A Proofs for Section 3

Observation 1. *The model shift in the surrogate-to-target model is equivalent to the covariance shift model (Mallinar et al., 2024). Formally, given $\boldsymbol{\beta}_\star \in \mathbb{R}^p$ and the covariance matrix $\boldsymbol{\Sigma}_t \in \mathbb{R}^{p\times p}$, there exists a unique $\boldsymbol{\beta}^s \in \mathbb{R}^p$ such that the risk of the surrogate-to-target problem $\mathcal{R}(\boldsymbol{\beta}^{s2t})$ with $(\boldsymbol{\beta}_\star, \boldsymbol{\beta}^s, \boldsymbol{\Sigma}_t)$ is equivalent to the risk of the covariance shift model $\mathcal{R}^{cs}(\hat{\boldsymbol{\beta}})$ with $(\boldsymbol{\beta}_\star, \boldsymbol{\Sigma}_s, \boldsymbol{\Sigma}_t)$ for any $\boldsymbol{\Sigma}_s \in \mathbb{R}^{p\times p}$ that is jointly diagonalizable with $\boldsymbol{\Sigma}_t$.*

Proof. By Observation 2, we assume that $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\Sigma}_s$ are diagonal matrices. As $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\Sigma}_s$ are jointly diagonalizable, there exists a unique diagonal matrix $A \in \mathbb{R}^{p\times p}$ such that

$$\boldsymbol{\Sigma}_s = \boldsymbol{A}^\top \boldsymbol{\Sigma}_t \boldsymbol{A}.$$

Then, consider the model shift discussed in Section 3. Take the case where $\boldsymbol{\beta}^s = \boldsymbol{A}\boldsymbol{\beta}_\star$ and labels are generated as $y = \boldsymbol{x}^\top \boldsymbol{\beta}^s + z$, where $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_t)$ and $z \sim \mathcal{N}(0, \sigma_t^2)$. This is equivalent to the case where $y = (\boldsymbol{x}^\top \boldsymbol{A})\boldsymbol{\beta}_\star + z = \bar{\boldsymbol{x}}^\top \boldsymbol{\beta}_\star + z$ such that $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_s)$ and $z \sim \mathcal{N}(0, \sigma_t^2)$. Note that *(i)* the transformed inputs and the labels are identical in both scenarios, and *(ii)* the estimators are computed in the same way. Thus, it follows that the risks $\mathcal{R}(\boldsymbol{\beta}^{s2t})$ and $\mathcal{R}^{\text{cs}}(\hat{\boldsymbol{\beta}})$ are equivalent. The other way follows from an almost identical argument. □

Observation 2. *For any covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, there exists an orthonormal matrix $\boldsymbol{U} \in \mathbb{R}^{p\times p}$ such that the transformation of $\boldsymbol{x} \to \boldsymbol{U}^\top \boldsymbol{x}$ and $\boldsymbol{\beta} \to \boldsymbol{U}^\top \boldsymbol{\beta}$ does not affect the labels $\boldsymbol{y}$ but ensures that the covariance matrix is diagonal.*

Proof. Since the covariance matrix $\boldsymbol{\Sigma}$ is PSD, its unit-norm eigenvectors are orthogonal. Consider the matrix $\boldsymbol{U}$ whose columns are the eigenvectors of $\boldsymbol{\Sigma}$. Then, $\boldsymbol{\Sigma}$ can be expressed as $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, where $\boldsymbol{\Lambda}$ is the diagonal matrix containing the eigenvalues of $\boldsymbol{\Sigma}$. Consider now the transformation

$$z = \boldsymbol{U}^\top \boldsymbol{x} \implies \mathbb{E}\left[zz^\top\right] = \mathbb{E}\left[\boldsymbol{U}^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{U}\right] = \boldsymbol{U}^\top \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right] \boldsymbol{U} = \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{\Lambda}.$$

In this way, the covariance matrix is diagonalized. Thus, the transformation $(\boldsymbol{x}, \boldsymbol{\beta}_\star) \to (\boldsymbol{U}^\top \boldsymbol{x}, \boldsymbol{U}^\top \boldsymbol{\beta}_\star)$ works as intended since the labels are preserved. □

Definition 1. *Let $\kappa_t = p/n > 1$ and $\tau_t \in \mathbb{R}$ be the unique solution of the following equation*

$$\kappa_t^{-1} = \frac{1}{p}\, tr\left((\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1} \boldsymbol{\Sigma}_t\right). \tag{7}$$

Define the function $\gamma_t : \mathbb{R}^p \to \mathbb{R}$ as

$$\gamma_t^2(\boldsymbol{\beta}^s) = \kappa_t \left(\sigma_s^2 + \mathbb{E}_{(\boldsymbol{x}, y^s)\sim \mathcal{D}_t(\boldsymbol{\beta}^s)}[\|\boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\beta}^{s2t} - \boldsymbol{\beta}^s)\|_2^2]\right). \tag{8}$$

Then, the asymptotic risk estimate is defined as

$$\begin{aligned}\bar{\mathcal{R}}^{s2t}_{\kappa_t, \sigma_t}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}_\star, \boldsymbol{\beta}^s) := (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star)^\top \boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) + \gamma_t^2(\boldsymbol{\beta}^s)\, \mathbb{E}_{\boldsymbol{g}_t}[\boldsymbol{\theta}_2^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_2] \\ + \boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t (\boldsymbol{I} - \boldsymbol{\theta}_1) \boldsymbol{\beta}_\star - 2\boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star),\end{aligned} \tag{9}$$

where $\boldsymbol{\theta}_1 := (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1} \boldsymbol{\Sigma}_t$, $\boldsymbol{\theta}_2 := (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1} \boldsymbol{\Sigma}_t^{1/2} \frac{\boldsymbol{g}_t}{\sqrt{p}}$, and $\boldsymbol{g}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$.

Theorem 1. *Suppose that, for some constant $M_t > 1$, we have $1/M_t \le \kappa_t, \sigma_t^2 \le M_t$ and $\|\boldsymbol{\Sigma}_t\|_{op}, \left\|\boldsymbol{\Sigma}_t^{-1}\right\|_{op} \le M_t$. Recall from (5) that $\mathcal{R}(\boldsymbol{\beta}^{s2t})$ represents the risk of the surrogate-to-target model given $\boldsymbol{\beta}^s$. Then, there exists a constant $C = C(M_t)$ such that, for any $\varepsilon \in (0, 1/2]$, the following holds with $R + 1 < M_t$:*

$$\sup_{\boldsymbol{\beta}_\star, \boldsymbol{\beta}^s \in \boldsymbol{B}_p(R)} \mathbb{P}\left(\left|\mathcal{R}(\boldsymbol{\beta}^{s2t}) - \bar{\mathcal{R}}^{s2t}_{\kappa_t, \sigma_t}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}_\star, \boldsymbol{\beta}^s)\right| \ge \varepsilon\right) \le C p e^{-p\varepsilon^4/C}. \tag{10}$$

Proof. Even though the claim readily follows from Theorem 2, we give a proof for the sake of completeness.

Define a function $f_1 : \mathbb{R}^p \to \mathbb{R}$ as $f_1(\boldsymbol{x}) = \|\boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{x} - \boldsymbol{\beta}_\star)\|_2^2$. The gradient of this function is

$$\|\nabla f_1(\boldsymbol{x})\|_2 = \|2\boldsymbol{\Sigma}_t(\boldsymbol{x} - \boldsymbol{\beta}_\star)\|_2 \le 2\,\|\boldsymbol{\Sigma}_t\|_{\text{op}}\, \|\boldsymbol{x} - \boldsymbol{\beta}_\star\|_2.$$

Using Corollary 2, there exists an event E with $\mathbb{P}(E^c) \le C_t e^{-p/C_t}$ where $C_t = C_t(M_t, \frac{M_t - R}{2})$ (with the definition of M_t in Corollary 2), such that $f_1(\boldsymbol{\beta}^{s2t})$ is $2M_t^2$-Lipschitz if $\boldsymbol{\beta}_\star, \boldsymbol{\beta}^s \in \boldsymbol{B}_p(R)$. Applying Theorem 3 on the target model, there exists a constant $\bar{C}_s = \bar{C}_s(M_t)$ such that for any $\varepsilon \in (0, 1/2]$, we obtain

$$\sup_{\boldsymbol{\beta}^s \in \boldsymbol{B}(\frac{M_t+R}{2})} \mathbb{P}\left(\left|f(\boldsymbol{\beta}^{s2t}) - \mathbb{E}_{\boldsymbol{g}_t}[f(X^t_{\kappa_t,\sigma_t^2}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}^s, \boldsymbol{g}_t))]\right| \ge \varepsilon\right) \le C p \mathrm{e}^{-p\varepsilon^4/C}, \tag{13}$$

where $f(\boldsymbol{\beta}^{s2t}) = \mathcal{R}(\boldsymbol{\beta}^{s2t})$ and

$$X^t_{\kappa_t,\sigma_t^2}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}^s, \boldsymbol{g}_t) = (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t \left[\boldsymbol{\beta}^s + \frac{\boldsymbol{\Sigma}_t^{-1/2}\gamma_t(\boldsymbol{\beta}^s)\boldsymbol{g}_t}{\sqrt{p}}\right].$$

Furthermore,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{g}_t}\left[f(X^s_{\kappa_t,\sigma_t^2}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}^s, \boldsymbol{g}_t))\right] &= \mathbb{E}_{\boldsymbol{g}_t}\left[\|\boldsymbol{\Sigma}_t^{1/2}\left(\boldsymbol{\theta}_1(\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) - (\boldsymbol{I} - \boldsymbol{\theta}_1)\boldsymbol{\beta}_\star + \boldsymbol{\theta}_2\gamma_t(\boldsymbol{\beta}^s)\right)\|_2^2\right] \\
&= (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star)^\top \boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) + \gamma_t^2(\boldsymbol{\beta}^s)\,\mathbb{E}_{\boldsymbol{g}_t}[\boldsymbol{\theta}_2^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_2] \\
&\quad + \boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t (\boldsymbol{I} - \boldsymbol{\theta}_1)\boldsymbol{\beta}_\star - 2\boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star),
\end{aligned} \tag{14}$$

where $\boldsymbol{\theta}_1 := (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t$ and $\boldsymbol{\theta}_2 := (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}\frac{\boldsymbol{g}_t}{\sqrt{p}}$. This completes the proof. □

Proposition 1. *Let* $\Omega = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_t^2(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-2})}{n}$. *The optimal surrogate* $\boldsymbol{\beta}^s$ *minimizing the asymptotic risk in* (9) *is*

$$\boldsymbol{\beta}^{s*} = \left((\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t + \frac{\Omega\tau_t^2}{1-\Omega}\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1}\right)^{-1}\boldsymbol{\beta}_\star.$$

Proof. We have that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{g}_t}\left[f(X^t_{\kappa_t,\sigma_t^2}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}^s, \boldsymbol{g}_t))\right] &= \mathbb{E}_{\boldsymbol{g}_t}\left[\|\boldsymbol{\Sigma}_t^{1/2}\left(\boldsymbol{\theta}_1(\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) - (\boldsymbol{I} - \boldsymbol{\theta}_1)\boldsymbol{\beta}_\star + \boldsymbol{\theta}_2\gamma_t(\boldsymbol{\beta}^s)\right)\|_2^2\right] \\
&= (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star)^\top \boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) + \gamma_t^2(\boldsymbol{\beta}^s)\,\mathbb{E}_{\boldsymbol{g}_t}[\boldsymbol{\theta}_2^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_2] \\
&\quad + \boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t (\boldsymbol{I} - \boldsymbol{\theta}_1)\boldsymbol{\beta}_\star - 2\boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star),
\end{aligned}$$

where $\boldsymbol{\theta}_1 := (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t$, $\boldsymbol{\theta}_2 := (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}\frac{\boldsymbol{g}_t}{\sqrt{p}}$, and

$$\gamma_s^2(\boldsymbol{\beta}^s) := \kappa_t \frac{\sigma_t^2 + \tau_t^2\|(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\beta}^s\|_2^2}{1 - \frac{1}{n}\mathrm{tr}\left((\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t^2\right)}.$$

In order to optimize this with respect to $\boldsymbol{\beta}^s$, let's take the derivative:

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\beta}^s}\,\mathbb{E}_{\boldsymbol{g}_t}\left[f(X^t_{\kappa_t,\sigma_t^2}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}^s, \boldsymbol{g}_t))\right] \\
&= 2\boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) - 2\boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t (\boldsymbol{I} - \boldsymbol{\theta}_1)\boldsymbol{\beta}_\star + 2\frac{\kappa_t\tau_t^2}{1-\Omega}\boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-2}\boldsymbol{\beta}^s \frac{\mathrm{tr}\left(\boldsymbol{\Sigma}_t^2(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-2}\right)}{p} \\
&= 2\boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 \boldsymbol{\beta}^s - 2\boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 \boldsymbol{\beta}_\star + 2\frac{\kappa_t\tau_t^2}{1-\Omega}\boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-2}\boldsymbol{\beta}^s \frac{\mathrm{tr}\left(\boldsymbol{\Sigma}_t^2(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-2}\right)}{p} \\
&= 2\boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 \boldsymbol{\beta}^s - 2\boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 \boldsymbol{\beta}_\star + 2\frac{\kappa_t\tau_t^2}{1-\Omega}\boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-2}\boldsymbol{\beta}^s \frac{n\Omega}{p} \\
&\implies \boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 \boldsymbol{\beta}^{s*} - \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 \boldsymbol{\beta}_\star + \frac{\Omega\tau_t^2}{1-\Omega}\boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-2}\boldsymbol{\beta}^{s*} = 0 \\
&\implies (\boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 + \frac{\Omega\tau_t^2}{1-\Omega}\boldsymbol{\theta}_1(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1})\boldsymbol{\beta}^{s*} = \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 \boldsymbol{\beta}_\star
\end{aligned}$$

Hence, the claimed result follows. □

Corollary 1. *Without loss of generality, suppose that $\boldsymbol{\Sigma}_t$ is diagonal.*[2] *Let $(\lambda_i)_{i=1}^p$ be the eigenvalues of $\boldsymbol{\Sigma}_t$ in non-increasing order and let $\zeta_i = \frac{\tau_t}{\lambda_i+\tau_t}$ for $i \in [p]$. Then, the following results hold:*

1. $\beta_i^{s*} = (\beta_*)_i \left((1-\zeta_i) + \zeta_i \frac{\Omega}{1-\Omega}\frac{\zeta_i}{1-\zeta_i}\right)^{-1}$ *for every* $i \in [p]$.
2. $|\beta_i^{s*}| > |(\beta_*)_i|$ *if and only if* $1-\zeta_i > \Omega = \frac{\sum_{j=1}^p (1-\zeta_j)^2}{\sum_{j=1}^p (1-\zeta_j)}$ *for every* $i \in [p]$.
3. $\boldsymbol{\beta}^{s*} = \boldsymbol{\beta}_\star$ *if and only if the covariance matrix* $\boldsymbol{\Sigma}_t = c\boldsymbol{I}$ *for some* $c \in \mathbb{R}$.

Proof. When the definition of ζ_i and Ω is plugged in Proposition 1, the first claim is obtained. Using the diagonalization assumption on $\boldsymbol{\Sigma}_t$, let's analyze only the i-th component of the optimal surrogate given in the Proposition 1:

$$\begin{aligned}
\beta_i^{s*} &= \frac{1}{\frac{\lambda_i}{\lambda_i+\tau_t} + \frac{\Omega}{1-\Omega}\frac{\tau_t^2}{\lambda_i(\lambda_i+\tau_t)}}(\beta_*)_i \\
\iff \beta_i^{s*} &= \frac{\frac{\lambda_i}{\lambda_i+\tau_t}}{\left(\frac{\lambda_i}{\lambda_i+\tau_t}\right)^2 + \frac{\Omega}{1-\Omega}\left(\frac{\tau_t}{\lambda_i+\tau_t}\right)^2}(\beta_*)_i \\
\iff \beta_i^{s*} &= (\beta_*)_i \frac{(1-\zeta_i)}{(1-\zeta_i)^2 + \frac{\Omega}{1-\Omega}\zeta_i^2} \\
\iff \beta_i^{s*} &= (\beta_*)_i \frac{1}{(1-\zeta_i) + \frac{\Omega}{1-\Omega}\frac{\zeta_i}{1-\zeta_i}\zeta_i}.
\end{aligned}$$

It's now clear that $\zeta_i > 1-\Omega$ if and only if $|\beta_i^{s*}| < |(\beta_*)_i|$.

Let's now check when the ratio between them is 1. Algebraic manipulations give:

$$\begin{aligned}
&\frac{(1-\zeta_i)}{(1-\zeta_i)^2 + \frac{\Omega}{1-\Omega}\zeta_i^2} = 1 \\
&\iff (1-\zeta_i) - (1-\zeta_i)^2 = \frac{\Omega}{1-\Omega}\zeta_i^2 \\
&\iff \zeta_i = 1-\Omega \iff 1-\zeta_i = \Omega \text{ where } \Omega = \frac{\sum_{i=1}^p (1-\zeta_i)^2}{\sum_{i=1}^p (1-\zeta_i)}.
\end{aligned}$$

This suggests $\boldsymbol{\beta}^{s*} = \boldsymbol{\beta}_\star$ if all ζ_i's are equal, which implies that all λ_i's are equal. Concluding, the covariance matrix is a multiple of the identity if and only if $\boldsymbol{\beta}^{s*} = \boldsymbol{\beta}_\star$. □

Proposition 2. *Consider the target model in (6), assume that $\boldsymbol{\Sigma}_t$ is diagonal, and recall the definitions of ζ_i and Ω. Then, the following results hold:*

1. *If the mask operation $\mathcal{M}$ selects all the features that satisfy $1-\zeta_i^2 > \Omega$, then the surrogate-to-target model outperforms the standard target model in the asymptotic risk in (9).*
2. *Let $\boldsymbol{M}$ represent the set of all possible $\mathcal{M}$, where $|\boldsymbol{M}| = 2^p$. The optimal $\mathcal{M}^*$ for the asymptotic risk in (9) within $\boldsymbol{M}$ is the one that selects all features satisfying $1-\zeta_i^2 > \Omega$.*

Proof. For the purposes of analysis, we assume, without loss of generality, that the first p_s dimensions are selected from $\boldsymbol{\beta}_\star$ in $\mathcal{M}(\boldsymbol{\beta}_\star) = \boldsymbol{\beta}^s \in \mathbb{R}^{p_s}$. Based on this, we no longer need to have the decreasing order for the corresponding λ_i's. From the excess test risk formula in Definition 2, we have that

$$\mathcal{R}(\boldsymbol{\beta}^t) = \mathbb{E}\left[\left(y - \boldsymbol{x}^\top \boldsymbol{\beta}^t\right)^2\right] - \sigma_t^2 = \frac{\mathcal{B}(\boldsymbol{\beta}_\star) + \sigma_t^2\Omega}{1-\Omega}. \tag{15}$$

[2] If not, there exists an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{p\times p}$ s.t. $\boldsymbol{U}\boldsymbol{\Sigma}_t\boldsymbol{U}^\top$ is diagonal. Then, we can consider the covariance matrix as $\boldsymbol{U}\boldsymbol{\Sigma}_t\boldsymbol{U}^\top$ and the ground truth parameter as $\boldsymbol{U}\boldsymbol{\beta}_\star$, which behaves the same as the original parameters, see Observation 2.

Next, we write the excess test risk formula for the surrogate-to-target model with respect to the original ground truth labels:

$$\mathcal{R}(\boldsymbol{\beta}^{s2t}) = \mathbb{E}\left[\left(y - \boldsymbol{x}^\top \boldsymbol{\beta}^{s2t}\right)^2\right] - \sigma_t^2 = \mathbb{E}\left[\left(\boldsymbol{\beta}^{s2t} - \boldsymbol{\beta}_\star\right)^\top \boldsymbol{\Sigma}\left(\boldsymbol{\beta}^{s2t} - \boldsymbol{\beta}_\star\right)\right].$$

Now, consider the zero-padded vector $\bar{\boldsymbol{\beta}}^s = \begin{bmatrix} \boldsymbol{\beta}^s \\ \mathbf{0}_{p-p_s} \end{bmatrix} \in \mathbb{R}^p$, and define $(\bar{\boldsymbol{\beta}}^s)' = \boldsymbol{\beta}_\star - \bar{\boldsymbol{\beta}}^s \in \mathbb{R}^p$ of which the first p_s dimensions are zero. In this way, we can consider the labels in the second training phase as $y^s = \boldsymbol{x}^\top \bar{\boldsymbol{\beta}}^s + z$, where $z \sim \mathcal{N}(0, \sigma_t^2)$. Applying the test risk estimate in Definition 2, we obtain:

$$\mathbb{E}\left[\left(y^s - \boldsymbol{x}^\top \boldsymbol{\beta}^{s2t}\right)^2\right] = \mathbb{E}\left[\left(\boldsymbol{\beta}^{s2t} - \boldsymbol{\beta}^s\right)^\top \boldsymbol{\Sigma}\left(\boldsymbol{\beta}^{s2t} - \boldsymbol{\beta}^s\right)\right] = \frac{\mathcal{B}(\bar{\boldsymbol{\beta}}^s) + \sigma_t^2 \Omega}{1 - \Omega}.$$

We further derive

$$\begin{aligned}
\mathbb{E}\left[\left(y - \boldsymbol{x}^\top \boldsymbol{\beta}^{s2t}\right)^2\right] &= \mathbb{E}\left[\left(y^s - \boldsymbol{x}^\top \boldsymbol{\beta}^{s2t} + \boldsymbol{x}^\top (\bar{\boldsymbol{\beta}}^s)'\right)^2\right] \\
&= \mathbb{E}\left[\left(y^s - \boldsymbol{x}^\top \boldsymbol{\beta}^{s2t}\right)^2\right] - 2\mathbb{E}\left[\left(y^s - \boldsymbol{x}^\top \boldsymbol{\beta}^{s2t}\right)\left(\boldsymbol{x}^\top (\bar{\boldsymbol{\beta}}^s)'\right)\right] + \mathbb{E}\left[\left(\boldsymbol{x}^\top (\bar{\boldsymbol{\beta}}^s)'\right)^2\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\left(y^s - \boldsymbol{x}^\top \boldsymbol{\beta}^{s2t}\right)^2\right] - 2\mathbb{E}\left[y^s - \boldsymbol{x}^\top \boldsymbol{\beta}^{s2t}\right] \underbrace{\mathbb{E}\left[\boldsymbol{x}^\top (\bar{\boldsymbol{\beta}}^s)'\right]}_{=0} + \mathbb{E}\left[\left(\boldsymbol{x}^\top (\bar{\boldsymbol{\beta}}^s)'\right)^2\right] \\
&= \mathbb{E}\left[\left(y^s - \boldsymbol{x}^\top \boldsymbol{\beta}^{s2t}\right)^2\right] + \mathbb{E}\left[\left(\boldsymbol{x}^\top (\bar{\boldsymbol{\beta}}^s)'\right)^2\right] \\
&= \mathbb{E}\left[\left(y^s - \boldsymbol{x}^\top \boldsymbol{\beta}^{s2t}\right)^2\right] + \sum_{i=p_s+1}^{p} \lambda_i \beta_i^2 \\
&= \frac{\mathcal{B}(\bar{\boldsymbol{\beta}}^s) + \sigma_t^2 \Omega}{1 - \Omega} + \sum_{i=p_s+1}^{p} \lambda_i \beta_i^2.
\end{aligned} \tag{16}$$

where in the above equality (a) follows from the fact that the components x_i are independent as the covariance matrix is diagonal. Thus, the risk difference between the target and surrogate-to-target models is

$$\begin{aligned}
\mathcal{R}(\boldsymbol{\beta}^t) - \mathcal{R}(\boldsymbol{\beta}^{s2t}) &= \frac{\mathcal{B}(\boldsymbol{\beta}_\star) - \mathcal{B}(\bar{\boldsymbol{\beta}}^s)}{1 - \Omega} - \sum_{i=p_s+1}^{p} \lambda_i \beta_i^2 \\
&= \frac{\sum_{i=p_s+1}^{p} \lambda_i \zeta_i^2 \beta_i^2}{1 - \Omega} - \sum_{i=p_s+1}^{p} \lambda_i \beta_i^2.
\end{aligned}$$

We observe that each dimension's contribution to the excess test risk can be analyzed individually. Therefore, if

$$\zeta_i^2 > 1 - \Omega, \tag{17}$$

excluding feature i in the feature selection reduces the overall risk $\mathcal{R}(\boldsymbol{\beta}^{s2t})$. Along the same lines, the projection $\mathcal{M}$ that selects all the features i that satisfy $\zeta_i^2 < 1 - \Omega$ minimizes the asymptotic excess test risk. □

B Proofs for Section 4

Definition 2 (Omniscient test risk estimate)**.** *Fix $p > n \geq 1$. Given a covariance $\boldsymbol{\Sigma} = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{U}^\top$, $\boldsymbol{\beta}_\star$, and the noise term σ, set $\bar{\boldsymbol{\beta}} = \boldsymbol{U}^\top \boldsymbol{\beta}_\star$ and define $\tau \in \mathbb{R}$ as the unique non-negative solution of $n = \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + \tau}$. Then, the excess test risk estimate is the following:*

$$\mathcal{R}(\hat{\boldsymbol{\beta}}) \approx \mathbb{E}_{\hat{\boldsymbol{\beta}} \sim D(\boldsymbol{\beta}_\star)}\left[(y - \boldsymbol{x}^\top \hat{\boldsymbol{\beta}})^2\right] - \sigma^2 = \frac{\sigma^2 \Omega + \mathcal{B}(\bar{\boldsymbol{\beta}})}{1 - \Omega}, \tag{11}$$

$$\textit{where} \quad \zeta_i = \frac{\tau}{\lambda_i + \tau}, \quad \Omega = \frac{1}{n}\sum_{i=1}^{p} (1 - \zeta_i)^2, \quad \mathcal{B}(\bar{\boldsymbol{\beta}}) = \sum_{i=1}^{p} \lambda_i \zeta_i^2 \bar{\beta_i}^2.$$

In the following proof, we suppose that the empirical distributions of $\bar{\boldsymbol{\beta}}$ and $\boldsymbol{\lambda}$ converge as $p \to \infty$ having fixed the ratio $p/n = \kappa$. Then, we will prove that the omniscient risk converges to the asymptotic risk defined in (9).

Proof for the proportional asymptotic case. Using Theorem 2.3 of Han & Xu (2023), we can estimate $\hat{\boldsymbol{\beta}}$ as follows:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{\Sigma} + \tau \boldsymbol{I})^{-1} \boldsymbol{\Sigma} \left(\boldsymbol{\beta}_\star + \frac{\boldsymbol{\Sigma}^{-1/2} \gamma(\boldsymbol{\beta}_\star) \boldsymbol{g}}{\sqrt{p}} \right),$$

where

$$\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_p), \quad \gamma(\boldsymbol{\beta}_\star)^2 = \kappa \frac{\sigma + \tau^2 \|(\boldsymbol{\Sigma} + \tau \boldsymbol{I})^{-1} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}_\star\|_2^2}{1 - \frac{1}{n} \mathtt{tr}\left((\boldsymbol{\Sigma} + \tau \boldsymbol{I})^{-2} \boldsymbol{\Sigma}^2\right)}, \quad \tau \text{ is the solution to } n = \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + \tau}.$$

Let

$$\boldsymbol{X}_1 = (\boldsymbol{\Sigma} + \tau \boldsymbol{I})^{-1} \boldsymbol{\Sigma} \quad , \quad \boldsymbol{X}_2 = \frac{(\boldsymbol{\Sigma} + \tau \boldsymbol{I})^{-1} \boldsymbol{\Sigma}^{1/2} \gamma(\boldsymbol{\beta}_\star)}{\sqrt{p}}.$$

Using this estimate, we can calculate the excess test risk as

$$\begin{aligned} \mathcal{R}(\hat{\boldsymbol{\beta}}) = & \quad \mathbb{E}\left[((\boldsymbol{X}_1 - \boldsymbol{I}) \boldsymbol{\beta}_\star + \boldsymbol{X}_2 \boldsymbol{g})^\top \boldsymbol{\Sigma} ((\boldsymbol{X}_1 - \boldsymbol{I}) \boldsymbol{\beta}_\star + \boldsymbol{X}_2 \boldsymbol{g}) \right] \\ = & \quad \boldsymbol{\beta}_\star{}^\top (\boldsymbol{X}_1 - \boldsymbol{I})^\top \boldsymbol{\Sigma} (\boldsymbol{X}_1 - \boldsymbol{I}) \boldsymbol{\beta}_\star + \mathbb{E}\left[\boldsymbol{g}^\top \boldsymbol{X}_2^\top \boldsymbol{\Sigma} \boldsymbol{X}_2 \boldsymbol{g} \right] \\ = & \quad \boldsymbol{\beta}_\star{}^\top (\boldsymbol{X}_1 - \boldsymbol{I})^\top \boldsymbol{\Sigma} (\boldsymbol{X}_1 - \boldsymbol{I}) \boldsymbol{\beta}_\star + \mathtt{tr}\left(\boldsymbol{X}_2^\top \boldsymbol{\Sigma} \boldsymbol{X}_2 \right). \end{aligned} \tag{18}$$

Then by recalling the eigendecomposition for the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top$, we have

$$\begin{aligned} \boldsymbol{X}_1 &= (\boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top + \tau \boldsymbol{U} \boldsymbol{U}^\top)^{-1} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \\ &= \boldsymbol{U} (\boldsymbol{\Lambda} + \tau \boldsymbol{I})^{-1} \boldsymbol{U}^\top \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \\ &= \boldsymbol{U} \operatorname{diag}\left(\frac{\boldsymbol{\lambda}}{\boldsymbol{\lambda} + \tau} \right) \boldsymbol{U}^\top. \end{aligned}$$

Using the diagonalization of $\boldsymbol{I}$, $\boldsymbol{X}_1 - \boldsymbol{I}$ can now be computed as

$$\boldsymbol{X}_1 - \boldsymbol{I} = \boldsymbol{U} \operatorname{diag}\left(\frac{-\tau}{\boldsymbol{\lambda} + \tau} \right) \boldsymbol{U}^\top.$$

Let's now compute

$$\begin{aligned} \boldsymbol{\beta}_\star{}^\top (\boldsymbol{X}_1 - \boldsymbol{I})^\top \boldsymbol{\Sigma} (\boldsymbol{X}_1 - \boldsymbol{I}) \boldsymbol{\beta}_\star &= \boldsymbol{\beta}_\star{}^\top \boldsymbol{U} \operatorname{diag}\left(\frac{-\tau}{\boldsymbol{\lambda} + \tau} \right) \boldsymbol{U}^\top \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U} \operatorname{diag}\left(\frac{-\tau}{\boldsymbol{\lambda} + \tau} \right) \boldsymbol{U}^\top \boldsymbol{\beta}_\star \\ &= \boldsymbol{\beta}_\star{}^\top \boldsymbol{U} \operatorname{diag}\left(\frac{\boldsymbol{\lambda} \tau^2}{(\boldsymbol{\lambda} + \tau)^2} \right) \boldsymbol{U}^\top \boldsymbol{\beta}_\star. \end{aligned}$$

As $\bar{\boldsymbol{\beta}} = \boldsymbol{U}^\top \boldsymbol{\beta}_\star$, we obtain that the RHS of the previous expression equals

$$\sum_{i=1}^{p} \frac{\lambda_i \tau^2 \bar{\beta}_i^2}{(\lambda_i + \tau)^2} = \mathcal{B}(\bar{\boldsymbol{\beta}}).$$

Next, we write more compactly the terms $\mathtt{tr}\left(\boldsymbol{X}_2^\top \boldsymbol{\Sigma} \boldsymbol{X}_2 \right)$ and $\gamma(\boldsymbol{\beta}_\star)^2$. By defining the short-hand notation $\Omega = \frac{1}{n} \mathtt{tr}\left((\boldsymbol{\Sigma} + \tau \boldsymbol{I})^{-2} \boldsymbol{\Sigma}^2 \right) = \frac{1}{n} \sum_{i=1}^{p} (1 - \zeta_i)^2$, we have

$$\mathtt{tr}\left(\boldsymbol{X}_2^\top \boldsymbol{\Sigma} \boldsymbol{X}_2 \right) = \frac{\gamma(\boldsymbol{\beta}_\star)^2}{p} \sum_{i=1}^{p} \left(\frac{\lambda_i}{\lambda_i + \tau} \right)^2 = \frac{\gamma(\boldsymbol{\beta}_\star)^2 n \Omega}{p}$$

$$\gamma(\boldsymbol{\beta}_\star)^2 = \kappa \frac{\sigma^2 + \tau^2 \|(\boldsymbol{\Sigma} + \tau \boldsymbol{I})^{-1} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}_\star\|_2^2}{1 - \Omega} = \kappa \frac{\sigma^2 + \sum_{i=1}^{p} \frac{\lambda_i \tau^2 \bar{\beta}_i^2}{(\lambda_i + \tau)^2}}{1 - \Omega} = \kappa \frac{\sigma^2 + \mathcal{B}(\bar{\boldsymbol{\beta}})}{1 - \Omega},$$

where $\kappa = \frac{p}{n}$. Hence, putting it all together in (18) gives the desired result. □

Proposition 3 (Asymptotic analysis of τ_t and Ω)**.** *Let $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ be diagonal and $\mathbf{\Sigma}_{i,i} = \lambda_i = i^{-\alpha}$ for $1 < \alpha$. If τ_t and Ω satisfy the equations*

$$\sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \tau_t} = n, \qquad n\Omega = \sum_{i=1}^{\infty} \left(\frac{i^{-\alpha}}{i^{-\alpha} + \tau_t}\right)^2,$$

then the following results hold

$$\begin{aligned} \tau_t &= cn^{-\alpha}\left(1 + O(n^{-1})\right), \qquad \text{for } c = \left(\frac{\pi}{\alpha \sin(\pi/\alpha)}\right)^{\alpha}, \\ \Omega &= \frac{\alpha - 1}{\alpha} - O(n^{-1}). \end{aligned} \tag{12}$$

Proof. We start with the asymptotic analysis of τ_t. Along the same lines as Simon et al. (2024), since $\frac{i^{-\alpha}}{i^{-\alpha}+\tau_t}$ is a monotonically decreasing function, we have:

$$n = \sum_{i=1}^{\infty} \frac{i^{-\alpha}}{i^{-\alpha} + \tau_t} \leq \int_0^{\infty} \frac{x^{-\alpha}}{x^{-\alpha} + \tau_t}\, dx = \frac{\pi}{\alpha \sin(\pi/\alpha)} \tau_t^{-1/\alpha}.$$

Furthermore,

$$\frac{\pi}{\alpha \sin(\pi/\alpha)} \tau_t^{-1/\alpha} - 1 = \int_0^{\infty} \frac{x^{-\alpha}}{x^{-\alpha} + \tau_t}\, dx - 1 \leq \int_1^{\infty} \frac{x^{-\alpha}}{x^{-\alpha} + \tau_t}\, dx \leq \sum_{i=1}^{\infty} \frac{i^{-\alpha}}{i^{-\alpha} + \tau_t} = n.$$

Hence, combining these two facts gives

$$\begin{aligned} &\frac{\pi}{\alpha \sin(\pi/\alpha)} \tau_t^{-1/\alpha} - 1 \leq n \leq \frac{\pi}{\alpha \sin(\pi/\alpha)} \tau_t^{-1/\alpha} \\ \iff &\left(\frac{(n+1)\,\alpha \sin(\pi/\alpha)}{\pi}\right)^{-\alpha} \leq \tau_t \leq \left(\frac{n\alpha \sin(\pi/\alpha)}{\pi}\right)^{-\alpha}, \end{aligned}$$

which leads to the desired result.

Next, we move to the asymptotic analysis of Ω. We have that

$$n\Omega = \sum_{i=1}^{\infty} \left(\frac{i^{-\alpha}}{i^{-\alpha} + \tau_t}\right)^2 \leq \int_0^{\infty} \left(\frac{x^{-\alpha}}{x^{-\alpha} + \tau_t}\right)^2 dx = \frac{\pi(\alpha - 1)}{\alpha^2 \sin(\pi/\alpha)} \tau_t^{-1/\alpha}.$$

Besides, since the summand is monotonically decreasing, we also have

$$\frac{\pi(\alpha - 1)}{\alpha^2 \sin(\pi/\alpha)} \tau_t^{-1/\alpha} - 1 \leq \int_0^{\infty} \left(\frac{x^{-\alpha}}{x^{-\alpha} + \tau_t}\right)^2 dx - 1 \leq \int_1^{\infty} \left(\frac{x^{-\alpha}}{x^{-\alpha} + \tau_t}\right)^2 dx \leq \sum_{i=1}^{\infty} \left(\frac{i^{-\alpha}}{i^{-\alpha} + \tau_t}\right)^2 = n\Omega.$$

Hence,

$$\frac{\pi(\alpha - 1)}{\alpha^2 \sin(\pi/\alpha)} \tau_t^{-1/\alpha} - 1 \leq n\Omega \leq \frac{\pi(\alpha - 1)}{\alpha^2 \sin(\pi/\alpha)} \tau_t^{-1/\alpha}. \tag{19}$$

By the hypothesis on τ_t, we have that

$$\tau_t^{-1/\alpha} = n \frac{\alpha \sin(\pi/\alpha)}{\pi} \left(1 - O\left(n^{-1}\right)\right), \tag{20}$$

and plugging this in (19) gives the desired result. □

Proposition 4. *Set the constants $C_1 := \dfrac{\alpha \sin(\pi/\alpha)}{\pi(\alpha - 1)^{1/\alpha}}$ and $C_2 := \dfrac{\alpha \sin(\pi/\alpha)}{\pi(\sqrt{\alpha} - 1)^{1/\alpha}}$ and assume the power-law eigenstructure $\lambda_i = i^{-\alpha}$ for $1 < \alpha$. Then, the indices i for which $\zeta_i < 1 - \Omega$ are $i < nC_1 + O(1)$; while the indices i for which is $\zeta_i^2 < 1 - \Omega$ are $i < nC_2 + O(1)$.*

Proof. Recall from Proposition 2 that we should identify indices i which satisfy the condition $\zeta_i^2 > 1 - \Omega$ to decide if we're better off not selecting this dimension i in the surrogate model. Furthermore, Proposition 3 gives that $\Omega = \dfrac{\alpha - 1}{\alpha} - O(n^{-1})$. Putting these together, we have

$$
\begin{aligned}
& \zeta_i^2 > 1 - \Omega \\
\iff & \zeta_i^2 > c' \quad \text{where } c' = \frac{1}{\alpha} + O(n^{-1}) \\
\iff & \frac{\tau_t^2}{(\tau_t + i^{-\alpha})^2} = \frac{\tau_t^2 i^{2\alpha}}{(\tau_t i^{\alpha} + 1)^2} > c' \\
\iff & (1 - c')\tau_t^2 i^{2\alpha} > 2c'\tau_t i^{\alpha} + c' \\
\iff & \left(\sqrt{1 - c'}\tau_t i^{\alpha} - \frac{c'}{\sqrt{1 - c'}}\right)^2 > \frac{c'}{1 - c'} \\
\iff & i^{\alpha} > \frac{\sqrt{c'}}{\tau_t(1 - \sqrt{c'})} \\
\iff & i > \tau_t^{-1/\alpha}\left(\frac{\sqrt{c'}}{1 - \sqrt{c'}}\right)^{1/\alpha}
\end{aligned}
$$

As $c' = \frac{1}{\alpha} + O(n^{-1})$, we get $\left(\frac{\sqrt{c'}}{1-\sqrt{c'}}\right)^{1/\alpha} = \frac{1}{(\sqrt{\alpha}-1)^{1/\alpha}}(1 + O(n^{-1}))$. Incorporating (20), we achieve that

$$
\tau_t^{-1/\alpha}\left(\frac{\sqrt{c'}}{1 - \sqrt{c'}}\right)^{1/\alpha} = n\frac{\alpha \sin(\pi/\alpha)}{\pi(\sqrt{\alpha} - 1)^{1/\alpha}}\left(1 + O(n^{-1})\right) = nC_2 + O(1).
$$

Similarly, by following the same procedure with the initial inequality $\zeta_i > 1 - \Omega$, we get

$$
\zeta_i > 1 - \Omega \iff i > nC_1 + O(1), \quad \text{where} \quad C_1 = \frac{\alpha \sin(\pi/\alpha)}{\pi(\alpha - 1)^{1/\alpha}}.
$$

□

In Figure 3, we compare the empirical results with theoretical predictions for the number of features that meet the selection criteria in the optimal mask $\mathcal{M}^*$ ($\zeta_i^2 < 1 - \Omega$). The theoretical value, calculated as $n\dfrac{\alpha \sin(\pi/\alpha)}{\pi(\sqrt{\alpha} - 1)^{1/\alpha}}$ ignoring the $O(1)$ term, aligns well with the experimental data and the accuracy in estimation increases with α.

Proposition 6 (Scaling law for masked surrogate-to-target model)**.** *Together with the eigenvalues, also assume now power-law form for $\lambda_i\beta_i^2$, that is $\lambda_i\beta_i^2 = i^{-\beta}$ for $\beta > 1$. Then, in the limit of $p \to \infty$, the excess test risk for the masked surrogate-to-target model with the optimal dimensionality has the same scaling law as the reference (target) model:*

$$
\mathcal{R}(\boldsymbol{\beta}^{s2t}) = \Theta(n^{-(\beta-1)}) \quad \text{if } \beta < 2\alpha + 1,
$$

and

$$
\mathcal{R}(\boldsymbol{\beta}^{s2t}) = \Theta(n^{-2\alpha}) \quad \text{if } \beta > 2\alpha + 1.
$$

Proof. As discussed in Section 4, in order to analyze the model's inherent error, we need to set $\sigma_t^2 = O(n^{-\gamma})$ where γ is the exponent characterizing the scaling law of the test risk in the noiseless setting. We will work on this proof in two cases depending on β and $2\alpha + 1$.

Case 1: $\beta < 2\alpha + 1$. In this case, it is previously stated by Cui et al. (2022); Simon et al. (2024) that the test risk of ridgeless overparameterized linear regression can be described in the scaling sense as $\mathbf{err} = \Theta(n^{-\beta+1})$ when $\beta < 2\alpha + 1$. Consider the optimal mask operation $\mathcal{M}$ mentioned in Proposition 2 that selects all features

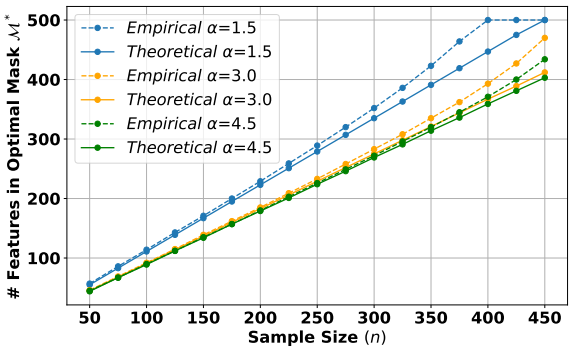

Figure 3: Comparison of the empirical and theoretical number of features satisfying the feature selection condition in the optimal mask $\mathcal{M}^*$ ($\zeta_i^2 < 1-\Omega$). The theoretical value is calculated as $n\dfrac{\alpha \sin(\pi/\alpha)}{\pi(\sqrt{\alpha}-1)^{1/\alpha}}$, ignoring the $O(1)$ in Proposition 4. **Setting:** The feature size is $p = 500$, and the feature covariance follows the power-law structure $\lambda_i = i^{-\alpha}$ for $\alpha = 1.5, 3.0$, and 4.5.

satisfying $1-\zeta_i^2 > \Omega$. Let p_s be the number of selected features. We can then decompose the risk estimate in Definition 2 as follows:

$$\frac{\mathcal{B}(\bar{\boldsymbol{\beta}}_\star) + \sigma_t^2\Omega}{1-\Omega} = \frac{\sum_{i=1}^{p_s}\lambda_i\zeta_i^2\beta_i^2 + \sum_{i=p_s+1}^{p}\lambda_i\zeta_i^2\beta_i^2 + \sigma_t^2\Omega}{1-\Omega} = \frac{\mathbf{err1} + \mathbf{err2} + \sigma_t^2\Omega}{1-\Omega},$$

where **err1** and **err2** are the contributions to the total risk of the target model from dimensions selected and omitted in the surrogate model, respectively. Therefore, we express the total error as:

$$\frac{\mathbf{err1} + \mathbf{err2} + \sigma_t^2\Omega}{1-\Omega} = \mathbf{err} = \Theta(n^{-\beta+1}).$$

Going back to Proposition 4, we know that, as $p \to \infty$, the criterion for selecting a feature i in the optimal masked surrogate model is given by

$$i > nC_2 + O(1), \quad \text{where} \quad C_2 = \frac{\alpha \sin(\pi/\alpha)}{\pi(\sqrt{\alpha}-1)^{1/\alpha}}.$$

Define now $\omega_n = nC_2 + O(1)$. The equation (16) tells us that after the optimal mask operation $\mathcal{M}$, **err2** is replaced by **err2′**, which is calculated as follows

$$\mathbf{err2}' = \sum_{i=\omega_n+1}^{p}\lambda_i\beta_i^2 = \sum_{i=\omega_n+1}^{p} i^{-\beta}.$$

Since $x^{-\beta}$ is a monotonically decreasing function, we can bound the summation by the following two integrals:

$$\int_{\omega_n+1}^{p+1} x^{-\beta}\,dx \le \sum_{\omega_n+1}^{p} i^{-\beta} \le \int_{\omega_n}^{p} x^{-\beta}\,dx$$

$$\frac{(\omega_n+1)^{-\beta+1} - p^{-\beta+1}}{\beta-1} \le \sum_{\omega_n+1}^{p} i^{-\beta} \le \frac{(\omega_n)^{-\beta+1} - p^{-\beta+1}}{\beta-1}.$$

In the limit of $p \to \infty$, we obtain,

$$\mathbf{err2}' = \Theta(n^{-\beta+1}).$$

Thus, we have tightly estimated **err2′**. Using the fact from Proposition 3 that $\Omega = \Theta(1)$, and our assumption on noise variance $\sigma_t^2 = O(n^{-\beta+1})$, we conclude that the scaling law doesn't change for the surrogate-to-target model as

$$\mathcal{R}(\boldsymbol{\beta}^{s2t}) = \frac{\mathbf{err1} + \sigma_t^2\Omega}{1-\Omega} + \mathbf{err2}' = \Theta(n^{-\beta+1}).$$

Case 2: $\beta > 2\alpha + 1$. In this case, we show that the scaling law is determined by **err1**, hence changing **err2** to **err2′** has no effect in the scaling sense. From Proposition 3, we have the asymptotic expression $\tau_t = cn^{-\alpha}\left(1 + O(n^{-1})\right)$, for $c = \left(\frac{\pi}{\alpha \sin(\pi/\alpha)}\right)^{\alpha}$. We can argue that there exists positive constants $c_1 < \frac{1}{c} < c_2$, such that $c_1 n^{\alpha} \leq \frac{1}{\tau_t} \leq c_2 n^{\alpha}$. We have that

$$\begin{aligned}\textbf{err1} = \sum_{i=1}^{\omega_n} \frac{i^{-\beta}}{(1 + \frac{1}{\tau_t} i^{-\alpha})^2} &\leq \sum_{i=1}^{\omega_n} \frac{i^{-\beta}}{(1 + c_1 n^{\alpha} i^{-\alpha})^2} \\ &= \sum_{i=1}^{\omega_n} \frac{i^{2\alpha-\beta}}{(i^{\alpha} + c_1 n^{\alpha})^2} \leq \sum_{i=1}^{\omega_n} \frac{i^{2\alpha-\beta}}{c_1^2 n^{2\alpha}}.\end{aligned}$$

This implies $\textbf{err1} = O(n^{-2\alpha})$. At the same time,

$$\begin{aligned}\textbf{err1} = \sum_{i=1}^{\omega_n} \frac{i^{-\beta}}{(1 + \frac{1}{\tau_t} i^{-\alpha})^2} &\geq \sum_{i=1}^{\omega_n} \frac{i^{2\alpha-\beta}}{(i^{\alpha} + c_2 n^{\alpha})^2} \\ &\geq \sum_{i=1}^{\omega_n} \frac{i^{2\alpha-\beta}}{((\omega_n)^{\alpha} + c_2 n^{\alpha})^2} = \sum_{i=1}^{\omega_n} \frac{i^{2\alpha-\beta}}{n^{2\alpha}((\omega_n/n)^{\alpha} + c_2)^2}.\end{aligned}$$

Using $\omega_n/n = \Theta(1)$ gives $\textbf{err1} = \Omega(n^{-2\alpha})$ and we can conclude that $\textbf{err1} = \Theta(n^{-2\alpha})$. From Cui et al. (2022), we already know that $\textbf{err} = \Theta(n^{-2\alpha})$ when $\beta > 2\alpha + 1$. Using $\Omega = \Theta(1)$, and our assumption on the noise variance $\sigma_t^2 = O(n^{-2\alpha})$ allows us to conclude that the scaling is dominated by **err1**, and thus, the scaling law remains unchanged. □

Proposition 5 (Scaling law)**.** *Assume that both eigenvalues λ_i and signal coefficients $\lambda_i \beta_i^2$ follow a power-law decay, i.e., $\lambda_i \beta_i^2 = i^{-\beta}$ and $\lambda_i = i^{-\alpha}$ for $\alpha, \beta > 1$. Let the optimal surrogate parameter $\boldsymbol{\beta}^{s*}$ be given by Proposition 1 and define the minimum surrogate-to-target risk attained by $\boldsymbol{\beta}^{s*}$ as $\mathcal{R}^*(\boldsymbol{\beta}^{s2t}) = \min \mathcal{R}(\boldsymbol{\beta}^{s2t})$. Then, in the limit of $p \to \infty$, the excess test risk of the surrogate-to-target model with an optimal surrogate parameter scales the same as that of the standard target model. Specifically, we have*

$$\begin{aligned}\mathcal{R}^*(\boldsymbol{\beta}^{s2t}) &= \Theta(n^{-(\beta-1)}) = \mathcal{R}(\boldsymbol{\beta}^t), \qquad \text{if } \beta < 2\alpha + 1, \\ \mathcal{R}^*(\boldsymbol{\beta}^{s2t}) &= \Theta(n^{-2\alpha}) = \mathcal{R}(\boldsymbol{\beta}^t), \qquad \text{if } \beta > 2\alpha + 1.\end{aligned}$$

Proof. From asymptotic risk decomposition in (26), we can write

$$\begin{aligned}\mathbb{E}_{\boldsymbol{g}_t}\left[f(X^t_{\kappa_t, \sigma_t^2}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}^s, \boldsymbol{g}_t))\right] &= (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star)^\top \boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) + \gamma_t^2(\boldsymbol{\beta}^s)\, \mathbb{E}_{\boldsymbol{g}_t}[\boldsymbol{\theta}_2^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_2] \\ &\quad + \boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t (\boldsymbol{I} - \boldsymbol{\theta}_1) \boldsymbol{\beta}_\star - 2\boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) \\ &\geq \gamma_t^2(\boldsymbol{\beta}^s)\, \mathbb{E}_{\boldsymbol{g}_t}[\boldsymbol{\theta}_2^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_2],\end{aligned}$$

since we can put in the form of $(a-b)^2 + c^2 \geq c^2$. At the same time, we know that

$$\begin{aligned}\gamma_t^2(\boldsymbol{\beta}^s)\, \mathbb{E}_{\boldsymbol{g}_t}[\boldsymbol{\theta}_2^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_2] &= \kappa_t \frac{\sigma_t^2 + \tau_t^2 \|(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1} \boldsymbol{\Sigma}_t^{1/2} \boldsymbol{\beta}_\star\|_2^2}{1 - \frac{1}{n} \mathtt{tr}\left((\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-2} \boldsymbol{\Sigma}_t^2\right)} \frac{\mathtt{tr}\left(\boldsymbol{\Sigma}_t^2 (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-2}\right)}{p} \\ &= \kappa_t \frac{\sigma_t^2 + \tau_t^2 \|(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1} \boldsymbol{\Sigma}_t^{1/2} \boldsymbol{\beta}_\star\|_2^2}{1 - \Omega} \frac{n\Omega}{p} \\ &= \frac{\Omega}{1 - \Omega}\left(\sigma_t^2 + \sum_{i=1}^{p} \lambda_i \beta_i^2 \zeta_i^2\right).\end{aligned}$$

Recall the optimal surrogate vector discussed in Proposition 1 and the corresponding minimal surrogate-to-target risk $\mathcal{R}^*(\boldsymbol{\beta}^{s2t})$. In this case, we can write

$$\sum_{i=1}^{p} \lambda_i \beta_i^{s*2} \zeta_i^2 = \sum_{i=1}^{p} \lambda_i \beta_i^2 \frac{(1-\zeta_i)^2 \zeta_i^2}{\left((1-\zeta_i)^2 + \frac{\Omega}{1-\Omega}\zeta_i^2\right)^2}.$$

Similar to the previous proposition and as discussed in Section 4, to analyze the model's inherent error, we set $\sigma_t^2 = O(n^{-\gamma})$ where γ is the exponent characterizing the scaling law of the test risk in the noiseless setting. It is previously stated by Cui et al. (2022); Simon et al. (2024) that the test risk of ridgeless overparameterized linear regression can be described in the scaling sense as $\mathbf{err} = \Theta(n^{-\beta+1})$ when $\beta < 2\alpha + 1$. We will proceed by considering two cases based on the relationship between β and $2\alpha + 1$.

Case 1: $\beta < 2\alpha + 1$

Consider the interval of i's satisfying $\zeta_i > 1 - \Omega$ and $\zeta_i^2 < 1 - \Omega$. By Proposition 4, we have

$$\zeta_i > 1 - \Omega \iff i > nC_1 + O(1), \quad \text{where} \quad C_1 = \frac{\alpha \sin(\pi/\alpha)}{\pi(\alpha - 1)^{1/\alpha}}.$$

$$\zeta_i^2 > 1 - \Omega \iff i > nC_2 + O(1), \quad \text{where} \quad C_2 = \frac{\alpha \sin(\pi/\alpha)}{\pi(\sqrt{\alpha} - 1)^{1/\alpha}}.$$

Let ω_n be defined as in the previous proposition and define $\phi_n = nC_1 + O(1)$. Then, the interval of interest corresponds to the set of indices i such that $\phi_n < i < \omega_n$. Within this interval, we observe

$$(1-\zeta_i)^2 \zeta_i^2 \geq \min\left((1-\Omega)^2\Omega^2, \left(1 - \sqrt{(1-\Omega)}\right)^2 (1-\Omega)\right) = k_1$$

$$(1-\zeta_i)^2 + \frac{\Omega}{1-\Omega}\zeta_i^2 \leq 1 + \frac{\Omega}{1-\Omega} = k_2$$

Using the fact from Proposition 3 that $\Omega = \dfrac{\alpha - 1}{\alpha} - O(n^{-1})$ tells us $k_1 = \Theta(1)$ and $k_2 = \Theta(1)$. Utilizing these bounds, we obtain

$$\begin{aligned}
\mathcal{R}^*(\boldsymbol{\beta}^{s2t}) \geq \sum_{i=1}^{p} \lambda_i \beta_i^2 \frac{(1-\zeta_i)^2\zeta_i^2}{\left((1-\zeta_i)^2 + \frac{\Omega}{1-\Omega}\zeta_i^2\right)^2} \geq \sum_{i=\phi_n}^{\omega_n} i^{-\beta} \frac{(1-\zeta_i)^2\zeta_i^2}{\left((1-\zeta_i)^2 + \frac{\Omega}{1-\Omega}\zeta_i^2\right)^2} \geq \sum_{i=\phi_n}^{\omega_n} i^{-\beta} \frac{k_1}{k_2} \\
= n^{-\beta+1} \frac{k_1}{k_2} \left((\omega_n/n)^{-\beta+1} - (\phi_n/n)^{-\beta+1}\right) \\
= \Theta(n^{-\beta+1}).
\end{aligned}$$

Recalling $\omega_n/n = \Theta(1)$ and $\phi_n/n = \Theta(1)$, we obtain that $\mathcal{R}^*(\boldsymbol{\beta}^{s2t}) = \Omega(n^{-\beta+1})$, and thus,

$$\mathcal{R}(\boldsymbol{\beta}^{s2t}) \geq \mathcal{R}^*(\boldsymbol{\beta}^{s2t}) \implies \mathcal{R}(\boldsymbol{\beta}^{s2t}) = \Omega(n^{-\beta+1}).$$

Case 2: $\beta > 2\alpha + 1$

In this case, we have

$$\begin{aligned}
\mathcal{R}^*(\boldsymbol{\beta}^{s2t}) \geq \sum_{i=1}^{p} \lambda_i \beta_i^2 \frac{(1-\zeta_i)^2\zeta_i^2}{\left((1-\zeta_i)^2 + \frac{\Omega}{1-\Omega}\zeta_i^2\right)^2} = \sum_{i=1}^{p} \lambda_i \beta_i^2 \zeta_i^2 \frac{(1-\zeta_i)^2}{\left((1-\zeta_i)^2 + \frac{\Omega}{1-\Omega}\zeta_i^2\right)^2} \\
\geq \sum_{i:\zeta_i < 1-\Omega} \lambda_i \zeta_i^2 \beta_i^2 \frac{\Omega^2}{\left(1 + \frac{\Omega}{1-\Omega}\right)^2} = \sum_{i=1}^{\phi_n} \frac{i^{-\beta}}{(1 + \frac{1}{\tau_t} i^{-\alpha})^2} k_3,
\end{aligned}$$

where $k_3 = \frac{\Omega^2}{\left(1+\frac{\Omega}{1-\Omega}\right)^2} = \Theta(1)$. From Case 2 in Proposition 6, we already know that the same summation – with upper bound ω_n rather than ϕ_n – scales as $\Theta(n^{-2\alpha})$. Yet, since ϕ_n and ω_n have the same order $\Theta(n)$, the result remains. This suggests $\mathcal{R}^*(\boldsymbol{\beta}^{s2t}) = \Omega(n^{-2\alpha})$, which eventually yields

$$\mathcal{R}(\boldsymbol{\beta}^{s2t}) \geq \mathcal{R}^*(\boldsymbol{\beta}^{s2t}) \implies \mathcal{R}(\boldsymbol{\beta}^{s2t}) = \Omega(n^{-2\alpha}).$$

Hence, this allows us to say that the scaling law doesn't improve even with the freedom to choose any $\boldsymbol{\beta}^s$. □

Proposition 7 (Non-asymptotic analysis of τ)**.** *Suppose that* $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ *is diagonal and* $\mathbf{\Sigma}_{i,i} = \lambda_i = i^{-\alpha}$ *for* $1 < \alpha$*. Assume that* $n < pk$ *for* $k = \dfrac{3 + \dfrac{1}{2^\alpha}}{4 + \dfrac{1}{2^{\alpha-2}}}$*. If* ξ *satisfies*

$$\sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + \frac{1}{\xi}} = n,$$

then $cn^\alpha \le \xi \le c\left(n + 1 + \frac{p+1}{\alpha-1}\right)^\alpha$ *for* $c = \left(\dfrac{\alpha \sin(\pi/\alpha)}{\pi}\right)^\alpha$*. Note that* ξ *is defined for the sake of the analysis, and it corresponds to* $\frac{1}{\tau_t}$*.*

Proof. From Simon et al. (2024), we have:

$$n = \sum_{i=1}^{p} \frac{i^{-\alpha}}{i^{-\alpha} + \frac{1}{\xi}} \le \int_0^p \frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\, dx = \int_0^\infty \frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\, dx - \int_p^\infty \frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\, dx = \frac{\pi}{\alpha \sin(\pi/\alpha)} \xi^{1/\alpha} - \int_p^\infty \frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\, dx.$$

Using $1 + x^\alpha \le (1 + x)^\alpha$ for $x \ge 0$,

$$\int_p^\infty \frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\, dx = \int_p^\infty \frac{1}{1 + \frac{x^\alpha}{\xi}}\, dx \ge \int_p^\infty \frac{1}{(1 + \frac{x}{\xi^{1/\alpha}})^\alpha}\, dx = \frac{\xi(\xi^{1/\alpha} + p)^{-\alpha+1}}{\alpha - 1},$$

which implies that

$$n \le \frac{\pi}{\alpha \sin(\pi/\alpha)} \xi^{1/\alpha} - \frac{\xi(\xi^{1/\alpha} + p)^{-\alpha+1}}{\alpha - 1}.$$

Since the summand is decreasing, we can bound the Riemann sum by an integral, thus:

$$\begin{aligned}
n = \sum_{i=1}^{p} \frac{i^{-\alpha}}{i^{-\alpha} + \frac{1}{\xi}} &\ge \int_1^{p+1} \frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\, dx \\
&= \int_0^\infty \frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\, dx - \int_0^1 \frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\, dx - \int_{p+1}^\infty \frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\, dx \\
&= \frac{\pi}{\alpha \sin(\pi/\alpha)} \xi^{1/\alpha} - \int_0^1 \frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\, dx - \int_{p+1}^\infty \frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\, dx \\
&\ge \frac{\pi}{\alpha \sin(\pi/\alpha)} \xi^{1/\alpha} - 1 - \int_{p+1}^\infty \frac{1}{1 + \frac{1}{\xi} x^\alpha}\, dx \\
&\ge \frac{\pi}{\alpha \sin(\pi/\alpha)} \xi^{1/\alpha} - 1 - \int_{p+1}^\infty \frac{1}{\frac{1}{\xi} x^\alpha}\, dx \\
&= \frac{\pi}{\alpha \sin(\pi/\alpha)} \xi^{1/\alpha} - 1 - \left[\frac{\xi x^{-\alpha+1}}{-\alpha + 1}\right]_{p+1}^\infty \\
&= \frac{\pi}{\alpha \sin(\pi/\alpha)} \xi^{1/\alpha} - \frac{\xi(p+1)^{-\alpha+1}}{\alpha - 1} - 1.
\end{aligned}$$

Recalling that $\alpha > 1$ and assuming $\xi < p^\alpha$, we derive:

$$\begin{aligned}
&\frac{\pi}{\alpha \sin(\pi/\alpha)} \xi^{1/\alpha} - 1 - \frac{p+1}{\alpha - 1} \le n \le \frac{\pi}{\alpha \sin(\pi/\alpha)} \xi^{1/\alpha} \\
\iff &\left(\frac{n\alpha \sin(\pi/\alpha)}{\pi}\right)^\alpha \le \xi \le \left(\frac{\left(n + 1 + \frac{p+1}{\alpha-1}\right) \alpha \sin(\pi/\alpha)}{\pi}\right)^\alpha.
\end{aligned}$$

We conclude by proving that $\xi < p^\alpha$. For the sake of contradiction, assume that $\xi \geq p^\alpha$. Then,

$$\begin{aligned} n = \sum_{i=1}^{p} \frac{1}{1 + \frac{i^\alpha}{\xi}} &= \sum_{i=1}^{p/2} \frac{1}{1 + \frac{i^\alpha}{\xi}} + \sum_{i=p/2+1}^{p} \frac{1}{1 + \frac{i^\alpha}{\xi}} \\ &\geq \sum_{i=1}^{p/2} \frac{1}{1 + \frac{1}{2^\alpha}} + \sum_{i=p/2+1}^{p} \frac{1}{1+1} \\ &= p\left(\frac{3 + \frac{1}{2^\alpha}}{4 + \frac{1}{2^{\alpha-2}}}\right), \end{aligned}$$

which contradicts our assumption that $n < pk$. □

Proposition 8 (Non-asymptotic analysis of Ω)**.** *Suppose that $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ is diagonal and $\mathbf{\Sigma}_{i,i} = \lambda_i = i^{-\alpha}$ for $1 < \alpha$. Let τ_t be defined as in Proposition 7 and Ω be the solution to*

$$n\Omega = \sum_{i=1}^{p} \left(\frac{\lambda_i}{\lambda_i + \frac{1}{\xi}}\right)^2.$$

Then,

$$\Omega > \frac{\alpha - 1}{\alpha} - \frac{1}{2\alpha - 1}\left(\frac{n + 1 + \frac{p+1}{\alpha-1}}{p+1}\right)^{2\alpha-1} - \frac{1}{n}.$$

Proof. We have that

$$\sum_{i=1}^{p} \left(\frac{i^{-\alpha}}{i^{-\alpha} + \frac{1}{\xi}}\right)^2 \leq \int_0^\infty \left(\frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\right)^2 dx = \frac{\pi(\alpha-1)}{\alpha^2 \sin(\pi/\alpha)} \xi^{1/\alpha}$$

Besides, since the summand is monotonically decreasing:

$$\begin{aligned} n\Omega = \sum_{i=1}^{p} \left(\frac{i^{-\alpha}}{i^{-\alpha} + \frac{1}{\xi}}\right)^2 &\geq \int_1^{p+1} \left(\frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\right)^2 dx \\ &= \int_0^\infty \left(\frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\right)^2 dx - \int_0^1 \left(\frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\right)^2 dx - \int_{p+1}^\infty \left(\frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\right)^2 dx \\ &= \frac{\pi(\alpha-1)}{\alpha^2 \sin(\pi/\alpha)} \xi^{1/\alpha} - \int_0^1 \left(\frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\right)^2 dx - \int_{p+1}^\infty \left(\frac{x^{-\alpha}}{x^{-\alpha} + \frac{1}{\xi}}\right)^2 dx \\ &\geq \frac{\pi(\alpha-1)}{\alpha^2 \sin(\pi/\alpha)} \xi^{1/\alpha} - 1 - \int_{p+1}^\infty \left(\frac{1}{1 + \frac{1}{\xi}x^\alpha}\right)^2 dx \\ &\geq \frac{\pi(\alpha-1)}{\alpha^2 \sin(\pi/\alpha)} \xi^{1/\alpha} - 1 - \int_{p+1}^\infty \frac{1}{(\frac{1}{\xi}x^\alpha)^2} dx \\ &= \frac{\pi(\alpha-1)}{\alpha^2 \sin(\pi/\alpha)} \xi^{1/\alpha} - 1 - \left[\frac{\xi^2 x^{-2\alpha+1}}{-2\alpha+1}\right]_{p+1}^\infty \\ &= \frac{\pi(\alpha-1)}{\alpha^2 \sin(\pi/\alpha)} \xi^{1/\alpha} - \frac{\xi^2 (p+1)^{-2\alpha+1}}{2\alpha - 1} - 1. \end{aligned}$$

Let's now utilize the upper and lower bounds for ξ from Proposition 7. Then, we have

$$\begin{aligned} n\Omega &\geq \frac{\pi(\alpha-1)}{\alpha^2 \sin(\pi/\alpha)} \frac{n\alpha \sin(\pi/\alpha)}{\pi} - \frac{\xi^2 (p+1)^{-2\alpha+1}}{2\alpha-1} - 1 \\ &= \frac{n(\alpha-1)}{\alpha} - \frac{\xi^2 (p+1)^{-2\alpha+1}}{2\alpha-1} - 1 \\ &\geq \frac{n(\alpha-1)}{\alpha} - \left(\frac{\left(n+1+\frac{p+1}{\alpha-1}\right)\alpha \sin(\pi/\alpha)}{\pi(p+1)} \right)^{2\alpha} \frac{p+1}{2\alpha-1} - 1 \\ \implies \Omega &> \frac{\alpha-1}{\alpha} - \left(\frac{\left(n+1+\frac{p+1}{\alpha-1}\right)\alpha \sin(\pi/\alpha)}{\pi(p+1)} \right)^{2\alpha} \frac{p+1}{n(2\alpha-1)} - \frac{1}{n} \\ &> \frac{\alpha-1}{\alpha} - \frac{1}{2\alpha-1} \left(\frac{n+1+\frac{p+1}{\alpha-1}}{p+1} \right)^{2\alpha-1} - \frac{1}{n}, \end{aligned}$$

since $\dfrac{\alpha \sin(\pi/\alpha)}{\pi} < 1$ for $\alpha > 1$. □

Proposition 9. *Under the assumption that* $n < \min\left((p+1)\dfrac{\alpha-2}{\alpha}, p\left(\dfrac{3+\frac{1}{2^\alpha}}{4+\frac{1}{2^{\alpha-2}}}\right), p\dfrac{\pi\left(\sqrt{\frac{\alpha}{2}}-1\right)^{1/\alpha}}{\alpha \sin(\pi/\alpha)} - \dfrac{p+1}{\alpha-1} \right) - 1$ *and* $\alpha > 3$*, we can find a masked surrogate-to-target setting that improves over the risk of the standard target model by selecting all features* i *such that* $\zeta_i^2 > 1 - \Omega$*.*

Proof. From Proposition 8, we have

$$\Omega > \frac{\alpha-1}{\alpha} - \frac{1}{2\alpha-1} \left(\frac{n+1+\frac{p+1}{\alpha-1}}{p+1} \right)^{2\alpha-1} - \frac{1}{n}.$$

It's then enough to show that we can find a set of i's such that

$$\zeta_i^2 > \frac{1}{\alpha} + \frac{1}{2\alpha-1} \left(\frac{n+1+\frac{p+1}{\alpha-1}}{p+1} \right)^{2\alpha-1} + \frac{1}{n}.$$

From proof of Proposition 3, we know that

$$\zeta_i^2 > c' \iff i > \tau_t^{-1/\alpha} \left(\frac{\sqrt{c'}}{1-\sqrt{c'}} \right)^{1/\alpha}.$$

Hence, using the bound on $\frac{1}{\tau_t} = \xi$ from Proposition 7, it's enough to find indices i such that

$$i > \frac{\alpha \sin(\pi/\alpha)}{\pi} \left(n+1+\frac{p+1}{\alpha-1} \right) \left(\frac{\sqrt{c'}}{1-\sqrt{c'}} \right)^{1/\alpha} \quad \text{where } c' = \frac{1}{\alpha} + \frac{1}{2\alpha-1} \left(\frac{n+1+\frac{p+1}{\alpha-1}}{p+1} \right)^{2\alpha-1} + \frac{1}{n}. \tag{21}$$

By our assumption $p+1 > n+1+\frac{p+1}{\alpha-1}$, we obtain that $\frac{2}{\alpha} > c'$. Since $\left(\frac{\sqrt{x}}{1-\sqrt{x}}\right)^{1/\alpha}$ in increasing with x when $0 \leq x \leq 1$, we have

$$\left(\frac{1}{\sqrt{\frac{\alpha}{2}}-1} \right)^{1/\alpha} \geq \left(\frac{\sqrt{c'}}{1-\sqrt{c'}} \right)^{1/\alpha}.$$

Then, to ensure the existence of an interval of i's satisfying the above inequality, we choose

$$\begin{aligned} & p - (p+1)\frac{\alpha \sin(\pi/\alpha)}{\pi(\alpha-1)} \left(\frac{1}{\sqrt{\frac{\alpha}{2}}-1} \right)^{1/\alpha} \geq (n+1)\frac{\alpha \sin(\pi/\alpha)}{\pi} \left(\frac{1}{\sqrt{\frac{\alpha}{2}}-1} \right)^{1/\alpha} \\ \iff & p\frac{\pi\left(\sqrt{\frac{\alpha}{2}}-1\right)^{1/\alpha}}{\alpha \sin(\pi/\alpha)} - \frac{p+1}{\alpha-1} \geq n+1 \end{aligned}$$

One can verify that the LHS expression is always positive when $\alpha > 3$. Thus, discarding the features i provided in the interval (21) will strictly improve the test risk of the masked surrogate-to-target model over the standard target model. □

C Proofs for Section 5

Theorem 3 (Distributional characterization, Han & Xu (2023)). *Let $\kappa_s = p/m > 1$ and suppose that, for some $M > 1$, $1/M \le \kappa_s, \sigma_s^2 \le M$ and $\|\mathbf{\Sigma}_s\|_{op}, \left\|\mathbf{\Sigma}_s^{-1}\right\|_{op} \le M$. Let $\tau_s \in \mathbb{R}$ be the unique solution of the following equation:*

$$\kappa_s^{-1} = \frac{1}{p} tr\left((\mathbf{\Sigma}_s + \tau_s \boldsymbol{I})^{-1}\mathbf{\Sigma}_s\right). \tag{22}$$

We define the function $\gamma_s : \mathbb{R}^p \to \mathbb{R}$ and the random variable based on $\boldsymbol{g}_s \sim \mathcal{N}(0, \boldsymbol{I})$ as follows:

$$\begin{aligned}\gamma_s^2(\boldsymbol{\beta}_\star) &:= \kappa_s\left(\sigma_s^2 + \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_s}[\|\mathbf{\Sigma}_s^{1/2}(\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star)\|_2^2]\right) = \kappa_s \frac{\sigma_s^2 + \tau_s^2\|(\mathbf{\Sigma}_s + \tau_s\boldsymbol{I})^{-1}\mathbf{\Sigma}_s^{1/2}\boldsymbol{\beta}_\star\|_2^2}{1 - \frac{1}{m} tr((\mathbf{\Sigma}_s + \tau_s\boldsymbol{I})^{-2}\mathbf{\Sigma}_s^2)} \\ X^s_{\kappa_s,\sigma_s^2}(\mathbf{\Sigma}_s, \boldsymbol{\beta}_\star, \boldsymbol{g}_s) &:= (\mathbf{\Sigma}_s + \tau_s\boldsymbol{I})^{-1}\mathbf{\Sigma}_s\left[\boldsymbol{\beta}_\star + \frac{\mathbf{\Sigma}_s^{-1/2}\gamma_s(\boldsymbol{\beta}_\star)\boldsymbol{g}_s}{\sqrt{p}}\right].\end{aligned} \tag{23}$$

Then, for any L-Lipschitz function $f : \mathbb{R}^p \to \mathbb{R}$ where $L < L(M)$, there exists a constant $C = C(M)$ such that for any $\varepsilon \in (0, 1/2]$, we have the following:

$$\mathbb{P}\Big(\sup_{\boldsymbol{\beta}_\star \in \boldsymbol{B}(R)} \left|f(\boldsymbol{\beta}^s) - \mathbb{E}_{\boldsymbol{g}_s}[f(X^w_{\kappa_s,\sigma_s^2}(\mathbf{\Sigma}_s, \boldsymbol{\beta}_\star, \boldsymbol{g}_s))]\right| \ge \varepsilon\Big) \le Cpe^{-p\varepsilon^4/C}, \tag{24}$$

where $R < M$.

Definition 3. *Recall the definition of τ_t and γ_t in Theorem 1. Let $\kappa_s = p/m > 1$ and define $\tau_s \in \mathbb{R}$ similar to τ_t. We define the function $\gamma_s : \mathbb{R}^p \to \mathbb{R}$ and the random variable $X^s_{\kappa_s,\sigma_s^2}$ based on $\boldsymbol{g}_s \sim \mathcal{N}(0, \boldsymbol{I})$ as follows:*

$$\gamma_s^2(\boldsymbol{\beta}_\star) = \kappa_s\left(\sigma_s^2 + \mathbb{E}_{(\tilde{\boldsymbol{x}},\tilde{y})\sim\mathcal{D}_s}[\|\mathbf{\Sigma}_s^{1/2}(\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star)\|_2^2]\right), \quad X^s_{\kappa_s,\sigma_s^2}(\mathbf{\Sigma}_s, \boldsymbol{\beta}_\star, \boldsymbol{g}_s) := (\mathbf{\Sigma}_s + \tau_s\boldsymbol{I})^{-1}\mathbf{\Sigma}_s\left[\boldsymbol{\beta}_\star + \frac{\mathbf{\Sigma}_s^{-1/2}\gamma_s(\boldsymbol{\beta}_\star)\boldsymbol{g}_s}{\sqrt{p}}\right].$$

Let $\dot{\kappa} = (\kappa_s, \kappa_t)$, $\dot{\mathbf{\Sigma}} = (\mathbf{\Sigma}_s, \mathbf{\Sigma}_t)$, and $\dot{\sigma} = (\sigma_s^2, \sigma_t^2)$. Then, we define the asymptotic risk estimate as

$$\begin{aligned}\bar{\mathcal{R}}_{\dot{\kappa},\dot{\sigma}}(\dot{\mathbf{\Sigma}}, \boldsymbol{\beta}_\star) &= \|\mathbf{\Sigma}_t^{1/2}\left(\boldsymbol{I} - (\mathbf{\Sigma}_t + \tau_t\boldsymbol{I})^{-1}\mathbf{\Sigma}_t(\mathbf{\Sigma}_s + \tau_s\boldsymbol{I})^{-1}\mathbf{\Sigma}_s\right)\boldsymbol{\beta}_\star\|_2^2 + \frac{\mathbb{E}_{\boldsymbol{\beta}^s\sim X^s_{\kappa_s,\sigma_s^2}}[\gamma_t^2(\boldsymbol{\beta}^s)]}{p} tr\left(\mathbf{\Sigma}_t^2(\mathbf{\Sigma}_t + \tau_t\boldsymbol{I})^{-2}\right) \\ &+ \frac{\gamma_s^2(\boldsymbol{\beta}_\star)}{p} tr\left(\mathbf{\Sigma}_s^{1/2}(\mathbf{\Sigma}_s + \tau_s\boldsymbol{I})^{-1}\mathbf{\Sigma}_t(\mathbf{\Sigma}_t + \tau_t\boldsymbol{I})^{-1}\mathbf{\Sigma}_t(\mathbf{\Sigma}_t + \tau_t\boldsymbol{I})^{-1}\mathbf{\Sigma}_t(\mathbf{\Sigma}_s + \tau_s\boldsymbol{I})^{-1}\mathbf{\Sigma}_s^{1/2}\right).\end{aligned}$$

Theorem 2. *Suppose that, for some constant $M_t > 1$, we have $1/M_t \le \kappa_s, \sigma_s^2, \kappa_t, \sigma_t^2 \le M_t$ and $\|\mathbf{\Sigma}_s\|_{op}, \left\|\mathbf{\Sigma}_s^{-1}\right\|_{op}, \|\mathbf{\Sigma}_t\|_{op}, \left\|\mathbf{\Sigma}_s^{-1}\right\|_{op} \le M_t$. Consider the surrogate-to-target model defined in Section 2, and let $\mathcal{R}(\boldsymbol{\beta}^{s2t})$ represent its risk when $\boldsymbol{\beta}_\star$ is given. Recall the definition of $\dot{\mathbf{\Sigma}}, \dot{\kappa}, \dot{\sigma}$ and $\bar{\mathcal{R}}_{\dot{\kappa},\dot{\sigma}}$ in Definition 3. Then, there exists a constant $C = C(M_t)$ such that for any $\varepsilon \in (0, 1/2]$, the following holds when $R + 1 < M$:*

$$\sup_{\boldsymbol{\beta}_\star \in \boldsymbol{B}_p(R)} \mathbb{P}\left(\left|\mathcal{R}(\boldsymbol{\beta}^{s2t}) - \bar{\mathcal{R}}_{\dot{\kappa},\dot{\sigma}}(\dot{\mathbf{\Sigma}}, \boldsymbol{\beta}_\star)\right| \ge \varepsilon\right) \le Cpe^{-p\varepsilon^4/C}.$$

Proof. Define a function $f_1 : \mathbb{R}^p \to \mathbb{R}$ as $f_1(\boldsymbol{x}) = \|\mathbf{\Sigma}_t^{1/2}(\boldsymbol{x} - \boldsymbol{\beta}_\star)\|_2^2$. The gradient of this function is

$$\|\nabla f_1(\boldsymbol{x})\|_2 = \|2\mathbf{\Sigma}_t(\boldsymbol{x} - \boldsymbol{\beta}_\star)\|_2 \le 2\|\mathbf{\Sigma}_t\|_{\text{op}}\|\boldsymbol{x} - \boldsymbol{\beta}_\star\|_2.$$

Using Proposition 11, there exists an event E with $\mathbb{P}(E^c) \le C_t e^{-p/C_t}$ where $C_t = C_t(M_t, \frac{M_t - R}{2})$ with the definition of M_t in Proposition 11, such that $f_1(\boldsymbol{\beta}^{s2t})$ is $2M_t^2$-Lipschitz if $\boldsymbol{\beta}_\star \in \boldsymbol{B}_p(R)$. Applying Theorem 3 on the target model, there exists a constant $\bar{C}_s = \bar{C}_s(M_t)$ such that for any $\varepsilon \in (0, 1/2]$, we obtain

$$\sup_{\boldsymbol{\beta}^s \in \boldsymbol{B}(\frac{M_t+R}{2})} \mathbb{P}\left(\left|f(\boldsymbol{\beta}^{s2t}) - \mathbb{E}_{\boldsymbol{g}_t}[f(X^t_{\kappa_t,\sigma_t^2}(\mathbf{\Sigma}_t, \boldsymbol{\beta}^s, \boldsymbol{g}_t))]\right| \ge \varepsilon\right) \le Cpe^{-p\varepsilon^4/C}, \tag{25}$$

where $f(\boldsymbol{\beta}^{s2t}) = \mathcal{R}(\boldsymbol{\beta}^{s2t})$ and

$$X^t_{\kappa_t,\sigma_t^2}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}^s, \boldsymbol{g}_t) = (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t \left[\boldsymbol{\beta}^s + \frac{\boldsymbol{\Sigma}_t^{-1/2}\gamma_t(\boldsymbol{\beta}^s)\boldsymbol{g}_t}{\sqrt{p}}\right].$$

Furthermore,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{g}_t}\left[f(X^s_{\kappa_t,\sigma_t^2}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}^s, \boldsymbol{g}_t))\right] &= \mathbb{E}_{\boldsymbol{g}_t}\left[\|\boldsymbol{\Sigma}_t^{1/2}\left(\boldsymbol{\theta}_1(\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) - (\boldsymbol{I} - \boldsymbol{\theta}_1)\boldsymbol{\beta}_\star + \boldsymbol{\theta}_2\gamma_t(\boldsymbol{\beta}^s)\right)\|_2^2\right] \\ &= (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star)^\top \boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) + \gamma_t^2(\boldsymbol{\beta}^s)\,\mathbb{E}_{\boldsymbol{g}_t}[\boldsymbol{\theta}_2^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_2] \\ &\quad + \boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t (\boldsymbol{I} - \boldsymbol{\theta}_1)\boldsymbol{\beta}_\star - 2\boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star), \end{aligned} \tag{26}$$

where $\boldsymbol{\theta}_1 := (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t$ and $\boldsymbol{\theta}_2 := (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}\frac{\boldsymbol{g}_t}{\sqrt{p}}$. Let $E(M_t, \frac{M_t - R}{2})$ be the event defined in Proposition 10. Let $f_2 : \mathbb{R}^p \to \mathbb{R}$ be defined as $f_2(\boldsymbol{x}) := (\boldsymbol{x} - \boldsymbol{\beta}_\star)^\top \boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{x} - \boldsymbol{\beta}_\star)$. By Proposition 12, the function f_2 is $2M_t^2$–Lipschitz if $\boldsymbol{\beta}_\star \in \boldsymbol{B}_p(R)$ on the event $E(M_t, \frac{M_t - R}{2})$. Applying Theorem 3 on the surrogate model, there exists a constant $\bar{C}_{w,1} = \bar{C}_{w,1}(M_t)$ such that for any $\varepsilon \in (0, 1/2]$, we obtain

$$\sup_{\boldsymbol{\beta}_\star \in \boldsymbol{B}_p(R)} \mathbb{P}\left(\left|f_2(\boldsymbol{\beta}^s) - \boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\Phi}_1)^\top \boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{I} - \boldsymbol{\Phi}_1)\boldsymbol{\beta}_\star - \gamma_s^2(\boldsymbol{\beta}_\star)\,\mathbb{E}_{\boldsymbol{g}_s}[\boldsymbol{\Phi}_2^\top \boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 \boldsymbol{\Phi}_2]\right| > \varepsilon\right) \le \bar{C}_{w,1} p e^{-p\varepsilon^4/\bar{C}_{w,1}}, \tag{27}$$

where $\boldsymbol{\Phi}_1 := (\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_s$ and $\boldsymbol{\Phi}_2 := (\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_s^{1/2}\frac{\boldsymbol{g}_s}{\sqrt{p}}$.

Let $f_3 : \mathbb{R}^p \to \mathbb{R}$ be defined as $f_3(\boldsymbol{x}) := \gamma_t^2(\boldsymbol{x})\boldsymbol{\theta}_2^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_2$. By Proposition 13 and Proposition 2.1 in Han & Xu (2023), the function f_3 is $4M_t^2$-Lipschitz if $\boldsymbol{\beta}_\star \in \boldsymbol{B}_p(R)$ on the event $E(M_t, \frac{M_t - R}{2})$. Applying Theorem 3 on the surrogate model, there exists a constant $\bar{C}_{w,2} = \bar{C}_{w,2}(M_t)$ such that for any $\varepsilon \in (0, 1/2]$, we obtain

$$\sup_{\boldsymbol{\beta}_\star \in \boldsymbol{B}_p(R)} \mathbb{P}\left(\left|f_3(\boldsymbol{\beta}^s) - \mathbb{E}_{\boldsymbol{\beta}^s \sim X^s}[\gamma_t^2(\boldsymbol{\beta}^s)]\,\mathbb{E}_{\boldsymbol{g}_t}[\boldsymbol{\theta}_2^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_2]\right| > \varepsilon\right) \le \bar{C}_{w,2} p e^{-p\varepsilon^4/\bar{C}_{w,2}}. \tag{28}$$

Let $f_4 : \mathbb{R}^p \to \mathbb{R}$ as $f_4(\boldsymbol{x}) := -2\boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{x} - \boldsymbol{\beta}_\star)$. By Proposition 14 and Proposition 2.1 in Han & Xu (2023), the function f_4 is $2M_t^2$–Lipschitz if $\boldsymbol{\beta}_\star \in \boldsymbol{B}_p(R)$ on the event $E(M_t, \frac{M_t - R}{2})$. Applying Theorem 3 on the surrogate model, there exists a constant $\bar{C}_{w,3} = \bar{C}_{w,3}(M_t)$ such that for any $\varepsilon \in (0, 1/2]$, we obtain

$$\sup_{\boldsymbol{\beta}_\star \in \boldsymbol{B}_p(R)} \mathbb{P}\left(\left|f_4(\boldsymbol{\beta}^s) - 2\left[\boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{\Phi}_1 - \boldsymbol{I})\boldsymbol{\beta}_\star\right]\right| > \varepsilon\right) \le \bar{C}_{w,3} p e^{-p\varepsilon^4/\bar{C}_{w,3}}. \tag{29}$$

By the definition of these functions, we have

$$\mathbb{E}_{\boldsymbol{g}_t}\left[f(X^s_{\kappa_t,\sigma_t^2}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}^s, \boldsymbol{g}_t))\right] - \boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t (\boldsymbol{I} - \boldsymbol{\theta}_1)\boldsymbol{\beta}_\star = f_2(\boldsymbol{\beta}^s) + f_3(\boldsymbol{\beta}^s) - f_4(\boldsymbol{\beta}^s) \tag{30}$$

By the definition of $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, $\boldsymbol{\Phi}_1$, and $\boldsymbol{\Phi}_2$, we have

$$\begin{aligned}\bar{\mathcal{R}}_{k,\dot{\sigma}}(\dot{\boldsymbol{\Sigma}}, \boldsymbol{\beta}_\star) - \boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t (\boldsymbol{I} - \boldsymbol{\theta}_1)\boldsymbol{\beta}_\star &= \boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\Phi}_1)^\top \boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{I} - \boldsymbol{\Phi}_1)\boldsymbol{\beta}_\star + \gamma_s^2(\boldsymbol{\beta}_\star)\,\mathbb{E}_{\boldsymbol{g}_s}[\boldsymbol{\Phi}_2^\top \boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 \boldsymbol{\Phi}_2] \\ &\quad + \mathbb{E}_{\boldsymbol{\beta}^s \sim X^s}[\gamma_t^2(\boldsymbol{\beta}^s)]\,\mathbb{E}_{\boldsymbol{g}_t}[\boldsymbol{\theta}_2^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_2] - 2\left[\boldsymbol{\beta}_\star{}^\top (\boldsymbol{I} - \boldsymbol{\theta}_1)^\top \boldsymbol{\Sigma}_t \boldsymbol{\theta}_1 (\boldsymbol{\Phi}_1 - \boldsymbol{I})\boldsymbol{\beta}_\star\right]. \end{aligned} \tag{31}$$

Using (30)-(31) and applying a union bound on (25), (27), (28), and (29), we obtain the advertised claim. □

Proposition 10. *Suppose that, for some $M_t > 1$, $1/M_t \le \kappa_s, \sigma_s^2 \le M_t$ and $\|\boldsymbol{\Sigma}_s\|_{op}, \left\|\boldsymbol{\Sigma}_s^{-1}\right\|_{op} \le M_t$. For every $c_s > 0$, there exists an event $E(M_t, c_s)$ with $\mathbb{P}((E(M_t, c_s))^c) \le C_s \mathrm{e}^{-p/C_s}$ where $C_s = C_s(M_t, c_s)$ such that*

$$\|\boldsymbol{\beta}^s\|_2 \le \|\boldsymbol{\beta}_\star\|_2 + c_s \quad \textit{and} \quad \|\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star\|_2 \le \|\boldsymbol{\beta}_\star\|_2 + c_s.$$

Proof. By the definition of $\boldsymbol{\beta}^s$, we have

$$\begin{aligned}\boldsymbol{\beta}^s &= \tilde{\boldsymbol{X}}^\top (\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{y}} \\ &= \tilde{\boldsymbol{X}}^\top (\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{X}}\boldsymbol{\beta}_\star + \tilde{\boldsymbol{X}}^\top (\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{z}}, \end{aligned} \tag{32}$$

where $\tilde{z} \sim \mathcal{N}(\mathbf{0}, \sigma_s^2 \boldsymbol{I})$. By triangle inequality, we obtain

$$\begin{aligned}\|\boldsymbol{\beta}^s\|_2 &\leq \|\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{X}}\boldsymbol{\beta}_\star\|_2 + \|\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{z}\|_2 \\ &\overset{(a)}{\leq} \|\boldsymbol{\beta}_\star\|_2 + \|\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{z}\|_2, \end{aligned} \tag{33}$$

where (a) in above follows from the fact that $\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{X}}$ is a projection matrix, and so all of its eigenvalues are either 0 or 1. Focusing on the second term of the RHS, we derive

$$\begin{aligned}\|\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{z}\|_2^2 = \tilde{z}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{z} &= \frac{\tilde{z}^\top}{\sqrt{p}}\left(\frac{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top}{p}\right)^{-1}\frac{\tilde{z}}{\sqrt{p}} \\ &\overset{(a)}{\leq} \frac{\tilde{z}^\top\tilde{z}}{p}\left\|\left(\frac{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top}{p}\right)^{-1}\right\|_{\text{op}}, \end{aligned} \tag{34}$$

where (a) in the above inequality follows from Cauchy-Schwarz inequality. Using Bernstein's inequality, there exists an absolute constant $C_0 > 0$ that depends on σ_s^2 such that

$$\mathbb{P}\left(\frac{\tilde{z}^\top\tilde{z}}{p} - \sigma_s^2 > t\right) \leq \exp\left\{-c\min\left\{\frac{pt^2}{4C_0^2}, \frac{pt}{2C_0}\right\}\right\}.$$

On the other hand, let $\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{X}}\boldsymbol{\Sigma}_s^{-1/2}$, which means that the entries of $\tilde{\boldsymbol{Z}}$ are independent and normally distributed with zero mean and unit variance. Then,

$$\left\|\left(\frac{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top}{p}\right)^{-1}\right\|_{\text{op}} = \left\|\left(\frac{\tilde{\boldsymbol{Z}}\boldsymbol{\Sigma}_s\tilde{\boldsymbol{Z}}^\top}{p}\right)^{-1}\right\|_{\text{op}} \leq \left\|\boldsymbol{\Sigma}_s^{-1}\right\|_{\text{op}}\left\|\left(\frac{\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^\top}{p}\right)^{-1}\right\|_{\text{op}}. \tag{35}$$

Using Theorem 1.1 in Rudelson & Vershynin (2009), there exist absolute constants $C_1, C_2 > 0$ such that we have the following for every $\varepsilon > 0$

$$\mathbb{P}\left(\left\|\left(\frac{\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^\top}{p}\right)^{-1}\right\|_{\text{op}} \leq \varepsilon^2\left(1 - \frac{1}{\kappa_s}\right)^2\right) \leq (C_1\varepsilon)^{p-m+1} + e^{-pC_2}. \tag{36}$$

By combining (34), (35), and (36), we obtain that

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{\beta}^s\|_2 \leq \|\boldsymbol{\beta}_\star\|_2 + \varepsilon(1 - \frac{1}{\kappa_s})\sqrt{(t + \sigma_s^2)\left\|\boldsymbol{\Sigma}_s^{-1}\right\|_{\text{op}}}\right) & \\ \leq (C_1\varepsilon)^{p-m+1} + e^{-pC_2} + e^{-c\min\left\{\frac{pt^2}{4C_0^2}, \frac{pt}{2C_0}\right\}}& \end{aligned}$$

The advertised claim for $\|\boldsymbol{\beta}^s\|_2$ follows when ε is selected as $\varepsilon < \frac{1}{C_1 e}$. For $\|\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star\|_2$, using the definition of $\boldsymbol{\beta}^s$, we write as follows:

$$\begin{aligned}\|\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star\|_2 &= \|\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{X}}\boldsymbol{\beta}_\star + \tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{z} - \boldsymbol{\beta}_\star\|_2 \\ &\leq \|\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{X}} - \boldsymbol{I}\|_2\|\boldsymbol{\beta}_\star\|_2 + \|\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{z}\|_2 \\ &\overset{(a)}{\leq} \|\boldsymbol{\beta}_\star\|_2 + \|\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{z}\|_2, \end{aligned} \tag{37}$$

where (a) in the above inequalities follows from the fact that the eigenvalues of $\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{X}} - \boldsymbol{I}$ are either 1 or 0 as the eigenvalues of $\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{X}}$ are either 1 or 0. The remaining part of this proof is identical to the previous part. □

Corollary 2. *Suppose that $\boldsymbol{\beta}^s \in \mathbb{R}^p$ is given, and for some $M_t > 1$, we have $1/M_t \leq \kappa_t, \sigma_t^2 \leq M_t$ and $\|\boldsymbol{\Sigma}_t\|_{op}, \left\|\boldsymbol{\Sigma}_t^{-1}\right\|_{op} \leq M_t$. For every $c_t > 0$, there exists an event $E(M_t, c_t)$ with $\mathbb{P}((E(M_t, c_t))^c) \leq C_t e^{-p/C_t}$ where $C_t = C_t(M_t, c_t)$ such that*

$$\|\boldsymbol{\beta}^{s2t}\|_2 \leq \|\boldsymbol{\beta}^s\|_2 + c_t \quad \textit{and} \quad \|\boldsymbol{\beta}^{s2t} - \boldsymbol{\beta}^s\|_2 \leq \|\boldsymbol{\beta}^s\|_2 + c_t.$$

Proof. The result directly follows from the proof of Proposition 10. □

Proposition 11. *Suppose that, for some $M_t > 1$, $1/M_t \leq \kappa_t, \sigma_t^2 \leq M_t$ and $\|\boldsymbol{\Sigma}_t\|_{op}, \left\|\boldsymbol{\Sigma}_t^{-1}\right\|_{op} \leq M_t$. For every $c_t > 0$, there exists an event $E(M_t, c_t)$ with $\mathbb{P}((E(M_t, c_t))^c) \leq C_t e^{-p/C_t}$ where $C_t = C_t(M_t, c_t)$ such that we have the following on this event $E(M_t, c_t)$:*

$$\|\boldsymbol{\beta}^{s2t}\|_2 \leq \|\boldsymbol{\beta}_\star\|_2 + c_t \quad \textit{and} \quad \|\boldsymbol{\beta}^{s2t} - \boldsymbol{\beta}_\star\|_2 \leq \|\boldsymbol{\beta}_\star\|_2 + c_t$$

Proof. By the definition of $\boldsymbol{\beta}^{s2t}$, we have the following:

$$\boldsymbol{\beta}^{s2t} = \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\boldsymbol{\beta}^s + \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{z} \tag{38}$$

where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \sigma_t^2 \boldsymbol{I})$. Plugging (32) into (38), we obtain

$$\boldsymbol{\beta}^{s2t} = \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\left(\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{X}}\boldsymbol{\beta}_\star + \tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{z}}\right) + \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{z} \tag{39}$$

Note that $\boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{X}}$ are projection matrices. Multiplication of two projection matrices results in a projection matrix. Using the fact that the eigenvalues of a projection matrix are either 1 or 0 in (39), we have

$$\|\boldsymbol{\beta}^{s2t}\|_2 \leq \|\boldsymbol{\beta}_\star\|_2 + \|\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{z}}\|_2 + \|\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{z}\|_2 \tag{40}$$

By a similar reasoning used in (34),(35), and (36); there exist absolute constants $C_0, C_1, C_2, c > 0$ such that we have the following for every $\varepsilon, t > 0$:

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{z}\|_2 \leq \varepsilon(1 - \frac{1}{\kappa_t})\sqrt{(t+\sigma_t^2)\left\|\boldsymbol{\Sigma}_t^{-1}\right\|_{\text{op}}}\right) \\ \leq (C_1\varepsilon)^{p-n+1} + e^{-pC_2} + e^{-c\min\left\{\frac{pt^2}{4C_0^2}, \frac{pt}{2C_0}\right\}}\end{aligned} \tag{41}$$

Similarly, for every $\tilde{\varepsilon} > 0$, there exist absolute constants $\tilde{C}_0, \tilde{C}_1, \tilde{C}_2, \tilde{c} > 0$ such that we have the following for every $\tilde{\varepsilon}, \tilde{t}$:

$$\begin{aligned}\mathbb{P}\left(\|\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)^{-1}\tilde{\boldsymbol{z}}\|_2 \leq \tilde{\varepsilon}(1 - \frac{1}{\kappa_s})\sqrt{(\tilde{t}+\sigma_s^2)\left\|\boldsymbol{\Sigma}_s^{-1}\right\|_{\text{op}}}\right) \\ \leq (\tilde{C}_1\tilde{\varepsilon})^{p-m+1} + e^{-p\tilde{C}_2} + e^{-\tilde{c}\min\left\{\frac{p\tilde{t}^2}{4\tilde{C}_0^2}, \frac{p\tilde{t}}{2\tilde{C}_0}\right\}}\end{aligned} \tag{42}$$

Note that $\boldsymbol{X}, \boldsymbol{z}, \tilde{\boldsymbol{X}}$, and $\tilde{\boldsymbol{z}}$ are independent of each other. Therefore, we can apply union bound on (41) and (42) with selecting $\varepsilon, t, \tilde{\varepsilon}$, and $\tilde{t}$ such that $\varepsilon < \frac{1}{C_1 e}$, $\frac{c_t}{2} < \varepsilon(1 - \frac{1}{\kappa_t})\sqrt{(t+\sigma_t^2)\left\|\boldsymbol{\Sigma}_t^{-1}\right\|_{\text{op}}}$, $\tilde{\varepsilon} < \frac{1}{\tilde{C}_1}$, and $\frac{c_t}{2} < \varepsilon(1 - \frac{1}{\kappa_s})\sqrt{(\tilde{t}+\sigma_s^2)\left\|\boldsymbol{\Sigma}_s^{-1}\right\|_{\text{op}}}$. As a result, there exists an event E with $\mathbb{P}(E^c) \leq C_t(M_t, c_t)$ such that

$$\|\boldsymbol{\beta}^{s2t}\|_2 \leq \|\boldsymbol{\beta}_\star\|_2 + c_t.$$

Using a similar argument in (37), we derive the following on the same event E_1

$$\|\boldsymbol{\beta}^{s2t} - \boldsymbol{\beta}_\star\|_2 \leq \|\boldsymbol{\beta}_\star\|_2 + c_t.$$

This completes the proof.

□

Proposition 12. *Let $g : \mathbb{R}^p \to \mathbb{R}$ be a function such that*

$$g(\boldsymbol{\beta}^s) := \|\boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t(\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star)\|_2^2$$

Then, on the same event $E(M_t, c_s)$ in Proposition 10, the function g is $(\|\boldsymbol{\beta}_\star\|_2 + c_s)\frac{2\lambda_1^3}{(\lambda_1+\tau_t)^2}$−Lipschitz where λ_1 is the largest eigenvalue of $\boldsymbol{\Sigma}_t$.

Proof. We take the gradient of the function g:

$$\begin{aligned}\|\nabla g(\boldsymbol{\beta}^s)\|_2 &= 2\|\boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t(\boldsymbol{\beta}^s-\boldsymbol{\beta}_\star)\|_2\\ &\leq \|\boldsymbol{\beta}^s-\boldsymbol{\beta}_\star\|_2 \max_i \frac{2\lambda_i^3}{(\lambda_i+\tau_t)^2}\\ &= \|\boldsymbol{\beta}^s-\boldsymbol{\beta}_\star\|_2 \max_i 2\lambda_i\left(1-\frac{\tau_t}{\lambda_i+\tau_t}\right)^2\\ &= \|\boldsymbol{\beta}^s-\boldsymbol{\beta}_\star\|_2 \frac{2\lambda_1^3}{(\lambda_1+\tau_t)^2}.\end{aligned}$$

Combining Proposition 10 on the event $E(M_t, c_s)$ with the above inequality provides the advertised claim. □

Proposition 13. *Let $g : \mathbb{R}^p \to \mathbb{R}$ be a function such that*

$$g(\boldsymbol{\beta}^s) := \frac{1}{p}\|\boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}\gamma_t(\boldsymbol{\beta}^s)\|_F^2$$

Then, on the same event $E(M_t, c_s)$ in Proposition 10, the function g is $L-$Lipschitz where $(\lambda_i)_{i=1}^p$ are the eigenvalues of $\boldsymbol{\Sigma}_t$ with a descending order and

$$L = \frac{4\tau_t^2}{m}\frac{\lambda_1^3}{(\lambda_1+\tau_t)^4}\frac{\|\boldsymbol{\beta}_\star\|_2+c_s}{1-\frac{1}{m}\sum_{i=1}^p\left(\frac{\lambda_i}{\lambda_i+\tau_t}\right)^2}.$$

Proof. We take the gradient of the function g:

$$\nabla g(\boldsymbol{\beta}^s) = \frac{2}{p}\boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}\nabla\gamma_t^2(\boldsymbol{\beta}^s).$$

Note that

$$\begin{aligned}\gamma_t^2(\boldsymbol{\beta}^s) &= \kappa_t\left(\sigma_t^2+\mathbb{E}_{\boldsymbol{\beta}^{s2t}}[\|\boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\beta}^{s2t}-\boldsymbol{\beta}^s)\|_2^2]\right)\\ &= \kappa_t\frac{\sigma_t^2+\tau_t^2\|(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\beta}^s\|_2^2}{1-\frac{1}{m}\mathtt{tr}\left((\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t^2\right)}.\end{aligned}$$

Then, we have

$$\nabla\gamma_t^2(\boldsymbol{\beta}^s) = 2\kappa_t\frac{\tau_t^2\boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\beta}^s}{1-\frac{1}{m}\mathtt{tr}\left((\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t^2\right)}.$$

Plugging $\nabla\gamma_t^2(\boldsymbol{\beta}^s)$ into $\nabla g(\boldsymbol{\beta}^s)$, we obtain that

$$\begin{aligned}\|\nabla g(\boldsymbol{\beta}^s)\|_2 &= \frac{4\tau_t^2}{m}\frac{\boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\beta}^s}{1-\frac{1}{m}\mathtt{tr}\left((\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t^2\right)}\\ &\leq \frac{4\tau_t^2}{m}\frac{\lambda_1^3}{(\lambda_1+\tau_t)^4}\frac{\|\boldsymbol{\beta}^s\|_2}{1-\frac{1}{m}\sum_{i=1}^p\left(\frac{\lambda_i}{\lambda_i+\tau_t}\right)^2}.\end{aligned}$$

Combining Proposition 10 on the event $E(M_t, c_s)$ with the above inequality provides the advertised claim. □

Proposition 14. *Let $g : \mathbb{R}^p \to \mathbb{R}$ be a function such that*

$$g(\boldsymbol{\beta}^s) := 2\boldsymbol{\beta}_\star{}^\top\left(\boldsymbol{I}-(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t\right)^\top\boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t(\boldsymbol{\beta}^s-\boldsymbol{\beta}_\star).$$

Then, the function g is $2\|\boldsymbol{\beta}_\star\|_2\tau_t\left(\frac{\lambda_1}{\lambda_1+\tau_t}\right)^2-$Lipschitz where λ_1 is the largest eigenvalue of $\boldsymbol{\Sigma}_t$.

Proof. We take the gradient of the function g:

$$\begin{aligned}\|\nabla g(\boldsymbol{\beta}^s)\|_2 &= 2\|\boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t\left(\boldsymbol{I}-(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t\right)\boldsymbol{\beta}_\star\|_2\\ &\le 2\|\boldsymbol{\beta}_\star\|_2\tau_t\max_i\left(1-\frac{\tau_t}{\lambda_i+\tau_t}\right)^2\\ &= 2\|\boldsymbol{\beta}_\star\|_2\tau_t\left(\frac{\lambda_1}{\lambda_1+\tau_t}\right)^2,\end{aligned}$$

and the desired result readily follows. □

Lemma 1. *We have that*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\beta}^s\sim X^s_{\kappa_s,\sigma_s^2}}[\gamma_t^2(\boldsymbol{\beta}^s)] = \kappa_t\frac{\sigma_t^2+\tau_t^2\|(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}((\boldsymbol{\Sigma}_s+\tau_s\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_s\boldsymbol{\beta}_\star\|_2^2}{1-\frac{1}{n}\mathtt{tr}\left((\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t^2\right)}\\ +\frac{\kappa_t\tau_t^2\gamma_s^2(\boldsymbol{\beta}_\star)}{p}\frac{\mathtt{tr}\left(\boldsymbol{\Sigma}_s^{1/2}(\boldsymbol{\Sigma}_s+\tau_s\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\Sigma}_s+\tau_s\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_s^{1/2}\right)}{1-\frac{1}{n}\mathtt{tr}\left((\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t^2\right)}.\end{aligned}$$

Proof. The desired claim follows from the following manipulations using the definition of $X^s_{\kappa_s,\sigma_s^2}$ in (13):

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\beta}^s\sim X^s_{\kappa_s,\sigma_s^2}}[\gamma_t^2(\boldsymbol{\beta}^s)] &= \mathbb{E}_{\boldsymbol{\beta}^s\sim X^s_{\kappa_s,\sigma_s^2}}\left[\kappa_t\frac{\sigma_t^2+\tau_t^2\|(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\beta}^s\|_2^2}{1-\frac{1}{n}\mathtt{tr}\left((\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t^2\right)}\right]\\ &= \mathbb{E}_{\boldsymbol{g}_s}\left[\kappa_t\frac{\sigma_t^2+\tau_t^2\|(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}\left((\boldsymbol{\Sigma}_s+\tau_s\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_s\boldsymbol{\beta}_\star+(\boldsymbol{\Sigma}_s+\tau_s\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_s^{1/2}\gamma_s(\boldsymbol{\beta}_\star)\boldsymbol{g}_s/\sqrt{p}\right)\|_2^2}{1-\frac{1}{n}\mathtt{tr}\left((\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t^2\right)}\right]\\ &= \kappa_t\frac{\sigma_t^2+\tau_t^2\|(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}((\boldsymbol{\Sigma}_s+\tau_s\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_s\boldsymbol{\beta}_\star\|_2^2}{1-\frac{1}{n}\mathtt{tr}\left((\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t^2\right)}\\ &+\frac{\kappa_t\tau_t^2\gamma_s^2(\boldsymbol{\beta}_\star)}{p}\frac{\mathtt{tr}\left(\boldsymbol{\Sigma}_s^{1/2}(\boldsymbol{\Sigma}_s+\tau_s\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\Sigma}_s+\tau_s\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_s^{1/2}\right)}{1-\frac{1}{n}\mathtt{tr}\left((\boldsymbol{\Sigma}_t+\tau_t\boldsymbol{I})^{-2}\boldsymbol{\Sigma}_t^2\right)}.\end{aligned}$$

□