Precise Performance of Linear Denoisers in the Proportional Regime

Reza Ghane* RGHANEKH@CALTECH.EDU
Danil Akhtiamov* DAKHTIAM@CALTECH.EDU
Babak Hassibi HASSIBI@CALTECH.EDU
California Institute of Technology, Pasadena, CA, 91125

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

In the present paper we study the performance of linear denoisers for noisy data of the form $\mathbf{x} + \mathbf{z}$, where $\mathbf{x} \in \mathbb{R}^d$ is the desired data with zero mean and unknown covariance $\mathbf{\Sigma}$, and $\mathbf{z} \sim \mathcal{N}(0, \mathbf{\Sigma_z})$ is additive noise. Since the covariance $\mathbf{\Sigma}$ is not known, the standard Wiener filter cannot be employed for denoising. Instead we assume we are given samples $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ from the true distribution. A standard approach would then be to estimate $\mathbf{\Sigma}$ from the samples and use it to construct an "empirical" Wiener filter. However, in this paper, motivated by the denoising step in diffusion models, we take a different approach whereby we train a linear denoiser $\mathbf{W}$ from the data itself. In particular, we synthetically construct noisy samples $\hat{\mathbf{x}}_i$ of the data by injecting the samples with Gaussian noise with covariance $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma_z}$ and find the best $\mathbf{W}$ that approximates $\mathbf{W}\hat{\mathbf{x}}_i \approx \mathbf{x}_i$ in a least-squares sense. In the proportional regime $\frac{n}{d} \to \kappa > 1$ we use the *Convex Gaussian Min-Max Theorem (CGMT)* to analytically find the closed form expression for the generalization error of the denoiser obtained from this process. Using this expression one can optimize over $\mathbf{\Sigma}_1$ to find the best possible denoiser. Our numerical simulations show that our denoiser outperforms the "empirical" Wiener filter in many scenarios and approaches the optimal Wiener filter as $\kappa \to \infty$.

Keywords: Denoising, Linear Denoisers, Linear Filters, Convex Gaussian Min-Max Theorem, Gaussian Comparison Inequalities, Proportional Regime, Wiener Filter

1. Introduction and Related Works

Data denoising is a fundamental problem arising in many areas of data science, signal processing, control theory, and machine learning. Denoisers have been successfully applied to signal processing (Kailath et al., 2000), image processing (Hirakawa and Parks, 2006; Zhang et al., 2017; Zamir et al., 2022) and generative AI (Song et al.; Song and Ermon, 2019) .

Arguably the most widely known denoiser, the Wiener Filter (Wiener, 1964; Kolmogorov, 1941), dates back to the 1940s. However, constructing such filter in practice requires precise knowledge of the covariance of the data $\mathbf{x} \in \mathbb{R}^d$. Noting that reliable estimation of the second-order statistics requires access to plethora of samples, other works, such as Hirakawa and Parks (2006), attempt to circumvent such need by *learning* denoisers *directly from data*.

To provide a pertinent and opportune motivation, denoisers are used as central building blocks in modern diffusion models (Song et al.; Song and Ermon, 2019). In the context of diffusion, the denoiser $D_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^d$ is a neural network parametrized by its weights $\boldsymbol{\theta} \in \mathbb{R}^D$, such as UNet (Ho et al., 2020) or Diffusion Transformer (DiT) (Peebles and Xie, 2023). Simplifying and omitting certain technical details for ease of explanation, one can illustrate the core idea of denoising diffusion models as follows. Suppose we trained the denoiser $D_{\boldsymbol{\theta}}$ to learn a "good quality"

*Equal Contribution

approximation $D_{\boldsymbol{\theta}}(\mathbf{x}_i + \mathbf{z}_i) \approx \mathbf{x}_i$ where $\mathbf{z}_i \sim \mathcal{N}(0, \sigma_{\mathbf{z}}^2 \mathbf{I}_d)$ is artificially injected noise such that the signal to noise ratio (SNR) is very low, i.e.

$$d\sigma_{\mathbf{z}}^2 \approx \|\mathbf{z}_i\|^2 \gg \|\mathbf{x}_i\|_2^2 \tag{1}$$

Then, under (1), it would be reasonable to attempt generating new samples by dropping dependence of the denoiser on $\mathbf{x}$ as $\mathbf{x} \sim D_{\boldsymbol{\theta}}(\mathbf{z})$ where $\mathbf{z} \sim \mathcal{N}(0, \sigma_{\mathbf{z}}^2 \mathbf{I}_d)$.

Motivated by classical works on learning denoisers as well as more recent works on diffusion, we would like to initiate a systematic study of the performances of denoisers. As a starting point, we propose to study linear denoisers trained via the least squares objective. If the number of samples n and the dimensionality of data d satisfy $n \gg d$, it is known that the least squares solution will converge to the Wiener Filter, whose performance has been studied extensively in the literature. As such, to make the problem interesting, we assume that n is proportional to d.

In the proportional regime n, the number of samples, and d, the number of features, grow together such that $\frac{n}{d} \to \kappa$. This is different from classical works in statistics and control theory, where, usually, d is fixed and $n \to \infty$. Arguably, the main technical difference between the two is that one can reliably estimate the covariance of the data reliably when $n \gg d^2$, but, generally speaking, cannot do so if $n = \theta(d)$. The latter in particular means that the Wiener Filter cannot be recovered precisely in the proportional regime, as it is defined based on the covariance of the data.

Understanding the precise asymptotics of various statistical inference problems in the proportional regime has been the subject of study of many works in the past decade (Thrampoulidis et al., 2018; Mallory et al., 2025; Hastie et al., 2022; Li, 2025; Wang et al., 2025; Huang, 2025; Loureiro et al., 2021; Deng et al., 2022; Akhtiamov et al., 2024; Ghane et al., 2025; Akhtiamov et al., 2025b; Ghane et al., 2024). Analyzing the performances of learning algorithms in the proportional regime requires a different toolbox from classical statistical work, the main two of which have proven to be the Approximate Message Passing (AMP) (Donoho et al., 2009) and the Convex Gaussian Min-Max (CGMT) (Thrampoulidis et al., 2014) frameworks. For the purposes of the present work, we employ the latter (CGMT) approach.

2. Problem Setting and Preliminaries

We use bold font for vectors and matrices and normal font for scalars in this exposition. Consider the following data denoising problem:

Given n i.i.d data points sampled from the normal distribution $\mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ with **an unknown** covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ and a noise distribution $\mathbf{z} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\mathbf{z}})$ with a **known** covariance $\boldsymbol{\Sigma}_{\mathbf{z}} \in \mathbb{R}^{d \times d}$, design a linear denoiser $\mathbf{W} \in \mathbb{R}^{d \times d}$, that takes noisy data of the form $\hat{\mathbf{x}} = \mathbf{x} + \mathbf{z}$ as input and "suppresses the noise" and outputs $\mathbf{W}\hat{\mathbf{x}} \approx \mathbf{x}$ as accurately as possible. More formally, $\mathbf{W}$ should minimize the following generalization error objective:

$$\mathcal{E}(\mathbf{W}) := \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\mathbf{z}})} \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})} \|\mathbf{W}^T(\mathbf{x} + \mathbf{z}) - \mathbf{x}\|_2^2 \tag{2}$$

Note that (2) can be simplified as:

$$\mathcal{E}(\mathbf{W}) = \mathbb{E}_{\mathbf{z}} \mathbb{E}_{\mathbf{x}} \|(\mathbf{W}^T - \mathbf{I})\mathbf{x} + \mathbf{W}^T\mathbf{z}\|_2^2 = \mathrm{Tr}\left((\mathbf{W}^T - \mathbf{I})\boldsymbol{\Sigma}(\mathbf{W}^T - \mathbf{I})^T\right) + \mathrm{Tr}\left(\mathbf{W}^T \boldsymbol{\Sigma}_{\mathbf{z}} \mathbf{W}\right) \tag{3}$$

Minimizing (3) over $\mathbf{W}$ directly leads to

$$\mathbf{W} = (\mathbf{\Sigma} + \mathbf{\Sigma_z})^{-1} \mathbf{\Sigma} \tag{4}$$

The denoiser (4) is known as the *Wiener Filter* in the literature. By construction, (4) achieves the optimal generalization error among all linear denoisers. However, finding (4) requires precise knowledge of $\mathbf{\Sigma}$. Recovering $\mathbf{\Sigma}$ from data with good accuracy might not be feasible when n and d are proportional to each other, unlike the classical case $n \gg d$. As such, one natural approach to the denoising problem would be to take the *empirical Wiener Filter*

$$\mathbf{W} = \left(\hat{\mathbf{\Sigma}} + \mathbf{\Sigma_z}\right)^{-1} \hat{\mathbf{\Sigma}} \tag{5}$$

Where the empirical covariance $\hat{\mathbf{\Sigma}}$ is defined as follows via the sample data matrix $\mathbf{X}^T = \begin{pmatrix} \mathbf{X}^1 & \mathbf{X}^2 & \dots \mathbf{X}^n \end{pmatrix} \in \mathbb{R}^{d \times n}$ with $\mathbf{X}^i \in \mathbb{R}^d$ being the rows generated independently:

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-1} \mathbf{X}^T \left(\mathbf{I}_n - \frac{\mathbf{1}_n \mathbf{1}_n^T}{n} \right) \mathbf{X} \tag{6}$$

Naturally, the quality of the denoiser (5) depends on the quality of approximation of the true covariance $\mathbf{\Sigma}$ by the empirical covariance (6). The latter does not have to be good in general, as $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ has $\Theta(d^2)$ degrees of freedom and (6) attempts to estimate $\mathbf{\Sigma}$ from $n = \Theta(d)$ samples. As such, we would like to consider an alternative approach to minimizing (2) as well. The approach we propose is outlined below and is based on learning $\mathbf{W}$ directly from data, but the main technicality lies in how the training data should be generated from $\mathbf{X}$.

We divide the training data $\mathbf{X}$ into N batches $\mathbf{X}_1 \in \mathbb{R}^{n_1 \times d}, \dots, \mathbf{X}_N \in \mathbb{R}^{n_N \times d}$ In other words, $\mathbf{X}_1 \in \mathbb{R}^{n_1 \times d}, \dots, \mathbf{X}_N \in \mathbb{R}^{n_N \times d}$ are defined via $\mathbf{X}^T = \begin{pmatrix} \mathbf{X}_1^T & \mathbf{X}_2^T & \dots & \mathbf{X}_N^T \end{pmatrix}$. Using $\mathbf{X}_t^i$ to denote the ith row of $\mathbf{X}_t$, note that the data points $\mathbf{X}_t^i$ and $\mathbf{X}_{t'}^{i'}$ are independent whenever $(i,t) \neq (i',t')$ and each row satisfies

$$\mathbf{X}_t^i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \text{ for } t = 1, \dots, N \text{ and } i = 1, \dots, n_t$$

Furthermore, consider independent noise vectors

$$\mathbf{Z}_t^i \in \mathbb{R}^d \text{ where } \mathbf{Z}_t^i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t),\ t = 1, \dots, N, i = 1, \dots, n_t$$

We construct noisy data batches $\hat{\mathbf{X}}_t := \mathbf{X}_t + \mathbf{Z}_t,\ t = 1, \dots, N$ and train a denoiser $\mathbf{W}_{\text{lsq}}$ as:

$$\mathbf{W}_{\text{lsq}} := \arg\min_{\mathbf{W}} \left\| \hat{\mathbf{X}} \mathbf{W} - \mathbf{X} \right\|_F^2 \tag{7}$$

Here, $\hat{\mathbf{X}}$ is the full noisy data matrix defined as $\hat{\mathbf{X}}^T = \begin{pmatrix} \hat{\mathbf{X}}_1^T & \hat{\mathbf{X}}_2^T & \dots & \hat{\mathbf{X}}_N^T \end{pmatrix}$. The main results of the present work, namely Theorems 1 and 5 in Section 3, address the following questions:

What is the generalization error (2) of the denoiser $\mathbf{W}_{\text{lsq}}$ found via (7) for given $N, n_1, \dots, n_t, d, \mathbf{\Sigma}_1, \dots, \mathbf{\Sigma}_N$?

Focusing on the case of $N = 1$ and relying on Theorem 1, we also conduct extensive mathematical (Corollary 3) as well as numerical (Section 5) investigations of the following questions:

What is the optimal $\mathbf{\Sigma}_1$ leading to the minimal generalization error (2) among linear denoisers trained via (7) in terms of $\mathbf{\Sigma}$ and $\mathbf{\Sigma_z}$? How can we approximate this optimal $\mathbf{\Sigma}_1$ algorithmically given access only to $\mathbf{x}_1, \dots, \mathbf{x}_n$? Can we outperform the empirical Wiener filter (5) using such an approximation of the optimal $\mathbf{\Sigma}_1$?

We will also impose the following technical assumptions that are necessary for our analyses.

Assumptions 1

1. $\mathbf{\Sigma}$ *is invertible.*
2. $\frac{n_t}{d} \to \kappa_t > 0$ *as* $n_t, d \to \infty$ *for each* $t = 1, \dots, N$ *and* $\kappa = \kappa_1 + \dots + \kappa_N > 1$.

Our Main Contributions:

- We characterize the generalization error of $\mathbf{W}_{\text{lsq}}$ found from (7) for arbitrary $N, n_1, \dots, n_t, d$, $\mathbf{\Sigma}_1, \dots, \mathbf{\Sigma}_N$ satisfying Assumptions 1 in Theorem 5.
- In the case of $N = 1$, we derive much more explicit results, which are formulated in Theorem 1, again assuming assumptions 1 hold.
- We use Theorem 1 to find the optimal $\mathbf{\Sigma}_1$ minimizing the generalization error for the specialized case when $N = 1$ and both $\mathbf{\Sigma}$ and $\mathbf{\Sigma_z}$ are scalar. Perhaps surprisingly, even in this case we see that the optimal $\mathbf{\Sigma}_1$ is in general different from $\mathbf{\Sigma_z}$, meaning it's better to take different noise distributions for training and testing. The result is presented in Corollary 3.
- In Section 5, we propose an algorithm for approximating the optimal $\mathbf{\Sigma}_1$ for the case $N = 1$ and arbitrary $\mathbf{\Sigma}, \mathbf{\Sigma_z}$. This algorithm is based on Theorem 1. We then proceed to verify that in certain cases the $\mathbf{W}_{\text{lsq}}$ resulting from (7) indeed outperforms the empirical Wiener filter (5).

3. Main Results

Since we use CGMT, our results are asymptotic, meaning that we characterize the error $\mathcal{E}(\mathbf{W}_{\text{lsq}})$ via certain quantities $f_{n,d}$, such that $\frac{\mathcal{E}(\mathbf{W}_{\text{lsq}})}{f_{n,d}} \to 1$ as $n, d \to \infty$ and $\frac{n}{d} \to \kappa > 1$. For the ease of exposition, we just write equalities within the results of this section, but keep the reference to the asymptotic regime in the formulation to avoid confusion. We would like to begin with stating the results we obtained for $N = 1$, as the expressions we get in this case are more explicit and thus allow for more insight.

Theorem 1 *Assume that* $N = 1$ *and Assumptions 1 hold. Then, for the denoiser* $\mathbf{W}_{lsq}$ *found via* (7)*, its generalization error* (2) *can be characterized asymptotically as follows:*

$$\mathcal{E}(\mathbf{W}_{lsq}) = \Big(1 + \frac{1}{n-d} Tr\Big[\Big(\mathbf{\Sigma} + \mathbf{\Sigma}_1\Big)^{-1}\Big(\mathbf{\Sigma} + \mathbf{\Sigma_z}\Big)\Big]\Big)\Big(Tr\Big[\mathbf{\Sigma}\Big] - Tr\Big[\Big(\mathbf{\Sigma} + \mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}^2\Big)\Big]\Big) + Tr\Big[\Big(\mathbf{\Sigma} + \mathbf{\Sigma}_1\Big)^{-1}\Big(\mathbf{\Sigma_z} - \mathbf{\Sigma}_1\Big)\Big(\mathbf{\Sigma} + \mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}^2\Big] \quad (8)$$

Specifying Theorem 1 to the case $\mathbf{\Sigma}_1 = \mathbf{\Sigma_z}$ leads to the following corollary:

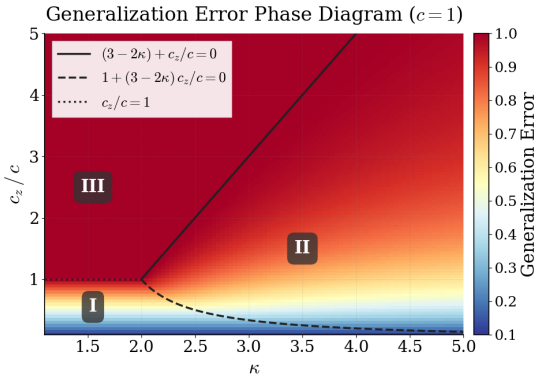


Figure 1: Phase Transition explained in Remark 4

Corollary 2 *Under the setting of Theorem 1, assume in addition that* $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_\mathbf{z}$ *and* $\mathbf{\Sigma}_\mathbf{z}$ *and* $\mathbf{\Sigma}$ *commute. Then, for the denoiser found via (7), its generalization error (2) asymptotically equals the following, where* $\mathcal{E}_{Wiener}$ *denotes the generalization error (2) minimized over* $\mathbf{W}$*, which is attained by the Wiener Filter (4):*

$$\mathcal{E}(\mathbf{W}_{lsq}) = \frac{\kappa}{\kappa - 1}\mathcal{E}_{Wiener}$$

Note that taking $\kappa \to \infty$*, i.e. assuming* $n \gg d$*, recovers the performance of the Wiener filter. As expected,* $\mathcal{E}(\mathbf{W}_{lsq}) \geq \mathcal{E}_{Wiener}$*.*

We obtain another corollary of Theorem 1 by treating the case of scalar $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_\mathbf{z}$ in greater detail:

Corollary 3 *Under the setting of Theorem 1, assume in addition that* $\mathbf{\Sigma}$ *and* $\mathbf{\Sigma}_\mathbf{z}$ *are scalar matrices and define:*

$$\mathbf{\Sigma} = \frac{c}{d}\mathbf{I}_d \text{ and } \mathbf{\Sigma}_\mathbf{z} = \frac{c_\mathbf{z}}{d}\mathbf{I}_d$$

Then minimizing the generalization error (2) of the denoiser found via (7) over $\mathbf{\Sigma}_1$ *leads to the following optimal generalization error:*

$$\begin{cases} \frac{-(c+c_\mathbf{z})^2+4(\kappa-1)^2 c c_\mathbf{z}}{4(\kappa-2)(\kappa-1)(c+c_\mathbf{z})} & (3-2\kappa)c + c_\mathbf{z} < 0, \quad c + (3-2\kappa)c_\mathbf{z} < 0 \\ \min\{c, c_\mathbf{z}\} & \textit{otherwise} \end{cases} \tag{9}$$

Remark 4 *Corollary 3 reveals an interesting phase transition phenomenon, whose phase diagram is depicted in Fig. 1. In region I, defined by* $3 - 2\kappa + \frac{c_\mathbf{z}}{c} > 0$*, the generalization error equals* $c_\mathbf{z}$ *and the optimal* $\mathbf{W}_{lsq} = \mathbf{I}_d$ *regardless of the choice of* $\mathbf{\Sigma}_1$*. In region III, defined by* $3 - 2\kappa + \frac{c}{c_\mathbf{z}} > 0$*, the generalization error equals* c *and the optimal* $\mathbf{W}_{lsq} = 0$*, again regardless of the choice of* $\mathbf{\Sigma}_1$*. The interesting region II, defined as the complement of I and III, is where there exists a unique* $\mathbf{\Sigma}_1$ *achieving the best performance and the resulting* $\mathbf{W}_{lsq} \notin \{0, \mathbf{I}_d\}$*. We fixed* $c = 1$ *for Fig. 1.*

Finally, we formulate the expression that we have derived for the case of $N > 1$. The expression we obtain for $N > 1$ is more complicated and we leave analyzing it further and making it more explicit as a subject for future work.

Theorem 5 *Suppose that Assumptions 1 hold. Consider $\{\theta_t\}_{t=1}^N$ found as the unique fixed points of the following system of equations*

$$\theta_t = 2 - 2\theta_t Tr\Big[\Big(\sum_{t=1}^{N}\theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}_t\Big]$$

Define

$$\theta = 2 - 4Tr\Big[\Big(\sum_{t=1}^{N}\theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}\Big]$$

Finally, consider the matrix $\mathbf{A}\in\mathbb{R}^{(N+2)\times(N+2)}$ defined in (25) and vector $\mathbf{b}\in\mathbb{R}^{N+2}$ defined in (27), deferred to the Appendix due to their longer descriptions. Then, for the denoiser $\mathbf{W}_{lsq}$ found via (7), its generalization error (2) can be characterized asymptotically as follows:

$$\mathcal{E}(\mathbf{W}_{lsq}) = \frac{1}{4(n-d)^2}(\mathbf{A}^{-1}\mathbf{b})_{N+2} + Tr\Big[\Big(\sum_{t=1}^{N}\theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}_{\mathbf{z}}\Big(\sum_{t=1}^{N}\theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}$$
$$\cdot\Big(4\theta^2(\mathbf{A}^{-1}\mathbf{b})_1\mathbf{\Sigma} + 16(n-d)^2\mathbf{\Sigma}^2 + \theta^2\sum_{t=1}^{N}\theta_t^2\mathbf{\Sigma}_t(\mathbf{A}^{-1}\mathbf{b})_{t+1}\Big)\Big]$$

Where $(\mathbf{A}^{-1}\mathbf{b})_j$ denotes the jth entry of the vector $\mathbf{A}^{-1}\mathbf{b}\in\mathbb{R}^{N+2}$.

4. Proof Technique and Sketches

Both proofs rely heavily on a result known in the literature as the Convex Gaussian Min-Max Theorem (CGMT). We refer the reader to Thrampoulidis et al. (2014); Akhtiamov et al. (2025a) for a formal detailed treatment of the CGMT, but we will outline it briefly for completeness. The goal of the CGMT is to analyze properties of solutions of the objectives of the following form, called the *Primary Optimization (PO)* objective:

$$\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\mathbf{u}^T\mathbf{G}\mathbf{w} + \psi(\mathbf{u},\mathbf{w}) \quad \textbf{(PO)} \tag{10}$$

Here, $\mathbf{u}\in\mathbb{R}^n$, $\mathbf{w}\in\mathbb{R}^d$, $\mathcal{S}_{\mathbf{u}}\subset\mathbb{R}^n$, $\mathcal{S}_{\mathbf{w}}\subset\mathbb{R}^d$ are compact convex sets, ψ is convex in $\mathbf{u}$ and concave in $\mathbf{w}$ and $\mathbf{G}$ is an i.i.d. $\mathcal{N}(0,1)$ *matrix*. To analyze (10), the CGMT framework introduces another objective, called *Auxillary Objective (AO)*, where $\mathbf{g}\in\mathbb{R}^d$, $\mathbf{h}\in\mathbb{R}^n$ are i.i.d. $\mathcal{N}(0,1)$ *vectors*:

$$\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\mathbf{g}^T\mathbf{w}\|\mathbf{u}\| + \mathbf{h}^T\mathbf{u}\|\mathbf{w}\| + \psi(\mathbf{u},\mathbf{w}) \quad \textbf{(AO)} \tag{11}$$

Consider an arbitrary set $\mathbf{S}\subset\mathbf{S}_{\mathbf{w}}$. Let $\mathbf{w}_{AO}$ and $\mathbf{w}_{PO}$ be any minimizers of (11) and (10) respectively. Assume that $\mathbf{w}_{AO}\in\mathbf{S}$ holds with probability approaching 1 (w.p.a. 1), i.e. :

$$\mathbb{P}\left(\mathbf{w}_{AO}\in\mathbf{S}\right)\to 1 \text{ as } n,d\to\infty$$

Then $\mathbf{w}_{PO}\in\mathbf{S}$ holds w.p.a. 1 as well, i.e.:

$$\mathbb{P}\left(\mathbf{w}_{PO}\in\mathbf{S}\right)\to 1 \text{ as } n,d\to\infty$$

Note that, formally speaking, we should introduce sequences of PO and AO to discuss convergence in probability. However, we suppress these formalities to ease exposition and refer the interested reader to Thrampoulidis et al. (2014); Akhtiamov et al. (2025a) for a rigorous exposition.

Sketch of the proof of Theorems 1 and 5: Recall that the linear denoiser $\mathbf{W}_{\text{lsq}}$ is defined via:

$$\min_{\mathbf{W}} \left\| \hat{\mathbf{X}}\mathbf{W} - \mathbf{X} \right\|_F^2 = \min_{\mathbf{W}} \|\mathbf{X}(\mathbf{W} - \mathbf{I}) + \mathbf{Z}\mathbf{W}\|_F^2 \tag{12}$$

Denoting the i-th column of $\mathbf{W}$ by $\mathbf{w}_i$, we can rewrite (12) as:

$$\sum_{i=1}^{d} \min_{\mathbf{w}_i} \|\mathbf{X}(\mathbf{w}_i - \mathbf{e}_i) + \mathbf{Z}\mathbf{w}_i\|^2 \tag{13}$$

Introducing a set of Fenchel dual variables $\mathbf{v}_i \in \mathbb{R}^d$ for $i = 1, \dots, d$ and writing $\mathbf{X} = \mathbf{G}\mathbf{\Sigma}^{1/2}$, where $\mathbf{G} \in \mathbb{R}^{n \times d}$ is i.i.d. standard normal, we obtain:

$$\sum_{i=1}^{d} \min_{\mathbf{w}_i} \max_{\mathbf{v}_i} \mathbf{v}_i^T \mathbf{G}\mathbf{\Sigma}^{1/2}(\mathbf{w}_i - \mathbf{e}_i) + \mathbf{v}_i^T \mathbf{Z}\mathbf{w}_i - \frac{\|\mathbf{v}_i\|_2^2}{4}$$

With a slight abuse of notation we drop the indices from $\mathbf{w}_i$ and $\mathbf{v}_i$ and rewrite:

$$\sum_{i=1}^{d} \min_{\mathbf{w}} \max_{\mathbf{v}} \mathbf{v}^T \mathbf{G}\mathbf{\Sigma}^{1/2}(\mathbf{w} - \mathbf{e}_i) + \mathbf{v}^T \mathbf{Z}\mathbf{w} - \frac{\|\mathbf{v}\|_2^2}{4}$$

Let $\mathbf{u} := \mathbf{\Sigma}^{1/2}(\mathbf{w} - \mathbf{e}_i)$. Using a Lagrange multiplier and leveraging Sion's minimax theorem:

$$\sum_{i=1}^{d} \max_{\boldsymbol{\lambda}} \min_{\mathbf{w},\mathbf{u}} \max_{\mathbf{v}} \mathbf{v}^T \mathbf{G}\mathbf{u} + \mathbf{v}^T \mathbf{Z}\mathbf{w} + \boldsymbol{\lambda}^T \mathbf{u} - \boldsymbol{\lambda}^T \mathbf{\Sigma}^{1/2}(\mathbf{w} - \mathbf{e}_i) - \frac{\|\mathbf{v}\|_2^2}{4} \tag{14}$$

Treating each term of (14) as a separate PO with an i.i.d. standard normal $\mathbf{G}$, we arrive at the following sum of AOs:

$$\sum_{i=1}^{d} \max_{\boldsymbol{\lambda}} \min_{\mathbf{w},\mathbf{u}} \max_{\mathbf{v}} \|\mathbf{u}\|_2 \mathbf{h}^T \mathbf{v} + \|\mathbf{v}\|_2 \mathbf{g}^T \mathbf{u} + \mathbf{v}^T \mathbf{Z}\mathbf{w} + \boldsymbol{\lambda}^T \mathbf{u} - \boldsymbol{\lambda}^T \mathbf{\Sigma}^{1/2}(\mathbf{w} - \mathbf{e}_i) - \frac{\|\mathbf{v}\|_2^2}{4}$$

Denoting $\beta = \|\mathbf{v}\|_2$ and performing optimization over the direction $\frac{\mathbf{v}}{\|\mathbf{v}\|}$, we have

$$\sum_{i=1}^{d} \max_{\boldsymbol{\lambda}} \min_{\mathbf{w},\mathbf{u}} \max_{\beta > 0} \beta \mathbf{g}^T \mathbf{u} + \beta \left\| \mathbf{Z}\mathbf{w} + \|\mathbf{u}\|_2 \mathbf{h} \right\|_2 + \boldsymbol{\lambda}^T \mathbf{u} - \boldsymbol{\lambda}^T \mathbf{\Sigma}^{1/2}(\mathbf{w} - \mathbf{e}_i) - \frac{\beta^2}{4}$$

Denoting $\eta = \|\mathbf{u}\|_2$ and optimizing over $\frac{\mathbf{u}}{\|\mathbf{u}\|}$, we obtain

$$\sum_{i=1}^{d} \max_{\boldsymbol{\lambda}} \min_{\mathbf{w},\eta > 0} \max_{\beta > 0} -\eta \|\beta \mathbf{g} + \boldsymbol{\lambda}\|_2 + \beta \left\| \mathbf{Z}\mathbf{w} + \eta \mathbf{h} \right\|_2 - \boldsymbol{\lambda}^T \mathbf{\Sigma}^{1/2}(\mathbf{w} - \mathbf{e}_i) - \frac{\beta^2}{4}$$

After swapping around min and max due to convexity-concavity by Lemma 6, we employ the square-root trick $\sqrt{x} = \min_{s>0} \frac{s}{2} + \frac{x}{2s}$ and obtain:

$$\sum_{i=1}^{d} \max_{\boldsymbol{\lambda}} \min_{\eta>0} \max_{\beta>0} -\eta \|\beta \mathbf{g} + \boldsymbol{\lambda}\|_2 - \frac{\beta^2}{4} + \min_{\mathbf{w},\tau>0} \frac{\beta\tau}{2} + \frac{\beta}{2\tau} \left\| \mathbf{Z}\mathbf{w} + \eta \mathbf{h} \right\|_2^2 - \boldsymbol{\lambda}^T \boldsymbol{\Sigma}^{1/2} (\mathbf{w} - \mathbf{e}_i)$$

Now, recall the block structure we have for $\mathbf{Z}$ and introduce a matching structure for $\mathbf{h}$, where each $\mathbf{h}^{(t)} \in \mathbb{R}^{n_t}$: $\mathbf{Z}^T = \begin{pmatrix} \mathbf{Z}_1^T & \mathbf{Z}_2^T & \dots & \mathbf{Z}_N^T \end{pmatrix}$ and $\mathbf{h}^T = \begin{pmatrix} \mathbf{h}^{(1)T} & \mathbf{h}^{(2)T} & \dots & \mathbf{h}^{(N)T} \end{pmatrix}$. We arrive at:

$$\sum_{i=1}^{d} \max_{\boldsymbol{\lambda}} \min_{\eta>0} \max_{\beta>0} -\eta \|\beta \mathbf{g} + \boldsymbol{\lambda}\|_2 - \frac{\beta^2}{4} + \min_{\mathbf{w},\tau>0} \frac{\beta\tau}{2} + \frac{\beta}{2\tau} \sum_{t=1}^{N} \left\| \mathbf{Z}_t\mathbf{w} + \eta \mathbf{h}^{(t)} \right\|_2^2 - \boldsymbol{\lambda}^T \boldsymbol{\Sigma}^{1/2} (\mathbf{w} - \mathbf{e}_i)$$

Using a Fenchel Dual $\mathbf{v}_t \in \mathbb{R}^{n_t}$ for each $t = 1, \dots, N$, we obtain:

$$\sum_{i=1}^{d} \min_{\mathbf{w}} \max_{\mathbf{v}_t} \frac{\beta}{2\tau} \Big(\sum_{t=1}^{N} \mathbf{v}_t^T \mathbf{Z}_t \mathbf{w} + \eta \mathbf{v}_t^T \mathbf{h}^{(t)} - \frac{\|\mathbf{v}_t\|_2^2}{4} \Big) - \boldsymbol{\lambda}^T \boldsymbol{\Sigma}^{1/2} (\mathbf{w} - \mathbf{e}_i) \tag{15}$$

Applying CGMT again, but this time with respect to the randomness in $\mathbf{Z}^{(t)}$, we can replace the inner optimization part of (15) by the following:

$$\min_{\mathbf{w}} \max_{\mathbf{v}_t} \frac{\beta}{2\tau} \Big(\sum_{t=1}^{N} \|\boldsymbol{\Sigma}_t^{1/2}\mathbf{w}\|_2 \mathbf{h}_t^T \mathbf{v}_t + \|\mathbf{v}_t\|_2 \mathbf{g}_t^T \boldsymbol{\Sigma}_t^{1/2} \mathbf{w} + \eta \mathbf{v}_t^T \mathbf{h}^{(t)} - \frac{\|\mathbf{v}_t\|_2^2}{4} \Big) - \boldsymbol{\lambda}^T \boldsymbol{\Sigma}^{1/2} (\mathbf{w} - \mathbf{e}_i) \tag{16}$$

We defer further analyses of (16) to the Appendix due to the lack of space, but we wanted to illustrate that the proofs of Theorems 1 and 5 proceed by a two-stage application of the CGMT followed by a meticuolous detailed analysis of (16).

5. Experiments

In this section we provide numerical experiments to corroborate our findings and discuss the proposed denoising algorithm.

5.1. Verifying Corollary 3 and Theorem 1

We begin with verifying Corollary 3. We take $\boldsymbol{\Sigma} = \mathbf{I}$ and $\boldsymbol{\Sigma}_{\mathbf{z}} = \boldsymbol{\Sigma}_1 = c_{\mathbf{z}}\mathbf{I}$ and plot the performance of the linear denoiser against κ. We also use the closed-form expression obtained in (9) to plot the optimal generalization error from Corollary 3. Furthermore, we investigate how the error changes if we replace the Gaussian features with Rademacher with the same the first and second moments. The results are reported in Fig. 2, where cross marks denote the simulated values and solid-lines depict the predictions coming from Corollary 3. We observe a close match between our predictions and the simulated errors for the Gaussian features as well as a close match between the generalization errors for Gaussian and Rademacher features. The latter suggests a form of universality, which we propose as a direction for future work.

To verify Theorem 1 for a non-scalar $\boldsymbol{\Sigma}$, we generated n synthetic data points of dimension $d = 50$ distributed i.i.d. according to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}_{ij} = \delta_{ij} i^{-4}, i, j = 1, \dots, d$ (the power law

exponent 4 is chosen arbitrarily without any specific consideration in mind). For simplicity, we set $\mathbf{\Sigma_z} = \mathbf{\Sigma}_1 = \mathbf{I}$. We vary the number of samples n by changing κ and train the denoiser via solving the least-squares objective through the linear system of equations by Karush–Kuhn–Tucker (KKT) conditions. For each κ, we evaluate the generalization error of the trained denoiser and compare it with our predictions from Theorem 1. Finally, we also approximated the optimal $\mathbf{\Sigma}_1$ dictated by the expression (8) from Theorem 1 via the following heuristics: we initialize $\mathbf{\Sigma}_1 = \mathbf{\Sigma_z}$ and then performed projected gradient descent on $\mathbf{\Sigma}_1$. To make sure $\mathbf{\Sigma}_1$ stays PSD, we fixed the basis of $\mathbf{\Sigma}_1$ to be the basis of $\mathbf{\Sigma}$ and trained vector of the eigenvalues of $\mathbf{\Sigma}_1$ via gradient descent with projection on the positive orthant. The results are reported in Fig. 3 where cross marks denote the simulated values and solid-lines depict the predictions coming from Theorem 1. We observe a close match between the theory and the simulated values as well as that optimizing over $\mathbf{\Sigma}_1$ does increase performance.

5.2. Heuristics for using the result of Theorem 1 for training denoisers

Since, in practice, $\mathbf{\Sigma}$ is not known, we propose to use the sample covariance (6) and perform the following optimization as a surrogate of the expression (8) from Theorem 1:

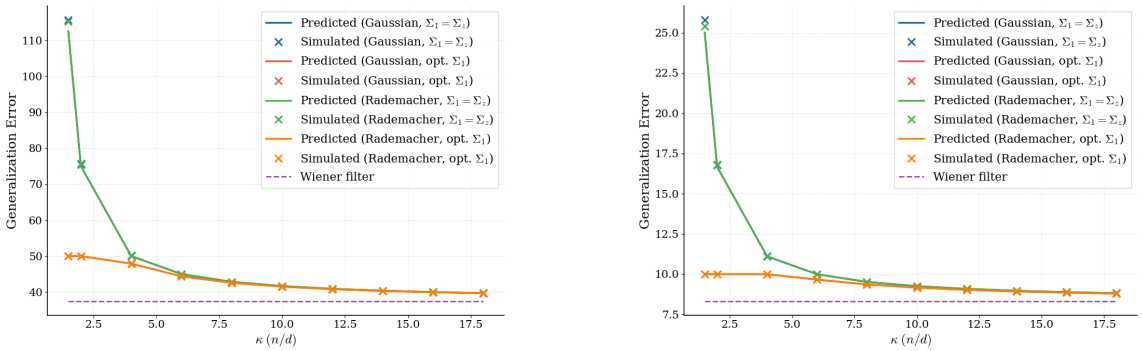
$$\min_{\mathbf{\Sigma}_1 \succeq \mathbf{0}} \Big(1 + \frac{1}{n-d} \mathrm{Tr}\Big(\Big(\hat{\mathbf{\Sigma}} + \mathbf{\Sigma}_1\Big)^{-1}\Big(\hat{\mathbf{\Sigma}} + \mathbf{\Sigma_z}\Big)\Big)\Big)\Big(\mathrm{Tr}\Big(\hat{\mathbf{\Sigma}}\Big) - \mathrm{Tr}\Big(\Big(\hat{\mathbf{\Sigma}} + \mathbf{\Sigma}_1\Big)^{-1}\hat{\mathbf{\Sigma}}^2\Big)\Big)\Big) \tag{17}$$

Since the problem (17) is not convex in $\mathbf{\Sigma}_1$, solving it precisely appears out of reach, so we suggest the following heuristics. We initialize $\mathbf{\Sigma}_1 = \mathbf{\Sigma_z}$ and, to make sure $\mathbf{\Sigma}_1$ stays PSD, we fix the basis of $\mathbf{\Sigma}_1$ to be the basis of $\hat{\mathbf{\Sigma}}$ and train the vector of the eigenvalues of $\mathbf{\Sigma}_1$ via gradient descent with projection on the positive orthant. The performance of the true Wiener filter is depicted using dashed-lines and serves as a baseline for performance. To demonstrate that it is possible to outperform the empirical Wiener filter, we consider the case where $\mathbf{\Sigma_z} = \mathbf{I}$ and $\mathbf{\Sigma} = c^2\mathbf{G}\mathbf{G}^T$, with $c > 0$ being the scaling factor and $\mathbf{G} \in \mathbb{R}^{d\times d}$ having i.i.d. standard normal entries. We choose a non-diagonal $\mathbf{\Sigma}$ to ensure that it cannot be estimated reliably from $\Theta(d)$ samples. We vary c and plot the performance of the empirical filter against the performance of $\mathbf{W}_{\text{lsq}}$ with optimized $\mathbf{\Sigma}_1$ in Figure 4. We also have included the performance of the Wiener filter and the $\mathbf{W} = \mathbf{I}$ denoiser. As we increase c, we observe that performance of the empirical and true Wiener filters and our proposed filter approaches that of the identity denoiser. We see that, for large values of c, i.e. the high SNR regime, $\mathbf{W}_{\text{lsq}}$ outperforms the empirical Wiener filter.

6. Conclusion and Future Directions

In this work we rigorously derived the precise asymptotics of a linear denoiser trained using the least-squares objective in the proportional regime. Possible future directions include investigating the expressions obtained in Theorem 5 to understand how using different batches of data may help with the performance. Another promising direction is to extend the analysis to the overparameterized regime where the number of parameters exceed the number of samples and the denoiser is trained via SGD or via a regularized objective.

While we maintained the assumption of data gaussianity throughout the present work, we suspect it is not necessary for Theorems 1 and 5. As has been observed in the literature for many statistical inference problems, the generalization error often depends only on the first and second order statistics of the data, a phenomenon commonly referred to as Gaussian Universality (see



(a) $c_{\mathbf{z}} = 3$ **(b)** $c_{\mathbf{z}} = 0.2$

Figure 2: Verification of Corollary 3 with $c = 1$

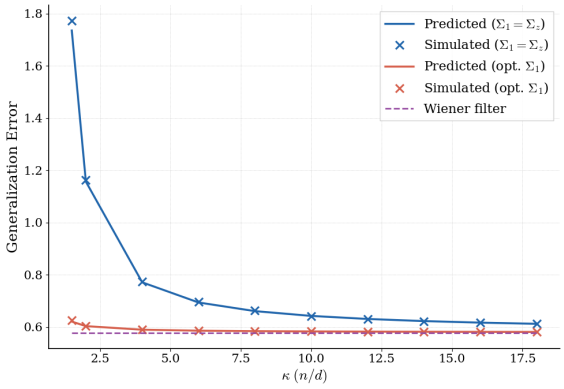
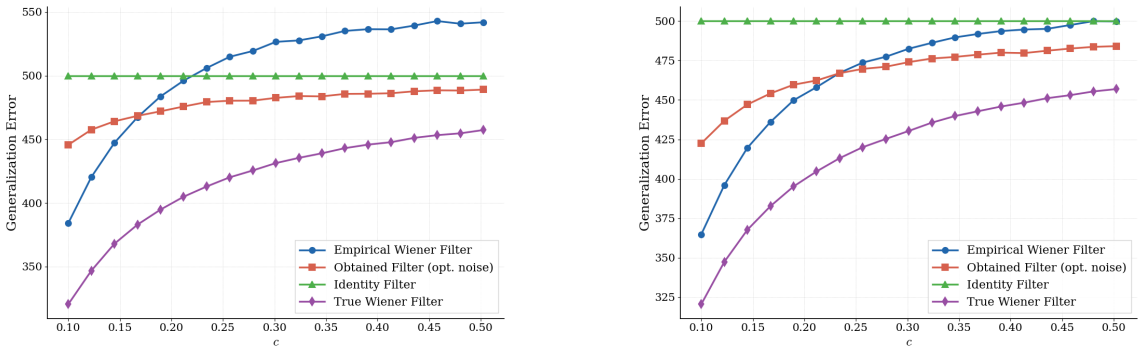


Figure 3: Numerical verification of Theorem 1

(Pesce et al., 2023; Ghane et al., 2024) and references therein). In addition, Figure 2 suggests that a form of universality might hold for the denoisers trained through the least-squares objective too. We leave the latter as an interesting direction for future work.



(a) $n = 800$ and $d = 500$ **(b)** $n = 1000$ and $d = 500$

Figure 4: Comparison of our approach to the empirical Wiener and true Wiener filters

Acknowledgments

R.G. and D.A. would like to thank Anthony Bao for many stimulating conversations on diffusion models that we have had together.

References

Danil Akhtiamov, Reza Ghane, and Babak Hassibi. Regularized linear regression for binary classification. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 202–207. IEEE, 2024.

Danil Akhtiamov, David Bosch, Reza Ghane, Nithin K Varma, and Babak Hassibi. A novel gaussian min-max theorem and its applications. *IEEE Transactions on Information Theory*, 2025a.

Danil Akhtiamov, Reza Ghane, and Babak Hassibi. A precise performance analysis of the randomized singular value decomposition. *arXiv preprint arXiv:2510.06490*, 2025b.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2): 435–495, 2022.

Ivan Dokmanic and Rémi Gribonval. Concentration of the frobenius norm of generalized matrix inverses. *SIAM Journal on matrix analysis and applications*, 40(1):92–121, 2019.

David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

Reza Ghane, Danil Akhtiamov, and Babak Hassibi. Universality in transfer learning for linear models. *Advances in Neural Information Processing Systems*, 37:125729–125779, 2024.

Reza Ghane, Danil Akhtiamov, and Babak Hassibi. One-bit quantization and sparsification for multiclass linear classification with strong regularization. *IEEE Transactions on Signal Processing*, 2025.

Hamed Hassani and Adel Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression. *The Annals of Statistics*, 52(2):441–465, 2024.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

Keigo Hirakawa and Thomas W Parks. Image denoising using total least squares. *IEEE Transactions on image processing*, 15(9):2730–2742, 2006.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Kevin Han Huang. *Universality beyond the classical asymptotic regime*. PhD thesis, UCL (University College London), 2025.

Thomas Kailath, Ali H. Sayed, and Babak Hassibi. *Linear estimation*. Prentice Hall, 2000.

Andrei Nikolaevitch Kolmogorov. Stationary sequences in hilbert space. *Bull. Math. Univ. Moscow*, 2(6):1–40, 1941.

Yufan Li. *Physics, Information and Inference: High-Dimensional Models Under Structured Dependencies*. PhD thesis, Harvard University, 2025.

Bruno Loureiro, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, 34:10144–10157, 2021.

Matthew Esmaili Mallory, Kevin Han Huang, and Morgane Austern. Universality of high-dimensional logistic regression and a novel cgmt under dependence with applications to data augmentation. *arXiv preprint arXiv:2502.15752*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

Luca Pesce, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Are gaussian data all you need? the extents and limits of universality in high-dimensional generalized linear estimation. In *International Conference on Machine Learning*, pages 27680–27708. PMLR, 2023.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

Longlin Wang, Yanke Song, Kuanhao Jiang, and Pragya Sur. Glamp: An approximate message passing framework for transfer learning with applications to lasso-based estimators. *arXiv preprint arXiv:2505.22594*, 2025.

Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*. The MIT press, 1964.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26 (7):3142–3155, 2017.

Appendix A. Proof of Theorems 1 and 5: common part

Here we describe the proof in full detail. We start with the objective (16) obtained in the main body and take $i = 1$ WLOG:

$$\min_{\mathbf{w}} \max_{\mathbf{v}_t} \frac{\beta}{2\tau}\Big(\sum_{t=1}^{N} \|\mathbf{\Sigma}_t^{1/2}\mathbf{w}\|_2 \mathbf{h}_t^T \mathbf{v}_t + \|\mathbf{v}_t\|_2 \mathbf{g}_t^T \mathbf{\Sigma}_t^{1/2}\mathbf{w} + \eta \mathbf{v}_t^T \mathbf{h}^{(t)} - \frac{\|\mathbf{v}_t\|_2^2}{4}\Big) - \boldsymbol{\lambda}^T \mathbf{\Sigma}^{1/2}(\mathbf{w} - \mathbf{e}_1) \tag{18}$$

Doing the optimization over the direction of $\mathbf{v}_t$ yields:

$$\min_{\mathbf{w}} \max_{\beta_t > 0} \frac{\beta}{2\tau}\Big(\sum_{t=1}^{N} \beta_t \Big\|\eta \mathbf{h}^{(t)} + \|\mathbf{\Sigma}_t^{1/2}\mathbf{w}\|_2 \mathbf{h}_t\Big\|_2 + \beta_t \mathbf{g}_t^T \mathbf{\Sigma}_t^{1/2}\mathbf{w} - \frac{\beta_t^2}{4}\Big) - \boldsymbol{\lambda}^T \mathbf{\Sigma}^{1/2}(\mathbf{w} - \mathbf{e}_1)$$

Using the square-root trick, we introduce the variables τ_t for $1 \le t \le N$:

$$\min_{\mathbf{w}} \max_{\beta_t > 0} \min_{\tau_t > 0} \frac{\beta}{2\tau}\Big(\sum_{t=1}^{N} \frac{\beta_t \tau_t}{2} + \frac{\beta_t}{2\tau_t}\Big(\eta^2 n_t + \|\mathbf{\Sigma}_t^{1/2}\mathbf{w}\|_2^2 n_t\Big) + \beta_t \mathbf{g}_t^T \mathbf{\Sigma}_t^{1/2}\mathbf{w} - \frac{\beta_t^2}{4}\Big) - \boldsymbol{\lambda}^T \mathbf{\Sigma}^{1/2}(\mathbf{w} - \mathbf{e}_1)$$

Using the Sion's minimax theorem and convex-concavity of objective, we exchange the order of $\min_{\mathbf{w}}$ with $\max_{\beta_t > 0} \min_{\tau_t > 0}$. Now we observe that the optimization over $\mathbf{w}$ is quadratic, after performing the optimization over $\mathbf{w}$, we obtain:

$$\begin{aligned}
\max_{\boldsymbol{\lambda}} \min_{\eta > 0} \max_{\beta > 0} & -\eta \|\beta \mathbf{g} + \boldsymbol{\lambda}\|_2 + \boldsymbol{\lambda}^T \mathbf{\Sigma}^{1/2} \mathbf{e}_1 - \frac{\beta^2}{4} + \min_{\tau > 0} \frac{\beta \tau}{2} \\
& + \max_{\beta_t > 0} \min_{\tau_t > 0} \frac{\beta}{2\tau}\Big(\sum_{t=1}^{N} \frac{\beta_t \tau_t}{2} + \frac{\beta_t}{2\tau_t}\eta^2 n_t - \frac{\beta_t^2}{4}\Big) \\
& - \frac{1}{4}\Big(-\mathbf{\Sigma}^{1/2}\boldsymbol{\lambda} + \frac{\beta}{2\tau}\sum_{t=1}^{N} \beta_t \mathbf{\Sigma}_t^{1/2} \mathbf{g}_t\Big)^T \Big(\frac{\beta}{2\tau}\sum_{t=1}^{N} \frac{\beta_t}{2\tau_t} n_t \mathbf{\Sigma}_t\Big)^{-1} \\
& \cdot \Big(-\mathbf{\Sigma}^{1/2}\boldsymbol{\lambda} + \frac{\beta}{2\tau}\sum_{t=1}^{N} \beta_t \mathbf{\Sigma}_t^{1/2} \mathbf{g}_t\Big)
\end{aligned}$$

Where the optimal $\mathbf{w}$ can be written as

$$\mathbf{w} = -\frac{1}{2}\Big(\frac{\beta}{2\tau}\sum_{t=1}^{N} \frac{\beta_t}{2\tau_t} n_t \mathbf{\Sigma}_t\Big)^{-1}\Big(-\mathbf{\Sigma}^{1/2}\boldsymbol{\lambda} + \frac{\beta}{2\tau}\sum_{t=1}^{N} \beta_t \mathbf{\Sigma}_t^{1/2} \mathbf{g}_t\Big)$$

Hence, we need to find $\boldsymbol{\lambda} \in \mathbb{R}^d$. Employing the square-root trick on the first term:

$$\begin{aligned}\max_{\boldsymbol{\lambda}} \max_{\tau_\lambda>0} \min_{\eta>0} \max_{\beta>0} &- \frac{\eta\tau_\lambda}{2} - \frac{\eta}{2\tau_\lambda}\|\beta\mathbf{g}+\boldsymbol{\lambda}\|_2^2 + \boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{1/2}\mathbf{e}_1 - \frac{\beta^2}{4} + \min_{\tau>0}\frac{\beta\tau}{2} \\ &+ \max_{\beta_t>0}\min_{\tau_t>0}\frac{\beta}{2\tau}\Big(\sum_{t=1}^N \frac{\beta_t\tau_t}{2} + \frac{\beta_t}{2\tau_t}\eta^2 n_t - \frac{\beta_t^2}{4}\Big) \\ &- \frac{1}{4}\Big(-\boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda} + \frac{\beta}{2\tau}\sum_{t=1}^N \beta_t\boldsymbol{\Sigma}_t^{1/2}\mathbf{g}_t\Big)^T\Big(\frac{\beta}{2\tau}\sum_{t=1}^N\frac{\beta_t}{2\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1} \\ &\cdot\Big(-\boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda} + \frac{\beta}{2\tau}\sum_{t=1}^N \beta_t\boldsymbol{\Sigma}_t^{1/2}\mathbf{g}_t\Big)\end{aligned}$$

The optimal $\boldsymbol{\lambda}$ is

$$\begin{aligned}\boldsymbol{\lambda} := &\Big(\frac{\eta}{2\tau_\lambda}\mathbf{I} + \frac{1}{4}\boldsymbol{\Sigma}^{1/2}\Big(\frac{\beta}{2\tau}\sum_{t=1}^N\frac{\beta_t}{2\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\Big)^{-1} \\ &\cdot\Big[-\frac{\eta\beta}{2\tau_\lambda}\mathbf{g} + \frac{1}{4}\boldsymbol{\Sigma}^{1/2}\Big(\frac{\beta}{2\tau}\sum_{t=1}^N\frac{\beta_t}{2\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1}\frac{\beta}{2\tau}\sum_{t=1}^N\beta_t\boldsymbol{\Sigma}_t^{1/2}\mathbf{g}_t + \frac{1}{2}\boldsymbol{\Sigma}^{1/2}\mathbf{e}_1\Big]\end{aligned}$$

This implies the optimal $\mathbf{w}$ is

$$\begin{aligned}&\Big(\frac{\beta}{2\tau}\sum_{t=1}^N\frac{\beta_t}{2\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\Big(\frac{\eta}{2\tau_\lambda}\mathbf{I} + \frac{1}{4}\boldsymbol{\Sigma}^{1/2}\Big(\frac{\beta}{2\tau}\sum_{t=1}^N\frac{\beta_t}{2\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\Big)^{-1}\frac{\eta\beta}{2\tau_\lambda}\mathbf{g} \\ &- \Big(\frac{\beta}{2\tau}\sum_{t=1}^N\frac{\beta_t}{2\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\Big(\frac{\eta}{2\tau_\lambda}\mathbf{I} + \frac{1}{4}\boldsymbol{\Sigma}^{1/2}\Big(\frac{\beta}{2\tau}\sum_{t=1}^N\frac{\beta_t}{2\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\Big)^{-1}\frac{1}{2}\boldsymbol{\Sigma}^{1/2}\mathbf{e}_1 \\ &- \Big(\frac{\beta}{2\tau}\sum_{t=1}^N\frac{\beta_t}{2\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\Big(\frac{\eta}{2\tau_\lambda}\mathbf{I} + \frac{1}{4}\boldsymbol{\Sigma}^{1/2}\Big(\frac{\beta}{2\tau}\sum_{t=1}^N\frac{\beta_t}{2\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\Big)^{-1} \\ &\cdot\frac{1}{4}\boldsymbol{\Sigma}^{1/2}\Big(\frac{\beta}{2\tau}\sum_{t=1}^N\frac{\beta_t}{2\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1}\frac{\beta}{2\tau}\sum_{t=1}^N\beta_t\boldsymbol{\Sigma}_t^{1/2}\mathbf{g}_t \\ &+ \Big(\frac{\beta}{2\tau}\sum_{t=1}^N\frac{\beta_t}{2\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1}\frac{\beta}{2\tau}\sum_{t=1}^N\beta_t\boldsymbol{\Sigma}_t^{1/2}\mathbf{g}_t\end{aligned}$$

Using matrix inversion lemma

$$\begin{aligned}\mathbf{w}^* &= -\frac{1}{2}\Big[4\beta\Big(\frac{\beta}{\tau}\sum_{t=1}^N\frac{\beta_t}{\tau_t}n_t\boldsymbol{\Sigma}_t + \frac{2\tau_\lambda}{\eta}\boldsymbol{\Sigma}\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\mathbf{g} - \frac{4\tau_\lambda}{\eta}\Big(\frac{\beta}{\tau}\sum_{t=1}^N\frac{\beta_t}{\tau_t}n_t\boldsymbol{\Sigma}_t + \frac{2\tau_\lambda}{\eta}\boldsymbol{\Sigma}\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1 \\ &\quad + 4\Big(\frac{\beta}{\tau}\sum_{t=1}^N\frac{\beta_t}{\tau_t}n_t\boldsymbol{\Sigma}_t + \frac{2\tau_\lambda}{\eta}\boldsymbol{\Sigma}\Big)^{-1}\frac{\beta}{2\tau}\sum_{t=1}^N\beta_t\boldsymbol{\Sigma}_t^{1/2}\mathbf{g}_t\Big] \\ &= \Big(\frac{\beta}{\tau}\sum_{t=1}^N\frac{\beta_t}{\tau_t}n_t\boldsymbol{\Sigma}_t + \frac{2\tau_\lambda}{\eta}\boldsymbol{\Sigma}\Big)^{-1}\Big(-2\beta\boldsymbol{\Sigma}^{1/2}\mathbf{g} + \frac{2\tau_\lambda}{\eta}\boldsymbol{\Sigma}\mathbf{e}_1 - \frac{\beta}{\tau}\sum_{t=1}^N\beta_t\boldsymbol{\Sigma}_t^{1/2}\mathbf{g}_t\Big)\end{aligned}$$

Performing the optimization over $\boldsymbol{\lambda}$ yields the following scalar optimization:

$$\begin{aligned}
\max_{\tau_\lambda>0}\min_{\eta>0}\max_{\beta>0} &-\frac{\eta\tau_\lambda}{2}-\frac{\beta^2}{4}+\min_{\tau>0}\frac{\beta\tau}{2}+\max_{\beta_t}\min_{\tau_t}\frac{\beta}{2\tau}\Big(\sum_{t=1}^N\frac{\beta_t\tau_t}{2}+\frac{\beta_t}{2\tau_t}\eta^2 n_t-\frac{\beta_t^2}{4}\Big)\\
&-\beta^2\mathbf{g}^T\Big(\frac{2\tau_\lambda}{\eta}\mathbf{I}+\frac{\beta}{\tau}\sum_{t=1}^N\frac{\beta_t}{\tau_t}n_t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}^{-1/2}\Big)^{-1}\mathbf{g}\\
&-\frac{\beta^2}{4\tau^2}\Big(\sum_{t=1}^N\beta_t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_t^{1/2}\mathbf{g}_t\Big)^T\Big(\frac{2\tau_\lambda}{\eta}\mathbf{I}+\frac{\beta}{\tau}\sum_{t=1}^N\frac{\beta_t}{\tau_t}n_t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}^{-1/2}\Big)^{-1}\\
&\cdot\sum_{t=1}^N\beta_t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_t^{1/2}\mathbf{g}_t\\
&+\frac{1}{4}\mathbf{e}_1^T\boldsymbol{\Sigma}^{1/2}\Big(\frac{\eta}{2\tau_\lambda}\mathbf{I}+\frac{1}{4}\boldsymbol{\Sigma}^{1/2}\Big(\frac{\beta}{2\tau}\sum_{t=1}^N\frac{\beta_t}{2\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\mathbf{e}_1
\end{aligned}$$

We apply the matrix inversion lemma

$$\begin{aligned}
&\mathbf{e}_1^T\boldsymbol{\Sigma}^{1/2}\Big(\frac{\eta}{2\tau_\lambda}\mathbf{I}+\frac{1}{4}\boldsymbol{\Sigma}^{1/2}\Big(\frac{\beta}{2\tau}\sum_{t=1}^N\frac{\beta_t}{2\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\mathbf{e}_1\\
&=\frac{2\tau_\lambda}{\eta}\mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1-\frac{4\tau_\lambda^2}{\eta^2}\mathbf{e}_1^T\boldsymbol{\Sigma}^{1/2}\Big(\frac{2\tau_\lambda}{\eta}\mathbf{I}+\frac{\beta}{\tau}\sum_{t=1}^N\frac{\beta_t}{\tau_t}n_t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}^{-1/2}\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\mathbf{e}_1
\end{aligned}$$

By Hanson-Wright inequality, we observe that:

$$\begin{aligned}
&\mathbf{g}^T\Big(\frac{2\tau_\lambda}{\eta}\mathbf{I}+\frac{\beta}{\tau}\sum_{t=1}^N\frac{\beta_t}{\tau_t}n_t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}^{-1/2}\Big)^{-1}\mathbf{g}\xrightarrow{\mathbb{P}}\mathrm{Tr}\Big(\frac{2\tau_\lambda}{\eta}\mathbf{I}+\frac{\beta}{\tau}\sum_{t=1}^N\frac{\beta_t}{\tau_t}n_t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}^{-1/2}\Big)^{-1}\\
&\Big(\sum_{t=1}^N\beta_t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_t^{1/2}\mathbf{g}_t\Big)^T\Big(\frac{2\tau_\lambda}{\eta}\mathbf{I}+\frac{\beta}{\tau}\sum_{t=1}^N\frac{\beta_t}{\tau_t}n_t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}^{-1/2}\Big)^{-1}\sum_{t=1}^N\beta_t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_t^{1/2}\mathbf{g}_t\\
&\xrightarrow{\mathbb{P}}\mathrm{Tr}\Big[\Big(\frac{2\tau_\lambda}{\eta}\mathbf{I}+\frac{\beta}{\tau}\sum_{t=1}^N\frac{\beta_t}{\tau_t}n_t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}^{-1/2}\Big)^{-1}\sum_{t=1}^N\beta_t^2\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}^{-1/2}\Big]
\end{aligned}$$

Thus the final scalar optimization turns into

$$\max_{\tau_\lambda>0}\min_{\eta>0}\max_{\beta>0}\min_{\tau>0}\max_{\beta_t>0}\min_{\tau_t>0}-\frac{\eta\tau_\lambda}{2}-\frac{\beta^2}{4}+\frac{\beta\tau}{2}+\frac{\beta}{2\tau}\Big(\sum_{t=1}^N\frac{\beta_t\tau_t}{2}+\frac{\beta_t}{2\tau_t}\eta^2 n_t-\frac{\beta_t^2}{4}\Big)+\frac{\tau_\lambda}{2\eta}\mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 \tag{19}$$

$$-\mathrm{Tr}\Big[\Big(\frac{2\tau_\lambda}{\eta}\boldsymbol{\Sigma}+\frac{\beta}{\tau}\sum_{t=1}^N\frac{\beta_t}{\tau_t}n_t\boldsymbol{\Sigma}_t\Big)^{-1}\Big(\beta^2\boldsymbol{\Sigma}+\frac{\beta^2}{4\tau^2}\sum_{t=1}^N\beta_t^2\boldsymbol{\Sigma}_t+\frac{\tau_\lambda^2}{\eta^2}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big)\Big] \tag{20}$$

We define

$$F\Big(\frac{\beta}{\tau},\frac{\beta_t}{\tau_t},\frac{\eta}{\tau_\lambda},\beta,\beta_t\Big) := \mathrm{Tr}\Big[\Big(\frac{2\tau_\lambda}{\eta}\mathbf{\Sigma}+\frac{\beta}{\tau}\sum_{t=1}^{N}\frac{\beta_t}{\tau_t}n_t\mathbf{\Sigma}_t\Big)^{-1}\Big(\beta^2\mathbf{\Sigma}+\frac{\beta^2}{4\tau^2}\sum_{t=1}^{N}\beta_t^2\mathbf{\Sigma}_t+\frac{\tau_\lambda^2}{\eta^2}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big)\Big]$$

Now we compute the stationary coditions for the scalar optimization in (19). For the derivatives w.r.t β and τ

$$\begin{aligned}\frac{\partial}{\partial\beta}=0\Rightarrow 0&=-\frac{\beta}{2}+\frac{\tau}{2}+\frac{1}{2\tau}\Big(\sum_{t=1}^{N}\frac{\beta_t\tau_t}{2}+\frac{\beta_t}{2\tau_t}\eta^2n_t-\frac{\beta_t^2}{4}\Big)\\&\quad-\frac{1}{\tau}\partial_1F\Big(\frac{\beta}{\tau},\frac{\beta_t}{\tau_t},\frac{\eta}{\tau_\lambda},\beta,\beta_t\Big)-2\beta\mathrm{Tr}\Big[\Big(\frac{2\tau_\lambda}{\eta}\mathbf{\Sigma}+\frac{\beta}{\tau}\sum_{t=1}^{N}\frac{\beta_t}{\tau_t}n_t\mathbf{\Sigma}_t\Big)^{-1}\mathbf{\Sigma}\Big]\\\frac{\partial}{\partial\tau}=0\Rightarrow 0&=\frac{\beta}{2}-\frac{\beta}{2\tau^2}\Big(\sum_{t=1}^{N}\frac{\beta_t\tau_t}{2}+\frac{\beta_t}{2\tau_t}\eta^2n_t-\frac{\beta_t^2}{4}\Big)+\frac{\beta}{\tau^2}\partial_1F\Big(\frac{\beta}{\tau},\frac{\beta_t}{\tau_t},\frac{\eta}{\tau_\lambda},\beta,\beta_t\Big)\end{aligned}$$

Thus for $\theta := \frac{\beta}{\tau}$ we have

$$\begin{aligned}\frac{\theta}{2}&=1-2\theta\mathrm{Tr}\Big(\frac{2\tau_\lambda}{\eta}\mathbf{\Sigma}+\theta\sum_{t=1}^{N}\frac{\beta_t}{\tau_t}n_t\mathbf{\Sigma}_t\Big)^{-1}\mathbf{\Sigma}\\\tau^2&=\sum_{t=1}^{N}\frac{\beta_t\tau_t}{2}+\frac{\beta_t}{2\tau_t}\eta^2n_t-\frac{\beta_t^2}{4}-2\partial_1F\Big(\frac{\beta}{\tau},\frac{\beta_t}{\tau_t},\frac{\eta}{\tau_\lambda},\beta,\beta_t\Big)\end{aligned}$$

Where

$$\begin{aligned}\partial_1F\Big(\frac{\beta}{\tau},\frac{\beta_t}{\tau_t},\frac{\eta}{\tau_\lambda},\beta,\beta_t\Big)&=-\mathrm{Tr}\Big[\Big(\frac{2\tau_\lambda}{\eta}\mathbf{\Sigma}+\frac{\beta}{\tau}\sum_{t=1}^{N}\frac{\beta_t}{\tau_t}n_t\mathbf{\Sigma}_t\Big)^{-1}\Big(\sum_{t=1}^{N}\frac{\beta_t}{\tau_t}n_t\mathbf{\Sigma}_t\Big)\\&\qquad\cdot\Big(\frac{2\tau_\lambda}{\eta}\mathbf{\Sigma}+\frac{\beta}{\tau}\sum_{t=1}^{N}\frac{\beta_t}{\tau_t}n_t\mathbf{\Sigma}_t\Big)^{-1}\Big(\beta^2\mathbf{\Sigma}+\frac{\beta^2}{4\tau^2}\sum_{t=1}^{N}\beta_t^2\mathbf{\Sigma}_t+\frac{\tau_\lambda^2}{\eta^2}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big)\Big]\\&\quad+\frac{\beta}{2\tau}\mathrm{Tr}\Big[\Big(\frac{2\tau_\lambda}{\eta}\mathbf{\Sigma}+\frac{\beta}{\tau}\sum_{t=1}^{N}\frac{\beta_t}{\tau_t}n_t\mathbf{\Sigma}_t\Big)^{-1}\sum_{t=1}^{N}\beta_t^2\mathbf{\Sigma}_t\Big]\end{aligned}$$

Similarly for η and τ_λ

$$\begin{aligned}\frac{\partial}{\partial\eta}=0\Rightarrow 0&=-\frac{\tau_\lambda}{2}+\frac{\beta\eta n_t}{\tau}\sum_{t=1}^{N}\frac{\beta_t}{2\tau_t}-\frac{\tau_\lambda}{2\eta^2}\mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1-\frac{1}{\tau_\lambda}\partial_3F\Big(\frac{\beta}{\tau},\frac{\beta_t}{\tau_t},\frac{\eta}{\tau_\lambda},\beta,\beta_t\Big)\\\frac{\partial}{\partial\tau_\lambda}=0\Rightarrow 0&=-\frac{\eta}{2}+\frac{1}{2\eta}\mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1+\frac{\eta}{\tau_\lambda^2}\partial_3F\Big(\frac{\beta}{\tau},\frac{\beta_t}{\tau_t},\frac{\eta}{\tau_\lambda},\beta,\beta_t\Big)\end{aligned}$$

Thus for $\zeta := \frac{\eta}{\tau_\lambda}$ we have

$$\begin{aligned}\frac{1}{\zeta}&=\theta n_t\sum_{t=1}^{N}\frac{\beta_t}{2\tau_t}\\\tau_\lambda^2&=\frac{1}{\zeta^2}\mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1+2\partial_3F\Big(\frac{\beta}{\tau},\frac{\beta_t}{\tau_t},\frac{\eta}{\tau_\lambda},\beta,\beta_t\Big)\end{aligned}$$

Where

$$\partial_3 F\Big(\frac{\beta}{\tau}, \frac{\beta_t}{\tau_t}, \frac{\eta}{\tau_\lambda}, \beta, \beta_t\Big) = \frac{2\tau_\lambda^2}{\eta^2}\mathrm{Tr}\Big[\Big(\frac{2\tau_\lambda}{\eta}\mathbf{\Sigma} + \frac{\beta}{\tau}\sum_{t=1}^{N}\frac{\beta_t}{\tau_t}n_t\mathbf{\Sigma}_t\Big)^{-1}\mathbf{\Sigma} \\ \Big(\frac{2\tau_\lambda}{\eta}\mathbf{\Sigma} + \frac{\beta}{\tau}\sum_{t=1}^{N}\frac{\beta_t}{\tau_t}n_t\mathbf{\Sigma}_t\Big)^{-1}\Big(\beta^2\mathbf{\Sigma} + \frac{\beta^2}{4\tau^2}\sum_{t=1}^{N}\beta_t^2\mathbf{\Sigma}_t + \frac{\tau_\lambda^2}{\eta^2}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big)\Big] \\ - 2\frac{\tau_\lambda^3}{\eta^3}\mathrm{Tr}\Big[\Big(\frac{2\tau_\lambda}{\eta}\mathbf{\Sigma} + \frac{\beta}{\tau}\sum_{t=1}^{N}\frac{\beta_t}{\tau_t}n_t\mathbf{\Sigma}_t\Big)^{-1}\mathbf{\Sigma}^{1/2}\mathbf{e}_1\mathbf{e}_1\mathbf{\Sigma}^{1/2}\Big]$$

Similarly for β_t and τ_t

$$\frac{\partial}{\partial\beta_t} = 0 \Rightarrow 0 = \frac{\beta}{2\tau}\Big(\frac{\tau_t}{2} + \frac{\eta^2 n_t}{2\tau_t} - \frac{\beta_t}{2}\Big) - \frac{1}{\tau_t}\partial_2 F\Big(\frac{\beta}{\tau}, \frac{\beta_t}{\tau_t}, \frac{\eta}{\tau_\lambda}, \beta, \beta_t\Big) \\ - \frac{\beta^2\beta_t}{2\tau^2}\mathrm{Tr}\Big[\Big(\frac{2\tau_\lambda}{\eta}\mathbf{\Sigma} + \frac{\beta}{\tau}\sum_{t=1}^{N}\frac{\beta_t}{\tau_t}n_t\mathbf{\Sigma}_t\Big)^{-1}\mathbf{\Sigma}_t\Big]$$

$$\frac{\partial}{\partial\tau_t} = 0 \Rightarrow 0 = \frac{\beta}{2\tau}\Big(\frac{\beta_t}{2} - \frac{\beta_t\eta^2 n_t}{2\tau_t^2}\Big) + \frac{\beta_t}{\tau_t^2}\partial_2 F\Big(\frac{\beta}{\tau}, \frac{\beta_t}{\tau_t}, \frac{\eta}{\tau_\lambda}, \beta, \beta_t\Big)$$

This implies for $\theta_t := \frac{\beta_t}{\tau_t}$

$$\theta_t = 2 - 2\theta\theta_t\mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\mathbf{\Sigma} + \theta\sum_{t=1}^{N}\theta_t n_t\mathbf{\Sigma}_t\Big)^{-1}\mathbf{\Sigma}_t\Big]$$

$$\tau_t^2 = \zeta^2\tau_\lambda^2 n_t - \frac{4}{\theta}\partial_2 F\Big(\frac{\beta}{\tau}, \frac{\beta_t}{\tau_t}, \frac{\eta}{\tau_\lambda}, \beta, \beta_t\Big)$$

Where

$$\partial_2 F\Big(\frac{\beta}{\tau}, \frac{\beta_t}{\tau_t}, \frac{\eta}{\tau_\lambda}, \beta, \beta_t\Big) = -\frac{\beta n_t}{\tau}\mathrm{Tr}\Big[\Big(\frac{2\tau_\lambda}{\eta}\mathbf{\Sigma} + \frac{\beta}{\tau}\sum_{t=1}^{N}\frac{\beta_t}{\tau_t}n_t\mathbf{\Sigma}_t\Big)^{-1}\mathbf{\Sigma}_t \\ \cdot\Big(\frac{2\tau_\lambda}{\eta}\mathbf{\Sigma} + \frac{\beta}{\tau}\sum_{t=1}^{N}\frac{\beta_t}{\tau_t}n_t\mathbf{\Sigma}_t\Big)^{-1}\Big(\beta^2\mathbf{\Sigma} + \frac{\beta^2}{4\tau^2}\sum_{t=1}^{N}\beta_t^2\mathbf{\Sigma}_t + \frac{\tau_\lambda^2}{\eta^2}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big)\Big]$$

Now we summarize the stationary conditions as follows:

$$
\begin{aligned}
\frac{\theta}{2} &= 1 - 2\theta \mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\mathbf{\Sigma} + \theta\sum_{t=1}^{N}\theta_t n_t \mathbf{\Sigma}_t\Big)^{-1}\mathbf{\Sigma}\Big] \\
\frac{2}{\zeta} &= \theta n_t \sum_{t=1}^{N}\theta_t \\
\theta_t &= 2 - 2\theta\theta_t \mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\mathbf{\Sigma} + \theta\sum_{t=1}^{N}\theta_t n_t \mathbf{\Sigma}_t\Big)^{-1}\mathbf{\Sigma}_t\Big] \\
\tau^2 &= \sum_{t=1}^{N}\frac{\theta_t\tau_t^2}{2} + \frac{\theta_t}{2}\zeta^2\tau_\lambda^2 n_t - \frac{\tau_t^2\theta_t^2}{4} - 2\partial_1 F\Big(\frac{\beta}{\tau}, \frac{\beta_t}{\tau_t}, \frac{\eta}{\tau_\lambda}, \beta, \beta_t\Big) \\
\tau_t^2 &= \zeta^2\tau_\lambda^2 n_t - \frac{4}{\theta}\partial_2 F\Big(\theta, \theta_t, \zeta, \beta, \beta_t\Big) \\
\tau_\lambda^2 &= \frac{1}{\zeta^2}\mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1 + 2\partial_3 F\Big(\theta, \theta_t, \zeta, \beta, \beta_t\Big) \\
\partial_1 F\Big(\theta, \theta_t, \zeta, \beta, \beta_t\Big) &= -\mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\mathbf{\Sigma} + \theta\sum_{t=1}^{N}\theta_t n_t \mathbf{\Sigma}_t\Big)^{-1}\Big(\sum_{t=1}^{N}\theta_t n_t \mathbf{\Sigma}_t\Big)\Big(\frac{2}{\zeta}\mathbf{\Sigma} + \theta\sum_{t=1}^{N}\theta_t n_t \mathbf{\Sigma}_t\Big)^{-1} \\
&\quad \cdot \Big(\theta^2\tau^2\mathbf{\Sigma} + \frac{\theta^2}{4}\sum_{t=1}^{N}\theta_t^2\tau_t^2\mathbf{\Sigma}_t + \frac{1}{\zeta^2}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big)\Big] \\
&\quad + \frac{\theta}{2}\mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\mathbf{\Sigma} + \theta\sum_{t=1}^{N}\theta_t n_t \mathbf{\Sigma}_t\Big)^{-1}\sum_{t=1}^{N}\theta_t^2\tau_t^2\mathbf{\Sigma}_t\Big] \\
\partial_2 F\Big(\theta, \theta_t, \zeta, \beta, \beta_t\Big) &= -\theta n_t \mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\mathbf{\Sigma} + \theta\sum_{t=1}^{N}\theta_t n_t \mathbf{\Sigma}_t\Big)^{-1}\mathbf{\Sigma}_t\Big(\frac{2}{\zeta}\mathbf{\Sigma} + \theta\sum_{t=1}^{N}\theta_t n_t \mathbf{\Sigma}_t\Big)^{-1} \\
&\quad \cdot \Big(\theta^2\tau^2\mathbf{\Sigma} + \frac{\theta^2}{4}\sum_{t=1}^{N}\theta_t^2\tau_t^2\mathbf{\Sigma}_t + \frac{1}{\zeta^2}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big)\Big] \\
\partial_3 F\Big(\theta, \theta_t, \zeta, \beta, \beta_t\Big) &= \frac{2}{\zeta^2}\mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\mathbf{\Sigma} + \theta\sum_{t=1}^{N}\theta_t n_t \mathbf{\Sigma}_t\Big)^{-1}\mathbf{\Sigma}\Big(\frac{2}{\zeta}\mathbf{\Sigma} + \theta\sum_{t=1}^{N}\theta_t n_t \mathbf{\Sigma}_t\Big)^{-1} \\
&\quad \cdot \Big(\theta^2\tau^2\mathbf{\Sigma} + \frac{\theta^2}{4}\sum_{t=1}^{N}\theta_t^2\tau_t^2\mathbf{\Sigma}_t + \frac{1}{\zeta^2}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big)\Big] \\
&\quad - \frac{2}{\zeta^3}\mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\mathbf{\Sigma} + \theta\sum_{t=1}^{N}\theta_t n_t \mathbf{\Sigma}_t\Big)^{-1}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big]
\end{aligned} \tag{21}
$$

At the optimal points, we have for $\mathbf{w}$

$$
\mathbf{w}^* = \Big(\theta\sum_{t=1}^{N}\theta_t n_t \mathbf{\Sigma}_t + \frac{2}{\zeta}\mathbf{\Sigma}\Big)^{-1}\Big(-2\theta\tau\mathbf{\Sigma}^{1/2}\mathbf{g} + 2\mathbf{\Sigma}\mathbf{e}_1 - \theta\sum_{t=1}^{N}\theta_t\tau_t\mathbf{\Sigma}_t^{1/2}\mathbf{g}_t\Big)
$$

The generalization error corresponding to this term ($i = 1$) is

$$\begin{aligned}
&\eta^2 + \mathbf{w}^T \mathbf{\Sigma_z} \mathbf{w} \\
&\xrightarrow{\mathbb{P}} \eta^2 + \mathrm{Tr}\Big[\Big(\theta \sum_{t=1}^{N} \theta_t n_t \mathbf{\Sigma}_t + \frac{2}{\zeta}\mathbf{\Sigma}\Big)^{-1} \mathbf{\Sigma_z} \Big(\theta \sum_{t=1}^{N} \theta_t n_t \mathbf{\Sigma}_t + \frac{2}{\zeta}\mathbf{\Sigma}\Big)^{-1} \\
&\quad \cdot \Big(4\theta^2 \tau^2 \mathbf{\Sigma} + \frac{4}{\zeta^2}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma} + \theta^2 \sum_{t=1}^{N} \theta_t^2 \tau_t^2 \mathbf{\Sigma}_t\Big)\Big]
\end{aligned}$$

Lemma 6 *The function $f(\beta) := \|\beta\mathbf{g} + \boldsymbol{\lambda}\|_2$ for $\beta \in \mathbb{R}$ and any $\mathbf{g}, \boldsymbol{\lambda} \in \mathbb{R}^d$ is convex in $\beta \in [0, \infty)$.*

Proof We proceed by verifying the definition of convexity. Let $0 \le t \le 1$, then for any $\beta_1, \beta_2 \ge 0$ we have

$$\begin{aligned}
tf(\beta_1) + (1-t)f(\beta_2) &= \|t\beta_1\mathbf{g} + t\boldsymbol{\lambda}\|_2 + \|(1-t)\beta_1\mathbf{g} + (1-t)\boldsymbol{\lambda}\|_2 \\
&\ge \|((t\beta_1 + (1-t)\beta_2)\mathbf{g} + \boldsymbol{\lambda}\|_2
\end{aligned}$$

Where the inequality follows by the convexity of ℓ_2 on vectors $\beta_1\mathbf{g} + \boldsymbol{\lambda}$ and $\beta_2\mathbf{g} + \boldsymbol{\lambda}$. Hence

$$tf(\beta_1) + (1-t)f(\beta_2) \ge f(t\beta_1 + (1-t)\beta_2)$$

And the result follows. ■

We now consider $N = 1$.

A.1. End of the proof of Theorem 1

The system in (21) turns into

$$\begin{aligned}
\frac{\theta}{2} &= 1 - 2\theta \mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\mathbf{\Sigma} + \theta\theta_1\|\mathbf{h}_1\|_2^2\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}\Big] \\
\frac{2}{\zeta} &= \theta\|\mathbf{h}\|_2^2\theta_1 \\
\theta_1 &= 2 - 2\theta\theta_1 \mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\mathbf{\Sigma} + \theta\theta_1\|\mathbf{h}_1\|_2^2\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}_1\Big]
\end{aligned}$$

Which implies

$$\begin{aligned}
\theta_1 &= 2 - \frac{2}{\|\mathbf{h}\|_2^2}\mathrm{Tr}\Big[\Big(\mathbf{\Sigma} + \mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}_1\Big] = \frac{2n - 2\mathrm{Tr}\Big[\Big(\mathbf{\Sigma} + \mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}_1\Big]}{n} \\
\theta\zeta &= \frac{2}{\theta_1\|\mathbf{h}\|_2^2} \\
\frac{\theta}{2} &= 1 - \zeta\theta\mathrm{Tr}\Big[\Big(\mathbf{\Sigma} + \mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}\Big]
\end{aligned}$$

Hence, by plugging in for θ_1, θ would be

$$\theta = 2 - \frac{2\mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}\Big]}{\|\mathbf{h}\|_2^2 - \mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}_1\Big]} = \frac{2n-2d}{n - \mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}_1\Big]}$$
$$\zeta = \frac{1}{2(n-d)}$$

Furthermore, we have for the other variables

$$\tau^2 = \frac{\theta_1\tau_1^2}{2} + \frac{\theta_1}{2}\zeta^2\tau_\lambda^2\|\mathbf{h}\|_2^2 - \frac{\theta_1^2\tau_1^2}{4} - 2\partial_1 F\Big(\frac{\beta}{\tau}, \frac{\beta_t}{\tau_t}, \frac{\eta}{\tau_\lambda}, \beta, \beta_t\Big)$$
$$\tau_1^2 = \zeta^2\tau_\lambda^2\|\mathbf{h}\|_2^2 - \frac{4}{\theta}\partial_2 F\Big(\theta, \theta_t, \zeta, \beta, \beta_1\Big)$$
$$\tau_\lambda^2 = \frac{1}{\zeta^2}\mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1 + 2\partial_3 F\Big(\theta, \theta_1, \zeta, \beta, \beta_t\Big)$$

Plugging in

$$\tau^2 = \frac{\theta_1\tau_1^2}{2} + \frac{\theta_1}{2}\zeta^2\tau_\lambda^2\|\mathbf{h}\|_2^2 - \frac{\theta_1^2\tau_1^2}{4} - \frac{\theta\zeta\theta_1^2\tau_1^2}{2}\mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}_1\Big]$$
$$+ \frac{\zeta^2\theta_1\|\mathbf{h}_1\|_2^2}{2}\mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}_1\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\Big(\theta^2\tau^2\mathbf{\Sigma} + \frac{\theta^2\theta_1^2\tau_1^2}{4}\mathbf{\Sigma}_1 + \frac{1}{\zeta^2}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big)\Big]$$
$$\tau_1^2 = \zeta^2\tau_\lambda^2\|\mathbf{h}\|_2^2 + \zeta^2\|\mathbf{h}\|_2^2\mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}_1\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\Big(\theta^2\tau^2\mathbf{\Sigma} + \frac{\theta^2}{4}\theta_1^2\tau_1^2\mathbf{\Sigma}_1 + \frac{1}{\zeta^2}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big)\Big]$$
$$\tau_\lambda^2 = \frac{1}{\zeta^2}\mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1 + \mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\Big(\theta^2\tau^2\mathbf{\Sigma} + \frac{\theta^2}{4}\theta_1^2\tau_1^2\mathbf{\Sigma}_1 + \frac{1}{\zeta^2}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big)\Big]$$
$$- \frac{2}{\zeta^2}\mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big]$$

Eliminating τ_λ^2 yields

$$\tau^2 = \frac{\theta_1\tau_1^2}{2} - \frac{\theta_1^2\tau_1^2}{4} + \frac{\zeta^2\theta_1\|\mathbf{h}_1\|_2^2\theta^2}{2}\mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}\Big]\tau^2 + \frac{\zeta^2\theta_1\|\mathbf{h}_1\|_2^2}{2}\frac{\theta^2\theta_1^2}{4}\mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}_1\Big]\tau_1^2$$
$$+ \frac{\theta_1\|\mathbf{h}_1\|_2^2}{2}\mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1 - \frac{\theta\zeta\theta_1^2\tau_1^2}{2}\mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}_1\Big] - \frac{\theta_1\|\mathbf{h}\|_2^2}{2}\mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big]$$
$$\tau_1^2 = \zeta^2\|\mathbf{h}\|_2^2\theta^2\mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}\Big]\tau^2 + \frac{\theta^2\theta_1^2\zeta^2\|\mathbf{h}\|_2^2}{4}\mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}_1\Big]\tau_1^2 + \|\mathbf{h}\|_2^2\mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1$$
$$- \|\mathbf{h}\|_2^2\mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big]$$
$$\tau_\lambda^2 = \frac{1}{\zeta^2}\mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1 + \mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\Big(\theta^2\tau^2\mathbf{\Sigma} + \frac{\theta^2}{4}\theta_1^2\tau_1^2\mathbf{\Sigma}_1 + \frac{1}{\zeta^2}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big)\Big]$$
$$- \frac{2}{\zeta^2}\mathrm{Tr}\Big[\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_1\Big)^{-1}\mathbf{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\mathbf{\Sigma}\Big]$$

By simplifying the expressions, we obtain

$$\frac{\theta}{2}\tau^2 = \frac{\theta_1\|\mathbf{h}_1\|_2^2}{2}\Big(\mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 - \mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big]\Big)$$
$$\frac{\theta_1}{2}\tau_1^2 = \frac{2-\theta}{\theta_1}\tau^2 + \|\mathbf{h}\|_2^2\mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 - \|\mathbf{h}\|_2^2\mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big]$$

Hence τ_1^2 can be written as follows

$$\tau_1^2 = \frac{4\|\mathbf{h}\|_2^2}{\theta\theta_1}\mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 - \frac{4\|\mathbf{h}\|_2^2}{\theta\theta_1}\mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big]$$

Furthermore, we have for τ_λ^2

$$\begin{aligned}\zeta^2\tau_\lambda^2 &= \mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 + \mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big] - 2\mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big]\\ &\quad + 2\zeta\Big(\mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 - \mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big]\Big)\mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\Big]\\ &= \frac{2}{\theta}\mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 - \frac{\theta+2}{\theta}\mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big] + \mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big]\end{aligned}$$

Summarizing for $\eta^2 = \zeta^2\tau_\lambda^2$

$$\begin{aligned}\mathrm{Tr}\Big[(\mathbf{W}^T-\mathbf{I})^T\boldsymbol{\Sigma}(\mathbf{W}^T-\mathbf{I})\Big] &\xrightarrow{\mathbb{P}} \frac{2}{\theta}\mathrm{Tr}\Big[\boldsymbol{\Sigma}\Big] - \frac{\theta+2}{\theta}\mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}^2\Big]\\ &\quad + \mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}^2\Big]\\ &= \mathrm{Tr}\Big[\Big(\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}-\mathbf{I}\Big)\Big(\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma} - \frac{n}{n-d+\mathrm{Tr}\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}}\mathbf{I}\Big)\boldsymbol{\Sigma}\Big]\end{aligned}$$

Denoting $\mathbf{R}_W := \Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}$ as the Wiener filter, we obtain

$$\mathrm{Tr}\Big[(\mathbf{W}^T-\mathbf{I})^T\boldsymbol{\Sigma}(\mathbf{W}^T-\mathbf{I})\Big] \xrightarrow{\mathbb{P}} \mathrm{Tr}\Big[\Big(\mathbf{R}_W-\mathbf{I}\Big)\Big(\mathbf{R}_W - \frac{n}{n-d+\mathrm{Tr}\mathbf{R}_W}\mathbf{I}\Big)\boldsymbol{\Sigma}\Big]$$

If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_z$ and $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}$ commute, we have for the generalization error

$$\mathcal{E}(\mathbf{W}_{\text{lsq}}) \xrightarrow{\mathbb{P}} \frac{n}{n-d}\Big(\mathrm{Tr}\boldsymbol{\Sigma} - \mathrm{Tr}\mathbf{R}_W\boldsymbol{\Sigma}\Big) = \frac{n}{n-d}\mathcal{E}_{\text{Wiener}}$$

More generally, we may write

$$\mathcal{E}(\mathbf{W}_{\text{lsq}}) - \mathcal{E}_{\text{Wiener}} \xrightarrow{\mathbb{P}} \mathrm{Tr}\Big[\boldsymbol{\Sigma}\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}_1\Big]\Big[\frac{n}{n-d}\mathbf{I} - \Big(\boldsymbol{\Sigma}_1\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1} + \Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\Big)\Big]$$

The error per term is

$$\begin{aligned}
&\zeta^2\tau_\lambda^2 + \zeta^2 \mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}\Big)^{-1}\boldsymbol{\Sigma}_{\mathbf{z}}\Big(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}\Big)^{-1}\Big(\theta^2\tau^2\boldsymbol{\Sigma}+\frac{1}{\zeta^2}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}+\frac{\theta^2\theta_1^2\tau_1^2}{4}\boldsymbol{\Sigma}_1\Big)\Big] \\
&\quad = \zeta^2\tau_\lambda^2 + \mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}\Big)^{-1}\boldsymbol{\Sigma}_{\mathbf{z}}\Big(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big] \\
&\quad + 2\zeta\Big(\mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 - \mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big]\Big)\mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}_{\mathbf{z}}\Big] \\
&\quad = \mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 + \mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_{\mathbf{z}}\Big)\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big] - 2\mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big] \\
&\quad + 2\zeta\Big(\mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 - \mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big]\Big)\mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_{\mathbf{z}}\Big)\Big]
\end{aligned}$$

Thus the total error is

$$\begin{aligned}
\mathcal{E}(\mathbf{W}_{\text{lsq}}) \xrightarrow{\mathbb{P}} \Big(1+2\zeta\mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_{\mathbf{z}}\Big)\Big]\Big)\Big(\mathrm{Tr}\Big[\boldsymbol{\Sigma}\Big] - \mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}^2\Big]\Big) \\
+\mathrm{Tr}\Big[\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\Big(\boldsymbol{\Sigma}_{\mathbf{z}}-\boldsymbol{\Sigma}_1\Big)\Big(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_1\Big)^{-1}\boldsymbol{\Sigma}^2\Big]
\end{aligned}$$

A.2. End of the Proof of Theorem 5.

Recall that we had from (21)

$$\begin{aligned}
\frac{\theta}{2} &= 1 - 2\theta\mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\boldsymbol{\Sigma}+\theta\sum_{t=1}^N\theta_t n_t\boldsymbol{\Sigma}_t\Big)^{-1}\boldsymbol{\Sigma}\Big] \\
\frac{2}{\zeta} &= \theta\sum_{t=1}^N n_t\theta_t \\
\theta_t &= 2 - 2\theta\theta_t\mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\boldsymbol{\Sigma}+\theta\sum_{t=1}^N\theta_t n_t\boldsymbol{\Sigma}_t\Big)^{-1}\boldsymbol{\Sigma}_t\Big]
\end{aligned}$$

This implies

$$\begin{aligned}
\theta_t &= 2 - 2\theta_t\mathrm{Tr}\Big[\Big(\sum_{t=1}^N\theta_t n_t(\boldsymbol{\Sigma}_t+\boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}_t\Big] \\
\zeta &= \frac{1}{2(n-d)} \\
\frac{\theta}{2} &= 1 - 2\theta\mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\boldsymbol{\Sigma}+\theta\sum_{t=1}^N\theta_t n_t\boldsymbol{\Sigma}_t\Big)^{-1}\boldsymbol{\Sigma}\Big]
\end{aligned}$$

Hence

$$\begin{aligned}\frac{\theta}{2}\tau^2 &= \frac{1}{\zeta\theta}\mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 - \frac{2}{\zeta^2\theta}\mathrm{Tr}\Big[\Big(\frac{2}{\zeta}\boldsymbol{\Sigma} + \theta\sum_{t=1}^N \theta_t n_t \boldsymbol{\Sigma}_t\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big] \\ &= \frac{1}{\zeta\theta}\mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 - \frac{2}{\zeta^2\theta^2}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big] \end{aligned} \tag{22}$$

Furthermore, τ_i^2 can be found through the following system of linear equations:

$$\begin{aligned}&- 4\tau^2\|\mathbf{h}^{(i)}\|_2^2\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma})\Big)^{-1}\Big(\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_t\Big)\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}\Big] \\ &+ \Big(1 - \|\mathbf{h}^{(i)}\|_2^2\theta_i^2\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\Big(\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_i\Big)\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}_i\Big]\Big)\tau_i^2 \\ &- \sum_{t\neq i}^N n_t\theta_t^2\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\Big(\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_i\Big)\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}_t\Big]\tau_t^2 \\ &= 4\frac{\|\mathbf{h}^{(i)}\|_2^2}{\theta^2\zeta^2}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma})\Big)^{-1}\Big(\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_t\Big)\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big] \\ &- \frac{4\|\mathbf{h}^{(i)}\|_2^2}{\zeta\theta}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big] \end{aligned} \tag{23}$$

Furthermore, we have for τ_λ^2

$$\begin{aligned}\tau_\lambda^2 = &\frac{1}{\zeta^2}\mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 + \frac{4}{\zeta^2\theta^2}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1} \\ &\cdot\Big(\theta^2\tau^2\boldsymbol{\Sigma} + \frac{\theta^2}{4}\sum_{t=1}^N \theta_t^2\tau_t^2\boldsymbol{\Sigma}_t + \frac{1}{\zeta^2}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big)\Big] \\ &- \frac{4}{\zeta^3\theta}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big]\end{aligned}$$

Arranging the terms, we obtain

$$\begin{aligned}&- \frac{4}{\zeta^2}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}\Big]\tau \\ &- \frac{1}{\zeta^2}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\sum_{t=1}^N \theta_t^2\boldsymbol{\Sigma}_t\Big]\tau_t^2 + \tau_\lambda^2 \\ &= \frac{1}{\zeta^2}\mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 + \frac{4}{\zeta^4\theta^2}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big] \\ &- \frac{4}{\zeta^3\theta}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}\mathbf{e}_1\mathbf{e}_1^T\boldsymbol{\Sigma}\Big] \end{aligned} \tag{24}$$

So far all the computation presented was for the case $j = 1$ as the scalar parameters depend on $\mathbf{e}_1$. However, we note the coefficient matrix for τ_t^2, τ, and τ_λ^2 is independent of $\mathbf{e}_1$ as θ and θ_t do not depend on $\mathbf{e}_1$. This implies that for every $j = 1, \cdots, d$, can be expressed as the solution to the following linear system of equations

$$\mathbf{A} \begin{pmatrix} \tau^{(j)2} \\ \tau_1^{(j)2} \\ \vdots \\ \tau_N^{(j)2} \\ \tau_\lambda^{(j)2} \end{pmatrix} = \mathbf{b}_j$$

Where $\mathbf{A}$ is constructed according to the equations (22), (23), and (24) as follows:

$$(\mathbf{A})_{ij} = \begin{cases} \frac{\theta}{2} & i = j = 1 \\ 0 & i = 1, \quad j = 2, \cdots N+2 \\ -4n_i \mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma})\Big)^{-1}\Big(\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_t\Big) \\ \cdot\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\Big]\boldsymbol{\Sigma} & i = 2, \cdots N+1, \quad j = 1 \\ -n_j\theta_j^2 \mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\Big(\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_i\Big) \\ \cdot\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}_j\Big] & i = 2, \cdots N+1, \quad j = 2, \cdots N+1, \quad j \neq i \\ 1 - n_i\theta_i^2 \mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\Big(\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_i\Big) \\ \cdot\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}_i\Big] & i = 2, \cdots N+1, \quad j = i \\ 0 & i = 2, \cdots N+1, \quad j = N+2 \\ -16(n-d)^2 \mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma} \\ \cdot\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}\Big] & i = N+2, \quad j = 1 \\ -16(n-d)^2\theta_j^2 \mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma} \\ \cdot\Big(\sum_{t=1}^N \theta_t n_t(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma})\Big)^{-1}\boldsymbol{\Sigma}_j\Big] & i = N+2, \quad j = 2, \cdots N+1 \\ 1 & i = N+2, \quad j = N+2 \end{cases} \tag{25}$$

And we have for $\mathbf{b}_j$

$$\mathbf{b}_j = \begin{pmatrix} \frac{1}{\zeta\theta}\mathbf{e}_j^T\mathbf{\Sigma}\mathbf{e}_j - \frac{2}{\zeta^2\theta^2}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}\mathbf{e}_j\mathbf{e}_j^T\mathbf{\Sigma}\Big] \\ \begin{pmatrix} 4\frac{n_i}{\theta^2\zeta^2}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_i+\mathbf{\Sigma})\Big)^{-1}\Big(\mathbf{\Sigma}+\mathbf{\Sigma}_t\Big)\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}\mathbf{e}_j\mathbf{e}_j^T\mathbf{\Sigma}\Big] \\ -\frac{4n_i}{\zeta\theta}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}\mathbf{e}_j\mathbf{e}_j^T\mathbf{\Sigma}\Big] \end{pmatrix}_{i=1}^N \\ \begin{pmatrix} \frac{1}{\zeta^2}\mathbf{e}_j^T\mathbf{\Sigma}\mathbf{e}_j + \frac{4}{\zeta^4\theta^2}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}\mathbf{e}_j\mathbf{e}_j^T\mathbf{\Sigma}\Big] \\ -\frac{4}{\zeta^3\theta}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}\mathbf{e}_j\mathbf{e}_j^T\mathbf{\Sigma}\Big] \end{pmatrix} \end{pmatrix}_{j=1} \tag{26}$$

And we define from (26)

$$\mathbf{b} := \sum_{j=1}^d \mathbf{b}_j \tag{27}$$

We note the total generalization error can be written as

$$\begin{aligned} &\zeta^2\sum_{j=1}^d \tau_\lambda^{(j)2} + 4\theta^2\sum_{j=1}^d \tau^{(j)2}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}_{\mathbf{z}}\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}\Big] \\ &+\frac{4}{\zeta^2}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}_{\mathbf{z}}\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}^2\Big] \\ &+\theta^2\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}_{\mathbf{z}}\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\sum_{t=1}^N \theta_t^2\mathbf{\Sigma}_t\Big]\sum_{j=1}^d \tau_t^{(j)2} \end{aligned}$$

We observe that the generalization error is a function of $\sum_{j=1}^d \tau_t^{(j)2}$, $\sum_{j=1}^d \tau_\lambda^{(j)2}$, $\sum_{j=1}^d \tau^{(j)2}$. Hence we can write the generalization error as

$$\begin{aligned} &\zeta^2(\mathbf{A}^{-1}\mathbf{b})_{N+2} + 4\theta^2(\mathbf{A}^{-1}\mathbf{b})_1\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}_{\mathbf{z}}\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}\Big] \\ &\qquad+\frac{4}{\zeta^2}\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}_{\mathbf{z}}\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}^2\Big] \\ &\qquad+\theta^2\mathrm{Tr}\Big[\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\mathbf{\Sigma}_{\mathbf{z}}\Big(\sum_{t=1}^N \theta_t n_t(\mathbf{\Sigma}_t+\mathbf{\Sigma})\Big)^{-1}\sum_{t=1}^N \theta_t^2\mathbf{\Sigma}_t(\mathbf{A}^{-1}\mathbf{b})_{t+1}\Big] \end{aligned}$$

Where $(\mathbf{A}^{-1}\mathbf{b})_j$ denotes the jth entry of the vector $\mathbf{A}^{-1}\mathbf{b} \in \mathbb{R}^{N+2}$.

A.3. Proof of Concentration of the Generalization Error

In this section, we would like to show that one can use the concentration of some scalar functions of $\mathbf{w}_j^*$ for $j = 1, \ldots, d$ to prove the concentration of the total generalization error of $\mathbf{W}_{\text{lsq}}$. Namely,

by the prior arguments, we have observed that $\|\mathbf{\Sigma}_j^{1/2}(\mathbf{w}_j^* - \mathbf{e}_j)\|_2^2 \xrightarrow{\mathbb{P}} \zeta^2 \tau_\lambda^{(j)2}$ for each $j = 1, \dots, d$. Moreover, one can specifiy the concentration rate, similar to Dokmanic and Gribonval (2019); Hassani and Javanmard (2024), where one can show that there exist $c_1, c_2 = \theta(1)$ such that for $t_j > 0$ for $i \in [d]$:

$$\mathbb{P}\Big(\Big|d\|\mathbf{\Sigma}_j^{1/2}(\mathbf{w}_j^* - \mathbf{e}_j)\|_2^2 - d\zeta^2\tau_\lambda^{(j)2}\Big| > t_j\Big) \le c_1 \exp(-c_2 d t_j)$$

Furthermore, let

$$\mathcal{L}_j := \mathrm{Tr}\Bigg[\Big(\theta\sum_{t=1}^N \theta_t n_t \mathbf{\Sigma}_t + \frac{2}{\zeta}\mathbf{\Sigma}\Big)^{-1}\mathbf{\Sigma}_{\mathbf{z}}\Big(\theta\sum_{t=1}^N \theta_t n_t \mathbf{\Sigma}_t + \frac{2}{\zeta}\mathbf{\Sigma}\Big)^{-1} \cdot \Big(4\theta^2\tau^2\mathbf{\Sigma} + \frac{4}{\zeta^2}\mathbf{\Sigma}\mathbf{e}_j\mathbf{e}_j^T\mathbf{\Sigma} + \theta^2\sum_{t=1}^N \theta_t^2 \tau_t^{(j)2}\mathbf{\Sigma}_t\Big)\Bigg]$$

We have similarly for $c_3, c_4 = \theta(1)$

$$\mathbb{P}\Big(\Big|d\mathbf{w}_j^{*T}\mathbf{\Sigma}_{\mathbf{z}}\mathbf{w}_j^* - d\mathcal{L}_j\Big| > t_j\Big) \le c_3 \exp(-c_4 d t_j)$$

Note that we had

$$\mathcal{E}(\mathbf{W}_{\text{lsq}}) = \sum_{j=1}^d \mathbf{w}^{*T}\mathbf{\Sigma}_{\mathbf{z}}\mathbf{w}_j^* + (\mathbf{w}_j^* - \mathbf{e}_j)^T\mathbf{\Sigma}_j(\mathbf{w}_j^* - \mathbf{e}_j)$$

The concentration of the solutions follows by a union bound argument, from:

$$\mathbb{P}\Big(\Big|\mathcal{E}(\mathbf{W}_{\text{lsq}}) - \sum_{j=1}^d \mathcal{L}_j + \zeta^2\tau_\lambda^{(j)2}\Big| > t\Big) \le \sum_{j=1}^d \mathbb{P}\Big(\Big|\mathbf{w}_j^{*T}\mathbf{\Sigma}_{\mathbf{z}}\mathbf{w}_j^* - \mathcal{L}_j + \|\mathbf{\Sigma}_j^{1/2}(\mathbf{w}_j^* - \mathbf{e}_j)\|_2^2 - \zeta^2\tau_\lambda^{(j)2}\Big| > \frac{t}{d}\Big)$$
$$\le 2d\max\{c_1, c_3\}\exp(-\min\{c_2, c_4\}dt)$$

Thus the statement follows as $t = \omega(1)$.

Appendix B. Proof of Corollary 3

We consider $\mathbf{\Sigma} = c\mathbf{I}$ and $\mathbf{\Sigma}_{\mathbf{z}} = c_{\mathbf{z}}\mathbf{I}$ and write

$$\min_{\mathbf{\Sigma}_1 \succeq \mathbf{0}}\Big(1 + \frac{c + c_{\mathbf{z}}}{n - d}\mathrm{Tr}\Big(c\mathbf{I} + \mathbf{\Sigma}_1\Big)^{-1}\Big)\Big(cd - c^2\mathrm{Tr}\Big(c\mathbf{I} + \mathbf{\Sigma}_1\Big)^{-1}\Big) + c^2\mathrm{Tr}\Bigg[\Big(c\mathbf{I} + \mathbf{\Sigma}_1\Big)^{-1}\Big(c_{\mathbf{z}}\mathbf{I} - \mathbf{\Sigma}_1\Big)\Big(c\mathbf{I} + \mathbf{\Sigma}_1\Big)^{-1}\Bigg]$$

Note that, since the KKT condition for the optimal $\mathbf{\Sigma}_1$ is symmetric w.r.t. the eigenvalues of $\mathbf{\Sigma}_1$ in this case, we can take $\mathbf{\Sigma}_1 = \sigma\mathbf{I}$, which leads to

$$\begin{aligned}
&c\cdot \min_{\sigma\geq 0}\Big(1+\frac{d}{n-d}\frac{c+c_{\mathbf{z}}}{c+\sigma}\Big)\Big(d-d\frac{c}{c+\sigma}\Big)+cd\frac{c_{\mathbf{z}}-\sigma}{(c+\sigma)^2}\\
&=cd\cdot\min_{\sigma\geq 0}1+\frac{\frac{d(c+c_{\mathbf{z}})}{n-d}-c}{c+\sigma}+\frac{-c\frac{d(c+c_{\mathbf{z}})}{n-d}+c(c_{\mathbf{z}}-\sigma)}{(c+\sigma)^2}\\
&=cd+cd\cdot\min_{\sigma\geq 0}\frac{-c\frac{d(c+c_{\mathbf{z}})}{n-d}+c(c_{\mathbf{z}}-\sigma)+\Big(\frac{d(c+c_{\mathbf{z}})}{n-d}-c\Big)(c+\sigma)}{(c+\sigma)^2}\\
&=cd+cd\cdot\min_{\sigma\geq 0}\frac{\Big(\frac{d(c+c_{\mathbf{z}})}{n-d}-2c\Big)\sigma+cc_{\mathbf{z}}-c^2}{(c+\sigma)^2}
\end{aligned}$$

Then we observe for the optimal σ

$$\sigma^*=\begin{cases}
0 & (3-2\kappa)c+c_{\mathbf{z}}<0, \quad c+(3-2\kappa)c_{\mathbf{z}}>0\\
c\frac{c+(3-2\kappa)c_{\mathbf{z}}}{(3-2\kappa)c+c_{\mathbf{z}}} & (3-2\kappa)c+c_{\mathbf{z}}<0, \quad c+(3-2\kappa)c_{\mathbf{z}}<0\\
\infty & (3-2\kappa)c+c_{\mathbf{z}}>0, \quad c+(3-2\kappa)c_{\mathbf{z}}<0\\
\infty & (3-2\kappa)c+c_{\mathbf{z}}>0, \quad c+(3-2\kappa)c_{\mathbf{z}}>0, \quad c_{\mathbf{z}}>c\\
0 & (3-2\kappa)c+c_{\mathbf{z}}>0, \quad c+(3-2\kappa)c_{\mathbf{z}}>0, \quad c_{\mathbf{z}}<c
\end{cases}$$

In each case, the best generalization error turns out to be

$$g.e\xrightarrow{\mathbb{P}}\begin{cases}
dc_{\mathbf{z}} & (3-2\kappa)c+c_{\mathbf{z}}<0, \quad c+(3-2\kappa)c_{\mathbf{z}}>0\\
d\frac{-(c+c_{\mathbf{z}})^2+4(\kappa-1)^2cc_{\mathbf{z}}}{4(\kappa-2)(\kappa-1)(c+c_{\mathbf{z}})} & (3-2\kappa)c+c_{\mathbf{z}}<0, \quad c+(3-2\kappa)c_{\mathbf{z}}<0\\
dc & (3-2\kappa)c+c_{\mathbf{z}}>0, \quad c+(3-2\kappa)c_{\mathbf{z}}<0\\
dc & (3-2\kappa)c+c_{\mathbf{z}}>0, \quad c+(3-2\kappa)c_{\mathbf{z}}>0, \quad c_{\mathbf{z}}>c\\
dc_{\mathbf{z}} & (3-2\kappa)c+c_{\mathbf{z}}>0, \quad c+(3-2\kappa)c_{\mathbf{z}}>0, \quad c_{\mathbf{z}}<c
\end{cases}$$