*Submitted to Bernoulli*

# Supplement to "Design of c-Optimal Experiments for High-dimensional Linear Models"

HAMID EFTEKHARI*, MOULINATH BANERJEE and YA'ACOV RITOV

*Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA*
*E-mail:* hamidef@umich.edu; moulib@umich.edu; yritov@umich.edu

The technical proofs and background material for the main paper are collected in this supplement.

*Keywords:* Optimal Design; Inference; Sparsity; Compressed Sensing

## 1. Proofs

For completeness, we record here Elfving's theorem [3] which has been alluded to in the main text.

**Theorem 2** (Elfving (1952)). *Let $(x_i)_1^N \in \mathbf{R}^p$ be the available design points and let $c \in \mathbf{R}^p$. Define the Elfving set to be the convex hull of $(\pm x_i)_1^N$:*

$$\mathcal{E} := \operatorname{conv}\left(\{x_i : 1 \le i \le N\} \cup \{-x_i : 1 \le i \le N\}\right).$$

*Let $x_c$ be the point on the boundary of $\mathcal{E}$ that intersects the half-line passing through the origin and $c$:*

$$x_c = \partial\mathcal{E} \cap \{tc : t \ge 0\}.$$

*If we write $x_c = \sum_{i=1}^N v_i x_i$, then the c-optimal design is given by $w_i^\star := |v_i| / \sum_1^N |v_j|$.*

***Definition* 1** (Restricted Eigenvalue Condition). A matrix $A$ is said to satisfy the restricted eigenvalue condition $RE(s_0, k_0, A)$with parameter $\lambda_{RE}$ if

$$\lambda_{RE} := \min_{\substack{J \subset \{1,\dots,P\} \\ |J| \le s_0}} \min_{\substack{\|v_{J^c}\|_1 \le k_0 \|v_J\|_1 \\ v \ne 0}} \frac{\|Av\|_2}{\|v_J\|_2} > 0.$$

We also denote this quantity by $\lambda_{RE}(s_0, k_0, A)$.

We record here a theorem by Rudelson and Zhou [4] that relates the restricted eigenvalues of random matrices to the (restricted) eigenvalues of the corresponding population covariance matrices. In the theorem the smallest $k$-sparse eigenvalue of $A$ is defined as

$$\rho_{\min}(k, A) = \min_{\substack{\|t\|_0 \le k \\ t \ne 0}} \frac{\|At\|_2}{\|t\|_2}.$$

**Theorem 3** (Rudelson & Zhou, Theorem 8)**.** *Let $0 < \delta < 1$ and $0 < s_0 < p$. Let $X \in \mathbf{R}^p$ be a random vector such that $\|X\|_\infty \le M$ a.s. and denote $\Sigma = \mathbf{E}XX^T$. Let $\mathbb{X}$ be an $n \times p$ matrix whose rows $X_1, X_2, \dots, X_n$ are independent copies of $X$. Let $\Sigma$ satisfy the $RE(s_0, 3k_0, \Sigma^{\frac{1}{2}})$ condition as in Definition 1. Define*

$$d = s_0 \left(1 + \max_j \|\Sigma^{\frac{1}{2}} e_j\|_2^2 \frac{16(3k_0)^2(3k_0+1)}{\delta^2 \cdot \lambda_{RE}^2(s_0, k_0, \Sigma^{\frac{1}{2}})}\right).$$

*Assume that $d \le p$ and $\rho = \rho_{\min}(d, \Sigma^{\frac{1}{2}}) > 0$. Assume that the sample size $n$ satisfies*

$$n \ge n_0 := \frac{C_{RZ} M^2 d \cdot \log p}{\rho^2 \delta^2} \cdot \log^3 \left(\frac{C_{RZ} M^2 d \cdot \log p}{\rho^2 \delta^2}\right),$$

*for an absolute constant $C_{RZ}$. Then with probability at least $1 - \exp\left(-\delta\rho^2 n/(6M^2 d)\right)$, the $RE(s_0, k_0, \mathbb{X}/\sqrt{n})$ condition holds for matrix $\mathbb{X}/\sqrt{n}$ with $\lambda_{RE}(s_0, k_0, \mathbb{X}/\sqrt{n}) \ge (1-\delta) \cdot \lambda_{RE}(s_0, k_0, \Sigma^{\frac{1}{2}})$.*

Note that this theorem concerns observations obtained via i.i.d sampling. The next proposition shows that a similar result holds for Poisson sampling as used in our work. Since the usual Lasso guarantees require $RE(s_0, k_0 = 3, \mathbb{X}/\sqrt{n})$, we will be using the above theorem with $k_0 = 3$.

**Proposition 1.** *Suppose that $K_1, \dots, K_N$ are independent Poisson random variables with $K_j \sim \mathrm{Poisson}(nw_j)$ and $\sum_1^N w_j = 1$, so that $K := \sum_1^N K_j \sim \mathrm{Poisson}(n)$. Let $\mathbb{X}$ be a $K \times p$ matrix where $x_j^T$ is repeated in the rows of $\mathbb{X}$ exactly $K_j$ times. Suppose that*

- *The population covariance matrix $\Sigma = \sum_{j=1}^N w_j x_j x_j^T$ satisfies*

$$\lambda_\star \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le \lambda^\star.$$

- *The expected sample size $n$ satisfies*

$$n \ge \frac{5}{4}\tilde{n}_0 \quad \text{where } \tilde{n}_0 = \frac{\tilde{C} M^2 \lambda^\star s_0 \log p}{\lambda_\star^2} \log^3 \left(\frac{\tilde{C} M^2 \lambda^\star s_0 \log p}{\lambda_\star^2}\right),$$

$$\text{and } \tilde{C} = 4 \times 51841 C_{RZ}.$$

*Then with probability at least $1 - e^{-\frac{\tilde{n}_0}{4}} - e^{-\frac{n_0 \lambda_\star}{12 M^2 d}}$ we have $\lambda_{RE}(s_0, 3, \mathbb{X}/\sqrt{K}) \ge \lambda_{RE}(s_0, 3, \Sigma^{\frac{1}{2}})/2$.*

***Proof.*** The basic idea is that conditioned on the total number of samples $K$, the conditional distribution of $(K_1, \dots, K_N)$ is multinomial with success probabilities $(w_1, \dots, w_N)$. Thus conditionally, the rows of $\mathbb{X}$ form an i.i.d. sample of the population $(x_i)_{i=1}^N$ with probabilities $(w_i)_{i=1}^N$. Therefore, the result of Theorem 3 can be applied to get a lower bound on the restricted eigenvalue of $\mathbb{X}/\sqrt{K}$. For this, we first find upper bounds on $d, n_0$ and a lower bound on $\rho, \lambda_{RE}$ as needed in the theorem.

From the assumption on the spectrum of $\Sigma$ and the definitions of sparse and restricted eigenvalues it is clear that $\rho^2 \ge \lambda_\star$ and $\lambda_{RE}^2(s_0, 3, \Sigma^{\frac{1}{2}}) \ge \lambda_\star$. From these inequalities, and using $\delta = 1/2$ and $k_0 = 3$, we obtain an upper bound on $d$:

$$d \le s_0 \left(1 + \frac{\lambda^\star}{\lambda_\star} 4 \cdot 64 \cdot 9^2 \cdot 10\right)$$

$$\leq 51841 \cdot s_0 \frac{\lambda^\star}{\lambda_\star}.$$

Next, writing the $n_0$ in Theorem 3 as $m_0 \log^3(m_0)$, we can bound $m_0$ by

$$\begin{aligned}
m_0 &= \frac{C_{RZ} M^2 d \cdot \log p}{\rho^2 \delta^2} \\
&\leq \frac{4 \times 51841 C_{RZ} M^2 \lambda^\star s_0 \log p}{\lambda_\star^2} \\
&= \frac{\tilde{C} M^2 \lambda^\star s_0 \log p}{\lambda_\star^2}
\end{aligned}$$

It follows that

$$\tilde{n}_0 \geq m_0 \log^3 m_0 = n_0.$$

Next, we show that with high probability, the sample size $K$ is not smaller than $n_0$. We have

$$\mathbf{P}(K < n_0) \leq \mathbf{P}(K < \tilde{n}_0) = e^{-n} \sum_{j=0}^{\tilde{n}_0 - 1} \frac{n^j}{j!} \leq e^{\tilde{n}_0 - n} \leq e^{-\frac{\tilde{n}_0}{4}}.$$

Now we proceed by conditioning on $K = k$ for $k \geq n_0$. Note that as mentioned before, given $K = k$, the rows of $\mathbb{X}$ have the same distribution as a weighted i.i.d. sample from $(x_i)_1^N$ with probabilities $(w_i)_1^N$, and since $k \geq n_0$, by Theorem 3 the probability that the $RE(s_0, 3, \mathbb{X}/\sqrt{k})$ does not hold is at most $\exp(-k\lambda_\star/(12M^2 d))$. Denote by $B$ the event that $\lambda_{RE}(s_0, 3, \mathbb{X}/\sqrt{K}) < \lambda_{RE}(s_0, 3, \Sigma^{\frac{1}{2}})/2$. Then we have

$$\begin{aligned}
\mathbf{P}(B) &\leq \mathbf{P}(K < n_0) + \mathbf{P}(B \cap [K \geq n_0]) \\
&\leq e^{-\frac{\tilde{n}_0}{4}} + \sum_{k=n_0}^{\infty} \mathbf{P}(B \mid K = k) \cdot \mathbf{P}(K = k) \\
&\leq e^{-\frac{\tilde{n}_0}{4}} + \sum_{k=n_0}^{\infty} \exp\left(-\frac{k\lambda_\star}{12M^2 d}\right) \cdot \mathbf{P}(K = k) \\
&\leq e^{-\frac{\tilde{n}_0}{4}} + e^{-\frac{n_0 \lambda_\star}{12M^2 d}}.
\end{aligned}$$

□

Next we record here a conditional central limit theorem due to [2, Theorem 1 and Corollary 3] that will be used in the proof of theorem 1. Let $\{U_{n,k}\}_{n,k}$ be a triangular array and $\mathcal{A}_n$ be the $\sigma$-algebra that can change with $n$. Denote by $\mathbf{E}^{\mathcal{A}_n}[\cdot] = \mathbf{E}[\cdot \mid \mathcal{A}_n]$ the conditional expectation with respect to $\mathcal{A}_n$ and define

$$S_n := \sum_{k=1}^{N} U_{n,k}, \qquad (\sigma_n^{\mathcal{A}_n})^2 := \mathbf{Var}^{\mathcal{A}_n} S_n = \mathbf{E}^{\mathcal{A}_n}(S_n - \mathbf{E}^{\mathcal{A}_n} S_n)^2$$

**Theorem 4** (Theorem 1 and Corollary 3 of Bulinski [2]). *Let $\{U_{n,k} : k = 1, \ldots, k_n$ and $n \in \mathbf{N}\}$ be an array of random variables, which are $\mathcal{A}_n$-independent (i.e. independent given $\mathcal{A}_n$) in each row (for some $\sigma$-algbera $\mathcal{A}_n \subset \mathcal{F}$, where $n \in \mathbf{N}$), and $\mathbf{Var}^{\mathcal{A}_n}(U_{n,k}) < \infty$ (a.s.) for $k = 1, \ldots, k_n, n \in \mathbf{N}$. Assume that $(\sigma_n^{\mathcal{A}_n})^2 := \mathbf{Var}^{\mathcal{A}_n} S_n > 0$ (a.s.) for all n large enough. Then the two relations*

$$\max_{k=1,\ldots,k_n} \frac{\mathbf{Var}^{\mathcal{A}_n} U_{n,k}}{(\sigma_n^{\mathcal{A}_n})^2} \to_p 0 \tag{1}$$

*and*

$$\mathbf{E}^{\mathcal{A}_n} \exp\left\{ it \frac{S_n - \mathbf{E}^{\mathcal{A}_n} S_n}{\sigma_n^{\mathcal{A}_n}} \right\} \to_p \exp\left\{ -\frac{t^2}{2} \right\}, \quad n \to \infty.$$

*hold if and only if the $\mathcal{A}_n$-Lindeberg condition is satisfied in a weak form: for any $t > 0$*

$$T_n := \frac{1}{(\sigma_n^{\mathcal{A}})^2} \sum_{i=1}^{N} \mathbf{E}^{A} \left[ (U_k - \mathbf{E}^{\mathcal{A}} U_k)^2 \mathbf{1}\{|U_k - \mathbf{E}^{\mathcal{A}} U_k| > t\sigma_n^{\mathcal{A}}\} \right] \to_p 0.$$

*Furthermore, if the above $\mathcal{A}_n$-Lindeberg condition holds, then we have*

$$\frac{S_n - \mathbf{E}^{\mathcal{A}_n} S_n}{\sigma_n^{\mathcal{A}_n}} \to_d Z \sim N(0,1), \quad \textit{as } n \to \infty.$$

***Proof of Theorem 1 of main paper.*** To lighten notation, in the following we write $\Sigma$ instead of $\Sigma_w$. Also, since the size of the Poisson sample $K = \sum_1^N N_i$ is itself a Poisson random variable with mean $n$, it follows that $\mathbf{P}(K = 0) = e^{-n} \to 0$. Therefore in the following analysis it is implicit that $K > 0$ (more formally, the analysis is restricted to the event $[K > 0]$ that occurs with probability $1 - e^{-n}$.) Finally, in what follows we repeatedly use the fact that conditionally given $K = k$, the random variables $X_1, \ldots, X_k$ form an i.i.d. sample from $(x_i)_1^N$ with weights $(w_i)_1^N$. This is true because of the well-known fact that given $\sum_1^N N_i = k$, the distribution of $(N_i)_1^N$ is multinomial with parameters $k$ and $(w_i)_1^N$.

We present the proof in three parts:

**Part 1. (Bias Bound)** First note that using weighted lasso with weights $\widehat{W}_j = \sqrt{\sum_{i=1}^K X_{ij}^2 / K}$ for $1 \le j \le p$ is equivalent to normalizing the columns of $\mathbb{X}$ before applying the lasso. Furthermore, since $\mathbf{E}\widehat{W}_j^2 = \mathbf{E}[\mathbf{E}[\sum_{i=1}^K X_{ij}^2 / K \mid K]] = \Sigma_{jj}$ and each $|X_{ij}|$ is bounded by $M$ by assumption, we can use Hoeffding's concentration inequality to write for each $j$ and every $t > 0$

$$\mathbf{P}\left( |\widehat{W}_j^2 - \Sigma_{jj}| \ge t \,\middle|\, K \right) \le 2 \exp\left( -\frac{2Kt^2}{4M^2} \right).$$

Setting $t = 2M\sqrt{\log(p)/K}$ in the above inequality yields

$$\mathbf{P}\left( |\widehat{W}_j^2 - \Sigma_{jj}| \ge 2M\sqrt{\frac{\log(p)}{K}} \,\middle|\, K \right) \le 2p^{-2}.$$

Taking the expectations on both sides with respect to the distribution of $K$ we obtain

$$\mathbf{P}\left(|\widehat{W}_j^2 - \Sigma_{jj}| \geq 2M\sqrt{\frac{\log(p)}{K}}\right) \leq 2p^{-2}.$$

Using a union bound over $j = 1, \dots, p$ and dividing by $\Sigma_{jj} \geq \lambda_\star > 0$ we obtain

$$\mathbf{P}\left(\max_{1\leq j\leq p}\left|\frac{\widehat{W}_j^2}{\Sigma_{jj}} - 1\right| \geq \frac{2M}{\lambda_\star}\sqrt{\frac{\log(p)}{K}}\right) \leq \frac{2}{p} \to 0.$$

We have $K/n \to_p 1$ since $\mathbf{Var}[K/n] = n^{-1} \to 0$. Moreover, by assumption we have $M\sqrt{\log(p)} = o(\sqrt{n})$. Together, these imply that $M\sqrt{\log(p)/K} \to 0$ and since $\lambda_\star$ is bounded away from zero by assumption, we obtain

$$\max_{1\leq j\leq p}\left|\frac{\widehat{W}_j^2}{\Sigma_{jj}} - 1\right| \to_p 0 \quad \text{as } n \to \infty.$$

Therefore with high probability the weights $\widehat{W}_j$ are bounded away from $0, \infty$ and thus the standard (unweighted) lasso guarantees apply [1, see Section 6.9 for the error analysis of the weighted lasso]. Next, observe that since $0 < \lambda_\star \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \lambda^\star < \infty$ and because by assumption $\sqrt{n} \gg s\log^{3/2}(p)$, Proposition 1 guarantees that the scaled design matrix $\mathbb{X}/\sqrt{K}$ satisfies the RE condition with a restricted eigenvalue larger than $\sqrt{\lambda_\star}/2$ with probability $1 - o(1)$, where $K$ is the number of samples obtained via Poisson sampling. Calling this event $G$, we have $P(G) = 1 - o(1)$. Given $G$, we have[1]

$$\mathbf{P}\left(\|\widehat{\beta} - \beta\|_1 \lesssim \frac{\sigma_\varepsilon s}{\lambda_\star}\sqrt{\frac{\log(p)}{K}}\,\middle|\, G\right) = 1 - o(1).$$

(for a proof see for example Theorem 6.1 and Corollary 6.6 of Bühlmann and Van De Geer [1]). It follows that

$$\begin{aligned}
\mathbf{P}\left(\|\widehat{\beta} - \beta\|_1 \lesssim \frac{\sigma_\varepsilon s}{\lambda_\star}\sqrt{\frac{\log(p)}{K}}\right) &\geq \mathbf{P}\left(\|\widehat{\beta} - \beta\|_1 \lesssim \frac{\sigma_\varepsilon s}{\lambda_\star}\sqrt{\frac{\log(p)}{K}}\,\middle|\, G\right) \cdot \mathbf{P}(G) \\
&= (1 - o(1)) \cdot (1 - o(1)) \\
&= (1 - o(1)).
\end{aligned}$$

Next, note that since $K/n \to_p 1$, we can substitute $n$ for $K$ in the above bound and write

$$\mathbf{P}\left(\|\widehat{\beta} - \beta^\star\|_1 \lesssim \frac{\sigma_\varepsilon s}{\lambda_\star}\sqrt{\frac{\log(p)}{n}}\right) \to 1. \tag{2}$$

Let $\widehat{w}_i = N_i/n$ where $N_i$ are obtained using Poisson sampling. Then the debiased lasso estimator can be written as

$$\widehat{\gamma} = \langle c, \widehat{\beta}\rangle + u^T\widehat{\Sigma}(\beta^\star - \widehat{\beta}) + \frac{1}{n}u^T X^T \varepsilon$$

[1] note that the lower bound $\sqrt{\lambda_\star}/2$ on the restricted eigenvalue is uniform over $G$

$$= \gamma + c^T(\Sigma^{-1}\widehat{\Sigma} - I)(\beta^\star - \widehat{\beta}) + \frac{1}{n}c^T\Sigma^{-1}X^T\varepsilon.$$

Subtracting $\gamma$ from both sides and multiplying by $\sqrt{n}$ we obtain

$$\sqrt{n}(\widehat{\gamma} - \gamma) = \sqrt{n}c^T(\widehat{\Sigma}^{-1}\widehat{\Sigma} - I)(\beta^\star - \widehat{\beta}) + \frac{1}{\sqrt{n}}c^T\Sigma^{-1}X^T\varepsilon.$$

We show that the first term is $o_p(1)$. Using an $\ell_1 - \ell_\infty$ bound and the error rate of the lasso estimate, with probability $1 - o(1)$ we have

$$\sqrt{n}|c^T(\widehat{\Sigma}^{-1}\widehat{\Sigma} - I)(\beta^\star - \widehat{\beta})| \leq \sqrt{n}\|c^T(\widehat{\Sigma}^{-1}\widehat{\Sigma} - I)\|_\infty \cdot \|\widehat{\beta} - \beta^\star\|_1$$

Thus we need first to upper bound $\|c^T(\widehat{\Sigma}^{-1}\widehat{\Sigma} - I)\|_\infty$. Let $\widehat{w}_i = N_i/n$ and note that

$$c^T(\Sigma^{-1}\widehat{\Sigma} - I) = \sum_{i=1}^{N}(\widehat{w}_i - w_i)c^T\Sigma^{-1}x_ix_i^T.$$

Recall that $N_i$ is a Poisson random variable with mean $nw_i$, and therefore $N_i - nw_i$ is subexponential with

$$\begin{aligned}
\mathbf{E}\exp(t(N_i - nw_i)) &= \exp(nw_i(e^t - 1 - t)) \\
&\leq \exp(nw_i(e^{|t|}t^2/2) \\
&\leq \exp(nw_i(t^2)) \quad \text{for } |t| \leq \frac{1}{2}.
\end{aligned}$$

This shows that $\|N_i - nw_i\|_{\psi_1} \lesssim \sqrt{nw_i}$ [5, Proposition 2.7.1], and therefore, $\|\widehat{w}_i - w_i\|_{\psi_1} \lesssim \sqrt{w_i/n}$. Define $V_{ij} = (\widehat{w}_i - w_i)c^T\Sigma^{-1}x_ix_{ij}$. Using Bernstein's inequality for subexponential random variables [5, Theorem 2.8.1], for some absolute constant $b > 0$ and all $j = 1, \ldots, p$ we have

$$\mathbf{P}\left(\left|\sum_{i=1}^{N}V_{ij}\right| > t\right) \leq 2\exp\left(-b\min\left\{\frac{t^2}{\sum_{i=1}^N\|V_{ij}\|_{\psi_1}^2}, \frac{t}{\max_i\|V_{ij}\|_{\psi_1}}\right\}\right). \tag{3}$$

Using the bound $\max_{i,j}|x_{ij}| \leq M$, we have

$$\begin{aligned}
\sum_i\|V_{ij}\|_{\psi_1}^2 &\leq \frac{M^2}{n}\sum_i w_i(c^T\Sigma^{-1}x_i)^2 \\
&= \frac{M^2}{n}\sum_i w_ic^T\Sigma^{-1}x_ix_i^T\Sigma^{-1}c \\
&= \frac{M^2}{n}c^T\Sigma^{-1}(\sum_i w_ix_ix_i^T)\Sigma^{-1}c \\
&= \frac{M^2}{n}c^T\Sigma^{-1}c.
\end{aligned}$$

Similarly,

$$\max_i \|V_{ij}\|_{\psi_1} \le M \cdot \max_i \sqrt{w_i/n}|c^T\Sigma^{-1}x_i| \le M\sqrt{\frac{c^T\Sigma^{-1}c}{n}}.$$

Using these bounds and for $t = \sqrt{2\log(p)/(nb)}$ the Bernstein bound (3) implies

$$\mathbf{P}\left(\left|\sum_{i=1}^{N} V_{ij}\right| > \frac{2M\log(p)}{b}\sqrt{\frac{c^T\Sigma^{-1}c}{n}}\right) \le 2\exp\left(-\min\left\{\frac{4\log^2(p)}{b}, 2\log(p)\right\}\right).$$

For $p > \exp(b/2)$, the exponential tail prevailes, and we obtain

$$\mathbf{P}\left(\left|\sum_{i=1}^{N} V_{ij}\right| > \frac{2M\log(p)}{b}\sqrt{\frac{c^T\Sigma^{-1}c}{n}}\right) \le 2p^{-2}, \text{ for all } j = 1, \ldots, p.$$

A union bound over all $j = 1, \ldots, p$ now yields

$$\mathbf{P}\left(\max_{1\le j\le p}\left|\sum_{i=1}^{N} V_{ij}\right| > \frac{2M\log(p)}{b}\sqrt{\frac{c^T\Sigma^{-1}c}{n}}\right) \le 2p^{-1} \quad , \text{ for } p > e^{b/2}. \tag{4}$$

Continuing with the $\ell_1 - \ell_\infty$ bound and using the upper bound (4) and the error rate of the lasso estimate (2), with probability $1 - o(1)$ we have

$$\begin{aligned}
|\sqrt{n}|c^T(\widehat{\Sigma}^{-1}\widehat{\Sigma} - I)(\beta^\star - \widehat{\beta})| &\le \sqrt{n}\|c^T(\widehat{\Sigma}^{-1}\widehat{\Sigma} - I)\|_\infty \cdot \|\widehat{\beta} - \beta^\star\|_1 \\
&\lesssim \sqrt{n} \cdot \left(M\log(p)\sqrt{\frac{c^T\Sigma^{-1}c}{n}}\right) \cdot \frac{\sigma_\varepsilon}{\lambda_\star} s\sqrt{\frac{\log(p)}{n}} \\
&= \frac{M\sigma_\varepsilon\sqrt{c^T\Sigma^{-1}c}}{\lambda_\star} \cdot \frac{s\log^{\frac{3}{2}}(p)}{\sqrt{n}}.
\end{aligned}$$

**Part 2.(Variance Approximation)** Next, we prove the third part of the theorem as the argument used here will be useful in the proof of asymptotic normality. The conditional variance of the noise term is

$$\mathbf{Var}(\frac{1}{\sqrt{n}}c^T\Sigma^{-1}X^T\varepsilon \mid X) = c^T\Sigma^{-1}\widehat{\Sigma}\Sigma^{-1}c.$$

We show that this variance can be approximated by $c^T\Sigma^{-1}c$, i.e.

$$\frac{c^T\Sigma^{-1}\widehat{\Sigma}\Sigma^{-1}c}{c^T\Sigma^{-1}c} \to_p 1.$$

We assume $N = |\{i : w_i \neq 0\}|$ as otherwise one can throw out the zero weights. We want to show:

$$A := \frac{\sum_i w_i(c^T\Sigma^{-1}x_i)^4}{n(c^T\Sigma^{-1}c)^2} \le \frac{1}{n}.$$

Let $d_i = \Sigma^{-1/2} x_i$ and $v = \Sigma^{-1/2} c$. Then $A$ can be written as

$$A = \frac{\sum_i w_i (d_i^T v)^4}{n\|v\|_2^4}.$$

**Step 1.** First suppose that $N = p$. We have

$$\sum_i w_i d_i d_i^T = \Sigma^{-\frac{1}{2}} \left( \sum_i w_i x_i x_i^T \right) \Sigma^{-\frac{1}{2}} = I_p.$$

(This is true for $N > p$ too.) Let $d_j^\star$ be the projection of $d_j$ on the ortho-complement of the span of $\{d_i \mid i \neq j\}$. Then multiplying both sides by $d_j^\star$ we obtain

$$w_j (d_j^T d_j^\star) d_j = d_j^\star.$$

Note that $d_j^T d_j^\star \neq 0$ for all $j = 1, \ldots, p$ as $d_1, \ldots, d_p$ form a basis for $R^p$ (since $\Sigma$ is non-singular by construction.) From this equation it follows that $d_1, \ldots, d_p$ are orthogonal, and after multiplying both sides by $d_j$ one also obtains $\|d_j\|_2^2 = w_j^{-1}$.

Now since $A$ does not depend on $\|v\|_2$, we have

$$A \leq \max_{u \neq 0} \frac{\sum_i w_i (d_i^T u)^4}{n\|u\|_2^4} = \frac{1}{n} \max_{\|u\|_2^2 = 1} \sum_i w_i (d_i^T u)^4.$$

The Lagrangian for the last maximization problem is

$$L(u, \lambda) = \sum_i w_i (d_i^T u)^4 - 2\lambda(u^T u - 1).$$

Taking derivative w.r.t. $u$ and setting to zero yields

$$\sum_i w_i (d_i^T \widehat{u})^3 d_i = \lambda \widehat{u}. \tag{5}$$

Changing variables to $\tilde{u} = \sqrt{1/\lambda}\,\widehat{u}$, we can rewrite (5) as

$$\sum_i w_i (d_i^T \tilde{u})^3 d_i = \tilde{u}. \tag{6}$$

Multiplying on the left once by $\tilde{u}^T$ and once by $d_j^T$ and using $d_j^T d_j = w_j^{-1}$ gives

$$\sum_i w_i (d_i^T \tilde{u})^4 = \tilde{u}^T \tilde{u} \quad \text{and} \quad (d_j^T \tilde{u})^2 = 1. \tag{7}$$

Note that $\sum_i w_i (d_i^T u)^4 / \|u\|_2^4$ does not depend on the norm of $u$, so that any nonzero multiple of $\widehat{u}$, and in particular $\tilde{u}$, is a maximizer. Plugging $\tilde{u}$ in this expression and using (7) we obtain

$$\begin{aligned} A &\leq \frac{\sum_i w_i (d_i^T \tilde{u})^4}{n(\tilde{u}^T \tilde{u})^2} \\ &\leq \frac{1}{n \cdot \sum_i w_i (d_i^T \tilde{u})^4} = \frac{1}{n}. \end{aligned}$$

This finishes the proof for the $N = p$ case.

**Step 2.** Now consider the $N > p$ case. The idea is to reduce this case to the $N = p$ case by appropriately extending the length of $d_i, v \in R^p$ from $p$ to $N$.

Note that the identity $\sum_i w_i d_i d_i^T = I_p$ is still valid. Define the vectors $\tilde{d}_i \in R^N$ by $\tilde{d}_i^T = (d_i^T, f_i^T)$ for some vectors $f_i \in R^{N-p}$ such that

$$\sum_i w_i \tilde{d}_i \tilde{d}_i^T = I_N.$$

A construction of these $f_i$'s is given in the Appendix.

For any $u \in R^N$ use a similar decomposition $u^T = (u_1^T, u_2^T)$ with $u_1 \in R^p$ and $u_2 \in R^{N-p}$. Then

$$\begin{aligned} n \cdot A &\leq \max_{\|v\|_2^2=1} \sum_i w_i (d_i^T v)^4 \\ &= \max_{\substack{\|u\|_2^2=1 \\ u_2=0}} \sum_i w_i (\tilde{d}_i^T u)^4 \\ &\leq \max_{\|u\|_2^2=1} \sum_i w_i (\tilde{d}_i^T u)^4 \\ &\leq 1, \end{aligned}$$

where the last inequality follows from the argument in **Step 1** (since now $\tilde{p} := \dim(\tilde{d}_i) = N$). This finishes the proof of the $N > p$ case.

**Part 3. (Asymptotic Normality)** Before we establish asymptotic normality, let us introduce some notation. Let $\mathcal{A}$ be the $\sigma$-algebra generated by $(N_i)_{i=1}^N$ (note that we are suppressing the dependence of $\mathcal{A} = \mathcal{A}_n$ on $n$ to lighten notation). As in Theorem 4, denote by $\mathbf{E}^{\mathcal{A}}[\cdot] = \mathbf{E}[\cdot \mid \mathcal{A}]$ the conditional expectation with respect to $\mathcal{A}$ and define

$$U_{n,k} := \begin{cases} \frac{1}{\sqrt{n}} c^T \Sigma^{-1} x_k (\varepsilon_1^{(k)} + \cdots + \varepsilon_{N_k}^{(k)}), & N_k > 0 \\ 0, & N_k = 0. \end{cases}$$

$$S_n := \sum_{i=1}^N U_{n,k}, \quad (\sigma_n^{\mathcal{A}})^2 := \mathbf{E}^{\mathcal{A}} (S_n - \mathbf{E}^{\mathcal{A}} S_n)^2,$$

where $\varepsilon_i^{(j)} \sim \varepsilon_n$ are iid mean-zero random variables with variance equal to $\mathbf{E}\varepsilon_n^2 = \sigma^2$ and sub-Guassian norm $\|\varepsilon_i^{(j)}\|_{\psi_2} \leq \sigma_\varepsilon$. It follows that in general $\sigma \lesssim \sigma_\varepsilon$. Observe that since $\mathbf{E}^{\mathcal{A}} S_n = 0$, we have

$$\begin{aligned} (\sigma_n^{\mathcal{A}})^2 = \mathbf{E}^{\mathcal{A}} S_n^2 &= \frac{1}{n} \sum_{k=1}^N (c^T \Sigma^{-1} x_k)^2 \mathbf{E}[(\varepsilon_1^{(k)} + \cdots + \varepsilon_{N_k}^{(k)})^2 \mid N_k] \\ &= \sum_{k=1}^N \left( \frac{N_k}{n} \right) (c^T \Sigma^{-1} x_i)^2 \sigma^2 \\ &= \sigma^2 \cdot c^T \Sigma^{-1} \widehat{\Sigma} \Sigma^{-1} c. \end{aligned}$$

Thus we want to show

$$\frac{c^T\Sigma^{-1}\mathbb{X}^T\varepsilon}{\sqrt{n\sigma^2\cdot c^T\Sigma^{-1}\widehat{\Sigma}\Sigma^{-1}c}} = \frac{S_n - \mathbf{E}^{\mathcal{A}_n}S_n}{\sigma_n^{\mathcal{A}_n}} \to_d Z \sim N(0,1), \quad \text{as } n\to\infty.$$

In light of Theorem 4, it suffices to prove the following conditional Lindeberg condition is satisfied:

$$\forall t>0: \quad T_n := \frac{1}{(\sigma_n^{\mathcal{A}})^2}\sum_{i=1}^N \mathbf{E}^{A}\left[(U_k - \mathbf{E}^{\mathcal{A}}U_k)^2\mathbf{1}\{|U_k - \mathbf{E}^{\mathcal{A}}U_k| > t\sigma_n^{\mathcal{A}}\}\right] \to_p 0.$$

Next we find appropriate upper bounds on the summands in the Lindeberg condition. Also note that by assumption the noise variance $\sigma^2$ is bounded away from 0 and $\infty$, we can assume without loss of generality that $\sigma^2 = 1$ in what follows. Using the Cauchy-Schawrz inequality,

$$\begin{aligned}\mathbf{E}^{\mathcal{A}}[U_{n,k}^2\mathbf{1}\{|U_{n,k}| > t\sigma_n^{\mathcal{A}}\}] &\le \left((\mathbf{E}U_{n,k}^4)(\mathbf{E}^{\mathcal{A}}\mathbf{1}\{|U_{n,k}| > t\sigma_n^{\mathcal{A}}\})\right)^{\frac{1}{2}}\\ &\le \left((\mathbf{E}^{\mathcal{A}}U_{n,k}^4)\cdot\frac{\mathbf{E}^{\mathcal{A}}U_{n,k}^2}{t^2(\sigma_n^{\mathcal{A}})^2}\right)^{\frac{1}{2}}.\end{aligned}$$

The fourth moment of $U_{n,k}$ can be bounded as follows

$$\begin{aligned}\mathbf{E}^{\mathcal{A}}U_{n,k}^4 &= \frac{1}{n^2}(c^T\Sigma^{-1}x_i)^4\mathbf{E}^{\mathcal{A}}(\epsilon_1^{(k)} + \cdots + \varepsilon_{N_k}^{(k)})^4\\ &= \frac{1}{n^2}(c^T\Sigma^{-1}x_i)^4\cdot(N_k\mathbf{E}\varepsilon^4 + N_k(N_k-1)(\mathbf{E}\varepsilon^2)^2)\\ &\le \frac{C}{n^2}(c^T\Sigma^{-1}x_i)^4\cdot(N_k^2\sigma_\varepsilon^4),\end{aligned}$$

where the last inequality follows because $N_k \le N_k^2$ and by sub-Gaussianity of $\varepsilon$ we have $\mathbf{E}\varepsilon^4 \le C\cdot\sigma_\varepsilon^4$ for an absolute constant $C>0$. Combined with the equality $\mathbf{E}^{\mathcal{A}}U_{n,k}^2 \lesssim N_k(c^T\Sigma^{-1}x_i)^2\sigma_\varepsilon^2$, we find the upper bound

$$\begin{aligned}\mathbf{E}^{\mathcal{A}}[U_{n,k}^2\mathbf{1}\{|U_{n,k}| > t\sigma_n^{\mathcal{A}}\}] &\le \left(\frac{C}{t^2(\sigma_n^{\mathcal{A}})^2}\left(\frac{N_k}{n}\right)^3\cdot(c^T\Sigma^{-1}x_k)^6\right)^{\frac{1}{2}}\\ &\le \frac{1}{\sigma_n^{\mathcal{A}}}\left(\frac{C}{t^2}\frac{\sum_{i=1}^N N_i}{n}\right)^{\frac{1}{2}}\left(\frac{N_k}{n}\right)|c^T\Sigma^{-1}x_k|^3.\end{aligned}$$

Therefore the expression in the Lindeberg condition has the following upper bound:

$$T_n \le \left(\frac{C}{t^2}\frac{\sum_{i=1}^N N_i}{n}\right)^{\frac{1}{2}}\left(\frac{(c^T\Sigma^{-1}c)}{(\sigma_n^{\mathcal{A}})^2}\right)^{\frac{3}{2}}\frac{1}{(c^T\Sigma^{-1}c)^{\frac{3}{2}}}\sum_{k=1}^N|c^T\Sigma^{-1}x_k|^3\widehat{w}_k$$

Since, as shown before, $\sum_i N_i/n \to_p 1$ and $(\sigma_n^{\mathcal{A}})^2/(c^T\Sigma^{-1}c) \to_p 1$, it suffices to show that

$$\frac{1}{(c^T\Sigma^{-1}c)^{\frac{3}{2}}}\sum_{k=1}^{N}|c^T\Sigma^{-1}x_k|^3\widehat{w}_k \to_p 0.$$

We will show the variance of this term converges to zero:

$$\frac{1}{n\cdot(c^T\Sigma^{-1}c)^3}\sum_{k=1}^{N}|c^T\Sigma^{-1}x_k|^6 w_k \to 0.$$

A similar technique as before is applicable here. Let $d_k = \Sigma^{-\frac{1}{2}}x_k$ and $v = \Sigma^{-\frac{1}{2}}c$. The variance can be rewritten as

$$\begin{aligned}\frac{1}{n\|v\|_2^6}\sum_k w_k(d_k^T v)^6 &\leq \frac{1}{n}\cdot \max_{\|u\|_2^2=1}\left\{\sum_k w_k(d_k^T u)^6\right\}\\ &\leq \frac{1}{n}\to 0, \quad \text{ as } n\to\infty.\end{aligned}$$

where the last inequality follows because the value of the maximization problem is seen to be 1 using a similar argument as the one used in Part 2 (Variance Approximation). □

**Proposition 2.** *Problem $P2$ can be recast as a semidefinite program (SDP).*

***Proof.*** We can write problem $\mathbf{P2}$ as

$$\begin{aligned}\mathbf{P2}': \min_{t\in\mathbb{R},w\in\mathbb{R}^N} \quad & t\\ \text{s.t.} \quad & \Sigma=\sum_{i=1}^{N}w_i x_i x_i^T, \quad \sum_{i=1}^{N}w_i=1\\ & \lambda_\star I \preccurlyeq \Sigma \preccurlyeq \lambda^\star \star I, \quad c^T\Sigma^{-1}c\leq t\\ & w\geq 0.\end{aligned}$$

The constraint $c^T\Sigma^{-1}c\leq t$ is equivalent to

$$\begin{bmatrix} t & c^T\\ c & \Sigma\end{bmatrix}\succcurlyeq 0$$

since, given that $\Sigma$ is positive definite (guaranteed by the constraint $\Sigma-\alpha I\succcurlyeq 0$), the above matrix is positive semidefinite if and only if the Schur complement $t-c^T\Sigma^{-1}c$ is positive semidefinite, by the following decomposition:

$$\begin{bmatrix} t & c^T\\ c & \Sigma\end{bmatrix}=\begin{bmatrix} 1 & c^T\Sigma^{-1}\\ 0 & I\end{bmatrix}\cdot\begin{bmatrix} t-c^T\Sigma^{-1}c & 0\\ 0 & \Sigma\end{bmatrix}\cdot\begin{bmatrix} 1 & c^T\Sigma^{-1}\\ 0 & I\end{bmatrix}^T.$$

□

## 2. Appendix

The following provides the details left out in example 1:

**Example 1 (continued).** First we show that on event $E$, the lasso estimate with the theoretical value of the tuning parameter $\lambda = \sqrt{(2+\eta)\log(p)/n}$ for some $\eta > 0$ vanishes with high probability. Let $L(\beta)$ be the objective of the weighted lasso and $\widehat{W}$ be a diagonal matrix with $\widehat{W}_j$'s on its diagonal. For any $\beta$, on $E$ we have

$$\begin{aligned} L(\beta) - L(0_p) &= \frac{1}{2n}\|\varepsilon - \mathbb{X}\beta\|_2^2 + \lambda\sum_{j=1}^{p}\widehat{W}_j|\beta_j| - \|\varepsilon\|_2^2 \\ &= \|\mathbb{X}\beta\|_2^2 + \frac{-1}{n}\varepsilon^T\mathbb{X}\beta + \lambda\|\widehat{W}\beta\|_1 \\ &\geq \|\mathbb{X}\beta\|_2^2 + \left(\lambda - \frac{\left\|\varepsilon^T X W^{+}\right\|_\infty}{n}\right)\cdot\|\widehat{W}\beta\|_1. \end{aligned}$$

A standard union argument shows that we have

$$\mathbf{P}(\lambda > \|\varepsilon^T\mathbb{X}W^{+}\|_\infty/n) \to 1.$$

This implies that with probability $1 - o(1)$ and for any $\beta$, we have $L(\beta) \geq L(0)$, so that $0_p \in \arg\min_\beta L(\beta)$. In fact, for all $j \in [p]$ we have $\widehat{W}_j\widehat{\beta}_j = 0$.

Next, we sketch an explicit construction of the experimental domain $(x_i)_1^p$ used in Example 1 (recall that in this example $N = p$). We start with the matrix $D$ of the discrete cosine transform (DCT) defined by equation (5) in Section 3 of the article. The matrix $B = \sqrt{p}D^T$ satisfies $B^TB = p\cdot I_p$ and $B_{i1} = 1$ for all $1 \leq i \leq p$ and $\max_{i,j}|B_{ij}| \leq \sqrt{2}$. Denote by $B_i$ the $i$-th row of $B$ and define

$$x_i = \begin{cases} \frac{1}{\sqrt{2}}(B_i + B_{i+1}) & : i \text{ is odd,} \\ \frac{1}{\sqrt{2}}(B_{i-1} - B_i) & : i \text{ is even.} \end{cases}$$

Then it is straightforward to check that $p^{-1}\sum_i x_i x_i^T = I_p$ and $\|x_i\|_\infty \leq 2$ for all $i \leq p$. Furthermore, we have

$$x_{i1} = \begin{cases} \sqrt{2} & : i \text{ is odd,} \\ 0 & : i \text{ is even.} \end{cases}$$

$\square$

**Constructing $f_i$'s.** To construct the $f_i$'s alluded to in the proof of the second part of Theorem 1, consider the following matrix (with the $f_i$'s to be specified shortly)

$$D = \begin{pmatrix} d_1^T & f_1^T \\ \vdots & \vdots \\ d_N^T & f_N^T \end{pmatrix} = \begin{pmatrix} | & \dots & | \\ g_1 & \dots & g_N \\ | & \dots & | \end{pmatrix} \in R^{N\times N}.$$

In our notation we write $d_i = (d_{ij})_{j=1}^p$ and $g_i = (g_{ij})_{j=1}^N$, so that $d_{ij}$ is the $j$-th coordinate of $d_i$, etc. Then we know that

$$\langle g_i, g_j \rangle_w := \sum_k w_k g_{ik} g_{jk} = \sum_k w_k d_{ki} d_{kj} = \left[ \sum_k w_k d_k d_k^T \right]_{ij} = \delta_{ij}$$

for $1 \le i, j \le p$. In other words, $\{g_1, \dots, g_p\}$ forms an orthonormal basis (w.r.t. $\langle \cdot, \cdot \rangle_w$) of its span. All we need is to choose $g_{p+1}, \dots, g_N$ in such a way that $\{g_1, \dots, g_N\}$ is an orthonormal basis of $R^N$ under $\langle \cdot, \cdot \rangle_w$, which is easy to construct using e.g. the Gram-Schmidt procedure. This will ensure that $\langle g_i, g_j \rangle_w = \delta_{ij}$ for all $1 \le i, j \le N$.

With this choice, the $(i, j)$-th coordinate of $\sum_k w_k \tilde{d}_k \tilde{d}_k^T$ is given by

$$\begin{aligned} \sum_k w_k [\tilde{d}_k \tilde{d}_k^T]_{ij} &= \sum_k w_k \tilde{d}_{ki} \tilde{d}_{kj} \\ &= \sum_k w_k g_{ik} g_{jk} \\ &= \langle g_i, g_j \rangle_w = \delta_{ij} \quad \text{for all } 1 \le i, j \le N. \end{aligned}$$

It follows that $\sum_k w_k \tilde{d}_k \tilde{d}_k^T = I_N$. □

## References


[1] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

[2] BULINSKI, A. (2017). Conditional Central Limit Theorem. *Theory of Probability and Its Applications* **61** 613-631.

[3] ELFVING, G. (1952). Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics* **23** 255–262.

[4] RUDELSON, M. and ZHOU, S. (2012). Reconstruction from anisotropic random measurements. In *Conference on Learning Theory* 10–1.

[5] VERSHYNIN, R. (2018). *High-dimensional probability: An introduction with applications in data science* **47**. Cambridge University Press.