Online Appendix

Private Optimal Inventory Policy Learning for Feature-based Newsvendor with Unknown Demand

Tuoyi Zhao, Wen-Xin Zhou and Lan Wang

Appendix A: Technical Lemmas

This section presents several technical lemmas that are essential for certifying that events $\mathcal{E}_0(B)$, $\mathcal{E}_1(\delta_0,\delta_1)$ and $\mathcal{E}_2(R)$ defined in (12)–(14) hold with high probability for properly chosen (B,δ_0,δ_1,R). For $\boldsymbol{\beta}\in\mathbb{R}^p$, after a change of variable $\boldsymbol{\delta}=\boldsymbol{\beta}-\boldsymbol{\beta}^*$ we define

$$\widehat{R}_{\varpi}(\boldsymbol{\delta})=\widehat{Q}_{\varpi}(\boldsymbol{\beta})-\widehat{Q}_{\varpi}(\boldsymbol{\beta}^*)-\langle\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^*),\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle, \tag{1}$$

$$\widehat{B}_{\varpi}(\boldsymbol{\delta})=\widehat{Q}_{\varpi}(\boldsymbol{\beta}^*)-\widehat{Q}_{\varpi}(\boldsymbol{\beta})-\langle\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}),\boldsymbol{\beta}^*-\boldsymbol{\beta}\rangle, \tag{2}$$

and their population counterparts $R_{\varpi}(\boldsymbol{\delta})=\mathrm{E}\{\widehat{R}_{\varpi}(\boldsymbol{\delta})\}$ and $B_{\varpi}(\boldsymbol{\delta})=\mathrm{E}\{\widehat{B}_{\varpi}(\boldsymbol{\delta})\}$.

Lemma 1 provides lower bounds for the first-order Taylor series remainder of the population (smoothed) loss in a neighborhood of $\boldsymbol{\beta}^*$, that is, $R_{\varpi}(\cdot)$ and $B_{\varpi}(\cdot)$. Lemma 2 provides upper bounds (in high probability) for the fluctuation of the empirical loss $|\widehat{D}_{\varpi}(\boldsymbol{\delta})-D_{\varpi}(\boldsymbol{\delta})|$ uniformly over $\boldsymbol{\delta}$ in a compact set excluding a local neighborhood of the origin. Lemma 3 establishes uniform convergence of the gradient and Hessian of the smoothed quantile loss to their population counterparts, as long as the sample size scales linearly in the number of features.

Note that under Condition 2, $m_q=\sup_{\mathbf{u}\in\mathbb{S}^{p-1}}\mathrm{E}|\langle\mathbf{w},\mathbf{u}\rangle|^q$ $(q=3,4)$ are bounded and only depend on υ_1.

Lemma 1. *Conditions 1–3 ensure that*

$$\min\big\{R_{\varpi}(\boldsymbol{\delta}),B_{\varpi}(\boldsymbol{\delta})\big\}\ge 0.5\big(f_l-l_0\kappa_1\varpi\big)\|\boldsymbol{\delta}\|_{\Sigma}^2 \quad \textit{for all } \boldsymbol{\delta}\in\Theta_{\Sigma}(1).$$

Lemma 2. *Let $r>\delta>0$. For any $u\ge 0$, the following bounds hold:*

(i) With probability at least $1-e^{-u}$,

$$\widehat{D}_{\varpi}(\boldsymbol{\delta})-D_{\varpi}(\boldsymbol{\delta})\le C_0\upsilon_1 r\sqrt{\frac{p+u}{n}} \quad \textit{for all } \boldsymbol{\delta}\in\Theta_{\Sigma}(r); \tag{3}$$

(ii) With probability at least $1-e^{-u}$,

$$\widehat{D}_{\varpi}(\boldsymbol{\delta})-D_{\varpi}(\boldsymbol{\delta})\le C_1\upsilon_1\sqrt{\frac{p+\log_2(r/\delta)+u}{n}}\cdot\|\boldsymbol{\delta}\|_{\Sigma} \quad \textit{for all } \boldsymbol{\delta}\in\Theta_{\Sigma}(r)\setminus\Theta_{\Sigma}(\delta), \tag{4}$$

where $C_0,C_1>0$ are absolute constants. The same upper bounds in (3) and (4) also apply to $D_{\varpi}(\boldsymbol{\delta})-\widehat{D}_{\varpi}(\boldsymbol{\delta})$.

Lemma 3. *Assume Conditions 1–3 hold and let $R > 0$ be a fixed constant. For any $u \geq 0$, the following bounds hold with probability at least $1 - 3e^{-u}$ as long as $n \gtrsim \upsilon_1^4(p+u)$: $\max_{1 \leq i \leq n} \|\Sigma^{-1/2}\mathbf{x}_i\|_2 \leq 1 + \upsilon_1\sqrt{p-1} + \upsilon_1\sqrt{2(u+\log n)}$,*

$$\sup_{\boldsymbol{\beta} \in \Theta^*_\Sigma(R)} \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}) - \nabla Q_\varpi(\boldsymbol{\beta})\|_{\Sigma^{-1}} \leq C_2 \upsilon_1 \sqrt{\frac{p\log(n/\varpi) + u}{n}} + (f_u\varpi + 2\kappa_u)\frac{R}{n} \tag{5}$$

and

$$\begin{aligned} &\sup_{\boldsymbol{\beta} \in \Theta^*_\Sigma(R)} \|\nabla^2 \widehat{Q}_\varpi(\boldsymbol{\beta}) - \nabla^2 Q_\varpi(\boldsymbol{\beta})\|_{\Sigma^{-1}} \\ &\lesssim \upsilon_1^2 \left\{ \sqrt{\frac{p\log(n/\varpi)+u}{n\varpi}} + \frac{p\log(n/\varpi)+u}{n\varpi} \right\} + \left\{ \upsilon_1(p + \log n + u)^{1/2} + l_0 m_3 \varpi^2 \right\} \frac{R}{n^2}. \end{aligned} \tag{6}$$

Moreover, $\|\nabla Q_\varpi(\boldsymbol{\beta}^*)\|_{\Sigma^{-1}} \leq 0.5 l_0 \kappa_2 \varpi^2$ and $\sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\nabla Q_\varpi(\boldsymbol{\beta})\|_{\Sigma^{-1}} \leq \bar{\tau} = \max(\tau, 1-\tau)$.

The following lemma provides upper bounds for the i.i.d. standard normal random vectors $\{\mathbf{g}_t\}_{t=0}^{T-1}$ in the noisy gradient descent algorithm. In particular, inequality (8) slightly improves the tail bound in Lemma 11 of Cai et al. (2021).

Lemma 4. *Let $\mathbf{g}_0, \mathbf{g}_1, \ldots, \mathbf{g}_{T-1}$ ($T \geq 1$) be independent random vectors from $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. Then, for any $z \geq 0$,*

$$P\left\{ \max_{0 \leq t \leq T-1} \|\mathbf{g}_t\|_2 \geq \sqrt{p} + \sqrt{2(\log T + z)} \right\} \leq e^{-z}. \tag{7}$$

Moreover, for any $\rho \in (0,1)$, we have with probability at least $1 - e^{-z}$ that

$$\sum_{t=0}^{T-1} \rho^t \|\mathbf{g}_t\|_2^2 \leq \frac{p}{1-\rho} + 2\sqrt{\frac{pz}{1-\rho^2}} + 2z \leq \frac{2p}{1-\rho} + \left(\frac{1}{1+\rho} + 2\right)z. \tag{8}$$

Lemma 5. *Let $z \geq 0$. In addition to (18), assume*

$$n \geq B_T \sigma / \Delta, \tag{9}$$

where $B_T := \sqrt{p} + \sqrt{2(\log T + z)}$ and $\Delta = \phi_1 - \delta_0 - b^* \in (0, f_l/2)$. Then, conditioning on $\mathcal{E}_0(B) \cap \mathcal{E}_1(\delta_0, \delta_1) \cap \mathcal{E}_2(2)$, the noisy gradient descent iterates satisfy $\boldsymbol{\beta}^{(t)} \in \Theta^*_\Sigma(1)$ for all $t = 1, \ldots, T$ with probability (over $\{\mathbf{g}_t\}_{t=0}^{T-1}$) at least $1 - e^{-z}$.

Lemma 6. *Let $R_0 = \|\boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma$, and assume $\|\widehat{\boldsymbol{\beta}}_\varpi - \boldsymbol{\beta}^*\|_\Sigma \leq r_0$ and $\eta_0 \in (0, 1/(2f_u)]$. For any $T_0 \geq 2$ and $z \geq 0$, assume the sample size satisfies*

$$n \geq \frac{e-1}{4-e}(2\bar{\tau}B + 1)\max\{1, \eta_0/R_0\}^2 T_0 B_{T_0} \sigma,$$

where $B_{T_0} := \sqrt{p} + \sqrt{2(\log T_0 + z)}$. Then, conditioned on the event $\mathcal{E}_0(B) \cap \mathcal{E}_2(R)$ with $R \geq 2R_0 + r_0$, we have

$$\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2 \leq (1 + 1/T_0)^t R_0^2 + \{(1+1/T_0)^t - 1\}(2\bar{\tau}B + 1)\eta_0^2 \frac{T_0 B_{T_0}\sigma}{n}, \quad 1 \leq t \leq T_0$$

with probability (over i.i.d. normal vectors $\{\mathbf{g}_t\}_{t=0}^{T_0-1}$) at least $1 - e^{-z}$. In particular, $\max_{1 \leq t \leq T_0} \|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma \leq 2R_0$.

Appendix B: Proof of Theorems

B.1. Proof of Theorems 2 and 3

The high probability bound in Theorem 2 is a direct consequence of Theorem 4 and Proposition 4 by taking $z = 2\log n$ therein. To obtain the second bound under expectation, note that

$$\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)}\|_\Sigma \le \left\| \frac{\eta_0}{n} \sum_{i=1}^n \{\bar{K}_\varpi(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(0)}) - \tau\}\overline{\mathbf{w}}_i + \frac{\eta_0\sigma}{n}\mathbf{g}_0 \right\|_2 \le \bar{\tau}\eta_0 B + \frac{\eta_0\sigma}{n}\|\mathbf{g}_0\|_2,$$

where we have used the property that $\max_{1\le i\le n}\|\overline{\mathbf{w}}_i\|_2 \le B$. Similarly, it can be shown that for each $t \ge 1$, $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|_\Sigma \le \bar{\tau}\eta_0 B + \frac{\eta_0\sigma}{n}\|\mathbf{g}_t\|_2$. With $\sigma = 2\bar{\tau}BT^{1/2}/\mu$, applying the triangle inequality repeatedly yields

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^{(0)}\|_\Sigma \le \sum_{t=1}^T \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|_\Sigma \le \bar{\tau}\eta_0 BT + 2\bar{\tau}\eta_0 BT^{1/2}\frac{1}{\mu n}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|_2.$$

Initialized at any $\boldsymbol{\beta}^{(0)}$ satisfying $\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_\Sigma \le 1$, it holds almost surely that

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma \le 1 + \bar{\tau}\eta_0 BT + 2\bar{\tau}\eta_0 BT^{1/2}\frac{1}{\mu n}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|_2. \tag{10}$$

From the first high probability bound, we see that there exists an event $\mathcal{E}$ such that $P(\mathcal{E}) \ge 1 - C_1 n^{-2}$ and on $\mathcal{E}$ it holds

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma \le r_n \asymp T^{1/2}\eta_0\frac{p + \log n}{\mu n} + \frac{1}{f_l}\sqrt{\frac{p\log n}{n}}.$$

By combining the above two bounds, with one being almost surely true and the other holding with high probability, we further derive that

$$\begin{aligned}
\mathrm{E}\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma &= \mathrm{E}\{\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma \mathbb{1}(\mathcal{E})\} + \mathrm{E}\{\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma \mathbb{1}(\mathcal{E}^{\mathrm{c}})\} \\
&\le r_n + (1 + \bar{\tau}\eta_0 BT)P(\mathcal{E}^{\mathrm{c}}) + 2\bar{\tau}\eta_0 BT^{1/2}\frac{1}{\mu n}\sum_{t=0}^{T-1}\mathrm{E}\{\|\mathbf{g}_t\|_2\mathbb{1}(\mathcal{E}^{\mathrm{c}})\} \\
&\le r_n + C_1\frac{1 + \bar{\tau}\eta_0 BT}{n^2} + 2\bar{\tau}\eta_0 BT^{1/2}\frac{1}{\mu n}\sum_{t=0}^{T-1}(\mathrm{E}\|\mathbf{g}_t\|_2^2)^{1/2}P(\mathcal{E}^{\mathrm{c}})^{1/2} \\
&\le r_n + C_1\frac{1 + \bar{\tau}\eta_0 BT}{n^2} + 2\bar{\tau}\eta_0\sqrt{C_1 p}\,\frac{BT^{3/2}}{\mu n^2}.
\end{aligned}$$

Taking $B \asymp \sqrt{p + \log n}$ and $T \asymp \log n$, the second and third terms on the right-hand side are dominated by r_n in order. This proves the claimed upper bound under expectation.

To prove the regret bound in Theorem 3, recall that the optimal coefficient $\boldsymbol{\beta}^*$ in the linear demand model satisfies the first-order condition $\nabla Q(\boldsymbol{\beta}^*) = 0$. Under Condition 3, the Hessian matrix $\nabla^2 Q(\boldsymbol{\beta}) = \mathrm{E}\{f_{\varepsilon|\mathbf{x}}(\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*))\mathbf{x}\mathbf{x}^{\mathrm{T}}\}$ satisfies $\|\Sigma^{-1/2}\nabla^2 Q(\boldsymbol{\beta})\Sigma^{-1/2}\|_2 \le f_u\|\mathrm{E}(\mathbf{w}\mathbf{w}^{\mathrm{T}})\|_2 = f_u$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$. Therefore, applying the mean value theorem yields $Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}^*) = Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}^*) - \langle \nabla Q(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle \le 0.5 f_u\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\Sigma^2$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$. This, joint with the property that $\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma \lesssim f_l^{-1}\sqrt{(p\log n)/n} + \sqrt{\log n}\,\eta_0(p + \log n)/(\mu n)$ holds with high probability, proves the claimed regret bound.

To bound the expected regret, by (10) and the assumption $\eta_0 \le 1/\max\{2f_u, f_l + \bar{\tau}\}$ we have

$$\begin{aligned}
\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma^2 &\le 2(1 + \bar{\tau}\eta_0 BT)^2 + 8\bar{\tau}^2\eta_0^2 B^2 T\frac{1}{(\mu n)^2}\left(\sum_{t=0}^{T-1}\|\mathbf{g}_t\|_2\right)^2 \\
&\le 2(1 + 0.5\bar{\tau}f_u^{-1}BT)^2 + 2\bar{\tau}^2\left(\frac{BT}{f_u\mu n}\right)^2\sum_{t=0}^{T-1}\|\mathbf{g}_t\|_2^2.
\end{aligned}$$

This, together with the inequality $\mathrm{E}\|\mathbf{g}_t\|_2^4 \le p\sum_{j=1}^p \mathrm{E} g_{t,j}^4 = 3p^2$, implies

$$\mathrm{E}\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma^2 \le r_n^2 + (1 + 0.5\bar{\tau} f_u^{-1} BT)^2 \frac{2C_1}{n^2} + 2\bar{\tau}^2 \sqrt{3C_1}\frac{p}{n}\left(\frac{BT}{f_u \mu n}\right)^2.$$

Once again, the leading term on the right-hand side is r_n^2 (in order). This proves the claimed expected regret bound. $\square$

B.2. Proof of Theorem 4

For simplicity, we write $\mathcal{E}_0 = \mathcal{E}_0(B)$, $\mathcal{E}_1 = \mathcal{E}_1(\delta_0, \delta_1)$ and $\mathcal{E}_2 = \mathcal{E}_2(2)$ for the events defined in (12)–(14), where $\delta_0, \delta_1 > 0$ satisfy the constraints in (18).

To control the random perturbations in noisy gradient descent, for any $z > 0$ we define

$$\mathcal{G} = \mathcal{G}(z) = \left\{ \max_{0 \le t \le T-1} \|\mathbf{g}_t\|_2 \le B_T \right\} \bigcap \left\{ \sum_{t=0}^{T-1} (1-\epsilon)^t \|\mathbf{g}_{T-1-t}\|_2^2 \le 2p/\epsilon + 3z \right\} \tag{11}$$

where $B_T := \sqrt{p} + \sqrt{2(\log T + z)}$ and $\epsilon := \eta_0 \phi_1$. It thus follows from Lemma 4 that $P\{\mathcal{G}(z)\} \ge 1 - 2e^{-z}$. In the following, we prove (20) by conditioning on these events.

Lemma 5 shows that starting at an initial estimate $\boldsymbol{\beta}^{(0)} \in \Theta_\Sigma^*(1)$ and conditioning on $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{G}$, all the successive iterates will stay in $\Theta_\Sigma^*(1)$. Next, we establish a contraction property for the noisy gradient descent iterates. Define

$$\widetilde{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_0 \Sigma^{-1} \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}) \quad \text{and} \quad \mathbf{h}_t = \frac{\eta_0 \sigma}{n} \mathbf{g}_t, \quad t = 0, 1, \ldots, T-1,$$

and note that $\boldsymbol{\beta}^{(t+1)} = \widetilde{\boldsymbol{\beta}}^{(t+1)} + \Sigma^{-1/2}\mathbf{h}_t$. Note that (9) is guaranteed by (19). Lemma 5 ensures that $\boldsymbol{\beta}^{(t)} \in \Theta_\Sigma^*(1)$ for all $t = 0, 1, \ldots, T$. Similarly, it can be shown that the non-private gradient descent iterates $\widetilde{\boldsymbol{\beta}}^{(t)}$ also stay in the ball $\Theta_\Sigma^*(1)$ for $t = 1, \ldots, T$. For simplicity, set

$$\boldsymbol{\delta}^{(t)} = \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*, \quad \widetilde{\boldsymbol{\delta}}^{(t)} = \widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^* \quad \text{for} \quad t = 0, 1, \ldots, T.$$

To upper bound $\|\boldsymbol{\delta}^{(t+1)}\|_\Sigma$ at each iteration, we split the analysis into two cases.

Case 1: Suppose the non-private gradient descent iterate $\widetilde{\boldsymbol{\beta}}^{(t+1)}$ is in $\Theta_\Sigma(r_0)$. Then

$$\|\boldsymbol{\delta}^{(t+1)}\|_\Sigma \le r_0 + \|\mathbf{h}_t\|_2. \tag{12}$$

Case 2: Assume otherwise $\widetilde{\boldsymbol{\beta}}^{(t+1)} \notin \Theta_\Sigma(r_0)$. For the previously defined $\epsilon = \eta_0\phi_1 \in (0, 1/2]$,

$$\begin{aligned}
\|\boldsymbol{\delta}^{(t+1)}\|_\Sigma^2 &= \|\widetilde{\boldsymbol{\delta}}^{(t+1)}\|_\Sigma^2 + \|\mathbf{h}_t\|_2^2 + 2\langle \Sigma^{1/2}\widetilde{\boldsymbol{\delta}}^{(t+1)}, \mathbf{h}_t \rangle \\
&\le (1+\epsilon)\|\widetilde{\boldsymbol{\delta}}^{(t+1)}\|_\Sigma^2 + (1 + 1/\epsilon)\|\mathbf{h}_t\|_2^2 \\
&= (1+\epsilon)\|\boldsymbol{\delta}^{(t)}\|_\Sigma^2 + (1+1/\epsilon)\|\mathbf{h}_t\|_2^2 \\
&\quad + 2\eta_0(1+\epsilon)\underbrace{\left\{ \frac{\eta_0}{2}\|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)})\|_{\Sigma^{-1}}^2 - \langle \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*, \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}) \rangle \right\}}_{\Pi}.
\end{aligned} \tag{13}$$

To bound Π, we use the local strong convexity and smoothness properties of $\widehat{Q}_\varpi(\cdot)$ as in (16) and (17). The latter implies

$$\widehat{Q}_\varpi(\widetilde{\boldsymbol{\beta}}^{(t+1)}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}) \le \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}), \widetilde{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^{(t)} \rangle + f_u \|\widetilde{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_\Sigma^2.$$

For the former, we assume without loss generality that $r_1 = (\delta_0 + \delta_1)/\phi_1 \geq 1/n$. Otherwise if $r_1 < 1/n$, (16) implies $\widehat{Q}_\varpi(\boldsymbol{\beta}^*) - \widehat{Q}_\varpi(\boldsymbol{\beta}) - \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}), \boldsymbol{\beta}^* - \boldsymbol{\beta} \rangle \geq \phi_1(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\Sigma^2 - n^{-1}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\Sigma)$ for all $\boldsymbol{\beta} \in \Theta_\Sigma^*(1) \setminus \Theta_\Sigma^*(1/n)$. Then we obtain (20) with r_1 replaced by $1/n$.

Since $r_1 \geq 1/n$, it holds for every $\boldsymbol{\beta}^{(t)} \in \Theta_\Sigma^*(1/n)$ that

$$\widehat{Q}_\varpi(\boldsymbol{\beta}^*) - \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}) - \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)} \rangle \geq 0 \geq \phi_1 \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma^2 - (\delta_0 + \delta_1)\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma.$$

Otherwise if $\boldsymbol{\beta}^{(t)} \in \Theta_\Sigma^*(1) \setminus \Theta_\Sigma^*(1/n)$,

$$\widehat{Q}_\varpi(\boldsymbol{\beta}^*) - \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}) - \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)} \rangle \geq \phi_1 \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma^2 - (\delta_0 + \delta_1)\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma.$$

Together, the above upper and lower bounds imply

$$\begin{aligned}
&\widehat{Q}_\varpi(\widetilde{\boldsymbol{\beta}}^{(t+1)}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^*) \\
&\leq \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}), \widetilde{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^* \rangle + f_u \|\widetilde{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_\Sigma^2 - \phi_1 \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma^2 + (\delta_0 + \delta_1)\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma \\
&= \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}), \widetilde{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^* \rangle + f_u \eta_0^2 \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)})\|_{\Sigma^{-1}}^2 - \phi_1 \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma^2 + (\delta_0 + \delta_1)\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma \\
&= \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^* \rangle - \eta_0(1 - f_u \eta_0) \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)})\|_{\Sigma^{-1}}^2 - \phi_1 \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma^2 + (\delta_0 + \delta_1)\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma \\
&\leq \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^* \rangle - \frac{\eta_0}{2} \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)})\|_{\Sigma^{-1}}^2 - \phi_1 \|\boldsymbol{\delta}^{(t)}\|_\Sigma^2 + (\delta_0 + \delta_1)\|\boldsymbol{\delta}^{(t)}\|_\Sigma,
\end{aligned}$$

where the last inequality holds when $\eta_0 \leq 1/(2f_u)$. Since $\widetilde{\boldsymbol{\beta}}^{(t+1)} \in \Theta_\Sigma^*(1) \setminus \Theta_\Sigma^*(r_0)$, the restricted lower bound (15) yields

$$\widehat{Q}_\varpi(\widetilde{\boldsymbol{\beta}}^{(t+1)}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^*) \geq \phi_1 \|\widetilde{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_\Sigma \cdot \big(\|\widetilde{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_\Sigma - r_0\big) \geq 0.$$

We hence conclude that $\Pi \leq -\phi_1 \|\boldsymbol{\delta}^{(t)}\|_\Sigma^2 + (\delta_0 + \delta_1)\|\boldsymbol{\delta}^{(t)}\|_\Sigma$. Substituting this into (13) implies

$$\begin{aligned}
&\|\boldsymbol{\delta}^{(t+1)}\|_\Sigma^2 \\
&\leq \big\{1 + \epsilon - 2\eta_0\phi_1(1+\epsilon)\big\}\|\boldsymbol{\delta}^{(t)}\|_\Sigma^2 + (1 + 1/\epsilon)\|\mathbf{h}_t\|_2^2 + 2\eta_0(1+\epsilon)(\delta_0 + \delta_1)\|\boldsymbol{\delta}^{(t)}\|_\Sigma \\
&\leq \big\{1 + \epsilon - 2\eta_0\phi_1(1+\epsilon)\big\}\|\boldsymbol{\delta}^{(t)}\|_\Sigma^2 + (1 + 1/\epsilon)\|\mathbf{h}_t\|_2^2 \\
&\qquad\qquad\qquad + (1+\epsilon)(\eta_0\phi_1)^2\|\boldsymbol{\delta}^{(t)}\|_\Sigma^2 + (1+\epsilon)(\delta_0 + \delta_1)^2/\phi_1^2 \\
&= (1+\epsilon)(1 - \eta_0\phi_1)^2\|\boldsymbol{\delta}^{(t)}\|_\Sigma^2 + (1 + 1/\epsilon)\|\mathbf{h}_t\|_2^2 + (1+\epsilon)r_1^2. \qquad (14)
\end{aligned}$$

On the other hand, observe that

$$(r_0 + \|\mathbf{h}_t\|_2)^2 = r_0^2 + \|\mathbf{h}_t\|_2^2 + 2r_0\|\mathbf{h}_t\|_2 \leq (1 + 1/\epsilon)\|\mathbf{h}_t\|_2^2 + (1+\epsilon)r_0^2.$$

Recall that $\epsilon = \eta_0\phi_1 \in (0, 1/2]$. It then follows from (12) and (14) that

$$\|\boldsymbol{\delta}^{(t+1)}\|_\Sigma^2 \leq (1-\epsilon)\|\boldsymbol{\delta}^{(t)}\|_\Sigma^2 + (1 + 1/\epsilon)\|\mathbf{h}_t\|_2^2 + (1+\epsilon)(r_0 \vee r_1)^2,$$

for $t = 0, 1, \ldots, T-1$. This recursive bound further implies

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma^2 \leq (1-\epsilon)^T + (1 + 1/\epsilon)\bigg\{\Big(\frac{\eta_0\sigma}{n}\Big)^2 \underbrace{\sum_{t=0}^{T-1}(1-\epsilon)^t\|\mathbf{g}_{T-1-t}\|_2^2}_{\leq 2p/\epsilon + 3z \text{ on } \mathcal{G}} + (r_0 \vee r_1)^2\bigg\}. \qquad (15)$$

Let $T = T_1 + T_2$ with $T_1 \geq 2\log(n)/\log((1-\epsilon)^{-1})$, and r^* be as in (20). Then, conditioned on $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{G}$, we obtain $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma \leq r^*$ for all $T_1 \leq t \leq T$, as desired. In other words, r^* given in (20) characterizes the convergence rate of the noisy gradient descent iterators $\boldsymbol{\beta}^{(t)}$ to the true parameter $\boldsymbol{\beta}^*$.

Turning to the non-private empirical risk minimizer $\widehat{\boldsymbol{\beta}}_\varpi \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \widehat{Q}_\varpi(\boldsymbol{\beta})$, we claim that $\|\widehat{\boldsymbol{\beta}}_\varpi - \boldsymbol{\beta}^*\|_\Sigma \leq r_0 = (\delta_0 + b^*)/\phi_1$ conditioning on $\mathcal{E}_1$. To see this, note that for any $\boldsymbol{\delta} \in \Theta_\Sigma(1)$,

$$\widehat{D}_\varpi(\boldsymbol{\delta}) = R_\varpi(\boldsymbol{\delta}) + \langle \nabla Q_\varpi(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle + \{\widehat{D}_\varpi(\boldsymbol{\delta}) - D_\varpi(\boldsymbol{\delta})\} \geq \phi_1 \|\boldsymbol{\delta}\|_\Sigma^2 - (\delta_0 + b^*)\|\boldsymbol{\delta}\|_\Sigma.$$

By (18), $\widehat{D}_\varpi(\boldsymbol{\delta}) \geq \phi_1 - \delta_0 - b^* > 0$ for all $\boldsymbol{\delta} \in \partial\Theta_\Sigma(1)$–the boundary of $\Theta_\Sigma(1)$. On the other hand, note that $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\beta}}_\varpi - \boldsymbol{\beta}^*$ satisfies $\widehat{D}_\varpi(\widehat{\boldsymbol{\delta}}_\varpi) \leq 0$. By the convexity of $\widehat{Q}_\varpi(\cdot)$, we must have $\widehat{\boldsymbol{\beta}}_\varpi \in \Theta_\Sigma^*(1)$. Otherwise if $\|\widehat{\boldsymbol{\beta}}_\varpi - \boldsymbol{\beta}^*\|_\Sigma > 1$, there exists some $\alpha \in (0,1)$ such that $\widetilde{\boldsymbol{\beta}} = \alpha\widehat{\boldsymbol{\beta}}_\varpi + (1-\alpha)\boldsymbol{\beta}^*$ lies on the boundary of $\Theta_\Sigma^*(1)$, that is, $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\Sigma = 1$. Then $\widetilde{\boldsymbol{\delta}} := \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ satisfies $\widehat{D}_\varpi(\widetilde{\boldsymbol{\delta}}) > 0$. However, by the convexity of $\widehat{D}_\varpi$,

$$\widehat{D}_\varpi(\widetilde{\boldsymbol{\delta}}) = \widehat{D}_\varpi(\alpha\widehat{\boldsymbol{\delta}}_\varpi + (1-\alpha)\mathbf{0}) \leq \alpha\widehat{D}_\varpi(\widehat{\boldsymbol{\delta}}) + (1-\alpha)\widehat{D}_\varpi(\mathbf{0}) \leq 0.$$

This leads to a contradiction. As a result, from (15) it can be deduced that

$$0 \geq \widehat{D}_\varpi(\widehat{\boldsymbol{\delta}}) \geq \phi_1\|\widehat{\boldsymbol{\delta}}\|_\Sigma^2 - (\delta_0 + b^*)\|\widehat{\boldsymbol{\delta}}\|_\Sigma,$$

thereby verifying the claim.

Finally, we prove the convergence of $\boldsymbol{\beta}^{(t)}$ towards the non-private estimator $\widehat{\boldsymbol{\beta}}_\varpi$, for which it has been shown that $\|\widehat{\boldsymbol{\beta}}_\varpi - \boldsymbol{\beta}^*\|_\Sigma \leq r_0$ conditioned on $\mathcal{E}_1$. By the triangle inequality, $\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma \leq \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma + \|\widehat{\boldsymbol{\beta}}_\varpi - \boldsymbol{\beta}^*\|_\Sigma \leq r^* + r_0$ for all $T_1 \leq t \leq T$ conditioned on $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{G}$. To improve this "crude" bound, the key element is the refined restricted strong convexity property, guaranteed by the event $\mathcal{E}_3 = \mathcal{E}_3(r, \phi_2)$ with $r = r^* + r_0$.

Thus far we have shown that $\{\widetilde{\boldsymbol{\beta}}^{(t)}\}_{t=0,1,\ldots,T} \subseteq \Theta_\Sigma^*(1)$ and $\{\boldsymbol{\beta}^{(t)}\}_{t=T_1,\ldots,T} \subseteq \Theta_\Sigma^*(r^*)$. Moreover, note that inequality (13) remains valid if $\boldsymbol{\beta}^*$ is replaced by $\widehat{\boldsymbol{\beta}}_\varpi$, that is,

$$\begin{aligned}\|\boldsymbol{\beta}^{(t+1)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2 &\leq (1+\epsilon)\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2 + (1+1/\epsilon)\|\mathbf{h}_t\|_2^2 \\ &\quad + 2\eta_0(1+\epsilon)\left\{\frac{\eta_0}{2}\|\nabla\widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)})\|_{\Sigma^{-1}}^2 - \langle \boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi, \nabla\widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)})\rangle\right\}.\end{aligned}$$

Conditioned on $\mathcal{E}_3(r, \phi_2)$ with $r = r^* + r_0 \geq 2r_0$, we have, for $t = T_1, T_1+1, \ldots, T-1$,

$$\begin{aligned}\widehat{Q}_\varpi(\widetilde{\boldsymbol{\beta}}^{(t+1)}) = \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)} &\leq \langle \nabla\widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}), \widetilde{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^{(t)}\rangle + f_u\|\widetilde{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_\Sigma^2 \\ \text{and}\ \ \widehat{Q}_\varpi(\widehat{\boldsymbol{\beta}}_\varpi) - \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}) &\geq \langle \nabla\widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}), \widehat{\boldsymbol{\beta}}_\varpi - \boldsymbol{\beta}^{(t)}\rangle + \phi_2\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2.\end{aligned}$$

Following a similar argument, and recall that $\widehat{\boldsymbol{\beta}}_\varpi$ minimizes $\widehat{Q}_\varpi(\cdot)$, we obtain

$$\begin{aligned}0 &\leq \widehat{Q}_\varpi(\widetilde{\boldsymbol{\beta}}^{(t+1)} - \widehat{Q}_\varpi(\widehat{\boldsymbol{\beta}}_\varpi) \\ &\leq \langle \nabla\widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi\rangle - \frac{\eta_0}{2}\|\nabla\widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}\|_{\Sigma^{-1}}^2 - \phi_2\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2.\end{aligned}$$

Putting together the pieces yields the recursive bound

$$\|\boldsymbol{\beta}^{(t+1)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2 \leq (1-\epsilon)\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2 + (1+1/\epsilon)\|\mathbf{h}_t\|_2^2, \quad t = T_1, T_1+1, \ldots, T.$$

Provided $T_2 \geq \log(n)/\log((1-\epsilon)^{-1})$, conditioning on $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{G}$ it follows that

$$\begin{aligned} \|\boldsymbol{\beta}^{(T)} - \widehat{\boldsymbol{\beta}}_{\varpi}\|_{\Sigma}^2 &\leq (1-\epsilon)^{T_2}\|\boldsymbol{\beta}^{(T_1)} - \widehat{\boldsymbol{\beta}}_{\varpi}\|_{\Sigma}^2 + (1+1/\epsilon)\left(\frac{\eta_0\sigma}{n}\right)^2 \sum_{t=0}^{T_2-1}(1-\epsilon)^t\|\mathbf{g}_{T-1-t}\|_2^2 \\ &\leq \frac{r^2}{n} + (1+1/\epsilon)(2p/\epsilon + 3z)\left(\frac{\eta_0\sigma}{n}\right)^2, \end{aligned}$$

as claimed. This completes the proof of the theorem. $\square$

B.3. Proof of Theorem 5

To begin with, recall from the proof of Theorem 4 that conditioning on $\mathcal{E}_0(B) \cap \mathcal{E}_1(\delta_0, \delta_1) \cap \mathcal{E}_2(2)$, the non-private empirical risk minimizer $\widehat{\boldsymbol{\beta}}_{\varpi}$ satisfies $\|\widehat{\boldsymbol{\beta}}_{\varpi} - \boldsymbol{\beta}^*\|_{\Sigma} \leq r_0 = (\delta_0 + b^*)/\phi_1 < 1$. For any $T_0 \geq 2$ and $z \geq 0$, set

$$\mathcal{G}_0(z) = \left\{\max_{0 \leq t \leq T_0 - 1}\|\mathbf{g}_t\|_2 \leq B_{T_0} := \sqrt{p} + \sqrt{2(\log T_0 + z)}\right\},$$

so that $P\{\mathcal{G}_0(z)\} \geq 1 - e^{-z}$. From Lemma 6 we see that conditioned on $\mathcal{E}_0(B) \cap \mathcal{E}_2(R) \cap \mathcal{G}_0(z)$ with $R = 2R_0 + r_0$, the iterates $\{\boldsymbol{\beta}^{(t)}\}_{t=1,\ldots,T_0}$ satisfy

$$\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_{\varpi}\|_{\Sigma}^2 \leq eR_0^2 + (e-1)(2\bar{\tau}B + 1/2)\frac{T_0 B_{T_0}\sigma}{n}\eta_0^2 \leq 4R_0^2 \tag{16}$$

as long as $n \geq \frac{e-1}{4-e}(2\bar{\tau}B + 1/2)T_0 B_{T_0}\sigma$. Consequently, $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_{\Sigma} \leq \|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_{\varpi}\|_{\Sigma} + \|\widehat{\boldsymbol{\beta}}_{\varpi} - \boldsymbol{\beta}^*\|_{\Sigma} \leq R$, and hence $\boldsymbol{\beta}^{(t)} \in \Theta_{\Sigma}^*(R)$ for all $t = 0, 1, \ldots, T_0$.

Keep the notation from the proof of Theorem 4, and note that $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_0\Sigma^{-1}\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)}) - \Sigma^{-1/2}\mathbf{h}_t$, where $\mathbf{h}_t = \frac{\eta_0\sigma}{n}\mathbf{g}_t$ and $\eta_0 \leq 1/(2f_u)$. Conditioned on $\mathcal{G}_0(z)$,

$$\max_{0 \leq t \leq T_0 - 1}\|\mathbf{h}_t\|_2 \leq e_{\text{priv}} := \frac{\eta_0 B_{T_0}\sigma}{n} < \frac{\eta_0}{2}. \tag{17}$$

By (17) and the convexity of $\widehat{Q}_{\varpi}(\cdot)$, for each $t = 0, 1, \ldots, T_0 - 1$ we have

$$\begin{aligned} \widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t+1)}) &\leq \widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)}) + \langle\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\rangle + f_u\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_{\Sigma}^2 \\ &= \widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)}) - \langle\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)}), \eta_0\Sigma^{-1}\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)}) + \Sigma^{-1/2}\mathbf{h}_t\rangle \\ &\quad + f_u\|\eta_0\Sigma^{-1}\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)}) + \Sigma^{-1/2}\mathbf{h}_t\|_{\Sigma}^2 \\ &\leq \widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)}) - \eta_0(1 - f_u\eta_0)\|\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)})\|_{\Sigma^{-1}}^2 \\ &\quad + (1 + 2f_u\eta_0)\|\mathbf{h}_t\|_2\|\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)})\|_{\Sigma^{-1}} + f_u\|\mathbf{h}_t\|_2^2 \\ &\leq \widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)}) - \frac{\eta_0}{2}\|\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)})\|_{\Sigma^{-1}}^2 + (2\bar{\tau}B + f_u e_{\text{priv}})e_{\text{priv}} \qquad (18) \\ &\leq \widehat{Q}_{\varpi}(\widehat{\boldsymbol{\beta}}_{\varpi}) + \underbrace{\langle\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_{\varpi}\rangle - \frac{\eta_0}{2}\|\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)})\|_{\Sigma^{-1}}^2}_{=:\,\Pi_t} + (2\bar{\tau}B + f_u e_{\text{priv}})e_{\text{priv}}. \end{aligned}$$

Moreover,

$$\begin{aligned} \Pi_t &= \frac{1}{2\eta_0}\left\{2\eta_0\langle\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_{\varpi}\rangle - \eta_0^2\|\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)})\|_{\Sigma^{-1}}^2\right\} \\ &= \frac{1}{2\eta_0}\left\{\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_{\varpi}\|_{\Sigma}^2 - \|\boldsymbol{\beta}^{(t)} - \eta_0\Sigma^{-1}\nabla\widehat{Q}_{\varpi}(\boldsymbol{\beta}^{(t)}) - \widehat{\boldsymbol{\beta}}\|_{\Sigma}^2\right\} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2\eta_0}\Big\{\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2 - \|\boldsymbol{\beta}^{(t+1)} + \Sigma^{-1/2}\mathbf{h}_t - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2\Big\} \\
&= \frac{1}{2\eta_0}\Big\{\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2 - \|\boldsymbol{\beta}^{(t+1)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2 - \|\mathbf{h}_t\|_2^2 - 2\langle \mathbf{h}_t, \Sigma^{1/2}(\boldsymbol{\beta}^{(t+1)} - \widehat{\boldsymbol{\beta}}_\varpi)\rangle\Big\} \\
&\le \frac{1}{2\eta_0}\Big\{\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2 - \|\boldsymbol{\beta}^{(t+1)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2\Big\} + \frac{1}{\eta_0}\|\mathbf{h}_t\|_2\|\boldsymbol{\beta}^{(t+1)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma.
\end{aligned}
$$

Summing over $t = 0, 1, \ldots, T_0 - 1$ gives

$$
\begin{aligned}
&\sum_{t=0}^{T_0-1}\{\widehat{Q}_\varpi(\boldsymbol{\beta}^{(t+1)}) - \widehat{Q}_\varpi(\widehat{\boldsymbol{\beta}}_\varpi)\} \\
&\le \frac{1}{2\eta_0}\Big\{\|\boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2 - \|\boldsymbol{\beta}^{(T_0)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2\Big\} \\
&\quad + \frac{1}{\eta_0}\sum_{t=0}^{T_0-1}\|\boldsymbol{\beta}^{(t+1)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma \cdot \|\mathbf{h}_t\|_2 + T_0(2\bar{\tau}B + f_u e_{\mathrm{priv}})e_{\mathrm{priv}},
\end{aligned}
$$

On the other hand, applying (18) repeatedly yields

$$
T_0\widehat{Q}_\varpi(\boldsymbol{\beta}^{(T_0)}) \le \sum_{t=1}^{T_0}\widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}) + \frac{T_0(T_0-1)}{2}(2\bar{\tau}B + f_u e_{\mathrm{priv}})e_{\mathrm{priv}}.
$$

Combining the last two bounds with (16) and (17) gives that

$$
\begin{aligned}
\widehat{Q}_\varpi(\boldsymbol{\beta}^{(T_0)}) - \widehat{Q}_\varpi(\widehat{\boldsymbol{\beta}}_\varpi) &\le \frac{1}{T_0}\sum_{t=1}^{T_0}\{\widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}) - \widehat{Q}_\varpi(\widehat{\boldsymbol{\beta}}_\varpi)\} + \frac{T_0-1}{2}(2\bar{\tau}B + f_u e_{\mathrm{priv}})e_{\mathrm{priv}} \\
&\le \frac{1}{2T_0\eta_0}\|\boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma^2 + \frac{2R_0 e_{\mathrm{priv}}}{\eta_0} + (T_0+1)(\bar{\tau}B + f_u e_{\mathrm{priv}}/2)e_{\mathrm{priv}} \\
&< \frac{R_0^2}{2T_0\eta_0} + \frac{2R_0 B_{T_0}\sigma}{n} + (\bar{\tau}B + 1/4)(T_0+1)\eta_0\frac{B_{T_0}\sigma}{n}.
\end{aligned}
$$

Under condition (22) on (T_0, n), it follows that

$$
\widehat{Q}_\varpi(\boldsymbol{\beta}^{(T_0)}) < \widehat{Q}_\varpi(\widehat{\boldsymbol{\beta}}_\varpi) + \Delta \le \widehat{Q}_\varpi(\boldsymbol{\beta}^*) + \Delta.
$$

Finally, define $\widehat{\Delta}_1 = \inf_{\boldsymbol{\beta}\notin\Theta_\Sigma^*(1)}\widehat{Q}_\varpi(\boldsymbol{\beta}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^*)$, so that any $\boldsymbol{\beta}$ such that $\widehat{Q}_\varpi(\boldsymbol{\beta}) < \widehat{Q}_\varpi(\boldsymbol{\beta}^*) + \widehat{\Delta}_1$ must satisfy $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\Sigma \le 1$. Recall that $\widehat{\boldsymbol{\beta}}_\varpi = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\widehat{Q}_\varpi(\boldsymbol{\beta}) \in \Theta_\Sigma^*(r_0) \subseteq \Theta_\Sigma^*(1)$. By the convexity of $\widehat{Q}_\varpi(\cdot)$, the infimum over $\mathbb{R}^p \setminus \Theta_\Sigma^*(1)$ must be achieved on the boundary $\partial\Theta_\Sigma^*(1)$. On the other hand, Proposition (3) ensures that conditioned on $\mathcal{E}_1(\delta_0, \delta_1)$, $\widehat{Q}_\varpi(\boldsymbol{\beta}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^*) \ge \Delta = \phi_1 - \delta_0 - b^*$ for any $\boldsymbol{\beta} \in \partial\Theta_\Sigma^*(1)$, thus implying

$$
\widehat{\Delta}_1 \ge \Delta > \widehat{Q}_\varpi(\boldsymbol{\beta}^{(T_0)}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^*).
$$

Therefore, we must have $\|\boldsymbol{\beta}^{(T_0)} - \boldsymbol{\beta}^*\|_\Sigma \le 1$, as claimed. □

Appendix C: Proof of Propositions

C.1. Proof of Proposition 3

PROOF OF (15). For $\boldsymbol{\beta} \in \mathbb{R}^p$, write $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$, and note that $\widehat{D}_\varpi(\cdot)$ given in (11) satisfies

$$
\begin{aligned}
\widehat{D}_\varpi(\boldsymbol{\delta}) &= \langle \nabla Q_\varpi(\boldsymbol{\beta}^*), \boldsymbol{\delta}\rangle + R_\varpi(\boldsymbol{\delta}) - \{D_\varpi(\boldsymbol{\delta}) - \widehat{D}_\varpi(\boldsymbol{\delta})\} \\
&\ge R_\varpi(\boldsymbol{\delta}) - \|\nabla Q_\varpi(\boldsymbol{\beta}^*)\|_{\Sigma^{-1}} \cdot \|\boldsymbol{\delta}\|_\Sigma - \{D_\varpi(\boldsymbol{\delta}) - \widehat{D}_\varpi(\boldsymbol{\delta})\} \\
&= R_\varpi(\boldsymbol{\delta}) - b^*\|\boldsymbol{\delta}\|_\Sigma - \{D_\varpi(\boldsymbol{\delta}) - \widehat{D}_\varpi(\boldsymbol{\delta})\}. \qquad (19)
\end{aligned}
$$

By Lemma 1, $R_\varpi(\boldsymbol{\delta}) \geq \phi_1 \|\boldsymbol{\delta}\|_\Sigma^2$ for all $\boldsymbol{\delta} \in \Theta_\Sigma(1)$; and conditioned on $\mathcal{E}_1$, $D_\varpi(\boldsymbol{\delta}) - \widehat{D}_\varpi(\boldsymbol{\delta}) \leq \delta_0 \|\boldsymbol{\delta}\|_\Sigma$ for all $\delta \in \Theta_\Sigma(1) \setminus \Theta_\Sigma(1/n)$. Substituting these two bounds into (19) proves the first part of (15).

To prove the second part of (15), consider the decomposition

$$\widehat{Q}_\varpi(\boldsymbol{\beta}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^*) = \widehat{R}_\varpi(\boldsymbol{\delta}) + \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle,$$

where $\widehat{R}_\varpi(\boldsymbol{\delta}) = \widehat{D}_\varpi(\boldsymbol{\delta}) - \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle$. By the mean value theorem,

$$\widehat{R}_\varpi(\boldsymbol{\delta}) = \int_0^1 \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^* + t\boldsymbol{\delta}) - \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle \mathrm{d}t.$$

By Lemma 8 in Loh and Wainwright (2015) and the convexity of $\widehat{Q}_\varpi(\cdot)$, we have

$$\langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^* + t\boldsymbol{\delta}) - \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^*), t\boldsymbol{\delta} \rangle \geq \frac{1}{s} \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^* + t \cdot s\boldsymbol{\delta}) - \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^*), t \cdot s\boldsymbol{\delta} \rangle$$

for any $s \in (0, 1]$, hence implying $\widehat{R}_\varpi(\boldsymbol{\delta}) \geq s^{-1} \widehat{R}_\varpi(s\boldsymbol{\delta})$. Consequently, for any $\boldsymbol{\beta} \in \Theta_\Sigma^*(1)^{\mathrm{c}}$, we set $s = 1/\|\boldsymbol{\delta}\|_\Sigma \in (0, 1)$, $\boldsymbol{\delta}_1 = \boldsymbol{\delta}/\|\boldsymbol{\delta}\|_\Sigma \in \partial\Theta_\Sigma(1)$, and obtain that

$$\begin{aligned}\widehat{Q}_\varpi(\boldsymbol{\beta}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^*) &\geq \|\boldsymbol{\delta}\|_\Sigma \cdot \widehat{R}_\varpi(\boldsymbol{\delta}_1) + \|\boldsymbol{\delta}\|_\Sigma \cdot \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^*), \boldsymbol{\delta}_1 \rangle \\ &= \|\boldsymbol{\delta}\|_\Sigma \cdot \big\{ \widehat{Q}_\varpi(\boldsymbol{\beta}^* + \boldsymbol{\delta}_1) - \widehat{Q}_\varpi(\boldsymbol{\beta}^*) \big\}.\end{aligned}$$

Combining this with the earlier lower bound when $\boldsymbol{\beta} \in \Theta_\Sigma^*(1)$ proves the second part of (15).

Proof of (16). The goal is to bound $\widehat{B}_\varpi(\boldsymbol{\delta})$ defined in (2) from below uniformly over $\boldsymbol{\delta}$ in a compact set. Note that

$$\begin{aligned}\widehat{B}_\varpi(\boldsymbol{\delta}) &= -\widehat{D}_\varpi(\boldsymbol{\delta}) + \langle \nabla Q_\varpi(\boldsymbol{\beta}), \boldsymbol{\delta} \rangle + \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}) - \nabla Q_\varpi(\boldsymbol{\beta}), \boldsymbol{\delta} \rangle \\ &= \underbrace{-D_\varpi(\boldsymbol{\delta}) + \langle \nabla Q_\varpi(\boldsymbol{\beta}), \boldsymbol{\delta} \rangle}_{= B_\varpi(\boldsymbol{\delta})} - \{ \widehat{D}_\varpi(\boldsymbol{\delta}) - D_\varpi(\boldsymbol{\delta}) \} + \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}) - \nabla Q_\varpi(\boldsymbol{\beta}), \boldsymbol{\delta} \rangle \\ &\geq B_\varpi(\boldsymbol{\delta}) - \{ \widehat{D}_\varpi(\boldsymbol{\delta}) - D_\varpi(\boldsymbol{\delta}) \} - \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}) - \nabla Q_\varpi(\boldsymbol{\beta})\|_{\Sigma^{-1}} \cdot \|\boldsymbol{\delta}\|_\Sigma.\end{aligned}$$

Again, from Lemma 1 we see that $B_\varpi(\boldsymbol{\delta}) \geq \phi_1 \|\boldsymbol{\delta}\|_\Sigma^2$ for all $\boldsymbol{\delta} \in \Theta_\Sigma(1)$. Conditioned on $\mathcal{E}_1 = \mathcal{E}_1(\delta_0, \delta_1)$, the following bounds

$$\widehat{D}_\varpi(\boldsymbol{\delta}) - D_\varpi(\boldsymbol{\delta}) \leq \delta_0 \|\boldsymbol{\delta}\|_\Sigma \quad \text{and} \quad \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}) - \nabla Q_\varpi(\boldsymbol{\beta})\|_{\Sigma^{-1}} \leq \delta_1$$

hold uniformly over $\boldsymbol{\delta} \in \Theta_\Sigma(1) \setminus \Theta_\Sigma(1/n)$ and $\boldsymbol{\beta} \in \Theta_\Sigma^*(1)$, respectively. Putting together the pieces yields (16).

Proof of (17). For $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \Theta_\Sigma^*(2)$, applying a second-order Taylor series expansion yields

$$\widehat{Q}_\varpi(\boldsymbol{\beta}_2) - \widehat{Q}_\varpi(\boldsymbol{\beta}_1) - \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}_1), \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1 \rangle = \frac{1}{2} (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)^{\mathrm{T}} \nabla^2 \widehat{Q}_\varpi(t\boldsymbol{\beta}_1 + (1-t)\boldsymbol{\beta}_2)(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)$$

for some $t \in [0, 1]$. Since $t\boldsymbol{\beta}_1 + (1-t)\boldsymbol{\beta}_2 \in \Theta_\Sigma^*(2)$, the right-hand side is further bounded by

$$\frac{1}{2} \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|_\Sigma^2 \cdot \sup_{\boldsymbol{\beta} \in \Theta_\Sigma^*(2)} \|\Sigma^{-1/2} \nabla^2 \widehat{Q}_\varpi(\boldsymbol{\beta}) \Sigma^{-1/2}\|_2 = \frac{1}{2} \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|_\Sigma^2 \cdot \sup_{\boldsymbol{\beta} \in \Theta_\Sigma^*(2)} \|\nabla^2 \widehat{Q}_\varpi(\boldsymbol{\beta})\|_{\Sigma^{-1}}.$$

By the triangle inequality,

$$\sup_{\boldsymbol{\beta} \in \Theta_\Sigma^*(2)} \|\nabla^2 \widehat{Q}_\varpi(\boldsymbol{\beta})\|_{\Sigma^{-1}} \leq \underbrace{\sup_{\boldsymbol{\beta} \in \Theta_\Sigma^*(2)} \|\nabla^2 \widehat{Q}_\varpi(\boldsymbol{\beta}) - \nabla^2 Q_\varpi(\boldsymbol{\beta})\|_{\Sigma^{-1}}}_{\leq f_u \text{ conditioned on } \mathcal{E}_2(2)} + \sup_{\boldsymbol{\beta} \in \Theta_\Sigma^*(2)} \|\nabla^2 Q_\varpi(\boldsymbol{\beta})\|_{\Sigma^{-1}}.$$

For the population Hessian, note that $\Sigma^{-1/2} \nabla^2 Q_\varpi(\boldsymbol{\beta}) \Sigma^{-1/2} = \mathrm{E}\{K_\varpi(\varepsilon - \mathbf{x}^{\mathrm{T}}\boldsymbol{\delta}) \mathbf{w}\mathbf{w}^{\mathrm{T}}\}$, where $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ and $\mathrm{E}\{K_\varpi(\varepsilon - \mathbf{x}^{\mathrm{T}}\boldsymbol{\delta}) | \mathbf{x}\} = \int_{-\infty}^{\infty} K(v) f_{\varepsilon|\mathbf{x}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\delta} + hv) \mathrm{d}v$. Hence, it follows from Condition 3 that $\|\nabla^2 Q_\varpi(\boldsymbol{\beta})\|_{\Sigma^{-1}} \leq f_u$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$. Combining this with the above two bounds proves (17). □

C.2. Proof of Proposition 4

The claimed bounds follow immediately from Lemmas 2 and 3. ☐

C.3. Proof of Proposition 5

For each pair $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ of parameters, it follows from a second-order Taylor series expansion that

$$
\begin{aligned}
&\widehat{Q}_\varpi(\boldsymbol{\beta}_1) - \widehat{Q}_\varpi(\boldsymbol{\beta}_2) - \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle \\
&= \frac{1}{2} \boldsymbol{\delta}^{\mathrm{T}} \nabla \widehat{Q}_\varpi\big((1-t)\boldsymbol{\beta}_1 + t\boldsymbol{\beta}_2\big)\boldsymbol{\delta} = \frac{1}{2n}\sum_{i=1}^n K_\varpi\big(\varepsilon_i - \langle \mathbf{x}_i, (1-t)\boldsymbol{\beta}_1 + t\boldsymbol{\beta}_2 - \boldsymbol{\beta}^* \rangle\big)\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle^2 \\
&= \frac{1}{2n}\sum_{i=1}^n K_\varpi\big(\varepsilon_i - t\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle - \langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle\big)\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle^2
\end{aligned}
$$

for some $t \in [0,1]$, where $\boldsymbol{\delta} = \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$. For each i, define the event

$$
\mathcal{F}_i = \big\{|\varepsilon_i| \le \varpi/4\big\} \cap \big\{|\langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle| \le \varpi/4\big\} \cap \big\{|\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle| \le \|\boldsymbol{\delta}\|_\Sigma \cdot \varpi/(2r)\big\},
$$

on which $|t\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle + \langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle| \le \varpi/2 + \varpi/4 = 3\varpi/4$ for all $\boldsymbol{\beta}_2 \in \boldsymbol{\beta}_1 + \Theta_\Sigma(r)$. Consequently.

$$
\widehat{Q}_\varpi(\boldsymbol{\beta}_1) - \widehat{Q}_\varpi(\boldsymbol{\beta}_2) - \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle \ge \frac{\kappa_l}{2n\varpi}\sum_{i=1}^n \langle \mathbf{x}_i, \boldsymbol{\delta}\rangle^2 \mathbb{1}(\mathcal{F}_i),
$$

where $\kappa_l = \min_{|u|\le 1} K(u)$. For $R > 0$, define functions $\varphi_R(u) = u^2\mathbb{1}(|u| \le R/2) + \{u\,\mathrm{sign}(u) - R\}^2\mathbb{1}(R/2 < |u| \le R)$ and $\psi_R(u) = \mathbb{1}(|u| \le R/2) + \{2 - (2/R)\,\mathrm{sign}(u)\}\mathbb{1}(R/2 < |u| \le R)$, which are smoothed versions of $u \mapsto u^2\mathbb{1}(|u| \le R)$ and $u \mapsto \mathbb{1}(|u| \le R)$. Moreover, note that $\varphi_{cR}(cu) = c^2\varphi_R(u)$ for any $c > 0$ and $\varphi_0(u) = 0$. The right-hand side of the above inequality can be further bounded from below by

$$
\begin{aligned}
&\frac{\kappa_l}{2n\varpi}\sum_{i=1}^n \mathbb{1}\big(|\varepsilon_i| \le \varpi/4\big) \cdot \varphi_{\|\boldsymbol{\delta}\|_\Sigma \cdot \varpi/(2r)}\big(\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle\big) \cdot \psi_{\varpi/4}\big(\langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle\big) \\
&= \kappa_l \|\boldsymbol{\delta}\|_\Sigma^2 \cdot \frac{1}{2}\underbrace{\frac{1}{n\varpi}\sum_{i=1}^n \mathbb{1}\big(|\varepsilon_i| \le \varpi/4\big) \cdot \varphi_{\varpi/(2r)}\big(\langle \mathbf{x}_i, \boldsymbol{\delta}/\|\boldsymbol{\delta}\|_\Sigma\rangle\big) \cdot \psi_{\varpi/4}\big(\langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle\big)}_{=:V_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)}. \qquad (20)
\end{aligned}
$$

In the following, we bound $\mathrm{E}V_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $V_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) - \mathrm{E}V_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, respectively. Write $\mathbf{v} = \Sigma^{1/2}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_\Sigma \in \mathbb{S}^{p-1}$, we have

$$
\begin{aligned}
&\mathrm{E}\big\{\varphi_{\varpi/(2r)}\big(\langle \mathbf{w}_i, \mathbf{v}\rangle\big) \cdot \psi_{\varpi/4}\big(\langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle\big) \cdot \mathbb{1}\big(|\varepsilon_i| \le \varpi/4\big)\big\} \\
&\ge \frac{1}{2} f_l' \varpi \cdot \mathrm{E}\big\{\langle \mathbf{w}_i, \mathbf{v}\rangle^2 \mathbb{1}\big(|\langle \mathbf{w}_i, \mathbf{v}\rangle| \le \varpi/(4r)\big) \cdot \psi_{\varpi/4}\big(\langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle\big)\big\} \\
&\ge \frac{1}{2} f_l' \varpi \cdot \Big\{1 - \mathrm{E}\langle \mathbf{w}_i, \mathbf{v}\rangle^2 \mathbb{1}\big(|\langle \mathbf{w}_i, \mathbf{v}\rangle| > \varpi/(4r)\big) - \mathrm{E}\langle \mathbf{w}_i, \mathbf{v}\rangle^2 \mathbb{1}\big(|\langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle| > \varpi/8\big)\Big\} \\
&\ge \frac{1}{2} f_l' \varpi \cdot \Big\{1 - m_4^{1/2} P\big(|\langle \mathbf{w}_i, \mathbf{v}\rangle| > \varpi/(4r)\big)^{1/2} - m_4^{1/2} P\big(|\langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle| > \varpi/8\big)^{1/2}\Big\}.
\end{aligned}
$$

Under Condition 2 with $\upsilon_1 \ge 1$, it can be shown that for any $\mathbf{v} \in \mathbb{S}^{p-1}$ and $\boldsymbol{\beta}_1 \in \Theta^*_\Sigma(r/2)$,

$$
P\big(|\langle \mathbf{w}_i, \mathbf{v}\rangle| > \varpi/(4r)\big) \le 2\exp\left\{\frac{1}{2} - \frac{1}{2}\frac{\varpi^2}{(4\upsilon_1 r)^2}\right\} \quad \text{and}
$$

$$
P\big(|\langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle| > \varpi/8\big) \le 2\exp\left\{\frac{1}{2} - \frac{1}{2}\frac{\varpi^2}{(4\upsilon_1 r)^2}\right\}.
$$

Let $\varpi/r \geq 16(m_4 \vee 3)^{1/4}\upsilon_1$. It then follows from a numerical calculation that

$$\frac{1}{4}\Big\{1 - m_4^{1/2} P\big(|\langle \mathbf{w}_i, \mathbf{v}\rangle| > \varpi/(4r)\big)^{1/2} - m_4^{1/2} P\big(|\langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle| > \varpi/8\big)^{1/2}\Big\} \geq c_0 \approx 0.248$$

holds uniformly over $\mathbf{v} \in \mathbb{S}^{p-1}$ and $\boldsymbol{\beta}_1 \in \boldsymbol{\beta}^* + \Theta_\Sigma(r)$. Putting together the pieces yields

$$\inf_{\boldsymbol{\beta}_1 \in \boldsymbol{\beta}^* + \Theta_\Sigma(r/2),\, \boldsymbol{\beta}_2 \in \boldsymbol{\beta}_1 + \Theta_\Sigma(r)} \frac{1}{2} \mathrm{E} V_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \geq c_0 f_l'. \tag{21}$$

Next we upper bound the supremum

$$\Omega(r) := \sup_{\boldsymbol{\beta}_1 \in \boldsymbol{\beta}^* + \Theta_\Sigma(r/2),\, \boldsymbol{\beta}_2 \in \boldsymbol{\beta}_1 + \Theta_\Sigma(r)} \big\{ - V_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) + \mathrm{E} V_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\big\}.$$

Write $V_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (1/n)\sum_{i=1}^n v_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, where $v_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \varpi^{-1} \mathbb{1}\big(|\varepsilon_i| \leq \varpi/4\big) \cdot \varphi_{\varpi/(2r)}(\langle \mathbf{w}_i, \mathbf{v}\rangle) \cdot \psi_{\varpi/4}(\langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle)$ satisfies

$$0 \leq v_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \leq \frac{\varpi}{(4r)^2} \quad \text{and} \quad \mathrm{E} v_i^2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \leq \frac{f_u m_4}{2\varpi}.$$

Applying Theorem 7.3 in Bousquet (2003), a refined version of Talagrand's inequality, we obtain that for any $t > 0$,

$$\Omega(r) \leq \mathrm{E}\Omega(r) + \big\{\mathrm{E}\Omega(r)\big\}^{1/2} \frac{1}{2r}\sqrt{\frac{\varpi t}{n}} + (f_u m_4)^{1/2}\sqrt{\frac{t}{n\varpi}} + \frac{\varpi}{(4r)^2}\frac{t}{3n} \tag{22}$$

holds with probability at least $1 - e^{-t}$. To bound the expectation $\mathrm{E}\Omega(r)$, using Rademacher symmetrization and Lemma 4.5 in Ledoux and Talagrand (1991), we obtain

$$\mathrm{E}\Omega(r) \leq 2 \cdot \sqrt{\frac{\pi}{2}} \cdot \mathrm{E}\bigg(\sup_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \mathbb{G}_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2}\bigg), \tag{23}$$

where the supremum is taken over $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \Theta_\Sigma^*(r/2) \times \{\boldsymbol{\beta}_1 + \Theta_\Sigma(r)\}$—including $(\boldsymbol{\beta}^*, \boldsymbol{\beta}^*)$,

$$\mathbb{G}_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} = \frac{1}{n\varpi}\sum_{i=1}^n g_i \cdot \chi_i \cdot \varphi_{\varpi/(2r)}(\langle \mathbf{w}_i, \mathbf{v}\rangle) \cdot \psi_{\varpi/4}(\langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle) \quad \text{with} \quad \chi_i = \mathbb{1}\big(|\varepsilon_i| \leq \varpi/4\big),$$

and $g_1, \ldots, g_n$ are independent standard normal variables. Conditional on $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, $\mathbb{G}_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2}$ is a centered Gaussian process and $\mathbb{G}_{\boldsymbol{\beta}^*, \boldsymbol{\beta}^*} = 0$. For any two admissible pairs $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$, write $\mathbf{v} = \Sigma^{1/2}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)/\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|_\Sigma$, $\mathbf{v}' = \Sigma^{1/2}(\boldsymbol{\beta}_2' - \boldsymbol{\beta}_1')/\|\boldsymbol{\beta}_2' - \boldsymbol{\beta}_1'\|_\Sigma$, and note that

$$\begin{aligned}
\mathbb{G}_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} - \mathbb{G}_{\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'} &= \mathbb{G}_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} - \mathbb{G}_{\boldsymbol{\beta}_1', \boldsymbol{\beta}_1' + \boldsymbol{\delta}} + \mathbb{G}_{\boldsymbol{\beta}_1', \boldsymbol{\beta}_1' + \boldsymbol{\delta}} - \mathbb{G}_{\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'} \\
&= \frac{1}{n\varpi}\sum_{i=1}^n g_i \cdot \chi_i \cdot \varphi_{\varpi/(2r)}(\langle \mathbf{w}_i, \mathbf{v}\rangle)\big\{\psi_{\varpi/4}(\langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\rangle) - \psi_{\varpi/4}(\langle \mathbf{x}_i, \boldsymbol{\beta}_1' - \boldsymbol{\beta}^*\rangle)\big\} \\
&\quad + \frac{1}{n\varpi}\sum_{i=1}^n g_i \cdot \chi_i \cdot \psi_{\varpi/4}(\langle \mathbf{x}_i, \boldsymbol{\beta}_1' - \boldsymbol{\beta}^*\rangle)\big\{\varphi_{\varpi/(2r)}(\langle \mathbf{w}_i, \mathbf{v}\rangle) - \varphi_{\varpi/(2r)}(\langle \mathbf{w}_i, \mathbf{v}'\rangle)\big\}.
\end{aligned}$$

Recall that φ_R and ψ_R are, respectively, R- and $(2/R)$-Lipschitz continuous, and $\varphi_R(u) \leq (R/2)^2$. It follows that

$$\begin{aligned}
&\mathrm{E}^*\big(\mathbb{G}_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} - \mathbb{G}_{\boldsymbol{\beta}_1', \boldsymbol{\beta}_1' + \boldsymbol{\delta}}\big)^2 \\
&\leq \frac{1}{(n\varpi)^2}\left(\frac{8}{\varpi}\right)^2\left(\frac{\varpi}{4r}\right)^4 \sum_{i=1}^n \chi_i \langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_1'\rangle^2 = \frac{1}{4r^4 n^2}\sum_{i=1}^n \chi_i \langle \mathbf{x}_i, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_1'\rangle^2
\end{aligned}$$

and

$$\begin{aligned}&\mathrm{E}^*\big(\mathbb{G}_{\boldsymbol{\beta}_1',\boldsymbol{\beta}_1'+\boldsymbol{\delta}}-\mathbb{G}_{\boldsymbol{\beta}_1',\boldsymbol{\beta}_2'}\big)^2\\&\le\frac{1}{(n\varpi)^2}\bigg(\frac{\varpi}{2r}\bigg)^2\sum_{i=1}^n\chi_i\big(\langle\mathbf{w}_i,\mathbf{v}\rangle-\langle\mathbf{w}_i,\mathbf{v}'\rangle\big)^2=\frac{1}{4r^2n^2}\sum_{i=1}^n\chi_i\langle\mathbf{w}_i,\mathbf{v}-\mathbf{v}'\rangle^2.\end{aligned}$$

Together, the last two displays imply

$$\mathrm{E}^*\big(\mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}-\mathbb{G}_{\boldsymbol{\beta}_1',\boldsymbol{\beta}_2'}\big)^2\le\frac{1}{2r^4n^2}\sum_{i=1}^n\chi_i\langle\mathbf{x}_i,\boldsymbol{\beta}_1-\boldsymbol{\beta}_1'\rangle^2+\frac{1}{2r^2n^2}\sum_{i=1}^n\chi_i\langle\mathbf{w}_i,\mathbf{v}-\mathbf{v}'\rangle^2.$$

Define another (conditional) Gaussian process $\{\mathbb{Z}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\}$ as

$$\mathbb{Z}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}=\frac{1}{2^{1/2}r^2n}\sum_{i=1}^n g_i'\chi_i\langle\mathbf{x}_i,\boldsymbol{\beta}_1-\boldsymbol{\beta}^*\rangle+\frac{1}{2^{1/2}rn}\sum_{i=1}^n g_i''\chi_i\frac{\langle\mathbf{x}_i,\boldsymbol{\beta}_2-\boldsymbol{\beta}_1\rangle}{\|\boldsymbol{\beta}_2-\boldsymbol{\beta}_1\|_\Sigma},$$

where $\{g_i'\}$ and $\{g_i''\}$ are two dependent copies of $\{g_i\}$. The above calculations show that

$$\mathrm{E}^*\big(\mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}-\mathbb{G}_{\boldsymbol{\beta}_1',\boldsymbol{\beta}_2'}\big)^2\le\mathrm{E}^*\big(\mathbb{Z}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}-\mathbb{Z}_{\boldsymbol{\beta}_1',\boldsymbol{\beta}_2'}\big)^2.$$

Using Sudakov-Fernique's Gaussian comparison inequality (see, e.g. Theorem 7.2.11 in Vershynin (2018)) gives

$$\mathrm{E}^*\bigg(\sup_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\bigg)\le\mathrm{E}^*\bigg(\sup_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\mathbb{Z}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\bigg).\tag{24}$$

The same bound also applies to unconditional expectations of the two suprema. For the random process $\{\mathbb{Z}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\}$, using the bound $P(|\varepsilon_i|\le\varpi/4|\mathbf{x}_i)\le f_u\varpi/2$ (a.s.) we derive that

$$\begin{aligned}&\mathrm{E}\bigg(\sup_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\mathbb{Z}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\bigg)\\&\le\frac{1}{2^{1/2}r^2}\sup_{\boldsymbol{\beta}_1\in\Theta^*_\Sigma(r/2)}\|\boldsymbol{\beta}_1-\boldsymbol{\beta}^*\|_\Sigma\cdot\mathrm{E}\bigg\|\frac{1}{n}\sum_{i=1}^n g_i'\chi_i\mathbf{w}_i\bigg\|_2+\frac{1}{2^{1/2}r}\mathrm{E}\bigg\|\frac{1}{n}\sum_{i=1}^n g_i''\chi_i\mathbf{w}_i\bigg\|_2\\&\le f_u^{1/2}\frac{1}{4r}\sqrt{\frac{\varpi p}{n}}+f_u^{1/2}\frac{1}{2r}\sqrt{\frac{\varpi p}{n}}=f_u^{1/2}\frac{3}{4r}\sqrt{\frac{\varpi p}{n}}.\end{aligned}$$

This, joint with (22), (23) and (24), implies that, with probability at least $1-e^{-t}$,

$$\begin{aligned}\Omega(r)&\le\frac{5}{4}\mathrm{E}\Omega(r)+(f_um_4)^{1/2}\sqrt{\frac{t}{n\varpi}}+(4+1/3)\frac{\varpi t}{(4r)^2n}\\&\le(f_um_4)^{1/2}\sqrt{\frac{t}{n\varpi}}+15(\pi/2)^{1/2}f_u^{1/2}\frac{1}{8r}\sqrt{\frac{\varpi p}{n}}+(4+1/3)\frac{\varpi t}{16r^2n}.\end{aligned}\tag{25}$$

Finally, combining (20), (21) and (25), we conclude that for all $\boldsymbol{\beta}_1\in\Theta^*_\Sigma(r/2)$ and $\boldsymbol{\beta}_2\in\boldsymbol{\beta}_1+\Theta_\Sigma(r)$ with $r=\varpi/(16\max\{m_4,3\}^{1/4}\upsilon_1)$,

$$\begin{aligned}&\frac{\widehat{Q}_h(\boldsymbol{\beta}_1)-\widehat{Q}_h(\boldsymbol{\beta}_2)-\langle\nabla\widehat{Q}_h(\boldsymbol{\beta}_2),\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\rangle}{\kappa_l\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_\Sigma^2}\\&\qquad\ge c_0f_l'-\frac{1}{2}(f_um_4)^{1/2}\sqrt{\frac{t}{n\varpi}}-C_1f_u^{1/2}m_4^{1/4}\upsilon_1\sqrt{\frac{p}{n\varpi}}-C_2m_4^{1/2}\upsilon_1^2\frac{t}{n\varpi}\end{aligned}$$

holds with probability at least $1-e^{-t}$, where $C_1,C_2>0$ are absolute constants. Let $n\varpi$ be sufficiently large—$n\varpi\gtrsim m_4^{1/2}\upsilon_1^2(p+t)$—so that the right-hand side is bounded from below by $c_0f_l'/2$. This completes the proof. □

Appendix D: Proofs of Technical Lemmas

D.1. Proof of Lemma 1

For simplicity, define the re-scaled kernel $W(u) = K_\varpi(u) = K(u/\varpi)/\varpi$, which is also symmetric and non-negative. Note that $\rho_\tau(u) = |u|/2 + (\tau - 1/2)u$ and $(\rho_\tau * W)(u) = (1/2)\int_{-\infty}^{\infty} |v| W(u-v)\mathrm{d}v + (\tau - 1/2)u$. It thus suffices to prove that for any $u \in \mathbb{R}$,

$$|u| \le \int_{-\infty}^{\infty} |v| W(u-v)\mathrm{d}v \le |u| + \kappa_1 \varpi.$$

The upper bound follows from the triangle inequality and a change of variable that

$$\begin{aligned}\int_{-\infty}^{\infty} |v| W(u-v)\mathrm{d}v &\le \int_{-\infty}^{\infty} |u-v| W(u-v)\mathrm{d}v + \int_{-\infty}^{\infty} |u| W(u-v)\mathrm{d}v \\ &= |u| + \frac{1}{\varpi}\int_{-\infty}^{\infty} |u-v| K\left(\frac{u-v}{\varpi}\right)\mathrm{d}v = |u| + \varpi \int_{-\infty}^{\infty} |s| K(s)\mathrm{d}s = |u| + \kappa_1 \varpi.\end{aligned}$$

To prove the lower bound, we first assume $u \ge 0$. Utilizing the change of variable and the properties $\int W(u)\mathrm{d}u = 1$ and $\int uW(u)\mathrm{d}u = 0$, we derive that

$$\begin{aligned}&\int_{-\infty}^{\infty} |v| W(u-v)\mathrm{d}v = \int_{-\infty}^{\infty} |u+w| W(w)\mathrm{d}w \\ &= \int_{-u}^{\infty} (u+w)W(w)\mathrm{d}w - \int_{-\infty}^{-u} (u+w)W(w)\mathrm{d}w \\ &= \int_{-u}^{\infty} (u+w)W(w)\mathrm{d}w + \int_{u}^{\infty} (-u+w)W(w)\mathrm{d}w \\ &= u\int_{-u}^{u} W(w)\mathrm{d}w + \int_{-u}^{\infty} wW(w)\mathrm{d}w + \int_{u}^{\infty} wW(w)\mathrm{d}w \\ &= u - \int_{u}^{\infty} uW(w)\mathrm{d}w - \int_{-\infty}^{-u} uW(w)\mathrm{d}w + \underbrace{\int_{-u}^{\infty} wW(w)\mathrm{d}w}_{= -\int_{-\infty}^{-u} wW(w)\mathrm{d}w \text{ by symmetry}} + \int_{u}^{\infty} wW(w)\mathrm{d}w \\ &= u + \underbrace{\int_{u}^{\infty} (w-u)W(w)\mathrm{d}w}_{\ge 0} - \underbrace{\int_{-\infty}^{-u} (w+u)W(w)\mathrm{d}w}_{\le 0} \ge u.\end{aligned}$$

This proves the lower bound when $u \ge 0$. The case where $u < 0$ can be proven using a similar argument, and therefore we omit the details for brevity. $\square$

D.2. Proof of Lemma 1

Starting with $R_\varpi(\boldsymbol{\delta})$, using a second-order Taylor series expansion yields

$$R_\varpi(\boldsymbol{\delta}) = \frac{1}{2}\mathrm{E}\big\{K_\varpi(\varepsilon - t\langle \mathbf{w}, \mathbf{v}\rangle)\langle \mathbf{w}, \mathbf{v}\rangle^2\big\} \quad \text{for some } t \in [0,1],$$

where $\mathbf{w} = \Sigma^{-1/2}\mathbf{x}$ and $\mathbf{v} = \Sigma^{1/2}\boldsymbol{\delta} = \Sigma^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$. By a change of variable and the Lipschitz continuity of $f_{\varepsilon|\mathbf{x}}(\cdot)$, we have

$$\mathrm{E}\big\{K_\varpi(\varepsilon - t\langle \mathbf{w}, \mathbf{v}\rangle)|\mathbf{x}\big\} = \int_{-\infty}^{\infty} K(v) f_{\varepsilon|\mathbf{x}}(t\langle \mathbf{w}, \mathbf{v}\rangle + \varpi v)\mathrm{d}v \ge f_{\varepsilon|\mathbf{x}}(t\langle \mathbf{w}, \mathbf{v}\rangle) - l_0(\mathbf{x})\kappa_1 \varpi.$$

It follows that for any $\boldsymbol{\delta} \in \mathbb{R}^p$ satisfying $\|\boldsymbol{\delta}\|_\Sigma \le 1$ (so that $\mathbf{v} \in \mathbb{B}^p(1)$),

$$
\begin{aligned}
2R_\varpi(\boldsymbol{\delta}) &\ge \mathrm{E}\big\{ f_{\varepsilon|\mathbf{x}}(t\langle \mathbf{w},\mathbf{v}\rangle)\langle \mathbf{w},\mathbf{v}\rangle^2\big\} - l_0\kappa_1\varpi \cdot \mathrm{E}\langle \mathbf{w},\mathbf{v}\rangle^2 \\
&= \mathrm{E}\big\{ f_{\varepsilon|\mathbf{x}}(\underbrace{t\|\mathbf{v}\|_2}_{\in[0,1]}\langle \mathbf{w},\mathbf{v}/\|\mathbf{v}\|_2\rangle)\langle \mathbf{w},\mathbf{v}/\|\mathbf{v}\|_2\rangle^2\big\} \cdot \|\mathbf{v}\|_2^2 - l_0\kappa_1\varpi \cdot \|\mathbf{v}\|_2^2 \\
&\ge (f_l - l_0\kappa_1\varpi)\|\mathbf{v}\|_2^2,
\end{aligned}
$$

where the second inequality follows from Condition 3. The same argument also applies to $B_\varpi(\boldsymbol{\delta})$. We thus omit the details. □

D.3. Proof of Lemma 2

We only need to derive an upper bound for $\widehat{D}_\varpi(\boldsymbol{\delta}) - D_\varpi(\boldsymbol{\delta})$ uniformly over $\boldsymbol{\delta} \in \mathbb{R}^p$ in a compact subset. The same argument applies to $D_\varpi(\boldsymbol{\delta}) - \widehat{D}_\varpi(\boldsymbol{\delta})$. For each sample $\mathbf{z}_i = (\mathbf{w}_i, \varepsilon_i)$ with $\mathbf{w}_i = \Sigma^{-1/2}\mathbf{x}_i$, define $d_\varpi(\mathbf{v};\mathbf{z}_i) = \ell_\varpi(\varepsilon_i - \langle \mathbf{w}_i,\mathbf{v}\rangle) - \ell_\varpi(\varepsilon_i)$, so that $\widehat{D}_\varpi(\boldsymbol{\delta}) = (1/n)\sum_{i=1}^n d_\varpi(\mathbf{v};\mathbf{z}_i)$ for $\mathbf{v} = \Sigma^{1/2}\boldsymbol{\delta}$. Note that $\ell_\varpi(\cdot)$ is continuously differentiable with $|\ell'_\varpi(u)| \le \bar{\tau} = \max(\tau, 1-\tau)$ for all u. Hence, for any $\mathbf{z}_i$ and $\mathbf{v},\mathbf{v}' \in \mathbb{R}^p$, $|d_\varpi(\mathbf{v};\mathbf{z}_i) - d_\varpi(\mathbf{v}';\mathbf{z}_i)| \le \bar{\tau}|\langle \mathbf{w}_i,\mathbf{v}\rangle - \langle \mathbf{w}_i,\mathbf{v}'\rangle|$. In other words, $d_\varpi(\mathbf{v};\mathbf{z}_i)$ is $\bar{\tau}$-Lipschitz continuous in $\langle \mathbf{w}_i,\mathbf{v}\rangle$.

For any fixed $r > 0$ and some $\epsilon \in (0,1)$ to be determined, define the random variable

$$
\Delta_\epsilon(r) = \frac{1-\epsilon}{2\bar{\tau}r} n^{1/2} \sup_{\mathbf{v}\in\mathbb{B}^p(r)} \big\{\widehat{D}_\varpi(\mathbf{v}) - D_\varpi(\mathbf{v})\big\} = \frac{1-\epsilon}{2\bar{\tau}r} \sup_{\mathbf{v}\in\mathbb{B}^p(r)} \bigg\{ \frac{1}{n^{1/2}} \sum_{i=1}^n (1-\mathrm{E}) d_\varpi(\mathbf{v};\mathbf{z}_i)\bigg\}.
$$

By Chernoff's inequality, for any $z \ge 0$,

$$
P\{\Delta_\epsilon(r) \ge z\} \le \exp\bigg[-\sup_{\lambda \ge 0}\big\{\lambda z - \log \mathrm{E} e^{\lambda \Delta_\epsilon(r)}\big\}\bigg]. \tag{26}
$$

To control the moment generating function $\mathrm{E}e^{\lambda\Delta_\epsilon(r)}$, by Rademacher symmetrization we have

$$
\mathrm{E}e^{\lambda\Delta_\epsilon(r)} \le \mathrm{E}\exp\bigg\{(1-\epsilon)\frac{2\lambda}{2\bar{\tau}r} \sup_{\mathbf{v}\in\mathbb{B}^p(r)} \frac{1}{n^{1/2}}\sum_{i=1}^n e_i d_\varpi(\mathbf{v};\mathbf{z}_i)\bigg\},
$$

where $e_1, \ldots, e_n$ are independent Rademacher random variables. Recall that $d_\varpi(\mathbf{v};\mathbf{z}_i)$ is $\bar{\tau}$-Lipschitz continuous in $\langle \mathbf{w}_i,\mathbf{v}\rangle$, and $d_\varpi(\mathbf{v};\mathbf{z}_i) = 0$ if $\langle \mathbf{w}_i,\mathbf{v}\rangle = 0$. Then, applying the Ledoux-Talagrand contraction inequality (see Theorem 4.12 and inequality (4.20) in Ledoux and Talagrand (1991)) yields

$$
\begin{aligned}
&\mathrm{E}\exp\bigg\{(1-\epsilon)\frac{\lambda}{\bar{\tau}r} \sup_{\mathbf{v}\in\mathbb{B}^p(r)} \frac{1}{n^{1/2}}\sum_{i=1}^n e_i d_\varpi(\mathbf{v};\mathbf{z}_i)\bigg\} \\
&\le \mathrm{E}\exp\bigg\{(1-\epsilon)\frac{\lambda}{r} \sup_{\mathbf{v}\in\mathbb{B}^p(r)} \frac{1}{n^{1/2}}\sum_{i=1}^n e_i \langle \mathbf{w}_i,\mathbf{v}\rangle\bigg\} \le \mathrm{E}\exp\bigg\{(1-\epsilon)\lambda\bigg\|\frac{1}{n^{1/2}}\sum_{i=1}^n e_i\mathbf{w}_i\bigg\|_2\bigg\}.
\end{aligned}
$$

For this $\epsilon \in (0,1)$, there exists an ϵ-net $\{\mathbf{u}_1, \ldots, \mathbf{u}_{N_\epsilon}\}$ of $\mathbb{S}^{p-1}$ with cardinality $N_\epsilon \le (1+2/\epsilon)^p$ such that $\|\sum_{i=1}^n e_i\mathbf{w}_i\|_2 \le (1-\epsilon)^{-1}\max_{1\le j\le N_\epsilon}\sum_{i=1}^n e_i\mathbf{u}_j^{\mathrm{T}}\mathbf{w}_i$. Therefore,

$$
\mathrm{E}\exp\bigg\{(1-\epsilon)\lambda\bigg\|\frac{1}{n^{1/2}}\sum_{i=1}^n e_i\mathbf{w}_i\bigg\|_2\bigg\} \le \sum_{j=1}^{N_\epsilon}\mathrm{E}\exp\bigg(\frac{\lambda}{n^{1/2}}\sum_{i=1}^n e_i\mathbf{u}_j^{\mathrm{T}}\mathbf{w}_i\bigg).
$$

Write $S_j = n^{-1/2}\sum_{i=1}^n e_i\mathbf{u}_j^{\mathrm{T}}\mathbf{w}_i$, which is a sum of zero-mean random variables. Note that $e_i \in \{-1,1\}$ is symmetric, and Condition 2 ensures that $\log \mathrm{E}\exp(ce_i\mathbf{u}_j^{\mathrm{T}}\mathbf{w}_i) \le c^2\upsilon_1^2/2$ for all $c \in \mathbb{R}$. Consequently,

$$
\mathrm{E}\exp(\lambda S_j) = \prod_{i=1}^n \mathrm{E}\exp(\lambda n^{-1/2}e_i\mathbf{u}_j^{\mathrm{T}}\mathbf{w}_i) \le \prod_{i=1}^n e^{\lambda^2\upsilon_1^2/(2n)} = e^{\lambda^2\upsilon_1^2/2},
$$

from which it follows that

$$\log \mathrm{E} e^{\lambda \Delta_\epsilon(r)} \le \log N_\epsilon + \frac{1}{2}\upsilon_1^2 \lambda^2.$$

For any $u \ge 0$, note that

$$\sup_{\lambda \ge 0} \big\{ \lambda z - \log \mathrm{E} e^{\lambda \Delta_\epsilon(r)} \big\} \ge -\log N_\epsilon + \sup_{\lambda \ge 0} \bigg(\lambda z - \frac{1}{2}\upsilon_1^2 \lambda^2 \bigg) = -\log N_\epsilon + z^2/(2\upsilon_1^2).$$

Substituting this into (26) and taking $v = z^2/(2\upsilon_1^2)$, we obtain that with probability at least $1 - \exp\{p\log(1 + 2/\epsilon) - v\}$,

$$\sup_{\mathbf{v} \in \mathbb{B}^p(r)} \big\{ \widehat{D}_\varpi(\mathbf{v}) - D_\varpi(\mathbf{v}) \big\} \le \frac{2\bar{\tau}\upsilon_1}{1-\epsilon} r \sqrt{\frac{2v}{n}}. \tag{27}$$

This proves (3) by setting $\epsilon = 2/(e^4 - 1)$ and $v = 4p + u$.

Via a peeling/slicing argument, next, we prove a uniform version of (27), which holds for all $\mathbf{v} \in \mathbb{B}^p(r_l, r_u) = \{\mathbf{v} \in \mathbb{R}^p : r_l \le \|\mathbf{v}\|_2 \le r_u\}$. For some $\gamma > 1$ to be determined, and positive integers $k = 1, \ldots, N := \lceil \log(\frac{r_u}{r_l})/\log(\gamma) \rceil$, define the sets $\Theta_k = \{\mathbf{v} \in \mathbb{R}^p : \gamma^{k-1} r_l \le \|\mathbf{v}\|_2 \le \gamma^k r_l\}$, so that $\mathbb{B}^p(r_l, r_u) \subseteq \cup_{k=1}^N \Theta_k$. Then,

$$\begin{aligned}
&P\bigg\{ \exists \mathbf{v} \in \mathbb{B}^p(r_l, r_u) \text{ s.t. } \widehat{D}_\varpi(\mathbf{v}) - D_\varpi(\mathbf{v}) > \frac{2\sqrt{2}\gamma}{1-\epsilon} \bar{\tau}\upsilon_1 \|\mathbf{v}\|_2 \sqrt{\frac{v}{n}} \bigg\} \\
&\le \sum_{k=1}^N P\bigg\{ \exists \mathbf{v} \in \Theta_k \text{ s.t. } \widehat{D}_\varpi(\mathbf{v}) - D_\varpi(\mathbf{v}) > \frac{2\sqrt{2}\gamma}{1-\epsilon} \bar{\tau}\upsilon_1 \gamma^{k-1} r_l \sqrt{\frac{v}{n}} \bigg\} \\
&\le \sum_{k=1}^N P\bigg\{ \sup_{\mathbf{v} \in \mathbb{B}^p(\gamma^k r_l)} \widehat{D}_\varpi(\mathbf{v}) - D_\varpi(\mathbf{v}) > \frac{2\sqrt{2}}{1-\epsilon} \bar{\tau}\upsilon_1 \gamma^k r_l \sqrt{\frac{v}{n}} \bigg\} \\
&\overset{\text{(i)}}{\le} \sum_{k=1}^N \exp\big\{ p\log(1 + 2/\epsilon) - v \big\} \le \lceil \log(\tfrac{r_u}{r_l})/\log(\gamma) \rceil \exp\big\{ p\log(1 + 2/\epsilon) - v \big\},
\end{aligned}$$

where inequality (i) follows from (27) with $r = \gamma^k r_l$ for $k = 1, \ldots, N$. Taking $\epsilon = 2/(e^4 - 1)$, $\gamma = e^{1/e}$ and $v = 4p + \log\{e \log(\frac{r_u}{r_l})\} + u$ yields that with probability at least $1 - e^{-u}$,

$$\widehat{D}_\varpi(\mathbf{v}) - D_\varpi(\mathbf{v}) \le 4.25 \bar{\tau}\upsilon_1 \cdot \|\mathbf{v}\|_2 \sqrt{\frac{4p + \log\{e\log(r_u/r_l)\} + u}{n}}.$$

This proves (4) by taking $(r_l, r_u) = (\delta, r)$. □

D.4. Proof of Lemma 3

PROOF OF (5). By a change of variable $\mathbf{v} = \Sigma^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$, define the centered gradient process

$$G_\varpi(\mathbf{v}) = \Sigma^{-1/2}\{\widehat{Q}_\varpi(\boldsymbol{\beta}) - Q_\varpi(\boldsymbol{\beta})\} = \frac{1}{n}\sum_{i=1}^n (1 - \mathrm{E}) \underbrace{\big\{ \bar{K}_\varpi(\mathbf{w}_i^{\mathrm{T}}\mathbf{v} - \varepsilon_i) - \tau \big\}}_{=:\xi_{i,\mathbf{v}}} \mathbf{w}_i.$$

For any $\epsilon \in (0, R)$, there exists an ϵ-net $\{\mathbf{v}_1, \ldots, \mathbf{v}_{N_\epsilon}\}$ of $\mathbb{B}^p(R)$ with $N_\epsilon \le (1 + 2R/\epsilon)^p$. For any $\mathbf{v} \in \mathbb{B}^p(R)$, there exists some $1 \le j \le N_\epsilon$ such that $\|\mathbf{v} - \mathbf{v}_j\|_2 \le \epsilon$. Then, by the triangle inequality,

$$\|G_\varpi(\mathbf{v})\|_2 \le \|G_\varpi(\mathbf{v}) - G_\varpi(\mathbf{v}_j)\|_2 + \|G_\varpi(\mathbf{v}_j)\|_2.$$

We first control the approximation error $\|G_\varpi(\mathbf{v}) - G_\varpi(\mathbf{v}_j)\|_2$. Using integration by parts and a change of variable, we have $\mathrm{E}(\xi_{i,\mathbf{v}}|\mathbf{x}_i) = \int_{-\infty}^{\infty} K(u) F_{\varepsilon_i|\mathbf{x}_i}(\mathbf{w}_i^{\mathrm{T}}\mathbf{v} - hu) \mathrm{d}u$. Condition 3 ensures that

$$\|\mathrm{E}(\xi_{i,\mathbf{v}}\mathbf{w}_i) - \mathrm{E}(\xi_{i,\mathbf{v}_j}\mathbf{w}_i)\|_2 \le \sup_{\mathbf{u} \in \mathbb{S}^{p-1}} \mathrm{E}\big\{ f_u(\mathbf{x}) |\langle \mathbf{w}_i, \mathbf{u} \rangle \cdot \langle \mathbf{w}_i, \mathbf{v} - \mathbf{v}_j \rangle| \big\} \le f_u \epsilon.$$

Turning to $\xi_{i,\mathbf{v}}\mathbf{w}_i$ and $\xi_{i,\mathbf{v}_j}\mathbf{w}_i$, since $K(\cdot)$ is bounded by κ_u, we have

$$\begin{aligned}\bigg\|\frac{1}{n}\sum_{i=1}^n \xi_{i,\mathbf{v}}\mathbf{w}_i - \frac{1}{n}\sum_{i=1}^n \xi_{i,\mathbf{v}_j}\mathbf{w}_i\bigg\|_2 &\le \sup_{\mathbf{u}\in\mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n |\bar{K}_\varpi(\mathbf{w}_i^{\mathrm{T}}\mathbf{v}-\varepsilon_i) - \bar{K}_\varpi(\mathbf{w}_i^{\mathrm{T}}\mathbf{v}_j-\varepsilon_i)|\cdot|\mathbf{w}_i^{\mathrm{T}}\mathbf{u}| \\ &\le \frac{\kappa_u}{\varpi}\sup_{\mathbf{u}\in\mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n |\mathbf{w}_i^{\mathrm{T}}(\mathbf{v}-\mathbf{v}_j)|\cdot|\mathbf{w}_i^{\mathrm{T}}\mathbf{u}| \\ &\le \frac{\kappa_u\epsilon}{\varpi}\cdot\lambda_{\max}\bigg(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_i\mathbf{w}_i^{\mathrm{T}}\bigg).\end{aligned}$$

Note that $\mathbf{w}_i = (1,\mathbf{w}_{i,-}^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{w}_{i,-}\in\mathbb{R}^{p-1}$ are zero-mean sub-Gaussian random vectors. By a standard covering argument, paired with Bernstein's inequality, it can be shown that (see, e.g. Theorem 4.6.1 in Vershynin (2018)) with probability at least $1-(1/2)e^{-u}$,

$$\lambda_{\max}\bigg(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_i\mathbf{w}_i^{\mathrm{T}}\bigg) \le 1 + C_1\upsilon_1^2\bigg(\sqrt{\frac{p+u}{n}}+\frac{p+u}{n}\bigg),$$

where $C_1>0$ is an absolute constant. Provided $n\gtrsim \upsilon_1^4(p+u)$, it follows that with probability at least $1-(1/2)e^{-u}$,

$$\|G_\varpi(\mathbf{v}) - G_\varpi(\mathbf{v}_j)\|_2 \le f_u\epsilon + 2\kappa_u\epsilon/\varpi \tag{28}$$

holds uniformly over all the pairs $(\mathbf{v},\mathbf{v}_j)$ satisfying $\|\mathbf{v}-\mathbf{v}_j\|_2\le\epsilon$.

It remains to deal with $\|G_\varpi(\mathbf{v}_j)\|_2$ for each $j=1,\ldots,N_\epsilon$. For simplicity, we write $\xi_{ij}=\xi_{i,\mathbf{v}_j}=\bar{K}_\varpi(\mathbf{w}_i^{\mathrm{T}}\mathbf{v}_j-\varepsilon_i)-\tau$, and define

$$S_j = \frac{1}{n}\sum_{i=1}^n(\xi_i-\mathrm{E}\xi_i) \quad\text{and}\quad \boldsymbol{S}_j = \frac{1}{n}\sum_{i=1}^n\{\xi_{ij}\mathbf{w}_{i,-}-\mathrm{E}(\xi_{ij}\mathbf{w}_{i,-})\}\in\mathbb{R}^{p-1}.$$

Then, $\|G_\varpi(\mathbf{v}_j)\|_2^2 = S_j^2 + \|\boldsymbol{S}_j\|_2^2$. For S_j, note that $\xi_{ij}\in[-\tau,1-\tau]$. As a direct application of Hoeffding's inequality, we have that for any $v\ge 0$,

$$P\bigg(|S_j|\ge\sqrt{\frac{\log 4+v}{2n}}\bigg)\le\frac{1}{2}e^{-v}. \tag{29}$$

Next we use a covering argument to bound $\|\boldsymbol{S}_j\|_2$. For any $\delta\in(0,1)$, there exists an δ-net $\mathcal{M}_\delta$ of $\mathbb{S}^{p-2}$ (unit sphere in $\mathbb{R}^{p-1}$) with cardinality $M_\delta\le(1+2/\delta)^{p-1}$ such that

$$\|\boldsymbol{S}_j\|_2 \le (1-\delta)^{-1}\frac{2\bar{\tau}\upsilon_1}{n^{1/2}}\underbrace{\max_{\mathbf{u}\in\mathcal{M}_\delta}\frac{1}{2\bar{\tau}\upsilon_1 n^{1/2}}\sum_{i=1}^n\langle\mathbf{u},\xi_{ij}\mathbf{w}_{i,-}-\mathrm{E}(\xi_{ij}\mathbf{w}_{i,-})\rangle}_{=:\Delta_\delta},$$

where $\bar{\tau}=\max(1-\tau,\tau)$. For any $z\ge0$, applying Chernoff's inequality gives

$$P(\Delta_\delta\ge z)\le\exp\bigg\{-\sup_{\lambda\ge0}\big(\lambda z-\log\mathrm{E}e^{\lambda\Delta_\delta}\big)\bigg\}.$$

By Rademacher symmetrization and independence,

$$\mathrm{E}e^{\lambda\Delta_\delta}\le\mathrm{E}\exp\bigg(\max_{\mathbf{u}\in\mathcal{M}_\delta}\frac{\lambda}{\bar{\tau}\upsilon_1 n^{1/2}}\sum_{i=1}^n e_i\xi_{ij}\mathbf{u}^{\mathrm{T}}\mathbf{w}_{i,-}\bigg)\le\sum_{u\in\mathcal{M}_\delta}\prod_{i=1}^n\mathrm{E}\exp\bigg(\frac{\lambda}{\bar{\tau}\upsilon_1 n^{1/2}}e_i\xi_{ij}\mathbf{u}^{\mathrm{T}}\mathbf{w}_{i,-}\bigg),$$

where e_i's are independent Rademacher random variables. Fix $\mathbf{u} \in \mathcal{M}_\delta$, and set $\omega_i = \mathbf{u}^{\mathrm{T}} \mathbf{w}_{i,-}/\upsilon_1$. We further have

$$\mathrm{E}\exp\left(\frac{\lambda}{\bar{\tau} n^{1/2}} e_i \xi_{ij} \omega_i\right) = \sum_{k=0}^{\infty} \frac{1}{k!}\left(\frac{\lambda}{n^{1/2}}\right)^k \underbrace{\mathrm{E}(e_i \xi_{ij} \omega_i/\bar{\tau})^k}_{=0 \text{ if } k \text{ is odd}} \le 1 + \sum_{\ell=1}^{\infty} \frac{1}{(2\ell)!}\left(\frac{\lambda}{n^{1/2}}\right)^{2\ell} \mathrm{E}(\omega_i^{2\ell}).$$

Recall from Condition 2 that $\mathrm{E}(e^{t\omega_i}) \le e^{t^2/2}$ for all $t \in \mathbb{R}$, and $\mathrm{E}(\omega_i^2) \le 1$. For $\ell \ge 2$,

$$\mathrm{E}(\omega_i^{2\ell}) = 2\ell \int_0^\infty u^{2\ell-1} P(|\omega_i| \ge u)\mathrm{d}u \le 4\ell \int_0^\infty u^{2\ell-1} e^{-u^2/2}\mathrm{d}u = 2^{\ell+1}\ell!.$$

Substituting this into the previous exponential moment bound yields

$$\mathrm{E}\exp\left(\frac{\lambda}{\bar{\tau} n^{1/2}} e_i \xi_{ij} \omega_i\right) \le 1 + \frac{\lambda^2}{n} + \sum_{\ell=2}^{\infty} \frac{4/3}{\ell!}\left(\frac{\lambda^2}{n}\right)^{\ell} \le \exp\left(\frac{2}{\sqrt{3}n}\lambda^2\right),$$

which in turn implies

$$\log \mathrm{E} e^{\lambda \Delta_\delta} \le \log M_\delta + \frac{2}{\sqrt{3}}\lambda^2 \quad \text{and} \quad \sup_{\lambda \ge 0}\left\{\lambda z - \log \mathrm{E} e^{\lambda \Delta_\delta}\right\} \ge -\log M_\delta + \frac{\sqrt{3}}{8} z^2.$$

Combining this with Chernoff's bound and a change of variable, we obtain that with probability at least $1 - (1/2)e^{-v}$,

$$\|\boldsymbol{S}_j\|_2^2 \le \frac{32\bar{\tau}^2 \upsilon_1^2}{3^{1/2}(1-\delta)} \frac{(p-1)\log(1+2/\delta) + \log 2 + v}{n}. \tag{30}$$

Together, (29) and (30) with a properly chosen δ, say, $\delta = 2/(e^2-1)$, imply that $\|G_\varpi(\mathbf{v}_j)\|_2 \le C_2 \upsilon_1 \sqrt{(p+v)/n}$ holds with probability at least $1 - e^{-v}$, where $C_2 > 0$ is an absolute constant. Taking the union bound over $j = 1, \ldots, N_\epsilon$, we have that with probability at least $1 - \exp\{p\log(1+2R/\epsilon) - v\}$,

$$\max_{1 \le j \le N_\epsilon} \|G_\varpi(\mathbf{v}_j)\|_2 \le C_2 \upsilon_1 \sqrt{\frac{p+v}{n}}. \tag{31}$$

Taking $\epsilon = R\varpi/n$ and $v = \log(2) + p\log(1+2n/\varpi) + u$ in (28) and (31), we obtain that with probability at least $1 - e^{-u}$,

$$\sup_{\mathbf{v} \in \mathbb{B}^p(R)} \|G_\varpi(\mathbf{v})\|_2 \le (f_u \varpi + 2\kappa_u)\frac{R}{n} + C_2 \upsilon_1 \sqrt{\frac{\log 2 + p\log(e + 2en/\varpi) + u}{n}}$$

as long as $n \gtrsim \upsilon_1^4(p+u)$. This proves (5).

For the population gradient, noting that $|\bar{K}_\varpi(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \tau| \le \bar{\tau}$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$, we have $\sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\nabla Q_\varpi(\boldsymbol{\beta})\|_{\Sigma^{-1}} \le \bar{\tau} \sup_{\mathbf{u} \in \mathbb{S}^{p-1}} \mathrm{E}|\langle \mathbf{w}, \mathbf{u}\rangle| \le \bar{\tau}$. At $\boldsymbol{\beta}^*$, the bound $\|\nabla Q_\varpi(\boldsymbol{\beta}^*)\|_{\Sigma^{-1}} \le 0.5 l_0 \kappa_2 \varpi^2$ follows from (C.2) in He et al. (2023).

PROOF OF (6). For $\boldsymbol{\beta} \in \mathbb{R}^p$, define after a change of variable $\mathbf{v} = \Sigma^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$ that

$$\widehat{\mathbf{H}}_\varpi(\mathbf{v}) = \Sigma^{-1/2} \nabla^2 \widehat{Q}_\varpi(\boldsymbol{\beta}) \Sigma^{-1/2} = \frac{1}{n}\sum_{i=1}^n K_\varpi(\varepsilon_i - \langle \mathbf{w}_i, \mathbf{v}\rangle)\mathbf{w}_i \mathbf{w}_i^{\mathrm{T}} \quad \text{and} \quad \mathbf{H}_\varpi(\mathbf{v}) = \mathrm{E}\{\widehat{\mathbf{H}}_\varpi(\mathbf{v})\}.$$

To upper bound $\sup_{\mathbf{v} \in \mathbb{B}^p(R)} \|\widehat{\mathbf{H}}_\varpi(\mathbf{v}) - \mathbf{H}_\varpi(\mathbf{v})\|_2$, we use a similar discretization argument as in the proof of (5) and obtain

$$\begin{aligned} &\sup_{\mathbf{v} \in \mathbb{B}^p(R)} \|\widehat{\mathbf{H}}_\varpi(\mathbf{v}) - \mathbf{H}_\varpi(\mathbf{v})\|_2 \\ &\le \frac{l_K \epsilon}{\varpi^2} \max_{1 \le i \le n} \|\mathbf{w}_i\|_2 \cdot \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{w}_i \mathbf{w}_i^{\mathrm{T}}\right\|_2 \\ &\quad + \underbrace{\max_{1 \le j \le N_\epsilon} \|\widehat{\mathbf{H}}_\varpi(\mathbf{v}_j) - \widehat{\mathbf{H}}_\varpi(\mathbf{v}_j)\|_2}_{\text{sampling error}} + \underbrace{\max_{1 \le j \le N_\epsilon} \|\mathbf{H}_\varpi(\mathbf{v}_j) - \mathbf{H}_\varpi(\mathbf{v}_j)\|_2}_{\text{approximation error}}, \end{aligned}$$

where l_K is the Lipschitz constant of the kernel $K(\cdot)$, $\{\mathbf{v}_1,\ldots,\mathbf{v}_{N_\epsilon}\}$ is an ϵ-net of $\mathbb{B}^p(R)$ and $N_\epsilon \le (1+2R/\epsilon)^p$. For the last term on the right-hand side, the Lipschitz continuity of $f_{\varepsilon|\mathbf{x}}(\cdot)$ ensures that

$$\max_{1\le j\le N_\epsilon} \|\mathbf{H}_\varpi(\mathbf{v}_j)-\mathbf{H}_\varpi(\mathbf{v}_j)\|_2 \le l_0 m_3 \epsilon,$$

where $m_3=\sup_{\mathbf{u}\in\mathbb{S}^{p-1}} \mathrm{E}|\langle \mathbf{w},\mathbf{u}\rangle|^3$.

We have already shown that $\lambda_{\max}(n^{-1}\sum_{i=1}^n \mathbf{w}_i\mathbf{w}_i^{\mathrm{T}})\le 2$ with probability at least $1-(1/2)e^{-u}$ as long as $n\gtrsim \upsilon_1^4(p+u)$. Moreover, applying Theorem 2.1 in Hsu et al. (2012) to each $\mathbf{w}_{i,-}$ yields that with probability at least $1-e^{-z}$,

$$\|\mathbf{w}_{i,-}\|_2^2 \le \upsilon_1^2\Big\{p-1+2\sqrt{(p-1)z}+2z\Big\} \le \upsilon_1^2\Big(\sqrt{p-1}+\sqrt{2z}\Big)^2.$$

Taking the union bound over $i=1,\ldots,n$ and setting $z=\log n+u$, we obtain that with probability $1-e^{-u}$,

$$\max_{1\le i\le n}\|\Sigma^{-1/2}\mathbf{x}_i\|_2=\max_{1\le i\le n}\|\mathbf{w}_i\|_2\le 1+\upsilon_1\Big\{\sqrt{p-1}+\sqrt{2\log n+2u}\Big\}\le C\upsilon_1(p+\log n+u)^{1/2}.$$

Turning to $\max_{1\le j\le N_\epsilon}\|\widehat{\mathbf{H}}_\varpi(\mathbf{v}_j)-\widehat{\mathbf{H}}_\varpi(\mathbf{v}_j)\|_2$, it has been shown in the proof of Proposition 3.2 in He et al. (2023) (see, e.g. (C.61)) that with probability at least $1-e^{-u}$,

$$\max_{1\le j\le N_\epsilon}\|\widehat{\mathbf{H}}_\varpi(\mathbf{v}_j)-\widehat{\mathbf{H}}_\varpi(\mathbf{v}_j)\|_2\lesssim \upsilon_1^2\bigg\{\sqrt{\frac{p\log(R/\epsilon)+u}{n\varpi}}+\frac{p\log(R/\epsilon)+u}{n\varpi}\bigg\}.$$

Finally, as long as $n\gtrsim\upsilon_1^4(p+u)$, taking $\epsilon=(\varpi/n)^2R$ yields the bound

$$\begin{aligned}&\sup_{\mathbf{v}\in\mathbb{B}^p(R)}\|\widehat{\mathbf{H}}_\varpi(\mathbf{v})-\mathbf{H}_\varpi(\mathbf{v})\|_2\\ &\lesssim \upsilon_1^2\bigg\{\sqrt{\frac{p\log(n/\varpi)+u}{n\varpi}}+\frac{p\log(n/\varpi)+u}{n\varpi}\bigg\}+\big\{\upsilon_1(p+\log n+u)^{1/2}+l_0m_3\varpi^2\big\}\frac{R}{n^2}.\end{aligned}$$

This proves (6), and thus completes the proof. □

D.5. Proof of Lemma 4

For each $t=0,1,\ldots,T-1$, we apply the concentration inequality for Lipschitz functions of standard normal random variables and obtain that

$$P\big(\|\mathbf{g}_t\|_2\ge\sqrt{p}+\sqrt{2z}\big)\le e^{-z},\quad \text{valid for any } z\ge 0.$$

Combining this with the union bound (over $t=0,1,\ldots,T-1$) yields (7).

Note that $\|\mathbf{g}_t\|_2^2$ follows the chi-square distribution χ_p^2, which is a special case of the gamma distribution $\Gamma(p/2,1/2)$. The centered variable, $\|\mathbf{g}_t\|_2^2-p$, is known to be sub-gamma with parameters $v=2p$ and $c=2$ (Boucheron et al. 2013). Let $Z=\sum_{t=0}^{T-1}\rho^t(\|\mathbf{g}_t\|_2^2-p)$. For each t and $0<\lambda<1/c$,

$$\log \mathrm{E}e^{\lambda\rho^t(\|\mathbf{g}_t\|_2^2-p)}\le\frac{v\lambda^2\rho^{2t}}{2(1-c\lambda\rho^t)}\le\frac{v\lambda^2\rho^{2t}}{2(1-c\lambda)}.$$

By independence,

$$\log\mathrm{E}e^{\lambda Z}=\sum_{t=0}^{T-1}\log\mathrm{E}e^{\lambda\rho^t(\|\mathbf{g}_t\|_2^2-p)}\le\frac{v\lambda^2\sum_{t=0}^{T-1}\rho^{2t}}{2(1-c\lambda)}\le\frac{v}{1-\rho^2}\frac{\lambda^2}{2(1-c\lambda)}.$$

In other words, the centered variable Z is sub-gamma with parameters $(v/(1-\rho^2),c)=(2p/(1-\rho^2),2)$. Applying Chernoff's bound to Z (see, e.g. Section 2.4 of Boucheron et al. (2013)) yields that, for any $z>0$,

$$P\bigg(Z>2\sqrt{\frac{pz}{1-\rho^2}}+2z\bigg)\le e^{-z}.$$

This, combined with the elementary inequality $\sum_{t=0}^{T-1}\rho^t\le 1/(1-\rho)$, proves (8). □

D.6. Proof of Lemma 5

To begin with, note that conditioned on event $\mathcal{E}_0(B) \cap \mathcal{E}_1(\delta_0, \delta_1)$ with $\delta_0 + b^* < \phi_1$,

$$\sup_{\boldsymbol{\beta} \in \Theta^*_\Sigma(1)} \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta})\|_{\Sigma^{-1}} \leq \delta_1 + \bar{\tau},$$

$$\widehat{Q}_\varpi(\boldsymbol{\beta}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^*) \geq \underbrace{(\phi_1 - \delta_0 - b^*)}_{=\Delta} \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\Sigma \quad \text{for all} \quad \boldsymbol{\beta} \in \Theta^*_\Sigma(1)^{\mathrm{c}}$$

and $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_0 \Sigma^{-1} \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}) - \eta_0 \sigma \Sigma^{-1/2} \mathbf{g}_t / n$ for $t = 0, 1, \ldots, T-1$.

Now assume that $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma \leq 1$ for some $t \geq 0$. Recall that $\delta_1 < \phi_1 < 0.5 f_l$. Hence, conditioned further on $\mathcal{G}$ defined in (11), we have

$$\begin{aligned} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_\Sigma &\leq \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma + \eta_0 \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)})\|_{\Sigma^{-1}} + \frac{\eta_0 \sigma}{n} \|\mathbf{g}_t\|_2 \\ &\leq 1 + (f_l + \bar{\tau}) \eta_0 \leq 2, \end{aligned}$$

where we used the assumptions $\eta_0 \leq 1/(f_l + \bar{\tau})$ and $n \geq 2 B_T \sigma / f_l$ in the last inequality. Proceeding via proof by contradiction, suppose $\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_\Sigma > 1$ so that

$$\Delta \cdot \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_\Sigma \leq \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t+1)}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^*).$$

For the right-hand side, we have

$$\begin{aligned} &\widehat{Q}_\varpi(\boldsymbol{\beta}^{(t+1)}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^*) = \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t+1)}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}) + \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^*) \\ &\overset{\text{(i)}}{\leq} \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)} \rangle + f_u \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_\Sigma^2 - \langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)} \rangle \\ &= \frac{1}{\eta_0} \langle \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^* \rangle_\Sigma + f_u \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_\Sigma^2 + \frac{\sigma}{n} \langle \Sigma^{-1/2} \mathbf{g}_t, \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^* \rangle \\ &= \frac{1}{2\eta_0} \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma^2 - \frac{1}{2\eta_0} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_\Sigma^2 - \frac{1}{2\eta_0} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_\Sigma^2 \\ &\quad + f_u \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_\Sigma^2 + \frac{\sigma}{n} \langle \Sigma^{-1/2} \mathbf{g}_t, \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^* \rangle \\ &\overset{\text{(ii)}}{\leq} \frac{1}{2\eta_0} \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma^2 - \frac{1}{2\eta_0} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_\Sigma^2 + \frac{\sigma}{n} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_\Sigma \cdot \|\mathbf{g}_t\|_2 \\ &\overset{\text{(iii)}}{\leq} \frac{1}{2\eta_0} \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma^2 - \frac{1}{2\eta_0} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_\Sigma^2 + \frac{B_T \sigma}{n} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_\Sigma, \end{aligned}$$

where inequality (i) follows from the restricted smoothness property (17), inequality (ii) holds if $\eta_0 \leq 1/(2 f_u)$, and inequality (iii) uses conditioning on $\mathcal{G}$. Provided that $B_T \sigma / n \leq \Delta$, combining the above lower and upper bounds on $\widehat{Q}_\varpi(\boldsymbol{\beta}^{(t+1)}) - \widehat{Q}_\varpi(\boldsymbol{\beta}^*)$ yields

$$\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_\Sigma^2 \leq \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma^2 \leq 1,$$

which leads to a contradiction. Therefore, starting from an initial value $\boldsymbol{\beta}^{(0)} \in \Theta^*(1)$, and conditioning on event $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{G}$ with properly chosen parameters, we must have $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\Sigma \leq 1$ for all $t = 1, \ldots, T$. □

D.7. Proof of Lemma 6

Recall that conditioning on $\mathcal{E}_0 \cap \mathcal{E}_1$, $\widehat{\boldsymbol{\beta}}_\varpi \in \Theta^*_\Sigma(r_0)$ with $r_0 = (\delta_0 + b^*)/\phi_1$, $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_0 \Sigma^{-1} \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)}) + \eta_0 \sigma \Sigma^{-1/2} \mathbf{g}_t / n$ and $\|\nabla^2 \widehat{Q}_\varpi(\boldsymbol{\beta})\|_{\Sigma^{-1}} \le 2 f_u$ for all $\boldsymbol{\beta} \in \Theta^*_\Sigma(R)$. The upper bound on the Hessian also implies

$$\langle \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}_1) - \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle \ge \frac{1}{2 f_u} \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}_1) - \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}_2)\|^2_{\Sigma^{-1}}, \quad \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \Theta^*_\Sigma(R).$$

Provided $\eta_0 \le 1/(2 f_u)$ and $R \ge R_0 + r_0$, applying this bound with $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{\beta}^{(0)}, \widehat{\boldsymbol{\beta}}_\varpi)$ we obtain

$$\begin{aligned}
\|\boldsymbol{\beta}^{(1)} - \widehat{\boldsymbol{\beta}}_\varpi\|^2_\Sigma &= \|\boldsymbol{\beta}^{(0)} - \eta_0 \Sigma^{-1} \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(0)}) + \eta_0 \sigma \Sigma^{-1/2} \mathbf{g}_0 / n - \widehat{\boldsymbol{\beta}}_\varpi\|^2_\Sigma \\
&= \|\boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}_\varpi\|^2_\Sigma + \eta_0^2 \|\Sigma^{-1/2} \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(0)}) - \sigma \mathbf{g}_0 / n\|_2^2 \\
&\quad - 2\eta_0 \big\langle \boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}_\varpi, \nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(0)}) - \sigma \Sigma^{1/2} \mathbf{g}_0 / n \big\rangle \\
&\le \|\boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}_\varpi\|^2_\Sigma + \eta_0^2 \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(0)})\|^2_{\Sigma^{-1}} + \left(\frac{\eta_0 \sigma}{n}\right)^2 \|\mathbf{g}_0\|_2^2 + 2 \frac{\eta_0^2 \sigma}{n} \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(0)})\|_{\Sigma^{-1}} \|\mathbf{g}_0\|_2 \\
&\quad - \frac{\eta_0}{f_u} \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(0)})\|^2_{\Sigma^{-1}} + 2 \frac{\eta_0 \sigma}{n} \|\boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma \|\mathbf{g}_0\|_2 \\
&\le \|\boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}_\varpi\|^2_\Sigma - \frac{\eta_0}{2 f_u} \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(0)})\|^2_{\Sigma^{-1}} + 2(R_0 + \eta_0 \bar{\tau} B) \frac{\eta_0 \sigma}{n} \|\mathbf{g}_0\|_2 + \left(\frac{\eta_0 \sigma}{n}\right)^2 \|\mathbf{g}_0\|_2^2.
\end{aligned}$$

For any given $T_0 \ge 1$, write $R_t = \|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi\|_\Sigma$ for $t = 1, \ldots, T_0$. Provided $R \ge \max_{0 \le t \le T_0 - 1} R_t + r_0$, it can similarly shown that for any $0 \le t \le T_0 - 1$,

$$\begin{aligned}
&\|\boldsymbol{\beta}^{(t+1)} - \widehat{\boldsymbol{\beta}}_\varpi\|^2_\Sigma \\
&\le \|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}_\varpi\|^2_\Sigma - \frac{\eta_0}{2 f_u} \|\nabla \widehat{Q}_\varpi(\boldsymbol{\beta}^{(t)})\|^2_{\Sigma^{-1}} + 2(R_t + \eta_0 \bar{\tau} B) \frac{\eta_0 \sigma}{n} \|\mathbf{g}_t\|_2 + \left(\frac{\eta_0 \sigma}{n}\right)^2 \|\mathbf{g}_t\|_2^2.
\end{aligned}$$

Given $z \ge 0$, by Lemma 4 we have that with probability at least $1 - e^{-z}$,

$$\max_{0 \le t \le T_0 - 1} \|\mathbf{g}_t\|_2 \le B_{T_0} := \sqrt{p} + \sqrt{2(\log T_0 + z)}.$$

Define $e_{\rm priv} = \eta_0 B_{T_0} \sigma / n$. For some $\epsilon \in (0,1)$ to be determined, the above recursive bound implies

$$R_{t+1}^2 \le (1+\epsilon) R_t^2 + (1 + 1/\epsilon) e_{\rm priv}^2 + C_0 e_{\rm priv}, \quad t = 0, 1, \ldots, T_0 - 1$$

and hence

$$\begin{aligned}
R_t^2 &\le (1+\epsilon)^t R_0^2 + \big\{(1+1/\epsilon) e_{\rm priv}^2 + C_0 e_{\rm priv}\big\} \sum_{k=0}^{t-1} (1+\epsilon)^k \\
&\le (1+\epsilon)^t R_0^2 + \frac{(1+\epsilon)^t - 1}{\epsilon} \big\{(1+1/\epsilon) e_{\rm priv}^2 + C_0 e_{\rm priv}\big\}, \quad t = 1, \ldots, T_0,
\end{aligned}$$

where $C_0 = 2\eta_0 \bar{\tau} B$. Provided $T_0 \ge 2$ and

$$n \ge \frac{e-1}{4-e} (2\bar{\tau} B + 1/2) \max\{1, \eta_0 / R_0\}^2 T_0 B_{T_0} \sigma > 2(T_0 + 1) B_{T_0} \sigma,$$

taking $\epsilon = 1/T_0 \in (0,1)$ we obtain that

$$\begin{aligned}
R_t^2 &\le e R_0^2 + (e-1) \big\{(T_0 + 1) e_{\rm priv} + C_0\big\} T_0 e_{\rm priv} \\
&\le e R_0^2 + (e-1)(2\bar{\tau} B + 1/2) \frac{T_0 B_{T_0} \sigma}{n} \eta_0^2 \le e R_0^2 + (4-e) R_0^2 = 4 R_0^2
\end{aligned}$$

for all $t = 1, \ldots, T_0$, as claimed. $\square$

References

Boucheron S, Lugosi G, Massart P (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence* (Oxford University Press).

Bousquet O (2003) Concentration inequalities for sub-additive functions using the entropy method. *Stochastic Inequalities and Applications*, 213–247 (Birkhäuser, Basel).

Cai TT, Wang Y, Zhang L (2021) The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics* 49(5):2825–2850.

He X, Pan X, Tan KM, Zhou WX (2023) Smoothed quantile regression with large-scale inference. *Journal of Econometrics* 232(2):367–388.

Hsu D, Kakade S, Zhang T (2012) A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability* 17(52):1–6.

Ledoux M, Talagrand M (1991) *Probability in Banach Spaces: Isoperimetry and Processes* (Springer Berlin, Heidelberg).

Loh PL, Wainwright MJ (2015) Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research* 16(19):559–616.

Vershynin R (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press).