A. Omitted proof for minimax estimator with covariate shift

A.1. Pinsker's Theorem and covariate shift with linear model

Theorem A.1 (Pinsker's Theorem)**.** *Suppose the obervations follow sequence model $y_i = \theta_i^* + \epsilon_i z_i, \epsilon_i > 0, i \in [d]$, and Θ is an ellipsoid in $\mathbb{R}^d$: $\Theta = \Theta(a, C) = \{\theta : \sum_i a_i^2 \theta_i^2 \leq C^2\}$. Then the minimax linear risk*

$$\begin{aligned} R_L(\Theta) &:= \min_{\hat{\boldsymbol{\theta}} \text{ linear}} \max_{\boldsymbol{\theta}^* \in \Theta} \mathbb{E}\,\|\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}^*\|^2 \\ &= \sum_i \epsilon_i^2 (1 - a_i/\mu)_+, \end{aligned}$$

where $\mu = \mu(C)$ is determined by

$$\sum_{i=1}^d \epsilon_i^2 a_i (\mu - a_i)_+ = C^2.$$

The linear minimax estimator is given by

$$\hat{\theta}_i^*(y) = c_i^* y_i = (1 - a_i/\mu)_+ y_i, \tag{11}$$

and is Bayes for a Gaussian prior π_C having independent components $\theta_i \sim \mathcal{N}(0, \tau_i^2)$ with $\tau_i^* = \epsilon_i^2(\mu/a_i - 1)_+$.

Our theorem 3.2 is to connect our parameter $\boldsymbol{\beta}^*$ to the $\boldsymbol{\theta}^*$ in pinsker's theorem. First we show that reformulating the problem from a linear map of n dimensional observations $\boldsymbol{y}_S$ to a linear map on the d-dimensional statistic $\hat{\boldsymbol{\beta}}_{SS}$ is sufficient, i.e., Claim 3.1:

Proof of Claim 3.1. This is to show that if $\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S) := A\boldsymbol{y}_S$ is a minimax linear estimator, each row vector of $A \in \mathbb{R}^{d \times n}$ is in the column span of X_S. Write $A = A_1 X_S^\top + A_2 W^\top$ where $W \in \mathbb{R}^{n \times (n-d)}$, columns of which forms the orthonormal complement for the column space of X_S. Equivalently we want to show $A_2 = 0$. We have

$$\begin{aligned} R_L(\mathcal{B}) &\equiv \min_{\hat{\boldsymbol{\beta}} = A\boldsymbol{y}} \max_{\boldsymbol{\beta}^* \in \mathcal{B}} \mathbb{E}\,\|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2 \\ &= \min_{A_1, A_2} \max_{\boldsymbol{\beta}^* \in \mathcal{B}} \mathbb{E}\,\|\Sigma_T^{1/2}((A_1 X_S^\top + A_2 W^\top)\boldsymbol{y}_S - \boldsymbol{\beta}^*)\|^2 \\ &= \min_{A_1, A_2} \max_{\boldsymbol{\beta}^* \in \mathcal{B}} \mathbb{E}\,\|\Sigma_T^{1/2}(A_1 X_S^\top (X_S \boldsymbol{\beta}^* + \boldsymbol{z}) + A_2 W^\top \boldsymbol{z} - \boldsymbol{\beta}^*)\|^2 && \text{(Since } W^\top X_S = 0\text{)} \\ &= \min_{A_1, A_2} \max_{\boldsymbol{\beta}^* \in \mathcal{B}} \Big\{ \|\Sigma_T^{1/2}(A_1 X_S^\top X_S - I)\boldsymbol{\beta}^*\|^2 + \mathbb{E}\,\|\Sigma_T^{1/2} A_1 X_S^\top \boldsymbol{z}\|^2 \\ &\quad + \mathbb{E}\,\|\Sigma_T^{1/2} A_2 W^\top \boldsymbol{z}\|^2 + \mathbb{E}\left\langle \Sigma_T^{1/2} A_1 X_S^\top \boldsymbol{z}, \Sigma_T^{1/2} A_2 W^\top \boldsymbol{z} \right\rangle \Big\} && \text{(Other cross terms vanish since } \mathbb{E}[\boldsymbol{z}] = \mathbf{0}\text{)} \\ &= \min_{A_1, A_2} \max_{\boldsymbol{\beta}^* \in \mathcal{B}} \Big\{ \|\Sigma_T^{1/2}(A_1 X_S^\top X_S - I)\boldsymbol{\beta}^*\|^2 + \mathbb{E}\,\|\Sigma_T^{1/2} A_1 X_S^\top \boldsymbol{z}\|^2 + \mathbb{E}\,\|\Sigma_T^{1/2} A_2 W^\top \boldsymbol{z}\|^2, \Big\} \end{aligned}$$

where the last equation is because

$$\begin{aligned} &\mathbb{E}\left\langle \Sigma_T^{1/2} A_1 X_S^\top \boldsymbol{z}, \Sigma_T^{1/2} A_2 W^\top \boldsymbol{z} \right\rangle = \mathbb{E}\left[\operatorname{Tr}\left[\Sigma_T^{1/2} A_1 X_S^\top \boldsymbol{z}\boldsymbol{z} W A_2^\top \Sigma_T\right]\right] \\ =& \operatorname{Tr}\left[\Sigma_T^{1/2} A_1 X_S^\top \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] W A_2^\top \Sigma_T\right] = \sigma^2 \operatorname{Tr}\left[\Sigma_T^{1/2} A_1 X_S^\top W A_2^\top \Sigma_T\right] = 0. \end{aligned}$$

Clearly, at min-max point, without loss of generality we can take $A_2 = 0$. □

Formally the proof for Theorem 3.2 is presented here:

Proof of Theorem 3.2. To use Pinsker's theorem to prove Theorem 3.2, we simply need to transform the problem match its setting. Let $\boldsymbol{y}_T = \Sigma_T^{1/2} \hat{\Sigma}_S^{-1} X_S^\top \boldsymbol{y}_S / n_S = \boldsymbol{\theta}_T^* + \boldsymbol{z}_T$, where $\boldsymbol{\theta}_T^* = U^\top \Sigma_T^{1/2} \boldsymbol{\beta}^*$ and $\boldsymbol{z}_T \sim \mathcal{N}(0, \sigma^2 \text{diag}([t_i/s_i]_{i=1}^d)/n_S)$. The set for θ_T^* is $\Theta = \{\boldsymbol{\theta} | \|\Sigma_T^{-1/2} U \boldsymbol{\theta}\| \leq r\}$, i.e., $\Theta = \{\theta | \sum_i \boldsymbol{\theta}_i^2 / t_i \leq r^2\}$.

Now with Pinsker's theorem, $\hat{\boldsymbol{\theta}}(\boldsymbol{y}_T)_i = (1-1/(\mu\sqrt{t_i}))_+(y_T)_i$ is the best linear estimator for $\boldsymbol{\theta}_T^*$, where $\mu=\mu(r)$ solves

$$\frac{\sigma^2}{n_S}\sum_{i=1}^{d}\frac{\sqrt{t_i}}{s_i}(\mu-\frac{1}{\sqrt{t_i}})_+ = r^2. \tag{12}$$

Connecting to the original problem, we get that the best estimator for $\Sigma_T^{1/2}\boldsymbol{\beta}^*$ is $U(I-\frac{1}{\mu}\text{diag}([1/\sqrt{t_i}]_{i=1}^d))\boldsymbol{y}_T = U(I-\frac{1}{\mu}\text{diag}([1/\sqrt{t_i}]_{i=1}^d))U^\top\Sigma_T^{1/2}\Sigma_S^{-1}X_S^\top\boldsymbol{y}_S/n_S$.

□

A.2. Omitted proof for noncommute second-moment matrices

Convex program. Our estimator for $\boldsymbol{\beta}^*$ can be achieved through convex programming:

Proof of Proposition 3.3. First note the objective function is quadratic in C and linear in τ, therefore we only need to prove the constraint $S=\{(C,\tau)|(C-I)^\top\Sigma_T(C-I)\preceq\tau I\}$ is a convex set. Notice for $(C_1,\tau_1),(C_2,\tau_2)\in S$, i.e., $(C_i-I)^\top\Sigma_T(C_i-I)\preceq\tau_i I, i\in\{1,2\}$. We simply need to prove for $C_\alpha:=\alpha C_1+(1-\alpha)C_2, \tau_\alpha:=\tau_1\alpha+\tau_2(1-\alpha)$, $(C_\alpha-I)^\top\Sigma_T(C_\alpha-I)\preceq\tau_\alpha I$ for any $\alpha\in[0,1]$. First, notice $(C_1-C_2)^\top\Sigma_T(C_1-C_2)\succeq 0$. Next,

$$\begin{aligned}
&(C_\alpha-I)^\top\Sigma_T(C_\alpha-I)\\
=&\alpha(C_1-I)^\top\Sigma_T(C_1-I)+(1-\alpha)(C_2-I)^\top\Sigma_T(C_2-I)\\
&-\alpha(1-\alpha)(C_1-C_2)^\top\Sigma_T(C_1-C_2)\\
\preceq&\alpha(C_1-I)^\top\Sigma_T(C_1-I)+(1-\alpha)(C_2-I)^\top\Sigma_T(C_2-I)\\
\preceq&\tau_\alpha I.
\end{aligned}$$

□

Benefit of our estimator. Compared to ridge regression, our estimator could possibly achieve much better ($d^{-1/4}$) improvements:

Proof of Remark 3.1. We consider diagonal second-moment matrices $\hat{\Sigma}_S=\text{diag}(\boldsymbol{s}), \Sigma_T=\text{diag}(\boldsymbol{t})$, $\sigma=1$. First we calculate the expected risk obtained with ridge regression: $\hat{\boldsymbol{\beta}}_{\text{RR}}^\lambda=(X_S^\top X_S/n+\lambda I)^{-1}X_S^\top\boldsymbol{y}_S/n_S\sim\mathcal{N}((\hat{\Sigma}_S+\lambda I)^{-1}\Sigma_S\boldsymbol{\beta}^*, 1/n_S(\Sigma_S+\lambda I)^{-2}\Sigma_S)$.

$$\begin{aligned}
L_{\mathcal{B}}(\boldsymbol{\beta}_{\text{RR}}^\lambda)=&\max_{\boldsymbol{\beta}^*\in\mathcal{B}}\mathbb{E}_{\boldsymbol{y}_S}\|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}}_{\text{RR}}^\lambda(\boldsymbol{y}_S)-\boldsymbol{\beta}^*)\|^2\\
=&\max_{\boldsymbol{\beta}^*\in\mathcal{B}}\|\Sigma_T^{1/2}((\hat{\Sigma}_S+\lambda I)^{-1}\hat{\Sigma}_S-I)\boldsymbol{\beta}^*\|^2+\text{Tr}(\frac{1}{n_S}(\hat{\Sigma}_S+\lambda I)^{-2}\hat{\Sigma}_S\Sigma_T)\\
=&\max_i r^2\left(\frac{\sqrt{t_i}s_i}{s_i+\lambda}-\sqrt{t_i}\right)^2+\sum_i\frac{1}{n_S}\frac{t_is_i}{(s_i+\lambda)^2}.
\end{aligned}$$

Compared to our risk:

$$R_L(\mathcal{B})=\sum_i\frac{1}{n_S}\frac{t_i}{s_i}(1-\frac{1}{\sqrt{t_i}\mu})_+,$$

where $\frac{1}{n}\sum_{i=1}^d\frac{\sqrt{t_i}}{s_i}(\mu-\frac{1}{\sqrt{t_i}})_+=r^2$. Let $r^2=\frac{\sqrt{d}}{n_S}$, $s_i=1,\forall i, t_i=1,\forall i\in[d_0], t_i=d^{-1/2}, d_0<i\le d$, where $d_0=\frac{\sqrt{d}}{d^{1/4}-1}\approx d^{1/4}$. Then $\mu=1$, and $R_L(\mathcal{B})=\frac{d^{1/4}}{n}$. In this case,

$$\min_\lambda\max_i r^2\left(\frac{\sqrt{t_i}s_i}{s_i+\lambda}-\sqrt{t_i}\right)^2+\sum_i\frac{1}{n_S}\frac{t_is_i}{(s_i+\lambda)^2}$$

$$
\begin{aligned}
&= \min_\lambda \max_i \frac{\sqrt{d}}{n}\left(\frac{\sqrt{t_i}}{1+\lambda} - \sqrt{t_i}\right)^2 + \sum_i \frac{1}{n_S}\frac{t_i}{(1+\lambda)^2} \geq \qquad \min_\lambda \frac{\sqrt{d}}{n}\frac{\lambda^2}{(1+\lambda)^2} + \frac{\sqrt{d}}{n}\frac{1}{(1+\lambda)^2} \\
&\geq \frac{\sqrt{d}}{2n}.
\end{aligned}
$$

Therefore $\min_\lambda L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}_{\text{RR}}^\lambda) \geq d^{1/4} R_L(\mathcal{B})/2$. □

Near minimax risk. Even among all nonlinear estimators, our estimator is within 1.25 of the minimax risk:

Proof of Theorem 3.4. First we note that for both linear and nonlinear estimators, it is sufficient to use $\hat{\boldsymbol{\beta}}_{\text{SS}}$ instead of the original observations $\boldsymbol{y}_S$. See Lemma A.2 and its corollary. Therefore it suffices to do the following reformulations of the problem.

When Σ_S and Σ_T commute, we formulate the problem as the following Gaussian sequence model. Recall $\hat{\Sigma}_S = U\text{diag}(\boldsymbol{s})U^\top, \Sigma_T = U\text{diag}(\boldsymbol{t})U^\top$. Let $\boldsymbol{\theta}^* = U^\top \Sigma_T^{1/2}\boldsymbol{\beta}^*$, and $\boldsymbol{y} = U^\top \Sigma_T^{1/2}\hat{\boldsymbol{\beta}}_{\text{SS}} \sim \mathcal{N}(\boldsymbol{\theta}^*, \frac{\sigma^2}{n_S}\text{diag}(\boldsymbol{t}/\boldsymbol{s}))$. Our objective of minimizing $\|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S) - \hat{\boldsymbol{\beta}}^*)\|$ from linear estimator is equivalent to minimizing $\|U(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \hat{\boldsymbol{\theta}}^*)\| = \|\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \hat{\boldsymbol{\theta}}^*\|$ from linear estimator.

The set for the parameter that satisfies $\boldsymbol{\theta}^* = U^\top\Sigma_T^{1/2}\boldsymbol{\beta}^*, \|\boldsymbol{\beta}^*\| \leq r$ is equivalent to $\|\Sigma_T^{-1/2}U\boldsymbol{\theta}^*\| \leq r \Leftrightarrow \|\theta_i^*/\sqrt{t_i}\| \leq r$ is an axis-aligned ellipsoid. Then we could directly derive our result from Corollary 4.26 from (Johnstone, 2011). Note that this result is a special case of Theorem 5.2 and we have provided a detailed proof in Section C. Therefore here we save further descriptions.

For the case when $\Sigma_T = \boldsymbol{a}\boldsymbol{a}^\top$ is rank-1, the objective function becomes:

$$
R_L^*(\mathcal{B}) = \min_{\boldsymbol{\beta}^* \text{ linear}} \max_{\boldsymbol{\beta}\in\mathcal{B}} \mathbb{E}(\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S) - \boldsymbol{\beta}^*))^2.
$$

Then the result could be derived from Corollary 1 of (Donoho, 1994), which reformulate the problem to the hardest one-dimensional problem which becomes tractable.

□

In the proof above, we equate the best nonlinear estimator on $\boldsymbol{y}_S$ as the best nonlinear estimator on $\hat{\boldsymbol{\beta}}_{\text{SS}}$. The reasoning is as follows:

Lemma A.2 (Sufficient statistic is enough to achieve a best estimator)**.** *Consider the statistical problem of estimating $\boldsymbol{\beta}^* \in \mathcal{B}$ from observations $\boldsymbol{y} \in \mathcal{Y}$. $\mathcal{B}$ ℓ^2-compact. If $S(\boldsymbol{y})$ is a sufficient statistic of $\boldsymbol{\beta}^*$, then the best estimator that achieves $\min_{\hat{\boldsymbol{\beta}}} \max_{\mathcal{B}} \ell(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$ is of the form $\hat{\boldsymbol{\beta}} = f(S(\boldsymbol{y}))$ with some function f, for any loss $\ell : \mathcal{Y} \to [0, \infty)$.*

This Lemma is restated from Proposition 3.13 from (Johnstone, 2011).

Corollary A.3 (Corollary of Lemma A.2)**.** *Under the same setting of Lemma A.2, $R_N(\mathcal{B})$ is achieved with the form $\hat{\boldsymbol{\beta}} = f(S(\boldsymbol{y}))$.*

A.3. Omitted proof for utilizing source and target data jointly

Sufficient statistic.

Proof of Claim 3.9. Denote by $\bar{\boldsymbol{\beta}}_S := \hat{\Sigma}_S^{-1}X_S^\top \boldsymbol{y}_S/n_S \sim \mathcal{N}(\boldsymbol{\beta}^*, \frac{\sigma^2}{n_S}\hat{\Sigma}_S^{-1})$ and $\bar{\boldsymbol{\beta}}_T := \hat{\Sigma}_T^{-1}X_T^\top \boldsymbol{y}_T/n_T \sim \mathcal{N}(\boldsymbol{\beta}^*, \frac{\sigma^2}{n_T}\hat{\Sigma}_T^{-1})$. We use the Fisher–Neyman factorization theorem to derive the sufficient statistics. The likelihood of observing $\bar{\boldsymbol{\beta}}_S, \bar{\boldsymbol{\beta}}_T$ from parameter $\boldsymbol{\beta}^*$ is:

$$
\begin{aligned}
p(\bar{\boldsymbol{\beta}}_S, \bar{\boldsymbol{\beta}}_T; \boldsymbol{\beta}^*) =& ce^{-\frac{n_S}{\sigma^2}(\bar{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*)\hat{\Sigma}_S(\bar{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*) - \frac{n_T}{\sigma^2}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\bar{\Sigma}_T(\bar{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*)} \\
=& cg(\boldsymbol{\beta}^*, T(\boldsymbol{\beta}^*))h(\bar{\boldsymbol{\beta}}_S, \bar{\boldsymbol{\beta}}_T),
\end{aligned}
$$

where $g(\boldsymbol{\beta}^*, T(\boldsymbol{\beta}^*)) = e^{-(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_{\text{SS}})^\top(\frac{n_S}{\sigma^2}\hat{\Sigma}_S + \frac{n_T}{\sigma^2}\hat{\Sigma}_T)^{-1}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_{\text{SS}})}$, and c is some constant. Therefore it's easy to see that $T(\boldsymbol{\beta}^*) = \hat{\boldsymbol{\beta}}_{\text{SS}}$ is the sufficient statistic for $\boldsymbol{\beta}^*$. □

Proof of Claim 3.10. With similar procedure as before, and notice $\boldsymbol{z}_S$ and $\boldsymbol{z}_T$ are independent, we could first conclude that the optimal estimator is of the form $\hat{\boldsymbol{\beta}} = A\hat{\Sigma}_S^{-1}X_S^\top \boldsymbol{y}_S/n_S + B\hat{\Sigma}_T^{-1}X_T^\top \boldsymbol{y}_T/n_T \sim \mathcal{N}((A+B)\boldsymbol{\beta}^*, \frac{\sigma^2}{n_S}A\hat{\Sigma}_S^{-1}A^\top + \frac{\sigma^2}{n_T}B\hat{\Sigma}_T^{-1}B^\top)$.

$$\begin{aligned}
R_L(\mathcal{B}) =& \min_{A,B}\max_{\boldsymbol{\beta}^*\in\mathcal{B}} \mathbb{E}_{\boldsymbol{z}}\, \|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|^2\\
=& \min_{A,B}\max_{\boldsymbol{\beta}^*\in\mathcal{B}} \Big\{ \|\Sigma^{1/2}(A+B-I)\boldsymbol{\beta}^*\|^2\\
&+\sigma^2\mathrm{Tr}((\frac{1}{n_S}A\hat{\Sigma}_S^{-1}A^\top + \frac{1}{n_T}B\hat{\Sigma}_T^{-1}B^\top)\Sigma_T)\Big\}\\
=& \min_{A,B}\Big\{ \|\Sigma^{1/2}(A+B-I)\|_{op}^2 r^2 + \sigma^2\mathrm{Tr}((\frac{1}{n_S}A\hat{\Sigma}_S^{-1}A^\top + \frac{1}{n_T}B\hat{\Sigma}_T^{-1}B^\top)\Sigma_T)\Big\}
\end{aligned}$$

Take gradient w.r.t A and B respectively we have:

$$\begin{aligned}
&\nabla_A(\|\Sigma^{1/2}(A+B-I)\|_{op}^2 r^2) + \frac{\sigma^2}{n_S}\Sigma_T A\hat{\Sigma}_S^{-1} = 0\\
=&\nabla_B(\|\Sigma^{1/2}(A+B-I)\|_{op}^2 r^2) + \frac{\sigma^2}{n_T}\Sigma_T B\hat{\Sigma}_T^{-1} = 0
\end{aligned}$$

Notice the first terms are equivalent. Therefore $\frac{1}{n_S}A\hat{\Sigma}_S^{-1} = \frac{1}{n_T}B\hat{\Sigma}_T^{-1}$ thus the optimal $\hat{\boldsymbol{\beta}}$ is of the form $C(X_S^\top \boldsymbol{y}_S + X_T^\top \boldsymbol{y}_T)$ for some matrix C, thus finishing the proof. □

B. Omitted proof with approximation error

Unbiased estimator for $\hat{\boldsymbol{\beta}}_T^*$.

Proof of Claim 4.1.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta}_T^* =& (X_S^\top \mathrm{diag}(\boldsymbol{w})X_S)^{-1}(X_S^\top \mathrm{diag}(\boldsymbol{w})\boldsymbol{y}) - \boldsymbol{\beta}_T^*\\
=&(X_S^\top \mathrm{diag}(\boldsymbol{w})X_S)^{-1}(X_S^\top \mathrm{diag}(\boldsymbol{w})(X_S\boldsymbol{\beta}_T^* + \boldsymbol{a}_T + \boldsymbol{z})) - \boldsymbol{\beta}_T^*\\
=&(X_S^\top \mathrm{diag}(\boldsymbol{w})X_S)^{-1}(X_S^\top \mathrm{diag}(\boldsymbol{w})(\boldsymbol{a}_T + \boldsymbol{z}))
\end{aligned}$$

Notice $\mathbb{E}_{\boldsymbol{x}\sim p_S}[\boldsymbol{x}a_T(\boldsymbol{x})\frac{p_T(\boldsymbol{x})}{p_S(\boldsymbol{x})}] = \mathbb{E}_{\boldsymbol{x}\sim p_T}[\boldsymbol{x}a_T(\boldsymbol{x})] = 0$. This is due to the KKT condition for the minimizer of $l(\boldsymbol{\beta}) := \mathbb{E}_{\boldsymbol{x}\sim p_T}\|f^*(\boldsymbol{x}) - \boldsymbol{\beta}^\top\boldsymbol{x}\|^2$ at $\boldsymbol{\beta}_T^*$: $\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}^*) = 0 \to \mathbb{E}_{\boldsymbol{x}\sim p_T}[\boldsymbol{x}(f^* - \boldsymbol{x}^\top\boldsymbol{\beta}_T^*)] = 0$, i.e., $\mathbb{E}_{\boldsymbol{x}\sim p_T}[\boldsymbol{x}a_T(\boldsymbol{x})] = 0$. Next we have: $\mathbb{E}_{\boldsymbol{x}_i\sim p_S}[X_S^\top \mathrm{diag}(\boldsymbol{w})X_S] = \mathbb{E}_{\boldsymbol{x}_i\sim p_S}\sum_{i=1}^n \frac{p_T(\boldsymbol{x}_i)}{p_S(\boldsymbol{x}_i)}\boldsymbol{x}_i\boldsymbol{x}_i^\top = \mathbb{E}_{\boldsymbol{x}_j\sim p_T}\sum_{j=1}^n[\boldsymbol{x}_j\boldsymbol{x}_j^\top] = n_S\Sigma_T$. Therefore

$$\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta}_T^* \to \mathcal{N}(0, \frac{1}{n_S}\Sigma_T^{-1}\,\mathbb{E}_{\boldsymbol{x}\sim p_T}[p_T(\boldsymbol{x})/p_S(\boldsymbol{x})(a_T(\boldsymbol{x})^2+\sigma^2)\boldsymbol{x}\boldsymbol{x}^\top]\Sigma_T^{-1}).$$

□

Proof of Claim 4.2. Recall $X_S = [\boldsymbol{x}_1^\top|\boldsymbol{x}_2^\top|\cdots|\boldsymbol{x}_n^\top]^\top \in \mathbb{R}^{n\times d}$, with $\boldsymbol{x}_i, \forall i\in[n]$ drawn from p_S, and $\boldsymbol{a}_T = [a_T(\boldsymbol{x}_1), a_T(\boldsymbol{x}_2), \cdots a_T(\boldsymbol{x}_n)]^\top \in \mathbb{R}^n$, $\boldsymbol{y} = [y(\boldsymbol{x}_1), y(\boldsymbol{x}_2), \cdots, y(\boldsymbol{x}_n)]^\top \in \mathbb{R}^n$, noise $\boldsymbol{z} = \boldsymbol{y} - f^*(X)$. $\boldsymbol{w} = [p_T(\boldsymbol{x}_i)/p_S(\boldsymbol{x}_i)]^\top$.

To prove the, we only need to show the minimax linear estimator $A\boldsymbol{y}$ is achieved of the form $A_1X^\top\mathrm{diag}(\boldsymbol{w})$, i.e., the row span of A is in the row span of $X^\top\mathrm{diag}(\boldsymbol{w})$.

$$\begin{aligned}
R_L(\mathcal{B}) \equiv& \min_A \max_{\boldsymbol{\beta}_T^*\in\mathcal{B}, a_T\in\mathcal{F}} \mathbb{E}_{\boldsymbol{x}_i\sim p_s,\boldsymbol{z}}[\|\Sigma_T^{1/2}(A\boldsymbol{y} - \boldsymbol{\beta}_T^*)\|^2]\\
=& \min_A \max_{\boldsymbol{\beta}_T^*\in\mathcal{B}, a_T\in\mathcal{F}} \mathbb{E}\,\|\Sigma_T^{1/2}((AX-I)\boldsymbol{\beta}_T^* + A\boldsymbol{a}_T + Az)\|^2\\
=& \min_A \max_{\boldsymbol{\beta}_T^*\in\mathcal{B}, a_T\in\mathcal{F}} \Big\{\|\Sigma_T^{1/2}((\mathbb{E}[AX]-I)\boldsymbol{\beta}_T^* + \mathbb{E}[A\boldsymbol{a}_T])\|_2^2
\end{aligned}$$

$$+\mathbb{E}\,\|\Sigma_T^{1/2}(AX-\mathbb{E}[AX])\boldsymbol{\beta}_T^*\|^2+\mathbb{E}\,\|\Sigma_T^{1/2}(A\boldsymbol{a}_T-\mathbb{E}[A\boldsymbol{a}_T])\|^2+\mathbb{E}\,\|\Sigma_T^{1/2}A\boldsymbol{z}\|^2\Big\}$$

Write $A=A_1X^\top\mathrm{diag}(\boldsymbol{w})+A_2W^\top$, where $X\in\mathbb{R}^{n\times d}$ and $W\in\mathbb{R}^{n\times(n-d)}$ forms the orthogonal complement for the column span of $\mathrm{diag}(\boldsymbol{w})X$. Therefore $X^\top\mathrm{diag}(\boldsymbol{w})W=0$, and $W^\top W=I_{n-d}$. Also, notice $\mathbb{E}_{\boldsymbol{x}_i\sim p_S}[X^\top\mathrm{diag}(\boldsymbol{w})\boldsymbol{a}_T]=n\,\mathbb{E}_{\boldsymbol{x}\sim p_T}[\boldsymbol{x}a_T(\boldsymbol{x})]=0$. Therefore plugging it in $R_L(\mathcal{B})$, we have:

$$\begin{aligned}R_L(\mathcal{B})=&\min_A\max_{\boldsymbol{\beta}_T^*\in\mathcal{B},f^*\in\mathcal{F}}\Big\{\|\Sigma_T^{1/2}((A_1\,\mathbb{E}_{p_S}[X^\top\mathrm{diag}(\boldsymbol{w})X]-I)\boldsymbol{\beta}_T^*+A_2\,\mathbb{E}[W^\top\boldsymbol{a}_T])\|_2^2\\&+\mathbb{E}\,\|\Sigma_T^{1/2}A_1(X^\top\mathrm{diag}(\boldsymbol{w})X-\mathbb{E}[X^\top\mathrm{diag}(\boldsymbol{w})X])\boldsymbol{\beta}_T^*\|^2\\&+\mathbb{E}\,\|\Sigma_T^{1/2}A_2(W^\top\boldsymbol{a}_T-\mathbb{E}[W^\top\boldsymbol{a}_T])\|^2\\&+\sigma^2\,\mathbb{E}\,\|\Sigma_T^{1/2}A_1X^\top\mathrm{diag}(\boldsymbol{w})\|^2+\sigma^2\,\mathbb{E}\,\|\Sigma_T^{1/2}A_2\|^2\Big\}\\=&\min_{A_1,A_2}\max_{\boldsymbol{\beta}_T^*\in\mathcal{B},f^*\in\mathcal{F}}\Big\{\|\Sigma_T^{1/2}((A_1n_S\Sigma_T-I)\boldsymbol{\beta}_T^*+A_2\,\mathbb{E}[W^\top\boldsymbol{a}_T])\|_2^2\\&+\mathbb{E}\,\|\Sigma_T^{1/2}A_1(X^\top\mathrm{diag}(\boldsymbol{w})X-\Sigma_T)\boldsymbol{\beta}_T^*\|^2+\mathbb{E}\,\|\Sigma_T^{1/2}A_2(W^\top\boldsymbol{a}_T-\mathbb{E}[W^\top\boldsymbol{a}_T])\|^2\\&+\sigma^2\,\mathbb{E}\,\|\Sigma_T^{1/2}A_1X^\top\mathrm{diag}(\boldsymbol{w})\|^2+\sigma^2\,\mathbb{E}\,\|\Sigma_T^{1/2}A_2\|^2\Big\}\end{aligned}$$

We could view $\mathbb{E}[W^\top\boldsymbol{a}_T]$ and $W^\top\boldsymbol{a}_T-\mathbb{E}[W^\top\boldsymbol{a}_T]$ separately. First notice at min-max point, if $\mathbb{E}[W^\top\boldsymbol{a}_T]=0$, the minimizer A_2 should be 0 since it only appears in the third and last non-negative terms. If $\mathbb{E}[W^\top\boldsymbol{a}_T]\neq 0$, the cross term of the bias should be non-negative, or otherwise since both f^* and $-f^*$ are in the set, $a_T,\boldsymbol{\beta}_T^*$ could be replaced by $-a_T,-\boldsymbol{\beta}_T^*$ and the loss increases. Clearly in this case A_2 should also be 0 at min-max point. □

On estimating p_T/p_S.

Proof of Proposition 4.3.

$$\begin{aligned}\mathbb{E}_{x,y\sim q}(y-f(x))^2=&\mathbb{E}_{y\sim Ber(1/2)}[\mathbb{E}[x\sim q_{X|Y}](y-f(x))^2|y]\\=&1/2\,\mathbb{E}_{x\sim p_T}(1-f(x))^2+1/2\,\mathbb{E}_{x\sim p_S}(0-f(x))^2\\=&\int_x p_T(\boldsymbol{x})(1-f(\boldsymbol{x}))^2+p_S(\boldsymbol{x})f(\boldsymbol{x})^2d\boldsymbol{x}.\end{aligned}$$

For any $\boldsymbol{x}$, the optimal value for $a:=f(\boldsymbol{x})$ is obtained by taking the derivative of $p_T(\boldsymbol{x})(1-a)^2+p_S(\boldsymbol{x})a^2$, i.e., $a=\frac{p_T(\boldsymbol{x})}{p_S(\boldsymbol{x})+p_T(\boldsymbol{x})}$. Therefore, the optimal function $f(\boldsymbol{x})\equiv\frac{p_T(\boldsymbol{x})}{p_S(\boldsymbol{x})+p_T(\boldsymbol{x})}$ for all $\boldsymbol{x}$. □

C. Omitted Proof with Model Shift

Definition C.1 (Orthosymmetry). *A set Θ is said to be solid and orthosymmetric if $\boldsymbol{\theta}\in\Theta$ and $|\zeta_i|\leq|\theta_i|$ for all i implies that $\boldsymbol{\zeta}\in\Theta$. If a solid, orthosymmetric Θ contains a point $\boldsymbol{\tau}$, then it contains the entire hyperrectangle that $\boldsymbol{\tau}$ defines:* $\Theta(\boldsymbol{\tau})\equiv\{\boldsymbol{\theta}||\theta_i|\leq\tau_i,\forall i\}\subset\Theta$.

Proof of Claim 5.1. First notice for any estimator $\hat{\boldsymbol{\beta}}$, it all satisfies

$$L_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}})\leq r_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}})\leq 2L_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}}).\tag{13}$$

The first inequality is straightforward with the same reasoning of AM-GM as the derivation of (8). As for the second inequality, we take a closer look at (8). Notice that when $\max_{\boldsymbol{\beta}_T^*\in\mathcal{B},\boldsymbol{\delta}\in\Delta}$ is achieved, the cross term has to be non-negative, or otherwise one could flip the sign of $\boldsymbol{\beta}_T^*$ to make the value larger. Therefore at maximum $\|\Sigma_T^{1/2}((A_1+A_2-I)\boldsymbol{\beta}_T^*\|^2+\|\Sigma_T^{1/2}A_1\boldsymbol{\delta}\|^2\leq\|\Sigma_T^{1/2}((A_1+A_2-I)\boldsymbol{\beta}_T^*+\Sigma_T^{1/2}A_1\boldsymbol{\delta}\|^2$, and notice the remaining parts are all non-negative. Therefore $r_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}})\leq 2L_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}})$.

Now let $\hat{\boldsymbol{\beta}}^*=\arg\min_{\hat{\boldsymbol{\beta}}=A_1\bar{\boldsymbol{y}}_S+A_2\bar{\boldsymbol{y}}_S}L_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}})$. We have:

$$R_L(\mathcal{B},\Delta)=L_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}}^*)\overset{(a)}{\leq}L_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})$$

$$\overset{(13)}{\le} r_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}}_{\mathrm{MM}}) \overset{(b)}{\le} r_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}}^*) \overset{(13)}{\le} 2L_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}}^*) = 2R_L(\mathcal{B},\Delta).$$

The inequality (a) is by definition of $\hat{\boldsymbol{\beta}}^*$ while (b) is from the definition of $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$. □

C.1. Lower Bound with Model Shift

In order to derive the lower bound, we abstract the problem to the following more general one:

Problem 1. *For arbitrary diagonal matrix $D \in \mathbb{R}^{d\times d}$, two ℓ_2-compact, solid, orthosymmetric, and quadratically convex sets $\Theta, \Delta \subset \mathbb{R}^d$, let*

$$\mathcal{P}_{\Theta,\Delta,D} = \left\{ \mathcal{N}\left(\left[\begin{array}{c} D\boldsymbol{\theta} + \boldsymbol{\delta} \\ \boldsymbol{\theta} \end{array} \right], \left[\begin{array}{c|c} I & 0 \\ \hline 0 & I \end{array} \right] \right) \middle| \boldsymbol{\theta} \in \Theta, \boldsymbol{\delta} \in \Delta \right\}$$

Let $R_L(\Theta,\Delta,D)$ and $R_N(\Theta,\Delta,D)$ be the minimax linear risk and minimax risk respectively for estimating $\boldsymbol{\theta}$ within the distribution class $\mathcal{P}_{\Theta,\Delta,D}$:

$$R_L(\Theta,\Delta,D) = \min_{\hat{\boldsymbol{\theta}}:\mathbb{R}^d\to\Theta \textit{ linear}} \max_{P\in\mathcal{P}_{\Theta,\Delta,D}} r_P(\hat{\boldsymbol{\theta}}),$$
$$R_N(\Theta,\Delta,D) = \min_{\hat{\boldsymbol{\theta}}:\mathbb{R}^d\to\Theta} \max_{P\in\mathcal{P}_{\Theta,\Delta,D}} r_P(\hat{\boldsymbol{\theta}}).$$

Here $r_P(\hat{\boldsymbol{\theta}}) := \mathbb{E}_{\boldsymbol{x}\sim P} \|\hat{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\theta}(P)\|_2^2$. We want to derive a uniform lower bound for R_N with R_L, i.e., $R_N \ge \mu^* R_L$, where μ^* is universal and doesn't depend on the choices of D, Θ or Δ.

Before proving the lower bound, we establish its connection to our considered problem:

Remark C.1. *Suppose $\Sigma_S = U\mathrm{diag}(\boldsymbol{s})U^\top$ and $\Sigma_T = U\mathrm{diag}(\boldsymbol{t})U^\top$ share the same eigenspace. Recall our samples $\boldsymbol{a} \sim \mathcal{N}(\Sigma_S^{1/2}(\boldsymbol{\beta}_T^* + \boldsymbol{\delta}), \sigma^2 I), \boldsymbol{b} \sim \mathcal{N}(\Sigma_T^{1/2}\boldsymbol{\beta}_T^*, \sigma^2 I)$. Our goal to uniformly lower bound $R_N(r,\gamma)$ by $R_L(r,\gamma)$ is essentially Problem 1, where*

$$R_L(r,\gamma) := \min_{\hat{\boldsymbol{\beta}} \textit{ linear}} \max_{\|\boldsymbol{\beta}_T^*\|\le r, \|\boldsymbol{\delta}\|\le\gamma} \mathbb{E}\,\|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}}(\boldsymbol{a},\boldsymbol{b}) - \boldsymbol{\beta}^*)\|^2,$$
$$R_N(r,\gamma) := \min_{\hat{\boldsymbol{\beta}}} \max_{\|\boldsymbol{\beta}_T^*\|\le r, \|\boldsymbol{\delta}\|\le\gamma} \mathbb{E}\,\|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}}(\boldsymbol{a},\boldsymbol{b}) - \boldsymbol{\beta}^*)\|^2.$$

Proof of Remark C.1. Our target considers samples drawn from distributions $\boldsymbol{x} \sim \mathcal{N}(\Sigma_S^{1/2}(\boldsymbol{\beta}_T^* + \boldsymbol{\delta}), \sigma^2 I), \boldsymbol{y} \sim \mathcal{N}(\Sigma_T^{1/2}\boldsymbol{\beta}_T^*, \sigma^2 I)$.

$$\left[\begin{array}{c} \boldsymbol{a} \\ \boldsymbol{b} \end{array} \right] \sim \mathcal{N}\left(\left[\begin{array}{c} U\mathrm{diag}(\boldsymbol{s}^{1/2})U^\top(\boldsymbol{\beta}_T^* + \boldsymbol{\delta}) \\ U\mathrm{diag}(\boldsymbol{t}^{1/2})U^\top\boldsymbol{\beta}_T^* \end{array} \right], \left[\begin{array}{c|c} \sigma^2 I & 0 \\ \hline 0 & \sigma^2 I \end{array} \right] \right), \boldsymbol{\theta} \in \Theta, \boldsymbol{\delta} \in \Delta$$
$$\iff \left[\begin{array}{c} U^\top\boldsymbol{a}/\sigma \\ U^\top\boldsymbol{b}/\sigma \end{array} \right] \sim \mathcal{N}\left(\left[\begin{array}{c} \mathrm{diag}(\boldsymbol{s}^{1/2})U^\top(\boldsymbol{\beta}_T^* + \boldsymbol{\delta}) \\ \mathrm{diag}(\boldsymbol{t}^{1/2})U^\top\boldsymbol{\beta}_T^* \end{array} \right], \left[\begin{array}{c|c} I & 0 \\ \hline 0 & I \end{array} \right] \right), \|\boldsymbol{\beta}_T^*\| \le r, \|\boldsymbol{\delta}\| \in \gamma$$

Let $\bar{\boldsymbol{a}} = U^\top\boldsymbol{a}/\sigma, \bar{\boldsymbol{b}} = U^\top\boldsymbol{b}/\sigma, \Theta = \{\boldsymbol{\theta} | \|\mathrm{diag}(\boldsymbol{t}^{-1/2})\boldsymbol{\theta}\| \le r\}$, $\Delta = \{\|\mathrm{diag}(\boldsymbol{s}^{-1/2})\boldsymbol{\delta}\| \le \gamma\}$. $\bar{\boldsymbol{\theta}} = U^\top\Sigma_T^{1/2}\boldsymbol{\beta}_T^*, \boldsymbol{\delta} = U^\top\Sigma_S^{1/2}\boldsymbol{\delta}$, and $D = \mathrm{diag}(\boldsymbol{s}^{1/2}\boldsymbol{t}^{-1/2})$. We get:

$$\left[\begin{array}{c} U^\top\boldsymbol{a}/\sigma \\ U^\top\boldsymbol{b}/\sigma \end{array} \right] \sim \mathcal{N}\left(\left[\begin{array}{c} \mathrm{diag}(\boldsymbol{s}^{1/2})U^\top(\boldsymbol{\beta}_T^* + \boldsymbol{\delta}) \\ \mathrm{diag}(\boldsymbol{t}^{1/2})U^\top\boldsymbol{\beta}_T^* \end{array} \right], \left[\begin{array}{c|c} I & 0 \\ \hline 0 & I \end{array} \right] \right), \|\boldsymbol{\beta}_T\| \le r, \|\boldsymbol{\delta}\| \in \gamma$$
$$\iff \left[\begin{array}{c} \bar{\boldsymbol{a}} \\ \bar{\boldsymbol{b}} \end{array} \right] \sim P_{\boldsymbol{\theta},\boldsymbol{\delta},D} := \mathcal{N}\left(\left[\begin{array}{c} D\bar{\boldsymbol{\theta}} + \bar{\boldsymbol{\delta}} \\ \bar{\boldsymbol{\theta}} \end{array} \right], \left[\begin{array}{c|c} I & 0 \\ \hline 0 & I \end{array} \right] \right), \bar{\boldsymbol{\theta}} \in \Theta, \bar{\boldsymbol{\delta}} \in \Delta.$$

Let $\mathcal{P}_{\Theta,\Delta,D} := \left\{ P_{\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\delta}},D} \middle| \bar{\boldsymbol{\theta}} \in \Theta, \bar{\boldsymbol{\delta}} \in \Delta \right\}$. Since U is an invertible matrices, observing $U^\top\boldsymbol{a}/\sigma, U^\top\boldsymbol{b}/\sigma$ instead of $\boldsymbol{a}, \boldsymbol{b}$ has no affect on the performance of the best estimator. Also Θ, Δ are axis-aligned ellipsoid and thus satisfy orthosymmetry. Therefore our problem is essentially reduced to Problem 1. □

Lemma C.2. *Let $\Theta(\boldsymbol{\tau}) = \{\boldsymbol{\theta}|\theta_i \leq \tau_i, \forall i, \boldsymbol{\theta} \in \Theta\}$ and similarly for $\Delta(\boldsymbol{\zeta}) = \{\boldsymbol{\delta}|\boldsymbol{\delta}_i \leq \zeta_i, \boldsymbol{\delta} \in \Delta\}$, D is some diagonal matrix.*

$$R_L(\Theta, \Delta, D) = \sup_{\boldsymbol{\tau}\in\Theta,\boldsymbol{\zeta}\in\Delta} R_L(\Theta(\boldsymbol{\tau}), \Delta(\boldsymbol{\zeta}), D), \text{ and}$$

$$R_N(\Theta, \Delta, D) \geq \sup_{\boldsymbol{\tau}\in\Theta,\boldsymbol{\zeta}\in\Delta} R_N(\Theta(\boldsymbol{\tau}), \Delta(\boldsymbol{\zeta}), D).$$

Write samples drawn from some $P_{\boldsymbol{\theta},\boldsymbol{\delta},D} \in \mathcal{P}_{\Theta,\Delta,D}$ as $(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{x} \sim \mathcal{N}(D\boldsymbol{\theta} + \boldsymbol{\delta}, I), \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\theta}, I)$.

Lemma C.3. *The minimax linear estimator $\hat{\boldsymbol{\theta}} : (\boldsymbol{x}, \boldsymbol{y}) \to A\boldsymbol{x} + B\boldsymbol{y}$ has the form $\hat{\boldsymbol{\theta}}_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_i a_i x_i + \sum_i b_i y_i$ for some $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$. Namely,*

$$R_L(\Theta, \Delta, D) = \inf_{\hat{\boldsymbol{\theta}}_{\boldsymbol{a},\boldsymbol{b}}} \max_{P\in\mathcal{P}_{\Theta,\Delta,D}} r_P(\hat{\boldsymbol{\theta}}_{\boldsymbol{a},\boldsymbol{b}}).$$

Proof. According to the proof of Proposition C.4.a, by discarding off-diagonal terms, the maximum risk of any linear estimator $\hat{\boldsymbol{\theta}}_{A,B}$ over any hyperrectangles $\Theta(\boldsymbol{\tau}), \Delta(\boldsymbol{\zeta})$ is reduced.

$$\max_{\boldsymbol{\theta}\in\Theta(\boldsymbol{\tau}),\boldsymbol{\delta}\in\Delta(\boldsymbol{\zeta})} r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{A,B}) \geq \max_{\boldsymbol{\theta}\in\Theta(\boldsymbol{\tau}),\boldsymbol{\delta}\in\Delta(\boldsymbol{\zeta})} r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{\mathrm{diag}(A),\mathrm{diag}(B)}).$$

Further we have:

$$\begin{aligned}\min_{A,B} \max_{\boldsymbol{\theta}\in\Theta,\boldsymbol{\delta}\in\Delta} r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{A,B}) \geq& \min_{A,B} \max_{\boldsymbol{\tau}\in\Theta,\boldsymbol{\zeta}\in\Delta} \max_{\boldsymbol{\theta}\in\Theta(\boldsymbol{\tau}),\boldsymbol{\delta}\in\Delta(\boldsymbol{\zeta})} r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{\mathrm{diag}(A),\mathrm{diag}(B)})\\ =& \min_{\boldsymbol{a},\boldsymbol{b}} \max_{\boldsymbol{\theta}\in\Theta,\boldsymbol{\zeta}\in\Delta} r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{a},\boldsymbol{b}})\\ \geq& \min_{C} \max_{\boldsymbol{\theta}\in\Theta,\boldsymbol{\delta}\in\Delta} r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{A,B}).\end{aligned}$$

Therefore all four terms have to be equal, thus finishing the proof. □

Notice $\Theta(\boldsymbol{\tau})$ and $\Delta(\boldsymbol{\zeta})$ are hyperrectangles in $\mathbb{R}^d$. Therefore we could decompose the problem to some 2-d problems:

Proposition C.4. *Under the same setting as Problem 1,*

$$a).\ \ R_L(\Theta(\boldsymbol{\tau}), \Delta(\boldsymbol{\zeta}), D) = \sum_i R_L(\tau_i, \zeta_i, D_{ii}).$$

If $\hat{\boldsymbol{\theta}}_{A,B}(\boldsymbol{x}, \boldsymbol{y}) = A\boldsymbol{x} + B\boldsymbol{y}$ is minimax linear estimator over $P_{\Theta(\boldsymbol{\tau}),\Delta(\boldsymbol{\zeta}),D}$, then necessarily A, B must be diagonal.

$$b).\ \ R_N(\Theta(\boldsymbol{\tau}), \Delta(\boldsymbol{\zeta}), D) = \sum_i R_N(\tau_i, \zeta_i, D_{ii}).$$

Proof of Proposition C.4.a . First review our notation:

$$\begin{aligned}r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{A,B}) =& \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim P_{\boldsymbol{\theta},\boldsymbol{\delta},D}} \|\hat{\boldsymbol{\theta}}_{A,B}(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\theta}\|^2\\ =& \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(D\boldsymbol{\theta}+\boldsymbol{\delta},I),\boldsymbol{y}\sim\mathcal{N}(\boldsymbol{\theta},I)} \|A\boldsymbol{x} + B\boldsymbol{y} - \boldsymbol{\theta}\|^2\\ =& \|A(D\boldsymbol{\theta} + \boldsymbol{\delta}) + B\boldsymbol{\theta} - \boldsymbol{\theta}\|^2 + \mathrm{Tr}(AA^\top) + \mathrm{Tr}(BB^\top)\\ =& \|(AD + B - I)\boldsymbol{\theta} + A\boldsymbol{\delta}\|^2 + \mathrm{Tr}(AA^\top) + \mathrm{Tr}(BB^\top).\end{aligned}$$

Our objective is

$$R_L(\Theta(\boldsymbol{\tau}), \Delta(\boldsymbol{\zeta}), D) := \min_{A,B} \max_{\boldsymbol{\theta}\in\Theta(\boldsymbol{\tau}),\boldsymbol{\delta}\in\Delta(\boldsymbol{\zeta})} r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{A,B})$$

We will show that restricting A, B to be diagonal will not include the RHS value.

For any $\bar{\boldsymbol{\tau}} \in \Theta(\boldsymbol{\tau}), \bar{\boldsymbol{\zeta}} \in \Delta(\boldsymbol{\zeta})$, let set $V(\bar{\boldsymbol{\tau}}, \bar{\boldsymbol{\zeta}}) = \{(\boldsymbol{\theta}, \boldsymbol{\delta})|(\theta_i, \delta_i) \in \{(\bar{\tau}_i, \bar{\zeta}_i), (-\bar{\tau}_i, -\bar{\zeta}_i)\}\}$ be the subset of vertices of $\Theta(\bar{\boldsymbol{\tau}}) \times \Delta(\bar{\boldsymbol{\zeta}})$. Let $\pi(\bar{\boldsymbol{\tau}}, \bar{\boldsymbol{\zeta}})$ be uniform distribution on this finite set. Due to the symmetry of this distribution, we have

$$\mathbb{E}_{\pi(\bar{\boldsymbol{\tau}},\bar{\boldsymbol{\zeta}})} \theta_i = 0, i \in [d],$$

$$\begin{aligned}
\mathbb{E}_{\pi(\bar{\boldsymbol{\tau}},\bar{\boldsymbol{\zeta}})}\,\delta_i &= 0, i\in[d],\\
\mathbb{E}_{\pi(\bar{\boldsymbol{\tau}},\bar{\boldsymbol{\zeta}})}\,\theta_i\theta_j &= \mathbf{1}_{i=j}\bar{\tau}_i^2, i\in[d],\\
\mathbb{E}_{\pi(\bar{\boldsymbol{\tau}},\bar{\boldsymbol{\zeta}})}\,\delta_i\delta_j &= \mathbf{1}_{i=j}\bar{\zeta}_i^2, i\in[d],\\
\mathbb{E}_{\pi(\bar{\boldsymbol{\tau}},\bar{\boldsymbol{\zeta}})}\,\theta_i\delta_j &= \mathbf{1}_{i=j}\bar{\tau}_i\bar{\zeta}_i, i\in[d].
\end{aligned}$$

We utilize the distribution to find the explicit value of the maximum (in fact the maximum will only be obtained inside the vertices set $V(\bar{\boldsymbol{\tau}},\bar{\boldsymbol{\zeta}})$):

$$\begin{aligned}
&\max_{(\boldsymbol{\theta},\boldsymbol{\delta})\in V(\bar{\boldsymbol{\tau}},\bar{\boldsymbol{\zeta}})} r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{A,B}) \ge \mathbb{E}_{\pi(\bar{\boldsymbol{\tau}},\bar{\boldsymbol{\zeta}})}\, r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{A,B})\\
=&\mathbb{E}_{\pi(\bar{\boldsymbol{\tau}},\bar{\boldsymbol{\zeta}})}\,\|(AD+B-I)\boldsymbol{\theta}+A\boldsymbol{\delta}\|^2+\mathrm{Tr}(AA^\top)+\mathrm{Tr}(BB^\top)\\
=&\mathrm{Tr}((AD+B-I)\,\mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^\top](AD+B-I)^\top)+\mathrm{Tr}(A\,\mathbb{E}[\boldsymbol{\delta}\boldsymbol{\delta}^\top]A^\top)+\\
&2\mathrm{Tr}((AD+B-I)\,\mathbb{E}[\boldsymbol{\theta}\boldsymbol{\delta}^\top]A^\top)+\mathrm{Tr}(AA^\top)+\mathrm{Tr}(BB^\top)\\
=&\mathrm{Tr}((AD+B-I)^\top(AD+B-I)\mathrm{diag}(\bar{\boldsymbol{\tau}}^2))+\mathrm{Tr}(A^\top A\mathrm{diag}(\bar{\boldsymbol{\zeta}}^2))\\
&+\mathrm{Tr}((AD+B-I)^\top A\mathrm{diag}(\bar{\boldsymbol{\tau}}\bar{\boldsymbol{\zeta}}))+\mathrm{Tr}(AA^\top)+\mathrm{Tr}(BB^\top)\\
=&\sum_i\|(AD+B-I)_{:,i}\bar{\tau}_i+A_{:,i}\bar{\zeta}_i\|^2+\mathrm{Tr}(AA^\top)+\mathrm{Tr}(BB^\top)\\
\ge&\sum_i((A_{ii}D_{ii}+B_{ii}-1)\bar{\tau}_i+A_{ii}\bar{\zeta}_i)^2+A_{ii}^2+B_{ii}^2\\
=&\|(\mathrm{diag}(A)D+\mathrm{diag}(B)-I)\boldsymbol{\theta}+\mathrm{diag}(A)\boldsymbol{\delta}\|^2+\mathrm{Tr}(\mathrm{diag}(A)^2)+\mathrm{Tr}(\mathrm{diag}(B)^2), \qquad (\forall(\boldsymbol{\theta},\boldsymbol{\delta})\in V(\bar{\boldsymbol{\tau}},\bar{\boldsymbol{\zeta}}))\\
=&\max_{V(\bar{\boldsymbol{\tau}},\bar{\boldsymbol{\zeta}})}\|(\mathrm{diag}(A)D+\mathrm{diag}(B)-I)\boldsymbol{\theta}+\mathrm{diag}(A)\boldsymbol{\delta}\|^2+\mathrm{Tr}(\mathrm{diag}(A)^2)+\mathrm{Tr}(\mathrm{diag}(B)^2)
\end{aligned}$$

Therefore we have:

$$\begin{aligned}
R_L(\Theta(\boldsymbol{\tau}),\Delta(\boldsymbol{\zeta}),D):=&\min_{A,B}\max_{\boldsymbol{\theta}\in\Theta(\boldsymbol{\tau}),\boldsymbol{\delta}\in\Delta(\boldsymbol{\zeta})} r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{A,B})\\
=&\min_{A,B}\max_{\bar{\boldsymbol{\tau}}\in\Theta(\boldsymbol{\tau}),\bar{\boldsymbol{\zeta}}\in\Delta(\boldsymbol{\zeta})}\max_{\boldsymbol{\theta}\in V(\bar{\boldsymbol{\tau}},\bar{\boldsymbol{\zeta}})} r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{A,B})\\
\ge&\min_{A,B}\max_{\bar{\boldsymbol{\tau}}\in\Theta(\boldsymbol{\tau}),\bar{\boldsymbol{\zeta}}\in\Delta(\boldsymbol{\zeta})}\max_{(\boldsymbol{\theta},\boldsymbol{\delta})\in V(\bar{\boldsymbol{\tau}},\bar{\boldsymbol{\zeta}})} r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{\mathrm{diag}(A),\mathrm{diag}(B)})\\
=&\min_{\boldsymbol{a}\in\mathbb{R}^d,\boldsymbol{b}\in\mathbb{R}^d}\max_{\boldsymbol{\theta}\in\Theta(\boldsymbol{\tau}),\boldsymbol{\delta}\in\Delta(\boldsymbol{\zeta})} r_{P_{\boldsymbol{\theta},\boldsymbol{\delta},D}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{a},\boldsymbol{b}}).
\end{aligned}$$

Next, since the optimal solution on the minimizer is always obtained by diagonal A, B, it becomes straightforward that each axis could be viewed in separation, thus finishing the proof for part a.

The nonlinear part is a straightforward extension of Proposition 4.16 from (Johnstone, 2011).

□

Theorem C.5 (Restated Le Cam Two Point Theorem (Wainwright, 2019))**.** *Let $\mathcal{P}$ be a family of distribution, and $\theta:\mathcal{P}\to\Theta$ is some associated parameter. Let $\rho:\Theta\times\Theta\to\mathbb{R}^+$ be some metric defined on Θ and $\Phi:\mathbb{R}_+\to\mathbb{R}_+$ is a monotone non-decreasing function with $\Phi(0)=0$. For any $\alpha\in(0,1)$,*

$$\begin{aligned}
\inf_{\hat{\theta}}\sup_{P\in\mathcal{P}}[\Phi(\rho(\hat{\theta},\theta(P)))]\ge\max_{P_1,P_2\in\mathcal{P}}\frac{1}{2}\Phi(\frac{1}{2}\rho(\theta(P_1),\theta(P_2)))(1-\alpha),\\
\textit{s.t. } \|P_1^n-P_2^n\|_{TV}\le\alpha.
\end{aligned}$$

Lemma C.6. *Consider a class of distribution $\mathcal{P}_{\tau,\zeta,s}=\{P_{\theta,\delta,s}|P_{\theta,\delta,s}:=\mathcal{N}([s\theta+\delta,\theta]^\top,I_2),|\theta|\le\tau,|\delta|\le\zeta\}$. Define*

$$R_L(\tau,\zeta,s)=\min_{\hat{\theta}\textit{ linear}}\max_{|\theta|\le\tau,|\delta|\le\zeta}\mathbb{E}_{\boldsymbol{x}\sim P_{\theta,\delta,s}}(\hat{\theta}(\boldsymbol{x})-\theta)^2,$$

$$\textit{and } R_N(\tau,\zeta,s) = \min_{\hat\theta} \max_{|\theta|\le\tau,|\delta|\le\zeta} \mathbb{E}_{\boldsymbol{x}\sim P_{\theta,\delta,s}}(\hat\theta(\boldsymbol{x})-\theta)^2$$

We have

$$R_L(\tau,\zeta,s) \le 27/2 R_N(\tau,\zeta,s), \forall \zeta, s>0, \tau>0.$$

Proof of Lemma C.6. We first calculate an upper bound of R_L and connect it to a lower bound of R_N.

$$\begin{aligned} R_L(\tau,\zeta,s) &= \min_{a,b} \max_{|\theta|\le\tau,|\delta|\le\zeta} [(as+b-1)\theta + a\delta]^2 + a^2 + b^2 \\ &= \min_{a,b} (|as+b-1|\tau + |a|\zeta)^2 + a^2 + b^2 \\ &\le \min_{a,b} 2(as+b-1)^2\tau^2 + 2a^2\zeta^2 + a^2 + b^2. \end{aligned}$$

By some detailed calculations, we get the RHS is equal to:

$$\begin{aligned} &\frac{2\tau^2(2\zeta^2+1)}{2\tau^2(s^2+2\zeta^2+1)+2\zeta^2+1} \\ \le &\min\{1, 2\tau^2, \frac{1+4\zeta^2}{s^2+1}\}. \end{aligned}$$

For simplify this form, we could see that

Next, we use Le cam two point theorem to lower bound $R_N(\tau,\zeta,s)$ where the metric ρ is Euclidean distance and Φ is squared function. Therefore

$$\begin{aligned} R_N(\tau,\zeta,s) \ge &\max_{|\theta_i|\le\tau,|\delta_i|\le\zeta, i\in\{1,2\}} \frac{1}{2}(\frac{1}{2}(\theta_1-\theta_2))^2(1-\alpha) \\ &\text{s.t. } \|\mathcal{N}([s\theta_1+\delta_1,\theta_1]^\top, I_2), \mathcal{N}([s\theta_2+\delta_2,\theta_2]^\top, I_2)\|_{TV} \le \alpha. \end{aligned}$$

Since the total variation distance is related to Kullback-Leibler divergence by Pinsker's inequality: $\|\cdot,\cdot\|_{TV} \le \sqrt{\frac{1}{2}D_{KL}(\cdot\|\cdot)}$, it's sufficient to replace the constraint as:

$$D_{KL}\left(\mathcal{N}([s\theta_1+\delta_1,\theta_1]^\top, I_2) \middle\| \mathcal{N}([s\theta_2+\delta_2,\theta_2]^\top, I_2)\right) \le 2\alpha^2.$$

$$\begin{aligned} &\max_{|\theta_i|\le\tau,|\delta_i|\le\zeta, i\in\{1,2\}} \frac{1}{8}(\theta_1-\theta_2)^2(1-\alpha) \\ &\text{s.t. } (s\theta_1+\delta_1-(s\theta_2+\delta_2))^2 + (\theta_1-\theta_2)^2 \le 2\alpha^2 \\ \Leftrightarrow &\max_{|c|\le 2\tau,|d|\le 2\zeta} \frac{c^2}{8}(1-\alpha) \\ &\text{s.t. } (sc+d)^2 + c^2 \le 2\alpha^2. \end{aligned}$$

Recall $R_L \le \min\{1, 2\tau^2, \frac{1+4\zeta}{s^2+1}\}$.

We first note that $c^2 \le 4\tau^2$ and setting $\alpha=0$ we have $R_N \ge \tau^2/2 \ge 1/4R_L$. For In the following we look at other cases when the bound for c^2 is smaller.

When $2\zeta \ge sc$, will set $d=-sc$ and $c^2 = 2\alpha^2$. Let $\alpha = 2/3$ for large τ we get : $c^2(1-\alpha)/8 = 2/27 \ge 2/27R_L$.

When $2\zeta \le sc$ we set $d=-2\zeta$ and require $(sc-2\zeta)^2 + c^2 \le 2\alpha^2$. We have $(sc-2\zeta)^2 + c^2 = s^2c^2 + 4\zeta^2 - 4\zeta sc + c^2 \le s^2c^2 + 4\zeta^2 - 8\zeta^2 + c^2 = (s^2+1)c^2 - 4\zeta^2$. Therefore as we set $c^2 = \frac{2\alpha^2+4\zeta^2}{s^2+1}$, the original inequality is satisfied. Again by setting $\alpha = 2/3$ we have $c^2 \ge 8/9\frac{1+4\zeta^2}{s^2+1} \ge 8/9R_L$. Therefore in this case $R_N \ge \frac{2}{27}R_L$.

□

D. Discussions on Random Design under Covariate Shift.

In the main text, we present the results where we consider X_S as fixed and Σ_T to be known. In this section, we view both source and target input data as random, and generalize the results of Section 3 while training is on finite observations and testing is on the (worst case) population loss, under some light-tail properties of the input data samples.

Proof of Theorem 3.8. The proof relies on the two technical claims D.1, D.2.

Let $\hat{\boldsymbol{\beta}}_R$ be the optimal linear estimator on $L_{\mathcal{B}}$, i.e., $L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}_R) = \min_{\boldsymbol{\beta} \text{ linear in } \boldsymbol{y}_S} L_{\mathcal{B}}(\boldsymbol{\beta}) = R_L(\mathcal{B})$.

$$
\begin{aligned}
& L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) \leq (1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n}}))\hat{L}_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) && \text{(Claim D.2)}\\
\leq & (1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n}}))\hat{L}_{\mathcal{B}}(\hat{\boldsymbol{\beta}}_R) && \text{(from definition of } \hat{\boldsymbol{\beta}})\\
\leq & (1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n}}))^2 L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}_R) && \text{(Claim D.2)}\\
\leq & (1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n}}))L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) = (1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n}}))R_L(\mathcal{B}).\\
& && \text{(from } \tfrac{\rho^4(d+\log(1/\delta))}{n} \ll 1 \text{, and definition of } \hat{\boldsymbol{\beta}}_R)
\end{aligned}
$$

From Theorem 3.4 we know $R_L(\mathcal{B}) \leq 1.25 R_N(\mathcal{B})$ when Σ_T is rank-1 matrix or commute with $\hat{\Sigma}_S$ which further finishes the whole proof. □

Claim D.1 (Restated Claim A.6 from (Du et al., 2020))**.** *Fix a failure probability $\delta \in (0,1)$, and assume $n \gg \rho^4(d+\log(1/\delta))$* [11]*. Then with probability at least $1-\frac{\delta}{10}$ over the inputs $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$, if $\boldsymbol{x}_i \sim p$ and p is a ρ^2-subgaussian distribution, we have*

$$
(1-O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n}}))\Sigma \preceq \frac{1}{n}X^\top X \preceq (1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n}}))\Sigma, \tag{14}
$$

where $\Sigma = \mathbb{E}_{\boldsymbol{x}\sim p}[\boldsymbol{x}\boldsymbol{x}^\top]$.

With the help of Claim D.1 we directly get:

Claim D.2. *Fix a failure probability $\delta \in (0,1)$, and assume $n_U \gg \rho^4(d+\log(1/\delta))$, $X_T = [\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n_U}]^\top \in \mathbb{R}^{n_U\times d}$ satisfies $\boldsymbol{x}_i \sim p_T$ where p_T is ρ^2-subgaussian. We have for any estimator $\boldsymbol{\beta}$:*

$$
(1-O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_U}}))L_{\mathcal{B}}(\boldsymbol{\beta}) \leq \hat{L}_{\mathcal{B}}(\boldsymbol{\beta}) \leq (1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_U}}))L_{\mathcal{B}}(\boldsymbol{\beta}),
$$

with high probability $1-\delta/10$ over the random samples X_T.

Proof of Claim D.2. Recall

$$
\hat{L}_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) = \max_{\boldsymbol{\beta}^*\in\mathcal{B}} \mathbb{E}_{\boldsymbol{y}_S} \frac{1}{n_U}\|X_T(\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S)-\boldsymbol{\beta}^*)\|^2,
$$

$$
L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) = \max_{\boldsymbol{\beta}^*\in\mathcal{B}} \mathbb{E}_{\boldsymbol{y}_S} \|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S)-\boldsymbol{\beta}^*)\|^2.
$$

Therefore for any estimator $\hat{\boldsymbol{\beta}}$, it satisfies

$$
\begin{aligned}
& L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) - \hat{L}_{\mathcal{B}}(\hat{\boldsymbol{\beta}})\\
= & (\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S)-\boldsymbol{\beta}^*)^\top(\Sigma_S - \hat{\Sigma}_S)(\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S)-\boldsymbol{\beta}^*)
\end{aligned}
$$

[11] When this is not satisfied the result is still satisfied by replacing $O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n}})$ with $O(\max\{\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n}}, \frac{\rho^2(d+\log(1/\delta))}{n}\})$. For cleaner presentation, we assume n is large enough and simplify the results.

$$\begin{aligned}&\lesssim O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_U}})(\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S)-\boldsymbol{\beta}^*)^\top\Sigma_S(\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S)-\boldsymbol{\beta}^*)\\&=O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_U}})L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}),\end{aligned}$$

which finishes the proof. □

D.1. Random design on source domain.

In the main text or the subsection above, the worst case excess risk is upper bounded by $1.25R_N$, which is achieved by best estimator that is using the same set of training data $(X_S, \boldsymbol{y}_S)$. Here we would like to take into consideration the randomness of X_S and compare the worst case excess risk using our estimator with a stronger notion of linear estimator.

For this purpose, we consider estimators that are linear functionals of $\boldsymbol{y}_R := \Sigma_S^{1/2}\boldsymbol{\beta}^* + \boldsymbol{z} \in \mathbb{R}^d, \boldsymbol{z} \sim \mathcal{N}(0, \sigma^2/n_S I_d)$ (this σ^2/n_S is the correct scaling since $X_S^\top X_S/n_S$ is comparable to Σ_S). We consider the minimax linear estimator with $\boldsymbol{y}_R$ and with access to Σ_S, and we compare our estimator against this oracle linear estimator. This estimator is not computable in practice since Σ_S must be estimated, but we will show that our estimator is within an absolute multiplicative constant in minimax risk of the oracle linear estimator.

To recap the notations and setup, let

$$\begin{aligned}\hat{L}_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) &:= \max_{\boldsymbol{\beta}^*}\mathbb{E}_{\boldsymbol{y}_S}\frac{1}{n_U}\|X_T(\hat{\beta}(\boldsymbol{y}_S)-\boldsymbol{\beta}^*)\|^2,\\ L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) &:= \max_{\boldsymbol{\beta}^*}\mathbb{E}_{\boldsymbol{y}_S}\mathbb{E}_{\boldsymbol{x}\sim p_T}\|\boldsymbol{x}^\top(\hat{\beta}(\boldsymbol{y}_S)-\boldsymbol{\beta}^*)\|^2,\\ L_{\mathcal{B},R}(\hat{\boldsymbol{\beta}}) &:= \max_{\boldsymbol{\beta}^*}\mathbb{E}_{\boldsymbol{y}_R}\mathbb{E}_{\boldsymbol{x}\sim p_T}\|\boldsymbol{x}^\top(\hat{\beta}(\boldsymbol{y}_R)-\boldsymbol{\beta}^*)\|^2.\end{aligned}$$

Our target is to find the best linear estimator using $\hat{L}_{\mathcal{B}}(\hat{\boldsymbol{\beta}})$ (trained with X_T) and prove its performance on the population (worst-case) excess risk $L_{\mathcal{B}}(\hat{\boldsymbol{\beta}})$ is no much worse compared to the minimax linear risk trained on $\boldsymbol{y}_R$ and Σ_S.

Theorem D.3. *Fix a failure probability $\delta \in (0,1)$. Suppose both target and source distributions p_S and p_T are ρ^2-subgaussian, and the sample sizes in source domain and target domain satisfies $n_S, n_U \gg \rho^4(d+\log\frac{1}{\delta})$. Let $\hat{C}$ be the solution for Eqn.(4), and set $\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S) \leftarrow \hat{C}\hat{\Sigma}_S^{-1}X_S^\top\boldsymbol{y}_S$. Then with probability at least $1-\delta$ over all the unlabeled samples from target domain and all the labeled samples X_S from source domain, our estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{y}_R)$ yields the worst case expected excess risk that satisfies:*

$$L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) \le \left(1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_U}})+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_T}})\right)\min_{\boldsymbol{\beta}\text{ linear in }\boldsymbol{y}_R}L_{R,\mathcal{B}}(\boldsymbol{\beta}).$$

Proof of Theorem D.3. For each matrix $C \in \mathbb{R}^{d\times d}$, we first conduct bias-variance decomposition and rewrite each worst-case risk with linear estimator in terms of a matrix C. When $\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S) = C\hat{\Sigma}_S^{-1}X_S^\top\boldsymbol{y}_S$, we have:

$$\begin{aligned}\hat{L}_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) &= \|\hat{\Sigma}_T^{1/2}(C-I)\|_{op}^2 r^2 + \frac{\sigma^2}{n}\mathrm{Tr}(\hat{\Sigma}_T C\hat{\Sigma}_S^{-1}C^\top) =: \hat{l}(C),\\ L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) &= \|\Sigma_T^{1/2}(C-I)\|_{op}^2 r^2 + \frac{\sigma^2}{n}\mathrm{Tr}(\Sigma_T C\hat{\Sigma}_S^{-1}C^\top) =: l(C),\end{aligned}$$

Similarly, when $\hat{\boldsymbol{\beta}}_R = C\Sigma_S^{-1/2}\boldsymbol{y}_R$, we have:

$$L_{R,\mathcal{B}}(\hat{\boldsymbol{\beta}}) = \|\Sigma_T^{1/2}(C-I)\|_{op}^2 r^2 + \frac{\sigma^2}{n}\mathrm{Tr}(\Sigma_T C\Sigma_S^{-1}C^\top) =: l_R(C).$$

Claim D.4. *Fix a failure probability $\delta \in (0,1)$, and assume $n_U, n_S \gg \rho^4(d+\log(1/\delta))$, $X_S \in \mathbb{R}^{n_S\times d}$, $X_T \in \mathbb{R}^{n_U\times d}$ are respectively from p_S p_T which are both ρ^2-subgaussian. We have for any matrix $C \in \mathbb{R}^{d\times d}$:*

$$(1-O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_U}}))\hat{l}(C) \le l(C) \le (1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_U}}))\hat{l}(C),$$

with high probability $1-\delta/10$ *over the random samples* X_T.

$$(1-O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_S}}))l(C)\le l_R(C)\le(1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_S}}))l(C),$$

with high probability $1-\delta/10$ *over the random samples* X_S.

Proof of Claim D.4. We omit the proof of the first inequality since it's exactly the same as proof of Claim D.2.

For the second line, we have:

$$\begin{aligned}
l_R(C)-l(C)=&\frac{\sigma^2}{n_S}\mathrm{Tr}(\Sigma_T C(\Sigma_S^{-1}-\hat{\Sigma}_S^{-1})C^\top)\\
\le&O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_S}})\frac{\sigma^2}{n_S}\mathrm{Tr}(\Sigma_T C\hat{\Sigma}_S^{-1}C^\top)\\
\le&O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_S}})l(C).
\end{aligned}$$

Therefore we prove the RHS of the second inequality. The LHS follows with the same proof techniques. □

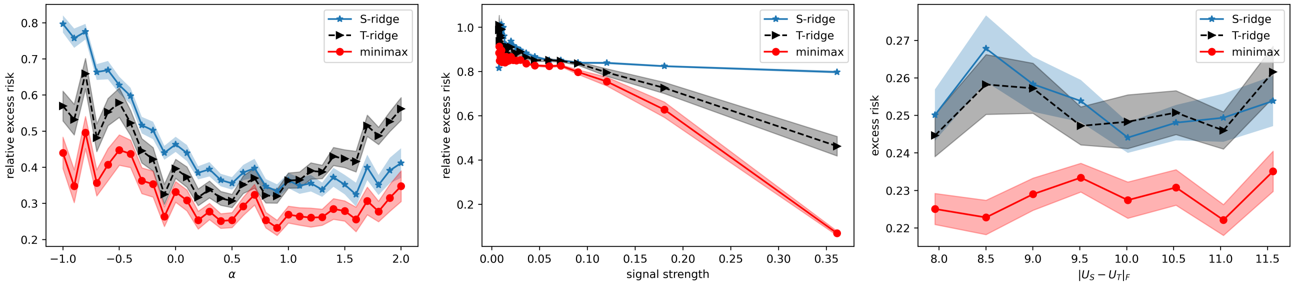
Now let $\hat{C}$ be the minimizer for $\hat{l}(C)$, and C_R be the minimizer for $l_R(C)$.

$$\begin{aligned}
l(\hat{C})\le&(1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_U}}))\hat{l}(\hat{C}) && \text{(w.p. } 1-\delta/10\text{; due to Claim D.4)}\\
\le&(1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_U}}))\hat{l}(C_R) && \text{(Due to the definition of } \hat{C}\text{)}\\
\le&(1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_U}}))^2 l(C_R) && \text{(w.p. } 1-\delta/5\text{; due to Claim D.4)}\\
=&(1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_U}}))l(C_R) && \text{(since } n_U \text{ is large enough)}\\
\le&(1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_U}}))(1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_T}}))l_R(C_R) && \text{(w.p. } 1-3\delta/10\text{; due to Claim D.4)}\\
=&\left(1+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_U}})+O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n_T}})\right)\min_C l_R(C).
\end{aligned}$$

This finishes the proof. □

E. More empirical results

We include some more empirical studies. In the main text our results have small noise. Here we show some more results with larger noise, and also the case with varied eigenspace. For the following results, we use $\sigma=10$ and $r=0.2\sqrt{d}$. Other meta data remains the same as presented in the main text. Figure 4 (a)(b) show similar phenomenon as the small noise setting presented in the main text. From Figure 4 (c) we see no particular relationship between the performance of each algorithm with eigenspace shift.



(a) covariate eigen-spectrum

(b) signal strength

(c) covariate eigenspace

Figure 4: (a): The x-axis α defines the spread of eigen-spectrum of Σ_S: $s_i \propto 1/i^\alpha, t_i \propto 1/i$. (b) x-axis is the normalized value of signal strength: $\|\Sigma_T \boldsymbol{\beta}^*\|/r$. (c) X-axis is the covariate shift due to eigenspace shift measured by $\|U_S - U_T\|_F$.