Appendix

To enhance clarity, we begin by summarizing the notations used throughout the main text and the appendix. For every integer $k \geq 1$, let $\mathbb{R}^k$ denote the k-dimensional Euclidean space, and write $[k] := \{1, \ldots, k\}$. For any vector $\mathbf{u} \in \mathbb{R}^k$ and integer $q \geq 1$, we denote its ℓ_q-norm by $\|\mathbf{u}\|_q$ and its support by $\mathrm{supp}(\mathbf{u})$. The (pseudo) ℓ_0-norm of $\mathbf{u} = (u_1, \ldots, u_k)^\top$ is defined as $\|\mathbf{u}\|_0 = \sum_{i=1}^k \mathbb{1}(u_i \neq 0)$, where $\mathbb{1}(\cdot)$ denotes the indicator function. We define the unit sphere and the closed ℓ_2-ball of radius r in $\mathbb{R}^k$ as

$$\mathbb{S}^{k-1} := \{\mathbf{u} \in \mathbb{R}^k : \|\mathbf{u}\|_2 = 1\}\,, \quad \text{and} \quad \mathbb{B}^k(r) := \{\mathbf{u} \in \mathbb{R}^k : \|\mathbf{u}\|_2 \leq r\}\,.$$

For a subset $S \subseteq [k]$ with cardinality $|S|$ and a k-dimensional vector $\mathbf{u} \in \mathbb{R}^k$, let $\mathbf{u}_S \in \mathbb{R}^{|S|}$ denote the subvector of $\mathbf{u}$ consisting of the coordinates indexed by S, and let $\mathbf{u}_{|S} \in \mathbb{R}^k$ denote the vector obtained from $\mathbf{u}$ by setting all coordinates outside S to zero. The inner product of any two vectors $\mathbf{u} = (u_1, \ldots, u_k)^\top$ and $\mathbf{v} = (v_1, \ldots, v_k)^\top$ is given by $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^k u_i v_i$.

For a symmetric matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$, we denote its smallest and largest eigenvalues by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$, respectively, and its spectral norm by $\|\mathbf{A}\|$. The vector norm induced by a positive semi-definite matrix $\mathbf{A}$ is defined as $\|\mathbf{u}\|_{\mathbf{A}} := \|\mathbf{A}^{1/2}\mathbf{u}\|_2$, $\mathbf{u} \in \mathbb{R}^k$. For two sequences of non-negative numbers $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, we write $a_n \lesssim b_n$ or equivalently $b_n \gtrsim a_n$ if there exists a constant $C > 0$, independent of n, such that $a_n \leq C b_n$. We write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold. For $x \in \mathbb{R}$, we use the notations $\lceil x \rceil = \inf\{y \in \mathbb{Z} : y \geq x\}$, $\lfloor x \rfloor = \sup\{y \in \mathbb{Z} : y \leq x\}$ and $x_+ = \max\{x, 0\}$. For any given smoothing parameter $\varpi > 0$, we define $\bar{K}_\varpi(\cdot) := \bar{K}(\cdot/\varpi)$. Finally, for any algorithm $\mathcal{M}$ mapping a dataset $\mathbf{Z}$ to $\mathbb{R}^p$, the ℓ_q-sensitivity (for $q \geq 1$) of $\mathcal{M}$ is defined as

$$\mathrm{sens}_q(\mathcal{M}) := \sup_{\mathbf{Z}, \mathbf{Z}'} \|\mathcal{M}(\mathbf{Z}) - \mathcal{M}(\mathbf{Z}')\|_q\,, \tag{S.1}$$

where the supremum is taken over all pairs of neighboring datasets $\mathbf{Z}$ and $\mathbf{Z}'$.

EC.1. Basic Mechanisms and Properties of (ϵ, δ)-DP

(ϵ, δ)-DP, as characterized in Definition 1, is particularly appealing due to its simplicity and flexibility in the design of private algorithms. In practice, such algorithms are typically constructed by adding random noise to the output of a non-private procedure. Among the various noise distributions, Gaussian and Laplace noise are the most commonly used. In addition, the post-processing and composition properties provide a foundation for designing multi-step DP algorithms. We review these basic mechanisms and properties of (ϵ, δ)-DP below. Recall that the ℓ_q-sensitivity of an algorithm $\mathcal{M}$ is defined in (S.1).

LEMMA L1 **(Gaussian mechanism)**. *Assume* $\mathrm{sens}_2(\mathcal{M}) \le B$ *for some* $B > 0$*. Then the mechanism* $\mathcal{M}(\mathbf{Z}) + \mathbf{g}$*, with* $\mathbf{g} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_p)$ *and* $\sigma = B\epsilon^{-1}\sqrt{2\log(1.25/\delta)}$*, is* (ϵ, δ)*-DP.*

Unlike the Gaussian mechanism, whose noise calibration is governed by the ℓ_2-sensitivity of the underlying algorithm, the Laplace mechanism is based on ℓ_1-sensitivity and provides a guarantee of pure DP.

LEMMA L2 **(Laplace mechanism)**. *Assume* $\mathrm{sens}_1(\mathcal{M}) \le B$ *for some* $B > 0$*. Let* $\mathbf{w} = (w_1, \ldots, w_p)^\top \in \mathbb{R}^p$ *have i.i.d. entries drawn from* $\mathrm{Laplace}(B/\epsilon)$*. Then* $\mathcal{M}(\mathbf{Z}) + \mathbf{w}$ *is* $(\epsilon, 0)$*-DP.*

LEMMA L3 **(Post-processing property)**. *Let* $\mathcal{M}_1$ *be an* (ϵ, δ)*-DP algorithm, and let* $\mathcal{M}_2$ *be an arbitrary deterministic mapping that takes the output of* $\mathcal{M}_1$ *as input. Then* $\mathcal{M}_2(\mathcal{M}_1(\mathbf{Z}))$ *is* (ϵ, δ)*-DP.*

Intuitively, the post-processing property implies that no external procedure can reverse the effects of privatization. The following composition properties describe how the privacy parameters evolve under repeated composition.

LEMMA L4 **(Composition property)**. *Let* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *be* (ϵ_1, δ_1)*-DP and* (ϵ_2, δ_2)*-DP algorithms, respectively. The composition* $\mathcal{M}_1 \circ \mathcal{M}_2$ *is* $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$*-DP.*

LEMMA L5 **(Advanced composition property)**. *If* T *algorithms are run sequentially, each satisfying* (ϵ, δ)*-DP, then for any* $\delta' > 0$*, the combined procedure is* $(\epsilon\sqrt{2T\log(1/\delta')} + T\epsilon(e^\epsilon - 1), T\delta + \delta')$*-DP.*

EC.2. Beyond Centralized Privacy

EC.2.1. Local Differentially Private Sparsity-Constrained ERM

Our proposed Algorithm 3 is developed under a centralized differential privacy (CDP) setting, which assumes a trusted central server with access to users' unperturbed data and enforces privacy only at the point of release. When this trust assumption is untenable, for instance, due to concerns about data storage, misuse, or security breaches, one may instead adopt local differential privacy (LDP) (Kasiviswanathan et al. 2011, Duchi et al. 2013), under which each user privatizes their information before communication. We begin by formally defining LDP.

DEFINITION EC.1. A randomized algorithm $\mathcal{M}:\mathcal{X}\to\mathcal{R}$ satisfies (ϵ,δ)-LDP if, for every pair of individual inputs $x,x'\in\mathcal{X}$ and every measurable subset $S\subseteq\mathcal{R}$, $\mathbb{P}\{\mathcal{M}(x)\in S\}\le e^{\epsilon}\cdot\mathbb{P}\{\mathcal{M}(x')\in S\}+\delta$.

Under LDP, privacy is enforced at the individual level: for any two possible values $x,x'\in\mathcal{X}$ that a single user may hold, the output distributions $\mathcal{M}(x)$ and $\mathcal{M}(x')$ must remain close for every measurable set S. In contrast, CDP (Definition 1) is defined with respect to neighboring datasets that differ in a single individual's record. LDP also satisfies the following parallel and sequential composition properties.

LEMMA L6. *Suppose n mechanisms $\mathcal{M}_1,\ldots,\mathcal{M}_n$ each satisfy (ϵ_i,δ_i)-LDP and are applied to disjoint subsets of the data. Then the joint mechanism $(\mathcal{M}_1,\ldots,\mathcal{M}_n)$ is $(\max_{i\in[n]}\epsilon_i,\max_{i\in[n]}\delta_i)$-LDP.*

LEMMA L7. *Suppose n mechanisms $\mathcal{M}_1,\ldots,\mathcal{M}_n$ each satisfy (ϵ_i,δ_i)-LDP and are applied sequentially to the same data. Then the combined mechanism $(\mathcal{M}_1,\ldots,\mathcal{M}_n)$ is $(\sum_{i=1}^n\epsilon_i,\sum_{i=1}^n\delta_i)$-LDP.*

As in the CDP setting, the noise magnitude required under LDP is determined by the sensitivity of the underlying local mapping. For a mapping $\mathcal{M}:\mathcal{X}\to\mathbb{R}^p$, we define its ℓ_q-sensitivity under LDP as

$$\widetilde{\mathrm{sens}}_q(\mathcal{M})=\sup_{x,x'\in\mathcal{X}}\|\mathcal{M}(x)-\mathcal{M}(x')\|_q\,.$$

We next state the standard Gaussian mechanism under LDP.

LEMMA L8. *Assume $\widetilde{\mathrm{sens}}_2(\mathcal{M})\le B$ for some $B>0$. Then the mechanism $\mathcal{M}(x)+\mathbf{g}$, with $\mathbf{g}\sim N(\mathbf{0},\sigma^2\mathbf{I}_p)$ and $\sigma=B\epsilon^{-1}\sqrt{2\log(1.25/\delta)}$, is (ϵ,δ)-LDP.*

In what follows, we present a locally differentially private, sparsity-constrained ERM procedure that integrates the above LDP notions with an explicit sparsity constraint. Let $\mathbf{Z}=\{\mathbf{z}_i=(d_i,\mathbf{x}_i)\}_{i=1}^n$ and define

$$g(\mathbf{z}_i,\boldsymbol{\beta})=\big\{\bar{K}\big((\mathbf{x}_i^\top\boldsymbol{\beta}-d_i)/\varpi\big)-\tau\big\}w_\gamma(\mathbf{x}_i)\,. \tag{S.2}$$

Starting from a common initialization $\boldsymbol{\beta}^{(0)}$, our CDP-based procedure (Algorithm 3) performs two steps at each iteration. First, we form an intermediate update using the gradients $\{g(\mathbf{z}_i,\boldsymbol{\beta}^{(t)})\}_{i=1}^n$: $\tilde{\boldsymbol{\beta}}_{\mathrm{cdp}}^{(t+1)}=\boldsymbol{\beta}_{\mathrm{cdp}}^{(t)}-\eta_0 n^{-1}\sum_{i=1}^n g(\mathbf{z}_i,\boldsymbol{\beta}_{\mathrm{cdp}}^{(t)})$. We then apply NoisyHT (Algorithm 2) to $\tilde{\boldsymbol{\beta}}_{\mathrm{cdp}}^{(t+1)}$ to obtain a sparse, differentially private update $\boldsymbol{\beta}_{\mathrm{cdp}}^{(t+1)}$. This design is justified by the bound $\sup_{\mathbf{z}\in\mathbb{R}\times\mathbb{R}^p,\boldsymbol{\beta}\in\mathbb{R}^p}\|g(\mathbf{z},\boldsymbol{\beta})\|_\infty \le \gamma\bar{\tau}$, which controls the sensitivity of the intermediate update and enables NoisyHT to enforce sparsity while preserving differential privacy.

Relative to the CDP update, the LDP counterpart differs primarily in that privatization is performed at the user level prior to aggregation. Since the ℓ_∞-sensitivity of g in (S.2) is bounded by $2\gamma\bar{\tau}$, under the LDP framework we construct $\boldsymbol{\beta}_{\mathrm{ldp}}^{(t+1)}$ via gradient perturbation, whereby each user locally perturbs $g(\mathbf{z}_i,\boldsymbol{\beta}_{\mathrm{ldp}}^{(t)})$. However, because $g(\mathbf{z}_i,\boldsymbol{\beta}_{\mathrm{ldp}}^{(t)})\in\mathbb{R}^p$, naively adding noise to all coordinates typically leads to a noise magnitude that scales polynomially with p, which in turn yields estimation error bounds with polynomial dependence on p; see, e.g., Wang and Xu (2021). To mitigate this issue, we restrict the perturbation to a prescribed subset of coordinates. In particular, let $S_0=\mathrm{supp}(\boldsymbol{\beta}^{(0)})$ and $s=|S_0|$. The ℓ_2-sensitivity of $\{g(\mathbf{z}_i,\boldsymbol{\beta}_{\mathrm{ldp}}^{(t)})\}_{S_0}$ is then bounded by $B_{\mathrm{ldp}}=2\gamma\bar{\tau}\sqrt{s}$. Accordingly, for each $i\in[n]$, we release $\tilde{g}(\mathbf{z}_i,\boldsymbol{\beta}_{\mathrm{ldp}}^{(t)})\in\mathbb{R}^p$ such that

$$\{\tilde{g}(\mathbf{z}_i,\boldsymbol{\beta}_{\mathrm{ldp}}^{(t)})\}_{S_0}=\{g(\mathbf{z}_i,\boldsymbol{\beta}_{\mathrm{ldp}}^{(t)})\}_{S_0}+\boldsymbol{\zeta}_i^t \ \text{ and } \ \{\tilde{g}(\mathbf{z}_i,\boldsymbol{\beta}_{\mathrm{ldp}}^{(t)})\}_{[p]\setminus S_0}=\mathbf{0}\,, \tag{S.3}$$

where $\boldsymbol{\zeta}_i^t\sim N(\mathbf{0},\sigma_{\mathrm{ldp}}^2\mathbf{I}_s)$ with $\sigma_{\mathrm{ldp}}=B_{\mathrm{ldp}}T\epsilon^{-1}\sqrt{2\log(1.25T/\delta)}$. By Lemma L8, the local release $\tilde{g}(\mathbf{z}_i,\boldsymbol{\beta}_{\mathrm{ldp}}^{(t)})$ is $(\epsilon T^{-1},\delta T^{-1})$-LDP for each fixed iteration. Since the n users perturb their gradients independently and in parallel within iteration t, the joint release $\{\tilde{g}(\mathbf{z}_i,\boldsymbol{\beta}_{\mathrm{ldp}}^{(t)})\}_{i=1}^n$ is also $(\epsilon T^{-1},\delta T^{-1})$-LDP by Lemma L6. Based on this construction, the $(t+1)$-th update under LDP is given by

$$\boldsymbol{\beta}_{\mathrm{ldp}}^{(t+1)}=\boldsymbol{\beta}_{\mathrm{ldp}}^{(t)}-\frac{\eta_0}{n}\sum_{i=1}^n\tilde{g}(\mathbf{z}_i,\boldsymbol{\beta}_{\mathrm{ldp}}^{(t)})\,. \tag{S.4}$$

Lemma L7 then implies that the resulting estimator after T iterations, $\boldsymbol{\beta}_{\mathrm{ldp}}^{(T)}$, is (ϵ,δ)-LDP. The complete LDP sparsity-constrained ERM procedure is summarized in Algorithm A1.

While Algorithm A1 provides a feasible locally private alternative to the CDP-based procedure, its design introduces several structural constraints that have important implications for both variable selection and statistical efficiency. We highlight two key limitations below.

- The first limitation concerns the support-set updating mechanism, which makes the LDP implementation more sensitive to initialization. Under the CDP procedure (Algorithm 3), each iteration first forms an intermediate update and then applies NoisyHT to adaptively screen and update the support, allowing the support set to evolve across iterations and gradually correct variable-selection errors. In contrast, the LDP construction (Algorithm A1) restricts the

Algorithm A1 Local Differentially Private Sparsity-Constrained ERM

Input: Dataset $\{\mathbf{z}_i = (d_i, \mathbf{x}_i)\}_{i=1}^n$, learning rate η_0, number of iterations T, truncation parameter γ, quantile level $\tau = b/(b+h)$, smoothing parameter ϖ, privacy levels (ϵ, δ), and initial value $\boldsymbol{\beta}_{\text{ldp}}^{(0)}$ with sparsity level s.

1: Let $S_0 = \text{supp}(\boldsymbol{\beta}_{\text{ldp}}^{(0)})$.

2: **for** $t = 0, \ldots, T-1$ **do**

3: Compute $\{g(\mathbf{z}_i, \boldsymbol{\beta}_{\text{ldp}}^{(t)})\}_{i=1}^n$ via (S.2), and draw $\{\boldsymbol{\zeta}_i^t\}_{i=1}^n$ independently, where $\boldsymbol{\zeta}_i^t \sim N(\mathbf{0}, \sigma_{\text{ldp}}^2 \mathbf{I}_s)$ with $\sigma_{\text{ldp}} = 2\gamma\bar{\tau}T\epsilon^{-1}\sqrt{2s\log(1.25T/\delta)}$.

4: Compute $\{\tilde{g}(\mathbf{z}_i, \boldsymbol{\beta}_{\text{ldp}}^{(t)})\}_{i=1}^n$ via (S.3) and update

$$\boldsymbol{\beta}_{\text{ldp}}^{(t+1)} = \boldsymbol{\beta}_{\text{ldp}}^{(t)} - \frac{\eta_0}{n}\sum_{i=1}^n \tilde{g}(\mathbf{z}_i, \boldsymbol{\beta}_{\text{ldp}}^{(t)})\,.$$

5: **end for**

Output: $\boldsymbol{\beta}_{\text{ldp}}^{(T)}$.

release and aggregation of individual gradients to $S_0 = \text{supp}(\boldsymbol{\beta}^{(0)})$ in order to keep the magnitude of local perturbations under control, while all remaining coordinates are set to zero. As a result, every update is confined to the fixed index set S_0, and the support remains contained within S_0 throughout. This dependence imposes a stronger requirement on the initial support coverage: to avoid missing true signals and to ensure valid estimation, it is necessary that $\text{supp}(\boldsymbol{\beta}^*) \subseteq S_0$.

- The second limitation stems from the fact that the stronger, user-level privacy guarantee of LDP requires a different noise-injection scheme than that used under CDP, and this stronger protection leads to a slower convergence rate, reflecting an inherent statistical cost of local privacy (Duchi et al. 2018). To illustrate this, note that each CDP iteration can be decomposed into a non-private update plus an additive privacy-noise term:

$$\boldsymbol{\beta}_{\text{cdp}}^{(t+1)} = \tilde{P}_s\left(\boldsymbol{\beta}_{\text{cdp}}^{(t)} - \frac{\eta_0}{n}\sum_{i=1}^n g(\mathbf{z}_i, \boldsymbol{\beta}_{\text{cdp}}^{(t)})\right) + \tilde{\mathbf{w}}_{|S^{t+1}}^t\,,$$

where S^{t+1} denotes the index set selected by Algorithm 2 for $\mathbf{v} = \tilde{\boldsymbol{\beta}}_{\text{cdp}}^{(t+1)}$, and $\tilde{\mathbf{w}}_{|S^{t+1}}^t$ denotes the restriction of $\tilde{\mathbf{w}}^t$ to S^{t+1}, whose entries are drawn independently from $\text{Laplace}(\lambda_{\text{cdp}})$, with $\lambda_{\text{cdp}} \asymp \eta_0\gamma\bar{\tau}T(n\epsilon)^{-1}\sqrt{s\log(T/\delta)}$. Using standard moment properties of the Laplace distribution, we have $\mathbb{E}(\|\tilde{\mathbf{w}}_{|S^{t+1}}^t\|_2^2) \asymp \{s\eta_0\gamma\bar{\tau}T(n\epsilon)^{-1}\}^2$ up to logarithmic factors, showing that the mean-squared ℓ_2 magnitude of CDP noise is governed by a denominator of order $(n\epsilon)^2$. In contrast, each LDP iteration can be rewritten as

$$\boldsymbol{\beta}_{\mathrm{ldp}}^{(t+1)} = \left\{\boldsymbol{\beta}_{\mathrm{ldp}}^{(t)} - \frac{\eta_0}{n}\sum_{i=1}^{n} g(\mathbf{z}_i, \boldsymbol{\beta}_{\mathrm{ldp}}^{(t)})\right\}_{|S_0} - \tilde{\boldsymbol{\zeta}}_{|S_0}^{t},$$

where $\tilde{\boldsymbol{\zeta}}^t = \eta_0 n^{-1}\sum_{i=1}^{n}\mathring{\boldsymbol{\zeta}}_i^t$, $\mathring{\boldsymbol{\zeta}}_i^t \sim N(\mathbf{0}, \sigma_{\mathrm{ldp}}^2\mathbf{I}_p)$, and $\sigma_{\mathrm{ldp}} \asymp \gamma\bar{\tau}T\epsilon^{-1}\sqrt{s\log(T/\delta)}$. Consequently, $\mathbb{E}(\|\tilde{\boldsymbol{\zeta}}_{|S_0}^{t}\|_2^2)$ is governed by a denominator of order $n\epsilon^2$, rather than $(n\epsilon)^2$ as in CDP. The slower decay of LDP noise with respect to n therefore leads to a larger cumulative noise level and, in turn, a slower convergence rate.

EC.2.2. Hierarchical Differentially Private Sparsity-Constrained ERM

The CDP and LDP procedures represent two extremes in terms of trust assumptions: CDP presumes a fully trusted central curator, whereas LDP assumes that no entity beyond individual users is trusted. In many practical settings, however, this dichotomy is overly restrictive (Lowy and Razaviyayn 2023). Users may be unwilling to disclose their raw information to a central server, yet still be comfortable sharing updates within a trusted local or organization domain, such as a hospital, branch office, or edge server. Motivated by this intermediate trust structure, we introduce a hierarchical differential privacy (HDP) framework, in which users within each trusted group contribute updates through secure aggregation (Bonawitz et al. 2017), and privacy protection is enforced at the group level when generating the group-level messages used for model updating.

We consider a hierarchical distributed setting with K disjoint groups $\{\mathcal{G}_k\}_{k=1}^K$ such that $\cup_{k=1}^K\mathcal{G}_k = [n]$, where each $\mathcal{G}_k$ denotes the set of users in group $k \in [K]$. For simplicity, we assume $|\mathcal{G}_k| \asymp n/K$. Within each iteration, the parameter is updated sequentially across the groups. Specifically, we initialize the within-iteration update by setting $\boldsymbol{\beta}_{\mathrm{hdp}}^{(t,1)} = \boldsymbol{\beta}_{\mathrm{hdp}}^{(t)}$. For each group $k \in [K]$, we form a group-wise intermediate update

$$\tilde{\boldsymbol{\beta}}_{\mathrm{hdp}}^{(t,k+1)} = \boldsymbol{\beta}_{\mathrm{hdp}}^{(t,k)} - \frac{\eta_0}{|\mathcal{G}_k|}\sum_{i\in\mathcal{G}_k}\left\{\bar{K}\big((\mathbf{x}_i^\top\boldsymbol{\beta}_{\mathrm{hdp}}^{(t,k)} - d_i)/\varpi\big) - \tau\right\}w_\gamma(\mathbf{x}_i),$$

whose ℓ_∞-sensitivity is bounded by $2\eta_0\gamma\bar{\tau}|\mathcal{G}_k|^{-1}$. According to Lemma 1, setting the privacy parameters to $(\epsilon T^{-1},\, \delta T^{-1})$ and choosing $\lambda = 2\eta_0\gamma\bar{\tau}|\mathcal{G}_k|^{-1}$ ensures that each group-wise update

$$\boldsymbol{\beta}_{\mathrm{hdp}}^{(t,k+1)} = \mathrm{NoisyHT}(\tilde{\boldsymbol{\beta}}_{\mathrm{hdp}}^{(t,k+1)})$$

is $(\epsilon T^{-1}, \delta T^{-1})$-DP with respect to the data in $\mathcal{G}_k$. Because the groups $\{\mathcal{G}_k\}_{k=1}^K$ are disjoint, the within-iteration sweep over $k \in [K]$ does not incur additional privacy loss beyond this per-iteration budget. After cycling through all K groups, we set $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}_{\text{hdp}}^{(t,K+1)}$ and proceed to the next iteration. Applying standard composition across the T outer iterations yields an overall (ϵ, δ)-DP guarantee for the final estimator. The complete HDP sparsity-constrained ERM procedure is presented in Algorithm A2.

Algorithm A2 Hierarchical Differentially Private Sparsity-Constrained ERM

Input: Dataset $\{(d_i, \mathbf{x}_i)\}_{i=1}^n$, disjoint partition $\{\mathcal{G}_k\}_{k=1}^K$ of $[n]$, learning rate η_0, number of iterations T, truncation parameter γ, quantile level $\tau = b/(b+h)$, smoothing parameter ϖ, privacy levels (ϵ, δ), sparsity level s, and initial value $\boldsymbol{\beta}_{\text{hdp}}^{(0)}$.

1: **for** $t = 0, \ldots, T-1$ **do**
2: Set $\boldsymbol{\beta}_{\text{hdp}}^{(t,1)} = \boldsymbol{\beta}_{\text{hdp}}^{(t)}$.
3: **for** $k = 1, \ldots, K$ **do**
4: Compute the group-wise update

$$\tilde{\boldsymbol{\beta}}_{\text{hdp}}^{(t,k+1)} = \boldsymbol{\beta}_{\text{hdp}}^{(t,k)} - \frac{\eta_0}{|\mathcal{G}_k|} \sum_{i \in \mathcal{G}_k} \left\{ \bar{K}\left((\mathbf{x}_i^\top \boldsymbol{\beta}_{\text{hdp}}^{(t,k)} - d_i)/\varpi \right) - \tau \right\} w_\gamma(\mathbf{x}_i)\,,$$

$$\boldsymbol{\beta}_{\text{hdp}}^{(t,k+1)} = \mathrm{NoisyHT}\left(\tilde{\boldsymbol{\beta}}_{\text{hdp}}^{(t,k+1)}, \{(d_i, \mathbf{x}_i)\}_{i \in \mathcal{G}_k}, s, \epsilon T^{-1}, \delta T^{-1}, 2\eta_0 \gamma \bar{\tau} |\mathcal{G}_k|^{-1} \right).$$

5: **end for**
6: Set $\boldsymbol{\beta}_{\text{hdp}}^{(t+1)} = \boldsymbol{\beta}_{\text{hdp}}^{(t,K+1)}$.
7: **end for**

Output: $\boldsymbol{\beta}_{\text{hdp}}^{(T)}$.

Compared with the CDP procedure (Algorithm 3), the HDP design performs updates on a group-by-group basis. With $|\mathcal{G}_k| \asymp n/K$, each group-wise step effectively uses a batch of size n/K, while a full sweep across all K groups still leverages the entire sample. This grouped implementation changes the scale at which privacy noise is introduced: the mean-squared ℓ_2 magnitude of the injected noise is governed by a denominator of order $(n\epsilon/K)^2$. Consequently, in terms of the noise-induced error component, HDP exhibits convergence behavior that lies between those of CDP and LDP; see Figure EC.1[A1]. This comparison highlights a clear privacy-efficiency trade-off: LDP provides the strongest user-level protection but requires the largest perturbation and therefore

[A1] In this implementation, we generate the training data $\{(d_i, \mathbf{x}_i)\}_{i=1}^n$ as described in Section 5.1 and evaluate the excess risk on an independent test set of size $30,000$. For Algorithm A2, we consider two choices of the number of groups, $K \in \{2, 4\}$, to examine the impact of group partitioning on performance.

converges the slowest; CDP is the most statistically efficient when a trusted curator is available; and HDP offers an intermediate compromise, improving upon LDP whenever $K = o(n^{1/2})$. Moreover, as in the CDP setting, HDP retains a practically important advantage over LDP in terms of variable selection. Because privatization occurs at the group-aggregate level, the active set can be updated adaptively via $\mathrm{NoisyHT}$ throughout the iterations, rather than being restricted to the initial support $S_0 = \mathrm{supp}(\boldsymbol{\beta}^{(0)})$.

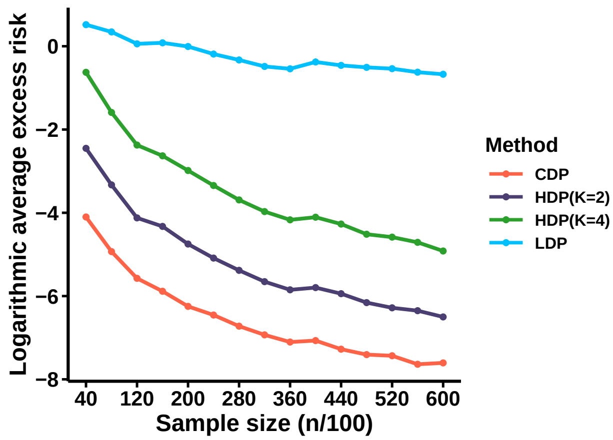

Figure EC.1 Logarithmic average excess risk versus sample size n, averaged over 500 replications, for three privacy-preserving procedures (CDP, HDP, and LDP). Results are shown for $(p, s^*) = (100, 5)$, $(b, h) = (2, 1)$, and $(\epsilon, \delta) = (0.5, 10^{-4})$.

EC.3. Additional Numerical Studies

EC.3.1. Computational Scalability across Sample Sizes and Dimensions

In this section, we compare the running times of ℓ_0-ERM (Algorithm 1), DP-ℓ_0-ERM (Algorithm 3), and ERM[A2] under varying sample sizes and ambient dimensions. For each (n, p), all algorithmic parameters are prespecified and kept fixed. The mean running time (in seconds) is reported over 100 replications. As shown in Table EC.1, the running time of all methods increases with both the sample size and the ambient dimension. When p is small, vanilla ERM has running times comparable to, and in some cases even slightly faster than, ℓ_0-ERM and DP-ℓ_0-ERM. However, its computational cost increases rapidly with p, because it is computed over the full feature space. In contrast, ℓ_0-ERM and DP-ℓ_0-ERM exploit sparsity and remain computationally efficient in higher-dimensional settings. Moreover, although DP-ℓ_0-ERM is consistently slower than ℓ_0-ERM, the additional computational cost induced by privacy protection remains moderate.

[A2] Here, ERM refers to the vanilla empirical risk minimizer based on the original cost function, computed using the `rq` function in the `quantreg` package, and used as a baseline for computational comparison.

Table EC.1 Mean running time (in seconds), averaged over 100 replications, for ℓ_0-ERM, DP-ℓ_0-ERM, and ERM under varying sample sizes and ambient dimensions. All computations were performed on a Windows platform with an Intel(R) Core(TM) i7-14700F CPU at 2.10 GHz.

	$n=1000$					$n=2000$				
Method	$p=25$	$p=50$	$p=100$	$p=200$	$p=400$	$p=25$	$p=50$	$p=100$	$p=200$	$p=400$
ℓ_0-ERM	0.004	0.006	0.009	0.019	0.035	0.009	0.013	0.025	0.036	0.076
DP-ℓ_0-ERM	0.006	0.011	0.016	0.032	0.064	0.013	0.021	0.040	0.063	0.137
ERM	0.004	0.009	0.038	0.119	0.419	0.008	0.028	0.088	0.340	1.116

In what follows, we present additional numerical studies to investigate the framework's performance under non-linearity through a two-pronged approach. The first study (Section EC.3.2) evaluates the proposed ℓ_0-ERM algorithm (Algorithm 1) under a structured nonlinear data-generating model and compares it with two nonparametric methods to illustrate how the proposed linear framework performs when the underlying decision rule is nonlinear. The second study (Section EC.3.3) develops a nonlinear extension via a two-layer feed-forward neural network to accommodate unstructured non-linearity. By comparing this extension with several classical nonlinear benchmarks, we demonstrate the flexibility of our privacy-preserving mechanisms beyond linear models, while also clarifying that such extensions are currently most feasible in moderate-dimensional settings.

EC.3.2. Sparse Linear Methods under Nonlinear Demand Models

In this section, we depart from the linear data-generating models considered in Section 5 and instead consider a nonlinear alternative. Specifically, we generate i.i.d. observations $\{(d_i, \mathbf{x}_i)\}_{i=1}^n$ according to $d_i = f_0(\mathbf{x}_i) + \varepsilon_i$, where $\mathbf{x}_i = (1, \mathbf{x}_{i,-1}^\top)^\top$ with $\mathbf{x}_{i,-1} \sim N(\mathbf{0}, \mathbf{I}_{p-1})$, and $\varepsilon_i \sim N(-\Phi^{-1}(b/(b+h)), 1)$. Here, $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, and

$$f_0(\mathbf{x}_i) = 0.5 + x_{i,2} - x_{i,3} + 0.8\exp(x_{i,5}) + x_{i,4}x_{i,5} + 0.8x_{i,2}x_{i,3}x_{i,4}\,,$$

where $x_{i,j}$ denotes the j-th component of $\mathbf{x}_i$ for $j \in \{2,3,4,5\}$. Under this construction, the τ-quantile of ε_i is zero for $\tau = b/(b+h)$, and hence $F^{-1}_{d_i\,|\,\mathbf{x}_i}(\tau) = f_0(\mathbf{x}_i)$. Therefore, in this experiment, the true conditional τ-quantile function $f_0(\mathbf{x})$ coincides with the population-optimal order-up-to function.

We consider two grid-based nonparametric benchmark methods, namely, Binning and the K-nearest neighbors (KNN) method. Since developing privacy-preserving versions of these two benchmarks is nontrivial, we focus on a non-private comparison with our ℓ_0-ERM approach for fairness. All methods are trained on a sample of size $n = 1000$, and their out-of-sample newsvendor costs are evaluated on an independent test set of size 5000. For both benchmarks, the action space is discretized over a common grid constructed from the training-sample demand values, consisting of 100 equally spaced candidate decisions ranging from the 1st to the 99th empirical percentiles of the training demand. For Binning, each covariate is partitioned into three bins using empirical training-sample quantiles, and the conditional demand distribution at a test point is approximated by the empirical distribution of training observations falling into the same multivariate bin. When the corresponding bin is empty, a prespecified fallback rule based on the global training distribution is used. For KNN, the number of neighbors is fixed at $K = 30$. Prior to applying KNN, all covariates are standardized using the sample means and standard deviations computed from the training data, and Euclidean distance in the standardized feature space is used to identify the nearest neighbors. In addition, the sparse framework can accommodate a richer class of decision rules through a predetermined expansion of the raw covariates. As a simple illustration, we also include an expanded-feature version, Expanded-ℓ_0-ERM, which applies Algorithm 1 after augmenting the feature space with all two-way interaction terms.

We fix per-unit costs $(b, h) = (2, 1)$ and vary p from 5 to 30. The comparison results are reported in Figure EC.2. Among the two nonparametric benchmarks, KNN is competitive only in low dimensions but deteriorates as p increases, whereas Binning remains less competitive throughout. By contrast, both ℓ_0-ERM and Expanded-ℓ_0-ERM remain comparatively stable as p increases. Moreover, Expanded-ℓ_0-ERM improves upon ℓ_0-ERM, indicating that modest feature expansion can enhance performance by providing a better approximation to the underlying order-up-to function when nonlinear structure is present. Overall, these findings suggest that even when the true decision rule is nonlinear, the sparse linear framework remains effective when combined with appropriate feature expansion.

EC.3.3. A Nonlinear Extension via a Two-Layer Neural Network

In this section, we extend our privacy-preserving smoothed loss to a canonical feed-forward neural network architecture. This extension demonstrates that our gradient-based perturbation mechanism can be effectively integrated with complex, unstructured nonlinear models with moderate dimensions beyond the sparse linear setting.

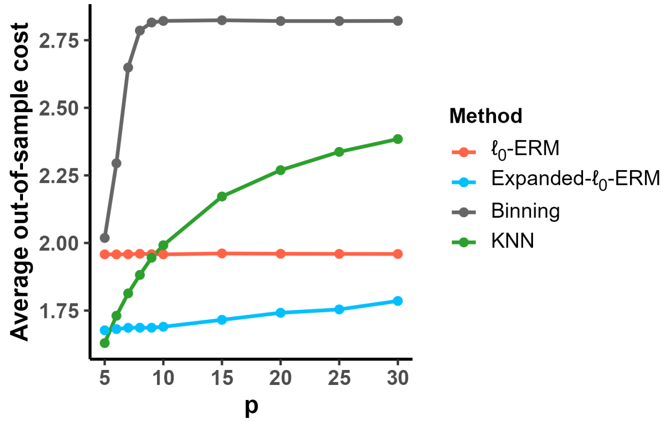

Figure EC.2 Average out-of-sample cost of ℓ_0-ERM, Expanded-ℓ_0-ERM, Binning, and KNN, based on 500 replications, under varying dimension p.

EC.3.3.1 Restricted ERM over a Two-Layer Neural Network Class

We assume that the true decision rule f_0 satisfies certain regularity properties, such as continuity or smoothness, so that it can be well approximated by a feed-forward neural network with sufficiently many hidden units (Cybenko 1989, Hornik et al. 1989, Mhaskar 1996). Motivated by this approximation perspective, we consider a standard one-hidden-layer (two-layer) feed-forward neural network class $\mathcal{F}_{\mathrm{NN}}$ to approximate f_0. Specifically, define

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = a + \sum_{r=1}^{m} b_r \cdot \psi(\boldsymbol{\xi}_r^\top \mathbf{x}), \tag{S.5}$$

where $\boldsymbol{\theta} = (a, \mathbf{b}^\top, \mathrm{vec}(\Xi)^\top)^\top \in \mathbb{R}^{m(p+1)+1}$, m denotes the network width (i.e., the number of hidden units), $\psi(\cdot)$ is a bounded and Lipschitz activation function, $\mathbf{b} = (b_1, \ldots, b_m)^\top \in \mathbb{R}^m$, and $\Xi = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m) \in \mathbb{R}^{p\times m}$ is the first-layer weight matrix. Note that, under the specification $\boldsymbol{\xi}_r = (\xi_{r,1}, \boldsymbol{\xi}_{r,-1}^\top)^\top$ and $\mathbf{x} = (1, \mathbf{x}_{-1}^\top)^\top$, the term $\boldsymbol{\xi}_r^\top \mathbf{x}$ in (S.5) can be written as $\boldsymbol{\xi}_{r,-1}^\top \mathbf{x}_{-1} + \xi_{r,1}$, so the bias term is already incorporated. In addition, we impose a coordinatewise boundedness constraint on $\boldsymbol{\theta}$ to control the capacity of the network. Specifically, for a fixed $B > 0$, define

$$\Theta(B) := \left\{ \boldsymbol{\theta} \in \mathbb{R}^{m(p+1)+1} : \|\boldsymbol{\theta}\|_\infty \le B \right\}.$$

We then set $\mathcal{F}_{\mathrm{NN}} := \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta(B)\}$. Based on $\mathcal{F}_{\mathrm{NN}}$, we estimate the decision rule by solving the constrained ERM

$$\min_{f\in\mathcal{F}_{\mathrm{NN}}} \left\{ \widehat{\mathcal{E}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \rho_{\tau,\varpi}(d_i - f(\mathbf{x}_i)) \right\} = \min_{\boldsymbol{\theta}\in\Theta(B)} \left\{ \widehat{\mathcal{E}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_{\tau,\varpi}(d_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)) \right\},$$

where $\rho_{\tau,\varpi}(\cdot)$ is the convolution-smoothed check function defined in Section 2.3.

EC.3.3.2 Noise-Perturbed Gradient Updates

Analogously to the clipping-based gradient update in Equation (5) of the main text, we define the intermediate iterate as

$$\tilde{\boldsymbol{\theta}}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \frac{\eta_0}{n}\sum_{i=1}^{n}\left\{\bar{K}\left(\frac{f_{\boldsymbol{\theta}^{(t)}}(\mathbf{x}_i)-d_i}{\varpi}\right)-\tau\right\}\nabla_{\boldsymbol{\theta}}^{\text{clip}} f_{\boldsymbol{\theta}^{(t)}}(\mathbf{x}_i)\,,\ \ t\in\{0\}\cup[T-1]\,, \tag{S.6}$$

where

$$\nabla_{\boldsymbol{\theta}}^{\text{clip}} f_{\boldsymbol{\theta}}(\mathbf{x}) := \big(1,\psi(\boldsymbol{\xi}_1^\top\mathbf{x}),\ldots,\psi(\boldsymbol{\xi}_m^\top\mathbf{x}), b_1\psi'(\boldsymbol{\xi}_1^\top\mathbf{x})\,\tilde{\mathbf{x}}^\top,\ldots,b_m\psi'(\boldsymbol{\xi}_m^\top\mathbf{x})\,\tilde{\mathbf{x}}^\top\big)^\top\,.$$

Here, $\tilde{\mathbf{x}} = \mathbf{x}\cdot\min\{\gamma/\|\mathbf{x}\|_2,1\}$ is introduced to ensure bounded ℓ_2-sensitivity of the intermediate iterate in (S.6). Define

$$G_0 := \sup_{u\in\mathbb{R}}|\psi(u)| \quad \text{and} \quad G_1 := \sup_{u\in\mathbb{R}}|\psi'(u)|\,. \tag{S.7}$$

Since

$$\sup_{(d,\mathbf{x})\in\mathbb{R}\times\mathbb{R}^p,\,\boldsymbol{\theta}\in\Theta(B)}\left|\bar{K}\left(\frac{f_{\boldsymbol{\theta}}(\mathbf{x})-d}{\varpi}\right)-\tau\right| \le \bar{\tau} := \max\{\tau,1-\tau\},$$

we have

$$\sup_{(d,\mathbf{x})\in\mathbb{R}\times\mathbb{R}^p,\,\theta\in\Theta(B)}\left\|\left\{\bar{K}\left(\frac{f_{\boldsymbol{\theta}}(\mathbf{x})-d}{\varpi}\right)-\tau\right\}\nabla_{\boldsymbol{\theta}}^{\text{clip}} f_{\boldsymbol{\theta}}(\mathbf{x})\right\|_2 \le \Delta := \bar{\tau}\sqrt{1+mG_0^2+mB^2G_1^2\gamma^2}\,,$$

which implies that the ℓ_2-sensitivity of each intermediate iterate is bounded by $2\eta_0\Delta n^{-1}$. By the Gaussian mechanism (Lemma L1), each noisy update

$$\check{\boldsymbol{\theta}}^{(t+1)} = \tilde{\boldsymbol{\theta}}^{(t+1)} + \mathbf{g}\,,$$

with $\mathbf{g}\sim N(\mathbf{0},\sigma^2\mathbf{I}_{m(p+1)+1})$, is $(\epsilon T^{-1},\delta T^{-1})$-DP, where $\sigma = 2\eta_0\Delta T(n\epsilon)^{-1}\sqrt{2\log(1.25T/\delta)}$. As a post-processing step, we enforce the coordinatewise boundedness constraint by setting $\boldsymbol{\theta}^{(t+1)} = w_B(\check{\boldsymbol{\theta}}^{(t+1)})$, where the truncation function $w_B(\cdot)$ is defined in Section 3.2. After T iterations, the standard composition property (Lemma L4) implies that the T-th iterate $\boldsymbol{\theta}^{(T)}$ is (ϵ,δ)-DP. The complete procedure is summarized in Algorithm A3. Unlike the standard literature on DP stochastic gradient descent (DP-SGD) for neural network training (Abadi et al. 2016, Bu et al. 2020, Park et al. 2023), Algorithm A3 is developed for decision learning and constructs a DP estimator via a sensitivity control argument combined with gradient clipping and noise injection. To seamlessly align the neural network optimization architecture with our analysis, we adopt a full-sample update to keep the sensitivity analysis and privacy calibration transparent. As discussed in Section EC.2.1, this framework can also be adapted to local DP, for which an SGD-type implementation is more natural.

Algorithm A3 Privatized Gradient Descent for One-Hidden-Layer Networks

Input: Dataset $\mathbf{Z} = \{(d_i, \mathbf{x}_i)\}_{i=1}^n$, stepsize η_0, number of iterations T, smoothing parameter ϖ, quantile level τ, clipping level γ, activation function $\psi(\cdot)$, width m, bound B, privacy parameters (ϵ, δ), initialization $\boldsymbol{\theta}^{(0)} \in \Theta(B)$.

1: **for** $t = 0, \ldots, T-1$ **do**

2: Compute the clipped intermediate iterate $\tilde{\boldsymbol{\theta}}^{(t+1)}$ according to (S.6);

3: Generate $\mathbf{g} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{m(p+1)+1})$ with $\sigma = 2\eta_0 \Delta T (n\epsilon)^{-1} \sqrt{2\log(1.25T/\delta)}$, and compute

$$\check{\boldsymbol{\theta}}^{(t+1)} = \tilde{\boldsymbol{\theta}}^{(t+1)} + \mathbf{g}\,,$$

where $\Delta = \bar{\tau}\sqrt{1 + mG_0^2 + mB^2 G_1^2 \gamma^2}$ with G_0, G_1 defined in (S.7);

4: Set $\boldsymbol{\theta}^{(t+1)} = w_B(\check{\boldsymbol{\theta}}^{(t+1)})$.

5: **end for**

Output: $\boldsymbol{\theta}^{(T)}$.

EC.3.3.3 Numerical Performance

Simulation Studies. To evaluate the finite-sample performance of the proposed feed-forward neural network (FNN), we consider two nonlinear data-generating models with $p = 5$. Specifically, we generate i.i.d. observations $\{(d_i, \mathbf{x}_i)\}_{i=1}^n$ according to

$$d_i = f_0(\mathbf{x}_i) + \varepsilon_i\,,$$

where $\mathbf{x}_i = (1, \mathbf{x}_{i,-1}^\top)^\top$ with $\mathbf{x}_{i,-1} \sim N(\mathbf{0}, \mathbf{I}_4)$, and $\varepsilon_i \sim \mathcal{N}(-\Phi^{-1}(b/(b+h)), 1)$. The true decision function $f_0(\cdot)$ is specified according to one of the following two models:

- Model 1: Define $g(u) = u\Phi(u)$ and $s(u) = \Phi(u) - 0.5$. Let

$$f_0(\mathbf{x}_i) = 0.5 + 0.5\{g(z_{i,1}) + g(z_{i,2})\} + s(z_{i,1})s(z_{i,2})\,,$$

where $\mathbf{x}_i = (1, x_{i,2}, x_{i,3}, x_{i,4}, x_{i,5})^\top$, $z_{i,1} = (x_{i,2} + x_{i,3})/\sqrt{2}$, and $z_{i,2} = (x_{i,4} - x_{i,5})/\sqrt{2}$.

- Model 2: Define $h(u) = (u^3 - 12u)/8$. Let

$$f_0(\mathbf{x}_i) = 0.5 + 0.5\{h(z_{i,1}) + h(z_{i,2})\}\,.$$

In our implementation, we use the activation function $\psi(u) = \tanh(u) = (e^u - e^{-u})/(e^u + e^{-u})$, for which $G_0 = G_1 = 1$. We set $(b, h) = (2, 1)$, $m = 200$, $T = 100$, $\eta_0 = 0.1$, $B = 0.1\sqrt{\log n}$, $\varpi = 0.1\{(p + \log n)/n\}^{2/5}$, and $\gamma = 0.1\sqrt{p + \log n}$. The initial value $\boldsymbol{\theta}^{(0)}$ is randomized via $\boldsymbol{\theta}^{(0)} = w_B(\tilde{\boldsymbol{\theta}}^{(0)})$, where $\tilde{\boldsymbol{\theta}}^{(0)} \sim N(\mathbf{0}, 0.1\mathbf{I}_{m(p+1)+1})$. Moreover, the privacy parameter δ is fixed at $\delta = 10^{-4}$. We evaluate the excess risk on an independent test set of size $N_{\text{test}} = 30{,}000$ as

$$\frac{1}{N_{\text{test}}}\sum_{i=1}^{N_{\text{test}}}\left\{b(d_i - \hat{q}_i)_+ + h(\hat{q}_i - d_i)_+\right\} - \frac{1}{N_{\text{test}}}\sum_{i=1}^{N_{\text{test}}}\left\{b(d_i - q_i^*)_+ + h(q_i^* - d_i)_+\right\},$$

where $\hat{q}_i = \hat{f}(\mathbf{x}_i)$ and $q_i^* = f_0(\mathbf{x}_i)$. Here, $\hat{f}(\cdot)$ denotes the estimated decision rule.

We first compare the proposed method with several nonlinear quantile estimation benchmarks in the non-private setting. Specifically, we consider a non-private version of the FNN estimator obtained without gradient clipping and noise injection. As benchmarks, we include two feature-based competitors and one feature-free method. The first feature-based benchmark is quantile regression forests (QRF), a tree-based ensemble method for conditional quantile estimation proposed by Meinshausen (2006) and implemented via the R function `quantregForest`. The second is the quantile regression neural network (QRNN), a neural-network-based method implemented via the R function `qrnn` following Cannon (2011, 2018). Relative to the standard `qrnn` implementation, the proposed method admits a DP extension through sensitivity-controlled gradient clipping and noise injection. To assess the role of feature information, we also include Sample Average Approximation (SAA) as a feature-free benchmark; see Ban and Rudin (2019). Figure EC.3 presents boxplots of excess risk for the four methods across different sample sizes under the two decision models. Overall, the feature-based methods consistently outperform the feature-free SAA benchmark, highlighting the importance of incorporating covariate information. Among the feature-based competitors, the proposed FNN exhibits strong and stable finite-sample performance, with boxplots that are generally lower and tighter than those of QRF and often comparable to or better than those of QRNN. QRNN becomes more competitive as the sample size increases, whereas QRF is generally less effective. Overall, the results indicate that the proposed FNN provides a competitive and stable approach for nonlinear decision learning.

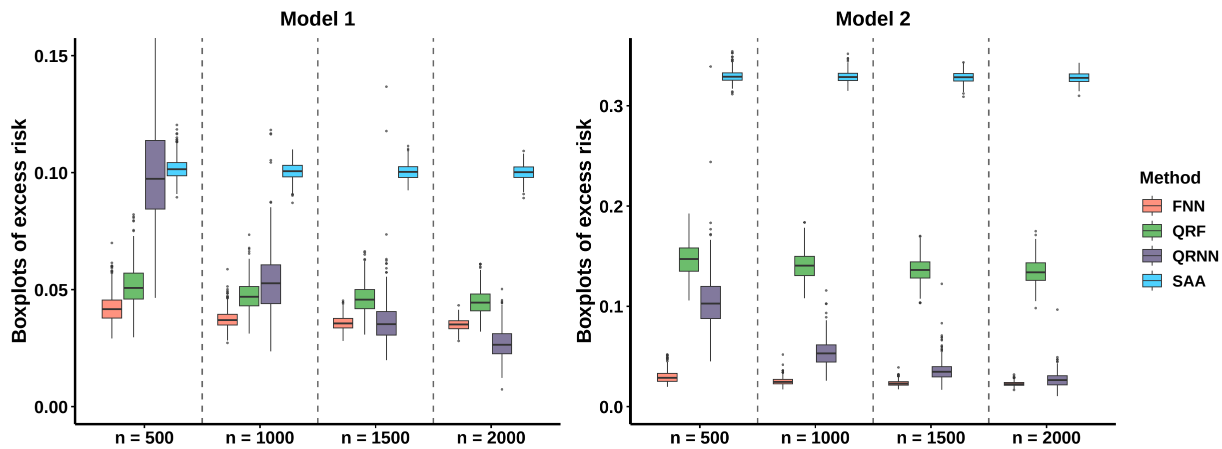

Figure EC.3 Boxplots of excess risk for FNN, QRF, QRNN, and SAA, over 500 replications, across four sample sizes under two decision models.

Next, we examine the finite-sample performance of the proposed (ϵ,δ)-DP FNN estimator relative to its non-private counterpart, with training sample sizes n ranging from $10,000$ to $100,000$. As shown in Figure EC.4, the excess risk of the proposed (ϵ,δ)-DP FNN estimator gradually approaches that of the corresponding non-private FNN estimator as the sample size increases. Moreover, weaker privacy protection, corresponding to a larger value of ϵ, leads to improved excess risk performance.

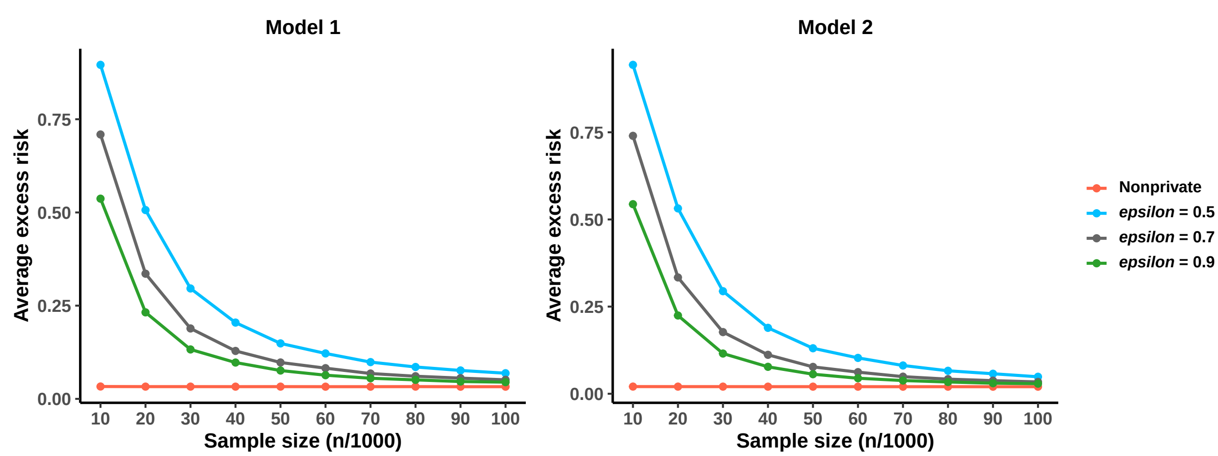

Figure EC.4 Plots of the average excess risk, averaged over 500 replications, as a function of the sample size under two decision models.

Real Data Analysis. Using the Grocery Sales Forecasting dataset described in Section 6, we compare the performance of the proposed FNN with that of QRF, QRNN, and SAA. In this analysis, we consider four continuous features: recent demand trends, the proportion of promotional sales, the total number of transactions, and the average weekly oil price. The design matrix is standardized using Z-scores prior to analysis. We randomly partition the dataset into a training set $\mathcal{D}_{\text{train}}$ containing 70% of the observations and a test set $\mathcal{D}_{\text{test}}$ containing the remaining 30%. We fix the holding cost at $h = 1$ and vary the underage cost over $b \in \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$. This procedure is repeated 500 times. Performance is evaluated by the average out-of-sample cost across replications, where the cost in each replication is computed as $|\mathcal{D}_{\text{test}}|^{-1} \sum_{i \in \mathcal{D}_{\text{test}}} \{b(d_i - \hat{q}_i)_+ + h(\hat{q}_i - d_i)_+\}$, where $\hat{q}_i = \hat{f}(\mathbf{x}_i)$ and $\hat{f}(\cdot)$ denotes the estimated decision rule. The comparison results are presented in Figure EC.5, which indicate that the proposed method achieves competitive out-of-sample performance relative to the alternative approaches across different values of the underage cost b.

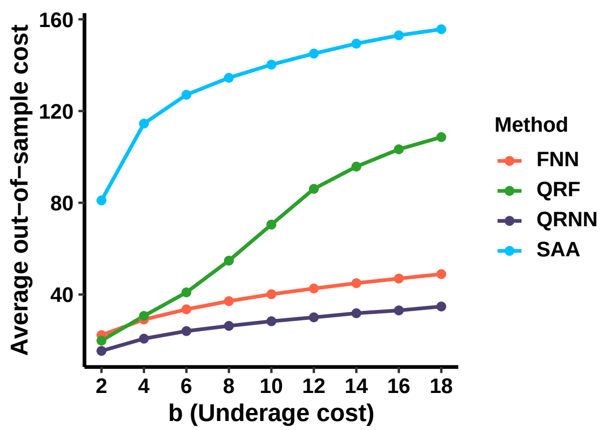

Figure EC.5 Average out-of-sample cost under varying underage cost *b*, averaged over 500 replications.

These numerical results demonstrate that the proposed framework can be extended to nonlinear model classes while maintaining competitive performance relative to standard nonparametric methods. At the same time, as shown in Algorithm A3, the noisy updates are performed over the full dimension $m(p+1)+1$. This suggests that the current extension is limited to moderate-dimensional settings. Developing a privacy-preserving neural-network framework that can effectively accommodate sparsity remains an important direction for future research.

Proofs of Statements

EC.4. Proof of Lemma 1

We first present the 'Private Max' algorithm (Algorithm A4), which identifies and returns the index of the value with the largest noisy magnitude.

Algorithm A4 Private Max

Input: Dataset $\mathbf{Z}$, vector-valued function $\mathbf{v} = \mathbf{v}(\mathbf{Z}) = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$ with each $\text{sens}_1(v_j)$ bounded by Δ, and privacy parameter ϵ.

1: **for** $j \in [p]$ **do**
2: Set $\tilde{v}_j = |v_j(\mathbf{Z})| + w_j$, where w_j is independently sampled from Laplace$(3\Delta/\epsilon)$;
3: **end for**

Output: Return $j^* = \arg\max_{j \in [p]} \tilde{v}_j$ and $v^*_{j^*}(\mathbf{Z}) := v_{j^*}(\mathbf{Z}) + w$, where w is a fresh draw from Laplace$(3\Delta/\epsilon)$.

Based on Algorithm A4, Algorithm 2 is equivalent to the 'Peeling' algorithm (Algorithm A5) with Laplace noise scale $\Lambda = 2\epsilon^{-1}\lambda\sqrt{5s\log(1/\delta)}$, where the parameter λ satisfies $\|\mathbf{v}(\mathbf{Z}) - \mathbf{v}(\mathbf{Z}')\|_\infty < \lambda$ for all neighboring datasets $\mathbf{Z}$ and $\mathbf{Z}'$.

Algorithm A5 Peeling

Input: Dataset $\mathbf{Z}$, vector-valued function $\mathbf{v} = \mathbf{v}(\mathbf{Z}) = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$, number of invocations s, and the Laplace noise scale Λ.

1: Initialize $\mathbf{v}_1 = \mathbf{v}_1(\mathbf{Z}) = \mathbf{v}(\mathbf{Z})$;
2: **for** $j \in [s]$ **do**
3: Let $(i_j, v^*_{i_j}(\mathbf{Z}))$ be returned by 'Private Max' applied to $(\mathbf{Z}, \mathbf{v}_j)$ with Laplace noise scale Λ;
4: Set $\mathbf{v}_{j+1} = \text{remove}(v_{i_j}, \mathbf{v}_j) \in \mathbb{R}^{p-j}$, i.e., remove the element v_{i_j} from $\mathbf{v}_j$;
5: **end for**

Output: The s-tuple $\{(i_1, v^*_{i_1}(\mathbf{Z})), \ldots, (i_s, v^*_{i_s}(\mathbf{Z}))\}$.

We first establish that

$$\text{Algorithm A4 (Private Max) is } (\epsilon, 0)\text{-DP}\,. \tag{S.8}$$

The proof of (S.8) will be given later. Let $\Lambda = 2\epsilon^{-1}\lambda\sqrt{5s\log(1/\delta)}$. By (S.8), each invocation of Private max within the Peeling algorithm is $(3\epsilon/\{2\sqrt{5s\log(1/\delta)}\}, 0)$-DP. Applying Lemma L5, the complete Peeling algorithm with noise scale Λ achieves $(\tilde{\epsilon}, \delta)$-DP with

$$\begin{aligned}\tilde{\epsilon} &= \frac{3\epsilon}{2\sqrt{5s\log(1/\delta)}}\sqrt{2s\log(1/\delta)} + \frac{3s\epsilon}{2\sqrt{5s\log(1/\delta)}}\left[\exp\left\{\frac{3\epsilon}{2\sqrt{5s\log(1/\delta)}}\right\} - 1\right] \\ &= \frac{3\sqrt{2}}{2\sqrt{5}}\epsilon + \frac{3s\epsilon}{2\sqrt{5s\log(1/\delta)}}\left[\exp\left\{\frac{3\epsilon}{2\sqrt{5s\log(1/\delta)}}\right\} - 1\right].\end{aligned}$$

Note that the function $f(x) = ax^{1/2}(e^{ax^{-1/2}} - 1)$ is monotonically decreasing in $x \geq 0$ for any $a > 0$. Under the parameter conditions $\epsilon \leq 0.5$, $\delta \leq 0.01$ and $s \geq 3$, we have

$$\frac{3s}{2\sqrt{5s\log(1/\delta)}}\left[\exp\left\{\frac{3\epsilon}{2\sqrt{5s\log(1/\delta)}}\right\} - 1\right] < 0.0512\,.$$

Since $3\sqrt{2}/(2\sqrt{5}) + 0.0512 < 1$, it follows that $\tilde{\epsilon} \leq \epsilon$.

To complete the proof of Lemma 1, it remains to verify the claim (S.8). Now we begin to prove (S.8). For any index $j \in [p]$ and any measurable set $S \subseteq \mathbb{R}$, it suffices to show that

$$\frac{\mathbb{P}(\tilde{v}_j \text{ is the largest and } v_j(\mathbf{Z}) + w \in S)}{\mathbb{P}(\tilde{v}_j' \text{ is the largest and } v_j(\mathbf{Z}') + w \in S)} \leq e^{\epsilon}\,,$$

where $\tilde{v}_j'$ corresponds to $\tilde{v}_j$ but is evaluated on an neighboring database $\mathbf{Z}'$. This inequality is equivalent to

$$\frac{\mathbb{P}(\tilde{v}_j \text{ is largest})\mathbb{P}(v_j(\mathbf{Z}) + w \in S \,|\, \tilde{v}_j \text{ is largest})}{\mathbb{P}(\tilde{v}_j' \text{ is largest})\mathbb{P}(v_j(\mathbf{Z}') + w \in S \,|\, \tilde{v}_j' \text{ is largest})} \leq e^{\epsilon}\,. \tag{S.9}$$

First, observe that by assumption $|v_j(\mathbf{Z}) - v_j(\mathbf{Z}')| \leq \Delta$. Then, Lemma 2.3 of Dwork et al. (2021) shows that

$$\frac{\mathbb{P}(v_j(\mathbf{Z}) + w \in S \,|\, \tilde{v}_j \text{ is largest})}{\mathbb{P}(v_j(\mathbf{Z}') + w \in S \,|\, \tilde{v}_j' \text{ is largest})} \leq e^{\epsilon/3}\,. \tag{S.10}$$

Second, let $\{w_j\}_{j=1}^p$ be the Laplace random variables generated in Algorithm A4 and $\mathbf{w}_{-j} = \mathrm{remove}(w_j, \mathbf{w})$, where $\mathbf{w} = (w_1, \ldots, w_p)^\top$. Then, proving

$$\frac{\mathbb{P}(\tilde{v}_j \text{ is largest})}{\mathbb{P}(\tilde{v}_j' \text{ is largest})} \leq e^{2\epsilon/3}\,,$$

is equivalent to showing

$$\frac{\mathbb{P}(\tilde{v}_j \text{ is largest} \,|\, \mathbf{Z}, \mathbf{w}_{-j})}{\mathbb{P}(\tilde{v}_j' \text{ is largest} \,|\, \mathbf{Z}', \mathbf{w}_{-j})} \leq e^{2\epsilon/3}\,. \tag{S.11}$$

Recall $\tilde{v}_j = |v_j(\mathbf{Z})| + w_j$ and $\tilde{v}'_j = |v_j(\mathbf{Z}')| + w_j$. Define

$$w_* = \min\left[w \in \mathbb{R} : |v_j(\mathbf{Z})| + w > |v_k(\mathbf{Z})| + w_k \text{ for any } k \in [p]\backslash\{j\}\right].$$

Then, $\tilde{v}_j$ is the largest value in the database $\mathbf{Z}$ if and only if $w_j \geq w_*$. Due to $\max_{j\in[p]} |v_j(\mathbf{Z}) - v_j(\mathbf{Z}')| \leq \Delta$, we have $|v_j(\mathbf{Z}')| + \Delta + w_* \geq |v_j(\mathbf{Z})| + w_* > |v_k(\mathbf{Z})| + w_k \geq |v_k(\mathbf{Z}')| - \Delta + w_k$ for any $k \in [p]\backslash\{j\}$, which implies $|v_j(\mathbf{Z}')| + 2\Delta + w_* > |v_k(\mathbf{Z}')| + w_k$ for any $k \in [p]\backslash\{j\}$. Thus, $\tilde{v}'_j$ is the largest value in the database $\mathbf{Z}'$ if $w_j \geq 2\Delta + w_*$. Since $w_j \sim \text{Laplace}(3\Delta/\epsilon)$, it holds that

$$\begin{aligned}
\mathbb{P}(\tilde{v}'_j \text{ is largest} \,|\, \mathbf{Z}', \mathbf{w}_{-j}) &\geq \mathbb{P}(w_j \geq 2\Delta + w_*) \\
&\geq e^{-2\epsilon/3}\mathbb{P}(w_j \geq w_*) = e^{-2\epsilon/3}\mathbb{P}(\tilde{v}_j \text{ is largest} \,|\, \mathbf{Z}, \mathbf{w}_{-j}),
\end{aligned}$$

which yields (S.11). Combining this with (S.10) establishes (S.9), as desired. Hence, we complete the proof of Lemma 1. □

EC.5. Proofs of Theoretical Results in Section 4

To derive the desired results, we first establish the convergence properties of the private decision estimator $\boldsymbol{\beta}^{(T)}$ obtained from Algorithm 3, along with those of its non-private counterpart $\boldsymbol{\beta}^{(T)}_{\mathrm{IHT}}$ obtained from Algorithm 1. The detailed proofs of the corresponding results can be found in Section EC.6.

Theorem T1 provides the estimation ℓ_2-error of the decision parameter $\boldsymbol{\beta}^{(T)}$ obtained from Algorithm 3. Its proof is given in Section EC.6.2.

THEOREM T1. *Let Assumptions* 1–3 *hold, and the parameter quadruple* (γ, s, ϖ, T) *in Algorithm* 3 *satisfy*

$$\gamma \asymp \sqrt{\log(pn)}\,,\ s \gtrsim_{f_u/f_l, \lambda_1/\lambda_p} s^*\,,\ \left\{\frac{s\log(pn)}{n}\right\}^{1/2} \lesssim \varpi \lesssim \left\{\frac{s\log(pn)}{n}\right\}^{1/4} \text{ and } T \asymp \log\left(\frac{n\epsilon}{\log p}\right). \tag{S.12}$$

Moreover, let the learning rate and sample size satisfy, respectively, that $\eta_0 \leq f_l\lambda_p/(40f_u^2\lambda_1^2)$ *and* $n\epsilon \gtrsim sT\log^{1/2}(T/\delta)\log^{3/2}(pn)$. *Then, the differentially private decision estimator* $\boldsymbol{\beta}^{(T)}$ *obtained from Algorithm* 3 *initialized at any* $\boldsymbol{\beta}^{(0)} \in \mathbb{H}(s) \cap \Theta(c_0)$ *satisfies*

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \frac{1}{f_l}\sqrt{\frac{s\log(pn)}{n}} + \frac{sT}{n\epsilon}\sqrt{\log\left(\frac{T}{\delta}\right)\log^3(pn)} \text{ with probability at least } 1 - \frac{C}{n^2}\,, \tag{S.13}$$

and

$$\mathbb{E}(\|\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*\|_2) \le \{\mathbb{E}(\|\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*\|_2^2)\}^{1/2} \lesssim \frac{1}{f_l}\sqrt{\frac{s\log(pn)}{n}} + \frac{sT}{n\epsilon}\sqrt{\log\left(\frac{T}{\delta}\right)\log^3(pn)}\,. \quad \text{(S.14)}$$

COROLLARY C1. *Let Assumptions* 1–3 *hold, and the triplet of parameters* (T,ϖ,s) *in Algorithm* 1 *satisfy*

$$s \gtrsim_{f_u/f_l,\lambda_1/\lambda_p} s^*\,,\ \left\{\frac{s\log(pn)}{n}\right\}^{1/2} \lesssim \varpi \lesssim \left\{\frac{s\log(pn)}{n}\right\}^{1/4} \ \textit{and}\ T \asymp \log\left(\frac{n}{\log p}\right).$$

Moreover, let the learning rate and sample size satisfy, respectively, that $\eta_0 \le f_l\lambda_p/(40 f_u^2\lambda_1^2)$ *and* $n \gtrsim s\log(pn)$. *Then, the non-private sparse decision estimator* $\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)}$ *obtained from Algorithm* 1 *initialized at any* $\boldsymbol{\beta}^{(0)} \in \mathbb{H}(s)\cap\Theta(c_0)$ *satisfies*

$$\|\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)}-\boldsymbol{\beta}^*\|_2 \lesssim \frac{1}{f_l}\sqrt{\frac{s\log(pn)}{n}} \ \textit{with probability at least}\ 1-\frac{C}{n^2}\,, \quad \text{(S.15)}$$

and

$$\mathbb{E}(\|\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)}-\boldsymbol{\beta}^*\|_2) \le \{\mathbb{E}(\|\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)}-\boldsymbol{\beta}^*\|_2^2)\}^{1/2} \lesssim \frac{1}{f_l}\sqrt{\frac{s\log(pn)}{n}}\,. \quad \text{(S.16)}$$

EC.5.1. Proof of Theorem 2

Recall that $q(\mathbf{x}_{n+1},\boldsymbol{\beta}^{(T)}) - q^*(\mathbf{x}_{n+1}) = \mathbf{x}_{n+1}^\top(\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*)$. According to (S.13) in Theorem T1, the event

$$\tilde{\mathcal{E}}_1 := \left[\|\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*\|_2 \le r_{n,1} \asymp \left\{\frac{1}{f_l}\sqrt{\frac{s\log(pn)}{n}} + \frac{sT}{n\epsilon}\sqrt{\log\left(\frac{T}{\delta}\right)\log^3(pn)}\right\}\right]$$

occurs with probability at least $1-Cn^{-2}$. Based on such defined $\tilde{\mathcal{E}}_1$ and Assumption 1, we have

$$\begin{aligned}
&\mathbb{P}\{|q(\mathbf{x}_{n+1},\boldsymbol{\beta}^{(T)}) - q^*(\mathbf{x}_{n+1})| > r_{n,1}z\} \le \mathbb{P}(|\langle \mathbf{x}_{n+1},\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*\rangle| > r_{n,1}z, \tilde{\mathcal{E}}_1) + \mathbb{P}(\tilde{\mathcal{E}}_1^{\mathrm{c}})\\
&\le \mathbb{P}\left\{\left|\left\langle \mathbf{x}_{n+1}, \frac{\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*\|_2}\right\rangle\right| > z\right\} + \frac{C}{n^2} \le \sup_{\mathbf{u}\in\mathbb{S}^{p-1}} \mathbb{P}(|\langle \mathbf{x}_{n+1},\mathbf{u}\rangle| > z) + \frac{C}{n^2}\\
&\le 2\exp\left(-\frac{z^2}{2\upsilon_1^2\lambda_1}\right) + \frac{C}{n^2}
\end{aligned}$$

for all $\boldsymbol{\beta}^{(T)} \ne \boldsymbol{\beta}^*$, which implies that, for all $\boldsymbol{\beta}^{(T)} \ne \boldsymbol{\beta}^*$,

$$|q(\mathbf{x}_{n+1},\boldsymbol{\beta}^{(T)}) - q^*(\mathbf{x}_{n+1})| \lesssim r_{n,1}\sqrt{\log n}$$

holds with probability at least $1-Cn^{-2}$. For $\boldsymbol{\beta}^{(T)}=\boldsymbol{\beta}^*$, we also have $|q(\mathbf{x}_{n+1},\boldsymbol{\beta}^{(T)})-q^*(\mathbf{x}_{n+1})|=0\lesssim r_{n,1}\sqrt{\log n}$. Then, we can conclude that

$$|q(\mathbf{x}_{n+1},\boldsymbol{\beta}^{(T)})-q^*(\mathbf{x}_{n+1})|\lesssim \frac{1}{f_l}\sqrt{\frac{s\log(pn)\log n}{n}}+\frac{sT}{n\epsilon}\sqrt{\log\left(\frac{T}{\delta}\right)\log^3(pn)\log n}$$

holds with probability at least $1-Cn^{-2}$. Plugging in $T\asymp\log(n\epsilon/\log p)$ yields (6) in Theorem 2.

Notice that $\mathbf{x}_{n+1}$ is independent of $\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*$. By (S.14) and Assumption 1, it holds that

$$\begin{aligned}\mathbb{E}\{|q(\mathbf{x}_{n+1},\boldsymbol{\beta}^{(T)})-q^*(\mathbf{x}_{n+1})|\}&=\mathbb{E}\{|\mathbf{x}_{n+1}^\top(\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*)|\}\le[\mathbb{E}\{|\mathbf{x}_{n+1}^\top(\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*)|^2\}]^{1/2}\\&=\big[\mathbb{E}\big\{(\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*)^\top\boldsymbol{\Sigma}(\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*)\big\}\big]^{1/2}\le\lambda_1^{1/2}\cdot\{\mathbb{E}(\|\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*\|_2^2)\}^{1/2}\\&\lesssim\frac{1}{f_l}\sqrt{\frac{s\log(pn)}{n}}+\frac{sT}{n\epsilon}\sqrt{\log\left(\frac{T}{\delta}\right)\log^3(pn)}\,.\end{aligned}\tag{S.17}$$

Since $T\asymp\log(n\epsilon/\log p)$, we complete the proof of (7) in Theorem 2. □

EC.5.2. Proof of Corollary 1

Recall that $q(\mathbf{x}_{n+1},\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)})-q^*(\mathbf{x}_{n+1})=\mathbf{x}_{n+1}^\top(\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)}-\boldsymbol{\beta}^*)$. Based on (S.15) in Corollary C1 and by applying similar arguments as those used in the proof of (6) in Section EC.5.1, we also obtain

$$|q(\mathbf{x}_{n+1},\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)})-q^*(\mathbf{x}_{n+1})|\lesssim\frac{1}{f_l}\sqrt{\frac{s\log(pn)\log n}{n}}$$

holds with probability at least $1-Cn^{-2}$.

Notice that $\mathbf{x}_{n+1}$ is independent of $\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)}-\boldsymbol{\beta}^*$. Analogous to (S.17), by (S.16), it holds that

$$\mathbb{E}\{|q(\mathbf{x}_{n+1},\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)})-q^*(\mathbf{x}_{n+1})|\}\le\lambda_1^{1/2}\cdot\{\mathbb{E}(\|\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)}-\boldsymbol{\beta}^*\|_2^2)\}^{1/2}\lesssim\frac{1}{f_l}\sqrt{\frac{s\log(pn)}{n}}\,.$$

We complete the proof of Corollary 1. □

EC.5.3. Proof of Theorem 3

Recall that $\boldsymbol{\beta}^*$ satisfies the first-order condition $\nabla R(\boldsymbol{\beta}^*)=\mathbf{0}$. Define $\bar{\boldsymbol{\beta}}=m\boldsymbol{\beta}^*+(1-m)\boldsymbol{\beta}^{(T)}\in\mathbb{R}^p$ for some $m\in[0,1]$. Then,

$$\begin{aligned}R(\boldsymbol{\beta}^{(T)})-R(\boldsymbol{\beta}^*)&=R(\boldsymbol{\beta}^{(T)})-R(\boldsymbol{\beta}^*)-\langle\nabla R(\boldsymbol{\beta}^*),\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*\rangle\\&=\frac{1}{2}(\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*)^\top\{\nabla^2R(\bar{\boldsymbol{\beta}})\}(\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*)\\&\le\frac{\|\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*\|_2^2}{2}\sup_{\boldsymbol{\beta}\in\mathbb{R}^p}\|\nabla^2R(\boldsymbol{\beta})\|\,.\end{aligned}\tag{S.18}$$

Under Assumption 3, we know that $\nabla^2 R(\boldsymbol{\beta}) = \mathbb{E}\{f_{\varepsilon\,|\,\mathbf{x}}(\langle \mathbf{x}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle)\mathbf{x}\mathbf{x}^\top\}$. Then, by Assumptions 1 and 3, we have

$$\sup_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\nabla^2 R(\boldsymbol{\beta})\| = \sup_{\boldsymbol{\beta}\in\mathbb{R}^p} \sup_{\mathbf{u}\in\mathbb{S}^{p-1}} \left|\mathbb{E}\{f_{\varepsilon\,|\,\mathbf{x}}(\langle \mathbf{x}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle)(\mathbf{x}^\top\mathbf{u})^2\}\right| \le f_u \lambda_1 \,.$$

This, together with (S.18), yields

$$R(\boldsymbol{\beta}^{(T)}) - R(\boldsymbol{\beta}^*) \lesssim f_u \|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2^2 \,. \tag{S.19}$$

Combining (S.19) with (S.13), we obtain (8) in Theorem 3. Moreover, (9) in Theorem 3 holds by appling (S.14) and (S.19). □

EC.5.4. Proof of Corollary 2

By the proof of (S.19) in Section EC.5.3, we also have

$$R(\boldsymbol{\beta}^{(T)}_{\mathrm{IHT}}) - R(\boldsymbol{\beta}^*) \lesssim f_u \|\boldsymbol{\beta}^{(T)}_{\mathrm{IHT}} - \boldsymbol{\beta}^*\|_2^2 \,. \tag{S.20}$$

Combining (S.20) with (S.15) and (S.16), we obtain Corollary 2. □

EC.6. Proofs of Theorem T1 and Corollary C1

Our analysis proceeds in two parts. First, we derive the convergence rates of $\boldsymbol{\beta}^{(T)}$ and $\boldsymbol{\beta}^{(T)}_{\mathrm{IHT}}$ by conditioning on a collection of ‘good’ events associated with the empirical smoothed check loss. We then demonstrate that these events occur with high probability under Assumptions 1–3.

EC.6.1. Intermediate Results for Theorem T1 and Corollary C1

To begin with, we define some ‘good’ events that are essential for establishing the convergence of $\boldsymbol{\beta}^{(T)}$ and $\boldsymbol{\beta}^{(T)}_{\mathrm{IHT}}$. Let

$$\mathcal{S} = \{S \subseteq [p] : 1 \le |S| \le 2s\}\,,\ \ \mathbb{H}(s) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 \le s\} \ \text{ and } \ \Theta(r) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le r\}$$

for any $r > 0$. Moreover, for any $\boldsymbol{\beta} \in \mathbb{H}(s)$, let $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$, and write

$$\widehat{D}_\varpi(\boldsymbol{\delta}) = \widehat{R}_\varpi(\boldsymbol{\beta}) - \widehat{R}_\varpi(\boldsymbol{\beta}^*)\,,\ \ D_\varpi(\boldsymbol{\delta}) = \mathbb{E}\{\widehat{D}_\varpi(\boldsymbol{\delta})\}\,,$$
$$A_\varpi(\boldsymbol{\delta}) = R_\varpi(\boldsymbol{\beta}) - R_\varpi(\boldsymbol{\beta}^*) - \langle \nabla R_\varpi(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle,\ B_\varpi(\boldsymbol{\delta}) = R_\varpi(\boldsymbol{\beta}^*) - R_\varpi(\boldsymbol{\beta}) - \langle \nabla R_\varpi(\boldsymbol{\beta}), \boldsymbol{\beta}^* - \boldsymbol{\beta} \rangle\,,$$

where $R_\varpi(\boldsymbol{\beta}) = \mathbb{E}\{\widehat{R}_\varpi(\boldsymbol{\beta})\}$. With the above preparations in place, we now define the 'good' events:

$$
\begin{aligned}
\mathcal{E}_0 = \left\{ \max_{i\in[n]} \|\mathbf{x}_i\|_\infty \le \gamma \right\}, \quad \mathcal{E}_1 = \left\{ \inf_{\boldsymbol{\delta}\in\mathbb{H}(2s)\cap\mathbb{B}^p(c_0)} \frac{\min\{A_\varpi(\boldsymbol{\delta}), B_\varpi(\boldsymbol{\delta})\}}{\|\boldsymbol{\delta}\|_2^2} \ge b_1 \right\}, \\
\mathcal{E}_2 = \left\{ \sup_{\boldsymbol{\delta}\in\mathbb{H}(2s)\cap\{\mathbb{B}^p(c_0)\setminus\mathbb{B}^p(n^{-1})\}} \frac{|\widehat{D}_\varpi(\boldsymbol{\delta}) - D_\varpi(\boldsymbol{\delta})|}{\|\boldsymbol{\delta}\|_2} \le b_2 \right\}, \\
\mathcal{E}_3 = \left\{ \sup_{S\in\mathcal{S}} \left\| \{\nabla \widehat{R}_\varpi(\boldsymbol{\beta}^*)\}_S \right\|_2 \le b_3 \right\}, \\
\mathcal{E}_4 = \left\{ \sup_{\boldsymbol{\beta}\in\mathbb{H}(s)\cap\Theta(c_0),\, S\in\mathcal{S}} \left\| (1-\mathbb{E})[\{\nabla \widehat{R}_\varpi(\boldsymbol{\beta})\}_S] \right\|_2 \le b_4 \right\}, \\
\mathcal{E}_5 = \left\{ \sup_{\boldsymbol{\beta}\in\mathbb{H}(2s)\cap\Theta(c_0),\, S\in\mathcal{S}} \|\{\nabla^2 \widehat{R}_\varpi(\boldsymbol{\beta})\}_{SS}\| \le b_5 \right\}, \\
\mathcal{E}_6 = \left\{ \sup_{\boldsymbol{\beta}\in\mathbb{H}(s)\cap\Theta(c_0),\, S\in\mathcal{S}} \left\| \{\nabla \widehat{R}_\varpi(\boldsymbol{\beta}) - \nabla \widehat{R}_\varpi(\boldsymbol{\beta}^*)\}_S \right\|_2 \le \frac{3b_5 c_0}{4} \right\}, \\
\mathcal{E}_7 = \left\{ \sup_{\boldsymbol{\beta}\in\mathbb{H}(s)\cap\Theta(n^{-1}),\, S\in\mathcal{S}} \left\| \{\nabla \widehat{R}_\varpi(\boldsymbol{\beta}) - \nabla \widehat{R}_\varpi(\boldsymbol{\beta}^*)\}_S \right\|_2 \le \frac{3b_5}{4n} \right\},
\end{aligned}
\tag{S.21}
$$

where $\gamma > 0$ is a truncation parameter, c_0 is the constant specified in Assumption 3, and $b_1, b_2, b_3, b_4, b_5 > 0$ are parameters associated with statistical error bounds.

The event $\mathcal{E}_0$ ensures that the ℓ_∞-sensitivity of each clipped gradient descent step is bounded. For unbounded covariates following a sub-Gaussian distribution, it suffices to set the truncation parameter as $\gamma \asymp \sqrt{\log(p \vee n)}$ to ensure that $\mathcal{E}_0$ holds with high probability. The events $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_4$ and $\mathcal{E}_5$ pertain to the restricted strong convexity (RSC) and restricted strong smoothness (RSM) of the smoothed empirical check loss over a sparse parameter set (see Proposition P1 for details). In addition, the events $\mathcal{E}_3$, $\mathcal{E}_6$, and $\mathcal{E}_7$ ensure that the empirical gradients $\nabla \widehat{R}_\varpi(\boldsymbol{\beta})$ are well-controlled within a sparse neighborhood centered at $\boldsymbol{\beta}^*$. With appropriate choices of $(b_1, b_2, b_3, b_4, b_5)$, Theorem T3 guarantees that the event $\mathcal{E} := \mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_5 \cap \mathcal{E}_6 \cap \mathcal{E}_7$ holds with high probability.

Theorem T2 establishes a high-probability bound on $\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2$ conditioned on $\mathcal{E}$, where the probability is taken over the Laplace noise variables $\{\mathbf{w}_1^t, \ldots, \mathbf{w}_s^t, \tilde{\mathbf{w}}^t\}_{t=0}^{T-1}$ injected into the gradient descent updates. The proof of Theorem T2 is provided in Section EC.7. Write

$$
\bar{b}_3 = \sup_{S\in\mathcal{S}} \|\{\nabla R_\varpi(\boldsymbol{\beta}^*)\}_S\|_2 \,. \tag{S.22}
$$

THEOREM T2. *Assume that $\boldsymbol{\beta}^{(0)} \in \mathbb{H}(s) \cap \Theta(c_0)$ and the learning rate η_0 satisfies $\eta_0 \le 2b_1/(9b_5^2)$, where the constants b_1 and b_5 are specified in* (S.21) *and satisfy $b_1 \le b_5/2$. For any $z > 0$, suppose that $\lambda = 2\eta_0\gamma\bar{\tau}/n$ satisfies*

$$\left(\frac{sT\lambda}{\epsilon}\right)^2 \{\log(psT)+z\}^2 \log\left(\frac{T}{\delta}\right) \lesssim \eta_0^2 b_1^2 c_0^2\,. \tag{S.23}$$

Then, conditioned on $\mathcal{E}$, the DP ℓ_0-ERM estimator $\boldsymbol{\beta}^{(T)}$ obtained from Algorithm 3*, with $T \ge 2\log(c_0 n\epsilon/\log p)/\log\{(1-\rho)^{-1}\}$, satisfies*

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2^2 \lesssim \frac{(b_2+b_3+b_4)^2}{b_1^2} + \frac{b_5}{b_1}\left(\frac{\log p}{n\epsilon}\right)^2 + \left(\frac{sT\lambda}{\epsilon}\right)^2 \{\log(psT)+z\}^2 \log\left(\frac{T}{\delta}\right) \max\left\{\frac{C_{\eta_0}}{b_1\rho}, 1\right\},$$

with probability at least $1-e^{-z}$ (over $\{\mathbf{w}_1^t,\ldots,\mathbf{w}_s^t,\tilde{\mathbf{w}}^t\}_{t=0}^{T-1}$), provided that $s \ge 38(\eta_0 b_1)^{-2}s^*$, $\bar{b}_3 \le b_3$ and $(b_2+b_3+b_4) \le b_1c_0/24$, where $\rho = 4\eta_0 b_5 s^*/s + \eta_0^2 b_1 b_5/12$, $C_{\eta_0} = 4\{b_5(5-\eta_0 b_5)(2-2\eta_0 b_5)^{-1} + 3(1-\eta_0 b_5)\eta_0^{-1}\}$ and $\bar{b}_3$ is specified in (S.22).

Let $\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)}$ denote the non-private ℓ_0-ERM estimator obtained from Algorithm 1. The following result, Corollary C2, which is of independent interest, establishes the convergence rate of $\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)}$ under conditions analogous to those in Theorem T2. The proof of Corollary C2 follows closely the argument used in the proof of Theorem T2, with the key distinction that the noise terms $\{\mathbf{w}_1^t,\ldots,\mathbf{w}_s^t,\tilde{\mathbf{w}}^t\}_{t=0}^{T-1}$ vanish.

COROLLARY C2. *Assume that $\boldsymbol{\beta}^{(0)} \in \mathbb{H}(s) \cap \Theta(c_0)$ and the learning rate satisfies $\eta_0 \le 2b_1/(9b_5^2)$ with $b_1 \le b_5/2$. Then, conditioned on $\mathcal{E}$, the non-private ℓ_0-ERM estimator $\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)}$ obtained from Algorithm* 1*, with $T \gtrsim \log(c_0 n/\log p)$, satisfies*

$$\|\boldsymbol{\beta}_{\mathrm{IHT}}^{(T)} - \boldsymbol{\beta}^*\|_2^2 \lesssim \frac{(b_2+b_3+b_4)^2}{b_1^2} + \frac{b_5}{b_1}\left(\frac{\log p}{n}\right)^2,$$

provided that $s \gtrsim (\eta_0 b_1)^{-2}s^*$, $\bar{b}_3 \le b_3$ and $(b_2+b_3+b_4) \lesssim b_1 c_0$.

Theorem T3 below establishes that the event $\mathcal{E}$ occurs with high probability under Assumptions 1, 2 and 3. The proof of Theorem T3 is provided in Section EC.8.

THEOREM T3. *Let Assumptions* 1–3 *hold. Then, for any $z \ge 1/2$, the event $\mathcal{E}$ holds with probability at least $1-7e^{-z}$, under the choice of parameters $(\gamma, b_1, b_2, b_3, b_4, b_5)$ in* (S.21) *specified as*

$$\gamma \asymp \upsilon_1 \lambda_1^{1/2}\sqrt{\log p + \log n + z}\,, \quad b_1 = \frac{f_l\lambda_p - l_f k_1 \lambda_1 \varpi}{2}\,, \quad b_2 \asymp \upsilon_1 \lambda_1^{1/2}\sqrt{\frac{s\log p + \log\log(c_0 n) + z}{n}}\,,$$

$$b_3 \asymp \upsilon_1 \lambda_1^{1/2} \bigg\{ C_{\tau,\varpi}^{1/2} \sqrt{\frac{s\log p + z}{n}} + \frac{s\log p + z}{n} \bigg\} + \lambda_1^{1/2} l_f k_2 \varpi^2 ,$$
$$b_4 \asymp \upsilon_1 \lambda_1^{1/2} \bigg\{ \sqrt{\frac{s\log p + s\log(n/\varpi) + z}{n}} + \frac{s\log p + s\log(n/\varpi) + z}{n} \bigg\} + \frac{f_u \lambda_1 \varpi c_0}{n} + \frac{k_{\max} \lambda_1 c_0}{n} ,$$

and $b_5 = 2f_u\lambda_1$, *provided that* $n\varpi \geq C_*(s\log p + s\log n + z)$ *and* $1 \gtrsim \varpi \geq C_{**}\sqrt{(s+z)/n}$, *where* $C_{\tau,\varpi} = \tau(1-\tau) + (1+\tau)l_f k_2 \varpi^2$, *and* $C_*, C_{**} > 0$ *are two constants depending only on* $(l_K, l_f, f_u, k_{\max}, \kappa_4, \upsilon_1, \lambda_1)$.

EC.6.2. Proof of Theorem T1

Throughout the sequel, constants depending on $(l_K, l_f, f_u, k_{\max}, \kappa_4, \upsilon_1, \lambda_1)$ are implicitly absorbed into the notation '$\lesssim$', '$\gtrsim$' and '$\asymp$'.

EC.6.2.1 Proof of (S.13)

Note that $\eta_0 \leq f_l\lambda_p/(40f_u^2\lambda_1^2)$, $b_1 = (f_l\lambda_p - l_f k_1\lambda_1\varpi)/2$ and $b_5 = 2f_u\lambda_1$. Provided that $\varpi \leq f_l\lambda_p/(10l_f k_1\lambda_1)$, it follows that $\eta_0 \leq 2b_1/(9b_5^2)$ and $b_1 \leq b_5/2$. Under this choice of (η_0, b_1, b_5), the quantity $\rho = 4\eta_0 b_5 s^*/s + \eta_0^2 b_1 b_5/12 \in (0,1)$ is uniformly bounded away from 1. Since c_0 is a constant, choosing $T \asymp \log(n\epsilon/\log p)$ ensures that $T \geq 2\log(c_0 n\epsilon/\log p)/\log\{(1-\rho)^{-1}\}$. Letting $z = 2\log n$ and noting that $\gamma \asymp \sqrt{\log(pn)}$, condition (S.23) becomes equivalent to

$$n\epsilon \gtrsim s\{\log(pn)\}^{3/2} T\sqrt{\log(T/\delta)} . \tag{S.24}$$

Under (S.24), the smoothing parameter ϖ specified in (S.12) satisfies $\varpi \to 0$ as $n \to \infty$. For the event $\mathcal{E} = \mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_5 \cap \mathcal{E}_6 \cap \mathcal{E}_7$ with $\mathcal{E}_0, \ldots, \mathcal{E}_7$ defined in (S.21), by selecting

$$b_2 \asymp \sqrt{\frac{s\log p + \log n}{n}} , \quad b_3 \asymp \sqrt{\frac{s\log p + \log n}{n}} + \varpi^2 \ \text{ and } \ b_4 \asymp \sqrt{\frac{s\log(pn)}{n}} , \tag{S.25}$$

and applying Theorem T3, it holds that $\mathbb{P}(\mathcal{E}) \geq 1 - 7n^{-2}$. With (b_2, b_3, b_4) selected as in (S.25), the restriction $(b_2 + b_3 + b_4) \leq b_1 c_0/24$ in Theorem T2 holds automatically. Moreover, by (S.88) in Lemma L14, we know that $\bar{b}_3 \leq b_3$ holds. Then, it follows from Theorem T2 and (S.25) that

$$\begin{aligned} \|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2^2 &\lesssim \bigg(\frac{b_2 + b_3 + b_4}{b_1}\bigg)^2 + \frac{b_5}{b_1}\bigg(\frac{\log p}{n\epsilon}\bigg)^2 + \bigg[\frac{s\{\log(pn)\}^{3/2}}{n\epsilon}\bigg]^2 T^2 \log\bigg(\frac{T}{\delta}\bigg) \\ &\lesssim \frac{1}{f_l^2}\frac{s\log(pn)}{n} + \bigg[\frac{s\{\log(pn)\}^{3/2}}{n\epsilon}\bigg]^2 T^2 \log\bigg(\frac{T}{\delta}\bigg) \end{aligned} \tag{S.26}$$

conditioned on $\mathcal{E}$. We thus complete the proof of (S.13) in Theorem T1. □

EC.6.2.2 Proof of (S.14)

To prove (S.14) in Theorem T1, we require Lemma L9 below, which extends Lemma 1 from Jain et al. (2014) and Lemma A.3 from Cai et al. (2021) by accounting for the noise introduced during private support set selection. The proof of Lemma L9 is provided in Section EC.9.1.

LEMMA L9. *Let $\tilde{P}_s(\cdot)$ be defined as in Algorithm 2. For any $\boldsymbol{\xi} \in \mathbb{R}^p$, let S be the index set obtained by Algorithm 2 with the sparsity level s and taking $\boldsymbol{\xi}$ as the input vector. Moreover, let $\hat{\boldsymbol{\xi}} \in \mathbb{R}^p$ satisfy $\|\hat{\boldsymbol{\xi}}\|_0 = \hat{s} \le s$. Then, for any $c > 0$ and any index set $\tilde{S} \subseteq [p]$ such that $S \subseteq \tilde{S}$, the following inequality holds:*

$$\|\{\tilde{P}_s(\boldsymbol{\xi}) - \boldsymbol{\xi}\}_{\tilde{S}}\|_2^2 \le (1 + c^{-1}) \frac{|\tilde{S}| - s}{|\tilde{S}| - \hat{s}} \|(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi})_{\tilde{S}}\|_2^2 + 4(1 + c) \sum_{i \in [s]} \|\mathbf{w}_i\|_\infty^2 \,, \tag{S.27}$$

where $\{\mathbf{w}_1, \ldots, \mathbf{w}_s\}$ are specified in Algorithm 2.

Let $\mathbf{w}_1^t, \ldots, \mathbf{w}_s^t, \tilde{\mathbf{w}}^t \in \mathbb{R}^p$ for $t \in \{0\} \cup [T-1]$, with some $T \ge 1$, be independent random vectors generated in the noisy update step of Algorithm 3, where each entry independently follows the Laplace distribution $\mathrm{Laplace}(2\epsilon^{-1} T \lambda \sqrt{5s \log(T/\delta)})$ with $\lambda = 2\eta_0 \gamma \bar{\tau}/n$. Define

$$W_t := \sum_{i=1}^{s} \|\mathbf{w}_i^t\|_\infty^2 + \|\tilde{\mathbf{w}}_{|S^{t+1}}^t\|_2^2 \,, \quad t \in \{0\} \cup [T-1] \,, \tag{S.28}$$

where $S^{t+1} \subseteq [p]$ is the associated index set S determined by Algorithm 2 when $\mathbf{v}$ is chosen as $\tilde{\boldsymbol{\beta}}^{(t+1)}$ specified in Algorithm 3.

For each $t \in \{0\} \cup [T-1]$, notice that $\boldsymbol{\beta}^{(t+1)} = \tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) + \tilde{\mathbf{w}}_{|S^{t+1}}^t = \{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)})\}_{|I^t} + \tilde{\mathbf{w}}_{|S^{t+1}}^t$, where $I^t = S^{t+1} \cup S^t$ with $S^0 = \mathrm{supp}(\boldsymbol{\beta}^{(0)})$. Recall $\bar{K}_\varpi(\cdot) = \bar{K}(\cdot/\varpi)$. Since $\mathrm{supp}(\boldsymbol{\beta}^{(t)}) \subseteq S^t$, by (5), for each $t \in \{0\} \cup [T-1]$, we have

$$\begin{aligned}
\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2 &= \big\| \{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) - \tilde{\boldsymbol{\beta}}^{(t+1)}\}_{|I^t} + \tilde{\boldsymbol{\beta}}_{|I^t}^{(t+1)} - \boldsymbol{\beta}^{(t)} + \tilde{\mathbf{w}}_{|S^{t+1}}^t \big\|_2 \\
&= \Bigg\| \{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) - \tilde{\boldsymbol{\beta}}^{(t+1)}\}_{|I^t} + \underbrace{\boldsymbol{\beta}_{|I^t}^{(t)} - \boldsymbol{\beta}^{(t)}}_{=\mathbf{0}} - \frac{\eta_0}{n} \sum_{i=1}^{n} \{\bar{K}_\varpi(\langle \mathbf{x}_i, \boldsymbol{\beta}^{(t)} \rangle - d_i) - \tau\} w_\gamma(\mathbf{x}_i)_{|I^t} + \tilde{\mathbf{w}}_{|S^{t+1}}^t \Bigg\|_2 \\
&\le \big\| \{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) - \tilde{\boldsymbol{\beta}}^{(t+1)}\}_{|I^t} \big\|_2 + \eta_0 \bar{\tau} \sqrt{2s} \gamma + \|\tilde{\mathbf{w}}_{|S^{t+1}}^t\|_2 \,,
\end{aligned} \tag{S.29}$$

where the last inequality follows from the facts that $|\bar{K}_\varpi(\langle \mathbf{x}_i, \boldsymbol{\beta}^{(t)} \rangle - d_i) - \tau| \le \bar{\tau} := \max\{\tau, 1 - \tau\}$ and $\|w_\gamma(\mathbf{x}_i)_{|I^t}\|_2 \le \sqrt{2s}\gamma$. By Lemma L9 with $c = 1$, it holds that

$$\big\| \{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) - \tilde{\boldsymbol{\beta}}^{(t+1)}\}_{I^t} \big\|_2^2 \le 2 \frac{|I^t| - s}{|I^t| - \|\boldsymbol{\beta}^{(t)}\|_0} \big\| \{\boldsymbol{\beta}^{(t)} - \tilde{\boldsymbol{\beta}}^{(t+1)}\}_{I^t} \big\|_2^2 + 8 \sum_{i=1}^{s} \|\mathbf{w}_i^t\|_\infty^2$$

$$\overset{\text{(i)}}{\leq} 4\eta_0^2\bar{\tau}^2 s\gamma^2 + 8\sum_{i=1}^{s}\|\mathbf{w}_i^t\|_\infty^2\,, \tag{S.30}$$

where (i) follows from the facts that $|I^t| \leq 2s$, $\|\boldsymbol{\beta}^{(t)}\|_0 \leq s$ and $\|\{\boldsymbol{\beta}^{(t)} - \tilde{\boldsymbol{\beta}}^{(t+1)}\}_{I^t}\|_2^2 \leq (\eta_0\bar{\tau}\sqrt{2s}\gamma)^2$. Combining (S.29) and (S.30) yields

$$\begin{aligned}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2 &\leq 2\eta_0\bar{\tau}\sqrt{s}\gamma + 2\sqrt{2}\sqrt{\sum_{i=1}^{s}\|\mathbf{w}_i^t\|_\infty^2} + \sqrt{2}\eta_0\bar{\tau}\sqrt{s}\gamma + \|\tilde{\mathbf{w}}^t_{|S^{t+1}}\|_2 \\ &\leq 4\eta_0\bar{\tau}\sqrt{s}\gamma + 4\sqrt{W_t}\end{aligned} \tag{S.31}$$

for each $t \in \{0\}\cup[T-1]$, where W_t is specified in (S.28). Therefore, by (S.31), we have

$$\begin{aligned}\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2^2 &\leq \left(\sum_{t=0}^{T-1}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2 + \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2\right)^2 \\ &\leq 2c_0^2 + 64T^2\eta_0^2\bar{\tau}^2 s\gamma^2 + 64\left(\sum_{t=0}^{T-1}\sqrt{W_t}\right)^2,\end{aligned} \tag{S.32}$$

where the last inequality follows from the initialization condition $\boldsymbol{\beta}^{(0)} \in \mathbb{H}(s)\cap\Theta(c_0)$.

Notice that $\tilde{w}_j^t \sim \text{Laplace}(B)$ and $w_{i,j}^t \sim \text{Laplace}(B)$ for any $i \in [s]$, $j \in S^{t+1}$ and $t \in \{0\}\cup[T-1]$, where

$$B = \frac{4\eta_0\gamma\bar{\tau}}{n\epsilon}T\sqrt{5s\log\left(\frac{T}{\delta}\right)}\,.$$

Then, for any $j \in S^{t+1}$ and $t \in \{0\}\cup[T-1]$, we have

$$\mathbb{E}(|\tilde{w}_j^t|^4) = \int_{-\infty}^{\infty}|z|^4\frac{1}{2B}e^{-|z|/B}\,\mathrm{d}z = \frac{1}{B}\int_0^{\infty}z^4e^{-z/B}\,\mathrm{d}z = B^4\Gamma(5) = 24B^4\,. \tag{S.33}$$

Moreover, for any $i \in [s]$ and $t \in \{0\}\cup[T-1]$, it holds that

$$\begin{aligned}\mathbb{E}(\|\mathbf{w}_i^t\|_\infty^4) = &\int_0^{B^4\log^4 p}\mathbb{P}\left(\max_{j\in[p]}|w_{i,j}^t| > z^{1/4}\right)\mathrm{d}z + \int_{B^4\log^4 p}^{\infty}\mathbb{P}\left(\max_{j\in[p]}|w_{i,j}^t| > z^{1/4}\right)\mathrm{d}z \\ &\leq B^4\log^4 p + 4p\int_{B\log p}^{\infty}z^3\max_{j\in[p]}\mathbb{P}(|w_{i,j}^t| > z)\,\mathrm{d}z \\ &\leq B^4\log^4 p + 4p\int_{B\log p}^{\infty}z^3e^{-z/B}\,\mathrm{d}z = B^4\log^4 p + 4pB^4\int_{\log p}^{\infty}z^3e^{-z}\,\mathrm{d}z \\ &\leq B^4\log^4 p + 64B^4\log^3 p \leq 65B^4\log^4 p\,.\end{aligned} \tag{S.34}$$

Therefore, by the Minkowski inequality, along with (S.33) and (S.34), it holds that

$$\begin{aligned}\bigg\{\mathbb{E}\bigg(\bigg|\sum_{t=0}^{T-1}\sqrt{W_t}\bigg|^4\bigg)\bigg\}^{1/4} &\leq \sum_{t=0}^{T-1}\big\{\mathbb{E}(W_t^2)\big\}^{1/4} = \sum_{t=0}^{T-1}\bigg\{\mathbb{E}\bigg(\bigg|\sum_{i=1}^{s}\|\mathbf{w}_i^t\|_\infty^2 + \sum_{j\in S^{t+1}}|\tilde{w}_j^t|^2\bigg|^2\bigg)\bigg\}^{1/4} \\ &\leq \sum_{t=0}^{T-1}\bigg[\sum_{i=1}^{s}\{\mathbb{E}(\|\mathbf{w}_i^t\|_\infty^4)\}^{1/2} + \sum_{j\in S^{t+1}}\big\{\mathbb{E}(|\tilde{w}_j^t|^4)\big\}^{1/2}\bigg]^{1/2} \\ &\leq 4T\sqrt{s}B\log p\,. \end{aligned}\tag{S.35}$$

From (S.26), we deduce that there exists an event $\tilde{\mathcal{E}}$ such that $\mathbb{P}(\tilde{\mathcal{E}}) \geq 1 - C_0 n^{-2}$, where $C_0 > 0$ is an absolute constant, such that on this event, we have

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2^2 \leq r_n \asymp \frac{1}{f_l^2}\frac{s\log(pn)}{n} + \left(\frac{sT}{n\epsilon}\right)^2 \log\left(\frac{T}{\delta}\right)\log^3(pn)\,.$$

Therefore, by the Cauchy–Schwarz inequality, along with (S.32) and (S.35), it holds that

$$\begin{aligned}\mathbb{E}(\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2^2) &= \mathbb{E}\{\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2^2\mathbb{1}(\tilde{\mathcal{E}})\} + \mathbb{E}\{\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2^2\mathbb{1}(\tilde{\mathcal{E}}^{\mathrm{c}})\} \\ &\leq r_n + \mathbb{E}\bigg[\bigg\{2c_0^2 + 64T^2\eta_0^2\bar{\tau}^2 s\gamma^2 + 64\bigg(\sum_{t=0}^{T-1}\sqrt{W_t}\bigg)^2\bigg\}\mathbb{1}(\tilde{\mathcal{E}}^{\mathrm{c}})\bigg] \\ &\leq r_n + (2c_0^2 + 64T^2\eta_0^2\bar{\tau}^2 s\gamma^2)\cdot\frac{C_0}{n^2} + 64\bigg\{\mathbb{E}\bigg(\bigg|\sum_{t=0}^{T-1}\sqrt{W_t}\bigg|^4\bigg)\bigg\}^{1/2}\cdot\frac{C_0^{1/2}}{n} \\ &\leq r_n + (2c_0^2 + 64T^2\eta_0^2\bar{\tau}^2 s\gamma^2)\cdot\frac{C_0}{n^2} + 64\cdot 16T^2 sB^2\log^2 p\cdot\frac{C_0^{1/2}}{n}\,. \end{aligned}\tag{S.36}$$

Since $\gamma \asymp \sqrt{\log(pn)}$ and $T \asymp \log(n\epsilon/\log p)$, we can conclude that

$$(c_0^2 + T^2\eta_0^2\bar{\tau}^2 s\gamma^2)\cdot\frac{1}{n^2} + T^2 sB^2\log^2 p\cdot\frac{1}{n} \lesssim r_n\,.\tag{S.37}$$

Combining (S.36) and (S.37), we have

$$\mathbb{E}(\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2^2) \lesssim r_n\,.\tag{S.38}$$

This completes the proof of (S.14) in Theorem T1, by applying the Cauchy–Schwarz inequality. □

EC.6.3. Proof of Corollary C1

EC.6.3.1 Proof of (S.15)

Following from the proof of Theorem T1, and noting that $\eta_0 \leq f_l\lambda_p/(40f_u^2\lambda_1^2)$, we conclude that with if $\varpi \leq f_l\lambda_p/(10l_f k_1\lambda_1)$, it holds that $\eta_0 \leq 2b_1/(9b_5^2)$ and $b_1 \leq b_5/2$. Since c_0 is a constant, by choosing $T \asymp \log(n/\log p)$ appropriately, we have $T \gtrsim \log(c_0 n/\log p)$. Given that $n \gtrsim s\log(pn)$, with (b_2, b_3, b_4) selected as in (S.25), the restriction $(b_2 + b_3 + b_4) \lesssim b_1 c_0$ in Corollary C2 holds automatically. Selecting $z = 2\log n$. Then, for such selection of (b_2, b_3, b_4), by Theorem T3, the event $\mathcal{E} = \mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_5 \cap \mathcal{E}_6 \cap \mathcal{E}_7$ with $\mathcal{E}_0, \ldots, \mathcal{E}_7$ defined in (S.21) satisfies $\mathbb{P}(\mathcal{E}) \geq 1 - 7n^{-2}$. Finally, according to Corollary C2 and (S.25),

$$\|\boldsymbol{\beta}^{(T)}_{\mathrm{IHT}} - \boldsymbol{\beta}^*\|_2 \lesssim \frac{b_2 + b_3 + b_4}{b_1} + \sqrt{\frac{b_5}{b_1}}\frac{\log p}{n} \lesssim \frac{1}{f_l}\sqrt{\frac{s\log(pn)}{n}} \tag{S.39}$$

conditioned on $\mathcal{E}$. We complete the proof of (S.15) in Corollary C1. □

EC.6.3.2 Proof of (S.16)

To prove (S.16) in Corollary C1, we need Lemma L10 below, which extends Lemma 1 of Jain et al. (2014). Its proof is given in Section EC.9.2.

LEMMA L10. *Let P_s denote a nonlinear operator on $\mathbb{R}^p$ that preserves the s largest entries (in magnitude) of a vector $\boldsymbol{\xi} \in \mathbb{R}^p$ and zeros out the rest. For any $\boldsymbol{\xi} \in \mathbb{R}^p$, let S be the index set obtained by P_s with the sparsity level s and taking $\boldsymbol{\xi}$ as the input vector. Moreover, let $\hat{\boldsymbol{\xi}} \in \mathbb{R}^p$ satisfy $\|\hat{\boldsymbol{\xi}}\|_0 = \hat{s} \leq s$. Then, for any index set $\tilde{S} \subseteq [p]$ such that $S \subseteq \tilde{S}$, the following inequality holds:*

$$\|\{P_s(\boldsymbol{\xi}) - \boldsymbol{\xi}\}_{\tilde{S}}\|_2^2 \leq \frac{|\tilde{S}| - s}{|\tilde{S}| - \hat{s}}\|(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi})_{\tilde{S}}\|_2^2\,. \tag{S.40}$$

According to Algorithm 1, we know that $\boldsymbol{\beta}^{(t+1)} = P_s(\boldsymbol{\beta}^{(t)} - \eta_0 \mathbf{g}^{(t)})$, where $\mathbf{g}^{(t)} = \nabla \widehat{R}_{\varpi}(\boldsymbol{\beta}^{(t)}) = n^{-1}\sum_{i=1}^n \{\bar{K}_{\varpi}(\langle \mathbf{x}_i, \boldsymbol{\beta}^{(t)}\rangle - d_i) - \tau\}\mathbf{x}_i$. For each $t \in \{0\} \cup [T-1]$, write $I_*^t = S_*^t \cup S_*^{t+1}$, where $S_*^t = \mathrm{supp}(\boldsymbol{\beta}^{(t)}) \subseteq \mathbb{R}^p$. Since $P_s(\boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)}) = \{P_s(\boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)})\}_{|I_*^t}$ and $\boldsymbol{\beta}^{(t)}_{|I_*^t} = \boldsymbol{\beta}^{(t)}$, we have

$$\begin{aligned}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2 &= \big\|\{P_s(\boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)}) - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_{|I_*^t} + \boldsymbol{\beta}^{(t)}_{|I_*^t} - \boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)}_{|I_*^t}\big\|_2 \\ &\leq \big\|\{P_s(\boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)}) - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_{|I_*^t}\big\|_2 + \eta_0\|\mathbf{g}^{(t)}_{|I_*^t}\|_2\,.\end{aligned} \tag{S.41}$$

By Lemma L10, it holds that

$$\big\|\{P_s(\boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)}) - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_{|I_*^t}\big\|_2^2 \leq \big\|\{\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_{|I_*^t}\big\|_2^2 \leq \eta_0^2\|\mathbf{g}^{(t)}_{|I_*^t}\|_2^2\,. \tag{S.42}$$

Combining (S.41) and (S.42) yields

$$\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\|_2 \le 2\eta_0\|\mathbf{g}^{(t)}_{|I^t_*}\|_2 \tag{S.43}$$

for each $t\in\{0\}\cup[T-1]$. Therefore, by (S.43), we have

$$\|\boldsymbol{\beta}^{(T)}_{\mathrm{IHT}}-\boldsymbol{\beta}^*\|_2^2 \le \bigg(\sum_{t=0}^{T-1}\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\|_2+\|\boldsymbol{\beta}^{(0)}-\boldsymbol{\beta}^*\|_2\bigg)^2 \le 2c_0^2+8\eta_0^2\bigg(\sum_{t=0}^{T-1}\|\mathbf{g}^{(t)}_{|I^t_*}\|_2\bigg)^2, \tag{S.44}$$

where the last inequality follows from the initialization condition $\boldsymbol{\beta}^{(0)}\in\mathbb{H}(s)\cap\Theta(c_0)$.

Let $\mathbf{u}=\boldsymbol{\Sigma}^{1/2}\mathbf{e}_j/\|\boldsymbol{\Sigma}^{1/2}\mathbf{e}_j\|_2\in\mathbb{S}^{p-1}$, where $\mathbf{e}_j$ is the unit vector with 1 in the j-th position and 0 elsewhere. By Assumption 1, due to $\|\boldsymbol{\Sigma}^{1/2}\mathbf{e}_j\|_2\le\lambda_1^{1/2}$, for any $i\in[n]$, $j\in[p]\backslash\{1\}$ and $z\ge 0$,

$$\mathbb{P}(|x_{i,j}|\ge \upsilon_1\lambda_1^{1/2}z)\le\mathbb{P}(|x_{i,j}|\ge\|\boldsymbol{\Sigma}^{1/2}\mathbf{e}_j\|_2\upsilon_1 z)=\mathbb{P}(|\langle\mathbf{u},\boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i\rangle|\ge\upsilon_1 z)\le 2e^{-z^2/2}. \tag{S.45}$$

Then, by (S.45), for any $i\in[n]$, $j\in[p]\backslash\{1\}$, it holds that

$$\begin{aligned}\mathbb{E}(|x_{i,j}|^6) &= \int_0^\infty \mathbb{P}(|x_{i,j}|\ge z^{1/6})\,\mathrm{d}z = 6\int_0^\infty z^5\mathbb{P}(|x_{i,j}|\ge z)\,\mathrm{d}z\\ &\le 12\int_0^\infty z^5 e^{-z^2/(2\upsilon_1^2\lambda_1)}\,\mathrm{d}z = 48(\upsilon_1\lambda_1^{1/2})^6\Gamma(3)=96(\upsilon_1\lambda_1^{1/2})^6\,. \end{aligned}\tag{S.46}$$

Since $x_{i,1}=1$ for all $i\in[n]$, by (S.46), we have

$$\mathbb{E}(|x_{i,j}|^6)\le\max\{1,96(\upsilon_1\lambda_1^{1/2})^6\} \tag{S.47}$$

for all $i\in[n]$ and $j\in[p]$. Then, by (S.47), for each $t\in\{0\}\cup[T-1]$, it holds that

$$\begin{aligned}\mathbb{E}(\|\mathbf{g}^{(t)}_{|I^t_*}\|_2^6) &= \mathbb{E}\bigg[\bigg\|\frac{1}{n}\sum_{i=1}^n\{\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\beta}^{(t)}\rangle-d_i)-\tau\}\mathbf{x}_{i,I^t_*}\bigg\|_2^6\bigg]\\ &\overset{\text{(i)}}{\le}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\big[\|\{\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\beta}^{(t)}\rangle-d_i)-\tau\}\mathbf{x}_{i,I^t_*}\|_2^6\big]\\ &\overset{\text{(ii)}}{\le}\frac{\bar{\tau}^6}{n}\sum_{i=1}^n\mathbb{E}\bigg\{\bigg(\sum_{j\in I^t_*}|x_{i,j}|^2\bigg)^3\bigg\}\overset{\text{(iii)}}{\le}\frac{\bar{\tau}^6}{n}\sum_{i=1}^n\bigg[\sum_{j\in I^t_*}\{\mathbb{E}(|x_{i,j}|^6)\}^{1/3}\bigg]^3\\ &\le\bar{\tau}^6(2s)^3\max\{1,96(\upsilon_1\lambda_1^{1/2})^6\}\,,\end{aligned}\tag{S.48}$$

where (i) follows from the Jensen inequality, (ii) follows from the fact that $|\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\beta}^{(t)}\rangle-d_i)-\tau|\le\bar{\tau}:=\max\{\tau,1-\tau\}$, (iii) follows from the Minkowski inequality, and the last inequality follows from the fact that $|I^t_*|\le 2s$. Therefore, by the Minkowski inequality, along with (S.48), it holds that

$$\left\{\mathbb{E}\left(\left|\sum_{t=0}^{T-1}\|\mathbf{g}^{(t)}_{|I^t_*}\|_2\right|^6\right)\right\}^{1/6}\le\sum_{t=0}^{T-1}\{\mathbb{E}(\|\mathbf{g}^{(t)}_{|I^t_*}\|_2^6)\}^{1/6}\le T\bar{\tau}\sqrt{2s}\max\{1,96^{1/6}\upsilon_1\lambda_1^{1/2}\}\,. \tag{S.49}$$

From (S.39), we deduce that there exists an event $\tilde{\mathcal{E}}_*$ such that $\mathbb{P}(\tilde{\mathcal{E}}_*)\ge 1-C_*n^{-2}$, where $C_*>0$ is an absolute constant, such that on this event, we have

$$\|\boldsymbol{\beta}^{(T)}_{\mathrm{IHT}}-\boldsymbol{\beta}^*\|_2^2\le r_{n,*}\asymp\frac{1}{f_l^2}\frac{s\log(pn)}{n}\,.$$

Therefore, by the Hölder inequality, along with (S.44) and (S.49), it holds that

$$\begin{aligned}\mathbb{E}(\|\boldsymbol{\beta}^{(T)}_{\mathrm{IHT}}-\boldsymbol{\beta}^*\|_2^2)&=\mathbb{E}\{\|\boldsymbol{\beta}^{(T)}_{\mathrm{IHT}}-\boldsymbol{\beta}^*\|_2^2\mathbb{1}(\tilde{\mathcal{E}}_*)\}+\mathbb{E}\{\|\boldsymbol{\beta}^{(T)}_{\mathrm{IHT}}-\boldsymbol{\beta}^*\|_2^2\mathbb{1}(\tilde{\mathcal{E}}_*^{\mathrm{c}})\}\\&\le r_{n,*}+\mathbb{E}\left[\left\{2c_0^2+8\eta_0^2\left(\sum_{t=0}^{T-1}\|\mathbf{g}^{(t)}_{|I^t_*}\|_2\right)^2\right\}\mathbb{1}(\tilde{\mathcal{E}}_*^{\mathrm{c}})\right]\\&\le r_{n,*}+2c_0^2\cdot\frac{C_*}{n^2}+8\eta_0^2\left\{\mathbb{E}\left(\left|\sum_{t=0}^{T-1}\|\mathbf{g}^{(t)}_{|I^t_*}\|_2\right|^6\right)\right\}^{1/3}\cdot\frac{C_*^{2/3}}{n^{4/3}}\\&\le r_{n,*}+2c_0^2\cdot\frac{C_*}{n^2}+16\eta_0^2T^2\bar{\tau}^2s\max\{1,5\upsilon_1^2\lambda_1\}\cdot\frac{C_*^{2/3}}{n^{4/3}}\,.\end{aligned} \tag{S.50}$$

Since $T\asymp\log(n/\log p)$, we can conclude that

$$c_0^2\cdot\frac{C_*}{n^2}+\eta_0^2T^2\bar{\tau}^2s\max\{1,5\upsilon_1^2\lambda_1\}\cdot\frac{C_*^{2/3}}{n^{4/3}}\lesssim r_{n,*}\,. \tag{S.51}$$

Combining (S.50) and (S.51), we have

$$\mathbb{E}(\|\boldsymbol{\beta}^{(T)}_{\mathrm{IHT}}-\boldsymbol{\beta}^*\|_2^2)\lesssim r_{n,*}\,. \tag{S.52}$$

This completes the proof of (S.16) in Corollary C1, by applying the Cauchy–Schwarz inequality. □

EC.7. Proof of Theorem T2

To prove Theorem T2, we first introduce Lemma L11, which provides a high probability bound on the maximum scale of noise injected per iteration due to the ‘peeling’ process of Algorithm 3. The proof of Lemma L11 is given in Section EC.9.3.

LEMMA L11. *For any $z>0$, the sequence $\{W_t\}_{t=0}^{T-1}$ defined in* (S.28) *satisfies*

$$\max_{t\in\{0\}\cup[T-1]}W_t\le C_0\left(\frac{sT\lambda}{\epsilon}\right)^2\{\log(psT)+z\}^2\log\left(\frac{T}{\delta}\right)=:\Delta_n(z) \tag{S.53}$$

with probability at least $1-e^{-z}$ *(over* $\{\mathbf{w}_1^t,\ldots,\mathbf{w}_s^t,\tilde{\mathbf{w}}^t\}_{t=0}^{T-1}$*), where* $C_0>0$ *is an absolute constant, and* $\lambda=2\eta_0\gamma\bar{\tau}/n$.

Proposition P1 establishes the RSC and RSM properties of the empirical smoothed quantile loss, whose proof is provided in Section EC.7.3.

PROPOSITION P1 (**Restricted Strong Convexity and Smoothness**). *For* $\bar{b}_3$ *defined in* (S.22)*, it holds that*

$$\widehat{R}_{\varpi}(\boldsymbol{\beta})-\widehat{R}_{\varpi}(\boldsymbol{\beta}^*)\geq b_1\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2-(b_2+\bar{b}_3)\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2$$
$$\text{conditioned on } \mathcal{E}_1\cap\mathcal{E}_2 \text{ for all } \boldsymbol{\beta}\in\mathbb{H}(s)\cap\{\Theta(c_0)\backslash\Theta(n^{-1})\}\,, \tag{S.54}$$
$$\widehat{R}_{\varpi}(\boldsymbol{\beta}^*)-\widehat{R}_{\varpi}(\boldsymbol{\beta})-\langle\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta}),\boldsymbol{\beta}^*-\boldsymbol{\beta}\rangle\geq b_1\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2-(b_2+b_4)\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2$$
$$\text{conditioned on } \mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_4 \text{ for all } \boldsymbol{\beta}\in\mathbb{H}(s)\cap\{\Theta(c_0)\backslash\Theta(n^{-1})\}\,, \tag{S.55}$$
$$\widehat{R}_{\varpi}(\boldsymbol{\beta}_1)-\widehat{R}_{\varpi}(\boldsymbol{\beta}_2)-\langle\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta}_2),\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\rangle\leq\frac{b_5}{2}\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_2^2$$
$$\text{conditioned on } \mathcal{E}_5 \text{ for all } \boldsymbol{\beta}_1,\,\boldsymbol{\beta}_2\in\mathbb{H}(s)\cap\Theta(c_0)\,. \tag{S.56}$$

Based on Proposition P1, Proposition P2 asserts that, for any $r_0\in(n^{-1},c_0]$, if $\boldsymbol{\beta}^{(t)}\in\mathbb{H}(s)\cap\{\Theta(r_0)\backslash\Theta(n^{-1})\}$ for some $t\in\{0\}\cup[T-1]$, then the $(t+1)$-th DP iterate $\boldsymbol{\beta}^{(t+1)}$ obtained by Algorithm 3 will remain within $\Theta(r_0)$. The proof of Proposition P2 is given in Section EC.7.4.

PROPOSITION P2. *Assume that* $\boldsymbol{\beta}^{(t)}\in\mathbb{H}(s)\cap\{\Theta(r_0)\backslash\Theta(n^{-1})\}$ *for some* $t\in\{0\}\cup[T-1]$ *and* $n^{-1}<r_0\leq c_0$*, and the learning rate satisfies* $\eta_0\leq 2b_1/(9b_5^2)$ *with* $b_1\leq b_5/2$*. Then, conditioned on* $\mathcal{E}_0\cap\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4\cap\mathcal{E}_6$*, the* $(t+1)$*-th DP iterate* $\boldsymbol{\beta}^{(t+1)}$ *obtained by Algorithm* 3 *satisfies* $\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^*\|_2<r_0$*, provided that*

$$(b_2+b_3+b_4)\leq\frac{b_1r_0}{24}\,,\ s\geq 38s^*(\eta_0b_1)^{-2}\,,\ \textit{and}\ \max_{t\in\{0\}\cup[T-1]}W_t\leq\frac{\eta_0^2b_1^2r_0^2}{4000}\,, \tag{S.57}$$

where the quantities (b_1,b_2,b_3,b_4,b_5) *and* W_t *are specified in* (S.21) *and* (S.28)*, respectively.*

While Proposition P2 imposes the condition $\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\|_2>n^{-1}$ as a technical assumption, Proposition P3 shows that even when this condition is violated, the $(t+1)$-th DP iterate $\boldsymbol{\beta}^{(t+1)}$ obtained by Algorithm 3 still remains within a suitable neighborhood of $\boldsymbol{\beta}^*$. The proof of Proposition P3 is given in Section EC.7.5.

PROPOSITION P3. *Assume that* $\boldsymbol{\beta}^{(t)}\in\mathbb{H}(s)\cap\Theta(n^{-1})$ *for some* $t\in\{0\}\cup[T-1]$*, and the learning rate satisfies* $\eta_0\leq 2b_1/(9b_5^2)$ *with* $b_1\leq b_5/2$*. Then, for any* $z>0$*, conditioned on* $\mathcal{E}_0\cap\mathcal{E}_3\cap\mathcal{E}_7$*,* *the* $(t+1)$*-th DP iterate* $\boldsymbol{\beta}^{(t+1)}$ *obtained by Algorithm* 3 *satisfies*

$$\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^*\|_2 \le \frac{2}{n}+\frac{b_3}{12b_1}+4\Delta_n^{1/2}(z)\,,$$

provided that $s \ge 38s^*(\eta_0 b_1)^{-2}$ *and* $\max_{t\in\{0\}\cup[T-1]} W_t \le \Delta_n(z)$*, where the quantities* (b_1,b_3,b_5)*,* W_t *and* $\Delta_n(z)$ *are specified in* (S.21)*,* (S.28) *and* (S.53)*, respectively.*

For any $z>0$, let

$$r(z)=\frac{24(b_2+b_3+b_4)}{b_1}+\frac{2}{n}+4\Delta_n^{1/2}(z)\,.$$

Assume $r(z)\le c_0$. Based on Propositions P2 and P3, given any $\boldsymbol{\beta}^{(0)}\in\mathbb{H}(s)\cap\Theta(c_0)$, we know that all subsequent iterates obtained by Algorithm 3 will remain within the region $\Theta(c_0)$. Let $T\ge 2\log(c_0 n\epsilon/\log p)/\log\{(1-\rho)^{-1}\}$ with $\rho=4\eta_0 b_5 s^*/s+\eta_0^2 b_1 b_5/12$. In the sequel, for any $z>0$, if $\max_{t\in\{0\}\cup[T-1]} W_t\le\Delta_n(z)$, we will show in Sections EC.7.1 and EC.7.2 that

$$\|\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*\|_2^2 \lesssim \frac{(b_2+b_3+b_4)^2}{b_1^2}+\frac{b_5}{b_1}\left(\frac{\log p}{n\epsilon}\right)^2+\Delta_n(z)\max\left\{\frac{C_{\eta_0}}{b_1\rho},1\right\} \tag{S.58}$$

holds in both cases: (i) $\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\|_2\le r(z)$ for some $t\in\{0\}\cup[T-1]$, and (ii) $\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\|_2> r(z)$ for all $t\in\{0\}\cup[T-1]$, where $C_{\eta_0}=4\{b_5(5-\eta_0 b_5)(2-2\eta_0 b_5)^{-1}+3(1-\eta_0 b_5)\eta_0^{-1}\}$. Then, together with Lemma L11, we have Theorem T2. □

EC.7.1. Case (i): $\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\|_2\le r(z)$ for some $t\in\{0\}\cup[T-1]$

If $\boldsymbol{\beta}^{(t)}\in\mathbb{H}(s)$ satisfies $R:=\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\|_2\le r(z)$ for some $t\in\{0\}\cup[T-1]$, we derive the upper bound of $\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^*\|_2$ under two scenarios: (a) $R\le n^{-1}$ and (b) $n^{-1}<R\le r(z)$. For Scenario (a), Proposition P3 guarantees that $\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^*\|_2\le r(z)$. For Scenario (b), since $n^{-1}<R\le r(z)$, Proposition P2 with $r_0=r(z)$ ensures that $\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^*\|_2\le r(z)$. Hence, if $\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\|_2\le r(z)$, then it always holds that $\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^*\|_2\le r(z)$. Moreover, since $b_1\le b_5/2$, then

$$\|\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*\|_2^2\le r^2(z)\lesssim \frac{(b_2+b_3+b_4)^2}{b_1^2}+\frac{b_5}{b_1}\left(\frac{\log p}{n\epsilon}\right)^2+\Delta_n(z)\max\left\{\frac{C_{\eta_0}}{b_1\rho},1\right\},$$

which indicates that (S.58) holds if there exists some $t\in\{0\}\cup[T-1]$ such that $\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\|_2\le r(z)$. □

EC.7.2. Case (ii): $\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\|_2> r(z)$ for all $t\in\{0\}\cup[T-1]$

Since $\|\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*\|_2\le c_0$, if $\|\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*\|_2\le r(z)$, then (S.58) holds automatically. In the following, we only need to show (S.58) under the scenario $r(z)<\|\boldsymbol{\beta}^{(T)}-\boldsymbol{\beta}^*\|_2\le c_0$. Conditioned on $\mathcal{E}_1\cap\mathcal{E}_2$, it follows from (S.54) in Proposition P1 that

$$\widehat{R}_{\varpi}(\boldsymbol{\beta}^{(T)}) - \widehat{R}_{\varpi}(\boldsymbol{\beta}^{*}) \geq -(b_2 + \bar{b}_3)\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^{*}\|_2 + b_1\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^{*}\|_2^2 .$$

Due to $|ab| \leq ca^2 + (4c)^{-1}b^2$ for all $a, b \in \mathbb{R}$ and $c > 0$, it holds that

$$\begin{aligned} b_1\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^{*}\|_2^2 &\leq (b_2 + \bar{b}_3)\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^{*}\|_2 + \widehat{R}_{\varpi}(\boldsymbol{\beta}^{(T)}) - \widehat{R}_{\varpi}(\boldsymbol{\beta}^{*}) \\ &\leq \frac{b_1}{2}\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^{*}\|_2^2 + \frac{(b_2 + \bar{b}_3)^2}{2b_1} + \widehat{R}_{\varpi}(\boldsymbol{\beta}^{(T)}) - \widehat{R}_{\varpi}(\boldsymbol{\beta}^{*}) . \end{aligned}$$

Provided that $\bar{b}_3 \leq b_3$, we have

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^{*}\|_2^2 \leq \frac{(b_2 + b_3)^2}{b_1^2} + \frac{2}{b_1}\{\widehat{R}_{\varpi}(\boldsymbol{\beta}^{(T)}) - \widehat{R}_{\varpi}(\boldsymbol{\beta}^{*})\} . \tag{S.59}$$

To complete the proof, it remains to establish an upper bound for $\widehat{R}_{\varpi}(\boldsymbol{\beta}^{(T)}) - \widehat{R}_{\varpi}(\boldsymbol{\beta}^{*})$. To this end, we need Proposition P4 below, which provides an iterative bound on the excess risk, $\widehat{R}_{\varpi}(\boldsymbol{\beta}^{(t)}) - \widehat{R}_{\varpi}(\boldsymbol{\beta}^{*})$. The proof of Proposition P4 is given in Section EC.7.6.

PROPOSITION P4. *Assume that $\boldsymbol{\beta}^{(t)} \in \mathbb{H}(s) \cap \{\Theta(c_0) \backslash \Theta(n^{-1})\}$ for some $t \in \{0\} \cup [T-1]$, and the learning rate satisfies $\eta_0 \leq 2b_1/(9b_5^2)$ with $b_1 \leq b_5/2$. Moreover, let the conditions* (S.57) *in Proposition* P2 *hold with $r_0 = c_0$. Then, conditioned on $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4 \cap \mathcal{E}_5$ and the restriction $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{*}\|_2 \geq 24(b_2 + b_4)/b_1$, the $(t+1)$-th iterate $\boldsymbol{\beta}^{(t+1)}$ obtained in Algorithm* 3 *satisfies*

$$\widehat{R}_{\varpi}(\boldsymbol{\beta}^{(t+1)}) - \widehat{R}_{\varpi}(\boldsymbol{\beta}^{*}) \leq (1-\rho)\{\widehat{R}_{\varpi}(\boldsymbol{\beta}^{(t)}) - \widehat{R}_{\varpi}(\boldsymbol{\beta}^{*})\} + C_{\eta_0}W_t ,$$

where $\rho = 4\eta_0 b_5 s^*/s + \eta_0^2 b_1 b_5/12$ and $C_{\eta_0} = 4\{b_5(5 - \eta_0 b_5)(2 - 2\eta_0 b_5)^{-1} + 3(1 - \eta_0 b_5)\eta_0^{-1}\}$.

Based on Propositions P2 and P3, given any $\boldsymbol{\beta}^{(0)} \in \mathbb{H}(s) \cap \Theta(c_0)$, we know that all subsequent iterates obtained by Algorithm 3 will remain within the region $\Theta(c_0)$. Due to $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{*}\|_2 > r(z) > n^{-1}$ for all $t \in \{0\} \cup [T-1]$, we know that $\boldsymbol{\beta}^{(t)} \in \mathbb{H}(s) \cap \{\Theta(c_0) \backslash \Theta(n^{-1})\}$ for all $t \in \{0\} \cup [T-1]$. Then, by Proposition P4, summing over $t \in \{0\} \cup [T-1]$ yields

$$\widehat{R}_{\varpi}(\boldsymbol{\beta}^{(T)}) - \widehat{R}_{\varpi}(\boldsymbol{\beta}^{*}) \leq (1-\rho)^T\{\widehat{R}_{\varpi}(\boldsymbol{\beta}^{(0)}) - \widehat{R}_{\varpi}(\boldsymbol{\beta}^{*})\} + C_{\eta_0}\sum_{t=0}^{T-1}(1-\rho)^{T-1-t}W_t .$$

Since $\boldsymbol{\beta}^{(0)} \in \mathbb{H}(s) \cap \Theta(c_0)$, conditioned on $\mathcal{E}_3 \cap \mathcal{E}_5$ with $b_3 \leq b_1 c_0/24$, we have

$$\begin{aligned} \widehat{R}_{\varpi}(\boldsymbol{\beta}^{(0)}) - \widehat{R}_{\varpi}(\boldsymbol{\beta}^{*}) &\leq \langle \nabla \widehat{R}_{\varpi}(\boldsymbol{\beta}^{*}), \boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^{*} \rangle + \frac{b_5}{2}\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^{*}\|_2^2 \\ &\leq \sup_{S \in \mathcal{S}} \|\{\nabla \widehat{R}_{\varpi}(\boldsymbol{\beta}^{*})\}_S\|_2 \cdot \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^{*}\|_2 + \frac{b_5}{2}\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^{*}\|_2^2 \end{aligned}$$

$$\leq b_3 c_0 + \frac{b_5 c_0^2}{2} \leq \bigg(\frac{b_5}{2} + \frac{b_1}{24}\bigg) c_0^2\,,$$

which implies

$$\widehat{R}_\varpi(\boldsymbol{\beta}^{(T)}) - \widehat{R}_\varpi(\boldsymbol{\beta}^*) \leq \bigg(\frac{b_5}{2} + \frac{b_1}{24}\bigg)(1-\rho)^T c_0^2 + C_{\eta_0}\sum_{t=0}^{T-1}(1-\rho)^{T-1-t}W_t\,.$$

For any $T \geq 2\log(c_0 n\epsilon/\log p)/\log\{(1-\rho)^{-1}\}$, we have $(1-\rho)^T c_0^2 \leq \{(n\epsilon)^{-1}\log p\}^2$. Due to $\sum_{t=0}^{T-1}(1-\rho)^{T-1-t} \leq \rho^{-1}$ and $b_1 \leq b_5/2$, it follows from (S.59) that

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_2^2 \lesssim \frac{(b_2+b_3+b_4)^2}{b_1^2} + \frac{b_5}{b_1}\bigg(\frac{\log p}{n\epsilon}\bigg)^2 + \Delta_n(z)\max\bigg\{\frac{C_{\eta_0}}{b_1\rho}, 1\bigg\}\,,$$

provided that $\max_{t\in\{0\}\cup[T-1]} W_t \leq \Delta_n(z)$. Hence, (S.58) holds if $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 > r(z)$ for all $t \in \{0\}\cup[T-1]$. $\square$

EC.7.3. Proof of Proposition P1

We will show (S.54), (S.55) and (S.56) in Sections EC.7.3.1, EC.7.3.2 and EC.7.3.3, respectively.

EC.7.3.1 Proof of (S.54)

Recall that $\widehat{D}_\varpi(\boldsymbol{\delta}) = \widehat{R}_\varpi(\boldsymbol{\beta}) - \widehat{R}_\varpi(\boldsymbol{\beta}^*)$, $D_\varpi(\boldsymbol{\delta}) = \mathbb{E}\{\widehat{D}_\varpi(\boldsymbol{\delta})\}$ and $R_\varpi(\boldsymbol{\beta}) = \mathbb{E}\{\widehat{R}_\varpi(\boldsymbol{\beta})\}$. Due to $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ with $\boldsymbol{\beta}, \boldsymbol{\beta}^* \in \mathbb{H}(s)$, then $\boldsymbol{\delta} \in \mathbb{H}(2s)$. Recall $\mathcal{S} = \{S \subseteq [p] : 1 \leq |S| \leq 2s\}$. Conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$\begin{aligned}
\widehat{D}_\varpi(\boldsymbol{\delta}) &= D_\varpi(\boldsymbol{\delta}) - \{D_\varpi(\boldsymbol{\delta}) - \widehat{D}_\varpi(\boldsymbol{\delta})\} \\
&= \underbrace{D_\varpi(\boldsymbol{\delta}) - \langle \nabla R_\varpi(\boldsymbol{\beta}^*), \boldsymbol{\delta}\rangle}_{A_\varpi(\boldsymbol{\delta})} + \langle \nabla R_\varpi(\boldsymbol{\beta}^*), \boldsymbol{\delta}\rangle - \{D_\varpi(\boldsymbol{\delta}) - \widehat{D}_\varpi(\boldsymbol{\delta})\} \\
&\geq A_\varpi(\boldsymbol{\delta}) - \sup_{S\in\mathcal{S}}\|\{\nabla R_\varpi(\boldsymbol{\beta}^*)\}_S\|_2\|\boldsymbol{\delta}\|_2 - \{D_\varpi(\boldsymbol{\delta}) - \widehat{D}_\varpi(\boldsymbol{\delta})\} \\
&\geq b_1\|\boldsymbol{\delta}\|_2^2 - (b_2 + \bar{b}_3)\|\boldsymbol{\delta}\|_2
\end{aligned}$$

holds for all $\boldsymbol{\delta} \in \mathbb{H}(2s) \cap \{\mathbb{B}^p(c_0)\backslash\mathbb{B}^p(n^{-1})\}$, which establishes (S.54). $\square$

EC.7.3.2 Proof of (S.55)

Similarly, conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4$, it holds that

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

$$
\begin{aligned}
&= -\widehat{D}_{\varpi}(\boldsymbol{\delta}) + \langle \nabla R_{\varpi}(\boldsymbol{\beta}), \boldsymbol{\delta}\rangle + \langle \nabla \widehat{R}_{\varpi}(\boldsymbol{\beta}) - \nabla R_{\varpi}(\boldsymbol{\beta}), \boldsymbol{\delta}\rangle \\
&= \underbrace{-D_{\varpi}(\boldsymbol{\delta}) + \langle \nabla R_{\varpi}(\boldsymbol{\beta}), \boldsymbol{\delta}\rangle}_{B_{\varpi}(\boldsymbol{\delta})} - \{\widehat{D}_{\varpi}(\boldsymbol{\delta}) - D_{\varpi}(\boldsymbol{\delta})\} + \langle \nabla \widehat{R}_{\varpi}(\boldsymbol{\beta}) - \nabla R_{\varpi}(\boldsymbol{\beta}), \boldsymbol{\delta}\rangle \\
&\geq B_{\varpi}(\boldsymbol{\delta}) - \sup_{\boldsymbol{\beta}\in\mathbb{H}(s)\cap\Theta(c_0),\, S\in\mathcal{S}} \|(1-\mathbb{E})[\{\nabla \widehat{R}_{\varpi}(\boldsymbol{\beta})\}_S]\|_2 \|\boldsymbol{\delta}\|_2 - \{\widehat{D}_{\varpi}(\boldsymbol{\delta}) - D_{\varpi}(\boldsymbol{\delta})\} \\
&\geq b_1\|\boldsymbol{\delta}\|_2^2 - (b_2+b_4)\|\boldsymbol{\delta}\|_2
\end{aligned}
$$

for all $\boldsymbol{\delta} \in \mathbb{H}(2s) \cap \{\mathbb{B}^p(c_0)\backslash \mathbb{B}^p(n^{-1})\}$. Then (S.55) holds. □

EC.7.3.3 Proof of (S.56)

Let $\mathbf{u} = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)/\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2$ and define $\bar{\boldsymbol{\beta}} = m\boldsymbol{\beta}_2 + (1-m)\boldsymbol{\beta}_1$ for some $m \in [0,1]$. Due to $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{H}(s) \cap \Theta(c_0)$, we have $\mathbf{u} \in \mathbb{H}(2s)$ and $\bar{\boldsymbol{\beta}} \in \mathbb{H}(2s) \cap \Theta(c_0)$. Then, conditioned on $\mathcal{E}_5$, it holds that

$$
\begin{aligned}
&\widehat{R}_{\varpi}(\boldsymbol{\beta}_1) - \widehat{R}_{\varpi}(\boldsymbol{\beta}_2) - \langle \nabla \widehat{R}_{\varpi}(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\rangle \\
&\quad = \frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^{\top}\{\nabla^2 \widehat{R}_{\varpi}(\bar{\boldsymbol{\beta}})\}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) = \frac{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2}{2}\mathbf{u}^{\top}\{\nabla^2 \widehat{R}_{\varpi}(\bar{\boldsymbol{\beta}})\}\mathbf{u} \\
&\quad \leq \frac{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2}{2} \sup_{\boldsymbol{\beta}\in\mathbb{H}(2s)\cap\Theta(c_0),\, S\in\mathcal{S}} \|\{\nabla^2 \widehat{R}_{\varpi}(\boldsymbol{\beta})\}_{SS}\| \leq \frac{b_5}{2}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2\,,
\end{aligned}
$$

which establishes (S.56). □

EC.7.4. Proof of Proposition P2

For ease of notation, we write $\mathbf{g}^{(t)} = \nabla \widehat{R}_{\varpi}(\boldsymbol{\beta}^{(t)})$ and $I^t = S^{t+1} \cup S^*$, where $S^* = \mathrm{supp}(\boldsymbol{\beta}^*)$ and S^{t+1} is the index set S determined by Algorithm 2 when $\mathbf{v}$ is selected as $\tilde{\boldsymbol{\beta}}^{(t+1)}$. Notice that $\boldsymbol{\beta}^{(t+1)} = \tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) + \tilde{\mathbf{w}}^t_{|S^{t+1}}$. Due to $\tilde{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)}$ conditioned on $\mathcal{E}_0$ and $\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) = \{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)})\}_{|I^t}$, we have

$$
\begin{aligned}
\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^* &= \{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) - \tilde{\boldsymbol{\beta}}^{(t+1)}\}_{|I^t} + \tilde{\boldsymbol{\beta}}^{(t+1)}_{|I^t} - \boldsymbol{\beta}^* + \tilde{\mathbf{w}}^t_{|S^{t+1}} \\
&= \underbrace{\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) - \tilde{\boldsymbol{\beta}}^{(t+1)}\}_{|I^t} + \tilde{\mathbf{w}}^t_{|S^{t+1}}}_{\Delta_{1,t}} - \eta_0\mathbf{g}^{(t)}_{|I^t} + \boldsymbol{\beta}^{(t)}_{|I^t} - \boldsymbol{\beta}^*\,.
\end{aligned}
$$

Since $\boldsymbol{\beta}^*_{|I^t} = \boldsymbol{\beta}^*$, it follows that

$$
\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_2^2 = \|\{\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\}_{|I^t}\|_2^2 - 2\eta_0\langle \boldsymbol{\beta}^{(t)}_{|I^t} - \boldsymbol{\beta}^*_{|I^t}, \mathbf{g}^{(t)}_{|I^t}\rangle
$$

$$+\underbrace{\|\Delta_{1,t}-\eta_0\mathbf{g}^{(t)}_{|I^t}\|_2^2+2\langle\boldsymbol{\beta}^{(t)}_{|I^t}-\boldsymbol{\beta}^*_{|I^t},\Delta_{1,t}\rangle}_{\Delta_{2,t}}\,. \tag{S.60}$$

Notice that $\|\{\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\}_{|I^t}\|_2\le\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\|_2\le r_0$. Letting $r_*=r_0/\sqrt{2}$, we will show in Sections EC.7.4.1 and EC.7.4.2 that $\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^*\|_2^2<r_0^2$ holds in both of the following cases: (i) $\|\{\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\}_{|I^t}\|_2\le r_*$, and (ii) $r_*<\|\{\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\}_{|I^t}\|_2\le r_0$. Then, we complete the proof of Proposition P2. □

EC.7.4.1 Case (i): $\|\{\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\}_{|I^t}\|_2\le r_*$

By Lemma L9 and conditioned on $\mathcal{E}_0$,

$$\begin{aligned}
\big\|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)})-\tilde{\boldsymbol{\beta}}^{(t+1)}\}_{|I^t}\big\|_2^2&=\big\|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)})-\tilde{\boldsymbol{\beta}}^{(t+1)}\}_{I^t}\big\|_2^2\\
&\le\frac{8}{7}\frac{|I^t|-s}{|I^t|-s^*}\big\|\{\boldsymbol{\beta}^*-\tilde{\boldsymbol{\beta}}^{(t+1)}\}_{I^t}\big\|_2^2+32\sum_{i=1}^{s}\|\mathbf{w}_i^t\|_\infty^2\\
&\overset{\text{(i)}}{\le}\frac{8s^*}{7s}\big\|\{\boldsymbol{\beta}^*-\boldsymbol{\beta}^{(t)}+\eta_0\mathbf{g}^{(t)}\}_{I^t}\big\|_2^2+32\sum_{i=1}^{s}\|\mathbf{w}_i^t\|_\infty^2\\
&\le\frac{8s^*}{7s}\big\{r_*^2+\eta_0^2\|\mathbf{g}^{(t)}_{I^t}\|_2^2+2\eta_0r_*\|\mathbf{g}^{(t)}_{I^t}\|_2\big\}+32\sum_{i=1}^{s}\|\mathbf{w}_i^t\|_\infty^2\,,
\end{aligned}$$

where (i) follows from $|I^t|\le|S^{t+1}|+|S^*|\le s+s^*$. Due to $\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\|_2\le r_0$, conditioned on $\mathcal{E}_3\cap\mathcal{E}_6$, we have $\|\mathbf{g}^{(t)}_{I^t}\|_2\le 3b_5r_0/4+b_3$. Define

$$\Omega_1(r_*):=\frac{8s^*}{7s}\left\{r_*+\eta_0\left(\frac{3b_5r_0}{4}+b_3\right)\right\}^2.$$

Then, we have $\|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)})-\tilde{\boldsymbol{\beta}}^{(t+1)}\}_{|I^t}\|_2^2\le\Omega_1(r_*)+32\sum_{i=1}^{s}\|\mathbf{w}_i^t\|_\infty^2$ and

$$\begin{aligned}
\|\Delta_{1,t}\|_2^2&\le\big[\big\|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)})-\tilde{\boldsymbol{\beta}}^{(t+1)}\}_{|I^t}\big\|_2+\big\|\tilde{\mathbf{w}}^t_{|S^{t+1}}\big\|_2\big]^2\\
&\le\Omega_1(r_*)+32W_t+2\sqrt{\Omega_1(r_*)W_t+32W_t^2}=:\Omega_2(r_*)\,,
\end{aligned}$$

where W_t is defined in (S.28). Since $r_*<r_0$ and $\Omega_2(r)$ is increasing in $r>0$, by (S.60), we obtain

$$\begin{aligned}
|\Delta_{2,t}|&\le\big(\|\Delta_{1,t}\|_2+\eta_0\|\mathbf{g}^{(t)}_{|I^t}\|_2\big)^2+2\|\{\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\}_{|I^t}\|_2\|\Delta_{1,t}\|_2\\
&\le\left\{\sqrt{\Omega_2(r_0)}+\eta_0\left(\frac{3b_5r_0}{4}+b_3\right)\right\}^2+2r_0\sqrt{\Omega_2(r_0)}
\end{aligned}\tag{S.61}$$

conditioned on $\mathcal{E}_0 \cap \mathcal{E}_3 \cap \mathcal{E}_6$. Moreover, conditioned on $\mathcal{E}_3 \cap \mathcal{E}_6$, it holds that

$$|2\eta_0 \langle \boldsymbol{\beta}^{(t)}_{|I^t} - \boldsymbol{\beta}^*_{|I^t}, \mathbf{g}^{(t)}_{|I^t} \rangle| \le 2\eta_0 \|\{\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\}_{|I^t}\|_2 \|\mathbf{g}^{(t)}_{|I^t}\|_2 \le \eta_0 r_* \bigg(\frac{3b_5 r_0}{2} + 2b_3 \bigg) .$$

Then, according to (S.60) and (S.61), we can conclude that

$$\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_2^2 \le r_*^2 + \bigg\{ \sqrt{\Omega_2(r_0)} + \eta_0 \bigg(\frac{3b_5 r_0}{4} + b_3 \bigg) \bigg\}^2 + 2r_0 \sqrt{\Omega_2(r_0)} + \eta_0 r_* \bigg(\frac{3b_5 r_0}{2} + 2b_3 \bigg) . \tag{S.62}$$

Due to $b_3 \le b_1 r_0/24$, $b_1 \le b_5/2$ and $\eta_0 \le 2b_1/(9b_5^2)$, we have

$$\eta_0 \bigg(\frac{3b_5 r_0}{4} + b_3 \bigg)^2 \le \frac{b_1 r_0^2}{8} + \frac{b_1^3 r_0^2}{2592 b_5^2} + \frac{b_1^2 r_0^2}{72 b_5} < \frac{b_1 r_0^2}{6} , \tag{S.63}$$

which implies $\eta_0(3b_5 r_0/4 + b_3) < r_0/6$. Combining with $s \ge 38 s^* (\eta_0 b_1)^{-2}$, it holds that

$$\Omega_1(r_0) = \frac{8s^*}{7s} \bigg\{ r_0 + \eta_0 \bigg(\frac{3b_5 r_0}{4} + b_3 \bigg) \bigg\}^2 < \frac{4\eta_0^2 b_1^2}{133} \bigg(\frac{7r_0}{6} \bigg)^2 < \frac{\eta_0^2 b_1^2 r_0^2}{24} .$$

Since $\max_{t \in \{0\} \cup [T-1]} W_t \le \eta_0^2 b_1^2 r_0^2 / 4000$, it follows that

$$\Omega_2(r_0) < \frac{\eta_0^2 b_1^2 r_0^2}{24} + \frac{\eta_0^2 b_1^2 r_0^2}{125} + 2\sqrt{\frac{\eta_0^4 b_1^4 r_0^4}{96000} + \frac{\eta_0^4 b_1^4 r_0^4}{500000}} < \frac{\eta_0^2 b_1^2 r_0^2}{16} .$$

Together with (S.63), we can conclude that

$$\begin{aligned} &\bigg\{ \sqrt{\Omega_2(r_0)} + \eta_0 \bigg(\frac{3b_5 r_0}{4} + b_3 \bigg) \bigg\}^2 + 2r_0 \sqrt{\Omega_2(r_0)} \\ &\quad < \bigg(\frac{\eta_0 b_1 r_0}{4} + \eta_0^{1/2} \sqrt{\frac{b_1 r_0^2}{6}} \bigg)^2 + \frac{\eta_0 b_1 r_0^2}{2} = \eta_0 b_1 r_0^2 \bigg(\frac{\eta_0 b_1}{16} + \frac{\eta_0^{1/2} b_1^{1/2}}{2\sqrt{6}} + \frac{2}{3} \bigg) . \end{aligned} \tag{S.64}$$

Since $r_* = r_0/\sqrt{2}$ and $\eta_0 \le 2b_1/(9b_5^2)$ with $b_1 \le b_5/2$, by (S.62), it holds that

$$\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_2^2 < \frac{r_0^2}{2} + \eta_0 b_1 r_0^2 \bigg(\frac{\eta_0 b_1}{16} + \frac{\eta_0^{1/2} b_1^{1/2}}{2\sqrt{6}} + \frac{2}{3} \bigg) + \frac{3\eta_0 b_5 r_0^2}{2\sqrt{2}} + \frac{\eta_0 b_1 r_0^2}{12\sqrt{2}} < r_0^2 .$$

This completes the proof. $\square$

EC.7.4.2 Case (ii): $r_* < \|\{\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\}_{|I^t}\|_2 \le r_0$

Conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4$, by (S.55) in Proposition P1, it holds that

$$\langle \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*, \mathbf{g}^{(t)} \rangle \ge \widehat{R}_\varpi(\boldsymbol{\beta}^{(t)}) - \widehat{R}_\varpi(\boldsymbol{\beta}^*) + b_1 r_*^2 - (b_2 + b_4) r_0 .$$

On the other hand, the convexity of $\widehat{R}_\varpi(\boldsymbol{\beta})$ implies that, given $\mathcal{E}_3$ holds,

$$\widehat{R}_\varpi(\boldsymbol{\beta}^{(t)}) - \widehat{R}_\varpi(\boldsymbol{\beta}^*) \geq \langle \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*, \nabla \widehat{R}_\varpi(\boldsymbol{\beta}^*)\rangle \geq -b_3 r_0\,,$$

which implies $\langle \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*, \mathbf{g}^{(t)}\rangle \geq b_1 r_*^2 - (b_2+b_3+b_4)r_0$. Then, conditioned on $\mathcal{E}_3 \cap \mathcal{E}_6$, we have

$$\begin{aligned}\langle \boldsymbol{\beta}^{(t)}_{|I^t} - \boldsymbol{\beta}^*_{|I^t}, \mathbf{g}^{(t)}_{|I^t}\rangle &= \langle \boldsymbol{\beta}^{(t)}_{|I^t} - \boldsymbol{\beta}^*, \mathbf{g}^{(t)}\rangle = \langle \boldsymbol{\beta}^{(t)}_{|I^t} - \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle + \langle \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*, \mathbf{g}^{(t)}\rangle \\ &\geq -\|\boldsymbol{\beta}^{(t)}_{\bar{S}\backslash I^t}\|_2 \left(\frac{3b_5 r_0}{4} + b_3\right) - (b_2+b_3+b_4)r_0 + b_1 r_*^2\,,\end{aligned}$$

where $\bar{S} = \mathrm{supp}(\boldsymbol{\beta}^{(t)})$. Due to $r_* = r_0/\sqrt{2}$ and $b_2+b_3+b_4 \leq b_1 r_0/24$, we can conclude that

$$\begin{aligned}-2\eta_0\langle \boldsymbol{\beta}^{(t)}_{|I^t} - \boldsymbol{\beta}^*_{|I^t}, \mathbf{g}^{(t)}_{|I^t}\rangle &\leq 2\eta_0 \|\boldsymbol{\beta}^{(t)}_{\bar{S}\backslash I^t}\|_2 \left(\frac{3b_5 r_0}{4} + b_3\right) + 2\eta_0(b_2+b_3+b_4)r_0 - 2\eta_0 b_1 r_*^2 \\ &\leq \|\boldsymbol{\beta}^{(t)}_{\bar{S}\backslash I^t}\|_2^2 + \eta_0^2\left(\frac{3b_5 r_0}{4} + b_3\right)^2 - \frac{11\eta_0 b_1 r_0^2}{12}\,. \end{aligned} \tag{S.65}$$

Analogously, we know that (S.61) also holds in Case (ii). Since $\eta_0 \leq 2b_1/(9b_5^2)$, by (S.61), (S.63) and (S.64), we have

$$\eta_0^2\left(\frac{3b_5 r_0}{4} + b_3\right)^2 + |\Delta_{2,t}| < \frac{\eta_0 b_1 r_0^2}{6} + \frac{3\eta_0 b_1 r_0^2}{4} = \frac{11\eta_0 b_1 r_0^2}{12}\,.$$

Notice that $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2^2 = \|\{\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\}_{|I^t}\|_2^2 + \|\boldsymbol{\beta}^{(t)}_{\bar{S}\backslash I^t}\|_2^2$. By (S.60) and (S.65), we have

$$\begin{aligned}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_2^2 &= \|\{\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\}_{|I^t}\|_2^2 - 2\eta_0\langle \boldsymbol{\beta}^{(t)}_{|I^t} - \boldsymbol{\beta}^*_{|I^t}, \mathbf{g}^{(t)}_{|I^t}\rangle + \Delta_{2,t} \\ &\leq \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2^2 + \eta_0^2\left(\frac{3b_5 r_0}{4} + b_3\right)^2 - \frac{11\eta_0 b_1 r_0^2}{12} + |\Delta_{2,t}| < \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2^2 \leq r_0^2\,.\end{aligned}$$

This completes the proof. □

EC.7.5. Proof of Proposition P3

Write $\mathbf{g}^{(t)} = \nabla \widehat{R}_\varpi(\boldsymbol{\beta}^{(t)})$ and $I^t = S^{t+1} \cup S^*$, where $S^* = \mathrm{supp}(\boldsymbol{\beta}^*)$ and S^{t+1} is the index set S determined by Algorithm 2 when $\mathbf{v}$ is selected as $\tilde{\boldsymbol{\beta}}^{(t+1)}$. Notice that $\boldsymbol{\beta}^{(t+1)} = \tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) + \tilde{\mathbf{w}}^t_{|S^{t+1}}$. Due to $\tilde{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_0 \mathbf{g}^{(t)}$ conditioned on $\mathcal{E}_0$, $\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) = \{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)})\}_{|I^t}$ and $\boldsymbol{\beta}^*_{|I^t} = \boldsymbol{\beta}^*$, we have

$$\begin{aligned}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_2 &= \|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) - \tilde{\boldsymbol{\beta}}^{(t+1)}\}_{|I^t} + \tilde{\boldsymbol{\beta}}^{(t+1)}_{|I^t} - \boldsymbol{\beta}^* + \tilde{\mathbf{w}}^t_{|S^{t+1}}\|_2 \\ &= \|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) - \tilde{\boldsymbol{\beta}}^{(t+1)}\}_{|I^t} + \tilde{\mathbf{w}}^t_{|S^{t+1}} - \eta_0 \mathbf{g}^{(t)}_{|I^t} + \boldsymbol{\beta}^{(t)}_{|I^t} - \boldsymbol{\beta}^*\|_2\end{aligned}$$

$$\leq \|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) - \tilde{\boldsymbol{\beta}}^{(t+1)}\}_{I^t}\|_2 + \|\tilde{\mathbf{w}}^t_{|S^{t+1}}\|_2 + \eta_0\|\mathbf{g}^{(t)}_{I^t}\|_2 + \|\{\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\}_{I^t}\|_2 \tag{S.66}$$

conditioned on $\mathcal{E}_0$. By Lemma L9 with $c=1$, conditioned on $\mathcal{E}_0$, it holds that

$$\begin{aligned}\big\|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) - \tilde{\boldsymbol{\beta}}^{(t+1)}\}_{I^t}\big\|_2^2 &\leq 2\frac{|I^t|-s}{|I^t|-s^*}\big\|\{\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}^{(t+1)}\}_{I^t}\big\|_2^2 + 8\sum_{i=1}^s \|\mathbf{w}_i^t\|_\infty^2 \\ &\overset{\text{(i)}}{\leq} \frac{2s^*}{s}\big\|\{\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_{I^t}\big\|_2^2 + 8\sum_{i=1}^s \|\mathbf{w}_i^t\|_\infty^2 \\ &\leq \frac{2s^*}{s}\bigg\{\|\{\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\}_{I^t}\|_2 + \eta_0\|\mathbf{g}^{(t)}_{I^t}\|_2\bigg\}^2 + 8\sum_{i=1}^s \|\mathbf{w}_i^t\|_\infty^2\,,\end{aligned}$$

where (i) follows from $|I^t| \leq |S^{t+1}| + |S^*| \leq s + s^*$. This, together with (S.66), yields

$$\begin{aligned}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_2 &\leq \left(1+\sqrt{\frac{2s^*}{s}}\right)\big\{\|\{\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\}_{I^t}\|_2 + \eta_0\|\mathbf{g}^{(t)}_{I^t}\|_2\big\} + \sqrt{8\sum_{i=1}^s \|\mathbf{w}_i^t\|_\infty^2} + \|\tilde{\mathbf{w}}^t_{|S^{t+1}}\|_2 \\ &\overset{\text{(ii)}}{\leq} \left(1+\sqrt{\frac{2s^*}{s}}\right)\big\{n^{-1} + \eta_0\|\mathbf{g}^{(t)}_{I^t}\|_2\big\} + 4W_t^{1/2}\end{aligned} \tag{S.67}$$

conditioned on $\mathcal{E}_0$, where (ii) follows from $\|\{\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\}_{I^t}\|_2 \leq \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 \leq n^{-1}$ and $W_t = \sum_{i=1}^s \|\mathbf{w}_i^t\|_\infty^2 + \|\tilde{\mathbf{w}}^t_{|S^{t+1}}\|_2^2$. Moreover, conditioned on $\mathcal{E}_3 \cap \mathcal{E}_7$, we have $\|\mathbf{g}^{(t)}_{I^t}\|_2 \leq 3b_5/(4n) + b_3$. Since $\eta_0 \leq 2b_1/(9b_5^2)$ with $b_1 \leq b_5/2$, we have $s^*/s \leq \eta_0^2 b_1^2/38 \leq 1/8$, $\eta_0 b_1 \leq 1/18$ and $\eta_0 b_5 \leq 1/9$. Then, by (S.67), it holds that

$$\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_2 \leq \frac{3}{2}\left(\frac{1}{n} + \frac{3\eta_0 b_5}{4n} + \eta_0 b_3\right) + 4W_t^{1/2} \leq \frac{13}{8n} + \frac{b_3}{12b_1} + 4W_t^{1/2}$$

conditioned on $\mathcal{E}_0 \cap \mathcal{E}_3 \cap \mathcal{E}_7$. Since $\max_{t\in\{0\}\cup[T-1]} W_t \leq \Delta_n(z)$, we complete the proof of Proposition P3. □

EC.7.6. Proof of Proposition P4

For ease of notation, we write $S^* = \mathrm{supp}(\boldsymbol{\beta}^*)$, and denote by S^t and S^{t+1}, respectively, the index set S determined by Algorithm 2 when $\mathbf{v}$ is selected as $\tilde{\boldsymbol{\beta}}^{(t)}$ and $\tilde{\boldsymbol{\beta}}^{(t+1)}$. Moreover, let $J^t = S^t \cup S^{t+1} \cup S^*$ and $\mathbf{g}^{(t)} = \nabla \widehat{R}_\varpi(\boldsymbol{\beta}^{(t)})$. By Proposition P2, we know $\boldsymbol{\beta}^{(t+1)} \in \mathbb{H}(s) \cap \{\Theta(c_0) \backslash \Theta(n^{-1})\}$. Then, conditioned on $\mathcal{E}_5$, it follows from (S.56) in Proposition P1 that

$$\begin{aligned}\widehat{R}_\varpi(\boldsymbol{\beta}^{(t+1)}) - \widehat{R}_\varpi(\boldsymbol{\beta}^{(t)}) &\leq \langle \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle + \frac{b_5}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 \\ &= \frac{b_5}{2}\big\|\{\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_{J^t}\big\|_2^2 - \frac{\eta_0^2 b_5}{2}\big\|\mathbf{g}^{(t)}_{J^t}\big\|_2^2\end{aligned}$$

$$+ (1-\eta_0 b_5)\langle \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle\,. \tag{S.68}$$

Recall $\tilde{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)}$ conditioned on $\mathcal{E}_0$. Then, conditioned on $\mathcal{E}_0$, it holds that $\boldsymbol{\beta}^{(t+1)} = \tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) + \tilde{\mathbf{w}}^t_{|S^{t+1}} = (\boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)} + \tilde{\mathbf{w}}^t)_{|S^{t+1}}$. This, together with $\boldsymbol{\beta}^{(t+1)}_{|S^t\setminus S^{t+1}} = \mathbf{0}$ and $S^{t+1}\cap(S^t\setminus S^{t+1}) = \emptyset$, leads to

$$\begin{aligned}\langle \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle &= \langle \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle_{S^{t+1}} + \langle \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle_{S^t\setminus S^{t+1}}\\ &= \langle \boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)} + \tilde{\mathbf{w}}^t - \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle_{S^{t+1}} - \langle \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle_{S^t\setminus S^{t+1}}\\ &= -\eta_0\langle \mathbf{g}^{(t)}, \mathbf{g}^{(t)}\rangle_{S^{t+1}} + \langle \tilde{\mathbf{w}}^t, \mathbf{g}^{(t)}\rangle_{S^{t+1}} - \langle \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle_{S^t\setminus S^{t+1}}\\ &\le -\{\eta_0 - (4c_1)^{-1}\}\langle \mathbf{g}^{(t)}, \mathbf{g}^{(t)}\rangle_{S^{t+1}} + c_1\langle \tilde{\mathbf{w}}^t, \tilde{\mathbf{w}}^t\rangle_{S^{t+1}} - \langle \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle_{S^t\setminus S^{t+1}}\end{aligned} \tag{S.69}$$

for any constant $c_1 > 0$. Here, we write $\langle \mathbf{a}, \mathbf{b}\rangle_S = \mathbf{a}_S^\top \mathbf{b}_S$ for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ and $S \subseteq [p]$.

Since $(S^{t+1}\setminus S^t) \subseteq S^{t+1}$, $(S^t\setminus S^{t+1}) \subseteq S^t$ and $|S^{t+1}\setminus S^t| = |S^t\setminus S^{t+1}|$, Lemma 3.4 in Cai et al. (2021) yields that, for any $c_2 > 0$,

$$\begin{aligned}\langle \tilde{\boldsymbol{\beta}}^{(t+1)}, \tilde{\boldsymbol{\beta}}^{(t+1)}\rangle_{S^t\setminus S^{t+1}} &\le (1+c_2)\langle \tilde{\boldsymbol{\beta}}^{(t+1)}, \tilde{\boldsymbol{\beta}}^{(t+1)}\rangle_{S^{t+1}\setminus S^t} + 4(1+c_2^{-1})\sum_{i=1}^s \|\mathbf{w}_i^t\|_\infty^2\\ &= (1+c_2)\langle \eta_0\mathbf{g}^{(t)}, \eta_0\mathbf{g}^{(t)}\rangle_{S^{t+1}\setminus S^t} + 4(1+c_2^{-1})\sum_{i=1}^s \|\mathbf{w}_i^t\|_\infty^2\end{aligned} \tag{S.70}$$

conditioned on $\mathcal{E}_0$, where the last step is based on the facts that $\tilde{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)}$ conditioned on $\mathcal{E}_0$ and $\boldsymbol{\beta}^{(t)}_{S^{t+1}\setminus S^t} = \mathbf{0}$. In the meanwhile, conditioned on $\mathcal{E}_0$,

$$\begin{aligned}\langle \tilde{\boldsymbol{\beta}}^{(t+1)}, \tilde{\boldsymbol{\beta}}^{(t+1)}\rangle_{S^t\setminus S^{t+1}} &= \langle \boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)}, \boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)}\rangle_{S^t\setminus S^{t+1}}\\ &= \langle \boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t)}\rangle_{S^t\setminus S^{t+1}} + \langle \eta_0\mathbf{g}^{(t)}, \eta_0\mathbf{g}^{(t)}\rangle_{S^t\setminus S^{t+1}} - 2\eta_0\langle \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle_{S^t\setminus S^{t+1}}\\ &\ge \langle \eta_0\mathbf{g}^{(t)}, \eta_0\mathbf{g}^{(t)}\rangle_{S^t\setminus S^{t+1}} - 2\eta_0\langle \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle_{S^t\setminus S^{t+1}}\,.\end{aligned} \tag{S.71}$$

Combining (S.70) and (S.71) yields

$$-\langle \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle_{S^t\setminus S^{t+1}} \le (1+c_2)\frac{\eta_0}{2}\big\|\mathbf{g}^{(t)}_{S^{t+1}\setminus S^t}\big\|_2^2 - \frac{\eta_0}{2}\big\|\mathbf{g}^{(t)}_{S^t\setminus S^{t+1}}\big\|_2^2 + (1+c_2^{-1})\frac{2}{\eta_0}\sum_{i=1}^s \|\mathbf{w}_i^t\|_\infty^2\,.$$

Plugging this to (S.69) yields

$$\begin{aligned}\langle \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle &\le (1+c_2)\frac{\eta_0}{2}\big\|\mathbf{g}^{(t)}_{S^{t+1}\setminus S^t}\big\|_2^2 - \frac{\eta_0}{2}\big\|\mathbf{g}^{(t)}_{S^t\setminus S^{t+1}}\big\|_2^2\\ &\quad - \{\eta_0 - (4c_1)^{-1}\}\big\|\mathbf{g}^{(t)}_{S^{t+1}}\big\|_2^2 + \{c_1 + 2(1+c_2^{-1})\eta_0^{-1}\}W_t\,,\end{aligned}$$

where $W_t = \sum_{i=1}^s \|\mathbf{w}_i^t\|_\infty^2 + \|\tilde{\mathbf{w}}_{|S^{t+1}}^t\|_2^2$ is defined in (S.28). Taking $c_1 = 2\eta_0^{-1}, c_2 = 1/4$, and using the equality $\|\mathbf{g}_{S^{t+1}}^{(t)}\|_2^2 = \|\mathbf{g}_{S^{t+1}\setminus S^t}^{(t)}\|_2^2 + \|\mathbf{g}_{S^t\cap S^{t+1}}^{(t)}\|_2^2$, it follows that

$$\begin{aligned}\langle\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle &\le -\frac{\eta_0}{4}\big\|\mathbf{g}_{S^{t+1}\setminus S^t}^{(t)}\big\|_2^2 - \frac{7\eta_0}{8}\big\|\mathbf{g}_{S^t\cap S^{t+1}}^{(t)}\big\|_2^2 - \frac{\eta_0}{2}\big\|\mathbf{g}_{S^t\setminus S^{t+1}}^{(t)}\big\|_2^2 + \frac{12}{\eta_0}W_t \\ &\le -\frac{\eta_0}{4}\big\|\mathbf{g}_{S^t\cup S^{t+1}}^{(t)}\big\|_2^2 + \frac{12}{\eta_0}W_t\end{aligned}$$

conditioned on $\mathcal{E}_0$. This, together with (S.68) and $\|\mathbf{g}_{J^t}^{(t)}\|_2^2 = \|\mathbf{g}_{J^t\setminus(S^t\cup S^*)}^{(t)}\|_2^2 + \|\mathbf{g}_{S^t\cup S^*}^{(t)}\|_2^2$, yields

$$\begin{aligned}\widehat{R}_\varpi(\boldsymbol{\beta}^{(t+1)}) - \widehat{R}_\varpi(\boldsymbol{\beta}^{(t)}) \le \frac{b_5}{2}&\underbrace{\big[\big\|\{\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_{J^t}\big\|_2^2 - \eta_0^2\big\|\mathbf{g}_{J^t\setminus(S^t\cup S^*)}^{(t)}\big\|_2^2\big]}_{\Pi_1} \\ &- \frac{\eta_0^2 b_5}{2}\big\|\mathbf{g}_{S^t\cup S^*}^{(t)}\big\|_2^2 - (1-\eta_0 b_5)\frac{\eta_0}{4}\big\|\mathbf{g}_{S^t\cup S^{t+1}}^{(t)}\big\|_2^2 + (1-\eta_0 b_5)\frac{12}{\eta_0}W_t\end{aligned}$$

conditioned on $\mathcal{E}_0\cap\mathcal{E}_5$. As we will show in Section EC.7.6.1, conditioned on $\mathcal{E}_0$, for any $c_3 > 0$,

$$\Pi_1 \le \frac{4s^*}{s}\big\|\{\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_{J^t}\big\|_2^2 + c_3\eta_0^2\big\|\mathbf{g}_{J^t\setminus(S^t\cup S^*)}^{(t)}\big\|_2^2 + 4(5+c_3^{-1})W_t \tag{S.72}$$

for $t\in\{0\}\cup[T-1]$. Hence, by (S.72), it follows that

$$\begin{aligned}\widehat{R}_\varpi(\boldsymbol{\beta}^{(t+1)}) - \widehat{R}_\varpi(\boldsymbol{\beta}^{(t)}) \le{}& \frac{2b_5 s^*}{s}\big\|\{\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_{J^t}\big\|_2^2 + \frac{c_3\eta_0^2 b_5}{2}\big\|\mathbf{g}_{J^t\setminus(S^t\cup S^*)}^{(t)}\big\|_2^2 \\ &- \frac{\eta_0^2 b_5}{2}\big\|\mathbf{g}_{S^t\cup S^*}^{(t)}\big\|_2^2 - (1-\eta_0 b_5)\frac{\eta_0}{4}\big\|\mathbf{g}_{S^t\cup S^{t+1}}^{(t)}\big\|_2^2 \\ &+ \bigg\{2b_5(5+c_3^{-1}) + (1-\eta_0 b_5)\frac{12}{\eta_0}\bigg\}W_t\,.\end{aligned} \tag{S.73}$$

Under the assumption that $\boldsymbol{\beta}^{(t)}\in\mathbb{H}(s)\cap\{\Theta(c_0)\setminus\Theta(n^{-1})\}$ and conditioned on $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_4$, we obtain from (S.55) in Proposition P1 that

$$\begin{aligned}\big\|\{\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_{J^t}\big\|_2^2 &= 2\eta_0\langle\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)}, \mathbf{g}^{(t)}\rangle + \|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)}\|_2^2 + \eta_0^2\big\|\mathbf{g}_{J^t}^{(t)}\big\|_2^2 \\ &\le 2\eta_0\{\widehat{R}_\varpi(\boldsymbol{\beta}^*) - \widehat{R}_\varpi(\boldsymbol{\beta}^{(t)})\} + (1-2\eta_0 b_1)\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)}\|_2^2 \\ &\quad + 2\eta_0(b_2+b_4)\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)}\|_2 + \eta_0^2\big\|\mathbf{g}_{J^t}^{(t)}\big\|_2^2\,.\end{aligned}$$

Combining this with (S.73) yields

$$\begin{aligned}\widehat{R}_\varpi(\boldsymbol{\beta}^{(t+1)}) - \widehat{R}_\varpi(\boldsymbol{\beta}^{(t)}) \le{}& \frac{4\eta_0 b_5 s^*}{s}\{\widehat{R}_\varpi(\boldsymbol{\beta}^*) - \widehat{R}_\varpi(\boldsymbol{\beta}^{(t)})\} + \frac{2b_5(1-2\eta_0 b_1)s^*}{s}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)}\|_2^2 \\ &+ \frac{4\eta_0 b_5(b_2+b_4)s^*}{s}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)}\|_2 + \frac{2\eta_0^2 b_5 s^*}{s}\big\|\mathbf{g}_{J^t}^{(t)}\big\|_2^2 - \frac{\eta_0^2 b_5}{2}\big\|\mathbf{g}_{S^t\cup S^*}^{(t)}\big\|_2^2\end{aligned}$$

$$+\frac{c_3\eta_0^2 b_5}{2}\big\|\mathbf{g}^{(t)}_{J^t\setminus(S^t\cup S^*)}\big\|_2^2-(1-\eta_0 b_5)\frac{\eta_0}{4}\big\|\mathbf{g}^{(t)}_{S^t\cup S^{t+1}}\big\|_2^2$$
$$+\bigg\{2b_5(5+c_3^{-1})+(1-\eta_0 b_5)\frac{12}{\eta_0}\bigg\}W_t\,. \tag{S.74}$$

Note that $\|\mathbf{g}^{(t)}_{J^t}\|_2^2=\|\mathbf{g}^{(t)}_{J^t\setminus(S^t\cup S^*)}\|_2^2+\|\mathbf{g}^{(t)}_{S^t\cup S^*}\|_2^2$ and $\|\mathbf{g}^{(t)}_{S^t\cup S^{t+1}}\|_2^2\geq\|\mathbf{g}^{(t)}_{J^t\setminus(S^t\cup S^*)}\|_2^2$. By taking $c_3=(1-\eta_0 b_5)/(4\eta_0 b_5)$, and noting that $\eta_0 b_5\leq 1$ and $b_2+b_4\leq b_1\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^{(t)}\|_2/24$, inequality (S.74) becomes

$$\begin{aligned}\widehat{R}_\varpi(\boldsymbol{\beta}^{(t+1)})-\widehat{R}_\varpi(\boldsymbol{\beta}^{(t)})\leq{}&\frac{4\eta_0 b_5 s^*}{s}\{\widehat{R}_\varpi(\boldsymbol{\beta}^*)-\widehat{R}_\varpi(\boldsymbol{\beta}^{(t)})\}+\frac{2b_5(1-2\eta_0 b_1)s^*}{s}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^{(t)}\|_2^2\\&+\frac{\eta_0 b_1 b_5 s^*}{6s}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^{(t)}\|_2^2+\eta_0\bigg(\frac{2\eta_0 b_5 s^*}{s}-\frac{1-\eta_0 b_5}{8}\bigg)\big\|\mathbf{g}^{(t)}_{J^t\setminus(S^t\cup S^*)}\big\|_2^2\\&+\frac{\eta_0^2 b_5}{2}\bigg(\frac{4s^*}{s}-1\bigg)\big\|\mathbf{g}^{(t)}_{S^t\cup S^*}\big\|_2^2+C_{\eta_0}W_t\,,\end{aligned}$$

where $C_{\eta_0}=4\{b_5(5-\eta_0 b_5)(2-2\eta_0 b_5)^{-1}+3(1-\eta_0 b_5)\eta_0^{-1}\}$. Under the assumption that $s\geq 38s^*(\eta_0 b_1)^{-2}$ and $\eta_0\leq 2b_1/(9b_5^2)$ with $b_1\leq b_5/2$, it holds that

$$\frac{2\eta_0 b_5 s^*}{s}\leq\frac{1-\eta_0 b_5}{8}\,,\quad\frac{4s^*}{s}-1\leq-\frac{1}{6}\,,\quad\text{and}\quad\frac{\{12(1-2\eta_0 b_1)+\eta_0 b_1\}b_5 s^*}{6s}\leq\frac{\eta_0^2 b_1^2 b_5}{18}\,.$$

We further have

$$\begin{aligned}\widehat{R}_\varpi(\boldsymbol{\beta}^{(t+1)})-\widehat{R}_\varpi(\boldsymbol{\beta}^{(t)})\leq{}&\frac{4\eta_0 b_5 s^*}{s}\{\widehat{R}_\varpi(\boldsymbol{\beta}^*)-\widehat{R}_\varpi(\boldsymbol{\beta}^{(t)})\}\\&-\frac{\eta_0^2 b_5}{12}\underbrace{\bigg(\big\|\mathbf{g}^{(t)}_{S^t\cup S^*}\big\|_2^2-\frac{2b_1^2}{3}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^{(t)}\|_2^2\bigg)}_{\Pi_2}+C_{\eta_0}W_t\,.\end{aligned}\tag{S.75}$$

As we will show in Section EC.7.6.2, conditioned on $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_4$ and $\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\|_2\geq 24(b_2+b_4)/b_1$,

$$\Pi_2\geq b_1\{\widehat{R}_\varpi(\boldsymbol{\beta}^{(t)})-\widehat{R}_\varpi(\boldsymbol{\beta}^*)\}\,. \tag{S.76}$$

Plugging this into (S.75) yields

$$\begin{aligned}\widehat{R}_\varpi(\boldsymbol{\beta}^{(t+1)})-\widehat{R}_\varpi(\boldsymbol{\beta}^*)&=\widehat{R}_\varpi(\boldsymbol{\beta}^{(t+1)})-\widehat{R}_\varpi(\boldsymbol{\beta}^{(t)})+\widehat{R}_\varpi(\boldsymbol{\beta}^{(t)})-\widehat{R}_\varpi(\boldsymbol{\beta}^*)\\&\leq\bigg(1-\frac{4\eta_0 b_5 s^*}{s}-\frac{\eta_0^2 b_1 b_5}{12}\bigg)\{\widehat{R}_\varpi(\boldsymbol{\beta}^{(t)})-\widehat{R}_\varpi(\boldsymbol{\beta}^*)\}+C_{\eta_0}W_t\,.\end{aligned}$$

We complete the proof of Proposition P4. $\square$

EC.7.6.1 Proof of (S.72)

Recall S^t is the index set S determined by Algorithm 2 when $\mathbf{v}$ is selected as $\tilde{\boldsymbol{\beta}}^{(t)}$. Based on the definition of $\boldsymbol{\beta}^{(t)}$, we know $\operatorname{supp}(\boldsymbol{\beta}^{(t)}) \subseteq S^t$, which implies $\boldsymbol{\beta}^{(t)}_{J^t\setminus(S^t\cup S^*)} = \mathbf{0}$. Since $\tilde{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)}$ conditioned on $\mathcal{E}_0$, we have

$$\eta_0^2\big\|\mathbf{g}^{(t)}_{J^t\setminus(S^t\cup S^*)}\big\|_2^2 = \big\|\{\boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)}\}_{J^t\setminus(S^t\cup S^*)}\big\|_2^2 = \big\|\tilde{\boldsymbol{\beta}}^{(t+1)}_{J^t\setminus(S^t\cup S^*)}\big\|_2^2 \tag{S.77}$$

conditioned on $\mathcal{E}_0$. Recall $J^t = S^t\cup S^{t+1}\cup S^*$ and $|S^t| = s = |S^{t+1}|$. Since

$$|S^t\setminus S^{t+1}| = |S^{t+1}\setminus S^t| \geq |S^{t+1}\setminus(S^t\cup S^*)| = |J^t\setminus(S^t\cup S^*)|\,,$$

there exists a set $R\subseteq S^t\setminus S^{t+1}$ with $|R| = |J^t\setminus(S^t\cup S^*)|$. Applying Lemma 3.4 in Cai et al. (2021) yields that, for any $c_3>0$,

$$\big\|\tilde{\boldsymbol{\beta}}^{(t+1)}_R\big\|_2^2 \leq (1+c_3)\big\|\tilde{\boldsymbol{\beta}}^{(t+1)}_{J^t\setminus(S^t\cup S^*)}\big\|_2^2 + 4(1+c_3^{-1})\sum_{i=1}^s\|\mathbf{w}_i^t\|_\infty^2\,. \tag{S.78}$$

As $R\subseteq S^t\setminus S^{t+1}$, we have $\boldsymbol{\beta}^{(t+1)}_R = \mathbf{0}$. By (S.77) and (S.78), it holds that

$$-\eta_0^2\big\|\mathbf{g}^{(t)}_{J^t\setminus(S^t\cup S^*)}\big\|_2^2 \leq -(1+c_3)^{-1}\big\|\{\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_R\big\|_2^2 + 4c_3^{-1}\sum_{i=1}^s\|\mathbf{w}_i^t\|_\infty^2 \tag{S.79}$$

conditioned on $\mathcal{E}_0$. By (S.79) and the facts that $(J^t\setminus R)\cup R = J^t$ and $(J^t\setminus R)\cap R = \emptyset$, we have

$$\begin{aligned}\Pi_1 &= \big\|\{\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_{J^t}\big\|_2^2 - \eta_0^2\big\|\mathbf{g}^{(t)}_{J^t\setminus(S^t\cup S^*)}\big\|_2^2\\ &\leq \big\|\{\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_{J^t\setminus R}\big\|_2^2 + c_3(1+c_3)^{-1}\big\|\{\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_R\big\|_2^2\\ &\quad + 4c_3^{-1}\sum_{i=1}^s\|\mathbf{w}_i^t\|_\infty^2\end{aligned} \tag{S.80}$$

conditioned on $\mathcal{E}_0$. In the sequel, we will bound the three terms on the right-hand side of the above inequality, respectively.

Conditioned on $\mathcal{E}_0$, it follows from (S.79) that

$$\big\|\{\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)} + \eta_0\mathbf{g}^{(t)}\}_R\big\|_2^2 \leq (1+c_3)\eta_0^2\big\|\mathbf{g}^{(t)}_{J^t\setminus(S^t\cup S^*)}\big\|_2^2 + 4c_3^{-1}(1+c_3)\sum_{i=1}^s\|\mathbf{w}_i^t\|_\infty^2\,. \tag{S.81}$$

For any $c_4>0$, due to $\boldsymbol{\beta}^{(t+1)} = \tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)}) + \tilde{\mathbf{w}}^t_{|S^{t+1}}$ and $\tilde{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_0\mathbf{g}^{(t)}$ conditioned on $\mathcal{E}_0$, it holds that

$$
\begin{aligned}
&\big\|\{\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}+\eta_0\mathbf{g}^{(t)}\}_{J^t\backslash R}\big\|_2^2 \\
&=\big\|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)})-\tilde{\boldsymbol{\beta}}^{(t+1)}+\tilde{\mathbf{w}}^t_{|S^{t+1}}\}_{|J^t\backslash R}\big\|_2^2 \overset{\text{(i)}}{=} \big\|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)})-\tilde{\boldsymbol{\beta}}^{(t+1)}\}_{|J^t\backslash R}+\tilde{\mathbf{w}}^t_{|S^{t+1}}\big\|_2^2 \\
&\le \big\|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)})-\tilde{\boldsymbol{\beta}}^{(t+1)}\}_{J^t\backslash R}\big\|_2^2+\|\tilde{\mathbf{w}}^t_{|S^{t+1}}\|_2^2+2\big\|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)})-\tilde{\boldsymbol{\beta}}^{(t+1)}\}_{J^t\backslash R}\big\|_2\|\tilde{\mathbf{w}}^t_{|S^{t+1}}\|_2 \\
&\le (1+c_4)\big\|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)})-\tilde{\boldsymbol{\beta}}^{(t+1)}\}_{J^t\backslash R}\big\|_2^2+(1+c_4^{-1})\|\tilde{\mathbf{w}}^t_{|S^{t+1}}\|_2^2\,,
\end{aligned}
\tag{S.82}
$$

where (i) follows from the fact that $S^{t+1}\subseteq J^t\backslash R$. By Lemma L9, for any $c_5>0$,

$$
\begin{aligned}
&\big\|\{\tilde{P}_s(\tilde{\boldsymbol{\beta}}^{(t+1)})-\tilde{\boldsymbol{\beta}}^{(t+1)}\}_{J^t\backslash R}\big\|_2^2 \\
&\le \big(1+c_5^{-1}\big)\frac{|J^t\backslash R|-s}{|J^t\backslash R|-s^*}\big\|\{\boldsymbol{\beta}^*-\tilde{\boldsymbol{\beta}}^{(t+1)}\}_{J^t\backslash R}\big\|_2^2+4(1+c_5)\sum_{i=1}^s\|\mathbf{w}_i^t\|_\infty^2 \\
&\le (1+c_5^{-1})\frac{s^*}{s}\big\|\{\boldsymbol{\beta}^*-\tilde{\boldsymbol{\beta}}^{(t+1)}\}_{J^t\backslash R}\big\|_2^2+4(1+c_5)\sum_{i=1}^s\|\mathbf{w}_i^t\|_\infty^2\,,
\end{aligned}
\tag{S.83}
$$

where the last step is based on the facts that

$$
|J^t\backslash R|=|J^t|-|R|=|J^t|-|J^t\backslash(S^t\cup S^*)|=|S^t\cup S^*|\le s+s^*
$$

and $|J^t\backslash R|\ge s\ge s^*$. Plugging (S.83) into (S.82) yields

$$
\begin{aligned}
\big\|\{\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}+\eta_0\mathbf{g}^{(t)}\}_{J^t\backslash R}\big\|_2^2 &\le (1+c_4)(1+c_5^{-1})\frac{s^*}{s}\big\|\{\boldsymbol{\beta}^*-\tilde{\boldsymbol{\beta}}^{(t+1)}\}_{J^t\backslash R}\big\|_2^2 \\
&\quad +4(1+c_4)(1+c_5)\sum_{i=1}^s\|\mathbf{w}_i^t\|_\infty^2+(1+c_4^{-1})\|\tilde{\mathbf{w}}^t_{|S^{t+1}}\|_2^2\,.
\end{aligned}
$$

As $\tilde{\boldsymbol{\beta}}^{(t+1)}=\boldsymbol{\beta}^{(t)}-\eta_0\mathbf{g}^{(t)}$ conditioned on $\mathcal{E}_0$, by taking $c_4=c_5=1$, we have

$$
\big\|\{\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}+\eta_0\mathbf{g}^{(t)}\}_{J^t\backslash R}\big\|_2^2\le\frac{4s^*}{s}\big\|\{\boldsymbol{\beta}^*-\boldsymbol{\beta}^{(t)}+\eta_0\mathbf{g}^{(t)}\}_{J^t\backslash R}\big\|_2^2+16\sum_{i=1}^s\|\mathbf{w}_i^t\|_\infty^2+2\|\tilde{\mathbf{w}}^t_{|S^{t+1}}\|_2^2\,.
\tag{S.84}
$$

Combining (S.80), (S.81) and (S.84), due to $W_t=\sum_{i=1}^s\|\mathbf{w}_i^t\|_\infty^2+\|\tilde{\mathbf{w}}^t_{|S^{t+1}}\|_2^2$, we conclude that

$$
\begin{aligned}
\Pi_1 &\le \frac{4s^*}{s}\big\|\{\boldsymbol{\beta}^*-\boldsymbol{\beta}^{(t)}+\eta_0\mathbf{g}^{(t)}\}_{J^t\backslash R}\big\|_2^2+c_3\eta_0^2\big\|\mathbf{g}^{(t)}_{J^t\backslash(S^t\cup S^*)}\big\|_2^2 \\
&\quad +20\sum_{i=1}^s\|\mathbf{w}_i^t\|_\infty^2+2\|\tilde{\mathbf{w}}^t_{|S^{t+1}}\|_2^2+4c_3^{-1}\sum_{i=1}^s\|\mathbf{w}_i^t\|_\infty^2 \\
&\le \frac{4s^*}{s}\big\|\{\boldsymbol{\beta}^*-\boldsymbol{\beta}^{(t)}+\eta_0\mathbf{g}^{(t)}\}_{J^t\backslash R}\big\|_2^2+c_3\eta_0^2\big\|\mathbf{g}^{(t)}_{J^t\backslash(S^t\cup S^*)}\big\|_2^2+4(5+c_3^{-1})W_t
\end{aligned}
$$

$$\leq \frac{4s^*}{s}\left\|\{\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)} + \eta_0 \mathbf{g}^{(t)}\}_{J^t}\right\|_2^2 + c_3\eta_0^2\left\|\mathbf{g}^{(t)}_{J^t\setminus(S^t\cup S^*)}\right\|_2^2 + 4(5+c_3^{-1})W_t$$

for any $c_3 > 0$. Hence, we obtain (S.72). □

EC.7.6.2 Proof of (S.76)

Under the assumption $\boldsymbol{\beta}^{(t)} \in \mathbb{H}(s) \cap \{\Theta(c_0)\setminus\Theta(n^{-1})\}$, and conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4$, it follows from (S.55) in Proposition P1 and $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 \geq 24(b_2+b_4)/b_1$ that

$$\begin{aligned}
\widehat{R}_\varpi(\boldsymbol{\beta}^{(t)}) - \widehat{R}_\varpi(\boldsymbol{\beta}^*) &\leq \langle \mathbf{g}^{(t)}, \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\rangle - b_1\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2^2 + (b_2+b_4)\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 \\
&\leq \frac{1}{b_1}\left\|\mathbf{g}^{(t)}_{S^t\cup S^*}\right\|_2^2 + \frac{b_1}{4}\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2^2 - b_1\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2^2 + \frac{b_1}{24}\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2^2 \\
&\leq \frac{1}{b_1}\left(\left\|\mathbf{g}^{(t)}_{S^t\cup S^*}\right\|_2^2 - \frac{2b_1^2}{3}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)}\|_2^2\right) = \frac{\Pi_2}{b_1},
\end{aligned}$$

which, in turn, implies $\Pi_2 \geq b_1\{\widehat{R}_\varpi(\boldsymbol{\beta}^{(t)}) - \widehat{R}_\varpi(\boldsymbol{\beta}^*)\}$. □

EC.8. Proof of Theorem T3

To show that $\mathcal{E} = \cap_{k=0}^7 \mathcal{E}_k$ holds with high probability, we rely on the technical lemmas presented in Section EC.8.1, with their proofs provided in Section EC.9.

EC.8.1. Technical Lemmas

Recall that $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$, $\widehat{D}_\varpi(\boldsymbol{\delta}) = \widehat{R}_\varpi(\boldsymbol{\beta}) - \widehat{R}_\varpi(\boldsymbol{\beta}^*)$, $D_\varpi(\boldsymbol{\delta}) = \mathbb{E}\{\widehat{D}_\varpi(\boldsymbol{\delta})\}$, and

$$A_\varpi(\boldsymbol{\delta}) = R_\varpi(\boldsymbol{\beta}) - R_\varpi(\boldsymbol{\beta}^*) - \langle \nabla R_\varpi(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle, B_\varpi(\boldsymbol{\delta}) = R_\varpi(\boldsymbol{\beta}^*) - R_\varpi(\boldsymbol{\beta}) - \langle \nabla R_\varpi(\boldsymbol{\beta}), \boldsymbol{\beta}^* - \boldsymbol{\beta}\rangle,$$

where $R_\varpi(\boldsymbol{\beta}) = \mathbb{E}\{\widehat{R}_\varpi(\boldsymbol{\beta})\}$.

Lemma L12 establishes lower bounds on the first-order Taylor remainder of the population (smoothed) loss in a neighborhood of $\boldsymbol{\beta}^*$, which is key to establishing the restricted strong convexity property of the smoothed empirical check loss. Its proof is provided in Section EC.9.4.

LEMMA L12. *Let Assumptions* 1–3 *hold. We have*

$$\min\{A_\varpi(\boldsymbol{\delta}), B_\varpi(\boldsymbol{\delta})\} \geq \frac{f_l\lambda_p - l_f k_1\lambda_1\varpi}{2}\|\boldsymbol{\delta}\|_2^2$$

for all $\boldsymbol{\delta} \in \mathbb{H}(2s) \cap \mathbb{B}^p(c_0)$.

Lemma L13 establishes a high-probability upper bound on $|\widehat{D}_\varpi(\boldsymbol{\delta}) - D_\varpi(\boldsymbol{\delta})|$. The proof is given in EC.9.5.

LEMMA L13. *Let Assumption* 1 *hold and* $r > a > 0$. *For any* $z \geq 0$, *we have*

(i) With probability at least $1 - e^{-z}$, *for all* $\boldsymbol{\delta} \in \mathbb{H}(2s) \cap \mathbb{B}^p(r)$,

$$\widehat{D}_{\varpi}(\boldsymbol{\delta}) - D_{\varpi}(\boldsymbol{\delta}) \lesssim \upsilon_1 \lambda_1^{1/2} r \sqrt{\frac{s \log p + z}{n}}\,; \tag{S.85}$$

(ii) With probability at least $1 - e^{-z}$, *for all* $\boldsymbol{\delta} \in \mathbb{H}(2s) \cap \{\mathbb{B}^p(r) \backslash \mathbb{B}^p(a)\}$,

$$\widehat{D}_{\varpi}(\boldsymbol{\delta}) - D_{\varpi}(\boldsymbol{\delta}) \lesssim \upsilon_1 \lambda_1^{1/2} \sqrt{\frac{s \log p + \log\log(r/a) + z}{n}} \|\boldsymbol{\delta}\|_2\,. \tag{S.86}$$

The same upper bounds in (S.85) *and* (S.86) *also apply to* $D_{\varpi}(\boldsymbol{\delta}) - \widehat{D}_{\varpi}(\boldsymbol{\delta})$.

Recall $\mathcal{S} = \{S \subseteq [p] : 1 \leq |S| \leq 2s\}$. Lemmas L14 and L15 establish the uniform convergence of the gradients of the smoothed empirical check loss to their population counterparts. The proofs of Lemmas L14 and L15 are given in Sections EC.9.6 and EC.9.7, respectively.

LEMMA L14. *Let Assumptions* 1–3 *hold. For any* $z \geq 0$,

$$\sup_{S \in \mathcal{S}} \big\|(1 - \mathbb{E})[\{\nabla \widehat{R}_{\varpi}(\boldsymbol{\beta}^*)\}_S]\big\|_2 \lesssim \upsilon_1 \lambda_1^{1/2} \bigg\{ C_{\tau,\varpi}^{1/2} \sqrt{\frac{s \log p + z}{n}} + \frac{s \log p + z}{n} \bigg\} \tag{S.87}$$

holds with probability at least $1 - e^{-z}$, *where* $C_{\tau,\varpi} = \tau(1-\tau) + (1+\tau) l_f k_2 \varpi^2$. *And*

$$\sup_{S \in \mathcal{S}} \big\|\mathbb{E}[\{\nabla \widehat{R}_{\varpi}(\boldsymbol{\beta}^*)\}_S]\big\|_2 \lesssim \lambda_1^{1/2} l_f k_2 \varpi^2\,. \tag{S.88}$$

LEMMA L15. *Let Assumptions* 1–3 *hold. For any* $z \geq 0$,

$$\begin{aligned} &\sup_{\boldsymbol{\beta} \in \mathbb{H}(s) \cap \Theta(r),\, S \in \mathcal{S}} \big\|(1 - \mathbb{E})[\{\nabla \widehat{R}_{\varpi}(\boldsymbol{\beta})\}_S]\big\|_2 \\ &\lesssim \upsilon_1 \lambda_1^{1/2} \bigg\{ \sqrt{\frac{s \log p + s \log(n/\varpi) + z}{n}} + \frac{s \log p + s \log(n/\varpi) + z}{n} \bigg\} + \frac{f_u \lambda_1 \varpi r}{n} + \frac{k_{\max} \lambda_1 r}{n} \end{aligned}$$

holds with probability at least $1 - e^{-z}$, *provided that* $n \gtrsim \upsilon_1^4 (s \log p + z)$.

Lemma L16 establishes the uniform bound for the Hessian of the smoothed empirical check loss, which plays a key role in establishing the restricted strong smoothness property of the smoothed empirical check loss. The proof of Lemma L16 is provided in Section EC.9.8.

LEMMA L16. *Let Assumptions* 1–3 *hold. For any* $z \geq 0$, *if* $n\varpi \gtrsim \bar{C} \upsilon_1^4 (s \log p + s \log n + z)$ *and* $n \gtrsim \tilde{C} \lambda_1^{1/4} r^{1/2} \{\upsilon_1^{1/2} (s \log p + \log n + z)^{1/4} + \varpi\}$ *for some* $r > 0$, *then, with probability at least*

$1-e^{-z}$,

$$\sup_{\boldsymbol{\beta}\in\mathbb{H}(2s)\cap\Theta(r),\,S\in\mathcal{S}} \|\{\nabla^2\widehat{R}_{\varpi}(\boldsymbol{\beta})\}_{SS}\| \le 2f_u\lambda_1\,,$$

where $\bar{C}>0$ is a constant depending only on $(k_{\max}, f_u)$, and $\tilde{C}>0$ is a constant depending only on $(l_K, l_f, f_u, \kappa_4)$.

Lemma L17 investigates the upper bound of $\|\{\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta})-\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta}^*)\}_S\|_2$. The proof of Lemma L17 is given in Section EC.9.9.

LEMMA L17. *Let Assumptions* 1–3 *hold. For any $z\ge 1/2$,*

$$\sup_{\boldsymbol{\beta}\in\mathbb{H}(s)\cap\Theta(r),\,S\in\mathcal{S}} \big\|\{\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta})-\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta}^*)\}_S\big\|_2 \le \frac{3}{2}f_u\lambda_1 r$$

holds with probability at least $1-e^{-z}$, provided that $\varpi\ge C_3\sqrt{(s+z)/n}$ and $n\varpi\ge C_4(s\log p+z)$, where $C_3>0$ and $C_4>0$ are two constants depending only on $(f_u, k_{\max}, \upsilon_1, \lambda_1)$.

EC.8.2. Proof of Theorem T3

Since $x_{i,1}=1$ for any $i\in[n]$, we have

$$\max_{i\in[n]}\|\mathbf{x}_i\|_\infty \le 1+\max_{i\in[n],\,j\in[p]\backslash\{1\}}|x_{i,j}|\,. \tag{S.89}$$

By (S.45), and taking the union bound over $i\in[n]$ and $j\in[p]\backslash\{1\}$, we have

$$\mathbb{P}\bigg\{\max_{i\in[n],j\in[p]\backslash\{1\}}|x_{i,j}|\ge \upsilon_1\lambda_1^{1/2}\sqrt{2\log\{2n(p-1)\}+2z}\bigg\}\le e^{-z}$$

for any $z\ge 0$. This, together with (S.89), yields, for any $z\ge 0$,

$$\max_{i\in[n]}\|\mathbf{x}_i\|_\infty \lesssim \upsilon_1\lambda_1^{1/2}\sqrt{\log p+\log n+z}$$

holds with probability at least $1-e^{-z}$. Then, for any $z\ge 0$, choosing $\gamma\asymp\upsilon_1\lambda_1^{1/2}\sqrt{\log p+\log n+z}$ ensures that the event $\mathcal{E}_0$ holds with probability at least $1-e^{-z}$. By Lemma L12, if $\varpi< f_l\lambda_p/(l_f k_1\lambda_1)$, the event $\mathcal{E}_1$ holds with $b_1=\frac{1}{2}(f_l\lambda_p-l_f k_1\lambda_1\varpi)$. By Lemma L13, for any $z\ge 0$, by selecting

$$b_2\asymp\upsilon_1\lambda_1^{1/2}\sqrt{\frac{s\log p+\log\log(c_0 n)+z}{n}}\,,$$

the event $\mathcal{E}_2$ holds with probability at least $1-e^{-z}$. By Lemma L14, for any $z \geq 0$, by selecting

$$b_3 \asymp \upsilon_1 \lambda_1^{1/2}\bigg\{C_{\tau,\varpi}^{1/2}\sqrt{\frac{s\log p + z}{n}} + \frac{s\log p + z}{n}\bigg\} + \lambda_1^{1/2} l_f k_2 \varpi^2\,,$$

the event $\mathcal{E}_3$ holds with probability at least $1-e^{-z}$, where $C_{\tau,\varpi} = \tau(1-\tau) + (1+\tau)l_f k_2 \varpi^2$. By Lemma L15, for any $z \geq 0$, by selecting

$$b_4 \asymp \upsilon_1 \lambda_1^{1/2}\bigg\{\sqrt{\frac{s\log p + s\log(n/\varpi) + z}{n}} + \frac{s\log p + s\log(n/\varpi) + z}{n}\bigg\} + \frac{f_u \lambda_1 \varpi c_0}{n} + \frac{k_{\max}\lambda_1 c_0}{n}\,,$$

the event $\mathcal{E}_4$ holds with probability at least $1-e^{-z}$, provided that $n \gtrsim \upsilon_1^4(s\log p + z)$. Note that $f_l \lambda_p \leq f_u \lambda_1$. By Lemmas L16 and L17, for any $z \geq 1/2$, by selecting $b_5 = 2f_u\lambda_1$ such that $b_1 \leq b_5/2$, the event $\mathcal{E}_5 \cap \mathcal{E}_6 \cap \mathcal{E}_7$ holds with probability at least $1-3e^{-z}$, provided that $n\varpi \gtrsim C_*(s\log p + s\log n + z)$ and $1 \gtrsim \varpi \gtrsim C_{**}\sqrt{(s+z)/n}$, where $C_*, C_{**} > 0$ are two constants depending only on $(l_K, l_f, f_u, k_{\max}, \kappa_4, \upsilon_1, \lambda_1)$. We complete the proof of Theorem T3. □

EC.9. Proofs of the Technical Lemmas

EC.9.1. Proof of Lemma L9

If $\tilde{S}\backslash S = \emptyset$ (i.e., $|\tilde{S}| = s$), then $\|\{\tilde{P}_s(\boldsymbol{\xi}) - \boldsymbol{\xi}\}_{\tilde{S}}\|_2^2 = 0$, and thus inequality (S.27) holds trivially. If $|\tilde{S}| > s$, let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)^\top$, and let H denote the index set of the top s coordinates (in magnitude) of $\boldsymbol{\xi}_{\tilde{S}}$. Then we have

$$\begin{aligned}
\|\{\tilde{P}_s(\boldsymbol{\xi}) - \boldsymbol{\xi}\}_{\tilde{S}}\|_2^2 &= \sum_{j\in\tilde{S}\backslash S}\xi_j^2 = \sum_{j\in(\tilde{S}\backslash S)\cap H^{\mathrm{c}}}\xi_j^2 + \sum_{j\in(\tilde{S}\backslash S)\cap H}\xi_j^2 \\
&\overset{\text{(i)}}{\leq} \sum_{j\in(\tilde{S}\backslash S)\cap H^{\mathrm{c}}}\xi_j^2 + (1+c^{-1})\sum_{j\in S\cap H^{\mathrm{c}}}\xi_j^2 + 4(1+c)\sum_{i\in[s]}\|\mathbf{w}_i\|_\infty^2 \\
&\leq (1+c^{-1})\sum_{j\in\tilde{S}\cap H^{\mathrm{c}}}\xi_j^2 + 4(1+c)\sum_{i\in[s]}\|\mathbf{w}_i\|_\infty^2\,,
\end{aligned} \tag{S.90}$$

where inequality (i) follows from Lemma 3.4 of Cai et al. (2021). Note that $\tilde{S}\cap H^{\mathrm{c}}$ corresponds to the indices of the smallest absolute values in $\boldsymbol{\xi}_{\tilde{S}}$. Let $\hat{\boldsymbol{\xi}} = (\hat{\xi}_1, \ldots, \hat{\xi}_p)^\top$, $\hat{S} = \mathrm{supp}(\hat{\boldsymbol{\xi}})$ and $\hat{S}_0 = \hat{S}\cap\tilde{S}$. Since $\hat{s}_0 := |\hat{S}_0| \leq \hat{s}$, it follows that

$$\frac{1}{|\tilde{S}|-s}\sum_{j\in\tilde{S}\cap H^{\mathrm{c}}}\xi_j^2 \leq \frac{1}{|\tilde{S}|-\hat{s}_0}\sum_{j\in\tilde{S}\cap\hat{S}_0^{\mathrm{c}}}\xi_j^2 \leq \frac{1}{|\tilde{S}|-\hat{s}}\sum_{j\in\tilde{S}\cap\hat{S}_0^{\mathrm{c}}}(\hat{\xi}_j - \xi_j)^2 \leq \frac{1}{|\tilde{S}|-\hat{s}}\|(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi})_{\tilde{S}}\|_2^2\,.$$

Combining this with (S.90) yields the desired bound in (S.27). □

EC.9.2. Proof of Lemma L10

If $\tilde{S}\backslash S=\emptyset$ (i.e., $|\tilde{S}|=s$), then $\|\{P_s(\boldsymbol{\xi})-\boldsymbol{\xi}\}_{\tilde{S}}\|_2^2=0$, and thus inequality (S.40) holds trivially. If $|\tilde{S}|>s$, we have

$$\|\{P_s(\boldsymbol{\xi})-\boldsymbol{\xi}\}_{\tilde{S}}\|_2^2=\sum_{j\in\tilde{S}\backslash S}\xi_j^2=\sum_{j\in\tilde{S}\cap S^{\mathrm{c}}}\xi_j^2\,. \tag{S.91}$$

Note that $\tilde{S}\cap S^{\mathrm{c}}$ corresponds to the indices of the smallest absolute values in $\boldsymbol{\xi}_{\tilde{S}}$. Let $\hat{\boldsymbol{\xi}}=(\hat{\xi}_1,\ldots,\hat{\xi}_p)^\top$, $\hat{S}=\mathrm{supp}(\hat{\boldsymbol{\xi}})$ and $\hat{S}_0=\hat{S}\cap\tilde{S}$. Since $\hat{s}_0:=|\hat{S}_0|\le\hat{s}$, it follows that

$$\frac{1}{|\tilde{S}|-s}\sum_{j\in\tilde{S}\cap S^{\mathrm{c}}}\xi_j^2\le\frac{1}{|\tilde{S}|-\hat{s}_0}\sum_{j\in\tilde{S}\cap\hat{S}_0^{\mathrm{c}}}\xi_j^2\le\frac{1}{|\tilde{S}|-\hat{s}}\sum_{j\in\tilde{S}\cap\hat{S}_0^{\mathrm{c}}}(\hat{\xi}_j-\xi_j)^2\le\frac{1}{|\tilde{S}|-\hat{s}}\|(\hat{\boldsymbol{\xi}}-\boldsymbol{\xi})_{\tilde{S}}\|_2^2\,.$$

Combining this with (S.91) yields the desired bound in (S.40). □

EC.9.3. Proof of Lemma L11

Write $\mathbf{w}_i^t=(w_{i,1}^t,\ldots,w_{i,p}^t)^\top$ and $\tilde{\mathbf{w}}^t=(\tilde{w}_1^t,\ldots,\tilde{w}_p^t)^\top$. Let $\lambda_{\mathrm{dp}}=2\epsilon^{-1}T\lambda\sqrt{5s\log(T/\delta)}$. For any $z>0$, we have $\mathbb{P}(|w_{i,j}^t|\ge\lambda_{\mathrm{dp}}z)\le e^{-z}$ and $\mathbb{P}(|\tilde{w}_j^t|\ge\lambda_{\mathrm{dp}}z)\le e^{-z}$ for all $i\in[s]$, $j\in[p]$, and $t\in\{0\}\cup[T-1]$. Taking a union bound over $i\in[s]$, $j\in[p]$, and $t\in\{0\}\cup[T-1]$, we have that for any $z>0$, with probability at least $1-e^{-z}$,

$$\max_{t\in\{0\}\cup[T-1]}\sum_{i=1}^{s}\|\mathbf{w}_i^t\|_\infty^2\le s\lambda_{\mathrm{dp}}^2\{z+\log(spT)\}^2\,,$$

and with probability at least $1-e^{-z}$,

$$\max_{t\in\{0\}\cup[T-1]}\|\tilde{\mathbf{w}}^t_{|S^{t+1}}\|_2^2\le s\lambda_{\mathrm{dp}}^2\{z+\log(sT)\}^2\,.$$

Combining the last two displays, there exists an absolute constant $C_0>0$ such that, for any $z>0$, with probability at least $1-e^{-z}$,

$$\max_{t\in\{0\}\cup[T-1]}W_t\le C_0\left(\frac{sT\lambda}{\epsilon}\right)^2\{\log(psT)+z\}^2\log\left(\frac{T}{\delta}\right),$$

as claimed. □

EC.9.4. Proof of Lemma L12

Recall that $A_\varpi(\boldsymbol{\delta})=R_\varpi(\boldsymbol{\beta})-R_\varpi(\boldsymbol{\beta}^*)-\langle\nabla R_\varpi(\boldsymbol{\beta}^*),\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle$ and $B_\varpi(\boldsymbol{\delta})=R_\varpi(\boldsymbol{\beta}^*)-R_\varpi(\boldsymbol{\beta})-\langle\nabla R_\varpi(\boldsymbol{\beta}),\boldsymbol{\beta}^*-\boldsymbol{\beta}\rangle$, where $R_\varpi(\cdot)=\mathbb{E}\{\widehat{R}_\varpi(\cdot)\}$ and $\boldsymbol{\delta}=\boldsymbol{\beta}-\boldsymbol{\beta}^*$. Based on the definition of $\widehat{R}_\varpi(\boldsymbol{\beta})$, applying a second-order Taylor expansion yields

$$A_\varpi(\boldsymbol{\delta}) = \frac{1}{2n}\sum_{i=1}^n \mathbb{E}\{K_\varpi(\varepsilon_i - m\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle)\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle^2\}$$

for some $m \in [0,1]$. By Assumptions 2 and 3, it holds that

$$\begin{aligned}
&\mathbb{E}\{K_\varpi(\varepsilon_i - m\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle)\,|\,\mathbf{x}_i\} = \int_{-\infty}^{\infty} K(v) f_{\varepsilon_i\,|\,\mathbf{x}_i}(m\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle + \varpi v)\,\mathrm{d}v \\
&\quad = f_{\varepsilon_i\,|\,\mathbf{x}_i}(m\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle) + \int_{-\infty}^{\infty} K(v)\{f_{\varepsilon_i\,|\,\mathbf{x}_i}(m\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle + \varpi v) - f_{\varepsilon_i\,|\,\mathbf{x}_i}(m\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle)\}\,\mathrm{d}v \\
&\quad \geq f_{\varepsilon_i\,|\,\mathbf{x}_i}(m\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle) - l_f k_1 \varpi\,.
\end{aligned}$$

Under the constraint $\|\boldsymbol{\delta}\|_2 \leq c_0$, it follows from Assumptions 1 and 3 that

$$\begin{aligned}
&2A_\varpi(\boldsymbol{\delta}) \\
&\geq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\bigg\{f_{\varepsilon_i\,|\,\mathbf{x}_i}\bigg(\underbrace{m\|\boldsymbol{\delta}\|_2}_{\leq c_0}\bigg\langle \mathbf{x}_i, \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2}\bigg\rangle\bigg)\bigg\langle \boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i, \frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}\|_2}\bigg\rangle^2 \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}\|_2^2 - l_f k_1 \varpi \langle \mathbf{x}_i, \boldsymbol{\delta}\rangle^2\bigg\} \\
&\geq (f_l\lambda_p - l_f k_1 \lambda_1 \varpi)\|\boldsymbol{\delta}\|_2^2
\end{aligned}$$

for all $\boldsymbol{\delta} \in \mathbb{H}(2s) \cap \mathbb{B}^p(c_0)$ with $\boldsymbol{\delta} \neq \mathbf{0}$. For $\boldsymbol{\delta} = \mathbf{0}$, we also have $2A_\varpi(\boldsymbol{\delta}) \geq 0 = (f_l\lambda_p - l_f k_1\lambda_1\varpi)\|\boldsymbol{\delta}\|_2^2$. Following the same arguments, we can also show $2B_\varpi(\boldsymbol{\delta}) \geq (f_l\lambda_p - l_f k_1\lambda_1\varpi)\|\boldsymbol{\delta}\|_2^2$ for all $\boldsymbol{\delta} \in \mathbb{H}(2s) \cap \mathbb{B}^p(c_0)$. We complete the proof of Lemma L12. □

EC.9.5. Proof of Lemma L13

Define $L_\varpi(u) = \int_{-\infty}^{\infty} \rho_\tau(v) K_\varpi(v-u)\,\mathrm{d}v$ with $\rho_\tau(u) = u\{\tau - \mathbb{1}(u<0)\}$, and let $d_\varpi(\mathbf{v}; \mathbf{z}_i) = L_\varpi(\varepsilon_i - \langle \mathbf{x}_i, \mathbf{v}\rangle) - L_\varpi(\varepsilon_i)$ with $\mathbf{z}_i = (\varepsilon_i, \mathbf{x}_i)$ for any $\mathbf{v} \in \mathbb{R}^p$. Then

$$\begin{aligned}
&\widehat{D}_\varpi(\boldsymbol{\delta}) - D_\varpi(\boldsymbol{\delta}) = (1-\mathbb{E})\{\widehat{R}_\varpi(\boldsymbol{\beta}) - \widehat{R}_\varpi(\boldsymbol{\beta}^*)\} \\
&\quad = \frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})\bigg[\int_{-\infty}^{\infty} \rho_\tau(u)\{K_\varpi(u + \mathbf{x}_i^\top\boldsymbol{\beta} - d_i) - K_\varpi(u + \mathbf{x}_i^\top\boldsymbol{\beta}^* - d_i)\}\,\mathrm{d}u\bigg] \\
&\quad = \frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})\{L_\varpi(\varepsilon_i - \langle \mathbf{x}_i, \boldsymbol{\delta}\rangle) - L_\varpi(\varepsilon_i)\} = \frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})\{d_\varpi(\boldsymbol{\delta}; \mathbf{z}_i)\}\,.
\end{aligned}$$

For any fixed $r > 0$, define the random variable

$$\aleph(r) = \frac{n^{1/2}}{4\bar{\tau}r} \sup_{\mathbf{v}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r)} \{\widehat{D}_\varpi(\mathbf{v}) - D_\varpi(\mathbf{v})\} = \frac{1}{4\bar{\tau}r} \sup_{\mathbf{v}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r)} \left[\frac{1}{n^{1/2}}\sum_{i=1}^n (1-\mathbb{E})\{d_\varpi(\mathbf{v};\mathbf{z}_i)\}\right], \tag{S.92}$$

where $\bar{\tau} = \max\{\tau, 1-\tau\}$. For any $z \geq 0$ and some $\lambda > 0$ to be determined later,

$$\mathbb{P}\{\aleph(r) \geq z\} \leq e^{-\lambda z}\mathbb{E}\big[\exp\{\lambda\aleph(r)\}\big]. \tag{S.93}$$

Then, it follows from Rademacher symmetrization that

$$\mathbb{E}\big[\exp\{\lambda\aleph(r)\}\big] \leq \mathbb{E}\left[\exp\left\{\frac{\lambda}{2\bar{\tau}r} \sup_{\mathbf{v}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r)} \frac{1}{n^{1/2}}\sum_{i=1}^n e_i d_\varpi(\mathbf{v};\mathbf{z}_i)\right\}\right], \tag{S.94}$$

where $e_1, \ldots, e_n$ are independent Rademacher random variables.

Note that $L_\varpi(\cdot)$ is $\bar{\tau}$-Lipschitz continuous. Therefore, for any $\mathbf{z}_i$ and $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^p$, it holds that $|d_\varpi(\mathbf{v};\mathbf{z}_i) - d_\varpi(\mathbf{v}';\mathbf{z}_i)| \leq \bar{\tau}|\langle\mathbf{x}_i, \mathbf{v}\rangle - \langle\mathbf{x}_i, \mathbf{v}'\rangle|$. Since $d_\varpi(\mathbf{v};\mathbf{z}_i) = 0$ if $\langle\mathbf{x}_i, \mathbf{v}\rangle = 0$, by (S.94), applying inequality (4.20) in Ledoux and Talagrand (2013) yields

$$\begin{aligned}
\mathbb{E}\big[\exp\{\lambda\aleph(r)\}\big] &\leq \mathbb{E}\left[\exp\left\{\frac{\lambda}{2\bar{\tau}r} \sup_{\mathbf{v}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r)} \frac{1}{n^{1/2}}\sum_{i=1}^n e_i d_\varpi(\mathbf{v};\mathbf{z}_i)\right\}\right] \\
&\leq \mathbb{E}\left[\exp\left\{\frac{\lambda}{2r} \sup_{\mathbf{v}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r)} \frac{1}{n^{1/2}}\sum_{i=1}^n e_i \langle\mathbf{x}_i, \mathbf{v}\rangle\right\}\right] \\
&\leq \mathbb{E}\left[\exp\left\{\frac{\lambda}{2} \sup_{\mathbf{u}\in\mathbb{H}(2s)\cap\mathbb{S}^{p-1}} \frac{1}{n^{1/2}}\sum_{i=1}^n e_i \langle\mathbf{x}_i, \mathbf{u}\rangle\right\}\right].
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{H}(2s)\cap\mathbb{S}^{p-1} &= \bigcup_{J\subseteq[p],\,|J|\leq 2s} \{\mathbf{u}\in\mathbb{R}^p : \|\mathbf{u}_J\|_2 = 1, \mathbf{u}_{J^c} = \mathbf{0}\} \\
&= \bigcup_{J\subseteq[p],\,|J|= 2s} \{\mathbf{u}\in\mathbb{R}^p : \|\mathbf{u}_J\|_2 = 1, \mathbf{u}_{J^c} = \mathbf{0}\}.
\end{aligned}$$

Since the cardinality of the ϵ-net of $\{\mathbf{u}\in\mathbb{R}^p : \|\mathbf{u}_J\|_2 = 1, \mathbf{u}_{J^c} = \mathbf{0}\}$ is equivalent to that of unit sphere $\mathbb{S}^{2s-1}$, Lemma 5.2 of Vershynin (2012) yields that the cardinality of the $1/2$-net of $\{\mathbf{u}\in\mathbb{R}^p : \|\mathbf{u}_J\|_2 = 1, \mathbf{u}_{J^c} = \mathbf{0}\}$ is at most 5^{2s}. Due to $\binom{p}{2s} \leq \{ep/(2s)\}^{2s}$, the total cardinality of the $1/2$-net of $\mathbb{H}(2s)\cap\mathbb{S}^{p-1}$ is at most $5^{2s}\{ep/(2s)\}^{2s} \leq (7p/s)^{2s}$. Hence, there exists a $1/2$-net

$\{\mathbf{u}_1, \ldots, \mathbf{u}_{N_{1/2}}\}$ of $\mathbb{H}(2s) \cap \mathbb{S}^{p-1}$ with $N_{1/2} \le (7p/s)^{2s}$ such that

$$\sup_{\mathbf{u} \in \mathbb{H}(2s) \cap \mathbb{S}^{p-1}} \frac{1}{n^{1/2}} \sum_{i=1}^n e_i \langle \mathbf{x}_i, \mathbf{u} \rangle \le 2 \max_{\ell \in [N_{1/2}]} \frac{1}{n^{1/2}} \sum_{i=1}^n e_i \langle \mathbf{x}_i, \mathbf{u}_\ell \rangle ,$$

which further implies

$$\mathbb{E}\big[\exp\{\lambda \aleph(r)\}\big] \le \mathbb{E}\bigg\{ \exp\bigg(\max_{\ell \in [N_{1/2}]} \frac{\lambda}{n^{1/2}} \sum_{i=1}^n e_i \langle \mathbf{x}_i, \mathbf{u}_\ell \rangle \bigg) \bigg\} \le \sum_{\ell=1}^{N_{1/2}} \mathbb{E}\bigg\{ \exp\bigg(\frac{\lambda}{n^{1/2}} \sum_{i=1}^n e_i \langle \mathbf{x}_i, \mathbf{u}_\ell \rangle \bigg) \bigg\} .$$

Write $S_\ell = n^{-1/2} \sum_{i=1}^n e_i \langle \mathbf{x}_i, \mathbf{u}_\ell \rangle$, which is a sum of zero-mean random variable. Since $\|\mathbf{u}_\ell\|_2 = 1$, by Assumption 1, we have

$$\mathbb{P}(|\langle \mathbf{x}_i, \mathbf{u}_\ell \rangle| \ge \upsilon_1 \lambda_1^{1/2} z) \le \mathbb{P}\bigg(\bigg| \bigg\langle \mathbf{\Sigma}^{-1/2} \mathbf{x}_i, \frac{\mathbf{\Sigma}^{1/2} \mathbf{u}_\ell}{\|\mathbf{\Sigma}^{1/2} \mathbf{u}_\ell\|_2} \bigg\rangle \bigg| \ge \upsilon_1 z \bigg) \le 2 e^{-z^2/2} .$$

Note that $e_i \in \{-1, 1\}$ is symmetric. Then, it holds that $\mathbb{E}\{\exp(c e_i \langle \mathbf{x}_i, \mathbf{u}_\ell \rangle)\} \le \exp(c^2 \upsilon_1^2 \lambda_1 / 2)$ for all $c \in \mathbb{R}$. Consequently, for each $\ell \in [N_{1/2}]$,

$$\mathbb{E}\{\exp(\lambda S_\ell)\} = \prod_{i=1}^n \mathbb{E}\bigg\{ \exp\bigg(\frac{\lambda}{n^{1/2}} e_i \langle \mathbf{x}_i, \mathbf{u}_\ell \rangle \bigg) \bigg\} \le \prod_{i=1}^n \exp\bigg(\frac{\lambda^2 \upsilon_1^2 \lambda_1}{2n} \bigg) = \exp\bigg(\frac{\lambda^2 \upsilon_1^2 \lambda_1}{2} \bigg) ,$$

which further implies $\log\big(\mathbb{E}\big[\exp\{\lambda \aleph(r)\}\big] \big) \le \log N_{1/2} + \frac{1}{2} \lambda^2 \upsilon_1^2 \lambda_1$. Then, by (S.93) and taking $\lambda = z / (\upsilon_1^2 \lambda_1)$, we obtain

$$\mathbb{P}\{\aleph(r) \ge z\} \le \exp\bigg(-\lambda z + \log N_{1/2} + \frac{\lambda^2 \upsilon_1^2 \lambda_1}{2} \bigg) = \exp\bigg(-\frac{z^2}{2 \upsilon_1^2 \lambda_1} + \log N_{1/2} \bigg) .$$

Since $N_{1/2} \le (7p/s)^{2s}$, we have $\aleph(r) \lesssim \upsilon_1 \lambda_1^{1/2} \sqrt{s \log(p/s) + z}$ with probability at least $1 - e^{-z}$. Combining this with (S.92) yields that, with probability at least $1 - e^{-z}$,

$$\sup_{\mathbf{v} \in \mathbb{H}(2s) \cap \mathbb{B}^p(r)} \{\widehat{D}_\varpi(\mathbf{v}) - D_\varpi(\mathbf{v})\} \le C_0 \upsilon_1 \lambda_1^{1/2} r \sqrt{\frac{s \log(p/s) + z}{n}} \tag{S.95}$$

where $C_0 > 0$ is an absolute constant. We complete the proof of (S.85).

Write $\Omega(r_l, r_u) = \{\mathbf{v} \in \mathbb{H}(2s) : r_l \le \|\mathbf{v}\|_2 \le r_u\}$. For some $M > 1$ to be determined later, and positive integers $k \in [N]$ with $N = \lceil \log(r_u / r_l) / \log M \rceil$, define the sets $\Omega_k = \{\mathbf{v} \in \mathbb{H}(2s) : M^{k-1} r_l \le \|\mathbf{v}\|_2 \le M^k r_l\}$, so that $\Omega(r_l, r_u) \subseteq \cup_{k=1}^N \Omega_k$. Then, by (S.95), for any $z \ge 0$,

$$\mathbb{P}\bigg[\exists\, \mathbf{v} \in \Omega(r_l, r_u) \;\text{ s.t. }\; \{\widehat{D}_\varpi(\mathbf{v}) - D_\varpi(\mathbf{v})\} > C_0 \upsilon_1 \lambda_1^{1/2} M \|\mathbf{v}\|_2 \sqrt{\frac{s \log(p/s) + z}{n}} \bigg]$$

$$
\begin{aligned}
&\le \sum_{k=1}^{N} \mathbb{P}\bigg[\exists\, \mathbf{v}\in\Omega_k \ \text{ s.t. } \ \{\widehat{D}_\varpi(\mathbf{v})-D_\varpi(\mathbf{v})\} > C_0\upsilon_1\lambda_1^{1/2}M^k r_l\sqrt{\frac{s\log(p/s)+z}{n}}\bigg] \\
&\le \sum_{k=1}^{N} \mathbb{P}\bigg[\sup_{\mathbf{v}\in\mathbb{H}(2s)\cap\mathbb{B}^p(M^k r_l)} \{\widehat{D}_\varpi(\mathbf{v})-D_\varpi(\mathbf{v})\} > C_0\upsilon_1\lambda_1^{1/2}M^k r_l\sqrt{\frac{s\log(p/s)+z}{n}}\bigg] \\
&\le e^{-z}\lceil \log(r_u/r_l)/\log M\rceil\,.
\end{aligned}
$$

Taking $M=e^{1/e}$ yields

$$
\sup_{\mathbf{v}\in\mathbb{H}(2s)\cap\{\mathbb{B}^p(r_u)\setminus\mathbb{B}^p(r_l)\}} \{\widehat{D}_\varpi(\mathbf{v})-D_\varpi(\mathbf{v})\} \le C_1\upsilon_1\lambda_1^{1/2}\sqrt{\frac{s\log(p/s)+\log\log(r_u/r_l)+z}{n}}\|\mathbf{v}\|_2
$$

with probability at least $1-e^{-z}$, where $C_1>0$ is an absolute constant. This proves (S.86) by taking $(r_l,r_u)=(r,a)$. Using analogous arguments, the bounds in (S.85) and (S.86) can also be established for $D_\varpi(\boldsymbol{\delta})-\widehat{D}_\varpi(\boldsymbol{\delta})$. We complete the proof of Lemma L13. □

EC.9.6. Proof of Lemma L14

Recall $\{\nabla\widehat{R}_\varpi(\boldsymbol{\beta}^*)\}_S = n^{-1}\sum_{i=1}^n\{\bar{K}_\varpi(-\varepsilon_i)-\tau\}\mathbf{x}_{i,S}$ for $S\in\mathcal{S}=\{S\subseteq[p]:1\le|S|\le 2s\}$. Due to $|S|\le 2s$, by Lemmas 5.2 and 5.3 of Vershynin (2012), there exists a $1/2$-net $\mathcal{N}_{1/2}$ of $\mathbb{S}^{|S|-1}$ with $|\mathcal{N}_{1/2}|\le 5^{2s}$ such that

$$
\begin{aligned}
\|(1-\mathbb{E})[\{\nabla\widehat{R}_\varpi(\boldsymbol{\beta}^*)\}_S]\|_2 &= \sup_{\mathbf{u}\in\mathbb{S}^{|S|-1}} \big((1-\mathbb{E})[\{\nabla\widehat{R}_\varpi(\boldsymbol{\beta}^*)\}_S]\big)^\top\mathbf{u} \\
&\le 2\max_{\mathbf{u}\in\mathcal{N}_{1/2}} \big((1-\mathbb{E})[\{\nabla\widehat{R}_\varpi(\boldsymbol{\beta}^*)\}_S]\big)^\top\mathbf{u}\,. \qquad \text{(S.96)}
\end{aligned}
$$

Note that $\bar{K}_\varpi(-\varepsilon_i)=\bar{K}(-\varepsilon_i/\varpi)=\int_{-\infty}^{-\varepsilon_i/\varpi}K(v)\,\mathrm{d}v$. Denote by $F_{\varepsilon_i\,|\,\mathbf{x}_i}(\cdot)$ the conditional distribution function of ε_i given $\mathbf{x}_i$. Then $\tau=F_{\varepsilon_i\,|\,\mathbf{x}_i}(0)$. By Assumptions 2 and 3, it holds that $|\bar{K}_\varpi(-\varepsilon_i)-\tau|\le\bar{\tau}:=\max\{1-\tau,\tau\}$, and

$$
\begin{aligned}
\mathbb{E}\{\bar{K}_\varpi^2(-\varepsilon_i)\,|\,\mathbf{x}_i\} &= \int_{-\infty}^{\infty}\bar{K}^2(-u/\varpi)\,\mathrm{d}F_{\varepsilon_i\,|\,\mathbf{x}_i}(u) \\
&\overset{\text{(i)}}{=} \int_{-\infty}^{\infty}F_{\varepsilon_i\,|\,\mathbf{x}_i}(-u\varpi)\,\mathrm{d}\bar{K}^2(u) = 2\int_{-\infty}^{\infty}K(u)\bar{K}(u)F_{\varepsilon_i\,|\,\mathbf{x}_i}(-u\varpi)\,\mathrm{d}u \\
&= 2\int_{-\infty}^{\infty}K(u)\bar{K}(u)F_{\varepsilon_i\,|\,\mathbf{x}_i}(0)\,\mathrm{d}u + 2\int_{-\infty}^{\infty}K(u)\bar{K}(u)\{F_{\varepsilon_i\,|\,\mathbf{x}_i}(-u\varpi)-F_{\varepsilon_i\,|\,\mathbf{x}_i}(0)\}\,\mathrm{d}u \\
&= 2\tau\int_{-\infty}^{\infty}K(u)\bar{K}(u)\,\mathrm{d}u + 2\int_{-\infty}^{\infty}K(u)\bar{K}(u)\bigg[\int_0^{-u\varpi}\{f_{\varepsilon_i\,|\,\mathbf{x}_i}(v)-f_{\varepsilon_i\,|\,\mathbf{x}_i}(0)\}\,\mathrm{d}v\bigg]\mathrm{d}u \\
&\quad - 2\varpi f_{\varepsilon_i\,|\,\mathbf{x}_i}(0)\int_{-\infty}^{\infty}uK(u)\bar{K}(u)\,\mathrm{d}u
\end{aligned}
$$

$$\overset{\text{(ii)}}{\leq} \tau + l_f \varpi^2 \int_{-\infty}^{\infty} u^2 K(u)\bar{K}(u)\,\mathrm{d}u \leq \tau + l_f k_2 \varpi^2\,, \tag{S.97}$$

where (i) follows from a change of variable and integration by parts, and (ii) follows from $\int_{-\infty}^{\infty} K(u)\bar{K}(u)\,\mathrm{d}u = 1/2$ and $\int_{-\infty}^{\infty} uK(u)\bar{K}(u)\,\mathrm{d}u = \int_0^{\infty} uK(u)\{\bar{K}(u)-\bar{K}(-u)\}\,\mathrm{d}u > 0$.

Due to $\tau = F_{\varepsilon_i\,|\,\mathbf{x}_i}(0)$, by Assumptions 2 and 3, we have

$$\begin{aligned}
&|\mathbb{E}\{\bar{K}_\varpi(-\varepsilon_i)-\tau\,|\,\mathbf{x}_i\}| = \bigg|\int_{-\infty}^{\infty} \bar{K}(-u/\varpi)\,\mathrm{d}F_{\varepsilon_i\,|\,\mathbf{x}_i}(u) - \tau\bigg| \\
&= \bigg|\int_{-\infty}^{\infty} K(u)F_{\varepsilon_i\,|\,\mathbf{x}_i}(-u\varpi)\,\mathrm{d}u - \tau\bigg| = \bigg|\int_{-\infty}^{\infty} K(u)\{F_{\varepsilon_i\,|\,\mathbf{x}_i}(-u\varpi) - F_{\varepsilon_i\,|\,\mathbf{x}_i}(0)\}\,\mathrm{d}u\bigg| \\
&\leq \bigg|\int_{-\infty}^{\infty} K(u)\int_0^{-u\varpi}\{f_{\varepsilon_i\,|\,\mathbf{x}_i}(v) - f_{\varepsilon_i\,|\,\mathbf{x}_i}(0)\}\,\mathrm{d}v\,\mathrm{d}u\bigg| + \varpi f_{\varepsilon_i\,|\,\mathbf{x}_i}(0)\bigg|\int_{-\infty}^{\infty} uK(u)\,\mathrm{d}u\bigg| \\
&\leq \frac{l_f k_2 \varpi^2}{2} + 0 = \frac{l_f k_2 \varpi^2}{2}\,,
\end{aligned} \tag{S.98}$$

where the last inequality follows from the symmetry of the kernel function $K(\cdot)$. Thus, by (S.97) and (S.98), we can conclude that

$$\begin{aligned}
\mathbb{E}\{|\bar{K}_\varpi(-\varepsilon_i)-\tau|^2\,|\,\mathbf{x}_i\} &= \tau^2 + \mathbb{E}\{\bar{K}_\varpi^2(-\varepsilon_i)\,|\,\mathbf{x}_i\} - 2\tau\mathbb{E}\{\bar{K}_\varpi(-\varepsilon_i)\,|\,\mathbf{x}_i\} \\
&= \tau^2 + \mathbb{E}\{\bar{K}_\varpi^2(-\varepsilon_i)\,|\,\mathbf{x}_i\} - 2\tau\mathbb{E}\{\bar{K}_\varpi(-\varepsilon_i)-\tau\,|\,\mathbf{x}_i\} - 2\tau^2 \\
&\leq \tau - \tau^2 + (1+\tau)l_f k_2\varpi^2 =: C_{\tau,\varpi}\,.
\end{aligned}$$

By Assumption 1, we have $\mathbb{P}(|\mathbf{x}_{i,S}^\top\mathbf{u}| > z) \leq 2e^{-z^2/(2\upsilon_1^2\lambda_1)}$ for any $\mathbf{u}\in\mathbb{S}^{|S|-1}$ and $z\geq 0$. Then, for any integer $q\geq 2$ and given $\mathbf{u}\in\mathbb{S}^{|S|-1}$, it holds that

$$\begin{aligned}
\mathbb{E}(|\mathbf{x}_{i,S}^\top\mathbf{u}|^q) &= \int_0^{\infty}\mathbb{P}(|\mathbf{x}_{i,S}^\top\mathbf{u}| > v^{1/q})\,\mathrm{d}v \\
&\leq 2q\int_0^{\infty} v^{q-1}e^{-v^2/(2\upsilon_1^2\lambda_1)}\,\mathrm{d}v = (\sqrt{2\lambda_1}\upsilon_1)^q q\Gamma(q/2) \leq (\sqrt{2\lambda_1}\upsilon_1)^q q!\,,
\end{aligned} \tag{S.99}$$

which implies that $\mathbb{E}[|\{\bar{K}_\varpi(-\varepsilon_i)-\tau\}\mathbf{x}_{i,S}^\top\mathbf{u}|^2] \leq 4C_{\tau,\varpi}\upsilon_1^2\lambda_1$. Due to $|\bar{K}_\varpi(-\varepsilon_i)-\tau|\leq\bar{\tau}$, we have

$$\begin{aligned}
\mathbb{E}[|\{\bar{K}_\varpi(-\varepsilon_i)-\tau\}\mathbf{x}_{i,S}^\top\mathbf{u}|^q] &\leq \bar{\tau}^{q-2}\cdot\mathbb{E}\{|\mathbf{x}_{i,S}^\top\mathbf{u}|^q\,\mathbb{E}(|\{\bar{K}_\varpi(-\varepsilon_i)-\tau\}|^2\,|\,\mathbf{x}_i)\} \\
&\leq \frac{q!}{2}(4C_{\tau,\varpi}\upsilon_1^2\lambda_1)(\sqrt{2\lambda_1}\upsilon_1\bar{\tau})^{q-2}
\end{aligned}$$

for all integer $q\geq 3$. Therefore, for any given $\mathbf{u}\in\mathbb{S}^{|S|-1}$ and $z\geq 0$, by Theorem 2.10 in Boucheron et al. (2013),

$$\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\big[\{\bar{K}_{\varpi}(-\varepsilon_i)-\tau\}\mathbf{x}_{i,S}^{\top}\mathbf{u}\big] \le \upsilon_1\lambda_1^{1/2}\left(2\sqrt{\frac{2C_{\tau,\varpi}z}{n}}+\frac{\sqrt{2}\bar{\tau}z}{n}\right) \tag{S.100}$$

holds with probability at least $1-e^{-z}$. For any fixed $\mathbf{u}\in\mathcal{N}_{1/2}$, we have

$$\big((1-\mathbb{E})[\{\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta}^*)\}_S]\big)^{\top}\mathbf{u} = \frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\big[\{\bar{K}_{\varpi}(-\varepsilon_i)-\tau\}\mathbf{x}_{i,S}^{\top}\mathbf{u}\big]\,.$$

Recall $|\mathcal{S}|\le\{ep/(2s)\}^{2s}$ and $|\mathcal{N}_{1/2}|\le 5^{2s}$. By (S.96) and (S.100), and taking union bound over all $S\in\mathcal{S}$ and $\mathbf{u}\in\mathcal{N}_{1/2}$, we have for any $z\ge 0$,

$$\sup_{S\in\mathcal{S}}\|(1-\mathbb{E})[\{\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta}^*)\}_S]\|_2 \lesssim \upsilon_1\lambda_1^{1/2}\left\{C_{\tau,\varpi}^{1/2}\sqrt{\frac{s\log(p/s)+z}{n}}+\bar{\tau}\,\frac{s\log(p/s)+z}{n}\right\}$$

holds with probability at least $1-e^{-z}$. This completes the proof of (S.87).

For any fixed $S\in\mathcal{S}$ and any given $\mathbf{u}\in\mathbb{S}^{|S|-1}$, by (S.98) and Assumption 1, we have

$$\mathbb{E}\big[\{\bar{K}_{\varpi}(-\varepsilon_i)-\tau\}\mathbf{x}_{i,S}^{\top}\mathbf{u}\big] \le \frac{l_f k_2\varpi^2}{2}\mathbb{E}(|\mathbf{x}_{i,S}^{\top}\mathbf{u}|) \le \frac{\lambda_1^{1/2}l_f k_2\varpi^2}{2}\,,$$

which further implies

$$\sup_{S\in\mathcal{S}}\|\mathbb{E}[\{\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta}^*)\}_S]\|_2 = \sup_{S\in\mathcal{S}}\sup_{\mathbf{u}\in\mathbb{S}^{|S|-1}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\{\bar{K}_{\varpi}(-\varepsilon_i)-\tau\}\mathbf{x}_{i,S}^{\top}\mathbf{u}\big] \le \frac{\lambda_1^{1/2}l_f k_2\varpi^2}{2}\,.$$

This completes the proof of (S.88). □

EC.9.7. Proof of Lemma L15

Recall $\{\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta})\}_S = n^{-1}\sum_{i=1}^{n}\{\bar{K}_{\varpi}(\langle\mathbf{x}_i,\boldsymbol{\delta}\rangle-\varepsilon_i)-\tau\}\mathbf{x}_{i,S}$, where $\boldsymbol{\delta}=\boldsymbol{\beta}-\boldsymbol{\beta}^*$ and $S\in\mathcal{S}=\{S\subseteq[p]:1\le|S|\le 2s\}$. Define a centered process

$$H_S(\boldsymbol{\delta}) = (1-\mathbb{E})[\{\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta})\}_S] = \frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})[\underbrace{\{\bar{K}_{\varpi}(\langle\mathbf{x}_i,\boldsymbol{\delta}\rangle-\varepsilon_i)-\tau\}}_{\zeta_{i,\boldsymbol{\delta}}}\mathbf{x}_{i,S}]\,.$$

Since $s^*\le s$, we have

$$\sup_{\boldsymbol{\beta}\in\mathbb{H}(s)\cap\Theta(r),\,S\in\mathcal{S}}\big\|(1-\mathbb{E})[\{\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta})\}_S]\big\|_2 \le \sup_{\boldsymbol{\delta}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r),\,S\in\mathcal{S}}\|H_S(\boldsymbol{\delta})\|_2\,. \tag{S.101}$$

By Lemma 3.3 of Plan and Vershynin (2013), there exists a ξ-net $\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{N_\xi}\}$ of $\mathbb{H}(2s)\cap\mathbb{B}^p(r)$ with $N_\xi\le\{9pr/(2s\xi)\}^{2s}$, such that for any $\boldsymbol{\delta}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r)$, there exists a $\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}\in\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{N_\xi}\}$

satisfying $\|\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}-\boldsymbol{\delta}\|_2\le\xi$. Furthermore, based on the proof for Lemma 3.3 of Plan and Vershynin (2013), we know $|\mathrm{supp}(\boldsymbol{\delta})\cup\mathrm{supp}(\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})})|\le 2s$ for any $\boldsymbol{\delta}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r)$. By the triangle inequality,

$$\sup_{S\in\mathcal{S}}\|H_S(\boldsymbol{\delta})\|_2\le\sup_{S\in\mathcal{S}}\|H_S(\boldsymbol{\delta})-H_S(\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})})\|_2+\sup_{S\in\mathcal{S}}\|H_S(\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})})\|_2\,.$$

This, together with (S.101), yields

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\mathbb{H}(s)\cap\Theta(r),\,S\in\mathcal{S}}\big\|(1-\mathbb{E})[\{\nabla\widehat{R}_\varpi(\boldsymbol{\beta})\}_S]\big\|_2\le&\sup_{\boldsymbol{\delta}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r)}\sup_{S\in\mathcal{S}}\|H_S(\boldsymbol{\delta})-H_S(\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})})\|_2\\&+\max_{\ell\in[N_\xi]}\sup_{S\in\mathcal{S}}\|H_S(\boldsymbol{\theta}_\ell)\|_2\,.\end{aligned}\tag{S.102}$$

We will show in Section EC.9.7.1 that, for any $z\ge0$,

$$\sup_{\boldsymbol{\delta}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r)}\sup_{S\in\mathcal{S}}\|H_S(\boldsymbol{\delta})-H_S(\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})})\|_2\le f_u\lambda_1\xi+\frac{2k_{\max}\lambda_1\xi}{\varpi}\tag{S.103}$$

holds with probability at least $1-e^{-z}$, provided that $n\gtrsim\upsilon_1^4(s\log p+z)$. Moreover, we will show in Section EC.9.7.2 that, for any $z\ge0$,

$$\max_{\ell\in[N_\xi]}\sup_{S\in\mathcal{S}}\|H_S(\boldsymbol{\theta}_\ell)\|_2\lesssim\upsilon_1\lambda_1^{1/2}\bigg[\sqrt{\frac{s\log\{pr/(s\xi)\}+z}{n}}+\frac{s\log\{pr/(s\xi)\}+z}{n}\bigg]\tag{S.104}$$

holds with probability at least $1-e^{-z}$, provided that $\xi\le r$. Taking $\xi\asymp r\varpi/n$ such that $\xi\le r$, together with (S.102), (S.103) and (S.104), for any $z\ge0$,

$$\begin{aligned}&\sup_{\boldsymbol{\beta}\in\mathbb{H}(s)\cap\Theta(r),\,S\in\mathcal{S}}\big\|(1-\mathbb{E})[\{\nabla\widehat{R}_\varpi(\boldsymbol{\beta})\}_S]\big\|_2\\&\quad\lesssim\upsilon_1\lambda_1^{1/2}\bigg\{\sqrt{\frac{s\log p+s\log(n/\varpi)+z}{n}}+\frac{s\log p+s\log(n/\varpi)+z}{n}\bigg\}+\frac{f_u\lambda_1\varpi r}{n}+\frac{k_{\max}\lambda_1 r}{n}\end{aligned}$$

holds with probability at least $1-e^{-z}$, provided that $n\gtrsim\upsilon_1^4(s\log p+z)$. This completes the proof of Lemma L15. □

EC.9.7.1 Proof of (S.103)

Recall $H_S(\boldsymbol{\delta})=n^{-1}\sum_{i=1}^n(1-\mathbb{E})(\zeta_{i,\boldsymbol{\delta}}\mathbf{x}_{i,S})$. By the triangle inequality,

$$\|H_S(\boldsymbol{\delta})-H_S(\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})})\|_2\le\frac{1}{n}\sum_{i=1}^n\|\mathbb{E}(\zeta_{i,\boldsymbol{\delta}}\mathbf{x}_{i,S})-\mathbb{E}(\zeta_{i,\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}}\mathbf{x}_{i,S})\|_2+\bigg\|\frac{1}{n}\sum_{i=1}^n(\zeta_{i,\boldsymbol{\delta}}\mathbf{x}_{i,S}-\zeta_{i,\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}}\mathbf{x}_{i,S})\bigg\|_2,$$

which implies that

$$\begin{aligned}
&\sup_{\boldsymbol{\delta}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r)}\sup_{S\in\mathcal{S}}\|H_S(\boldsymbol{\delta})-H_S(\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})})\|_2 \\
\le\ &\underbrace{\sup_{\boldsymbol{\delta}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r)}\sup_{S\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^n\|\mathbb{E}(\zeta_{i,\boldsymbol{\delta}}\mathbf{x}_{i,S})-\mathbb{E}(\zeta_{i,\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}}\mathbf{x}_{i,S})\|_2}_{\mathrm{I}_1} \\
&+\underbrace{\sup_{\boldsymbol{\delta}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r)}\sup_{S\in\mathcal{S}}\bigg\|\frac{1}{n}\sum_{i=1}^n(\zeta_{i,\boldsymbol{\delta}}\mathbf{x}_{i,S}-\zeta_{i,\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}}\mathbf{x}_{i,S})\bigg\|_2}_{\mathrm{I}_2}\,. \qquad \text{(S.105)}
\end{aligned}$$

In the sequel, we will derive the upper bounds for I_1 and I_2, respectively.

Denote by $F_{\varepsilon_i\,|\,\mathbf{x}_i}(\cdot)$ the conditional distribution function of ε_i given $\mathbf{x}_i$. Then

$$\mathbb{E}\{\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\delta}\rangle-\varepsilon_i)\,|\,\mathbf{x}_i\}=\int_{-\infty}^{\infty}\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\delta}\rangle-u)\,\mathrm{d}F_{\varepsilon_i\,|\,\mathbf{x}_i}(u)=\int_{-\infty}^{\infty}K(u)F_{\varepsilon_i\,|\,\mathbf{x}_i}(\langle\mathbf{x}_i,\boldsymbol{\delta}\rangle-u\varpi)\,\mathrm{d}u\,.$$

Due to $\|\boldsymbol{\delta}-\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}\|_2\le\xi$, by Assumptions 1 and 3, for any $S\in\mathcal{S}$ and any given $\mathbf{u}\in\mathbb{S}^{|S|-1}$, we have

$$\mathbb{E}[\{\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\delta}\rangle-\varepsilon_i)-\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}\rangle-\varepsilon_i)\}\langle\mathbf{x}_{i,S},\mathbf{u}\rangle]\le f_u\mathbb{E}(|\langle\mathbf{x}_i,\boldsymbol{\delta}-\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}\rangle\langle\mathbf{x}_{i,S},\mathbf{u}\rangle|)\le f_u\lambda_1\xi\,.$$

Since $\zeta_{i,\boldsymbol{\delta}}=\{\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\delta}\rangle-\varepsilon_i)-\tau\}$, for any $i\in[n]$ and $S\in\mathcal{S}$, we have

$$\begin{aligned}
&\|\mathbb{E}(\zeta_{i,\boldsymbol{\delta}}\mathbf{x}_{i,S})-\mathbb{E}(\zeta_{i,\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}}\mathbf{x}_{i,S})\|_2=\|\mathbb{E}[\{\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\delta}\rangle-\varepsilon_i)-\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}\rangle-\varepsilon_i)\}\mathbf{x}_{i,S}]\|_2 \\
=\ &\sup_{\mathbf{u}\in\mathbb{S}^{|S|-1}}\mathbb{E}\big[\{\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\delta}\rangle-\varepsilon_i)-\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}\rangle-\varepsilon_i)\}\langle\mathbf{x}_{i,S},\mathbf{u}\rangle\big]\le f_u\lambda_1\xi\,.
\end{aligned}$$

Since the above upper bound does not depend on $(\boldsymbol{\delta},S,i)$, we can conclude that

$$\mathrm{I}_1\le f_u\lambda_1\xi\,. \qquad \text{(S.106)}$$

Denote by $\widehat{\boldsymbol{\Sigma}}=n^{-1}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top$. Since $\|\boldsymbol{\delta}-\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}\|_2\le\xi$, by Assumption 2, for any $S\in\mathcal{S}$,

$$\begin{aligned}
&\bigg\|\frac{1}{n}\sum_{i=1}^n(\zeta_{i,\boldsymbol{\delta}}\mathbf{x}_{i,S}-\zeta_{i,\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}}\mathbf{x}_{i,S})\bigg\|_2 \\
=\ &\sup_{\mathbf{u}\in\mathbb{S}^{|S|-1}}\frac{1}{n}\sum_{i=1}^n\{\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\delta}\rangle-\varepsilon_i)-\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}\rangle-\varepsilon_i)\}\langle\mathbf{x}_{i,S},\mathbf{u}\rangle \\
\le\ &\sup_{\mathbf{u}\in\mathbb{S}^{|S|-1}}\frac{1}{n}\sum_{i=1}^n|\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\delta}\rangle-\varepsilon_i)-\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}\rangle-\varepsilon_i)|\cdot|\langle\mathbf{x}_{i,S},\mathbf{u}\rangle|
\end{aligned}$$

$$\leq \frac{k_{\max}}{\varpi} \sup_{\mathbf{u}\in\mathbb{S}^{|S|-1}} \frac{1}{n}\sum_{i=1}^{n} |\langle \mathbf{x}_i, \boldsymbol{\delta} - \boldsymbol{\theta}_{\ell(\boldsymbol{\delta})}\rangle \langle \mathbf{x}_{i,S}, \mathbf{u}\rangle| \leq \frac{k_{\max}\xi}{\varpi} \cdot \sup_{S\in\mathcal{S}} \|\widehat{\boldsymbol{\Sigma}}_{SS}\|\,,$$

which implies that

$$\mathrm{I}_2 \leq \frac{k_{\max}\xi}{\varpi} \cdot \sup_{S\in\mathcal{S}} \|\widehat{\boldsymbol{\Sigma}}_{SS}\|\,. \tag{S.107}$$

In the sequel, we will control $\sup_{S\in\mathcal{S}} \|\widehat{\boldsymbol{\Sigma}}_{SS}\|$.

Consider the decomposition

$$\sup_{S\in\mathcal{S}} \|\widehat{\boldsymbol{\Sigma}}_{SS}\| \leq \sup_{S\in\mathcal{S}} \|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{SS}\| + \sup_{S\in\mathcal{S}} \|\boldsymbol{\Sigma}_{SS}\| \leq \sup_{S\in\mathcal{S}} \|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{SS}\| + \lambda_1\,, \tag{S.108}$$

where the last inequality follows from Assumption 1. Notice that

$$\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{SS}\| = \sup_{\mathbf{v}\in\mathbb{S}^{|S|-1}} |\mathbf{v}^\top(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{SS}\mathbf{v}|\,. \tag{S.109}$$

Due to $|S| \leq 2s$, by Lemmas 5.2 and 5.4 of Vershynin (2012), there exists a $1/4$-net $\mathcal{N}_{1/4}(S)$ of $\mathbb{S}^{|S|-1}$ with $|\mathcal{N}_{1/4}(S)| \leq 9^{2s}$ such that

$$\sup_{\mathbf{v}\in\mathbb{S}^{|S|-1}} \left|\mathbf{v}^\top(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{SS}\mathbf{v}\right| \leq 2 \sup_{\mathbf{v}\in\mathcal{N}_{1/4}(S)} \left|\mathbf{v}^\top(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{SS}\mathbf{v}\right|. \tag{S.110}$$

By Assumption 1, we know $\mathbf{x}_i = (1, \mathbf{x}_{i,-1}^\top)^\top$ with $\mathbb{E}(\mathbf{x}_{i,-1}) = \mathbf{0}$. Then,

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top - \mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top) \\
&= \begin{bmatrix} 1 & n^{-1}\sum_{i=1}^{n} \mathbf{x}_{i,-1}^\top \\ n^{-1}\sum_{i=1}^{n} \mathbf{x}_{i,-1} & n^{-1}\sum_{i=1}^{n} \mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^\top \end{bmatrix} - \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathbb{E}(\mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^\top) \end{bmatrix} \\
&= \begin{bmatrix} 0 & n^{-1}\sum_{i=1}^{n} \mathbf{x}_{i,-1}^\top \\ n^{-1}\sum_{i=1}^{n} \mathbf{x}_{i,-1} & n^{-1}\sum_{i=1}^{n} \mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^\top - \mathbb{E}(\mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^\top) \end{bmatrix}.
\end{aligned}$$

For any $\mathbf{v} \in \mathbb{S}^{|S|-1}$, write $\mathbf{u}_{\mathbf{v}} = (u_{\mathbf{v},1}, \mathbf{u}_{\mathbf{v},-1}^\top)^\top \in \mathbb{S}^{p-1}$ such that $\mathbf{u}_{\mathbf{v},S} = \mathbf{v}$ and $\mathbf{u}_{\mathbf{v},S^c} = \mathbf{0}$, which implies

$$\begin{aligned}
\left|\mathbf{v}^\top(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{SS}\mathbf{v}\right| &= \left|\mathbf{u}_{\mathbf{v}}^\top(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\mathbf{u}_{\mathbf{v}}\right| = \left|(u_{\mathbf{v},1}, \mathbf{u}_{\mathbf{v},-1}^\top)(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})(u_{\mathbf{v},1}, \mathbf{u}_{\mathbf{v},-1}^\top)^\top\right| \\
&= \left|\frac{2}{n}\sum_{i=1}^{n} u_{\mathbf{v},1}\mathbf{x}_{i,-1}^\top\mathbf{u}_{\mathbf{v},-1} + \mathbf{u}_{\mathbf{v},-1}^\top\left\{\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^\top - \mathbb{E}(\mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^\top)\right\}\mathbf{u}_{\mathbf{v},-1}\right|
\end{aligned}$$

$$\leq \left|\frac{2}{n}\sum_{i=1}^{n} u_{\mathbf{v},1}\mathbf{x}_{i,-1}^{\top}\mathbf{u}_{\mathbf{v},-1}\right| + \left|\mathbf{u}_{\mathbf{v},-1}^{\top}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^{\top} - \mathbb{E}(\mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^{\top})\right\}\mathbf{u}_{\mathbf{v},-1}\right|.$$

Define $\mathcal{S}_* = \{S_* \subseteq [p-1] : 1 \leq |S_*| \leq 2s\}$. Therefore, by (S.109) and (S.110), we have

$$\begin{aligned}\sup_{S\in\mathcal{S}} \|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{SS}\| \leq & \sup_{S\in\mathcal{S}} \sup_{\mathbf{v}\in\mathcal{N}_{1/4}(S)} \left|\frac{4}{n}\sum_{i=1}^{n} u_{\mathbf{v},1}\mathbf{x}_{i,-1}^{\top}\mathbf{u}_{\mathbf{v},-1}\right| \\ & + 2\sup_{S_*\in\mathcal{S}_*}\left\|\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^{\top} - \mathbb{E}(\mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^{\top})\right\}_{S_*S_*}\right\|. \qquad \text{(S.111)}\end{aligned}$$

For any $\breve{\mathbf{u}} \in \mathbb{R}^{p-1}$ such that $0 < \|\breve{\mathbf{u}}\|_2 \leq 1$, define $\tilde{\mathbf{u}} = (0, \breve{\mathbf{u}}^{\top})^{\top} \in \mathbb{R}^p$. By Assumption 1, we have $\|\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{u}}\|_2 \leq \lambda_1^{1/2}$ and

$$\begin{aligned}&\mathbb{P}(|\langle\mathbf{x}_{i,-1}, \breve{\mathbf{u}}\rangle| \geq z) = \mathbb{P}(|\langle\boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i, \boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{u}}\rangle| \geq z) \\ &= \mathbb{P}\left(\left|\left\langle\boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i, \frac{\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{u}}}{\|\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{u}}\|_2}\right\rangle\right| \geq \frac{z}{\|\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{u}}\|_2}\right) \leq 2e^{-z^2/(2\upsilon_1^2\lambda_1)} \qquad \text{(S.112)}\end{aligned}$$

for any $z \geq 0$. Note that $|u_{\mathbf{v},1}| \leq 1$ and $\mathbf{u}_{\mathbf{v},-1} \in \mathbb{R}^{p-1}$ such that $\|\mathbf{u}_{\mathbf{v},-1}\|_2 \leq 1$. The tail bound in (S.112) also applies for $|u_{\mathbf{v},1}\mathbf{x}_{i,-1}^{\top}\mathbf{u}_{\mathbf{v},-1}|$. Since $\mathbb{E}(u_{\mathbf{v},1}\mathbf{x}_{i,-1}^{\top}\mathbf{u}_{\mathbf{v},-1}) = 0$, by Proposition 2.5.2 of Vershynin (2018), for all $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}\{\exp(\lambda u_{\mathbf{v},1}\mathbf{x}_{i,-1}^{\top}\mathbf{u}_{\mathbf{v},-1})\} \leq \exp\left(\frac{\lambda^2 C_0\upsilon_1^2\lambda_1}{2}\right),$$

where $C_0 > 0$ is an absolute constant. Then, for any given $\mathbf{v} \in \mathbb{S}^{|S|-1}$ and $z \geq 0$, by Proposition 2.5 of Wainwright (2019), it holds that

$$\left|\frac{4}{n}\sum_{i=1}^{n} u_{\mathbf{v},1}\mathbf{x}_{i,-1}^{\top}\mathbf{u}_{\mathbf{v},-1}\right| \lesssim \upsilon_1\lambda_1^{1/2}\sqrt{\frac{z}{n}} \qquad \text{(S.113)}$$

with probability at least $1 - e^{-z}$. Recall that $|\mathcal{N}_{1/4}(S)| \leq 9^{2s}$. By (S.113), for any fixed $S \in \mathcal{S}$, applying a union bound over all subsets in $\mathcal{N}_{1/4}(S)$ yields, for any $z \geq 0$,

$$\sup_{\mathbf{v}\in\mathcal{N}_{1/4}(S)}\left|\frac{4}{n}\sum_{i=1}^{n} u_{\mathbf{v},1}\mathbf{x}_{i,-1}^{\top}\mathbf{u}_{\mathbf{v},-1}\right| \lesssim \upsilon_1^2\lambda_1\sqrt{\frac{s+z}{n}} \qquad \text{(S.114)}$$

holds with probability at least $1 - e^{-z}$. Notice that $|\mathcal{S}| \leq \{ep/(2s)\}^{2s}$. By (S.114), taking a union bound over $S \in \mathcal{S}$ yields that, for any $z \geq 0$,

$$\sup_{S\in\mathcal{S}}\sup_{\mathbf{v}\in\mathcal{N}_{1/4}(S)}\left|\frac{4}{n}\sum_{i=1}^{n} u_{\mathbf{v},1}\mathbf{x}_{i,-1}^{\top}\mathbf{u}_{\mathbf{v},-1}\right| \lesssim \upsilon_1^2\lambda_1\sqrt{\frac{s\log p + z}{n}} \qquad \text{(S.115)}$$

holds with probability at least $1-e^{-z}$.

For a given $S_*\in\mathcal{S}_*$, since $\mathbb{E}(\mathbf{x}_{i,-1})=\mathbf{0}$, by (S.112) and Proposition 2.5.2 of Vershynin (2018), we know that for any $\mathbf{u}\in\mathbb{S}^{|S_*|-1}$ and $\lambda\in\mathbb{R}$, we have

$$\mathbb{E}\{\exp(\lambda\langle\mathbf{x}_{i,-1,S_*},\mathbf{u}\rangle)\}\le\exp\left(\frac{\lambda^2C_0\upsilon_1^2\lambda_1}{2}\right).$$

Therefore, due to $|S_*|\le 2s$, by Theorem 6.5 in Wainwright (2019), for any $z\ge 0$,

$$\left\|\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^{\top}-\mathbb{E}(\mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^{\top})\right\}_{S_*S_*}\right\|\lesssim\upsilon_1^2\lambda_1\left(\sqrt{\frac{s+z}{n}}+\frac{s+z}{n}\right)\tag{S.116}$$

holds with probability at least $1-e^{-z}$. Notice that $|\mathcal{S}_*|\le\{e(p-1)/(2s)\}^{2s}$. By (S.116) and applying a union bound over all subsets in $\mathcal{S}_*$ yields, for any $z\ge 0$,

$$\sup_{S_*\in\mathcal{S}_*}\left\|\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^{\top}-\mathbb{E}(\mathbf{x}_{i,-1}\mathbf{x}_{i,-1}^{\top})\right\}_{S_*S_*}\right\|\lesssim\upsilon_1^2\lambda_1\left(\sqrt{\frac{s\log p+z}{n}}+\frac{s\log p+z}{n}\right)\tag{S.117}$$

holds with probability at least $1-e^{-z}$. Combining (S.111), (S.115) and (S.117) yields that, for any $z\ge 0$,

$$\sup_{S\in\mathcal{S}}\|(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{SS}\|\lesssim\upsilon_1^2\lambda_1\left(\sqrt{\frac{s\log p+z}{n}}+\frac{s\log p+z}{n}\right)$$

holds with probability at least $1-e^{-z}$. Provided that $n\gtrsim\upsilon_1^4(s\log p+z)$, the right-hand side of the above inequality is bounded further by λ_1. Together with (S.108), we obtain that for any $z\ge 0$, as long as $n\gtrsim\upsilon_1^4(s\log p+z)$,

$$\sup_{S\in\mathcal{S}}\|\widehat{\boldsymbol{\Sigma}}_{SS}\|\le 2\lambda_1\tag{S.118}$$

holds with probability at least $1-e^{-z}$. By (S.107) and (S.118), for any $z\ge 0$,

$$\mathrm{I}_2\le\frac{2k_{\max}\lambda_1\xi}{\varpi}\tag{S.119}$$

holds with probability at least $1-e^{-z}$, provided that $n\gtrsim\upsilon_1^4(s\log p+z)$. Finally, by (S.105), (S.106) and (S.119), it holds that, for any $z\ge 0$,

$$\sup_{\boldsymbol{\delta}\in\mathbb{H}(2s)\cap\mathbb{B}^p(r)}\sup_{S\in\mathcal{S}}\|H_S(\boldsymbol{\delta})-H_S(\boldsymbol{\theta}_{\ell(\boldsymbol{\delta})})\|_2\le f_u\lambda_1\xi+\frac{2k_{\max}\lambda_1\xi}{\varpi}$$

holds with probability at least $1-e^{-z}$, provided that $n\gtrsim\upsilon_1^4(s\log p+z)$. We complete the proof of (S.103). $\square$

EC.9.7.2 Proof of (S.104)

Recall $H_S(\boldsymbol{\delta}) = n^{-1}\sum_{i=1}^n (1-\mathbb{E})(\zeta_{i,\boldsymbol{\delta}}\mathbf{x}_{i,S})$ for $\zeta_{i,\boldsymbol{\delta}} = \{\bar{K}_\varpi(\langle \mathbf{x}_i, \boldsymbol{\delta}\rangle - \varepsilon_i) - \tau\}$ and $S \in \mathcal{S} = \{S \subseteq [p] : 1 \le |S| \le 2s\}$. Due to $|S| \le 2s$, by Lemmas 5.2 and 5.3 of Vershynin (2012), there exists a $1/2$-net $\mathcal{N}_{1/2}$ of $\mathbb{S}^{|S|-1}$ with $|\mathcal{N}_{1/2}| \le 5^{2s}$ such that

$$\|H_S(\boldsymbol{\theta}_\ell)\|_2 = \sup_{\mathbf{u}\in\mathbb{S}^{|S|-1}} \left\{\frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})(\zeta_{i,\boldsymbol{\theta}_\ell}\mathbf{x}_{i,S})\right\}^\top \mathbf{u} \le 2\max_{\mathbf{u}\in\mathcal{N}_{1/2}} \left\{\frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})(\zeta_{i,\boldsymbol{\theta}_\ell}\mathbf{x}_{i,S})\right\}^\top \mathbf{u}\,. \tag{S.120}$$

Since $|\zeta_{i,\boldsymbol{\theta}_\ell}| \le \bar{\tau} = \max\{\tau, 1-\tau\}$, for any fixed $\mathbf{u}\in\mathbb{S}^{|S|-1}$, it follows by (S.99) in Section EC.9.6 that

$$\mathbb{E}\big(|\zeta_{i,\boldsymbol{\theta}_\ell}\langle \mathbf{x}_{i,S}, \mathbf{u}\rangle|^q\big) \le \frac{q!}{2}(4\upsilon_1^2\lambda_1\bar{\tau}^2)(\sqrt{2\lambda_1}\upsilon_1\bar{\tau})^{q-2}\,.$$

for any integer $q \ge 2$. Therefore, for any given $\mathbf{u}\in\mathbb{S}^{|S|-1}$ and $z \ge 0$, by Theorem 2.10 in Boucheron et al. (2013),

$$\frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})(\zeta_{i,\boldsymbol{\theta}_\ell}\langle \mathbf{x}_{i,S}, \mathbf{u}\rangle) \lesssim \upsilon_1\lambda_1^{1/2}\left(\sqrt{\frac{z}{n}} + \frac{z}{n}\right) \tag{S.121}$$

holds with probability at least $1 - e^{-z}$. For any fixed $\mathbf{u}\in\mathcal{N}_{1/2}$, we have

$$\left\{\frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})(\zeta_{i,\boldsymbol{\theta}_\ell}\mathbf{x}_{i,S})\right\}^\top \mathbf{u} = \frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})(\zeta_{i,\boldsymbol{\theta}_\ell}\langle \mathbf{x}_{i,S}, \mathbf{u}\rangle)\,.$$

Recall $N_\xi \le \{9pr/(2s\xi)\}^{2s}$, $|\mathcal{S}| \le \{ep/(2s)\}^{2s}$ and $|\mathcal{N}_{1/2}| \le 5^{2s}$. By (S.120) and (S.121), and taking union bound over all $\ell \in N_\xi$, $S \in \mathcal{S}$ and $\mathbf{u}\in\mathcal{N}_{1/2}$, we have for any $z \ge 0$,

$$\max_{\ell\in[N_\xi]}\sup_{S\in\mathcal{S}} \|H_S(\boldsymbol{\theta}_\ell)\|_2 \lesssim \upsilon_1\lambda_1^{1/2}\left[\sqrt{\frac{s\log\{pr/(s\xi)\} + z}{n}} + \frac{s\log\{pr/(s\xi)\} + z}{n}\right]$$

holds with probability at least $1 - e^{-z}$, provided that $\xi \le r$. We complete the proof of (S.104). □

EC.9.8. Proof of Lemma L16

Recall $\nabla^2\widehat{R}_\varpi(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n K_\varpi(\varepsilon_i - \langle \mathbf{x}_i, \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle)\mathbf{x}_i\mathbf{x}_i^\top$ and $K_\varpi(u) = \varpi^{-1}K(u/\varpi)$. Then,

$$\begin{aligned}\|\mathbb{E}[\{\nabla^2\widehat{R}_\varpi(\boldsymbol{\beta})\}_{SS}]\| &= \sup_{\mathbf{v}\in\mathbb{S}^{|S|-1}}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\{\int_{-\infty}^{\infty} K_\varpi(v - \langle \mathbf{x}_i, \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle) f_{\varepsilon_i\,|\,\mathbf{x}_i}(v)\,\mathrm{d}v\right\}\mathbf{v}^\top\mathbf{x}_{i,S}\mathbf{x}_{i,S}^\top\mathbf{v}\right]\right| \\ &= \sup_{\mathbf{v}\in\mathbb{S}^{|S|-1}}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\{\int_{-\infty}^{\infty} K(v) f_{\varepsilon_i\,|\,\mathbf{x}_i}(\varpi v + \langle \mathbf{x}_i, \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle)\,\mathrm{d}v\right\}\mathbf{v}^\top\mathbf{x}_{i,S}\mathbf{x}_{i,S}^\top\mathbf{v}\right]\right|\end{aligned}$$

$$\leq f_u\lambda_1\,,$$

where the last inequality follows from $\sup_{\mathbf{v}\in\mathbb{S}^{|S|-1}}\mathbb{E}(|\mathbf{v}^\top\mathbf{x}_{i,S}|^2)\leq\lambda_1$ (Assumption 1) and Assumption 3. As a result,

$$\sup_{\boldsymbol{\beta}\in\mathbb{H}(2s)\cap\Theta(r),\,S\in\mathcal{S}}\|\mathbb{E}[\{\nabla^2\widehat{R}_{\varpi}(\boldsymbol{\beta})\}_{SS}]\|\leq f_u\lambda_1\,. \tag{S.122}$$

We will show in Section EC.9.8.1 that, for any $z\geq 0$,

$$\begin{aligned}&\sup_{\boldsymbol{\beta}\in\mathbb{H}(2s)\cap\Theta(r),\,S\in\mathcal{S}}\|(1-\mathbb{E})[\{\nabla^2\widehat{R}_{\varpi}(\boldsymbol{\beta})\}_{SS}]\|\\&\quad\lesssim C_*\upsilon_1^2\lambda_1\bigg\{\sqrt{\frac{s\log(n/\varpi)+s\log p+z}{n\varpi}}+\frac{s\log(n/\varpi)+s\log p+z}{n\varpi}\bigg\}\\&\qquad+\big\{l_K\upsilon_1\lambda_1^{3/2}\sqrt{s\log p+z+\log n}+l_f\varpi^2\lambda_1^{3/2}\kappa_4^{3/4}\big\}\frac{r}{n^2}\end{aligned} \tag{S.123}$$

holds with probability at least $1-e^{-z}$, provided that $n\gtrsim\upsilon_1^4(s\log p+z)$, where $C_*>0$ is a constant depending only on $(k_{\max},f_u)$. By (S.123), we can conclude that, for any $z\geq 0$, it holds with probability at least $1-e^{-z}$ that

$$\sup_{\boldsymbol{\beta}\in\mathbb{H}(2s)\cap\Theta(r),\,S\in\mathcal{S}}\|(1-\mathbb{E})[\{\nabla^2\widehat{R}_{\varpi}(\boldsymbol{\beta})\}_{SS}]\|\leq f_u\lambda_1\,,$$

provided that $n\varpi\gtrsim\bar{C}\upsilon_1^4(s\log p+s\log n+z)$ and $n\gtrsim\tilde{C}\lambda_1^{1/4}r^{1/2}\{\upsilon_1^{1/2}(s\log p+\log n+z)^{1/4}+\varpi\}$, where $\bar{C}>0$ is a constant depending only on $(k_{\max},f_u)$, and $\tilde{C}>0$ is a constant depending only on (l_K,l_f,f_u,κ_4). Together with (S.122), for any $z\geq 0$, we have

$$\sup_{\boldsymbol{\beta}\in\mathbb{H}(2s)\cap\Theta(r),\,S\in\mathcal{S}}\|\{\nabla^2\widehat{R}_{\varpi}(\boldsymbol{\beta})\}_{SS}\|\leq 2f_u\lambda_1$$

holds with probability at least $1-e^{-z}$, provided that $n\varpi\gtrsim\bar{C}\upsilon_1^4(s\log p+s\log n+z)$ and $n\gtrsim\tilde{C}\lambda_1^{1/4}r^{1/2}\{\upsilon_1^{1/2}(s\log p+\log n+z)^{1/4}+\varpi\}$. We complete the proof of Lemma L16. □

EC.9.8.1 Proof of (S.123)

By Lemma 3.3 of Plan and Vershynin (2013), there exists a ξ-net $\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{N_\xi}\}$ of $\mathbb{H}(2s)\cap\Theta(r)$ with $N_\xi\leq\{9pr/(2s\xi)\}^{2s}$, such that for any $\boldsymbol{\beta}\in\mathbb{H}(2s)\cap\Theta(r)$, there exists a $\boldsymbol{\theta}_{\ell(\boldsymbol{\beta})}\in\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{N_\xi}\}$ satisfying $\|\boldsymbol{\theta}_{\ell(\boldsymbol{\beta})}-\boldsymbol{\beta}\|_2\leq\xi$. Furthermore, based on the proof for Lemma 3.3 of Plan and Vershynin (2013), we know $|\mathrm{supp}(\boldsymbol{\beta})\cup\mathrm{supp}(\boldsymbol{\theta}_{\ell(\boldsymbol{\beta})})|\leq 2s$ for any $\boldsymbol{\beta}\in\mathbb{H}(2s)\cap\Theta(r)$. By triangle inequality, for any given $\boldsymbol{\beta} \in \mathbb{H}(2s) \cap \Theta(r)$, it holds that

$$
\begin{aligned}
&\sup_{S \in \mathcal{S}} \|(1-\mathbb{E})[\{\nabla^2 \widehat{R}_\varpi(\boldsymbol{\beta})\}_{SS}]\| \\
&\quad \le \underbrace{\sup_{S \in \mathcal{S}} \|\{\nabla^2 \widehat{R}_\varpi(\boldsymbol{\beta}) - \nabla^2 \widehat{R}_\varpi(\boldsymbol{\theta}_{\ell(\boldsymbol{\beta})})\}_{SS}\|}_{\mathrm{I}_1(\boldsymbol{\beta})} + \underbrace{\sup_{S \in \mathcal{S}} \|\mathbb{E}[\{\nabla^2 \widehat{R}_\varpi(\boldsymbol{\beta}) - \nabla^2 \widehat{R}_\varpi(\boldsymbol{\theta}_{\ell(\boldsymbol{\beta})})\}_{SS}]\|}_{\mathrm{I}_2(\boldsymbol{\beta})} \\
&\qquad + \underbrace{\sup_{S \in \mathcal{S}} \|(1-\mathbb{E})[\{\nabla^2 \widehat{R}_\varpi(\boldsymbol{\theta}_{\ell(\boldsymbol{\beta})})\}_{SS}]\|}_{\mathrm{I}_3(\boldsymbol{\theta}_{\ell(\boldsymbol{\beta})})},
\end{aligned}
$$

which implies that

$$
\sup_{\boldsymbol{\beta} \in \mathbb{H}(2s) \cap \Theta(r)} \sup_{S \in \mathcal{S}} \|(1-\mathbb{E})[\{\nabla^2 \widehat{R}_\varpi(\boldsymbol{\beta})\}_{SS}]\| \le \sup_{\boldsymbol{\beta} \in \mathbb{H}(2s) \cap \Theta(r)} \mathrm{I}_1(\boldsymbol{\beta}) + \sup_{\boldsymbol{\beta} \in \mathbb{H}(2s) \cap \Theta(r)} \mathrm{I}_2(\boldsymbol{\beta}) + \max_{\ell \in [N_\xi]} \mathrm{I}_3(\boldsymbol{\theta}_\ell). \tag{S.124}
$$

Recall $\nabla^2 \widehat{R}_\varpi(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n K_\varpi(\varepsilon_i - \langle \mathbf{x}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle) \mathbf{x}_i \mathbf{x}_i^\top$ and $K_\varpi(u) = \varpi^{-1} K(u/\varpi)$. Due to $\|\boldsymbol{\beta} - \boldsymbol{\theta}_{\ell(\boldsymbol{\beta})}\|_2 \le \xi$, under Assumption 2, we have

$$
\begin{aligned}
\mathrm{I}_1(\boldsymbol{\beta}) &= \sup_{S \in \mathcal{S}} \left\| \frac{1}{n} \sum_{i=1}^n \{K_\varpi(\varepsilon_i - \langle \mathbf{x}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle) - K_\varpi(\varepsilon_i - \langle \mathbf{x}_i, \boldsymbol{\theta}_{\ell(\boldsymbol{\beta})} - \boldsymbol{\beta}^* \rangle)\} \mathbf{x}_{i,S} \mathbf{x}_{i,S}^\top \right\| \\
&= \sup_{S \in \mathcal{S}} \sup_{\mathbf{u} \in \mathbb{S}^{|S|-1}} \left| \frac{1}{n} \sum_{i=1}^n \{K_\varpi(\varepsilon_i - \langle \mathbf{x}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle) - K_\varpi(\varepsilon_i - \langle \mathbf{x}_i, \boldsymbol{\theta}_{\ell(\boldsymbol{\beta})} - \boldsymbol{\beta}^* \rangle)\} (\mathbf{x}_{i,S}^\top \mathbf{u})^2 \right| \\
&\le \frac{l_K}{\varpi^2} \cdot \|\boldsymbol{\theta}_{\ell(\boldsymbol{\beta})} - \boldsymbol{\beta}\|_2 \cdot \max_{i \in [n]} \|\mathbf{x}_{i, \mathrm{supp}(\boldsymbol{\beta}) \cup \mathrm{supp}(\boldsymbol{\theta}_{\ell(\boldsymbol{\beta})})}\|_2 \cdot \sup_{S \in \mathcal{S}} \|\widehat{\boldsymbol{\Sigma}}_{SS}\| \\
&\le \frac{l_K \xi}{\varpi^2} \cdot \max_{i \in [n]} \sup_{S \in \mathcal{S}} \|\mathbf{x}_{i,S}\|_2 \cdot \sup_{S \in \mathcal{S}} \|\widehat{\boldsymbol{\Sigma}}_{SS}\|,
\end{aligned} \tag{S.125}
$$

where $\widehat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$. This further implies that

$$
\sup_{\boldsymbol{\beta} \in \mathbb{H}(2s) \cap \Theta(r)} \mathrm{I}_1(\boldsymbol{\beta}) \le \frac{l_K \xi}{\varpi^2} \cdot \max_{i \in [n]} \sup_{S \in \mathcal{S}} \|\mathbf{x}_{i,S}\|_2 \cdot \sup_{S \in \mathcal{S}} \|\widehat{\boldsymbol{\Sigma}}_{SS}\|. \tag{S.126}
$$

By Lemmas 5.2 and 5.3 of Vershynin (2012), $\|\mathbf{x}_{i,S}\|_2 \le 2 \max_{\mathbf{u} \in \mathcal{N}_{1/2}} |\langle \mathbf{u}, \mathbf{x}_{i,S} \rangle|$, where $\mathcal{N}_{1/2}$ is the $1/2$-net of $\mathbb{S}^{|S|-1}$ satisfying $|\mathcal{N}_{1/2}| \le 5^{2s}$. For each $\mathbf{u} \in \mathcal{N}_{1/2}$, define $\tilde{\mathbf{u}} \in \mathbb{R}^p$ such that $\tilde{\mathbf{u}}_S = \mathbf{u}$ and $\tilde{\mathbf{u}}_{S^c} = \mathbf{0}$. Let $\widetilde{\mathcal{N}}_{1/2}$ denote the collection of all such $\tilde{\mathbf{u}}$ induced by $\mathcal{N}_{1/2}$. By Assumption 1, for any $z \ge 0$,

$$
\mathbb{P}(\|\mathbf{x}_{i,S}\|_2 > z) \le \mathbb{P}\bigg(2 \max_{\tilde{\mathbf{u}} \in \widetilde{\mathcal{N}}_{1/2}} |\langle \boldsymbol{\Sigma}^{1/2} \tilde{\mathbf{u}}, \boldsymbol{\Sigma}^{-1/2} \mathbf{x}_i \rangle| > z\bigg) \le 2 \cdot 5^{2s} e^{-z^2/(8 v_1^2 \lambda_1)},
$$

which implies that, for any $z \geq 0$, with probability at least $1 - e^{-z}$, $\|\mathbf{x}_{i,S}\|_2 \leq 2\sqrt{2}\upsilon_1\lambda_1^{1/2}\sqrt{2s\log 5 + z + \log 2}$.

Due to $|\mathcal{S}| \leq \{ep/(2s)\}^{2s}$, then, with probability at least $1-e^{-z}$, $\max_{i\in[n]}\sup_{S\in\mathcal{S}}\|\mathbf{x}_{i,S}\|_2 \lesssim \upsilon_1\lambda_1^{1/2}\sqrt{s\log p + z + \log n}$. Together with (S.118) in Section EC.9.7.1 and (S.126), for any $z \geq 0$, we have

$$\sup_{\boldsymbol{\beta}\in\mathbb{H}(2s)\cap\Theta(r)} \mathrm{I}_1(\boldsymbol{\beta}) \lesssim \frac{l_K\xi\upsilon_1\lambda_1^{3/2}}{\varpi^2}\sqrt{s\log p + z + \log n} \tag{S.127}$$

holds with probability at least $1-e^{-z}$, provided that $n \gtrsim \upsilon_1^4(s\log p + z)$.

Write $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ and $\boldsymbol{\delta}_{\ell(\boldsymbol{\beta})} = \boldsymbol{\theta}_{\ell(\boldsymbol{\beta})} - \boldsymbol{\beta}^*$. By Assumptions 2 and 3, it holds that

$$\begin{aligned}
&\big|\mathbb{E}\big[\big\{K_\varpi(\varepsilon_i - \langle\mathbf{x}_i,\boldsymbol{\delta}\rangle) - K_\varpi(\varepsilon_i - \langle\mathbf{x}_i,\boldsymbol{\delta}_{\ell(\boldsymbol{\beta})}\rangle)\big\}\,|\,\mathbf{x}_i\big]\big| \\
&= \left|\frac{1}{\varpi}\left\{\int_{-\infty}^{\infty} K\left(\frac{v - \langle\mathbf{x}_i,\boldsymbol{\delta}\rangle}{\varpi}\right) f_{\varepsilon_i\,|\,\mathbf{x}_i}(v)\,\mathrm{d}v - \int_{-\infty}^{\infty} K\left(\frac{v - \langle\mathbf{x}_i,\boldsymbol{\delta}_{\ell(\boldsymbol{\beta})}\rangle}{\varpi}\right) f_{\varepsilon_i\,|\,\mathbf{x}_i}(v)\,\mathrm{d}v\right\}\right| \\
&= \left|\int_{-\infty}^{\infty} K(v)\big\{f_{\varepsilon_i\,|\,\mathbf{x}_i}(\varpi v + \langle\mathbf{x}_i,\boldsymbol{\delta}\rangle) - f_{\varepsilon_i\,|\,\mathbf{x}_i}(\varpi v + \langle\mathbf{x}_i,\boldsymbol{\delta}_{\ell(\boldsymbol{\beta})}\rangle)\big\}\,\mathrm{d}v\right| \\
&\leq l_f|\langle\mathbf{x}_i,\boldsymbol{\delta} - \boldsymbol{\delta}_{\ell(\boldsymbol{\beta})}\rangle|\cdot\int_{-\infty}^{\infty} K(v)\,\mathrm{d}v = l_f|\langle\mathbf{x}_i,\boldsymbol{\delta} - \boldsymbol{\delta}_{\ell(\boldsymbol{\beta})}\rangle|\,.
\end{aligned}$$

Due to $\|\boldsymbol{\delta} - \boldsymbol{\delta}_{\ell(\boldsymbol{\beta})}\|_2 \leq \xi$ and $\sup_{\mathbf{u}\in\mathbb{S}^{p-1}}\mathbb{E}(|\langle\mathbf{x}_i,\mathbf{u}\rangle|^3) \leq \lambda_1^{3/2}\sup_{\mathbf{u}\in\mathbb{S}^{p-1}}\mathbb{E}(|\langle\boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i,\mathbf{u}\rangle|^3) \leq \lambda_1^{3/2}\kappa_4^{3/4}$, we have

$$\begin{aligned}
&\big\|\mathbb{E}\big[\{\nabla^2\widehat{R}_\varpi(\boldsymbol{\beta}) - \nabla^2\widehat{R}_\varpi(\boldsymbol{\theta}_{\ell(\boldsymbol{\beta})})\}_{SS}\big]\big\| \\
&= \left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\{K_\varpi(\varepsilon_i - \langle\mathbf{x}_i,\boldsymbol{\delta}\rangle) - K_\varpi(\varepsilon_i - \langle\mathbf{x}_i,\boldsymbol{\delta}_{\ell(\boldsymbol{\beta})}\rangle)\}\mathbf{x}_{i,S}\mathbf{x}_{i,S}^\top\big]\right\| \\
&= \sup_{\mathbf{v}\in\mathbb{S}^{|S|-1}}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\{K_\varpi(\varepsilon_i - \langle\mathbf{x}_i,\boldsymbol{\delta}\rangle) - K_\varpi(\varepsilon_i - \langle\mathbf{x}_i,\boldsymbol{\delta}_{\ell(\boldsymbol{\beta})}\rangle)\}\mathbf{v}^\top\mathbf{x}_{i,S}\mathbf{x}_{i,S}^\top\mathbf{v}\big]\right| \\
&\leq \frac{1}{n}\sum_{i=1}^n \sup_{\mathbf{v}\in\mathbb{S}^{|S|-1}}\big|\mathbb{E}\big(l_f|\langle\mathbf{x}_i,\boldsymbol{\delta} - \boldsymbol{\delta}_{\ell(\boldsymbol{\beta})}\rangle|\mathbf{v}^\top\mathbf{x}_{i,S}\mathbf{x}_{i,S}^\top\mathbf{v}\big)\big| \\
&\leq l_f\xi\sup_{\mathbf{u}\in\mathbb{S}^{p-1}}\mathbb{E}(|\langle\mathbf{x}_i,\mathbf{u}\rangle|^3) \leq l_f\xi\lambda_1^{3/2}\kappa_4^{3/4}
\end{aligned}$$

holds uniformly over $S\in\mathcal{S}$, which further implies that

$$\sup_{\boldsymbol{\beta}\in\mathbb{H}(2s)\cap\Theta(r)} \mathrm{I}_2(\boldsymbol{\beta}) \leq l_f\xi\lambda_1^{3/2}\kappa_4^{3/4}\,. \tag{S.128}$$

For any given $S \in \mathcal{S}$, by Lemmas 5.2 and 5.4 of Vershynin (2012), it holds that

$$\begin{aligned}\|(1-\mathbb{E})[\{\nabla^2\widehat{R}_\varpi(\boldsymbol{\theta}_\ell)\}_{SS}]\| &= \sup_{\mathbf{u}\in\mathbb{S}^{|S|-1}}\left|\mathbf{u}^\top\left[\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\{K_\varpi(d_i-\mathbf{x}_i^\top\boldsymbol{\theta}_\ell)\}\mathbf{x}_{i,S}\mathbf{x}_{i,S}^\top\right]\mathbf{u}\right| \\ &\le 2\sup_{\mathbf{u}\in\mathcal{N}_{1/4}}\left|\mathbf{u}^\top\left[\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\{K_\varpi(d_i-\mathbf{x}_i^\top\boldsymbol{\theta}_\ell)\}\mathbf{x}_{i,S}\mathbf{x}_{i,S}^\top\right]\mathbf{u}\right|, \quad \text{(S.129)}\end{aligned}$$

where $\mathcal{N}_{1/4}$ is the $1/4$-net of $\mathbb{S}^{|S|-1}$ satisfying $|\mathcal{N}_{1/4}|\le 9^{2s}$. Let $\phi_{i,\mathbf{u}} = K_\varpi(d_i-\mathbf{x}_i^\top\boldsymbol{\theta}_\ell)(\mathbf{x}_{i,S}^\top\mathbf{u})^2$, where $i\in[n]$ and $\mathbf{u}\in\mathcal{N}_{1/4}$. By (S.99) in Section EC.9.6, for any fixed $\mathbf{u}\in\mathbb{S}^{|S|-1}$ and any integer $q\ge 2$, it holds that

$$\mathbb{E}(|\mathbf{x}_{i,S}^\top\mathbf{u}|^{2q}) \le 2(2\upsilon_1^2\lambda_1)^q q\Gamma(q) = \frac{q!}{2}4(2\upsilon_1^2\lambda_1)^q\,.$$

By Assumptions 2 and 3, it holds that $|K_\varpi(d_i-\mathbf{x}_i^\top\boldsymbol{\theta}_\ell)|\le\varpi^{-1}k_{\max}$, and

$$\begin{aligned}\mathbb{E}\{K_\varpi^2(d_i-\mathbf{x}_i^\top\boldsymbol{\theta}_\ell)\,|\,\mathbf{x}_i\} &= \frac{1}{\varpi^2}\int_{-\infty}^{\infty}K^2\left(\frac{v-\langle\mathbf{x}_i,\boldsymbol{\theta}_\ell-\boldsymbol{\beta}^*\rangle}{\varpi}\right)f_{\varepsilon_i\,|\,\mathbf{x}_i}(v)\,\mathrm{d}v \\ &= \frac{1}{\varpi}\int_{-\infty}^{\infty}K^2(v)f_{\varepsilon_i\,|\,\mathbf{x}_i}(\varpi v+\langle\mathbf{x}_i,\boldsymbol{\theta}_\ell-\boldsymbol{\beta}^*\rangle)\,\mathrm{d}v \le \frac{k_{\max}f_u}{\varpi}\,.\end{aligned}$$

Thus, for any fixed $\mathbf{u}\in\mathbb{S}^{|S|-1}$, we have $\mathbb{E}(\phi_{i,\mathbf{u}}^2)\le\varpi^{-1}k_{\max}f_u(4\upsilon_1^2\lambda_1)^2$ and

$$\mathbb{E}(|\phi_{i,\mathbf{u}}|^q) \le \frac{q!}{2}\frac{k_{\max}f_u(4\upsilon_1^2\lambda_1)^2}{\varpi}\left(\frac{2k_{\max}\upsilon_1^2\lambda_1}{\varpi}\right)^{q-2}$$

for any integer $q\ge 3$. Therefore, for any fixed $\mathbf{u}\in\mathbb{S}^{|S|-1}$ and any $z\ge 0$, by Theorem 2.10 of Boucheron et al. (2013),

$$\left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\phi_{i,\mathbf{u}}\right| \lesssim \sqrt{\frac{k_{\max}f_u(\upsilon_1^2\lambda_1)^2 z}{n\varpi}} + \frac{k_{\max}\upsilon_1^2\lambda_1 z}{n\varpi} \qquad \text{(S.130)}$$

holds with probability at least $1-e^{-z}$. Note that, for any fixed $\mathbf{u}\in\mathbb{S}^{|S|-1}$,

$$\left|\mathbf{u}^\top\left[\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\{K_\varpi(d_i-\mathbf{x}_i^\top\boldsymbol{\theta}_\ell)\}\mathbf{x}_{i,S}\mathbf{x}_{i,S}^\top\right]\mathbf{u}\right| = \left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\phi_{i,\mathbf{u}}\right|.$$

Recall $|\mathcal{N}_{1/4}|\le 9^{2s}$. By (S.129) and (S.130), and taking a union bound over $\mathbf{u}\in\mathcal{N}_{1/4}$, we have

$$\|(1-\mathbb{E})[\{\nabla^2\widehat{R}_\varpi(\boldsymbol{\theta}_\ell)\}_{SS}]\| \le C_*\upsilon_1^2\lambda_1\left(\sqrt{\frac{s+z}{n\varpi}}+\frac{s+z}{n\varpi}\right).$$

holds with probability at least $1-e^{-z}$, where $C_*>0$ is a constant depending only on $(k_{\max},f_u)$. Recall $|\mathcal{S}|\le\{ep/(2s)\}^{2s}$ and $N_\xi\le\{9pr/(2s\xi)\}^{2s}$. Taking a union bound over $S\in\mathcal{S}$ and $\ell\in[N_\xi]$, it follows that for any $z \ge 0$, with probability at least $1-e^{-z}$,

$$\max_{\ell\in[N_\xi]} \mathrm{I}_3(\boldsymbol{\theta}_\ell) = \max_{\ell\in N_\xi}\sup_{S\in\mathcal{S}} \|(1-\mathbb{E})[\{\nabla^2 \widehat{R}_\varpi(\boldsymbol{\theta}_\ell)\}_{SS}]\| \lesssim C_* \upsilon_1^2\lambda_1\Bigg[\sqrt{\frac{s\log\{pr/(s\xi)\}+z}{n\varpi}} + \frac{s\log\{pr/(s\xi)\}+z}{n\varpi}\Bigg], \tag{S.131}$$

provided that $\xi \le r$.

Combining (S.124), (S.127), (S.128) and (S.131), with selecting $\xi \asymp r\varpi^2 n^{-2}$ such that $\xi \le r$, for any $z\ge 0$ and $n \gtrsim \upsilon_1^4(s\log p + z)$, it holds with probability at least $1-e^{-z}$ that

$$\begin{aligned}&\sup_{\boldsymbol{\beta}\in\mathbb{H}(2s)\cap\Theta(r),\, S\in\mathcal{S}} \|(1-\mathbb{E})[\{\nabla^2\widehat{R}_\varpi(\boldsymbol{\beta})\}_{SS}]\| \\ &\lesssim C_*\upsilon_1^2\lambda_1\Bigg\{\sqrt{\frac{s\log(n/\varpi)+s\log p+z}{n\varpi}} + \frac{s\log(n/\varpi)+s\log p + z}{n\varpi}\Bigg\} \\ &\quad + \big\{l_K \upsilon_1\lambda_1^{3/2}\sqrt{s\log p + z + \log n} + l_f\varpi^2\lambda_1^{3/2}\kappa_4^{3/4}\big\}\frac{r}{n^2}\,.\end{aligned}$$

We complete the proof of (S.123). □

EC.9.9. Proof of Lemma L17

Recall $\nabla\widehat{R}_\varpi(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n\{\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle - \varepsilon_i) - \tau\}\mathbf{x}_i$, where $\bar{K}_\varpi(u) = \bar{K}(u/\varpi)$ with $\bar{K}(u) = \int_{-\infty}^u K(v)\,\mathrm{d}v$. For any $\boldsymbol{\beta}\in\mathbb{H}(s)\cap\Theta(r)$ and $S\in\mathcal{S}$, it holds that

$$\begin{aligned}&\|\mathbb{E}[\{\nabla\widehat{R}_\varpi(\boldsymbol{\beta}) - \nabla\widehat{R}_\varpi(\boldsymbol{\beta}^*)\}_S]\|_2 \\ &= \bigg\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\{\bar{K}_\varpi(\langle\mathbf{x}_i,\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle - \varepsilon_i) - \bar{K}_\varpi(-\varepsilon_i)\}\mathbf{x}_{i,S}\big]\bigg\|_2 \\ &= \bigg\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\bigg[\bigg\{\int_{-\infty}^{\infty} f_{\varepsilon_i\,|\,\mathbf{x}_i}(v)\int_{-\frac{v}{\varpi}}^{\frac{\langle\mathbf{x}_i,\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle - v}{\varpi}} K(u)\,\mathrm{d}u\,\mathrm{d}v\bigg\}\mathbf{x}_{i,S}\bigg]\bigg\|_2 \\ &\overset{\text{(i)}}{=} \bigg\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\bigg[\bigg\{\int_{-\infty}^{\infty} K(u)\cdot\int_{-\varpi u}^{\langle\mathbf{x}_i,\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle - \varpi u} f_{\varepsilon_i\,|\,\mathbf{x}_i}(v)\mathrm{d}v\cdot\mathrm{d}u\bigg\}\mathbf{x}_{i,S}\bigg]\bigg\|_2 \\ &\overset{\text{(ii)}}{\le} f_u\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\sup_{\mathbf{u}\in\mathbb{S}^{p-1}}\frac{1}{n}\sum_{i=1}^n\mathbb{E}(\langle\mathbf{x}_i,\mathbf{u}\rangle^2) \le f_u\lambda_1 r\,,\end{aligned}$$

where (i) is from interchanging the order of integration, and (ii) follows from Assumptions 2 and 3. Hence, we conclude that

$$\sup_{\boldsymbol{\beta}\in\Theta(r)\cap\mathbb{H}(s),\, S\in\mathcal{S}} \|\mathbb{E}[\{\nabla\widehat{R}_\varpi(\boldsymbol{\beta}) - \nabla\widehat{R}_\varpi(\boldsymbol{\beta}^*)\}_S]\|_2 \le f_u\lambda_1 r\,. \tag{S.132}$$

Based on (S.132), to establish Lemma L17, it suffices to derive the an upper bound for

$$\sup_{\boldsymbol{\beta}\in\mathbb{H}(s)\cap\Theta(r),\, S\in\mathcal{S}} \|(1-\mathbb{E})\{\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta})-\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta}^*)\}_S\|_2 \,.$$

Let $\mathcal{S}'=\{S'\subseteq[p]:1\le|S'|\le s\}$. We first fix a subset $S'\in\mathcal{S}'$. Write $\tilde{S}=S'\cup\mathrm{supp}(\boldsymbol{\beta}^*)$. For any $\boldsymbol{\eta}\in\mathbb{R}^{|\tilde{S}|}$ and $S\in\mathcal{S}$, define

$$\mathcal{P}_{S'}(\boldsymbol{\eta};S)=(1-\mathbb{E})\bigg[\frac{1}{n}\sum_{i=1}^n\{\bar{K}_{\varpi}(\langle\mathbf{x}_{i,\tilde{S}},\boldsymbol{\eta}\rangle-\varepsilon_i)-\bar{K}_{\varpi}(-\varepsilon_i)\}\mathbf{x}_{i,S}\bigg]\,.$$

For any $\boldsymbol{\beta}\in\mathbb{H}(s)\cap\Theta(r)$ with $\mathrm{supp}(\boldsymbol{\beta})=S'$, write $\boldsymbol{\zeta}_{S'}(\boldsymbol{\beta})=(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\tilde{S}}$. Then $\langle\mathbf{x}_i,\boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle=\langle\mathbf{x}_{i,\tilde{S}},\boldsymbol{\zeta}_{S'}(\boldsymbol{\beta})\rangle$. Due to $\|\boldsymbol{\zeta}_{S'}(\boldsymbol{\beta})\|_2\le r$ and

$$(1-\mathbb{E})\{\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta})-\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta}^*)\}_S=\mathcal{P}_{S'}(\boldsymbol{\zeta}_{S'}(\boldsymbol{\beta});S)$$

for any $\boldsymbol{\beta}\in\mathbb{H}(s)\cap\Theta(r)$ with $\mathrm{supp}(\boldsymbol{\beta})=S'$, then

$$\sup_{\boldsymbol{\beta}\in\Theta(r):\,\mathrm{supp}(\boldsymbol{\beta})=S'}\big\|(1-\mathbb{E})\{\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta})-\nabla\widehat{R}_{\varpi}(\boldsymbol{\beta}^*)\}_S\big\|_2\le\sup_{\|\boldsymbol{\eta}\|_2\le r}\|\mathcal{P}_{S'}(\boldsymbol{\eta};S)\|_2\,. \tag{S.133}$$

Taking the gradient with respect to $\boldsymbol{\eta}$ yields

$$\nabla\mathcal{P}_{S'}(\boldsymbol{\eta};S)=\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\big(K_{i,\boldsymbol{\eta}}\mathbf{x}_{i,S}\mathbf{x}_{i,\tilde{S}}^\top\big)\,,$$

where $K_{i,\boldsymbol{\eta}}=K_{\varpi}(\langle\mathbf{x}_{i,\tilde{S}},\boldsymbol{\eta}\rangle-\varepsilon_i)$. By Assumptions 2 and 3, it holds that

$$\begin{aligned}
|\mathbb{E}(K_{i,\boldsymbol{\eta}}\,|\,\mathbf{x}_i)|&=\bigg|\frac{1}{\varpi}\int_{-\infty}^{\infty}K\bigg(\frac{\langle\mathbf{x}_{i,\tilde{S}},\boldsymbol{\eta}\rangle-v}{\varpi}\bigg)f_{\varepsilon_i\,|\,\mathbf{x}_i}(v)\,\mathrm{d}v\bigg|=\bigg|\int_{-\infty}^{\infty}K(v)f_{\varepsilon_i\,|\,\mathbf{x}_i}(\langle\mathbf{x}_{i,\tilde{S}},\boldsymbol{\eta}\rangle-\varpi v)\,\mathrm{d}v\bigg|\le f_u\,,\\
\mathbb{E}(K_{i,\boldsymbol{\eta}}^2\,|\,\mathbf{x}_i)&=\frac{1}{\varpi^2}\int_{-\infty}^{\infty}K^2\bigg(\frac{\langle\mathbf{x}_{i,\tilde{S}},\boldsymbol{\eta}\rangle-v}{\varpi}\bigg)f_{\varepsilon_i\,|\,\mathbf{x}_i}(v)\,\mathrm{d}v\\
&=\frac{1}{\varpi}\int_{-\infty}^{\infty}K^2(v)f_{\varepsilon_i\,|\,\mathbf{x}_i}(\langle\mathbf{x}_{i,\tilde{S}},\boldsymbol{\eta}\rangle-\varpi v)\,\mathrm{d}v\le\frac{k_{\max}f_u}{\varpi}\,.
\end{aligned}$$

By Assumption 1, we have for any $\mathbf{u}\in\mathbb{S}^{|S|-1}$, $\mathbf{v}\in\mathbb{S}^{|\tilde{S}|-1}$ and $z\ge0$,

$$\mathbb{P}(|\langle\mathbf{x}_{i,S},\mathbf{u}\rangle\langle\mathbf{x}_{i,\tilde{S}},\mathbf{v}\rangle|>z)\le\mathbb{P}(|\langle\mathbf{x}_{i,S},\mathbf{u}\rangle|>z^{1/2})+\mathbb{P}(|\langle\mathbf{x}_{i,\tilde{S}},\mathbf{v}\rangle|>z^{1/2})\le4e^{-z/(2\upsilon_1^2\lambda_1)}\,.$$

Let $w_{i,\mathbf{u},\mathbf{v}}=\langle\mathbf{x}_{i,S},\mathbf{u}\rangle\langle\mathbf{x}_{i,\tilde{S}},\mathbf{v}\rangle$. For any integer $q\ge1$, we have

$$\mathbb{E}(|w_{i,\mathbf{u},\mathbf{v}}|^q)=\int_0^{\infty}\mathbb{P}(|w_{i,\mathbf{u},\mathbf{v}}|>v^{1/q})\,\mathrm{d}v\le4q\int_0^{\infty}v^{q-1}e^{-v/(2\upsilon_1^2\lambda_1)}\,\mathrm{d}v=4(2\upsilon_1^2\lambda_1)^qq!\,. \tag{S.134}$$

Then, for any fixed $\mathbf{u}\in\mathbb{S}^{|S|-1}$, $\mathbf{v}\in\mathbb{S}^{|\tilde{S}|-1}$ and $\lambda\in\mathbb{R}$, it holds that

$$\begin{aligned}
\mathbb{E}\big[\exp\{\lambda n^{1/2}\langle\mathbf{u},\nabla\mathcal{P}_{S'}(\boldsymbol{\eta};S)\mathbf{v}\rangle\}\big] &= \prod_{i=1}^n\mathbb{E}\bigg[\exp\bigg\{\frac{\lambda}{\sqrt{n}}(1-\mathbb{E})(K_{i,\boldsymbol{\eta}}w_{i,\mathbf{u},\mathbf{v}})\bigg\}\bigg]\\
&\overset{\text{(i)}}{\le}\prod_{i=1}^n\bigg(1+\frac{\lambda^2}{2n}\exp\bigg(\frac{8\upsilon_1^2\lambda_1 f_u|\lambda|}{\sqrt{n}}\bigg)\mathbb{E}\bigg[\{(1-\mathbb{E})(K_{i,\boldsymbol{\eta}}w_{i,\mathbf{u},\mathbf{v}})\}^2\exp\bigg(\frac{k_{\max}|\lambda|}{\varpi\sqrt{n}}|w_{i,\mathbf{u},\mathbf{v}}|\bigg)\bigg]\bigg)\\
&\overset{\text{(ii)}}{\le}\prod_{i=1}^n\bigg[1+\frac{k_{\max}f_u\lambda^2}{n\varpi}\exp\bigg(\frac{8\upsilon_1^2\lambda_1 f_u|\lambda|}{\sqrt{n}}\bigg)\mathbb{E}\bigg\{|w_{i,\mathbf{u},\mathbf{v}}|^2\exp\bigg(\frac{k_{\max}|\lambda|}{\varpi\sqrt{n}}|w_{i,\mathbf{u},\mathbf{v}}|\bigg)\bigg\}\\
&\qquad+\frac{32\upsilon_1^4\lambda_1^2k_{\max}f_u\lambda^2}{n\varpi}\exp\bigg(\frac{8\upsilon_1^2\lambda_1 f_u|\lambda|}{\sqrt{n}}\bigg)\mathbb{E}\bigg\{\exp\bigg(\frac{k_{\max}|\lambda|}{\varpi\sqrt{n}}|w_{i,\mathbf{u},\mathbf{v}}|\bigg)\bigg\}\bigg]\,,
\end{aligned}$$

where (i) follows from the fact $e^u\le 1+u+u^2e^{|u|}/2$ for any $u\in\mathbb{R}$, and $\mathbb{E}(|w_{i,\mathbf{u},\mathbf{v}}|)\le 8\upsilon_1^2\lambda_1$, and (ii) follows from the facts $(a+b)^2\le 2a^2+2b^2$ for any $a,b\in\mathbb{R}$, $\mathbb{E}(K_{i,\boldsymbol{\eta}}^2\,|\,\mathbf{x}_i)\le k_{\max}f_u/\varpi$, and $\mathbb{E}(|w_{i,\mathbf{u},\mathbf{v}}|^2)\le 32\upsilon_1^4\lambda_1^2$. For any given $c>0$, by (S.134),

$$\mathbb{E}\{\exp(c|w_{i,\mathbf{u},\mathbf{v}}|)\}=\mathbb{E}\bigg\{1+\sum_{q=1}^\infty\frac{c^q|w_{i,\mathbf{u},\mathbf{v}}|^q}{q!}\bigg\}\le 1+\sum_{q=1}^\infty(8c\upsilon_1^2\lambda_1)^q=\frac{1}{1-8c\upsilon_1^2\lambda_1}\le e^{16c\upsilon_1^2\lambda_1}\,,$$

provided that $16c\upsilon_1^2\lambda_1\le 1$, where the last inequality follows from the fact $(1-x)^{-1}\le e^{2x}$ for any $x\in[0,1/2]$. Thus, we can conclude that

$$\mathbb{E}\bigg\{\exp\bigg(\frac{k_{\max}|\lambda|}{\varpi\sqrt{n}}|w_{i,\mathbf{u},\mathbf{v}}|\bigg)\bigg\}\le e$$

for $|\lambda|\le\varpi n^{1/2}(16k_{\max}\upsilon_1^2\lambda_1)^{-1}$. Analogously, it also holds that

$$\begin{aligned}
\mathbb{E}\bigg\{|w_{i,\mathbf{u},\mathbf{v}}|^2\exp\bigg(\frac{k_{\max}|\lambda|}{\varpi\sqrt{n}}|w_{i,\mathbf{u},\mathbf{v}}|\bigg)\bigg\}&\le\{\mathbb{E}(|w_{i,\mathbf{u},\mathbf{v}}|^4)\}^{1/2}\cdot\bigg[\mathbb{E}\bigg\{\exp\bigg(\frac{2k_{\max}|\lambda|}{\varpi\sqrt{n}}|w_{i,\mathbf{u},\mathbf{v}}|\bigg)\bigg\}\bigg]^{1/2}\\
&\le 48e\upsilon_1^4\lambda_1^2\,,
\end{aligned}$$

provided that $|\lambda|\le\varpi n^{1/2}(32k_{\max}\upsilon_1^2\lambda_1)^{-1}$, which further implies

$$\begin{aligned}
\mathbb{E}\big[\exp\{\lambda n^{1/2}\langle\mathbf{u},\nabla\mathcal{P}_{S'}(\boldsymbol{\eta};S)\mathbf{v}\rangle\}\big]&\le\prod_{i=1}^n\bigg\{1+\frac{80e\upsilon_1^4\lambda_1^2k_{\max}f_u\lambda^2}{n\varpi}\exp\bigg(\frac{8\upsilon_1^2\lambda_1 f_u|\lambda|}{\sqrt{n}}\bigg)\bigg\}\\
&\le\prod_{i=1}^n\bigg\{1+(C_1\upsilon_1^2\lambda_1)^2\frac{\lambda^2}{2n\varpi}\bigg\}\le\exp\bigg\{\frac{(C_1\upsilon_1^2\lambda_1)^2\lambda^2}{2\varpi}\bigg\}
\end{aligned}$$

for some constant $C_1>0$ depending only on $(f_u,k_{\max})$, provided that $\lambda^2\le\varpi^2n(32k_{\max}\upsilon_1^2\lambda_1)^{-2}$. Due to $|\tilde{S}|=|S'\cup\mathrm{supp}(\boldsymbol{\beta}^*)|\le 2s$, for any $z\ge 1/2$, by applying Theorem A.3 of Spokoiny (2013) with the parameters (ν_0,g) chosen as $\nu_0=C_1\upsilon_1^2\lambda_1\varpi^{-1/2}$ and $g^2=\varpi^2n(32\sqrt{2}k_{\max}\upsilon_1^2\lambda_1)^{-2}$, we obtain

$$\sup_{\|\boldsymbol{\eta}\|_2 \le r} \|\mathcal{P}_{S'}(\boldsymbol{\eta}; S)\|_2 \le 6C_1 \upsilon_1^2 \lambda_1 r \sqrt{\frac{8s+2z}{n\varpi}} \tag{S.135}$$

holds with probability at least $1-e^{-z}$, provided that $n\varpi^2 \ge (32\sqrt{2}k_{\max}\upsilon_1^2\lambda_1)^2(8s+2z)$. Notice that $|\mathcal{S}'| \le (ep/s)^s$. By (S.133) and (S.135), and taking union bound over all $S' \in \mathcal{S}'$, we have for any $S \in \mathcal{S}$ and $z \ge 1/2$,

$$\sup_{\boldsymbol{\beta} \in \mathbb{H}(s) \cap \Theta(r)} \left\|(1-\mathbb{E})\{\nabla \widehat{R}_\varpi(\boldsymbol{\beta}) - \nabla \widehat{R}_\varpi(\boldsymbol{\beta}^*)\}_S\right\|_2 \le C_* \upsilon_1^2 \lambda_1 r \sqrt{\frac{s\log(p/s)+z}{n\varpi}} \tag{S.136}$$

holds with probability at least $1-e^{-z}$, provided that $n\varpi^2 \gtrsim k_{\max}^2 \upsilon_1^4 \lambda_1^2 (s+z)$, where $C_* > 0$ is a constant depending only on $(f_u, k_{\max})$.

Recall $|\mathcal{S}| \le \{ep/(2s)\}^{2s}$. By (S.136), taking a union bound over all $S \in \mathcal{S}$ yields for any $z \ge 1/2$,

$$\sup_{\boldsymbol{\beta} \in \mathbb{H}(s) \cap \Theta(r),\, S \in \mathcal{S}} \left\|(1-\mathbb{E})\{\nabla \widehat{R}_\varpi(\boldsymbol{\beta}) - \nabla \widehat{R}_\varpi(\boldsymbol{\beta}^*)\}_S\right\|_2 \le C_{**} \upsilon_1^2 \lambda_1 r \sqrt{\frac{s\log(p/s)+z}{n\varpi}}$$

holds with probability at least $1-e^{-z}$, provided that $n\varpi^2 \gtrsim k_{\max}^2 \upsilon_1^4 \lambda_1^2 (s+z)$, where $C_{**} > 0$ is a constant depending only on $(f_u, k_{\max})$. Combining with (S.132), for any $z \ge 1/2$, we have

$$\sup_{\boldsymbol{\beta} \in \Theta(r) \cap \mathbb{H}(s),\, S \in \mathcal{S}} \left\|\{\nabla \widehat{R}_\varpi(\boldsymbol{\beta}) - \nabla \widehat{R}_\varpi(\boldsymbol{\beta}^*)\}_S\right\|_2 \le \frac{3}{2} f_u \lambda_1 r$$

holds with probability at least $1-e^{-z}$, provided that $n\varpi \ge 4C_{**}^2 f_u^{-2} \upsilon_1^4 (s\log p + z)$ and $n\varpi^2 \ge C_{***} \upsilon_1^4 \lambda_1^2 (s+z)$, where $C_{**} > 0$ and $C_{***} > 0$ are two constants depending only on $(f_u, k_{\max})$. Hence, we complete the proof of Lemma L17. □

References

Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016) Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318.

Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Operations Research* 67(1):90–108.

Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, Ramage D, Segal A, Seth K (2017) Practical secure aggregation for privacy-preserving machine learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175–1191.

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

Bu Z, Dong J, Long Q, Su W (2020) Deep learning with gaussian differential privacy. *Harvard Data Science Review* 2(3).

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.

Cannon AJ (2011) Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Computers & Geosciences* 37(9):1277–1284.

Cannon AJ (2018) Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environmental Research and Risk Assessment* 32:3207–3225.

Cybenko G (1989) Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* 2:303–314.

Duchi JC, Jordan MI, Wainwright MJ (2013) Local privacy and statistical minimax rates. *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 429–438.

Duchi JC, Jordan MI, Wainwright MJ (2018) Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association* 113(521):182–201.

Dwork C, Su W, Zhang L (2021) Differentially private false discovery rate control. *Journal of Privacy and Confidentiality* 11(2):1–44.

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. *Neural Networks* 2(5):359–366.

Jain P, Tewari A, Kar P (2014) On iterative hard thresholding methods for high-dimensional M-estimation. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, volume 1, 685–693.

Kasiviswanathan SP, Lee HK, Nissim K, Raskhodnikova S, Smith A (2011) What can we learn privately? *SIAM Journal on Computing* 40(3):793–826.

Ledoux M, Talagrand M (2013) *Probability in Banach Spaces: Isoperimetry and Processes* (Springer Berlin, Heidelberg).

Lowy A, Razaviyayn M (2023) Private federated learning without a trusted server: Optimal algorithms for convex losses. *Proceedings of the 11th International Conference on Learning Representations*.

Meinshausen N (2006) Quantile regression forests. *Journal of Machine Learning Research* 7:983–999.

Mhaskar HN (1996) Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation* 8(1):164–177.

Park J, Kim H, Choi Y, Lee J (2023) Differentially private sharpness-aware training. *Proceedings of the 40th International Conference on Machine Learning*, volume 202, 27204–27224.

Plan Y, Vershynin R (2013) One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics* 66(8):1275–1297.

Spokoiny V (2013) Bernstein-von Mises theorem for growing parameter dimension. ArXiv:1302.3430.

Vershynin R (2012) *Introduction to the Non-asymptotic Analysis of Random Matrices* (Cambridge University Press).

Vershynin R (2018) *High-dimensional Probability: An Introduction with Applications in Data Science* (Cambridge University Press).

Wainwright MJ (2019) *High-dimensional Statistics: A Non-asymptotic Viewpoint* (Cambridge University Press).

Wang D, Xu J (2021) On sparse linear regression in the local differential privacy model. *IEEE Transactions on Information Theory* 67(2):1182–1200.