*$ is $s$-sparse, but $s$ is not necessarily known. Then the approaches of both SLOPE and square-root SLOPE are identical to how we debias the constrained LASSO problem in Section 4.5, since the convex set $K = \{\beta : \|\beta\|_1 \leq \|\beta^*\|_1\}$ can be used in the same manner as in the constrained LASSO case. In step 1 we find a $v = v_s$ such that $\|\hat{\beta} - v_s\| + \sqrt{s \log(ep/s)} / n$ is the smallest given that $v_s$ is $s$-sparse and $\|v_s\|_1 = \|\beta^*\|_1$. Next we solve step 2 with such a vector $v$ and a convex set $K$.

Second, we consider the case when $\|\beta^*\|_1$ is unknown, but an upper bound on sparsity $\|\beta^*\|_0 \leq s^u$ is available. In this case we do not have a prior knowledge of the convex parameter space $K$ in which $\beta^*$ belongs to. Instead we will construct $K$ and the vector $v$ required in step 1 “from
scratch”. To find a vector $v$ which satisfies the condition in step 1, we propose to solve the following optimization program

$$\arg\max \|v\|_1, \text{ s.t. } \|v - \hat{\beta}\| \leq C \sqrt{\frac{s^u \log(2ep/s^u)}{n}} \text{ and } \|v\|_0 \leq s^u, \quad (5.6)$$

for a sufficiently large constant $C$. Since the function $s \mapsto s \log(2ep/s)$ is increasing in $s$, $\beta^*$ is a feasible point when $C$ is sufficiently large. Theorem 5.2 proves that the solution $v$ of the above optimization program (5.6) satisfies the condition in step 1 with the set

$$K = \{\beta : \|\beta\|_1 \leq \|v\|_1\}. \quad (5.7)$$

Notice that since $\beta^*$ is a feasible point of (5.6) with a proper choice of $C$, $v$ also satisfies $\|v\|_1 \geq \|\beta^*\|_1$ which implies $\beta^* \in K$. In order for us to state our next result we need to give a definition from (Bellec et al., 2018b, Page 10).

**Definition 5.1** (Weighted Restricted Eigenvalue (WRE) condition). For a design matrix $X \in \mathbb{R}^{n \times p}$ satisfying $\|Xe^{(j)}\| \leq \sqrt{n}$ for all $j \in [p]$ define

$$\vartheta(s, c_0) = \min_{\delta \in \{\delta : \sum_{j=1}^p \lambda_j/|\delta_{\delta,j}| \leq (1+c_0)||\delta||/\sqrt{\sum_{j=1}^p \lambda_j^2}, \delta \neq 0\}} \frac{1}{\sqrt{n}} \frac{\|X\delta\|}{||\delta||},$$

where $\lambda_j$ are given in (5.2) (or equivalently in (5.5)). A design matrix $X$ as above is said to satisfy WRE if $\vartheta(s, c_0) > 0$.

The next theorem will condition on the event that $X$ (the design matrix from the first split of the data) satisfies the WRE for $s^u$ and $c_0 = 3$ for SLOPE and $c_0 = 20$ for square-root SLOPE.

**Theorem 5.2.** Consider the same setting as Theorem 3.1. Suppose $\|\beta^*\|_0 \leq s^u$. Condition on the event that the matrix $X$ satisfies the WRE with $\vartheta^* := \vartheta(s^u, 3)$ for SLOPE and $\vartheta^* := \vartheta(s^u, 20)$ for square-root SLOPE. With a proper choice of $C \geq \frac{\vartheta^*}{\vartheta^*}$ satisfying $C\frac{s^u \log(ep/s^u)}{\sqrt{n}} = o(1)$, for $\hat{\beta}$ as a SLOPE estimator obtained via (5.1) or a square-root SLOPE estimator obtained via (5.4), the solution $v$ of (5.6), and the set $K = \{\beta : \|\beta\|_1 \leq \|v\|_1\}$ satisfy the condition needed in step 1 of Algorithm 1.

**Remark 5.3.** We now comment on the condition that $X$ satisfies the WRE with $\vartheta(s^u, c_0)$ for $c_0 = 3$ or $c_0 = 20$. By Theorem 8.3 of Bellec et al. (2018b) we know that for a large class of data generating mechanisms (including Gaussian and bounded mean-zero $X_i$ for $i \in [n]$) if $\Sigma$ has bounded from below by $\kappa > 0$ eigenvalue, and in addition $\max_i \Sigma_{ii} \leq \frac{1}{2}$ then if $n \gtrsim \frac{(1+c_0)^2}{\kappa^2} s^u \log(2ep/s^u)$ the matrix $X$ will satisfy WRE with $s^u$ and $c_0$ with $\vartheta(s^u, c_0) = \kappa/\sqrt{2}$ with high probability. It follows that when $\sigma$ is fixed, $C \geq \frac{\sqrt{2\kappa}}{\kappa}$ suffices to meet the requirements in Theorem 5.2. This is surely satisfied if one picks $C \gg 1$. Below we give an example of such a choice for $C$.

From the proof of Theorem 5.2 it becomes evident that in principle, we can select any small enough $C > \overline{C} \sigma$ in (5.3) since that will ensure that $\beta^*$ is a feasible point in (5.6). One might directly analyze an upper bound on $\overline{C}$ according to the high probability upper bounds on $\|\hat{\beta} - \beta^*\|$ given in (Bellec et al., 2018b, Corollary 6.2) and (Derumigny et al., 2018, Corollary 6.2). However, such an upper bound on $\overline{C}$ requires finding weighted restricted eigenvalues and may not be easily
computable. Here we suggest an alternative way to obtain a slightly larger \( C \) for the debiasing purpose. This is possible under the assumptions of Remark 5.3. We claim that \( C \) can be picked as

\[
C \sim \left( \frac{\sqrt{n}}{s^u \log(ep/s^u)} \right)^\gamma \quad \text{where } 0 < \gamma < 1 \text{ is a small number.} \tag{5.8}
\]

In this way, if \( s^u = o(\sqrt{n}/\log(ep/s^u)) \), the order of \( C \) in (5.8) is slightly larger than the constant in (5.3) which is \( O(1) \) (assuming \( \sigma = O(1) \)) under the assumptions of Remark 5.3. Thus \( \beta^* \) is guaranteed to be a feasible point of (5.6). At the same time, \( C \) is only moderately large so that \( \overline{w}(T_K(v) \cap S_p^{-1})\|v - \beta^*\| = o_p(1) \) still holds under the same assumptions as in Theorem 5.2. This is because in the proof of Theorem 5.2 we establish that with high probability

\[
\overline{w}(T_K(v) \cap S_p^{-1})\|v - \beta^*\| \lesssim C \frac{\log(ep/s^u)}{\sqrt{n}} \sim \left( \frac{s^u \log(ep/s^u)}{\sqrt{n}} \right)^{1-\gamma} = o(1).
\]

After picking a proper \( C \), there are no obstacles to compute a \( v \) in step 1 since the optimization program (5.6) actually has an analytical solution as shown in Lemma 5.4.

**Lemma 5.4.** The solution of (5.6) is

\[
v_{\#i} = \begin{cases} 
\tilde{\beta}_{\#i} + \text{sign}(\tilde{\beta}_{\#i})c, & \text{if } i = 1, \ldots, s^u \\
0, & \text{otherwise.}
\end{cases}
\]

where \( c = \sqrt{\left( C^2 s^u \log(2ep/s^u) - \sum_{i=s^u+1}^{p} \hat{\beta}_{\#i}^2 \right)/s^u} \), and ties in \( \tilde{\beta}_{\#i} \) are broken arbitrarily, and with a slight abuse of notation we assign the same index for \( v_{\#i} \) in \( v \) as \( \tilde{\beta}_{\#i} \) has in \( \hat{\beta} \).

Notice that \( C \) should be selected so that we are able to compute \( c \) as a positive real number, hence it should satisfy \( C \geq \sqrt{\frac{n \sum_{i=s^u}^{p} \hat{\beta}_{\#i}^2}{s^u \log(2ep/s^u)}} \). Observe that this does not imply that \( C \) is “too large”. From (5.1) we know that \( \sqrt{\|eta_{S_\ast} - \beta_{S_\ast, i}^0\|^2 + \|eta_{S_\ast, i}^0\|^2} \leq \frac{C \sigma \sqrt{s^u \log(ep/s^u)}}{n} \), where \( S_\ast \) denotes the support of \( \beta^* \). Next since \( s^u \geq s \) it follows that \( \sqrt{\sum_{i=s^u+1}^{p} \hat{\beta}_{\#i}^2} \leq \|\beta_{S_\ast, i}^0\| \), which shows that if \( C > C \sigma \) the condition will be met. After one finds \( v \) in step 1, one can compute the auxiliary vector \( \tilde{\eta} \) in step 2 based on \( v \) and \( K = \{\beta : \|\beta\|_1 \leq \|v\|_1\} \), and then use \( \tilde{\eta} \) to construct the debiased estimator \( \tilde{\beta}_d \) and the confidence interval as (3.1). When constructing the confidence interval, we estimate \( \sigma \) via \( \hat{\sigma} = \sqrt{n^{-1} \sum_{i \in [n]} (Y_i - X_i^\top \hat{\beta})^2} \) on the first sample split. The following Lemma 5.5 coupled with Theorem 3.4 together show that we are able to get a consistent estimator of \( \sigma \).

**Lemma 5.5.** Consider the same setting as Theorem 5.2 where \( \hat{\beta} \) is a SLOPE or square-root SLOPE estimator. Then under the conditions of Remark 5.3, Theorem 3.4 applies with

\[
\delta \asymp \sigma \kappa^{-1} \sqrt{s^u \log(2ep/s^u)}.
\]

Lemma 5.5 establishes that it is possible to consistently estimate \( \sigma \), and therefore we can construct confidence intervals as in (3.2). We end this section with two remarks regarding the choice of \( s^u \) and what “classical” debiasing methods can achieve in the SLOPE, or square-root SLOPE problems.
Remark 5.6. Since \( s^u \geq s \), assuming \( s^u \log(ep/s^u) = o(\sqrt{n}) \) implies that \( s \log(ep/s) = o(\sqrt{n}) \) for the true sparsity \( s \). By the work of Cai et al. (2017) we know the latter condition is nearly necessary in the case of sparse linear regression with unknown covariance. In fact Cai et al. (2017) show that the length of the confidence interval is \( \geq \max\left\{ \frac{1}{\sqrt{n}}, \frac{s \log(ep/s)}{n} \right\} \). Thus if \( s \log(ep/s) = O(\sqrt{n}) \) interval length of the order of \( \frac{1}{\sqrt{n}} \) is possible. However, in practice it is often assumed that \( s \log(ep/s) = o(\sqrt{n}) \) in order to achieve an exact asymptotic \((1 - \alpha)\)-level confidence interval. We now provide some guidance on selecting \( s^u \). In principle it is difficult if not impossible to estimate an upper bound on \( s \) from the data. However, in order for the debiasing to work we do need \( s \log(ep/s) = o(\sqrt{n}) \). If the practitioner has prior knowledge on the precise rate \( r_n := \frac{s \log(ep/s)}{\sqrt{n}} \), the practitioner can select any \( s^u \) such that \( \frac{s^u \log(ep/s^u)}{\sqrt{n}} = \sqrt{r_n} \), e.g. and this will work asymptotically.

On the other hand, if information on \( r_n \) is not available but it is known that \( \frac{s \log(ep/s)}{\sqrt{n}} = o(1) \), the practitioner may opt for devising a slightly conservative confidence interval, by selecting \( s^u \) such that \( \frac{s^u \log(ep/s^u)}{\sqrt{n}} = c \) for some small constant \( c \). It is not too hard to see that in such a setting, the term \(|\Delta_j|\) from Theorem 3.1 will be asymptotically smaller than

\[ |\Delta_j| \leq \sqrt{n} \rho c ||\beta^* - v||, \]

where \( \rho \) is the tuning parameter from (2.1) of Algorithm 1. Now by the triangle inequality \( ||\beta^* - v|| \leq ||\beta^* - \hat{\beta}|| + ||\hat{\beta} - v|| \leq 2C \sqrt{s^u \log(2ep/s^u)} \) since \( \beta^* \) is a feasible point of (5.6). Set \( K := 2C \rho \sqrt{s^u \log(ep/s^u)} \), where \( C \) (note that any fixed constant \( C \) here will do since \( s^u \gg s \)). is the constant from (5.6). Therefore the confidence interval from (3.2) widened by \( \pm \frac{K}{\sqrt{n}} \) will be a valid \( \frac{1}{\sqrt{n}} \)-confidence interval of \( \beta^{(j)} \).

Remark 5.7. Theorem 5.2 and Remark 5.6 point out that our debiasing algorithm works for SLOPE as long as \( s = o(\sqrt{n}/\log ep/s) \). Clearly this is less stringent than the condition \( s = o(\sqrt{n}/(\log ep/s)^{3/2}) \). Such a condition appears necessary if one opts for applying previous debiasing algorithms and their analysis such as the one proposed by (Javanmard and Montanari, 2014, Algorithm 1). To see why the condition \( s = o(\sqrt{n}/(\log ep/s)^{3/2}) \) arises, the reader is referred to (Javanmard et al., 2018, eq (9)) which summarizes well the standard argument for the analysis of why debiasing works. It relies on an \( \ell_1 - \ell_\infty \) Hölder’s inequality. While the SLOPE or square-root SLOPE do not have a direct \( \ell_1 \) guarantee for their \( \hat{\beta} \) estimates, a sub-optimal guarantee may be easily derived from (Bellec et al., 2018b; Derumigny et al., 2018). It is simple to see that

\[ \sigma ||\hat{\beta} - \beta^*||_1/\sqrt{n} \lesssim ||\hat{\beta} - \beta^*||_* \lesssim \sigma^2 s \log(ep/s)/n, \]

where \( ||v||_* = \sum_{j \in [p]} \lambda_j |v_{#j}| \), where \( \lambda_j \) are as in (5.2). In contrast, in the LASSO case one may bound \( ||\hat{\beta} - \beta^*||_1 \lesssim \sigma s \sqrt{\frac{\log(p)}{n}} \) (Wainwright, 2019, Section 7). One can see that SLOPE has an extra \( \sqrt{\log(ep/s)} \) factor in the \( \ell_1 \)-bound in comparison with LASSO, hence the extra \( \sqrt{\log(ep/s)} \) factor in the condition \( s = o(\sqrt{n}/(\log ep/s)^{3/2}) \).

In the following two subsections we give the detailed procedures about how to debias SLOPE and square-root SLOPE estimator, as specific instances of Algorithm 1.
5.1 Debiasing Algorithm for SLOPE

We start by briefly summarizing how to solve the SLOPE $\hat{\beta}$ in (5.1). The reader is encouraged to read the full details of the implementation which was first described in Bogdan et al. (2015). The SLOPE has a non-differentiable objective function, which can be solved by proximal gradient descent. A detailed introduction of the proximal gradient methods can be found in (Nesterov, 2003, Chapter 2). The basic idea is: the objective function in (5.1) can be written as the sum of a convex differentiable function $f_1(\beta) = \frac{1}{n} \| \tilde{Y} - \bar{X} \beta \|^2$ and a convex non-differentiable function $f_2(\beta) = \lambda_1|\beta_{#1}| + \lambda_2|\beta_{#2}| + \ldots + \lambda_p|\beta_{#p}|$. For a convex optimization program whose objective function can be written as $f(\beta) = f_1(\beta) + f_2(\beta)$, where $f_1$ is differentiable but $f_2$ is not, each step of the proximal gradient method can be written as

$$\beta_{n+1} = \text{prox}_{h_n} \left( \beta_n - h_n \nabla f_1(\beta_n) \right),$$

where $h_n$ is the step size, and $\text{prox}_{h_n}(\cdot)$ is the proximal mapping defined as

$$\text{prox}_{h}(x) = \text{argmin}_{z} \frac{1}{2h} \| x - z \|^2 + f_2(z).$$

One can see that the proximal mapping in (5.9) forces the new candidate $\beta_{n+1}$ to stay close to the gradient update of $f_1$, and also makes $f_2$ small. The proximal mapping can be solved with the PAVA algorithm for isotonic regression. See (Bogdan et al., 2015, Algorithm 3) for details.

After solving $\hat{\beta}$, we debias it. The vector $v$ in step 2 can be computed analytically by Lemma 5.4 with $C$ picked according to (5.8), and $K$ is constructed as $K = \{ \beta : \| \beta \|_1 \leq \| v \|_1 \}$.

Algorithm 6 Debias the $j$th Coefficient in SLOPE

**Input:** Two equal size partitions $(\bar{X}, \tilde{Y})$ and $(\hat{\bar{X}}, \hat{\tilde{Y}})$; $\hat{\beta}$ as a SLOPE estimator. $s^u$ upper bound on $s$, $C$ a sufficiently large tuning parameter.

**Initialize:** Empirical Gram matrix of the second partition $\hat{\Sigma} = \frac{1}{n} \hat{X}^\top \hat{X}$.

1. $c \leftarrow \sqrt{\left( C^2 s^u \log 2ep/s^u \right) - \sum_{i=s^u+1}^p i \beta_{#i}^2 / s^u}$, $v \leftarrow (0, \ldots, 0)$
   Assign $v_{#i} = \beta_{#i} + \text{sign}(\beta_{#i}) c$ for $i = 1, \ldots, s^u$

2. Run Algorithm 2. Compute $\Pi_{N_k(\psi)}(\cdot)$ by (4.9), (4.10), and apply Moreau’s decomposition to get $\Pi_{r_k(\psi)}(\cdot)$. For $\Pi_{r_k(\psi)}(\cdot)$ use (2.5)
   The debiased $j$th coefficient $\hat{\beta}_{d}^{(j)} \leftarrow v^{(j)} + n^{-1} \hat{\eta}^\top \hat{X}^\top (\hat{\tilde{Y}} - \hat{\bar{X}} v)$.

5.2 Debiasing Algorithm for Square-Root SLOPE

To solve the square-root SLOPE, the joint optimization (5.4) can be solved by alternatively minimizing $\beta$ and $\sigma$: the minimization in $\beta$ is the same as SLOPE in (5.1) with parameters $\sigma \lambda_1, \ldots, \sigma \lambda_p$, and after that setting $\hat{\sigma}$ to $\hat{\sigma} = \| \tilde{Y} - \bar{X} \beta \| / \sqrt{n}$. Details can be found in (Stucky and Van De Geer,
2017, Algorithm 1) and (Derumigny et al., 2018, Algorithm 2). The debiasing algorithm for square-root SLOPE is the same as Algorithm 6.

6 Non-Gaussian Errors

In this section we modify our Algorithm 1 to accommodate for sub-Gaussian noise. The modified procedure is presented in Algorithm 7. Algorithm 7 requires an additional condition in step 1, namely \( \|v - \beta^*\|\sqrt{\log n} = o_p(1) \). We view this as a fairly mild assumption, which in most relevant practical cases is dominated by the assumption \( \|v - \beta^*\|\mathbb{E}(\mathcal{W}(v) \cap S^{p-1}) = o_p(1) \). In step 2 of Algorithm 7, we have added an additional \( \ell_\infty \) constraint to the optimization. Observe that the modified program in step 2 is still a convex program, and can be solved by subgradient descent as before.

**Algorithm 7** Debias the \( j \)th Coordinate of A Non-Ordinary Least Squares Estimator

**Input:** Two equal size partitions \((\tilde{X}, \tilde{Y})\) and \((\tilde{X}, \tilde{Y})\), \( \hat{\beta} \) obtained using \((\tilde{X}, \tilde{Y})\).

**Initialize:** Empirical Gram matrix of the second partition \( \hat{\Sigma} = \frac{1}{n} \tilde{X}^\top \tilde{X} \).

1. Using the first data split, find a convex set \( K \) and a vector \( v \), such that: \( v, \beta^* \in K \) with high probability, and \( \|v - \beta^*\| \max\{\mathbb{E}(\mathcal{W}(v) \cap S^{p-1}), \sqrt{\log n} \} = o_p(1) \).

2. The debiased \( j \)th coefficient \( \hat{\beta}_d^{(j)} \leftarrow e^{(j)\top} v + n^{-1} \hat{\eta}^\top \tilde{X}^\top (\tilde{Y} - \tilde{X} v) \), where \( \hat{\eta} \) is computed by

\[
\hat{\eta} \leftarrow \arg\min_{\eta} \| \hat{\Sigma}^{\frac{1}{2}} \eta \| \text{ subject to } \sup_{u \in \mathcal{T}_K(v) \cap S^{p-1}} |(\eta^\top \hat{\Sigma} - e^{(j)\top}) u| \leq \rho \frac{\mathbb{E}(\mathcal{W}(v) \cap S^{p-1})}{\sqrt{n}}, \\
\|\tilde{X} \eta\|_\infty \leq \rho' \sqrt{\log n},
\]

for some sufficiently large tuning parameters \( \rho > 0, \rho' > 0 \).

To show that the new optimization program has a feasible point and consequently a non-empty interior, we evaluate the constraints at the point \( \eta = \Sigma^{-1} e^{(j)} \). By using a similar argument to that of (Javanmard and Montanari, 2014, p. 33) we are able to show that \( \|\tilde{X} \Sigma^{-1} e^{(j)}\|_\infty \leq \sqrt{\log n} \), and the argument of non-empty interior is similar to how we prove Lemma 2.4. Details are given in Lemma 6.1 and its proof.

**Lemma 6.1.** Suppose that \( \tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_n)^\top \) where every observation \( \tilde{X}_i \) is a zero-mean bounded or a zero-mean Gaussian random variable with covariance matrix \( \Sigma \), and the eigenvalues of \( \Sigma \) are bounded from above and below. For a sufficiently large constant \( \rho' > 0 \), the set

\[
\{ \eta : \|\tilde{X} \eta\|_\infty \leq \rho' \sqrt{\log n} \} \cap \left\{ \eta : \sup_{u \in \mathcal{T}_K(v) \cap S^{p-1}} |(\eta^\top \hat{\Sigma} - e^{(j)\top}) u| \leq \rho \frac{\mathbb{E}(\mathcal{W}(v) \cap S^{p-1})}{\sqrt{n}} \right\},
\]

has a non-empty interior.
Solving the optimization in step 2 of Algorithm 7 is similar to solving the optimization in step 2 of Algorithm 1, since both of them are convex programs with inequality constraints. The only difference is that the former has two constraints while the latter has only one. According to (Boyd et al., 2003, Section 7), the idea of solving optimization with multiple inequality constraints is: if the current point is feasible, subgradient descent is applied to the objective function; if the current point is not feasible, we pick any one of the violated constraints, and apply subgradient descent to it. Define

\[ \psi'(\eta) = \|X\eta\|_\infty - \rho' \sqrt{\log n}. \]

The second constraint in step 2 of Algorithm 7 can be written as \( Q' = \{ \eta : \psi'(\eta) \leq 0 \} \). To this end we remind the reader of the shorthand notations \( \psi(\eta) \) from (2.2), and \( Q = \{ \eta : \psi(\eta) \leq 0 \} \). The sequence \( \{\eta_n\} \) is generated as in (2.3), where \( g_n \) is the gradient of the objective function if \( \eta_n \in Q \) and \( \eta_n \in Q' \); is a subgradient of \( \psi(\eta_n) \) if \( \eta_n \notin Q \); otherwise is a subgradient of \( \psi'(\eta_n) \) if \( \eta_n \in Q \) and \( \eta_n \notin Q' \). In the following Lemma 6.2 we give the expression of a subgradient of \( \psi'(\eta_n) \).

**Lemma 6.2.** Let \( i^* = \arg\max_{i \in [n]} [X_i^\top \eta_n] \). Then \( \nabla \psi'(\eta_n) = \text{sign}(X_i^\top \eta_n)X_i^\top \) is a subgradient of \( \psi'(\eta_n) \).

After adding the new constraint \( \psi'(\eta) \leq 0 \), Algorithm 2 is modified to Algorithm 8. In terms of the convergence of Algorithm 8, it also takes \( n = O(1/\epsilon^2) \) iterations to get an \( \epsilon \)-suboptimal solution i.e. \( \|\hat{\Sigma}^z \eta_n\| - \|\hat{\Sigma}^z \eta^*\| \leq \epsilon \). The proof of Lemma 2.6 will remain unchanged since \( \psi'(\eta_n) \) is a Lipschitz function of \( \eta_n \) (since with probability 1, \( \sup_{i \in [n]} \|\bar{X}_i\| < \infty \)).

**Algorithm 8 Solve the Optimization in Step 2 of Algorithm 7**

**Input:** The convex set \( K \), the vector \( v \) from step 2, empirical Gram matrix of the second partition \( \hat{\Sigma} = \frac{1}{n} \bar{X}^\top \bar{X} \).

**Initialize:** \( \eta_1 \)

Run for sufficiently long time:

Compute \( P_+ = \Pi_{\mathbb{T}_K(v)}(\hat{\Sigma} \eta_n - e^{(j)}) \), \( P_- = \Pi_{-\mathbb{T}_K(v)}(\hat{\Sigma} \eta_n - e^{(j)}) \).

if \( \max\{\|P_+\|,\|P_-\|\} \leq \frac{\rho(T_K(v)\sqrt{n})}{\sqrt{n}} \) & \( \|X\eta_n\|_\infty \leq \rho' \sqrt{\log n} \)

\( \eta_{n+1} \leftarrow \eta_n - h_n \frac{\Sigma \eta_n}{\|\Sigma \eta_n\|} \)

else max\{\( \|P_+\|,\|P_-\|\) \( > \frac{\rho(T_K(v)\sqrt{n})}{\sqrt{n}} \)): \( \phi_0(\eta_n) \leftarrow P_+ / \|P_+\| \), \( \phi_1(\eta_n) \leftarrow P_- / \|P_-\| \).

\( \eta_{n+1} \leftarrow \eta_n - h_n \Sigma \phi_1((\eta_n^\top \Sigma - e^{(j)^\top})(\phi_0(\eta_{n-1}) - \phi_1(\eta_{n-1})) < 0) \) \( \eta_n \)

else:

\( \eta_{n+1} \leftarrow \eta_n - h_n \text{sign}(X_i^\top \eta_n)X_i^\top \), where \( i^* = \arg\max_{i \in [n]} |X_i^\top \eta_n| \)

\( \hat{\eta} \leftarrow \eta_{n+1} \)

We now state a result which establishes the confidence interval for non-Gaussian noise.
Theorem 6.3. Consider a linear model in (1.2) and with sub-Gaussian errors $\varepsilon_i$. Suppose the
eigenvalues of $\Sigma$ are bounded from both above and below. Recall that $\hat{\beta}^{(j)}_d$ is the debiased $j$th
coefficient obtained by Algorithm 7. Let $a_n = o(1)$ be any slowly converging to 0 rate such that
\[
\frac{1}{a_n} = o\left(\frac{n}{\log n}\right),
\]
and let $c$ be sufficiently large constant satisfying $c > C' \sqrt{\log n} \sqrt{[\|\beta^* - v\| \sqrt{\log n}] / a_n} = o_p(1)$, where $C'$ is a universal constant. Then the confidence interval
\[
\left(\hat{\beta}^{(j)}_d - z_{\frac{\alpha}{2}} \sigma_n \left(\|\hat{\Sigma}^{-\frac{1}{2}} \hat{\eta}\| \vee c\right), \hat{\beta}^{(j)}_d + z_{\frac{\alpha}{2}} \sigma_n \left(\|\hat{\Sigma}^{-\frac{1}{2}} \hat{\eta}\| \vee c\right)\right),
\]
contains $\beta^*$ with probability at least $1 - \alpha$ asymptotically.

It is worthwhile to mention that even though the length of the confidence interval (6.2) is always
of the order $O(1/\sqrt{n})$, when the quantity $\|\hat{\Sigma}^{-\frac{1}{2}} \hat{\eta}\|$ is very small such that
\[
\|\hat{\Sigma}^{-\frac{1}{2}} \hat{\eta}\| \lesssim \sqrt{\log n / \sqrt{([\|\beta^* - v\| \sqrt{\log n}] \vee a_n)}},
\]
as can be seen from our proof, the debiased estimator $\hat{\beta}^{(j)}_d$ actually converges faster than the rate
$1/\sqrt{n}$. In this case the confidence interval (6.2) is still valid, but not very efficient. And contrarily if
\[
\|\hat{\Sigma}^{-\frac{1}{2}} \hat{\eta}\| \gtrsim \sqrt{\log n / \sqrt{([\|\beta^* - v\| \sqrt{\log n}] \vee a_n)}},
\]
then a Central Limit Theorem applies to $\sqrt{n} (\hat{\beta}^{(j)}_d - \beta^*(j))$, and the variance would be exactly
$\sigma^2 \|\hat{\Sigma}^{-\frac{1}{2}} \hat{\eta}\|$. Thus the confidence interval is tight when $\|\hat{\Sigma}^{-\frac{1}{2}} \hat{\eta}\| \geq c$, and is slightly loose when $\|\hat{\Sigma}^{-\frac{1}{2}} \hat{\eta}\| < c$ since we are using a slightly larger variance.

Finally, we can also consistently estimate $\sigma$ as in Theorem 3.4 whose proof does not rely on the
Gaussian assumption on the noise.

7 Simulations

Now we examine the performance of the proposed debiasing procedure for the monotone cone
regression, positive monotone cone regression, non-negative least squares, constrained LASSO,
SLOPE and square-root SLOPE cases. We pick a single coordinate to debias. In all the experiments
of this section, the last coordinate of the signal vector is picked.

In terms of the construction of true coefficient $\beta^*$, for the monotone cone case, $\beta^*$ consists of
-1 and 1, where the first 70% coordinates are -1, and the remaining 30% are 1. For the positive
monotone cone case, the true coefficient $\beta^*$ consists of 0 and 1, where the first 70% coordinates are 0,
and the remaining 30% are 1. For the non-negative least squares case, we generate $\beta^*$ such that each
coordinate is $\max\{N(0, 3), 0\}$. For the LASSO case, $\beta^*$ consists of 0 and 1, where the first 99.5% of
the coordinates are 0, and the remaining 0.5% are 1. For the SLOPE and square-root SLOPE
cases, the first 99.5% of the true $\beta^*$ are 0, the remaining coordinates are formed by an increasing
series of integers with step size 1 starting from 1. In terms of the sample size $n$ and dimension $p$,
we use $n = 100, p = 100$ for the monotone cone and positive monotone cone cases. Note that this
conforms to our assumption that $w^2 (T_K(\beta^*) \cap \mathbb{S}^{p-1}) = o(\sqrt{n})$ since the vector $\beta^*$ is comprised only
of 2 constant pieces. For the non-negative least squares case, we pick \( n = 1000, p = 50 \) in order to make \( w^2(T_K(\beta^*) \cap \mathbb{S}^{p-1}) \approx p \) (see Lemma 4.12) approximately comparable to \( \sqrt{n} \). For LASSO, SLOPE and square-root SLOPE, we use \( n = 1000, p = 1000 \). Coupled with the small proportion of non-zero coordinates in \( \beta^* \) this guarantees that \( w^2(T_K(\beta^*) \cap \mathbb{S}^{p-1}) \approx s \log(ep/s) \) is smaller than \( \sqrt{n} \), where \( s \) denotes the sparsity of \( \beta^* \).

The predictors \( X \) are drawn from a mean-zero Gaussian distribution. In order to verify the compatibility of this debiasing procedure with different types of input data, three different covariance matrices \( \Sigma \) are used to generate different Gaussian distributions: an identity matrix, a random matrix with bounded eigenvalues, and a Toeplitz matrix whose \( i,j \)-th element is \( \rho^{|i-j|} \) where \( \rho \in (0,1) \) (we use \( \rho = 0.4 \)).

For each type of the predictor and covariance matrix \( \Sigma \), we generate the data \( X, Y, \beta^* \), we obtain the original estimator \( \hat{\beta} \), and perform Algorithm 1 to debias the last coordinate. The experiment is repeated 100 times. According to Theorem 3.1, for any coordinate \( j \), the debiased estimator \( \hat{\beta}_d^{(j)} \) should satisfy \( \sqrt{n}(\hat{\beta}_d^{(j)} - \beta^{* (j)}) \sim N(0, \sigma^2 \hat{\eta} \Sigma \hat{\eta}) \), which doesn’t necessarily hold for the non-debiased estimator \( \hat{\beta}^{(j)} \). In Figure 1, we examine the distribution of \( \frac{\sqrt{n}(\hat{\beta}_d^{(j)} - \beta^{* (j)})}{\hat{\eta} \Sigma \hat{\eta}} \) for \( j = p \), by plotting them against the standard Gaussian distribution in a Q-Q plot.

We can see from those plots that \( \frac{\sqrt{n}(\hat{\beta}_d^{(j)} - \beta^{* (j)})}{\hat{\eta} \Sigma \hat{\eta}} \) appears pretty close to \( N(0,1) \), which is not true for \( \frac{\hat{\beta}^{(j)} - \beta^{* (j)}}{sd(\beta^{* (j)})} \) in terms of both bias and variance difference. It is worth pointing out that for the SLOPE and square-root SLOPE cases, although the undebiased estimators points appear to align well on the Q-Q plot they are not centered at the correct value. Figure 1 only reports the results in the setting \( \Sigma = I \). Similar plots for the bounded eigenvalue and Toeplitz population covariance matrix settings are attached in the appendix.

### 8 Future Work

In this paper we proposed a novel abstract procedure for debiasing linear regressions. Our method is able to perform inference for some constrained and regularized problems for which inferential tools were not previously available.

An interesting further question to explore is whether we can prove lower bounds on confidence intervals obtained in the above way such as the work of Cai et al. (2017). In other words are the conditions \( w^2(T_K(\nu') \cap \mathbb{S}^{p-1}) = o(\sqrt{n}) \) and \( \|\nu' - \beta^*\| = o(1/\sqrt{n}) \) also necessary for the unknown covariance case?

Another open question is debiasing the constrained least squares using (4.2) in the unknown covariance case but without resorting to sample splitting. Our conjecture is that sample splitting is not required, but a proof of this fact will require carefully isolating the dependency of \( \hat{\beta} \) on \( X \). For this purpose, it may be necessary to employ a slightly different debiasing scheme as the one undertook by Bellec and Zhang (2019a).

Furthermore, the question of how can one solve the second optimization program if projecting on \( T_K(\nu) \) is hard is also interesting. In particular we are curious whether it is possible to apply interior point methods.

Finally, our main procedure requires us to split the data. Inevitably, this results in a loss of efficiency. One way to correct for that is to use a cross-fitted estimator as in Chernozhukov et al.
Figure 1: The Q-Q Plot of \( \frac{\sqrt{n}(\hat{\beta}^{(j)} - \beta^*(j))}{\hat{\sigma}||\hat{\Sigma}\hat{\eta}||} \) and \( \frac{\hat{\beta}^{(j)} - \beta^*(j)}{sd(\hat{\beta}^{(j)} - \beta^*(j))} \) where \( j = p \), against Standard Normal in the Identity Population Matrix Setting. The Upper Row: the scaled Difference between the Undebiased Estimator and the True Coefficient; the Lower Row: the scaled Difference between the Debiased Estimator and the True Coefficient. From Left to Right: Monotone Cone Regression, Positive Monotone Cone Regression, Non-negative Regression, LASSO, SLOPE, Square-root SLOPE.

(2018); Eftekhari et al. (2021). It is unclear to us at the moment whether this strategy will work in our case as the influence functions of the estimators on the two samples may not be independent.
A Additional Simulation Results

All the code for experiments can be found in: https://github.com/Pythongoras/debiascvgV2.

Figure 2: The Q-Q Plot of \( \frac{\sqrt{n}(\hat{\beta}^{(j)} - \beta^*(j))}{\hat{\sigma} \sqrt{\sum \eta}} \) and \( \frac{\hat{\beta}^{(j)} - \beta^*(j)}{\text{sd}(\hat{\beta}^{(j)} - \beta^*(j))} \) where \( j = p \), against Standard Normal in the Bounded Eigenvalue Population Matrix Setting. The Upper Row: the scaled Difference between the Undebiased Estimator and the True Coefficient; the Lower Row: the scaled Difference between the Debiased Estimator and the True Coefficient. From Left to Right: Monotone Cone Regression, Positive Monotone Cone Regression, Non-negative Regression, LASSO, SLOPE, Square-root SLOPE.

Figure 3: The Q-Q Plot of \( \frac{\sqrt{n}(\hat{\beta}^{(j)} - \beta^*(j))}{\hat{\sigma} \sqrt{\sum \eta}} \) and \( \frac{\hat{\beta}^{(j)} - \beta^*(j)}{\text{sd}(\hat{\beta}^{(j)} - \beta^*(j))} \) where \( j = p \), against Standard Normal in the Toeplitz Population Matrix Setting. The Upper Row: the scaled Difference between the Undebiased Estimator and the True Coefficient; the Lower Row: the scaled Difference between the Debiased Estimator and the True Coefficient. From Left to Right: Monotone Cone Regression, Positive Monotone Cone Regression, Non-negative Regression, LASSO, SLOPE, Square-root SLOPE.
B Preliminaries Used in the Proofs

We present several preliminary definitions and results which are needed in the proofs of the future sections.

**Definition B.1.** For a random variable $X \in \mathbb{R}$, define its $\psi_\ell$ norm by
\[
\|X\|_{\psi_\ell} = \sup_{p \geq 1} p^{-1/\ell} (\mathbb{E}|X|^p)^{1/p}.
\]
for $\ell \in \{1, 2\}$. For a random vector $X \in \mathbb{R}^d$ define
\[
\|X\|_{\psi_\ell} = \sup_{v \in \mathbb{R}^{d-1}} \|X^T v\|_{\psi_\ell}.
\]

Next is Gordon’s Escape Through Mesh which bounds the restricted operator norm of a Gaussian matrix over a convex set. Details can be found in (Gordon, 1988, Theorem A).

**Lemma B.2.** (Gordon’s Escape Through Mesh) Let $K \subset \mathbb{R}^n$ be a convex cone and $X$ be an $n \times p$ standard Gaussian matrix. Then for every $t \geq 0$,
\[
\mathbb{P}\left\{ \sup_{u \in K \cap \mathbb{S}^{p-1}} \|Xu\| \geq \sqrt{n} + w(K \cap \mathbb{S}^{p-1}) + t \right\} \leq e^{-\frac{t^2}{2}},
\]
\[
\mathbb{P}\left\{ \inf_{u \in K \cap \mathbb{S}^{p-1}} \|Xu\| \leq \sqrt{n} - 1 - w(K \cap \mathbb{S}^{p-1}) - t \right\} \leq e^{-\frac{t^2}{2}}.
\]

The next result Lemma B.4 gives an upper bound of the estimation error of the convex constrained least squares, which is an analogy of Corollary 2.6 in Neykov (2019). We give a proof here since the proof of Corollary 2.6 is eliminated in Neykov (2019). The proof is similar as the proof of (Neykov, 2019, Lemma 2.3). Lemma B.3 is an intermediate result needed in the proof of Lemma B.4.

**Lemma B.3.** (Neykov, 2019, Lemma A.1) For any $v \in K$ we have the following inequality
\[
\frac{1}{\sqrt{n}} \|X(\hat{\beta} - v)\| \leq \frac{4}{\sqrt{n}} \|X(\beta^* - v)\| + \sqrt{\left(\frac{4}{n} \langle X(\hat{\beta} - v), \varepsilon \rangle - \frac{2}{n} \|X(\hat{\beta} - v)\|^2\right) + 0}.
\]

**Lemma B.4.** For matrix $X$ and vector $\varepsilon$, let $X_i \sim N(0, \Sigma)$, and $\varepsilon_i$ be a zero-mean stochastic noise with finite variance $\sigma^2$. Let $K \subset \mathbb{R}^p$ be a convex cone. Fix any $\beta^*, \hat{\beta}$ and $v$ in $K$. Suppose $1 \leq w(\Sigma_2^{1/2} T_K(v) \cap \mathbb{S}^{p-1}) = o(\sqrt{n})$ and $\beta^* \in K$. Then with probability at least $1 - e^{-w(\Sigma_2^{1/2} T_K(v) \cap \mathbb{S}^{p-1}) - 3e^{-2}}\left(\frac{\sigma^2}{\sqrt{n}\sigma^4}\right) - \frac{\text{Var}(\varepsilon^2)}{\gamma \sigma^4}$ we have
\[
\|\Sigma^{1/2}_2 (\beta^* - \hat{\beta})\| \leq \|\Sigma^{1/2}_2 (\beta^* - v)\| + \frac{\sigma w(\Sigma_2^{1/2} T_K(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}}.
\]

**Remark B.5.** In the above Lemma, in the case when $\Sigma$ has bounded spectrum, one can substitute $w(\Sigma_2^{1/2} T_K(v) \cap \mathbb{S}^{p-1}) \leq \|\Sigma^{1/2}\|_{\text{op}} \|\Sigma^{-1/2}\|_{\text{op}} w(\Sigma_2^{1/2} T_K(v) \cap \mathbb{S}^{p-1})$ (see Remark 1.7 Plan and Vershynin (2016)). We may substitute $w(\Sigma_2^{1/2} T_K(v) \cap \mathbb{S}^{p-1})$ with any upper bound $\|\Sigma^{1/2}\|_{\text{op}} \|\Sigma^{-1/2}\|_{\text{op}} w(\Sigma_2^{1/2} T_K(v) \cap \mathbb{S}^{p-1})$, and the statement (including the high-probability guarantee) continues to hold with $w(\Sigma_2^{1/2} T_K(v) \cap \mathbb{S}^{p-1})$ in place of $w(\Sigma_2^{1/2} T_K(v) \cap \mathbb{S}^{p-1})$. 

33
Proof. Consider the “empirical process” term
\[
I = \frac{2}{n} \langle X(\hat{\beta} - v), \varepsilon \rangle - \frac{1}{n} \|X(\hat{\beta} - v)\|^2.
\]
Note that the unit vector \(\frac{\Sigma^{\frac{1}{2}} (\hat{\beta} - v)}{\|\Sigma^{\frac{1}{2}} (\hat{\beta} - v)\|} \in \Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}\), and \(X\Sigma^{-\frac{1}{2}}\) is a standard normal matrix.

By Gordon’s escape through mesh (Lemma B.2), with probability at least \(1 - e^{-\frac{t^2}{2}}\) we have
\[
\left\|X\Sigma^{-\frac{1}{2}} \frac{\Sigma^{\frac{1}{2}} (\hat{\beta} - v)}{\|\Sigma^{\frac{1}{2}} (\hat{\beta} - v)\|} \right\| \geq \inf_{w \in \Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}} \|X\Sigma^{-\frac{1}{2}} w\| \geq \sqrt{n - 1 - w(\Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}) - t}.
\]

Then
\[
I = \frac{2}{n} \langle X\Sigma^{-\frac{1}{2}} \Sigma^{\frac{1}{2}} (\hat{\beta} - v), \varepsilon \rangle - \frac{1}{n} \|X\Sigma^{-\frac{1}{2}} \Sigma^{\frac{1}{2}} (\hat{\beta} - v)\|^2
\leq \frac{2}{n} \left(\sqrt{n - 1 - w(\Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}) - t}\right) \|\Sigma^{\frac{1}{2}} (\hat{\beta} - v)\| \|\Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}\| - \frac{1}{n} \left(\sqrt{n - 1 - w(\Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}) - t}\right)^2 \|\Sigma^{\frac{1}{2}} (\hat{\beta} - v)\|^2.
\]

Using the fact \(-a^2 + 2ab \leq b^2\), with probability \(1 - e^{-\frac{t^2}{2}}\) we have
\[
I \leq \left(\sup_{u \in \Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}} \langle u, \frac{1}{\sqrt{n}} (X\Sigma^{-1})^\top \varepsilon \rangle\right)^2
\]
\[
I \leq \frac{\left(\sup_{u \in \Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}} \langle u, \frac{1}{\sqrt{n}} (X\Sigma^{-1})^\top \varepsilon \rangle\right)^2}{(\sqrt{n - 1 - w(\Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}) - t})^2}.
\]

Note that conditioning on the error term \(\varepsilon\), the vector \(\frac{1}{\sqrt{n}} (X\Sigma^{-\frac{1}{2}})^\top \varepsilon \sim N(0, I \|\varepsilon\|^2)\). Let
\[
I_{up} = \sup_{u \in \Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}} \langle u, \frac{1}{\sqrt{n}} (X\Sigma^{-\frac{1}{2}})^\top \varepsilon \rangle,
\]
by a concentration inequality of Gaussian process with finite variance (Boucheron et al., 2013, Theorem 5.8), we have
\[
\mathbb{P}(I_{up} - \mathbb{E}I_{up} \geq \sqrt{2t} \|\varepsilon\| \sqrt{n}) \leq e^{-t}.
\]

By the definition of Gaussian complexity \(\mathbb{E}I_{up} = w(\Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}) \|\varepsilon\| \sqrt{n}\) conditional on \(\varepsilon\). Then with probability \(1 - e^{-t}\) we have
\[
I_{up} \leq (w(\Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}) + \sqrt{2t} \|\varepsilon\| \sqrt{n}) \|\varepsilon\| \sqrt{n},
\]
thus with probability \(1 - e^{-t} - e^{-\frac{t^2}{2}}\) we have
\[
I \leq \left(\frac{(w(\Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}) + \sqrt{2t} \|\varepsilon\| \sqrt{n}) \|\varepsilon\| \sqrt{n}}{\sqrt{n - 1 - w(\Sigma^{\frac{1}{2}} T_K(v) \cap S^{p-1}) - t})^2}\right)^2.
\]
Then by lemma B.3 we have
\[
\frac{1}{\sqrt{n}}\|X(\hat{\beta} - v)\| \leq \frac{4}{\sqrt{n}}\|X(\beta^* - v)\| + \sqrt{2t}
\]
\[
\leq \frac{4}{\sqrt{n}}\|X(\beta^* - v)\| + \frac{\sqrt{2(w(\Sigma_T^t(v) \cap S^{p-1}) + \sqrt{2t})\|\varepsilon\|}}{\sqrt{n - 1 - w(\Sigma_T^t(v) \cap S^{p-1}) - t}}.
\]
(B.1)

The terms can be rewritten as
\[
\|X(\bar{\beta} - v)\| = \|X\Sigma^{-\frac{1}{2}}(\Sigma_T^\frac{1}{2}(\bar{\beta} - v))\| \|\Sigma_T^\frac{1}{2}(\bar{\beta} - v)\|
\]
\[
\|X(\beta^* - v)\| = \|X\Sigma^{-\frac{1}{2}}(\Sigma_T^\frac{1}{2}(\beta^* - v))\| \|\Sigma_T^\frac{1}{2}(\beta^* - v)\|
\]

Observe that both $\Sigma_T^\frac{1}{2}(\bar{\beta} - v)$ and $\Sigma_T^\frac{1}{2}(\beta^* - v)$ belong to $\Sigma_T^\frac{1}{2}T_K(v)$. We can bound the terms $\|X\Sigma^{-\frac{1}{2}}(\Sigma_T^\frac{1}{2}(\bar{\beta} - v))\|$ and $\|X\Sigma^{-\frac{1}{2}}(\Sigma_T^\frac{1}{2}(\beta^* - v))\|$ by Gordon’s escape through mesh (Lemma B.2), then with probability at least $1 - e^{-t} - 3e^{-\frac{t^2}{2}}$, we have
\[
\|\Sigma_T^\frac{1}{2}(\bar{\beta} - v)\| \leq \frac{4(\sqrt{n} + w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1}) + t)}{(\sqrt{n - 1} - w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1}) - t)} \|\Sigma_T^\frac{1}{2}(\beta^* - v)\|
\]
\[
+ \frac{\sqrt{2(w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1}) + \sqrt{2t})\|\varepsilon\|}}{\sqrt{n}}
\]
\[
\leq \frac{\sqrt{n - 1} - w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1}) + t}{\sqrt{n}}.
\]

Since $E\|\varepsilon\|^2 = \sigma^2$ and $Var(\|\varepsilon\|) = \frac{Var(\varepsilon^2)}{n}$, by Chebyshev’s inequality we have
\[
P\left(\left|\frac{\|\varepsilon\|^2}{n} - \sigma^2\right| \geq t\right) \leq \frac{Var(\varepsilon^2)}{nt^2}.
\]
(B.2)

Plug in $t = \sigma^2$ to get $\frac{\|\varepsilon\|}{\sqrt{n}} \leq \sqrt{2\sigma}$ with probability at least $1 - \frac{Var(\varepsilon^2)}{n\sigma^4}$. And by the triangle inequality $\|\Sigma_T^\frac{1}{2}(\bar{\beta} - \beta^*)\| - \|\Sigma_T^\frac{1}{2}(\beta^* - v)\| \leq \|\Sigma_T^\frac{1}{2}(\bar{\beta} - v)\|$, we can get
\[
\|\Sigma_T^\frac{1}{2}(\bar{\beta} - \beta^*)\| \leq \left(\frac{4(\sqrt{n} + w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1}) + t)}{(\sqrt{n - 1} - w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1}) - t)} + 1\right) \|\Sigma_T^\frac{1}{2}(\beta^* - v)\|
\]
\[
+ \frac{2\sigma(w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1}) + \sqrt{2t})}{\sqrt{n - 1} - w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1}) + t}
\]
\[
\leq \frac{\sqrt{n - 1} - w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1}) + t}{\sqrt{n}}.
\]

with probability at least $1 - e^{-t} - 3e^{-\frac{t^2}{2}} - \frac{Var(\varepsilon^2)}{n\sigma^4}$. Finally, given the assumption $w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1}) = o(\sqrt{n})$, we plug in $t = w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1})$ to get
\[
\|\Sigma_T^\frac{1}{2}(\bar{\beta} - \beta^*)\| \leq \|\Sigma_T^\frac{1}{2}(\beta^* - v)\| + \frac{w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1})}{\sqrt{n}}\sigma,
\]
with probability at least $1 - e^{-w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1})} - 3e^{-\frac{(w(\Sigma_T^\frac{1}{2}T_K(v) \cap S^{p-1})^2}{\sigma}} - \frac{Var(\varepsilon^2)}{n\sigma^4}$. 
\[\square\]
Next result bounds the supremum of a general covariance Gaussian process over a set $\mathcal{T}_K(v) \cap \mathbb{S}^{p-1}$. Notice that Lemma B.6 still holds if we replace $\mathcal{T}_K(v) \cap \mathbb{S}^{p-1}$ by any other set in $\mathbb{S}^p$.

**Lemma B.6.** For a convex set $K \subseteq \mathbb{R}^p$, $v \in K$, $g \sim N(0, I)$, and $\Sigma \in \mathbb{R}^{p \times p}$, we have

$$\mathbb{E} \sup_{u \in \mathcal{T}_K(v) \cap \mathbb{S}^{p-1}} |g^\top \Sigma^{1/2} u| \leq C\|\Sigma^{1/2}\|_{op} w(\mathcal{T}_K(v) \cap \mathbb{S}^{p-1}),$$

where $C \in \mathbb{R}$ is a constant.

**Proof.** First note that

$$\mathbb{E} \sup_{u \in \mathcal{T}_K(v) \cap \mathbb{S}^{p-1}} |g^\top \Sigma^{1/2} u| = \mathbb{E} \sup_{u \in (\mathcal{T}_K(v) \cup -\mathcal{T}_K(v)) \cap \mathbb{S}^{p-1}} g^\top \Sigma^{1/2} u.$$  

Now we will compare the process $X_u = g^\top \Sigma^{1/2} u$ to the process $Y_u = \|\Sigma^{1/2}\|_{op} g^\top u$. We have

$$\mathbb{E}(X_u - X_{u'})^2 = \mathbb{E}(g^\top \Sigma^{1/2} u - g^\top \Sigma^{1/2} u')^2 = (u - u')^\top \Sigma (u - u') \leq \|\Sigma\|_{op}\|u - u'\|^2,$$

and

$$\mathbb{E}(Y_u - Y_{u'})^2 = \|\Sigma^{1/2}\|^2_{op} \mathbb{E}(g^\top u - g^\top u')^2 = \|\Sigma\|_{op}\|u - u'\|^2 \geq \mathbb{E}(X_u - X_{u'})^2.$$  

Hence by Sudakov-Fernique’s inequality (Vershynin, 2018, Theorem 7.2.11), we can claim that

$$\mathbb{E} \sup_{u \in (\mathcal{T}_K(v) \cup -\mathcal{T}_K(v)) \cap \mathbb{S}^{p-1}} g^\top \Sigma^{1/2} u \leq \|\Sigma^{1/2}\|_{op} \mathbb{E} \sup_{u \in (\mathcal{T}_K(v) \cup -\mathcal{T}_K(v)) \cap \mathbb{S}^{p-1}} g^\top u$$

$$= \|\Sigma^{1/2}\|_{op} \mathbb{E} \sup_{u \in \mathcal{T}_K(v) \cap \mathbb{S}^{p-1}} |g^\top u|.$$  

Notice that the Gaussian complexity $w(\mathcal{T}_K(v) \cap \mathbb{S}^{p-1}) = \mathbb{E} \sup_{u \in \mathcal{T}_K(v) \cap \mathbb{S}^{p-1}} g^\top u$ has the same order as the quantity $\mathbb{E} \sup_{u \in \mathcal{T}_K(v) \cap \mathbb{S}^{p-1}} |g^\top u|$ (Vershynin, 2018, Exercise 7.6.9), so we get the desired result

$$\mathbb{E} \sup_{u \in \mathcal{T}_K(v) \cap \mathbb{S}^{p-1}} |g^\top \Sigma^{1/2} u| \leq C\|\Sigma^{1/2}\|_{op} w(\mathcal{T}_K(v) \cap \mathbb{S}^{p-1}).$$

\[\square\]

The next result demonstrates a property of the projection of a vector $y \in \mathbb{R}^p$ into the intersection of a convex cone $K$ and the unit sphere $\mathbb{S}^{p-1}$.

**Lemma B.7.** Let $K$ be a closed convex cone, and $\mathbb{S}^{p-1}$ be the unit sphere. For any vector $y \in \mathbb{R}^p$, we have

$$\arg\sup_{u \in K \cap \mathbb{S}^{p-1}} y^\top u = \frac{\Pi_K(y)}{\|\Pi_K(y)\|}.$$

**Proof.** Arbitrarily pick $u \in K$. By the characterization of the projection on a closed convex set (Moreau, 1962, Proposition 1),

$$(u - \Pi_K(y))(y - \Pi_K(y)) \leq 0.$$  

(B.3)
Since $K$ is a convex cone, $2\Pi_K(y)$ and $\frac{1}{2}\Pi_K(y)$ are in $K$. Plug them into (B.3) get
\[ y^\top \Pi_K(y) = \|\Pi_K(y)\|^2. \]  
(B.4)

Expand (B.3) and use the fact at (B.4) to get the following inequality
\[ u^\top y \leq u^\top \Pi_K(y), \]
thus
\[ \sup_{u \in K \cap S^{p-1}} u^\top y \leq \sup_{u \in K \cap S^{p-1}} u^\top \Pi_K(y). \]

By Cauchy-Schwartz inequality,
\[ \sup_{u \in K \cap S^{p-1}} u^\top \Pi_K(y) \leq \|\Pi_K(y)\|. \]

Combine the above two inequalities with (B.4), and the desired result is obtained
\[ \sup_{u \in K \cap S^{p-1}} u^\top y \leq \|\Pi_K(y)\| \leq y^\top \frac{\Pi_K(y)}{\|\Pi_K(y)\|}. \]

\[ \square \]

C Proof of Theorem 2.2

1.) Zero-mean Gaussian $X_i.$

When $\eta^\top = e^{(j)^\top} \Sigma^{-1}$, the LHS becomes
\[ \sup_{u \in T_K(v) \cap S^{p-1}} \left| \left( \sum_{i=1}^n n^{-1} e^{(j)^\top} \Sigma^{-1} X_i X_i^\top - e^{(j)^\top} u \right) \right|. \]

Set $b_j^\top = e^{(j)^\top} \Sigma^{-1/2}$, $\bar{b}_j = b_j/\|b_j\|$, $\tilde{X}_i = \Sigma^{-1/2} X_i$. Notice that $\tilde{X}_i \sim N(0, I)$. We obtain the equivalent expression:
\[ \|b_j\| \sup_{u \in T_K(v) \cap S^{p-1}} \left| \left( \sum_{i=1}^n n^{-1} \bar{b}_j^\top \tilde{X}_i \tilde{X}_i^\top - \bar{b}_j^\top \right) \right|. \]

Now by the triangle inequality the above can be decomposed as:
\[ \leq \|b_j\| \sup_{u \in T_K(v) \cap S^{p-1}} \left| \left( \sum_{i=1}^n n^{-1} \bar{b}_j^\top \tilde{X}_i \tilde{X}_i^\top \bar{b}_j - 1 \right) \bar{b}_j^\top \right| \]
\[ + \|b_j\| \sup_{u \in T_K(v) \cap S^{p-1}} \left| \left( \sum_{i=1}^n n^{-1} \bar{b}_j^\top \tilde{X}_i \tilde{X}_i^\top \left( I - \bar{b}_j \bar{b}_j^\top \right) \right) \right|. \]

Next the idea is to rearrange to get sub-exponential terms. Introduce new standard Gaussian vectors $\tilde{X}_i \sim N(0, I)$. Denote $Z_i = \bar{b}_j^\top \tilde{X}_i$, $W_i = \bar{b}_j^\top \tilde{X}_i$ and $\overline{X}_i = (I - \bar{b}_j \bar{b}_j^\top) \tilde{X}_i + \bar{b}_j \bar{b}_j^\top \tilde{X}_i$. Since

37
$\bar{b}_j$ is unit-norm, it is straightforward that $Z_i$ and $W_i$ are 1-dimensional standard Gaussian random variables. One can also see $\bar{X}_i \sim N(0, I)$ since

$$\mathbb{E}\bar{X}_i = (I - \bar{b}_j\bar{b}_j^T)\mathbb{E}X_i + \bar{b}_j\bar{b}_j^T \mathbb{E}X_i = 0,$$

$$\text{Var}(\bar{X}_i) = (I - \bar{b}_j\bar{b}_j^T)^2 + (\bar{b}_j\bar{b}_j^T)^2 = I.$$

and $Z_i$ is independent of $\bar{X}_i$ since $\mathbb{E}[Z_i \bar{X}_i] = 0$. By the triangle inequality we can further bound the above inequality as

$$\sup_{u \in \mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}}} \left| \sum_{i=1}^{n} n^{-1} e^{(j)^\top} \Sigma^{-1} X_i X_i^\top - e^{(j)^\top} u \right| \leq \|b_j\| \sup_{u \in \mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}}} \left| \sum_{i=1}^{n} n^{-1} Z_i^2 - 1 \right| \|\bar{b}_j\| \Sigma^{1/2} u\|

+ \|b_j\| \sup_{u \in \mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}}} \left| \sum_{i=1}^{n} n^{-1} Z_i \bar{X}_i \Sigma^{1/2} u \right|

+ \|b_j\| \sup_{u \in \mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}}} \left| \sum_{i=1}^{n} n^{-1} Z_i W_i \bar{b}_j^\top \Sigma^{1/2} u \right|.

(C.1)

1. We first control the first and third terms in (C.1).

Note that $\|b_j\| b_j^\top \Sigma^{1/2} = e^{(j)^\top}$, so that $\|b_j\| b_j^\top \Sigma^{1/2} u = u_j$. Thus the first and third term above can be re-written as

$$\sup_{u \in \mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}}} \left| \sum_{i=1}^{n} n^{-1} Z_i^2 - 1 \right| u_j \leq \left| \sum_{i=1}^{n} n^{-1} Z_i^2 - 1 \right|,

\sup_{u \in \mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}}} \left| \sum_{i=1}^{n} n^{-1} Z_i W_i \right| u_j \leq \left| \sum_{i=1}^{n} n^{-1} Z_i W_i \right|.

$$

Notice that $Z_i^2 \sim \chi_1^2$ is sub-exponential. The following tail bound holds (Wainwright, 2019, Example 2.11) for $t \in [0, 1]$

$$\mathbb{P}(\left| \sum_{i=1}^{n} n^{-1} Z_i^2 - 1 \right| \geq t) \leq 2 \exp(-\frac{nt^2}{8}).$$

$Z_i W_i$ is also a sub-exponential random variable having the same Orlicz norm with $Z_i^2 \sim \chi_1^2$, since they are both a product of two standard Gaussian random variables (Vershynin, 2018, Lemma 2.7.7). Then the same tail bound holds for $Z_i W_i$

$$\mathbb{P}(\left| \sum_{i=1}^{n} n^{-1} Z_i W_i \right| \geq t) \leq 2 \exp(-\frac{nt^2}{8}).$$

Given that $\overline{w}(\mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}}) = O(\sqrt{n})$, let $c_1, c_2$ be constants so that $c_i \overline{w}(\mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}}) \leq 1$ for $i \in \{1, 2\}$. Plugging in $t = \frac{c_1 \overline{w}(\mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}})}{\sqrt{n}}$ and $t = \frac{c_2 \overline{w}(\mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}})}{\sqrt{n}}$, the first and last term can be bounded by

$$\mathbb{P}\left( \|b_j\| \sup_{u \in \mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}}} \left| \sum_{i=1}^{n} n^{-1} Z_i^2 - 1 \right| b_j^\top \Sigma^{1/2} u \right) \leq 2 \exp(-\frac{c_1^2 \overline{w}^2(\mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}})}{8}),$$

(C.2)

$$\mathbb{P}\left( \|b_j\| \sup_{u \in \mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}}} \left| \sum_{i=1}^{n} n^{-1} Z_i W_i b_j^\top \Sigma^{1/2} u \right| \right) \leq 2 \exp(-\frac{c_2^2 \overline{w}^2(\mathcal{T}_K(v) \cap \mathcal{G}_{p^{-1}})}{8}).$$

(C.3)
2. Next, we bound the second term in (C.1). Condition on $Z = (Z_1, \ldots, Z_n)$ to get
\[
\mathbb{P}
\left(\left\|b_j\right\| \sup_{u \in \mathcal{T}_K(\nu) \cap \mathbb{S}^{p-1}} \left| \left( \sum_{i=1}^n n^{-1} Z_i \mathbf{X}_i^T \right) \Sigma^{1/2} u \right| \geq t \right)
\]
\[
= \int_{\frac{\sum Z_i^2}{n} > 2} \mathbb{P}
\left(\left\|b_j\right\| \sup_{u \in \mathcal{T}_K(\nu) \cap \mathbb{S}^{p-1}} \left| \left( \sum_{i=1}^n n^{-1} Z_i \mathbf{X}_i^T \right) \Sigma^{1/2} u \right| \geq t \left| Z \right| \right) p(Z) dZ + \int_{\frac{\sum Z_i^2}{n} \leq 2} \mathbb{P}
\left(\left\|b_j\right\| \sup_{u \in \mathcal{T}_K(\nu) \cap \mathbb{S}^{p-1}} \left| \left( \sum_{i=1}^n n^{-1} Z_i \mathbf{X}_i^T \right) \Sigma^{1/2} u \right| \geq t \left| Z \right| \right) p(Z) dZ
\]
\[
\leq \mathbb{P}\left( \frac{\sum Z_i^2}{n} > 2 \right) + \int_{\frac{\sum Z_i^2}{n} \leq 2} \mathbb{P}\left(\left\|b_j\right\| \sup_{u \in \mathcal{T}_K(\nu) \cap \mathbb{S}^{p-1}} \left| \left( \sum_{i=1}^n n^{-1} Z_i \mathbf{X}_i^T \right) \Sigma^{1/2} u \right| \geq t \left| Z \right| \right) p(Z) dZ.
\]

The first term above would be small. Formally, since $Z_i^2 \sim \chi_1^2$ is sub-exponential, the following tail bound holds (Wainwright, 2019, Example 2.11) for $t \in [0, \mathbb{I}]$
\[
\mathbb{P}(\left| \sum_{i=1}^n n^{-1} Z_i^2 - 1 \right| \geq t) \leq 2 \exp\left(-\frac{nt^2}{8}\right),
\]
and pick $t = 1$ to get
\[
\mathbb{P}\left( \frac{\sum Z_i^2}{n} > 2 \right) \leq \exp\left(\frac{n}{8}\right).
\]

For the second term above, here notice that $Z_i$ is independent of $\mathbf{X}_i$. Conditionally on the values of $Z_i$, we have that $\sum_{i=1}^n n^{-1} Z_i \mathbf{X}_i^T \sim N(0, (\sum_{i=1}^n Z_i^2) \mathbf{I})$. Thus $(\sum_{i=1}^n n^{-1} Z_i \mathbf{X}_i^T)^d = \frac{1}{\sqrt{n}} \sqrt{\sum Z_i^2} g^\top$, where $g \sim N(0, \mathbf{I})$. We can conclude that
\[
\left\|b_j\right\| \sup_{u \in \mathcal{T}_K(\nu) \cap \mathbb{S}^{p-1}} \left| \left( \sum_{i=1}^n n^{-1} Z_i \mathbf{X}_i^T \right) \Sigma^{1/2} u \right| \left| Z \right| \overset{d}{=} \left\|b_j\right\| \frac{1}{\sqrt{n}} \sqrt{\sum Z_i^2} \sup_{u \in \mathcal{T}_K(\nu) \cap \mathbb{S}^{p-1}} \left| g^\top \Sigma^{1/2} u \right|,
\]
so that
\[
\int_{\frac{\sum Z_i^2}{n} \leq 2} \mathbb{P}\left(\left\|b_j\right\| \sup_{u \in \mathcal{T}_K(\nu) \cap \mathbb{S}^{p-1}} \left| \left( \sum_{i=1}^n n^{-1} Z_i \mathbf{X}_i^T \right) \Sigma^{1/2} u \right| \geq t \left| Z \right| \right) p(Z) dZ
\]
\[
\leq \mathbb{P}\left(\left\|b_j\right\| \sqrt{\frac{2}{n}} \sup_{u \in \mathcal{T}_K(\nu) \cap \mathbb{S}^{p-1}} \left| g^\top \Sigma^{1/2} u \right| \geq t \right).
\]

Since the eigenvalues of $\Sigma^{1/2}$ are bounded, the following reasoning is valid. To bound this probability, we argue that $\sup_{u \in \mathcal{T}_K(\nu) \cap \mathbb{S}^{p-1}} \left| g^\top \Sigma^{1/2} u \right|$ is a $\|\Sigma^{1/2}\|_{\text{op}}$-Lipschitz function of $g$ in the Euclidean norm. Since $\left| g^\top \Sigma^{1/2} u \right|$ is a continuous function, and $\mathcal{T}_K(\nu) \cap \mathbb{S}^{p-1}$ is compact, the supremum is actually achieved. Let $u^*$ be the point where the supremum
of $\sup_{u \in T_K(v) \cap S^{p-1}} |g^T \Sigma^{1/2} u|$ is achieved. Then we have

$$\sup_{u \in T_K(v) \cap S^{p-1}} |g^T \Sigma^{1/2} u| - \sup_{u \in T_K(v) \cap S^{p-1}} |g^T \Sigma^{1/2} u^*| \leq |(g - g')^T \Sigma^{1/2} u^*| \leq \|g - g\|_{\text{op}} \|\Sigma^{1/2}\|_{\text{op}},$$

which completes the proof of $\|\Sigma^{1/2}\|_{\text{op}}$-Lipschitz. By the concentration of Lipschitz functions of Gaussian Variables (Wainwright, 2019, Theorem 2.26), we then obtain

$$\mathbb{P}\left( \sup_{u \in T_K(v) \cap S^{p-1}} |g^T \Sigma^{1/2} u| - \mathbb{E} \sup_{u \in T_K(v) \cap S^{p-1}} |g^T \Sigma^{1/2} u| \geq t \right) \leq \exp\left( - \frac{t^2}{2 \|\Sigma^{1/2}\|_{\text{op}}^2} \right). \quad (C.4)$$

Let $C$ be the constant such that $\mathbb{E} \sup_{u \in T_K(v) \cap S^{p-1}} |g^T \Sigma^{1/2} u| \leq C \|\Sigma^{1/2}\|_{\text{op}} \overline{w}(T_K(v) \cap S^{p-1})$ as in Lemma B.6. Plug in $t = C \|\Sigma^{1/2}\|_{\text{op}} \overline{w}(T_K(v) \cap S^{p-1})$, the above inequality becomes

$$\mathbb{P}\left( \sup_{u \in T_K(v) \cap S^{p-1}} |g^T \Sigma^{1/2} u| \geq 2C \|\Sigma^{1/2}\|_{\text{op}} \overline{w}(T_K(v) \cap S^{p-1}) \right) \leq \exp\left( - \frac{C^2 \overline{w}^2(T_K(v) \cap S^{p-1})}{2} \right),$$

noting that $\|b_j\| = \|\Sigma^{-\frac{1}{2}} e^{(j)}\| \leq \|\Sigma^{-\frac{1}{2}}\|_{\text{op}}$, so that

$$\mathbb{P}\left( \|b_j\| \sqrt{\frac{2}{n}} \sup_{u \in T_K(v) \cap S^{p-1}} |g^T \Sigma^{1/2} u| \geq 2\sqrt{2}C \|\Sigma^{1/2}\|_{\text{op}} \overline{w}(T_K(v) \cap S^{p-1}) \right) \leq \exp\left( - \frac{C^2 \overline{w}^2(T_K(v) \cap S^{p-1})}{2} \right).$$

Let $C' = C \|\Sigma^{\frac{1}{2}}\|_{\text{op}} \|\Sigma^{-\frac{1}{2}}\|_{\text{op}}$. Finally the second term of (C.1) can be bounded as

$$\mathbb{P}\left( \|b_j\| \sup_{u \in T_K(v) \cap S^{p-1}} \left| \left( \sum_{i=1}^n n^{-1} Z_i X_i^T \right) \Sigma^{\frac{1}{2}} u \right| \geq 2\sqrt{2}C' \overline{w}(T_K(v) \cap S^{p-1}) \right) \leq \exp\left( - \frac{n}{8} \right) + \exp\left( - \frac{C^2 \overline{w}^2(T_K(v) \cap S^{p-1})}{2} \right). \quad (C.5)$$

3. Bound (C.1)

Combine (C.2), (C.3), and (C.5) we can get

$$\mathbb{P}\left( \sup_{u \in T_K(v) \cap S^{p-1}} \left| \sum_{i=1}^n n^{-1} e^{(j)^T} \Sigma^{-1} X_i X_i^T - e^{(j)^T} u \right| \right) \leq c_1 \overline{w}(T_K(v) \cap S^{p-1}) \sqrt{\frac{n}{8}} + c_2 \overline{w}(T_K(v) \cap S^{p-1}) + \frac{2\sqrt{2}C' \overline{w}(T_K(v) \cap S^{p-1})}{\sqrt{n}} \geq 1 - 2 \exp\left( - \frac{c_1^2 \overline{w}^2(T_K(v) \cap S^{p-1})}{8} \right) - \exp\left( - \frac{n}{8} \right) - \exp\left( - \frac{C^2 \overline{w}^2(T_K(v) \cap S^{p-1})}{2} \right).$$

Since by assumption $\overline{w}(T_K(v) \cap S^{p-1}) \to \infty$ as $n$ increases we conclude that with probability converging to one we get

$$\sup_{u \in T_K(v) \cap S^{p-1}} \left| \sum_{i=1}^n n^{-1} e^{(j)^T} \Sigma^{-1} X_i X_i^T - e^{(j)^T} u \right| \leq \frac{\overline{w}(T_K(v) \cap S^{p-1})}{\sqrt{n}}.$$
2). Zero-mean Bounded $X_i$.

Observe the following identities:

$$Z := \sup_{u \in T_K(v) \cap \mathbb{S}^{p-1}} \left| \frac{1}{n} \sum_{i=1}^{n} e^{(j)^T} \Sigma^{-1} X_i X_i^T - e^{(j)^T} u \right|$$

$$= \sup_{u \in (T_K(v) \cap \mathbb{S}^{p-1}) \cup (-T_K(v) \cap \mathbb{S}^{p-1})} \left( \frac{1}{n} \sum_{i=1}^{n} e^{(j)^T} \Sigma^{-1} X_i X_i^T - e^{(j)^T} u \right)$$

$$= \sup_{u \in (T_K(v) \cap \mathbb{S}^{p-1}) \cup (-T_K(v) \cap \mathbb{S}^{p-1})} \frac{1}{n} \sum_{i=1}^{n} (e^{(j)^T} \Sigma^{-1} X_i X_i^T - e^{(j)^T} u).$$

Notice that $Z$ is the supremum of a bounded empirical processes. By Talagrand’s concentration inequality (Wainwright, 2019, Theorem 3.27)

$$\mathbb{P}(Z > \mathbb{E}Z + \delta) \leq 2 \exp\left(-\frac{n\delta^2}{8e\mathbb{E}\Sigma^2 + 4b\delta}\right), \quad (C.6)$$

where

$$b = \sup_{i \in [n]} \sup_{u \in (T_K(v) \cap \mathbb{S}^{p-1}) \cup (-T_K(v) \cap \mathbb{S}^{p-1})} \left( |e^{(j)^T} \Sigma^{-1} X_i X_i^T u| + |e^{(j)^T} u| \right),$$

and the definition of $\Sigma^2$ is given below. The second term of $b$ is bounded by 1, and the first term of $b$ can be bounded by Cauchy-Schwartz inequality as

$$|e^{(j)^T} \Sigma^{-1} X_i X_i^T u| \leq \|\Sigma^{-1} X_i\| \|X_i\| \leq \|\Sigma^{-1}\|_{op} \|X_i\|^2.$$ 

Since $\|X_i\|$ is finite and the eigenvalues of $\Sigma$ are bounded, the above quantity $|e^{(j)^T} \Sigma^{-1} X_i X_i^T u|$ is bounded from above. Thus $b$ is bounded from above. Also

$$\Sigma^2 = \sup_{u \in (T_K(v) \cap \mathbb{S}^{p-1}) \cup (-T_K(v) \cap \mathbb{S}^{p-1})} \frac{1}{n} \sum_{i=1}^{n} ((e^{(j)^T} \Sigma^{-1} X_i X_i^T - e^{(j)^T}) u)^2 \leq b^2.$$

Next, we would like to show that $\mathbb{E}Z$ cannot be large. $Z$ can be rewritten as

$$Z = \sup_{u \in T_K(v) \cap \mathbb{S}^{p}} \left| \frac{1}{n} \sum_{i=1}^{n} e^{(j)^T} \Sigma^{-1} X_i X_i^T u - \frac{1}{n} \sum_{i=1}^{n} e^{(j)^T} \Sigma^{-1} \mathbb{E}(\widetilde{X}_i \widetilde{X}_i^T) u \right|,$$

where $\widetilde{X}_i$ are independent copies of $X_i$. Notice that

$$f(M) = \sup_{u \in T_K(v) \cap \mathbb{S}^{p-1}} \left| \frac{1}{n} \sum_{i=1}^{n} e^{(j)^T} \Sigma^{-1} X_i X_i^T u - \frac{1}{n} \sum_{i=1}^{n} e^{(j)^T} \Sigma^{-1} Mu \right|,$$

is a convex function since it is a supremum over a set of convex functions. Thus by Jensen’s inequality

$$Z \leq \mathbb{E}_{\widetilde{X}} \sup_{u \in T_K(v) \cap \mathbb{S}^{p-1}} \left| \frac{1}{n} \sum_{i=1}^{n} e^{(j)^T} \Sigma^{-1} \widetilde{X}_i \widetilde{X}_i^T u - \frac{1}{n} \sum_{i=1}^{n} e^{(j)^T} \Sigma^{-1} \widetilde{X}_i \widetilde{X}_i^T u \right|.$$
Then we will use symmetrization and the contraction principle. Since $X, \tilde{X}$ have the same distribution, $e^{(j)\top} \Sigma^{-1} X_i X_i^\top u - e^{(j)\top} \Sigma^{-1} \tilde{X}_i \tilde{X}_i^\top u$ should be a symmetric random variable. This is because the difference between two random variables with the same distribution is symmetric. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ be a vector of Rademacher random variables. We have

$$E[Z] \leq E_{X, \tilde{X}} \sup_{u \in T_K(v) \cap S^{p-1}} \left| \frac{1}{n} \sum_{i=1}^{n} e^{(j)\top} \Sigma^{-1} X_i X_i^\top u - \frac{1}{n} \sum_{i=1}^{n} e^{(j)\top} \Sigma^{-1} \tilde{X}_i \tilde{X}_i^\top u \right|$$

$$= E_{X, \tilde{X}, \epsilon} \sup_{u \in T_K(v) \cap S^{p-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (e^{(j)\top} \Sigma^{-1} X_i X_i^\top u - e^{(j)\top} \Sigma^{-1} \tilde{X}_i \tilde{X}_i^\top u) \right|$$

$$\leq 2E_{X, \epsilon} \sup_{u \in T_K(v) \cap S^{p-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i e^{(j)\top} \Sigma^{-1} X_i X_i^\top u \right|$$

$$\leq E_{X} 2 \max_i |e^{(j)\top} \Sigma^{-1} X_i| E_{X, \epsilon} \sup_{u \in T_K(v) \cap S^{p-1}} \left| \left( \sum_{i=1}^{n} -1 \epsilon_i X_i^\top \right) u \right|,$$

where the last inequality follows by the contraction principle (Boucheron et al., 2013, Theorem 11.5) since the variable $\max_i |e^{(j)\top} \Sigma^{-1} X_i|$ is bounded. Next we will substitute the Rademachers with Gaussians in the following way

$$E_{X, \epsilon} \sup_{u \in T_K(v) \cap S^{p-1}} \left| \left( \sum_{i=1}^{n} -1 \epsilon_i X_i^\top \right) u \right|$$

$$= \sqrt{\frac{\pi}{2}} E_{X, \epsilon} \sup_{u \in T_K(v) \cap S^{p-1}} \left| \left( \sum_{i=1}^{n} -1 \epsilon_i E[|\epsilon_i| X_i^\top] \right) u \right|$$

$$\leq \sqrt{\frac{\pi}{2}} E_{X, \epsilon, \xi} \sup_{u \in T_K(v) \cap S^{p-1}} \left| \left( \sum_{i=1}^{n} -1 \epsilon_i |\epsilon_i| X_i^\top \right) u \right|$$

$$= \sqrt{\frac{\pi}{2}} E_{X, \xi} \sup_{u \in T_K(v) \cap S^{p-1}} \left| \left( \sum_{i=1}^{n} -1 \epsilon_i X_i^\top \right) u \right|.$$

Now conditional on $X$, the vector $\sum_{i=1}^{n} -1 \epsilon_i X_i^\top$ is Gaussian with zero mean and covariance $X^\top X/n^2$. Using Lemma B.6 we can upper bound the conditional expectation with

$$E_{\xi}( \sup_{u \in T_K(v) \cap S^{p-1}} \left| \left( \sum_{i=1}^{n} -1 \epsilon_i X_i^\top \right) u \right| |X) = E \sup_{u \in T_K(v) \cap S^{p-1}} \|g(X^\top X/n^2)^{1/2} u\| \leq C \|g(X^\top X/n^2)^{1/2}\|_{op} w(T_K(v) \cap S^{p-1}),$$

42
where \( g \sim N(0, I) \). And by Jensen’s inequality we have
\[
\mathbb{E}_X \|(X^\top X/n^2)^{1/2}\|_{\text{op}} = \mathbb{E}_X \sqrt{\sup_{u \in \mathbb{S}^{p-1}} |u^\top (X^\top X/n^2) u|}
\]
\[
= \mathbb{E}_X \sqrt{\frac{1}{n} \sup_{u \in \mathbb{S}^{p-1}} \left| \frac{1}{n} \sum_{i=1}^n u^\top X_i X_i^\top u \right|}
\]
\[
\leq \mathbb{E}_X \sqrt{\frac{1}{n} \sup_{u \in \mathbb{S}^{p-1}} \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \|u\|^2}
\]
\[
\leq \frac{1}{\sqrt{n}} \sup_i \|X_i\|.
\]
Since \( X_i \) is bounded, \( \mathbb{E}_X \|(X^\top X/n^2)^{1/2}\|_{\text{op}} \) is of the order \( 1/\sqrt{n} \). Hence the whole expectation will be bounded by
\[
\mathbb{E}Z \leq C \frac{\overline{w}(T_K(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}}.
\]
Pick \( \delta = \overline{w}(T_K(v) \cap \mathbb{S}^{p-1})/\sqrt{n} \) in (C.6), then with probability converging to one
\[
Z \lesssim \frac{\overline{w}(T_K(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}}.
\]

**D Proof of Corollary 2.4**

*Proof.* By Theorem 2.2, we know that the vector \( \eta^\top = e^{(j)^\top} \Sigma^{-1} \) is in \( Q \) with high probability. Now the idea is to show that there exists a small \( \delta > 0 \) such that \( B_\delta(e^{(j)^\top} \Sigma^{-1}) \) is inside of \( Q \) with high probability. Now let \( x \) be a unit vector. We have
\[
\sup_{u \in T_K(v) \cap \mathbb{S}^{p-1}} |(e^{(j)^\top} \Sigma^{-1} + \delta x^\top) X^\top X/n - e^{(j)^\top}) u| \leq \rho \frac{\overline{w}(T_K(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}} + \frac{\delta}{n} \sup_{x \in \mathbb{S}^{p-1}, u \in T_K(v) \cap \mathbb{S}^{p-1}} \frac{\|x\|}{n} \sup_{u \in T_K(v) \cap \mathbb{S}^{p-1}} \|X u\|.
\]
If \( X \) is bounded the above quantities are bounded with probability 1 hence the conclusion follows.

Next we consider the case when \( X \sim N(0, \Sigma) \). Let \( \overline{X} \) be an \( n \times p \) matrix with independent \( N(0, 1) \) entries. The last two terms \( \|\cdot\| \) are bounded as
\[
\sup_{x \in \mathbb{S}^{p-1}} \|X x\| = \sup_{x \in \mathbb{S}^{p-1}} \|\overline{X} \Sigma^{1/2} x\| \leq \|\overline{X}\|_{\text{op}} \|\Sigma^{1/2}\|_{\text{op}},
\]
\[
\sup_{u \in T_K(v) \cap \mathbb{S}^{p-1}} \|X u\| = \sup_{u \in T_K(v) \cap \mathbb{S}^{p-1}} \|\overline{X} \Sigma^{1/2} u\| \leq \|\overline{X}\|_{\text{op}} \|\Sigma^{1/2}\|_{\text{op}}.
\]
By the tail bound of the operator norm of Gaussian matrix (Vershynin, 2018, Corollary 7.3.3), \( \| \hat{X} \|_{op} \) is bounded by \( \sqrt{n} + \sqrt{p} \) with high probability, so that
\[
\sup_{u \in T_\kappa(v) \cap \mathbb{S}^{p-1}} |(e^{(j)}^T \Sigma^{-1} + \delta x^T)X/n - e^{(j)^T})u| < \frac{\rho(\mathbb{T}_k(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}} + \frac{\delta}{n} \| \mathbb{S}^{1/2} \|_{op}^2 (\sqrt{n} + \sqrt{p})^2.
\]

Let \( \epsilon = \frac{\delta}{n} \| \mathbb{S}^{1/2} \|_{op}^2 (\sqrt{n} + \sqrt{p})^2 \). Since we can find such a \( \delta \) for any \( \epsilon > 0 \), the ball \( B_\delta(e^{(j)^T} \Sigma^{-1}) \) is inside of \( Q \). Thus \( Q \) has a non-empty interior with high probability. \( \square \)

E Proof of Lemma 2.5

Proof. Let \( \lambda = \frac{\rho(\mathbb{T}_k(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}} \). We have
\[
\psi(\eta_n) = \sup_{u \in T_\kappa(v) \cap \mathbb{S}^{p-1}} |(\eta_n^T \hat{\Sigma} - e^{(j)^T})u - \lambda|
\]
\[
= \max\{ \sup_{u \in T_\kappa(v) \cap \mathbb{S}^{p-1}} (\eta_n^T \hat{\Sigma} - e^{(j)^T})u - \lambda, \sup_{u \in -T_\kappa(v) \cap \mathbb{S}^{p-1}} (\eta_n^T \hat{\Sigma} - e^{(j)^T})u - \lambda \}.
\]

Let \( \psi_0(\eta) = \sup_{u \in T_\kappa(v) \cap \mathbb{S}^{p-1}} (\eta^T \hat{\Sigma} - e^{(j)^T})u - \lambda \), and \( \psi_1(\eta) = \sup_{u \in -T_\kappa(v) \cap \mathbb{S}^{p-1}} (\eta^T \hat{\Sigma} - e^{(j)^T})u - \lambda \). The subgradient of \( \psi_0(\eta) \) is
\[
\partial \psi_0(\eta) = \hat{\Sigma} \arg \sup_{u \in T_\kappa(v) \cap \mathbb{S}^{p-1}} (\eta^T \hat{\Sigma} - e^{(j)^T})u,
\]
since for any \( y \in \mathbb{R}^p \),
\[
\psi_0(y) - \psi_0(x) = \sup_{u \in T_\kappa(v) \cap \mathbb{S}^{p-1}} (y^T \hat{\Sigma} - e^{(j)^T})u - \sup_{u \in T_\kappa(v) \cap \mathbb{S}^{p-1}} (x^T \hat{\Sigma} - e^{(j)^T})u
\]
\[
\geq (\hat{\Sigma}y - e^{(j)}, \arg \sup_{u \in T_\kappa(v) \cap \mathbb{S}^{p-1}} (x^T \hat{\Sigma} - e^{(j)^T})u) - \sup_{u \in T_\kappa(v) \cap \mathbb{S}^{p-1}} (x^T \hat{\Sigma} - e^{(j)^T})u
\]
\[
= (y - x, \hat{\Sigma} \arg \sup_{u \in T_\kappa(v) \cap \mathbb{S}^{p-1}} (x^T \hat{\Sigma} - e^{(j)^T})u).
\]

In the above observe that the “\( \arg \sup \)” is actually “\( \arg \max \)” since the set \( T_\kappa(v) \cap \mathbb{S}^{p-1} \) is compact an the function \( u \mapsto (\eta^T \hat{\Sigma} - e^{(j)^T})u \) is continuous. Similarly, the subgradient of \( \psi_1(\eta) \) is
\[
\partial \psi_1(\eta) = \hat{\Sigma} \arg \sup_{u \in -T_\kappa(v) \cap \mathbb{S}^{p-1}} (\eta^T \hat{\Sigma} - e^{(j)^T})u.
\]

By Lemma B.7, the subgradient of \( \psi_0 \) and \( \psi_1 \) are equivalent to
\[
\partial \psi_0(\eta) = \hat{\Sigma} \arg \sup_{u \in T_\kappa(v) \cap \mathbb{S}^{p-1}} (\eta^T \hat{\Sigma} - e^{(j)^T})u = \hat{\Sigma} \phi_0(\eta),
\]
\[
\partial \psi_1(\eta) = \hat{\Sigma} \arg \sup_{u \in -T_\kappa(v) \cap \mathbb{S}^{p-1}} (\eta^T \hat{\Sigma} - e^{(j)^T})u = \hat{\Sigma} \phi_1(\eta)
\]

By the pointwise maximum rule of subgradient (Shor, 2012, Theorem 1.13), the subgradient of \( \psi \) at \( \eta \) is \( \partial \psi_0(\eta) \) if \( \psi_0(\eta) > \psi_1(\eta) \), is \( \partial \psi_1(\eta) \) otherwise. \( \square \)
Proof of Lemma 2.6

Let \( \eta^* \in \text{argmin}_{\eta \in Q} \|\Sigma^{1/2} \eta\| \) be a constrained minima such that \( \|\eta^*\| \) is the smallest. Note that this implies that \( \eta^* \in \text{col}(\hat{\Sigma}^{1/2}) \). Let \( \eta_1 \) be the initial point with a finite \( \ell_2 \) norm. By Corollary 2.4 there exists a strictly feasible point \( \eta^{sf} \) such that \( \psi(\eta^{sf}) < 0 \). It is not hard to see that \( \|\eta^*\| \) is bounded, since \( \|\Sigma^{1/2} \eta^*\| \leq \|\Sigma^{1/2} \eta_1\| \) is bounded and \( \|\eta^*\| \leq \|\Sigma^{1/2} \eta^*\| \left(\lambda^{\min}_{\text{min}}(\hat{\Sigma}^{1/2})\right)^{-1} \) where \( \lambda^{\min}_{\text{min}}(\hat{\Sigma}^{1/2}) \) is the smallest positive eigenvalue of \( \hat{\Sigma}^{1/2} \). The latter holds by the definition of \( \eta^* \), and the fact that \( \eta^* \in \text{col}(\hat{\Sigma}^{1/2}) \). Furthermore, there exists at least one \( \eta^{sf} \) which is \( \|\eta^{sf}\| \) bounded, since according to Corollary 2.4 \( \eta^{sf} = \mathbf{e}^{(j)^\top} \hat{\Sigma}^{-1} \) is a choice of \( \eta^{sf} \). Thus \( \|\eta_1 - \eta^*\| \) and \( \|\eta_1 - \eta^{sf}\| \) are bounded. Let \( C_1 \) be such a constant satisfying \( \|\eta_1 - \eta^*\| \leq C_1 \) and \( \|\eta_1 - \eta^{sf}\| \leq C_1 \).

We also note that \( \|g_n\| \leq \|\Sigma^{1/2}\|_{\text{op}} \|\Sigma^{1/2} \eta_n\| \leq \|\Sigma^{1/2}\|_{\text{op}} \|\eta_n\| \) for \( \eta_n \in Q \); and obviously \( \|g_n\| \leq \|\Sigma\|_{\text{op}} \|\eta_n\| \) for \( \eta_n \notin Q \). Define a constant \( C_2 = \max\{\|\Sigma^{1/2}\|_{\text{op}}, \|\Sigma\|_{\text{op}}\} \), so that \( \|g_n\| \leq C_2 \).

Now we show that such a subgradient method converges in finite iterations. Let \( f(\eta) := \|\Sigma^{1/2} \eta\| \). At every step of iteration, we record the best candidate found so far as

\[
\eta_n^{\text{best}} = \text{argmin} \{ f(\eta_i) \mid \eta_i \in Q, i \in [n] \}.
\]

Arbitrarily choose \( \epsilon > 0 \). Let \( k \) be the iteration number such that after \( k \) the best value is \( \epsilon \)-suboptimal: \( f(\eta_n^{\text{best}}) < f(\eta^*) + \epsilon \) for \( n > k \). Also the best value before \( k \) is outside of the \( \epsilon \)-neighborhood: \( f(\eta_k^{\text{best}}) \geq f(\eta^*) + \epsilon \). Consequently \( f(\eta_n) \geq f(\eta^*) + \epsilon \) for \( n < k \) and \( \eta_n \in Q \).

1. Find a point \( \bar{\eta} \) and a constant \( c > 0 \) such that \( f(\bar{\eta}) \leq f(\eta^*) + c/2 \), and \( \psi(\bar{\eta}) \leq -c \).

Such a point \( \bar{\eta} \) can be chosen as

\[
\bar{\eta} = (1 - \theta)\eta^* + \theta \eta^{sf},
\]

where \( \theta = \min\{1, (\epsilon/2)/(f(\eta^{sf}) - f(\eta^*))\} \). One can see

\[
f(\bar{\eta}) \leq (1 - \theta)f(\eta^*) + \theta f(\eta^{sf}) \leq f(\eta^*) + \epsilon/2,
\]

\[
\psi(\bar{\eta}) \leq (1 - \theta)\psi(\eta^*) + \theta \psi(\eta^{sf}) \leq -\theta \psi(\eta^{sf}).
\]

so the constant \( c \) can be chosen as \( c = -\theta \psi(\eta^{sf}) \).

2. Show that before \( k \), for every iteration \( \|\eta_n + 1 - \bar{\eta}\|^2 \leq \|\eta_n - \bar{\eta}\|^2 - h \delta + h^2 \|g_n\|^2 \) where \( \delta = \min\{\epsilon, 2c\} \).

If \( \eta_n \in Q \), then \( g_n = \partial f(\eta_n) \), and by the definition of subgradient we have \( f(\bar{\eta}) - f(\eta_n) \geq g_n^\top (\bar{\eta} - \eta_n) \). Since \( f(\bar{\eta}) \leq f(\eta^*) + \epsilon/2 \) and \( f(\eta_n) \geq f(\eta^*) + \epsilon \), we have \( f(\eta_n) - f(\bar{\eta}) \geq \epsilon/2 \). Thus

\[
\|\eta_{n+1} - \bar{\eta}\|^2 = \|\eta_n - \eta_n g_n - \bar{\eta}\|^2
\]

\[
= \|\eta_n - \bar{\eta}\|^2 - 2h_n g_n^\top (\eta_n - \bar{\eta}) + h_n^2 \|g_n\|^2
\]

\[
\leq \|\eta_n - \bar{\eta}\|^2 - 2h_n (f(\eta_n) - f(\bar{\eta})) + h_n^2 \|g_n\|^2
\]

\[
\leq \|\eta_n - \bar{\eta}\|^2 - h_n \epsilon + h_n^2 \|g_n\|^2.
\]
If $\eta_n \notin Q$, then $g_n = \partial \psi(\eta_n)$, and by the definition of subgradient we have $\psi(\eta) - \psi(\eta_n) \geq g_n^\top (\eta - \eta_n)$. Since $\psi(\eta) \leq -c$ and $\psi(\eta_n) > 0$, we have $\psi(\eta_n) - \psi(\eta) \geq c$. Thus

$$
\|\eta_{n+1} - \tilde{\eta}\|^2 = \|\eta_n - h_n g_n - \tilde{\eta}\|^2 \\
= \|\eta_n - \tilde{\eta}\|^2 - 2 h_n g_n^\top (\eta_n - \tilde{\eta}) + h_n^2 \|g_n\|^2 \\
\leq \|\eta_n - \tilde{\eta}\|^2 - 2 h_n (\psi(\eta_n) - \psi(\tilde{\eta})) + h_n^2 \|g_n\|^2 \\
\leq \|\eta_n - \tilde{\eta}\|^2 - 2 h_n c + h_n^2 \|g_n\|^2.
$$

Define $\delta = \min\{\epsilon, 2c\}$ we have

$$
\|\eta_{n+1} - \tilde{\eta}\|^2 \leq \|\eta_n - \tilde{\eta}\|^2 - h_n \delta + h_n^2 \|g_n\|^2.
$$

(F.1)

3. Recursively apply (F.1) to get

$$
\|\eta_{n+1} - \tilde{\eta}\|^2 \leq \|\eta_1 - \tilde{\eta}\|^2 - \delta \sum_{n=1}^k h_n + \sum_{n=1}^k h_n^2 \|g_n\|^2,
$$

so that

$$
0 \leq C_1^2 - \delta \sum_{n=1}^k h_n + C_2^2 \sum_{n=1}^k h_n^2.
$$

When $\epsilon$ is chosen to be small, $\delta$ has the same order as $\epsilon$, since $\delta = \min\{\epsilon, 2c\}$ and $c = -\theta \psi(\eta^{sf}) = \epsilon \frac{\psi(\eta^{sf})}{2(f(\eta^{sf}) - f(\eta^*))}$. Thus we have

$$
\epsilon \lesssim \frac{C_1^2 + C_2^2 \sum_{n=1}^k h_n^2}{\sum_{n=1}^k h_n}.
$$

G Proof of Theorem 3.1

Proof. The debiased estimator $\hat{\beta}_d$ is constructed as

$$
\hat{\beta}_d = v + n^{-1} \tilde{\eta} \tilde{X}^\top (\tilde{Y} - \tilde{X} v).
$$

Using simple rearrangements the above can be seen to be equivalent to

$$
\sqrt{n}(\hat{\beta}_d - \beta^*) = \frac{1}{\sqrt{n}} \tilde{\eta} \tilde{X}^\top \varepsilon + \sqrt{n}(\tilde{\eta} \tilde{\Sigma} - I)(\beta^* - v).
$$

If we are interested in the $j^{th}$ coefficient $\sqrt{n}(\hat{\beta}_d(j) - \beta^*(j))$ we can multiply the above by $e(j)^\top = (0, \ldots, 1_j, \ldots, 0)$ to obtain

$$
\sqrt{n}(\hat{\beta}_d(j) - \beta^*(j)) = \frac{1}{\sqrt{n}} \tilde{\eta}^\top \tilde{X}^\top \varepsilon + \sqrt{n}(\tilde{\eta}^\top \tilde{\Sigma} - e(j)^\top)(\beta^* - v).
$$

(G.1)
In (G.1), we can see the first term is Gaussian conditional on $\overline{X}, \overline{Y}, \overline{X}$. The vector $\hat{\eta}$ depends on $\overline{X}, \overline{Y}$ since the constraint of the optimization (2.1) in step 2 involves $v$, which is obtained in step 1 and is dependent on $\overline{X}, \overline{Y}$. Since the noise $\varepsilon$ is assumed to be normal we have:

$$Z_j = \frac{1}{\sqrt{n}} \hat{\eta}^\top \overline{X} \varepsilon | \overline{X}, \overline{Y}, \overline{X} \sim N(0, \sigma^2 \hat{\eta}^\top \Sigma \hat{\eta}).$$

One can see that the solution of the optimization program (2.1) minimizes the variance of the first term in (G.1). Next, we would like the second term in (G.1) to converge to zero in order to achieve the asymptotic distribution of the debiased coefficient. Notice that the vector $\frac{\beta^* - v}{\Vert \beta^* - v \Vert} \in T_K(v) \cap \mathcal{S}^{p-1}$, so the second term $\Delta_j$ can be bounded as

$$|\Delta_j| = |\sqrt{n}(\hat{\eta}^\top \hat{\Sigma} - \mathbf{e}^{(j)^\top})(\beta^* - v)| \leq \sqrt{n} \sup_{u \in T_K(v) \cap \mathcal{S}^{p-1}} |(\hat{\eta}^\top \hat{\Sigma} - \mathbf{e}^{(j)^\top})u| \|v - \beta^*\|. \quad (G.2)$$

Since $\hat{\eta}$ is chosen so that the constraint in (2.1) is satisfied, the above will be at most

$$\sqrt{n} \rho \overline{w}(T_K(v) \cap \mathcal{S}^{p-1}) \|v - \beta^*\|. \quad (H.1)$$

Since $\overline{w}(T_K(v) \cap \mathcal{S}^{p-1}) \|v - \beta^*\| = o_p(1)$ as required in step 1, we have $\Delta_j = o_p(1)$. \qed

### H Proof of Theorem 3.4

**Proof.** By the triangle inequality we have

$$\left| \frac{1}{n} \sum_{i\in[n]} (Y_i - X_i^\top \hat{\beta})^2 - \sigma^2 \right| \leq \left| \frac{1}{n} \sum_{i\in[n]} \varepsilon_i^2 - \sigma^2 \right| + \left| \frac{1}{n} \sum_{i\in[n]} (Y_i - X_i^\top \beta) - \frac{1}{n} \sum_{i\in[n]} \varepsilon_i^2 \right|. \quad (H.1)$$

Let $T_n = \frac{\sqrt{n}}{\sqrt{\text{Var}(\varepsilon_i)}} \left( \frac{1}{n} \sum_{i\in[n]} \varepsilon_i^2 - \sigma^2 \right)$. Notice that $T_n$ converges to a standard normal distribution by central limit theorem. Suppose $\mathbb{E}e^6 < +\infty$. Let $\rho = \frac{\mathbb{E}[(e_i^2 - \sigma^2)^3]}{\text{Var}(e_i)^{3/2}}$, and $z \sim N(0, 1)$. By the Berry-Esseen central limit theorem (Vershynin, 2018, Theorem 2.1.3), we have

$$\mathbb{P}\{T_n > \delta\} - \mathbb{P}\{z > \delta\} \leq \frac{\rho}{\sqrt{n}}$$

$$\Rightarrow \mathbb{P}\{T_n > \delta\} \leq \mathbb{P}\{z > \delta\} + \frac{\rho}{\sqrt{n}}.$$

By a tail bound of a standard normal random variable (Wainwright, 2019, Example 2.1), the above inequality can be written as

$$\mathbb{P}\{T_n > \delta\} \leq e^{-\frac{\delta^2}{2}} + \frac{\rho}{\sqrt{n}}.$$

Thus plug in $T_n = \frac{\sqrt{n}}{\sqrt{\text{Var}(\varepsilon_i)}} \left( \frac{1}{n} \sum_{i\in[n]} \varepsilon_i^2 - \sigma^2 \right)$ we get

$$\mathbb{P}\left\{ \frac{1}{n} \sum_{i\in[n]} \varepsilon_i^2 - \sigma^2 \geq \sqrt{\frac{\text{Var}(\varepsilon_i)}{n}} \delta \right\} \leq e^{-\frac{\delta^2}{2}} + \frac{\rho}{\sqrt{n}}. \quad (H.1)$$
The second term can be bounded as
\[
\left|\frac{1}{n} \sum_{i \in [n]} (Y_i - X_i^T \hat{\beta})^2 - \frac{1}{n} \sum_{i \in [n]} \epsilon_i^2\right| = \left|\frac{1}{n} \sum_{i \in [n]} \left( (X_i^T \beta^* - X_i^T \hat{\beta} + \epsilon_i)^2 - \epsilon_i^2 \right) - \frac{1}{n} \sum_{i \in [n]} \left( (X_i^T \beta^* - X_i^T \hat{\beta})^2 - 2(X_i^T \beta^* - X_i^T \hat{\beta})\epsilon_i \right) \right|
\leq \frac{1}{n} \|X(\hat{\beta} - \beta^*)\|^2 + \frac{2}{n} \|X(\hat{\beta} - \beta^*)\| \|\epsilon\|.
\]
Since we have \(\frac{1}{\sqrt{n}}\|X(\hat{\beta} - \beta^*)\| \lesssim \frac{\sigma \delta}{\sqrt{n}}\), and \(\|\epsilon\|\) can be bounded by \(\sqrt{2}\sigma\) according to (B.2), so that
\[
\frac{1}{n} \|X(\hat{\beta} - \beta^*)\|^2 \lesssim \frac{\sigma^2 \delta^2}{n}, \quad \text{and} \quad \frac{2}{n} \|X(\hat{\beta} - \beta^*)\| \|\epsilon\| \lesssim \frac{\sigma^2 \delta}{\sqrt{n}}.
\]
By the fact \(\delta = o(\sqrt{n})\), we have \(\delta^2/n \leq \delta/\sqrt{n}\). Thus with probability converging to one we have
\[
\left|\frac{1}{n} \sum_{i \in [n]} (Y_i - X_i^T \hat{\beta})^2 - \frac{1}{n} \sum_{i \in [n]} \epsilon_i^2\right| \lesssim \frac{\sigma^2 \delta}{\sqrt{n}}.
\]
(H.2)
Combine (H.1) and (H.2), with probability converging to one
\[
|\hat{\sigma}^2 - \sigma^2| \lesssim \left(\frac{\sqrt{\text{Var}(\epsilon_i^2)} \vee \sigma^2}{\sqrt{n}}\right) \delta.
\]
\[\square\]

I Proof of Theorem 4.1

In the optimization program (4.1), \(v\) is the minima, so by the fact \(v' \in K\) is a feasible point, we have
\[
\|\hat{\beta} - v\| \leq \|\hat{\beta} - v'\| + \frac{\bar{w}(T_K(v') \cap S^{p-1})}{\sqrt{n}} - \frac{\bar{w}(T_K(v) \cap S^{p-1})}{\sqrt{n}}
\leq \|\hat{\beta} - v'\| + \frac{\bar{w}(T_K(v') \cap S^{p-1})}{\sqrt{n}},
\]
and by triangle inequality
\[
\|\hat{\beta} - v\| \leq \|\hat{\beta} - \beta^*\| + \|v' - \beta^*\|.
\]
Plug in \(v'\) in Lemma B.4 (and use Remark B.5 after it), to obtain with probability at least \(1 - e^{-\bar{w}(T_K(v') \cap S^{p-1})} - 3e^{-\left(\frac{\bar{w}(T_K(v') \cap S^{p-1})}{2}\right)} - \frac{\text{Var}(\epsilon_i^2)}{\sigma^2}\), we have
\[
\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\| \lesssim \|\Sigma^{1/2}(v' - \beta^*)\| + \frac{\sigma \bar{w}(\Sigma^{1/2}(T_K(v') \cap S^{p-1}))}{\sqrt{n}}.
\]
48
By Lemma B.6, Remark 1.7 of Plan and Vershynin (2016) and the fact that $\Sigma$ has bounded spectrum we conclude that

$$
\| \hat{\beta} - \beta^* \| \leq \| \nu' - \beta^* \| + \frac{\sigma \varpi(T_K(\nu'') \cap S^{p-1})}{\sqrt{n}},
$$

so that

$$
\| \hat{\beta} - \nu' \| \leq \| \nu' - \beta^* \| + \frac{\sigma \varpi(T_K(\nu'') \cap S^{p-1})}{\sqrt{n}} \quad \text{and} \quad \| \hat{\beta} - \nu \| \leq \| \nu' - \beta^* \| + \frac{(\sigma + 1) \varpi(T_K(\nu'') \cap S^{p-1})}{\sqrt{n}}.
$$

Again by triangle inequality

$$
\| \nu - \beta^* \| \leq \| \nu - \hat{\beta} \| + \| \hat{\beta} - \beta^* \|.
$$

Obviously the order of $\varpi(T_K(\nu) \cap S^{p-1})$ is also controlled by $\| \nu' - \beta^* \|$ and $\varpi(T_K(\nu'') \cap S^{p-1})$ since

$$
\frac{\varpi(T_K(\nu) \cap S^{p-1})}{\sqrt{n}} \leq \| \hat{\beta} - \nu' \| + \frac{\varpi(T_K(\nu'') \cap S^{p-1})}{\sqrt{n}} - \| \hat{\beta} - \nu \|
$$

$$
\leq \| \hat{\beta} - \nu' \| + \frac{\varpi(T_K(\nu'') \cap S^{p-1})}{\sqrt{n}}
$$

$$
\leq \| \nu' - \beta^* \| + \frac{(\sigma + 1) \varpi(T_K(\nu'') \cap S^{p-1})}{\sqrt{n}}.
$$

Finally

$$
\varpi(T_K(\nu) \cap S^{p-1}) \| \nu - \beta^* \| \leq \frac{1}{\sqrt{n}} \left[ \sqrt{n} \| \nu' - \beta^* \| + (\sigma + 1) \varpi(T_K(\nu'') \cap S^{p-1}) \right]^2 \leq \frac{1}{\sqrt{n}} \| \nu' - \beta^* \|^2 \vee \frac{(\sigma + 1)^2 \varpi^2(T_K(\nu') \cap S^{p-1})}{\sqrt{n}}.
$$

According to the condition of $\| \nu' - \beta^* \|$ and $\varpi(T_K(\nu'') \cap S^{p-1})$, with probability at least $1 - e^{-\varpi(T_K(\nu') \cap S^{p-1})} - 3e^{-\frac{\varpi(T_K(\nu'') \cap S^{p-1})}{2}} - \frac{\text{Var}(r^2)}{4\sigma^4}$

$$
\varpi(T_K(\nu) \cap S^{p-1}) \| \nu - \beta^* \| = o_p(1).
$$

### J Proof of Lemma 4.3

By an intermediate result (B.1) in the proof of Lemma B.4, with probability $1 - e^{-t} - e^{-\frac{t^2}{2}}$ we have

$$
\frac{1}{\sqrt{n}} \| x(\hat{\beta} - \nu') \| \leq \frac{4}{\sqrt{n}} \| x(\nu' - \beta^*) \| + \frac{\sqrt{2} (\varpi(T_K(\nu'') \cap S^{p-1}) + \sqrt{2} t \| e \| \sqrt{n}}{\sqrt{n} - 1 - \varpi(T_K(\nu'') \cap S^{p-1}) - t}.
$$
Set $t = w(\Sigma^{\frac{1}{2}} T_K(v') \cap S^{p-1})$, and $\frac{\|v\|}{\sqrt{n}}$ can be bounded by $\sqrt{2}\sigma$ according to (B.2). The above inequality becomes

$$\frac{1}{\sqrt{n}} \|X(\hat{\beta} - v')\| \leq \frac{4}{\sqrt{n}} \|X(v' - \beta^*)\| + \frac{w(\Sigma^{\frac{1}{2}} T_K(v') \cap S^{p-1})\sigma}{\sqrt{n}},$$

and by triangle inequality

$$\frac{1}{\sqrt{n}} \|X(\hat{\beta} - \beta^*)\| \leq \frac{1}{\sqrt{n}} \|X(\hat{\beta} - v')\| + \frac{1}{\sqrt{n}} \|X(v' - \beta^*)\| \leq \frac{5}{\sqrt{n}} \|X(v' - \beta^*)\| + \frac{w(\Sigma^{\frac{1}{2}} T_K(v') \cap S^{p-1})\sigma}{\sqrt{n}}.$$

Now what’s left is to bound $\frac{1}{\sqrt{n}} \|X(v' - \beta^*)\|$. For the Gaussian case $X_i \sim N(0, \Sigma)$, we can rewrite it as

$$\|X(\beta^* - v')\| = \|X\Sigma^{-\frac{1}{2}} \Sigma^{\frac{1}{2}} (\beta^* - v') \| \|\Sigma^{\frac{1}{2}} (\beta^* - v')\|.$$

By Gordon’s escape through mesh (Lemma B.2), since $\frac{\Sigma^{\frac{1}{2}}(\beta^* - v')}{\|\Sigma^{\frac{1}{2}} (\beta^* - v')\|} \in \Sigma^{\frac{1}{2}} T_K(v') \cap S^{p-1}$ with probability at least $1 - e^w(\Sigma^{\frac{1}{2}} T_K(v') \cap S^{p-1})/2$, we have

$$\|X\Sigma^{-\frac{1}{2}} \Sigma^{\frac{1}{2}} (\beta^* - v') \| \|\Sigma^{\frac{1}{2}} (\beta^* - v')\| \leq \sup_{u \in \Sigma^{\frac{1}{2}} T_K(v') \cap S^{p-1}} \|Xu\| \leq \sqrt{n} + 2 w(\Sigma^{\frac{1}{2}} T_K(v') \cap S^{p-1}).$$

Thus

$$\frac{1}{\sqrt{n}} \|X(\beta^* - v')\| \leq \frac{\sqrt{n} + 2 w(\Sigma^{\frac{1}{2}} T_K(v') \cap S^{p-1})}{\sqrt{n}} \|\Sigma^{\frac{1}{2}}\|_{op} \|v' - \beta^*\| \lesssim \|v' - \beta^*\|,$$

consequently

$$\frac{1}{\sqrt{n}} \|X(\hat{\beta} - \beta^*)\| \lesssim \|v' - \beta^*\| + \frac{w(\Sigma^{\frac{1}{2}} T_K(v') \cap S^{p-1})\sigma}{\sqrt{n}}.$$
K Proof of Lemma 4.4

Using simple rearrangement the equation
\[ \hat{\beta}_d = \hat{\beta} + n^{-1} \Sigma^{-1} \tilde{X}^\top (\tilde{Y} - \tilde{X}\hat{\beta}), \]
can be seen to be equivalent to
\[ \sqrt{n}(\hat{\beta}_d - \beta^*) = \frac{1}{\sqrt{n}} \Sigma^{-1} \tilde{X}^\top e + \sqrt{n}(\Sigma^{-1} \hat{\Sigma} - I)(\beta^* - \tilde{\beta}). \]  

(K.1)

The first term is Gaussian condition on \( \tilde{X} \):
\[ Z = \frac{1}{\sqrt{n}} \Sigma^{-1} \tilde{X}^\top e | \tilde{X} \sim N(0, \sigma^2 \Sigma^{-1} \hat{\Sigma} \Sigma^{-1}). \]

What remains to show is that the second term in (K.1) converges to zero with high probability. Let \( u = \beta^* - \tilde{\beta} \), and \( e^{(j)\top} = (0, \ldots, 1_j, \ldots, 0) \). The \( j \)th coordinate of the second term can be written as
\[ \sqrt{n}(e^{(j)\top} \Sigma^{-1} \tilde{\Sigma} - e^{(j)\top})u = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (e^{(j)\top} \Sigma^{-1} \tilde{X}_i \tilde{X}_i^\top u - u_j). \]

Let \( g_i = e^{(j)\top} \Sigma^{-1} \tilde{X}_i \tilde{X}_i^\top u - u_j \). Notice that \( (\tilde{X}, \tilde{Y}) \) is independent from \( (\tilde{X}, \tilde{Y}) \), and \( u \) is constant conditionally on \( (\tilde{X}, \tilde{Y}) \), so \( E(g_i|\tilde{X}, \tilde{Y}) = e^{(j)\top} \Sigma^{-1} \Sigma u - u_j = 0 \). Moreover, \( e^{(j)\top} \Sigma^{-1} \tilde{X}_i^\top u \) are Gaussian random variables condition on \( (\tilde{X}, \tilde{Y}) \). Let \( \| \cdot \|_{\psi_2} \) be the sub-gaussian norm defined in (Vershynin, 2018, Definition 2.5.6). The sub-gaussian norm of a Gaussian random variable is up to a constant of its standard deviation (Vershynin, 2018, Example 2.5.8), so we have
\[ \| e^{(j)\top} \Sigma^{-1} \tilde{X}_i \|_{\psi_2} \leq C_1 \| \Sigma^{-\frac{1}{2}} e^{(j)} \| \]
\[ \| \tilde{X}_i^\top u \|_{\psi_2} \leq C_2 \| \Sigma^{\frac{1}{2}} u \|. \]

Let \( \| \cdot \|_{\psi_1} \) be the sub-exponential norm defined in (Vershynin, 2018, Definition 2.7.5). The product of two sub-gaussian random variables is a sub-exponential random variable, and the corresponding sub-exponential norm is less than the product of sub-Gaussian norms (Vershynin, 2018, Lemma 2.7.7). Thus
\[ \| e^{(j)\top} \Sigma^{-1} \tilde{X}_i \tilde{X}_i^\top u \|_{\psi_1} \leq \| e^{(j)\top} \Sigma^{-1} \tilde{X}_i \|_{\psi_2} \| \tilde{X}_i^\top u \|_{\psi_2} \]
\[ \leq C_1 C_2 \| \Sigma^{-\frac{1}{2}} \|_{\text{op}} \| \Sigma^{\frac{1}{2}} u \|. \]

Additionally, the sub-exponential norm of a centered sub-exponential random variable is up to a constant to the original one (Vershynin, 2018, Exercise 2.7.10)
\[ \| e^{(j)\top} \Sigma^{-1} \tilde{X}_i \tilde{X}_i^\top u - u_j \|_{\psi_1} \leq C_3 \| e^{(j)\top} \Sigma^{-1} \tilde{X}_i \tilde{X}_i^\top u \|_{\psi_1} \]
\[ \leq C_1 C_2 C_3 \| \Sigma^{-\frac{1}{2}} \|_{\text{op}} \| \Sigma^{\frac{1}{2}} u \|. \]
Let $C = C_1C_2C_3$. Given the sub-exponential norm of $g_i = e^{(j)^T \Sigma^{-1} \hat{X}_i \hat{X}_i^T u - u_j}$, use Bernstein’s inequality (Vershynin, 2018, Theorem 2.8.1) to get the conditional concentration inequality

$$\mathbb{P} \left( \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i \right| \geq t \right| \mathbf{X}, \mathbf{Y} \right) \leq 2 \exp \left[ - c \min \left( \frac{t^2}{C^2 \| \Sigma^{-\frac{1}{2}} \|_{\text{op}}^2 \| \Sigma^{\frac{1}{2}} \mathbf{u} \|^2}, \frac{t \sqrt{n}}{C \| \Sigma^{-\frac{1}{2}} \|_{\text{op}} \| \Sigma^{\frac{1}{2}} \mathbf{u} \|} \right) \right].$$

The unconditioned concentration inequality can be obtained by

$$\mathbb{P} \left( \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i \right| \geq t \right) = \int \mathbb{P} \left( \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i \right| \geq t \right| \mathbf{X}, \mathbf{Y} \right) d\mu(\mathbf{X}, \mathbf{Y})$$

$$\leq \int 2 \exp \left[ - c \min \left( \frac{t^2}{C^2 \| \Sigma^{-\frac{1}{2}} \|_{\text{op}}^2 \| \Sigma^{\frac{1}{2}} \mathbf{u} \|^2}, \frac{t \sqrt{n}}{C \| \Sigma^{-\frac{1}{2}} \|_{\text{op}} \| \Sigma^{\frac{1}{2}} \mathbf{u} \|} \right) \right] d\mu(\mathbf{X}, \mathbf{Y})$$

$$= \int_{\| \Sigma^{\frac{1}{2}} \mathbf{u} \| \leq \theta} 2 \exp \left[ - c \min \left( \frac{t^2}{C^2 \| \Sigma^{-\frac{1}{2}} \|_{\text{op}}^2 \| \Sigma^{\frac{1}{2}} \mathbf{u} \|^2}, \frac{t \sqrt{n}}{C \| \Sigma^{-\frac{1}{2}} \|_{\text{op}} \| \Sigma^{\frac{1}{2}} \mathbf{u} \|} \right) \right] d\mu(\mathbf{X}, \mathbf{Y})$$

$$\mathbb{P} \left( \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i \right| \geq t \right) \cdot \mathbb{P} \left[ \| \Sigma^{\frac{1}{2}} \mathbf{u} \| > \theta \right].$$

The threshold $\theta = \| \Sigma^{\frac{1}{2}}(\mathbf{v}' - \mathbf{\beta}^*)\| + \frac{w(\Sigma^{\frac{1}{2}}T_K(\mathbf{v}')\cap \mathcal{S}^{p-1})}{\sqrt{n}}$ is chosen according to the result of Lemma B.4 in order to make the second term vanish. Apply Lemma B.4 with $\mathbf{v} = \mathbf{v}'$, one can see the second term of RHS vanishes as $n \to \infty$.

For the first term, take $t = \theta \| \Sigma^{-\frac{1}{2}} \|_{\text{op}} a_n$, we can see that $\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i \right|$ is bounded as

$$\mathbb{P} \left( \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i \right| \geq \theta \| \Sigma^{-\frac{1}{2}} \|_{\text{op}} a_n \right) \leq 2 \exp \left[ - c \min \left( \frac{a_n^2 \sqrt{na_n}}{C^2}, \frac{\sqrt{n}a_n}{C} \right) \right],$$

where $a_n$ is picked such that $\theta \| \Sigma^{-\frac{1}{2}} \|_{\text{op}} a_n = o(1)$ and $a_n \to \infty$. Specifically we have

$$\| \Sigma^{\frac{1}{2}}(\mathbf{v}' - \mathbf{\beta}^*)\| a_n = o(1), \ w(\Sigma^{\frac{1}{2}}T_K(\mathbf{v}')\cap \mathcal{S}^{p-1}) a_n = o(\sqrt{n}), \ a_n \to \infty.$$ 

The first condition reduces to $\| \mathbf{v}' - \mathbf{\beta}^*\| a_n = o(1)$ since $\lambda_{\min}(\Sigma^{1/2}) \| \mathbf{v}' - \mathbf{\beta}^*\| \leq \| \Sigma^{\frac{1}{2}}(\mathbf{v}' - \mathbf{\beta}^*)\| \leq \| \Sigma^{1/2} \|_{\text{op}} \| \mathbf{v}' - \mathbf{\beta}^*\|$. The condition $w(\Sigma^{\frac{1}{2}}T_K(\mathbf{v}')\cap \mathcal{S}^{p-1}) a_n = o(\sqrt{n})$ reduces to $w(T_K(\mathbf{v}')\cap \mathcal{S}^{p-1}) a_n = o(\sqrt{n})$ by the fact $w(\Sigma^{\frac{1}{2}}T_K(\mathbf{v}')\cap \mathcal{S}^{p-1}) \leq \| \Sigma^{-\frac{1}{2}} \|_{\text{op}} \| \Sigma^{\frac{1}{2}} \|_{\text{op}} w(T_K(\mathbf{v}')\cap \mathcal{S}^{p-1})$ (Plan and Vershynin, 2016, Remark 1.7).

### L Proof of Lemma 4.6

This argument is mostly repeating an argument from Cai et al. (2017). Before the proof, we need to introduce two definitions. The first is the $\chi^2$ distance between two density functions

$$\chi^2(f_1, f_0) = \int \frac{(f_1(z) - f_0(z))^2}{f_0(z)} dz = \int \frac{f_1^2(z)}{f_0(z)} dz - 1.$$ 

The second is the total variation distance (with a scaling factor 2 in front) between two density functions

$$TV(f_1, f_0) = \int |f_1(z) - f_0(z)| dz$$
A well-known fact is that $TV(f_1, f_0) \leq \sqrt{\chi^2(f_1, f_0)}$.

Now we start the proof. Let $K \subset \mathbb{R}^p$ be a convex set. The parameter space is defined as

$$
\mathcal{H} = \{ \beta \in \mathbb{R}^p : \| \beta - \nu \|^2 \leq \frac{R_n}{\sqrt{n}} \text{ for } \nu \in K \text{ and } \overline{w}^2(T_K(\nu) \cap \mathcal{S}^{p-1}) = R_n \sqrt{n} \},
$$

which is the space we are able to perform inference on, asymptotically, via the debiasing procedure proposed in this paper. Suppose we want to debias the $j$-th coordinate.

1. Let $\delta > 0$ be a small positive constant such that $\delta \sigma \sqrt{\| \Sigma^{\frac{1}{2}} \|_{op}^{-1}} = o(n^{\frac{1}{2}})$. Define

$$
\mathcal{H}_0 = \{ \beta \in K : \overline{w}^2(T_K(\beta) \cap \mathcal{S}^{p-1}) = 2r_n \sqrt{n} \text{ and } \beta(+ \text{ or } -) \frac{\delta \sigma \sqrt{\| \Sigma^{\frac{1}{2}} \|_{op}^{-1/2}}}{\sqrt{n}} e(j) \in K \},
$$

and $\mathcal{H}_1 = \mathcal{H}$. Since $R_n \geq 2r_n$, for and $r_n = o(1)$ for $n$ large enough it is not hard to see that $\mathcal{H}_0 \subset \mathcal{H}$. In addition, by the definition of $r_n$, it follows that the set $\mathcal{H}_0$ is not empty. For a given $\beta^* \in \mathcal{H}_0$, we find a $\beta'$ such that

$$
\| \beta' - \beta^* \| = | \beta'_j - \beta^*_j | = \frac{\delta \sigma \sqrt{\| \Sigma^{\frac{1}{2}} \|_{op}^{-1/2}}}{\sqrt{n}}.
$$

According to the definition of $\mathcal{H}$, we always have $\beta' \in \mathcal{H}$.

2. Let $f_0(Y | X)$ be the density of $Y$ given $X$ with the parameter $\beta^*$, and $f_1(Y | X)$ be the density of $Y$ given $X$ with the parameter $\beta'$. Such a conditional distribution of $Y$ is Gaussian since the noise has a Gaussian distribution with standard error $\sigma$. It can be shown that

$$
\chi^2(f_1(Y | X), f_0(Y | X)) = \exp\left(\frac{1}{\sigma^2} \| X(\beta' - \beta^*) \|^2 \right) - 1.
$$

With the fact $X_i \sim N(0, \Sigma)$, we have

$$
\chi^2(f_1(Y, X), f_0(Y, X)) = \mathbb{E}_X \exp\left(\frac{1}{\sigma^2} \| X(\beta' - \beta^*) \|^2 \right) - 1
\begin{align*}
&= \prod_{i=1}^n \mathbb{E}_X \exp\left(\frac{1}{\sigma^2} \| X_i^\top (\beta' - \beta^*) \|^2 \right) - 1 \\
&= \prod_{i=1}^n \mathbb{E}_X \exp\left(\frac{1}{\sigma^2} \| (\Sigma^{-\frac{1}{2}} X_i)^\top \Sigma^{\frac{1}{2}} (\beta' - \beta^*) \|^2 \right) - 1.
\end{align*}
$$

Since $(\Sigma^{-\frac{1}{2}} X_i)^\top \Sigma^{\frac{1}{2}} (\beta' - \beta^*) = \| \Sigma^{\frac{1}{2}} (\beta' - \beta^*) \| z_i$ where $z_i \sim N(0, 1)$, by the moment generating function of $\chi^2$ distribution, the above equation becomes

$$
\chi^2(f_1(Y, X), f_0(Y, X)) = \left( 1 - \frac{2\| \Sigma^{\frac{1}{2}} (\beta' - \beta^*) \|^2}{\sigma^2} \right)^{-\frac{n}{2}} - 1.
$$

If $\frac{2\| \Sigma^{\frac{1}{2}} (\beta' - \beta^*) \|^2}{\sigma^2} < \frac{\log 2}{2}$, by the inequality $\frac{1}{1-x} \leq \exp(2x)$ for $x \in [0, \frac{\log 2}{2}]$, we have

$$
\chi^2(f_1(Y, X), f_0(Y, X)) \leq \exp\left(\frac{2n \| \Sigma^{\frac{1}{2}} (\beta' - \beta^*) \|^2}{\sigma^2} \right) - 1.
$$
3. By Lemma 1 in Cai et al. (2017), for any \( CI_\alpha(\beta^j, Y, X) \in \mathcal{I}_\alpha(\mathcal{H}) \) we have

\[
L(CI_\alpha(\beta^j, Y, X)) \geq \frac{\delta \sigma \| \Sigma \beta \|^1_{op}}{\sqrt{n}} \left( 1 - 2\alpha - TV(f_1(Y, X), f_0(Y, X)) \right)
\]

\[
\geq \frac{\delta \sigma \| \Sigma \beta \|^1_{op}}{\sqrt{n}} \left( 1 - 2\alpha - \sqrt{\exp(2\delta^2) - 1} \right).
\]

M Proof of Lemma 4.8

The proof is the same as that of Lemma 4.6 modulo some small changes. For any \( \beta^* \in \mathcal{H}(R_n) \) let \( \beta = \beta^* + \frac{\delta \| \Sigma \beta \|^1_{op} w}{\sqrt{n}} \). We now argue that \( \beta \in \mathcal{H}(\nu_n) \). It is clear that \( \beta \in K \) by the definition of \( K \). Let \( \nu \) be such that \( \| \beta^* - \nu \|^2 \leq R_n/\sqrt{n} \), for \( \nu \in K \) and \( \mathbb{E}(\mathcal{M}(\nu) \cap S^{p-1}) \leq R_n/\sqrt{n} \). By the triangle inequality:

\[
\| \nu - \beta \| \leq \| \nu - \beta^* \| + \frac{\delta \| \Sigma \beta \|^1_{op} \nu}{\sqrt{n}} \leq \sqrt{\frac{R_n}{n}} + \frac{\delta \| \Sigma \beta \|^1_{op} \nu}{\sqrt{n}}.
\]

Squaring the inequality in the preceding display and using the elementary inequality \((a+b)^2 \leq 2a^2 + 2b^2\) shows that \( \beta \in \mathcal{H}(\nu_n) \). The rest of the proof is identical to that of Lemma 4.6 and we omit the details.

N Proof of Proposition 4.10

By definition, \( \mathcal{T}(M^{p+} (\nu)) = \{ u - tv : t \geq 0, u \in M^{p+} \} \). If \( \nu \) is a non-zero constant, it is trivial that \( \mathcal{T}(M^{p+} (\nu)) = M^{p} \). Moreover if all the coordinates of \( \nu \) are zeros, the positiveness is also preserved so that \( \mathcal{T}(M^{p+} (\nu)) = M^{p+} \). Now it is sufficient to consider the case where \( \nu \) has at least two constant pieces.

Firstly, suppose the first constant piece of \( \nu \) doesn’t consist of zeros. Within each constant piece, the monotonicity of \( u_i - tv_i \) is preserved, but not necessarily the positiveness, so that \( \mathcal{T}(M^{p+} (\nu)) \subset M^{p_1} \times M^{p_2} \times \ldots \times M^{p_l} \). To show the other direction, arbitrarily choose \( x \in M^{p_1} \times M^{p_2} \times \ldots \times M^{p_i} \). Let \( \epsilon_1 = \min_{i \in S} (v_{i+1} - v_i) \), where \( S = \{ i : v_i+1 > v_i \} \) and \( \epsilon_2 = 2 \min_{i \in [p]} v_i \). Pick \( t = \frac{2 \| x \|_{\infty}}{\epsilon_1 \wedge \epsilon_2} \), then for all \( i \in [p] \) we have

\[
x_i + tv_i \geq x_i + \| x \|_{\infty} \geq 0,
\]

and for \( i \in S \):

\[
t(v_i+1 - v_i) = \frac{2 \| x \|_{\infty}}{\epsilon_1 \wedge \epsilon_2} (v_i+1 - v_i) \geq x_i - x_{i+1} \quad \Rightarrow \quad x_i + tv_i \leq x_{i+1} + tv_{i+1}.
\]

For \( i \in [p-1] \setminus S \) we have \( v_i+1 = v_i \) and \( x_i \leq x_{i+1} \) so that \( x_i + tv_i \leq x_{i+1} + tv_{i+1} \) also holds. Thus for any \( x \in M^{p_1} \times M^{p_2} \times \ldots \times M^{p_i} \) there is a \( t \) such that \( x + tv \in M^{p+} \). The direction \( \mathcal{T}(M^{p+} (\nu)) \supset M^{p_1} \times M^{p_2} \times \ldots \times M^{p_i} \) holds.

When the first constant piece of \( \nu \) is zero valued, within it \( u_i - tv_i = u_i \) is always positive and monotone. For the other constant pieces, \( u_i - tv_i \) is still monotone, so that \( \mathcal{T}(M^{p+} (\nu)) \subset M^{p_1} \times M^{p_2} \times \ldots \times M^{p_i} \). For the other direction, let \( \epsilon_1 = \min_{i \in S} (v_{i+1} - v_i) \), and \( \epsilon_2 \) be two times the minimum non-zero \( v_i \). Also let \( t = \frac{2 \| x \|_{\infty}}{\epsilon_1 \wedge \epsilon_2} \), it is easy to verify that \( x + tv \in M^{p+} \).
Proof of Lemma 4.12

Proof. We first note that by Cauchy-Schwartz the Gaussian complexity is upper bounded by the statistical dimension, i.e.,
\[ \mathbb{E}_{g \sim N(0, I)} \sup_{x \in T_K(v)} \langle g, x \rangle = \mathbb{E}_{g \sim N(0, I)} \| \Pi_{T_K(v)}(g) \| \leq \sqrt{\mathbb{E}_{g \sim N(0, I)} \| \Pi_{T_K(v)}(g) \| ^2} \]
Now by Lemma 4.14, and \( g \sim \chi_1^2 \), the projection is
\[ \mathbb{E} \| \Pi_{T_K(v)}(g) \|^2 = \mathbb{E} \sum_{i: \mathbf{v}(i) \neq 0} g(i)^2 + \mathbb{E} \sum_{i: \mathbf{v}(i) = 0} g(i)^2 = p - |\{ i : \mathbf{v}(i) = 0 \}|/2. \]

Proof of Lemma 4.13

Proof. This statement is obvious and we omit the details.

Proof of Lemma 4.14

Proof. Let \( S \) be the set of zero coordinates of \( \mathbf{v} \). The tangent cone of \( K \) at \( \mathbf{v} \) can be written as
\[ T_K(v) = \{ x \in \mathbb{R}^p : x(i) \geq 0 \text{ for } i \in S \} \]
Then it’s straightforward that the projection takes the corresponding form.

Proof of Lemma 4.15

Proof. By definition, \( \mathbf{v}_s \in \arg\min_{\mathbf{w} \in T} \| \mathbf{w} - \hat{\beta} \| \). For brevity let \( \mathbf{v}' \) be any vector in \( \arg\min_{\mathbf{w} \in T} \| \mathbf{w} - \hat{\beta} \| \). First for each coordinate of \( \mathbf{v}' \), we have either \( \text{sign}(\mathbf{v}'(i)) = \text{sign}(\hat{\beta}(i)) \), or \( \text{sign}(\mathbf{v}'(i)) = 0 \), because otherwise we can always reverse the sign to make the \( \ell_2 \)-norm of difference \( \| \mathbf{v}' - \hat{\beta} \| \) smaller.

Fix a set \( S' \) of \( s \) coordinates which is the assumed support for the vector \( \mathbf{v}'(i) \). Consider the following optimization problem
\[ \min_{\mathbf{v}'} \sum_{i \in S'} (|\hat{\beta}(i)| - |\mathbf{v}'(i)|)^2 + \sum_{i \notin S'} \hat{\beta}^2(i) \quad \text{subject to} \quad \sum_{i \in S'} |\mathbf{v}'(i)| = \| \beta^* \|_1. \] (R.1)
Relax this to the following problem which can potentially get a smaller objective function value
\[ \min_{\mathbf{v}'} \sum_{i \in S'} (|\hat{\beta}(i)| - a_i)^2 + \sum_{i \notin S'} \hat{\beta}^2(i) \quad \text{subject to} \quad \sum_{i \in S'} a_i = \| \beta^* \|_1, \]
where \( a_i \in \mathbb{R} \) (here we lose the positivity of \( a_i \) from problem (R.1)).

Use Lagrange multipliers we obtain the Lagrangian
\[ L = \sum_{i \in S'} (|\hat{\beta}(i)| - a_i)^2 + \sum_{i \notin S'} \hat{\beta}^2(i) + \lambda(\sum_{i \in S'} a_i - \| \beta^* \|_1), \]

55
and solve \( \frac{\partial L}{\partial a_i} = 0 \) to get
\[
a_i = |\hat{\beta}_i| + \lambda \text{ for all } i \in S'.
\]
Combine it with the fact that \( \sum_{i \in S'} a_i = \| \beta^* \|_1 \), we have
\[
\lambda = \frac{\| \beta^* \|_1 - \sum_{i \in S'} |\hat{\beta}_i|}{s} > 0,
\]
where the last inequality follows since \( \| \beta^* \|_1 \geq \| \hat{\beta} \|_1 \). It follows that \( a_i \geq 0 \), and thus the minimum for problem (R.1) is also achieved at the same point. Hence at the optimal point we have \( \| v' - \hat{\beta} \| = \sqrt{s \lambda^2 + \sum_{i \in S'} \hat{\beta}^2_i} \). Note that when \( S' = S \) is the set of indices of the \( s \) most significant coordinates both \( \lambda \) and \( \sum_{i \notin S'} \hat{\beta}^2_i \) are minimized. This completes the proof.

\[\square\]

**S Proof of Theorem 5.2**

For the SLOPE estimator, we combine the results in Corollary 6.2 in Bellec et al. (2018b). With probability at least \( 1 - \frac{1}{2}(s^u/n)^{3/2} \) we have
\[
\| \hat{\beta} - \beta^* \| \lesssim \sigma \sqrt{s^u \log(2ep/s^u)} / n. \tag{S.1}
\]

For the square-root SLOPE estimator, we use the result in (Derumigny et al., 2018, Corollary 6.2). With probability at least \( 1 - (s^u/n)^{3/2} (1 + e^2) e^{-n/24} \) we have the same rate as (S.1). It follows that when \( C \geq \sigma/\beta^* \), \( \beta^* \) will be a feasible point. Hence \( \beta^* \in K = \{ \beta : \| \beta \|_1 \leq \| v \|_1 \} \). Next, since \( \| v - \hat{\beta} \| \leq C \sqrt{s^u \log(2ep/s^u)/n} \) is guaranteed in step 1, by triangle inequality we have
\[
\| \beta^* - v \| \leq \| \hat{\beta} - v \| + \| \beta - \beta^* \| \lesssim (C + \sigma/\beta^*) \sqrt{s^u \log(2ep/s^u)/n}.
\]

For \( \overline{w}(T_K(v) \cap S^{p-1}) \), since \( K = \{ \beta : \| \beta \|_1 \leq \| v \|_1 \} \) and \( v \) is at least \( s^u \) sparse, by (Chandrasekaran et al., 2012, Proposition 3.10) we have
\[
\overline{w}(T_K(v) \cap S^{p-1}) \lesssim \sqrt{s^u \log \frac{ep}{s^u}}.
\]
Finally since \( s^u = o(\sqrt{n} \log(ep/s^u)) \) we have
\[
\overline{w}(T_K(v) \cap S^{p-1}) \| v - \beta^* \| \lesssim \frac{s^u \log ep/s^u}{\sqrt{n}} C = o_p(1),
\]
by assumption. This completes the proof.
Proof of Lemma 5.4

Let \( \mathbf{v}_s \) be a vector the set of vectors with \( s \) non-zero coordinates. Recall the optimization problem

\[
\arg\max \| \mathbf{v} \|_1, \quad \text{s.t.} \quad \| \mathbf{v} - \hat{\beta} \| \leq C \sqrt{su \log 2ep/\nu} \quad \text{and} \quad \| \mathbf{v} \|_0 \leq su.
\]

Let \( \mathbf{v}_s \) be an \( s < su \) sparse vector candidate for being the solution of the program above. First for each coordinate of \( \mathbf{v}_s \), we have either \( \text{sign}(\mathbf{v}_s^{(i)}) = \text{sign}(\hat{\beta}^{(i)}) \), or \( \text{sign}(\mathbf{v}_s^{(i)}) = 0 \), because otherwise we can always change that coordinate to \( -\text{sign}(\mathbf{v}_s^{(i)})(|\mathbf{v}_s^{(i)}| + 2|\hat{\beta}^{(i)}|) \) to make \( \| \mathbf{v}_s \|_1 \) larger while keeping \( \| \mathbf{v}_s - \hat{\beta} \| \) unchanged.

Then we show that the non-zero indices in \( \mathbf{v}_s \) have the form \( \mathbf{v}_s^{(i)} = \hat{\beta}^{(i)} + \text{sign}(\hat{\beta}^{(i)})c \) for some \( c \geq 0 \). Let \( S' \) with \( |S'| = s \) be the set of non-zero coordinates of \( \mathbf{v}_s \). The optimization program becomes

\[
\arg\max_{\mathbf{v}_s} \sum_{i \in S'} |\mathbf{v}_s^{(i)}|, \quad \text{s.t.} \quad \sum_{i \in S'} (|\hat{\beta}^{(i)}| - |\mathbf{v}_s^{(i)}|)^2 + \sum_{i \notin S'} \hat{\beta}^2_{(i)} \leq C \sqrt{su \log 2ep/\nu}.
\]

Relax the above problem to

\[
\arg\max_{\mathbf{a}_i} \sum_{i \in S'} a_i, \quad \text{s.t.} \quad \sum_{i \in S'} (|\hat{\beta}^{(i)}| - a_i)^2 + \sum_{i \notin S'} \hat{\beta}^2_{(i)} \leq C \sqrt{su \log 2ep/\nu},
\]

where \( a_i \) need not be positive. Using Lagrange multipliers we obtain

\[
L = \sum_{i \in S'} a_i + \lambda \sum_{i \in S'} (|\hat{\beta}^{(i)}| - a_i)^2,
\]

and solve \( \frac{\partial L}{\partial a_i} = 0 \) to get

\[
a_i = |\hat{\beta}^{(i)}| + \frac{1}{2\lambda} \quad \text{for all} \quad i \in S'.
\]

Let \( c = \frac{1}{2\lambda} \). We have that

\[
sc^2 + \sum_{i \notin S'} \hat{\beta}^2_{(i)} \leq C \sqrt{su \log 2ep/\nu}.
\]

Hence the maximal value of \( c \) satisfies \( c^2 = C \sqrt{su \log 2ep/\nu} - \sum_{i \notin S'} \hat{\beta}^2_{(i)} \geq 0 \). The latter is \( \geq 0 \) if there exists a feasible point in the program. When \( C \sqrt{su \log 2ep/\nu} < \sum_{i \notin S'} \hat{\beta}^2_{(i)} \) then the vector with support \( S' \) can never be feasible in any case.

Note that our objective function is

\[
\sum_{i \in S'} a_i = \sum_{i \in S'} |\hat{\beta}^{(i)}| + sc,
\]

57
which is maximized when \( c = \sqrt{\frac{\log 2}{n}}. \) It is also clear that in the above, one should pick \( S' \) which minimizes the coefficients of \( \sum_{i \in S'} \beta_i^2 \) and at the same time, maximizes \( \sum_{i \in S'} |\beta_i| \). Clearly, this set corresponds to the maximal in magnitude elements in the vector \( \hat{\beta} \). Since \( a_i \) are positive then one can find the corresponding maximal values of \( |v_i| = a_i \), and \( v_i = \hat{\beta}_i + \text{sign}(\hat{\beta}_i)c \) on the set \( S' \) where the largest \( s \) coefficients of \( \hat{\beta} \) are located. Furthermore, the bigger the \( s \) the bigger the objective function. Hence we take \( s = s^* \). This completes the proof.

**U Proof of Lemma 5.5**

According to the results in (Bellec et al., 2018b, Corollary 6.2) and (Derumigny et al., 2018, Corollary 6.2), with probability converging to 1, the quantity \( \frac{1}{\sqrt{n}} \|\overline{X}(\hat{\beta} - \beta^*)\| \) can be bounded as

\[
\frac{1}{\sqrt{n}} \|\overline{X}(\hat{\beta} - \beta^*)\| \lesssim \frac{\sigma}{\vartheta^*} \sqrt{\frac{s^u \log(2e\varrho/s^u)}{n}}.
\]

conditional on \( \overline{X} \) satisfying the WRE with \( \vartheta^* \), where \( \vartheta^* \) is defined in the main text and is \( \vartheta(s^u,3) \) for the LASSO, and \( \vartheta(s^u,20) \) for square-root SLOPE. From (Bellec et al., 2018b, Theorem 8.3) and the assumptions of Remark 5.3, we know that \( \vartheta^* \geq \kappa/\sqrt{2} \) with high probability and \( \overline{X} \) satisfies the WRE condition. This is what we wanted to show.

**V Proof of Lemma 6.1**

It suffices to show that for a sufficiently large \( \rho' \) the constraint \( \|\overline{X}\Sigma^{-1}e^{(j)}\|_\infty \leq \rho' \sqrt{\log n} \) contains a \( \delta \ell_2 \)-ball, since we have proved that the other set contains a small ball around the point \( \Sigma^{-1}e^{(j)} \) in Corollary 2.4.

1. **Feasible Point:**

   We argue that \( \eta = \Sigma^{-1}e^{(j)} \) is a feasible point since \( \|\overline{X}\Sigma^{-1}e^{(j)}\|_\infty \leq \rho' \sqrt{\log n} \) with probability converging to one. Notice that each coordinate of \( \overline{X}\Sigma^{-1}e^{(j)} \) is a sub-Gaussian variable since

   \[
   \|\overline{X}\Sigma^{-1}e^{(j)}\|_{\psi_2} \leq \|e^{(j)}^T \Sigma^{-1}\| \|\overline{X}_i\|_{\psi_2} = \sqrt{\Sigma^{-2}} \|\overline{X}_i\|_{\psi_2} = O(1).
   \]

   Since \( \Sigma \) has bounded eigenvalues so does \( \Sigma^{-2} \), and hence all of its entries should be bounded, thus \( \Sigma^{-2} \) is bounded. And since \( \overline{X}_i \) is either a bounded or Gaussian, which both belong to the sub-Gaussian category, \( \overline{X}_i \) is sub-Gaussian. Therefore, \( \|\overline{X}\Sigma^{-1}e^{(j)}\|_{\psi_2} \) is bounded for all \( i \in [n] \), or in other words each coordinate of \( \overline{X}\Sigma^{-1}e^{(j)} \) is sub-Gaussian. By the concentration inequality of maximum sub-Gaussian variables (Duchi, 2017, p. 14), with probability converging to one

   \[
   \max_{i \in [n]} |(\overline{X}\Sigma^{-1}e^{(j)})_i| \lesssim \sqrt{\log n}.
   \]
Thus for a sufficiently large $\rho'$ we have

$$\|\bar{X} \Sigma^{-1} e^{(j)}\|_\infty \leq \rho' \sqrt{\log n}.$$  

2. Non-empty Interior:

We are able to find $\eta = \Sigma^{-1} e^{(j)}$ as a feasible point. Now the idea is to show that there exists a small $\delta > 0$ such that $B_\delta(\eta) \cap \Sigma^{-1}$ is still inside of the feasible region with high probability. Now let $\mathbf{x}$ be a unit vector. We have

$$\|\bar{X}(\Sigma^{-1} e^{(j)} + \delta \mathbf{x})\|_\infty \leq \rho' \sqrt{\log n} + \delta \|\bar{X}\mathbf{x}\|_\infty.$$  

Picking $\delta = \rho' \sqrt{\log n} / \sup_{\|\mathbf{x}\|_1} \|\bar{X}\mathbf{x}\|_\infty$ shows that for the value $2\rho'$ the set has non-empty interior. This completes the proof.

W Proof of Lemma 6.2

This fact follows by a direct calculation. We omit the details.

X Proof of Theorem 6.3

We state and prove the following result. Its proof rests on an argument from (Javanmard and Montanari, 2014, Lemma 3.1).

**Lemma X.1.** The following holds:

$$\left\| \widehat{\Sigma}^{1/2} \eta \right\|^2 \geq \sup_{\mathbf{u} \in \mathcal{T}_K(v) \cap \mathbb{S}^{p-1}} (|u_j| - \rho \lambda)^2 \mathbb{1}\{|u_j| \geq \rho \lambda\} \geq \frac{\left(\|u^*_j| - \rho \lambda\|^2 \mathbb{1}\{|u_j| \geq \rho \lambda\}\right)}{\mathbf{u}^* \widehat{\Sigma} \mathbf{u}^*},$$

where $\lambda = \frac{w(T_K(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}}$, and $u^* = \frac{\beta^*_v}{\|\beta^*_v\|}$.

**Proof.** Let $\lambda = \frac{w(T_K(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}}$. The constraint $\sup_{\mathbf{u} \in \mathcal{T}_K(v) \cap \mathbb{S}^{p-1}} \|\eta \widehat{\Sigma} - e^{(j)}\| \mathbf{u} \leq \rho \frac{w(T_K(v) \cap \mathbb{S}^{p-1})}{\sqrt{n}}$ implies

$$u_j - \langle \mathbf{u}, \widehat{\Sigma} \eta \rangle \leq \rho \lambda, \quad \mathbf{u} \in \mathcal{T}_K(v) \cap \mathbb{S}^{p-1}$$

or

$$-\rho \lambda \leq u_j - \langle \mathbf{u}, \widehat{\Sigma} \eta \rangle, \quad \mathbf{u} \in \mathcal{T}_K(v) \cap \mathbb{S}^{p-1}. $$

Consider the first case $u_j - \langle \mathbf{u}, \widehat{\Sigma} \eta \rangle \leq \rho \lambda$. Then for any feasible $\tilde{\eta}$ and $c \geq 0$, when $\mathbf{u} \in \mathcal{T}_K(v) \cap \mathbb{S}^{p-1}$ we have

$$\tilde{\eta} \widehat{\Sigma} \eta \geq \tilde{\eta} \widehat{\Sigma} \eta + c(u_j - \rho \lambda) - c(\mathbf{u}, \widehat{\Sigma} \eta)$$

$$\geq \min_{\eta: \sup_{\mathbf{u} \in \mathcal{T}_K(v) \cap \mathbb{S}^{p-1}} (u_j - \rho \lambda) - \langle \mathbf{u}, \widehat{\Sigma} \eta \rangle \leq 0} [\eta \widehat{\Sigma} \eta + c(u_j - \rho \lambda) - c(\mathbf{u}, \widehat{\Sigma} \eta)].$$

Thus the optimal value of the optimization (6.1) in step 2 satisfies

$$\left\| \widehat{\Sigma}^{1/2} \tilde{\eta} \right\|^2 \geq \min_{\eta: \sup_{\mathbf{u} \in \mathcal{T}_K(v) \cap \mathbb{S}^{p-1}} (u_j - \rho \lambda) - \langle \mathbf{u}, \widehat{\Sigma} \eta \rangle \leq 0} [\eta \widehat{\Sigma} \eta + c(u_j - \rho \lambda) - c(\mathbf{u}, \widehat{\Sigma} \eta)].$$

59
When $\eta = cu/2$, the RHS is minimized. Thus
\[ \|\hat{\Sigma}^{1/2}\hat{\eta}\|^2 \geq c(u_j - \rho\lambda) - \frac{c^2}{4} u^\top \hat{\Sigma} u, \quad \text{if } u_j - \rho\lambda \leq \frac{c}{2} u^\top \hat{\Sigma} u. \]

We then optimize over $c$. When $c = 2(u_j - \rho\lambda)/u^\top \hat{\Sigma} u$, the condition $u_j - \rho\lambda \leq \frac{c}{2} u^\top \hat{\Sigma} u$ holds for any $u$. And since we need $c \geq 0$, the condition $u_j \geq \rho\lambda$ should hold. Plug in the value of $c$ to the RHS, we get
\[ \|\hat{\Sigma}^{1/2}\hat{\eta}\|^2 \geq (u_j - \rho\lambda)^2 u^\top \hat{\Sigma} u / BD \{ u_j \geq \rho\lambda \}. \]

Similarly for the second case $-\rho\lambda \leq u_j - (u, \hat{\Sigma}\eta)$ we will get
\[ \|\hat{\Sigma}^{1/2}\hat{\eta}\|^2 \geq (-u_j - \rho\lambda)^2 u^\top \hat{\Sigma} u / BD \{-u_j \geq \rho\lambda \}. \]

Finally
\[ \|\hat{\Sigma}^{1/2}\hat{\eta}\|^2 \geq (|u_j| - \rho\lambda)^2 u^\top \hat{\Sigma} u / BD \{|u_j| \geq \rho\lambda \}. \]

\[ \square \]

**Lemma X.2.** Suppose $X_i$ has a covariance matrix $\Sigma$, and the eigenvalues of $\Sigma$ are bounded. $u^*$ is defined as in Lemma X.1. Then conditionally on $X$ we have that $\lambda_{\text{min}}(\Sigma)/2 \leq u^* \Sigma u^* \leq 3/2 \|\Sigma\|_{\text{op}}$ with high probability.

**Proof.** Since conditionally on $X$ we have that $v$ is independent of $\hat{\Sigma}$, and $(u^* X_i)^2$ is a sub-exponential random variable (with norm less than $K := \|X_i\|_{\psi_2}^2$ which is bounded by assumption), we can use a Bernstein type of concentration inequality to claim that
\[ \mathbb{P}\left( \left| \frac{1}{n} \sum_{i \in [n]} (u^* X_i)^2 - \mathbb{E}[(u^* X_i)^2 | X] \right| \geq t \right) \leq \exp(-c nt^2 / K^2 \wedge t / K). \]

Choose $t = \lambda_{\text{min}}(\Sigma)/2$, and note that $\lambda_{\text{max}}(\Sigma) \geq \mathbb{E}[(u^* X_i)^2 | X] \geq \lambda_{\text{min}}(\Sigma)$, completing the proof. \[ \square \]

**Theorem X.3** (Lindeberg-Feller CLT). (Greene, 2003, p. 901) Let $X_1, \ldots, X_n$ be independent but not necessarily identically distributed random variables with $\mathbb{E}[X_i] = \mu_i$ and $\text{Var}(X_i) = \sigma_i^2 < \infty$. Define $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $\bar{\sigma}^2_n = n^{-1} \sum_{i=1}^n \sigma_i^2$. Suppose
\[ \lim_{n \to \infty} \max_i \frac{\sigma_i^2}{n \bar{\sigma}^2_n} = 0, \quad \lim_{n \to \infty} \bar{\sigma}^2_n < \infty. \]

Then
\[ \sqrt{n} \left( \frac{\bar{X}_n - \bar{\mu}_n}{\bar{\sigma}_n} \right) \overset{d}{\to} Z \sim N(0, 1). \]
The proof of Theorem 6.3 starts here. We divide the proof into two cases in terms of the scale of \(\|\hat{\Sigma}^{1/2} \tilde{\eta}\|\). A sufficiently large \(\|\hat{\Sigma}^{1/2} \tilde{\eta}\|\) is required if one would like to use Lindeberg-Feller CLT to derive the limiting distribution of \(\sqrt{n}(\hat{\beta}_d^{(j)} - \beta^{(j)})\).

**Proof.** Let \(a_n = o(1)\) be any slowly converging to 0 rate such that \(\frac{1}{\sigma_n} = o\left(\frac{n}{\log n}\right)\).

1. Suppose now that

\[
\|\hat{\Sigma}^{1/2} \tilde{\eta}\| \leq C_1 \frac{\sqrt{\log n}/\sqrt{\|\beta^* - v\| \log n}}{\sqrt{n}} \vee a_n,
\]

for some constant \(C_1\). Then by Lemma X.1 and Lemma X.2, for some constant \(C'\) we have

\[
(|u_j^*| - p\lambda) \mathbb{I}(|u_j^*| > p\lambda) \leq C' \frac{\sqrt{\log n}/\sqrt{\|\beta^* - v\| \log n}}{\sqrt{n}} \vee a_n.
\]

Plug in \(u^* = \frac{\beta^* - v}{\|\beta^* - v\|}\) to get

\[
|\beta_j^* - v_j| \leq \|\beta^* - v\| C' \frac{\sqrt{\log n}/\sqrt{\|\beta^* - v\| \log n}}{\sqrt{n}} \vee a_n + \|\beta^* - v\| p\lambda.
\]

Given that \(\|\beta^* - v\| \max\{\bar{w}(T_K(v) \cap \mathbb{S}^{p-1}), \sqrt{\log n}\} = o_p(1)\), we have

\[
|\beta_j^* - v_j| = o_p(1/\sqrt{n}),
\]

so \(v_j\) is more precise than what we need already.

Then we show that the debiased estimator \(\tilde{\beta}_d^{(j)} \leftarrow e^{(j)\top} v + n^{-1} \eta^\top \bar{X}^\top (\bar{Y} - \bar{X} v)\) is still \(o_p(1/\sqrt{n})\) close to \(\beta_j^*\) since the correction term

\[
n^{-1} \eta^\top \bar{X}^\top (\bar{Y} - \bar{X} v) = o_p(1/\sqrt{n}).
\]

We have

\[
n^{-1} \eta^\top \bar{X}^\top (\bar{Y} - \bar{X} v) \leq \eta^\top \bar{X}^\top \frac{\|\beta^* - v\| \vee a_n}{\sqrt{n}} + n^{-1} \|\hat{\Sigma}^{1/2} \tilde{\eta}\| \sum_{i \in n} (\bar{X}_n \tilde{\eta}_{\cdot i}) \epsilon_i,
\]

where \(u^* = \frac{\beta^* - v}{\|\beta^* - v\|}\).

The first term can be bounded as follows. The first line uses the first constraint in step 2, and the second line uses Lemma X.1 and Lemma X.2. Suppose the upper bound of \(\sqrt{u^* \Sigma u^*}\) is \(C_3\) for a constant \(C_3 > 0\).

\[
|\tilde{\eta}^\top \bar{X}^\top u^*| \|\beta^* - v\| \leq (\rho \lambda + |u_j^*|) \|\beta^* - v\|
\]

\[
\leq (\rho \lambda + C_3(\|\hat{\Sigma}^{1/2} \eta\| + \rho \lambda)) \|\beta^* - v\|,
\]

61
Since \( \| \beta^* - v \| \mathbb{E}(T_K(v) \cap S^{p-1}) = o_p(1) \), we have \( \lambda \| \beta^* - v \| = o_p(1/\sqrt{n}) \). And by the condition of \( \| \hat{\Sigma}^{1/2} \theta \| \) we have \( \| \hat{\Sigma}^{1/2} \beta^* - v \| = o_p(1/\sqrt{n}) \) as well. Thus the above quantity is \( o_p(1/\sqrt{n}) \).

For the second term, by the condition \( \frac{1}{a_n} = o(\frac{n}{\log n}) \) we have \( \| \hat{\Sigma}^{1/2} \theta \| = o_p(1) \). Notice that \( \| X \eta_n \| \) is sub-Gaussian condition on \( X \). This is because \( \epsilon_i \) is sub-Gaussian, it is independent of \( X \) and the coefficients satisfies \( \sum_{i \in \mathbb{N}} ( \| X \eta_n \| )^2 = 1 \).

Hence we have established that

\[
\hat{\beta}(j) - \beta^* = o_p(1/\sqrt{n}),
\]
so any confidence interval centering at \( \hat{\beta}(j) \) with length \( O(1/\sqrt{n}) \) will contain \( \beta^* \). Even though such a confidence interval might not be very efficient since \( \hat{\beta}(j) \) converges faster than the rate \( 1/\sqrt{n} \).

To make sure the confidence interval is of the length \( O(1/\sqrt{n}) \), one can pick some small constant \( c > C'(\log n)^{1/2}/\sqrt{(\| \beta^* - v \| / \sqrt{\log n}) \lor a_n} \) and make the confidence intervals as \( (6.2) \).

2. Suppose now that

\[
\| \hat{\Sigma}^{1/2} \theta \| \geq C \frac{\log n / \sqrt{\| \beta^* - v \| / \sqrt{\log n}) \lor a_n}}{\sqrt{n}}.
\]

In that case it follows

\[
\| X \theta_n \| / (\sqrt{n} \| \hat{\Sigma}^{1/2} \theta \|) \lesssim (\| \beta^* - v \| / \sqrt{\log n}) \lor a_n = o_p(1),
\]
so we can apply the Lindeberg-Feller CLT (Theorem X.3). Let \( Z_j = \frac{1}{\sqrt{n}} \theta_n^\top X^\top \epsilon \), we have

\[
\sqrt{n}(\hat{\beta}(j) - \beta^*(j)) = Z_j + \Delta_j, \quad \Delta_j = \sqrt{n}(\theta_n^\top \hat{\Sigma} - e(j)^\top)(\beta^* - v).
\]

\( \Delta_j \) converges to zero with probability converging to one since \( \| \beta^* - v \| \mathbb{E}(T_K(v) \cap S^{p-1}) = o_p(1) \).

And \( Z_j \) is Gaussian conditional on \( X, \hat{\Sigma}, \hat{\eta} \) by the Lindeberg-Feller CLT (Theorem X.3)

\[
\frac{Z_j}{\sigma \| \hat{\Sigma}^{1/2} \theta \|} \xrightarrow{d} N(0,1).
\]

Thus the confidence interval \( (6.2) \) also applies in this case.

\[
\square
\]

References

Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. Information and Inference: A Journal of the IMA, 3(3):224–294.
Bellec, P. C. et al. (2018a). Sharp oracle inequalities for least squares estimators in shape restricted regression. The Annals of Statistics, 46(2):745–780.

Bellec, P. C., Lecué, G., Tsybakov, A. B., et al. (2018b). Slope meets lasso: improved oracle bounds and optimality. The Annals of Statistics, 46(6B):3603–3642.

Bellec, P. C. and Tsybakov, A. B. (2015). Sharp oracle bounds for monotone and convex regression through aggregation. J. Mach. Learn. Res., 16:1879–1892.

Bellec, P. C. and Zhang, C.-H. (2019a). De-biasing the lasso with degrees-of-freedom adjustment. arXiv preprint arXiv:1902.08885.

Bellec, P. C. and Zhang, C.-H. (2019b). Second order poincaré inequalities and de-biasing arbitrary convex regularizers when $p/n \to \gamma$. arXiv preprint arXiv:1912.11943.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81(2):608–650.

Belloni, A., Chernozhukov, V., and Kato, K. (2015). Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. Biometrika, 102(1):77–94.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., et al. (2009). Simultaneous analysis of lasso and dantzig selector. The Annals of statistics, 37(4):1705–1732.

Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). Slope—adaptive variable selection via convex optimization. The annals of applied statistics, 9(3):1103.

Boucheron, S., Lugosi, G., and Massart, P. (2013). Concentration inequalities: A nonasymptotic theory of independence. Oxford university press.

Boyd, S., Xiao, L., and Mutapcic, A. (2003). Subgradient methods. lecture notes of EE392o, Stanford University, Autumn Quarter, 2004:2004–2005.

Bradic, J., Fan, J., and Zhu, Y. (2018). Testability of high-dimensional linear models with non-sparse structures. arXiv preprint arXiv:1802.09117.

Bühlmann, P. and Van De Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications, Springer Science & Business Media.

Bunea, F., Tsybakov, A., Wegkamp, M., et al. (2007). Sparsity oracle inequalities for the lasso. Electronic Journal of Statistics, 1:169–194.

Cai, T. T., Guo, Z., et al. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. The Annals of statistics, 45(2):615–646.

Celentano, M., Montanari, A., and Wei, Y. (2020). The lasso with general gaussian designs with applications to hypothesis testing. arXiv preprint arXiv:2007.13716.

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

Chernoff, H. (1954). On the distribution of the likelihood ratio. The Annals of Mathematical Statistics, pages 573–578.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Derumigny, A. et al. (2018). Improved bounds for square-root lasso and square-root slope. Electronic Journal of Statistics, 12(1):741–766.

Duchi, J. (2017). Stanford cs229t, lecture notes: Concentration inequalities and tail bounds. URL: https://web.stanford.edu/class/cs229t/2017/Lectures/concentration-slides.pdf.

Eftekhari, H., Banerjee, M., and Ritov, Y. (2021). Inference in high-dimensional single-index models under symmetric designs. Journal of Machine Learning Research, 22(27):1–63.
Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5):849–911.

Feng, H. and Ning, Y. (2019). High-dimensional mixed graphical model with ordinal data: Parameter estimation and statistical inference. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 654–663.

Gao, C., Han, F., and Zhang, C.-H. (2017). Minimax risk bounds for piecewise constant models. arXiv preprint arXiv:1705.06386.

Genzel, M. and Kipp, C. (2020). Generic error bounds for the generalized lasso with sub-exponential data. arXiv preprint arXiv:2004.05361.

Geyer, C. J. et al. (1994). On the asymptotics of constrained $m$-estimation. The Annals of Statistics, 22(4):1993–2010.

Gordon, Y. (1988). On milman’s inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In Geometric Aspects of Functional Analysis, pages 84–106. Springer.

Greene, W. H. (2003). Econometric analysis. Pearson Education India.

Hettich, R. and Kortanek, K. O. (1993). Semi-infinite programming: theory, methods, and applications. SIAM review, 35(3):380–429.

Jankova, J., Van De Geer, S., et al. (2015). Confidence intervals for high-dimensional inverse covariance estimation. Electronic Journal of Statistics, 9(1):1205–1229.

Jankova, J., Van De Geer, S., et al. (2018). Semiparametric efficiency bounds for high-dimensional models. Annals of Statistics, 46(5):2336–2359.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. The Journal of Machine Learning Research, 15(1):2869–2909.

Javanmard, A., Montanari, A., et al. (2018). Debiasing the lasso: Optimal sample size for gaussian designs. The Annals of Statistics, 46(6A):2593–2622.

Kiefer, J. (1953). Sequential minimax search for a maximum. Proceedings of the American mathematical society, 4(3):502–506.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. Annals of statistics, pages 1356–1378.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8):30–37.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. (2016). Exact post-selection inference, with application to the lasso. Annals of Statistics, 44(3):907–927.

Lee, J. D. and Taylor, J. E. (2014). Exact post model selection inference for marginal screening. arXiv preprint arXiv:1402.5596.

Li, Y.-H., Hsieh, Y.-P., Zerbib, N., and Cevher, V. (2015). A geometric view on constrained $m$-estimators. arXiv preprint arXiv:1506.08163.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. Annals of statistics, 42(2):413.

Lustig, M., Donoho, D. L., Santos, J. M., and Pauly, J. M. (2008). Compressed sensing mri. IEEE signal processing magazine, 25(2):72–82.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4):417–473.
Meinshausen, N., Bühlmann, P., et al. (2006). High-dimensional graphs and variable selection with the lasso. The annals of statistics, 34(3):1436–1462.

Mendelson, S. (2010). Empirical processes with a bounded $\psi$ 1 diameter. Geometric and Functional Analysis, 20(4):988–1027.

Moreau, J. J. (1962). Decomposition orthogonale d’un espace hilbertien selon deux cones mutuellement polaires. In Comptes rendus hebdomadaires des seances de l’Academie des sciences, pages 238–240.

Németh, A. and Németh, S. (2012). How to project onto the monotone nonnegative cone using pool adjacent violators type algorithms. arXiv preprint arXiv:1201.2343.

Nesterov, Y. (2003). Introductory lectures on convex optimization: A basic course, volume 87. Springer Science & Business Media.

Neykov, M. (2019). Gaussian regression with convex constraints. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 31–38.

Neykov, M., Ning, Y., Liu, J. S., Liu, H., et al. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. Statistical Science, 33(3):427–443.

Ning, Y., Liu, H., et al. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. The Annals of Statistics, 45(1):158–195.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. The annals of applied statistics, 4(1):53.

Plan, Y. and Vershynin, R. (2016). The generalized lasso with non-linear observations. IEEE Transactions on information theory, 62(3):1528–1537.

Polyak, B. T. (1967). A general method for solving extremal problems. In Doklady Akademii Nauk, volume 174, pages 33–36. Russian Academy of Sciences.

Robertson, T. (1988). Order restricted statistical inference. Technical report, No. 04; QA278. 7, R6.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association, 82(398):605–610.

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(1):55–80.

Shor, N. Z. (2012). Minimization methods for non-differentiable functions, volume 3. Springer Science & Business Media.

Stucky, B. and Van De Geer, S. (2017). Sharp oracle inequalities for square root regularization. The Journal of Machine Learning Research, 18(1):2256–2284.

Thrampoulidis, C., Oymak, S., and Hassibi, B. (2014). Simple error bounds for regularized noisy linear inverse problems. In 2014 IEEE International Symposium on Information Theory, pages 3007–3011. IEEE.

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

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. The Annals of Statistics, 42(3):1166–1202.
Van de Geer, S. A. et al. (2008). High-dimensional generalized linear models and the lasso. The Annals of Statistics, 36(2):614–645.

Van der Vaart, A. W. (2000). Asymptotic statistics, volume 3. Cambridge university press.

Vershynin, R. (2018). High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge University Press.

Wainwright, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. IEEE Transactions on Information Theory, 55(12):5728–5741.

Wainwright, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge University Press.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. Annals of statistics, 37(5A):2178.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics, 38(2):894–942.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):217–242.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. Journal of Machine learning research, 7(Nov):2541–2563.

Zhu, Y. and Bradic, J. (2018). Linear hypothesis testing in dense high-dimensional linear models. Journal of the American Statistical Association, 113(524):1583–1600.

Zou, H. and Hastie, T. J. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. B, 67(2):301–320.