7 Technical lemmas, propositions and proofs

Lemma 4. *Suppose* $\mathbb{E}(\mathbf{v}^\top \mathbf{x}_i)^4 \le R$ *for any* $\mathbf{v} \in \mathcal{S}^{d-1}$. *Define the* ℓ_4*-norm shrunk samples*

$$\widetilde{\mathbf{x}}_i := \frac{\min(\|\mathbf{x}_i\|_4, \tau)}{\|\mathbf{x}_i\|_4}\mathbf{x}_i,$$

where τ *is a threshold value. Then we have the following:*

1. $\|\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top - \mathbb{E}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\|_{\mathrm{op}} \le \|\widetilde{\mathbf{x}}_i\|_2^2 + \sqrt{R} \le \sqrt{d}\tau^2 + \sqrt{R}$;
2. $\|\mathbb{E}((\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top - \mathbb{E}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top)^\top(\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top - \mathbb{E}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top))\|_{\mathrm{op}} \le R(d+1)$;
3. *For all* $\xi > 0$, $\mathbb{P}\big\{\|\widetilde{\mathbf{\Sigma}}_n(\tau) - \mathbf{\Sigma}\|_{\mathrm{op}} \ge \xi\big(\frac{Rd\log n}{n}\big)^{1/2}\big\} \le n^{1-C\xi}$, *where* $\tau \asymp \big(nR/(\log n)\big)^{1/4}$ *and* C *is a universal constant.*

Proof. This result is from Fan et al. (2020+). For convenience of adapting the lemma to other settings, we present its proof here. Notice that

$$\|\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top - \mathbb{E}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\|_{\mathrm{op}} \le \|\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\|_{\mathrm{op}} + \|\mathbb{E}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\|_{\mathrm{op}} = \|\widetilde{\mathbf{x}}_i\|_2^2 + \sqrt{R} \le \sqrt{d}\tau^2 + \sqrt{R}. \tag{14}$$

Also for any $\mathbf{v} \in \mathcal{S}^{d-1}$, we have

$$\begin{aligned}\mathbb{E}(\mathbf{v}^\top\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\mathbf{v}) &= \mathbb{E}(\|\widetilde{\mathbf{x}}_i\|_2^2(\mathbf{v}^\top\widetilde{\mathbf{x}}_i)^2) \le \mathbb{E}(\|\mathbf{x}_i\|_2^2(\mathbf{v}^\top\mathbf{x}_i)^2)\\ &= \sum_{j=1}^d \mathbb{E}(x_{ij}^2(\mathbf{v}^\top\mathbf{x}_i)^2) \le \sum_{j=1}^d\sqrt{\mathbb{E}(x_{ij}^4)\mathbb{E}(\mathbf{v}^\top\mathbf{x}_i)^4} \le Rd\end{aligned}$$

Then it follows that $\|\mathbb{E}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\|_{\mathrm{op}} \le Rd$. Since $\|(\mathbb{E}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top)^\top\mathbb{E}(\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top)\|_{\mathrm{op}} \le R$,

$$\|\mathbb{E}((\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top - \mathbb{E}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top)^\top(\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top - \mathbb{E}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top))\|_{\mathrm{op}} \le R(d+1). \tag{15}$$

By the matrix Bernstein's inequality (Theorem 5.29 in Vershynin (2010)), we have for some constant c_1,

$$\mathbb{P}\Big(\|\frac{1}{n}\sum_{i=1}^n\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top - \mathbb{E}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\|_{\mathrm{op}} > t\Big) \le 2d\exp\Big(-c_1\big(\frac{nt^2}{R(d+1)} \wedge \frac{nt}{\sqrt{d}\tau^2 + \sqrt{R}}\big)\Big). \tag{16}$$

For any $\mathbf{v} \in \mathcal{S}^{d-1}$, it holds that

$$\mathbb{E}(\mathbf{v}^\top(\mathbf{x}_i\mathbf{x}_i^\top)\mathbf{v}1_{\{\|\mathbf{x}_i\|_4\ge\tau\}}) \le \sqrt{\mathrm{E}(\mathbf{v}^\top\mathbf{x}_i)^4P(\|\mathbf{x}_i\|_4 > \tau)} \le \left(\frac{R^2d}{\tau^4}\right)^{1/2} = \frac{R\sqrt{d}}{\tau^2}. \tag{17}$$

Therefore we have

$$\|\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top - \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top)\|_{\mathrm{op}} \le R\sqrt{d}/\tau^2. \tag{18}$$

Choose $\tau \asymp (nR/\log d)^{1/4}$ and substitute t with $\xi\sqrt{Rd\log n/n}$. Then we reach the final conclusion by combining the concentration bound and bias bound.

□

Proof of Lemma 1. Define a contraction function

$$\phi(x;\theta) = x^2 1_{\{|x|\le\theta\}} + (x-2\theta)^2 1_{\{\theta<x\le 2\theta\}} + (x+2\theta)^2 1_{\{-2\theta\le x<-\theta\}}.$$

One can verify that $\phi(x;\theta) \le x^2$ for any θ. This contraction function was used in a preliminary version of Negahban et al. (2012) to establish the RSC of negative log-likelihood. Given any $\mathbf{\Delta} \in \mathcal{B}_2(\mathbf{0}, r)$, by the Taylor expansion, we can find $v \in (0,1)$ such that

$$\begin{aligned}\delta\widetilde{\ell}_n(\boldsymbol{\beta}^*+\boldsymbol{\Delta};\boldsymbol{\beta}^*) &= \widetilde{\ell}_n(\boldsymbol{\beta}^*+\boldsymbol{\Delta}) - \widetilde{\ell}_n(\boldsymbol{\beta}^*) - \nabla\widetilde{\ell}_n(\boldsymbol{\beta}^*)^\top\boldsymbol{\Delta} = \frac{1}{2}\boldsymbol{\Delta}^\top\widetilde{\mathbf{H}}_n(\boldsymbol{\beta}^*+v\boldsymbol{\Delta})\boldsymbol{\Delta} \\ &= \frac{1}{2n}\sum_{i=1}^n b''(\widetilde{\mathbf{x}}_i^\top(\boldsymbol{\beta}^*+v\boldsymbol{\Delta}))(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i)^2 \geq \frac{1}{2n}\sum_{i=1}^n b''(\widetilde{\mathbf{x}}_i^\top(\boldsymbol{\beta}^*+v\boldsymbol{\Delta}))\phi(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i;\alpha_1 r)1_{\{|\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i|\leq\alpha_2\}} \\ &\geq \frac{m(\omega)}{2n}\sum_{i=1}^n \phi(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i;\alpha_1 r)1_{\{|\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i|\leq\alpha_2\}},\end{aligned} \tag{19}$$

where we choose $\omega = \alpha_1+\alpha_2 > \alpha_1 r + \alpha_2$ so that the last inequality holds by Condition (1) . For ease of notation, let $\mathcal{A}_i := \{|\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i| \leq \alpha_1 r\}$ and $\mathcal{B}_i := \{|\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i| \leq \alpha_2\}$. We have

$$\begin{aligned}\mathbb{E}[\phi(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i;\alpha_1 r)1_{\mathcal{B}_i}] &\geq \mathbb{E}[(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i)^2 1_{\mathcal{A}_i\cap\mathcal{B}_i}] \\ &\geq \boldsymbol{\Delta}^\top\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top 1_{\mathcal{A}_i\cap\mathcal{B}_i}]\boldsymbol{\Delta} - \boldsymbol{\Delta}^\top\mathbb{E}[(\mathbf{x}_i\mathbf{x}_i^\top - \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top)1_{\mathcal{A}_i\cap\mathcal{B}_i}]\boldsymbol{\Delta} \\ &\geq \boldsymbol{\Delta}^\top\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top 1_{\mathcal{A}_i\cap\mathcal{B}_i}]\boldsymbol{\Delta} - \boldsymbol{\Delta}^\top\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top - \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top]\boldsymbol{\Delta} \\ &\geq \boldsymbol{\Delta}^\top\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top)\boldsymbol{\Delta} - \boldsymbol{\Delta}^\top\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top 1_{\mathcal{A}_i^c\cup\mathcal{B}_i^c})\boldsymbol{\Delta} - \boldsymbol{\Delta}^\top\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top - \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top]\boldsymbol{\Delta} \\ &\geq \kappa_0\|\boldsymbol{\Delta}\|_2^2 - \sqrt{\mathbb{E}(\boldsymbol{\Delta}^\top\mathbf{x}_i)^4(\mathbb{P}(\mathcal{A}_i^c)+\mathbb{P}(\mathcal{B}_i^c))} - \boldsymbol{\Delta}^\top\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top - \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top]\boldsymbol{\Delta}.\end{aligned}$$

By the Markov Inequality,

$$\mathbb{P}(\mathcal{A}_i^c) \leq \frac{\mathbb{E}(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i)^4}{\alpha_1^4 r^4} \leq \frac{R}{\alpha_1^4} \quad \text{and} \quad \mathbb{P}(\mathcal{B}_i^c) \leq \frac{\mathbb{E}(\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i)^4}{\alpha_2^4} \leq \frac{R\|\boldsymbol{\beta}^*\|_2^4}{\alpha_2^4} \leq \frac{RL^4}{\alpha_2^4}.$$

Besides, according to (18),

$$\boldsymbol{\Delta}^\top\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top - \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top]\boldsymbol{\Delta} \leq \frac{R\sqrt{d}\|\boldsymbol{\Delta}\|_2^2}{\tau_1^2} \leq C_1 R\|\boldsymbol{\Delta}\|_2^2\left(\frac{d\log d}{n}\right)^{1/2},$$

where C_1 is certain constant. Therefore, for sufficiently large α_1, α_2, n and d,

$$\mathbb{E}[\phi(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i;\alpha_1 r)1_{\mathcal{B}_i}] \geq \frac{\kappa_0}{2}\|\boldsymbol{\Delta}\|_2^2. \tag{20}$$

For notational convenience, define $Z_i := \phi(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i;\alpha_1 r)1_{\mathcal{B}_i} = \phi(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i 1_{\mathcal{B}_i};\alpha_1 r)$ and $\Gamma_r := \sup_{\|\boldsymbol{\Delta}\|_2\leq r}\big|n^{-1}\sum_{i=1}^n Z_i - \mathbb{E}Z_i\big|$. Then an application of Massart's inequality (Massart (2000)) delivers that

$$\mathbb{P}\left\{|\Gamma_r - \mathbb{E}\Gamma_r| \geq \alpha_1^2 r^2\left(\frac{t}{n}\right)^{1/2}\right\} \leq 2\exp\big(-\frac{t}{8}\big). \tag{21}$$

The remaining job is to derive the order of $\mathbb{E}\Gamma_r$. Note that $|\phi(x_1;\theta) - \phi(x_2;\theta)| \leq 2\theta|x_1 - x_2|$ for any $x_1, x_2 \in \mathbb{R}$. By the symmetrization argument and then Ledoux-Talagrand contraction inequality (see Ledoux and Talagrand (2013), p. 112), for a sequence of i.i.d. Rademacher variables $\{\gamma_i\}_{i=1}^n$,

$$\begin{aligned}\mathbb{E}\Gamma_r &\leq 2\mathbb{E}\sup_{\|\boldsymbol{\Delta}\|_2\leq r}\Big|\frac{1}{n}\sum_{i=1}^n\gamma_i Z_i\Big| \leq 8\alpha_1 r\mathbb{E}\sup_{\|\boldsymbol{\Delta}\|_2\leq r}\Big|\langle\frac{1}{n}\sum_{i=1}^n\gamma_i\widetilde{\mathbf{x}}_i 1_{\{|\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i|\leq\alpha_2\}},\boldsymbol{\Delta}\rangle\Big| \\ &\leq 8\alpha_1 r^2\mathbb{E}\Big\|\frac{1}{n}\sum_{i=1}^n\gamma_i\widetilde{\mathbf{x}}_i 1_{\{|\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i|\leq\alpha_2\}}\Big\|_2 \leq 8\alpha_1 r^2\left(\mathbb{E}\Big\|\frac{1}{n}\sum_{i=1}^n\gamma_i\widetilde{\mathbf{x}}_i 1_{\{|\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i|\leq\alpha_2\}}\Big\|_2^2\right)^{1/2} \\ &\leq 8\alpha_1 r^2\left(\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\|\widetilde{\mathbf{x}}_i\|_2^2\right)^{1/2} \leq 8\alpha_1 r^2 R^{1/4}\left(\frac{d}{n}\right)^{1/2}.\end{aligned}$$

Combining the above inequality with (19), (20) and (21) yields that for any $t > 0$, with probability at least $1 - 2\exp(-t)$, for all $\boldsymbol{\Delta} \in \mathbb{R}^d$ such that $\|\boldsymbol{\Delta}\|_2 \leq r$,

$$\delta\widetilde{\ell}_n(\boldsymbol{\beta};\boldsymbol{\beta}^*) \geq \frac{m\kappa_0}{4}\|\boldsymbol{\Delta}\|_2^2 - \alpha_1^2\left(\frac{8t}{n}\right)^{1/2} r^2 - 8\alpha_1 R^{1/4}\left(\frac{d}{n}\right)^{1/2} r^2.$$

□

Proof of Theorem 1. Construct an intermediate estimator $\widetilde{\boldsymbol\beta}_\eta$ between $\widetilde{\boldsymbol\beta}$ and $\boldsymbol\beta^*$:

$$\widetilde{\boldsymbol\beta}_\eta = \boldsymbol\beta^* + \eta(\widetilde{\boldsymbol\beta} - \boldsymbol\beta^*),$$

where $\eta = 1$ if $\|\widetilde{\boldsymbol\beta} - \boldsymbol\beta^*\|_2 \le r$ and $\eta = r/\|\widetilde{\boldsymbol\beta} - \boldsymbol\beta^*\|_2$ if $\|\widetilde{\boldsymbol\beta} - \boldsymbol\beta^*\|_2 > r$. Write $\widetilde{\boldsymbol\beta}_\eta - \boldsymbol\beta^*$ as $\widetilde{\boldsymbol\Delta}_\eta$. By Lemma 1, it holds with probability at least $1 - 2\exp(-t)$ that

$$\kappa\|\widetilde{\boldsymbol\Delta}_\eta\|_2^2 - Cr^2\left\{\left(\frac{t}{n}\right)^{1/2} + \left(\frac{d}{n}\right)^{1/2}\right\} \le \delta\widetilde\ell_n(\widetilde{\boldsymbol\beta}_\eta;\boldsymbol\beta^*) \le -\nabla\widetilde\ell_n(\boldsymbol\beta^*)^\top\widetilde{\boldsymbol\Delta}_\eta \le \|\nabla\widetilde\ell_n(\boldsymbol\beta^*)\|_2\|\widetilde{\boldsymbol\Delta}_\eta\|_2,$$

which further implies that

$$\|\widetilde{\boldsymbol\Delta}_\eta\|_2 \le \frac{3\|\nabla\widetilde\ell_n(\boldsymbol\beta^*)\|_2}{\kappa} + \left(\frac{3c_1r^2}{\kappa}\right)^{1/2}\left(\frac{t}{n}\right)^{1/4} + \left(\frac{3c_2r^2}{\kappa}\right)^{1/2}\left(\frac{d}{n}\right)^{1/4}. \tag{22}$$

Now we derive the rate of $\|\nabla\widetilde\ell_n(\boldsymbol\beta^*)\|_2$.

$$\begin{aligned}\nabla\widetilde\ell_n(\boldsymbol\beta^*) &= \frac{1}{n}\sum_{i=1}^n(\widetilde z_i - b'(\widetilde{\mathbf x}_i^\top\boldsymbol\beta^*))\widetilde{\mathbf x}_i \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^n \widetilde z_i\widetilde{\mathbf x}_i - \mathbb E\widetilde z_i\widetilde{\mathbf x}_i}_{T_1} + \underbrace{\mathbb E(\widetilde z_i - b'(\widetilde{\mathbf x}_i^\top\boldsymbol\beta^*))\widetilde{\mathbf x}_i}_{T_2} + \underbrace{\frac{1}{n}\sum_{i=1}^n b'(\widetilde{\mathbf x}_i^\top\boldsymbol\beta^*)\widetilde{\mathbf x}_i - \mathbb E(b'(\widetilde{\mathbf x}_i^\top\boldsymbol\beta^*)\widetilde{\mathbf x}_i)}_{T_3}.\end{aligned} \tag{23}$$

where $\overline{\mathbf x}_i$ is between $\mathbf x_i$ and $\widetilde{\mathbf x}_i$ by the mean value theorem. In the following we will bound T_1, T_2 and T_3 respectively.

Bound for T_1: Define the Hermitian dilation matrix

$$\widetilde{\mathbf Z}_i := \widetilde z_i\begin{pmatrix} 0 & \widetilde{\mathbf x}_i^\top \\ \widetilde{\mathbf x}_i & \mathbf 0\end{pmatrix}$$

Note that

$$\|\mathbb E\widetilde{\mathbf Z}_i^2\|_{\mathrm{op}} = \|\mathbb E\Big[\widetilde z_i^2\begin{pmatrix}\widetilde{\mathbf x}_i^\top\widetilde{\mathbf x}_i & \mathbf 0^\top \\ \mathbf 0 & \widetilde{\mathbf x}_i\widetilde{\mathbf x}_i^\top\end{pmatrix}\Big]\|_{\mathrm{op}} = \max(\mathbb E(\widetilde z_i^2\widetilde{\mathbf x}_i^\top\widetilde{\mathbf x}_i), \|\mathbb E(\widetilde z_i^2\widetilde{\mathbf x}_i\widetilde{\mathbf x}_i^\top)\|_{\mathrm{op}})$$

For any $j \in [d]$,

$$\mathrm E\big(\widetilde z_i^2\widetilde x_{ij}^2\big) \le \sqrt{\mathrm E\, z_i^4\,\mathrm E\, x_{ij}^4} \le \sqrt{M_1R},$$

so $\mathbb E[\widetilde z_i^2\widetilde{\mathbf x}_i^\top\widetilde{\mathbf x}_i] \le d\sqrt{M_1R}$. In addition, for any $\mathbf v \in \mathbb R^d$ such that $\|\mathbf v\|_2 = 1$,

$$\mathrm E\big(\widetilde z_i^2(\mathbf v^\top\widetilde{\mathbf x}_i)^2\big) \le \sqrt{M_1R}.$$

We thus have $\|\mathbb E\widetilde{\mathbf Z}_i^2\|_{\mathrm{op}} \le d\sqrt{M_1R}$. In addition, $\|\mathbb E\widetilde{\mathbf Z}_i\|_{\mathrm{op}} = \mathbb E(\widetilde z_i\|\widetilde{\mathbf x}_i\|_2) \le \sqrt{\mathbb E z_i^2\mathbb E\|\mathbf x_i\|_2^2} \le \sqrt d(M_1R)^{1/4}$, which further implies that $\|\mathbb E(\widetilde{\mathbf Z}_i - \mathbb E\widetilde{\mathbf Z}_i)^2\|_{\mathrm{op}} \le 2d\sqrt{M_1R}$. Also notice that since $\|\widetilde{\mathbf x}_i\|_4 \le \tau_1$ and $\widetilde z_i \le \tau_2$, $\|\widetilde{\mathbf Z}_i\|_{\mathrm{op}} \le \frac12 d^{1/4}\tau_1\tau_2$. By the matrix Bernstein's inequality,

$$P\big(\|\frac1n\sum_{i=1}^n\widetilde{\mathbf Z}_i - \mathbb E\widetilde{\mathbf Z}_i\|_{\mathrm{op}} \ge t\big) \le d\exp\left\{-c_1\min\left(\frac{nt^2}{2d\sqrt{M_1R}}, \frac{2nt}{d^{1/4}\tau_1\tau_2}\right)\right\}.$$

Given that $\|T_1\|_2 = 2\|n^{-1}\sum_{i=1}^n\widetilde{\mathbf Z}_i - \mathbb E\widetilde{\mathbf Z}_i\|_{\mathrm{op}}$, it thus holds that

$$\mathbb P\Big(\|T_1\|_2 \ge 2t\Big) \le d\exp\left\{-c_1\min\left(\frac{nt^2}{2d\sqrt{M_1R}}, \frac{2nt}{d^{1/4}\tau_1\tau_2}\right)\right\}. \tag{24}$$

Bound for T_2: We decompose T_2 as follows:

$$\|T_2\|_2 \le \underbrace{\|\mathbb{E}(\widetilde{z}_i - z_i)\widetilde{\mathbf{x}}_i\|_2}_{T_{21}} + \underbrace{\|\mathbb{E}(z_i - y_i)\widetilde{\mathbf{x}}_i\|_2}_{T_{22}} + \underbrace{\|\mathbb{E}(y_i - b'(\mathbf{x}_i^\top \boldsymbol{\beta}^*))\widetilde{\mathbf{x}}_i\|_2}_{T_{23}}$$
$$+ \underbrace{\|\mathbb{E}(b'(\mathbf{x}_i^\top \boldsymbol{\beta}^*) - b'(\widetilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^*))\widetilde{\mathbf{x}}_i\|_2}_{T_{24}}.$$

Now we work on $\{T_{2i}\}_{i=1}^4$ one by one. For any $\mathbf{v} \in \mathbb{R}^d$ such that $\|\mathbf{v}\|_2 = 1$,

$$|\mathbb{E}(\widetilde{z}_i - z_i)(\mathbf{v}^\top \widetilde{\mathbf{x}}_i)| \le \mathbb{E}(|z_i|(\mathbf{v}^\top \mathbf{x}_i)1_{\{|z_i|>\tau_2\}}) \le \sqrt{\mathbb{E}(z_i^2(\mathbf{v}^\top \mathbf{x}_i)^2)\mathbb{P}(|z_i| > \tau_2)}.$$
$$\le (M_1 R)^{1/4} \frac{\sqrt{M_1}}{\tau_2^2},$$

thus we have $\|T_{21}\|_2 \le M_1^{3/4} R^{1/4}/\tau_2^2$. Again, for any $\mathbf{v} \in \mathbb{R}^d$ such that $\|\mathbf{v}\|_2 = 1$, since $\|\mathbb{E}\epsilon_i \mathbf{x}_i\|_2 \le M_2\sqrt{d/n}$,

$$\begin{aligned}\mathbb{E}[\epsilon_i(\widetilde{\mathbf{x}}_i^\top \mathbf{v})] &= \mathbb{E}[\epsilon_i((\widetilde{\mathbf{x}}_i - \mathbf{x}_i)^\top \mathbf{v})] + \mathbb{E}[\epsilon_i(\mathbf{x}_i^\top \mathbf{v})] \le \mathbb{E}[\epsilon_i|\mathbf{x}_i^\top \mathbf{v}|1_{\{\|\mathbf{x}_i\|_4 \ge \tau_1\}}] + M_2\Big(\frac{d}{n}\Big)^{1/2} \\ &\le \sqrt{\mathbb{E}(\epsilon_i(\mathbf{x}_i^\top \mathbf{v}))^2\mathbb{P}\big(\|\mathbf{x}_i\|_4 \ge \tau_1\big)} + M_2\Big(\frac{d}{n}\Big)^{1/2} \\ &\le (M_1 R)^{1/4}\frac{\sqrt{dR}}{\tau_1^2} + M_2\Big(\frac{d}{n}\Big)^{1/2}.\end{aligned}$$

Therefore $\|T_{22}\|_2 \le (M_1R)^{1/4}\sqrt{dR}/\tau_1^2 + M_2\sqrt{d/n}$. For T_{23}, since $\mathbb{E}[y_i - b'(\mathbf{x}_i^\top \boldsymbol{\beta}^*)|\mathbf{x}_i] = 0$, $T_{23} = 0$. Finally we bound T_{24}. For any $\mathbf{v} \in \mathbb{R}^d$ such that $\|\mathbf{v}\|_2 = 1$,

$$\begin{aligned}\|T_{24}\|_2 &\le M\,\mathrm{E}(\boldsymbol{\beta^*}^\top(\mathbf{x}_i - \widetilde{\mathbf{x}}_i))(\mathbf{v}^\top \widetilde{\mathbf{x}}_i) \le M\,\mathrm{E}[(\boldsymbol{\beta^*}^\top \mathbf{x}_i)(\mathbf{v}^\top \mathbf{x}_i)1_{\{\|\mathbf{x}_i\|_4 \ge \tau_1\}}] \\ &\le M\sqrt{\mathrm{E}(\boldsymbol{\beta^*}^\top \mathbf{x}_i)^2(\mathbf{v}^\top \mathbf{x}_i)^2 P\big(\|\mathbf{x}_i\|_4 \ge \tau_1\big)} \le ML\sqrt{dR}/\tau_1^2.\end{aligned}$$

To summarize here, we have

$$\|T_2\|_2 \le (M_1R)^{1/4}\Big(\frac{\sqrt{M_1}}{\tau_2^2} + \frac{\sqrt{dR}}{\tau_1^2}\Big) + ML\frac{\sqrt{dR}}{\tau_1^2} + M_2\Big(\frac{d}{n}\Big)^{1/2}. \tag{25}$$

Bound for T_3: We apply a similar proof strategy as in the bound for T_1. Define the following Hermitian dilation matrix:

$$\widetilde{\mathbf{X}}_i := b'(\widetilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^*)\begin{pmatrix} 0 & \widetilde{\mathbf{x}}_i^\top \\ \widetilde{\mathbf{x}}_i & \mathbf{0}\end{pmatrix}.$$

First,

$$\|\mathbb{E}\widetilde{\mathbf{X}}_i^2\|_{\mathrm{op}} = \max(\mathbb{E}(b'(\widetilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^*)\widetilde{\mathbf{x}}_i^\top \widetilde{\mathbf{x}}_i), \|\mathbb{E}b'(\widetilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^*)^2\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\|_{\mathrm{op}}).$$

Write $|b'(1)|$ as b_1. For any $j \in [d]$,

$$\begin{aligned}\mathbb{E}\big(b'(\widetilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^*)^2\widetilde{x}_{ij}^2\big) &\le \mathbb{E}[(b_1 + M|\widetilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^* - 1|)^2\widetilde{x}_{ij}^2] \le 2\mathbb{E}[((b_1+M)^2 + M^2(\widetilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^*)^2)\widetilde{x}_{ij}^2] \\ &\le 2M^2R\|\boldsymbol{\beta}^*\|_2^2 + 2(b_1+M)^2\sqrt{R} =: V,\end{aligned}$$

so $\mathbb{E}[b'(\widetilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^*)^2\widetilde{\mathbf{x}}_i^\top \widetilde{\mathbf{x}}_i] \le dV$. In addition, for any $\mathbf{v} \in \mathbb{R}^d$ such that $\|\mathbf{v}\|_2 = 1$,

$$\mathbb{E}\big(b'(\widetilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^*)^2(\mathbf{v}^\top \widetilde{\mathbf{x}}_i)^2\big) \le \mathbb{E}((b_1 + M|\widetilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^* - 1|)^2(\mathbf{v}^\top \widetilde{\mathbf{x}}_i)^2) \le V.$$

We thus have $\|\mathbb{E}\widetilde{\mathbf{X}}_i^2\|_{\mathrm{op}} \le dV$. In addition, $\|\mathbb{E}\widetilde{\mathbf{X}}_i\|_{\mathrm{op}} = \mathbb{E}(b'(\widetilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^*)\|\widetilde{\mathbf{x}}_i\|_2) \le \sqrt{\mathbb{E}b'(\widetilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^*)^2\mathbb{E}\|\widetilde{\mathbf{x}}_i\|_2^2} \le \sqrt{d}V$, which further implies that $\|\mathbb{E}(\widetilde{\mathbf{X}}_i - \mathbb{E}\widetilde{\mathbf{X}}_i)^2\|_{\mathrm{op}} \le (d+\sqrt{d})V$. Also notice that $\|\widetilde{\mathbf{X}}_i\|_{\mathrm{op}} \le ((b_1+M) + M\|\boldsymbol{\beta}^*\|_2 d^{1/4}\tau_1)d^{1/4}\tau_1$. By the matrix Bernstein's inequality,

$$\mathbb{P}\big(\|\frac{1}{n}\sum_{i=1}^n \widetilde{\mathbf{X}}_i - \mathrm{E}\,\widetilde{\mathbf{X}}_i\|_{\mathrm{op}} \ge t\big) \le d\exp\Big(-c_1\min\big(\frac{nt^2}{(d+\sqrt{d})V}, \frac{nt}{(b_1 + M + M\|\boldsymbol{\beta}^*\|_2 d^{1/4}\tau_1)d^{1/4}\tau_1}\big)\Big).$$

Given that $\|T_3\|_2 = 2\|n^{-1}\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i - \mathrm{E}\,\widetilde{\mathbf{X}}_i\|_{\mathrm{op}}$, it thus holds that

$$\mathbb{P}\Big(\|T_3\|_2 \geq 2t\Big) \leq d\exp\Big(-c_1\min\big(\frac{nt^2}{(d+\sqrt{d})V}, \frac{nt}{(b_1+M+M\|\boldsymbol{\beta}^*\|_2 d^{1/4}\tau_1)d^{1/4}\tau_1}\big)\Big). \tag{26}$$

Finally, choose $\tau_1, \tau_2 \asymp (n/\log n)^{1/4}$. Combining (24), (25) and (26) delivers that for some constant C_1 any $\xi > 1$,

$$\mathbb{P}\bigg\{\|\nabla\widetilde{\ell}_n(\boldsymbol{\beta}^*)\|_2 \geq C_1\xi\bigg(\frac{d\log n}{n}\bigg)^{1/2}\bigg\} \leq n^{1-\xi}. \tag{27}$$

Choose $t = \xi\log n$ and let r be larger than the RHS of (22). When d/n is sufficiently small and n is sufficiently large, we can obtain that

$$r \geq C_2\xi\bigg(\frac{d\log n}{n}\bigg)^{1/2} =: r_0,$$

where C_2 is a constant. Choose $r = r_0$. Then by (22), $\|\boldsymbol{\Delta}_\eta\|_2 \leq r_0$ and thus $\widetilde{\boldsymbol{\Delta}} = \widetilde{\boldsymbol{\Delta}}_\eta$. Finally, we reach the conclusion that

$$\mathbb{P}\bigg\{(\|\widetilde{\boldsymbol{\Delta}}\|_2 \geq C_2\xi\bigg(\frac{d\log n}{n}\bigg)^{1/2}\bigg\} \leq n^{1-\xi} + 2n^{-\xi} \leq 3n^{1-\xi}.$$

□

Proof of Corollary 1. The proof strategy is nearly the same as that for deriving Theorem 1, so we provide a roadmap here and do not dive into great details. For ease of notation, write $n^{-1}\sum_{i=1}^{n}\ell^w(\widetilde{\mathbf{x}}_i, z_i; \boldsymbol{\beta})$ as $\widetilde{\ell}^w(\boldsymbol{\beta})$ and denote the hessian matrix of $\widetilde{\ell}_n^w(\boldsymbol{\beta})$ by $\widetilde{\mathbf{H}}_n^w(\boldsymbol{\beta})$. Since $\widetilde{\mathbf{H}}_n^w(\boldsymbol{\beta}) = \nabla^2\widetilde{\ell}_n(\boldsymbol{\beta}) = \widetilde{\mathbf{H}}_n(\boldsymbol{\beta})$, we can directly obtain the uniform strong convexity of $\widetilde{\mathbf{H}}_n^w(\boldsymbol{\beta})$ from Lemma 1. In addition,

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_n^w(\boldsymbol{\beta}^*) &= \frac{1-p}{1-2p}\underbrace{\frac{1}{n}\sum_{i=1}^{n}(b'(\widetilde{\mathbf{x}}_i^\top\boldsymbol{\beta}^*) - z_i)\widetilde{\mathbf{x}}_i}_{T_1} - \frac{p}{1-2p}\underbrace{\frac{1}{n}\sum_{i=1}^{n}(b'(\widetilde{\mathbf{x}}_i^\top\boldsymbol{\beta}^*) - (1-z_i))\widetilde{\mathbf{x}}_i}_{T_2} \\
&= \frac{1-p}{1-2p}(T_1 - \mathbb{E}T_1) - \frac{p}{1-2p}(T_2 - \mathbb{E}T_2) + \frac{1-p}{1-2p}\mathbb{E}T_1 - \frac{p}{1-2p}\mathbb{E}T_2 \\
&= \frac{1-p}{1-2p}(T_1 - \mathbb{E}T_1) - \frac{p}{1-2p}(T_2 - \mathbb{E}T_2) + \mathbb{E}(b'(\widetilde{\mathbf{x}}_i^\top\boldsymbol{\beta}^*) - y_i)\widetilde{\mathbf{x}}_i.
\end{aligned}$$

Since $|b'(\widetilde{\mathbf{x}}_i^\top\boldsymbol{\beta}^*) - z_i| \leq 1$ and $|b'(\widetilde{\mathbf{x}}_i^\top\boldsymbol{\beta}^*) - (1-z_i)| \leq 1$, following the bound for T_1 in Theorem 1, we will obtain

$$\mathbb{P}\bigg\{\|\frac{1-p}{1-2p}(T_1 - \mathbb{E}T_1) - \frac{p}{1-2p}(T_2 - \mathbb{E}T_2)\|_2 \geq c_1\xi\bigg(\frac{d\log n}{n}\bigg)^{1/2}\bigg\} \leq n^{1-\xi},$$

where $c_1 > 0$ depends on R and p and $\xi > 1$. In addition, following the bound for T_{23} and T_{24} in Theorem 1, we shall obtain

$$\|\mathbb{E}(b'(\widetilde{\mathbf{x}}_i^\top\boldsymbol{\beta}^*) - y_i)\widetilde{\mathbf{x}}_i\|_2 \leq M_2 L\frac{\sqrt{dR}}{\tau_1^2} \leq c_2 M_2\bigg(\frac{dR\log n}{n}\bigg)^{1/2}.$$

where $c_2 > 0$ is a constant. Therefore, for some constant c_3 depending on R, p, M_2, R, we have

$$\mathbb{P}\bigg\{\|\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_n^w(\boldsymbol{\beta}^*)\|_2 \geq c_3\xi\bigg(\frac{d\log n}{n}\bigg)^{1/2}\bigg\} \leq n^{1-\xi}.$$

Combining this with the uniform strong convexity of $\widetilde{\mathbf{H}}_n^w(\boldsymbol{\beta})$ delivers the final conclusion. □

Proof of Lemma 2. According to (3), $[\nabla_{\boldsymbol\beta}\widetilde\ell(\boldsymbol\beta^*)]_j = (b'(\widetilde{\mathbf{x}}_i^\top\boldsymbol\beta^*) - \widetilde z_i)\widetilde x_{ij}$. Then we have

$$
\begin{aligned}
\Big|\frac{1}{n}\sum_{i=1}^n (b'(\widetilde{\mathbf{x}}_i^\top\boldsymbol\beta^*) - \widetilde z_i)\widetilde x_{ij}\Big| &\le \underbrace{\Big|\frac{1}{n}\sum_{i=1}^n b'(\widetilde{\mathbf{x}}_i^\top\boldsymbol\beta^*)\widetilde x_{ij} - \mathbb{E}b'(\widetilde{\mathbf{x}}_i^\top\boldsymbol\beta^*)\widetilde x_{ij}\Big|}_{T_1} + \underbrace{\big|\mathbb{E}(b'(\widetilde{\mathbf{x}}_i^\top\boldsymbol\beta^*) - \widetilde z_i)\widetilde x_{ij}\big|}_{T_2} \\
&\quad + \underbrace{\Big|\frac{1}{n}\sum_{i=1}^n \widetilde z_i\widetilde x_{ij} - \mathbb{E}\widetilde z_i\widetilde x_{ij}\Big|}_{T_3}.
\end{aligned} \tag{28}
$$

We start with the upper bound of T_1. By the Mean Value Theorem, for any $i\in[n]$, there exists ξ_i between 1 and $\widetilde x_i^\top\boldsymbol\beta^*$ such that $b'(\widetilde{\mathbf{x}}_i^\top\boldsymbol\beta^*) = b'(1) + b''(\xi_i)(\widetilde{\mathbf{x}}_i^\top\boldsymbol\beta^* - 1)$. Therefore we have

$$
\begin{aligned}
T_1 &\le \Big|\frac{1}{n}\sum_{i=1}^n b'(1)\widetilde x_{ij} - \mathbb{E}(b'(1)\widetilde x_{ij})\Big| + \Big|\frac{1}{n}\sum_{i=1}^n b''(\xi_i)\widetilde x_{ij}(\widetilde{\mathbf{x}}_i^\top\boldsymbol\beta^* - 1) - \mathbb{E}(b''(\xi_i)\widetilde x_{ij}(\widetilde{\mathbf{x}}_i^\top\boldsymbol\beta^* - 1))\Big| \\
&\le \Big|\frac{1}{n}\sum_{i=1}^n b'(1)\widetilde x_{ij} - \mathbb{E}(b'(1)\widetilde x_{ij})\Big| + \sum_{k=1}^d|\beta_k^*|\Big|\frac{1}{n}\sum_{i=1}^n b''(\xi_i)\widetilde x_{ij}\widetilde x_{ik} - \mathbb{E}b''(\xi_i)\widetilde x_{ij}\widetilde x_{ik}\Big| \\
&\quad + \Big|\frac{1}{n}\sum_{i=1}^n b''(\xi_i)\widetilde x_{ij} - \mathbb{E}(b''(\xi_i)\widetilde x_{ij})\Big|.
\end{aligned}
$$

Since $\mathrm{var}(\widetilde x_{ij}) \le \sqrt{R}$ and $|\widetilde x_{ij}| \le \tau_1$, an application of Bernstein's inequality (Theorem 2.10 in Boucheron et al. (2013)) yields that

$$
\mathbb{P}\bigg[\Big|\frac{1}{n}\sum_{i=1}^n b'(1)\widetilde x_{ij} - \mathbb{E}(b'(1)\widetilde x_{ij})\Big| \ge |b'(1)|\bigg\{\bigg(\frac{\sqrt{R}2t}{n}\bigg)^{1/2} + \frac{c_1\tau_1 t}{n}\bigg\}\bigg] \le 2\exp(-t),
$$

where $c_1>0$ is some universal constant. In addition, $b''(\xi_i)\widetilde x_{ij}\widetilde x_{ik} \le M\tau_1^2$ and $\mathrm{var}(b''(\xi_i)\widetilde x_{ij}\widetilde x_{ik}) \le \mathrm{E}(b''(\xi_i)\widetilde x_{ij}\widetilde x_{ik})^2 \le M^2R$. Again by Bernstein's inequality,

$$
\mathbb{P}\bigg\{\Big|\frac{1}{n}\sum_{i=1}^n b''(\xi_i)\widetilde x_{ij}\widetilde x_{ik} - \mathbb{E}(b''(\xi_i)\widetilde x_{ij}\widetilde x_{ik})\Big| \ge \bigg(\frac{2M^2Rt}{n}\bigg)^{1/2} + \frac{c_1M\tau_1^2t}{n}\bigg\} \le 2\exp(-t).
$$

Similarly,

$$
\mathbb{P}\bigg\{\Big|\frac{1}{n}\sum_{i=1}^n b''(\xi_i)\widetilde x_{ij} - \mathbb{E}(b''(\xi_i)\widetilde x_{ij})\Big| \ge \bigg(\frac{M^2\sqrt{R}t}{n}\bigg)^{1/2} + \frac{M\tau_1 t}{n}\bigg\} \le 2\exp(-t).
$$

Combining the above three inequalities delivers that

$$
\begin{aligned}
\mathbb{P}\bigg[T_1 \ge |b'(1)|\bigg\{\bigg(\frac{\sqrt{R}2t}{n}\bigg)^{1/2} + \frac{c_1\tau_1t}{n}\bigg\} + \bigg(\frac{2M^2Rt}{n}\bigg)^{1/2} + \frac{c_1M\tau_1^2t}{n} + \bigg(\frac{M^2\sqrt{R}t}{n}\bigg)^{1/2} + \frac{M\tau_1t}{n}\bigg] \\
\le 6\exp(-t).
\end{aligned} \tag{29}
$$

Now we bound T_2.

$$
\begin{aligned}
T_2 &= \mathbb{E}[(z_i - \widetilde z_i)\widetilde x_{ij}] + \mathbb{E}\epsilon_i\widetilde x_{ij} + \mathbb{E}[(b'(\mathbf{x}_i^\top\boldsymbol\beta^*) - b'(\widetilde{\mathbf{x}}_i^\top\boldsymbol\beta^*))\widetilde x_{ij}] \\
&\le \mathbb{E}[|z_i\widetilde x_{ij}|1_{\{|z_i|\ge\tau_2\}}] + \mathbb{E}(\epsilon_i x_{ij}) + \mathbb{E}\epsilon_i(x_{ij} - \widetilde x_{ij}) + M\sum_{k=1}^d|\beta_k^*|\mathbb{E}|\widetilde x_{ik}(\widetilde x_{ij} - x_{ij})| \\
&\le (M_1R)^{1/4}\frac{\sqrt{M_1}}{\tau_2^2} + \frac{M_3}{\sqrt{n}} + \frac{(M_1R)^{1/4}}{\tau_1^2} + MM_2\frac{\sqrt{R}}{\tau_1^2}.
\end{aligned} \tag{30}
$$

Finally we bound T_3. Note that $|\widetilde z_i\widetilde x_{ij}| \le \tau_1\tau_2$, $\mathrm{var}(\widetilde x_{ij}\widetilde z_i) \le \mathrm{E}\,|\widetilde z_i\widetilde x_{ij}|^2 \le \sqrt{M_1R}$. According to the Bernstein's inequality,

$$
\mathbb{P}\bigg\{|T_3| \ge \bigg(\frac{2t\sqrt{M_1R}}{n}\bigg)^{1/2} + \frac{c_1\tau_1\tau_2t}{n}\bigg\} \le 2\exp(-t). \tag{31}
$$

Choose $\tau_1, \tau_2 \asymp (n/\log d)^{1/4}$. Combining (29), (30) and (31) delivers that for some constant $C_1 > 0$ that depends on $M, R, \{M_i\}_{i=1}^3, b'(1)$ and any $\xi > 1$,

$$\mathbb{P}\bigg\{|[\nabla_{\boldsymbol{\beta}}\widetilde{\ell}(\boldsymbol{\beta}^*)]_j| \geq C_1\xi\bigg(\frac{\log d}{n}\bigg)^{1/2}\bigg\} \leq 2d^{-\xi}.$$

Then by the union bound for all $j \in [d]$, it holds that

$$\mathbb{P}\bigg\{\max_{j\in[d]}|[\nabla_{\boldsymbol{\beta}}\widetilde{\ell}(\boldsymbol{\beta}^*)]_j| \geq C_1\xi\bigg(\frac{\log d}{n}\bigg)^{1/2}\bigg\} \leq 2d^{1-\xi}.$$

□

Proof of Lemma 3. The proof strategy is quite similar to that for Lemma 1, except that we need to take advantage of the restricted cone $\mathcal{C}(\mathcal{S})$ that $\boldsymbol{\Delta}$ lies in. First of all, for any $1 \leq j, k \leq d$,

$$|\mathbb{E}(\widetilde{x}_{ij}\widetilde{x}_{ik} - x_{ij}x_{ik})| \leq \sqrt{\mathbb{E}(x_{ij}x_{ik})^2(\mathbb{P}(|x_{ij}| \geq \tau_1) + \mathbb{P}(|x_{ik}| \geq \tau_1))} \leq \frac{\sqrt{2}R}{\tau_1^2}.$$

We thus have

$$\|\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top - \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top]\|_{\max} \leq \frac{\sqrt{2}R}{\tau_1^2} \leq CR\bigg(\frac{2\log d}{n}\bigg)^{1/2}, \tag{32}$$

where $C > 0$ is some constant. Again, define a contraction function

$$\phi(x;\theta) = x^2 1_{\{|x|\leq\theta\}} + (x-2\theta)^2 1_{\{\theta<x\leq 2\theta\}} + (x+2\theta)^2 1_{\{-2\theta\leq x<-\theta\}}.$$

Given any $\boldsymbol{\Delta} \in \mathcal{B}_2(\mathbf{0}, r) \cap \mathcal{C}(\mathcal{S})$, by the Taylor expansion, we can find $v \in (0,1)$ such that

$$\begin{aligned}
&\delta\widetilde{\ell}_n(\boldsymbol{\beta}^* + \boldsymbol{\Delta}; \boldsymbol{\beta}^*) = \widetilde{\ell}_n(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - \widetilde{\ell}_n(\boldsymbol{\beta}^*) - \nabla\widetilde{\ell}_n(\boldsymbol{\beta}^*)^\top\boldsymbol{\Delta} = \frac{1}{2}\boldsymbol{\Delta}^\top\widetilde{\mathbf{H}}_n(\boldsymbol{\beta}^* + v\boldsymbol{\Delta})\boldsymbol{\Delta} \\
&= \frac{1}{2n}\sum_{i=1}^n b''(\widetilde{\mathbf{x}}_i^\top(\boldsymbol{\beta}^* + v\boldsymbol{\Delta}))(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i)^2 \geq \frac{1}{2n}\sum_{i=1}^n b''(\widetilde{\mathbf{x}}_i^\top(\boldsymbol{\beta}^* + v\boldsymbol{\Delta}))\phi(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i; \alpha_1 r)1_{\{|\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i|\leq\alpha_2\}} \\
&\geq \frac{m(\omega)}{2n}\sum_{i=1}^n \phi(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i; \alpha_1 r)1_{\{|\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i|\leq\alpha_2\}},
\end{aligned} \tag{33}$$

where we choose $\omega = \alpha_1 + \alpha_2 > \alpha_1 r + \alpha_2$ so that the last inequality holds by Condition (1). For ease of notation, let $\mathcal{A}_i := \{|\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i| \leq \alpha_1 r\}$ and $\mathcal{B}_i := \{|{\boldsymbol{\beta}^*}^\top\widetilde{\mathbf{x}}_i| \leq \alpha_2\}$. We have

$$\begin{aligned}
&\mathbb{E}[\phi(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i; \alpha_1 r)1_{\mathcal{B}_i}] \geq \mathbb{E}[(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i)^2 1_{\mathcal{A}_i\cap\mathcal{B}_i}] \\
&\quad\geq \boldsymbol{\Delta}^\top\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top 1_{\mathcal{A}_i\cap\mathcal{B}_i}]\boldsymbol{\Delta} - \boldsymbol{\Delta}^\top\mathbb{E}[(\mathbf{x}_i\mathbf{x}_i^\top - \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top)1_{\mathcal{A}_i\cap\mathcal{B}_i}]\boldsymbol{\Delta} \\
&\quad\geq \boldsymbol{\Delta}^\top\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top 1_{\mathcal{A}_i\cap\mathcal{B}_i}]\boldsymbol{\Delta} - \boldsymbol{\Delta}^\top\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top - \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top]\boldsymbol{\Delta} \\
&\quad\geq \boldsymbol{\Delta}^\top\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top)\boldsymbol{\Delta} - \boldsymbol{\Delta}^\top\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top 1_{\mathcal{A}_i^c\cup\mathcal{B}_i^c})\boldsymbol{\Delta} - \boldsymbol{\Delta}^\top\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top - \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top]\boldsymbol{\Delta} \\
&\quad\geq \kappa_0\|\boldsymbol{\Delta}\|_2^2 - \sqrt{\mathbb{E}(\boldsymbol{\Delta}^\top\mathbf{x}_i)^4(\mathbb{P}(\mathcal{A}_i^c) + \mathbb{P}(\mathcal{B}_i^c))} - \boldsymbol{\Delta}^\top\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top - \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top]\boldsymbol{\Delta} \\
&\quad\geq \kappa_0\|\boldsymbol{\Delta}\|_2^2 - \sqrt{R(\mathbb{P}(\mathcal{A}_i^c) + \mathbb{P}(\mathcal{B}_i^c))}\|\boldsymbol{\Delta}\|_2^2 - \|\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top - \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top]\|_{\max}\|\boldsymbol{\Delta}\|_1^2
\end{aligned}$$

By the Markov Inequality and (32),

$$\begin{aligned}
\mathbb{P}(\mathcal{A}_i^c) &\leq \frac{\mathbb{E}(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i)^2}{\alpha_1^2 r^2} \leq \frac{\mathbb{E}(\boldsymbol{\Delta}^\top\mathbf{x}_i)^2 + \boldsymbol{\Delta}^\top\mathbb{E}(\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top - \mathbf{x}_i\mathbf{x}_i^\top)\boldsymbol{\Delta}}{\alpha_1^2 r^2} \\
&\leq \frac{\sqrt{R}\|\boldsymbol{\Delta}\|_2^2 + CRs\|\boldsymbol{\Delta}\|_2^2\sqrt{2\log d/n}}{\alpha_1^2 r^2} \leq \frac{\sqrt{R} + CRs\sqrt{\log d/n}}{\alpha_1^2}
\end{aligned}$$

and

$$\mathbb{P}(\mathcal{B}_i^c) \le \frac{\mathbb{E}(\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i)^2}{\alpha_2^2} \le \frac{\mathbb{E}(\boldsymbol{\beta}^{*\top}\mathbf{x}_i)^2 + \boldsymbol{\beta}^{*\top}\mathbb{E}(\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top - \mathbf{x}_i\mathbf{x}_i^\top)\boldsymbol{\beta}^*}{\alpha_2^2}$$
$$\le \frac{\sqrt{R}\|\boldsymbol{\beta}^*\|_2^2 + CRs\|\boldsymbol{\beta}^*\|_2^2\sqrt{2\log d/n}}{\alpha_2^2} \le \frac{\sqrt{R}L^2 + CRL^2 s\sqrt{2\log d/n}}{\alpha_2^2}.$$

Overall, as long as α_1, α_2 are sufficiently large and $s\sqrt{\log d/n}$ is not large,

$$\mathbb{E}[\phi(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i;\alpha_1 r)1_{\mathcal{B}_i}] \ge \frac{\kappa_0}{2}\|\boldsymbol{\Delta}\|_2^2. \tag{34}$$

For notational convenience, define $Z_i := \phi(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i;\alpha_1 r)1_{\mathcal{B}_i} = \phi(\boldsymbol{\Delta}^\top\widetilde{\mathbf{x}}_i 1_{\mathcal{B}_i};\alpha_1 r)$ and $\Gamma_r := \sup_{\|\boldsymbol{\Delta}\|_2\le r, \boldsymbol{\Delta}\in\mathcal{C}(\mathcal{S})}\big|n^{-1}\sum_{i=1}^n Z_i - \mathbb{E}Z_i\big|$. Then an application of Massart's inequality (Massart (2000)) delivers that

$$\mathbb{P}\bigg\{|\Gamma_r - \mathbb{E}\Gamma_r| \ge \alpha_1^2 r^2\Big(\frac{t}{n}\Big)^{1/2}\bigg\} \le 2\exp\big(-\frac{t}{8}\big). \tag{35}$$

The remaining job is to derive the order of $\mathbb{E}\Gamma_r$. By the symmetrization argument and Ledoux-Talagrand contraction inequality, for a sequence of i.i.d. Rademacher variables $\{\gamma_i\}_{i=1}^n$,

$$\mathbb{E}\Gamma_r \le 2\mathbb{E}\sup_{\|\boldsymbol{\Delta}\|_2\le r,\boldsymbol{\Delta}\in\mathcal{C}(\mathcal{S})}\Big|\frac{1}{n}\sum_{i=1}^n\gamma_i Z_i\Big| \le 8\alpha_1 r\mathbb{E}\sup_{\|\boldsymbol{\Delta}\|_2\le r,\boldsymbol{\Delta}\in\mathcal{C}(\mathcal{S})}\Big|\langle\frac{1}{n}\sum_{i=1}^n\gamma_i\widetilde{\mathbf{x}}_i 1_{\{|\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i|\le\alpha_2\}}, \boldsymbol{\Delta}\rangle\Big|$$
$$\le 8\alpha_1\sqrt{s}r^2\mathbb{E}\|\frac{1}{n}\sum_{i=1}^n\gamma_i\widetilde{\mathbf{x}}_i 1_{\{|\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i|\le\alpha_2\}}\|_{\max}.$$

For any $1 \le j \le d$, by Bernstein inequality,

$$\mathbb{P}\bigg\{\Big|\frac{1}{n}\sum_{i=1}^n\gamma_i\widetilde{x}_{ij}1_{\{|\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i|\le\alpha_2\}}\Big| \ge \bigg(\frac{2\sqrt{R}t}{n}\bigg)^{1/2} + \frac{C_1\tau_1 t}{n}\bigg\} \le 2\exp(-t),$$

where C_1 is some constant. By the union bound, we can deduce that for some constant C_2,

$$\mathbb{P}\bigg\{\|\frac{1}{n}\sum_{i=1}^n\gamma_i\widetilde{\mathbf{x}}_i 1_{\{|\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i|\le\alpha_2\}}\|_{\max} \ge C_2\Big(\frac{t\log d}{n}\Big)^{1/2}\bigg\} \le 2d^{1-t},$$

which further implies that

$$\mathbb{E}\Gamma_r \le 8\alpha_1\sqrt{s}r^2\mathbb{E}\|\frac{1}{n}\sum_{i=1}^n\gamma_i\widetilde{\mathbf{x}}_i 1_{\{|\boldsymbol{\beta}^{*\top}\widetilde{\mathbf{x}}_i|\le\alpha_2\}}\|_{\max} \le 8C_3\alpha_1 r^2\Big(\frac{s\log d}{n}\Big)^{1/2}.$$

for some constant C_3. Combining the above inequality with (33), (34) and (35) yields that for any $t > 0$, with probability at least $1 - 2\exp(-t)$,

$$\delta\widetilde{\ell}_n(\boldsymbol{\beta};\boldsymbol{\beta}^*) \ge \frac{m\kappa_0}{4}\|\boldsymbol{\Delta}\|_2^2 - \alpha_1^2 r^2\Big(\frac{8t}{n}\Big)^{1/2} - 8C_3\alpha_1 r^2\Big(\frac{s\log d}{n}\Big)^{1/2}.$$

□

Proof of Theorem 2. According to Lemma 1 in Negahban et al. (2012), as long as $\lambda \ge 2\|\nabla\widetilde{\ell}_n(\boldsymbol{\beta})\|_{\max}$, $\widetilde{\boldsymbol{\Delta}} \in \mathcal{C}(\mathcal{S})$. We construct an intermediate estimator $\widetilde{\boldsymbol{\beta}}_\eta$ between $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$:

$$\widetilde{\boldsymbol{\beta}}_\eta = \boldsymbol{\beta}^* + \eta(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*),$$

where $\eta = 1$ if $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le r$ and $\eta = r/\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ if $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > r$. Choose $\lambda = 2C\xi\sqrt{\log d/n}$, where C and ξ are the same as in Lemma 2. By Lemmas 2 and 3, it holds with probability at least $1 - 2\exp(-t)$ that

$$\begin{aligned}
\kappa\|\widetilde{\boldsymbol{\Delta}}_\eta\|_2^2 - C_0 r^2\left\{\left(\frac{t}{n}\right)^{1/2} + \left(\frac{s\log d}{n}\right)^{1/2}\right\} &\le \delta\widetilde{\ell}_n(\widetilde{\boldsymbol{\beta}}_\eta; \boldsymbol{\beta}^*) \\
&= \widetilde{\ell}_n(\widetilde{\boldsymbol{\beta}}_\eta) - \widetilde{\ell}_n(\boldsymbol{\beta}^*) - \nabla\widetilde{\ell}_n(\boldsymbol{\beta}^*)^\top\widetilde{\boldsymbol{\Delta}}_\eta \\
&\le \lambda\|\widetilde{\boldsymbol{\Delta}}_\eta\|_1 + \|\nabla\widetilde{\ell}_n(\boldsymbol{\beta}^*)\|_{\max}\|\widetilde{\boldsymbol{\Delta}}_\eta\|_1 \\
&\le (\lambda + \|\nabla\widetilde{\ell}_n(\boldsymbol{\beta}^*)\|_{\max})\|\widetilde{\boldsymbol{\Delta}}_\eta\|_1 \\
&\le 4(\lambda + \|\nabla\widetilde{\ell}_n(\boldsymbol{\beta}^*)\|_{\max})\|[\widetilde{\boldsymbol{\Delta}}_\eta]_{\mathcal{S}}\|_1 \\
&\le 4\sqrt{s}(\lambda + \|\nabla\widetilde{\ell}_n(\boldsymbol{\beta}^*)\|_{\max})\|\widetilde{\boldsymbol{\Delta}}_\eta\|_2.
\end{aligned} \tag{36}$$

Some algebra delivers that

$$\begin{aligned}
\|\widetilde{\boldsymbol{\Delta}}_\eta\|_2 &\le \frac{4\sqrt{s}(\lambda + \|\nabla\widetilde{\ell}_n(\boldsymbol{\beta}^*)\|_{\max})}{\kappa} + r\left[\frac{C_0}{\kappa}\left\{\left(\frac{t}{n}\right)^{1/2} + \left(\frac{s\log d}{n}\right)^{1/2}\right\}\right]^{1/2} \\
&= \frac{4\sqrt{s}\|\nabla\widetilde{\ell}_n(\boldsymbol{\beta}^*)\|_{\max}}{\kappa} + \frac{8C\xi}{\kappa}\left(\frac{s\log d}{n}\right)^{1/2} + r\left[\frac{C_0}{\kappa}\left\{\left(\frac{t}{n}\right)^{1/2} + \left(\frac{s\log d}{n}\right)^{1/2}\right\}\right]^{1/2}.
\end{aligned} \tag{37}$$

Choose $t = \xi\log d$ above. Let r be greater than the RHS of the inequality above. For sufficiently sufficiently small $s\log d/n$, we have $r \ge 5\sqrt{s}\|\nabla\widetilde{\ell}_n(\boldsymbol{\beta}^*)\|_{\max}/\kappa$. Define $r_0 := 5\sqrt{s}\|\nabla\widetilde{\ell}_n(\boldsymbol{\beta}^*)\|_{\max}/\kappa$ and choose $r = r_0$. Therefore, $\|\widetilde{\boldsymbol{\Delta}}_\eta\|_2 \le r$ and $\widetilde{\boldsymbol{\Delta}}_\eta = \widetilde{\boldsymbol{\Delta}}$. By Lemma 2, we reach the conclusion. □