TRANSFER LEARNING FOR HIGH-DIMENSIONAL QUANTILE REGRESSION VIA CONVOLUTION SMOOTHING

Yijiao Zhang and Zhongyi Zhu

Department of Statistics and Data Science, Fudan University

Supplementary Materials

Section S1 includes auxiliary lemmas and the proofs of the theories in the main text. Section S2 includes additional simulation results mentioned in the main text. A distributed version of the Oracle-Trans-SQR algorithm as well as its theoretical properties is provided in Section S3. The proofs of theories in Section S3 are placed in Section S4. We put the proofs of the auxiliary lemmas in Section S5.

S1 Proof of Main Results

We first introduce some notations, which will be repeatedly used in the sequel. Denote the empirical smoothed quantile loss on the pooled data by

$$\widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}) = \frac{1}{(n_{\mathcal{A}_\eta}+n_0)h_{\boldsymbol{w}}} \sum_{k\in\mathcal{A}_\eta\cup\{0\}} \sum_{i=1}^{n_k} \int_{-\infty}^{\infty} \rho_\tau(u) K\left(\frac{u+\boldsymbol{w}^\top \boldsymbol{x}_i^{(k)} - y_i^{(k)}}{h_{\boldsymbol{w}}}\right) \mathrm{d}u.$$

Further define the empirical smoothed quantile losses on the target data

and the pooled target data with the kth source data respectively as

$$\widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{w}) = 1/(n_0 h_{\boldsymbol{\delta}}) \sum_{i=1}^{n_0} \int_{-\infty}^{\infty} \rho_\tau(u) K((u + \boldsymbol{w}^\top \boldsymbol{x}_i^{(0)} - y_i^{(0)})/h_{\boldsymbol{\delta}}) \mathrm{d}u,$$

and

$$\widehat{Q}^{(k)}_{h^{(k)}}(\boldsymbol{w}) = \frac{1}{(n_k + n_0)h^{(k)}} \sum_{k' \in \{0,k\}} \sum_{i=1}^{n_{k'}} \int_{-\infty}^{\infty} \rho_\tau(u) K\left(\frac{u + \boldsymbol{w}^\top \boldsymbol{x}_i^{(k')} - y_i^{(k')}}{h^{(k')}}\right) \mathrm{d}u.$$

Define the integrated kernel function $\bar{K} : \mathbb{R} \to [0,1]$ as $\bar{K}(u) = \int_{-\infty}^{u} K(t)\mathrm{d}t$. The gradient of $\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w})$, $\widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{w})$, and $\widehat{Q}^{(k)}_{h^{(k)}}(\boldsymbol{w})$ are given respectively by

$$\nabla \widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}) = 1/(n_{\mathcal{A}_\eta} + n_0) \sum_{k \in \mathcal{A}_\eta \cup \{0\}} \sum_{i=1}^{n_k} \{\bar{K}(((\boldsymbol{x}_i^{(k)})^\top \boldsymbol{w} - y_i^{(k)})/h_{\boldsymbol{w}}) - \tau\} \boldsymbol{x}_i^{(k)},$$

$$\nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{w}) = 1/n_0 \sum_{i=1}^{n_0} \{\bar{K}(((\boldsymbol{x}_i^{(0)})^\top \boldsymbol{w} - y_i^{(0)})/h_{\boldsymbol{w}}) - \tau\} \boldsymbol{x}_i^{(0)},$$

and

$$\nabla \widehat{Q}^{(k)}_{h^{(k)}}(\boldsymbol{w}) = 1/(n_k + n_0) \sum_{k' \in \{0,k\}} \sum_{i=1}^{n_{k'}} \{\bar{K}(((\boldsymbol{x}_i^{(k')})^\top \boldsymbol{w} - y_i^{(k')})/h^{(k')}) - \tau\} \boldsymbol{x}_i^{(k')}.$$

Further define $Q^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}) = \mathbb{E}[\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w})]$, $\widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{w}) = \mathbb{E}[\widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{w})]$ and $Q^{(k)}_{h^{(k)}}(\boldsymbol{w}) = \mathbb{E}[\widehat{Q}^{(k)}_{h^{(k)}}(\boldsymbol{w})]$. The gradients of $Q^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w})$, $Q^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{w})$, and $Q^{(k)}_{h^{(k)}}(\boldsymbol{w})$ are given by $\nabla Q^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}) = \mathbb{E}[\nabla \widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w})]$, $\nabla Q^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{w}) = \mathbb{E}[\nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{w})]$ and $\nabla Q^{(k)}_{h^{(k)}}(\boldsymbol{w}) = \mathbb{E}[\nabla \widehat{Q}^{(k)}_{h^{(k)}}(\boldsymbol{w})]$ respectively. In addition, define the ℓ_1-ball, ℓ_2-ball and ℓ_1-cone as $\mathbb{B}_1(r) = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}\|_1 \leq r\}$, $\mathbb{B}_2(r) = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}\|_2 \leq r\}$ and $\mathbb{C}(l) = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\delta\|_1 \leq l\|\delta\|_2\}$ respectively.

S1.1 Auxiliary Lemmas

To facilitate the proof of the main results, some technical lemmas are given here, of which the proofs are given in Section S5.

Lemma S1.1. *Assume Condition 4 holds. Then we have*

$$\left\|\boldsymbol{\delta}^{\mathcal{A}_\eta}\right\|_1 = \left\|\boldsymbol{\beta} - \boldsymbol{w}^{\mathcal{A}_\eta}\right\|_1 \leq c_M \eta.$$

Lemma S1.2. *Under Conditions 1-3, we have that*

$$\|\nabla Q^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta})\|_2 \leq l_0\mu_1\kappa_2 h^2_{\boldsymbol{w}}/2, \quad \|\nabla Q^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)})\|_2 \leq l_0\mu_1\kappa_2 (h^{(k)})^2/2,$$

and $\|\nabla Q^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta})\|_2 \leq l_0\mu_1\kappa_2 h^2_{\boldsymbol{\delta}}/2.$

Lemma S1.3. *Under Conditions 2-3, there exists a positive constan C such that*

$$\mathbb{P}\left(\|\nabla \widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta}) - \nabla Q^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta})\|_\infty \leq 2\sqrt{\frac{C\log p}{n_{\mathcal{A}_\eta}+n_0}}\right) \geq 1-4p^{-1}-(n_{\mathcal{A}}+n_0)^{-1},$$

$$\mathbb{P}\left(\|\nabla \widehat{Q}^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)}) - \nabla Q^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)})\|_\infty \leq 2\sqrt{\frac{C\log p}{n_k+n_0}}\right) \geq 1-4p^{-1},$$

and

$$\mathbb{P}\left(\|\nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta}) - \nabla Q^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta})\|_\infty \leq 2\sqrt{\frac{c_x\log p}{n_0}}\right) \geq 1-4p^{-1}.$$

The following two lemmas provide a core result for establishing error bounds for our smoothed two-step QR estimators. They are related to the local RSC property of the empirical smoothed quantile loss, which may be

of independent interest. The RSC property plays a critical role in the theoretical analysis of regularized M-estimation in high dimensions (Negahban et al., 2012; Loh and Wainwright, 2013; Loh, 2017) as well as in recent literature on high-dimensional transfer learning (Li et al., 2022; Tian and Feng, 2022).

Lemma S1.4. *Assume Conditions 1-3 and 5 hold. Suppose* $n_{\mathcal{A}_\eta} \gtrsim h_{\boldsymbol{w}} r^{-2} \log p$ *and* $32v_1^2 r \le h_{\boldsymbol{w}} \le f_l/(2l_0)$. *Then there exist positive constants* a_1 *and* a_2 *depending only on* $(\kappa_l, f_l, f_u, \gamma_p, v_1)$, *such that for all* $\boldsymbol{\Delta} \in \mathbb{B}_2(r)$, *we have*

$$\langle \nabla \widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta}+\boldsymbol{\Delta}) - \nabla \widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta}), \boldsymbol{\Delta} \rangle \ge a_1(\|\boldsymbol{\Delta}\|_2^2 - a_2\sqrt{\frac{h_{\boldsymbol{w}} \log p}{(n_{\mathcal{A}_\eta}+n_0)r^2}}\|\boldsymbol{\Delta}\|_1\|\boldsymbol{\Delta}\|_2), \tag{S1.1}$$

with probability as least $1 - c_1\exp(-c_2 \log p)$ *for some* c_1 *and* $c_2 > 0$.

Lemma S1.5. *Assume Conditions 1-3 hold. Suppose* $n_0 \gtrsim h_{\boldsymbol{\delta}} r^{-2} \log p$, $32v_1^2 r \le h_{\boldsymbol{\delta}} \le f_l/(2l_0)$. *Then there exist positive constants* a_1 *and* a_2 *depending only on* $(\kappa_l, f_l, f_u, \gamma_p, v_1)$, *such that for all* $\boldsymbol{\Delta} \in \mathbb{B}_2(r)$, *we have*

$$\langle \nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta}+\boldsymbol{\Delta}) - \nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta}), \boldsymbol{\Delta} \rangle \ge a_1(\|\boldsymbol{\Delta}\|_2^2 - a_2\sqrt{\frac{h_{\boldsymbol{\delta}} \log p}{n_0 r^2}}\|\boldsymbol{\Delta}\|_1\|\boldsymbol{\Delta}\|_2), \tag{S1.2}$$

with probability as least $1 - c_1\exp(-c_2 \log p)$ *for some* c_1 *and* $c_2 > 0$.

Corollary S1.1. *Assume Conditions 1-3 and 5 hold. Suppose* $n_k \gtrsim h^{(k)}/r^2 \log p$ *and* $32v_1^2 r \le h^{(k)} \le f_l/(2l_0)$. *Then there exist positive constants* a_1 *and* a_2

depending only on $(\kappa_l, f_l, f_u, \gamma_p, v_1)$*, such that for all* $\boldsymbol{\Delta} \in \mathbb{B}_2(r)$*, we have*

$$\langle \nabla \widehat{Q}^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)}+\boldsymbol{\Delta}) - \nabla \widehat{Q}^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)}), \boldsymbol{\Delta} \rangle \geq a_1(\|\boldsymbol{\Delta}\|_2^2 - a_2\sqrt{\frac{h^{(k)}\log p}{(n_k+n_0)r^2}}\|\boldsymbol{\Delta}\|_1\|\boldsymbol{\Delta}\|_2), \tag{S1.3}$$

with probability as least $1 - c_1\exp(-c_2 \log p)$ *for some* c_1 *and* $c_2 > 0$*.*

Remark 1. *Lemma S1.4 and Lemma S1.5 characterize the curvature of the smoothed quantile loss* $\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w})$ *and* $\widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{w})$ *respectively.*

Remark 2. *These two lemmas can be seen as a refinement of the local RSC properties investigated in Tan et al. (2022), where they restricted* $\boldsymbol{\Delta} \in \mathbb{B}_2(r) \cap \mathbb{C}_l$ *for some* l¿0*. By employing a peeling technique (van der Vaart and Wellner, 1996; Van de Geer, 2000), we remove this constraint to make the local RSC properties hold uniformly in the ratio* l*. As we can see from the proof of Theorem 1, the establishment of our estimation error bounds depends crucially on this uniformity.*

S1.2 Proof of Theorem 1

Let $\hat{\boldsymbol{u}} = \hat{\boldsymbol{w}}^{\mathcal{A}_\eta} - \boldsymbol{w}^{\mathcal{A}_\eta}$ and $\hat{\boldsymbol{v}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ denote the estimation bias of the transferring step and final estimator respectively. We first derive the estimation bound for $\hat{\boldsymbol{u}}$ on which we base to derive the bound for $\hat{\boldsymbol{v}}$.

Proof. **Step 1: Bounds for** $\hat{\boldsymbol{u}}$**.**

We establish bound for $\hat{\boldsymbol{u}}$ by providing upper and lower bounds for the symmetric Bregman divergence $\langle \nabla \widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\hat{\boldsymbol{w}}^{\mathcal{A}_\eta}) - \nabla \widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta}), \hat{\boldsymbol{u}} \rangle$.

Upped bound: Firstly, by the optimality of $\hat{\boldsymbol{w}}^{\mathcal{A}_\eta}$, we have

$$\nabla \widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\hat{\boldsymbol{w}}^{\mathcal{A}_\eta}) + \lambda_{\boldsymbol{w}} \mathrm{sgn}(\hat{\boldsymbol{w}}^{\mathcal{A}_\eta}) = 0. \tag{S1.4}$$

The convexity of $\widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w})$, together with the optimal condition in (S1.4) implies that

$$\begin{aligned}
0 \leq & \langle \nabla \widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\hat{\boldsymbol{w}}^{\mathcal{A}_\eta}) - \nabla \widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta}), \hat{\boldsymbol{u}} \rangle \\
= & \langle -\lambda_{\boldsymbol{w}} \mathrm{sgn}(\hat{\boldsymbol{w}}^{\mathcal{A}_\eta}), \hat{\boldsymbol{u}} \rangle - \langle \nabla \widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta}) - \nabla Q_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta}), \hat{\boldsymbol{u}} \rangle - \langle \nabla Q_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta}), \hat{\boldsymbol{u}} \rangle \\
\leq & \langle -\lambda_{\boldsymbol{w}} \mathrm{sgn}(\hat{\boldsymbol{w}}^{\mathcal{A}_\eta}), \hat{\boldsymbol{u}} \rangle + \| \nabla \widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta}) - \nabla Q_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta}) \|_\infty \| \hat{\boldsymbol{u}} \|_1 \\
& + \| \nabla Q_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta}) \|_2 \| \hat{\boldsymbol{u}} \|_2,
\end{aligned} \tag{S1.5}$$

where we used Hölder's inequality in the last inequality.

By the convexity of $\| \cdot \|_1$, we have

$$\langle -\lambda_{\boldsymbol{w}} \mathrm{sgn}(\hat{\boldsymbol{w}}^{\mathcal{A}_\eta}), \hat{\boldsymbol{u}} \rangle \leq \lambda_{\boldsymbol{w}} \| \boldsymbol{w}^{\mathcal{A}_\eta} \|_1 - \lambda_{\boldsymbol{w}} \| \hat{\boldsymbol{w}}^{\mathcal{A}_\eta} \|_1. \tag{S1.6}$$

Define the deterministic quantity $q_{\mathcal{A}_\eta} = \| \nabla Q_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta}) \|_2$. Plugging (S1.6) into (S1.5) and conditioning on the event $\mathcal{E}_{\boldsymbol{w}} = \{ \lambda_{\boldsymbol{w}} \geq 2 \| \nabla \widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta}) -$

$\nabla Q_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta})\|_\infty\}$, we have

$$\begin{aligned}
0 &\leq \lambda_{\boldsymbol{w}}\|\boldsymbol{w}^{\mathcal{A}_\eta}\|_1 - \lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{w}}^{\mathcal{A}_\eta}\|_1 + \frac{1}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}\|_1 + q_{\mathcal{A}_\eta}\|\hat{\boldsymbol{u}}\|_2 \\
&\leq \lambda_{\boldsymbol{w}}\|\boldsymbol{w}_{\mathcal{S}^c}^{\mathcal{A}_\eta}\|_1 - \lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{w}}_{\mathcal{S}^c}^{\mathcal{A}_\eta}\|_1 + \frac{3}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_{\mathcal{S}}\|_1 + \frac{1}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_{\mathcal{S}^c}\|_1 + q_{\mathcal{A}_\eta}\|\hat{\boldsymbol{u}}\|_2 \\
&\leq 2\lambda_{\boldsymbol{w}}\|\boldsymbol{w}_{\mathcal{S}^c}^{\mathcal{A}_\eta}\|_1 + \frac{3}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_{\mathcal{S}}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_{\mathcal{S}^c}\|_1 + q_{\mathcal{A}_\eta}\|\hat{\boldsymbol{u}}\|_2 \\
&\leq 2\lambda_{\boldsymbol{w}}c_M\eta + \frac{3}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_{\mathcal{S}}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_{\mathcal{S}^c}\|_1 + q_{\mathcal{A}_\eta}\|\hat{\boldsymbol{u}}\|_2,
\end{aligned} \tag{S1.7}$$

where we used the triangle inequality in the third and the fourth inequalities and the last inequality follows from Lemma S1.1.

We can derive from (S1.7) that $\hat{\boldsymbol{u}}$ satisfies the cone-like constraint $\|\hat{\boldsymbol{u}}_{\mathcal{S}^c}\|_1 \leq 3\|\hat{\boldsymbol{u}}_{\mathcal{S}}\|_1 + 2\lambda_{\boldsymbol{w}}^{-1}q_{\mathcal{A}_\eta}\|\hat{\boldsymbol{u}}\|_2 + 4c_M\eta$, from which it follows that

$$\|\hat{\boldsymbol{u}}\|_1 \leq 4\|\hat{\boldsymbol{u}}_{\mathcal{S}}\|_1 + 2\lambda_{\boldsymbol{w}}^{-1}q_{\mathcal{A}_\eta}\|\hat{\boldsymbol{u}}\|_2 + 4\lambda_{\boldsymbol{w}}c_M\eta \leq (4\sqrt{s} + 2\lambda_{\boldsymbol{w}}^{-1}q_{\mathcal{A}_\eta})\|\hat{\boldsymbol{u}}\|_2 + 4c_M\eta.$$

By Lemma S1.2, we have $q_{\mathcal{A}_\eta} \leq l_0\mu_1\kappa_2 h_{\boldsymbol{w}}^2/2$. Let $h_{\boldsymbol{w}}^2 \leq \lambda_{\boldsymbol{w}}\sqrt{s}/(l_0\mu_1\kappa_2)$, then conditioning on $\boldsymbol{\mathcal{E}}_{\boldsymbol{w}}$, $\hat{\boldsymbol{u}}$ falls into the set $\mathbb{U}$ defined as

$$\mathbb{U} = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_1 \leq 5\sqrt{s}\|\boldsymbol{u}\|_2 + 4c_M\eta\}. \tag{S1.8}$$

Lower bound: Define $D(\boldsymbol{\Delta}) = \langle \nabla\widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta} + \boldsymbol{\Delta}) - \nabla\widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta}), \boldsymbol{\Delta}\rangle$. To derive a lower bound, we consider the event $\boldsymbol{\mathcal{E}}'_{\boldsymbol{w}}(r)$ defined as

$$\left\{ \frac{D(\boldsymbol{\Delta})}{\|\boldsymbol{\Delta}\|_2^2} \geq a_1\left(1 - a_2\sqrt{\frac{h_{\boldsymbol{w}}\log p}{(n_{\mathcal{A}_\eta} + n_0)r^2}}\frac{\|\boldsymbol{\Delta}\|_1}{\|\boldsymbol{\Delta}\|_2}\right), \quad \text{for all} \quad \boldsymbol{\Delta} \in \mathbb{B}_2(r) \right\}.$$

We set $r = h_{\boldsymbol{w}}/c_0$ with $c_0 = 32v_1^2$. We use proof by contradiction to show

that conditioning on $\mathcal{E}_{\boldsymbol{w}} \cap \mathcal{E}'_{\boldsymbol{w}}(r)$,

$$\|\hat{\boldsymbol{u}}\|_2 \leq 8a_2c_0c_M\sqrt{\frac{\log p}{(n_{\mathcal{A}_\eta}+n_0)h_{\boldsymbol{w}}}}\eta + 4\lambda_{\boldsymbol{w}}\sqrt{s} + 2\sqrt{\frac{\lambda_{\boldsymbol{w}}c_M\eta}{a_1}} =: c_u. \quad \text{(S1.9)}$$

In order to make use of the result on $\mathcal{E}'_{\boldsymbol{w}}(r)$, we let

$$h_{\boldsymbol{w}} \gtrsim (\log p/(n_{\mathcal{A}_\eta}+n_0))^{1/4}\sqrt{\eta} + \lambda_{\boldsymbol{w}}\sqrt{s} + \sqrt{\log p/((n_{\mathcal{A}_\eta}+n_0)h_{\boldsymbol{w}})\eta} \quad \text{(S1.10)}$$

such that $r > c_u$. Consider $\tilde{\boldsymbol{u}} = t\hat{\boldsymbol{u}}$ for some $t \in (0,1)$. Choose t such that $\|\tilde{\boldsymbol{u}}\|_2 \leq r$ and $\|\tilde{\boldsymbol{u}}\|_2 \geq c_u$. We can verify that $\tilde{\boldsymbol{u}} \in \mathbb{U}$. Denote $\tilde{\boldsymbol{w}}^{\mathcal{A}_\eta} = \boldsymbol{w}^{\mathcal{A}_\eta} + \tilde{\boldsymbol{u}}$. The optimality of $\hat{\boldsymbol{u}}$ implies that $G(\tilde{\boldsymbol{u}}) \geq G(\hat{\boldsymbol{u}})$ with $G(\boldsymbol{u}) = \widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta}+\boldsymbol{u}) + \lambda_{\boldsymbol{w}}\|\boldsymbol{w}^{\mathcal{A}_\eta}+\boldsymbol{u}\|_1$. This together with the convexity of the function $G(\boldsymbol{u})$ leads to

$$\langle \nabla\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\tilde{\boldsymbol{w}}^{\mathcal{A}_\eta}) + \lambda_{\boldsymbol{w}}\mathrm{sgn}(\tilde{\boldsymbol{w}}^{\mathcal{A}_\eta}), \tilde{\boldsymbol{u}}\rangle = \frac{t}{1-t}\langle \nabla\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\tilde{\boldsymbol{w}}^{\mathcal{A}_\eta}) + \lambda_{\boldsymbol{w}}\mathrm{sgn}(\tilde{\boldsymbol{w}}^{\mathcal{A}_\eta}), \hat{\boldsymbol{u}} - \tilde{\boldsymbol{u}}\rangle \leq 0.$$

Therefore, we have

$$\langle \nabla\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\tilde{\boldsymbol{w}}^{\mathcal{A}_\eta}) - \nabla\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta}), \hat{\boldsymbol{u}}\rangle \leq \langle -\lambda_{\boldsymbol{w}}\mathrm{sgn}(\tilde{\boldsymbol{w}}^{\mathcal{A}_\eta}) - \nabla\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta}), \hat{\boldsymbol{u}}\rangle$$

Then conditioning on $\mathcal{E}_{\boldsymbol{w}} \cap \mathcal{E}'_{\boldsymbol{w}}(r)$ and by the same argument in (S1.7), it can be shown that

$$a_1\|\tilde{\boldsymbol{u}}\|_1^2 - a_1a_2c_0\sqrt{\frac{\log p}{(n_{\mathcal{A}_\eta}+n_0)h_{\boldsymbol{w}}}}\|\tilde{\boldsymbol{u}}\|_1\|\tilde{\boldsymbol{u}}\|_2 \leq 2\lambda_{\boldsymbol{w}}\sqrt{s}\|\tilde{\boldsymbol{u}}\|_2 + 2\lambda_{\boldsymbol{w}}c_M\eta. \quad \text{(S1.11)}$$

As long as $5a_2c_0c\sqrt{s\log p/((n_{\mathcal{A}_\eta}+n_0)h_{\boldsymbol{w}})} \leq 1/2$, using the cone-like con-

straint for $\tilde{\boldsymbol{u}} \in \mathbb{U}$, we can derive from (S1.11) that

$$\frac{1}{2}a_1\|\tilde{\boldsymbol{u}}\|_2^2 - \left(4a_1a_2c_0c_M\sqrt{\frac{\log p}{(n_{\mathcal{A}_\eta}+n_0)h_{\boldsymbol{w}}}}\eta + 2\lambda_{\boldsymbol{w}}\sqrt{s}\right)\|\tilde{\boldsymbol{u}}\|_2 - 2\lambda_{\boldsymbol{w}}c_M\eta \leq 0.$$

Consequently, we have a contradiction under the assumption that $\|\tilde{\boldsymbol{u}}\|_2 \geq c_u$. Thus we can conclude from (S1.9) that conditioning on $\mathcal{E}_{\boldsymbol{w}} \cap \mathcal{E}'_{\boldsymbol{w}}(r)$,

$$\|\hat{\boldsymbol{u}}\|_2 \lesssim \sqrt{\frac{\log p}{(n_{\mathcal{A}_\eta}+n_0)h_{\boldsymbol{w}}}}\eta + \lambda_{\boldsymbol{w}}\sqrt{s} + \sqrt{\lambda_{\boldsymbol{w}}\eta}.$$

With the stated choices $\lambda_{\boldsymbol{w}} \asymp \sqrt{\log p/(n_{\mathcal{A}_\eta}+n_0)}$ and $h_{\boldsymbol{w}}^2 \asymp \lambda_{\boldsymbol{w}}\sqrt{s}$, we can verify that (S1.10) holds as long as $\eta \lesssim \sqrt{s}$. Then we obtain

$$\|\hat{\boldsymbol{u}}\|_2 \lesssim \sqrt{\frac{s\log p}{n_{\mathcal{A}_\eta}+n_0}} + \left(\frac{\log p}{n_{\mathcal{A}_\eta}+n_0}\right)^{\frac{1}{4}}\sqrt{\eta} + \left(\frac{\log p}{n_{\mathcal{A}_\eta}+n_0}\right)^{\frac{3}{8}}s^{-\frac{1}{8}}\eta, \quad \text{(S1.12)}$$

and

$$\begin{aligned}\|\hat{\boldsymbol{u}}\|_1 \leq 5\sqrt{s}\|\hat{\boldsymbol{u}}\|_2 + 4c_M\eta &\lesssim s\sqrt{\frac{\log p}{n_{\mathcal{A}_\eta}+n_0}} + \left(\frac{\log p}{n_{\mathcal{A}_\eta}+n_0}\right)^{\frac{1}{4}}\sqrt{s\eta} + \eta \\ &\lesssim s\sqrt{\frac{\log p}{n_{\mathcal{A}_\eta}+n_0}} + \eta,\end{aligned} \quad \text{(S1.13)}$$

conditioning on $\mathcal{E}_{\boldsymbol{w}} \cap \mathcal{E}'_{\boldsymbol{w}}(r)$.

It remains to bound the probability of the event $\mathcal{E}_{\boldsymbol{w}} \cap \mathcal{E}'_{\boldsymbol{w}}(r)$ when $\lambda_{\boldsymbol{w}} \asymp \sqrt{\log p/(n_{\mathcal{A}_\eta}+n_0)}$, which follows directly from Lemma S1.3 and Lemma S1.4. Pulling these components together, we can finally conclude that the bounds in (S1.12) and (S1.13) hold with probability at least $1 - c_1\exp(-c_2\log p)$ for some positive constant c_1 and c_2.

Step 2: Bounds for $\hat{\boldsymbol{v}}$.

Upper bound: By the optimality of $\hat{\boldsymbol{\delta}}^{\mathcal{A}_\eta}$, we have

$$\nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\hat{\boldsymbol{w}}^{\mathcal{A}_\eta} + \hat{\boldsymbol{\delta}}^{\mathcal{A}_\eta}) + \lambda_{\boldsymbol{\delta}} \mathrm{sgn}(\hat{\boldsymbol{\delta}}^{\mathcal{A}_\eta}) = 0. \tag{S1.14}$$

Consider the event $\mathcal{E}_{\boldsymbol{\delta}} = \{\lambda_{\boldsymbol{\delta}} \geq 2\|\nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta}) - \nabla Q^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta})\|_\infty\}$ and define the deterministic quantity $q_0 = \|\nabla Q^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta})\|_2$. Conditioning on $\mathcal{E}_{\boldsymbol{\delta}}$, the convexity of $\widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{w})$ and $\|\cdot\|_1$, the optimal condition (S1.14) and Hölder's inequality together imply that

$$\begin{aligned}
0 \leq & \langle \nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\hat{\boldsymbol{\beta}}) - \nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta}), \hat{\boldsymbol{v}} \rangle \\
= & \langle -\lambda_{\boldsymbol{\delta}} \mathrm{sgn}(\hat{\boldsymbol{\delta}}^{\mathcal{A}_\eta}), \hat{\boldsymbol{v}} \rangle - \langle \nabla Q^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta}), \hat{\boldsymbol{v}} \rangle - \langle \nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta}) - \nabla Q^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta}), \hat{\boldsymbol{v}} \rangle \\
\leq & \lambda_{\boldsymbol{\delta}}(\|\check{\boldsymbol{\delta}}^{\mathcal{A}_\eta}\|_1 - \|\hat{\boldsymbol{\delta}}^{\mathcal{A}_\eta}\|_1) + q_0 \|\hat{\boldsymbol{v}}\|_2 + \|\nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta}) - \nabla Q^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta})\|_\infty \|\hat{\boldsymbol{v}}\|_1 \\
\leq & \lambda_{\boldsymbol{\delta}}(2\|\check{\boldsymbol{\delta}}^{\mathcal{A}_\eta}\|_1 - \|\hat{\boldsymbol{v}}\|_1) + q_0 \|\hat{\boldsymbol{v}}\|_2 + \frac{\lambda_{\boldsymbol{\delta}}}{2}\|\hat{\boldsymbol{v}}\|_1,
\end{aligned} \tag{S1.15}$$

where $\check{\boldsymbol{\delta}}^{\mathcal{A}_\eta} = \boldsymbol{\beta} - \hat{\boldsymbol{w}}^{\mathcal{A}} = \hat{\boldsymbol{\delta}}^{\mathcal{A}_\eta} - \hat{\boldsymbol{v}} = \boldsymbol{\delta}^{\mathcal{A}_\eta} - \hat{\boldsymbol{w}}^{\mathcal{A}_\eta}$.

By Lemma S1.2, we have $q_0 \leq l_0 \mu_1 \kappa_2 h_{\boldsymbol{\delta}}^2/2$. Choose $h_{\boldsymbol{\delta}} \lesssim \sqrt{\lambda_{\boldsymbol{\delta}}}$ such that $l_0 \mu_1 \kappa_2 h_{\boldsymbol{\delta}}^2/2 < \lambda_{\boldsymbol{\delta}}/4$. Now combining this with (S1.15) leads to

$$\|\hat{\boldsymbol{v}}\|_1 \leq 8\|\check{\boldsymbol{\delta}}^{\mathcal{A}_\eta}\|_1 \leq 8(c_M \eta + \|\hat{\boldsymbol{u}}\|_1). \tag{S1.16}$$

Lower bound: To derive a lower bound for $\langle \nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\hat{\boldsymbol{\beta}}) - \nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta}), \hat{\boldsymbol{v}} \rangle$, define $D^{(0)}(\boldsymbol{\Delta}) = \langle \nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta} + \boldsymbol{\Delta}) - \nabla \widehat{Q}^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta}), \boldsymbol{\Delta} \rangle$ and consider the event $\mathcal{E}'_{\boldsymbol{\delta}}(r')$

defined as

$$\left\{ \frac{D^{(0)}(\boldsymbol{\Delta})}{\|\boldsymbol{\Delta}\|_2^2} \geq a_1 \left(1 - a_2 \sqrt{\frac{h_{\boldsymbol{\delta}} \log p}{n_0 r^2}} \frac{\|\boldsymbol{\Delta}\|_1}{\|\boldsymbol{\Delta}\|_2} \right), \quad \text{for all} \quad \boldsymbol{\Delta} \in \mathbb{B}_2(r') \right\}.$$

We set $r' = h_{\boldsymbol{\delta}}/c_0$ with $c_0 = 32 v_1^2$. We use proof by contradiction to show that conditioning on $\mathcal{E}_{\boldsymbol{\delta}} \cap \mathcal{E}'_{\boldsymbol{\delta}}(r')$,

$$\|\hat{\boldsymbol{v}}\|_2 \leq 8 a_2 c_0 \sqrt{\frac{\log p}{n_0 h_{\boldsymbol{\delta}}}} (\|\hat{\boldsymbol{u}}\|_1 + c_M \eta) + \sqrt{2 \frac{\lambda_{\boldsymbol{\delta}} (\|\hat{\boldsymbol{u}}\|_1 + c_M \eta)}{a_1}} =: c_v. \quad \text{(S1.17)}$$

In order to make use of the result on $\mathcal{E}'_{\boldsymbol{\delta}}(r')$, we let

$$h_{\boldsymbol{\delta}} \gtrsim (\log p / n_0)^{1/4} \|\hat{\boldsymbol{u}}\|_1 + \sqrt{\log p / (n_0 h_{\boldsymbol{\delta}})} \|\hat{\boldsymbol{u}}\|_1 \quad \text{(S1.18)}$$

such that $r' > c_v$.

Choose $t \in (0, 1)$ such that $t\|\hat{\boldsymbol{v}}\|_2 \leq r'$ and $t\|\hat{\boldsymbol{v}}\|_2 \geq c_v$. Denote $\tilde{\boldsymbol{v}} = t\hat{\boldsymbol{v}}$ and $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \tilde{\boldsymbol{v}}$. By the same arguments that lead to (S1.13) and (S1.16), and conditioning on $\mathcal{E}'_{\boldsymbol{\delta}}(r')$, we deduce that

$$a_1 (\|\tilde{\boldsymbol{v}}\|_2^2 - a_2 c_0 \sqrt{\frac{\log p}{n_0 h_{\boldsymbol{\delta}}}} \|\tilde{\boldsymbol{v}}\|_1 \|\tilde{\boldsymbol{v}}\|_2) - 2\lambda_{\boldsymbol{\delta}} \|\boldsymbol{\delta}^{\mathcal{A}_\eta}\|_1 \leq 0,$$

which contradicts with the assumption that $\|\tilde{\boldsymbol{v}}\|_2 = t\|\hat{\boldsymbol{v}}\|_2 \geq c_v$.

Combining (S1.16), (S1.17), and (S1.13) gives us

$$\|\hat{\boldsymbol{v}}\|_1 \lesssim \eta + s \sqrt{\log p / (n_{\mathcal{A}_\eta} + n_0)} \quad \text{(S1.19)}$$

and

$$\|\hat{\boldsymbol{v}}\|_2 \lesssim \sqrt{\frac{\log p}{n_0 h_{\boldsymbol{\delta}}}} \|\hat{\boldsymbol{u}}\|_1 + \sqrt{\lambda_{\boldsymbol{\delta}} \|\hat{\boldsymbol{u}}\|_1}. \quad \text{(S1.20)}$$

With the stated choices $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\log p/n_0}$ and $h_{\boldsymbol{\delta}}^2 \asymp \lambda_{\boldsymbol{\delta}}$, we can verify that (S1.18) holds as long as $\|\hat{\boldsymbol{u}}\|_1 \lesssim 1$. Then we obtain

$$\|\hat{\boldsymbol{v}}\|_2 \lesssim \sqrt{s}\left(\frac{\log p}{n_0}\right)^{\frac{1}{4}}\left(\frac{\log p}{n_{\mathcal{A}_\eta}+n_0}\right)^{\frac{1}{4}} + \sqrt{\eta}\left(\frac{\log p}{n_0}\right)^{\frac{1}{4}}. \tag{S1.21}$$

It remains to investigate the probability of the event $\mathcal{E}_{\boldsymbol{\delta}} \cap \mathcal{E}'_{\boldsymbol{\delta}}(r')$, which follows directly from Lemma S1.3 and Lemma S1.5. This completes the proof of Theorem 1. □

S1.3 Proof of Theorem 2

Proof. Recall that $\widehat{T}^{(k)} = \widehat{Q}^{(0)}(\hat{\boldsymbol{\beta}}^{(k)}; \mathcal{I}_0^{\mathrm{va}}) - \widehat{Q}^{(0)}(\hat{\boldsymbol{\beta}}^{(0)}; \mathcal{I}_0^{\mathrm{va}})$, where $\widehat{Q}^{(0)}(\boldsymbol{w}; \mathcal{I}) = 1/|\mathcal{I}| \sum_{i\in\mathcal{I}} \rho_\tau(y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top \boldsymbol{w})$. We proof the result in Theorem 2 by showing that under a proper choice of t, we have

$$\mathbb{P}\big(\inf_{k\in\mathcal{A}_\eta^c} \widehat{T}^{(k)} \geq t(\widehat{Q}^{(0)}(\hat{\boldsymbol{\beta}}^{(0)}; \mathcal{I}_0^{\mathrm{va}}) \vee 0.01) \geq \sup_{k\in\mathcal{A}_\eta} \widehat{T}^{(k)}\big) \to 1. \tag{S1.22}$$

In order to establish (S1.22), we first investigate the bound of $\widehat{T}^{(k)}$. Define $Q^{(0)}(\boldsymbol{w}) = \mathbb{E}[\widehat{Q}^{(0)}(\boldsymbol{w}; \mathcal{I}_0^{\mathrm{va}})]$. We have the following decomposition:

$$\begin{aligned}
\widehat{T}^{(k)} =& \Big\{Q^{(0)}(\boldsymbol{\beta}^{(k)}) - Q^{(0)}(\boldsymbol{\beta})\Big\} + \Big\{\widehat{Q}^{(0)}(\hat{\boldsymbol{\beta}}^{(k)}; \mathcal{I}_0^{\mathrm{va}}) - \widehat{Q}^{(0)}(\boldsymbol{\beta}^{(k)}; \mathcal{I}_0^{\mathrm{va}})\Big\} \\
&+ \Big\{\widehat{Q}^{(0)}(\boldsymbol{\beta}; \mathcal{I}_0^{\mathrm{va}}) - \widehat{Q}^{(0)}(\hat{\boldsymbol{\beta}}^{(0)}; \mathcal{I}_0^{\mathrm{va}})\Big\} \\
&+ \Big\{\widehat{Q}^{(0)}(\boldsymbol{\beta}^{(k)}; \mathcal{I}_0^{\mathrm{va}}) - \widehat{Q}^{(0)}(\boldsymbol{\beta}; \mathcal{I}_0^{\mathrm{va}}) - \Big(Q^{(0)}(\boldsymbol{\beta}^{(k)}) - Q^{(0)}(\boldsymbol{\beta})\Big)\Big\} \\
\equiv& I_1^{(k)} + I_2^{(k)} + I_3^{(0)} + I_4^{(k)}.
\end{aligned} \tag{S1.23}$$

We analyze the above four terms separately. We start with $I_1^{(k)}$. By the mean-value theorem, we have

$$I_1^{(k)} = (\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta})^\top \mathbb{E}\left[\int_0^1 \boldsymbol{x}^{(0)}(\boldsymbol{x}^{(0)})^\top f_{\epsilon^{(0)}|\boldsymbol{x}^{(0)}}(t(\boldsymbol{x}^{(0)})^\top(\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}))\mathrm{dt}\right](\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}),$$

for some $t \in (0,1)$. Therefore, under Condition 7, we have

$$\sup_{k\in\mathcal{A}_\eta} I_1^{(k)} \le \bar{\lambda} \sup_{k\in\mathcal{A}_\eta} \|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2^2 \quad \text{and} \quad \inf_{k\in\mathcal{A}_\eta^c} I_1^{(k)} \ge \underline{\lambda} \inf_{k\in\mathcal{A}_\eta^c} \|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2^2. \tag{S1.24}$$

Now we investigate $I_2^{(k)}$. Let $u_i^{(k)} = (\boldsymbol{x}_i^{(0)})^\top(\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\beta}}^{(k)})$ and $\Psi_\tau(\epsilon) = \tau - \mathbf{I}\{\epsilon < 0\}$. By Knight's Identity, we have

$$\begin{aligned} I_2^{(k)} =& \frac{1}{|\mathcal{I}_0^{\mathrm{va}}|}\sum_{i\in\mathcal{I}_0^{\mathrm{va}}}\left[\rho_\tau(y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top\hat{\boldsymbol{\beta}}^{(k)}) - \rho_\tau(y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top\boldsymbol{\beta}^{(k)})\right] \\ =& \frac{1}{|\mathcal{I}_0^{\mathrm{va}}|}\sum_{i\in\mathcal{I}_0^{\mathrm{va}}}(\boldsymbol{x}_i^{(0)})^\top(\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\beta}}^{(k)})\Psi_\tau(\epsilon_i^{(0)} + u_i^{(k)}) \\ &+ \frac{1}{|\mathcal{I}_0^{\mathrm{va}}|}\sum_{i\in\mathcal{I}_0^{\mathrm{va}}}\int_0^{(\boldsymbol{x}_i^{(0)})^\top(\hat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)})} \mathbf{I}\{\epsilon_i^{(0)} \le t - u_i^{(k)}\} - \mathbf{I}\{\epsilon_i^{(0)} \le -u_i^{(k)}\}\mathrm{dt}. \end{aligned}$$

Since $|\Psi_\tau(\cdot)| \le \max\{\tau, 1-\tau\}$ and $|\mathbf{I}\{\epsilon_i^{(0)} \le t - u_i^{(k)}\} - \mathbf{I}\{\epsilon_i^{(0)} \le -u_i^{(k)}\}| \le 1$, we have

$$|I_2^{(k)}| \le \frac{2}{|\mathcal{I}_0^{\mathrm{va}}|}\sum_{i\in\mathcal{I}_0^{\mathrm{va}}}|(\boldsymbol{x}_i^{(0)})^\top(\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\beta}}^{(k)})| := \frac{2}{|\mathcal{I}_0^{\mathrm{va}}|}\sum_{i\in\mathcal{I}_0^{\mathrm{va}}} z_i^{(k)}.$$

Conditioning on $\mathcal{I}_0^{\mathrm{tr}}$, $\{z_i^{(k)}\}_{i\in\mathcal{I}_0^{\mathrm{va}}}$ are independent sub-gaussian random variables with parameter no more than $v_1^2\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)}\|_2^2$. By tail bounds and

noting that $|\mathcal{I}_0^{\mathrm{va}}| = n_0/2$, we obtain that

$$\mathbb{P}\left(|I_2^{(k)}| \leq \left(\mu_1 + 2v_1\sqrt{\frac{\log p}{n_0}}\right)\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)}\|_2\right) \geq 1 - p^{-1}. \tag{S1.25}$$

Similarly, we can derive that

$$\mathbb{P}\left(|I_3^{(0)}| \leq \left(\mu_1 + 2v_1\sqrt{\frac{\log p}{n_0}}\right)\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}\|_2\right) \geq 1 - p^{-1}. \tag{S1.26}$$

From the result in Tan et al. (2022), we have $\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}\|_2 \lesssim \Omega_0 = \sqrt{s\log p/n_0}$ with probability at least $1 - p^{-1}$. It remains to bound $\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)}\|_2$ for $k \in [K]$. The following lemma provides a high probability upper bound for the ℓ_2 estimation error of $\hat{\boldsymbol{\beta}}^{(k)}$. The proof of this lemma is similar to the proof in the first step of Theorem 1 and is relegated to Section S5.6.

Lemma S1.6. *Under the conditions in Theorem 2, then for the estimator* $\hat{\boldsymbol{\beta}}^{(k)}$ *obtained from the Trans-SQR algorithm, then we have*

$$\mathbb{P}\left(\|\hat{\boldsymbol{u}}^{(k)}\|_2 \lesssim \Omega_k\right) \geq 1 - 5p^{-1},$$

where $\Omega_k = \sqrt{s'\log p/(n_k + n_0)} + (\log p/(n_k + n_0))^{1/4}\sqrt{\eta'} + (\log p/(n_k + n_0))^{3/8}(s')^{-1/8}\eta'$.

By Lemma S1.6 and union bounds, we can conclude from (S1.25) and (S1.26) that there exist a constant C such that

$$\mathbb{P}\left(|I_2^{(k)}| + |I_3^{(0)}| \leq \left(\mu_1 + 2v_1\sqrt{\frac{\log p}{n_0}}\right)C(\Omega_{\max} + \Omega_0)\right) \geq 1 - 8p^{-1}. \tag{S1.27}$$

Lastly, we analyze $I_4^{(k)}$. Again by Knight's Identity, we obtain that

$$\begin{aligned} I_4^{(k)} =& \frac{1}{|\mathcal{I}_0^{\mathrm{va}}|}\sum_{i\in\mathcal{I}_0^{\mathrm{va}}}\Big[\rho_\tau(y_i^{(0)}-(\boldsymbol{x}_i^{(0)})^\top\boldsymbol{\beta}^{(k)})-\rho_\tau(y_i^{(0)}-(\boldsymbol{x}_i^{(0)})^\top\boldsymbol{\beta})\Big] \\ & -\mathbb{E}\Big[\rho_\tau(y_i^{(0)}-(\boldsymbol{x}_i^{(0)})^\top\boldsymbol{\beta}^{(k)})-\rho_\tau(y_i^{(0)}-(\boldsymbol{x}_i^{(0)})^\top\boldsymbol{\beta})\Big] \\ =& \frac{1}{|\mathcal{I}_0^{\mathrm{va}}|}\sum_{i\in\mathcal{I}_0^{\mathrm{va}}}\{z_{i1}-\mathbb{E}z_{i1}\}+\frac{1}{|\mathcal{I}_0^{\mathrm{va}}|}\sum_{i\in\mathcal{I}_0^{\mathrm{va}}}\{z_{i2}-\mathbb{E}z_{i2}\}, \end{aligned}$$

where $z_{i1} = (\boldsymbol{x}_i^{(0)})^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)})\Psi_\tau(\epsilon_i^{(0)})$ and $z_{i2} = \int_0^{(\boldsymbol{x}_i^{(0)})^\top(\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta})}\mathbf{I}\{\epsilon_i^{(0)}\le t\}-\mathbf{I}\{\epsilon_i^{(0)}\le 0\}\mathrm{dt}$. Note that $\mathbb{E}z_{i1}=0$ and $\Psi_\tau(\cdot)$ is bounded and hence sub-gaussian with parameter less than 1. Therefore, $\{z_{i1}\}_{i\in\mathcal{I}_0^{\mathrm{va}}}$ are $v_1^2\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2^2$-sub-exponential. Besides, $|z_{i2}|\le|(\boldsymbol{x}_i^{(0)})^\top(\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta})|$. Therefore, $\{z_{i2}\}_{i\in\mathcal{I}_0^{\mathrm{va}}}$ are independent sub-gaussian random variables with parameter almost $v_1^2\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2^2$. By tails bounds and union bounds, we obtain that

$$\mathbb{P}\left(|I_4^{(k)}|\le 6v_1\sqrt{\frac{\log p}{n_0}}\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2\right)\ge 1-3p^{-1}. \tag{S1.28}$$

Combining (S1.23), (S1.24), (S1.27), and (S1.28) leads to

$$\sup_{k\in\mathcal{A}_\eta}\widehat{T}^{(k)}\le\bar{\lambda}\sup_{k\in\mathcal{A}_\eta}\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2^2+6v_1\sqrt{\frac{\log p}{n_0}}\sup_{k\in\mathcal{A}_\eta}\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2+Cc_n,\quad\text{and}$$

$$\inf_{k\in\mathcal{A}_\eta^c}\widehat{T}^{(k)}\ge\underline{\lambda}\inf_{k\in\mathcal{A}_\eta^c}\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2^2-6v_1\sqrt{\frac{\log p}{n_0}}\sup_{k\in\mathcal{A}_\eta^c}\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2-Cc_n,$$

with probability $1-11p^{-1}$, where $c_n=(\mu_1+2v_1\sqrt{\log p/n_0})(\Omega_{\max}+\Omega_0)$.

Under Condition 7, $\inf_{k\in\mathcal{A}_\eta^c}\widehat{T}^{(k)}\ge\underline{\lambda}/2\inf_{k\in\mathcal{A}_\eta^c}\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2^2$, and then we can choose $\bar{\lambda}\sup_{k\in\mathcal{A}_\eta}\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2^2+\sqrt{\log p/n_0}\sup_{k\in[K]}\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2+c_n\lesssim$

$t \lesssim \underline{\lambda}/2 \inf_{k \in \mathcal{A}_\eta^c} \|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}\|_2^2$ such that (S1.22) is satisfied. This proves the claimed result. □

S2 Additional Simulation results

S2.1 QR Transfer with More Heterogeneous Designs

In this subsection, we conduct additional simulations to investigate the impact of heterogeneous designs on the performance of the transferred estimators. Specifically, we consider the same setting as that in Section 4 of the main paper under Gaussian errors with the only difference in the generation of covariates in the sources. We consider $\boldsymbol{x}^{(0)} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \left(0.7^{|i-j|}\right)_{1 \le i,j \le p}$ and $\boldsymbol{x}^{(k)} \sim \mathcal{N}\left(\mathbf{0}_p, \boldsymbol{\Sigma} + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right)$ with $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}_p, \delta^2 \boldsymbol{I}_p\right)$ for $k = 1, \ldots, K$. We fix $h = 10$, $\tau = 0.5$ and consider $A \in \{4, 8, 12\}$. We vary δ from 0.5 to 1.5 with stepsize 0.2 and evaluate the change of performance with respect to δ. The average estimation errors of the five methods (ℓ_1-SQR, Oracle-TSQR, Oracle-TQR, Naive-TSQR, and TSQR) over 100 replications are displayed in Figure S2.1.

As we can see from Figure S2.1, the performance of our Oracle-TSQR as well as the TSQR remains stable as the heterogeneity parameter δ of the designs increases, while the performance of the Naive-TSQR and the

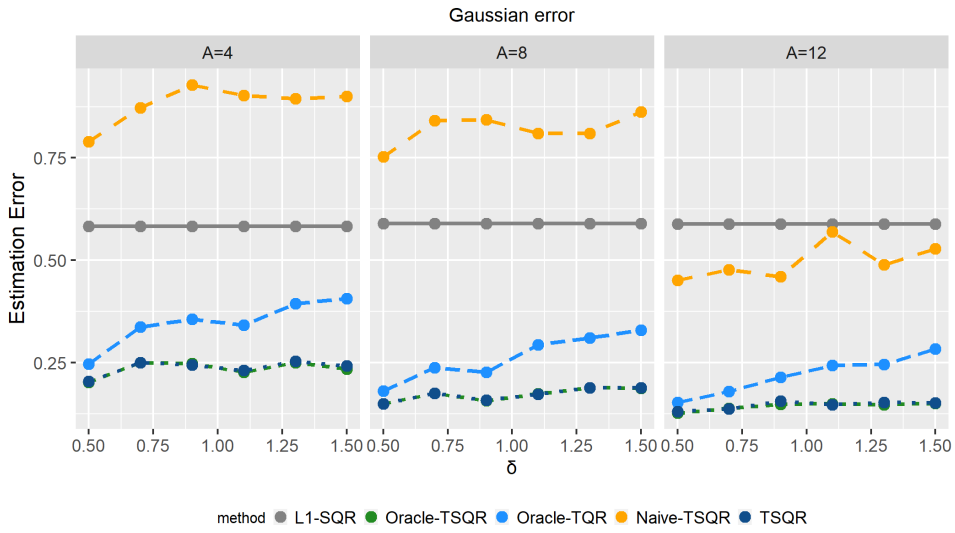

Figure S2.1: ℓ_2 estimation errors of various methods with respect to δ under Gaussian errors at quantile level $\tau = 0.5$ and $\eta = 10$, averaged over 100 replications. Here the horizontal axis δ represents the heterogeneity parameter of the designs.

non-smoothed Oracle-TQR becomes slightly worse as δ increases. This illustrates the stability of our smoothed QR transferring algorithms.

S2.2 Sensitivity to the smoothing bandwidths

We investigate the sensitivity of our Oracle-TSQR to the smoothing bandwidths in this subsection. Specifically, we consider the same setting as that in Section 4 of the main paper under Gaussian errors with η fixed at 10 and A fixed at 8. For simulations in the main text, we choose the bandwidths as $\max\{0.05, \sqrt{\tau(1-\tau)}\{\log(p)/n\}^{1/4}\}$ as recommended in Tan

et al. (2022) for a specific n in the corresponding problem. For example, for $A = 8$, we have $h_{\boldsymbol{w}} \approx 0.14$ and $h_{\boldsymbol{\delta}} \approx 0.07$ at the quantile level $\tau = 0.5$. Here we consider different choices of bandwidths $(h_{\boldsymbol{w}}, h_{\boldsymbol{\delta}})$, both of which take values in $\{0.05, 0.10, ..., 0.3\}$. There are 36 different combinations in total. For each combination choice of bandwidths, we replicate the simulation 100 times, and here in Figure S2.2 we present the average ℓ_2-estimation errors for each combination. For better comparison, we also report the average estimation errors of the five methods (ℓ_1-SQR, Oracle-TQR, Oracle-TSQR, Naive-TSQR, and TSQR) with the smoother ones using bandwidths $\max\{0.05, \sqrt{\tau(1-\tau)}\{\log(p)/n\}^{1/4}\}$ at $\eta = 10$ and $A = 8$ under different quantile levels.

As we can see from Figure S2.2, the estimation errors are not very sensitive to the choice of the smoothing bandwidths $h_{\boldsymbol{w}}$ and $h_{\boldsymbol{\delta}}$ under each quantile level. Let us take $\tau = 0.5$ for illustration. As reported in Table S2.1, the average estimation error of Oracle-TSQR chosen by the recommended bandwidths in Tan et al. (2022) is 0.1307 with a standard deviation of 0.0471. We can see from the middle panel that the estimation errors under various choices of $h_{\boldsymbol{w}}$ and $h_{\boldsymbol{\delta}}$ fall between 0.1314 and 0.1449, the range of which is approximately 1/3 of its standard deviation (0.0471). This suggests that the performance of our proposed Oracle-TSQR is not really sensitive

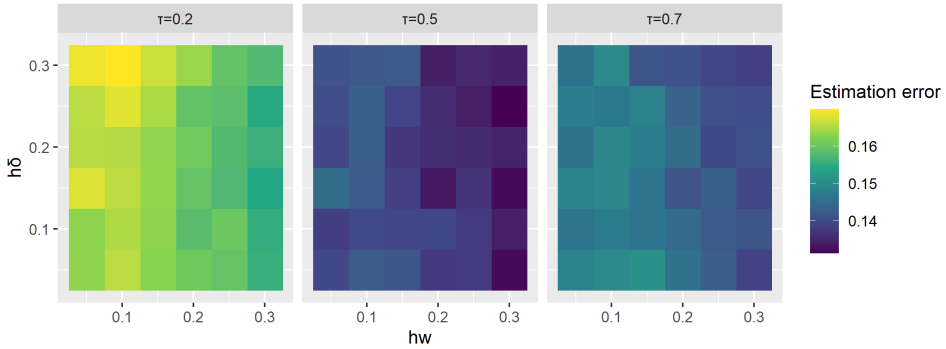

Figure S2.2: ℓ_2 estimation errors of Oracle-TSQR with different choice of bandwidths under Gaussian errors at $\eta = 10$, $A = 8$, and quantile levels $\tau = 0.2, 0.5, 0.7$.

Table S2.1: Estimation errors and standard deviations of various methods at $\eta = 10$ and $A = 8$ under different quantile levels, with the smoothed estimators based on the recommended bandwidths.

	L1-SQR	Oracle-TSQR	Oracle-TQR	Naive-TSQR	TSQR
τ	ℓ_2-estimation error (standard deviation)				
0.2	0.7530 (0.2530)	**0.1631 (0.0479)**	0.1918 (0.0532)	0.6804 (0.1935)	0.1609 (0.0488)
0.5	0.5897 (0.1840)	**0.1307 (0.0471)**	0.1466 (0.0471)	0.6202 (0.1859)	0.1322 (0.0485)
0.7	0.6153 (0.1903)	**0.1406 (0.0408)**	0.1648 (0.0516)	0.6353 (0.1767)	0.1462 (0.0482)

to the smoothing bandwidths. We also note that under various choices of bandwidths, the estimation errors of our Oracle-TSQR all perform slightly better under the non-smoothed Oracle-TQR, which is 0.1466. This again illustrates the benefit of our convolution smoothing.

Similar conclusions also apply for the quantile levels $\tau = 0.2$ and $\tau =$

0.7 and thus omitted here. In conclusion, we recommend to choose the bandwidths as that in Tan et al. (2022), although in practice one can also use CV for selecting the bandwidths for optimal numeric performance.

S3 Distributed QR Transfer

S3.1 Distributed QR Transfer Algorithm

Here we adopt the approximate Newton-type method proposed by Shamir et al. (2014) and further examined in Jordan et al. (2019) and Wang et al. (2017) to solve the transferring step in Algorithm 1 in a distributed manner.

With our loss of generality, we set $\mathcal{A}_\eta = \{1, \ldots, |\mathcal{A}_\eta|\}$. Let $\boldsymbol{\alpha}$ denote a $|\mathcal{A}_\eta| + 1$-dimensional vector with the $k+1$-th element being α_k. We first generate the pilot sample sizes $\{n_k^*\}_{k\in\mathcal{A}\cup\{0\}}$ from multinomial distribution $\mathcal{M}(n_*, \boldsymbol{\alpha})$ with $n_* = \rho_0(n_{\mathcal{A}_\eta} + n_0)$ for some $\rho_0 \in (0,1)$. For each $k \in \mathcal{A}\cup\{0\}$, we randomly select n_k^* samples from the k-th site, with the index set denoted by $\mathcal{D}_k^*$. Transfer $\{((\boldsymbol{x}_i^{(k)})^\top, y_i^{(k)})\}_{i\in\mathcal{D}_k^*}$ from the k-th site to the target site.

Denote the empirical smoothed quantile loss on the pilot pooled data by

$$\widehat{Q}_{h_*}^*(\boldsymbol{w}) = \frac{1}{n_* h_*} \sum_{k\in\mathcal{A}_\eta\cup\{0\}} \sum_{i\in\mathcal{D}_k^*} \int_{-\infty}^{\infty} \rho_\tau(u) K\left(\frac{u + \boldsymbol{w}^\top \boldsymbol{x}_i^{(k)} - y_i^{(k)}}{h_*}\right) \mathrm{d}u$$

. The gradient vectors of $\widehat{Q}^*_{h_*}(\boldsymbol{w})$ are given respectively by

$$\nabla \widehat{Q}^*_{h_*}(\boldsymbol{w}) = 1/n^* \sum_{k\in\mathcal{A}\cup\{0\}} \sum_{i\in\mathcal{D}^*_k} \{\bar{K}(((\boldsymbol{x}_i^{(k)})^\top \boldsymbol{w} - y_i^{(k)})/h_*) - \tau\}\boldsymbol{x}_i^{(k)}.$$

Given an initial estimator $\tilde{\boldsymbol{w}}^{(0)}$, consider the first-order Taylor expansion of $\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w})$ around $\tilde{\boldsymbol{w}}^{(0)}$:

$$\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}) = \widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\tilde{\boldsymbol{w}}^{(0)}) + \langle \nabla \widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\tilde{\boldsymbol{w}}^{(0)}), \boldsymbol{w} - \tilde{\boldsymbol{w}}^{(0)}\rangle + R^{\mathcal{A}_\eta}(\boldsymbol{w}), \tag{S3.1}$$

where $R^{\mathcal{A}_\eta}(\boldsymbol{w})$ is the linear approximation error. In the distributed environment, the gradient $\nabla \widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\tilde{\boldsymbol{w}}^{(0)})$ can be easily be communicated. Therefore, it suffices to find a good replacement of $R^{\mathcal{A}_\eta}(\boldsymbol{w})$. Here we propose to use the analogous approximation error from the loss of the pilot pooled sample, that is, we approximate $R^{\mathcal{A}_\eta}(\boldsymbol{w})$ by

$$R^{\mathcal{A}_\eta}(\boldsymbol{w}) \approx R^*(\boldsymbol{w}) = \widehat{Q}^*_{h_*}(\boldsymbol{w}) - \widehat{Q}^*_{h_*}(\tilde{\boldsymbol{w}}^{(0)}) - \langle \nabla \widehat{Q}^*_{h_*}(\tilde{\boldsymbol{w}}^{(0)}), \boldsymbol{w} - \tilde{\boldsymbol{w}}^{(0)}\rangle. \tag{S3.2}$$

Plugging (S3.2) into (S3.1) motivates us to consider the surrogate smoothed quantile loss:

$$\tilde{Q}(\boldsymbol{w}) = \widehat{Q}^*_{h_*}(\boldsymbol{w}) - \langle \nabla \widehat{Q}^*_{h_*}(\tilde{\boldsymbol{w}}^{(0)}) - \nabla \widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\tilde{\boldsymbol{w}}^{(0)}), \boldsymbol{w}\rangle.$$

Consequently, a communication-efficient penalized estimator for $\boldsymbol{w}^{\mathcal{A}_\eta}$ can be obtained by solving

$$\tilde{\boldsymbol{w}}^{(1)} \in \underset{\boldsymbol{w}\in\mathbb{R}^p}{\arg\min} \quad \tilde{Q}(\boldsymbol{w}) + \lambda_1 \|\boldsymbol{w}\|_1. \tag{S3.3}$$

The above procedure could be done iteratively. In fact, we can define shifted losses $\tilde{Q}^{(t)}(\boldsymbol{w}) = \widehat{Q}^*_{h_*}(\boldsymbol{w}) - \langle \nabla \widehat{Q}^*_{h_*}(\tilde{\boldsymbol{w}}^{(t-1)}) - \nabla \widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\tilde{\boldsymbol{w}}^{(t-1)}), \boldsymbol{w} \rangle$ and obtain a sequence of estimators by solving $\tilde{\boldsymbol{w}}^{(t)} \in \underset{\boldsymbol{w}\in\mathbb{R}^p}{\arg\min} \tilde{Q}^{(t)}(\boldsymbol{w}) + \lambda_t \|\boldsymbol{w}\|_1$ for $t = 2, \ldots, T$. The details for our distributed Trans-SQR are presented in Algorithm 3.

Algorithm 3 Distributed-Oracle-Trans-SQR Algorithm

Input: Target data $(\boldsymbol{X}^{(0)}, \boldsymbol{y}^{(0)})$, source data $\{(\boldsymbol{X}^{(k)}, \boldsymbol{y}^{(k)})\}_{k\in\mathcal{A}_\eta}$, pilot pooled sample $\{\{(\boldsymbol{x}_i^{(k)}, y_i^{(k)})\}_{i\in\mathcal{D}_k^*}\}_{k\in\mathcal{A}_\eta\cup\{0\}}$, an initial estimate $\tilde{\boldsymbol{w}}^{(0)}$, number of iterations T, penalty parameters $(\lambda_{\boldsymbol{w}}, \lambda_{\boldsymbol{\delta}}, \{\lambda_t\}_{t=1}^T)$ and bandwidths $(h_{\boldsymbol{w}}, h_{\boldsymbol{\delta}}, h_*)$.

1: **Distributed Transferring:** For $t = 1, \ldots, T$, compute

$$\tilde{\boldsymbol{w}}^{(t)} \leftarrow \underset{\boldsymbol{w}\in\mathbb{R}^p}{\arg\min} \quad \tilde{Q}^{(t)}(\boldsymbol{w}) + \lambda_t \|\boldsymbol{w}\|_1 \tag{S3.4}$$

2: **Debiasing:** Compute

$$\tilde{\boldsymbol{\delta}}^{(T)} \leftarrow \ell_1\text{-SQR}(\{(\boldsymbol{x}_i^{(0)}, y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top \tilde{\boldsymbol{w}}^{(T)})\}_{i=1}^{n_0}; \lambda_{\boldsymbol{\delta}}, h_{\boldsymbol{\delta}})$$

Output:

$$\tilde{\boldsymbol{\beta}}^{(T)} = \tilde{\boldsymbol{w}}^{(T)} + \tilde{\boldsymbol{\delta}}^{(T)}$$

As we will show in Theorem S3.1, a "good" estimator $\tilde{\boldsymbol{w}}^{(0)}$ is needed to guarantee the theoretical properties of $\tilde{\boldsymbol{w}}^{(T)}$. Taking the heterogeneity among the target and sources into consideration, we propose to obtain $\tilde{\boldsymbol{w}}^{(0)}$

on the pilot pooled sample by solving

$$\tilde{\boldsymbol{w}}^{(0)} \in \underset{\boldsymbol{w}\in\mathbb{R}^p}{\arg\min} \quad \widehat{Q}^*_{h_*}(\boldsymbol{w}) + \lambda_*\|\boldsymbol{w}\|_1. \tag{S3.5}$$

The optimization problem in (S3.4) could be solved by the local adaptive majorize-minimize (LAMM) algorithm (Fan et al., 2018; Tan et al., 2022).

S3.2 Theory for Distributed-Oracle-Trans-SQR

Here we establish the analogous estimation error bounds for our distributed QR transfer estimator. In addition to Condition 3, we impose the following boundedness condition on the covariate vectors.

Condition S3.1. *There exists some constant* $B \geq 1$ *such that* $\max_{j\in[p]} |x_j^{(k)}| \leq B$ *almost surely for all* $k = 0, \ldots, K$.

To better understand the mechanism of the distributed QR transfer estimator, we first present a deterministic result based on some "good" events. Define the ℓ_2-ball $\mathbb{B}_2(r) = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}\|_2 \leq r\}$ and the cone-like set $\Lambda = \Lambda(s, \eta) = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_1 \leq 5\sqrt{s}\|\boldsymbol{u}\|_2 + 4c_M\eta\}$. Consider the events $\mathcal{E}_0(r) = \{\tilde{\boldsymbol{w}}^{(0)} : \tilde{\boldsymbol{w}}^{(0)} - \boldsymbol{w}^{\mathcal{A}_\eta} \in \mathbb{B}_2(r) \cap \Lambda\}$ and $\mathcal{E}_{\boldsymbol{w}}(\lambda_{\boldsymbol{w}}) = \{\lambda_{\boldsymbol{w}} \geq 2\|\nabla\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta}) - \nabla Q^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta})\|_\infty\}$.

Proposition S3.1. *Assume Conditions 1-5 and S3.1 hold. Define* $\rho_* = \sqrt{\log p/(n_{\mathcal{A}_\eta} + n_0)}/h_{\boldsymbol{w}} + \sqrt{\log p/n_*}/h_*$ *for some* $h_* > 0$. *Let* $0 < h_{\boldsymbol{w}} \leq$

$h_* \lesssim 1$ and $\lambda_1 = \lambda_{\boldsymbol{w}} + \rho$ satisfy $\rho \asymp \max\{s^{-1/2}(h_{\boldsymbol{w}}^2 + h_* r_*), r_*\sqrt{s}\rho_* + \eta\rho_*\}$ and $h_* \gtrsim \sqrt{\log p/(n_* h_*)}\eta + \lambda_1\sqrt{s} + \sqrt{\lambda_1\eta}$. Then conditioning on the event $\mathcal{E}_0(r_*) \cap \mathcal{E}_{\boldsymbol{w}}(\lambda_{\boldsymbol{w}})$, the one-step estimator $\tilde{\boldsymbol{w}}^{(1)}$ obtained by (S3.3) satisfies $\tilde{\boldsymbol{w}}^{(1)} \in \Lambda$ and

$$\begin{aligned} &\inf_{\boldsymbol{B}\in\Theta(s,\eta)} \mathbb{P}\left(\|\tilde{\boldsymbol{w}}^{(1)} - \boldsymbol{w}^{\mathcal{A}_\eta}\|_2 \lesssim \lambda_{\boldsymbol{w}}\sqrt{s} + h_{\boldsymbol{w}}^2 + \varphi_0 r_* + \varphi_1\sqrt{\eta} + \varphi_2\eta\right) \\ &\geq 1 - c_1\exp(-c_2\log p), \end{aligned} \tag{S3.6}$$

where $\varphi_0 = s\rho_* + h_*$, $\varphi_1 = (s^{-1/4}\sqrt{h_*} + s^{1/4}\sqrt{\rho_*})\sqrt{r_*} + s^{-1/4}h_{\boldsymbol{w}} + \sqrt{\lambda_{\boldsymbol{w}}}$ and $\varphi_2 = \sqrt{s}\rho_* + \sqrt{\rho_*} + \sqrt{\log p/(n_* h_*)}$.

The upper bound in (S3.6) can be decomposed into three parts: (i) the first two terms $\lambda_{\boldsymbol{w}}\sqrt{s} + h_{\boldsymbol{w}}^2$ is the nearly optimal rate when all transferable sources are used and there's no heterogeneity among the sources and target; (ii) the third term $\varphi_0 r_*$ is a contraction of the initial estimation error given by $\tilde{\boldsymbol{w}}^{(0)}$ and (iii) the last two terms could be seen as the price we pay for the heterogeneity among used data sets.

With large enough samples in the pilot and pooled sources, that is, $n_* \gtrsim s^2\log p$ and $n_{\mathcal{A}_\eta} + n_0 \gtrsim s^3\log p$, the contraction factor φ_0 can be strictly less than 1, which will consequently improve the convergence rate of $\tilde{\boldsymbol{w}}^{(0)}$.

When homogeneity is assumed among the sources and target—namely, $\eta = 0$, Proposition S3.1 degenerates to Theorem 11 in Tan et al. (2022).

Therefore, it can be seen as an extension of the established results for distributed smoothing QR estimators in Tan et al. (2022) to allow for the existence of heterogeneity in the high-dimensional setting.

We are now ready to present the estimation error bounds for $\tilde{\boldsymbol{\beta}}^{(T)}$from Algorithm 3.

Theorem S3.1. *Assume Conditions 1-5 and S3.1 hold. Suppose that* $n_* \gtrsim s^2 \log p$, $n_{\mathcal{A}_\eta} \gtrsim s^3 \log p$, $\eta \lesssim s\sqrt{\log p/n_*}$ *and* $\eta \lesssim (s^5 \log p/n_*)^{1/8}$. *Choose the regularization parameters* $(\lambda_{\boldsymbol{w}}, \lambda_{\boldsymbol{\delta}})$ *and bandwidths* $(h_{\boldsymbol{w}}, h_{\boldsymbol{\delta}})$ *as that in Theorem 1. Further choose* $h_* \asymp s^{1/2}(\log p/n_*)^{1/4}$ *and* $\lambda_t (t \geq 1)$ *as*

$$\lambda_t \asymp \sqrt{\frac{\log p}{n_{\mathcal{A}_\eta} + n_0}} + \max\left\{\frac{s^2 \log p}{n_*}, \frac{s^3 \log p}{n_{\mathcal{A}_\eta} + n_0}\right\} \sqrt{\frac{\log p}{n_*}}.$$

With the initial estimator $\tilde{\boldsymbol{w}}^{(0)}$ *given by (S3.5) and the number of iterations* $T \asymp \lceil \log((n_{\mathcal{A}_\eta} + n_0)/n_*) \rceil$, *the distributed QR transfer estimator* $\tilde{\boldsymbol{\beta}}^{(T)}$ *obtained from Algorithm 3 satisfies the error bounds*

$$\inf_{\boldsymbol{B} \in \Theta(s,\eta)} \mathbb{P}\left(\|\tilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}\|_1 \lesssim s\left(\frac{\log p}{n_{\mathcal{A}_\eta} + n_0}\right)^{1/2} + a_\eta + \eta \right) \geq 1 - c_1 \exp\left(-c_2 \log p\right),$$

$$\inf_{\boldsymbol{B} \in \Theta(s,\eta)} \mathbb{P}\left(\|\tilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}\|_2 \lesssim \left(\frac{\log p}{n_0}\right)^{1/4} \left(\sqrt{s}\left(\frac{\log p}{n_{\mathcal{A}_\eta} + n_0}\right)^{1/4} + \sqrt{\eta} + \sqrt{a_\eta} \right) \right)$$

$$\geq 1 - c_1 \exp\left(-c_2 \log p\right),$$

where $a_\eta = (s \log p/n_*)^{3/8} \sqrt{s\eta}$.

Remark 3. *In comparison to the non-distributed results in Theorem 1, a*

stronger condition $\eta \ll (s\sqrt{\log p/n_0}) \wedge (n_*^3 s \log p/n_0^4)^{1/4}$ *is needed in the distributed setting for the improvement of estimation error.*

S4 Proof of Results in Section S1

S4.1 Proof of Proposition S3.1

Proof. For simplicity, we write $\tilde{\boldsymbol{w}} = \tilde{\boldsymbol{w}}^{(1)}$ and $\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{w}} - \boldsymbol{w}^{\mathcal{A}_\eta}$. By the optimality of $\tilde{\boldsymbol{w}}$, we have $\nabla\tilde{Q}(\tilde{\boldsymbol{w}}) + \lambda_1 \mathrm{sgn}(\tilde{\boldsymbol{w}}) = 0$. Next, the convexity of $\tilde{Q}(\cdot)$ and $\|\cdot\|_1$ implies that

$$\begin{aligned} 0 &\le \langle \nabla\tilde{Q}(\boldsymbol{w}) - \nabla\tilde{Q}(\boldsymbol{w}^{\mathcal{A}_\eta}), \tilde{\boldsymbol{z}}\rangle = \langle -\lambda_1 \mathrm{sgn}(\tilde{\boldsymbol{w}}), \tilde{\boldsymbol{z}}\rangle - \langle \nabla\tilde{Q}(\boldsymbol{w}^{\mathcal{A}_\eta}), \tilde{\boldsymbol{z}}\rangle \\ &\le \lambda_1 \|\boldsymbol{w}^{\mathcal{A}_\eta}\|_1 - \lambda_1 \|\tilde{\boldsymbol{w}}\|_1 - \langle \nabla\tilde{Q}(\boldsymbol{w}^{\mathcal{A}_\eta}), \tilde{\boldsymbol{z}}\rangle \end{aligned} \tag{S4.1}$$

Define the vector-index random processes

$$H_0(\boldsymbol{w}) = \nabla\widehat{Q}^*_{h_*}(\boldsymbol{w}) - \nabla\widehat{Q}^*_{h_*}(\boldsymbol{w}^{\mathcal{A}_\eta}), \quad \text{and} \quad H(\boldsymbol{w}) = \nabla\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}) - \nabla\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta}).$$

Note that $\mathbb{E}\widehat{Q}^*_{h_*}(\boldsymbol{w}) = Q^{\mathcal{A}_\eta}_{h_*}(\boldsymbol{w})$. Recall the definition of $\mathbb{U}$ in (S1.8). For $r > 0$, define the suprema of random processes over our interested region $\mathbb{W} := \{\boldsymbol{w} : \boldsymbol{w} - \boldsymbol{w}^{\mathcal{A}_\eta} \in \mathbb{B}_2(r) \cap \mathbb{U}\}$

$$\Pi_0(r) = \sup_{\boldsymbol{w}\in\mathbb{W}} \|H_0(\boldsymbol{w}) - \mathbb{E}H_0(\boldsymbol{w})\|_\infty, \quad \Pi(r) = \sup_{\boldsymbol{w}\in\mathbb{W}} \|H(\boldsymbol{w}) - \mathbb{E}H(\boldsymbol{w})\|_\infty$$

and the deterministic quantities

$$q(r) = \sup_{\boldsymbol{w}\in\mathbb{W}} \|\mathbb{E}H(\boldsymbol{w}) - \mathbb{E}H_0(\boldsymbol{w})\|_\infty, \quad q_{\mathcal{A}_\eta} = \|\nabla Q^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta})\|_2, \quad \text{and} \quad q_* = q(r_*) + q_{\mathcal{A}_\eta}.$$

Following the proof of Theorem 11 in Tan et al. (2022), we can show that

$$\left|\langle \nabla \tilde{Q}(\boldsymbol{w}^{\mathcal{A}_\eta}), \tilde{\boldsymbol{z}} \rangle\right| \leq \left\{ \Pi(r_*) + \Pi_0(r_*) + \frac{\lambda_{\boldsymbol{w}}}{2} \right\} \|\tilde{\boldsymbol{z}}\|_1 + q_* \|\tilde{\boldsymbol{w}}\|_2, \tag{S4.2}$$

conditioning on the event $\mathcal{E}_0(r_*) \cap \mathcal{E}_{\boldsymbol{w}}(\lambda_{\boldsymbol{w}})$. Combing (S4.1) and (S4.2) and using similar arguments that lead to (S1.7), we arrive at

$$0 \leq 2\lambda_1 c_M \eta + (\frac{3}{2}\lambda_{\boldsymbol{w}}\sqrt{s} + q_*)\|\tilde{\boldsymbol{z}}\|_2 - \frac{1}{2}\lambda_{\boldsymbol{w}}\|\tilde{\boldsymbol{z}}_{\mathcal{S}^c}\|_1, \tag{S4.3}$$

which implies $\|\tilde{\boldsymbol{z}}\|_1 \leq (4\sqrt{s} + 2(\lambda_1)^{-1} q_*)\|\tilde{\boldsymbol{z}}\|_2 + 4c_M\eta \leq 5\sqrt{s}\|\tilde{\boldsymbol{z}}\|_2 + 4c_M\eta$, provided that $\lambda_1 = \lambda_{\boldsymbol{w}} + \rho$ with ρ being chosen such that

$$q_* s^{-\frac{1}{2}} \leq \frac{1}{2}\lambda_1 \quad \text{and} \quad \frac{1}{2}\lambda_{\boldsymbol{w}} + \Pi(r_*) + \Pi_0(r_*) \leq \frac{1}{2}\lambda_1. \tag{S4.4}$$

Consequently, we have $\tilde{\boldsymbol{w}} \in \mathbb{W}$.

The proof for the upper bound is similar to that used in Theorem 1. Define $\tilde{D}(\boldsymbol{\Delta}) = \langle \nabla \tilde{Q}(\boldsymbol{w}^{\mathcal{A}_\eta} + \boldsymbol{\Delta}) - \nabla \tilde{Q}(\boldsymbol{w}^{\mathcal{A}_\eta}), \boldsymbol{\Delta} \rangle$. We have that $\tilde{D}(\boldsymbol{\Delta}) = \langle \nabla \widehat{Q}^*_{h_*}(\boldsymbol{w}^{\mathcal{A}_\eta} + \boldsymbol{\Delta}) - \nabla \widehat{Q}^*_{h_*}(\boldsymbol{w}^{\mathcal{A}_\eta}), \boldsymbol{\Delta} \rangle$. With $(\widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\cdot), h_{\boldsymbol{w}}, n_{\mathcal{A}_\eta} + n_0)$ in Lemma S1.4 replaced by $(\widehat{Q}^*_{h_*}(\cdot), h_*, n_*)$, we can show that with probability $1 - c_1 \exp(-c_2 \log p)$ occurs the event

$$\mathcal{E}_*(r_*) = \left\{ \tilde{D}(\boldsymbol{\Delta}) \geq a_1 \left(\|\boldsymbol{\Delta}\|_2^2 - a_2 \sqrt{\frac{h_* \log p}{n_* r_*^2}} \|\boldsymbol{\Delta}\|_1 \|\boldsymbol{\Delta}\|_2 \right), \quad \text{for all} \quad \boldsymbol{\Delta} \in \mathbb{B}_2(r_*) \right\}.$$

We set $r_* = h_*/c_0$ with $c_0 = 32 v_1^2$. Further choose $t \in (0, 1)$ such that $t\|\tilde{\boldsymbol{z}}\|_2 \in \mathbb{B}_2(r_*)$. Let $\check{\boldsymbol{z}} = t\tilde{\boldsymbol{z}}$ and $\check{\boldsymbol{w}} = \boldsymbol{w}^{\mathcal{A}_\eta} + \check{\boldsymbol{z}}$. Then using the same

arguments that lead to (S4.3), we can show that

$$\tilde{D}(\check{\boldsymbol{z}}) \leq 2\lambda_1 c_M \eta + 2\lambda_{\boldsymbol{w}}\sqrt{s}\|\check{\boldsymbol{z}}\|_2,$$

as long as $2q_* s^{-1/2} \leq \lambda_1$.

Conditioning on $\mathcal{E}_*(r_*)$, which provides lower bound for $\tilde{D}(\check{\boldsymbol{z}})$, we obtain that

$$a_1\|\check{\boldsymbol{z}}\|_2^2 - a_1 a_2 c_0\sqrt{\frac{\log p}{n_* h_*}}\|\check{\boldsymbol{z}}\|_1\|\check{\boldsymbol{z}}\|_2 \leq 2\lambda_1 c_M\eta + 2\lambda_{\boldsymbol{w}}\sqrt{s}\|\check{\boldsymbol{z}}\|_2. \tag{S4.5}$$

Via proof by contradiction, we can deduce from (S4.5) that

$$\|\check{\boldsymbol{z}}\|_2 \leq 10 a_2 c_0 c_M\sqrt{\frac{\log p}{n_* h_*}}\eta + 4\lambda_1\sqrt{s} + 2\sqrt{\frac{\lambda_1 c_M \eta}{a_1}} =: c_u^*, \tag{S4.6}$$

by choosing

$$h_* \gtrsim \sqrt{\log p/(n_* h_*)}\eta + \lambda_1\sqrt{s} + \sqrt{\lambda_1\eta} \tag{S4.7}$$

such that $r_* > c_u^*$.

It suffices to choose a λ_1 large enough such that (S4.4) holds. To facilitate the proof, we present upper bounds for $\Pi_0(r), \Pi(r)$, and $q(r)$ in the following two lemmas, with their proofs relegated to Section S5.7 and S5.8.

Lemma S4.1. *Assume Conditions 1-S3.1 hold. For any $r > 0$ and $x > 0$, with probability $1 - e^{-x}$,*

$$\begin{aligned}\Pi(r) \leq & r\left(C_1\frac{1}{h_{\boldsymbol{w}}}\sqrt{\frac{2s\log(2p)}{n_{\mathcal{A}_\eta}+n_0}} + C_2\sqrt{\frac{\log(2p)+x}{(n_{\mathcal{A}_\eta}+n_0)h_{\boldsymbol{w}}}} + C_3\frac{\sqrt{s}(\log(2p)+x)}{(n_{\mathcal{A}_\eta}+n_0)h_{\boldsymbol{w}}}\right)\\ & + \eta\left(C_4\frac{1}{h_{\boldsymbol{w}}}\sqrt{\frac{2\log(2p)}{n_{\mathcal{A}_\eta}+n_0}} + C_5\frac{\log(2p)+x}{(n_{\mathcal{A}_\eta}+n_0)h_{\boldsymbol{w}}}\right),\end{aligned}$$

where $C_1 = 25k_uB^2$, $C_2 = (2k_uf_u\mu_4)^{1/2}$, $C_3 = 65k_uB^2/3$, $C_4 = 4C_1c_M/5$ *and* $C_5 = 4C_3c_M/5$. *The same upper bound holds for* $\Pi_0(r)$ *with* $(n_{\mathcal{A}_\eta} + n_0, h_{\boldsymbol{w}})$ *replaced by* (n_*, h_*).

Lemma S4.2. *Assume Conditions 1-S3.1 hold. For* $r > 0$, *we have* $q(r) \leq \mu_2 l_0 k_1 |h_* - h_{\boldsymbol{w}}| r$.

Lemma S4.1 implies that with probability at least $1 - 2/p$,

$$\begin{aligned}\Pi(r_*) + \Pi_0(r_*) \lesssim & r_*\sqrt{s}\left(\frac{1}{h_{\boldsymbol{w}}}\sqrt{\frac{\log p}{n_{\mathcal{A}_\eta} + n_0}} + \frac{1}{h_*}\sqrt{\frac{\log p}{n_*}}\right) \\ & + \eta\left(\frac{1}{h_{\boldsymbol{w}}}\sqrt{\frac{\log p}{n_{\mathcal{A}_\eta} + n_0}} + \frac{1}{h_*}\sqrt{\frac{\log p}{n_*}}\right),\end{aligned}$$

provided that $0 < h_{\boldsymbol{w}} \leq h_* \lesssim 1$.

Moreover, Lemma S1.2 and Lemma S4.2 together imply that $q_* \leq l_0(\mu_1 k_2 h_{\boldsymbol{w}}^2/2 + \mu_2 k_1 h_* r_*)$. Therefore, with $h_{\boldsymbol{w}}^{-1}\sqrt{\log p/(n_{\mathcal{A}_\eta} + n_0)} + h_*^{-1}\sqrt{\log p/n_*}$ denoted by ρ_*, we can choose

$$\rho \asymp \max\left\{s^{-1/2}(h_{\boldsymbol{w}}^2 + h_* r_*), r_*\sqrt{s}\rho_* + \eta\rho_*\right\},$$

such that (S4.4) holds with high probability. The claimed result follows directly by plugging the rate of λ_1 into (S4.6). □

S4.2 Proof of Theorem S3.1

Proof. The proof of Theorem S3.1 is by verifying the conditions of Proposition S3.1 and applying it repeatedly. We start with the first iterate $\tilde{\boldsymbol{w}}^{(1)}$.

Let $\lambda_1 = \lambda_{\boldsymbol{w}} + \rho_1$ with $\rho_1 \asymp \max\{s^{-1/2}(h_{\boldsymbol{w}}^2 + h_* r_*), r_*\sqrt{s}\rho_* + \eta\rho_*\}$. With bandwidths $h_* \asymp \sqrt{s}(\log p/n_*)^{1/4}$ and $h_{\boldsymbol{w}} \asymp (s\log p/(n_{\mathcal{A}_\eta} + n_0))^{1/4}$, we have

$$\rho_1 \asymp \max\left\{ s^{-\frac{1}{2}}\varphi_0 r_*, \sqrt{\frac{\log p}{n_{\mathcal{A}_\eta} + n_0}} + \left(\frac{\log p}{s n_*}\right)^{\frac{1}{4}} \eta \right\},$$

with

$$\varphi_0 \asymp \left(\frac{s^3 \log p}{(n_{\mathcal{A}_\eta} + n_0)}\right)^{\frac{1}{4}} + \left(\frac{s^2 \log p}{n_*}\right)^{\frac{1}{4}}.$$

Then we can verify that (S4.7) is satisfied provided that

$$r_* \lesssim \min\left\{1, \left(\frac{n_{\mathcal{A}_\eta} + n_0}{s n_*}\right)^{\frac{1}{4}}\right\}, \eta \lesssim \min\left\{s^{\frac{1}{4}}, \left(\frac{s^5 \log p}{n_*}\right)^{\frac{1}{8}}\right\}, r_*\eta \lesssim \left(\frac{s^4 \log p}{n_*}\right)^{\frac{1}{4}}. \tag{S4.8}$$

Examining the proof of Theorem 1, we know that with high probability $\tilde{\boldsymbol{w}}^{(0)} - \boldsymbol{w}^{\mathcal{A}_\eta} \in \mathbb{B}_2(r_*) \cap \mathbb{U}$ with $r_* \asymp \sqrt{s\log/n_*} + (\log p/n_*)^{1/4}\sqrt{\eta} + (\log p/n_*)^{3/8} s^{-1/8}\eta$. With the imposed conditions on η, we can verify that $r_* \gtrsim \sqrt{s\log p/n_*}$ and that (S4.8) is satisfied for the first iteration. Therefore, by Proposition S3.1 and the first part of Lemma S1.3, we obtain

$$\|\tilde{\boldsymbol{w}}^{(1)} - \boldsymbol{w}^{\mathcal{A}_\eta}\|_2 \le C_0\varphi_0 r_* + C_1(r_{\boldsymbol{w}} + h_{\boldsymbol{w}}^2) + C_2\varphi_1\sqrt{\eta} + C_3\varphi_2\eta =: r_1, \tag{S4.9}$$

where $r_{\boldsymbol{w}} = \lambda_{\boldsymbol{w}}\sqrt{s}$ and $(\varphi_0, \varphi_1, \varphi_2)$ are given in Theorem S3.1. Define $\gamma = C_0\varphi_0$. With sufficiently large sample sizes, i.e., $n_* \gtrsim s^2 \log p$ and $n_{\mathcal{A}_\eta} + n_0 \gtrsim s^3 \log p$, γ can be strictly less than 1. Consequently, $\tilde{\boldsymbol{w}}^{(1)}$ reduces the estimation error of $\tilde{\boldsymbol{w}}^{(0)}$ when the remaining terms are relatively small.

To move on to more iterations, we first note that

$$\varphi_1 \lesssim \left(\frac{s\log p}{n_*}\right)^{\frac{1}{8}}\sqrt{r_*}+\left(\frac{\log p}{n_{\mathcal{A}_\eta}+n_0}\right)^{\frac{1}{4}}, \quad \varphi_2 \lesssim \left(\frac{s\log p}{n_*}\right)^{\frac{1}{4}}+s^{-\frac{1}{4}}\left(\frac{\log p}{n_*}\right)^{\frac{3}{8}}.$$

For $t = 2, \ldots, T$, consider the event $\mathcal{E}_t(r_t) = \{\tilde{\boldsymbol{w}}^{(t)} - \boldsymbol{w}^{\mathcal{A}_\eta} \in \mathbb{B}_2(r_t) \cap \mathbb{U}\}$ with

$$r_t := \gamma^t r_* + \left\{C_1(r_{\boldsymbol{w}} + h_{\boldsymbol{w}}^2) + C_2\varphi_1\sqrt{\eta} + C_3\varphi_2\eta\right\}\frac{1-\gamma^t}{1-\gamma}.$$

For $t \geq 2$, we set $\lambda_t = \lambda_{\boldsymbol{w}} + \rho_t$ with

$$\begin{aligned}\rho_t &\asymp \max\left\{s^{-\frac{1}{2}}\varphi_0 r_{t_1}, \sqrt{\frac{\log p}{n_{\mathcal{A}_\eta}+n_0}} + \left(\frac{\log p}{sn_*}\right)^{\frac{1}{4}}\eta\right\} \\ &\asymp \max\left\{s^{-\frac{1}{2}}\varphi_0^t r_* + s^{-\frac{1}{2}}\varphi_0\left\{(r_{\boldsymbol{w}} + h_{\boldsymbol{w}}^2) + \varphi_1\sqrt{\eta} + \varphi_2\eta\right\},\right. \\ &\qquad \left.\sqrt{\frac{\log p}{n_{\mathcal{A}_\eta}+n_0}} + \left(\frac{\log p}{sn_*}\right)^{\frac{1}{4}}\eta\right\}.\end{aligned}$$

As long as (S4.8) is satisfied, it can be verified that

$$h_* \gtrsim \sqrt{\frac{\log p}{n_* h_*}}\eta + \lambda_t\sqrt{s} + \sqrt{\lambda_t}\eta, \quad \text{for all} \quad t \geq 2.$$

Then a repeated application of Proposition S3.1 gives us that, conditioning on $\mathcal{E}_{t-1}(r_{t-1}) \cap \mathcal{E}_{\boldsymbol{w}}(r_{\boldsymbol{w}})$, $\tilde{\boldsymbol{w}}^{(t)} - \boldsymbol{w}^{\mathcal{A}_\eta} \in \mathbb{U}$ and $\|\tilde{\boldsymbol{w}}^{(t)} - \boldsymbol{w}^{\mathcal{A}_\eta}\|_2 \leq r_t$ with probability at least $1 - c_1\exp(-c_2\log p)$.

Let the number of iterations T be chosen as $\lceil \log(r_*/r_{\boldsymbol{w}})/\log(1/\gamma)\rceil$ such that $\gamma^t r_* \leq r_{\boldsymbol{w}}$, we have

$$\|\tilde{\boldsymbol{w}}^{(T)} - \boldsymbol{w}^{\mathcal{A}_\eta}\|_2 \lesssim \lambda_{\boldsymbol{w}}\sqrt{s} + h_{\boldsymbol{w}}^2 + \left(\frac{s\log p}{n_*}\right)\sqrt{r_*\eta} + s^{-\frac{1}{4}}\left(\frac{\log p}{n_*}\right)^{\frac{3}{8}}\eta.$$

Plugging the rates of $\lambda_{\boldsymbol{w}}$ and r_* leads to

$$\|\tilde{\boldsymbol{w}}^{(T)} - \boldsymbol{w}^{\mathcal{A}_\eta}\|_2 \lesssim \sqrt{\frac{s\log p}{n_{\mathcal{A}_\eta} + n_0}} + \left(\frac{s\log p}{n_*}\right)^{\frac{3}{8}} \sqrt{\eta} + s^{-\frac{1}{4}} \left(\frac{\log p}{n_*}\right)^{\frac{3}{8}} \eta. \quad \text{(S4.10)}$$

This together with the fact $\tilde{\boldsymbol{w}}^{(t)} - \boldsymbol{w}^{\mathcal{A}_\eta} \in \mathbb{U}$ gives us

$$\|\tilde{\boldsymbol{w}}^{(T)} - \boldsymbol{w}^{\mathcal{A}_\eta}\|_1 \lesssim s\sqrt{\frac{\log p}{n_{\mathcal{A}_\eta} + n_0}} + s^{\frac{7}{8}} \left(\frac{\log p}{n_*}\right)^{\frac{3}{8}} \sqrt{\eta} + \eta. \quad \text{(S4.11)}$$

The claimed estimation error bounds for $\tilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}$ follows by combining (S4.10), (S4.11), (S1.16), and (S1.20). This completes the proof. □

S5 Proof of Auxiliary Lemmas

Here we provide the proofs of all the auxiliary lemmas mentioned in Section S1 and Section S4.

S5.1 Proof of Lemma S1.1

Proof. Since $\boldsymbol{\delta}^{\mathcal{A}_\eta} = \boldsymbol{M}_{\mathcal{A}_\eta}^{-1} \sum_{k\in\mathcal{A}_\eta\cup\{0\}} \alpha_k \boldsymbol{M}_k \boldsymbol{\delta}^{(k)}$, by Condition 4, we have

$$\|\boldsymbol{\delta}^{\mathcal{A}_\eta}\|_1 \leq \sum_{k\in\mathcal{A}_\eta\cup\{0\}} \alpha_k \|\boldsymbol{M}_{\mathcal{A}_\eta}^{-1}\boldsymbol{M}_k\|_1 \|\boldsymbol{\delta}^{(k)}\|_1 < c_M\eta.$$

□

S5.2 Proof of Lemma S1.2

Proof. Note that

$$\nabla Q_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}) = \sum_{k\in\mathcal{A}\cup\{0\}} \alpha_k \mathbb{E}\left\{\bar{K}\left(\frac{d^{(k)}-\epsilon^{(k)}}{h_{\boldsymbol{w}}}\right) - \tau\right\}\boldsymbol{x}^{(k)}$$

where $d^{(k)} = {\boldsymbol{x}^{(k)}}^\top(\boldsymbol{w}^{\mathcal{A}_\eta} - \boldsymbol{w}^{(k)})$. By integration by parts, we obtain

$$\begin{aligned}
&\mathbb{E}\left\{\bar{K}\left(\frac{d^{(k)}-\epsilon^{(k)}}{h_{\boldsymbol{w}}}\right)\middle|\boldsymbol{x}^{(k)}\right\}\\
=&\int_{-\infty}^{\infty} K(u) F_{\boldsymbol{\epsilon}^{(k)}|\boldsymbol{x}^{(k)}}(-h_{\boldsymbol{w}}u + d^{(k)})\mathrm{d}u\\
=&F_{\boldsymbol{\epsilon}^{(k)}|\boldsymbol{x}^{(k)}}(d^{(k)}) + \int_{-\infty}^{\infty} K(u)\int_{d^{(k)}}^{-h_{\boldsymbol{w}}u+d^{(k)}} (f_{\boldsymbol{\epsilon}^{(k)}|\boldsymbol{x}^{(k)}}(t) - f_{\boldsymbol{\epsilon}^{(k)}|\boldsymbol{x}^{(k)}}(0))\mathrm{d}t\mathrm{d}u,
\end{aligned}$$

which together with the Lipschitz condition on $f_{\boldsymbol{\epsilon}^{(k)}|\boldsymbol{x}^{(k)}}(\cdot)$ leads to

$$\left|\mathbb{E}\left\{\left|\bar{K}\left(\frac{d^{(k)}-\epsilon^{(k)}}{h_{\boldsymbol{w}}}\right)\right|\middle|\boldsymbol{x}^{(k)}\right\} - F_{\boldsymbol{\epsilon}^{(k)}|\boldsymbol{x}^{(k)}}(d^{(k)})\right| \le \frac{l_0}{2}\kappa_2 h_{\boldsymbol{w}}^2. \tag{S5.1}$$

The moment condition on $\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}|{(\boldsymbol{x}^{(k)})}^\top\boldsymbol{u}|$ and (S5.1) together imply that

$$\begin{aligned}
\|\nabla Q_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta})\|_2 &= \left\|\sum_{k\in\mathcal{A}\cup\{0\}} \alpha_k\mathbb{E}\left\{\bar{K}\left(\frac{d^{(k)}-\epsilon^{(k)}}{h_{\boldsymbol{w}}}\right) - F_{\boldsymbol{\epsilon}^{(k)}|\boldsymbol{x}^{(k)}}(d^{(k)})\right\}\boldsymbol{x}^{(k)}\right\|_2\\
&\le \frac{l_0\kappa_2 h_{\boldsymbol{w}}^2}{2}\left\|\sum_{k\in\mathcal{A}\cup\{0\}}\alpha_k\mathbb{E}\boldsymbol{x}^{(k)}\right\|_2\\
&\le \frac{l_0\kappa_2 h_{\boldsymbol{w}}^2}{2}\sup_{k\in\mathcal{A}\cup\{0\}}\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}\left|{(\boldsymbol{x}^{(k)})}^\top\boldsymbol{u}\right| \le \frac{l_0\mu_1\kappa_2 h_{\boldsymbol{w}}^2}{2},
\end{aligned} \tag{S5.2}$$

where the first equality follows from the definition of $\boldsymbol{w}^{\mathcal{A}_\eta}$ in (2.3).

The proof for $\|\nabla Q^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)})\|_2 \leq l_0\mu_1\kappa_2(h^{(k)})^2/2$, and $\|\nabla Q^{(0)}_{h_{\boldsymbol{\delta}}}(\boldsymbol{\beta})\|_2 \leq l_0\mu_1\kappa_2 h_{\boldsymbol{\delta}}^2/2$ is similar and thus omitted here. □

S5.3 Proof of Lemma S1.3

Proof. Let $\zeta^{(k)}_{i,1} = \bar{K}(((\boldsymbol{x}^{(k)}_i)^\top \boldsymbol{w}^{(k)} - y^{(k)}_i)/h_{\boldsymbol{w}}) - \tau$ and $\zeta^{(k)}_{i,2} = \bar{K}(((\boldsymbol{x}^{(k)}_i)^\top \boldsymbol{w}^{\mathcal{A}_\eta} - y^{(k)}_i)/h_{\boldsymbol{w}}) - \bar{K}(((\boldsymbol{x}^{(k)}_i)^\top \boldsymbol{w}^{(k)} - y^{(k)}_i)/h_{\boldsymbol{w}})$. Then we have

$$\begin{aligned}
&\nabla \widehat{Q}^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta}) - \nabla Q^{\mathcal{A}_\eta}_{h_{\boldsymbol{w}}}(\boldsymbol{w}^{\mathcal{A}_\eta}) \\
=&\frac{1}{n_{\mathcal{A}_\eta} + n_0} \sum_{k\in\mathcal{A}\cup\{0\}} \sum_{i=1}^{n_k} \left\{ \zeta^{(k)}_{i,1} \boldsymbol{x}^{(k)}_i - \mathbb{E}[\zeta^{(k)}_{i,1} \boldsymbol{x}^{(k)}_i] \right\} \\
&+ \frac{1}{n_{\mathcal{A}_\eta} + n_0} \sum_{k\in\mathcal{A}\cup\{0\}} \sum_{i=1}^{n_k} \left\{ \zeta^{(k)}_{i,2} \boldsymbol{x}^{(k)}_i - \mathbb{E}[\zeta^{(k)}_{i,2} \boldsymbol{x}^{(k)}_i] \right\}.
\end{aligned} \tag{S5.3}$$

We start by analyzing the second line in (S5.3). Note that since $|\zeta^{(k)}_{i,1}| \leq (1-\tau) \vee \tau$, $\{\{\zeta^{(k)}_{i,1}\}_{i=1}^{n_k}\}_{k\in\mathcal{A}\cup\{0\}}$ are independent sub-Gaussain random variables with parameter bounded by 1.

In addition, $\boldsymbol{x}^{(k)}$ is sub-Gaussian with $\mathbb{E}[(x^{(k)}_{ij})^2] \leq 4v_1^2$ for each $j = 1, \ldots, p$. For a mean zero sub-exponential random variable z with parameter (v, α), we have $\mathbb{E}(e^{tx}) \leq \exp(v^2t^2/2)$ for $|t| \leq 1/\alpha$. Therefore, $\{\{(x^{(k)}_{ij})^2 - \mathbb{E}[(x^{(k)}_{ij})^2]\}_{i=1}^{n_k}\}_{k\in\mathcal{A}\cup\{0\}}$ are sub-exponential variables with param-

eter $(256v_1^4, 16v_1^2)$. Applying Bernstein's inequality gives us

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{(n_{\mathcal{A}_\eta}+n_0)}\sum_{k\in\mathcal{A}\cup\{0\}}\left|\sum_{i=1}^{n_k}(x_{ij}^{(k)})^2-\mathbb{E}[(x_{ij}^{(k)})^2]\right|>t\right)\\
&\leq 2\exp\left\{-\frac{n_{\mathcal{A}_\eta}+n_0}{2}\min\left(\frac{t^2}{256v_1^4},\frac{t}{16v_1^2}\right)\right\}.
\end{aligned}$$

By choosing $t=32v_1^2\sqrt{\log p/(n_{\mathcal{A}_\eta}+n_0)}$ and the union bound, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{j\in[p]}\frac{1}{(n_{\mathcal{A}_\eta}+n_0)}\sum_{k\in\mathcal{A}\cup\{0\}}\left|\sum_{i=1}^{n_k}(x_{ij}^{(k)})^2-\mathbb{E}[(x_{ij}^{(k)})^2]\right|>32v_1^2\sqrt{\frac{\log p}{n_{\mathcal{A}_\eta}+n_0}}\right)\\
&\leq 2\exp(-\log p),
\end{aligned}$$

which implies the event $\mathcal{E}_x:=\left\{\max_{j\in[p]}1/(n_{\mathcal{A}_\eta}+n_0)\sum_{k\in\mathcal{A}\cup\{0\}}\sum_{i=1}^{n_k}(x_{ij}^{(k)})^2<c_x\right\}$ occurs with probability at least $1-2p^{-1}$ for some constant $c_x>4v_1^2+32v_1^2\sqrt{\log p/(n_{\mathcal{A}_\eta}+n_0)}$.

Therefore, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{n_{\mathcal{A}_\eta}+n_0}\sum_{k\in\mathcal{A}\cup\{0\}}\sum_{i=1}^{n_k}\left\{\zeta_{i,1}^{(k)}\boldsymbol{x}_i^{(k)}-\mathbb{E}[\zeta_{i,1}^{(k)}\boldsymbol{x}_i^{(k)}]\right\}\right\|_\infty>2\sqrt{\frac{c_x\log p}{n_{\mathcal{A}_\eta}+n_0}}\right)\\
&=\mathbb{P}\left(\left\|\frac{1}{n_{\mathcal{A}_\eta}+n_0}\sum_{k\in\mathcal{A}\cup\{0\}}\sum_{i=1}^{n_k}\left\{\zeta_{i,1}^{(k)}\boldsymbol{x}_i^{(k)}-\mathbb{E}[\zeta_{i,1}^{(k)}\boldsymbol{x}_i^{(k)}]\right\}\right\|_\infty>2\sqrt{\frac{c_x\log p}{n_{\mathcal{A}_\eta}+n_0}}\mid\mathcal{E}_x\right)+\mathbb{P}(\mathcal{E}_x^c)\\
&\leq p\max_{j\in[p]}\mathbb{P}\left(\left|\frac{1}{n_{\mathcal{A}_\eta}+n_0}\sum_{k\in\mathcal{A}\cup\{0\}}\sum_{i=1}^{n_k}\left\{\zeta_{i,1}^{(k)}x_{ij}^{(k)}-\mathbb{E}[\zeta_{i,1}^{(k)}x_{ij}^{(k)}]\right\}\right|>2\sqrt{\frac{c_x\log p}{n_{\mathcal{A}_\eta}+n_0}}\right)+\mathbb{P}(\mathcal{E}_x^c)\\
&\leq 2p\mathbb{E}_{\boldsymbol{X}}\left[\exp\left(-\frac{4c_x(n_{\mathcal{A}_\eta}+n_0)\log p}{\sum_{k\in\mathcal{A}\cup\{0\}}\sum_{i=1}^{n_k}\left(x_{ij}^{(k)}\right)^2}\right)\right]+\frac{2}{p}\leq\frac{4}{p}.
\end{aligned}$$

Here we use $\mathbb{E}_{\boldsymbol{X}}$ to denote the expectation with respect to $\{\boldsymbol{x}^{(k)}\}_{k\in\mathcal{A}\cup\{0\}}$.

It remains to bound the third line in (S5.3). By the mean value theorem,

$$\zeta_{i,2}^{(k)} x_{ij}^{(k)} = \frac{1}{h_{\boldsymbol{w}}} K\left(\frac{(\boldsymbol{x}_i^{(K)})^\top \boldsymbol{w}^{(k)} + v_i^{(k)} (\boldsymbol{x}_i^{(K)})^\top (\boldsymbol{w}_{\mathcal{A}} - \boldsymbol{w}^{(k)})}{h_{\boldsymbol{w}}}\right) x_{ij}^{(k)} (\boldsymbol{x}_i^{(K)})^\top (\boldsymbol{w}_{\mathcal{A}} - \boldsymbol{w}^{(k)}).$$

Under Condition 2(c) and Condition 3, we note that $\zeta_{i,2}^{(k)} x_{ij}^{(k)}$ is a $(c_M + 1)^2 \eta^2 M_K^2 v_1^2$-subexponential variable. By tail bounds of subexponential variables and union bounds, we have

$$\left\| \frac{1}{n_{\mathcal{A}_\eta} + n_0} \sum_{k \in \mathcal{A} \cup \{0\}} \sum_{i=1}^{n_k} \left\{ \zeta_{i,2}^{(k)} \boldsymbol{x}_i^{(k)} - \mathbb{E}[\zeta_{i,2}^{(k)} \boldsymbol{x}_i^{(k)}] \right\} \right\|_\infty > (c_M + 1) M_K v_1 \eta \sqrt{\frac{\log p}{n_{\mathcal{A}_\eta} + n_0}}$$

with probability less than $(n_{\mathcal{A}_\eta} + n_0)^{-1}$. This completes the proof of the first part by union bounds and choosing $C = 2 \max\{c_x, (c_M + 1) M_K v_1\}$.

The proof for the second and the third part is similar and thus omitted.

□

S5.4 Proof of Lemma S1.4

Proof. Define $d_i^{(k)} = (\boldsymbol{x}_i^{(k)})^\top (\boldsymbol{w}^{\mathcal{A}_\eta} - \boldsymbol{w}^{(k)})$. Note that

$$\begin{aligned} D(\boldsymbol{\Delta}) &= \frac{1}{n_{\mathcal{A}_\eta} + n_0} \sum_{k \in \mathcal{A}_\eta \cup \{0\}} \sum_{i=1}^{n_k} \left\{ \bar{K}\left(\frac{(\boldsymbol{x}_i^{(k)})^\top (\boldsymbol{w}^{\mathcal{A}_\eta} + \boldsymbol{\Delta}) - y_i^{(k)}}{h_{\boldsymbol{w}}}\right) \right. \\ &\qquad \left. - \bar{K}\left(\frac{d_i^{(k)} - \epsilon_i^{(k)}}{h_{\boldsymbol{w}}}\right) \right\} \boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(k)} \\ &\geq \frac{\kappa_l}{(n_{\mathcal{A}_\eta} + n_0) h_{\boldsymbol{w}}} \sum_{k \in \mathcal{A} \cup \{0\}} \sum_{i=1}^{n_k} (\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(k)})^2 \mathbf{I}\{E_i^{(k)}\}, \end{aligned}$$

where the event $E_i^{(k)}$ is defined as $E_i^{(k)} = \{|\epsilon_i^{(k)} - d_i^{(k)}| \leq h_{\boldsymbol{w}}/2\} \cap \{|\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(k)}| \leq \|\boldsymbol{\Delta}\|_2 h_{\boldsymbol{w}}/(2r)\}$ on which $|y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\boldsymbol{w}^{\mathcal{A}_\eta} + \boldsymbol{\Delta})| \leq h_{\boldsymbol{w}}$ for all $\boldsymbol{\Delta} \in \mathbb{B}_2(r)$.

The last inequality follows from the mean value theorem along with Condition 2.

Define the function

$$\phi_R(u) = \begin{cases} u^2, & \text{if } |u| \le \frac{R}{2}, \\ \{u - R\,\mathrm{sign}(u)\}^2, & \text{if } \frac{R}{2} \le |u| \le R, \\ 0, & \text{if } |u| > R, \end{cases} \tag{S5.4}$$

Note that $\phi_R(\cdot)$ is R-Lipschitz with homogeneity property $\phi_c(cu) = c^2\phi(u)$. We also note that

$$|u|\mathbf{I}\{R/2 \le |u| \le R\} \le u^2\mathbf{I}\{|u| \le R\}, \tag{S5.5}$$

which implies

$$\frac{D(\boldsymbol{\Delta})}{\|\boldsymbol{\Delta}\|_2^2} \ge \underbrace{\frac{1}{(n_{\mathcal{A}_\eta} + n_0)h_{\boldsymbol{w}}} \sum_{k \in \mathcal{A}\cup\{0\}} \sum_{i=1}^{n_k} \phi_{h_{\boldsymbol{w}}/(2r)}(\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(k)}/\|\boldsymbol{\Delta}\|_2)\chi_i^{(k)}}_{D_0(\boldsymbol{\Delta})}, \tag{S5.6}$$

with $\chi_i^{(k)}$ defined as $\mathbf{I}\{|\epsilon_i^{(k)} - d_i^{(k)}| \le h_{\boldsymbol{w}}/2\}$.

Under Condition 1, we have

$$\begin{aligned} &|\mathbb{E}(\chi_i^{(k)} \mid \boldsymbol{x}_i^{(k)}) - h_{\boldsymbol{w}} f_{\boldsymbol{\epsilon}^{(k)}|\boldsymbol{x}^{(k)}}(d_i^{(k)})| \\ \le &\int_{d_i^{(k)} - h_{\boldsymbol{w}}/2}^{d_i^{(k)} + h_{\boldsymbol{w}}/2} |f_{\boldsymbol{\epsilon}^{(k)}|\boldsymbol{x}^{(k)}}(t) - f_{\boldsymbol{\epsilon}^{(k)}|\boldsymbol{x}^{(k)}}(d_i^{(k)})| \le l_0 h_{\boldsymbol{w}}^2/4. \end{aligned}$$

Under Condition 5 and applying Hölder's inequality, we have $|d_i^{(k)}| < \|\boldsymbol{x}_i^{(k)}\|_\infty \|\boldsymbol{\delta}^{\mathcal{A}_\eta} - \boldsymbol{\delta}^{(k)}\|_1 < b_0$. Given that $h_{\boldsymbol{w}} \le f_l/(2l_0)$, we conclude that

$$\frac{7}{8} f_l h_{\boldsymbol{w}} \le \mathbb{E}(\chi_i^{(k)} \mid \boldsymbol{x}_i^{(k)}) \le \frac{9}{8} f_u h_{\boldsymbol{w}} \quad \text{almost surely.} \tag{S5.7}$$

Combining (S5.5)-(S5.7) leads to

$$\begin{aligned}&\mathbb{E}\left[\phi_{h_{\boldsymbol{w}}/(2r)}\left(\frac{\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(k)}}{\|\boldsymbol{\Delta}\|_2}\right)\chi_i^{(k)}\right]\\ &\geq \frac{7f_l h_{\boldsymbol{w}}}{8\|\boldsymbol{\Delta}\|_2^2}\mathbb{E}\left[(\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(k)})^2\mathbf{I}\left\{|\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(k)}|\leq \frac{h_{\boldsymbol{w}}\|\boldsymbol{\Delta}\|_2}{4r}\right\}\right].\end{aligned} \tag{S5.8}$$

Cauchy-Schwartz inequality and tail bounds for sub-Gaussians imply that

$$\begin{aligned}&\mathbb{E}\left[\frac{(\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(k)})^2}{\|\boldsymbol{\Delta}\|_2^2}\mathbf{I}\left\{|\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(k)}|\geq \frac{h_{\boldsymbol{w}}\|\boldsymbol{\Delta}\|_2}{4r}\right\}\right]\\ \leq&\frac{\sqrt{\mathbb{E}[(\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(k)})^4]}}{\|\boldsymbol{\Delta}\|_2^2}\sqrt{\mathbb{P}\left(|\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(k)}|\geq \frac{h_{\boldsymbol{w}}\|\boldsymbol{\Delta}\|_2}{4r}\right)}\\ \leq&4v_1^2\exp\left(-\frac{h_{\boldsymbol{w}}^2}{128r^2v_1^2}\right)\leq\frac{\gamma_p}{4},\end{aligned} \tag{S5.9}$$

as long as $r\leq h_{\boldsymbol{w}}/(32v_1^2)$.

Plugging (S5.8) and (S5.9) into $D_0(\boldsymbol{\Delta})$ gives us

$$\mathbb{E}[D_0(\boldsymbol{\Delta})]\geq\frac{21}{32}f_l\gamma_p,\quad\text{for all}\quad\boldsymbol{\Delta}\in\mathbb{B}_2(r), \tag{S5.10}$$

provided that $32v_1^2r\leq h_{\boldsymbol{w}}\leq f_l/(2l_0)$.

Now we bound the random fluctuation $D_0(\boldsymbol{\Delta})-\mathbb{E}[D_0(\boldsymbol{\Delta})]$ over $\boldsymbol{\Delta}\in\mathbb{B}_2(r)$. We first consider the set $B_l=\mathbb{B}_2(r)\cap\mathbb{C}(l)$. Further define $Z_l=\sup_{\boldsymbol{\Delta}\in B_l}|D_0(\boldsymbol{\Delta})-\mathbb{E}[D_0(\boldsymbol{\Delta})]|$. Following a similar argument in the proof of Proposition 4.2 in Tan et al. (2022), it can be shown that there exist positive constants C_1 and C_2 such that

$$Z_l\leq C_1v_1l\sqrt{\frac{f_uh_{\boldsymbol{w}}\log p}{(n_{\mathcal{A}_\eta}+n_0)r^2}}+C_2\sqrt{\frac{f_uh_{\boldsymbol{w}}\log p}{(n_{\mathcal{A}_\eta}+n_0)r^2}}, \tag{S5.11}$$

with probability at least $1-p^{-1}$, as long as $h_{\boldsymbol{w}} \gtrsim \log p/(n_{\mathcal{A}_\eta}+n_0)$. Combing (S5.6),(S5.10) and (S5.11), we can conclude that

$$\frac{D(\boldsymbol{\Delta})}{\|\boldsymbol{\Delta}\|_2^2} \geq \frac{1}{2}\kappa_l f_l \gamma_p - C_1 v_1 l \kappa_l \sqrt{\frac{f_u h_{\boldsymbol{w}} \log p}{(n_{\mathcal{A}_\eta}+n_0)r^2}}, \quad \text{for all} \quad \boldsymbol{\Delta} \in B_l, \quad \text{(S5.12)}$$

as long as $C_2\sqrt{f_u h_{\boldsymbol{w}} \log p/((n_{\mathcal{A}_\eta}+n_0)r^2)} \leq 5 f_l \gamma_p/32$.

Now we employ a peeling technique (van der Vaart and Wellner, 1996; Van de Geer, 2000) to extend the bound in (S5.12) to one that is uniform in $\|\boldsymbol{\Delta}\|_1/\|\boldsymbol{\Delta}\|_2$. Consider the event

$$\mathcal{E} = \left\{ \frac{D(\boldsymbol{\Delta})}{\|\boldsymbol{\Delta}\|_2^2} \geq \frac{1}{2}\kappa_l f_l \gamma_p - C_1 v_1 \kappa_l \sqrt{\frac{f_u h_{\boldsymbol{w}} \log p}{(n_{\mathcal{A}_\eta}+n_0)r^2}} \frac{\|\boldsymbol{\Delta}\|_1}{\|\boldsymbol{\Delta}\|_2}, \quad \text{for all} \quad \boldsymbol{\Delta} \in S(r) \right\},$$

where

$$S(r) = \{\boldsymbol{\Delta} \in \mathbb{B}_2(r) : \frac{\|\boldsymbol{\Delta}\|_1}{\|\boldsymbol{\Delta}\|_2} \leq \frac{f_l}{2C_1 v_1} \frac{(n_{\mathcal{A}_\eta}+n_0)r^2}{f_u h_{\boldsymbol{w}} \log p}^{1/2} \}.$$

Define $\mathcal{D}^{\mathcal{A}_\eta} = \{(\boldsymbol{X}^{(k)}, \boldsymbol{y}^{(k)})\}_{k\in\mathcal{A}_\eta}$ and the functions

$$f(\boldsymbol{\Delta}; \mathcal{D}^{\mathcal{A}_\eta}) = \frac{1}{2}\kappa_l f_l \gamma_p - \frac{D(\boldsymbol{\Delta})}{\|\boldsymbol{\Delta}\|_2^2}$$

along with

$$g(l) = C_1 v_1 l \kappa_l \sqrt{\frac{f_u h_{\boldsymbol{w}} \log p}{(n_{\mathcal{A}_\eta}+n_0)r^2}}, \quad \text{and} \quad h(\boldsymbol{\Delta}) = \frac{\|\boldsymbol{\Delta}\|_1}{\|\boldsymbol{\Delta}\|_2}.$$

The inequality (S5.12) implies that

$$\mathbb{P}\left(\sup_{\boldsymbol{\Delta}\in S(r), h(\boldsymbol{\Delta})\leq \ell} f(\boldsymbol{\Delta}; \mathcal{D}^{\mathcal{A}_\eta}) \geq g(\ell) \right) \leq \frac{1}{p}, \quad \text{for any} \quad \ell > 0.$$

Since $1 \leq h(\boldsymbol{\Delta}) \leq f_l/(2C_1v_1)\sqrt{(n_{\mathcal{A}_\eta}+n_0)r^2/(f_u h_{\boldsymbol{w}} \log p)}$, we have

$$g(h(\boldsymbol{\Delta})) \in \left[C_1 v_1 \kappa_l \sqrt{\frac{f_u h_{\boldsymbol{w}} \log p}{(n_{\mathcal{A}_\eta}+n_0)r^2}}, \frac{\kappa_l f_l \gamma_p}{2} \right]$$

over the region of interest. Define the set

$$V_m = \{\boldsymbol{\Delta} \mid 2^{m-1} b \leq g(h(\boldsymbol{\Delta})) \leq 2^m b\}, \quad m = 1, \ldots, M,$$

where $b = C_1 v_1 \kappa_l \sqrt{f_u h_{\boldsymbol{w}} \log p/((n_{\mathcal{A}_\eta}+n_0)r^2)}$ and M is taken as the smallest integer such that $2^M \geq f_l \gamma_p/(2C_1v_1)(n_{\mathcal{A}_\eta}+n_0)^{1/2} r/(f_u h_{\boldsymbol{w}} \log p)^{1/2}$. Since $32v_1^2 r \leq h_{\boldsymbol{w}} \leq f_l/(2l_0)$, we can take $M = \lceil \log\{c\sqrt{(n_{\mathcal{A}_\eta}+n_0)/\log p}\} \rceil$.

By a union bound, there exist some positive constants c_1 and c_2 such that

$$\begin{aligned} \mathbb{P}(\mathcal{E}^c) &\leq \sum_{m=1}^{M} \mathbb{P}(\exists \boldsymbol{\Delta} \in V_m, s.t. f(\boldsymbol{\Delta}; \mathcal{D}^{\mathcal{A}_\eta}) \geq 2g(h(\boldsymbol{\Delta}))) \\ &\leq \sum_{m=1}^{M} \mathbb{P}\left(\sup_{h(\boldsymbol{\Delta}) \leq g^{-1}(2^m b)} f(\boldsymbol{\Delta}; \mathcal{D}^{\mathcal{A}_\eta}) \geq 2^m b \right) \\ &\leq M/p \leq c_1 \exp(-c_2 \log p). \end{aligned} \tag{S5.13}$$

This proves the claimed result with by taking $a_1 = \frac{1}{2}\kappa_l f_l \gamma_p$ and $a_2 = 2C_1 v_1/(f_l \gamma_p)$. □

S5.5 Proof of Lemma S1.5

Proof. Lemma S1.5 is a special case of Lemma S1.4 and hence we omit its proof here. □

S5.6 Proof of Lemma S1.6

Proof. Define $\hat{\boldsymbol{u}}^{(k)} = \hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)}$ and $q^{(k)} = \|\nabla Q^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)})\|_2$. By Lemma S1.2, we have $q^{(k)} \le l_0\mu_1\kappa_2(h^{(k)})^2/2$. For $k \in \mathcal{A}$, consider the event $\mathcal{E}_k = \{2\|\nabla \widehat{Q}^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)}) - \nabla Q^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)})\|_\infty \le \lambda^{(k)}\}$. Now following the same arguments used in (S1.6) and (S1.7), we obtain that

$$\begin{aligned}
0 \le& \langle \nabla \widehat{Q}^{(k)}_{h^{(k)}}(\hat{\boldsymbol{\beta}}^{(k)}) - \nabla \widehat{Q}^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)}), \hat{\boldsymbol{u}}^{(k)}\rangle \\
=& \langle -\lambda^{(k)}\mathrm{sgn}(\hat{\boldsymbol{\beta}}^{(k)}), \hat{\boldsymbol{u}}^{(k)}\rangle - \langle \nabla \widehat{Q}^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)}) - \nabla Q^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)}), \hat{\boldsymbol{u}}^{(k)}\rangle \\
& - \langle \nabla Q^{(k)}_{h^{(k)}}(\boldsymbol{w}^{(k)}), \hat{\boldsymbol{u}}^{(k)}\rangle \\
\le& \lambda^{(k)}\|\boldsymbol{\beta}^{(k)}\|_1 - \lambda^{(k)}\|\hat{\boldsymbol{\beta}}^{(k)}\|_1 + \frac{1}{2}\lambda^{(k)}\|\hat{\boldsymbol{u}}^{(k)}\|_1 + q^{(k)}\|\hat{\boldsymbol{u}}^{(k)}\|_2 \\
\le& 2\lambda^{(k)}\|\boldsymbol{\beta}^{(k)}_{\mathcal{S}_k^c}\|_1 + \frac{3}{2}\lambda^{(k)}\|\hat{\boldsymbol{u}}^{(k)}_{\mathcal{S}_k}\|_1 - \frac{1}{2}\lambda^{(k)}\|\hat{\boldsymbol{u}}^{(k)}_{\mathcal{S}_k^c}\|_1 + q^{(k)}\|\hat{\boldsymbol{u}}^{(k)}\|_2.
\end{aligned}$$

Provided that $(h^{(k)})^2 \le \lambda^{(k)}\sqrt{s}/(l_0\mu_1\kappa_2)$ and together with Condition 6, this leads to the cone-like constraint for $\hat{\boldsymbol{u}}^{(k)}$:

$$\begin{aligned}
\|\hat{\boldsymbol{u}}^{(k)}_{\mathcal{S}_k^c}\|_1 &\le 3\|\hat{\boldsymbol{u}}^{(k)}_{\mathcal{S}_k}\|_1 + 2(\lambda^{(k)})^{-1}q^{(k)}\|\hat{\boldsymbol{u}}^{(k)}\|_2 + 4\|\boldsymbol{\beta}^{(k)}_{\mathcal{S}_k^c}\|_1 \\
&\le 4\sqrt{s'}\|\hat{\boldsymbol{u}}^{(k)}\|_2 + 4\eta',
\end{aligned} \tag{S5.14}$$

from which it follows that

$$\|\hat{\boldsymbol{u}}^{(k)}\|_1 \le 5\sqrt{s}\|\hat{\boldsymbol{u}}^{(k)}\|_2 + 4\eta'. \tag{S5.15}$$

Next, we provide a lower bound for $D^{(k)}(\boldsymbol{\Delta}) := \langle \nabla \widehat{Q}^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)} + \boldsymbol{\Delta}) - \nabla \widehat{Q}^{(k)}_{h^{(k)}}(\boldsymbol{\beta}^{(k)}), \boldsymbol{\Delta}\rangle$. Consider the event

$$\mathcal{E}'_k(r^{(k)}) = \left\{ \frac{D^{(k)}(\boldsymbol{\Delta})}{\|\boldsymbol{\Delta}\|_2^2} \ge a_1\left(1 - a_2\sqrt{\frac{h^{(k)}\log p}{(n_k+n_0)(r^{(k)})^2}}\frac{\|\boldsymbol{\Delta}\|_1}{\|\boldsymbol{\Delta}\|_2}\right), \quad \text{for all} \quad \boldsymbol{\Delta} \in \mathbb{B}_2(r^{(k)}) \right\}.$$

We set $r^{(k)} = h^{(k)}/c_0$ with $c_0 = 32v_1^2$.

Conditioning on $\mathcal{E}_k \cap \mathcal{E}_k'(r^{(k)})$ and using the same techinique for deriving (S1.9), we can deduce from (S5.15) that

$$\|\hat{\boldsymbol{u}}^{(k)}\|_2 \le 8a_2c_0\sqrt{\frac{\log p}{(n_k+n_0)h^{(k)}}}\eta' + 4\lambda^{(k)}\sqrt{s'} + 2\sqrt{\frac{\lambda^{(k)}\eta'}{a_1}} =: c_u^{(k)},$$

by choosing $h^{(k)} \gtrsim \sqrt{\log p/((n_k+n_0)h^{(k)})}\eta' + \lambda^{(k)}\sqrt{s'} + \sqrt{\lambda^{(k)}\eta'}$ such that $r^{(k)} > c_u^{(k)}$. With the stated choices $\lambda^{(k)} \asymp \sqrt{\log p/(n_k+n_0)}$ and $h^{(k)} \asymp \sqrt{\lambda^{(k)}}(s')^{1/4}$, as long as $\eta' \lesssim \sqrt{s'}$, we have

$$\|\hat{\boldsymbol{u}}^{(k)}\|_2 \lesssim \sqrt{\frac{s'\log p}{n_k+n_0}} + \left(\frac{\log p}{n_k+n_0}\right)^{1/4}\sqrt{\eta'} + \left(\frac{\log p}{n_k+n_0}\right)^{\frac{3}{8}} s^{-\frac{1}{8}}\eta' = \Omega_k. \tag{S5.16}$$

It remains to bound the probability of the event $\mathcal{E}_k \cap \mathcal{E}_k'(r^{(k)})$ when $\lambda^{(k)} \asymp \sqrt{\log p/(n_k+n_0)}$, which follows from Lemma S1.3 and Corollary S1.1. □

S5.7 Proof of Lemma S4.1

Proof. Recall that $d_i^{(k)} = (\boldsymbol{x}_i^{(k)})^\top(\boldsymbol{w}^{\mathcal{A}_\eta} - \boldsymbol{w}^{(k)})$, $H_0(\boldsymbol{w}) = \nabla\widehat{Q}_{h_*}^*(\boldsymbol{w}) - \nabla\widehat{Q}_{h_*}^*(\boldsymbol{w}^{\mathcal{A}_\eta})$, and $H(\boldsymbol{w}) = \nabla\widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}) - \nabla\widehat{Q}_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta})$. To facilitate proof, for $\boldsymbol{z} \in \mathbb{R}^p$, define

$$V_{ij}^{(k)}(\boldsymbol{z}) := \left\{\bar{K}\left(\frac{(\boldsymbol{x}_i^{(k)})^\top\boldsymbol{z} + d_i^{(k)} - \epsilon_i^{(k)}}{h_{\boldsymbol{w}}}\right) - \bar{K}\left(\frac{d_i^{(k)} - \epsilon_i^{(k)}}{h_{\boldsymbol{w}}}\right)\right\} x_{ij}^{(k)}$$

and $V_j := \sup_{\boldsymbol{z}\in\Theta_{\boldsymbol{z}}(r_1,r_2)} |1/(n_{\mathcal{A}_\eta}+n_0)\sum_{k\in\mathcal{A}_\eta\cup\{0\}}\sum_{i=1}^{n_k} V_{ij}^{(k)}(\boldsymbol{z}) - \mathbb{E}V_{ij}^{(k)}(\boldsymbol{z})|$ where $\Theta_{\boldsymbol{z}}(r_1,r_2) = \{\boldsymbol{z}\in\mathbb{R}^p : \|\boldsymbol{z}\|_1 \le r_1, \|\boldsymbol{z}\|_2 \le r_2\}$.

With the notations above, we have

$$\sup_{\boldsymbol{w}\in\Theta_{\boldsymbol{z}}(r_1,r_2)} \|H(\boldsymbol{w}) - \mathbb{E}H(\boldsymbol{w})\|_\infty = \max_{j\in[p]} V_j \tag{S5.17}$$

By similar arguments as used in the proof of Lemma 19 in Tan et al. (2022), it can be shown that

$$\sup_{\boldsymbol{z}\in\Theta_{\boldsymbol{z}}(r_1,r_2)} \left|V_{ij}^{(k)}(\boldsymbol{z})\right| \le k_u B^2 r_1/h_{\boldsymbol{w}}, \tag{S5.18}$$

$$\mathbb{E}\left[V_{ij}^{(k)}\right]^2 \le k_u f_u h_{\boldsymbol{w}}^{-1}\left[\mathbb{E}(x_{ij}^{(k)})^4\right]^{\frac{1}{2}}\left[\mathbb{E}\langle \boldsymbol{x}_i^{(k)}, \boldsymbol{z}\rangle^4\right]^{\frac{1}{2}} \le k_u f_u \mu_4 h_{\boldsymbol{w}}^{-1} r_2, \tag{S5.19}$$

for all $\boldsymbol{z}\in\Theta_{\boldsymbol{z}}(r_1,r_2)$ and $k\in\mathcal{A}\cup\{0\}$, and also

$$\mathbb{E}V_j \le 4k_u B^2 r_1/h_{\boldsymbol{w}}\sqrt{2\log(2p)/N}, \quad \text{for all} \quad j\in[p]. \tag{S5.20}$$

By applying Bousquet's version of Talagrand's inequality Bousquet (2003), we obtain that for any $t>0$,

$$V_j \le \frac{5}{4}\mathbb{E}V_j + \sup_{\boldsymbol{z}\in\Theta_{\boldsymbol{z}}(r_1,r_2)}\left[\mathbb{E}\left\{V_{ij}^{(k)}(\boldsymbol{z})\right\}^2\right]^{\frac{1}{2}}\sqrt{\frac{2t}{n_{\mathcal{A}_\eta}+n_0}} + \frac{13}{3}k_u B^2\frac{r_1 t}{(n_{\mathcal{A}_\eta}+n_0)h_{\boldsymbol{w}}}, \tag{S5.21}$$

with probability at least $1-2e^{-t}$. Finally, the claimed result follows by combining (S5.17)-(S5.21) and taking $t=\log(2p)/x$, $r_2=r$, and $r_1 = 5\sqrt{s}r + 4c_M\eta$. □

S5.8 Proof of Lemma S4.2

Proof. The Hessian matrix of $Q_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w})$ is given by

$$\nabla^2 Q_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}) = \sum_{k\in\mathcal{A}\cup\{0\}} \alpha_k \mathbb{E}\left\{ \frac{1}{h_{\boldsymbol{w}}} K\left(\frac{(\boldsymbol{x}^{(k)})^\top \boldsymbol{w}^{\mathcal{A}_\eta} - y^{(k)}}{h_{\boldsymbol{w}}} \right) \boldsymbol{x}^{(k)} (\boldsymbol{x}^{(k)})^\top \right\}.$$

Define $\boldsymbol{z} = \boldsymbol{w} - \boldsymbol{w}^{\mathcal{A}_\eta}$. Recall that $d^{(k)} = (\boldsymbol{x}^{(k)})^\top (\boldsymbol{w}^{\mathcal{A}_\eta} - \boldsymbol{w}^{(k)})$.By the mean value theorem, we have

$$\begin{aligned}
\mathbb{E}H(\boldsymbol{w}) &= \int_0^1 \nabla^2 Q_{h_{\boldsymbol{w}}}^{\mathcal{A}_\eta}(\boldsymbol{w}^{\mathcal{A}_\eta} + t\boldsymbol{z}) \mathrm{d}t (\boldsymbol{w} - \boldsymbol{w}^{\mathcal{A}_\eta}) \\
&= \int_0^1 \sum_{k\in\mathcal{A}_\eta\cup\{0\}} \alpha_k \mathbb{E}\left\{ \frac{1}{h_{\boldsymbol{w}}} K\left(\frac{(\boldsymbol{x}^{(k)})^\top (\boldsymbol{w}^{\mathcal{A}_\eta} + t\boldsymbol{z}) - y^{(k)}}{h_{\boldsymbol{w}}} \right) \boldsymbol{x}^{(k)} (\boldsymbol{x}^{(k)})^\top \right\} \mathrm{d}t \boldsymbol{z} \\
&= \sum_{k\in\mathcal{A}_\eta\cup\{0\}} \alpha_k \mathbb{E}\left\{ \int_0^1 \int_{-\infty}^{\infty} K(u) f_{\epsilon^{(k)}|\boldsymbol{x}^{(k)}}(t\boldsymbol{z} + d^{(k)} - h_{\boldsymbol{w}} u) \mathrm{d}u \mathrm{d}t \boldsymbol{x}^{(k)} (\boldsymbol{x}^{(k)})^\top \right\} \boldsymbol{z}.
\end{aligned}$$

Similarly, we can show that

$$\mathbb{E}H_0(\boldsymbol{w}) = \sum_{k\in\mathcal{A}_\eta\cup\{0\}} \alpha_k \mathbb{E}\left\{ \int_0^1 \int_{-\infty}^{\infty} K(u) f_{\epsilon^{(k)}|\boldsymbol{x}^{(k)}}(t\boldsymbol{z} + d^{(k)} - h_* u) \mathrm{d}u \mathrm{d}t \boldsymbol{x}^{(k)} (\boldsymbol{x}^{(k)})^\top \right\} \boldsymbol{z}.$$

The above two equalities and the lipschitz continuity of $f_{\epsilon^{(k)}|\boldsymbol{x}^{(k)}}(\cdot)$ imply

$$\begin{aligned}
&\|\mathbb{E}H(\boldsymbol{w}) - \mathbb{E}H_0(\boldsymbol{w})\|_2 \\
\leq& \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \sum_{k\in\mathcal{A}_\eta\cup\{0\}} \alpha_k \mathbb{E}\Big\{ \int_0^1 \int_{-\infty}^{\infty} K(u) \Big| f_{\epsilon^{(k)}|\boldsymbol{x}^{(k)}}(t\boldsymbol{z} + d^{(k)} - h_{\boldsymbol{w}} u) \\
&\quad - f_{\epsilon^{(k)}|\boldsymbol{x}^{(k)}}(t\boldsymbol{z} + d^{(k)} - h_* u) \Big| \mathrm{d}u \mathrm{d}t \Big| (\boldsymbol{x}^{(k)})^\top \boldsymbol{z} \Big| \Big| (\boldsymbol{x}^{(k)})^\top \boldsymbol{u} \Big| \Big\} \\
\leq& l_0 \int_{-\infty}^{\infty} |u| K(u) \mathrm{d}u |h_{\boldsymbol{w}} - h_*| \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\{((\boldsymbol{x}^{(k)})^\top \boldsymbol{u})\}^2 \|\boldsymbol{z}\|_2 \\
\leq& l_0 \kappa_1 |h_{\boldsymbol{w}} - h_*| \mu_2 r_2.
\end{aligned}$$

This completes the proof. □

References

Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic inequalities and applications*, pp. 213–247. Springer.

Fan, J., Liu, H., Sun, Q., Zhang, T. (2018). I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Annals of statistics 46*(2), 814.

Jordan, M. I., Lee, J. D., Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association 114*(526), 668–681.

Li, S., Cai, T. T., Li, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 84*(1), 149–173.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics 45*(2), 866–896.

Loh, P.-L., Wainwright, M. J. (2013). Regularized m-estimators with non-

convexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learing Research 16*, 559–616.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science 27*(4), 538–557.

Shamir, O., Srebro, N., Zhang, T. (2014). Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pp. 1000–1008. PMLR.

Tan, K. M., Battey, H., Zhou, W.-X. (2022). Communication-constrained distributed quantile regression with optimal statistical guarantees. *Journal of Machine Learning Research 23*, 1–61.

Tan, K. M., Wang, L., Zhou, W. (2022). High-dimensional quantile regression: Convolution smoothing and concave regularization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 84*(1), 205–233.

Tian, Y., Feng, Y. (2022). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association* (to appear).

Van de Geer, S. A. (2000). *Empirical Processes in M-estimation*, Volume 6. Cambridge university press.

van der Vaart, A. W., Wellner, J. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Wang, J., Kolar, M., Srebro, N., Zhang, T. (2017). Efficient distributed learning with sparsity. In *International conference on machine learning*, pp. 3636–3645. PMLR.