A. Proofs of Main Results

In this section we present proofs of the results from Section 5.

A.1. Proof of Theorem 5.5 and Corollaries

We first demonstrate how we decompose the estimation error in the unbiased Lasso Granger estimator $\widehat{\boldsymbol{\theta}}^u$ into the sum of two components:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^u - \boldsymbol{\theta}^* &= \widehat{\boldsymbol{\theta}} + \frac{1}{T-p}\mathbf{M}\widetilde{\mathbf{X}}^\top(\widetilde{\mathbf{X}}\boldsymbol{\theta}^* + \boldsymbol{\epsilon} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}^* \\ &= \frac{1}{T-p}\mathbf{M}\widetilde{\mathbf{X}}^\top\boldsymbol{\epsilon} + \frac{1}{T-p}\mathbf{M}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}) - (\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}) \\ &= \frac{1}{T-p}\mathbf{M}\widetilde{\mathbf{X}}^\top\boldsymbol{\epsilon} + (\mathbf{M}\widetilde{\boldsymbol{\Sigma}}_n - \mathbf{I})(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}).\end{aligned}$$

Letting $\boldsymbol{Z} = \mathbf{M}\widetilde{\mathbf{X}}^\top\boldsymbol{\epsilon}/\sqrt{T-p}$ and $\boldsymbol{\Delta} = \sqrt{T-p}(\mathbf{M}\widetilde{\boldsymbol{\Sigma}}_n - \mathbf{I})(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})$, we have

$$\sqrt{T-p}(\widehat{\boldsymbol{\theta}}^u - \boldsymbol{\theta}^*) = \boldsymbol{Z} + \boldsymbol{\Delta}. \tag{A.1}$$

Note that, clearly, $\mathbb{E}[\boldsymbol{Z}] = \mathbf{0}$. Thus, $\boldsymbol{\Delta}$ encapsulates the bias in $\widehat{\boldsymbol{\theta}}^u$. We divide this proof into two parts. We first establish in Lemma A.1 that $\widehat{\boldsymbol{\theta}}^u$ is an asymptotically unbiased estimator of $\boldsymbol{\theta}^*$ by proving that $\|\boldsymbol{\Delta}\|_\infty = o(1)$. We then proceed to prove in Lemma A.2 that $\boldsymbol{Z}$ is asymptotically normally distributed.

Lemma A.1. Suppose Assumptions 5.3 and 5.4 are satisfied. Let $s_0 = \text{supp}(\widehat{\boldsymbol{\theta}}) \asymp (\sqrt{T-p})/\log(pd)$ and $\mu \asymp \sqrt{\log(pd)/(T-p)}$. Then $\|\boldsymbol{\Delta}\|_\infty = o(1)$, where $\boldsymbol{\Delta} = \sqrt{T-p}(\mathbf{M}\widetilde{\boldsymbol{\Sigma}}_n - \mathbf{I})(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})$.

The proof of Lemma A.1, presented in Appendix B.1, uses Hőlder's inequality to decompose $\|\boldsymbol{\Delta}\|_1$ into the product $\sqrt{T-p}\|\mathbf{M}\widetilde{\boldsymbol{\Sigma}}_n - \mathbf{I}\|_\infty \cdot \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\|_1$. We bound $\|\mathbf{M}\widetilde{\boldsymbol{\Sigma}}_n - \mathbf{I}\|_\infty$ by constructing a martingale difference sequence (see Definition G.1 in Appendix G) and applying a Bernstein-type inequality (Lemma F.8) to this sequence. We then bound $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\|_1$ via a standard argument for Lasso-type estimators that relies on the restricted eigenvalue condition. We amend this argument to work in our non-i.i.d. setting by appealing to martingale theory and present a restricted eigenvalue condition for martingale difference sequences in Appendix F.

Lemma A.2. Suppose Assumptions 5.3 and 5.4 are satisfied. Let $s_0 = \text{supp}(\widehat{\boldsymbol{\theta}}) \asymp (\sqrt{T-p})/\log(pd)$ and $\mu \asymp \sqrt{\log(pd)/(T-p)}$. Then we have $\boldsymbol{Z}/(\sigma[\mathbf{M}\widetilde{\boldsymbol{\Sigma}}_n\mathbf{M}^\top]^{1/2}) = \mathbf{M}\widetilde{\mathbf{X}}^\top\boldsymbol{\epsilon}/(\sigma\sqrt{T-p}[\mathbf{M}\widetilde{\boldsymbol{\Sigma}}_n\mathbf{M}^\top]^{1/2}) \xrightarrow{D} N(0, \mathbf{I}_{pd\times pd})$.

The proof of Lemma A.2, deferred to Appendix B.2, relies on constructing a martingale difference sequence equal to $\boldsymbol{Z}/(\sigma[\mathbf{M}\widetilde{\boldsymbol{\Sigma}}_n\mathbf{M}^\top]^{1/2})$, and applying the Martingale Central Limit Theorem (Hall & Heyde, 1980).

Having established Lemmas A.1 and A.2, we are now ready to present a proof of Theorem 5.5.

Proof of Theorem 5.5. We write the estimation error of the unbiased Lasso Granger estimator as

$$\sqrt{T-p}(\widehat{\boldsymbol{\theta}}^u - \boldsymbol{\theta}^*) = \boldsymbol{Z} + \boldsymbol{\Delta}.$$

where $\boldsymbol{Z} = \mathbf{M}\widetilde{\mathbf{X}}^\top\boldsymbol{\epsilon}/\sqrt{T-p}$ and $\boldsymbol{\Delta} = \sqrt{T-p}(\mathbf{M}\widetilde{\boldsymbol{\Sigma}}_n - \mathbf{I})(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})$. Then, by Lemma A.1 we have that $\boldsymbol{\Delta} \xrightarrow{P} 0$. By Lemma A.2, we have that $\boldsymbol{Z}/(\sigma[\mathbf{M}\widetilde{\boldsymbol{\Sigma}}_n\mathbf{M}^\top]^{1/2}) \xrightarrow{D} N(0, \mathbf{I})$. Therefore, by the Slutsky Theorem (Van der Vaart, 2000), we have that $\sqrt{T-p}(\widehat{\boldsymbol{\theta}}^u - \boldsymbol{\theta}^*)/(\sigma[\mathbf{M}\widetilde{\boldsymbol{\Sigma}}_n\mathbf{M}^\top]^{1/2}) \xrightarrow{D} N(0, \mathbf{I}_{pd\times pd})$, as desired. □

Theorem 5.5 allows us to demonstrate the asymptotic validity of the confidence intervals we present in Corollary 5.6 as follows:

Proof of Corollary 5.6. By Theorem 5.5, the asymptotic normality of $\widehat{\theta}_i^u$ implies

$$\begin{aligned}\lim_{T-p\to\infty}\mathbb{P}\Big(\theta_i^*\in I_i\Big) &= \lim_{T-p\to\infty}\mathbb{P}\Big(\widehat{\theta}_i^u-\theta_i^*\le\delta(\alpha,T-p)\Big)-\lim_{T-p\to\infty}\mathbb{P}\Big(\widehat{\theta}_i^u-\theta_i^*\le-\delta(\alpha,T-p)\Big)\\ &= \lim_{T-p\to\infty}\mathbb{P}\bigg(\frac{\sqrt{T-p}(\widehat{\theta}_i^u-\theta_i^*)}{\sigma[\mathbf{M}\widetilde{\mathbf{\Sigma}}_n\mathbf{M}^\top]_{i,i}^{1/2}}\le\Phi^{-1}(1-\alpha/2)\bigg)\\ &\quad-\lim_{T-p\to\infty}\mathbb{P}\bigg(\frac{\sqrt{T-p}(\widehat{\theta}_i^u-\theta_i^*)}{\sigma[\mathbf{M}\widetilde{\mathbf{\Sigma}}_n\mathbf{M}^\top]_{i,i}^{1/2}}\le-\Phi^{-1}(1-\alpha/2)\bigg)\\ &= 1-\alpha. \end{aligned}\tag{A.2}$$

□

In a similar manner, Theorem 5.5 also permits us to prove the several desirable properties of hypothesis test $\Psi_Z(\alpha)$, which we introduce in (4.7), that we present in Corollary 5.7.

Proof of Corollary 5.7. By (4.3), we have

$$\mathbb{P}(\Psi_Z(\alpha)=1|H_0^i)=\mathbb{P}(-|\widehat{Z_i}|<z_{\alpha/2})=\alpha$$

where $z_{\alpha/2}=\Phi^{-1}(\alpha/2)$, since $\widehat{Z_i}$ converges in distribution to the standard normal distribution. Similarly, for any $u\in(0,1)$, we see that

$$\begin{aligned}\mathbb{P}(P_i<u)&=\mathbb{P}(2(1-\Phi(|\widehat{Z_i}|))<u)=\mathbb{P}\bigg(\Phi(|\widehat{Z_i}|)>1-\frac{u}{2}\bigg)\\ &=\mathbb{P}\bigg(|\widehat{Z_i}|>\Phi^{-1}\Big(1-\frac{u}{2}\Big)\bigg)\xrightarrow{(T-p)\to\infty}u\end{aligned}$$

since, again, $\widehat{Z_i}$ converges in distribution to the standard normal distribution.

□

In Section 4.1, we claim that the Scaled Lasso noise estimator (Sun & Zhang, 2012) $\widehat{\sigma}$, as given by (4.4), is a consistent estimator of the true noise level σ. We note that while Sun & Zhang (2012) prove the consistency in the i.i.d. case, $\widehat{\sigma}$ is nevertheless still consistent in our non-i.i.d. case as well. This result follows directly from Theorem 1 in Sun & Zhang (2012), which we paraphrase in the following lemma.

Lemma A.3. Let $(\widehat{\boldsymbol{\theta}}(\lambda),\widehat{\sigma}(\lambda))$ be the Scaled Lasso estimator from (4.4) and $\lambda=8C\sigma\sqrt{\log(pd)/(T-p)}$. Furthermore, let the assumptions of Theorem 5.5 hold. Then,

$$\mathbb{P}\bigg(\bigg|\frac{\widehat{\sigma}(\lambda)}{\sigma}-1\bigg|>\epsilon\bigg)\to 0,$$

for all $\epsilon>0$ as $(T-p,pd)\to\infty$.

We present a proof of this lemma in Appendix B.3.

A.2. Proof of Theorem 5.9

We first present a property and three lemmas that will allow us to prove Theorem 5.9. For ease of presentation, let $G(t)=2(1-\Phi(t))$ and $G^{-1}(t)=\Phi^{-1}(1-t/2)$.

Property 1. Recall that $\widetilde{\mathbf{\Sigma}}=[\sigma_{i,j}]\in\mathbb{R}^{pd\times pd}$ is the true covariance matrix of our design matrix $\widetilde{\mathbf{X}}$. Now let $\rho_{i,j}=\sigma_{i,j}/\sqrt{\sigma_{i,i}\sigma_{j,j}}$, and for $\delta,\epsilon>0$, let $\mathcal{B}(\delta)=\{(i,j)||\rho_{i,j}|\ge\delta,i\ne j\}$ and $\mathcal{A}(\epsilon)=\mathcal{B}([\log(pd)]^{-2-\epsilon})$. By a similar argument to that made by Liu et al. (2013b), since we assume that $\text{supp}(\widehat{\boldsymbol{\theta}})\asymp\sqrt{T-p}/\log(pd)$ in Theorem 5.5, we have

$$\sum_{(i,j)\in\mathcal{A}(\epsilon)}(pd)^{a_1}=O((pd)^2/(\log(pd)^2),$$

where $a_1 = 2|\rho_{i,j}|/(1+|\rho_{i,j}|) + \delta$.

Property 1 amounts to a sparsity assumption on the true covariance matrix and allows us to cope with the correlation among test statistics.

Lemma A.4. Suppose Assumptions 5.3, 5.4, the conditions of Theorem 5.9, and Property 1 are satisfied. Then we can bound $\widehat{\nu}$ from (4.9) as follows:

$$\mathbb{P}(0 \le \widehat{\nu} \le x_{pd}) \to 1$$

where $x_{pd} = G^{-1}(y_{pd}/(pd))$, and y_{pd} is a sequence such that $y_{pd} \xrightarrow{pd} \infty$ and $y_{pd} = o(pd)$.

Lemma A.4 bounds $\widehat{\nu}$ with high probability. We use this bound in the following lemma:

Lemma A.5. Suppose Assumptions 5.3, 5.4, and Property 1 are satisfied. Then for x_{pd} as defined in Lemma A.4, we have:

$$\sup_{0\le\nu\le x_{pd}} \left| \frac{\sum_{i\in\mathcal{H}_0} \mathbb{1}(|\widehat{Z_i}| \ge \nu)}{|\mathcal{H}_0|G(\nu)} - 1 \right| \xrightarrow{P} 0.$$

Lemma A.6. Suppose Assumptions 5.3 and 5.4 are satisfied. Then we have:

$$\frac{(pd)G(\widehat{\nu})}{\max\{\sum_{1\le j\le pd} \mathbb{1}(|\widehat{Z_i} \ge \widehat{\nu}), 1\}} = \alpha.$$

In the interest of clarity, we defer the proofs of Lemmas A.4, A.5, and A.6 to Appendices B.4, B.5, and B.6, respectively.

The proof of Theorem 5.9 proceeds in three parts. We first bound $\widehat{\nu}$ with high probability in Lemma A.4, and then prove in Lemma A.5 that for any ν within those bounds

$$\frac{\sum_{i\in\mathcal{H}_0} \mathbb{1}(|\widehat{Z_i}| \ge \widehat{\nu})}{|\mathcal{H}_0|G(\widehat{\nu})} \xrightarrow{P} 1.$$

The result of Theorem 5.9 then follows naturally by Lemma B.6 .

Proof of Theorem 5.9. By Lemma A.4, $\mathbb{P}(0 \le \widehat{\nu} \le x_{pd}) \to 1$. Then by Lemma A.5, we have

$$\left| \frac{\sum_{i\in\mathcal{H}_0} \mathbb{1}(|\widehat{Z_i}| \ge \widehat{\nu})}{|\mathcal{H}_0|G(\widehat{\nu})} - 1 \right| \le \sup_{0\le\nu\le x_{pd}} \left| \frac{\sum_{i\in\mathcal{H}_0} \mathbb{1}(|\widehat{Z_i}| \ge \nu)}{|\mathcal{H}_0|G(\nu)} - 1 \right| \xrightarrow{P} 0.$$

Thus, we have

$$\frac{\sum_{i\in\mathcal{H}_0} \mathbb{1}(|\widehat{Z_i}| \ge \widehat{\nu})}{|\mathcal{H}_0|G(\widehat{\nu})} \xrightarrow{P} 1.$$

From this result, we see that by the definition of FDP(ν),

$$\frac{\text{FDP}(\widehat{\nu})}{\alpha|\mathcal{H}_0|/(pd)} = \frac{(pd)\sum_{i\in\mathcal{H}_0} \mathbb{1}(|\widehat{Z_i}| \ge \widehat{\nu})}{\alpha|\mathcal{H}_0|\max\{\sum_{1\le j\le pd} \mathbb{1}(|\widehat{Z_i} \ge \widehat{\nu}), 1\}} \xrightarrow{P} \frac{(pd)|\mathcal{H}_0|G(\widehat{\nu})}{\alpha|\mathcal{H}_0|\max\{\sum_{1\le j\le pd} \mathbb{1}(|\widehat{Z_i} \ge \widehat{\nu}), 1\}}. \tag{A.3}$$

By Lemma A.6,

$$\frac{(pd)G(\widehat{\nu})}{\max\{\sum_{1\le j\le pd} \mathbb{1}(|\widehat{Z_i} \ge \widehat{\nu}), 1\}} = \alpha. \tag{A.4}$$

Thus, by (A.3) and (A.4)

$$\frac{\text{FDP}(\widehat{\nu})}{\alpha|\mathcal{H}_0|/(pd)} \xrightarrow{P} 1,$$

as $(T-p, pd) \to \infty$. This result clearly then implies that

$$\frac{\text{FDR}(\widehat{\nu})}{\alpha|\mathcal{H}_0|/(pd)} \xrightarrow{P} 1,$$

as $(T-p, pd) \to \infty$, as desired. □

B. Proofs of Technical Lemmas in Appendix A

In this section we present the proofs of technical lemmas introduced in Appendix A.

B.1. Proof of Lemma A.1

We first present two auxiliary lemmas that we will use in the proof of Lemma A.1.

Lemma B.1. If Assumption 5.3 holds and we additionally assume that the rows of $\widetilde{\mathbf{X}}\widetilde{\mathbf{\Sigma}}^{-1/2}$ are sub-Gaussian with sub-Gaussian norm of $\kappa = \|\widetilde{\mathbf{\Sigma}}^{-1/2}\widetilde{\boldsymbol{X}}_i\|_{\psi_2}$, then

$$\|\mathbf{M}\widetilde{\mathbf{\Sigma}}_n - \mathbf{I}\|_\infty \leq a\sqrt{\frac{\log(pd)}{T-p}},$$

holds with probability at least $1 - 2(pd)^{-c_2}$, where $c_2 = \dfrac{a^2 C_{\min}}{24e^2\kappa^4 C_{\max}} - 2$ and a is some constant.

The proof of Lemma B.1, which we defer to Appendix C.1, relies constructing a martingale difference sequence (see Definition G.1 in Appendix G) and applying a Bernstein-type inequality (Lemma F.8) to this sequence.

Lemma B.2. Let $\lambda = 8C\sigma\sqrt{\log(pd)/(T-p)}$ for some constant C. Then

$$\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\|_1 \leq \frac{12\lambda s_0}{\kappa_\ell},$$

with probability at least

$$1 - b_1\exp[-b_2\sigma^2\log(pd)] - 2\exp(-c_0^2c_1^2c_2\omega^2(B)) - L\exp\left[-4\frac{(\omega(A))^2}{\alpha^2}\right],$$

where $\lambda_{\min}(\widetilde{\mathbf{\Sigma}}_n|A) = \inf_{\mathbf{u}\in A}\dfrac{1}{T-p}\|\widetilde{\mathbf{X}}\mathbf{u}\|_2^2$ is the restricted minimum eigenvalue of $\widetilde{\mathbf{\Sigma}}_n$ restricted to $A \subseteq S^{pd-1}$ (the unit sphere in $\mathbb{R}^{pd}$ space), $B = \{\widetilde{\mathbf{u}} : \widetilde{\mathbf{u}} = \widetilde{\mathbf{\Sigma}}^{1/2}\mathbf{u}/\|\widetilde{\mathbf{\Sigma}}^{1/2}\mathbf{u}\|_2, \mathbf{u} \in A\}$ is the normalized set of A, $\alpha = \operatorname{diam}(A) = \sup_{\mathbf{u},\mathbf{v}\in A} d(\mathbf{u},\mathbf{v}) = \sup_{\mathbf{u},\mathbf{v}\in A}\|\mathbf{u}-\mathbf{v}\|_2$, $\omega(A)$ is the Gaussian width of set A as defined in Definition F.5, and $\kappa_\ell, b_1, b_2, c_0, c_1, c_2, L > 0$ are constants.

The proof of Lemma B.2, which we present in Appendix C.2, relies on the restricted eigenvalue condition to bound $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\|_1$ with high probability.

We now present a proof of Lemma A.1.

Proof of Lemma A.1. Hőlder's inequality allows us to decompose $\|\mathbf{\Delta}\|_1$ as follows:

$$\|\mathbf{\Delta}\|_\infty \leq \|\mathbf{\Delta}\|_1 \leq \sqrt{T-p}\|\mathbf{M}\widetilde{\mathbf{\Sigma}}_n - \mathbf{I}\|_\infty \cdot \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\|_1.$$

We now bound $\|\mathbf{M}\widetilde{\mathbf{\Sigma}}_n - \mathbf{I}\|_\infty$ and $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\|_1$ separately. By Lemma B.1, we find that $\|\mathbf{M}\widetilde{\mathbf{\Sigma}}_n - \mathbf{I}\|_\infty \leq a\sqrt{\log(pd)/T-p}$ with high probability . Additionally, by Lemma B.2, $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\|_1 \leq 12\lambda s_0/\kappa_\ell$ with high probability

Combining these two high-probability bounds yields the following result:

$$\sqrt{T-p}\|\mathbf{M}\widetilde{\mathbf{\Sigma}}_n - \mathbf{I}\|_\infty \cdot \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\|_1 < as_0\frac{96\sigma}{\kappa_\ell}\cdot\frac{\log(pd)}{\sqrt{T-p}}, \tag{B.1}$$

with high probability. Thus, by (B.1), $\|\mathbf{\Delta}\|_\infty = o(s_0\log(pd)/\sqrt{T-p})$. Recall that by assumption, $s_0 \asymp \sqrt{T-p}/\log(pd)$. Therefore, $\|\mathbf{\Delta}\|_\infty = o(1)$. □

B.2. Proof of Lemma A.2

We first present an auxiliary lemma that we will use in the proof of Lemma A.2.

Lemma B.3 (Lindeberg condition from Hall & Heyde (1980))**.** Denote the martingale difference sequence $\zeta_{i,t} = (\epsilon_t \widetilde{\boldsymbol{X}}_t^\top \mathbf{m}_i)/(\sigma[\mathbf{m}_i^\top \widetilde{\boldsymbol{\Sigma}}_n \mathbf{m}_i]^{1/2})$ and filtration $\mathcal{F}_t = \sigma(\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_t, \epsilon_1, \ldots, \epsilon_t)$. Then, $\sum_{t=p+1}^T \mathbb{E}[\zeta_{i,t}^2 \mathbb{1}(\|\zeta_t| \geq \delta) | \mathcal{F}_{t-1}] \to 0$.

In the interest of clarity, we defer the proof of Lemma B.3 to Appendix C.3.

Proof of Lemma A.2. To prove this lemma, we need to show that

$$\frac{\widehat{Z}_i}{\sigma[\mathbf{M}\widetilde{\boldsymbol{\Sigma}}_n \mathbf{M}^\top]_{i,i}^{1/2}} = \frac{\mathbf{m}_i^\top \widetilde{\mathbf{X}}^\top \boldsymbol{\epsilon}}{\sigma[\mathbf{m}_i^\top \widetilde{\boldsymbol{\Sigma}}_n \mathbf{m}_i]^{1/2}} \xrightarrow{D} N(0,1).$$

Note that

$$\widehat{Z}_i = \mathbf{m}_i^\top \widetilde{\mathbf{X}}^\top \boldsymbol{\epsilon} = (\mathbf{m}_i^\top \widetilde{\mathbf{X}}^\top \boldsymbol{\epsilon})^\top = \boldsymbol{\epsilon}^\top \widetilde{\mathbf{X}} \mathbf{m}_i = \sum_{t=p+1}^T \epsilon_t \widetilde{\boldsymbol{X}}_t^\top \mathbf{m}_i.$$

Now define filtration $\mathcal{F}_t = \sigma(\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_t, \epsilon_1, \ldots, \epsilon_t)$. By (3.1), ϵ_t is independent of $\mathcal{F}_{t-1}$ and conditionally independent of $\widetilde{\boldsymbol{X}}_t$ given $\mathcal{F}_{t-1}$. Furthermore, by (4.2), ϵ_t is conditionally independent $\mathbf{m}_i$ given $\mathcal{F}_{t-1}$. Therefore,

$$\begin{aligned}\mathbb{E}[\epsilon_t \widetilde{\boldsymbol{X}}_t^\top \mathbf{m}_i | \mathcal{F}_{t-1}] &= \mathbb{E}[\epsilon_t | \mathcal{F}_{t-1}] \cdot \mathbb{E}[\widetilde{\boldsymbol{X}}_t^\top \mathbf{m}_i | \mathcal{F}_{t-1}] \\ &= \mathbb{E}[\epsilon_t] \cdot \mathbb{E}[\widetilde{\boldsymbol{X}}_t^\top \mathbf{m}_i | \mathcal{F}_{t-1}] \\ &= 0,\end{aligned}$$

where the last equality follows since $\epsilon_t \sim N(0, \sigma^2)$. Thus,

$$\{\zeta_{i,t}\}_{t=p+1}^T = \left\{ \frac{\epsilon_t \widetilde{\boldsymbol{X}}_t^\top \mathbf{m}_i}{\sigma[\mathbf{m}_i^\top \widetilde{\boldsymbol{\Sigma}}_n \mathbf{m}_i]^{1/2}} \right\}_{t=p+1}^T,$$

is a martingale difference sequence by Definition G.1 in Appendix G.

Since $\widehat{Z}_i/(\sigma[\mathbf{m}_i^\top \widetilde{\boldsymbol{\Sigma}}_n \mathbf{m}_i]^{1/2}) = \sum_{t=p+1}^T \zeta_{i,t}$, if we can show that we can apply the Martingale Central Limit Theorem (MCLT) (Hall & Heyde, 1980) to $\zeta_{i,t}$, the result of this lemma will follow. To demonstrate that we can apply the MCLT to $\zeta_{i,t}$, we must prove that the Lindeberg condition holds for this sequence. By Lemma B.3, the Lindeberg condition holds for $\zeta_{i,t}$. Therefore, by the MCLT

$$\sum_{t=p+1}^T \zeta_{i,t} = \frac{\widehat{Z}_i}{\sigma[\mathbf{m}_i^\top \widetilde{\boldsymbol{\Sigma}}_n \mathbf{m}_i]^{1/2}} \xrightarrow{D} N(0,1),$$

as desired.

□

B.3. Proof of Lemma A.3

We present a variation on the argument made in the Proof of Theorem 1 in Sun & Zhang (2012). The proof of Lemma A.3 requires the following two supporting lemmas from Sun & Zhang (2012), which in turn necessitate introducing some new notation. Denote the penalized least-squares loss function $L(\boldsymbol{\theta}) = (2(T-p))^{-1}\|\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$. We distinguish $L(\cdot)$ from the Scaled Lasso loss function from (4.4), which we denote $L_\lambda(\boldsymbol{\theta}, \sigma) = (2(T-p))^{-1}\|\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}\|_2^2 + \sigma/2 + \lambda\|\boldsymbol{\theta}\|_1$. As Sun & Zhang (2012) note, $\boldsymbol{\theta}$ is a critical point of L if and only if it satisfies:

$$\begin{cases} \widetilde{\boldsymbol{X}}_{\cdot,j}(\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta})/(T-p) = \lambda \mathrm{sign}(\boldsymbol{\theta}_j), \boldsymbol{\theta}_j \neq 0 \\ \widetilde{\boldsymbol{X}}_{\cdot,j}(\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta})/(T-p) \in [-\lambda, \lambda], \boldsymbol{\theta}_j \neq 0 \end{cases} \tag{B.2}$$

where $\widetilde{\boldsymbol{X}}_{\cdot,j}$ is the j-th column of the design matrix $\widetilde{\mathbf{X}}$. Importantly, Sun & Zhang (2012) note that B.2 is the Karush-Kuhn-Tucker (KKT) condition for the minimization of $L(\cdot)$ when $L(\cdot)$ is convex in $\boldsymbol{\theta}$. This property will prove important in the discussion of Lemma B.3 below. We now present the first of two supporting lemmas for the proof of Lemma A.3.

Lemma B.4 (Proposition 1 from Sun & Zhang (2012))**.** Let $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\lambda)$ be a solution path of B.2 and $\lambda_0 = \lambda/\sigma = 8C\sqrt{\log(pd)/(T-p)}$. Then the loss function $L_\lambda(\cdot,\cdot)$ is jointly convex in $(\boldsymbol{\theta},\sigma)$. Furthermore, the derivative of $L_\lambda(\cdot,\cdot)$ with respect to σ is

$$\frac{\partial}{\partial t}L_{\lambda_0}(\widehat{\boldsymbol{\theta}}(t\lambda_0),t) = \frac{1}{2} - \frac{\|\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}(t\lambda_0)\|_2^2}{2(T-p)t^2}. \tag{B.3}$$

Lemma B.4 does not rely on the i.i.d.-ness of the rows of $\widetilde{\mathbf{X}}$, and so we refer readers to the proof of Proposition 1 in Sun & Zhang (2012) for a proof of Lemma B.4.

The second supporting lemma requires additional notation from Sun & Zhang (2012). Let $\eta(\lambda,\xi,\mathbf{w},Q)$ be a prediction error bound for the estimation of $\boldsymbol{\theta}^*$ via the Scaled Lasso. Let $\mathbf{w}\in\mathbb{R}^{pd}$ $Q\subset 1\ldots,pd$, and

$$\eta(\lambda,\xi,\mathbf{w},Q) = \|\widetilde{\mathbf{X}}(\boldsymbol{\theta}^* - \mathbf{w})\|_2^2/(T-p) + 2\lambda\|\mathbf{w}_{Q^c}\|_1(2 - \mathbb{1}(\mathbf{w}=\boldsymbol{\theta}, Q=\emptyset)) + \frac{4\xi^2\lambda^2|Q|}{(\xi+1)^2\kappa(\xi,Q)}, \tag{B.4}$$

where $\mathbf{v}_S = [v_i]_{i\in S}) \in \mathbb{R}^{|S|}$ and

$$\kappa(\xi,Q) = \min\left\{\frac{|Q|^{1/2}\|\widetilde{\mathbf{X}}\mathbf{u}\|_2}{\|\mathbf{u}_Q\|_1\sqrt{T-p}} : \mathbf{u}\in E(\xi,Q), \mathbf{u}\neq 0\right\}, \tag{B.5}$$

where $E(\xi,Q) = \{\mathbf{u} : \|\mathbf{u}_{Q^c}\|_1 \leq \xi\|\mathbf{u}_Q\|_1\}$. Now let $\sigma^* = \|\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}^*\|_2/\sqrt{T-p}$, which Sun & Zhang (2012) call the oracle estimator for σ. Based on these definitions, let the minimum prediction error bound be $\eta_*(\lambda,\xi) = \inf_{\mathbf{w},Q}\eta(\lambda,\xi,\mathbf{w},Q)$ and define the following related quantity $\tau_0 = \eta_*^{1/2}(\sigma^*\lambda_0,\xi)/\sigma^*$, where recall that $\lambda_0 = \lambda/\sigma = 8C\sqrt{\log(pd)/(T-p)}$. As Sun & Zhang (2012) note, since in (3.2) $\boldsymbol{\epsilon}$ in a Gaussian random vector, σ^* is the maximum likelihood estimator for σ when $\boldsymbol{\theta}$ is known. Thus, in the proof of Lemma A.3, we bound the quantity $\widehat{\sigma}/\sigma^* - 1$ by τ_0 in order to prove the consistency of $\widehat{\sigma}$. However, we first require the following intermediate result.

Lemma B.5 (Theorem 4 from Sun & Zhang (2012))**.** Let $\widehat{\boldsymbol{\theta}}(\lambda)$ minimize $L(\cdot)$, $\xi > 0$, and define $\eta^*(\lambda,\xi) = \min_Q\left\{(1/2)(\eta(\lambda,\xi,\boldsymbol{\theta}^*,Q) + (\eta(\lambda,\xi,\boldsymbol{\theta}^*,Q) - 16\lambda^2\|\boldsymbol{\theta}_{Q^c}\|_1^2)^{1/2})\right\}$ (not to be confused with $\eta_*(\lambda,\xi)$ defined above). If $\|\widetilde{\mathbf{X}}^{top}(\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}^*)\|_\infty/(T-p) \leq \lambda(\xi-1)/(\xi+1)$, then

$$\|\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2/(T-p) \leq \min\left\{\eta_*(\lambda,\xi),\eta^*(\lambda,\xi)\right\}.$$

The proof of Lemma B.3 follows directly from the proof of Theorem 4 from Sun & Zhang (2012). The only point of contention in that proof is that B.2 must be the KKT condition for the minimization of $L(\cdot)$, which as we state above requires that $L(\cdot)$ is convex in $theta$. In Lemma C.1 in Appendix C.2, we prove that the restricted eigenvalue condition holds with high probability for the sample covariance matrix $\widetilde{\boldsymbol{\Sigma}}_n$. As we demonstrate in the proof of Lemma B.2 in Appendix C.2, the restricted eigenvalue condition implies that the unpenalized least-squares loss function $(2(T-p))^{-1}\|\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}\|_2^2$ is strictly convex for all $\boldsymbol{\theta}$ such that the error vector $\boldsymbol{\theta} - \boldsymbol{\theta}^*$ falls in the error cone $E(3,S)$, where S is the support of $\boldsymbol{\theta}^*$. Since $E(3,S)$ actually encompasses all possible error vectors, a property we prove in the proof of Lemma B.2, we see that the unpenalized loss function is convex with respect to $\boldsymbol{\theta}$. Since the derivative of penalty term $\lambda\|\boldsymbol{\theta}\|_1$ with respect to $\boldsymbol{\theta}$ is a strictly positive vector, given that the unpenalized loss function is convex in $\boldsymbol{\theta}$, so is the pealized loss function. Thus, B.2 is the KKT condition for the minimization of $L(\cdot)$, and Lemma follows immediately from the proof of Theorem 4 in Sun & Zhang (2012). We refer the reader to that paper for the full proof.

We are now ready to present the proof of Lemma A.3.

Proof of Lemma A.3. We present an amended version of the Proof of Theorem 1 from Sun & Zhang (2012). Let $z^* = \|\widetilde{\mathbf{X}}^\top(\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}^*)/(T-p)\|_\infty/\sigma^*$. Without loss of generality, assume $\tau_0 < 1$ and let $t \geq \sigma^*(1-\tau_0)$ and let $\lambda_1 = t\lambda_0$, where λ_0 is as defined above. Now note that

$$z^*\sigma^* \leq \sigma^*(1-\tau_0)\lambda_0\frac{\xi-1}{\xi+1} \leq t\lambda_0\frac{\xi-1}{\xi+1} = \lambda_1\frac{\xi-1}{\xi+1}.$$

Then by this inequality, the definition of σ^*, the Cauchy-Schwarz inequality, and Lemma B.3, we have

$$\left|\frac{\|\boldsymbol{Y}-\widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}(\lambda_1)\|_2}{\sqrt{T-p}}-\sigma^*\right| \leq \frac{\|\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\theta}}(\lambda_1)-\boldsymbol{\theta}^*)\|_2}{\sqrt{T-p}} \leq \eta_*^{1/2}(\lambda_1,\xi). \tag{B.6}$$

Now observe that B.3 in Lemma B.4 yields

$$2t^2\frac{\partial}{\partial t}L_{\lambda_0}(\widehat{\boldsymbol{\theta}}(t\lambda_0),t) = t^2 - \frac{\|\boldsymbol{Y}-\widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}(t\lambda_0)\|_2^2}{(T-p)} \leq t^2-(\sigma^*)^2(1-\tau_0)^2,$$

where the last inequality follows from B.6 and the definition of τ_0, which implies $\eta_*^{1/2}(\lambda_1,\xi)\leq\sigma^*\tau_0$ for $t<\sigma^*$, since $\lambda_1=t\lambda_0$. Note that when $t=\sigma^*(1-\tau_0)$, the expression on the right-hand side of the last inequality equals zero. Note further that $L_\lambda(\cdot,\cdot)$ is strictly convex in σ. Then the negativity of $2t^2\partial/(\partial t)L_{\lambda_0}(\widehat{\boldsymbol{\theta}}(t\lambda_0),t)$ implies that $\widehat{\sigma}\geq\sigma^*(1-\tau_0)$. On the other hand, at $t=\sigma^*/(1-\tau_0)>\sigma^*$, we have

$$t^2-\frac{\|\boldsymbol{Y}-\widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}(t\lambda_0)\|_2^2}{(T-p)}\geq t^2-(\sigma+t\tau_0)^2\geq 0,$$

since for $t>\sigma^*$, we have $\eta_*^{1/2}(\lambda_1,\xi)\leq t\tau_0$. This result implies $\sigma^*\geq\widehat{\sigma}(1-\tau_0)$ by the strict convexity of $L_\lambda(\cdot,\cdot)$ in σ. Therefore,

$$\max\left(1-\frac{\widehat{\sigma}}{\sigma^*},1-\frac{\sigma^*}{\widehat{\sigma}}\right)\leq\tau_0. \tag{B.7}$$

As Sun & Zhang (2012) argue, $\eta_*^{1/2}(\lambda_1,\xi)/\sigma\to 0$ as $(T-p,pd)\to\infty$. Thus,

$$\mathbb{P}\left(\left|\frac{\widehat{\sigma}(\lambda)}{\sigma}-1\right|>\epsilon\right)\to 0,$$

for all $\epsilon>0$ as $(T-p,pd)\to\infty$. □

B.4. Proof of Lemma A.4

We will need the following lemma, which is a modified version on Lemma 6.1 from Liu et al. (2013b) to prove Lemma A.4.

Lemma B.6. Let $\xi_1,\ldots,\xi_n\in\mathbb{R}^p$ have mean zero. Suppose that $p\leq n^r$, $\log(p)=o(\sqrt{n})$, and $\mathbb{E}[\|\xi_i\|_2^{bpr+2+\epsilon}]\leq\infty$, for $r,b,\epsilon>0$. Furthermore, assume that $\|\operatorname{Cov}(\xi_i)-\mathbf{I}_{p\times p}\|_2\leq C(\log(p))^{-2-\gamma}$, where $\operatorname{Cov}(\xi_i)=\mathbb{E}[(1/n)\sum_{i=1}^n\xi_i\xi_i^\top]$ and $\gamma>0$. Define $\|\cdot\|_{\min}$ as $\|\mathbf{v}\|_{\min}=\min_{1\leq i\leq p}\{|v_i|\}$. Then,

$$\sup_{0\leq t\leq b\sqrt{\log(p)}}\left|\frac{\mathbb{P}(\|\sum_{i=1}^n\xi_i\|_{\min}\geq t\sqrt{n})}{(G(t))^p}-1\right|\leq C(\log(p))^{-1-\gamma_1},$$

where $\gamma_1=\min\{\gamma,1/2\}$.

Note that we make the assumptions $p\leq n^r$, $r>0$ and $\log(p)=o(\sqrt{n})$, in the statement of Theorem 5.9. The former assumption is clearly satisfied in our setting, as we can have $r>1$. As noted by Liu & Luo (2014), given that $p\leq n^r$, for $r>0$, our assumption 5.4 of $\widetilde{\mathbf{X}}$ having bounded sub-Gaussian rows is equivalent to $\mathbb{E}[\|\xi_i\|_2^{bpr+2+\epsilon}]\leq\infty$, for $b,\epsilon>0$. Additionally, the assumption of $\|\operatorname{Cov}(\xi_i)-\mathbf{I}_{p\times p}\|_2\leq C(\log(p)^{-2-\gamma}$ is satisfied by Property 1, the sparsity property of the covariance matrix. Whereas Liu et al. (2013b) prove Lemma B.6 for the i.i.d. case, we present a proof that draws on the Bernstein inequality for martingale difference sequences (Lemma F.8) to prove this lemma when $\xi_1,\ldots,\xi_n\in\mathbb{R}^p$ are not independent. We defer the proof of Lemma B.6 to Appendix C.4. Having established Lemma B.6, we now present the proof of Lemma A.4.

Proof of Lemma A.4. By Lemma B.6, we know that

$$\max_{1\leq i\leq pd}\sup_{0\leq\nu\leq 4\sqrt{\log(pd)}}\left|\frac{\mathbb{P}(\widehat{Z_i}\geq\nu)}{G(\nu)}-1\right|\leq C(\log(pd))^{-1-\gamma_1}, \tag{B.8}$$

for positive constants C and γ_1. Then by (B.8), we have that $\mathbb{P}(|\widehat{Z}_i| \geq \sqrt{2\log(pd)}) \rightarrow 1$ for all $i \in \mathcal{B} = \{i||\theta_i^*|/(\sigma\widetilde{\mathbf{\Sigma}}_{i,i}^{-1/2}) \geq \sqrt{c\log(pd)/(T-p)}\}$. Thus,

$$\frac{\sum_{i\in\mathcal{B}}\mathbb{P}(|\widehat{Z}_i| \geq \sqrt{2\log(pd)})}{|\mathcal{B}|} \xrightarrow{P} 1, \tag{B.9}$$

since we assumed by Assumption 5.8 that $|\mathcal{B}| = \sum_{i\in\mathcal{H}_1}\mathbb{1}\{|\theta_i^*|/(\sigma\widetilde{\mathbf{\Sigma}}_{i,i}^{-1/2}) \geq \sqrt{c\log(pd)/(T-p)}\} \rightarrow \infty$. If the number of true alternatives were fixed, this convergence would clearly not occur. Now note that by Markov's inequality

$$\begin{aligned}\mathbb{P}\bigg(\sum_{i\in\mathcal{B}}\mathbb{1}\{|\widehat{Z}_i| \geq \sqrt{2\log(pd)})\} \geq |\mathcal{B}|\bigg) &\leq \frac{\mathbb{E}\Big[\sum_{i\in\mathcal{B}}\mathbb{1}\{|\widehat{Z}_i| \geq \sqrt{2\log(pd)})\}\Big]}{|\mathcal{B}|} \\ &= \frac{\sum_{i\in\mathcal{B}}\mathbb{P}(|\widehat{Z}_i| \geq \sqrt{2\log(pd)}))}{|\mathcal{B}|} \\ &\xrightarrow{P} 1,\end{aligned}$$

where the convergence follows by (B.9). Therefore, since $\sum_{i\in\mathcal{B}}\mathbb{1}\{|\widehat{Z}_i| \geq \sqrt{2\log(pd)})\} \leq |\mathcal{B}|$, we have

$$\frac{\sum_{i\in\mathcal{B}}\mathbb{1}\{|\widehat{Z}_i| \geq \sqrt{2\log(pd)})\}}{|\mathcal{B}|} \xrightarrow{P} 1.$$

This line implies that for $0 \leq \widehat{\nu} \leq x_{pd}$, our FDR control procedure will correctly identify all true positives that meet a certain minimum signal condition. The result of this lemma then follows from the definition of $\widehat{\nu}$ Section 4.2. □

B.5. Proof of Lemma A.5

We will need the following lemma, which is a modified version on Lemma 6.2 from Liu et al. (2013b) to prove Lemma A.5.

Lemma B.7. Let $\boldsymbol{\eta_1}, \ldots, \boldsymbol{\eta_n}$ have mean zero, where $\boldsymbol{\eta_t} = (\eta_{t,1}, \eta_{t,2})^\top$. Suppose that $p \leq n^r$, $\log(p) = o(\sqrt{n})$, and $\mathbb{E}[\|\xi_i\|_2^{bpr+2+\epsilon}] \leq \infty$, for $r, b, \epsilon > 0$. Furthermore, assume that $V[\eta_{t,1}] = V[\eta_{t,2}]$ and $|Cov(\eta_{t,1}, \eta_{t,2})| \leq \delta$, for some $0 \leq \delta \leq 1$. Then

$$\mathbb{P}\bigg(\bigg|\sum_{i=1}^n \eta_{i,1}\bigg| \geq t\sqrt{n}, \bigg|\sum_{i=1}^n \eta_{i,2}\bigg| \geq t\sqrt{n}\bigg) \leq C(t+1)^{-2}\exp[-t^2/(1+\delta)],$$

for $0 \leq t \leq b\log(2)$, where C depends only on b, r, ϵ, δ.

The proof of Lemma B.7 follows almost exactly the Proof of Lemma 6.2 in Liu et al. (2013b). The only difference is that whereas Lemma 6.2 in Liu et al. (2013b) requires i.i.d. $\boldsymbol{\eta_t}$ vectors in order to cite the Proof of Lemma 6.1 in Liu et al. (2013b), we do not require i.i.d. $\boldsymbol{\eta_t}$ vectors and instead appeal to the proof of Lemma B.6. We refer the reader to Liu et al. (2013b) for more details. Having established Lemma B.7, we now present the proof of Lemma A.5.

Proof of Lemma A.5. Let $b_0 < b_1 < \ldots < b_k$ and $\nu_i = G^{-1}(b_i)$, where $b_0 = y_{pd}/(pd)$, $b_i = y_{pd}/(pd) + y_{pd}^{2/3}e^{i^\delta}/(pd)$, $k = [\log((pd - y_{pd})/y_{pd}^{2/3})]^{1/\delta}$, and $0 < \delta < 1$. Then we have $G(\nu_i)/G(\nu_{i+1}) = 1 + o(1)$ for all $0 \leq i \leq k$, and $\nu_0/\sqrt{2\log(pd/y_{pd})} = 1 + o(1)$. One can easily verify that $0 \leq j \leq k \leftrightarrow 0 \leq \nu \leq y_{pd}$. So we see that to prove this lemma it suffices to show that

$$\max_{0\leq j\leq k}\bigg|\frac{\sum_{i\in\mathcal{H}_0}[\mathbb{1}(|\widehat{Z}_i| \geq \nu_j) - G(\nu_j)]}{|\mathcal{H}_0|G(\nu_j)}\bigg| \xrightarrow{P} 0. \tag{B.10}$$

Observe that for all $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\bigg(\max_{0\leq j\leq k}\bigg|\frac{\sum_{i\in\mathcal{H}_0}[\mathbb{1}(|\widehat{Z}_i| \geq \nu_j) - G(\nu_j)]}{|\mathcal{H}_0|G(\nu_j)}\bigg| \geq \epsilon\bigg) &\leq \mathbb{P}\bigg(\max_{0\leq j\leq k}\bigg|\frac{\sum_{i\in\mathcal{H}_0}[\mathbb{1}(|\widehat{Z}_i| \geq \nu_j) - G(\nu_j)]}{|\mathcal{H}_0|G(\nu_j)}\bigg| \geq \frac{\epsilon}{2}\bigg) \\ &\leq \sum_{j=0}^k \mathbb{P}\bigg(\bigg|\frac{\sum_{i\in\mathcal{H}_0}[\mathbb{1}(|\widehat{Z}_i| \geq \nu_j) - G(\nu_j)]}{|\mathcal{H}_0|G(\nu_j)}\bigg| \geq \frac{\epsilon}{2}\bigg).\end{aligned} \tag{B.11}$$

Now let

$$I(\nu) = \frac{\sum_{i\in\mathcal{H}_0}[\mathbb{1}(|\widehat{Z}_i| \geq \nu) - \mathbb{P}(|\widehat{Z}_i| \geq \nu)]}{|\mathcal{H}_0|G(\nu)}.$$

Note that

$$I(\nu) = \frac{\sum_{i\in\mathcal{H}_0}[\mathbb{1}(|\widehat{Z}_i| \geq \nu) - \mathbb{P}(|\widehat{Z}_i| \geq \nu)]}{|\mathcal{H}_0|G(\nu)} \xrightarrow{P} \frac{\sum_{i\in\mathcal{H}_0}[\mathbb{1}(|\widehat{Z}_i| \geq \nu_j) - G(\nu_j)]}{|\mathcal{H}_0|G(\nu_j)},$$

by (B.8) in the proof of Lemma A.4 in Appendix B.4. Clearly, $\mathbb{E}[I(\nu)] = 0$, so if we can show that $V[I(\nu)] = \mathbb{E}[I^2(\nu)] \to 0$, we will have that $I(\nu) \xrightarrow{P} 0$. By (B.11), this convergence will prove (B.10). We now decompose $V[I(\nu)] = \mathbb{E}[I^2(\nu)]$ as follows:

$$\begin{aligned}\mathbb{E}[I^2(\nu)] &= \frac{\sum_{i\in\mathcal{H}_0}[\mathbb{P}(|\widehat{Z}_i| \geq \nu) - \mathbb{P}^2(|\widehat{Z}_i| \geq \nu)]}{|\mathcal{H}_0|^2G^2(\nu)} + \frac{\sum_{i,j\in\mathcal{H}_0, i\neq j}[\mathbb{P}(|\widehat{Z}_i| \geq \nu, |\widehat{Z}_j| \geq \nu) - \mathbb{P}(|\widehat{Z}_i| \geq \nu)\mathbb{P}(|\widehat{Z}_j| \geq \nu)]}{|\mathcal{H}_0|^2G^2(\nu)} \\ &\leq \frac{C|\mathcal{H}_0|G(\nu)}{(|\mathcal{H}_0|G(\nu))^2} + \frac{1}{G^2(\nu)|\mathcal{H}_0|^2}\sum_{(i,j)\in\mathcal{A}(\epsilon)}\mathbb{P}(|\widehat{Z}_i| \geq \nu, |\widehat{Z}_j| \geq \nu) + \frac{1}{|\mathcal{H}_0|^2}\sum_{i,j\in\mathcal{H}_0\bigcap\mathcal{A}(\epsilon)^c}\left(\frac{\mathbb{P}(|\widehat{Z}_i| \geq \nu, |\widehat{Z}_j| \geq \nu)}{G^2(\nu)} - 1\right),\end{aligned} \tag{B.12}$$

where the equality follows by direct computation, and the inequality holds by (B.8) in the proof of Lemma A.4 in Appendix B.4. If we let

$$I_{1,1}(\nu) = \frac{1}{G^2(\nu)|\mathcal{H}_0|^2}\sum_{(i,j)\in\mathcal{A}(\epsilon)}\mathbb{P}(|\widehat{Z}_i| \geq \nu, |\widehat{Z}_j| \geq \nu),$$

and

$$I_{1,2}(\nu) = \frac{1}{|\mathcal{H}_0|^2}\sum_{i,j\in\mathcal{H}_0\bigcap\mathcal{A}(\epsilon)^c}\left(\frac{\mathbb{P}(|\widehat{Z}_i| \geq \nu, |\widehat{Z}_j| \geq \nu)}{G^2(\nu)} - 1\right),$$

then (B.12) yields

$$\mathbb{E}[I^2(\nu)] \leq \frac{C}{|\mathcal{H}_0|G(\nu)} + I_{1,1}(\nu) + I_{1,2}(\nu). \tag{B.13}$$

Applying Lemma B.6 to $I_{1,2}(\nu)$ yields the following result for all $0 \leq t \leq \sqrt{2\log(pd)}$ for some $\delta > 0$:

$$|I_{1,2}(\nu)| \leq C(\log(pd))^{-1-\delta}. \tag{B.14}$$

Furthermore, Lemma B.7 yields

$$\mathbb{P}(|\widehat{Z}_i| \geq \nu, |\widehat{Z}_j| \geq \nu) \leq C\exp\left[\frac{-\nu^2}{1 + |\rho_{i,j}| + \delta_1}\right], \tag{B.15}$$

for all $(i,j) \in \mathcal{A}(\epsilon)$ and $i,j \in \mathcal{H}_0$, where $\delta_1, C > 0$. Lastly, the proceeding result follows from (B.15) and Property 1

$$I_{1,1}(\nu) \leq C(\log(pd))^{-2}. \tag{B.16}$$

Then by (F.2), (B.14), and (B.16), we have that

$$\begin{aligned}\sum_{j=0}^{k}\mathbb{E}[I(\nu_j)]^2 &\leq Ck[(\log(pd))^{1-\delta} + (\log(pd))^{-2}] + C\sum_{j=0}^{k}(pdG(\nu_j))^{-1} \\ &\leq C\sum_{j=0}^{k}\frac{1}{y_{pd} + y_{pd}^{2/3}e^{j^\delta}} + o(1) \\ &= o(1).\end{aligned}$$

The second inequality holds by the definition of the sequence ν_i and since $k = o(\log(pd))$. The third inequality follows since $1/(y_{pd} + y_{pd}^{2/3}e^{j^\delta}) = o(1/y_{pd}) = o(1/(pd))$. Our desired result (B.10) follows naturally from this last set of inequalities. □

B.6. Proof of Lemma A.6

We present a variation on the argument made in the Proof of Theorem 3.1 in Liu et al. (2013b).

Proof of Lemma A.6. Since by the definition of $\widehat{\nu}$ in 4.9, $\widehat{\nu}$ is the infimum of all values $\nu > 0$ such that FDP$(\nu) \leq \alpha$, for $\nu < \widehat{\nu}$

$$\frac{\sum_{i\in\mathcal{H}_0} \mathbb{1}(|\widehat{Z_i}| \geq \nu)}{\max\{\sum_{1\leq j\leq pd} \mathbb{1}(|\widehat{Z_i} \geq \nu), 1\}} > \alpha.$$

Using the asymptotic normality of $\widehat{Z_i}$, which holds by Theorem 5.5, we can approximate $\sum_{i\in\mathcal{H}_0} \mathbb{1}(|\widehat{Z_i}| \geq \nu)$ by $(pd)G(\nu)$, yielding

$$\frac{(pd)G(\nu)}{\max\{\sum_{1\leq j\leq pd} \mathbb{1}(|\widehat{Z_i} \geq \nu), 1\}} > \alpha,$$

for $\nu < \widehat{\nu}$. Note that $(pd)G(\nu)/\max\{\sum_{1\leq j\leq pd} \mathbb{1}(|\widehat{Z_i} \geq \nu), 1\}$ is decreasing in ν. Then by letting ν approach $\widehat{\nu}$, we obtain

$$\frac{(pd)G(\widehat{\nu})}{\max\{\sum_{1\leq j\leq pd} \mathbb{1}(|\widehat{Z_i} \geq \widehat{\nu}), 1\}} \geq \alpha. \tag{B.17}$$

Now to prove the reverse bound, we note that the definition of infimum implies the existence of a sequence ν_k, where $\nu_k \geq \widehat{\nu}$ and $\nu_k \xrightarrow{k\to\infty} \widehat{\nu}$. Since $\nu_k \geq \widehat{\nu}$, we have

$$\frac{(pd)G(\nu_k)}{\max\{\sum_{1\leq j\leq pd} \mathbb{1}(|\widehat{Z_i} \geq \nu_k), 1\}} \leq \alpha.$$

Thus, by letting $\nu_k \to \widehat{\nu}$, we see that

$$\frac{(pd)G(\widehat{\nu})}{\max\{\sum_{1\leq j\leq pd} \mathbb{1}(|\widehat{Z_i} \geq \widehat{\nu}), 1\}} \leq \alpha. \tag{B.18}$$

Therefore, by (B.17) and (B.18)

$$\frac{(pd)G(\widehat{\nu})}{\max\{\sum_{1\leq j\leq pd} \mathbb{1}(|\widehat{Z_i} \geq \widehat{\nu}), 1\}} = \alpha,$$

as desired. □

C. Proofs of Auxiliary Lemmas in Appendix B

In this section we present the proofs of auxiliary lemmas introduced in Appendix B.

C.1. Proof of Lemma B.1

Here we present a modified version of the proof for Theorem 7.(b) from Javanmard & Montanari (2014).

Proof of Lemma B.1. Clearly $\|\mathbf{M}\widetilde{\mathbf{\Sigma}}_n - \mathbf{I}\|_\infty \leq \|\widetilde{\mathbf{\Sigma}}^{-1}\widetilde{\mathbf{\Sigma}}_n - \mathbf{I}\|_\infty$. Let $\overline{\boldsymbol{X}}_t = \widetilde{\mathbf{\Sigma}}^{-1/2}\widetilde{\boldsymbol{X}}_t$, where $\widetilde{\boldsymbol{X}}_t$ is as defined in Section 3. Now define $\mathbf{Z} \in \mathbb{R}^{pd\times pd}$ as follows:

$$\mathbf{Z} = \widetilde{\mathbf{\Sigma}}^{-1}\widetilde{\mathbf{\Sigma}}_n - \mathbf{I} = \frac{1}{T-p}\sum_{t=p+1}^{T}\left(\widetilde{\mathbf{\Sigma}}^{-1}\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{X}}_t^\top - \mathbf{I}\right) = \frac{1}{T-p}\sum_{t=p+1}^{T}\left(\widetilde{\mathbf{\Sigma}}^{-1/2}\overline{\boldsymbol{X}}_t\overline{\boldsymbol{X}}_t^\top\widetilde{\mathbf{\Sigma}}^{1/2} - \mathbf{I}\right).$$

For any given pair $1 \leq i,j \leq pd$, denote $\gamma_t^{(ij)} = \langle \widetilde{\mathbf{\Sigma}}_{i,\cdot}^{-1/2}, \overline{\boldsymbol{X}}_t\rangle \cdot \langle \widetilde{\mathbf{\Sigma}}_{j,\cdot}^{1/2}, \overline{\boldsymbol{X}}_t\rangle - \delta_{i,j}$, where $p+1 \leq t \leq T$, and $\delta_{i,j}$ represents the Kronecker delta: $\delta_{i,j} = \mathbb{1}(i=j)$. Let $\mathcal{F}_t$ be the filtration $\mathcal{F}_t = \sigma(\overline{\boldsymbol{X}}_1, \ldots, \overline{\boldsymbol{X}}_t)$. Then note that

$\mathbb{E}[\gamma_t^{(ij)}|\mathcal{F}_{t-1}] = 0$, so by Definition G.1 in Appendix G, $\gamma_t^{(ij)}$ forms a martingale difference sequence. Note further that $\mathbf{Z}_{i,j} = (T-p)^{-1}\sum_{t=p+1}^{T}\gamma_t^{(ij)}$. Thus, to bound $\|\widetilde{\mathbf{\Sigma}}^{-1}\widetilde{\mathbf{\Sigma}}_n - \mathbf{I}\|_\infty$, and by transitivity $\|\mathbf{M}\widetilde{\mathbf{\Sigma}}_n - \mathbf{I}\|_\infty$, we need only bound $\mathbf{Z}_{i,j}$.

We refer the reader to Definition G.2 in Appendix G for the definition of the sub-exponential norm. By Remark 5.18 (Centering) from Vershynin (2012), we can bound the sub-exponential norm of $\gamma_t^{(ij)}$ as follows:

$$\|\gamma_t^{(ij)}\|_{\psi_1} \le 2\|\langle\widetilde{\mathbf{\Sigma}}_{i,\cdot}^{-1/2}, \overline{\boldsymbol{X}}_t\rangle \cdot \langle\widetilde{\mathbf{\Sigma}}_{j,\cdot}^{1/2}, \overline{\boldsymbol{X}}_t\rangle\|_{\psi_1}. \tag{C.1}$$

Now note that, as shown by Javanmard & Montanari (2014) we can bound the sub-exponential norm of the product of two random variables X and Y by:

$$\begin{aligned}
\|XY\|_{\psi_1} &\le \sup_{q\ge 1} q^{-1}\left(\mathbb{E}\left[|XY|^q\right]\right)^{1/q}\\
&\le \sup_{q\ge 1} q^{-1}\left(\mathbb{E}\left[|X|^{2q}\right]\right)^{1/2q}\left(\mathbb{E}\left[|Y|^{2q}\right]\right)^{1/2q}\\
&\le 2\left(\sup_{r\ge 2} r^{-1/2}\left(\mathbb{E}\left[|X|^{r}\right]\right)^{1/r}\right)\left(\sup_{r\ge 2} r^{-1/2}\left(\mathbb{E}\left[|Y|^{r}\right]\right)^{1/r}\right)\\
&\le 2\|X\|_{\psi_2}\cdot\|Y\|_{\psi_2}.
\end{aligned}$$

Therefore, by (C.1)

$$\|\gamma_t^{(ij)}\|_{\psi_1} \le 2\|\langle\widetilde{\mathbf{\Sigma}}_{i,\cdot}^{-1/2}, \overline{\boldsymbol{X}}_t\rangle \cdot \langle\widetilde{\mathbf{\Sigma}}_{j,\cdot}^{1/2}, \overline{\boldsymbol{X}}_t\rangle\|_{\psi_1} \le 2\|\langle\widetilde{\mathbf{\Sigma}}_{i,\cdot}^{-1/2}, \overline{\boldsymbol{X}}_t\rangle\|_{\psi_2}\cdot\|\langle\widetilde{\mathbf{\Sigma}}_{j,\cdot}^{1/2}, \overline{\boldsymbol{X}}_t\rangle\|_{\psi_2},$$

and by assumption

$$2\|\langle\widetilde{\mathbf{\Sigma}}_{i,\cdot}^{-1/2}, \overline{\boldsymbol{X}}_t\rangle\|_{\psi_2}\cdot\|\langle\widetilde{\mathbf{\Sigma}}_{j,\cdot}^{1/2}, \overline{\boldsymbol{X}}_t\rangle\|_{\psi_2} \le 2\|\widetilde{\mathbf{\Sigma}}_{i,\cdot}^{-1/2}\|_{\psi_2}\cdot\|\widetilde{\mathbf{\Sigma}}_{j,\cdot}^{1/2}\|_{\psi_2}\kappa^2 \le 2\kappa^2\sqrt{\frac{C_{\max}}{C_{\min}}}.$$

Thus, if we let $\kappa' = 2\kappa^2\sqrt{C_{\max}/C_{\min}}$, then $\|\gamma_t^{(ij)}\|_{\psi_1} \le \kappa'$. Now, since $\gamma_t^{(ij)}$ is a martingale difference sequence, we can apply the Bernstein inequality for martingale difference sequences (Lemma F.8 in Appendix F) to obtain:

$$\mathbb{P}\left(\frac{1}{T-p}\left|\sum_{t=p+1}^{T}\gamma_t^{(ij)}\right| \ge \epsilon\right) \le 2\exp\left[-\frac{T-p}{6}\min\left(\left(\frac{\epsilon}{e\kappa'}\right)^2, \frac{\epsilon}{e\kappa'}\right)\right].$$

Let $\epsilon = a\sqrt{\log(pd)/(T-p)}$, and assume that $T-p \ge (a/(e\kappa'))^2\log(pd)$ so that $(\epsilon/(e\kappa'))^2 \le (\epsilon/(e\kappa')) \le 1$. Then,

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{T-p}\left|\sum_{t=p+1}^{T}\gamma_t^{(ij)}\right| \ge a\sqrt{\frac{\log(pd)}{T-p}}\right) &= \mathbb{P}\left(\left|\mathbf{Z}_{i,j}\right| \ge a\sqrt{\frac{\log(pd)}{T-p}}\right)\\
&\le 2(pd)^{-a^2/(6e^2\kappa'^2)}\\
&= 2(pd)^{-(a^2C_{\min})/(24e^2\kappa^4C_{\max})}.
\end{aligned}$$

Taking the union over all $(pd)^2$ pairs and letting $c_2 = (a^2C_{\min})/(24e^2\kappa^4C_{\max}) - 2$ yields the result:

$$\mathbb{P}\left(\|\mathbf{M}\widetilde{\mathbf{\Sigma}}_n - \mathbf{I}\|_\infty \le a\sqrt{\frac{\log(pd)}{T-p}}\right) \ge 1 - 2(pd)^{-c_2}.$$

□

C.2. Proof of Lemma B.2

The proof of Lemma B.2 relies on two lemmas. First, the following lemma asserts that the *restricted eigenvalue condition* (RE condition) holds true for our design matrix $\widetilde{\mathbf{X}}$. As we will see in the proof of Lemma B.2 below, the restricted eigenvalue condition implies the *restricted strong convexity condition* when the loss function is the least squares loss function. This property will prove instrumental in bounding $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1$.

Lemma C.1. Under Assumptions 5.3 and 5.4, we have

$$\inf_{\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \in E_r} \frac{1}{T-p} \|\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2 > 0$$

with probability at least

$$1 - 2\exp(-c_0^2 c_1^2 c_2 \omega^2(B)) - L\exp\left[-4\frac{(\omega(A))^2}{\alpha^2}\right],$$

where where $\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_n | A) = \inf_{\mathbf{u} \in A} \frac{1}{T-p}\|\widetilde{\mathbf{X}}\mathbf{u}\|_2^2$ is the restricted minimum eigenvalue of $\widetilde{\boldsymbol{\Sigma}}_n$ restricted to $A \subseteq S^{pd-1}$ (the unit sphere in $\mathbb{R}^{pd}$ space), $B = \{\widetilde{\mathbf{u}} : \widetilde{\mathbf{u}} = \widetilde{\boldsymbol{\Sigma}}^{1/2}\mathbf{u}/\|\widetilde{\boldsymbol{\Sigma}}^{1/2}\mathbf{u}\|_2, \mathbf{u} \in A\}$ is the normalized set of A, $\alpha = \text{diam}(A) = \sup_{\mathbf{u},\mathbf{v} \in A} d(\mathbf{u}, \mathbf{v}) = \sup_{\mathbf{u},\mathbf{v} \in A} \|\mathbf{u} - \mathbf{v}\|_2$, and $c_0, c_1, c_2, L > 0$ are constants.

The RE condition has been studied extensively in the setting where the rows of the design matrix are independent. However, since the rows of the design matrix are dependent in this setting, we must appeal to martingale theory to prove the RE condition. We construct a martingale difference sequence equal to $1/(T-p)\|\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2$, and then bound the minimum restricted eigenvalue of that sequence using the results from Appendix F. We defer the proof of Lemma C.1 to Appendix D.1.

Second, the following lemma establishes with high-probability a property of the regularization parameter λ.

Lemma C.2. Denote the least-squares loss function $\mathcal{L}(\boldsymbol{\theta}) = (2(T-p))^{-1}\|\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}\|_2^2$. If Assumption 5.4 holds and $\lambda = 8C\sigma\sqrt{\log(pd)/(T-p)}$ for some constant C, then

$$\lambda \geq 2\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty,$$

holds with probability at least $1 - b_1 \exp[-b_2\sigma^2 \log(pd)]$, for constants b_1 and b_2.

We defer the proof of Lemma C.2 to Appendix D.2.

We now present a proof of Lemma B.2.

Proof of Lemma B.2. To prove this lemma, we first recall the form of the biased Lasso Granger estimator from (3.3):

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2(T-p)}\|\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1.$$

For brevity, we denote the loss function $\mathcal{L}(\boldsymbol{\theta}) = (2(T-p))^{-1}\|\boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}\|_2^2$. We immediately realize that the optimality of $\widehat{\boldsymbol{\theta}}$ yields the following inequality:

$$\mathcal{L}(\widehat{\boldsymbol{\theta}}) + \lambda\|\widehat{\boldsymbol{\theta}}\|_1 \leq \mathcal{L}(\boldsymbol{\theta}^*) + \lambda\|\boldsymbol{\theta}^*\|_1. \tag{C.2}$$

To establish a lower bound on $\mathcal{L}(\widehat{\boldsymbol{\theta}})$, we will appeal to the *restricted strong convexity condition* (RSC condition), which provides a lower bound on the first-degree Taylor approximation of $\mathcal{L}(\widehat{\boldsymbol{\theta}})$:

$$\delta\mathcal{L}(\widehat{\boldsymbol{\theta}}) := \mathcal{L}(\widehat{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^*) - \langle\nabla\mathcal{L}(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\rangle \geq \kappa_\ell\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 > 0, \tag{C.3}$$

for some constant κ'_ℓ. As noted by Negahban et al. (2012), when $\mathcal{L}(\cdot)$ is the least-squares loss function, as it is in our setting, we obtain

$$\delta\mathcal{L}(\widehat{\boldsymbol{\theta}}) = \frac{1}{T-p}\|\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2 \geq \kappa'_\ell\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2. \tag{C.4}$$

Note that

$$\inf_{\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\in E_r} \frac{1}{T-p}\|\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\|_2^2 = \lambda_{\min}\left(\frac{1}{T-p}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}|E_r\right) = \lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_n|E_r), \tag{C.5}$$

where E_r is the set the error vector $\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*$ can fall in, and $\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_n|E_r)$ is the minimum restricted eigenvalue of the sample covariance matrix. So we see that when $\mathcal{L}(\cdot)$ is the least-squares loss function, the restricted strong convexity condition collapses into the RE condition. The RE condition holds for $\widetilde{\mathbf{X}}$ with high probability by Lemma C.1. Thus, $\lambda_m in(\widetilde{\boldsymbol{\Sigma}}_n|E_r) > 0$, and so the RSC condition holds. Then, rearranging (C.3), we see that

$$\mathcal{L}(\widehat{\boldsymbol{\theta}}) \geq \mathcal{L}(\boldsymbol{\theta}^*) + \langle \nabla\mathcal{L}(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\rangle + \frac{\kappa_\ell}{2}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2, \tag{C.6}$$

where $\kappa_\ell = 2\kappa_\ell'$.

As a consequence of (C.2) and (C.6), we have

$$\mathcal{L}(\boldsymbol{\theta}^*) + \langle \nabla\mathcal{L}(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\rangle + \lambda\|\widehat{\boldsymbol{\theta}}\|_1 \leq \mathcal{L}(\widehat{\boldsymbol{\theta}}) + \lambda\|\widehat{\boldsymbol{\theta}}\|_1 \leq \mathcal{L}(\boldsymbol{\theta}^*) + \lambda\|\boldsymbol{\theta}^*\|_1.$$

Furthermore, since $\langle \nabla\mathcal{L}(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\rangle \leq \|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty \cdot \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_1$ by Holder's inequality, we achieve the following result:

$$-\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty \cdot \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_1 + \lambda\|\widehat{\boldsymbol{\theta}}\|_1 \leq \lambda\|\boldsymbol{\theta}^*\|_1. \tag{C.7}$$

We apply Lemma C.2 to establish that $\lambda \geq 2\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_\infty$ with probability at least $1 - b_1\exp[-b_2\sigma^2\log(pd)]$, for constants b_1 and b_2. Thus, since $\lambda \geq 2\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty$ with high probability, (C.7) implies

$$-\frac{1}{2}\lambda\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_1 + \lambda\|\widehat{\boldsymbol{\theta}}\|_1 \leq \lambda\|\boldsymbol{\theta}^*\|_1.$$

Now denote S to be the support of $\boldsymbol{\theta}^*$, so that $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*_S$ and $\boldsymbol{\theta}^*_{S^c} = \mathbf{0}$. Then, based on the previous inequality, we have

$$-\frac{1}{2}\lambda\|(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)_S\|_1 - \frac{1}{2}\lambda\|(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)_{S^c}\|_1 + \lambda\|\widehat{\boldsymbol{\theta}}_S\|_1 + \lambda\|\widehat{\boldsymbol{\theta}}_{S^c}\|_1 \leq \lambda\|\boldsymbol{\theta}^*_S\|_1.$$

Rearranging these terms yields:

$$\begin{aligned}-\frac{1}{2}\lambda\|(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)_S\|_1 - \frac{1}{2}\lambda\|(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)_{S^c}\|_1 + \lambda\|(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)_{S^c}\|_1 &\leq \lambda\|\boldsymbol{\theta}^*_S\|_1 - \lambda\|\widehat{\boldsymbol{\theta}}_S\|_1 \\ &\leq \lambda\|(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)_S\|_1,\end{aligned}$$

where the second inequality follows by the triangle inequality. Rearranging terms once more produces the following result:

$$\lambda\|(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)_{S^c}\|_1 \leq 3\lambda\|(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)_S\|_1. \tag{C.8}$$

We use (C.8) to bound $\lambda\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_1$. Note that (C.2) and (C.6) together imply

$$-\frac{1}{2}\lambda\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_1 + \lambda\|\widehat{\boldsymbol{\theta}}\|_1 + \frac{\kappa_\ell}{2}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2 \leq \lambda\|\boldsymbol{\theta}^*\|_1,$$

and thus,

$$\frac{\kappa_\ell}{2}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2 \leq \lambda\|\boldsymbol{\theta}^*\|_1 - \lambda\|\widehat{\boldsymbol{\theta}}\|_1 + \frac{1}{2}\lambda\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_1.$$

Applying the triangle inequality yields

$$\begin{aligned}\frac{\kappa_\ell}{2}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2 &\leq \lambda\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_1 + \frac{1}{2}\lambda\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_1 \\ &= \frac{3}{2}\lambda\|(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\|_1.\end{aligned} \tag{C.9}$$

Note that by (C.8), we have $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq 4\|(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)_S\|_1$. Substituting this result into (C.9) allows us to conclude that

$$\frac{\kappa_\ell}{2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 \leq 6\lambda\|(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)_S\|_1 \leq 6\lambda\sqrt{s_0}\|(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)_S\|_2 \leq 6\lambda\sqrt{s_0}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2.$$

Thus, $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq 12\lambda\sqrt{s_0}/\kappa_\ell$, which offers us the result of this lemma:

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \sqrt{s_0}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \frac{12\lambda s_0}{\kappa_\ell}.$$

We note that this result holds with high probability by Lemma C.2.

□

C.3. Proof of Lemma B.3

Here we verify that the Lindeberg condition (Hall & Heyde, 1980) holds for $\zeta_{i,t}$.

Proof of Lemma B.3. Note that for some random variable $Q \sim N(0,1)$, $C = [(\epsilon_t\widetilde{\boldsymbol{X}}_t^\top\mathbf{m}_i)^2/(\sigma^2[\mathbf{m}_i^\top\widetilde{\boldsymbol{\Sigma}}_n\mathbf{m}_i])]^{-1/2}$, some positive, fixed constant $\delta > 0$, and $t \geq p$:

$$\mathbb{E}[\zeta_t^2\,\mathbb{1}(\|\zeta_t| \geq \delta)|\mathcal{F}_{t-1}] = \frac{(\epsilon_t\widetilde{\boldsymbol{X}}_t^\top\mathbf{m}_i)^2\mathbb{E}[Q^2\,\mathbb{1}(|Q| > \delta C)]}{\sigma^2[\mathbf{m}_i^\top\widetilde{\boldsymbol{\Sigma}}_n\mathbf{m}_i]}. \tag{C.10}$$

By the properties of the truncated standard normal, we can see that $\mathbb{E}[Q^2|Q > c] = 1 + c(\phi(c)/\Phi(c))$, where $\phi(c)$ is the PDF of the standard normal. Thus,

$$\mathbb{E}[Q^2\,\mathbb{1}(Q > c)] = 2\Phi(c)\left(1 + c\frac{\phi(c)}{\Phi(c)}\right) = 2(\Phi(c) + c\phi(c)) \leq 2\phi(c)\left(\frac{c + c^2}{c}\right).$$

Now the proceeding inequality follows from the union bound and a standard bound on the normal CDF:

$$\max_{p+1\leq t\leq T}|\epsilon_t\widetilde{\boldsymbol{X}}_t^\top\mathbf{m}_i| > v,$$

with probability at most $2(T-p)\exp[-v^2/(2\mathbf{m}_i^\top\widetilde{\boldsymbol{\Sigma}}\mathbf{m}_i)]$. If we let $v = 2\sqrt{\log(T-p)\mathbf{m}_i^\top\widetilde{\boldsymbol{\Sigma}}\mathbf{m}_i}$, the we obtain the following bound:

$$\max_{p+1\leq t\leq T}|\epsilon_t\widetilde{\boldsymbol{X}}_t^\top\mathbf{m}_i| \leq 2\sqrt{\log(T-p)\mathbf{m}_i^\top\widetilde{\boldsymbol{\Sigma}}\mathbf{m}_i},$$

with probability at least $1 - 2/(T-p)$. Returning to (C.10), we now see that for $D = \sqrt{(T-p)/(4\log(T-p)}$:

$$\mathbb{E}[\zeta_t^2\,\mathbb{1}(\|\ zeta_t\| \geq \delta)|\mathcal{F}_{t-1}] \leq \frac{8\log(T-P)\phi(\delta D)((\delta D)^{-1} + \delta D)}{T-p},$$

since $\phi(z)(z^{-1} + z)$ is a decreasing function. Therefore, if we sum over all t, we attain

$$\sum_{t=p+1}^{T}\mathbb{E}[\zeta_t^2\,\mathbb{1}(\|\zeta_t| \geq \delta)|\mathcal{F}_{t-1}] \leq 8\log(T-P)\phi(\delta D)((\delta D)^{-1} + \delta D) \to 0,$$

which demonstrates that the Lindeberg condition does indeed hold for ζ_t. □

C.4. Proof of Lemma B.6

We present a modified version of the proof of Lemma 6.1 from Liu et al. (2013b).

Proof of Lemma B.6. Define filtration $\mathcal{F}_t = \sigma(\xi_1, \ldots, \xi_t)$. For $1 \leq \leq p$, let:

$$\begin{aligned}\widehat{\xi_t} &= \xi_t \,\mathbb{1}\{\|\xi_t\|_2 \leq \sqrt{n}/(\log(p))^4\} - \mathbb{E}[\xi_t \,\mathbb{1}\{\|\xi_t\|_2 \leq \sqrt{n}/(\log(p))^4\}|\mathcal{F}_t],\\ \widetilde{\xi_t} &= \xi_t - \widehat{\xi_t}.\end{aligned} \tag{C.11}$$

Note that by Definition G.1, $\widehat{\xi_i}$ forms a martingale difference sequence (MDS). Now we have by the triangle inequality

$$\begin{aligned}\mathbb{P}\left(\left\|\sum_{t=1}^n \xi_t\right\|_{\min} \geq t\sqrt{n}\right) \leq &\mathbb{P}\left(\left\|\sum_{t=1}^n \widehat{\xi_t} \geq t\sqrt{n} - \sqrt{n}/(\log(p))^2\right\|_{\min}\right)\\ &+ \mathbb{P}\left(\left\|\sum_{t=1}^n \widetilde{\xi_t} \geq \sqrt{n}/(\log(p))^2\right\|_{\min}\right),\end{aligned} \tag{C.12}$$

and

$$\begin{aligned}\mathbb{P}\left(\left\|\sum_{t=1}^n \xi_t\right\|_{\min} \geq t\sqrt{n}\right) \geq &\mathbb{P}\left(\left\|\sum_{t=1}^n \widehat{\xi_t} \geq t\sqrt{n} + \sqrt{n}/(\log(p))^2\right\|_{\min}\right)\\ &- \mathbb{P}\left(\left\|\sum_{t=1}^n \widetilde{\xi_t} \geq \sqrt{n}/(\log(p))^2\right\|_{\min}\right).\end{aligned} \tag{C.13}$$

By our assumption that $log(p) = o(\sqrt{n})$, we have that,

$$\sum_{t=1}^n \mathbb{E}[\xi_t \,\mathbb{1}\{\|\xi_t\|_2 \leq \sqrt{n}/(\log(p))^4\}|\mathcal{F}_t] = o(\sqrt{n}/(\log(p))^2).$$

So by our assumption that $\mathbb{E}[\|\xi_t\|_2^{bpr+2+\epsilon}] \leq \infty$, for $r, b, \epsilon > 0$, the previous line implies that:

$$\mathbb{P}\left(\left\|\sum_{t=1}^n \widetilde{\xi_t} \geq \sqrt{n}/(\log(p))^2\right\|_{\min}\right) \leq n\mathbb{P}(\max_{1\leq t\leq n} \|\xi_t\| \geq \sqrt{n}/(\log(p))^4) \leq C(\log(p))^{-3/2}(G(t))^p,$$

for $0 \leq t \leq b\sqrt{\log(p)}$. Thus, by (C.13) and (C.12), we obtain:

$$\mathbb{P}\left(\left\|\sum_{t=1}^n \xi_t\right\|_{\min} \geq t\sqrt{n}\right) \geq \mathbb{P}\left(\left\|\sum_{t=1}^n \widehat{\xi_t} \geq t\sqrt{n} + \sqrt{n}/(\log(p))^2\right\|_{\min}\right) - C(\log(p))^{-3/2}(G(t))^p,$$

and

$$\mathbb{P}\left(\left\|\sum_{t=1}^n \xi_t\right\|_{\min} \geq t\sqrt{n}\right) \leq \mathbb{P}\left(\left\|\sum_{t=1}^n \widehat{\xi_t} \geq t\sqrt{n} - \sqrt{n}/(\log(p))^2\right\|_{\min}\right) + C(\log(p))^{-3/2}(G(t))^p,$$

for constant $C > 0$ Therefore, it suffices to prove

$$\sup_{0\leq t\leq b\sqrt{\log(p)}} \left|\frac{\mathbb{P}(\|\sum_{t=1}^n \widehat{\xi_t}\|_{\min} \geq (t \pm (\log(p)^{-2})\sqrt{n})}{(G(t))^p} - 1\right| \leq C(\log(p))^{-1-\gamma_1} \tag{C.14}$$

in order to prove this lemma. To prove this line, we appeal to Theorem 1.1 from Zaitsev (1987). The original version of Theorem 1.1 requires i.i.d. vectors only in order to leverage the Bernstein inequality. However, since in this application $\widehat{\xi_t}$ form a MDS, the proof of Theorem 1.1 holds by our Bernstein inequality for MDS (Lemma F.8). In the interest of clarity, since Theorem 1.1 requires introducing a significant amount of new material from probability theory, we refer the reader to Zaitsev (1987) for more details. Thus, by application of Theorem 1.1 from Zaitsev (1987) to $\widehat{\xi}_t$, we obtain

$$\mathbb{P}\bigg(\bigg\|\sum_{t=1}^{n}\widehat{\xi}_t\bigg\|_{\min} \geq (t + (\log(p)^{-2})\sqrt{n}\bigg) \leq \mathbb{P}(\|\boldsymbol{W}\|_{\min} \geq t - 2(\log(p))^{-2}) + c_{1,p}\exp[-c_{2,d}(\log(p))^2],$$

and

$$\mathbb{P}\bigg(\bigg\|\sum_{t=1}^{n}\widehat{\xi}_t\bigg\|_{\min} \geq (t - (\log(p)^{-2})\sqrt{n}\bigg) \leq \mathbb{P}(\|\boldsymbol{W}\|_{\min} \geq t + 2(\log(p))^{-2}) - c_{1,p}\exp[-c_{2,d}(\log(p))^2],$$

where $c_{1,p}, c_{2,p} > 0$ are constants that depend only on p and $\boldsymbol{W} \sim N\bigg(0, \operatorname{Cov}\Big(\sum_{t=1}^{n}\widehat{\xi}_i/\sqrt{n}\Big)\bigg)$. By our assumption that $\|\operatorname{Cov}(\xi_i) - \mathbf{I}_{p\times p}\|_2 \leq C(\log(p)^{-2-\gamma}$, one can easily show that,

$$\mathbb{P}(\|\boldsymbol{W}\|_{\min} \geq t - 2(\log(p))^{-2}) \leq (1 + C(\log(p))^{-1-\gamma_1})(G(t))^p,$$

and

$$\mathbb{P}(\|\boldsymbol{W}\|_{\min} \geq t + 2(\log(p))^{-2}) \leq (1 - C(\log(p))^{-1-\gamma_1})(G(t))^p,$$

for $0 \leq t \leq b\sqrt{\log(p)}$. Since for $0 \leq t \leq b\sqrt{\log(p)}$ we have $c_{1,p}\exp[-c_{2,p}(\log(p))^2] \leq C(\log(p))^{-1-\gamma_1}(G(t))^p$, the following holds:

$$\mathbb{P}\bigg(\bigg\|\sum_{t=1}^{n}\widehat{\xi}_t\bigg\|_{\min} \geq (t - (\log(p)^{-2})\sqrt{n}\bigg) \leq (1 + C(\log(p))^{-1-\gamma_1})(G(t))^p, \tag{C.15}$$

and

$$\mathbb{P}\bigg(\bigg\|\sum_{t=1}^{n}\widehat{\xi}_t\bigg\|_{\min} \geq (t + (\log(p)^{-2})\sqrt{n}\bigg) \leq (1 - C(\log(p))^{-1-\gamma_1})(G(t))^p, \tag{C.16}$$

for $0 \leq t \leq b\sqrt{\log(p)}$. Therefore, (C.15) and (C.16) imply (C.14), which concludes the proof. □

D. Proofs of Supporting Lemmas in Appendix C

In this section we present proofs for the supporting Lemmas of Lemma B.2.

D.1. Proof of Lemma C.1

We present a variation of the proof of Theorem 5 from Johnson et al. (2016). This proof will rely on lower bounding $\inf_{\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\in E_r}\dfrac{1}{T-p}\|\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\|_2^2$ by $\inf_{\mathbf{u}\in A} 1/(T-p)\sum_{t=p+1}^{T}\langle\widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2 - \sup_{\mathbf{u}\in A} 2/(T-p)\sum_{t=p+1}^{T}\langle\widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle$. We lower bound $\inf_{\mathbf{u}\in A} 1/(T-p)\sum_{t=p+1}^{T}\langle\widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2$ in Lemma D.1 and upper bound $\sup_{\mathbf{u}\in A} 2/(T-p)\sum_{t=p+1}^{T}\langle\widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle$ in Lemma D.2 below.

Lemma D.1. Let $\boldsymbol{Z}_t = \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t$, for $\boldsymbol{\mu}_t = \mathbb{E}[\widetilde{\boldsymbol{X}}_t|\mathcal{F}_{t-1}]$, where $\mathcal{F}_t$ is the filtration $\mathcal{F}_t = \sigma(\widetilde{\boldsymbol{X}}_{p+1}, \ldots, \widetilde{\boldsymbol{X}}_t)$, and $\mathbf{Z} = [\boldsymbol{Z}_{p+1}, \ldots, \boldsymbol{Z}_T]^\top$. Furthermore, let $\mathbb{E}[\mathbf{Z}^\top\mathbf{Z}/(T-p)] = \boldsymbol{\Sigma}_Z$ and $\|\boldsymbol{\Sigma}_Z^{-1/2}\boldsymbol{Z}_t\|_{\psi_2} \leq \kappa'$. Then,

$$\mathbb{P}\bigg(\inf_{\mathbf{u}\in A}\frac{1}{T-p}\sum_{t=p+1}^{T}\langle\widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2 \geq \lambda_{\min}(\boldsymbol{\Sigma}_Z|A)\bigg(1 - \frac{2c_0c_1\kappa'^2\omega(B)}{\sqrt{T-p}}\bigg) > 0\bigg) \geq 1 - 2\exp(-c_0^2c_1^2c_2\omega^2(B))$$

where $\lambda_{\min}(\boldsymbol{\Sigma}|A) = \inf_{\mathbf{u}\in A}\mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u}$ is the restricted minimum eigenvalue of $\boldsymbol{\Sigma}$ restricted to $A \subseteq S^{pd-1}$ (the unit sphere in $\mathbb{R}^{pd}$ space), and $B = \{\tilde{\mathbf{u}} : \tilde{\mathbf{u}} = \boldsymbol{\Sigma}^{1/2}\mathbf{u}/\|\boldsymbol{\Sigma}^{1/2}\mathbf{u}\|_2, \mathbf{u} \in A\}$ is the normalized set of A.

The proof of Lemma D.1 relies on the restricted eigenvalue condition for martingale difference sequences from Appendix F.

Lemma D.2. Let $\boldsymbol{\mu}_t = \mathbb{E}[\widetilde{\boldsymbol{X}}_t|\mathcal{F}_{t-1}]$, where $\mathcal{F}_t$ is the filtration $\mathcal{F}_t = \sigma(\widetilde{\boldsymbol{X}}_{p+1}, \dots, \widetilde{\boldsymbol{X}}_t)$, $A \subseteq S^{pd-1}$ (the unit sphere in $\mathbb{R}^{pd}$ space), and $\alpha = \text{diam}(A) = \sup_{\mathbf{u},\mathbf{v}\in A} d(\mathbf{u},\mathbf{v})$. Then,

$$\mathbb{P}\bigg(\sup_{\mathbf{u}\in A} \frac{2}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle \le 0\bigg) \ge 1 - L\exp\bigg[-4\frac{(\omega(A))^2}{\alpha^2}\bigg].$$

The proof of Lemma D.2 employs a generic chaining argument (Talagrand, 2006). We defer the proofs of Lemmas D.1 and D.2 to Appendicies E.1 and E.2. We now present the proof of Lemma C.1.

Proof of Lemma C.1. We seek to prove that,

$$\inf_{\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\in E_r} \frac{1}{T-p}\|\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\|_2^2 > 0. \tag{D.1}$$

From Negahban et al. (2012), we note that the error set E_r is actually a cone, and that the magnitude of the error vector $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ in (D.1) does not matter, only the direction does. Thus, we consider set $A = S^{pd-1} \cap E_r$ and reformulate our problem as

$$\inf_{\mathbf{u}\in A} \frac{1}{T-p}\|\widetilde{\mathbf{X}}\mathbf{u}\|_2^2 > 0. \tag{D.2}$$

It suffices to prove (D.2) to prove the result of this lemma.

We now construct a martingale difference sequence that we will bound in order to prove (D.2). Let $\boldsymbol{\mu}_t = \mathbb{E}[\widetilde{\boldsymbol{X}}_t|\mathcal{F}_{t-1}]$, where $\mathcal{F}_t$ is the filtration $\mathcal{F}_t = \sigma(\widetilde{\boldsymbol{X}}_{p+1}, \dots, \widetilde{\boldsymbol{X}}_t)$, so that by Definition G.1 in Appendix G, $\widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t$ forms a martingale difference sequence (MDS). Then we have,

$$\begin{aligned}
\frac{1}{T-p}\|\widetilde{\mathbf{X}}\mathbf{u}\|_2^2 &= \frac{1}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t, \mathbf{u}\rangle \\
&= \frac{1}{T-p}\sum_{t=p+1}^{T} \Big((\widetilde{\boldsymbol{X}}_t^\top \mathbf{u})^2 - 2(\widetilde{\boldsymbol{X}}_t^\top \mathbf{u})(\boldsymbol{\mu}_t^\top \mathbf{u}) + 2(\widetilde{\boldsymbol{X}}_t^\top \mathbf{u})(\boldsymbol{\mu}_t^\top \mathbf{u}) - (\boldsymbol{\mu}_t^\top \mathbf{u})^2 + (\boldsymbol{\mu}_t^\top \mathbf{u})^2 \Big) \\
&= \frac{1}{T-p}\sum_{t=p+1}^{T} \Big((\widetilde{\boldsymbol{X}}_t^\top \mathbf{u} - \boldsymbol{\mu}_t^\top \mathbf{u})^2 - (\boldsymbol{\mu}_t^\top \mathbf{u})^2 + 2(\widetilde{\boldsymbol{X}}_t^\top \mathbf{u})(\boldsymbol{\mu}_t^\top \mathbf{u}) \Big).
\end{aligned}$$

Distributing the summation in the last line then yields:

$$\begin{aligned}
\frac{1}{T-p}\|\widetilde{\mathbf{X}}\mathbf{u}\|_2^2 &= \frac{1}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2 - \frac{1}{T-p}\sum_{t=p+1}^{T} \langle \boldsymbol{\mu}_t, \mathbf{u}\rangle^2 + \frac{2}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t, \mathbf{u}\rangle\langle \boldsymbol{\mu}_t, \mathbf{u}\rangle \\
&= \frac{1}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2 + \frac{1}{T-p}\sum_{t=p+1}^{T} \langle \boldsymbol{\mu}_t, \mathbf{u}\rangle^2 + \frac{2}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle\langle \boldsymbol{\mu}_t, \mathbf{u}\rangle \\
&\ge \frac{1}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2 + \frac{2}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle\langle \boldsymbol{\mu}_t, \mathbf{u}\rangle.
\end{aligned}$$

Hence,

$$\begin{aligned}
\inf_{\mathbf{u}\in A} \frac{1}{T-p}\|\widetilde{\mathbf{X}}\mathbf{u}\|_2^2 &\ge \inf_{\mathbf{u}\in A} \frac{1}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2 + \inf_{\mathbf{u}\in A} \frac{2}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle\langle \boldsymbol{\mu}_t, \mathbf{u}\rangle && \text{(D.3)} \\
&\ge \inf_{\mathbf{u}\in A} \frac{1}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2 - \sup_{\mathbf{u}\in A} \frac{2}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle, && \text{(D.4)}
\end{aligned}$$

where the second inequality holds since we can scale the design matrix to fall in the L_2 unit ball so that $\langle \boldsymbol{\mu}_t, \mathbf{u}\rangle <= 1$. Thus, to prove (D.2), we must lower bound $\inf_{\mathbf{u}\in A} 1/(T-p)\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2$ and upper bound $\sup_{\mathbf{u}\in A} 2/(T-p)\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle$. We bound $\inf_{\mathbf{u}\in A} 1/(T-p)\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2$ with Lemma D.1, and $\sup_{\mathbf{u}\in A} 2/(T-p)\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle$ with Lemma D.2. Thus, by application of Lemmas D.1 and D.2 to (D.3), we have that

$$\inf_{\mathbf{u}\in A}\frac{1}{T-p}\|\widetilde{\mathbf{X}}\mathbf{u}\|_2^2 \geq \inf_{\mathbf{u}\in A}\frac{1}{T-p}\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2 - \sup_{\mathbf{u}\in A}\frac{2}{T-p}\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle$$
$$\geq \lambda_{\min}(\boldsymbol{\Sigma}_Z|A)\left(1 - \frac{2c_0c_1\kappa'^2\omega(B)}{\sqrt{T-p}}\right) > 0,$$

with probability at least

$$1 - 2\exp(-c_0^2c_1^2c_2\omega^2(B)) - L\exp\left[-4\frac{(\omega(A))^2}{\alpha^2}\right].$$

Thus, $\widetilde{\mathbf{X}}$ satisfies that restricted eigenvalue condition with high probability. □

D.2. Proof of Lemma C.2

Here we prove Lemma C.2.

Proof of Lemma C.2. To prove this lemma, we first note that $2\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty = 4\left\|(T-p)^{-1}\widetilde{\mathbf{X}}^\top\boldsymbol{\epsilon}\right\|_\infty$. Let $\boldsymbol{\epsilon}$ be the error vector from (3.2), so $\boldsymbol{\epsilon} = \boldsymbol{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}^*$. Then Assumption 5.4 implies that $\boldsymbol{\epsilon}$ (3.2) is sub-Gaussian. Note that since Assumption 5.4 ensures that $\|\widetilde{\boldsymbol{X}}_i\|_{\psi_2} \leq \kappa$, for $i \in \{1, 2, \ldots, T-p\}$, we can scale the columns of any $\widetilde{\boldsymbol{X}}$ so that $\|\widetilde{\mathbf{X}}_{\cdot,j}\|_2/\sqrt{T-p} \leq 1$, for $1 \leq k \leq pd$. Then the sub-Gaussian tails of $\boldsymbol{\epsilon}$ guarantee that for all $t > 0$

$$\frac{1}{T-p}\|\langle\widetilde{\mathbf{X}}, \boldsymbol{\epsilon}\rangle\|_2 < t,$$

with probability at least $1 - 2\exp[-(T-p)t^2/(2\sigma^2)]$. Bounding over all pd columns yields:

$$\left\|\frac{1}{T-p}\widetilde{\mathbf{X}}^\top\boldsymbol{\epsilon}\right\|_\infty < t,$$

holds with probability at least $1 - 2\exp[-(T-p)t^2/(2\sigma^2) + \log(pd)]$. Setting $t = 2\sigma\sqrt{\log(pd)/(T-p)}$ allows us to conclude that

$$\lambda = 8\sigma\sqrt{\frac{\log(pd)}{T-p}} \geq 2\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty = 4\left\|\frac{1}{T-p}\widetilde{\mathbf{X}}^\top\boldsymbol{\epsilon}\right\|_\infty,$$

holds with probability at least $1 - b_1\exp[-b_2\sigma^2\log(pd)]$. □

E. Proofs of Supporting Lemmas in Appendix D

In this section, we present proofs of the supporting lemmas for Lemma C.1.

E.1. Proof of Lemma D.1

We seek to bound $\inf_{\mathbf{u}\in A} 1/(T-p)\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2$. Since $\{\widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t\}$ is a martingale difference sequence (MDS) by Definition G.1 in Appendix G, we will appeal to the restricted eigenvalue condition for MDS that we present in Theorem F.6 in Appendix F. We reproduce Theorem F.6 here for the convenience of the reader:

Theorem E.1. Let $\mathbf{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^\top$ be a $n \times d$ design matrix whose anisotropic sub-Gaussian rows form a vector valued martingale difference sequence. Let $\mathbb{E}[\mathbf{X}^\top \mathbf{X}/n] = \mathbf{\Sigma}$ and $\|\mathbf{\Sigma}^{-1/2}\boldsymbol{X}_i\|_{\psi_2} \leq \kappa$. Then for absolute constants $c_0, c_1, c_2 > 0$, with probability at least $1 - 2\exp(-c_0^2 c_1^2 c_2 \omega^2(B))$, we have

$$\lambda_{\min}(\mathbf{\Sigma}|A)\left(1 - \frac{2c_0 c_1 \kappa^2 \omega(B)}{\sqrt{n}}\right) \leq \inf_{\mathbf{u}\in A} \frac{1}{n}\|\mathbf{X}\mathbf{u}\|_2^2,$$

where $\lambda_{\min}(\mathbf{\Sigma}|A) = \inf_{\mathbf{u}\in A} \mathbf{u}^\top \mathbf{\Sigma}\mathbf{u}$ is the restricted minimum eigenvalues of $\mathbf{\Sigma}$ restricted to $A \subseteq S^{d-1}$ (the unit sphere in $\mathbb{R}^d$ space), and $B = \{\widetilde{\mathbf{u}} : \widetilde{\mathbf{u}} = \mathbf{\Sigma}^{1/2}\mathbf{u}/\|\mathbf{\Sigma}^{1/2}\mathbf{u}\|_2, \mathbf{u} \in A\}$ is the normalized set of A.

Proof of Lemma D.1. Let $\boldsymbol{Z}_t = \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t$ for $\boldsymbol{\mu}_t = \mathbb{E}[\widetilde{\boldsymbol{X}}_t|\mathcal{F}_{t-1}]$, where $\mathcal{F}_t$ is the filtration $\mathcal{F}_t = \sigma(\widetilde{\boldsymbol{X}}_{p+1}, \ldots, \widetilde{\boldsymbol{X}}_t)$, as defined in Appendix D.1. Then let $\mathbf{Z} = [\boldsymbol{Z}_{p+1}, \ldots, \boldsymbol{Z}_T]^\top$. Clearly the rows of $\mathbf{Z}$ form a vector-values MDS, and by Assumption 5.4, they are sub-Gaussian as well. The definition of $\widetilde{\mathbf{X}}$ in Section 3 implies that the rows or $\mathbf{Z}$ are anisotropic. Let $\mathbb{E}[\mathbf{Z}^\top\mathbf{Z}/(T-p)] = \mathbf{\Sigma}_Z$ be the true covariance matrix of $\mathbf{Z}$, and $\|\mathbf{\Sigma}_Z^{-1/2}\boldsymbol{Z}_t\|_{\psi_2} \leq \kappa'$. Then by Theorem F.6 we have that,

$$\inf_{\mathbf{u}\in A} \frac{1}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2 = \inf_{\mathbf{u}\in A}\frac{1}{n}\|\mathbf{Z}\mathbf{u}\|_2^2 \geq \lambda_{\min}(\mathbf{\Sigma}_Z|A)\left(1 - \frac{2c_0c_1\kappa'^2\omega(B)}{\sqrt{T-p}}\right),$$

with high probability. Thus, to prove Lemma D.1, it suffices to demonstrate that $\lambda_{\min}(\mathbf{\Sigma}_Z|A)$, the restricted minimum eigenvalue of $\mathbf{\Sigma}_Z$, is positive. In this endeavor, we will draw upon the argument made by Johnson et al. (2016) in a similar context.

Let $B_2^{pd}(\boldsymbol{x}, \epsilon)$ be a L_2 ball centered at $\boldsymbol{x}$ with radius ϵ. Clearly, we can scale the orignial design matrix $\widetilde{\mathbf{X}}$ so that its rows fall in a L_2 unit ball, in which case the rows of $\mathbf{Z}$ fall in a L_2 unit ball centered at the origin. So then the set, $\mathcal{A} = \{\widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t\}$, where $\widetilde{\boldsymbol{X}}_t$ is drawn from the aforementioned L_2 ball, is a subset of the L_2 ball centered at the origin. Now note that by definition $\lambda_{\min}(\mathbf{\Sigma}_Z|A) \geq \lambda_{\min}(\mathbf{\Sigma}_Z)$, where $\lambda_{\min}(\mathbf{\Sigma}_Z)$ is the unrestricted minimum eigenvalue of $\mathbf{\Sigma}_Z$. So it suffices to show that $\lambda_{\min}(\mathbf{\Sigma}_Z) > 0$. By way of contradiction, assume that $\lambda_{\min}(\mathbf{\Sigma}_Z) = 0$. Then let the eigendecomposition of $\mathbf{\Sigma}_Z$ be $\mathbf{\Sigma}_Z = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$, where $\mathbf{Q} = [\mathbf{v}_1, \ldots, \mathbf{v}_{pd}]$ has the eigenvectors of $\mathbf{\Sigma}_Z$ for columns, and $\Lambda = \text{diag}(\lambda_i)_{i=1}^{pd}$ is a diagonal matrix of the eigenvalues of $\mathbf{\Sigma}_Z$ in descending order. Observe that $\lambda_{pd} = 0$ implies

$$\mathbb{E}_{\mathbf{a}\sim\mathcal{A}}[\langle \mathbf{a}, \mathbf{v}_{pd}\rangle] = 0, \tag{E.1}$$

since $\mathbf{v}_{pd}$ is the eigenvector corresponding to the minimum eigenvalue of 0. Let $\mathcal{A}_{\mathbf{v}_{pd}} = \{\mathbf{a} \in \mathcal{A} | \langle \mathbf{a}, \mathbf{v}_{pd}\rangle = 0\}$. Then clearly, (E.1) implies,

$$\mathbb{P}(\mathbf{a} \in \mathcal{A}_{\mathbf{v}_{pd}}) = 1. \tag{E.2}$$

Note that by (E.2), since there is no probability density outside $\mathcal{A}_{\mathbf{v}_{pd}}$, this density is thus concentrated on a subspace of $\mathbb{R}^{pd}$. Here we have a contradiction, since the span of $\mathcal{A}_{\mathbf{v}_{pd}}$ is $\mathbb{R}^{pd}$. Therefore, $\lambda_{\min}(\mathbf{\Sigma}_Z|A) \geq \lambda_{\min}(\mathbf{\Sigma}_Z) > 0$, and we have

$$\inf_{\mathbf{u}\in A} \frac{1}{T-p}\sum_{t=p+1}^{T} \langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle^2 = \inf_{\mathbf{u}\in A}\frac{1}{n}\|\mathbf{Z}\mathbf{u}\|_2^2 \geq \lambda_{\min}(\mathbf{\Sigma}_Z|A)\left(1 - \frac{2c_0c_1\kappa'^2\omega(B)}{\sqrt{T-p}}\right) > 0,$$

with probability at least $1 - 2\exp(-c_0^2c_1^2c_2\omega^2(B))$, as desired. □

E.2. Proof of Lemma D.2

In this section we present a proof for Lemma D.2. This proof will make a generic chaining argument, and will thus rely on the following standard lemmas from Talagrand (2006) and Talagrand (2014).

Lemma E.2 (Theorem 2.1.5 from Talagrand (2006))**.** Consider two processes $\{X_t\}_{t\in A}$ and $\{Y_t\}_{t\in A}$, indexed by the same set A. Assume $\{X_t\}_{t\in A}$ is Gaussian, and $\{Y_t\}_{t\in A}$ satisfies the condition:

$$\forall\delta > 0, \forall \mathbf{u}, \mathbf{v} \in A, \mathbb{P}(|Y_u - Y_v| > \delta) \leq 2\exp\left(\frac{-\,\delta^2}{d(\mathbf{u}, \mathbf{v})^2}\right),$$

where $d(\mathbf{u},\mathbf{v})$ is the distance metric associated with X_t (we assume $d(\mathbf{u},\mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|_2$). Then, for some constant L,

$$\mathbb{E}[\sup_{\mathbf{u},\mathbf{v}\in A} |Y_u - Y_v|] \leq L\mathbb{E}[\sup_{\mathbf{v}\in A} X_v]]$$

Lemma E.3 (Lemma 1.2.8 from Talagrand (2006))**.** If the process $\{X_t\}_{t\in A}$ is symmetric, then

$$\mathbb{E}[\sup_{\mathbf{u},\mathbf{v}\in A} |X_u - X_v|] = 2\mathbb{E}[\sup_{\mathbf{u}\in A} X_u]].$$

Lemma E.4 (Theorem 2.2.27 from Talagrand (2014))**.** Let $\{X_t\}_{t\in A}$ be a process that satisfies Lemma E.2. Then for any $\delta > 0$ and constant L,

$$\mathbb{P}\left(\sup_{\mathbf{u},\mathbf{v}\in A} |X_u - X_v| \geq L(\gamma_2(A, d(\mathbf{u},\mathbf{v})) + \delta\alpha)\right) \leq L\exp(-\delta^2),$$

where $\alpha = \text{diam}(A) = \sup_{\mathbf{u},\mathbf{v}\in A} d(\mathbf{u},\mathbf{v})$, and $\gamma_2(\cdot,\cdot)$ is the γ_2-functional defined in F.4.

Lemma E.5 (Theorem 2.2.27 from Talagrand (2014))**.** For constant L,

$$\frac{1}{L}\gamma_2(A, d(\mathbf{u},\mathbf{v})) \leq \mathbb{E}\sup_{\mathbf{u}\in A} X_u \leq L\gamma_2(A, d(\mathbf{u},\mathbf{v})).$$

Having presented these lemmas, we now give the proof of Lemma D.2.

Proof of Lemma D.2. We seek to bound $\sup_{\mathbf{u}\in A} \frac{2}{T-p}\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle$. Recall that by Assumption 5.4, the sub-Gaussian norm of each row $\widetilde{\boldsymbol{X}}_t$ is bounded by κ, which implies that $\|\widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t\|_{\psi_2} \leq \kappa$. Thus, $\widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t$ forms a sub-Gaussian bounded MDS, and so we can apply the Azuma-Hoeffding inequality to obtain

$$\mathbb{P}\left(\frac{1}{\sqrt{T-p}}\left|\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle\right| \geq \epsilon\right) \leq 2\exp\left[\frac{-\epsilon^2}{2\|\mathbf{u}\|_2^2\kappa^2}\right], \tag{E.3}$$

for any $\mathbf{u} \in A$. Then for $\mathbf{u},\mathbf{v} \in A$, we have that

$$\mathbb{P}\left(\frac{1}{\sqrt{T-p}}\left|\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}-\mathbf{v}\rangle\right| \geq \epsilon\right) \leq 2\exp\left[\frac{-\epsilon^2}{2\|\mathbf{u}-\mathbf{v}\|_2^2\kappa^2}\right]. \tag{E.4}$$

We now make a generic chaining argument to bound $\sup_{\mathbf{u}\in A} 1/\sqrt{T-p}\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle$ with high probability.

Let $\mathbf{Z} = [(\widetilde{\boldsymbol{X}}_{p+1} - \boldsymbol{\mu}_{p+1}), \ldots, (\widetilde{\boldsymbol{X}}_T - \boldsymbol{\mu}_T)]^\top$. We first note that by (E.3), the process $Z_u = \langle \mathbf{Z}, \mathbf{u}\rangle = 1/\sqrt{T-p}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle$ has sub-Gaussian concentration. Similarly, by (E.4), $Z_u - Z_v$ is a sub-Gaussian process $\forall \mathbf{u},\mathbf{v} \in A$. We now bound $\mathbb{E}[\sup_{\mathbf{u}\in A} 1/\sqrt{T-p}\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle]$ in terms of the Gaussian width of set A (see Definition F.5), and then prove that $\sup_{\mathbf{u}\in A} 1/\sqrt{T-p}\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle$ concentrates around its expectation with high probability.

To bound $\mathbb{E}[\sup_{\mathbf{u}\in A} 1/\sqrt{T-p}\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle]$, we appeal to Lemma E.2. In our setting, $\mathbb{E}[\sup_{\mathbf{v}\in A} X_v] = \omega(A)$, the Gaussian width of A. Thus, by Lemma E.2 and (E.4), we achieve the following bound for some constant L:

$$\mathbb{E}[\sup_{\mathbf{u},\mathbf{v}\in A} |Z_u - Z_v|] \leq L\kappa\omega(A). \tag{E.5}$$

Note that by (E.3), the process Z_u is symmetric, and so Lemma E.3 applies. Thus, by Lemma E.3 and (E.5), we have that,

$$\mathbb{E}[\sup_{\mathbf{u},\mathbf{v}\in A} |Z_u - Z_v|] = 2\mathbb{E}[\sup_{\mathbf{u}\in A} Z_u]] \leq L\kappa\omega(A). \tag{E.6}$$

Thus, the definition of Z_u, we can bound $\mathbb{E}[\sup_{\mathbf{u}\in A} 2/\sqrt{T-p}\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle]$ as follows:

$$2\mathbb{E}\left[\sup_{\mathbf{u}\in A}\frac{2}{T-p}\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle\right] = 2\mathbb{E}[\sup_{\mathbf{u}\in A} Z_u]] \leq L\kappa\omega(A).$$

Having bound the expectation of $\mathbb{E}[\sup_{\mathbf{u}\in A} 1/\sqrt{T-p}\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle]$ in terms of the Gaussian width of A, we now seek to bound $\sup_{\mathbf{u}\in A} 1/\sqrt{T-p}\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle$ around its expectation with high-probability. To do so, we appeal to lemma E.4. In this setting, $d(\mathbf{u},\mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|_2$. Lemma E.4 motivates the following result:

$$\mathbb{P}(\sup_{\mathbf{u},\mathbf{v}\in A}|Z_u - Z_v| \geq L(\gamma_2(A, d(\mathbf{u},\mathbf{v})) + \delta\alpha) = \mathbb{P}(\sup_{\mathbf{u},\mathbf{v}\in A}|Z_u - Z_v| \geq L(\gamma_2(A, d(\mathbf{u},\mathbf{v})) + \epsilon), \tag{E.7}$$

where the right-hand side of the inequality follows since, as Taylor et al. (2014) notes, $\gamma_2(A, d(\mathbf{u},\mathbf{v})) \geq \alpha$ from the definition of the γ_2 functional. We now bound $\gamma_2(A, d(\mathbf{u},\mathbf{v}))$ with Lemma E.5. So by Lemma E.5 and (E.5),

$$\begin{aligned}\mathbb{P}(\sup_{\mathbf{u},\mathbf{v}\in A}|Z_u - Z_v| \geq L(\gamma_2(A, d(\mathbf{u},\mathbf{v})) + \epsilon) &\leq \mathbb{P}(\sup_{\mathbf{u},\mathbf{v}\in A}|Z_u - Z_v| \geq \mathbb{E}[\sup_{\mathbf{u},\mathbf{v}\in A}|Z_u - Z_v|] + \epsilon)\\ &= \mathbb{P}(\sup_{\mathbf{u}\in A}|Z_u| \geq 2\mathbb{E}[\sup_{\mathbf{u}\in A}|Z_u|] + \epsilon),\end{aligned}$$

where the equality follows from Lemma E.3. Now substituting in the definition of Z_u, and applying (E.6) and Lemma E.4, we have:

$$\mathbb{P}\left(\sup_{\mathbf{u}\in A}\frac{1}{\sqrt{T-p}}\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle \geq 2L\kappa\omega(A) + \epsilon\right) \leq L\exp\left[-\left(\frac{\epsilon}{L\kappa\alpha}\right)^2\right].$$

Dividing through by $\sqrt{T-p}$, multiplying by 2, and letting $\epsilon = -2L\kappa\alpha\omega(A)$, we achieve the following bound on the desired quantity

$$\mathbb{P}\left(\sup_{\mathbf{u}\in A}\frac{2}{T-p}\sum_{t=p+1}^{T}\langle \widetilde{\boldsymbol{X}}_t - \boldsymbol{\mu}_t, \mathbf{u}\rangle \geq 0\right) \leq L\exp\left[-4\frac{(\omega(A))^2}{\alpha^2}\right]. \tag{E.8}$$

□

F. Restricted Eigenvalue Condition for Martingale Difference Sequences

In this section, we prove that under mild conditions, the restricted eigenvalue condition will hold for martingale difference sequences (MDS), which we define in Definition G.1 in Appendix G. We first present the following definitions:

Definition F.1. An *isotropic* design matrix is one for which the covariance matrix of each row $\boldsymbol{\Sigma} = \mathbb{E}[X_i X_i^T] = \mathbf{I}$.

Definition F.2. An *anisotropic* design matrix has rows with a general covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[X_i X_i^T]$, but with corresponding isotropic rows $\overline{\boldsymbol{X}}_i = X_i\boldsymbol{\Sigma}^{-1/2}$.

Definition F.3. For a finite set $A \subset T$, denote the cardinality of A by $|A|$. An *admissible sequence* of T is a collection of subsets of T, $\{T_s : s \geq 0\}$, such that for every $s \geq 1$, $|T_s| = 2^{2^s}$ and $|T_0| = 1$.

Definition F.4. (Talagrand, 2006) For a metric space (T, d) and $k = 1, 2$, define

$$\gamma_k(T, d) = \inf\sup_{t\in T}\sum_{s=0}^{\infty} 2^{s/k} d(t, T_s),$$

where $d(t, T_s)$ is the distance between the set T_s and t, and the infimum is taken with respect to all admissible sequences of T. In cases where the metric is clear from the context, we will denote the γ_k functional by $\gamma_k(T)$.

Definition F.5. (Gordon, 1988; Chandrasekaran et al., 2012) The *Gaussian width* of a set $A \in \mathbb{R}^p$ is

$$\omega(A) = \sup_{\boldsymbol{u}\in A}\mathbb{E}[\langle \boldsymbol{g}, \boldsymbol{u}\rangle],$$

where we take the expectation over random vector $\boldsymbol{g} \sim N(0, I_{p\times p})$. The Gaussian width is a measure of the size of set A.

We now present the restricted eigenvalue condition for MDS.

Theorem F.6. Let $\mathbf{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^\top$ be a $n \times d$ design matrix whose anisotropic sub-Gaussian rows form a vector valued martingale difference sequence. Let $\mathbb{E}[\mathbf{X}^\top \mathbf{X}/n] = \boldsymbol{\Sigma}$ and $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\|_{\psi_2} \leq \kappa$. Then for absolute constants $c_0, c_1, c_2 > 0$, with probability at least $1 - 2\exp(-c_0^2 c_1^2 c_2 \omega^2(B))$, we have

$$\lambda_{\min}(\boldsymbol{\Sigma}|A)\left(1 - \frac{2c_0 c_1 \kappa^2 \omega(B)}{\sqrt{n}}\right) \leq \inf_{\mathbf{u}\in A} \frac{1}{n}\|\mathbf{X}\mathbf{u}\|_2^2,$$

where $\lambda_{\min}(\boldsymbol{\Sigma}|A) = \inf_{\mathbf{u}\in A} \mathbf{u}^\top \boldsymbol{\Sigma}\mathbf{u}$ is the restricted minimum eigenvalues of $\boldsymbol{\Sigma}$ restricted to $A \subseteq S^{d-1}$ (the unit sphere in $\mathbb{R}^d$ space), and $B = \{\widetilde{\mathbf{u}} : \widetilde{\mathbf{u}} = \boldsymbol{\Sigma}^{1/2}\mathbf{u}/\|\boldsymbol{\Sigma}^{1/2}\mathbf{u}\|_2, \mathbf{u} \in A\}$ is the normalized set of A.

In the proof of this theorem, we will use the following lemma, which is a MDS version of the sub-Gaussian concentration in Mendelson et al. (2007).

Lemma F.7. (Mendelson et al., 2007) Let (Ω, μ) be a probability space, and $F \subset S_{L_2}$ be a set of functions, where $S_{L_2} := \{f : \|f\|_{L_2} = 1\}$ is the unit sphere in $L_2(\mu)$ space. Assume that $\mathrm{diam}(F, \|\cdot\|_{\psi_2}) = \alpha$. Then, for any $\theta > 0$ and $n \geq 1$ satisfying

$$c_1 \alpha \gamma_2(F, \|\cdot\|_{\psi_2}) \leq \theta\sqrt{n},$$

we have with probability at least $1 - \exp(-c_2 n\theta^2/\alpha^4)$ that

$$\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{k} f^2(\boldsymbol{X}_i) - \mathbb{E}[f^2]\right| \leq \theta,$$

where c_1, c_2 are absolute constants.

The detailed proof of Lemma F.7 can be found in Mendelson et al. (2007). We outline the proof of this lemma in Appendix F.1.

Proof of Theorem F.6. We will apply Lemma F.7 to this proof. First, we define the following class of functions:

$$F = \left\{f_u : \mathbf{u} \in A, f_u(\cdot) = \frac{1}{\sqrt{\mathbf{u}^\top \boldsymbol{\Sigma}\mathbf{u}}}\langle \cdot, \mathbf{u}\rangle\right\}. \tag{F.1}$$

We need to verify that $F \subset S_{L_2}$. In fact, for any $f_u \in F$, we have

$$\|f_u(\mathbf{X})\|_{L_2} = \frac{1}{\mathbf{u}^\top \boldsymbol{\Sigma}\mathbf{u}}\mathbb{E}_X[\langle \mathbf{X}, \mathbf{u}\rangle^2] = \frac{1}{\mathbf{u}^\top \boldsymbol{\Sigma}\mathbf{u}}\mathbb{E}_X[\mathbf{u}^\top \mathbf{X}^\top \mathbf{X}\mathbf{u}] = 1.$$

Note that

$$\mathrm{diam}(F, \|\cdot\|_{\psi_2}) = \sup_{f_u, f_v \in F}\|f_u - f_v\|_{\psi_2} \leq 2\sup_{f_u\in F}\|f_u\|_{\psi_2}.$$

In order to bound the diameter of F according to $\|\cdot\|_{\psi_2}$, we only need to get a bound on the following term

$$\sup_{f_u\in F}\|f_u\|_{\psi_2} = \sup_{\mathbf{u}\in A}\left\|\frac{1}{\sqrt{\mathbf{u}^\top \boldsymbol{\Sigma}\mathbf{u}}}\langle \mathbf{X}, \mathbf{u}\rangle\right\|_{\psi_2} = \sup_{\mathbf{u}\in A}\left\|\left\langle \boldsymbol{\Sigma}^{-1/2}\mathbf{X}, \frac{\boldsymbol{\Sigma}^{1/2}\mathbf{u}}{\|\boldsymbol{\Sigma}^{1/2}\mathbf{u}\|_2}\right\rangle\right\|_{\psi_2}.$$

Thus, we have $\sup_{f_u\in F}\|f_u\|_{\psi_2} \leq \|\boldsymbol{\Sigma}^{-1/2}\mathbf{X}\ |_{\psi_2} \leq \kappa$. By similar argument, we have

$$\|f_u - f_v\|_{\psi_2} = \left\|\left\langle \boldsymbol{\Sigma}^{-1/2}\mathbf{X}, \frac{\boldsymbol{\Sigma}^{1/2}\mathbf{u}}{\|\boldsymbol{\Sigma}^{1/2}\mathbf{u}\|_2} - \frac{\boldsymbol{\Sigma}^{1/2}\mathbf{v}}{\|\boldsymbol{\Sigma}^{1/2}\mathbf{v}\|_2}\right\rangle\right\|_{\psi_2} \leq \kappa\left\|\frac{\boldsymbol{\Sigma}^{1/2}\mathbf{u}}{\|\boldsymbol{\Sigma}^{1/2}\mathbf{u}\|_2} - \frac{\boldsymbol{\Sigma}^{1/2}\mathbf{v}}{\|\boldsymbol{\Sigma}^{1/2}\mathbf{v}\|_2}\right\|_2.$$

By definition, we also have

$$\|f_u - f_v\|_{L_2} = \mathbb{E}\left[\left\langle \boldsymbol{\Sigma}^{-1/2}\mathbf{X}, \frac{\boldsymbol{\Sigma}^{1/2}\mathbf{u}}{\|\boldsymbol{\Sigma}^{1/2}\mathbf{u}\|_2} - \frac{\boldsymbol{\Sigma}^{1/2}\mathbf{v}}{\|\boldsymbol{\Sigma}^{1/2}\mathbf{v}\|_2}\right\rangle^2\right] = \left\|\frac{\boldsymbol{\Sigma}^{1/2}\mathbf{u}}{\|\boldsymbol{\Sigma}^{1/2}\mathbf{u}\|_2} - \frac{\boldsymbol{\Sigma}^{1/2}\mathbf{v}}{\|\boldsymbol{\Sigma}^{1/2}\mathbf{v}\|_2}\right\|_2.$$

This equality immediately implies $\|f_u - f_v\|_{\psi_2} \le \kappa \|f_u - f_v\|_{L_2}$. Then the γ_2-functional in Lemma F.7 can be bounded as

$$\gamma_2(F \cap S_{L_2}, \|\cdot\|_{\psi_2}) \le \kappa \gamma_2(F \cap S_{L_2}, \|\cdot\|_{L_2}) \le \kappa c_0 \omega(B),$$

where the last inequality is due to the majorizing measure theorem in Talagrand (2006), $B := \{\widetilde{\mathbf{u}} : \widetilde{\mathbf{u}} = \mathbf{\Sigma}^{1/2}\mathbf{u}/\|\mathbf{\Sigma}^{1/2}\mathbf{u}\|_2, \mathbf{u} \in A\}$ is the normalized set of A and $c_0 > 0$ is an absolute constant. Now we choose the parameter θ in Theorem F.7 as

$$\theta = \frac{2c_0 c_1 \kappa^2 \omega(B)}{\sqrt{n}} \ge \frac{2c_1 \kappa \gamma_2(F, \|\cdot\|_{\psi_2})}{\sqrt{n}}.$$

Therefore, we have with probability at least $1 - \exp(-c_0^2 c_1^2 c_2 \omega(B)^2/4)$ that

$$\sup_{\mathbf{u}\in A} \left| \frac{1}{n} \frac{1}{\mathbf{u}^\top \mathbf{\Sigma} \mathbf{u}} \sum_{i=1}^n \langle \boldsymbol{X}_i, \mathbf{u} \rangle^2 - 1 \right| \le \frac{2c_0 c_1 \kappa^2 \omega(B)}{\sqrt{n}},$$

where c_0, c_1, c_2 are absolute constants and $B := \{\widetilde{\mathbf{u}} : \widetilde{\mathbf{u}} = \mathbf{\Sigma}^{1/2}\mathbf{u}/\|\mathbf{\Sigma}^{1/2}\mathbf{u}\|_2, \mathbf{u} \in A\}$. It follows that

$$\sup_{\mathbf{u}\in A} \left(1 - \frac{1}{n} \frac{1}{\mathbf{u}^\top \mathbf{\Sigma} \mathbf{u}} \|\mathbf{X}\mathbf{u}\|_2^2 \right) \le \frac{2c_0 c_1 \kappa^2 \omega(B)}{\sqrt{n}},$$

and

$$1 - \frac{2c_0 c_1 \kappa^2 \omega(B)}{\sqrt{n}} \le \inf_{\mathbf{u}\in A} \frac{1}{n} \frac{1}{\mathbf{u}^\top \mathbf{\Sigma} \mathbf{u}} \|\mathbf{X}\mathbf{u}\|_2^2 \le \frac{1}{\lambda_{\min}(\mathbf{\Sigma}|A)} \inf_{\mathbf{u}\in A} \frac{1}{n} \|\mathbf{X}\mathbf{u}\|_2^2.$$

Thus we obtain that

$$\lambda_{\min}(\mathbf{\Sigma}|A) \left(1 - \frac{2c_0 c_1 \kappa^2 \omega(B)}{\sqrt{n}} \right) \le \inf_{\mathbf{u}\in A} \frac{1}{n} \|\mathbf{X}\mathbf{u}\|_2^2$$

holds with probability at least $1 - \exp(-c_0^2 c_1^2 c_2 \omega(B)^2/4)$, with c_0, c_1, c_2 being absolute constants. □

F.1. Sketch of Proof of Lemma F.7

Here we lay the outline of the proof for Lemma F.7, and show that it can be extended to bounded martingale difference sequence. The only difference in the proof for our MDS version of this lemma from the independent case is the Bernstein inequality. Whereas the original result leveraged the canonical Bernstein inequality, we here use the following MDS version of the Bernstein inequality:

Lemma F.8 (Bernstein-Type Inequality for Martingale Difference Sequences)**.** Let $X_1, \ldots, X_n$ form a sub-exponential Martingale Difference Sequence (MDS) such that $\max_{1\le i\le n} \|X_i\|_{\psi_1} \le \kappa$. Here $\|\cdot\|_{\psi_1}$ is the sub-Exponential norm defined in Definition G.2 in Appendix G. Then

$$\mathbb{P}\left(\left| \sum_i^n a_i X_i \right| \ge t \right) \le 2\exp\left[-C \min\left\{ \frac{t^2}{\kappa^2 \|\boldsymbol{a}\|^2}, \frac{t}{\kappa \|\boldsymbol{a}\|_\infty} \right\} \right],$$

where C is a constant.

We defer the proof of Lemma F.8 to Appendix F.2. With the Bernstein inequality for MDS, we can now outline the proof of Lemma F.7.

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a bounded martingale difference sequence. We first define the empircal processes

$$Z_f = 1/n \sum_{i=1}^n f^2(\boldsymbol{X}_i) - \mathbb{E}[f^2] \qquad W_f = \left(1/n \sum_{i=1}^n f^2(\boldsymbol{X}_i) \right)^2.$$

The following lemma from Mendelson et al. (2007) can be easily obtained from Lemma F.8.

Lemma F.9. There exists an absolute constant $c_1 > 0$ for which the following holds. Let $F \subset S_{L_2}$, $\alpha = \text{diam}(F, \|\cdot\|_{\psi_2})$ and set $n \geq 1$. For every $f, g \in F$ and every $u \geq 2$ we have

$$\mathbb{P}(W_{f-g} \geq u\|f-g\|_{\psi_2}) \leq \exp(-c_1 n u^2).$$

Also, for every $u > 0$,

$$\mathbb{P}(|Z_f - Z_g| \geq u\alpha\|f-g\|_{\psi_2}) \leq \exp(-c_1 n u^2),$$

and

$$\mathbb{P}(|Z_f| \geq u\alpha^2) \leq 2\exp(-c_1 n u^2).$$

The following two lemmas from Mendelson et al. (2007) hold in our setting since they do not require i.i.d. observations.

Lemma F.10. There exists an absolute constant C for which the following holds. Let $F \subset S_{L_2}$, $\alpha = \text{diam}(F, \|\cdot\|_{\psi_2})$ and $n \geq 1$. There is $F' \subset F$ such that $|F'| \leq 4^n$ and with probability at least $1 - \exp(-n)$, we have, for every $f \in F$,

$$W_{f-\pi_{F'}(f)} \leq \frac{C\gamma_2(F, \|\cdot\|_{\psi_2})}{\sqrt{n}},$$

where $\pi_{F'}(f)$ is a nearest point to f in F' with respect to the ψ_2 metric.

Lemma F.11. There exist absolute constants C and $c' > 0$ for which the following holds. Let $F \subset S_{L_2}$ and $\alpha = \text{diam}(F, \|\cdot\|_{\psi_2})$. Let $n \geq 1$ and $F' \subset F$ such that $|F'| \leq 4^n$. Then for every $w > 0$,

$$\sup_{f \in F'} |Z_f| \leq C\alpha\frac{\gamma_2(F, \|\cdot\|_{\psi_2})}{\sqrt{n}} + \alpha^2 w,$$

with probability at least $1 - 3\exp(-c' n \min(w, w^2))$.

Based on the above lemmas, the rest of proof of Theorem F.7 is stated as the proof of Theorem 1.4 in Mendelson et al. (2007).

F.2. Proof of Berstein Inequality for MDS (Lemma F.8)

We first present the following lemma from Vershynin (2012).

Lemma F.12 (Lemma 5.15 from Vershynin (2012))**.** If X is a sub-exponential random variable such that $\mathbb{E}[X] = 0$, then for every t such that $|t| \leq c/\|X\|_{\psi_1}$, we have

$$\mathbb{E}[\exp(tX)] \leq \exp(Ct^2\|X\|_{\psi_1}],$$

where $C, c > 0$ are constants.

We now prove the Berstein Inequality for MDS.

Proof of Lemma F.8. We begin by bounding the moment-generating function of $\sum_i^n a_i X_i$ as follows:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(t \cdot \sum_i^n a_i X_i\right)\right] &= \mathbb{E}\left[\prod_i^n \exp[ta_i X_i]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp[ta_n X_n]\prod_i^{n-1}\exp[ta_i X_i]\middle| X_1, \ldots, X_{n-1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp[ta_n X_n]\middle| X_1, \ldots, X_{n-1}\right] \cdot \mathbb{E}\left[\prod_i^{n-1}\exp[ta_i X_i]\middle| X_1, \ldots, X_{n-1}\right]\right] \\
&\leq \mathbb{E}\left[\exp(Ct^2 a_n^2 \kappa^2) \cdot \mathbb{E}\left[\prod_i^{n-1}\exp[ta_i X_i]\middle| X_1, \ldots, X_{n-1}\right]\right],
\end{aligned}$$

where the second inequality holds by the Law of Iterated Expectations. Furthermore, the last inequality holds by Lemma F.12 since $\mathbb{E}[X_n|X_1, \dots, X_{n-1}] = \mathbb{E}[X_n|\mathcal{F}_{n-1}] = 0$ by the definition of a MDS (recall Definition G.1 in Appendix G). By iteratively repeating this process, we obtain

$$\begin{aligned}\mathbb{E}\left[\exp\left(t \cdot \sum_i^n a_i X_i\right)\right] &\le \mathbb{E}\left[\prod_i^n \exp(Ct^2 a_i^2 \kappa^2)\right] \\ &= \exp\left(Ct^2 \sum_i^n a_i^2 \kappa^2\right) \\ &= \exp\left(Ct^2 \|\boldsymbol{a}\|_2^2 \kappa^2\right). \end{aligned} \tag{F.2}$$

Now note that by the Chernoff bound, for all λ such that $|\lambda| \le C/\|\boldsymbol{a}\|_\infty$ we have

$$\begin{aligned}\mathbb{P}\left(\sum_i^n a_i X_i \ge t\right) &= \mathbb{P}\left(\exp\left(\lambda \sum_i^n a_i X_i\right) \ge \exp(\lambda t)\right) \\ &\le \frac{\mathbb{E}[\exp[\lambda \sum_i^n a_i X_i]]}{\exp[\lambda t]} \\ &\le \frac{\exp[C\lambda^2 \|\boldsymbol{a}\|_2^2 \kappa^2]}{\exp[\lambda t]} \\ &= \exp[C\lambda^2 \|\boldsymbol{a}\|_2^2 \kappa^2 - \lambda t],\end{aligned}$$

where the last inequality holds by F.2. Now if we let $\lambda = \min\{t/(2C\|\boldsymbol{a}\|_2^2\kappa^2), c/(\|\boldsymbol{a}\|_\infty \kappa)\}$, then we see that if $t/(2C\|\boldsymbol{a}\|_2^2\kappa^2) < c/(\|\boldsymbol{a}\|_\infty \kappa)$, then

$$\mathbb{P}\left(\sum_i^n a_i X_i \ge t\right) \le \exp\left[\frac{t^2}{4C\|\boldsymbol{a}\|_2^2\kappa^2} - \frac{t^2}{2\kappa^2 C\|\boldsymbol{a}\|_2^2}\right] = \exp\left[\frac{-t^2}{4C\|\boldsymbol{a}\|_2^2\kappa^2}\right]. \tag{F.3}$$

Similarly, if $c/(\|\boldsymbol{a}\|_\infty \kappa) < t/(2C\|\boldsymbol{a}\|_2^2\kappa^2)$, then

$$\mathbb{P}\left(\sum_i^n a_i X_i \ge t\right) \le \exp\left[\frac{Cc\|\boldsymbol{a}\|_2^2\kappa}{\|\boldsymbol{a}\|_\infty} \cdot \frac{c}{\|\boldsymbol{a}\|_\infty \kappa} - \frac{ct}{\|\boldsymbol{a}\|_\infty \kappa}\right] \le \exp\left[-\frac{ct}{2\|\boldsymbol{a}\|_\infty \kappa}\right], \tag{F.4}$$

where the second inequality holds since $Cc\|\boldsymbol{a}\|_2^2\kappa/\|\boldsymbol{a}\|_\infty \le t/2$. Combining F.3 and F.4 yields

$$\mathbb{P}\left(\sum_i^n a_i X_i \ge t\right) \le \exp\left(\min\left\{\frac{-t^2}{4C\|\boldsymbol{a}\|_2^2\kappa^2}, \frac{-ct}{2\|\boldsymbol{a}\|_\infty \kappa}\right\}\right).$$

Note that we can repeat this process and replace each X_i with $-X_i$ to obtain this same bound for $\mathbb{P}(-\sum_i^n a_i X_i \ge t)$. This lemma then follows as a result of these two bounds. □

G. Auxiliary Definitions

In this section we present definitions used in the Appendix sections.

Definition G.1. A stochastic process $\{\xi_t\}$ is a *martingale difference sequence* with respect to filtration $\mathcal{F}_t$ if:

1. ξ_t is $\mathcal{F}_t$-measurable, and
2. $\mathbb{E}[\xi_t|\mathcal{F}_{t-1}] = 0$.

Definition G.2. The *sub-Exponential* norm of a random scalar variable X , $\|X\|_{\psi_1}$, is:

$$\|X\|_{\psi_1} = \sup_{q \ge 1} q^{-1}\left(\mathbb{E}\left[|X|^q\right]\right)^{1/q}.$$

The sub-Exponential norm of a random vector $\boldsymbol{X} \in \mathbb{R}^n$ is:

$$\|\boldsymbol{X}\|_{\psi_1} = \sup_{\mathbf{u} \in S^{n-1}} \|\langle \boldsymbol{X}, \mathbf{u} \rangle\|_{\psi_1},$$

where S^{n-1} is the unit sphere in $\mathbb{R}^n$ space.