Supplementary Material

In this supplementary, we first provide an overview of our proof techniques in Appendix A and then in Appendix B provide the proofs of theorems and technical lemmas stated in the main paper.

A Overview of proof techniques

Our analysis of the generalization error is based on an extension of Gordon's Gaussian process inequality [13], called Convex-Gaussian Minimax Theorem (CGMT) [30]. Here, we outline the general steps of this framework and refer to the supplementary for complete details and derivations.

Consider the following two Gaussian processes:

$$\begin{aligned} \boldsymbol{X}_{\boldsymbol{u},\boldsymbol{v}} &:= \boldsymbol{u}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{v} + \psi(\boldsymbol{u},\boldsymbol{v})\,, \\ \boldsymbol{Y}_{\boldsymbol{u},\boldsymbol{v}} &:= \|\boldsymbol{u}\|_{\ell_2}\boldsymbol{g}^{\mathsf{T}}\boldsymbol{v} + \|\boldsymbol{v}\|_{\ell_2}\boldsymbol{h}^{\mathsf{T}}\boldsymbol{u} + \psi(\boldsymbol{u},\boldsymbol{v})\,, \end{aligned}$$

where $\boldsymbol{G} \in \mathbb{R}^{n\times d}$, $\boldsymbol{g} \in \mathbb{R}^n$ and $\boldsymbol{h} \in \mathbb{R}^d$, all have i.i.d standard normal entries. Further, $\psi : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ is a continuous function, which is convex in the first argument and concave in the second argument.

Given the above two processes, consider the following min-max optimization problems, which are respectively referred to as the *Primary Optimization (PO)* and the *Auxiliary Optimization (AO)* problems:

$$\Phi_{\rm PO}(\boldsymbol{G}) := \min_{\boldsymbol{u}\in\boldsymbol{S}_{\boldsymbol{u}}}\max_{\boldsymbol{v}\in\boldsymbol{S}_{\boldsymbol{v}}} \boldsymbol{X}_{\boldsymbol{u},\boldsymbol{v}}\,, \tag{A.1}$$

$$\Phi_{\rm AO}(\boldsymbol{g},\boldsymbol{h}) := \min_{\boldsymbol{u}\in\boldsymbol{S}_{\boldsymbol{u}}}\max_{\boldsymbol{v}\in\boldsymbol{S}_{\boldsymbol{v}}} \boldsymbol{Y}_{\boldsymbol{u},\boldsymbol{v}}\,. \tag{A.2}$$

The main result of CGMT is to connect the above two random optimization problems. As shown in [30](Theorem 3), if $\boldsymbol{S}_{\boldsymbol{u}}$ and $\boldsymbol{S}_{\boldsymbol{v}}$ are compact and convex then, for any $\lambda \in \mathbb{R}$ and $t > 0$,

$$\mathbb{P}\left(|\Phi_{\rm PO}(\boldsymbol{G}) - \lambda| > t\right) \le 2\mathbb{P}\left(|\Phi_{\rm AO}(\boldsymbol{g},\boldsymbol{h}) - \lambda| > t\right)\,.$$

An immediate corollary of this result (by choosing $\lambda = \mathbb{E}[\Phi_{\rm AO}(\boldsymbol{g},\boldsymbol{h})]$) is that if the optimal cost of AO problem concentrates in probability, then the optimal cost of the corresponding PO problem also concentrates, in probability, around the same value. In addition, as shown in part (iii) of [30](Theorem 3), concentration of the optimal solution of the AO problem implies concentration of the optimal solution of the PO around the same value. Therefore, the two optimization are intimately connected and by analyzing the AO problem, which is substantially simpler, one can derive corresponding properties of the PO problem.

The CGMT framework has been used to infer statistical properties of estimators in certain high-dimensional asymptotic regime. The intermediate steps in the CGMT framework can be summarized as follows: First form an PO problem in the form of (A.1) and construct the corresponding AO problem. Second, derive the point-wise limit of the AO objective in terms of a convex-concave optimization problem, over only few scalar variables. This step is called 'scalarization'. Next, it is possible to establish uniform convergence of the scalarized AO to the (deterministic) min-max optimization problem using convexity conditions. Finally, by analyzing the latter deterministic problem, one can derive the desired asymptotic characterizations.

Of course implementing the above steps involved problem-specific intricate calculations. Our proofs of Theorems 3.1, 3.2, 3.3 in the supplementary follow this general strategy.

B Proof of theorems and technical lemmas

B.1 Proof of Lemma 2.1

By substituting for y from (2.1) in the definition of risk we obtain

$$\begin{aligned}
\mathrm{Risk}(\boldsymbol{\theta}) &= \mathbb{E}[(y - \boldsymbol{x}^\top \boldsymbol{\theta})^2] \\
&= \mathbb{E}[(\boldsymbol{x}^\top(\boldsymbol{\theta}_0 - \boldsymbol{\theta}))^2] + \mathbb{E}[\boldsymbol{\varepsilon}^2] \\
&\overset{(a)}{=} \sum_{\ell\in[k]} \pi_\ell \, \mathbb{E}[((\boldsymbol{\mu}_\ell + \boldsymbol{z})^\top(\boldsymbol{\theta}_0 - \boldsymbol{\theta}))^2] + \mathbb{E}[\boldsymbol{\varepsilon}^2] \\
&= \sum_{\ell\in[k]} \pi_\ell \, \mathbb{E}[(\boldsymbol{\mu}_\ell^\top(\boldsymbol{\theta}_0 - \boldsymbol{\theta}))^2] + \sum_{\ell\in[k]} \pi_\ell \, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2}^2 + \sigma^2 \\
&\overset{(b)}{=} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \boldsymbol{M}\mathrm{diag}(\boldsymbol{\pi})\boldsymbol{M}^\top(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) + \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|_{\ell_2} + \sigma^2\,,
\end{aligned}$$

where (a) follows from the Gaussian-Mixture model (2.2) and (b) holds since $\sum_{\ell\in[k]} \pi_\ell = 1$.

B.2 Proof of Theorem 3.1 and Theorem 3.2

Recall that the look-alike estimator is defined as the min-norm estimator over the feature matrix X_L, where the look-alike representations are used instead of individual sensitive features; see (3.2).

To analyze risk of $\widehat{\boldsymbol{\theta}}_L$, we consider the ridge regression estimator given by

$$\widehat{\boldsymbol{\theta}}_\lambda = \arg\min_{\boldsymbol{\theta}} \frac{1}{2n} \left\|\boldsymbol{y} - \boldsymbol{X}_L^\top \boldsymbol{\theta}\right\|_{\ell_2}^2 + \lambda \|\boldsymbol{\theta}\|_{\ell_2}^2 \,.$$

The minimum-norm estimator is given by $\widehat{\boldsymbol{\theta}}_L = \lim_{\lambda\to 0^+} \widehat{\boldsymbol{\theta}}_\lambda$.

We follow the CGMT framework explained in Section A. Recall that

$$\boldsymbol{X}_L = \begin{bmatrix} \boldsymbol{M}_{\mathrm{s}}\boldsymbol{\Lambda} \\ \boldsymbol{M}_{\mathrm{ns}}\boldsymbol{\Lambda} + \boldsymbol{Z}_{\mathrm{ns}} \end{bmatrix},$$

and therefore by substituting for $\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{X}_L$, we get

$$\begin{aligned}
\frac{1}{2n}\left\|\boldsymbol{y} - \boldsymbol{X}_L^\top \boldsymbol{\theta}\right\|_{\ell_2}^2 &= \frac{1}{2n}\left\|\boldsymbol{\varepsilon} + \boldsymbol{X}^\top\boldsymbol{\theta}_0 - \boldsymbol{X}_L^\top \boldsymbol{\theta}\right\|_{\ell_2}^2 \\
&= \frac{1}{2n}\left\|\boldsymbol{\varepsilon} + \boldsymbol{\Lambda}^\top \boldsymbol{M}_{\mathrm{s}}^\top(\boldsymbol{\theta}_{0,\mathrm{s}} - \boldsymbol{\theta}_{\mathrm{s}}) + \boldsymbol{Z}_{\mathrm{s}}^\top\boldsymbol{\theta}_{0,\mathrm{s}} + (\boldsymbol{\Lambda}^\top\boldsymbol{M}_{\mathrm{ns}}^\top + \boldsymbol{Z}_{ns}^\top)(\boldsymbol{\theta}_{0,\mathrm{ns}} - \boldsymbol{\theta}_{\mathrm{ns}})\right\|_{\ell_2}^2 .
\end{aligned}$$

We define the primary optimization loss as follows:

$$\mathcal{L}_{PO}(\boldsymbol{\theta}_{\mathrm{s}}, \boldsymbol{\theta}_{\mathrm{ns}}) := \frac{1}{2n}\left\|\boldsymbol{\varepsilon} + \boldsymbol{\Lambda}^\top \boldsymbol{M}_{\mathrm{s}}^\top(\boldsymbol{\theta}_{0,\mathrm{s}} - \boldsymbol{\theta}_{\mathrm{s}}) + \boldsymbol{Z}_{\mathrm{s}}^\top\boldsymbol{\theta}_{0,\mathrm{s}} + (\boldsymbol{\Lambda}^\top\boldsymbol{M}_{\mathrm{ns}}^\top + \boldsymbol{Z}_{ns}^\top)(\boldsymbol{\theta}_{0,\mathrm{ns}} - \boldsymbol{\theta}_{\mathrm{ns}})\right\|_{\ell_2}^2 + \lambda\|\boldsymbol{\theta}_{\mathrm{s}}\|_{\ell_2}^2 + \lambda\|\boldsymbol{\theta}_{\mathrm{ns}}\|_{\ell_2}^2$$

We continue by deriving the auxiliary optimization (AO) problem. By duality, we have

$$\begin{aligned}
\mathcal{L}_{PO}(\boldsymbol{\theta}_{\mathrm{s}}, \boldsymbol{\theta}_{\mathrm{ns}}) = \max_{\boldsymbol{v}} \frac{1}{n}\Big(&\boldsymbol{v}^\top\boldsymbol{\varepsilon} + \boldsymbol{v}^\top\boldsymbol{\Lambda}^\top\boldsymbol{M}_{\mathrm{s}}^\top(\boldsymbol{\theta}_{0,\mathrm{s}} - \boldsymbol{\theta}_{\mathrm{s}}) + \boldsymbol{v}^\top\boldsymbol{Z}_{\mathrm{s}}^\top\boldsymbol{\theta}_{0,\mathrm{s}} + \boldsymbol{v}^\top(\boldsymbol{\Lambda}^\top\boldsymbol{M}_{\mathrm{ns}}^\top + \boldsymbol{Z}_{ns}^\top)(\boldsymbol{\theta}_{0,\mathrm{ns}} - \boldsymbol{\theta}_{\mathrm{ns}}) - \frac{\|\boldsymbol{v}\|_{\ell_2}^2}{2}\Big) \\
&+ \lambda\|\boldsymbol{\theta}_{\mathrm{s}}\|_{\ell_2}^2 + \lambda\|\boldsymbol{\theta}_{\mathrm{ns}}\|_{\ell_2}^2
\end{aligned}$$

Note that the above is jointly convex in $(\boldsymbol{\theta}_{\mathrm{s}}, \boldsymbol{\theta}_{\mathrm{ns}})$ and concave in $\boldsymbol{v}$, and the Gaussian matrix $\boldsymbol{Z}$ is independent of everything else. Therefore, the AO problem reads:

$$\begin{aligned}
\mathcal{L}_{AO}(\boldsymbol{\theta}_{\mathrm{s}}, \boldsymbol{\theta}_{\mathrm{ns}}) = \max_{\boldsymbol{v}} \frac{1}{n}\Big(&\boldsymbol{v}^\top\boldsymbol{\varepsilon} + \boldsymbol{v}^\top\boldsymbol{\Lambda}^\top\boldsymbol{M}_{\mathrm{s}}^\top(\boldsymbol{\theta}_{0,\mathrm{s}} - \boldsymbol{\theta}_{\mathrm{s}}) \\
&+ \|\boldsymbol{\theta}_{0,\mathrm{s}}\|_{\ell_2}\boldsymbol{g}_{\mathrm{s}}^\top\boldsymbol{v} + \|\boldsymbol{v}\|_{\ell_2}\boldsymbol{h}_{\mathrm{s}}^\top\boldsymbol{\theta}_{0,\mathrm{s}} \\
&+ \|\boldsymbol{\theta}_{0,\mathrm{ns}} - \boldsymbol{\theta}_{\mathrm{ns}}\|_{\ell_2}\boldsymbol{g}_{\mathrm{ns}}^\top\boldsymbol{v} + \|\boldsymbol{v}\|_{\ell_2}\boldsymbol{h}_{\mathrm{ns}}^\top(\boldsymbol{\theta}_{0,\mathrm{ns}} - \boldsymbol{\theta}_{\mathrm{ns}}) \\
&+ \boldsymbol{v}^\top\boldsymbol{\Lambda}^\top\boldsymbol{M}_{\mathrm{ns}}^\top(\boldsymbol{\theta}_{0,\mathrm{ns}} - \boldsymbol{\theta}_{\mathrm{ns}}) - \frac{\|\boldsymbol{v}\|_{\ell_2}^2}{2}\Big) + \lambda\|\boldsymbol{\theta}_{\mathrm{s}}\|_{\ell_2}^2 + \lambda\|\boldsymbol{\theta}_{\mathrm{ns}}\|_{\ell_2}^2\,,
\end{aligned}$$

where $\boldsymbol{g}_{\rm s}, \boldsymbol{g}_{\rm ns} \in \mathbb{R}^n$ and $\boldsymbol{h}_{\rm s} \in \mathbb{R}^p$, $\boldsymbol{h}_{\rm ns} \in \mathbb{R}^{d-p}$ are independent Gaussian random vectors with i.i.d $\mathsf{N}(0,1)$ entries.

We next fix norm of $\|\boldsymbol{v}\|_{\ell_2} = \beta$, and maximize over its direction to obtain

$$
\begin{aligned}
\mathcal{L}_{AO}(\boldsymbol{\theta}_{\rm s}, \boldsymbol{\theta}_{\rm ns}) &= \max_{\beta \geq 0} \frac{1}{n} \Big(\beta \left\| \boldsymbol{\varepsilon} + \boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{M}_{\rm s}^{\mathsf{T}} (\boldsymbol{\theta}_{0,\rm s} - \boldsymbol{\theta}_{\rm s}) + \|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2} \boldsymbol{g}_{\rm s} + \|\boldsymbol{\theta}_{0,\rm ns} - \boldsymbol{\theta}_{\rm ns}\|_{\ell_2} \boldsymbol{g}_{\rm ns} + \boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{M}_{\rm ns}^{\mathsf{T}} (\boldsymbol{\theta}_{0,\rm ns} - \boldsymbol{\theta}_{\rm ns}) \right\|_{\ell_2} \\
&\qquad + \beta \boldsymbol{h}_{\rm s}^{\mathsf{T}} \boldsymbol{\theta}_{0,\rm s} + \beta \boldsymbol{h}_{\rm ns}^{\mathsf{T}} (\boldsymbol{\theta}_{0,\rm ns} - \boldsymbol{\theta}_{\rm ns}) - \frac{\beta^2}{2} \Big) + \lambda \|\boldsymbol{\theta}_{\rm s}\|_{\ell_2}^2 + \lambda \|\boldsymbol{\theta}_{\rm ns}\|_{\ell_2}^2 \\
&= \max_{\beta \geq 0} \frac{1}{n} \Big(\beta \left\| \boldsymbol{\varepsilon} + \boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{M}_{\rm s}^{\mathsf{T}} (\boldsymbol{\theta}_{0,\rm s} - \boldsymbol{\theta}_{\rm s}) + \boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{M}_{\rm ns}^{\mathsf{T}} (\boldsymbol{\theta}_{0,\rm ns} - \boldsymbol{\theta}_{\rm ns}) + \sqrt{\|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2 + \|\boldsymbol{\theta}_{0,\rm ns} - \boldsymbol{\theta}_{\rm ns}\|_{\ell_2}^2} \boldsymbol{g} \right\|_{\ell_2} \\
&\qquad + \beta \boldsymbol{h}_{\rm s}^{\mathsf{T}} \boldsymbol{\theta}_{0,\rm s} + \beta \boldsymbol{h}_{\rm ns}^{\mathsf{T}} (\boldsymbol{\theta}_{0,\rm ns} - \boldsymbol{\theta}_{\rm ns}) - \frac{\beta^2}{2} \Big) + \lambda \|\boldsymbol{\theta}_{\rm s}\|_{\ell_2}^2 + \lambda \|\boldsymbol{\theta}_{\rm ns}\|_{\ell_2}^2 ,
\end{aligned}
$$

where we used that $\boldsymbol{g}_{\rm s}, \boldsymbol{g}_{\rm ns} \in \mathbb{R}^n$ have independent Gaussian entries. Here, $\boldsymbol{g} \in \mathbb{R}^n$ has i.i.d entries from $\mathsf{N}(0,1)$. Next, note that the above optimization over β has a closed form. Using the identity $\max_{\beta \geq 0} (\beta x - \beta^2/2) = x_+^2/2$, with $x_+ = \max(x, 0)$, we get

$$
\begin{aligned}
\mathcal{L}_{AO}(\boldsymbol{\theta}_{\rm s}, \boldsymbol{\theta}_{\rm ns}) = \frac{1}{2n} \Big(& \left\| \boldsymbol{\varepsilon} + \boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{M}_{\rm s}^{\mathsf{T}} (\boldsymbol{\theta}_{0,\rm s} - \boldsymbol{\theta}_{\rm s}) + \boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{M}_{\rm ns}^{\mathsf{T}} (\boldsymbol{\theta}_{0,\rm ns} - \boldsymbol{\theta}_{\rm ns}) + \sqrt{\|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2 + \|\boldsymbol{\theta}_{0,\rm ns} - \boldsymbol{\theta}_{\rm ns}\|_{\ell_2}^2} \boldsymbol{g} \right\|_{\ell_2} \\
& + \boldsymbol{h}_{\rm s}^{\mathsf{T}} \boldsymbol{\theta}_{0,\rm s} + \boldsymbol{h}_{\rm ns}^{\mathsf{T}} (\boldsymbol{\theta}_{0,\rm ns} - \boldsymbol{\theta}_{\rm ns}) \Big)_+^2 + \lambda \|\boldsymbol{\theta}_{\rm s}\|_{\ell_2}^2 + \lambda \|\boldsymbol{\theta}_{\rm ns}\|_{\ell_2}^2 . \qquad \text{(B.1)}
\end{aligned}
$$

Scalarization of the auxiliary optimization (AO) problem. We next proceed to scalarize the AO problem. Consider the singular value decomposition

$$
\boldsymbol{M}_{\rm s} = \boldsymbol{U}_{\rm s} \boldsymbol{\Sigma}_{\rm s} \boldsymbol{V}_{\rm s}^{\mathsf{T}} ,
$$

with $\boldsymbol{U}_{\rm s} \in \mathbb{R}^{p \times r}$, $\boldsymbol{\Sigma}_{\rm s} \in \mathbb{R}^{r \times r}$, $\boldsymbol{V}_{\rm s} \in \mathbb{R}^{k \times r}$, where $r = {\rm rank}(\boldsymbol{M}_{\rm s}) \leq k$. Decompose $\boldsymbol{q}_{\rm s} := \boldsymbol{\theta}_{0,\rm s} - \boldsymbol{\theta}_{\rm s}$ in its projections onto the space spanned by the columns $\boldsymbol{u}_{1,\rm s}, \dots, \boldsymbol{u}_{r,\rm s}$ of $\boldsymbol{U}_{\rm s}$, and the orthogonal component:

$$
\boldsymbol{q}_{\rm s} = \sum_{i=1}^{r} \alpha_i \boldsymbol{u}_{i,\rm s} + \alpha_0 \boldsymbol{q}_{\rm s}^{\perp} ,
$$

where $\|\boldsymbol{q}_{\rm s}^{\perp}\|_{\ell_2} = 1$, $\alpha_0 \geq 0$, and $\boldsymbol{U}_{\rm s}^{\mathsf{T}} \boldsymbol{q}_{\rm s}^{\perp} = \boldsymbol{0}$. Using the shorthand $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_r)$, we write

$$
\boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{M}_{\rm s}^{\mathsf{T}} (\boldsymbol{\theta}_{0,\rm s} - \boldsymbol{\theta}_{\rm s}) = \boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{V}_{\rm s} \boldsymbol{\Sigma}_{\rm s} \boldsymbol{U}_{\rm s}^{\mathsf{T}} \boldsymbol{q}_{\rm s} = \boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{V}_{\rm s} \boldsymbol{\Sigma}_{\rm s} \boldsymbol{\alpha} .
$$

In addition,

$$
\begin{aligned}
\|\boldsymbol{\theta}_{\rm s}\|_{\ell_2}^2 &= \|\boldsymbol{\theta}_{0,\rm s} - (\boldsymbol{\theta}_{0,\rm s} - \boldsymbol{\theta}_{\rm s})\|_{\ell_2}^2 \\
&= \|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2 + \|\boldsymbol{q}_{\rm s}\|_{\ell_2}^2 - 2\langle \boldsymbol{\theta}_{0,\rm s}, \boldsymbol{q}_{\rm s} \rangle \\
&= \|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2 + \|\boldsymbol{q}_{\rm s}\|_{\ell_2}^2 - 2\langle \boldsymbol{\theta}_{0,\rm s}, \boldsymbol{U}_{\rm s} \boldsymbol{\alpha} \rangle - 2\alpha_0 \langle \boldsymbol{\theta}_{0,\rm s}, \boldsymbol{q}_{\rm s}^{\perp} \rangle \\
&= \|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2 + \|\boldsymbol{q}_{\rm s}\|_{\ell_2}^2 - 2\langle \boldsymbol{U}_{\rm s}^{\mathsf{T}} \boldsymbol{\theta}_{0,\rm s}, \boldsymbol{\alpha} \rangle - 2\alpha_0 \langle \boldsymbol{\theta}_{0,\rm s}, \boldsymbol{q}_{\rm s}^{\perp} \rangle \\
&= \|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2 + (\alpha_0^2 + \|\boldsymbol{\alpha}\|_{\ell_2}^2) - 2\langle \boldsymbol{U}_{\rm s}^{\mathsf{T}} \boldsymbol{\theta}_{0,\rm s}, \boldsymbol{\alpha} \rangle - 2\alpha_0 \langle \boldsymbol{\theta}_{0,\rm s}, \boldsymbol{q}_{\rm s}^{\perp} \rangle . \qquad \text{(B.2)}
\end{aligned}
$$

Similarly, we define $\boldsymbol{q}_{\rm ns} = \boldsymbol{\theta}_{0,\rm ns} - \boldsymbol{\theta}_{\rm ns}$ and consider the singular value decomposition

$$
\boldsymbol{M}_{\rm ns} = \boldsymbol{U}_{\rm ns} \boldsymbol{\Sigma}_{\rm ns} \boldsymbol{V}_{\rm ns}^{\mathsf{T}} ,
$$

with $\boldsymbol{U}_{\rm ns} \in \mathbb{R}^{(d-p) \times t}$, $\boldsymbol{\Sigma}_{\rm ns} \in \mathbb{R}^{t \times t}$, $\boldsymbol{V}_{\rm ns} \in \mathbb{R}^{k \times t}$, where $t = {\rm rank}(\boldsymbol{M}_{\rm ns}) \leq k$. Decomposing $\boldsymbol{q}_{\rm ns}$ in its projections on the orthogonal columns $\boldsymbol{u}_{1,\rm ns}, \dots, \boldsymbol{u}_{r,\rm ns}$ of $\boldsymbol{U}_{\rm ns}$, and the orthogonal component we write

$$
\boldsymbol{q}_{\rm ns} = \sum_{i=1}^{t} \gamma_i \boldsymbol{u}_{i,\rm ns} + \gamma_0 \boldsymbol{q}_{\rm ns}^{\perp} ,
$$

with $\|\boldsymbol{q}_{\rm ns}^{\perp}\|_{\ell_2} = 1$, $\gamma_0 \geq 0$, and $\boldsymbol{U}_{\rm ns}^{\mathsf{T}} \boldsymbol{q}_{\rm ns}^{\perp} = \boldsymbol{0}$. Define $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_t)$. In this notation, we have

$$
\boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{M}_{\rm ns}^{\mathsf{T}} (\boldsymbol{\theta}_{0,\rm ns} - \boldsymbol{\theta}_{\rm ns}) = \boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{V}_{\rm ns} \boldsymbol{\Sigma}_{\rm ns} \boldsymbol{U}_{\rm ns}^{\mathsf{T}} \boldsymbol{q}_{\rm ns} = \boldsymbol{\Lambda}^{\mathsf{T}} \boldsymbol{V}_{\rm ns} \boldsymbol{\Sigma}_{\rm ns} \boldsymbol{\gamma} .
$$

Also, $\|\boldsymbol{\theta}_{0,\mathrm{ns}} - \boldsymbol{\theta}_{\mathrm{ns}}\|_{\ell_2} = \|\boldsymbol{q}_{\mathrm{ns}}\|_{\ell_2} = \sqrt{\gamma_0^2 + \|\boldsymbol{\gamma}\|_{\ell_2}^2}$. In addition,

$$\boldsymbol{h}_{\mathrm{ns}}^{\mathsf{T}}(\boldsymbol{\theta}_{0,\mathrm{ns}} - \boldsymbol{\theta}_{\mathrm{ns}}) = \boldsymbol{h}_{\mathrm{ns}}^{\mathsf{T}}\boldsymbol{q}_{\mathrm{ns}} = \sum_{i=1}^{t} \gamma_i \boldsymbol{h}_{\mathrm{ns}}^{\mathsf{T}}\boldsymbol{u}_{i,\mathrm{ns}} + \gamma_0 \boldsymbol{h}_{\mathrm{ns}}^{\mathsf{T}}\boldsymbol{q}_{\mathrm{ns}}^{\perp} .$$

Using the above identities in (B.1), we have

$$\begin{aligned}\mathcal{L}_{AO}(\boldsymbol{\theta}_{\mathrm{s}}, \boldsymbol{\theta}_{\mathrm{ns}}) = \frac{1}{2n}\Big(&\Big\|\boldsymbol{\varepsilon} + \boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{V}_{\mathrm{s}}\boldsymbol{\Sigma}_{\mathrm{s}}\boldsymbol{\alpha} + \boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{V}_{\mathrm{ns}}\boldsymbol{\Sigma}_{\mathrm{ns}}\boldsymbol{\gamma} + \sqrt{\|\boldsymbol{\theta}_{0,\mathrm{s}}\|_{\ell_2}^2 + \gamma_0^2 + \|\boldsymbol{\gamma}\|_{\ell_2}^2}\,\boldsymbol{g}\Big\|_{\ell_2} \\ &+ \boldsymbol{h}_{\mathrm{s}}^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{s}} + \sum_{i=1}^{t}\gamma_i \boldsymbol{h}_{\mathrm{ns}}^{\mathsf{T}}\boldsymbol{u}_{i,\mathrm{ns}} + \gamma_0 \boldsymbol{h}_{\mathrm{ns}}^{\mathsf{T}}\boldsymbol{q}_{\mathrm{ns}}^{\perp}\Big)_+^2 \\ &+ \lambda\|\boldsymbol{\theta}_{0,\mathrm{s}}\|_{\ell_2}^2 + \lambda(\alpha_0^2 + \|\boldsymbol{\alpha}\|_{\ell_2}^2) - 2\lambda\langle \boldsymbol{U}_{\mathrm{s}}^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{s}}, \boldsymbol{\alpha}\rangle - 2\lambda\alpha_0\langle \boldsymbol{\theta}_{0,\mathrm{s}}, \boldsymbol{q}_{\mathrm{s}}^{\perp}\rangle \\ &+ \lambda\|\boldsymbol{\theta}_{0,\mathrm{ns}}\|_{\ell_2}^2 + \lambda(\gamma_0^2 + \|\boldsymbol{\gamma}\|_{\ell_2}^2) - 2\lambda\langle \boldsymbol{U}_{\mathrm{ns}}^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{ns}}, \boldsymbol{\gamma}\rangle - 2\lambda\gamma_0\langle \boldsymbol{\theta}_{0,\mathrm{ns}}, \boldsymbol{q}_{\mathrm{ns}}^{\perp}\rangle . \quad \text{(B.3)}\end{aligned}$$

By the above characterization, minimization over $\boldsymbol{\theta}_{\mathrm{s}}$ and $\boldsymbol{\theta}_{\mathrm{ns}}$ reduces to minimization over α_0, γ_0, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\boldsymbol{q}_{\mathrm{s}}^{\perp}$ and $\boldsymbol{q}_{\mathrm{ns}}^{\perp}$. Further, these variables are free from each other and can be optimized over separately. For $\boldsymbol{q}_{\mathrm{s}}^{\perp}$, there is only one term involving this variable and therefore, minimization over it reduces to

$$\min_{\boldsymbol{q}_{\mathrm{s}}^{\perp}, \|\boldsymbol{q}_{\mathrm{s}}^{\perp}\|_{\ell_2}=1} -\langle \boldsymbol{\theta}_{0,\mathrm{s}}, \boldsymbol{q}_{\mathrm{s}}^{\perp}\rangle = \min_{\boldsymbol{q}_{\mathrm{s}}^{\perp}, \|\boldsymbol{q}_{\mathrm{s}}^{\perp}\|_{\ell_2}=1} -\langle \boldsymbol{U}_{\mathrm{s}}^{\perp}(\boldsymbol{U}_{\mathrm{s}}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{s}}, \boldsymbol{q}_{\mathrm{s}}^{\perp}\rangle = -\left\|(\boldsymbol{U}_{\mathrm{s}}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{s}}\right\|_{\ell_2} .$$

For $\boldsymbol{q}_{\mathrm{ns}}^{\perp}$, we note that there are two terms involving this variable, namely $\langle \frac{\boldsymbol{h}_{\mathrm{ns}}}{\sqrt{n}}, \boldsymbol{q}_{\mathrm{ns}}^{\perp}\rangle$ and $\langle (\boldsymbol{U}_{\mathrm{ns}}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{ns}}, \boldsymbol{q}_{\mathrm{ns}}^{\perp}\rangle$. Since $\|\boldsymbol{q}_{\mathrm{ns}}^{\perp}\|_{\ell_2} = 1$, it is easy to see that the optimal $\boldsymbol{q}_{\mathrm{ns}}^{\perp}$ should be in the span of $\boldsymbol{h}_{\mathrm{ns}}^{\perp}$ and $(\boldsymbol{U}_{\mathrm{ns}}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{ns}}$. In addition,

$$\langle \frac{\boldsymbol{h}_{\mathrm{ns}}^{\perp}}{\sqrt{n}}, (\boldsymbol{U}_{\mathrm{ns}}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{ns}}\rangle \overset{(p)}{\to} 0 ,$$

by the law of large numbers. In words, these two vectors are asymptotically orthogonal. Hence, we can consider the following decomposition of the optimal $\boldsymbol{q}_{\mathrm{ns}}^{\perp}$:

$$\boldsymbol{q}_{\mathrm{ns}}^{\perp} = -\xi\frac{\boldsymbol{h}_{\mathrm{ns}}^{\perp}}{\|\boldsymbol{h}_{\mathrm{ns}}^{\perp}\|_{\ell_2}} + \sqrt{1-\xi^2}\frac{\boldsymbol{U}_{\mathrm{ns}}^{\perp}(\boldsymbol{U}_{\mathrm{ns}}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{ns}}}{\|(\boldsymbol{U}_{\mathrm{ns}}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{ns}}\|_{\ell_2}} ,$$

where $\xi \ge 0$ and $\boldsymbol{h}_{\mathrm{ns}}^{\perp}$ denotes the projection of $\boldsymbol{h}_{\mathrm{ns}}$ onto the (left) null space of $\boldsymbol{U}_{\mathrm{ns}}$. This brings us to

$$\begin{aligned}\min_{\boldsymbol{\theta}_{\mathrm{s}}, \boldsymbol{\theta}_{\mathrm{ns}}} \mathcal{L}_{AO}(\boldsymbol{\theta}_{\mathrm{s}}, \boldsymbol{\theta}_{\mathrm{ns}}) = \min_{\alpha_0, \gamma_0 \ge 0, \boldsymbol{\alpha}, \boldsymbol{\gamma}} \frac{1}{2}\Big(&\frac{1}{\sqrt{n}}\Big\|\boldsymbol{\varepsilon} + \boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{V}_{\mathrm{s}}\boldsymbol{\Sigma}_{\mathrm{s}}\boldsymbol{\alpha} + \boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{V}_{\mathrm{ns}}\boldsymbol{\Sigma}_{\mathrm{ns}}\boldsymbol{\gamma} + \sqrt{\|\boldsymbol{\theta}_{0,\mathrm{s}}\|_{\ell_2}^2 + \gamma_0^2 + \|\boldsymbol{\gamma}\|_{\ell_2}^2}\,\boldsymbol{g}\Big\|_{\ell_2} \\ &+ \frac{\boldsymbol{h}_{\mathrm{s}}^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{s}}}{\sqrt{n}} + \sum_{i=1}^{t}\gamma_i\frac{\boldsymbol{h}_{\mathrm{ns}}^{\mathsf{T}}\boldsymbol{u}_{i,\mathrm{ns}}}{\sqrt{n}} - \gamma_0\xi\frac{\|\boldsymbol{h}_{\mathrm{ns}}^{\perp}\|_{\ell_2}}{\sqrt{n}}\Big)_+^2 \\ &+ \lambda\|\boldsymbol{\theta}_{0,\mathrm{s}}\|_{\ell_2}^2 + \lambda(\alpha_0^2 + \|\boldsymbol{\alpha}\|_{\ell_2}^2) - 2\lambda\langle \boldsymbol{U}_{\mathrm{s}}^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{s}}, \boldsymbol{\alpha}\rangle - 2\lambda\alpha_0\left\|(\boldsymbol{U}_{\mathrm{s}}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{s}}\right\|_{\ell_2} \\ &+ \lambda\|\boldsymbol{\theta}_{0,\mathrm{ns}}\|_{\ell_2}^2 + \lambda(\gamma_0^2 + \|\boldsymbol{\gamma}\|_{\ell_2}^2) - 2\lambda\langle \boldsymbol{U}_{\mathrm{ns}}^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{ns}}, \boldsymbol{\gamma}\rangle - 2\lambda\gamma_0\sqrt{1-\xi^2}\left\|(\boldsymbol{U}_{\mathrm{ns}}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\mathrm{ns}}\right\|_{\ell_2} . \\ &\text{(B.4)}\end{aligned}$$

Note that at this stage, the AO problem is reduced to an optimization over $r + t + 3$ scalar variables ($\alpha_0, \gamma_0 \ge 0$, $0 \le \xi \le 1$ and $\boldsymbol{\alpha} \in \mathbb{R}^r$, $\boldsymbol{\gamma} \in \mathbb{R}^t$).

Convergence of the auxiliary optimization problem. We next continue to derive the point-wise in-probability limit of the AO problem.

First observe that since $\boldsymbol{\varepsilon}$ and $\boldsymbol{g}$ are independent with i.i.d $\mathsf{N}(0,1)$ entries, we have

$$\boldsymbol{\varepsilon} + \sqrt{\|\boldsymbol{\theta}_{0,\mathrm{s}}\|_{\ell_2}^2 + \gamma_0^2 + \|\boldsymbol{\gamma}\|_{\ell_2}^2}\,\boldsymbol{g} \overset{(d)}{=} \sqrt{\sigma^2 +^2 (\|\boldsymbol{\theta}_{0,\mathrm{s}}\|_{\ell_2}^2 + \gamma_0^2 + \|\boldsymbol{\gamma}\|_{\ell_2}^2)}\,\tilde{\boldsymbol{g}} ,$$

where $\tilde{\boldsymbol{g}} \in \mathbb{R}^n$ has i.i.d $\mathsf{N}(0,1)$ entries.

Second, by construction $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} = \mathrm{diag}(n_1, \dots, n_k) \in \mathbb{R}^{k\times k}$, where n_ℓ denotes the number of examples from cluster ℓ. Hence,

$$\begin{aligned}\frac{1}{n}\left\|\boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{V}_{\mathrm{s}}\boldsymbol{\Sigma}_{\mathrm{s}}\boldsymbol{\alpha} + \boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{V}_{\mathrm{ns}}\boldsymbol{\Sigma}_{\mathrm{ns}}\boldsymbol{\gamma}\right\|_{\ell_2}^2 &= (\boldsymbol{V}_{\mathrm{s}}\boldsymbol{\Sigma}_{\mathrm{s}}\boldsymbol{\alpha} + \boldsymbol{V}_{\mathrm{ns}}\boldsymbol{\Sigma}_{\mathrm{ns}}\boldsymbol{\gamma})^{\mathsf{T}}\mathrm{diag}(\tfrac{n_1}{n}, \dots, \tfrac{n_k}{k})(\boldsymbol{V}_{\mathrm{s}}\boldsymbol{\Sigma}_{\mathrm{s}}\boldsymbol{\alpha} + \boldsymbol{V}_{\mathrm{ns}}\boldsymbol{\Sigma}_{\mathrm{ns}}\boldsymbol{\gamma}) \\ &\overset{(p)}{\to} (\boldsymbol{V}_{\mathrm{s}}\boldsymbol{\Sigma}_{\mathrm{s}}\boldsymbol{\alpha} + \boldsymbol{V}_{\mathrm{ns}}\boldsymbol{\Sigma}_{\mathrm{ns}}\boldsymbol{\gamma})^{\mathsf{T}}\mathrm{diag}(\boldsymbol{\pi})(\boldsymbol{V}_{\mathrm{s}}\boldsymbol{\Sigma}_{\mathrm{s}}\boldsymbol{\alpha} + \boldsymbol{V}_{\mathrm{ns}}\boldsymbol{\Sigma}_{\mathrm{ns}}\boldsymbol{\gamma})\end{aligned}$$

Next, by using concentration of Lipschitz functions of Gaussian vectors, we obtain

$$\begin{aligned}&\frac{1}{\sqrt{n}}\left\|\boldsymbol{\varepsilon}+\boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{V}_{\rm s}\boldsymbol{\Sigma}_{\rm s}\boldsymbol{\alpha}+\boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{V}_{\rm ns}\boldsymbol{\Sigma}_{\rm ns}\boldsymbol{\gamma}+\sqrt{\|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2+\gamma_0^2+\|\boldsymbol{\gamma}\|_{\ell_2}^2}\,\boldsymbol{g}\right\|_{\ell_2}\\ &\overset{p}{\to}\sqrt{(\boldsymbol{V}_{\rm s}\boldsymbol{\Sigma}_{\rm s}\boldsymbol{\alpha}+\boldsymbol{V}_{\rm ns}\boldsymbol{\Sigma}_{\rm ns}\boldsymbol{\gamma})^{\mathsf{T}}{\rm diag}(\boldsymbol{\pi})(\boldsymbol{V}_{\rm s}\boldsymbol{\Sigma}_{\rm s}\boldsymbol{\alpha}+\boldsymbol{V}_{\rm ns}\boldsymbol{\Sigma}_{\rm ns}\boldsymbol{\gamma})+\sigma^2+(\|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2+\gamma_0^2+\|\boldsymbol{\gamma}\|_{\ell_2}^2)}\end{aligned}$$

Also, since $\|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}$ is bounded and $\|\boldsymbol{u}_{i,\rm s}\|_{\ell_2}=1$, we get

$$\frac{\boldsymbol{h}_{\rm s}^{\mathsf{T}}\boldsymbol{\theta}_{0,\rm s}}{\sqrt{n}},\ \frac{\boldsymbol{h}_{\rm ns}^{\mathsf{T}}\boldsymbol{u}_{i,\rm ns}}{\sqrt{n}}\overset{(p)}{\to}0\,.$$

In addition, $\|\boldsymbol{h}_{\rm ns}^{\perp}\|_{\ell_2}$ concentrates around $\sqrt{d-p-t}$ and $(d-p-t)/n\to\psi_d-\psi_p$, because $t\le k$ remains bounded as n diverges, and so

$$\frac{\|\boldsymbol{h}_{\rm ns}^{\perp}\|_{\ell_2}}{\sqrt{n}}\overset{(p)}{\to}\sqrt{\psi_d-\psi_p}\,.$$

Using the above limits, the objective in (B.4) converges in-probability to

$$\begin{aligned}&\mathcal{D}(\alpha_0,\gamma_0,\xi,\boldsymbol{\alpha},\boldsymbol{\gamma}):=\\ &\frac{1}{2}\left(\sqrt{(\boldsymbol{V}_{\rm s}\boldsymbol{\Sigma}_{\rm s}\boldsymbol{\alpha}+\boldsymbol{V}_{\rm ns}\boldsymbol{\Sigma}_{\rm ns}\boldsymbol{\gamma})^{\mathsf{T}}{\rm diag}(\boldsymbol{\pi})(\boldsymbol{V}_{\rm s}\boldsymbol{\Sigma}_{\rm s}\boldsymbol{\alpha}+\boldsymbol{V}_{\rm ns}\boldsymbol{\Sigma}_{\rm ns}\boldsymbol{\gamma})+\sigma^2+(\|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2+\gamma_0^2+\|\boldsymbol{\gamma}\|_{\ell_2}^2)}-\gamma_0\xi\sqrt{\psi_d-\psi_p}\right)_+^2\\ &+\lambda\|\boldsymbol{\theta}_0\|_{\ell_2}^2+\lambda(\alpha_0^2+\gamma_0^2+\|\boldsymbol{\alpha}\|_{\ell_2}^2+\|\boldsymbol{\gamma}\|_{\ell_2}^2)\\ &-2\lambda\left(\langle\boldsymbol{U}_{\rm s}^{\mathsf{T}}\boldsymbol{\theta}_{0,\rm s},\boldsymbol{\alpha}\rangle+\alpha_0\left\|(\boldsymbol{U}_{\rm s}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\rm s}\right\|_{\ell_2}+\langle\boldsymbol{U}_{\rm ns}^{\mathsf{T}}\boldsymbol{\theta}_{0,\rm ns},\boldsymbol{\gamma}\rangle+\gamma_0\sqrt{1-\xi^2}\left\|(\boldsymbol{U}_{\rm ns}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\rm ns}\right\|_{\ell_2}\right)\end{aligned}\tag{B.5}$$

We are now ready to prove the theorems.

B.2.1 Proof of Theorem 3.1

Using Lemma 2.1, we have

$$\begin{aligned}{\rm Risk}(\widehat{\boldsymbol{\theta}}_L)&=\sigma^2+\|\boldsymbol{\theta}_0-\boldsymbol{\theta}\|_{\ell_2}^2+(\boldsymbol{\theta}_0-\boldsymbol{\theta})^{\mathsf{T}}\boldsymbol{M}{\rm diag}(\boldsymbol{\pi})\boldsymbol{M}^{\mathsf{T}}(\boldsymbol{\theta}_0-\boldsymbol{\theta})\\ &=\sigma^2+\|\boldsymbol{q}_{\rm s}\|_{\ell_2}^2+\|\boldsymbol{q}_{\rm ns}\|_{\ell_2}^2+\boldsymbol{q}_{\rm s}^{\mathsf{T}}\boldsymbol{M}_{\rm s}{\rm diag}(\boldsymbol{\pi})\boldsymbol{M}_{\rm s}^{\mathsf{T}}\boldsymbol{q}_{\rm s}+\boldsymbol{q}_{\rm ns}^{\mathsf{T}}\boldsymbol{M}_{\rm ns}{\rm diag}(\boldsymbol{\pi})\boldsymbol{M}_{\rm ns}^{\mathsf{T}}\boldsymbol{q}_{\rm ns}\\ &=\sigma^2+(\alpha_0^2+\gamma_0^2+\|\boldsymbol{\alpha}\|_{\ell_2}^2+\|\boldsymbol{\gamma}\|_{\ell_2}^2)\\ &\quad+\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\Sigma}_{\rm s}\boldsymbol{V}_{\rm s}^{\mathsf{T}}{\rm diag}(\boldsymbol{\pi})\boldsymbol{V}_{\rm s}\boldsymbol{\Sigma}_{\rm s}\boldsymbol{\alpha}+\boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{\Sigma}_{\rm ns}\boldsymbol{V}_{\rm ns}^{\mathsf{T}}{\rm diag}(\boldsymbol{\pi})\boldsymbol{V}_{\rm ns}\boldsymbol{\Sigma}_{\rm ns}\boldsymbol{\gamma}\,.\end{aligned}\tag{B.6}$$

Since $\psi_d-\psi_p\le 1$, we are in the over- determined (a.k.a underparametrized) regime. As $\lambda\to 0^+$, the terms involving λ become negligible compared to the first term in (B.5) except those that include α_0, as α_0 is not present in the first term . Since $(x)_+^2$ is increasing, and

$$(\boldsymbol{V}_{\rm s}\boldsymbol{\Sigma}_{\rm s}\boldsymbol{\alpha}+\boldsymbol{V}_{\rm ns}\boldsymbol{\Sigma}_{\rm ns}\boldsymbol{\gamma})^{\mathsf{T}}{\rm diag}(\boldsymbol{\pi})(\boldsymbol{V}_{\rm s}\boldsymbol{\Sigma}_{\rm s}\boldsymbol{\alpha}+\boldsymbol{V}_{\rm ns}\boldsymbol{\Sigma}_{\rm ns}\boldsymbol{\gamma})+\|\boldsymbol{\gamma}\|_{\ell_2}^2\ge 0,$$

the minimum over $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ is achieved for $\boldsymbol{\alpha}=\boldsymbol{0}\in\mathbb{R}^r$ and $\boldsymbol{\gamma}=\boldsymbol{0}\in\mathbb{R}^t$. The optimization (B.5) then reduces to

$$\min_{\alpha_0,\gamma_0\ge0,0\le\xi\le1}\frac{1}{2}\left(\sqrt{\sigma^2+(\|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2+\gamma_0^2)}-\gamma_0\xi\sqrt{\psi_d-\psi_p}\right)_+^2+\lambda\alpha_0^2-2\lambda\alpha_0\left\|(\boldsymbol{U}_{\rm s}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\rm s}\right\|_{\ell_2}\,.\tag{B.7}$$

The optimal ξ is given by $\xi=1$. Also, setting derivative with respect to α_0 to zero we obtain the optimal $\alpha_0=\left\|(\boldsymbol{U}_{\rm s}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\rm s}\right\|_{\ell_2}$. Next, by setting derivative with respect to γ_0 we arrive at

$$\gamma_0^2=(\sigma^2+\|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2)\frac{\psi_d-\psi_p}{1-(\psi_d-\psi_p)}\,.$$

Using the optimal variables in (B.6) we obtain the risk of minimum-norm estimator as

$$\begin{aligned}{\rm Risk}(\widehat{\boldsymbol{\theta}}_L)&=\sigma^2+\left\|(\boldsymbol{U}_{\rm s}^{\perp})^{\mathsf{T}}\boldsymbol{\theta}_{0,\rm s}\right\|_{\ell_2}^2+(\sigma^2+\|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2)\frac{\psi_d-\psi_p}{1-(\psi_d-\psi_p)}\\ &=(\sigma^2+\|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2)\frac{1}{1-(\psi_d-\psi_p)}-\left\|\boldsymbol{U}_{\rm s}^{\mathsf{T}}\boldsymbol{\theta}_{0,\rm s}\right\|_{\ell_2}^2\,.\end{aligned}$$

Recall that by assumption, $r_{\rm s}=\|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}$ and $\left\|\boldsymbol{U}_{\rm s}^{\mathsf{T}}\boldsymbol{\theta}_{0,\rm s}\right\|_{\ell_2}=\sqrt{\rho}r_{\rm s}$, which completes the proof.

B.2.2 Proof of Theorem 3.2

We continue from (B.5). In the case of $\psi_d - \psi_p \le 1$, it is easy to see that the derivative of the first term of (B.5), in the active region is decreasing in γ_0. With the consideration $\lambda \to 0^+$, minimizing over γ_0 will push us into the non-active region. Therefore the optimization problem (B.5) reduces to

$$\begin{aligned}\text{minimize}\quad & \|\boldsymbol{\theta}_0\|_{\ell_2}^2 + \alpha_0^2 + \gamma_0^2 + \|\boldsymbol{\alpha}\|_{\ell_2}^2 + \|\boldsymbol{\gamma}\|_{\ell_2}^2 \\ & - 2\left(\langle \boldsymbol{U}_{\rm s}^{\mathsf T}\boldsymbol{\theta}_{0,\rm s}, \boldsymbol{\alpha}\rangle + \alpha_0 \left\|(\boldsymbol{U}_{\rm s}^{\perp})^{\mathsf T}\boldsymbol{\theta}_{0,\rm s}\right\|_{\ell_2} + \langle \boldsymbol{U}_{\rm ns}^{\mathsf T}\boldsymbol{\theta}_{0,\rm ns}, \boldsymbol{\gamma}\rangle + \gamma_0\sqrt{1-\xi^2}\left\|(\boldsymbol{U}_{\rm ns}^{\perp})^{\mathsf T}\boldsymbol{\theta}_{0,\rm ns}\right\|_{\ell_2}\right)\end{aligned}$$

subject to

$$(\boldsymbol{V}_{\rm s}\boldsymbol{\Sigma}_{\rm s}\boldsymbol{\alpha} + \boldsymbol{V}_{\rm ns}\boldsymbol{\Sigma}_{\rm ns}\boldsymbol{\gamma})^{\mathsf T}{\rm diag}(\boldsymbol{\pi})(\boldsymbol{V}_{\rm s}\boldsymbol{\Sigma}_{\rm s}\boldsymbol{\alpha} + \boldsymbol{V}_{\rm ns}\boldsymbol{\Sigma}_{\rm ns}\boldsymbol{\gamma}) + \sigma^2 + (\|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2 + \gamma_0^2 + \|\boldsymbol{\gamma}\|_{\ell_2}^2) \le \gamma_0^2\xi^2(\psi_d - \psi_p) \tag{B.8}$$

By Assumption 2, $\boldsymbol{\Sigma}_{\rm s} = \mu\boldsymbol{I}_k$, $\boldsymbol{V}_{\rm s} = \boldsymbol{I}_{k\times k}$, and $\boldsymbol{\Sigma}_{\rm ns} = \boldsymbol{0}$, $\boldsymbol{U}_{\rm ns} = \boldsymbol{0}$ (no cluster structure on non-sensitive features and an orthogonal, equal energy cluster centers on the sensitive features). Therefore, by fixing $\gamma := \|\boldsymbol{\gamma}\|_{\ell_2}$, the optimization problem (B.8) becomes:

$$\begin{aligned}\text{minimize}\quad & \alpha_0^2 + \gamma_0^2 + \|\boldsymbol{\alpha}\|_{\ell_2}^2 + \gamma^2 - 2\left(\langle \boldsymbol{U}_{\rm s}^{\mathsf T}\boldsymbol{\theta}_{0,\rm s}, \boldsymbol{\alpha}\rangle + \alpha_0\left\|(\boldsymbol{U}_{\rm s}^{\perp})^{\mathsf T}\boldsymbol{\theta}_{0,\rm s}\right\|_{\ell_2} + \gamma_0\sqrt{1-\xi^2}\,\|\boldsymbol{\theta}_{0,\rm ns}\|_{\ell_2}\right)\\ \text{subject to}\quad & \\ & \mu^2\boldsymbol{\alpha}^{\mathsf T}{\rm diag}(\boldsymbol{\pi})\boldsymbol{\alpha} + \sigma^2 + \|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2 + \gamma_0^2 + \gamma^2 \le \gamma_0^2\xi^2(\psi_d - \psi_p)\,.\end{aligned} \tag{B.9}$$

Since α_0 does not appear in the constraint, it is easy to see that its optimal value is given by $\alpha_0 = \left\|(\boldsymbol{U}_{\rm s}^{\perp})^{\mathsf T}\boldsymbol{\theta}_{0,\rm s}\right\|_{\ell_2}$. Also, note that by decreasing γ the objective value decreases and also by the constraint on the other variables become more relaxed. Consequently, the optimal value of γ is $\gamma = 0$. Removing α_0 from the objective function, we are left with

$$\begin{aligned}\text{minimize}\quad & \gamma_0^2 + \|\boldsymbol{\alpha}\|_{\ell_2}^2 - 2\left(\langle \boldsymbol{U}_{\rm s}^{\mathsf T}\boldsymbol{\theta}_{0,\rm s}, \boldsymbol{\alpha}\rangle + \gamma_0\sqrt{1-\xi^2}\,\|\boldsymbol{\theta}_{0,\rm ns}\|_{\ell_2}\right)\\ \text{subject to}\quad & \mu^2\boldsymbol{\alpha}^{\mathsf T}{\rm diag}(\boldsymbol{\pi})\boldsymbol{\alpha} + \sigma^2 + \|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2 + \gamma_0^2 \le \gamma_0^2\xi^2(\psi_d - \psi_p)\,.\end{aligned} \tag{B.10}$$

Optimal choice of ξ results in the constraint to become equality. Solving for ξ, the optimization reduces to

$$\text{minimize}\quad \gamma_0^2 + \|\boldsymbol{\alpha}\|_{\ell_2}^2 - 2\left(\langle \boldsymbol{U}_{\rm s}^{\mathsf T}\boldsymbol{\theta}_{0,\rm s}, \boldsymbol{\alpha}\rangle + \sqrt{\gamma_0^2 - \frac{\mu^2\boldsymbol{\alpha}^{\mathsf T}{\rm diag}(\boldsymbol{\pi})\boldsymbol{\alpha} + \sigma^2 + \|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2 + \gamma_0^2}{\psi_d - \psi_p}}\,\|\boldsymbol{\theta}_{0,\rm ns}\|_{\ell_2}\right)$$

Setting derivative with respect to γ_0 to zero, we obtain

$$\sqrt{\gamma_0^2 - \frac{\mu^2\boldsymbol{\alpha}^{\mathsf T}{\rm diag}(\boldsymbol{\pi})\boldsymbol{\alpha} + \sigma^2 + \|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2 + \gamma_0^2}{\psi_d - \psi_p}} = \left(1 - \frac{1}{\psi_d - \psi_p}\right)\|\boldsymbol{\theta}_{0,\rm ns}\|_{\ell_2}\,. \tag{B.11}$$

Setting derivative with respect to $\boldsymbol{\alpha}$ to zero and using the previous stationary equation, we get

$$\boldsymbol{\alpha} = \left(\boldsymbol{I} + \frac{\mu^2{\rm diag}(\boldsymbol{\pi})}{\psi_d - \psi_p - 1}\right)^{-1}\boldsymbol{U}_{\rm s}^{\mathsf T}\boldsymbol{\theta}_{0,\rm s}\,. \tag{B.12}$$

We next square both sides of (B.12) and rearrange the terms to get

$$\begin{aligned}\gamma_0^2 &= \frac{1}{\psi_d - \psi_p - 1}\left(\sigma^2 + \|\boldsymbol{\theta}_{0,\rm s}\|_{\ell_2}^2 + \mu^2\boldsymbol{\alpha}^{\mathsf T}{\rm diag}(\boldsymbol{\pi})\boldsymbol{\alpha}\right) + \left(1 - \frac{1}{\psi_d - \psi_p}\right)\|\boldsymbol{\theta}_{0,\rm ns}\|_{\ell_2}^2\\ &= \frac{1}{\psi_d - \psi_p - 1}\left(\sigma^2 + r_{\rm s}^2 + \mu^2\boldsymbol{\alpha}^{\mathsf T}{\rm diag}(\boldsymbol{\pi})\boldsymbol{\alpha}\right) + \left(1 - \frac{1}{\psi_d - \psi_p}\right)r_{\rm ns}^2\,,\end{aligned}$$

which are the same expressions for $\boldsymbol{\alpha}$ and γ_0 given in the theorem statement.

The final step is to write the risk of estimator in terms of $\boldsymbol{\alpha}$, γ_0. Invoke equation (B.6), and recall that in the current case, $\boldsymbol{\Sigma}_{\rm ns} = \boldsymbol{0}$, $\boldsymbol{\Sigma}_{\rm s} = \mu\boldsymbol{I}$. Also, as we showed in our derivation, $\gamma = \|\boldsymbol{\gamma}\|_{\ell_2} = 0$, $\alpha_0 = \left\|(\boldsymbol{U}_{\rm s}^{\perp})^{\mathsf T}\boldsymbol{\theta}_{0,\rm s}\right\|_{\ell_2}$, by which we arrive at

$$\begin{aligned}{\rm Risk}(\widehat{\boldsymbol{\theta}}_L) &= \mu^2\boldsymbol{\alpha}^{\mathsf T}{\rm diag}(\boldsymbol{\pi})\boldsymbol{\alpha} + \sigma^2 + \left(\left\|(\boldsymbol{U}_{\rm s}^{\perp})^{\mathsf T}\boldsymbol{\theta}_{0,\rm s}\right\|_{\ell_2}^2 + \gamma_0^2 + \|\boldsymbol{\alpha}\|_{\ell_2}^2\right)\\ &= \sigma^2 + (1-\rho)r_{\rm s}^2 + \gamma_0^2 + \boldsymbol{\alpha}^{\mathsf T}\left(\boldsymbol{I} + \mu^2{\rm diag}(\boldsymbol{\pi})\right)\boldsymbol{\alpha}\,.\end{aligned} \tag{B.13}$$

This concludes the proof.

B.3 Proof of Theorem 3.3

We follow the proof strategy used for Theorem 3.1-3.2. Here, we would like to characterize the risk of min-norm estimator $\widehat{\boldsymbol{\theta}}$. The features matrix has a clustering structure, but the learner is not using that (no look-alike clustering) and is just compute the min-norm estimator for fitting the responses to individual features. Therefore, one can think of this setting as a special case of our previous analysis when there is no sensitive features (so $\psi_p = 0$).

(a) By setting $\psi_p = 0$ and $r_{\rm s} = 0$ in the result of Theorem 3.1, we get that when $\psi_d \le 1$,

$$\mathrm{Risk}(\widehat{\boldsymbol{\theta}}) = \frac{\sigma^2}{1-\psi_d}\,.$$

(b) In this case, we specialize the proof of Theorem 3.2 to the case that $\psi_p = 0$. Continuing from (B.8), and removing the terms corresponding to sensitive features, we arrive at

$$\begin{aligned}&\text{minimize} \quad \gamma_0^2 + \|\boldsymbol{\gamma}\|_{\ell_2}^2 - 2\left(\langle \boldsymbol{U}_{\rm ns}^\top \boldsymbol{\theta}_{0,{\rm ns}}, \boldsymbol{\gamma}\rangle + \gamma_0\sqrt{1-\xi^2}\left\|(\boldsymbol{U}_{\rm ns}^\perp)^\top \boldsymbol{\theta}_{0,{\rm ns}}\right\|_{\ell_2}\right)\\ &\text{subject to}\\ &\qquad (\boldsymbol{V}_{\rm ns}\boldsymbol{\Sigma}_{\rm ns}\boldsymbol{\gamma})^\top \mathrm{diag}(\boldsymbol{\pi})(\boldsymbol{V}_{\rm ns}\boldsymbol{\Sigma}_{\rm ns}\boldsymbol{\gamma}) + \sigma^2 + \gamma_0^2 + \|\boldsymbol{\gamma}\|_{\ell_2}^2 \le \gamma_0^2\xi^2\psi_d \end{aligned} \tag{B.14}$$

We drop the index ‘ns’ as it is not relevant in this case. Also by Assumption 2, $\boldsymbol{\Sigma}_{\rm ns} = \mu \boldsymbol{I}_d$, $\boldsymbol{V}_{\rm ns} = \boldsymbol{I}_d$. Therefore, the above optimization can be written as

$$\begin{aligned}&\text{minimize} \quad \gamma_0^2 + \|\boldsymbol{\gamma}\|_{\ell_2}^2 - 2\left(\langle \boldsymbol{U}^\top \boldsymbol{\theta}_0, \boldsymbol{\gamma}\rangle + \gamma_0\sqrt{1-\xi^2}\left\|(\boldsymbol{U}^\perp)^\top \boldsymbol{\theta}_0\right\|_{\ell_2}\right)\\ &\text{subject to} \quad \boldsymbol{\gamma}^\top(\boldsymbol{I} + \mu^2\mathrm{diag}(\boldsymbol{\pi}))\boldsymbol{\gamma} + \sigma^2 + \gamma_0^2 \le \gamma_0^2\xi^2\psi_d\,. \end{aligned} \tag{B.15}$$

Optimal ξ makes the constraint equality. Solving for ξ, the above optimization can be written as so we have

$$\text{minimize} \quad \gamma_0^2 + \|\boldsymbol{\gamma}\|_{\ell_2}^2 - 2\left(\langle \boldsymbol{U}^\top \boldsymbol{\theta}_0, \boldsymbol{\gamma}\rangle + \sqrt{\gamma_0^2 - \frac{\boldsymbol{\gamma}^\top(\boldsymbol{I} + \mu^2\mathrm{diag}(\boldsymbol{\pi}))\boldsymbol{\gamma} + \sigma^2 + \gamma_0^2}{\psi_d}}\left\|(\boldsymbol{U}^\perp)^\top \boldsymbol{\theta}_0\right\|_{\ell_2}\right).$$

Setting the derivative with respect to γ_0 to zero, we get

$$\sqrt{\gamma_0^2 - \frac{\boldsymbol{\gamma}^\top(\boldsymbol{I} + \mu^2\mathrm{diag}(\boldsymbol{\pi}))\boldsymbol{\gamma} + \sigma^2 + \gamma_0^2}{\psi_d}} = \left(1 - \frac{1}{\psi_d}\right)\left\|(\boldsymbol{U}^\perp)^\top \boldsymbol{\theta}_0\right\|_{\ell_2}\,. \tag{B.16}$$

Setting derivative with respect to $\boldsymbol{\gamma}$ to zero and using the above equation, we obtain

$$\boldsymbol{\gamma} = \left(\boldsymbol{I} + \frac{\boldsymbol{I} + \mu^2\mathrm{diag}(\boldsymbol{\pi})}{\psi_d - 1}\right)^{-1}\boldsymbol{U}^\top\boldsymbol{\theta}_0\,. \tag{B.17}$$

We next square both sides of equation (B.16), and rearrange the terms to get:

$$\gamma_0^2 = \frac{1}{\psi_d - 1}\left(\sigma^2 + \boldsymbol{\gamma}^\top(\boldsymbol{I} + \mu^2\mathrm{diag}(\boldsymbol{\pi}))\boldsymbol{\gamma}\right) + \left(1 - \frac{1}{\psi_d}\right)\left\|(\boldsymbol{U}^\perp)^\top \boldsymbol{\theta}_0\right\|_{\ell_2}^2\,.$$

Under the simplifying Assumption 2, there is no cluster structure on the non-sensitive features and so $\boldsymbol{U}_{\rm ns} = 0$. Therefore,

$$\begin{aligned}\left\|\boldsymbol{U}^\top\boldsymbol{\theta}_0\right\|_{\ell_2} &= \left\|\boldsymbol{U}_{\rm s}^\top\boldsymbol{\theta}_{0,{\rm s}}\right\|_{\ell_2} = \sqrt{\rho}r_{\rm s}\,,\\ \left\|(\boldsymbol{U}^\perp)^\top\boldsymbol{\theta}_0\right\|_{\ell_2}^2 &= \|\boldsymbol{\theta}_0\|_{\ell_2}^2 - \left\|\boldsymbol{U}^\top\boldsymbol{\theta}_0\right\|_{\ell_2}^2 = (1-\rho)r_{\rm s}^2 + r_{\rm ns}^2\,.\end{aligned}$$

We next proceed to compute the risk of estimator in terms of $\boldsymbol{\gamma}$, γ_0. We use equation (B.6), which for the min-norm estimator with no look-alike clustering, reduces to

$$\mathrm{Risk}(\widehat{\boldsymbol{\theta}}) = \sigma^2 + \gamma_0^2 + \boldsymbol{\gamma}^\top(\boldsymbol{I} + \mu^2\mathrm{diag}(\boldsymbol{\pi}))\boldsymbol{\gamma}\,. \tag{B.18}$$

This concludes the proof. Note that in the theorem statement we made the change of variables $\gamma_0 \to \tilde{\gamma}_0$ and $\boldsymbol{\gamma} \to \tilde{\boldsymbol{\alpha}}$, for an easier comparison with the risk of look-alike estimator.)

B.4 Proof of Proposition 3.4

Consider singular value decompositions $\boldsymbol{X}_L = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ and $\tilde{\boldsymbol{X}}_L = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^T$. We then can write the estimators $\widetilde{\boldsymbol{\theta}}_L$ and $\widehat{\boldsymbol{\theta}}_L$ as follows:

$$\widehat{\boldsymbol{\theta}}_L = \boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^{\mathsf{T}}\boldsymbol{y}, \quad \widetilde{\boldsymbol{\theta}}_L = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{V}}^{\mathsf{T}}\boldsymbol{y}\,.$$

We first bound $\|\widehat{\boldsymbol{\theta}}_L - \widetilde{\boldsymbol{\theta}}_L\|$. We write

$$\|\widehat{\boldsymbol{\theta}}_L - \widetilde{\boldsymbol{\theta}}_L\| \le \|\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^{\mathsf{T}} - \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{V}}^{\mathsf{T}}\|\|\boldsymbol{y}\|\,. \tag{B.19}$$

We have

$$\|\boldsymbol{y}\| = \|\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}\| = \|\boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{M}^{\mathsf{T}}\boldsymbol{\theta}_0 + \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}\|\,.$$

Note that $\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon} \overset{(d)}{=} \sqrt{\|\boldsymbol{\theta}_0\|^2 + \sigma^2}\boldsymbol{g}$ where $\boldsymbol{g} \sim \mathsf{N}(0, I_n)$. In addition,

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{M}^{\mathsf{T}}\boldsymbol{\theta}_0\|^2 &= \frac{1}{n}\boldsymbol{\theta}_0^{\mathsf{T}}\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{M}^{\mathsf{T}}\boldsymbol{\theta}_0\\ &= \boldsymbol{\theta}_0^{\mathsf{T}}\boldsymbol{M}\text{diag}(\tfrac{n_1}{n}, \dots, \tfrac{n_k}{n})\boldsymbol{M}^{\mathsf{T}}\boldsymbol{\theta}_0 \xrightarrow{p} \boldsymbol{\theta}_0^{\mathsf{T}}\boldsymbol{M}\text{diag}(\pi_1, \dots, \pi_k)\boldsymbol{M}^{\mathsf{T}}\boldsymbol{\theta}_0\,.\end{aligned}$$

Therefore by using concentration of Lipschitz functions of Gaussian vectors, we get

$$\frac{1}{\sqrt{n}}\|\boldsymbol{y}\| \xrightarrow{p} \sqrt{\boldsymbol{\theta}_0^{\mathsf{T}}\boldsymbol{M}\text{diag}(\boldsymbol{\pi})\boldsymbol{M}^{\mathsf{T}}\boldsymbol{\theta}_0 + \|\boldsymbol{\theta}_0\|^2 + \sigma^2}\,.$$

This shows that

$$\frac{1}{\sqrt{n}}\|\boldsymbol{y}\| \to C \le \sqrt{(\mu+1)(r_{\mathrm{s}}^2 + r_{\mathrm{ns}}^2) + \sigma^2}. \tag{B.20}$$

We next use the result of [28, Theorem 3.3], by which we obtain

$$\|\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^{\mathsf{T}} - \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{V}}^{\mathsf{T}}\| \le \frac{1+\sqrt{5}}{2}\max\left(\frac{1}{\sigma_{\min}(\boldsymbol{\Sigma})^2}, \frac{1}{\sigma_{\min}(\widetilde{\boldsymbol{\Sigma}})^2}\right)\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}} - \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^{\mathsf{T}}\|\,. \tag{B.21}$$

Note that

$$\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}} - \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^{\mathsf{T}}\| = \|\boldsymbol{X}_L - \tilde{\boldsymbol{X}}_L\| = \|\boldsymbol{M}_{\mathrm{s}}\boldsymbol{\Lambda} - \widetilde{\boldsymbol{M}}_{\mathrm{s}}\widetilde{\boldsymbol{\Lambda}}\| \le \delta\sqrt{n}\,, \tag{B.22}$$

by the assumption of the theorem statement. We next lower bound $\sigma_{\min}(\boldsymbol{\Sigma}) = \sigma_{\min}(\boldsymbol{X}_L)$. Recall that $\boldsymbol{X}_L^{\mathsf{T}} = (\boldsymbol{M}\boldsymbol{\Lambda})^{\mathsf{T}} + [\boldsymbol{0}_{n\times p}, \boldsymbol{Z}_{n\times(d-p)}]$, with $\boldsymbol{Z}$ having i.i.d $\mathsf{N}(0,1)$ entries.

Next suppose that Condition (i) holds true, namely $\delta < \sqrt{1-(\psi_d - \psi_p)} - \sqrt{\psi_d - \psi_p}$, with $\psi_d - \psi_p < 0.5$. Using the result of [31, Theorem 2.1], we have with probability at least $1 - n^{-1}$,

$$\sigma_{\min}(\boldsymbol{X}_L) \ge \sqrt{n}\left(\sqrt{\psi_d - \psi_p - 1} - 1 - \sqrt{\frac{2\log n}{n}}\right)\,.$$

Furthermore,

$$\begin{aligned}\sigma_{\min}(\tilde{\boldsymbol{X}}_L) &\ge \sigma_{\min}(\boldsymbol{X}_L) - \|\boldsymbol{X}_L - \tilde{\boldsymbol{X}}_L\|\\ &\ge \sqrt{n}\left(\sqrt{1-(\psi_d - \psi_p)} - \sqrt{\psi_d - \psi_p} - \sqrt{\frac{2\log n}{n}} - \delta\right)\\ &\ge c'\sqrt{n}\left(\sqrt{1-(\psi_d - \psi_p)} - \sqrt{\psi_d - \psi_p}\right)\,,\end{aligned}$$

using the assumption on the estimation error rate δ. Therefore, using the above bound along with (B.22) in (B.21) we get

$$\|\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^{\mathsf{T}} - \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{V}}^{\mathsf{T}}\| \le \frac{1+\sqrt{5}}{2c'^2}\frac{1}{\sqrt{n}\left(\sqrt{1-(\psi_d - \psi_p)} - \sqrt{\psi_d - \psi_p}\right)^2}\delta\,.$$

Combining the above bound with (B.20), we get

$$\|\widehat{\boldsymbol{\theta}}_L - \widetilde{\boldsymbol{\theta}}_L\| \le \frac{1+\sqrt{5}}{2c'^2} \frac{C}{\left(\sqrt{1-(\psi_d-\psi_p)} - \sqrt{\psi_d-\psi_p}\right)^2}\delta\,. \tag{B.23}$$

We next note that by triangle inequality, the above bound implies that

$$\|\widetilde{\boldsymbol{\theta}}_L - \boldsymbol{\theta}_0\| - \|\widehat{\boldsymbol{\theta}}_L - \boldsymbol{\theta}_0\| \le \|\widehat{\boldsymbol{\theta}}_L - \widetilde{\boldsymbol{\theta}}_L\| = O(\delta)\,.$$

Therefore, by invoking Lemma 2.1, we obtain the desired result on $\mathrm{Risk}(\widetilde{\boldsymbol{\theta}}_L)$.

Next suppose that Condition (ii) holds, namely $\delta < \sqrt{\psi_d - \psi_p - 1} - 1$ with $\psi_d - \psi_p > 2$. Using the result of [31, Theorem 2.1] for $\boldsymbol{X}^\top$, we have with probability at least $1 - n^{-1}$,

$$\sigma_{\min}(\boldsymbol{X}_L) \ge \sqrt{n}\left(\sqrt{\psi_d - \psi_p - 1} - 1 - \sqrt{\frac{2\log n}{n}}\right)\,.$$

By following a similar argument we prove the claim under Condition (ii).

B.5 Proof of Theorem 5.1

We use Theorem 3.3 (b) to characterize $\mathrm{Risk}(\widehat{\boldsymbol{\theta}})$ in the regime of $\psi_d \ge 1$. Specializing to the case of balanced cluster priors, the risk depends on $\tilde{\boldsymbol{\alpha}}$ only through its norm $\tilde{\alpha} := \|\tilde{\boldsymbol{\alpha}}\|_{\ell_2}$, and is given by

$$\begin{aligned}\mathrm{Risk}(\widehat{\boldsymbol{\theta}}) &\xrightarrow{\mathcal{P}} \sigma^2 + \tilde{\gamma}_0^2 + \left(\frac{\mu^2}{k}+1\right)\tilde{\alpha}^2\\ &= \frac{\psi_d}{\psi_d-1}\left(\sigma^2 + \left(\frac{\mu^2}{k}+1\right)\tilde{\alpha}^2\right) + \left(1-\frac{1}{\psi_d}\right)((1-\rho)r_{\mathrm{s}}^2 + r_{\mathrm{ns}}^2),\end{aligned}$$

with

$$\tilde{\alpha} = \left(1 + \frac{\frac{\mu^2}{k}+1}{\psi_d-1}\right)^{-1}\sqrt{\rho}\, r_{\mathrm{s}}\,.$$

In addition, by Theorem 3.1 we have

$$\mathrm{Risk}(\widehat{\boldsymbol{\theta}}_L) \xrightarrow{\mathcal{P}} \frac{\sigma^2 + r_{\mathrm{s}}^2}{1-\psi_d+\psi_p} - \rho r_{\mathrm{s}}^2\,.$$

Note that $\mathrm{Risk}(\widehat{\boldsymbol{\theta}}_L)$ in this regime does not depend on μ^2/k. Also, it is easy to verify that $\mathrm{Risk}(\widehat{\boldsymbol{\theta}})$ is decreasing in μ^2/k. Therefore the gain Δ is decreasing in μ^2/k.

Also observe that $\mathrm{Risk}(\widehat{\boldsymbol{\theta}})$ is increasing in r_{ns}, while $\mathrm{Risk}(\widehat{\boldsymbol{\theta}}_L)$ does not depend on r_{ns}. Therefore, the gain Δ is increasing in r_{ns}.

To understand the dependence of Δ on ρ, we write

$$\begin{aligned}\Delta - 1 &= \frac{\mathrm{Risk}(\widehat{\boldsymbol{\theta}})}{\mathrm{Risk}(\widehat{\boldsymbol{\theta}}_L)} - 1\\ &= \frac{\frac{\psi_d}{\psi_d-1}\left(\sigma^2 + \left(\frac{\mu^2}{k}+1\right)\left(1+\frac{\mu^2/k+1}{\psi_d-1}\right)^{-2}\rho r_{\mathrm{s}}^2\right) + \left(1-\frac{1}{\psi_d}\right)((1-\rho)r_{\mathrm{s}}^2 + r_{\mathrm{ns}}^2)}{\frac{\sigma^2+r_{\mathrm{s}}^2}{1-\psi_d+\psi_p} - \rho r_{\mathrm{s}}^2} - 1\\ &= \frac{\frac{\psi_d}{\psi_d-1}\left(\sigma^2 + \left(\frac{\mu^2}{k}+1\right)\left(1+\frac{\mu^2/k+1}{\psi_d-1}\right)^{-2}\rho r_{\mathrm{s}}^2\right) + \left(1-\frac{1}{\psi_d}\right)(r_{\mathrm{s}}^2 + r_{\mathrm{ns}}^2) - \frac{\sigma^2+r_{\mathrm{s}}^2}{1-\psi_d+\psi_p} + \frac{\rho r_{\mathrm{s}}^2}{\psi_d}}{\frac{\sigma^2+r_{\mathrm{s}}^2}{1-\psi_d+\psi_p} - \rho r_{\mathrm{s}}^2}\end{aligned}$$

As we see the numerator is increasing in ρ and denominator is decreasing in ρ, which implies that the gain Δ is increasing in ρ.

We next show that $\Delta \geq 1$ if condition (5.1) holds. Since Δ is decreasing in μ^2/k and increasing in ρ, it suffices to show the claim assuming $\mu^2/k \to \infty$ and $\rho = 0$. In this case we have $\left(\frac{\mu^2}{k} + 1\right)\tilde{\alpha}^2 \to 0$ and so

$$\begin{aligned}
\Delta &\to \frac{\frac{\sigma^2 \psi_d}{\psi_d - 1} + \left(1 - \frac{1}{\psi_d}\right)(r_{\mathrm{s}}^2 + r_{\mathrm{ns}}^2)}{\frac{\sigma^2 + r_{\mathrm{s}}^2}{1 - \psi_d + \psi_p}} \\
&\geq \frac{\frac{\sigma^2 \psi_d}{\psi_d - 1} + \left(1 - \frac{1}{\psi_d}\right) r_{\mathrm{s}}^2}{\frac{\sigma^2 + r_{\mathrm{s}}^2}{1 - \psi_d + \psi_p}} \\
&= \frac{\frac{\psi_d}{\psi_d - 1} + \left(1 - \frac{1}{\psi_d}\right) \mathrm{SNR}^2}{\frac{1 + \mathrm{SNR}^2}{1 - \psi_d + \psi_p}} \geq 1,
\end{aligned}$$

where the last step follows from condition (5.1).