Supplementary Material for "Multitask Smoothed Support Vector Machines for High-Dimensional Heterogeneous Data"

Ling Peng[1], Yue Wang[2], Xiaohui Liu[1] and Heng Lian[3,4]

[1] *Jiangxi University of Finance and Economics*

[2] *University of Science and Technology of China*

[3] *City University of Hong Kong Shenzhen Research Institute*

[4] *City University of Hong Kong*

In this supplement, we first give some specific examples of smoothed hinge loss in Section S1, present the proximal-gradient algorithm in detail in Section S2, and then present the proofs for the main results in Section S4. For $m \geq 2$, let $\mathbb{S}^{m-1}$ denote the m-dimensional unit sphere. Denote P and P_n the expectation under the population measure and its empirical counterpart, respectively. That is, $Pf = \mathbb{E}f(x)$ and $P_n f = (1/n)\sum_{i=1}^{n} f(x_i)$. Throughout the proof, C denotes a generic positive constant whose value may change even on the same line.

S1 Specific Examples of Smoothed Hinge Loss

We give specific expressions for the smoothed hinge loss under various kernel functions in Table 1, and present the curve of $\phi_h(\cdot)$ and its first-order derivative $\phi_h'(\cdot)$ with Gaussian, Logistic, and Laplacian kernels as examples in Figure 1.

Table 1: Kernel functions and the corresponding smoothed hinge losses.

Kernel type	Kernel function $K(u)$	Smoothed hinge loss $\phi_h(u)$
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$	$(1-u)\Phi\left(\frac{1-u}{h}\right)+\frac{h}{\sqrt{2\pi}}e^{-\frac{(1-u)^2}{2h^2}}$
Logistic	$\frac{e^{-u}}{(1+e^{-u})^2}$	$-u+h\log(e^{1/h}+e^{u/h})$
Laplacian	$\frac{1}{2}e^{-\lvert u\rvert}$	$\begin{cases}1+\frac{h}{2}e^{(u-1)/h}-u, & u<1\\ \frac{h}{2}e^{(1-u)/h}, & u\geq 1\end{cases}$
Epanechnikov	$\frac{3}{4}(1-u^2)I\{-1\leq u\leq 1\}$	$\begin{cases}1-u, & u\leq 1-h\\ \frac{(1-u+h)^3(3h-1+u)}{16h^3}, & 1-h<u\leq 1+h\\ 0, & u\geq 1+h\end{cases}$

Here $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

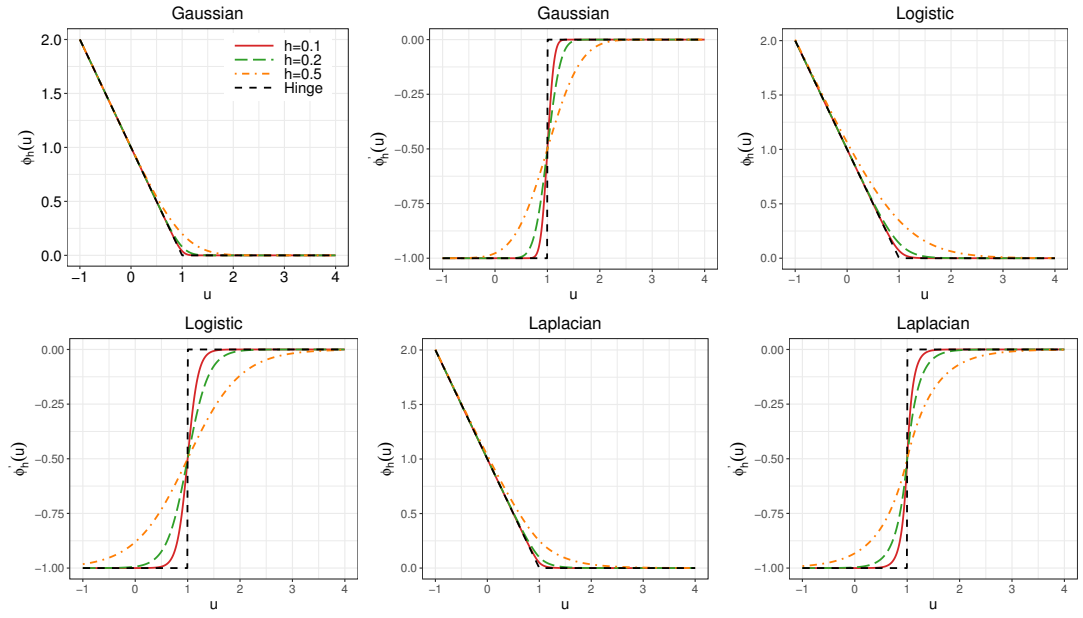


Figure 1: The smoothed hinge loss function $\phi_h(\cdot)$ and its first-order derivative $\phi_h'(\cdot)$ with Gaussian, Logistic, and Laplacian kernels under different bandwidths h. Note that the smoothed version approaches the hinge loss as $h\to 0$.

S2 Implementation via Projected Proximal Gradient Algorithm

We next describe the computational procedure for the proposed multitask smoothed SVM estimator. Recall that

$$
\begin{aligned}
&(\{\widehat{b}_h^{(m)}\}_{m=1}^M, \widehat{\boldsymbol{\mu}}_h, \{\widehat{\boldsymbol{\alpha}}_h^{(m)}\}_{m=1}^M) \\
&:= \underset{\substack{\{b^{(m)}\}_{m=1}^M, \boldsymbol{\mu}, \{\boldsymbol{\alpha}^{(m)}\}_{m=1}^M \\ \mathbf{1}^\top \boldsymbol{\alpha_j}=0, j=1,\ldots,p}}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \phi_h \left(r_i^{(m)}(b^{(m)}, \boldsymbol{\mu}, \boldsymbol{\alpha}^{(m)}) \right) + \lambda \Big(\|\boldsymbol{\mu}\|_1 + \lambda_g \|\mathbf{A}\|_{2,1} \Big) \right\},
\end{aligned}
$$

where $\mathbf{b} := (b^{(1)}, \ldots, b^{(M)})^\top \in \mathbb{R}^M$, and $\mathbf{A} := (\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(M)}) \in \mathbb{R}^{p \times M}$ with $\boldsymbol{\alpha}_j = (\alpha_j^{(1)}, \ldots, \alpha_j^{(M)})^\top$ collecting the j-th coordinates of $\{\boldsymbol{\alpha}^{(m)}\}_{m=1}^M$. For any $\boldsymbol{\Theta} := (\mathbf{b}, \boldsymbol{\mu}, \mathbf{A})$, define the task-specific loss

$$
\ell_h^{(m)}(\boldsymbol{\Theta}; \mathbf{x}, y) := \phi_h \left(y(\mathbf{x}^\top (\boldsymbol{\mu} + \boldsymbol{\alpha}^{(m)}) + b^{(m)}) \right),
$$

The pooled empirical loss is then

$$
\mathcal{L}_{N,h}(\boldsymbol{\Theta}) = \frac{1}{N} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \ell_h^{(m)}(\boldsymbol{\Theta}; \mathbf{x}_i^{(m)}, y_i^{(m)}),
$$

Thus, the objective consists of a smooth convex loss plus a nonsmooth composite penalty, suggesting a projected proximal-gradient algorithm. Recall that $\boldsymbol{\theta}^{(m)} :=$

$(b^{(m)}, (\boldsymbol{\mu}+\boldsymbol{\alpha}^{(m)})^\top)^\top$,

$$\begin{aligned}\widehat{\mathbf{s}}_{h,1}^{(m)}(\boldsymbol{\theta}^{(m)}) &= -\frac{1}{n_m}\sum_{i=1}^{n_m}\bar{K}\left(\frac{1-y_i^{(m)}(\widetilde{\mathbf{x}}_i^{(m)})^\top\boldsymbol{\theta}^{(m)}}{h}\right)y_i^{(m)} \quad \text{and}\\ \widehat{\mathbf{s}}_{h,-1}^{(m)}(\boldsymbol{\theta}^{(m)}) &= -\frac{1}{n_m}\sum_{i=1}^{n_m}\bar{K}\left(\frac{1-y_i^{(m)}(\widetilde{\mathbf{x}}_i^{(m)})^\top\boldsymbol{\theta}^{(m)}}{h}\right)y_i^{(m)}\mathbf{x}_i^{(m)}.\end{aligned}$$

Then, the gradient with respect to the task-specific intercept vector $\mathbf{b} = (b^{(1)}, \ldots, b^{(M)})^\top$ is given by

$$\nabla_{\mathbf{b}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}) = \left(\frac{n_1}{N}\widehat{\mathbf{s}}_{h,1}^{(1)}(\boldsymbol{\theta}^{(1)}), \ldots, \frac{n_M}{N}\widehat{\mathbf{s}}_{h,1}^{(M)}(\boldsymbol{\theta}^{(M)})\right)^\top.$$

The gradient with respect to the shared component $\boldsymbol{\mu}$ is

$$\nabla_{\boldsymbol{\mu}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}) = \frac{1}{N}\sum_{m=1}^{M} n_m\widehat{\mathbf{s}}_{h,-1}^{(m)}(\boldsymbol{\theta}^{(m)}).$$

The gradient with respect to the deviation matrix $\mathbf{A} = (\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(M)})$ is obtained columnwise:

$$\nabla_{\mathbf{A}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}) = \left(\frac{n_1}{N}\widehat{\mathbf{s}}_{h,-1}^{(1)}(\boldsymbol{\theta}^{(1)}), \ldots, \frac{n_M}{N}\widehat{\mathbf{s}}_{h,-1}^{(M)}(\boldsymbol{\theta}^{(M)})\right),$$

with its jth row being

$$\nabla_{\mathbf{A}_{j\cdot}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}) = \left(\frac{n_1}{N}\{\widehat{\mathbf{s}}_{h,-1}^{(1)}(\boldsymbol{\theta}^{(1)})\}_j, \ldots, \frac{n_M}{N}\{\widehat{\mathbf{s}}_{h,-1}^{(M)}(\boldsymbol{\theta}^{(M)})\}_j\right).$$

We now specify the proximal maps used in the algorithm. For $\tau > 0$, let

$$S_\tau(u) = \mathrm{sign}(u)(|u| - \tau)_+$$

denote the scalar soft-thresholding operator, applied componentwise to vectors. Define the centered subspace $\mathcal{C}_M = \left\{\mathbf{a} \in \mathbb{R}^M : \mathbf{1}^\top \mathbf{a} = 0\right\}$, and let

$$\Pi_{\mathcal{C}_M}(\mathbf{z}) = \mathbf{z} - \frac{1}{M}(\mathbf{1}^\top \mathbf{z})\mathbf{1}$$

be the orthogonal projection onto $\mathcal{C}_M$. The proximal map associated with the group penalty and the centering constraint is

$$\mathcal{G}_\tau(\mathbf{z}) = \left(1 - \frac{\tau}{\|\Pi_{\mathcal{C}_M}(\mathbf{z})\|_2}\right)_+ \Pi_{\mathcal{C}_M}(\mathbf{z}),$$

with $\mathcal{G}_\tau(\mathbf{z}) = \mathbf{0}$ when $\|\Pi_{\mathcal{C}_M}(\mathbf{z})\|_2 = 0$.

The algorithm proceeds as follows and we summarize it in Algorithm 1. Given the current iterate $\mathbf{\Theta}^t = (\mathbf{b}^t, \boldsymbol{\mu}^t, \mathbf{A}^t)$ and a step size $\eta > 0$, we first compute

$$\widetilde{\mathbf{b}}^{t+1} = \mathbf{b}^t - \eta \nabla_{\mathbf{b}} \mathcal{L}_{N,h}(\mathbf{\Theta}^t),$$

$$\widetilde{\boldsymbol{\mu}}^{t+1} = \boldsymbol{\mu}^t - \eta \nabla_{\boldsymbol{\mu}} \mathcal{L}_{N,h}(\mathbf{\Theta}^t),$$

$$\widetilde{\mathbf{A}}^{t+1} = \mathbf{A}^t - \eta \nabla_{\mathbf{A}} \mathcal{L}_{N,h}(\mathbf{\Theta}^t).$$

The proximal updates are then

$$\mathbf{b}^{t+1} = \widetilde{\mathbf{b}}^{t+1},$$

$$\boldsymbol{\mu}^{t+1} = S_{\eta\lambda}(\widetilde{\boldsymbol{\mu}}^{t+1}),$$

$$\mathbf{A}_{j\cdot}^{t+1} = \mathcal{G}_{\eta\lambda\lambda_g}\left(\widetilde{\mathbf{A}}_{j\cdot}^{t+1}\right), \qquad j \in [p].$$

Algorithm 1 Projected proximal-gradient algorithm for the proposed multitask smoothed SVM.

Input: Data $\{(\mathbf{x}_i^{(m)}, y_i^{(m)})\}_{i=1}^{n_m}$ for $m \in [M]$; step size $\eta > 0$; regularization parameters $\lambda, \lambda_g > 0$.

Initialize $\mathbf{b}^0 = \mathbf{0} \in \mathbb{R}^M$, $\boldsymbol{\mu}^0 = \mathbf{0} \in \mathbb{R}^p$, $\mathbf{A}^0 = \mathbf{0} \in \mathbb{R}^{p\times M}$.

For $t = 0, 1, \ldots, T-1$

1. Compute the intercept update: $\mathbf{b}^{t+1} \leftarrow \mathbf{b}^t - \eta\nabla_{\mathbf{b}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}^t)$.
2. Compute the gradient step for the common component: $\widetilde{\boldsymbol{\mu}}^{t+1} \leftarrow \boldsymbol{\mu}^t - \eta\nabla_{\boldsymbol{\mu}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}^t)$.
3. Compute the gradient step for each deviation row: $\widetilde{\mathbf{A}}_{j\cdot}^{t+1} \leftarrow \mathbf{A}_{j\cdot}^t - \eta\nabla_{\mathbf{A}_{j\cdot}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}^t)$, $j \in [p]$.
4. Apply the ℓ_1 proximal map to the common component:
$$\boldsymbol{\mu}^{t+1} \leftarrow S_{\eta\lambda}(\widetilde{\boldsymbol{\mu}}^{t+1}).$$
5. Apply the constrained group proximal map to each deviation row:
$$\mathbf{A}_{j\cdot}^{t+1} \leftarrow \mathcal{G}_{\eta\lambda\lambda_g}(\widetilde{\mathbf{A}}_{j\cdot}^{t+1}), \qquad j \in [p].$$

Output: The estimators $\widehat{b}_h^{(m)}$, $\widehat{\boldsymbol{\mu}}_h$, and $\widehat{\boldsymbol{\alpha}}_h^{(m)}$ for $m \in [M]$.

S3 Sensitivity of the choice of h

Our choice of $h = 0.2$ is somewhat arbitrary, but is of similar order to the bandwidth suggested by the empirical rule used in Tan et al. (2022). Their setting is sparse high-dimensional quantile regression, whereas ours

is multitask SVM classification, so this comparison does not establish an optimal bandwidth for our method. One could also select h by cross-validation, but this would increase the computational cost substantially because the regularization parameters must be tuned for each candidate bandwidth.

Finally, we assess sensitivity to the bandwidth by comparing $h \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.8\}$ for the proposed method under the baseline setting. We tune the regularization parameters for each bandwidth. Figure 2 reports the misclassification rate (MR), estimation error (EE), true positive rate (TPR), and false positive rate (FPR). We see that the results are not very sensitive to the choice of h over the range considered, although it also suggests room for improvement using cross-validation.

S4 Proofs for Section 3

Before presenting the proofs of the main theoretical results, we introduce some notation that will be used throughout this section. We first define the estimation errors and the relevant active sets. We then introduce compact notation for the empirical and population risks, together with their corresponding score functions. These definitions allow us to express the basic inequality and the subsequent stochastic and curvature arguments in a concise form.

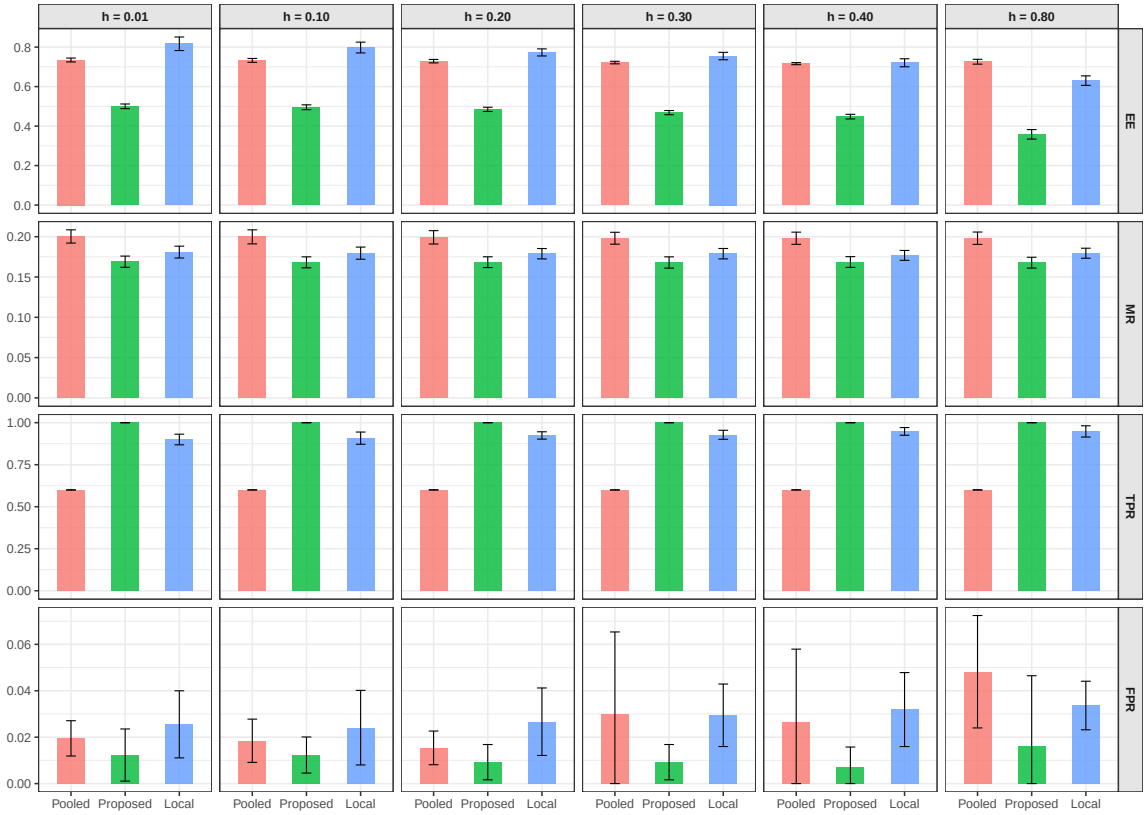


Figure 2: **Sensitivity to the bandwidth** $h \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.8\}$**.**

For each task m, define the estimation errors

$$\widehat{\delta}_{b^{(m)}} := \widehat{b}_h^{(m)} - b^{*(m)} \in \mathbb{R}, \quad \widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}} := \widehat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}^* \in \mathbb{R}^p, \quad \widehat{\boldsymbol{\delta}}_{\boldsymbol{\alpha}^{(m)}} := \widehat{\boldsymbol{\alpha}}_h^{(m)} - \boldsymbol{\alpha}^{*(m)} \in \mathbb{R}^p,$$

and collect them as

$$\widehat{\boldsymbol{\delta}}_{\mathbf{b}} := (\widehat{\delta}_{b^{(1)}}, \ldots, \widehat{\delta}_{b^{(M)}})^\top \in \mathbb{R}^M, \quad \widehat{\boldsymbol{\Delta}}_{\mathbf{A}} := (\widehat{\boldsymbol{\delta}}_{\boldsymbol{\alpha}^{(1)}}, \ldots, \widehat{\boldsymbol{\delta}}_{\boldsymbol{\alpha}^{(M)}}) \in \mathbb{R}^{p \times M}.$$

To describe the sparsity structure of the true parameter, let

$$\begin{aligned} S_\mu &:= \left\{ j \in [p] : \mu_j^* \neq 0 \right\}, \\ S_\alpha &:= \left\{ j \in [p] : \|\mathbf{A}_{j\cdot}^*\|_2 \neq 0 \right\}. \end{aligned}$$

Here, S_μ is the support of the shared component, whereas S_α is the active row set of the task-specific deviation matrix $\mathbf{A}^*$. These two sets will be used to decompose the estimation errors into their active and inactive components.

We next introduce notation for the loss and risk functions. Write

$$\mathbf{b} := \left(b^{(1)}, \ldots, b^{(M)}\right)^\top \in \mathbb{R}^M.$$

For any $\boldsymbol{\Theta} := (\mathbf{b}, \boldsymbol{\mu}, \mathbf{A})$, define the task-specific loss with a generic observation $(\mathbf{x}, y)$

from task m by

$$\ell_h^{(m)}(\boldsymbol{\Theta}; \mathbf{x}, y) := \phi_h \left(y(\mathbf{x}^\top(\boldsymbol{\mu} + \boldsymbol{\alpha}^{(m)}) + b^{(m)}) \right),$$

Accordingly, the pooled empirical risk is then given by

$$\mathcal{L}_{N,h}(\boldsymbol{\Theta}) := P_N \ell_h(\boldsymbol{\Theta}) = \frac{1}{N} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \ell_h^{(m)}(\boldsymbol{\Theta}; \mathbf{x}_i^{(m)}, y_i^{(m)}),$$

while its population counterpart is

$$\mathcal{L}_h(\boldsymbol{\Theta}) := P \ell_h(\boldsymbol{\Theta}) = \frac{1}{N} \sum_{m=1}^{M} n_m \mathbb{E}^{(m)} \left[\ell_h^{(m)}(\boldsymbol{\Theta}; \mathbf{x}^{(m)}, y^{(m)}) \right].$$

where $\mathbb{E}^{(m)}[\cdot]$ denotes the expectation with respect to the task-m distribution.

To express gradients compactly, recall that the augmented task-specific parameter is $\boldsymbol{\theta}^{(m)} := (b^{(m)}, (\boldsymbol{\mu} + \boldsymbol{\alpha}^{(m)})^\top)^\top$. Partition the corresponding population score vector as

$$\mathbf{s}_h^{(m)}(\boldsymbol{\theta}^{(m)}) = \left(\mathbf{s}_{h,1}^{(m)}(\boldsymbol{\theta}^{(m)}), \left(\mathbf{s}_{h,-1}^{(m)}(\boldsymbol{\theta}^{(m)}) \right)^\top \right)^\top,$$

where the first component corresponds to the intercept and the remaining p compo-

nents correspond to the slope coefficients. More explicitly,

$$\begin{aligned}
\mathbf{s}_{h,1}^{(m)}(\boldsymbol{\theta}^{(m)}) &= -\mathbb{E}^{(m)}\left[\bar{K}\left(\frac{1-y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top\boldsymbol{\theta}^{(m)}}{h}\right)y^{(m)}\right] \in \mathbb{R}, \\
\mathbf{s}_{h,-1}^{(m)}(\boldsymbol{\theta}^{(m)}) &= -\mathbb{E}^{(m)}\left[\bar{K}\left(\frac{1-y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top\boldsymbol{\theta}^{(m)}}{h}\right)y^{(m)}(\mathbf{x}^{(m)})^\top\right] \in \mathbb{R}^p.
\end{aligned}$$

Therefore, the gradient with respect to the task-specific intercept vector $\mathbf{b} = (b^{(1)}, \ldots, b^{(M)})^\top$ is given by

$$\nabla_{\mathbf{b}}\mathcal{L}_h(\boldsymbol{\Theta}) = \left(\frac{n_1}{N}\mathbf{s}_{h,1}^{(1)}(\boldsymbol{\theta}^{(1)}), \ldots, \frac{n_M}{N}\mathbf{s}_{h,1}^{(M)}(\boldsymbol{\theta}^{(M)})\right)^\top.$$

In contrast, the shared component $\boldsymbol{\mu}$ appears in the loss of every task. Its gradient therefore aggregates the task-specific slope scores:

$$\nabla_{\boldsymbol{\mu}}\mathcal{L}_h(\boldsymbol{\Theta}) = \frac{1}{N}\sum_{m=1}^{M} n_m \mathbf{s}_{h,-1}^{(m)}(\boldsymbol{\theta}^{(m)}).$$

Finally, because $\boldsymbol{\alpha}^{(m)}$ affects only the loss for task m, the gradient with respect to $\mathbf{A} = (\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(M)})$ is obtained columnwise as

$$\nabla_{\mathbf{A}}\mathcal{L}_h(\boldsymbol{\Theta}) = \left(\frac{n_1}{N}\mathbf{s}_{h,-1}^{(1)}(\boldsymbol{\theta}^{(1)}), \ldots, \frac{n_M}{N}\mathbf{s}_{h,-1}^{(M)}(\boldsymbol{\theta}^{(M)})\right).$$

The empirical score functions are obtained by replacing the expectations in the popu-

lation scores with the corresponding within-task sample averages. Specifically, define

$$\begin{aligned}
\nabla_{\mathbf{b}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}) &= \left(\frac{n_1}{N}\widehat{\mathbf{s}}_{h,1}^{(1)}(\boldsymbol{\theta}^{(1)}),\ldots,\frac{n_M}{N}\widehat{\mathbf{s}}_{h,1}^{(M)}(\boldsymbol{\theta}^{(M)})\right)^{\top},\\
\nabla_{\boldsymbol{\mu}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}) &= \frac{1}{N}\sum_{m=1}^{M} n_m\widehat{\mathbf{s}}_{h,-1}^{(m)}(\boldsymbol{\theta}^{(m)}),\\
\nabla_{\mathbf{A}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}) &= \left(\frac{n_1}{N}\widehat{\mathbf{s}}_{h,-1}^{(1)}(\boldsymbol{\theta}^{(1)}),\ldots,\frac{n_M}{N}\widehat{\mathbf{s}}_{h,-1}^{(M)}(\boldsymbol{\theta}^{(M)})\right),
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\mathbf{s}}_{h,1}^{(m)}(\boldsymbol{\theta}^{(m)}) &= -\frac{1}{n_m}\sum_{i=1}^{n_m}\bar{K}\left(\frac{1-y_i^{(m)}(\widetilde{\mathbf{x}}_i^{(m)})^{\top}\boldsymbol{\theta}^{(m)}}{h}\right)y_i^{(m)}\\
\widehat{\mathbf{s}}_{h,-1}^{(m)}(\boldsymbol{\theta}^{(m)}) &= -\frac{1}{n_m}\sum_{i=1}^{n_m}\bar{K}\left(\frac{1-y_i^{(m)}(\widetilde{\mathbf{x}}_i^{(m)})^{\top}\boldsymbol{\theta}^{(m)}}{h}\right)y_i^{(m)}\mathbf{x}_i^{(m)}.
\end{aligned}$$

With this notation in place, we now establish several auxiliary lemmas that control the score fluctuations, the smoothing bias, and the local curvature of the empirical loss. These results form the main ingredients in the proofs of the estimation error bound and the variable-selection consistency result.

S4.1 Lemmas and Their Proofs

In the task-wise conditional-density arguments below, select a largest-magnitude coordinate j_m and label it as coordinate 1 locally, without loss of generality. The symbol $\beta_1^{*(m)}$ denotes its coefficient; $\mathbf{x}_{-1}^{(m)}$ contains the remaining coordinates. Condition 4 applies in these local coordinates,

uniformly across tasks. This convention is used only for integration in the proofs of Lemmas 1 and 2; all multitask parameters retain their original global feature indexing. Euclidean score norms and Hessian quadratic-form bounds are unchanged by these permutations, so the resulting bounds combine across tasks as before.

Lemma 1. Under Conditions 2, 4 and 5, for each task m and $\boldsymbol{\theta}^{*(m)} := (b^{*(m)}, (\boldsymbol{\beta}^{*(m)})^\top)^\top$ with $\boldsymbol{\beta}^{*(m)} := \boldsymbol{\mu}^* + \boldsymbol{\alpha}^{*(m)}$, we have

$$\|\mathbf{s}_h^{(m)}(\boldsymbol{\theta}^{*(m)})\| \leq C(h/\beta_1^{*(m)})^2.$$

Proof of Lemma 1. For notational simplicity, we suppress the task-specific superscript $\cdot^{(m)}$ throughout the proof, since the same argument applies to each task $m = 1, \ldots, M$. Let $\pi_+ = \mathbb{P}(y = 1)$, $\pi_- = \mathbb{P}(y = -1)$, $\widetilde{\mathbf{x}} = (1, \mathbf{x}^\top)^\top$, $\mathbf{u} = (u_1, \mathbf{u}_{-1}^\top)^\top$ and $\widetilde{\mathbf{u}} = (u_0, \mathbf{u}^\top)^\top$. For any $\widetilde{\mathbf{u}} \in \mathbb{S}^p$, we get

$$\begin{aligned} \widetilde{\mathbf{u}}^\top \mathbf{s}_h(\boldsymbol{\theta}^*) &= -\widetilde{\mathbf{u}}^\top \mathbb{E}\left[\bar{K}\left(\frac{1 - y\widetilde{\mathbf{x}}^\top \boldsymbol{\theta}^*}{h}\right) y\widetilde{\mathbf{x}}\right] \\ = \; -\pi_+ \mathbb{E}\left[\bar{K}\left(\frac{1 - y\widetilde{\mathbf{x}}^\top \boldsymbol{\theta}^*}{h}\right) y\widetilde{\mathbf{u}}^\top \widetilde{\mathbf{x}} | y = 1\right] &- \pi_- \mathbb{E}\left[\bar{K}\left(\frac{1 - y\widetilde{\mathbf{x}}^\top \boldsymbol{\theta}^*}{h}\right) y\widetilde{\mathbf{u}}^\top \widetilde{\mathbf{x}} | y = -1\right]. \end{aligned}$$

It suffices to upper bound $\mathbb{E}\left[\bar{K}\left(\frac{1-y\widetilde{\mathbf{x}}^\top \boldsymbol{\theta}^*}{h}\right) \widetilde{\mathbf{u}}^\top \widetilde{\mathbf{x}} | y = 1\right]$, and the term conditional on $y = -1$ is handled analogously. Recall that $f(\mathbf{x})$ is the conditional density function

of $\mathbf{x}$ given $y = 1$. Then we have

$$\begin{aligned}
&\mathbb{E}\left[\bar{K}\left(\frac{1-y\widetilde{\mathbf{x}}^\top\boldsymbol{\theta}^*}{h}\right)y\widetilde{\mathbf{u}}^\top\widetilde{\mathbf{x}}|y=1\right]\\
=&\int_{\mathbb{R}^p}(u_0+\mathbf{u}^\top\mathbf{x})\bar{K}\left(\frac{1-b^*-\mathbf{x}^\top\boldsymbol{\beta}^*}{h}\right)f(\mathbf{x})d\mathbf{x}\\
=&\int_{\mathbb{R}^{p-1}}\int_{\mathbb{R}}(u_0+u_1x_1+\mathbf{u}_{-1}^\top\mathbf{x}_{-1})\bar{K}\left(\frac{1-b^*-x_1\beta_1^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*}{h}\right)f(x_1,\mathbf{x}_{-1})dx_1d\mathbf{x}_{-1}.
\end{aligned}$$

By a change of variable $z=\frac{1-b^*-x_1\beta_1^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*}{h}$, the above equation becomes

$$\begin{aligned}
&\mathbb{E}\left[\bar{K}\left(\frac{1-y\widetilde{\mathbf{x}}^\top\boldsymbol{\theta}^*}{h}\right)y\widetilde{\mathbf{u}}^\top\widetilde{\mathbf{x}}|y=1\right]\\
=&\frac{h}{\beta_1^*}\int_{\mathbb{R}^{p-1}}f_{-1}(\mathbf{x}_{-1})\int_{\mathbb{R}}\left(u_0+u_1\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*-hz}{\beta_1^*}+\mathbf{u}_{-1}^\top\mathbf{x}_{-1}\right)\times\\
&f\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*-hz}{\beta_1^*}|\mathbf{x}_{-1}\right)\bar{K}(z)dzd\mathbf{x}_{-1},
\end{aligned}\tag{S4.1}$$

where $f_{-1}(\mathbf{x}_{-1})$ denotes the joint density function of $\mathbf{x}_{-1}=(x_2,\ldots,x_p)^\top$ conditional on $y=1$. Define $H(t|\mathbf{x}_{-1})=\int_{-\infty}^t xf(x|\mathbf{x}_{-1})dx$. Applying integration by parts yields

$$\begin{aligned}
&\frac{h}{\beta_1^*}\int_{\mathbb{R}}u_1\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*-hz}{\beta_1^*}f\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*-hz}{\beta_1^*}|\mathbf{x}_{-1}\right)\bar{K}(z)dz\\
=&-u_1\int_{\mathbb{R}}\bar{K}(z)dH\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*-hz}{\beta_1^*}|\mathbf{x}_{-1}\right)\\
=&-u_1\left\{\bar{K}(z)H\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*-hz}{\beta_1^*}|\mathbf{x}_{-1}\right)\Big|_{-\infty}^{\infty}\right.\\
&\qquad\left.-\int_{\mathbb{R}}H\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*-hz}{\beta_1^*}|\mathbf{x}_{-1}\right)d\bar{K}(z)\right\}\\
=&u_1\int_{\mathbb{R}}H\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*-hz}{\beta_1^*}|\mathbf{x}_{-1}\right)K(z)dz
\end{aligned}$$

This, along with Taylor's expansion, further implies

$$\begin{aligned}
&\frac{h}{\beta_1^*}\int_{\mathbb{R}} u_1 \frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*-hz}{\beta_1^*} f\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*-hz}{\beta_1^*}|\mathbf{x}_{-1}\right)\bar{K}(z)dz \\
= \;& u_1\int_{\mathbb{R}} H\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*}{\beta_1^*}|\mathbf{x}_{-1}\right)K(z)dz \\
&+u_1\int_{\mathbb{R}} \frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*}{\beta_1^*} f\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*}{\beta_1^*}|\mathbf{x}_{-1}\right)\Delta(z)K(z)dz \\
&+\frac{u_1}{2}\int_{\mathbb{R}}\left\{f\left(\delta|\mathbf{x}_{-1}\right)+\delta f'\left(\delta|\mathbf{x}_{-1}\right)\right\}\Delta^2(z)K(z)dz \\
\\
= \;& u_1 H\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*}{\beta_1^*}|\mathbf{x}_{-1}\right) \\
&+\frac{u_1}{2}\int_{\mathbb{R}}\left\{f\left(\delta|\mathbf{x}_{-1}\right)+\delta f'\left(\delta|\mathbf{x}_{-1}\right)\right\}\Delta^2(z)K(z)dz, \qquad \text{(S4.2)}
\end{aligned}$$

where

$$\Delta(z)=\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*-hz}{\beta_1^*}-\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*}{\beta_1^*}=-\frac{hz}{\beta_1^*},\delta=\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*-\xi}{\beta_1^*}$$

with $\xi\in(0,hz)$, and we used the fact that $\int_{\mathbb{R}}K(u)du=1$, $\int_{\mathbb{R}}uK(u)du=0$ on the last line. Denote $F(t|\mathbf{x}_{-1})=\int_{-\infty}^t f(x|\mathbf{x}_{-1})dx$. Similarly, it holds that

$$\begin{aligned}
&\frac{h}{\beta_1^*}\int_{\mathbb{R}}(u_0+\mathbf{u}_{-1}^\top\mathbf{x}_{-1})f\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*-hz}{\beta_1^*}|\mathbf{x}_{-1}\right)\bar{K}(z)dz \\
= \;& (u_0+\mathbf{u}_{-1}^\top\mathbf{x}_{-1})F\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*}{\beta_1^*}|\mathbf{x}_{-1}\right) \\
&+\frac{u_0+\mathbf{u}_{-1}^\top\mathbf{x}_{-1}}{2}\int_{\mathbb{R}} f'\left(\delta|\mathbf{x}_{-1}\right)\Delta^2(z)K(z)dz. \qquad \text{(S4.3)}
\end{aligned}$$

Substituting (S4.2)-(S4.3) into (S4.1) yields

$$\begin{aligned}
& -\mathbb{E}\left[\bar{K}\left(\frac{1-y\widetilde{\mathbf{x}}^\top\boldsymbol{\theta}^*}{h}\right)\widetilde{\mathbf{u}}^\top\widetilde{\mathbf{x}}|y=1\right] \\
= & -\int_{\mathbb{R}^{p-1}} f_{-1}(\mathbf{x}_{-1})\Bigg\{u_1 H\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*}{\beta_1^*}|\mathbf{x}_{-1}\right) \\
& +(u_0+\mathbf{u}_{-1}^\top\mathbf{x}_{-1})F\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*}{\beta_1^*}|\mathbf{x}_{-1}\right)\Bigg\}d\mathbf{x}_{-1}+O((h/\beta_1^*)^2) \\
:= & A_1+A_2+O((h/\beta_1^*)^2), \qquad \text{(S4.4)}
\end{aligned}$$

where the final term $O((h/\beta_1^*)^2)$ is obtained by invoking Condition 4 and Condition 5 that $\int_{\mathbb{R}} u^2K(u)du \le C$.

Let $w=\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*}{\beta_1^*}$. It follows from direct calculations that

$$\begin{aligned}
A_1 := & -\int_{\mathbb{R}^{p-1}} f_{-1}(\mathbf{x}_{-1})u_1 H\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*}{\beta_1^*}|\mathbf{x}_{-1}\right)d\mathbf{x}_{-1} \\
= & -\int_{\mathbb{R}^{p-1}} f_{-1}(\mathbf{x}_{-1})\int_{-\infty}^{w} u_1x_1 f(x_1|\mathbf{x}_{-1})dx_1d\mathbf{x}_{-1} \\
= & -\int_{\mathbb{R}^{p-1}} f_{-1}(\mathbf{x}_{-1})\int_{\mathbb{R}} u_1x_1 I\{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*\ge x_1\beta_1^*\}f(x_1|\mathbf{x}_{-1})dx_1d\mathbf{x}_{-1} \\
= & -\int_{\mathbb{R}^{p-1}}\int_{\mathbb{R}} u_1x_1 I\{1-b^*-x_1\beta_1^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*\ge 0\}f(x_1,\mathbf{x}_{-1})dx_1d\mathbf{x}_{-1} \\
= & -\mathbb{E}[yu_1x_1I\{1-y\widetilde{\mathbf{x}}^\top\boldsymbol{\theta}^*\ge 0\}|y=1], \qquad \text{(S4.5)}
\end{aligned}$$

and

$$\begin{aligned} A_2 &:= -\int_{\mathbb{R}^{p-1}} f_{-1}(\mathbf{x}_{-1})(u_0+\mathbf{u}_{-1}^\top\mathbf{x}_{-1})F\left(\frac{1-b^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*}{\beta_1^*}|\mathbf{x}_{-1}\right)d\mathbf{x}_{-1} \\ &= -\int_{\mathbb{R}^{p-1}} f_{-1}(\mathbf{x}_{-1})(u_0+\mathbf{u}_{-1}^\top\mathbf{x}_{-1})\int_{-\infty}^{w} f(x_1|\mathbf{x}_{-1})dx_1 d\mathbf{x}_{-1} \\ &= -\int_{\mathbb{R}^{p-1}} f_{-1}(\mathbf{x}_{-1})(u_0+\mathbf{u}_{-1}^\top\mathbf{x}_{-1})\int_{\mathbb{R}} I\{1-b^*-x_1\beta_1^*-\mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}^*\geq 0\}f(x_1|\mathbf{x}_{-1})dx_1 d\mathbf{x}_{-1} \\ &= -\mathbb{E}[y(u_0+\mathbf{u}_{-1}^\top\mathbf{x}_{-1})I\{1-y\widetilde{\mathbf{x}}^\top\boldsymbol{\theta}^*\geq 0\}|y=1]. \end{aligned} \tag{S4.6}$$

Therefore, combining inequalities (S4.4)-(S4.6) together gives

$$\begin{aligned} &-\mathbb{E}\left[\bar{K}\left(\frac{1-y\widetilde{\mathbf{x}}^\top\boldsymbol{\theta}^*}{h}\right)\widetilde{\mathbf{u}}^\top\widetilde{\mathbf{x}}|y=1\right] \\ =\ &-\mathbb{E}[y\widetilde{\mathbf{u}}^\top\widetilde{\mathbf{x}}I\{1-y\widetilde{\mathbf{x}}^\top\boldsymbol{\theta}^*\geq 0\}|y=1]+O((h/\beta_1^*)^2)=O((h/\beta_1^*)^2), \end{aligned} \tag{S4.7}$$

where we used the first-order optimality condition of (2.4) in the last inequality. Thus

$$\|\mathbf{s}_h(\boldsymbol{\theta}^*)\|=\sup_{\widetilde{\mathbf{u}}\in\mathbb{S}^p}\widetilde{\mathbf{u}}^\top\mathbf{s}_h(\boldsymbol{\theta}^*)\leq C(h/\beta_1^*)^2, \tag{S4.8}$$

which completes the proof. □

Lemma 2. Under Conditions 2, 4 and 5, we have for any $\widetilde{\mathbf{v}}=(v_0,\mathbf{v}^\top)^\top\in\mathbb{R}^{p+1}$, $\boldsymbol{\theta}^{(m)}:=(b^{(m)},(\boldsymbol{\beta}^{(m)})^\top)^\top$ with $\boldsymbol{\beta}^{(m)}=\boldsymbol{\mu}+\boldsymbol{\alpha}^{(m)}$, we have

$$\widetilde{\mathbf{v}}^\top\mathbf{H}_h^{(m)}(\boldsymbol{\theta}^{(m)})\widetilde{\mathbf{v}}\geq\widetilde{\mathbf{v}}^\top\mathbf{H}^{(m)}(\boldsymbol{\theta}^{(m)})\widetilde{\mathbf{v}}-\frac{h}{(\beta_1^{(m)})^2}\|\widetilde{\mathbf{v}}\|^2. \tag{S4.9}$$

Proof of Lemma 2. For notational simplicity, we suppress the task-specific superscript $\cdot^{(m)}$ throughout the proof, since the same argument applies to each task $m = 1, \ldots, M$. It follows from direct definitions that

$$\begin{aligned}
&\widetilde{\mathbf{v}}^\top \mathbf{H}_h(\boldsymbol{\theta})\widetilde{\mathbf{v}} = \widetilde{\mathbf{v}}^\top \mathbb{E}[K_h(1 - y\widetilde{\mathbf{x}}^\top\boldsymbol{\theta})\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^\top]\widetilde{\mathbf{v}} \\
= \;& \pi_+ \mathbb{E}\left[(\widetilde{\mathbf{v}}^\top\widetilde{\mathbf{x}})^2 K_h(1 - y\widetilde{\mathbf{x}}^\top\boldsymbol{\theta})|y = 1\right] + \pi_- \mathbb{E}\left[(\widetilde{\mathbf{v}}^\top\widetilde{\mathbf{x}})^2 K_h(1 - y\widetilde{\mathbf{x}}^\top\boldsymbol{\theta})|y = -1\right] \\
= \;& \pi_+ \int_{\mathbb{R}^p} (\widetilde{\mathbf{v}}^\top\widetilde{\mathbf{x}})^2 K_h(1 - \widetilde{\mathbf{x}}^\top\boldsymbol{\theta}) f(\mathbf{x}) d\mathbf{x} + \pi_- \int_{\mathbb{R}^p} (\widetilde{\mathbf{v}}^\top\widetilde{\mathbf{x}})^2 K_h(1 + \widetilde{\mathbf{x}}^\top\boldsymbol{\theta}) g(\mathbf{x}) d\mathbf{x}.
\end{aligned}$$

For convenience, denote $\varrho(\mathbf{x}) = (\widetilde{\mathbf{v}}^\top\widetilde{\mathbf{x}})^2 = (v_0 + v_1 x_1 + \mathbf{v}_{-1}^\top \mathbf{x}_{-1})^2$, then we have

$$\begin{aligned}
&\int_{\mathbb{R}^p} (\widetilde{\mathbf{v}}^\top\widetilde{\mathbf{x}})^2 K_h(1 - \widetilde{\mathbf{x}}^\top\boldsymbol{\theta}) f(\mathbf{x}) d\mathbf{x} \\
= \;& \frac{1}{h} \int_{\mathbb{R}^p} K\left(\frac{1 - b - \mathbf{x}^\top\boldsymbol{\beta}}{h}\right) \varrho(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \\
= \;& \frac{1}{h} \int_{\mathbb{R}^{p-1}} \int_{\mathbb{R}} K\left(\frac{1 - b - x_1\beta_1 - \mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}}{h}\right) \varrho(x_1, \mathbf{x}_{-1}) f(x_1, \mathbf{x}_{-1}) dx_1 d\mathbf{x}_{-1}. \quad \text{(S4.10)}
\end{aligned}$$

By a change of variable $z = \frac{1 - b - x_1\beta_1 - \mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1}}{h}$, equation (S4.10) is written as

$$\begin{aligned}
&\int_{\mathbb{R}^p} (\widetilde{\mathbf{v}}^\top\widetilde{\mathbf{x}})^2 K_h(1 - \widetilde{\mathbf{x}}^\top\boldsymbol{\theta}) f(\mathbf{x}) d\mathbf{x} \\
= \;& \frac{1}{\beta_1} \int_{\mathbb{R}^{p-1}} \int_{\mathbb{R}} \varrho\left(\frac{1 - b - \mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1} - hz}{\beta_1}, \mathbf{x}_{-1}\right) f\left(\frac{1 - b - \mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1} - hz}{\beta_1}, \mathbf{x}_{-1}\right) K(z) dz d\mathbf{x}_{-1} \\
= \;& \frac{1}{\beta_1} \int_{\mathbb{R}^{p-1}} f_{-1}(\mathbf{x}_{-1}) \int_{\mathbb{R}} \varrho\left(\frac{1 - b - \mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1} - hz}{\beta_1}, \mathbf{x}_{-1}\right) \times \\
& f\left(\frac{1 - b - \mathbf{x}_{-1}^\top\boldsymbol{\beta}_{-1} - hz}{\beta_1} | \mathbf{x}_{-1}\right) K(z) dz d\mathbf{x}_{-1}. \quad \text{(S4.11)}
\end{aligned}$$

Regarding the inner integration with respect to z, it follows from Taylor's expansion that

$$\begin{aligned}
&\int_{\mathbb{R}} \varrho\left(\frac{1-b-\mathbf{x}_{-1}^{\top}\boldsymbol{\beta}_{-1}-hz}{\beta_1}, \mathbf{x}_{-1}\right) f\left(\frac{1-b-\mathbf{x}_{-1}^{\top}\boldsymbol{\beta}_{-1}-hz}{\beta_1}|\mathbf{x}_{-1}\right) K(z)dz \\
= &\int_{\mathbb{R}} \varrho\left(\frac{1-b-\mathbf{x}_{-1}^{\top}\boldsymbol{\beta}_{-1}}{\beta_1}, \mathbf{x}_{-1}\right) f\left(\frac{1-b-\mathbf{x}_{-1}^{\top}\boldsymbol{\beta}_{-1}}{\beta_1}|\mathbf{x}_{-1}\right) K(z)dz \\
&+\int_{\mathbb{R}} 2v_1\left(v_0+v_1\frac{1-b-\mathbf{x}_{-1}^{\top}\boldsymbol{\beta}_{-1}-\xi}{\beta_1}+\mathbf{v}_{-1}^{\top}\mathbf{x}_{-1}\right) f\left(\frac{1-b-\mathbf{x}_{-1}^{\top}\boldsymbol{\beta}_{-1}-\xi}{\beta_1}|\mathbf{x}_{-1}\right)\Delta(z)K(z)dz \\
&+\int_{\mathbb{R}} \left(v_0+v_1\frac{1-b-\mathbf{x}_{-1}^{\top}\boldsymbol{\beta}_{-1}-\xi}{\beta_1}+\mathbf{v}_{-1}^{\top}\mathbf{x}_{-1}\right)^2 f'\left(\frac{1-b-\mathbf{x}_{-1}^{\top}\boldsymbol{\beta}_{-1}-\xi}{\beta_1}|\mathbf{x}_{-1}\right)\Delta(z)K(z)dz,
\end{aligned}$$

where $\Delta(z)=-\frac{hz}{\beta_1}$ and $\xi\in(0,hz)$. By Condition 4 and Condition 5, we have

$$\begin{aligned}
&\int_{\mathbb{R}} \varrho\left(\frac{1-b-\mathbf{x}_{-1}^{\top}\boldsymbol{\beta}_{-1}-hz}{\beta_1}, \mathbf{x}_{-1}\right) f\left(\frac{1-b-\mathbf{x}_{-1}^{\top}\boldsymbol{\beta}_{-1}-hz}{\beta_1}|\mathbf{x}_{-1}\right) K(z)dz \\
= &\varrho\left(\frac{1-b-\mathbf{x}_{-1}^{\top}\boldsymbol{\beta}_{-1}}{\beta_1}, \mathbf{x}_{-1}\right) f\left(\frac{1-b-\mathbf{x}_{-1}^{\top}\boldsymbol{\beta}_{-1}}{\beta_1}|\mathbf{x}_{-1}\right) \\
&+O(h/\beta_1)(|v_0|+|v_1|+\|\mathbf{v}_{-1}\|)^2\mathcal{E}(\mathbf{x}_{-1}).
\end{aligned}$$

This together with (S4.11) yields

$$\begin{aligned}
&\int_{\mathbb{R}^p} (\widetilde{\mathbf{v}}^\top \widetilde{\mathbf{x}})^2 K_h(1-\widetilde{\mathbf{x}}^\top \boldsymbol{\theta}) f(\mathbf{x}) d\mathbf{x} \\
= \;& \frac{1}{\beta_1} \int_{\mathbb{R}^{p-1}} f_{-1}(\mathbf{x}_{-1}) \left[\varrho\left(\frac{1-b-\mathbf{x}_{-1}^\top \boldsymbol{\beta}_{-1}}{\beta_1}, \mathbf{x}_{-1} \right) f\left(\frac{1-b-\mathbf{x}_{-1}^\top \boldsymbol{\beta}_{-1}}{\beta_1} | \mathbf{x}_{-1} \right) \right] d\mathbf{x}_{-1} \\
& \frac{1}{\beta_1} \int_{\mathbb{R}^{p-1}} f_{-1}(\mathbf{x}_{-1}) \mathcal{E}(\mathbf{x}_{-1}) \left[O(h/\beta_1) \|\widetilde{\mathbf{v}}\|^2 \right] d\mathbf{x}_{-1} \\
= \;& \frac{1}{\beta_1} \int_{\mathbb{R}^{p-1}} \left[\varrho\left(\frac{1-b-\mathbf{x}_{-1}^\top \boldsymbol{\beta}_{-1}^*}{\beta_1}, \mathbf{x}_{-1} \right) f\left(\frac{1-b-\mathbf{x}_{-1}^\top \boldsymbol{\beta}_{-1}}{\beta_1}, \mathbf{x}_{-1} \right) \right] d\mathbf{x}_{-1} \\
& + O(h/\beta_1^2) \|\widetilde{\mathbf{v}}\|^2 \\
= \;& \widetilde{\mathbf{v}}^\top \mathbb{E}[\delta(1-y\widetilde{\mathbf{x}}^\top \boldsymbol{\theta}) \widetilde{\mathbf{x}} \widetilde{\mathbf{x}}^\top | y=1] \widetilde{\mathbf{v}} + O(h/\beta_1^2) \|\widetilde{\mathbf{v}}\|^2.
\end{aligned}$$

Similarly, we obtain $\int_{\mathbb{R}^p} (\widetilde{\mathbf{v}}^\top \widetilde{\mathbf{x}})^2 K_h(1+\widetilde{\mathbf{x}}^\top \boldsymbol{\theta}) g(\mathbf{x}) d\mathbf{x} = \widetilde{\mathbf{v}}^\top \mathbb{E}[\delta(1-y\widetilde{\mathbf{x}}^\top \boldsymbol{\theta}) \widetilde{\mathbf{x}} \widetilde{\mathbf{x}}^\top | y=-1] \widetilde{\mathbf{v}} + O(h/\beta_1^2) \|\widetilde{\mathbf{v}}\|^2$. Thus we have

$$\left| \widetilde{\mathbf{v}}^\top \mathbf{H}_h(b, \boldsymbol{\beta}) \widetilde{\mathbf{v}} - \widetilde{\mathbf{v}}^\top \mathbb{E}[\delta(1-y\widetilde{\mathbf{x}}^\top \boldsymbol{\theta}) \widetilde{\mathbf{x}} \widetilde{\mathbf{x}}^\top] \widetilde{\mathbf{v}} \right| \leq C(h/\beta_1^2) \|\widetilde{\mathbf{v}}\|^2.$$

This completes the proof. □

Lemma 3. Suppose that Condition 1 holds and that the task sample sizes are balanced in the sense that

$$\frac{c_-}{M} \leq \frac{n_m}{N} \leq \frac{c_+}{M}, \qquad m \in [M],$$

for some constants $c_-, c_+ > 0$. Then, with probability at least $1 - (p \vee N)^{-C}$,

$$\begin{aligned}
\|\nabla_{\boldsymbol{\mu}} \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*)\|_\infty &\leq C\left(\sqrt{\frac{\log(p \vee N)}{N}} + \left(\frac{h}{|\beta^*_{1,\min}|}\right)^2\right), \\
\|\nabla_{\mathbf{A}} \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*)\|_{2,\infty} &\leq C\left(\sqrt{\frac{M + \log(p \vee N)}{MN}} + \frac{1}{\sqrt{M}}\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2\right) \\
\|\nabla_{\mathbf{b}} \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*)\|_2 &\leq C\left(\sqrt{\frac{M + \log(p \vee N)}{MN}} + \frac{1}{\sqrt{M}}\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2\right).
\end{aligned}$$

Proof of Lemma 3. Let $\beta_1^{*(m)} = \beta_{j_m}^{*(m)}$ denote a largest-magnitude coordinate of $\boldsymbol{\beta}^{*(m)} = \boldsymbol{\mu}^* + \boldsymbol{\alpha}^{*(m)}$ under the task-wise local convention, and define

$$\beta^*_{1,\min} := \min_{1 \leq m \leq M} |\beta_1^{*(m)}| = \min_{\mathbf{1} \leq \mathbf{m} \leq \mathbf{M}} \|\boldsymbol{\beta}^{*(\mathbf{m})}\|_\infty > 0.$$

By Lemma 1, applied task-wise,

$$\|\mathbf{s}_h^{(m)}(\boldsymbol{\theta}^{*(m)})\| \leq C\left(\frac{h}{|\beta_1^{*(m)}|}\right)^2 \leq C\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2, \qquad m = 1, \ldots, M.$$

We first bound the population gradients. For the intercept block, it follows immediately that

$$\|\nabla_{\mathbf{b}} \mathcal{L}_h(\boldsymbol{\Theta}^*)\|^2 = \sum_{m=1}^M \left(\frac{n_m}{N}\right)^2 |\mathbf{s}_{h,1}^{(m)}(\boldsymbol{\theta}^{*(m)})|^2 \leq C\left(\frac{h}{|\beta^*_{1,\min}|}\right)^4 \sum_{m=1}^M \frac{1}{M^2},$$

where the last step uses the task-balance condition. Therefore,

$$\|\nabla_{\mathbf{b}}\mathcal{L}_h(\mathbf{\Theta}^*)\| \le \frac{C}{\sqrt{M}}\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2. \tag{S4.12}$$

For the shared block, we have

$$\begin{aligned}\|\nabla_{\boldsymbol{\mu}}\mathcal{L}_h(\mathbf{\Theta}^*)\|_\infty &\le \frac{1}{N}\sum_{m=1}^M n_m\|\mathbf{s}^{(m)}_{h,-1}(\boldsymbol{\theta}^{*(m)})\|_\infty \le \frac{1}{N}\sum_{m=1}^M n_m\|\mathbf{s}^{(m)}_{h,-1}(\boldsymbol{\theta}^{*(m)})\| \\ &\le \frac{1}{N}\sum_{m=1}^M n_m\|\mathbf{s}^{(m)}_{h}(\boldsymbol{\theta}^{*(m)})\| \le C\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2. \end{aligned}\tag{S4.13}$$

For the deviation block, for each row $j \in [p]$,

$$\begin{aligned}\|\left(\nabla_{\mathbf{A}}\mathcal{L}_h(\mathbf{\Theta}^*)\right)_{j\cdot}\|^2 &= \sum_{m=1}^M\left(\frac{n_m}{N}\right)^2|(\mathbf{s}^{(m)}_{h,-1}(\boldsymbol{\theta}^{*(m)}))_j|^2 \\ &\le \left(\max_{1\le m\le M}\|\mathbf{s}^{(m)}_{h,-1}(\boldsymbol{\theta}^{*(m)})\|^2_\infty\right)\sum_{m=1}^M\left(\frac{n_m}{N}\right)^2 \\ &\le \frac{C}{M}\left(\frac{h}{|\beta^*_{1,\min}|}\right)^4.\end{aligned}$$

Hence,

$$\|\nabla_{\mathbf{A}}\mathcal{L}_h(\mathbf{\Theta}^*)\|_{2,\infty} \le \frac{C}{\sqrt{M}}\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2. \tag{S4.14}$$

Next, we bound the empirical fluctuations for the three blocks. Consider the

shared block, for each $j \in \{1, \ldots, p\}$, define

$$\xi_{ij}^{(m)} := \bar{K}\left(\frac{1 - y_i^{(m)}(\widetilde{\mathbf{x}}_i^{(m)})^\top \boldsymbol{\theta}^{*(m)}}{h}\right) y_i^{(m)} x_{ij}^{(m)} - \mathbb{E}^{(m)}\left[\bar{K}\left(\frac{1 - y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top \boldsymbol{\theta}^{*(m)}}{h}\right) y^{(m)} x_j^{(m)}\right].$$

Then

$$\left(\nabla_{\boldsymbol{\mu}} \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) - \nabla_{\boldsymbol{\mu}} \mathcal{L}_h(\boldsymbol{\Theta}^*)\right)_j = -\frac{1}{N} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \xi_{ij}^{(m)}.$$

Since $|\bar{K}| \leq 1$ and $y \in \{\pm 1\}$, Condition 1 implies each $\xi_{ij}^{(m)}$ is centered sub-Gaussian with $\|\xi_{ij}^{(m)}\|_{\psi_2} \leq C$. Hence, by Hoeffding's inequality for sub-Gaussian variables (see, e.g., Theorem 2.6.2 in Vershynin (2018)), for any $u > 0$,

$$P\left(\left|\{\nabla_{\boldsymbol{\mu}} \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) - \nabla_{\boldsymbol{\mu}} \mathcal{L}_h(\boldsymbol{\Theta}^*)\}_j\right| > u\right) \leq 2\exp(-CNu^2).$$

Taking a union bound over $j \in [p]$ and taking $u \asymp \sqrt{\log(p \vee N)/N}$, with probability at least $1 - (p \vee N)^{-C}$,

$$\|\nabla_{\boldsymbol{\mu}} \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) - \nabla_{\boldsymbol{\mu}} \mathcal{L}_h(\boldsymbol{\Theta}^*)\|_\infty \leq C\sqrt{\frac{\log(p \vee N)}{N}}.$$

Combining with (S4.13) yields

$$\begin{aligned} \|\nabla_{\boldsymbol{\mu}} \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*)\|_\infty &\leq \|\left(\nabla_{\boldsymbol{\mu}} \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) - \nabla_{\boldsymbol{\mu}} \mathcal{L}_h(\boldsymbol{\Theta}^*)\right)\|_\infty + \|\nabla_{\boldsymbol{\mu}} \mathcal{L}_h(\boldsymbol{\Theta}^*)\|_\infty \\ &\leq C\left(\sqrt{\frac{\log(p \vee N)}{N}} + \left(\frac{h}{|\beta_{1,\min}^*|}\right)^2\right). \end{aligned}$$

Now consider the deviation block. For each $j \in [p]$, define $\mathbf{Z}_j := (Z_{j1}, \ldots, Z_{jM}) \in \mathbb{R}^M$ with

$$Z_{jm} := \left(\nabla_{\mathbf{A}} \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) - \nabla_{\mathbf{A}} \mathcal{L}_h(\boldsymbol{\Theta}^*)\right)_{jm} = -\frac{1}{N} \sum_{i=1}^{n_m} \xi_{ij}^{(m)}, \qquad m = 1, \ldots, M.$$

Then $\|\nabla_{\mathbf{A}} \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) - \nabla_{\mathbf{A}} \mathcal{L}_h(\boldsymbol{\Theta}^*)\|_{2,\infty} = \max_{1 \le j \le p} \|\mathbf{Z}_j\|_2$.

For a unit vector $\mathbf{u} := (u_1, \ldots, u_M)^\top \in \mathbb{S}^{M-1}$, $\langle \mathbf{u}, \mathbf{Z}_j \rangle$ is centered sub-Gaussian with

$$\begin{aligned} \|\langle \mathbf{u}, \mathbf{Z}_j \rangle\|_{\psi_2} &= \left\| -\frac{1}{N} \sum_{m=1}^{M} \sum_{i=1}^{n_m} u_m \xi_{ij}^{(m)} \right\|_{\psi_2} \le \frac{C}{N} \left(\sum_{m=1}^{M} \sum_{i=1}^{n_m} u_m^2 \right)^{1/2} = \frac{C}{N} \left(\sum_{m=1}^{M} n_m u_m^2 \right)^{1/2} \\ &\le \frac{C}{N} \sqrt{\max_{1 \le m \le M} n_m} \le \frac{C}{\sqrt{MN}}, \end{aligned}$$

where the first inequality is because $\|\xi_{ij}^{(m)}\|_{\psi_2} \le C$ and $\xi_{ij}^{(m)}$ are independent across and within each task, and the last inequality uses the task-balance condition. Therefore, by the tail bound for a sub-Gaussian variable (see, e.g., (2.14) in Vershynin (2018)), for any $u > 0$,

$$P\left(|\langle \mathbf{u}, \mathbf{Z}_j \rangle| > u\right) \le 2 \exp(-CMNu^2). \tag{S4.15}$$

Let $\mathcal{N}_{1/2}$ be a 1/2-net of $\mathbb{S}^{M-1}$ with cardinality $|\mathcal{N}_{1/2}| \le 5^M$, and let $\mathbf{u}^* = \mathbf{Z}_j / \|\mathbf{Z}_j\|_2 \in$

$\mathbb{S}^{M-1}$. There exists $\mathbf{v} \in \mathcal{N}_{1/2}$ such that $|\mathbf{u}^* - \mathbf{v}|_2 \leq 1/2$. Hence,

$$\|\mathbf{Z}_j\|_2 = \langle \mathbf{u}^*, \mathbf{Z}_j \rangle \leq |\langle \mathbf{v}, \mathbf{Z}_j \rangle| + |\langle \mathbf{u}^* - \mathbf{v}, \mathbf{Z}_j \rangle| \leq |\langle \mathbf{v}, \mathbf{Z}_j \rangle| + \frac{1}{2}\|\mathbf{Z}_j\|_2.$$

Rearranging yields

$$\|\mathbf{Z}_j\|_2 \leq 2 \max_{\mathbf{u} \in \mathcal{N}_{1/2}} |\langle \mathbf{u}, \mathbf{Z}_j \rangle|.$$

Hence, by (S4.15) and taking a union bound over $\mathbf{u} \in \mathcal{N}_{1/2}$,

$$P\left(\|\mathbf{Z}_j\|_2 > 2u\right) \leq \sum_{\mathbf{u} \in \mathcal{N}_{1/2}} P\left(|\langle \mathbf{u}, \mathbf{Z}_j \rangle| > u\right) \leq 2 \cdot 5^M \exp(-CMNu^2).$$

Taking a union bound over $j = 1, \ldots, p$ gives

$$P\left(\max_{1 \leq j \leq p} \|\mathbf{Z}_j\|_2 > 2u\right) \leq 2p \cdot 5^M \exp(-CMNu^2).$$

Choose $u \asymp \sqrt{(M + \log(p \vee N))/(MN)}$. Then, with probability at least $1 - (p \vee N)^{-C}$,

$$\max_{1 \leq j \leq p} \|\mathbf{Z}_j\|_2 \leq C\sqrt{\frac{M + \log(p \vee N)}{MN}}.$$

Combining with the bias bound (S4.14) yields

$$\begin{aligned}\|\nabla_{\mathbf{A}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}^*)\|_{2,\infty} &\leq \|\nabla_{\mathbf{A}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) - \nabla_{\mathbf{A}}\mathcal{L}_h(\boldsymbol{\Theta}^*)\|_{2,\infty} + \|\nabla_{\mathbf{A}}\mathcal{L}_h(\boldsymbol{\Theta}^*)\|_{2,\infty} \\ &\leq C\left(\sqrt{\frac{M+\log(p\vee N)}{MN}} + \frac{1}{\sqrt{M}}\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2\right).\end{aligned}$$

Finally, consider the intercept block. Define $\mathbf{g} := \nabla_{\mathbf{b}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) - \nabla_{\mathbf{b}}\mathcal{L}_h(\boldsymbol{\Theta}^*) \in \mathbb{R}^M$ and define

$$\zeta_i^{(m)} := \bar{K}\left(\frac{1 - y_i^{(m)}(\widetilde{\mathbf{x}}_i^{(m)})^\top \boldsymbol{\theta}^{*(m)}}{h}\right) y_i^{(m)} - \mathbb{E}^{(m)}\left[\bar{K}\left(\frac{1 - y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top \boldsymbol{\theta}^{*(m)}}{h}\right) y^{(m)}\right].$$

Then

$$g_m = -\frac{1}{N}\sum_{i=1}^{n_m} \zeta_i^{(m)}, \qquad m = 1, \ldots, M.$$

Since $|\zeta_i^{(m)}| \leq 2$, for any $\mathbf{u} = (u_1, \ldots, u_M)^\top \in \mathbb{S}^{M-1}$, $|-u_m\zeta_i^{(m)}/N|$ is bounded. By Hoeffding's inequality for bounded mean zero random variables, for any $u > 0$,

$$\begin{aligned}P\left(|\langle \mathbf{u}, \mathbf{g}\rangle| \geq u\right) &= P\left(\left|-\sum_{m=1}^{M}\sum_{i=1}^{n_m}\frac{u_m\zeta_i^{(m)}}{N}\right| > u\right) \leq 2\exp\left(-\frac{Cu^2}{\sum_{m,i}(2|u_m|/N)^2}\right) \\ &\leq 2e^{-CMNu^2},\end{aligned}$$

where the last step uses the task-balance condition and $\sum_{m=1}^M u_m^2 = 1$. Applying the

same 1/2-net argument as above, we obtain

$$P(\|\mathbf{g}\| > 2u) \leq 2 \cdot 5^M \exp(-CMNu^2).$$

Take $u \asymp \sqrt{(M + \log(p \vee N))/MN}$, then with probability at least $1 - (p \vee N)^{-C}$,

$$\|\nabla_{\mathbf{b}} \mathcal{L}_{N,h}(\mathbf{\Theta}^*) - \nabla_{\mathbf{b}} \mathcal{L}_h(\mathbf{\Theta}^*)\| = \|\mathbf{g}\|_2 \leq C\sqrt{\frac{M + \log N}{MN}}.$$

Combining with (S4.12) yields

$$\begin{aligned} \|\nabla_{\mathbf{b}} \mathcal{L}_{N,h}(\mathbf{\Theta}^*)\|_2 &\leq \|\nabla_{\mathbf{b}} \mathcal{L}_{N,h}(\mathbf{\Theta}^*) - \nabla_{\mathbf{b}} \mathcal{L}_h(\mathbf{\Theta}^*)\| + \|\nabla_{\mathbf{b}} \mathcal{L}_h(\mathbf{\Theta}^*)\| \\ &\leq C\left(\sqrt{\frac{M + \log(p \vee N)}{MN}} + \frac{1}{\sqrt{M}}\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2\right). \end{aligned}$$

This completes the proof. □

Lemma 4. Define the cone set

$$\begin{aligned} \mathbb{C} := \Big\{ &(\mathbf{u}, \mathbf{v}, \mathbf{\Delta}) \in \mathbb{R}^M \times \mathbb{R}^p \times \mathbb{R}^{p \times M} : \\ &\|\mathbf{v}_{S_\mu^c}\|_1 + \lambda_g \|\mathbf{\Delta}_{S_\alpha^c}\|_{2,1} \leq 3(\|\mathbf{v}_{S_\mu}\|_1 + \lambda_g \|\mathbf{\Delta}_{S_\alpha}\|_{2,1}) + M^{-1/2}\|\mathbf{u}\|_2 \Big\}. \end{aligned}$$

Suppose that

$$\lambda \geq 2 \max\left\{\sqrt{M}\|\nabla_{\mathbf{b}} \mathcal{L}_{N,h}(\mathbf{\Theta}^*)\|, \|\nabla_{\boldsymbol{\mu}} \mathcal{L}_{N,h}(\mathbf{\Theta}^*)\|_\infty\right\}, \quad \lambda\lambda_g \geq 2\|\nabla_{\mathbf{A}} \mathcal{L}_{N,h}(\mathbf{\Theta}^*)\|_{2,\infty},$$

then we have $(\widehat{\boldsymbol{\delta}}_{\mathbf{b}}, \widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Delta}}_{\mathbf{A}}) \in \mathbb{C}$.

Proof of Lemma 4. Since $(\widehat{\mathbf{b}}, \widehat{\boldsymbol{\mu}}, \widehat{\mathbf{A}})$ minimize the penalized objective in (2.3), we have

$$\mathcal{L}_{N,h}(\widehat{\boldsymbol{\Theta}}) + \lambda\Big(\|\widehat{\boldsymbol{\mu}}\|_1 + \lambda_g\|\widehat{\mathbf{A}}\|_{2,1}\Big) \leq \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) + \lambda\Big(\|\boldsymbol{\mu}^*\|_1 + \lambda_g\|\mathbf{A}^*\|_{2,1}\Big)$$

Rearranging gives

$$\mathcal{L}_{N,h}(\widehat{\boldsymbol{\Theta}}) - \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) \leq \lambda(\|\boldsymbol{\mu}^*\|_1 - \|\widehat{\boldsymbol{\mu}}\|_1) + \lambda\lambda_g(\|\mathbf{A}^*\|_{2,1} - \|\widehat{\mathbf{A}}\|_{2,1}).$$

By the convexity of the empirical loss $\mathcal{L}_{N,h}$, expanding the gradient in blocks $(\mathbf{b}, \boldsymbol{\mu}, \mathbf{A})$, we obtain

$$\begin{aligned}
\mathcal{L}_{N,h}(\widehat{\boldsymbol{\Theta}}) - \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) \;&\geq\; -\|\nabla_{\mathbf{b}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}^*)\|_2\|\widehat{\boldsymbol{\delta}}_{\mathbf{b}}\|_2 - \|\nabla_{\boldsymbol{\mu}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}^*)\|_\infty\|\widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\|_1 \\
&\quad -\|\nabla_{\mathbf{A}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}^*)\|_{2,\infty}\|\widehat{\boldsymbol{\Delta}}_{\mathbf{A}}\|_{2,1} \\
&\geq\; -\frac{\lambda}{2\sqrt{M}}\|\widehat{\boldsymbol{\delta}}_{\mathbf{b}}\|_2 - \frac{\lambda}{2}\|\widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\|_1 - \frac{\lambda\lambda_g}{2}\|\widehat{\boldsymbol{\Delta}}_{\mathbf{A}}\|_{2,1},
\end{aligned}$$

where the last inequality holds due to the assumption that

$$\lambda \geq 2\sqrt{M}\|\nabla_{\mathbf{b}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}^*)\|_2, \quad \lambda \geq 2\|\nabla_{\boldsymbol{\mu}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}^*)\|_\infty, \quad \lambda\lambda_g \geq 2\|\nabla_{\mathbf{A}}\mathcal{L}_{N,h}(\boldsymbol{\Theta}^*)\|_{2,\infty}.$$

Therefore, we have

$$\begin{aligned} & -\frac{\lambda}{2\sqrt{M}}\|\widehat{\boldsymbol{\delta}}_{\mathbf{b}}\|_2 - \frac{\lambda}{2}\|\widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\|_1 - \frac{\lambda\lambda_g}{2}\|\widehat{\boldsymbol{\Delta}}_{\mathbf{A}}\|_{2,1} \\ \leq \;& \lambda(\|\boldsymbol{\mu}^*\|_1 - \|\boldsymbol{\mu}^* + \widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\|_1) + \lambda\lambda_g(\|\mathbf{A}^*\|_{2,1} - \|\mathbf{A}^* + \widehat{\boldsymbol{\Delta}}_{\mathbf{A}}\|_{2,1}). \end{aligned} \tag{S4.16}$$

Since $\boldsymbol{\mu}^*_{S_\mu^c} = \mathbf{0}$, we have

$$\|\boldsymbol{\mu}^* + \widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\|_1 \geq \|\boldsymbol{\mu}^*_{S_\mu}\|_1 - \|\widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu},S_\mu}\|_1 + \|\widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu},S_\mu^c}\|_1.$$

Hence,

$$\|\boldsymbol{\mu}^*\|_1 - \|\boldsymbol{\mu}^* + \widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\|_1 \leq \|\widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu},S_\mu}\|_1 - \|\widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu},S_\mu^c}\|_1. \tag{S4.17}$$

Similarly, for the row-wise group norm, applying the triangle inequality row by row yields

$$\|\mathbf{A}^* + \widehat{\boldsymbol{\Delta}}_{\mathbf{A}}\|_{2,1} = \sum_{j=1}^{p} \|\mathbf{A}^*_{j\cdot} + (\widehat{\boldsymbol{\Delta}}_{\mathbf{A}})_{j\cdot}\|_2 \geq \sum_{j\in S_\alpha} \big(\|\mathbf{A}^*_{j\cdot}\|_2 - \|(\widehat{\boldsymbol{\Delta}}_{\mathbf{A}})_{j\cdot}\|_2\big) + \sum_{j\in S_\alpha^c} \|(\widehat{\boldsymbol{\Delta}}_{\mathbf{A}})_{j\cdot}\|_2,$$

since $\mathbf{A}^*_{j\cdot} = 0$ for all $j \in S_\alpha^c$. Therefore,

$$\|\mathbf{A}^*\|_{2,1} - \|\mathbf{A}^* + \widehat{\boldsymbol{\Delta}}_{\mathbf{A}}\|_{2,1} \leq \|\widehat{\boldsymbol{\Delta}}_{\mathbf{A},S_\alpha}\|_{2,1} - \|\widehat{\boldsymbol{\Delta}}_{\mathbf{A},S_\alpha^c}\|_{2,1}. \tag{S4.18}$$

Substituting (S4.17) and (S4.18) into the basic inequality (S4.16) and rearranging

terms gives

$$\|\widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu},S_\mu^c}\|_1 + \lambda_g \|\widehat{\boldsymbol{\Delta}}_{\mathbf{A},S_\alpha^c}\|_{2,1} \le 3(\|\widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu},S_\mu}\|_1 + \lambda_g \|\widehat{\boldsymbol{\Delta}}_{\mathbf{A},S_\alpha}\|_{2,1}) + \frac{1}{\sqrt{M}}\|\widehat{\boldsymbol{\delta}}_{\mathbf{b}}\|_2,$$

which completes the proof. □

Lemma 5. Suppose that Condition 1 holds, the task sizes are balanced in the sense that

$$\frac{c_-}{M} \le \frac{n_m}{N} \le \frac{c_+}{M}, \qquad m \in [M], \tag{S4.19}$$

for some constants $c_-, c_+ > 0$, and $\lambda_g \asymp M^{-1/2}$. Define $r_s = \sqrt{s_\mu + s_\alpha}$. Then, for any $t > 0$, if $\log^3(p \vee N) = O(N)$, with probability at least $1 - (p \vee N)^{-C}$

$$\sup_{\substack{(\boldsymbol{\delta}_{\mathbf{b}},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_{\mathbf{A}})\in\mathbb{C} \\ M^{-1}\|\boldsymbol{\delta}_{\mathbf{b}}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_{\mathbf{A}}\|_F^2\le t^2}} \left|(P_N - P)\big\{\ell_h(\boldsymbol{\Theta}) - \ell_h(\boldsymbol{\Theta}^*)\big\}\right| \le C t r_s \sqrt{\frac{M + \log(p \vee N)}{N}},$$

where $(\boldsymbol{\delta}_{\mathbf{b}}, \boldsymbol{\delta}_\mu, \boldsymbol{\Delta}_{\mathbf{A}}) := \boldsymbol{\Theta} - \boldsymbol{\Theta}^* = (\mathbf{b} - \mathbf{b}^*, \boldsymbol{\mu} - \boldsymbol{\mu}^*, \mathbf{A} - \mathbf{A}^*)$.

Proof of Lemma 5. Let $\{\sigma_i^{(m)}\}$ be i.i.d. Rademacher random variables independent

of the data. Since the observations $\{(\mathbf{x}_i^{(m)}, y_i^{(m)})\}$ are mutually independent, we have

$$
\begin{aligned}
& \mathbb{E}\left[\sup_{\substack{(\boldsymbol{\delta}_\mathbf{b},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_\mathbf{A})\in\mathbb{C} \\ M^{-1}\|\boldsymbol{\delta}_\mathbf{b}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_\mathbf{A}\|_F^2\le t^2}} \left|(P_N - P)\big\{\ell_h(\boldsymbol{\Theta}) - \ell_h(\boldsymbol{\Theta}^*)\big\}\right|\right] \\
\le\ & 2\mathbb{E}\left[\sup_{\substack{(\boldsymbol{\delta}_\mathbf{b},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_\mathbf{A})\in\mathbb{C} \\ M^{-1}\|\boldsymbol{\delta}_\mathbf{b}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_\mathbf{A}\|_F^2\le t^2}} \left|\frac{1}{N}\sum_{m=1}^{M}\sum_{i=1}^{n_m}\sigma_i^{(m)}\big\{\ell_h^{(m)}(\boldsymbol{\Theta};\mathbf{x}_i^{(m)},y_i^{(m)}) - \ell_h^{(m)}(\boldsymbol{\Theta}^*;\mathbf{x}_i^{(m)},y_i^{(m)})\big\}\right|\right] \\
\le\ & 2\mathbb{E}\left[\sup_{\substack{(\boldsymbol{\delta}_\mathbf{b},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_\mathbf{A})\in\mathbb{C} \\ M^{-1}\|\boldsymbol{\delta}_\mathbf{b}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_\mathbf{A}\|_F^2\le t^2}} \left|\frac{1}{N}\sum_{m=1}^{M}\sum_{i=1}^{n_m}\sigma_i^{(m)}y_i^{(m)}\Big((\mathbf{x}_i^{(m)})^\top(\boldsymbol{\delta}_{\boldsymbol{\mu}}+\boldsymbol{\delta}_{\boldsymbol{\alpha}^{(m)}})+\delta_{b^{(m)}}\Big)\right|\right],
\end{aligned}
$$

where the last inequality is by applying the contraction inequality for the Rademacher complexity (e.g., Theorem 2.2 of Koltchinskii (2011)) since ϕ_h is 1-Lipschitz continuous. It remains to bound the symmetrized linear process. We decompose

$$
\begin{aligned}
& \frac{1}{N}\sum_{m=1}^{M}\sum_{i=1}^{n_m}\sigma_i^{(m)}y_i^{(m)}\Big((\mathbf{x}_i^{(m)})^\top(\boldsymbol{\delta}_{\boldsymbol{\mu}}+\boldsymbol{\delta}_{\boldsymbol{\alpha}^{(m)}})+\boldsymbol{\delta}_\mathbf{b}^{(m)}\Big) \\
=\ & \left\langle\frac{1}{N}\sum_{m=1}^{M}\sum_{i=1}^{n_m}\sigma_i^{(m)}y_i^{(m)}\mathbf{x}_i^{(m)},\boldsymbol{\delta}_{\boldsymbol{\mu}}\right\rangle + \sum_{m=1}^{M}\left\langle\frac{1}{N}\sum_{i=1}^{n_m}\sigma_i^{(m)}y_i^{(m)}\mathbf{x}_i^{(m)},\boldsymbol{\delta}_{\boldsymbol{\alpha}^{(m)}}\right\rangle \\
& + \sum_{m=1}^{M}\left(\frac{1}{N}\sum_{i=1}^{n_m}\sigma_i^{(m)}y_i^{(m)}\right)\delta_{b^{(m)}} \\
\le\ & \left\|\frac{1}{N}\sum_{m,i}\sigma_i^{(m)}y_i^{(m)}\mathbf{x}_i^{(m)}\right\|_\infty\|\boldsymbol{\delta}_{\boldsymbol{\mu}}\|_1 + \max_{1\le j\le p}\left(\sum_{m=1}^{M}\left|\frac{1}{N}\sum_{i=1}^{n_m}\sigma_i^{(m)}y_i^{(m)}x_{i,j}^{(m)}\right|^2\right)^{1/2}\|\boldsymbol{\Delta}_\mathbf{A}\|_{2,1} \\
& + \left(\sum_{m=1}^{M}\left|\frac{1}{N}\sum_{i=1}^{n_m}\sigma_i^{(m)}y_i^{(m)}\right|^2\right)^{1/2}\|\boldsymbol{\delta}_\mathbf{b}\|.
\end{aligned}
$$

Taking supremum and expectation on both sides, we obtain

$$
\begin{aligned}
& \mathbb{E}\left[\sup_{\substack{(\boldsymbol{\delta}_{\mathbf{b}},\boldsymbol{\delta}_{\mu},\boldsymbol{\Delta}_{\mathbf{A}})\in\mathbb{C}\\ M^{-1}\|\boldsymbol{\delta}_{\mathbf{b}}\|^2+\|\boldsymbol{\delta}_{\mu}\|^2+M^{-1}\|\boldsymbol{\Delta}_{\mathbf{A}}\|_F^2\le t^2}} \left|(P_N-P)\big\{\ell_h(\boldsymbol{\Theta})-\ell_h(\boldsymbol{\Theta}^*)\big\}\right|\right] \\
\le\ & 2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{m,i}\sigma_i^{(m)}y_i^{(m)}\mathbf{x}_i^{(m)}\right\|_{\infty}\right] \sup_{\substack{(\boldsymbol{\delta}_{\mathbf{b}},\boldsymbol{\delta}_{\mu},\boldsymbol{\Delta}_{\mathbf{A}})\in\mathbb{C}\\ M^{-1}\|\boldsymbol{\delta}_{\mathbf{b}}\|^2+\|\boldsymbol{\delta}_{\mu}\|^2+M^{-1}\|\boldsymbol{\Delta}_{\mathbf{A}}\|_F^2\le t^2}} \|\boldsymbol{\delta}_{\boldsymbol{\mu}}\|_1 \\
& +2\mathbb{E}\left[\max_{1\le j\le p}\left(\sum_{m=1}^{M}\left|\frac{1}{N}\sum_{i=1}^{n_m}\sigma_i^{(m)}y_i^{(m)}x_{i,j}^{(m)}\right|^2\right)^{1/2}\right] \sup_{\substack{(\boldsymbol{\delta}_{\mathbf{b}},\boldsymbol{\delta}_{\mu},\boldsymbol{\Delta}_{\mathbf{A}})\in\mathbb{C}\\ M^{-1}\|\boldsymbol{\delta}_{\mathbf{b}}\|^2+\|\boldsymbol{\delta}_{\mu}\|^2+M^{-1}\|\boldsymbol{\Delta}_{\mathbf{A}}\|_F^2\le t^2}} \|\boldsymbol{\Delta}_{\mathbf{A}}\|_{2,1} \\
& +2\mathbb{E}\left[\left(\sum_{m=1}^{M}\left|\frac{1}{N}\sum_{i=1}^{n_m}\sigma_i^{(m)}y_i^{(m)}\right|^2\right)^{1/2}\right] \sup_{\substack{(\boldsymbol{\delta}_{\mathbf{b}},\boldsymbol{\delta}_{\mu},\boldsymbol{\Delta}_{\mathbf{A}})\in\mathbb{C}\\ M^{-1}\|\boldsymbol{\delta}_{\mathbf{b}}\|^2+\|\boldsymbol{\delta}_{\mu}\|^2+M^{-1}\|\boldsymbol{\Delta}_{\mathbf{A}}\|_F^2\le t^2}} \|\boldsymbol{\delta}_{\mathbf{b}}\|. \qquad \text{(S4.20)}
\end{aligned}
$$

We first bound the deterministic suprema under the cone constraint. For $(\boldsymbol{\delta}_{\mathbf{b}},\boldsymbol{\delta}_{\boldsymbol{\mu}},\boldsymbol{\Delta}_{\mathbf{A}})\in\mathbb{C}$, we have

$$
\|\boldsymbol{\delta}_{\boldsymbol{\mu}}\|_1+\lambda_g\|\boldsymbol{\Delta}_{\mathbf{A}}\|_{2,1}\le 4\left(\|\boldsymbol{\delta}_{\boldsymbol{\mu},S_\mu}\|_1+\lambda_g\|\boldsymbol{\Delta}_{\mathbf{A},S_\alpha}\|_{2,1}\right)+\frac{1}{\sqrt{M}}\|\boldsymbol{\delta}_{\mathbf{b}}\|_2.
$$

By the Cauchy–Schwarz inequality and the definition of the active sets S_μ and S_α, we have $\|\boldsymbol{\delta}_{\boldsymbol{\mu},S_\mu}\|_1\le\sqrt{s_\mu}\|\boldsymbol{\delta}_{\boldsymbol{\mu}}\|$, and $\|\boldsymbol{\Delta}_{\mathbf{A},S_\alpha}\|_{2,1}\le\sqrt{s_\alpha}\|\boldsymbol{\Delta}_{\mathbf{A}}\|_F$. Moreover, under

$$
M^{-1}\|\boldsymbol{\delta}_{\mathbf{b}}\|^2+\|\boldsymbol{\delta}_{\boldsymbol{\mu}}\|^2+M^{-1}\|\boldsymbol{\Delta}_{\mathbf{A}}\|_F^2\le t^2,
$$

we have $\|\boldsymbol{\delta}_{\mathbf{b}}\|_2 \leq \sqrt{M}t$, $\|\boldsymbol{\delta}_{\boldsymbol{\mu}}\| \leq t$ and $\|\boldsymbol{\Delta}_{\mathbf{A}}\|_F \leq \sqrt{M}t$. Hence,

$$\begin{aligned} \|\boldsymbol{\delta}_{\boldsymbol{\mu}}\|_1 + \lambda_g \|\boldsymbol{\Delta}_{\mathbf{A}}\|_{2,1} &\leq 4\sqrt{s_\mu}t + 4\lambda_g\sqrt{s_\alpha}\sqrt{M}t + t \\ &\leq Cr_s t. \end{aligned}$$

In particular,

$$\sup_{\substack{(\boldsymbol{\delta}_{\mathbf{b}},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_{\mathbf{A}})\in\mathbb{C} \\ M^{-1}\|\boldsymbol{\delta}_{\mathbf{b}}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_{\mathbf{A}}\|_F^2\leq t^2}} \|\boldsymbol{\delta}_{\boldsymbol{\mu}}\|_1 \leq Cr_s t, \tag{S4.21}$$

and

$$\sup_{\substack{(\boldsymbol{\delta}_{\mathbf{b}},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_{\mathbf{A}})\in\mathbb{C} \\ M^{-1}\|\boldsymbol{\delta}_{\mathbf{b}}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_{\mathbf{A}}\|_F^2\leq t^2}} \|\boldsymbol{\Delta}_{\mathbf{A}}\|_{2,1} \leq Cr_s\sqrt{M}t, \tag{S4.22}$$

since $\lambda_g \asymp M^{-1/2}$. Moreover,

$$\sup_{\substack{(\boldsymbol{\delta}_{\mathbf{b}},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_{\mathbf{A}})\in\mathbb{C} \\ M^{-1}\|\boldsymbol{\delta}_{\mathbf{b}}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_{\mathbf{A}}\|_F^2\leq t^2}} \|\boldsymbol{\delta}_{\mathbf{b}}\| \leq \sqrt{M}t. \tag{S4.23}$$

Next, we bound the three stochastic coefficients in (S4.20) For the shared block, by condition 1 and the standard maximal inequalities (see, e.g., Exercise 2.5.10 of Vershynin (2018)),

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{m,i}\sigma_i^{(m)}y_i^{(m)}\mathbf{x}_i^{(m)}\right\|_\infty\right] \leq CK_x\sqrt{\frac{\log p}{N}}. \tag{S4.24}$$

For the deviation block, let

$$Z_{jm} := \frac{1}{N}\sum_{i=1}^{n_m} \sigma_i^{(m)} y_i^{(m)} x_{ij}^{(m)}, \qquad \mathbf{Z}_{j\cdot} := (Z_{j1}, \ldots, Z_{jM})^\top.$$

For any $\mathbf{u} = (u_1, \ldots, u_M)^\top \in \mathbb{S}^{M-1}$,

$$\begin{aligned} \|\langle \mathbf{u}, \mathbf{Z}_{j\cdot}\rangle\|_{\psi_2} &= \left\| \frac{1}{N}\sum_{m=1}^{M}\sum_{i=1}^{n_m} u_m \sigma_i^{(m)} y_i^{(m)} x_{ij}^{(m)} \right\|_{\psi_2} \le \frac{CK_x}{N}\left(\sum_{m=1}^{M}\sum_{i=1}^{n_m} u_m^2\right)^{1/2} \\ &= \frac{CK_x}{N}\left(\sum_{m=1}^{M} n_m u_m^2\right)^{1/2} \le \frac{CK_x}{\sqrt{MN}}, \end{aligned}$$

where the last step uses (S4.19). Let $\mathcal{N}_{1/2}$ be a 1/2-net of $\mathbb{S}^{M-1}$ with $|\mathcal{N}_{1/2}| \le 5^M$. By the standard net argument,

$$P\left(\|\mathbf{Z}_{j\cdot}\|_2 > 2x\right) \le 2 \cdot 5^M \exp(-CMNx^2).$$

A union bound over $j \in [p]$ then gives

$$P\left(\max_{1\le j\le p} \|\mathbf{Z}_{j\cdot}\|_2 > 2x\right) \le 2p \cdot 5^M \exp(-CMNx^2).$$

Integrating this tail bound yields

$$\mathbb{E}\left[\max_{1\le j\le p}\left(\sum_{m=1}^{M}\left|\frac{1}{N}\sum_{i=1}^{n_m}\sigma_i^{(m)} y_i^{(m)} x_{i,j}^{(m)}\right|^2\right)^{1/2}\right] \le CK_x\sqrt{\frac{M+\log p}{MN}}. \tag{S4.25}$$

For the intercept block, by Jensen's inequality,

$$\mathbb{E}\left[\left(\sum_{m=1}^{M}\left|\frac{1}{N}\sum_{i=1}^{n_m}\sigma_i^{(m)}y_i^{(m)}\right|^2\right)^{1/2}\right] \leq \left(\sum_{m=1}^{M}\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{n_m}\sigma_i^{(m)}y_i^{(m)}\right)^2\right)^{1/2}$$
$$= \left(\sum_{m=1}^{M}\frac{n_m}{N^2}\right)^{1/2} = \frac{1}{\sqrt{N}}. \tag{S4.26}$$

Substituting (S4.21)–(S4.26) into (S4.20), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\substack{(\boldsymbol{\delta}_\mathbf{b},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_\mathbf{A})\in\mathbb{C}\\ M^{-1}\|\boldsymbol{\delta}_\mathbf{b}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_\mathbf{A}\|_F^2\leq t^2}}\left|(P_N-P)\big\{\ell_h(\boldsymbol{\Theta})-\ell_h(\boldsymbol{\Theta}^*)\big\}\right|\right]\\
\leq\ & CK_x r_s t\sqrt{\frac{\log p}{N}} + CK_x r_s\sqrt{M}t\sqrt{\frac{M+\log p}{MN}} + C\sqrt{M}t\cdot\frac{1}{\sqrt{N}}\\
\leq\ & Ctr_s\sqrt{\frac{M+\log p}{N}}.
\end{aligned}$$

We now turn to concentration. Applying Adamczak's concentration inequality (see, for example, Section 2.3 of Koltchinskii (2011)), we obtain that, with probability at least $1-e^{-z}$,

$$\begin{aligned}
&\sup_{\substack{(\boldsymbol{\delta}_\mathbf{b},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_\mathbf{A})\in\mathbb{C}\\ M^{-1}\|\boldsymbol{\delta}_\mathbf{b}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_\mathbf{A}\|_F^2\leq t^2}}\left|(P_N-P)\big\{\ell_h(\boldsymbol{\Theta})-\ell_h(\boldsymbol{\Theta}^*)\big\}\right|\\
\leq\ & C\mathbb{E}\left[\sup_{\substack{(\boldsymbol{\delta}_\mathbf{b},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_\mathbf{A})\in\mathbb{C}\\ M^{-1}\|\boldsymbol{\delta}_\mathbf{b}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_\mathbf{A}\|_F^2\leq t^2}}\left|(P_N-P)\big\{\ell_h(\boldsymbol{\Theta})-\ell_h(\boldsymbol{\Theta}^*)\big\}\right|\right]\\
&+C\sqrt{\frac{Bz}{N}}+C\frac{Uz}{N},
\end{aligned} \tag{S4.27}$$

where

$$B := \sup_{\substack{(\boldsymbol{\delta}_\mathbf{b},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_\mathbf{A})\in\mathbb{C}\\ M^{-1}\|\boldsymbol{\delta}_\mathbf{b}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_\mathbf{A}\|_F^2\le t^2}} \mathbb{E}\left[(\ell_h(\boldsymbol{\Theta})-\ell_h(\boldsymbol{\Theta}^*))^2\right],$$

$$U := \left\| \max_{\substack{1\le m\le M,\\ 1\le i\le n_m}} \sup_{\substack{(\boldsymbol{\delta}_\mathbf{b},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_\mathbf{A})\in\mathbb{C}\\ M^{-1}\|\boldsymbol{\delta}_\mathbf{b}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_\mathbf{A}\|_F^2\le t^2}} |\ell_h^{(m)}(\boldsymbol{\Theta};\mathbf{x}_i^{(m)},y_i^{(m)})-\ell_h^{(m)}(\boldsymbol{\Theta}^*;\mathbf{x}_i^{(m)},y_i^{(m)})| \right\|_{\psi_1}.$$

Since ϕ_h is 1-Lipschitz and Condition 1 implies $\lambda_{\max}(\mathbb{E}^{(m)}[\mathbf{x}^{(m)}(\mathbf{x}^{(m)})^\top]) \le C$ uniformly in m,

$$\begin{aligned} B &\le \sup_{\substack{(\boldsymbol{\delta}_\mathbf{b},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_\mathbf{A})\in\mathbb{C}\\ M^{-1}\|\boldsymbol{\delta}_\mathbf{b}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_\mathbf{A}\|_F^2\le t^2}} \frac{1}{N}\sum_{m=1}^M n_m\mathbb{E}^{(m)}\left[\left((\mathbf{x}^{(m)})^\top(\boldsymbol{\delta}_{\boldsymbol{\mu}}+\boldsymbol{\delta}_{\boldsymbol{\alpha}^{(m)}})+\delta_{b^{(m)}}\right)^2\right] \\ &\le C \sup_{\substack{(\boldsymbol{\delta}_\mathbf{b},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_\mathbf{A})\in\mathbb{C}\\ M^{-1}\|\boldsymbol{\delta}_\mathbf{b}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_\mathbf{A}\|_F^2\le t^2}} \left\{\|\boldsymbol{\delta}_{\boldsymbol{\mu}}\|^2+\frac{1}{N}\sum_{m=1}^M n_m\|\boldsymbol{\delta}_{\boldsymbol{\alpha}^{(m)}}\|^2+\frac{1}{N}\sum_{m=1}^M n_m\delta_{b^{(m)}}^2\right\} \\ &\le Ct^2, \end{aligned}$$

where the last step uses (S4.19).

To bound U, note that

$$\begin{aligned} |\ell_h^{(m)}(\boldsymbol{\Theta};\mathbf{x},y)-\ell_h^{(m)}(\boldsymbol{\Theta}^*;\mathbf{x},y)| &\le |\mathbf{x}^\top(\boldsymbol{\delta}_{\boldsymbol{\mu}}+\boldsymbol{\delta}_{\boldsymbol{\alpha}^{(m)}})+\delta_{b^{(m)}}| \\ &\le \|\mathbf{x}\|_\infty(\|\boldsymbol{\delta}_{\boldsymbol{\mu}}\|_1+\|\boldsymbol{\delta}_{\boldsymbol{\alpha}^{(m)}}\|_1)+|\delta_{b^{(m)}}| \\ &\le \|\mathbf{x}\|_\infty(\|\boldsymbol{\delta}_{\boldsymbol{\mu}}\|_1+\|\boldsymbol{\Delta}_\mathbf{A}\|_{2,1})+\|\boldsymbol{\delta}_\mathbf{b}\|. \end{aligned}$$

Using (S4.21), (S4.22), (S4.23), and by Lemma 2.2.2 of van der Vaart and Wellner

(1996), we have

$$\left\| \max_{1\le m\le M,\ 1\le i\le n_m} \|\mathbf{x}_i^{(m)}\|_\infty \right\|_{\psi_2} \le CK_x\sqrt{\log(pN)},$$

then,

$$\begin{aligned} U &\le CK_x\sqrt{\log(pN)}(r_s t + \sqrt{M} r_s t) + C\sqrt{M}t \\ &\le Ct\sqrt{M} r_s \sqrt{\log(p\vee N)}. \end{aligned}$$

Next, substituting the above upper bound for B and U into (S4.27) and setting $z \asymp \log(p\vee N)$, we obtain

$$\begin{aligned} &\sup_{\substack{(\boldsymbol{\delta}_\mathbf{b},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_\mathbf{A})\in\mathbb{C} \\ M^{-1}\|\boldsymbol{\delta}_\mathbf{b}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_\mathbf{A}\|_F^2\le t^2}} |(P_N - P)\{\ell_h(\boldsymbol{\Theta}) - \ell_h(\boldsymbol{\Theta}^*)\}| \\ \le\ & Ctr_s\sqrt{\frac{M+\log(p\vee N)}{N}} + Ct\sqrt{\frac{\log(p\vee N)}{N}} + Ct\sqrt{M} r_s \frac{\log^{3/2}(p\vee N)}{N}. \end{aligned}$$

Since $r_s \ge 1$, the second term is dominated by the first one. Moreover, under the condition $\log^3(p\vee N) = O(N)$, the third term is also dominated by the first one. Therefore

$$\sup_{\substack{(\boldsymbol{\delta}_\mathbf{b},\boldsymbol{\delta}_\mu,\boldsymbol{\Delta}_\mathbf{A})\in\mathbb{C} \\ M^{-1}\|\boldsymbol{\delta}_\mathbf{b}\|^2+\|\boldsymbol{\delta}_\mu\|^2+M^{-1}\|\boldsymbol{\Delta}_\mathbf{A}\|_F^2\le t^2}} |(P_N - P)\{\ell_h(\boldsymbol{\Theta}) - \ell_h(\boldsymbol{\Theta}^*)\}| \le Ctr_s\sqrt{\frac{M+\log(p\vee N)}{N}}$$

with probability at least $1-(p\vee N)^{-C}$. □

Lemma 6. Suppose that the task-balance condition holds, namely,

$$\frac{c_-}{M} \le \frac{n_m}{N} \le \frac{c_+}{M}, \qquad m \in [M],$$

for some constants $c_-, c_+ > 0$. If the task-specific components satisfy the centering condition

$$\sum_{m=1}^{M} \boldsymbol{\alpha}^{(m)} = 0,$$

then, for any $\mathbf{u} = (u_1, \ldots, u_M)^\top \in \mathbb{R}^M$, $\mathbf{v} \in \mathbb{R}^p$ and $\mathbf{A} = (\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(M)})$, we have

$$\begin{aligned} c\left(M^{-1}\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 + M^{-1}\|\mathbf{A}\|_F^2\right) &\le \sum_{m=1}^{M} \frac{n_m}{N}\left\{u_m^2 + \|\mathbf{v} + \boldsymbol{\alpha}^{(m)}\|_2^2\right\} \\ &\le C\left(M^{-1}\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 + M^{-1}\|\mathbf{A}\|_F^2\right), \end{aligned}$$

where $c > 0$ depends only on c_- and $C > 0$ depends only on c_+.

Proof of Lemma 6. By the task-balance condition,

$$\frac{c_-}{M}\sum_{m=1}^{M} u_m^2 \le \sum_{m=1}^{M} \frac{n_m}{N} u_m^2 \le \frac{c_+}{M}\sum_{m=1}^{M} u_m^2.$$

That is,

$$c_- M^{-1}\|\mathbf{u}\|_2^2 \le \sum_{m=1}^{M} \frac{n_m}{N} u_m^2 \le c_+ M^{-1}\|\mathbf{u}\|_2^2.$$

Next, by the centering condition,

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}\|\mathbf{v}+\boldsymbol{\alpha}^{(m)}\|_2^2 &= \frac{1}{M}\sum_{m=1}^{M}\left\{\|\mathbf{v}\|_2^2+2\mathbf{v}^\top\boldsymbol{\alpha}^{(m)}+\|\boldsymbol{\alpha}^{(m)}\|_2^2\right\}\\ &= \|\mathbf{v}\|_2^2+\frac{2}{M}\mathbf{v}^\top\sum_{m=1}^{M}\boldsymbol{\alpha}^{(m)}+\frac{1}{M}\sum_{m=1}^{M}\|\boldsymbol{\alpha}^{(m)}\|_2^2\\ &= \|\mathbf{v}\|_2^2+\frac{1}{M}\sum_{m=1}^{M}\|\boldsymbol{\alpha}^{(m)}\|_2^2\\ &= \|\mathbf{v}\|_2^2+M^{-1}\|\mathbf{A}\|_F^2.\end{aligned}$$

Therefore,

$$\sum_{m=1}^{M}\frac{n_m}{N}\|\mathbf{v}+\boldsymbol{\alpha}^{(m)}\|_2^2 \geq \frac{c_-}{M}\sum_{m=1}^{M}\|\mathbf{v}+\boldsymbol{\alpha}^{(m)}\|_2^2 = c_-\left(\|\mathbf{v}\|_2^2+M^{-1}\|\mathbf{A}\|_F^2\right),$$

and similarly,

$$\sum_{m=1}^{M}\frac{n_m}{N}\|\mathbf{v}+\boldsymbol{\alpha}^{(m)}\|_2^2 \leq \frac{c_+}{M}\sum_{m=1}^{M}\|\mathbf{v}+\boldsymbol{\alpha}^{(m)}\|_2^2 = c_+\left(\|\mathbf{v}\|_2^2+M^{-1}\|\mathbf{A}\|_F^2\right).$$

Combining the above bounds proves the lemma. □

Lemma 7. Suppose that Conditions 1–5 hold. Let

$$d_N = \sqrt{\frac{M+\log(p\vee N)}{N}}+\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2.$$

Then

$$\left\|\nabla_{\boldsymbol{\vartheta}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*)\right\|_{\mathcal{D},+} = O_p(d_N).$$

Proof of Lemma 7. The argument is the reparameterized version of Lemma 3. Indeed, the intercept, shared, and deviation components of $\nabla_{\boldsymbol{\vartheta}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*)$ are linear combinations of the corresponding taskwise empirical scores considered there. The norm $\|\cdot\|_{\mathcal{D},+}$ rescales these blocks exactly as in Lemma 3, with $\lambda_g \asymp M^{-1/2}$ for the deviation rows. Combining the same sub-Gaussian concentration bounds with the smoothing-bias bound from Lemma 1 gives $O_p(d_N)$. □

Lemma 8. Let $\mathcal{N}_N$ be the neighborhood of $\boldsymbol{\vartheta}^*$ defined by

$$\mathcal{N}_N = \left\{\boldsymbol{\vartheta} : M^{-1/2}\|\mathbf{b}-\mathbf{b}^*\|_2 + \|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|_2 + M^{-1/2}\|\mathbf{A}-\mathbf{A}^*\|_F \le Cr_s\lambda\right\},$$

where $C > 0$ is sufficiently large. Under Conditions 1–5 and the bandwidth conditions ensuring local smoothness of the smoothed loss, uniformly over $\boldsymbol{\vartheta} \in \mathcal{N}_N$,

$$\left\|\nabla_{\boldsymbol{\vartheta}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}) - \nabla_{\boldsymbol{\vartheta}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) - \boldsymbol{\Gamma}_h(\boldsymbol{\vartheta}^*)(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^*)\right\|_{\mathcal{D},+} = o_p(\lambda).$$

Consequently, if $\boldsymbol{\vartheta}_{\mathcal{I}^c} = \boldsymbol{\vartheta}^*_{\mathcal{I}^c} = 0$, then, by taking the corresponding subvectors of the preceding expansion,

$$\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}) = \nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*)(\boldsymbol{\vartheta}_{\mathcal{I}} - \boldsymbol{\vartheta}^*_{\mathcal{I}}) + R_{\mathcal{I}}(\boldsymbol{\vartheta}),$$

and

$$\nabla_{\mathcal{I}^c}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}) = \nabla_{\mathcal{I}^c}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*)(\boldsymbol{\vartheta}_{\mathcal{I}} - \boldsymbol{\vartheta}^*_{\mathcal{I}}) + R_{\mathcal{I}^c}(\boldsymbol{\vartheta}),$$

where

$$\|R_{\mathcal{I}}(\boldsymbol{\vartheta})\|_{\mathcal{D},+} = o_p(\lambda), \qquad \|R_{\mathcal{I}^c}(\boldsymbol{\vartheta})\|_{\mathcal{D}} = o_p(\lambda).$$

Proof of Lemma 8. Let $B_N = C\sqrt{\log(p \vee N)}$ be such that, with probability tending to one,

$$\max_{1\le m\le M}\max_{1\le i\le n_m} \|\widetilde{\mathbf{x}}_i^{(m)}\|_\infty \le B_N.$$

This event follows from Condition 1 by a union bound for sub-Gaussian variables. In the proof of Theorem 2, the remainder is applied only to the restricted/oracle estimator satisfying $\boldsymbol{\vartheta}_{\mathcal{I}^c} = \boldsymbol{\vartheta}^*_{\mathcal{I}^c} = 0$. Hence $\boldsymbol{\delta} = \boldsymbol{\vartheta} - \boldsymbol{\vartheta}^*$ is supported on the active block, whose dimension is $q_N = M + s_\mu + (M-1)s_\alpha$. If one wants the stronger uniform expansion over all of $\mathcal{N}_N$, the same argument requires replacing q_N by the ambient dimension $M(p+1)$. By the bandwidth condition imposed in Theorem 2,

$$\frac{B_N^3 M^2 q_N r_s^2 \lambda}{h} = o(1). \tag{S4.28}$$

For each observation in task m, write the margin as a linear function of the

reparameterized vector:

$$y_i^{(m)}(\widetilde{\mathbf{x}}_i^{(m)})^\top \boldsymbol{\theta}^{(m)}(\boldsymbol{\vartheta}) = (\boldsymbol{\psi}_i^{(m)})^\top \boldsymbol{\vartheta},$$

where $\boldsymbol{\psi}_i^{(m)}$ is the corresponding reparameterized design vector multiplied by $y_i^{(m)}$. More explicitly, $\boldsymbol{\psi}_i^{(m)}$ has the intercept component $y_i^{(m)}$ in the mth intercept coordinate, the shared block $y_i^{(m)}\mathbf{x}_i^{(m)}$, and deviation blocks equal to $-y_i^{(1)}\mathbf{x}_i^{(1)}$ for all $\mathbf{a}_j$-blocks when $m = 1$, and equal to $y_i^{(m)}\mathbf{x}_i^{(m)}$ in the mth deviation block when $m \geq 2$, with zeros elsewhere. Hence, on the event above,

$$\|\boldsymbol{\psi}_i^{(m)}\|_\infty \leq B_N, \qquad 1 \leq i \leq n_m,\ 1 \leq m \leq M.$$

Let $\boldsymbol{\delta} = \boldsymbol{\vartheta} - \boldsymbol{\vartheta}^*$. By Taylor's formula applied to ϕ_h' for each observation,

$$\begin{aligned}
&\nabla_{\boldsymbol{\vartheta}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}) - \nabla_{\boldsymbol{\vartheta}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) - \boldsymbol{\Gamma}_h(\boldsymbol{\vartheta}^*)\boldsymbol{\delta} \\
= \;& \frac{1}{N}\sum_{m=1}^{M}\sum_{i=1}^{n_m} \boldsymbol{\psi}_i^{(m)} \left[\phi_h'\{(\boldsymbol{\psi}_i^{(m)})^\top\boldsymbol{\vartheta}\} - \phi_h'\{(\boldsymbol{\psi}_i^{(m)})^\top\boldsymbol{\vartheta}^*\} - \phi_h''\{(\boldsymbol{\psi}_i^{(m)})^\top\boldsymbol{\vartheta}^*\}(\boldsymbol{\psi}_i^{(m)})^\top\boldsymbol{\delta}\right] \\
= \;& \frac{1}{N}\sum_{m=1}^{M}\sum_{i=1}^{n_m} \boldsymbol{\psi}_i^{(m)} \left\{((\boldsymbol{\psi}_i^{(m)})^\top\boldsymbol{\delta})^2 \int_0^1 (1-t)\phi_h'''\left((\boldsymbol{\psi}_i^{(m)})^\top\boldsymbol{\vartheta}^* + t(\boldsymbol{\psi}_i^{(m)})^\top\boldsymbol{\delta}\right) dt\right\}.
\end{aligned}$$

Denote the preceding vector by $R(\boldsymbol{\vartheta})$. Then, for every coordinate ℓ,

$$|R_\ell(\boldsymbol{\vartheta})| \leq \frac{C}{Nh}\sum_{m=1}^{M}\sum_{i=1}^{n_m} |\psi_{i\ell}^{(m)}||(\boldsymbol{\psi}_i^{(m)})^\top\boldsymbol{\delta}|^2. \tag{S4.29}$$

If $\boldsymbol{\delta}$ has at most q_N nonzero coordinates, then

$$|(\boldsymbol{\psi}_i^{(m)})^\top \boldsymbol{\delta}| \le \|(\boldsymbol{\psi}_i^{(m)})_{\mathrm{supp}(\boldsymbol{\delta})}\|_2 \|\boldsymbol{\delta}\|_2 \le B_N \sqrt{q_N} \|\boldsymbol{\delta}\|_2.$$

Moreover, since $\boldsymbol{\vartheta} \in \mathcal{N}_N$,

$$\|\boldsymbol{\delta}\|_2^2 \le C \left(\|\mathbf{b} - \mathbf{b}^*\|_2^2 + \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|_2^2 + \sum_{m=2}^{M} \|\boldsymbol{\alpha}^{(m)} - \boldsymbol{\alpha}^{*(m)}\|_2^2 \right) \le C M r_s^2 \lambda^2.$$

Substituting these two bounds and $|\psi_{i\ell}^{(m)}| \le B_N$ into (S4.29) gives the coordinatewise bound

$$\|R(\boldsymbol{\vartheta})\|_\infty \le C \frac{B_N^3 q_N M r_s^2 \lambda^2}{h}, \qquad \boldsymbol{\vartheta} \in \mathcal{N}_N. \tag{S4.30}$$

It remains only to translate this coordinatewise bound into the mixed norm $\|\cdot\|_{\mathcal{D},+}$. The intercept part satisfies

$$M^{-1/2} \|R_{\mathbf{b}}(\boldsymbol{\vartheta})\|_2 \le \|R(\boldsymbol{\vartheta})\|_\infty.$$

For the shared component,

$$\|R_{\boldsymbol{\mu}}(\boldsymbol{\vartheta})\|_\infty \le \|R(\boldsymbol{\vartheta})\|_\infty.$$

For a deviation row j, using $\lambda_g \asymp M^{-1/2}$ and the equivalence of norms induced by $\mathbf{T}$

on $\mathbb{R}^{M-1}$,

$$\lambda_g^{-1}\|R_{\mathbf{a},j}(\boldsymbol{\vartheta})\|_{\widetilde{\mathbf{T}}} \le C\sqrt{M}\|R_{\mathbf{a},j}(\boldsymbol{\vartheta})\|_2 \le CM\|R(\boldsymbol{\vartheta})\|_\infty.$$

Combining the three displays with (S4.30), we obtain

$$\|R(\boldsymbol{\vartheta})\|_{\mathcal{D},+} \le C\frac{B_N^3M^2q_Nr_s^2\lambda^2}{h^2} = \lambda\cdot O\left(\frac{B_N^3M^2q_Nr_s^2\lambda}{h^2}\right) = o(\lambda),$$

where the last equality follows from (S4.28). This proves the required sparse active-block remainder expansion used in Theorem 2. The full-neighborhood uniform version follows from the same display after replacing q_N by $M(p+1)$.

Now suppose additionally that $\boldsymbol{\vartheta}_{\mathcal{I}^c} = \boldsymbol{\vartheta}^*_{\mathcal{I}^c} = 0$. Then $\boldsymbol{\delta}_{\mathcal{I}^c} = 0$, and hence

$$\boldsymbol{\Gamma}_h(\boldsymbol{\vartheta}^*)\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*)(\boldsymbol{\vartheta}_{\mathcal{I}} - \boldsymbol{\vartheta}^*_{\mathcal{I}}) \\ \boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*)(\boldsymbol{\vartheta}_{\mathcal{I}} - \boldsymbol{\vartheta}^*_{\mathcal{I}}) \end{pmatrix}$$

after arranging the coordinates according to $(\mathcal{I}, \mathcal{I}^c)$. Taking the corresponding sub-vectors in the full expansion gives

$$\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}) = \nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*)(\boldsymbol{\vartheta}_{\mathcal{I}} - \boldsymbol{\vartheta}^*_{\mathcal{I}}) + R_{\mathcal{I}}(\boldsymbol{\vartheta}),$$

and

$$\nabla_{\mathcal{I}^c}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}) = \nabla_{\mathcal{I}^c}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*)(\boldsymbol{\vartheta}_{\mathcal{I}} - \boldsymbol{\vartheta}^*_{\mathcal{I}}) + R_{\mathcal{I}^c}(\boldsymbol{\vartheta}).$$

Because the active and inactive mixed norms are restrictions of $\|\cdot\|_{\mathcal{D},+}$, the preceding bound immediately implies

$$\|R_{\mathcal{I}}(\boldsymbol{\vartheta})\|_{\mathcal{D},+} = o_p(\lambda), \qquad \|R_{\mathcal{I}^c}(\boldsymbol{\vartheta})\|_{\mathcal{D}} = o_p(\lambda).$$

This completes the proof. □

Lemma 9. Let

$$\widetilde{\boldsymbol{\vartheta}} = \underset{\substack{\mu_j=0,\ j\in S_\mu^c \\ \mathbf{a}_j=\mathbf{0},\ j\in S_\alpha^c}}{\operatorname{argmin}} \left\{ \widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}) + \lambda \mathcal{P}(\boldsymbol{\vartheta}) \right\}.$$

Under Conditions 1–5 and 6, and the tuning condition $\lambda \gtrsim d_N$,

$$P\left(\widetilde{\boldsymbol{\vartheta}} \in \mathcal{N}_N\right) \to 1.$$

Proof of Lemma 9. $\widetilde{\boldsymbol{\vartheta}}$ is actually the oracle estimator that minimizes the penalized loss over the true active set. Thus the result directly follows from Theorem 1. This completes the proof. □

S4.2 Proof of Theorem 1

Let

$$\widehat{\boldsymbol{\delta}}_{\mathbf{b}} := \widehat{\mathbf{b}}_h - \mathbf{b}^*, \qquad \widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}} := \widehat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}^*, \qquad \widehat{\boldsymbol{\Delta}}_{\mathbf{A}} := \widehat{\mathbf{A}}_h - \mathbf{A}^*.$$

By Lemma 3 and the choice of λ and λ_g, the conditions required in Lemma 4 hold with probability at least $1-(p\vee N)^{-C}$. Hence,

$$(\widehat{\boldsymbol{\delta}}_{\mathbf{b}}, \widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Delta}}_{\mathbf{A}}) \in \mathbb{C}.$$

In the following proof, we work on this high-probability event.

Suppose, on the contrary, that

$$\left(\frac{1}{M}\|\widehat{\boldsymbol{\delta}}_{\mathbf{b}}\|^2 + \|\widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\|^2 + \frac{1}{M}\|\widehat{\boldsymbol{\Delta}}_{\mathbf{A}}\|_F^2\right)^{1/2} > L\lambda r_s,$$

for some sufficiently large constant $L>0$. Let

$$\eta := \frac{L\lambda r_s}{\left(M^{-1}\|\widehat{\boldsymbol{\delta}}_{\mathbf{b}}\|^2 + \|\widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\|^2 + M^{-1}\|\widehat{\boldsymbol{\Delta}}_{\mathbf{A}}\|_F^2\right)^{1/2}} \in (0,1),$$

and define

$$\bar{\boldsymbol{\delta}}_{\mathbf{b}} := \eta\widehat{\boldsymbol{\delta}}_{\mathbf{b}}, \qquad \bar{\boldsymbol{\delta}}_{\boldsymbol{\mu}} := \eta\widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}}, \qquad \bar{\boldsymbol{\Delta}}_{\mathbf{A}} := \eta\widehat{\boldsymbol{\Delta}}_{\mathbf{A}}.$$

Let

$$\bar{\boldsymbol{\Theta}}_h := \boldsymbol{\Theta}^* + (\bar{\boldsymbol{\delta}}_{\mathbf{b}}, \bar{\boldsymbol{\delta}}_{\boldsymbol{\mu}}, \bar{\boldsymbol{\Delta}}_{\mathbf{A}}).$$

Then, we have

$$\big(M^{-1}\|\bar{\boldsymbol{\delta}}_{\mathbf{b}}\|^2 + \|\bar{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\|^2 + M^{-1}\|\bar{\boldsymbol{\Delta}}_{\mathbf{A}}\|_F^2\big)^{1/2} = t, \qquad t := L\lambda r_s. \tag{S4.31}$$

Since $\mathbb{C}$ is a cone,

$$(\bar{\boldsymbol{\delta}}_{\mathbf{b}}, \bar{\boldsymbol{\delta}}_{\boldsymbol{\mu}}, \bar{\boldsymbol{\Delta}}_{\mathbf{A}}) \in \mathbb{C}.$$

Denote

$$\mathcal{Q}_{N,h}(\boldsymbol{\Theta}) := \mathcal{L}_{N,h}(\boldsymbol{\Theta}) + \lambda\|\boldsymbol{\mu}\|_1 + \lambda\lambda_g\|\mathbf{A}\|_{2,1},$$

By convexity of the empirical objective and the fact that $\widehat{\boldsymbol{\Theta}}_h$ minimizes $\mathcal{Q}_{N,h}(\boldsymbol{\Theta})$, we have

$$\mathcal{Q}_{N,h}(\bar{\boldsymbol{\Theta}}_h) \le \eta\mathcal{Q}_{N,h}(\widehat{\boldsymbol{\Theta}}_h) + (1-\eta)\mathcal{Q}_{N,h}(\boldsymbol{\Theta}^*) \le \mathcal{Q}_{N,h}(\boldsymbol{\Theta}^*).$$

Hence,

$$\mathcal{L}_{N,h}(\bar{\boldsymbol{\Theta}}_h) - \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) \le \lambda\big(\|\boldsymbol{\mu}^*\|_1 - \|\bar{\boldsymbol{\mu}}_h\|_1\big) + \lambda\lambda_g\big(\|\mathbf{A}^*\|_{2,1} - \|\bar{\mathbf{A}}_h\|_{2,1}\big). \tag{S4.32}$$

By decomposability of the ℓ_1 and $\ell_{2,1}$ norms,

$$\|\boldsymbol{\mu}^*\|_1 - \|\boldsymbol{\mu}^* + \bar{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\|_1 \le \|\bar{\boldsymbol{\delta}}_{\boldsymbol{\mu},S_\mu}\|_1 - \|\bar{\boldsymbol{\delta}}_{\boldsymbol{\mu},S_\mu^c}\|_1 \le \|\bar{\boldsymbol{\delta}}_{\boldsymbol{\mu},S_\mu}\|_1,$$

and

$$\begin{aligned}\|\mathbf{A}^*\|_{2,1} - \|\mathbf{A}^* + \bar{\boldsymbol{\Delta}}_{\mathbf{A}}\|_{2,1} &= \sum_{j\in S_\alpha}\big(\|\mathbf{A}^*_{j\cdot}\|_2 - \|\mathbf{A}^*_{j\cdot} + \bar{\boldsymbol{\Delta}}_{\mathbf{A},j\cdot}\|_2\big) - \sum_{j\in S_\alpha^c}\|\bar{\boldsymbol{\Delta}}_{\mathbf{A},j\cdot}\|_2 \\ &\le \sum_{j\in S_\alpha}\|\bar{\boldsymbol{\Delta}}_{\mathbf{A},j\cdot}\|_2 - \sum_{j\in S_\alpha^c}\|\bar{\boldsymbol{\Delta}}_{\mathbf{A},j\cdot}\|_2 \\ &= \|\bar{\boldsymbol{\Delta}}_{\mathbf{A},S_\alpha}\|_{2,1} - \|\bar{\boldsymbol{\Delta}}_{\mathbf{A},S_\alpha^c}\|_{2,1} \le \|\bar{\boldsymbol{\Delta}}_{\mathbf{A},S_\alpha}\|_{2,1}.\end{aligned}$$

Substituting these bounds into (S4.32) yields

$$\mathcal{L}_{N,h}(\bar{\boldsymbol{\Theta}}_h) - \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) \le \lambda\|\bar{\boldsymbol{\delta}}_{\boldsymbol{\mu},S_\mu}\|_1 + \lambda\lambda_g\|\bar{\boldsymbol{\Delta}}_{\mathbf{A},S_\alpha}\|_{2,1}. \tag{S4.33}$$

By Lemma 5 and (S4.33), with probability at least $1-(p\vee N)^{-C}$,

$$\begin{aligned}\mathcal{L}_h(\bar{\boldsymbol{\Theta}}_h) - \mathcal{L}_h(\boldsymbol{\Theta}^*) &\le \mathcal{L}_{N,h}(\bar{\boldsymbol{\Theta}}_h) - \mathcal{L}_{N,h}(\boldsymbol{\Theta}^*) + Ctr_s\sqrt{\frac{M+\log(p\vee N)}{N}} \\ &\le \lambda\|\bar{\boldsymbol{\delta}}_{\boldsymbol{\mu},S_\mu}\|_1 + \lambda\lambda_g\|\bar{\boldsymbol{\Delta}}_{\mathbf{A},S_\alpha}\|_{2,1} + Ctr_s\sqrt{\frac{M+\log(p\vee N)}{N}} \\ &\le \lambda\sqrt{s_\mu}\|\bar{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\| + \lambda\lambda_g\sqrt{s_\alpha}\|\bar{\boldsymbol{\Delta}}_{\mathbf{A}}\|_F + Ctr_s\sqrt{\frac{M+\log(p\vee N)}{N}} \\ &\le C\lambda\big(\sqrt{s_\mu}+\sqrt{s_\alpha}\big)t + Ctr_s\sqrt{\frac{M+\log(p\vee N)}{N}} \\ &\le C\lambda r_s t, \end{aligned}\tag{S4.34}$$

where the penultimate inequality uses $\lambda\lambda_g\sqrt{s_\alpha}\|\bar{\boldsymbol{\Delta}}_{\mathbf{A}}\|_F \le \lambda\lambda_g\sqrt{s_\alpha}\sqrt{M}t \le C\lambda\sqrt{s_\alpha}t$

since $\lambda_g \asymp M^{-1/2}$, and the last step uses the choice

$$\lambda \geq C\left(\sqrt{\frac{M+\log(p\vee N)}{N}}+\left(\frac{h}{\beta^*_{1,\min}}\right)^2\right).$$

Next we derive the lower bound. For each task m, define

$$\bar{\boldsymbol{\theta}}_h^{(m)} := \left(\bar{b}_h^{(m)}, (\bar{\boldsymbol{\mu}}_h + \bar{\boldsymbol{\alpha}}_h^{(m)})^\top\right)^\top, \qquad \boldsymbol{\theta}^{*(m)} := \left(b^{*(m)}, (\boldsymbol{\mu}^* + \boldsymbol{\alpha}^{*(m)})^\top\right)^\top,$$

and

$$\bar{\boldsymbol{\delta}}_\theta^{(m)} := \bar{\boldsymbol{\theta}}_h^{(m)} - \boldsymbol{\theta}^{*(m)} = \left(\bar{\delta}_b^{(m)}, (\bar{\boldsymbol{\delta}}_{\boldsymbol{\mu}} + \bar{\boldsymbol{\delta}}_{\boldsymbol{\alpha}}^{(m)})^\top\right)^\top.$$

Applying Taylor's expansion task-wise, we have

$$\begin{aligned}
& \mathcal{L}_h(\bar{\boldsymbol{\Theta}}_h) - \mathcal{L}_h(\boldsymbol{\Theta}^*) \\
= \;& \frac{1}{N}\sum_{m=1}^M n_m \left(\mathbb{E}^{(m)}\left[\phi_h\left(y\widetilde{\mathbf{x}}^\top \bar{\boldsymbol{\theta}}_h^{(m)}\right)\right] - \mathbb{E}^{(m)}\left[\phi_h\left(y\widetilde{\mathbf{x}}^\top \boldsymbol{\theta}^{*(m)}\right)\right]\right) \\
= \;& \frac{1}{N}\sum_{m=1}^M n_m \mathbf{s}_h^{(m)}(\boldsymbol{\theta}^{*(m)})^\top \bar{\boldsymbol{\delta}}_\theta^{(m)} + \frac{1}{2N}\sum_{m=1}^M n_m (\bar{\boldsymbol{\delta}}_\theta^{(m)})^\top \mathbf{H}_h^{(m)}(\boldsymbol{\theta}'^{(m)})\bar{\boldsymbol{\delta}}_\theta^{(m)}, \quad \text{(S4.35)}
\end{aligned}$$

where $\boldsymbol{\theta}'^{(m)}$ lies on the line segment joining $\boldsymbol{\theta}^{*(m)}$ and $\bar{\boldsymbol{\theta}}_h^{(m)}$.

We first control the score term. By Lemma 1,

$$\|\mathbf{s}_h^{(m)}(\boldsymbol{\theta}^{*(m)})\| \leq C\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2, \qquad m \in [M].$$

Hence, by Cauchy's inequality and Lemma 6,

$$\begin{aligned}\frac{1}{N}\sum_{m=1}^{M} n_m \mathbf{s}_h^{(m)}(\boldsymbol{\theta}^{*(m)})^\top \bar{\boldsymbol{\delta}}_\theta^{(m)} &\geq -C\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2 \cdot \frac{1}{N}\sum_{m=1}^{M} n_m \|\bar{\boldsymbol{\delta}}_\theta^{(m)}\| \\ &\geq -C\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2 \left(\frac{1}{N}\sum_{m=1}^{M} n_m \|\bar{\boldsymbol{\delta}}_\theta^{(m)}\|^2\right)^{1/2} \\ &\geq -C\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2 t. \end{aligned} \tag{S4.36}$$

Since $(\bar{\boldsymbol{\delta}}_{\mathbf{b}}, \bar{\boldsymbol{\delta}}_{\boldsymbol{\mu}}, \bar{\boldsymbol{\Delta}}_{\mathbf{A}}) \in \mathbb{C}$ and

$$\left(M^{-1}\|\bar{\boldsymbol{\delta}}_{\mathbf{b}}\|^2 + \|\bar{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\|^2 + M^{-1}\|\bar{\boldsymbol{\Delta}}_{\mathbf{A}}\|_F^2\right)^{1/2} = t = L\lambda r_s,$$

the line segment between $\boldsymbol{\Theta}^*$ and $\bar{\boldsymbol{\Theta}}_h$ lies in the local neighborhood in Condition 3, provided that the constant C_1 therein is sufficiently large. **Keep each j_m fixed at its population choice. For every parameter on this segment, the normalized neighborhood radius gives**

$$\begin{aligned}\|\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{*(m)}\|_2 &\leq \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|_2 + \|\mathbf{A} - \mathbf{A}^*\|_F \\ &\leq (1+\sqrt{M})t \leq 2L\sqrt{M}\lambda r_s.\end{aligned}$$

Consequently,

$$\begin{aligned}|\beta_{j_m}^{(m)}| &\geq |\beta_{j_m}^{*(m)}| - \|\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{*(m)}\|_2 \\ &\geq \beta^*_{1,\min} - C\sqrt{M}\lambda r_s \geq \tfrac{1}{2}\beta^*_{1,\min},\end{aligned}$$

uniformly across tasks for sufficiently large N, since Condition 2 assumes

$\sqrt{M}\lambda r_s = o(\beta^*_{1,\min})$**. Thus the Hessian approximation error is uniformly bounded by a constant times** $h/(\beta^*_{1,\min})^2$**.** By Lemma 2 and Condition 3, for $h/|\beta^*_{1,\min}|^2$ sufficiently small,

$$\left(\bar{\boldsymbol{\delta}}_\theta^{(m)}\right)^\top \mathbf{H}_h^{(m)}(\boldsymbol{\theta}'^{(m)})\bar{\boldsymbol{\delta}}_\theta^{(m)} \geq C\|\bar{\boldsymbol{\delta}}_\theta^{(m)}\|_2^2, \qquad m \in [M].$$

Therefore, by Lemma 6,

$$\begin{aligned}
\frac{1}{2}\sum_{m=1}^{M}\frac{n_m}{N}\left(\bar{\boldsymbol{\delta}}_\theta^{(m)}\right)^\top \mathbf{H}_h^{(m)}(\boldsymbol{\theta}'^{(m)})\bar{\boldsymbol{\delta}}_\theta^{(m)} &\geq C\sum_{m=1}^{M}\frac{n_m}{N}\|\bar{\boldsymbol{\delta}}_\theta^{(m)}\|_2^2 \\
&\geq C\left(M^{-1}\|\bar{\boldsymbol{\delta}}_{\mathbf{b}}\|_2^2 + \|\bar{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\|_2^2 + M^{-1}\|\bar{\boldsymbol{\Delta}}_{\mathbf{A}}\|_F^2\right) \\
&= Ct^2. \qquad \text{(S4.37)}
\end{aligned}$$

Combining (S4.35), (S4.36), and (S4.37), we get

$$\mathcal{L}_h(\bar{\boldsymbol{\Theta}}_h) - \mathcal{L}_h(\boldsymbol{\Theta}^*) \geq -C\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2 t + Ct^2. \qquad \text{(S4.38)}$$

Finally, combining (S4.34) and (S4.38) yields

$$Ct^2 \leq C\lambda r_s t + C\left(\frac{h}{|\beta^*_{1,\min}|}\right)^2 t \leq C\lambda r_s t,$$

by the choice of λ. Hence

$$t \leq C\lambda r_s.$$

This contradicts (S4.31) for sufficiently large L. Therefore,

$$\left(\frac{1}{M}\|\widehat{\boldsymbol{\delta}}_{\mathbf{b}}\|^2+\|\widehat{\boldsymbol{\delta}}_{\boldsymbol{\mu}}\|^2+\frac{1}{M}\|\widehat{\boldsymbol{\Delta}}_{\mathbf{A}}\|_F^2\right)^{1/2}\leq C\lambda r_s.$$

This completes the proof.

□

S4.3 Proof of Theorem 2

We prove the theorem by a primal-dual witness argument. Recall the restricted estimator is defined as

$$\widetilde{\boldsymbol{\vartheta}}=\underset{\substack{\mu_j=0,\ j\in S_\mu^c\\ \mathbf{a}_j=0,\ j\in S_\alpha^c}}{\operatorname{argmin}}\left\{\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta})+\lambda\mathcal{P}(\boldsymbol{\vartheta})\right\}.$$

By Lemma 9,

$$P\left(\widetilde{\boldsymbol{\vartheta}}\in\mathcal{N}_N\right)\to 1.$$

Thus, with probability tending to one,

$$\begin{aligned}
&\|\widetilde{\boldsymbol{\mu}}_{S_\mu}-\boldsymbol{\mu}^*_{S_\mu}\|_\infty+M^{-1/2}\max_{j\in S_\alpha}\|\mathbf{T}(\widetilde{\mathbf{a}}_j-\mathbf{a}_j^*)\|_2\\
=\ &\|\widetilde{\boldsymbol{\mu}}_{S_\mu}-\boldsymbol{\mu}^*_{S_\mu}\|_\infty+M^{-1/2}\max_{j\in S_\alpha}\|\widetilde{\mathbf{A}}_{j\cdot}-\mathbf{A}^*_{j\cdot}\|_2\\
\leq\ &Cr_s\lambda, \qquad \text{(S4.39)}
\end{aligned}$$

since $\mathbf{T}(\widetilde{\mathbf{a}}_j - \mathbf{a}_j^*) = \widetilde{\mathbf{A}}_{j\cdot}^\top - \mathbf{A}_{j\cdot}^{*\top}$. For any $j \in S_\mu$, it is clear that $|\widetilde{\mu}_j| \geq |\mu_j^*| - |\widetilde{\mu}_j - \mu_j^*|$. Therefore, combining (S4.39), with probability tending to one,

$$\begin{aligned} \min_{j\in S_\mu} |\widetilde{\mu}_j| &\geq \min_{j\in S_\mu} |\mu_j^*| - \|\widetilde{\boldsymbol{\mu}}_{S_\mu} - \boldsymbol{\mu}_{S_\mu}^*\|_\infty \\ &\geq \nu_N - Cr_s\lambda \\ &> 0, \end{aligned}$$

where the last inequality holds since Condition 8 assumes $r_s\lambda = o(\nu_N)$. Thus,

$$\widetilde{\mu}_j \neq 0, \qquad j \in S_\mu.$$

Similarly,

$$M^{-1/2} \min_{j\in S_\alpha} \|\widetilde{\mathbf{A}}_{j\cdot}\|_2 \geq M^{-1/2} \min_{j\in S_\alpha} \|\mathbf{A}_{j\cdot}^*\|_2 - M^{-1/2} \max_{j\in S_\alpha} \|\widetilde{\mathbf{A}}_{j\cdot} - \mathbf{A}_{j\cdot}^*\|_2.$$

Therefore,

$$M^{-1/2} \min_{j\in S_\alpha} \|\widetilde{\mathbf{A}}_{j\cdot}\|_2 \geq \nu_N - Cr_s\lambda > 0$$

with probability tending to one. Hence,

$$\|\widetilde{\mathbf{A}}_{j\cdot}\|_2 \neq 0, \qquad j \in S_\alpha.$$

This proves that no truly active variable is missed.

Let $\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}}$ be an active-block subgradient of the penalty at $\widetilde{\boldsymbol{\vartheta}}$. On the active block $\mathcal{I}$, the restricted estimator satisfies

$$\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\widetilde{\boldsymbol{\vartheta}}) + \lambda\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}} = 0,$$

with $\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}}$ well-defined from the KKT condition of the restricted optimization problem. By the definition of the subgradient of $\mathcal{P}$, its intercept component is zero and

$$\|\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}}\|_{\mathcal{D},+} \leq 1. \tag{S4.40}$$

Indeed, $|\widetilde{\boldsymbol{\zeta}}_{\mu,j}| \leq 1$ for $j \in S_\mu$, and $\|\widetilde{\boldsymbol{\zeta}}_{a,j}\|_{\widetilde{\mathbf{T}}} \leq \lambda_g$ for $j \in S_\alpha$.

By Lemma 8, on the event $\widetilde{\boldsymbol{\vartheta}} \in \mathcal{N}_N$, we have

$$\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\widetilde{\boldsymbol{\vartheta}}) = \nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*)\left(\widetilde{\boldsymbol{\vartheta}}_{\mathcal{I}} - \boldsymbol{\vartheta}^*_{\mathcal{I}}\right) + \widetilde{R}_{\mathcal{I}},$$

where

$$\|\widetilde{R}_{\mathcal{I}}\|_{\mathcal{D},+} = o_p(\lambda). \tag{S4.41}$$

Combining this expansion with the active KKT equation gives

$$\widetilde{\boldsymbol{\vartheta}}_{\mathcal{I}} - \boldsymbol{\vartheta}^*_{\mathcal{I}} = -\boldsymbol{\Gamma}^{-1}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*)\left\{\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \widetilde{R}_{\mathcal{I}} + \lambda\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}}\right\}. \tag{S4.42}$$

By Lemma 7,

$$\left\|\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*)\right\|_{\mathcal{D},+} = O_p(d_N).$$

Since $d_N = o(\eta_0\lambda)$, we have $d_N = o(\lambda)$. Together with (S4.40) and (S4.41), by the triangle inequality of the mixed dual norm, we obtain

$$\left\|\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \widetilde{R}_{\mathcal{I}} + \lambda\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}}\right\|_{\mathcal{D},+} = O_p(\lambda).$$

Next, we construct the inactive subgradient. Define

$$\lambda\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}^c} = -\nabla_{\mathcal{I}^c}\widetilde{\mathcal{L}}_{N,h}(\widetilde{\boldsymbol{\vartheta}}).$$

We verify strict dual feasibility, namely, $\|\widetilde{\boldsymbol{\zeta}}_{\mu,S_\mu^c}\|_\infty < 1$ and $\max_{j\in S_\alpha^c}\|\widetilde{\boldsymbol{\zeta}}_{a,j}\|_{\widetilde{\mathbf{T}}} < \lambda_g$. By Lemma 8, on the event $\widetilde{\boldsymbol{\vartheta}} \in \mathcal{N}_N$,

$$\nabla_{\mathcal{I}^c}\widetilde{\mathcal{L}}_{N,h}(\widetilde{\boldsymbol{\vartheta}}) = \nabla_{\mathcal{I}^c}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*)\left(\widetilde{\boldsymbol{\vartheta}}_{\mathcal{I}} - \boldsymbol{\vartheta}^*_{\mathcal{I}}\right) + \widetilde{R}_{\mathcal{I}^c},$$

where $\|\widetilde{R}_{\mathcal{I}^c}\|_{\mathcal{D}} = o_p(\lambda)$. Substituting the expression for $\widetilde{\boldsymbol{\vartheta}}_{\mathcal{I}} - \boldsymbol{\vartheta}^*_{\mathcal{I}}$ in (S4.42), we obtain

$$\lambda\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}^c} = \lambda\boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*)\boldsymbol{\Gamma}^{-1}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*)\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}} + \widetilde{\Delta}_N,$$

where

$$\widetilde{\Delta}_N := \mathbf{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*)\mathbf{\Gamma}^{-1}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*)\left\{\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \widetilde{R}_{\mathcal{I}}\right\} - \nabla_{\mathcal{I}^c}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) - \widetilde{R}_{\mathcal{I}^c}.$$

Dividing both sides by λ, we get

$$\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}^c} = \mathbf{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*)\mathbf{\Gamma}^{-1}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*)\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}} + \lambda^{-1}\widetilde{\Delta}_N. \tag{S4.43}$$

Taking the $\mathcal{D}$-norm and using the triangle inequality yields

$$\begin{aligned}\|\widetilde{\Delta}_N\|_{\mathcal{D}} \;\leq\; & \left\|\mathbf{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*)\mathbf{\Gamma}^{-1}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*)\left\{\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \widetilde{R}_{\mathcal{I}}\right\}\right\|_{\mathcal{D}} \\ & + \left\|\nabla_{\mathcal{I}^c}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*)\right\|_{\mathcal{D}} + \|\widetilde{R}_{\mathcal{I}^c}\|_{\mathcal{D}}.\end{aligned}$$

Let

$$\mathbf{u}_{\mathcal{I}} = \nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \widetilde{R}_{\mathcal{I}},$$

and decompose

$$\mathbf{u}_{\mathcal{I}} = \mathbf{u}_{\mathcal{I},p} + \mathbf{u}_{\mathcal{I},b},$$

where $\mathbf{u}_{\mathcal{I},p}$ is supported on the penalized active components $(\boldsymbol{\mu}_{S_\mu}, \mathbf{a}_{S_\alpha})$, and $\mathbf{u}_{\mathcal{I},b}$ is supported on the unpenalized intercept component.

For the penalized part, define

$$\boldsymbol{\zeta}_{\mathcal{I}}^{(u)} = \frac{\mathbf{u}_{\mathcal{I},p}}{\|\mathbf{u}_{\mathcal{I},p}\|_{\mathcal{D},+}}.$$

Then $\boldsymbol{\zeta}_b^{(u)} = 0$, $|\boldsymbol{\zeta}_{\mu,j}^{(u)}| \leq 1$ for $j \in S_\mu$, and $\|\boldsymbol{\zeta}_{a,j}^{(u)}\|_{\widetilde{\mathbf{T}}} \leq \lambda_g$ for $j \in S_\alpha$. Then, by Condition 7,

$$\left\| \boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*) \boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}^{-1}(\boldsymbol{\vartheta}^*) \boldsymbol{\zeta}_{\mathcal{I}}^{(u)} \right\|_{\mathcal{D}} \leq 1,$$

which implies

$$\left\| \boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*) \boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}^{-1}(\boldsymbol{\vartheta}^*) \mathbf{u}_{\mathcal{I},p} \right\|_{\mathcal{D}} \leq \|\mathbf{u}_{\mathcal{I},p}\|_{\mathcal{D},+},$$

For the unpenalized intercept part, by Condition 7, there exists a constant $C_b > 0$ such that

$$\left\| \boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*) \boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}^{-1}(\boldsymbol{\vartheta}^*) \mathbf{u}_{\mathcal{I},b} \right\|_{\mathcal{D}} \leq C_b M^{-1/2} \|\mathbf{u}_b\|_2 \leq C_b \|\mathbf{u}_{\mathcal{I}}\|_{\mathcal{D},+},$$

since the intercept component is included in $\|\cdot\|_{\mathcal{D},+}$. Hence,

$$\begin{aligned}
\left\| \boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*) \boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}^{-1}(\boldsymbol{\vartheta}^*) \mathbf{u}_{\mathcal{I}} \right\|_{\mathcal{D}} &\leq \left\| \boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*) \boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}^{-1}(\boldsymbol{\vartheta}^*) \mathbf{u}_{\mathcal{I},p} \right\|_{\mathcal{D}} + \left\| \boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*) \boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}^{-1}(\boldsymbol{\vartheta}^*) \mathbf{u}_{\mathcal{I},b} \right\|_{\mathcal{D}} \\
&\leq \|\mathbf{u}_{\mathcal{I},p}\|_{\mathcal{D},+} + C_b \|\mathbf{u}_{\mathcal{I}}\|_{\mathcal{D},+} \\
&\leq C \|\mathbf{u}_{\mathcal{I}}\|_{\mathcal{D},+}.
\end{aligned}$$

Recall that $\mathbf{u}_{\mathcal{I}} = \nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \widetilde{R}_{\mathcal{I}}$, since $\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*)$ is obtained by restricting the full score to the active block $\mathcal{I}$, the definition of the block norms gives

$$\left\|\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*)\right\|_{\mathcal{D},+} \leq \left\|\nabla_{\boldsymbol{\vartheta}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*)\right\|_{\mathcal{D},+}.$$

Hence, by Lemma 7,

$$\left\|\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*)\right\|_{\mathcal{D},+} = O_p(d_N).$$

Moreover, by Lemma 8,

$$\|\widetilde{R}_{\mathcal{I}}\|_{\mathcal{D},+} = o_p(\lambda).$$

Therefore,

$$\left\|\boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*)\boldsymbol{\Gamma}^{-1}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*)\left\{\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \widetilde{R}_{\mathcal{I}}\right\}\right\|_{\mathcal{D}} = O_p(d_N) + o_p(\lambda).$$

Next, since $\nabla_{\mathcal{I}^c}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*)$ is the inactive subvector of the full score, Lemma 7 also implies

$$\left\|\nabla_{\mathcal{I}^c}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*)\right\|_{\mathcal{D}} \leq \left\|\nabla_{\boldsymbol{\vartheta}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*)\right\|_{\mathcal{D},+} = O_p(d_N).$$

Finally, by Lemma 8,

$$\|\widetilde{R}_{\mathcal{I}^c}\|_{\mathcal{D}} = o_p(\lambda).$$

Combining the preceding bounds, we obtain

$$\begin{aligned}
\|\widetilde{\Delta}_N\|_{\mathcal{D}} &\leq \left\|\boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*)\boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}^{-1}(\boldsymbol{\vartheta}^*)\left\{\nabla_{\mathcal{I}}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*) + \widetilde{R}_{\mathcal{I}}\right\}\right\|_{\mathcal{D}} \\
&\quad + \left\|\nabla_{\mathcal{I}^c}\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}^*)\right\|_{\mathcal{D}} + \|\widetilde{R}_{\mathcal{I}^c}\|_{\mathcal{D}} \\
&= O_p(d_N) + o_p(\lambda).
\end{aligned}$$

Since $d_N = o(\eta_0\lambda)$, it follows that

$$\lambda^{-1}\|\widetilde{\Delta}_N\|_{\mathcal{D}} = O_p\left(\frac{d_N}{\lambda}\right) + o_p(1) = o_p(\eta_0).$$

Since the mixed norm $\|\cdot\|_{\mathcal{D}}$ dominates the common inactive component, we have

$$\lambda^{-1}\|\widetilde{\Delta}_{N,\mu,S_\mu^c}\|_\infty \leq \lambda^{-1}\|\widetilde{\Delta}_N\|_{\mathcal{D}} = o_p(\eta_0).$$

By the empirical irrepresentable condition in Condition 7,

$$\left\|\left[\boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*)\boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}^{-1}(\boldsymbol{\vartheta}^*)\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}}\right]_{\mu,S_\mu^c}\right\|_\infty \leq 1 - \eta_0.$$

Hence, combining the above bounds and (S4.43), we have

$$\|\widetilde{\boldsymbol{\zeta}}_{\mu,S_\mu^c}\|_\infty \le 1-\eta_0+\lambda^{-1}\|\widetilde{\Delta}_{N,\mu,S_\mu^c}\|_\infty < 1$$

with probability tending to one.

Similarly, by the definition of the mixed norm $\|\cdot\|_{\mathcal{D}}$,

$$\lambda_g^{-1}\max_{j\in S_\alpha^c}\|\widetilde{\Delta}_{N,a,j}\|_{\widetilde{\mathbf{T}}} \le \|\widetilde{\Delta}_N\|_{\mathcal{D}}.$$

Therefore,

$$\lambda^{-1}\max_{j\in S_\alpha^c}\|\widetilde{\Delta}_{N,a,j}\|_{\widetilde{\mathbf{T}}} \le \lambda_g\lambda^{-1}\|\widetilde{\Delta}_N\|_{\mathcal{D}} = o_p(\lambda_g\eta_0).$$

By Condition 7,

$$\max_{j\in S_\alpha^c}\left\|\left[\boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*)\boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}^{-1}(\boldsymbol{\vartheta}^*)\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}}\right]_{a,j}\right\|_{\widetilde{\mathbf{T}}} \le \lambda_g(1-\eta_0).$$

Therefore, combining the above bounds and (S4.43), we have

$$\max_{j\in S_\alpha^c}\|\widetilde{\boldsymbol{\zeta}}_{a,j}\|_{\widetilde{\mathbf{T}}} \le \lambda_g(1-\eta_0)+\lambda^{-1}\max_{j\in S_\alpha^c}\|\widetilde{\Delta}_{N,a,j}\|_{\widetilde{\mathbf{T}}}.$$

Since the second term is $o_p(\lambda_g\eta_0)$,

$$\max_{j\in S_\alpha^c}\|\widetilde{\boldsymbol{\zeta}}_{a,j}\|_{\widetilde{\mathbf{T}}} < \lambda_g$$

with probability tending to one.

Thus, strict dual feasibility holds, and the restricted estimator $\widetilde{\boldsymbol{\vartheta}}$ satisfies the KKT conditions of the full penalized problem. By convexity, it is a global minimizer. Taking $\widehat{\boldsymbol{\vartheta}}_h = \widetilde{\boldsymbol{\vartheta}}$, we have

$$\widehat{\boldsymbol{\mu}}_{h,S_\mu^c} = 0, \qquad \widehat{\mathbf{A}}_{h,S_\alpha^c\cdot} = 0.$$

Combining this with the active lower bounds proved above gives

$$\widehat{S}_\mu = S_\mu, \qquad \widehat{S}_\alpha = S_\alpha$$

with probability tending to one. This completes the proof of Theorem 2. □

S4.4 Discussion and Gaussian verification of Condition 4

Condition 4 supplies sufficient conditional-density bounds for the score-bias and Hessian-approximation arguments. Its weighted integrable envelope allows dependence among covariates. The following calculation applies to each task and class, with constants uniform across tasks and classes. We suppress these indices and consider the distribution given $y = 1$; the other class is treated identically. The coordinate selected in Condition 2 is locally denoted by x_1.

Assume $\mathbf{x} \mid y = 1 \sim \mathcal{N}(\boldsymbol{\xi}, \Sigma)$, with

$$c_\Sigma I \preceq \Sigma \preceq C_\Sigma I, \qquad \|\boldsymbol{\xi}\|_2 \le C_\xi,$$

where c_Σ, C_Σ, and C_ξ are positive constants. Partition $\boldsymbol{\xi}$ and Σ according to x_1 and $\mathbf{x}_{-1}$. The conditional Gaussian distribution is

$$x_1 \mid \mathbf{x}_{-1}, y = 1 \sim \mathcal{N}(\eta(\mathbf{x}_{-1}), s),$$

where

$$\eta(\mathbf{x}_{-1}) = \xi_1 + \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}(\mathbf{x}_{-1} - \boldsymbol{\xi}_{-1}),$$

$$s = \sigma_{11} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}.$$

Since $s = \{(\Sigma^{-1})_{11}\}^{-1}$, we have $c_\Sigma \le s \le C_\Sigma$. Moreover, under the distribution given $y = 1$,

$$\mathbb{E}\eta^2 = \xi_1^2 + \sigma_{11} - s \le C_\xi^2 + C_\Sigma.$$

Let $f(x \mid \mathbf{x}_{-1})$ be the conditional density. For $x \neq 0$,

$$\log\{|x| f(x \mid \mathbf{x}_{-1})\} = \log|x| - \frac{(x-\eta)^2}{2s} - \frac{1}{2}\log(2\pi s).$$

Its stationary points satisfy $x(x-\eta) = s$, hence $|x| \le |\eta| + \sqrt{s}$. Consequently,

$$\sup_x |x| f(x \mid \mathbf{x}_{-1}) \le C(1 + |\eta|).$$

To cover every term of Condition 4, write $t = \boldsymbol{\alpha}^\top \mathbf{x}_{-1}$ for a fixed unit vector $\boldsymbol{\alpha}$. Using $z = (x - \eta)/\sqrt{s}$ and

$$f'(x \mid \mathbf{x}_{-1}) = -\frac{x - \eta}{s} f(x \mid \mathbf{x}_{-1}),$$

the bounded polynomial-weighted standard Gaussian density gives

$$\sup_x (1 + |x| + |t|) f(x \mid \mathbf{x}_{-1}) \le C(1 + |\eta| + |t|),$$

$$\sup_x (1 + |x| + |t| + x^2 + t^2) |f'(x \mid \mathbf{x}_{-1})| \le C(1 + \eta^2 + t^2).$$

Thus an admissible envelope is

$$\mathcal{E}_{\boldsymbol{\alpha}}(\mathbf{x}_{-1}) = C(1 + \eta(\mathbf{x}_{-1})^2 + t^2).$$

Since

$$\mathbb{E}t^2 = (\boldsymbol{\alpha}^\top \boldsymbol{\xi}_{-1})^2 + \boldsymbol{\alpha}^\top \Sigma_{-1,-1} \boldsymbol{\alpha} \le C_\xi^2 + C_\Sigma,$$

we obtain

$$\sup_{\|\boldsymbol{\alpha}\|_2 = 1} \int \mathcal{E}_{\boldsymbol{\alpha}}(\mathbf{x}_{-1}) f_{-1}(\mathbf{x}_{-1}) \, d\mathbf{x}_{-1} \le C(1 + 2C_\xi^2 + 2C_\Sigma) < \infty.$$

Thus its integral is bounded uniformly over unit directions, tasks, and classes.

Finally, we note that requiring $\sup_{x, \mathbf{x}_{-1}} |x| f(x \mid \mathbf{x}_{-1}) < \infty$ would be much

more restrictive. Evaluating at $x = \eta(\mathbf{x}_{-1})$ gives

$$\sup_x |x| f(x \mid \mathbf{x}_{-1}) \geq \frac{|\eta(\mathbf{x}_{-1})|}{\sqrt{2\pi s}}.$$

If $\Sigma_{1,-1} \neq 0$, the conditional mean is an unbounded affine function of $\mathbf{x}_{-1}$, so this joint supremum is infinite. For Gaussian covariates, a finite joint supremum therefore requires the selected coordinate to be independent of the remaining coordinates. Condition 4 avoids this restriction by requiring an integrable envelope instead.

Bibliography

Koltchinskii, V. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *Lecture Notes in Mathematics*. Berlin: Springer (2011).

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

van der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. New York: Springer (1996).

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