## Impact of Bottleneck Layers and Skip Connections on the Generalization of Linear Denoising Autoencoders

Jonghyun Ham\* University of Freiburg

Maximilian Fleissner
Technical University of Munich

**Debarghya Ghoshdastidar** Technical University of Munich

## **Abstract**

Modern deep neural networks exhibit strong generalization even in highly overparameterized regimes. Significant progress has been made to understand this phenomenon in the context of supervised learning, but for unsupervised tasks such as denoising, several open questions remain. While some recent works have successfully characterized the test error of the linear denoising problem, they are limited to linear models (one-layer network). In this work, we focus on two-layer linear denoising autoencoders trained under gradient flow, incorporating two key ingredients of modern deep learning architectures: A low-dimensional bottleneck layer that effectively enforces a rank constraint on the learned solution, as well as the possibility of a skip connection that bypasses the bottleneck. We derive closed-form expressions for all critical points of this model under product regularization, and in particular describe its global minimizer under the minimum-norm principle. From there, we derive the test risk formula in the overparameterized regime, both for models with and without skip connections. Our analysis reveals two interesting phenomena: Firstly, the bottleneck layer introduces an additional complexity measure akin to the classical bias-variance trade-off—increasing the bottleneck width reduces bias but introduces variance, and vice versa. Secondly, skip connection can mitigate the variance in denoising autoencoders—especially when the model is mildly overparameterized. We further analyze the impact of skip connections in denoising autoencoder using random matrix theory and support our claims with numerical evidence.

## 1 Introduction

Despite having a large number of parameters and achieving nearly zero training error, modern neural networks generalize remarkably well to unseen data. This phenomenon, often referred to as *benign overfitting* [9], challenges the classical understanding of generalization characterized by a U-shaped risk curve, where increasing model complexity is expected to eventually harm test performance. Extensive theoretical efforts have sought to explain this behavior—albeit almost exclusively in supervised learning. In contrast, little attention has been paid to understanding generalization in *unsupervised learning*, where contradictory statements are made based on numerical studies [25, 34].

A prominent example is the denoising autoencoder (DAE) [44]. Despite its distinct setting—which differs significantly from standard regression in that *noise is added to the input* rather than the output—and its widespread use in unsupervised representation learning, the generalization properties of DAEs remain underexplored. A pioneering study by [38] initiated the theoretical analysis of

<sup>\*</sup>Contact: hamj@informatik.uni-freiburg.de

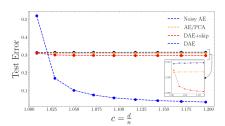


Figure 1: Test error curves for variants of linear autoencoders (AE). Unsupervised AE learns to reconstruct the input and is equivalent to PCA. Noisy AE maps clean input to a noisy version [2]. We study the generalization error in denoising AE (DAE) [44] that reconstructs clean samples from noisy version, and DAE with skip connection [35] that implicitly learns to generate pure noise from noisy data. While the generalization error of AE and noisy AE are not affected by over-parameterisation (in the linear case), linear DAE exhibits pronounced peak near  $c \approx 1$ , which is partly dampened in DAE with skip connection.

the denoising problem under a rank-1 data setting. This work was later extended to general low-rank data by [27]. However, both works are confined to single-layer linear architectures, limiting their applicability to modern neural networks. More specifically, a notable architectural feature of contemporary neural networks is the presence of bottleneck structures, where intermediate layers have significantly lower dimensionality than the overall parameter count. This form of architectural complexity is not adequately addressed by existing theory, which tends to treat input dimensionality as the only measure of model complexity. Our goal is to investigate the role of bottleneck layers as a complementary complexity measure and examine their impact on generalization behavior, specifically in the context of DAEs.

In general, there is little understanding of how the training dynamics in the presence of bottleneck layers influence the generalization error of neural networks. Several works study random feature models for noisy autoencoders [2] or investigate the effect of principal component analysis (PCA)-based dimensionality reduction in linear (one-layer) regression settings [40, 20, 14]. With dimensionality reduction, the number of retained principal components acts as a form of regularization and an appropriate small dimension suppresses the double descent phenomenon. In contrast, fully trained architectures with a bottleneck layer can still exhibit double descent behavior as illustrated in the generalization error curves of over-parameterized two-layer linear DAEs in Figure 1.

In addition, practical implementations of DAE (such as the U-Net architecture [35] that serves as a de facto denoiser in diffusion models [24]) routinely incorporate skip connections as a core architectural feature. While skip connections are widely acknowledged for enhancing training stability by mitigating vanishing or exploding gradients [23], their impact on test error remains poorly understood. [13] investigate the role of skip connections in improving generalization performance in undercomplete DAEs using two-layer nonlinear neural networks. However, their analysis is conducted under restrictive assumptions, including Gaussian input data and tied weights. Our work extends the more realistic data assumptions used in [27] and does not impose tied-weight constraints.

To this end, we analyze linear DAEs modeled as two-layer linear networks in the high-dimensional regime, where the input dimension d exceeds the number of training samples n. Our model includes a low-dimensional bottleneck layer of size  $k \ll n < d$  with or without a skip connection. To address the effect of bottleneck layers and skip connections, we extend the theoretical frameworks for DAEs [38, 27], which are built upon the assumption of low rank data, and derive analytical expressions that characterize the generalization error in these settings. Moreover, we offer a deeper theoretical understanding of the role of skip connections by explicitly performing a bias-variance decomposition, which was absent in previous studies. Our contributions are summarized as follows.

- 1. In Section 2, we obtain closed-form expressions for the critical points and global minimizers of linear DAEs with bottleneck layers, both with and without skip connections, under a reconstruction loss with product regularization. Furthermore, we derive the minimum-norm solution, and find that it is approached by the global minimizer of the regularized loss in the ridgeless limit.
- 2. In Section 3, we leverage the closed-form expressions obtained for the learned model to derive expressions for the test risk. To better understand this result, we perform a bias-variance decomposition that quantifies how the bottleneck dimension influences generalization. In particular, increasing the bottleneck dimension reduces bias but increases variance, and vice versa. While this trade-off echoes the classical interpretation of the bias-variance relationship, it arises within the modern high-dimensional regime, by virtue of the bottleneck layer.

- 3. We extend our analysis of the test risk to DAEs with skip connections. Notably, this results in a significantly smoother variance curve as a function of d/n, compared to the model without skip connections. The effect is particularly pronounced when the model is only mildly overparameterized.
- 4. In Section 4, we provide further insights into the origin of the smoother variance for models with skip connections, using tools from random matrix theory. By analyzing a slightly simplified model, we uncover the origin of the smoother variance curve induced by skip connections.

## 2 Setting

We begin by introducing the denoising autoencoders(DAEs) and specifying their associated loss functions. We then characterize the solutions learned under gradient flow by deriving all critical points and providing an explicit expression for the global minimizer. This analysis lays the groundwork for our subsequent study of generalization error in Section 3.

## 2.1 Training Setup

We consider two variants of two-layer linear networks. The first model contains the bottleneck structure, and does not include skip connections. Given an input matrix  $\mathbf{Z} \in \mathbb{R}^{d \times n}$ , where n denotes the number of training samples and each column of  $\mathbf{Z}$  represents a d-dimensional data point, the model is defined by two weight matrices:  $\mathbf{W}_1 \in \mathbb{R}^{k \times d}$  (the encoder) and  $\mathbf{W}_2 \in \mathbb{R}^{d \times k}$  (the decoder). The output of this model is given by  $\mathbf{W}_2\mathbf{W}_1\mathbf{Z}$ . The second model includes a skip connection. Its output is defined as  $(\mathbf{W}_2\mathbf{W}_1 + \mathbb{I})\mathbf{Z}$ , where the skip connection is implemented as an identity map  $\mathbb{I} \in \mathbb{R}^{d \times d}$ , directly linking the input to the output layer and bypassing the trainable weight matrices.

Loss Functions for the Denoising Setup In the denoising setting,  $\mathbf{X} \in \mathbb{R}^{d \times n}$  denotes the clean input data matrix, and  $\mathbf{A} \in \mathbb{R}^{d \times n}$  is the noise matrix. Then, the input to the network is the corrupted matrix  $\mathbf{X} + \mathbf{A}$ , while the target output is the clean matrix  $\mathbf{X}$ . The networks with and without skip connection are trained to minimize the reconstruction loss of  $\mathbf{X}$  from  $\mathbf{X} + \mathbf{A}$ . Additionally, a product regularization term [33] with regularization strength  $\lambda$  is added to the loss function. For the model without skip connections, the loss function  $\mathcal{L}_{\mathrm{DAE}}$  is given by

$$\mathcal{L}_{DAE}(\mathbf{W}_2, \mathbf{W}_1, \lambda) := \frac{1}{n} \|\mathbf{X} - \mathbf{W}_2 \mathbf{W}_1(\mathbf{X} + \mathbf{A})\|_F^2 + \lambda \|\mathbf{W}_2 \mathbf{W}_1\|_F^2. \tag{1}$$

For the model with skip connections, the corresponding loss function  $\mathcal{L}_{DAE+SC}$  is given by

$$\mathcal{L}_{\text{DAE+SC}}(\mathbf{W}_2, \mathbf{W}_1, \lambda) := \frac{1}{n} \|\mathbf{X} - (\mathbf{W}_2 \mathbf{W}_1 + \mathbb{I})(\mathbf{X} + \mathbf{A})\|_F^2 + \lambda \|\mathbf{W}_2 \mathbf{W}_1\|_F^2$$
$$= \frac{1}{n} \|-\mathbf{A} - \mathbf{W}_2 \mathbf{W}_1(\mathbf{X} + \mathbf{A})\|_F^2 + \lambda \|\mathbf{W}_2 \mathbf{W}_1\|_F^2. \tag{2}$$

Hence, the loss function for the skip-connected model,  $\mathcal{L}_{DAE+SC}$ , can be interpreted as a variant of  $\mathcal{L}_{DAE}$  where the target output  $\mathbf{X}$  is replaced by  $-\mathbf{A}$ . In other words, the model learns to reconstruct the noise from the noisy input—a setup commonly used in diffusion models [24]. This observation naturally leads to a more general formulation of the training objective, which we present next.

**Training with a General Input-Output Pair** Both (1) and (2) can be viewed more generally as training a two-layer linear network to predict an output  $\mathbf{Y} \in \mathbb{R}^{d \times n}$  from inputs  $\mathbf{Z} \in \mathbb{R}^{d \times n}$ . In that case, the loss function  $\mathcal{L}$  is given by

$$\mathcal{L}(\mathbf{W}_2, \mathbf{W}_1, \lambda) := \frac{1}{n} \|\mathbf{Y} - \mathbf{W}_2 \mathbf{W}_1 \mathbf{Z}\|_F^2 + \lambda \|\mathbf{W}_2 \mathbf{W}_1\|_F^2.$$
(3)

To analyze the generalization error in DAEs, we first need to characterize the critical points  $\hat{\mathbf{W}}_2 \in \mathbb{R}^{d \times k}$  and  $\hat{\mathbf{W}}_1 \in \mathbb{R}^{k \times d}$  that satisfy

$$\frac{d}{d\hat{\mathbf{W}}_2}\mathcal{L}(\hat{\mathbf{W}}_2,\hat{\mathbf{W}}_1,\lambda) = 0 \quad \text{and} \quad \frac{d}{d\hat{\mathbf{W}}_1}\mathcal{L}(\hat{\mathbf{W}}_2,\hat{\mathbf{W}}_1,\lambda) = 0.$$

Particularly, we analyze the ridgeless limit of the critical points  $\hat{\mathbf{W}}_c := \hat{\mathbf{W}}_2 \hat{\mathbf{W}}_1$ , that is  $\mathbf{W}_c = \lim_{\lambda \to 0} \hat{\mathbf{W}}_c$ . Ultimately, for  $\hat{\mathbf{W}}_2^*, \hat{\mathbf{W}}_1^* = \operatorname{argmin}_{\hat{\mathbf{W}}_1, \hat{\mathbf{W}}_2} \mathcal{L}(\hat{\mathbf{W}}_2, \hat{\mathbf{W}}_1, \lambda)$ , we are interested in the minimum-norm global minimizer  $\mathbf{W}_*$  which is given by  $\mathbf{W}_* := \lim_{\lambda \to 0} \hat{\mathbf{W}}_2^* \hat{\mathbf{W}}_1^*$ .

#### 2.2 General Expressions for Critical Points

We begin by introducing several notations. Given an input matrix  $\mathbf{Z} \in \mathbb{R}^{d \times n}$ , let  $\tilde{\mathbf{Z}} := \mathbf{Z}\mathbf{Z}^{\top} + \lambda \mathbb{I}$  and  $\mathbf{G} := \mathbf{Y}\mathbf{Z}^{\top}\tilde{\mathbf{Z}}^{-1}\mathbf{Z}\mathbf{Y}^{\top}$ . Let  $\mathbf{G} = \mathbf{U}_{\mathbf{G}}\Lambda_{\mathbf{G}}\mathbf{U}_{\mathbf{G}}^{\top}$  denote the eigendecomposition of  $\mathbf{G}$ . We use the shorthand [k] to denote the set of natural numbers  $\{1, 2, \dots, k\}$ . For any matrix  $\mathbf{M}$ , we write  $r_{\mathbf{M}}$  for its rank and  $\mathbf{M}^{\dagger}$  for its Moore-Penrose pseudo-inverse. For  $k \leq r_{\mathbf{M}}$ , let  $\mathcal{I}_{\mathbf{M},k}$  denote the collection of ordered index sets, where each  $I \in \mathcal{I}_{\mathbf{M},k}$  satisfies  $I \subseteq [r_{\mathbf{M}}]$  such that  $|I| \leq k$ . That is, for  $I = \{j_1, \dots, j_{|I|}\}$ , we require  $1 \leq j_1 < j_2 < \dots < j_{|I|} \leq r_{\mathbf{M}}$ . Then, we define the projection onto the corresponding rank-one components of  $\mathbf{M}$  by  $P_I(\mathbf{M}) := \sum_{j \in I} \sigma_j^{\mathbf{M}} \mathbf{u}_j^{\mathbf{M}} \mathbf{v}_j^{\mathbf{M}^{\top}}$ , where  $\sigma_j^{\mathbf{M}}$ ,  $\mathbf{u}_j^{\mathbf{M}}$ , and  $\mathbf{v}_j^{\mathbf{M}}$  denote j-th singular value, left singular vector, and right singular vector of  $\mathbf{M}$ , respectively. With this notation in place, we now state a simplified version of the critical point characterization; the full version and its proof are provided in Appendix D.

**Theorem 2.1** (Critical Points for General Input-Output Pairs). Assume that the multiplicity of non-zero eigenvalues of G is 1, i.e.,  $\lambda_1^{\mathbf{G}} > \lambda_2^{\mathbf{G}} > \cdots > \lambda_{r_{\mathbf{G}}}^{\mathbf{G}} > 0$ . Further, suppose that the bottleneck dimension k satisfies  $k \leq r_{\mathbf{G}}$ . Then, each index set  $I_{\mathbf{G}} \in \mathcal{I}_{\mathbf{G},k}$  characterizes a critical point  $\hat{\mathbf{W}}_c$  of the loss function  $\mathcal{L}$ , given by

$$\hat{\mathbf{W}}_c = \hat{\mathbf{W}}_2 \hat{\mathbf{W}}_1 = \mathbf{U}_{\mathbf{G}, I_{\mathbf{G}}} \mathbf{U}_{\mathbf{G}, I_{\mathbf{G}}}^{\top} \mathbf{Y} \mathbf{Z}^{\top} \tilde{\mathbf{Z}}^{-1}.$$
(4)

where  $U_{G,I_G}$  denotes the submatrix of  $U_G$  consisting of the columns indexed by  $I_G$ . Moreover, assume that  $r_Z = n$ . Then, in the ridgeless limit  $\lambda \to 0$ , it holds that

$$\mathbf{W}_c := \lim_{\lambda \to 0} \hat{\mathbf{W}}_c = P_{I_{\mathbf{G}}}(\mathbf{Y}) \mathbf{Z}^{\dagger}. \tag{5}$$

Furthermore, the minimum-norm global minimizer  $W_*$  is given by

$$\mathbf{W}_* = P_{[k]}(\mathbf{Y})\mathbf{Z}^{\dagger}. \tag{6}$$

Note that  $W_c$  is constructed by selecting a subset of rank-one components from the singular value decomposition of Y. The global minimizer corresponds to the choice  $I_G = [k]$ .

**Remark 2.2** (Comparison with [8]). This result is reminiscent of the classical analysis in [8], which characterizes critical points in the underparameterized regime (d < n). In contrast, our theorem extends this line of work to the overparameterized setting, while additionally incorporating a regularization term to find the minimum-norm solution.

**Remark 2.3** (Effect of Overparameterization and Minimum-Norm Principle). Observe that Eq. (4) involves the matrix  $U_{G,I}$ , whose columns are eigenvectors of  $\mathbf{G}$ . Its dependence on both Y and Z makes direct analysis challenging. However, in the overparameterized regime and in the ridgeless limit,  $\mathbf{G}$  simplifies to  $\mathbf{G} = \mathbf{YY}^{\top}$ . This leads to a more tractable expression for  $\mathbf{W}_c$ , which now depends only on a projection of  $\mathbf{Y}$  and the pseudo-inverse of  $\mathbf{Z}$ . This enable us to build on technical tools developed in [38, 27] to characterize the generalization risk, as detailed in Section 3.

## 2.3 Specification to the Denoising Setup

We now specialize the result of Theorem 2.1 to the denoising autoencoder setting, yielding closed-form solutions for the models introduced earlier. We assume that the eigenvalues of both  $\mathbf{X}\mathbf{X}^{\top}$  and  $\mathbf{A}\mathbf{A}^{\top}$  have multiplicity one. We also assume that the matrix  $\mathbf{X} + \mathbf{A}$  is full-rank, which holds almost surely (see [19]). Under this setup, we obtain the following corollaries for the two DAE models (1) and (2).

**Corollary 2.4** (Critical Points of the Model without Skip Connections). Let  $\mathcal{I}_{\mathbf{X}}$  be a family of index sets, where each element is an ordered set of distinct natural numbers from [k], where  $k \leq r_{\mathbf{X}}$ . Then, for each  $I^x \in \mathcal{I}_{\mathbf{X}}$ , a critical point is given by  $\mathbf{W}_c = P_{I^x}(\mathbf{X})(\mathbf{X} + \mathbf{A})^{\dagger}$ . Moreover, the global minimizer  $\mathbf{W}_*$  is given by  $\mathbf{W}_* = P_{[k]}(\mathbf{X})(\mathbf{X} + \mathbf{A})^{\dagger}$ .

**Corollary 2.5** (Critical Points of the Model with Skip Connections). Let  $\mathcal{I}_{\mathbf{A}}$  be a family of index sets, where each element is an ordered set of distinct natural numbers from [k], where  $k \leq r_{\mathbf{A}}$ . Then, for each  $I^a \in \mathcal{I}_{\mathbf{A}}$ , a critical point is given by  $\mathbf{W}_c^{sc} = -P_{I^a}(\mathbf{A})(\mathbf{X} + \mathbf{A})^{\dagger}$ . Moreover, the global minimizer  $\mathbf{W}_*^{sc}$  is given by  $\mathbf{W}_*^{sc} = -P_{[k]}(\mathbf{A})(\mathbf{X} + \mathbf{A})^{\dagger}$ .

These closed-form expressions enable us to derive formulas for the test risk of both models. This is the topic of the following section.

## 3 The Generalization Error of Linear DAEs

In this section, we analyze the test error corresponding to the critical points derived in Section 2.3. We begin by outlining the data assumptions of our analysis and then derive expressions for the test error of the models with and without skip connections. Next, we introduce a natural bias-variance decomposition that arises from the bottleneck structure, and examine how varying the bottleneck dimension influences bias and variance in both models. Finally, we highlight the surprising effect that adding skip connections produces a smoother test error curve. We first present the data model that forms the basis of our analysis. For training, we consider  $\mathbf{X} \in \mathbb{R}^{d \times n}$  as the clean (noise-free) data matrix and  $\mathbf{A} \in \mathbb{R}^{d \times n}$  as the associated additive Gaussian noise matrix. Likewise,  $\mathbf{X}_{\text{tst}} \in \mathbb{R}^{d \times N_{\text{tst}}}$  and  $\mathbf{A}_{\text{tst}} \in \mathbb{R}^{d \times N_{\text{tst}}}$  denote the clean and noise matrices used for testing, respectively. Note that in this section, we work on non-asymptotic setting where d, n are high-dimensional but finite.

## **Assumption 3.1** (Data Assumptions).

- 1. Normalized Low Rank:  $\mathbf{X}$  is normalized such that  $\|\mathbf{X}\|_2 = \Theta(1)$ . Its rank, denoted as r, satisfies  $r \ll d, n$ .
- 2. Well Conditioned: The ratio between the largest singular value and the smallest nonzero singular value of X is  $\Theta(1)$ .
- 3. Noise: The entries of  $\mathbf{A}$ ,  $\mathbf{A}_{tst}$ , are sampled independently from  $\mathcal{N}(0, \frac{\eta_{tm}^2}{d})$ ,  $\mathcal{N}(0, \frac{\eta_{tst}^2}{d})$  respectively, where  $\eta_{tm}$ ,  $\eta_{tst} = \Theta(1)$ .
- 4. Test Data:  $\mathbf{X}_{tst}$  is assumed to lie in the same low dimensional subspace as the training data. In other words, the test data  $\mathbf{X}_{tst}$  satisfies  $\mathbf{X}_{tst} = \mathbf{U}\mathbf{L}$ , for  $\mathbf{U} \in \mathbb{R}^{d \times r}$  the left singular vectors of  $\mathbf{X}$  and for some non-zero coefficient matrix  $\mathbf{L} \in \mathbb{R}^{r \times N_{tst}}$ .

The data scaling assumption for X ensures that the signal matrix and the noise matrix are comparable in magnitude. This is motivated by the fact that the spectral norm of the noise matrix satisfies  $\|\mathbf{A}\|_2 = O(1)$  with high probability, as established in [43, Theorem 4.4.5]. The low-rank assumption on  $\mathbf{X}$  is supported by empirical evidence that real-world datasets are approximately low-rank, as argued in [42] and adopted by [27]. A crucial point emphasized in [27] is that the training data X is treated as an arbitrary but deterministic low-rank matrix without any distributional assumptions. In particular, the observations need not be independent.

Given the critical points  $\mathbf{W}_c$  and  $\mathbf{W}_c^{sc}$ , we evaluate the test error. Following [38], the test error of the model without skip connections is given by

$$R(\mathbf{W}_{c}, \mathbf{X}_{tst}) := \frac{1}{N_{tst}} \mathbb{E}_{\mathbf{A}_{tm}, \mathbf{A}_{tst}} \left[ \| \mathbf{X}_{tst} - \mathbf{W}_{c} (\mathbf{X}_{tst} + \mathbf{A}_{tst}) \|_{F}^{2} \right]. \tag{7}$$

For the model with skip connections, it is given by

$$R_{\text{sc}}(\mathbf{W}_{\text{c}}^{\text{sc}}, \mathbf{X}_{\text{tst}}) := \frac{1}{N_{\text{tst}}} \mathbb{E}_{\mathbf{A}_{\text{tm}}, \mathbf{A}_{\text{tst}}} \left[ \|\mathbf{X}_{\text{tst}} - (\mathbf{W}_{\text{c}}^{\text{sc}} + \mathbb{I})(\mathbf{X}_{\text{tst}} + \mathbf{A}_{\text{tst}}) \|_F^2 \right]. \tag{8}$$

## 3.1 Effect of Bottleneck Layers on Generalization Error

To understand how the bottleneck dimension influences generalization, we start by analyzing the test error for the model without skip connections (7). To this end, we plug in the the critical points  $\mathbf{W}_c$  obtained in Corollary 2.4. We then explore a natural bias-variance decomposition of the generalization error, focusing on the global minimizer  $\mathbf{W}_*$ .

**Theorem 3.2** (Test Error for the Model without Skip Connections). Let  $\alpha_i := \sigma_i \eta_{tm}^{-1}$ , where  $\sigma_i$  denotes the *i*-th singular value of **X**. Let  $d \ge n + r$ , and  $c := \frac{d}{n}$ . Let  $\mathbf{J} \in \mathbb{R}^{r \times r}$  be the diagonal matrix

$$\mathbf{J}_{ii} = \left(\alpha_i^2 + 1\right)^{-2} \cdot \mathbb{1}_{i \in I^x} + \mathbb{1}_{i \notin I^x}$$

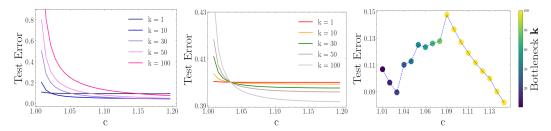


Figure 2: **Effect of Bottleneck.** Test errors on CIFAR-10 illustrating how the bottleneck dimension k influences generalization. The left plot corresponds to the model without skip connections. The center shows results for the model with skip connections. The right subfigure is constructed by jointly increasing both k and d, using the corresponding test errors from the left plot. As seen in the left and center plots, the optimal choice of k depends on the level of overparameterization, reflecting a distinct bias–variance trade-off in different regimes.

where  $\mathbb{1}_{(.)}$  denotes the indicator function. Then, for a critical point  $\mathbf{W}_c$  we have that

$$R(\mathbf{W}_c, \mathbf{X}_{tst}) = \frac{1}{N_{tst}} \operatorname{Tr}(\mathbf{J} \mathbf{L} \mathbf{L}^{\top}) + \frac{\eta_{tst}^2 c}{d(c-1)} \sum_{i \in I^x} \frac{\alpha_j^2}{1 + \alpha_j^2} + o\left(\frac{1}{d}\right). \tag{9}$$

See Appendix E.1 for the proof of this theorem.

**Remark 3.3** (Global Minimizer). The test risk for  $W_*$  is obtained by plugging in [k] for  $I^x$ .

**Remark 3.4** (Bias Term). Since  $\mathbf{L}$  depends only on the test data, the overall magnitude of  $\mathrm{Tr}(\mathbf{J}\mathbf{L}\mathbf{L}^{\top})$  is mainly influenced by the size of the diagonal entries of  $\mathbf{J}$ .

**Bottleneck Dimension as a Complexity Measure** Consider the case of the global minimizer, where  $I^x = [k]$ , and all parameters except k are fixed. Then, the first term of Eq. (9) decreases as k increases towards r. This is because the diagonal entries of  $\mathbf{J}$  are either 1 or of the form  $(\alpha_i^2 + 1)^{-2} < 1$ . As k increases, the number of 1s decreases, leading to a lower value. Conversely, the second term increases as k grows, due to the increase of the number of summands. This trade-off behavior aligns with the classical understanding of bias-variance trade-off, as adjusting k is directly related to varying the model complexity.

In light of this, we interpret the first term of Eq. (9) as the bias component and the second term as the variance. Our notion of bias and variance aligns with that of [38]: the bias term is derived from the expected reconstruction error on clean data, given by  $N_{\rm tst}^{-1}\mathbb{E}[\|\mathbf{X}_{\rm tst}-\mathbf{W}_{\rm c}\mathbf{X}_{\rm tst}\|_F^2]$ , while the variance term arises from  $d^{-1}\eta_{\rm tst}^2\mathbb{E}[\|\mathbf{W}_{\rm c}\|_F^2]$ , which relates to the norm of the estimator (see Appendix E.2).

Under this decomposition, fixing k and n while increasing d (thus increasing overparameterization) leads to a decrease in the overall test error due to the reduced variance term. Notably, this occurs without a corresponding increase in bias, and therefore, without a trade-off. This absence of trade-off reflects the *modern understanding* of the bias-variance relationship, in which larger models can generalize better. In this sense, Theorem 3.2 illustrates the *coexistence* of both classical and modern perspectives on the interplay between bias and variance in the DAE setting.

This trade-off is further illustrated in Figure 2, which shows that for fixed input dimension d, a smaller bottleneck dimension k can improve test error in certain regimes by reducing the variance term—dominant in the mildly overparameterized setting due to peaking behavior. Moreover, the right plot of Figure 2 demonstrates that jointly increasing input dimension d and bottleneck dimension k leads to a  $second\ peak$  in the test curve within the overparameterized regime. These findings underscore that both the degree of overparameterization in d and the choice of k shape the generalization of DAEs.

## 3.2 Including Additional Skip Connections

Having examined the effect of bottleneck layers, we now investigate how the inclusion of skip connections influences test performance, particularly through its effect on variance. As in the

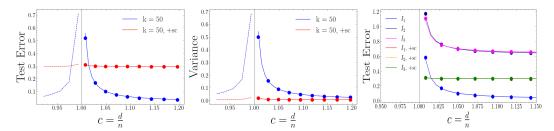


Figure 3: Effect of Skip Connections. Experiments on CIFAR-10 (Data rank is fixed at r=100). Solid lines represent theoretical predictions, while cross markers indicate empirical results. Dotted lines correspond to empirical values in the underparameterized regime (solutions derived from [8]). Red lines and markers denote results for the model with skip connections. The left subfigure shows the test error; the center subfigure displays the corresponding variance curve. The right subfigure compares test errors across different critical points:  $I_1=[50],\,I_2=[11,60],\,$  and  $I_3=[31,80].$ 

previous subsection, we first present a theoretical result and then interpret it through numerical experiments. Consider any critical point  $\mathbf{W}_{c}^{sc}$  from Corollary 2.5, with the corresponding index set  $I^{a}$ . Then, for  $d \geq n + r$ , we have the following theorem.

**Theorem 3.5** (Test Error for the Model with Skip Connections). Let  $\mathbf{J}^{sc}$  be a diagonal matrix defined as,  $\mathbf{J}^{sc}_{ii} = \frac{c + (c-1)\sigma_i^2}{c(1+\eta_m^2\sigma_i^2)^2}$ , for each  $i \in [r]$ . Then, for a critical point  $\mathbf{W}^{sc}_c$ , we have that

$$R_{sc}(\mathbf{W}_{c}^{sc}, \mathbf{X}_{tst}) = \eta_{tst}^{2} \left( 1 - \frac{|I^{a}|}{d} \right) + \frac{|I^{a}|}{dN_{tst}} \operatorname{Tr} \left( \mathbf{J}^{sc} \mathbf{L} \mathbf{L}^{\top} \right) + \frac{\eta_{tst}^{2} |I^{a}|}{d^{2}} \frac{c}{c - 1} \sum_{i=1}^{r} \frac{\sigma_{i}^{2}}{(\eta_{tm}^{2} + \sigma_{i}^{2})} + \frac{3\eta_{tst}^{2} |I^{a}|}{dn} \frac{1}{c} \sum_{i=1}^{r} \frac{\eta_{tm}^{2} \sigma_{i}^{2}}{(\eta_{tm}^{2} + \sigma_{i}^{2})} + O\left(\frac{1}{dN_{tst}}\right).$$
(10)

Similar to the previous theorem, with the global minimizer  $\mathbf{W}_*^{\mathrm{sc}}$ , we replace  $I^a$  to [k]. The proofs of this theorem can be found in Appendix E.1.

**Remark 3.6** (Bias and Variance Terms). Following the approach in the previous subsection, we interpret the norm term, i.e.,  $\eta^2_{tst}d^{-1}\|\mathbf{W}^{sc}_*\|_F^2$ , as the variance term, with the remaining terms attributed to bias. Importantly, the third term involving  $(c-1)^{-1}$  arises directly from the norm term. This decomposition is consistent with a unified definition of bias and variance across the models with&without skip connections, where bias decreases with increasing model complexity (measured by bottleneck size), and variance increases. We formalize this interpretation in Appendix E.2.

Remark 3.7 (Small Difference in Test Error Among Critical Points). From Eq.(10), the first term dominates the test error, while the remaining terms are comparatively small. Based on the decomposition in Remark 3.6, the leading term reflects the bias, with the remaining terms contributing to the variance (see Appendix E.2 for details). As a result, the overall test error exhibits only minor variation across different critical points. This contrasts with the model with no skip connections, where different critical points can significantly influence both the bias and variance components. The right subfigure of Figure 3 illustrates this.

The Impact of Skip Connections on the Test Error Curve Observe first that the variance term in Eq.(9) is responsible for the sharp increase in the test curves as the ratio c approaches 1, due to the inclusion of  $(c-1)^{-1}$ . A similar albeit less pronounced trend is observed in the model with skip connections. The term  $\frac{\eta_{\rm ls}^2|I^a|}{d^2}\frac{c}{c-1}\sum_{i=1}^r\frac{\sigma_i^2}{(\eta_{\rm lm}^2+\sigma_i^2)}$  also includes the factor  $(c-1)^{-1}$ . However, in contrast to the model without skip connections, the expression is multiplied with an additional factor of  $d^{-1}$ . This suggests that skip connections help mitigate the sharp rise in variance that typically occurs when the model is in the moderately overparameterized regime, leading to more stable generalization performance even in this regime. We now examine this intriguing phenomenon more closely to better understand its origin.

# 4 Explaining the Variance Discrepancy Between DAEs With and Without Skip Connections

As seen in the previous section, the variance term is scaled by an additional  $d^{-1}$  factor for models with skip connections. As Figure 3 illustrates, this difference becomes particularly pronounced as c = d/n approaches 1. However, the underlying cause of this behavior is not immediately clear.

In this section, we identify the source of the discrepancy. We remind the reader of Eq. (2), which shows that the skip connection effectively cancels the signal component, reorienting the learning task toward predicting the noise. We demonstrate that this shift leads to weaker alignment between certain singular vector pairs, which in turn yields a substantially smaller variance.

For the remainder of this section, let  $\bar{\lambda}_j$ ,  $\lambda_j$ , and  $\lambda_j^{\mathbf{A}}$  denote the squared j-th singular values of the matrices  $\mathbf{X} + \mathbf{A}$ ,  $\mathbf{X}$ , and  $\mathbf{A}$ , respectively. Similarly, let  $\bar{\mathbf{V}}$ ,  $\mathbf{V}$ , and  $\mathbf{V}_{\mathbf{A}}$  denote the corresponding matrices of right singular vectors for  $\mathbf{X} + \mathbf{A}$ ,  $\mathbf{X}$ , and  $\mathbf{A}$ . Then, the Frobenius norms of the global minimizers with and without skip connections are given by:

$$\|\mathbf{W}_{*}^{\text{sc}}\|_{F}^{2} = \sum_{j=1}^{n} \bar{\lambda}_{j}^{-1} \sum_{i=1}^{k} \lambda_{i}^{\mathbf{A}} (\mathbf{V}_{\mathbf{A}}^{\mathsf{T}} \bar{\mathbf{V}})_{ij}^{2} \quad \text{and} \quad \|\mathbf{W}_{*}\|_{F}^{2} = \sum_{j=1}^{n} \bar{\lambda}_{j}^{-1} \sum_{i=1}^{k} \lambda_{i} (\mathbf{V}^{\mathsf{T}} \bar{\mathbf{V}})_{ij}^{2}.$$
 (11)

Note that  $\bar{\lambda}_j$  is shared between both models. Moreover, under data assumptions 3.1, the k first squared singular values  $\lambda_i$  and  $\lambda_i^{\mathbf{A}}$  scale at the same rate. Therefore, any substantial difference between the two must be attributed to what we refer to as the **alignment terms**  $(\mathbf{V}^{\top}\bar{\mathbf{V}})_{ij}^2$  and  $(\mathbf{V}_{\mathbf{A}}^{\top}\bar{\mathbf{V}})_{ij}^2$ .

Specifically, for  $i \in \{1,\ldots,k\}$  and  $j \in \{1,\ldots,n\}$ , the term  $(\mathbf{V}^\top \bar{\mathbf{V}})_{ij}^2$  quantifies the squared inner product between the i-th singular vector of the signal matrix  $\mathbf{X}$  and the j-th singular vector of the signal-plus-noise matrix  $\mathbf{X} + \mathbf{A}$ . In contrast,  $(\mathbf{V}_{\mathbf{A}}^\top \bar{\mathbf{V}})_{ij}^2$  captures the corresponding alignment between the noise matrix  $\mathbf{A}$  and  $\mathbf{X} + \mathbf{A}$ . Ideally, we would study the alignment behavior through the "Information-plus-Noise" model [16], defined by  $(\mathbf{X} + \mathbf{A})(\mathbf{X} + \mathbf{A})^\top$ . However, this is technically challenging. Therefore, we instead focus on a simpler, but conceptually closely related model<sup>2</sup>.

**Definition 4.1** (Rank-1 Additive Model). Let  $\mathbf{X}$  and  $\mathbf{A}$  satisfy Assumptions 3.1. Further assume that  $\mathbf{X}$  is rank-1, i.e.,  $\mathbf{X} = \sqrt{\lambda_1}\mathbf{u}_1\mathbf{v}_1^{\mathsf{T}}$ , where  $\sqrt{\lambda_1}$  is the largest singular value of  $\mathbf{X}$ , and  $\mathbf{u}_1$ ,  $\mathbf{v}_1^{\mathsf{T}}$  are the corresponding singular vectors. We then define the additive model  $\mathbf{S} := \mathbf{X}\mathbf{X}^{\mathsf{T}} + \mathbf{A}\mathbf{A}^{\mathsf{T}}$ .

**Remark 4.2** (Connection between Additive Models and Information-plus-noise Models). The expected value of the Information-plus-Noise model coincides with that of the Additive model:  $\mathbb{E}[(\mathbf{X} + \mathbf{A})(\mathbf{X} + \mathbf{A})^{\top}] = \mathbb{E}[\mathbf{X}\mathbf{X}^{\top} + \mathbf{A}\mathbf{A}^{\top}]$ , since  $\mathbb{E}[\mathbf{X}\mathbf{A}^{\top}] = \mathbb{E}[\mathbf{A}\mathbf{X}^{\top}] = 0$ .

With this simplified model, our focus is on comparing the alignment between the top eigenvector of  $\mathbf{X}\mathbf{X}^{\top}$  and the j-th eigenvector of  $\mathbf{S}$  (Denoted as  $\mathbf{u}_{j}^{\mathbf{S}}$ ), and the alignment between the top eigenvectors of  $\mathbf{A}\mathbf{A}^{\top}$  (Denoted as  $\mathbf{u}_{i}^{\mathbf{A}}$  for  $i \in [k]$ ) and the same  $\mathbf{u}_{j}^{\mathbf{S}}$ . Denote by  $\lambda_{j}^{\mathbf{S}}$  the j-th eigenvalue of  $\mathbf{S}$ . The corresponding proofs of the theorem below are given in Appendix F.2.

**Theorem 4.3** (Skip Connections Cause Weaker Alignment). For  $c \in (0, \infty)$ ,  $i \in [k]$ , and  $j \in [2, n] \setminus \{i - 1, i\}$ ,

$$\mathbb{E}\left[\frac{\langle \mathbf{u}_i^{\mathbf{A}}, \mathbf{u}_j^{\mathbf{S}} \rangle^2}{\langle \mathbf{u}_1, \mathbf{u}_j^{\mathbf{S}} \rangle^2}\right] = \Theta\left(\frac{1}{d(\lambda_i^{\mathbf{A}} - \lambda_j^{\mathbf{S}})^2}\right). \tag{12}$$

This theorem suggests that when the eigenvalue gap is large, the alignment term in models with skip connections becomes significantly weaker than in models without them. To see this, the first point to notice is that the eigenvalues of **S** follow the *Marchenko–Pastur distribution*. This implies that as *c* approaches 1, an increasing number of eigenvalues of **S** concentrate near zero (see Appendix F.1).

Now consider the term  $(\lambda_i^{\mathbf{A}} - \lambda_j^{\mathbf{S}})^2$ . For example, take i=1. It is known that  $\lambda_1^{\mathbf{A}}$  converges almost surely to  $\eta_{\mathrm{tst}}^2 c^{-1} (1+\sqrt{c})^2$  as  $d,n\to\infty$  [6, Theorem 5.8]. For small  $\lambda_j^{\mathbf{S}}$ , which care plentiful when

<sup>&</sup>lt;sup>2</sup>This simplified setting belongs to a broader class of models where a low-rank signal is perturbed by additive noise, commonly referred to as "spiked models" [7]. We formally define this model class in Appendix C.3. For a comprehensive overview, see [12].

 $c\approx 1$  and concentrate around 0, we obtain  $(\lambda_i^{\mathbf{A}}-\lambda_j^{\mathbf{S}})^2=\Theta(1)$ . As a result, the ratio between the alignment terms is  $\Theta(d^{-1})$ . This is important because the Frobenius norms in Eq. (11) involve the *inverses* of the eigenvalues of the corrupted input covariance matrix. Hence, small eigenvalues dominate the variance. More precisely, it is known that with high probability, the smallest eigenvalue of  $\mathbf{A}\mathbf{A}^{\top}$  scales as  $\Theta((\sqrt{d}-\sqrt{n-1})^2)$  [36]. At c=d/n=1 this scales as  $\Theta(d^{-2})$ , implying that its inverse is of order  $\Theta(d^2)$ . Accordingly, the Frobenius norms are largely influenced by the smallest eigenvalues. Importantly, for those small eigenvalues, Theorem 4.3 shows that the corresponding alignment terms in skip-connected models are *suppressed* by a factor of  $\Theta(d^{-1})$  relative to models without skip connections. This directly accounts for the reduced variance and explains the smoother generalization curves observed in the previous section.

## 5 Discussion

Bottleneck Dimension as an Additional Complexity Measure In supervised settings, [14] empirically show that by controlling an additional complexity measure along with the input parameter count, the test error curve can take on diverse shapes, ranging from the traditional U-shaped curve to ones exhibiting multiple descents. Our work provides concrete theoretical evidence that this interpretation extends to unsupervised scenarios, identifying the bottleneck dimension as a key complexity measure in DAEs. Beyond this, our findings uncover a subtle yet crucial distinction for denoising autoencoders: the emergence of a bias—variance trade-off, governed by the number of neurons in the bottleneck layer. This phenomenon is not considered in [14], whose focus is on Principal Component Regression (PCR) rather than denoising. We elaborate on this further in the next paragraph.

**PCA-based Methods vs. Two-Layer Linear DAEs** The key difference lies in where the dimensionality reduction happens. While PCA-based methods (including PCR [40, 20, 14] and PCA-denoising [13]) identify the top-k input components, Theorem 2.4 shows that two-layer linear DAEs align with the top-k directions of the *output*. This is a consequence of the critical points identified in Corollaries 2.4–2.5, which lead to generalization behavior distinct from PCA (cf. Figure 1). Theorem 3.2 and Figure 3 highlight that the double descent phenomenon as a function of d/n persists, even for small bottleneck dimension k. In particular, the variance term becomes dominant near  $c \approx 1$ , significantly influencing the test error—an effect that diminishes with increasing overparameterization. Conversely, PCA-based methods suppress this peaking behavior by discarding small eigenvalues of the input data.

The Role of Skip Connections & Diffusion Models Instead of eliminating small eigenvalues, skip connections attenuate their contribution. This improves generalization in certain regimes of d/n, which is in line with previous works [13]. Theorems 3.5 and 4.3 extend this understanding by identifying the variance term as the primary driver of the improvement. Interestingly, the input–output structure in Eq. (2) mirrors that of diffusion models [24], where the network is trained to predict noise rather than clean signals—a design choice shown to improve performance empirically [24, Sec. 3.2]. In relation to this, our results suggest two explanations: first, reduced variance; and second, as shown in the right plot of Figure 3, the presence of non-global critical points that perform comparably to the global minimizer, potentially easing optimization.

**Interpolation & Double Descent Phenomenon** Broadly, the double descent phenomenon suggests that generalization error peaks near the "interpolation threshold" [10] where a network attains (nearly) zero training error, before decreasing again as the model enters the overparameterized regime. Empirical works such as [10], [32], as well as prior theoretical works such as [22], [5] show that this second descent occurs after the model interpolates the training data. However, when the bottleneck dimension is lower than the rank of the input and output data, *exact interpolation is not possible*. However, our results reveal that a peak can still exist in the regime  $d/n \approx 1$  (cf. k < r in Figure 2).

**Future Work & Limitations** Our paper opens several pathways for future work to explore. While our setting strictly extends prior works on single-layer networks [38, 27] to two-layer networks with bottleneck (and skip connection), it remains restricted to the linear regime. Although [13] explores a non-linear setting, their analysis relies on a tied-weights assumption, and their resulting formulas do not capture the mechanisms we highlight. Extending our work to non-linear models is therefore an important direction for future work. Furthermore, our study focuses on two-layer networks with a

single skip connection from input to output. Investigating deeper architectures with skip connections between intermediate layers may yield deeper insights into their role in shaping generalization.

## References

- [1] El Mehdi Achour, François Malgouyres, and Sébastien Gerchinovitz. *The loss landscape of deep linear neural networks: a second-order analysis*. 2024. arXiv: 2107.13289 [math.ST]. URL: https://arxiv.org/abs/2107.13289.
- [2] Ben Adlam, Jake A. Levinson, and Jeffrey Pennington. "A Random Matrix Perspective on Mixtures of Nonlinearities in High Dimensions". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, 28–30 Mar 2022, pp. 3434–3457. URL: https://proceedings.mlr.press/v151/adlam22a.html.
- [3] Lars Valerian Ahlfors and Lars V Ahlfors. *Complex analysis*. Vol. 3. McGraw-Hill New York, 1979.
- [4] Jimmy Ba et al. "Generalization of two-layer neural networks: An asymptotic viewpoint". In: *International conference on learning representations*. 2020.
- [5] Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. 2023. arXiv: 2303.01372 [cs.LG]. URL: https://arxiv.org/abs/2303. 01372.
- [6] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*. Vol. 20. Springer, 2010.
- [7] Jinho Baik and Jack W. Silverstein. *Eigenvalues of Large Sample Covariance Matrices of Spiked Population Models*. 2004. arXiv: math/0408165 [math.ST]. URL: https://arxiv.org/abs/math/0408165.
- [8] Pierre Baldi and Kurt Hornik. "Neural networks and principal component analysis: Learning from examples without local minima". In: *Neural networks* 2.1 (1989), pp. 53–58.
- [9] Peter L. Bartlett et al. "Benign overfitting in linear regression". In: *Proceedings of the National Academy of Sciences* 117.48 (Apr. 2020), pp. 30063–30070. ISSN: 1091-6490. DOI: 10.1073/pnas.1907378117. URL: http://dx.doi.org/10.1073/pnas.1907378117.
- [10] Mikhail Belkin et al. "Reconciling modern machine-learning practice and the classical bias-variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32 (2019). ISSN: 1091-6490. DOI: 10.1073/pnas.1903070116. URL: http://dx.doi.org/10.1073/pnas.1903070116.
- [11] George W. Bohrnstedt and Arthur S. Goldberger. "On the Exact Covariance of Products of Random Variables". In: *Journal of the American Statistical Association* 64.328 (1969), pp. 1439–1442. ISSN: 01621459, 1537274X. URL: http://www.jstor.org/stable/2286081 (visited on 08/10/2024).
- [12] Romain Couillet and Zhenyu Liao. Random Matrix Methods for Machine Learning. https://zhenyu-liao.github.io/book/. Cambridge University Press, 2022. DOI: 10.1017/9781009128490.
- [13] Hugo Cui and Lenka Zdeborová. "High-dimensional asymptotics of denoising autoencoders". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 11850–11890.
- [14] Alicia Curth, Alan Jeffares, and Mihaela van der Schaar. "A U-turn on Double Descent: Rethinking Parameter Counting in Statistical Learning". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. URL: https://openreview.net/forum?id=00Lz8XZT2b.
- [15] P. T. Davies and M. K-S. Tso. "Procedures for Reduced-Rank Regression". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31.3 (1982), pp. 244–255. ISSN: 00359254, 14679876. URL: http://www.jstor.org/stable/2347998 (visited on 01/30/2025).
- [16] R Brent Dozier and Jack W Silverstein. "On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices". In: *Journal of Multivariate Analysis* 98.4 (2007), pp. 678–694.
- [17] Rick Durrett. Probability: theory and examples. Vol. 49. Cambridge university press, 2019.

- [18] Carl Eckart and Gale Young. "The approximation of one matrix by another of lower rank". In: *Psychometrika* 1.3 (1936), pp. 211–218.
- [19] Xinlong Feng and Zhinan Zhang. "The rank of a random matrix". In: *Applied mathematics and computation* 185.1 (2007), pp. 689–694.
- [20] Daniel Gedon, Antonio H. Ribeiro, and Thomas B. Schön. "No Double Descent in Principal Component Regression: A High-Dimensional Analysis". In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, 21–27 Jul 2024, pp. 15271–15293. URL: https://proceedings.mlr.press/v235/gedon24a.html.
- [21] Cédric Gerbelot, Alia Abbara, and Florent Krzakala. Asymptotic errors for convex penalized linear regression beyond Gaussian matrices. 2020. arXiv: 2002.04372 [stat.ML]. URL: https://arxiv.org/abs/2002.04372.
- [22] Trevor Hastie et al. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. 2020. arXiv: 1903.08560 [math.ST]. URL: https://arxiv.org/abs/1903.08560.
- [23] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV]. URL: https://arxiv.org/abs/1512.03385.
- [24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG]. URL: https://arxiv.org/abs/2006.11239.
- [25] Dulhan Hansaja Jayalath, Alisia Maria Lupidi, and Yonatan Gideoni. No Double Descent in Self-Supervised Learning. 2023. URL: https://openreview.net/forum?id= qNJRvdKDGYg.
- [26] Iain M. Johnstone. "On the Distribution of the Largest Eigenvalue in Principal Components Analysis". In: *The Annals of Statistics* 29.2 (2001), pp. 295–327. ISSN: 00905364, 21688966. URL: http://www.jstor.org/stable/2674106 (visited on 01/26/2025).
- [27] Chinmaya Kausik, Kashvi Srivastava, and Rishi Sonthalia. *Double Descent and Overfitting under Noisy Inputs and Distribution Shift for Linear Denoisers*. 2024. arXiv: 2305.17297 [cs.LG]. URL: https://arxiv.org/abs/2305.17297.
- [28] Kenji Kawaguchi. *Deep Learning without Poor Local Minima*. 2016. arXiv: 1605.07110 [stat.ML]. URL: https://arxiv.org/abs/1605.07110.
- [29] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009).
- [30] Alisia Lupidi, Yonatan Gideoni, and Dulhan Jayalath. "Does Double Descent Occur in Self-Supervised Learning?" In: *arXiv preprint arXiv:2307.07872* (2023).
- [31] Ashin Mukherjee and Ji Zhu. "Reduced rank ridge regression and its kernel extensions". In: *Statistical analysis and data mining: the ASA data science journal* 4.6 (2011), pp. 612–622.
- [32] Preetum Nakkiran et al. *Deep Double Descent: Where Bigger Models and More Data Hurt.* 2019. arXiv: 1912.02292 [cs.LG]. URL: https://arxiv.org/abs/1912.02292.
- [33] Arnu Pretorius, Steve Kroon, and Herman Kamper. "Learning dynamics of linear denoising autoencoders". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 4141– 4150.
- [34] Kobi Rahimi, Tom Tirer, and Ofir Lindenbaum. "Multiple Descents in Unsupervised Learning: The Role of Noise, Domain Shift and Anomalies". In: *arXiv preprint arXiv:2406.11703* (2024).
- [35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV]. URL: https://arxiv.org/abs/1505.04597.
- [36] Mark Rudelson and Roman Vershynin. "Smallest singular value of a random rectangular matrix". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 62.12 (2009), pp. 1707–1739.
- [37] J.W. Silverstein and Z.D. Bai. "On the Empirical Distribution of Eigenvalues of a Class of Large Dimensional Random Matrices". In: *Journal of Multivariate Analysis* 54.2 (1995), pp. 175–192. ISSN: 0047-259X. DOI: https://doi.org/10.1006/jmva.1995.1051. URL: https://www.sciencedirect.com/science/article/pii/S0047259X85710512.
- [38] Rishi Sonthalia and Raj Rao Nadakuditi. *Training Data Size Induced Double Descent For Denoising Neural Networks and the Role of Training Noise Level*. 2022. URL: https://openreview.net/forum?id=5ALGcXpmFyC.
- [39] Terence Tao. *Topics in random matrix theory*. Vol. 132. American Mathematical Soc., 2012.

- [40] Ningyuan Teresa, David W. Hogg, and Soledad Villar. "Dimensionality Reduction, Regularization, and Generalization in Overparameterized Regressions". In: SIAM Journal on Mathematics of Data Science 4.1 (Feb. 2022), pp. 126–152. ISSN: 2577-0187. DOI: 10.1137/20m1387821. URL: http://dx.doi.org/10.1137/20M1387821.
- [41] Matthew Trager, Kathlén Kohn, and Joan Bruna. Pure and Spurious Critical Points: a Geometric Study of Linear Networks. 2020. arXiv: 1910.01671 [cs.LG]. URL: https://arxiv.org/abs/1910.01671.
- [42] Madeleine Udell and Alex Townsend. Why are Big Data Matrices Approximately Low Rank? 2018. arXiv: 1705.07474 [cs.LG]. URL: https://arxiv.org/abs/1705.07474.
- [43] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.
- [44] Pascal Vincent et al. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." In: *Journal of machine learning research* 11.12 (2010).
- [45] Yimin Wei. "The weighted Moore-Penrose inverse of modified matrices". In: Applied Mathematics and Computation 122.1 (2001), pp. 1-13. ISSN: 0096-3003. DOI: https://doi.org/10.1016/S0096-3003(00)00007-2. URL: https://www.sciencedirect.com/science/article/pii/S0096300300000072.
- [46] Shuo Xiang et al. "Optimal exact least squares rank minimization". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.* 2012, pp. 480–488.
- [47] Yi Zhou and Yingbin Liang. Critical Points of Neural Networks: Analytical Forms and Landscape Properties. 2017. arXiv: 1710.11205 [stat.ML]. URL: https://arxiv.org/abs/1710.11205.

## **Appendix**

We include additional material in the appendix. Section A discusses further related work. Section B introduces general notational conventions. Section C provides background on Random Matrix Theory to support the results in Section F. Section D presents the proofs for Section 3 and proposes a definition of bias—variance tailored to our setting. Section F proves the results from Section 4 and includes additional supporting results. Section G describes the experimental settings used to generate the figures in the main text.

## A Related Works

In this work, we focus on two-layer linear networks and emphasize that the current progress remains limited to linear models. Prior studies of two-layer non-linear architectures typically impose additional architectural constraints, such as random projection methods (where one layer is fixed) or weight-tying assumptions between the layers. These constraints hinder a complete analysis of two-layer networks. To the best of our knowledge, our work is the first to address the full two-layer linear network architecture in the overparameterized setting without such restrictions, although our analysis is confined to the denoising setting. Additional related works are discussed in detail, where we also clarify how our approach compares to prior literature.

Loss Landscape and Critical Points of Linear Neural Network While the loss landscape properties of training metric are well studied using standard linear algebra [8, 28], second-order analysis [1], and algebraic geometry [41], comparatively few works focused on the analytical characterization of solutions and the generalization error. [8] provided an analytical description of the critical points in two-layer linear networks under the underparameterized setting, and [47] extended this to multi-layer linear networks. Yet these efforts leave open questions regarding regularized solutions, and the corresponding generalization error. This paper builds on this body of work by deriving regularized solutions for two layer linear DAEs and analyzing their generalization error in the context of bottleneck layers.

**Characterizing Generalization Behavior in Overparameterized Setting** The double descent phenomenon has emerged as a key insight in understanding generalization in overparameterized

models, prompting significant interest in analyzing simple models within this regime. This behavior has been particularly well studied in linear regression with Gaussian inputs, often through the lens of minimum-norm solutions and regularized settings such as Lasso [21] and ridge regression [24, 5]. Extensions to nonlinear two-layer neural networks have been explored in [5, 4], where one of the weight matrices is trained while the other is fixed randomly. These studies show that double descent arises when the hidden layer dimension scales proportionally with the number of data points, in both nonlinear and linear settings—mirroring similar findings in the PCA setting [20, 40]. In contrast, our work takes the opposite approach in the context of denoising autoencoders (DAEs): we analyze the two-layer linear case without fixing either weight matrix and without scaling the hidden layer width with the dataset size. This allows for a complete characterization of the generalization behavior in a fully trainable two-layer linear architecture.

**Discussion of Double Descent Phenomenon in Auto-Encoder Setting** There is ongoing debate about the presence of the double descent phenomenon in autoencoder models. For example, [30] conjectured its absence in self-supervised learning tasks and provided empirical evidence using reconstruction autoencoders (RAEs). In contrast, [34] demonstrated that double descent does occur in deep undercomplete RAEs, particularly when the input data includes noisy measurements. Although these studies focus on RAEs, their findings—alongside ours—suggest that input noise plays a critical role in shaping the generalization behavior of undercomplete autoencoders.

## **B** Notations

#### **B.1** General Notations

For a matrix  $\mathbf{M} \in \mathbb{R}^{d \times n}$ , we use

- $\text{Tr}(\mathbf{M})$ ,  $\|\mathbf{M}\|_F$ ,  $\|\mathbf{M}\|_2$ ,  $\mathbf{M}^{\dagger}$ , and  $\mathbf{m}_i$  to denote the Trace, Frobenius norm, spectral norm, Moore-Penrose pseudoinverse, and i-th column vector of  $\mathbf{M}$ , respectively.
- $r_{\mathbf{M}}$ ,  $\mathbb{I}_d$  to denote the rank of M, and the identity matrix of size  $d \times d$ .
- For an index set  $I \subset \{1, ..., n\}$ ,  $\mathbf{M}_I$  denotes the submatrix of  $\mathbf{M}$  with its columns indexed by I.  $\mathbf{M}_p$  denotes the submatrix of  $\mathbf{M}$  with its first p columns.

Some set notations:

- $[p] = \{1, 2, \dots, p\}$  denotes the set of natural numbers from 1 to p.
- Similarly, [p, q] denotes  $\{p, p+1, \ldots, q\}$ , for  $p, q \in \mathbb{N}$ .
- |I| denotes the cardinality of a set I.

## Additionally,

- $\mathbf{v}$  denotes a vector, and  $\|\mathbf{v}\|_2$  denotes its Euclidean norm.
- For some vector  $\mathbf{v}$  and  $\mathbf{u}$ ,  $\langle \mathbf{v}, \mathbf{u} \rangle$  denotes the inner product of  $\mathbf{v}$  and  $\mathbf{u}$ .
- For  $a, b \in \mathbb{R}$ , we denote  $(a, b)^+ = \max(a, b)$  and  $(a, b)^- = \min(a, b)$ .
- Im(z) for  $z \in \mathbb{C}$  denotes the imaginary part of z.
- $a \simeq b$  denotes a converges almost surely to b.
- \$\mathbb{1}\_{\text{condition}}\$ denotes the indicator function, which is 1 if the condition is satisfied, and 0 otherwise.

**Big-O Notation** Throughout this work, we use standard asymptotic notation. Specifically,  $O(\cdot)$  denotes an upper bound up to constant factors—that is, a quantity that grows no faster than the reference function. In contrast,  $\Theta(\cdot)$  denotes a tight asymptotic bound, meaning the quantity grows at the same rate as the reference function, up to constant factors. Finally,  $o(\cdot)$  denotes a lower-order term that becomes negligible compared to the reference term in the asymptotic regime.

## **B.2** Singular Value Decomposition(SVD)

For an arbitrary matrix  $\mathbf{M} \in \mathbb{R}^{d \times n}$ ,

•  $\mathbf{U_M} \mathbf{\Sigma_M} \mathbf{V_M}^{\top}$  denotes the SVD of  $\mathbf{M}$ , where  $\mathbf{U_M} \in \mathbb{R}^{d \times d}$ ,  $\mathbf{\Sigma_M} \in \mathbb{R}^{d \times n}$ , and  $\mathbf{V_M} \in \mathbb{R}^{n \times n}$ 

- $\mathbf{u}_i^{\mathbf{M}}$  and  $\sigma_i^{\mathbf{M}}$  are the *i*-th column vector of  $\mathbf{U}_{\mathbf{M}}$  and the *i*-th singular value of  $\mathbf{M}$ , respectively.
- $\tilde{\mathbf{U}}_{\mathbf{M}}\mathbf{D}_{\mathbf{M}}\tilde{\mathbf{V}}_{\mathbf{M}}^{\top} = \mathbf{U}_{\mathbf{M},r_{\mathbf{M}}}\mathbf{D}_{\mathbf{M}}\mathbf{V}_{\mathbf{M},r_{\mathbf{M}}}$  is the *reduced SVD* of  $\mathbf{M}$ , where  $\tilde{\mathbf{U}}_{M} := \mathbf{U}_{\mathbf{M},r_{\mathbf{M}}}$ ,  $\tilde{\mathbf{V}}_{M} := \mathbf{V}_{\mathbf{M},r_{\mathbf{M}}}$ , and  $\mathbf{D}_{\mathbf{M}} \in \mathbb{R}^{r_{\mathbf{M}} \times r_{\mathbf{M}}}$  is the diagonal matrix with the singular values of  $\mathbf{M}$ , i.e.,  $(\mathbf{D}_{\mathbf{M}})_{ij} = \sigma_i^{\mathbf{M}} \delta_{ij}$ , for  $\delta$  the Dirac Delta function.

For an index set  $I \subset \{1, \dots, r_{\mathbf{M}}\}$ , we use

•  $\hat{\Sigma}_{\mathbf{M},I} \in \mathbb{R}^{d \times n}$  to denote the matrix with its diagonal part consists of the singular values of  $\mathbf{M}$  indexed by I. Precisely,

$$(\hat{\mathbf{\Sigma}}_{\mathbf{M},I})_{ij} = egin{cases} \sigma_i^{\mathbf{M}} & \text{if } i=j \text{ and } i \in I \\ 0 & \text{otherwise} \end{cases}$$

•  $P_I(\mathbf{M}) := \mathbf{U}_{\mathbf{M}} \hat{\boldsymbol{\Sigma}}_{\mathbf{M},I} \mathbf{V}_{\mathbf{M}}^{\top} = \mathbf{U}_{\mathbf{M},I} \mathbf{D}_{\hat{\boldsymbol{\Sigma}}_{\mathbf{M}}} \mathbf{V}_{\mathbf{M},I}^{\top}$ 

For example, when applying Eckart-Young Theorem [18], we need the best rank-q approximation of a matrix M. For this special situation, we denote  $P_{[q]}(\mathbf{M})$  as  $P_q(\mathbf{M})$ . Then, it can be written as  $P_q(\mathbf{M}) = \mathbf{U}_{\mathbf{M}} \hat{\mathbf{\Sigma}}_{\mathbf{M},q} \mathbf{V}_{\mathbf{M}}^{\top} = \mathbf{U}_{\mathbf{M},q} \mathbf{D}_{\mathbf{M},q} \mathbf{V}_{\mathbf{M},q}^{\top}.$ Additionally, for a symmetric matrix  $\mathbf{S} \in \mathbb{R}^{d \times d}$ 

- $\mathbf{S} = \mathbf{U_S} \mathbf{\Lambda_S} \mathbf{U_S}^{\top}$  denotes the eigendecomposition of  $\mathbf{S}$ , where  $\mathbf{U_S} \in \mathbb{R}^{d \times d}$  and  $\mathbf{\Lambda_S} \in \mathbb{R}^{d \times d}$ .  $\lambda_i(\mathbf{S})$  or  $\lambda_i^{\mathbf{S}}$  denotes the i-th eigenvalue of  $\mathbf{S}$ .

## **Background on Random Matrix Theory**

Here, we present the notations and key tools from Random Matrix Theory that are frequently referenced in Appendix F. We include only the essential definitions and theorems, and refer the reader to [12, 6] for more comprehensive treatments.

## C.1 Getting Eigenvalue Information

We consider a Gaussian random matrix  $\tilde{\mathbf{A}} \in \mathbb{R}^{d \times n}$ , whose elements are independently and identically distributed(i.i.d) following a gaussian distribution of mean 0 and variance  $\sigma^2$ , i.e.,  $\mathcal{N}(0, \sigma^2)$ . The goal is to analyze the behavior of the eigenvalues of  $\frac{1}{n}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T$  as d, n grows together toward infinity, with their ratio satisfying  $\frac{d}{n} \to c \in (0, \infty)$ . Notably, the normalized histogram of the eigenvalues of  $\frac{1}{\pi} \ddot{\mathbf{A}} \ddot{\mathbf{A}}^T$  exhibits deterministic behavior as d, n grows, converging weakly to the Marchenko-Pastur distribution, which is denoted as  $\mu_{MP}$ . This can be written as follows.

$$\frac{1}{d} \sum_{i=1}^{d} \delta_{\lambda_{i}(\frac{1}{n}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^{T})} \to \mu_{MP} \text{ weakly.}$$
 (13)

To characterize this distribution, the Stieltjes transform is one of the most commonly used tools. We introduce the following definitions, following [12, Chapter 2].

1. (Resolvent). For a symmetric matrix  $\mathbf{M} \in \mathbb{R}^{d \times d}$ , the resolvent of  $\mathbf{M}$  is **Definition C.1** defined as

$$\mathbf{Q}_{\mathbf{M}}(\alpha) = (\mathbf{M} - \alpha \mathbb{I}_d)^{-1}, \quad \alpha \in \mathbb{C} \setminus \{\lambda_1^{\mathbf{M}}, \dots, \lambda_d^{\mathbf{M}}\}.$$

2. (Empirical Spectral Measure). The above eq. (13) is formally known as Empirical Spectral Measure, and for a symmetric matrix  $\mathbf{M} \in \mathbb{R}^{d \times d}$ , it is defined as

$$\mu_{\mathbf{M}} = \frac{1}{d} \sum_{i=1}^{d} \delta_{\lambda_i^{\mathbf{M}}}.$$

3. (Stieltjes Transform). For a real probability measure  $\mu$  with support supp $(\mu)$ , the Stieltjes Transform of  $\mu$  is defined as

$$m_{\mu}(\alpha) = \int \frac{1}{x - \alpha} d\mu(x), \quad \alpha \in \mathbb{C} \backslash \text{supp}(\mu).$$

Note that the stieltjes transform of the *empirical spectral measure* is given by,

$$m_{\mu_{\mathbf{M}}}(\alpha) = \int \frac{1}{x - \alpha} d\mu_{\mathbf{M}}(x)$$
$$= \frac{1}{d} \sum_{j=1}^{d} \frac{1}{\lambda_{j}^{\mathbf{M}} - \alpha}$$
$$= \frac{1}{d} \operatorname{Tr}(\mathbf{Q}_{\mathbf{M}}).$$

The stielties transform is used to characterize the following convergence.

$$\frac{1}{d}\operatorname{Tr}(\mathbf{Q}_{\frac{1}{n}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T}) \to \int \frac{1}{z-\alpha} d\mu_{MP}(z) =: m_{\mu_{MP}}.$$

Once we characterize how the stieltjes transform  $m_{\mu_{MP}}$  looks like, using the *Inverse Stieltjes* Transform ([12, Theorem 2.1]), the Marchenko-Pastur density  $\mu_{MP}$  can be retrieved.

**Theorem C.2** (Marcenko-Pastur Law, [6, Theorem 3.1.1]). For a random matrix  $\tilde{\mathbf{A}} \in \mathbb{R}^{d \times n}$ , whose entries are i.i.d random variables with mean 0 and finite variance  $\sigma^2$ , the empirical spectral measure of  $\frac{1}{n}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T$  converges weakly to  $\mu_{MP}$ , where

$$\mu_{MP}(dx) = (1 - \frac{1}{c})^{+} \delta_{0}(x) + \frac{1}{2\pi\sigma^{2}cx} \sqrt{(b - x)(x - a)} \mathbb{1}_{x \in [a, b]},$$

for the ratio c satisfies  $\frac{d}{n} \to c \in (0, \infty)$ ,  $a := \sigma^2 (1 - \sqrt{c})^2$ ,  $b := \sigma^2 (1 + \sqrt{c})^2$ .

The corresponding stieltjes transform  $m_{\mu_{MP}}$  satisfies the following quadratic equation:

$$c\alpha\sigma^2 m_{\mu_{MP}}^2 - (\sigma^2(1-c) - \alpha) m_{\mu_{MP}} + 1 = 0.$$
 (14)

 $c\alpha\sigma^2 m_{\mu_{MP}}^2 - \left(\sigma^2(1-c) - \alpha\right) m_{\mu_{MP}} + 1 = 0. \tag{1}$  The following lemma connects the resolvent of the sample covariance matrix  $\mathbf{Q}_{\frac{1}{n}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T}$  to  $m_{\mu_{MP}}$ .

**Lemma C.3** (Due to [12, Theorem 2.4]) For unit vectors  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ , it satisfies that  $\mathbf{a}^T \mathbf{Q}_{\perp \tilde{\mathbf{A}} \tilde{\mathbf{A}}^T} \mathbf{b} \simeq m_{\mu_{MP}} \mathbf{a}^T \mathbf{b}.$ 

We state important lemmas further that are frequently referred in Appendix F.

**Lemma C.4** 1. (Resolvent Identity, [12, Chapter 2]). Let A, B invertible matrices. Then,  $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ .

- 2. (Sherman-Morrison Formula). Let  $\mathbf{A} \in \mathbb{R}^{p \times p}$ , and  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{p}$ . For  $1 + \mathbf{y}^{T} \mathbf{A} \mathbf{x} \neq 0$ , we have that  $(\mathbf{A} + \mathbf{x} \mathbf{y}^{T})^{-1} = \mathbf{A}^{-1} \frac{\mathbf{A}^{-1} \mathbf{x} \mathbf{y}^{T} \mathbf{A}^{-1}}{1 + \mathbf{v}^{T} \mathbf{A} \mathbf{x}}$ .
- 3. (Concentration of Quadratic Forms, Gaussian case). For  $\alpha \in \mathbb{C}\backslash\mathbb{R}_+$ , let  $\tilde{\mathbf{Q}}(\alpha) =$  $\left(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T - \alpha \mathbb{I}_d\right)^{-1} = \left(\sum_{i=1}^N \tilde{\mathbf{a}}_i \tilde{\mathbf{a}}_i^T - \alpha \mathbb{I}_d\right)^{-1}$ , where  $\tilde{A} \in \mathbb{R}^{d \times N}$  is a i.i.d real gaussian random matrix, whose entries are sampled from  $\mathcal{N}(0,1)$ . Let  $\tilde{\mathbf{Q}}_{-j}(\alpha)$  be  $\tilde{\mathbf{Q}}(\alpha)$  with j-th row and column removed. Then,  $\frac{1}{d}\tilde{\mathbf{a}}_{i}^{T}\tilde{\mathbf{Q}}_{-j}\tilde{\mathbf{a}}_{j} \xrightarrow{a.s} \frac{1}{d}\operatorname{Tr}(\tilde{\mathbf{Q}}_{-j})$ . (Proof.)

We have from Hanson-Wright Inequality [43], that for  $\epsilon > 0$ , there exists C > 0 such that

$$\begin{split} & \mathbb{P}\left(\left|\frac{1}{d}\tilde{\mathbf{a}}_{j}^{T}\tilde{\mathbf{Q}}_{-j}\tilde{\mathbf{a}}_{j} - \frac{1}{d}\mathbb{E}\left[\tilde{\mathbf{a}}_{j}^{T}\tilde{\mathbf{Q}}_{-j}\tilde{\mathbf{a}}_{j}\right]\right| > \epsilon\right) \leq 2\exp(-\epsilon dC\|\tilde{\mathbf{Q}}_{-j}\|_{2}^{-1}). \\ & \textit{Due to } \|\tilde{\mathbf{Q}}_{-j}\|_{2}^{-1} = \left(\min_{1 \leq k \leq d} \left|\lambda_{k}^{\tilde{\mathbf{A}}\tilde{\mathbf{A}}^{T}} - \alpha\right|\right)^{-1} < \infty, \\ & \lim_{d \to \infty} \sum_{n=1}^{d} \mathbb{P}\left(\left|\frac{1}{n}\tilde{\mathbf{a}}_{j}^{T}\tilde{\mathbf{Q}}_{-j}\tilde{\mathbf{a}}_{j} - \frac{1}{n}\mathbb{E}[\tilde{\mathbf{a}}_{j}^{T}\tilde{\mathbf{Q}}_{-j}\tilde{\mathbf{a}}_{j}]\right| > \epsilon\right) \\ & = \lim_{d \to \infty} \sum_{n=1}^{d} \left(2\exp(-\epsilon nC\|\tilde{\mathbf{Q}}_{-j}\|_{2}^{-1})\right) < \infty. \end{split}$$

The application of Borel-Cantelli Lemma[17, Theorem 2.3.1] gives us that  $\frac{1}{d}\tilde{\mathbf{a}}_{j}^{T}\tilde{\mathbf{Q}}_{-j}\tilde{\mathbf{a}}_{j} \xrightarrow{a.s} \frac{1}{d}\mathbb{E}[\tilde{\mathbf{a}}_{j}^{T}\tilde{\mathbf{Q}}_{-j}\tilde{\mathbf{a}}_{j}] = \frac{1}{d}\operatorname{Tr}(\tilde{\mathbf{Q}}_{-j}).$ 

4. (Minimal Effect of Finite-Rank Perturbations Inside a Trace, due to [37, Lemma 2.6]). For symmetric and positive semi-definite  $\mathbf{B}, \mathbf{M} \in \mathbb{R}^{d \times d}$ , and  $\alpha \in \mathbb{C} \setminus \mathbb{R}_+$ , let  $\mathbf{M} := \sum_{i=1}^h l_i \mathbf{x}_i \mathbf{x}_i^T$  for fixed h, and  $l_i \in \mathbb{R}$ ,  $\mathbf{x}_i \in \mathbb{R}^d$ . Then, we have that

$$|\operatorname{Tr}\left((\mathbf{B} - \alpha \mathbb{I})^{-1} - (\mathbf{B} + \mathbf{M} - \alpha \mathbb{I})^{-1}\right)| \le \frac{h}{Im(\alpha)}.$$

(Proof.)

The proof is a simple recursive argument of [37, Lemma 2.6], which states that

$$\left| \operatorname{Tr} \left( (\mathbf{B} - \alpha \mathbb{I})^{-1} - (\mathbf{B} + l \mathbf{x} \mathbf{x}^T - \alpha \mathbb{I})^{-1} \right) \right| \le \frac{1}{Im(\alpha)},$$

for some  $l \in \mathbb{R}$ , and  $\mathbf{x} \in \mathbb{R}^d$ . We repeat this h times to get the result.

5. (Weinstein-Aronszajn Identity). For  $\mathbf{A} \in \mathbb{R}^{p \times q}$ ,  $\mathbf{B} \in \mathbb{R}^{q \times p}$ , and  $\lambda \in \mathbb{R} \setminus \{0\}$ . Then, we have that

$$det(\mathbf{AB} - \lambda \mathbb{I}_p) = (-\lambda)^{p-q} det(\mathbf{BA} - \lambda \mathbb{I}_q).$$

6. (Woodbury Identity). For  $\mathbf{A} \in \mathbb{R}^{p \times p}$ ,  $\mathbf{U} \in \mathbb{R}^{p \times q}$ ,  $\mathbf{V} \in \mathbb{R}^{q \times p}$ , we have that

$$(\mathbf{A} + \mathbf{U}\mathbf{V}^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbb{I}_q + \mathbf{V}^T\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}^T\mathbf{A}^{-1}.$$

## C.2 Getting Subspace Information

One useful property of the resolvent and the Stieltjes transform is their connection to the eigenstructure of a random matrix. This connection stems from their resemblance to the Cauchy integral formula.

**Theorem C.5** (Cauchy Integral Formula [3, Thm. 6]). For a complex analytic function f(z) in a simply connected domain D with a simple pole at  $z_0 \in D$ , it satisfies that

$$f(z_0) = \frac{1}{2\pi i} \oint_{\gamma} \frac{f(z)}{z - z_0} dz,$$

where  $\gamma$  is a positively oriented simple closed curve around  $z_0$ .

By using the orthogonal decomposition of  $\frac{1}{n}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T = \mathbf{U}_{\tilde{\mathbf{A}}}\mathbf{\Lambda}_{\tilde{\mathbf{A}}}\mathbf{U}_{\tilde{\mathbf{A}}}^T$ ,  $\mathbf{Q}_{\frac{1}{n}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T}$  can be written down as

$$\mathbf{Q}_{\frac{1}{n}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T} = \sum_{j=1}^d \frac{1}{\lambda_j \left(\frac{1}{n}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T\right) - \alpha} \mathbf{u}_j^{\tilde{\mathbf{A}}} (\mathbf{u}_j^{\tilde{\mathbf{A}}})^T.$$

By defining  $\Gamma_{\lambda_j}$  as a positively oriented simple closed curve that *only* encloses  $\lambda_j\left(\frac{1}{n}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T\right)$ , we can express the eigenvector information of the random matrix as follows.

$$\mathbf{u}_{j}^{\tilde{\mathbf{A}}}(\mathbf{u}_{j}^{\tilde{\mathbf{A}}})^{T} = -\frac{1}{2\pi i} \oint_{\Gamma_{\lambda_{j}}} \mathbf{Q}_{\frac{1}{n}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^{T}} \, d\alpha.$$

For example, for a vector  $\mathbf{v} \in \mathbb{R}^d$ , the projection information of  $\mathbf{u}_i^{\tilde{\mathbf{A}}}$  onto  $\mathbf{v}$  is given by

$$\mathbf{v}^T \mathbf{u}_j^{\tilde{\mathbf{A}}} (\mathbf{u}_j^{\tilde{\mathbf{A}}})^T \mathbf{v} = \langle \mathbf{v}, \mathbf{u}_j^{\tilde{\mathbf{A}}} \rangle^2 = -\frac{1}{2\pi i} \oint_{\Gamma_{\lambda_j}} \mathbf{v}^T \mathbf{Q}_{\frac{1}{n} \tilde{\mathbf{A}} \tilde{\mathbf{A}}^T} \mathbf{v} \, d\alpha.$$

#### C.3 Spiked Models

The denoising problem considered in this work is closely related to the *Spiked Models* in Random Matrix Theory. The authors [12] have categorized the following models(among others) as *Additive Models*, emphasizing their common *additive structures*.

**Definition C.6** (Spiked-Covariance Model, due to [26]). For a rank r deterministic, and symmetric matrix  $\mathbf{P} \in \mathbb{R}^{d \times d}$ , we consider a covariance matrix  $\mathbf{C} \in \mathbb{R}^{d \times d}$ , which is defined as:

$$\mathbf{C} = \sum_{j=1}^{r} \lambda_i \mathbf{u}_j^{\mathbf{P}} (\mathbf{u}_j^{\mathbf{P}})^T + \mathbb{I}_d.$$

Then, for a random matrix  $\mathbf{A}_1 \in \mathbb{R}^{d \times n}$ , whose columns are i.i.d random vectors sampled from a distribution with mean 0 and covariance  $\mathbf{C}$ , the spiked-covariance model is defined as

$$\frac{1}{n}\mathbf{A}_1\mathbf{A}_1^T$$
.

**Definition C.7** (Information-plus-Noise Model) For a rank r deterministic matrix  $\mathbf{X}_2 \in \mathbb{R}^{d \times n}$ , and an i.i.d random matrix  $\mathbf{A}_2 \in \mathbb{R}^{d \times n}$ , the Information-plus-Noise model is defined as

$$\frac{1}{n}(\mathbf{X}_2 + \mathbf{A}_2)(\mathbf{X}_2 + \mathbf{A}_2)^T.$$

**Definition C.8** (Additive Model) For a rank r deterministic, and symmetric matrix  $\mathbf{P}_3 \in \mathbb{R}^{d \times d}$ , and an i.i.d random matrix  $\mathbf{A}_3 \in \mathbb{R}^{d \times n}$ , the Additive Model is defined as

$$\mathbf{P}_3 + \frac{1}{n} \mathbf{A}_3 \mathbf{A}_3^T.$$

These three models share a common structure, in that they all involve *adding the low-rank data to the random matrices*. Our denoising problem corresponds precisely to the Information-plus-Noise model. In Appendix F.2, we will leverage this structural similarity between these models to derive an interesting result, by alternatively considering the Additive Model.

## **D** Characterization of Critical Points

In this section, we first state and prove a detailed version of Theorem 2.1 (see Appendix D.1). In addition, we show that the global minimizer it identifies is equivalent to the solution obtained via reduced-rank regression under the minimum-norm principle, as shown in Appendix D.2.

#### D.1 Proofs of Section 2

Before stating the main theorem, we present two auxiliary lemmas. The first lemma establishes necessary conditions for critical points and their consequences, while the second lemma is important for characterizing each critical point. To facilitate the proofs of these lemmas, we introduce the following notation: let  $\mathbf{U_G} = [\mathbf{U_1} \quad \mathbf{U_2}]$ , where  $\mathbf{U_1} \in \mathbb{R}^{d \times r_{\mathbf{G}}}$  and  $\mathbf{U_2} \in \mathbb{R}^{d \times (\mathbf{D} - r_{\mathbf{G}})}$ . Additionally, for any matrix  $\mathbf{M}$ , we denote by  $P_{\mathbf{M}}$  the orthogonal projection matrix onto the column space of  $\mathbf{M}$ .

**Lemma D.1** (Necessary Conditions for Critical Points). *Consider the objective of minimizing (3).* Then, the necessary conditions for critical points  $\hat{\mathbf{W}}_2$ ,  $\hat{\mathbf{W}}_1$  are given as:

$$\mathbf{Y}\mathbf{Z}^{\top}\hat{\mathbf{W}}_{1}^{\top} = \hat{\mathbf{W}}_{2}\hat{\mathbf{W}}_{1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\top}\hat{\mathbf{W}}_{1}^{\top}.\tag{15}$$

$$\hat{\mathbf{W}}_{2}^{\mathsf{T}}\mathbf{Y}\mathbf{Z}^{\mathsf{T}} = \hat{\mathbf{W}}_{2}^{\mathsf{T}}\hat{\mathbf{W}}_{2}\hat{\mathbf{W}}_{1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathsf{T}}.$$
(16)

The following equations are implied by the necessary conditions:

$$\hat{\mathbf{W}} := \hat{\mathbf{W}}_2 \hat{\mathbf{W}}_1 = P_{\hat{\mathbf{W}}_2} \mathbf{Y} \mathbf{Z}^{\top} (\tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^{\top})^{-1}, \tag{17}$$

$$P_{\hat{\mathbf{W}}_2}\mathbf{G} = \mathbf{G}P_{\hat{\mathbf{W}}_2} = P_{\hat{\mathbf{W}}_2}\mathbf{G}P_{\hat{\mathbf{W}}_2}.$$
 (18)

*Proof.* The gradient flows of weight matrices  $W_1, W_2$  are:

$$\begin{split} \frac{d\mathbf{W}_1}{dt} &= \mathbf{W}_2^\top (\mathbf{Y} \mathbf{Z}^\top - \mathbf{W}_2 \mathbf{W}_1 \mathbf{Z} \mathbf{Z}^\top - \lambda \mathbf{W}_2 \mathbf{W}_1) \\ &= \mathbf{W}_2^\top \mathbf{Y} \mathbf{Z}^\top - \mathbf{W}_2^\top \mathbf{W}_2 \mathbf{W}_1 \tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^\top, \\ \frac{d\mathbf{W}_2}{dt} &= (\mathbf{Y} \mathbf{Z}^\top - \mathbf{W}_2 \mathbf{W}_1 \mathbf{Z} \mathbf{Z}^\top - \lambda \mathbf{W}_2 \mathbf{W}_1) \mathbf{W}_1^\top \\ &= \mathbf{Y} \mathbf{Z}^\top \mathbf{W}_1^\top - \mathbf{W}_2 \mathbf{W}_1 \tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^\top \mathbf{W}_1^\top. \end{split}$$

In order to find critical points, we set both of the gradient flows to zero. For the first condition, by taking the generalized inverse of  $\mathbf{W}_2^{\top}\mathbf{W}_2$ , we have that for any  $\mathbf{W}_1 \in \mathbb{R}^{k \times d}$ , the following holds:

$$\mathbf{W}_1 = (\mathbf{W}_2^\top \mathbf{W}_2)^\dagger \mathbf{W}_2^\top \mathbf{Y} \mathbf{Z}^\top (\tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^\top)^{-1} + (\mathbb{I} - (\mathbf{W}_2^\top \mathbf{W}_2)^\dagger \mathbf{W}_2^\top \mathbf{W}_2) \mathbf{L}_1$$

For the second condition  $\frac{d\mathbf{W}_2}{dt} = 0$ , this is equivalent to

$$\mathbf{Y}\mathbf{Z}^{\mathsf{T}}\mathbf{W}_{1}^{\mathsf{T}} = \mathbf{W}_{2}\mathbf{W}_{1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathsf{T}}\mathbf{W}_{1}^{\mathsf{T}}.\tag{19}$$

For a critical point  $\hat{\mathbf{W}} := \hat{\mathbf{W}}_2 \hat{\mathbf{W}}_1$ , it holds that

$$\hat{\mathbf{W}} = \hat{\mathbf{W}}_{2} \hat{\mathbf{W}}_{1} 
= \hat{\mathbf{W}}_{2} (\hat{\mathbf{W}}_{2}^{\top} \hat{\mathbf{W}}_{2})^{\dagger} \hat{\mathbf{W}}_{2}^{\top} \mathbf{Y} \mathbf{Z}^{\top} (\tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^{\top})^{-1} + (\hat{\mathbf{W}}_{2} - \hat{\mathbf{W}}_{2} (\hat{\mathbf{W}}_{2}^{\top} \hat{\mathbf{W}}_{2})^{\dagger} \hat{\mathbf{W}}_{2}^{\top} \hat{\mathbf{W}}_{2}) \mathbf{L}_{1} 
= P_{\hat{\mathbf{W}}_{2}} \mathbf{Y} \mathbf{Z}^{\top} (\tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^{\top})^{-1}.$$
(20)

Then, it follows from the second condition (19) that  $\mathbf{Y}\mathbf{Z}^{\top}\hat{\mathbf{W}}^{\top} = \hat{\mathbf{W}}(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\top})^{-1}\hat{\mathbf{W}}^{\top}$ , by multiplying  $\hat{\mathbf{W}}_{2}^{\top}$  on both sides. From eq. (20), we have that

$$P_{\hat{\mathbf{W}}_2}\mathbf{G}P_{\hat{\mathbf{W}}_2} = P_{\hat{\mathbf{W}}_2}\mathbf{G} = \mathbf{G}P_{\hat{\mathbf{W}}_2}.$$
 
$$P_{\hat{\mathbf{W}}_2}\mathbf{G} = \mathbf{G}P_{\hat{\mathbf{W}}_2} \text{ is due to } P_{\hat{\mathbf{W}}_2} = P_{\hat{\mathbf{W}}_2}^{\top}.$$

**Lemma D.2** Following from Lemma D.1, it holds that

$$P_{\mathbf{U}_{\mathbf{G}}^{\top}\hat{\mathbf{W}}_{2}} = \begin{bmatrix} \mathcal{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}, \tag{21}$$

where  $\mathcal{D}$  and  $\mathbf{W}$  satisfy the followings.

1. 
$$\mathbf{B} \in \mathbb{R}^{(\mathbf{D}-r_{\mathbf{G}})\times(\mathbf{D}-r_{\mathbf{G}})}$$
 is defined as  $\mathbf{B} := \mathbf{B}_1\mathbf{B}_1^{\top}$ , where  $\mathbf{B}_1 = ((\mathbf{U}_{\mathbf{G}}^{\top}\mathbf{U}_{\hat{\mathbf{W}}_2})_{i,j})_{r_{\mathbf{G}}+1 \leq i \leq d, r_{\mathbf{G}}+1 \leq j \leq d}$ .

Furthermore, B has  $r_{\mathbf{B}}$  eigenvalues equal to 1 and the rest equal to 0.

- 2.  $\mathcal{D} \in \mathbb{R}^{r_{\mathbf{G}} \times r_{\mathbf{G}}}$  is a diagonal matrix with its diagonal elements consist of  $r_{\mathcal{D}}$  number of 1's and  $r_{\mathbf{G}} r_{\mathcal{D}}$  number of 0's.
- 3.  $r_{\mathcal{D}} + r_{\mathbf{B}} = r_{\hat{\mathbf{W}}_2}$ .

*Proof.* We analyze the conditions (17) and (18) further. Consider the eigendecomposition of  $\mathbf{G}$ , given by  $\mathbf{G} = \mathbf{U}_{\mathbf{G}} \Lambda_{\mathbf{G}} \mathbf{U}_{\mathbf{G}}^{\top}$ , and the singular value decomposition of  $\hat{\mathbf{W}}_2$ , given by  $\hat{\mathbf{W}}_2 = \mathbf{U}_{\hat{\mathbf{W}}_2} \Sigma_{\hat{\mathbf{W}}_2} \mathbf{V}_{\hat{\mathbf{W}}_2}^{\top}$ . Then,

$$P_{\mathbf{U}_{\mathbf{G}}^{\top}\hat{\mathbf{W}}_{2}} = \mathbf{U}_{\mathbf{G}}^{\top}\hat{\mathbf{W}}_{2}(\mathbf{U}_{\mathbf{G}}^{\top}\hat{\mathbf{W}}_{2})^{\dagger} = \mathbf{U}_{\mathbf{G}}^{\top}\hat{\mathbf{W}}_{2}\hat{\mathbf{W}}_{2}^{\dagger}\mathbf{U}_{\mathbf{G}} = \mathbf{U}_{\mathbf{G}}^{\top}P_{\hat{\mathbf{W}}_{2}}\mathbf{U}_{\mathbf{G}}.$$
 (22)

Then, it satisfies that  $P_{\hat{\mathbf{W}}_2} = \hat{\mathbf{W}}_2 \hat{\mathbf{W}}_2^\dagger = \mathbf{U}_{\mathbf{G}} P_{\mathbf{U}_{\mathbf{G}}^\top \hat{\mathbf{W}}_2} \mathbf{U}_{\mathbf{G}}^\top$ . From (18), we have  $P_{\hat{\mathbf{W}}_2} \mathbf{G} = \mathbf{G} P_{\hat{\mathbf{W}}_2}$ . By combining the two equations, we obtain  $P_{\mathbf{U}_{\mathbf{G}}^\top \hat{\mathbf{W}}_2} \Lambda = \Lambda P_{\mathbf{U}_{\mathbf{G}}^\top \hat{\mathbf{W}}_2}$ . Combining this with the fact that  $P_{\hat{\mathbf{W}}_2} = U_{\hat{\mathbf{W}}_2} \begin{bmatrix} \mathbb{I}_{r_{\hat{\mathbf{W}}_2}} & 0 \\ 0 & 0 \end{bmatrix} U_{\hat{\mathbf{W}}_2}^\top$ , we obtain that  $P_{\mathbf{U}_{\mathbf{G}}^\top \hat{\mathbf{W}}_2} = \begin{bmatrix} \mathcal{D} & 0 \\ 0 & \mathbf{B} \end{bmatrix}$ , where  $\mathcal{D} \in \mathbb{R}^{r_{\mathbf{G}} \times r_{\mathbf{G}}}$ ,  $\mathbf{B} \in \mathbb{R}^{(\mathbf{D} - r_{\mathbf{G}}) \times (\mathbf{D} - r_{\mathbf{G}})}$ , and  $r_{\mathcal{D}} + r_{\mathbf{B}} = r_{\mathbf{W}_2} \leq k$ . Note that  $P_{\mathbf{U}_{\mathbf{G}}^\top \hat{\mathbf{W}}_2}$  has eigenvalues consist of  $r_{\hat{\mathbf{W}}_2}$  number of 1s and the rest 0. Using this fact, a straightforward calculation shows that  $\mathcal{D}$  is a diagonal matrix with  $r_{\mathcal{D}}$  number of 1s, and the rest 0s. For  $\mathbf{B} = \mathbf{B}_1 \mathbf{B}_1^\top$  where  $\mathbf{B}_1 = ((\mathbf{U}_{\mathbf{G}}^\top \mathbf{U}_{\hat{W}_2})_{i,j})_{r_{\mathbf{G}} + 1 \leq i \leq d, r_{\mathbf{G}} + 1 \leq j \leq d}$ , its eigenvalues consist of  $r_{\mathbf{B}}$  number of 1s, and the rest 0s.

Full Statement of Theorem 2.1 From Lemma D.2, we established that  $\mathbf{B}$  is a symmetric positive semi-definite matrix with its eigenvalues 1 or 0. Thus we can write down its eigendecomposition as  $\mathbf{B} = \tilde{\mathbf{U}}_{\mathbf{B}_1} \tilde{\mathbf{U}}_{\mathbf{B}_1}^{\mathsf{T}}$ . With this notation, we state the following theorem.

**Theorem D.3** (Full Version of Theorem 2.1). Assume that the multiplicity of non-zero eigenvalues of G is 1, i.e.,  $\sigma_1^G > \sigma_2^G > \cdots > \sigma_{r_G}^G > 0$ . We further assume that the bottleneck dimension k is chosen such that  $k \leq r_G^3$ . Let  $\mathcal{I}$  be a family of index sets, where each element is an ordered set of distinct natural numbers from [k]. Then, we establish the following results.

1. Let  $(I, \mathbf{B})$  a tuple, such that  $|I| + r_{\mathbf{B}} = k$ , for  $I \in \mathcal{I}$ . Then, for  $\hat{\mathbf{W}}_2$  and  $\hat{\mathbf{W}}_1$  define critical points for (3) if and only if there exist an invertible matrix  $\mathbf{C} \in \mathbb{R}^{k \times k}$  and a tuple  $(I, \mathbf{B})$  such that  $\hat{\mathbf{W}}_2$  and  $\hat{\mathbf{W}}_1$  satisfy that: for |I| < k,

$$\begin{split} \hat{\mathbf{W}}_2 &= \begin{bmatrix} \mathbf{U}_{\mathbf{G},I} & \mathbf{U}_2 \tilde{\mathbf{U}}_{\mathbf{B}_1} \end{bmatrix} \mathbf{C}, \\ \hat{\mathbf{W}}_1 &= \mathbf{C}^{-1} \begin{bmatrix} \mathbf{U}_{\mathbf{G},I} & \mathbf{U}_2 \tilde{\mathbf{U}}_{\mathbf{B}_1} \end{bmatrix}^\top \mathbf{Y} \mathbf{Z}^\top (\tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^\top)^{-1}, \end{split}$$

and for |I| = k,

$$\hat{\mathbf{W}}_2 = \mathbf{U}_{\mathbf{G},I}\mathbf{C},$$
  
 $\hat{\mathbf{W}}_1 = \mathbf{C}^{-1}\mathbf{U}_{\mathbf{G},I}^{\top}\mathbf{Y}\mathbf{Z}^{\top}(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\top})^{-1}.$ 

For the both cases, assume  $r_{\mathbf{Z}} = n$ . Then, in  $\lambda \to 0$  limit, it satisfies that

$$\mathbf{W}_c := \lim_{\lambda \to 0} \hat{\mathbf{W}}_2 \hat{\mathbf{W}}_1 = P_I(\mathbf{Y}) \mathbf{Z}^{\dagger}.$$

2. For some invertible matrix  $\mathbf{C} \in \mathbb{R}^{r_k \times r_k}$ , global minimizers  $\hat{\mathbf{W}}_2^*, \hat{\mathbf{W}}_1^*$  are given by

$$\begin{split} \hat{\mathbf{W}}_2^* &= \mathbf{U}_{\mathbf{G},k} \mathbf{C} \\ \hat{\mathbf{W}}_1^* &= \mathbf{C}^{-1} \mathbf{U}_{\mathbf{G},k}^{\top} \mathbf{Y} \mathbf{Z}^{\top} (\tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^{\top})^{-1}. \end{split}$$

Furthermore, assume  $r_{\mathbf{Z}} = n$ . Then the global minimizer  $\mathbf{W}_*$  in the ridgeless case is given uniquely by,

$$\mathbf{W}_* = \lim_{\lambda \to 0} \hat{\mathbf{W}}_2^* \hat{\mathbf{W}}_1^* = P_k(\mathbf{Y}) \mathbf{Z}^{\dagger}.$$

3. For  $\lambda > 0$ , we have an unique global minimizer in terms of  $\hat{\mathbf{W}}_* = \hat{\mathbf{W}}_2^* \hat{\mathbf{W}}_1^*$ . On the other hand, at  $\lambda = 0$ , there are multiple global minimizers. But for the both cases, it satisfies that all the critical points other than global minima are saddle points. In other words, all the local minima are global minima and other critical points are saddle points.

*Proof.* We begin by proving the first result, which characterizes general critical points. Note that the " $\Leftarrow$ " direction of the proof follows from a straightforward calculation, and we therefore omit the details. For the " $\Rightarrow$ " part, observe that for any critical point, the conditions (17) and

(18) must be satisfied, and we have that  $P_{\mathbf{W}_2} = \mathbf{U}_{\mathbf{G}} P_{\mathbf{U}_{\mathbf{G}}^{\top} \mathbf{W}_2} \mathbf{U}_{\mathbf{G}}^{\top}$ , where  $P_{\mathbf{U}_{\mathbf{G}}^{\top} \mathbf{W}_2} = \begin{bmatrix} \mathcal{D} & 0 \\ 0 & \mathbf{B} \end{bmatrix}$ . From Lemma D.2, we have that  $\mathbf{B}$  is a symmetric positive semi-definite matrix with eigenvalues 1 or 0. Thus its eigendecomposition can be represented as  $\mathbf{B} = \tilde{\mathbf{U}}_{\mathbf{B}_1} \tilde{\mathbf{U}}_{\mathbf{B}_1}^{\top}$ . With  $\mathbf{U}_{\mathbf{G}} = [\mathbf{U}_1 \quad \mathbf{U}_2]$ , for  $\mathbf{U}_1 \in \mathbb{R}^{d \times r_{\mathbf{G}}}$ ,  $\mathbf{U}_2 \in \mathbb{R}^{d \times (\mathbf{D} - r_{\mathbf{G}})}$ . With  $I_{\mathcal{D}} := \{i; \mathcal{D}_{ii} = 1\}$ , we denote  $\mathbf{M}_{\hat{\mathbf{W}}_2} := [\mathbf{U}_{\mathbf{G},I_{\mathcal{D}}} \quad \mathbf{U}_2 \tilde{\mathbf{U}}_{\mathbf{B}_1}]$ . Then,  $P_{\hat{\mathbf{W}}_2} = \mathbf{U}_{\mathbf{G}} P_{\mathbf{U}_{\mathbf{G}}^{\top} \hat{\mathbf{W}}_2} \mathbf{U}_{\mathbf{G}}^{\top} = \mathbf{M}_{\hat{\mathbf{W}}_2} \mathbf{M}_{\hat{\mathbf{W}}_2}^{\top}$ . Thus  $\hat{\mathbf{W}}_2$  is spanned by the columns of  $\mathbf{M}_{\hat{\mathbf{W}}_2}$ . Then, there exist an invertible coefficient matrix  $\mathbf{C}$  such that  $\hat{\mathbf{W}}_2 = \mathbf{M}_{\hat{\mathbf{W}}_2} \mathbf{C}$ . From this, we have that  $\hat{\mathbf{W}}_1 = \mathbf{C}^{-1} \mathbf{M}_{\hat{\mathbf{W}}_2}^{\top} \mathbf{Y} \mathbf{Z}^{\top} (\tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^{\top})^{-1}$ , from (17). For the

 $<sup>^3</sup>$ The condition  $k \le r_G$  is imposed to analyze the effect of the bottleneck layer, which is the focus of this work. This assumption can be relaxed to the general case, albeit at the cost of losing the uniqueness of the global minimizer.

<sup>&</sup>lt;sup>4</sup>Uniqueness in terms of there is an unique  $\hat{\mathbf{W}}_*$ .

full rank  $\mathbf{Z}$ (i.e.,  $r_{\mathbf{Z}}=n$ ), it satisfies that  $\hat{\mathbf{W}}_1=\mathbf{C}^{-1}\begin{bmatrix}\mathbf{U}_{I_{\mathcal{D}}}&\mathbf{0}\end{bmatrix}^{\top}\mathbf{Y}\mathbf{Z}^{\top}(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\top})^{-1}$ . To see this, consider  $\mathbf{G}=\mathbf{Y}\mathbf{Z}^{\top}(\mathbf{Z}\mathbf{Z}^{\top}+\lambda\mathbb{I})^{-1}\mathbf{Z}^{\top}\mathbf{Y}$ . Let  $\mathbf{Z}=\mathbf{U}_{\mathbf{Z}}\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{V}_{\mathbf{Z}}^{\top}$  be the SVD of  $\mathbf{Z}$ . Then, we have that  $\mathbf{G}=\mathbf{Y}\mathbf{V}_{\mathbf{Z}}\mathbf{\Sigma}_{\mathbf{Z}}^{\top}(\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{\Sigma}_{\mathbf{Z}}^{\top}+\lambda\mathbb{I}_{d})^{-1}\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{V}_{\mathbf{Z}}^{\top}\mathbf{Y}^{\top}$ . Let  $\mathbf{T}:=\mathbf{\Sigma}_{\mathbf{Z}}^{\top}(\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{\Sigma}_{\mathbf{Z}}^{\top}+\lambda\mathbb{I})^{-1}\mathbf{\Sigma}_{\mathbf{Z}}$ . Then  $\mathbf{T}$  is a diagonal matrix with  $\mathbf{T}_{ij}=(\sigma_{i}^{Z})^{2}((\sigma_{i}^{Z})^{2}+\lambda)^{-1}\mathbf{1}_{i=j}$  as its diagonal elements. Thus, for  $\mathbf{G}=\mathbf{Y}\mathbf{V}_{\mathbf{Z}}\mathbf{T}\mathbf{V}_{\mathbf{Z}}^{\top}\mathbf{Y}^{\top}=\mathbf{U}_{\mathbf{G}}\mathbf{\Lambda}_{\mathbf{G}}\mathbf{U}_{\mathbf{G}}^{\mathbf{T}}, \mathbf{U}_{\mathbf{G}}$  is a matrix of the left singular vectors of  $\mathbf{Y}\mathbf{V}_{\mathbf{Z}}\mathbf{V}^{\mathbf{T}}$ . Then,  $\mathbf{U}_{2}\mathbf{Y}=\mathbf{U}_{2}\mathbf{Y}\mathbf{V}_{\mathbf{Z}}\mathbf{V}^{\top}\mathbf{T}^{-1}\mathbf{V}_{\mathbf{Z}}^{\mathbf{T}}$ , and  $\mathbf{U}_{2}\mathbf{Y}\mathbf{V}_{\mathbf{Z}}\mathbf{V}^{\mathbf{T}}=0$ . Therefore, in any case of |I|, it holds that  $\hat{\mathbf{W}}_{c}=\hat{\mathbf{W}}_{2}\hat{\mathbf{W}}_{1}=\mathbf{U}_{\mathbf{G},I}\mathbf{U}_{\mathbf{G},I}^{\top}\mathbf{Y}\mathbf{Z}^{\top}(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\top})^{-1}$ . Finally, in the ridgeless limit  $\lambda\to 0$ , it holds that  $\mathbf{G}=\mathbf{Y}\mathbf{Y}^{\top}$ . Therefore, it satisfies that:

$$\begin{aligned} \mathbf{W}_c &= \lim_{\lambda \to 0} \hat{\mathbf{W}}_c = \lim_{\lambda \to 0} \mathbf{U}_{\mathbf{G},I} \mathbf{U}_{\mathbf{G},I}^\top \mathbf{Y} \mathbf{Z}^\top (\tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^\top)^{-1} \\ &= \mathbf{U}_{\mathbf{Y},I} \mathbf{U}_{\mathbf{Y},I}^\top \mathbf{Y} \mathbf{Z}^\dagger = P_I(\mathbf{Y}) \mathbf{Z}^\dagger. \end{aligned}$$

We now prove the second claim about the global minimizer. To do this, we analyze the loss function (3) further. The loss function with  $\lambda \to 0$  is given by

$$n\mathcal{L}(\mathbf{W}_{2}, \mathbf{W}_{1}) = \|\mathbf{Y} - \mathbf{W}_{2}\mathbf{W}_{1}\mathbf{Z}\|_{F}^{2}$$

$$= \operatorname{Tr}(\mathbf{Y}\mathbf{Y}^{\top}) - 2\operatorname{Tr}(\mathbf{W}_{2}\mathbf{W}_{1}\mathbf{Z}\mathbf{Y}^{\top}) + \operatorname{Tr}(\mathbf{W}_{2}\mathbf{W}_{1}\mathbf{Z}\mathbf{Z}^{\top}\mathbf{W}_{1}^{\top}\mathbf{W}_{2}^{\top})$$

$$= \operatorname{Tr}(\mathbf{Y}\mathbf{Y}^{\top}) - 2\operatorname{Tr}(P_{\mathbf{W}_{2}}\mathbf{G}) + \operatorname{Tr}(P_{\mathbf{W}_{2}}\mathbf{G}P_{\mathbf{W}_{2}})$$

$$= \operatorname{Tr}(\mathbf{Y}\mathbf{Y}^{\top}) - \operatorname{Tr}(P_{\mathbf{U}_{\mathbf{G}}^{\top}\mathbf{W}_{2}}\mathbf{G})$$

$$= \operatorname{Tr}(\mathbf{Y}\mathbf{Y}^{\top}) - \operatorname{Tr}(\mathcal{D}\Lambda_{\mathbf{G}}).$$

This result implies that the loss function depends solely on  $\mathrm{Tr}(\mathcal{D}\Lambda_{\mathbf{G}})$ . Consequently, minimizing the loss is equivalent to maximizing  $\mathrm{Tr}(\mathcal{D}\Lambda)$ . Then, for the optimal  $\mathcal{D}^*$ ,  $r_{\mathcal{D}^*}$  must be k, which is its upper bound. Furthermore,  $\mathcal{D}_{ii}^* = \mathbb{1}_{i \leq k}$ , as the first k components of  $\Lambda_{\mathbf{G}}$  will give the biggest trace. This is our global minimum. Thus, our optimal index set  $I^* = [k]$ . Note that in this case, where all the budget k is spent on  $\mathcal{D}^*$ ,  $\mathbf{B}$  becomes 0, which follows from the fact that all the eigenvalues of  $\mathbf{B}$  should be 0. As a result,  $P_{\mathbf{U}_{\mathbf{G}}^{\top}\mathbf{W}_{2}^{*}}$  becomes a diagonal matrix, with its i-th diagonal element  $(P_{\mathbf{U}_{\mathbf{G}}^{\top}\mathbf{W}_{2}^{*}})_{ii} = \mathbb{1}_{i \leq k}$ . Thus,

$$P_{\mathbf{W}_{2}^{*}} = \mathbf{U}_{\mathbf{G}} P_{\mathbf{U}_{\mathbf{G}}^{\top} \mathbf{W}_{2}} \mathbf{U}_{\mathbf{G}}^{\top} = \mathbf{U}_{\mathbf{G}, k} \mathbf{U}_{\mathbf{G}, k}^{\top}, \tag{23}$$

where  $\mathbf{U}_{\mathbf{G},k}$  consists of the first k columns of  $\mathbf{U}_{\mathbf{G}}$ , which correspond to the top-k eigenvalues of  $\mathbf{G}$ . This leads us to the conclusion that for the full rank  $\mathbf{Z}$ , the global minimum of our ridgeless estimator is indeed

$$\begin{split} \mathbf{W}_* &= \lim_{\lambda \to 0} P_{\mathbf{W}_2} \mathbf{Y} \mathbf{Z}^\top (\tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^\top)^{-1} \\ &= \lim_{\lambda \to 0} \mathbf{U}_{\mathbf{G},k} \mathbf{U}_{\mathbf{G},k}^\top \mathbf{Y} \mathbf{Z}^\top (\tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^\top)^{-1} \xrightarrow{\lambda \to 0} P_k(\mathbf{Y}) \mathbf{Z}^\dagger. \end{split}$$

The last part is to prove that these critical points are saddle points with descent direction. If  $\lambda>0$ , then the global minimum is uniquely defined with  $\hat{\mathbf{W}}^*$ . All the critical points other than the global minimum are saddle points, and it turns out that we can prove this by following the proof of [8]. We give a brief summary of this proof. The main idea of this proof is basically perturbing a column vector of  $\mathbf{U}_{\mathbf{G},I_{\mathcal{D}}}$  a bit into the direction of eigenvector  $\mathbf{u}_j^{\mathbf{G}}$  of  $\mathbf{U}_{\mathbf{G},k}$ , for  $j\notin I_{\mathcal{D}}$  and  $j\in [k]$ . Because this  $\mathbf{u}_j^{\mathbf{G}}$  corresponds to some bigger eigenvalue, we can construct a new perturbed  $\tilde{\mathcal{D}}$  matrix which contains an entry in its diagonal part that would make  $\mathrm{Tr}(\tilde{\mathcal{D}}\mathbf{\Lambda})>\mathrm{Tr}(\mathcal{D}\mathbf{\Lambda})$ . Thus this results in decreased loss function, which basically shows that this is indeed a saddle point with decreasing direction. For example in [8], they have constructed this direction, by picking a column vector  $\mathbf{u}_p^{\mathbf{G}}$  from  $\mathbf{U}_{\mathbf{G},I_{\mathcal{D}}}$ , that  $p\notin [k]$ . Then, they have replaced this  $\mathbf{u}_p^{\mathbf{G}}$  with  $\tilde{\mathbf{u}}_p^{\mathbf{G}}=(1+\epsilon^2)^{-\frac{1}{2}}(\mathbf{u}_p^{\mathbf{G}}+\epsilon\mathbf{u}_j^{\mathbf{G}})$ , for any  $\epsilon>0$ , where  $\mathbf{u}_j^{\mathbf{G}}$  is the eigenvector of  $\mathbf{U}_k$  that corresponds to the j-th eigenvalue. Further details of can be found in [8]. For  $\lambda=0$ , we have multiple global minima, but all the critical points other than global minima are saddle points. This fact has been already proven in [47, Thm. 1] and we refer to this work.

#### D.2 Equivalence to the Reduced-Rank Regression

We begin by examining the solution to a one-layer linear denoising autoencoder (DAE) with a rank constraint and no regularization. We then consider the corresponding minimum-norm solution. This form of linear regression with a rank constraint is commonly known in the literature as *reduced-rank regression* [31, 15].

**Training Objective** The training objective is defined as follows:

$$\mathbf{W}_* := \underset{\mathbf{W} \in \mathbb{R}^{d \times d}, r_{\mathbf{W}} \le k}{\operatorname{argmin}} \frac{1}{n} \|\mathbf{X} - \mathbf{W}(\mathbf{X} + \mathbf{A})\|_F^2.$$
 (24)

Since the loss function involves a non-smooth rank constraint, we derive the solution by following the method outlined in [46]. Note that the result is applicable for the overparameterized setting.

**Theorem D.4** (Solution of Rank-Constrained One-Layer Linear DAE). Consider the one-layer linear denoising autoencoder with the training objective defined in (24). Let  $\mathbf{X} + \mathbf{A} = \bar{\mathbf{U}}\bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^T$  denote the singular value decomposition of  $\mathbf{X} + \mathbf{A}$ . Then, for any  $\mathbf{C} \in \mathbb{R}^{n \times (d-n)}$ , the global minimizer  $\mathbf{W}_*$  is given by

$$\mathbf{W}_* = \begin{bmatrix} P_k(\mathbf{X})\bar{\mathbf{V}}\bar{\mathbf{D}}^{-1} & P_k(\mathbf{X})\bar{\mathbf{V}}\mathbf{C} \end{bmatrix} \bar{\mathbf{U}}^T.$$
 (25)

**Remark D.5** (Minimum-Norm Solution). Note that this holds for any C, and the minimum-norm solution corresponds to the case C = 0. In this case,  $W_* = P_k(X)(X + A)^{\dagger}$ , which matches the minimum-norm solution derived in Theorem D.3.

*Proof.* Let  $\mathbf{Z} := \mathbf{X}\bar{\mathbf{V}}$  and  $\mathbf{Y} := \mathbf{W}\bar{\mathbf{U}}$ . Using the invariance of the Frobenius norm under the unitary transformations, we can write:

$$\|\mathbf{X} - \mathbf{W}(\mathbf{X} + \mathbf{A})\|_F^2 = \|(\mathbf{X} - \mathbf{W}\bar{\mathbf{U}}\bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^T)\bar{\mathbf{V}}\|_F^2$$

$$= \|\mathbf{X}\bar{\mathbf{V}} - \mathbf{W}\bar{\mathbf{U}}\bar{\mathbf{\Sigma}}\|_F^2$$

$$= \|\mathbf{Y}\bar{\mathbf{\Sigma}} - \mathbf{Z}\|_F^2.$$
(26)

With  $\mathbf{Y}\bar{\mathbf{\Sigma}} = \mathbf{Z}$ , we have that  $\mathbf{Y}\bar{\mathbf{\Sigma}} = \begin{bmatrix} \mathbf{Y}_n & \mathbf{Y}_{d-n} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{D}}_n \\ \mathbf{0} \end{bmatrix} = \mathbf{Y}_n\bar{\mathbf{D}}_n = \mathbf{Z}$ . This equivalence tells us that finding  $\mathbf{Y}_n$  is equivalent to finding  $\mathbf{W}$  that satisfies the rank constraint. Thus, the problem of finding W can be reduced to solving the following optimization problem:

$$\underset{\mathbf{Y}_n \in \mathbb{R}^{d \times n}}{\operatorname{argmin}} \|\mathbf{Y}_n \bar{\mathbf{D}}_n - \mathbf{Z}\|_F^2, \quad \text{s.t.} \quad rank(\mathbf{Y}_n \bar{\mathbf{D}}_n) \le k. \tag{27}$$

Now applying the Eckart-Young Theorem[18] gives us that  $\mathbf{Y}_n^* \bar{\mathbf{D}}_n := P_k(\mathbf{Z})$ . Therefore,  $\mathbf{Y}_n^* = P_k(\mathbf{Z}) \mathbf{D}_n^{-1}$ .

For  $\mathbf{Y}_{d-n}^*$ , there are no additional constraints other than the fact that this term cannot introduce additional rank, since all the rank k were spent for  $\mathbf{Y}_n$ . In other words, columns of this matrix should be linear combinations of columns of the  $\mathbf{Y}_n$ . Thus there exists a coefficient matrix  $\mathbf{C} \in \mathbb{R}^{n \times (d-n)}$ , such that  $\mathbf{Y}_{d-n}^* = \mathbf{Y}_n^* \mathbf{C}$ . Substituting these results, we have that

$$\begin{aligned} \mathbf{W}_* &= \mathbf{Y}^* \mathbf{U}^T \\ &= \left[ P_k(\mathbf{X} \bar{\mathbf{V}}) \bar{\mathbf{D}}_n^{-1} & P_k(\mathbf{X} \bar{\mathbf{V}}) \mathbf{C} \right] \bar{\mathbf{U}}^T \\ &= \left[ P_k(\mathbf{X}) \bar{\mathbf{V}} \bar{\mathbf{D}}_n^{-1} & P_k(\mathbf{X}) \bar{\mathbf{V}} \mathbf{C} \right] \bar{\mathbf{U}}^T. \end{aligned}$$

## **E** Proofs and Bias-Variance Definition for Section 3

In this section, we first provide detailed proofs of the theorems stated in Section 3. Using these results, we then define a notion of bias and variance that applies to the two models studied in this paper, as presented in Appendix E.2.

#### E.1 Proofs

The main idea of the proof, originally introduced in [38] and extended in [27], is to first handle the pseudo-inverse term and then apply a concentration argument. To deal with the pseudo-inverse, [27] applies results from [45], which allow expansion of the pseudo-inverse under certain rank conditions.

For the concentration step, assume X and Y are real-valued random variables that are concentrated around their means, with variances scaling as o(1). In this case, we have the bound  $|\mathbb{E}[XY]|$  $\mathbb{E}[X]\mathbb{E}[Y]| \leq \sqrt{Var(X)Var(X)}$ , which implies that  $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] + o(1)$ . A similar approximation holds for products involving more than two random variables, using results of [11] regarding the covariance of four random variables. Importantly, this concentration argument allows us to approximate expectations of products by products of expectations. For example,  $\mathbb{E}\left[\operatorname{Tr}(XY)\right] =$  $\operatorname{Tr}(\mathbb{E}[X]\mathbb{E}[Y]) + o(1)$ , which plays a critical role in the proof. For further details, we refer the reader to [27].

We first prove Theorem 3.2, followed by Theorem 3.5. Before presenting the proofs, we introduce the following notation, adapted from [27]. Let  $\tilde{\mathbf{U}}\mathbf{D}\tilde{\mathbf{V}}^{\top}$  be the reduced SVD of  $\mathbf{X}$ . Then, we denote

- $\mathbf{P} := -(\mathbb{I} \mathbf{A} \mathbf{A}^{\dagger}) \tilde{\mathbf{U}} \mathbf{D} \in \mathbb{R}^{d \times r}$

- $\mathbf{H} := \tilde{\mathbf{V}}^{\top} \mathbf{A}^{\dagger} \in \mathbb{R}^{r \times d}$   $\mathbf{Z} := \mathbb{I} + \tilde{\mathbf{V}}^{\top} \mathbf{A}^{\dagger} \tilde{\mathbf{U}} \mathbf{D} \in \mathbb{R}^{r \times r}$   $\mathbf{K}_1 := \mathbf{H} \mathbf{H}^{\top} + \mathbf{Z} (\mathbf{P}^{\top} \mathbf{P})^{-1} \mathbf{Z}^{\top} \in \mathbb{R}^{r \times r}$ .

#### E.1.1 Proof of Theorem 3.2

Observe the following decomposition of the test metric (Eq. (7)):

$$\frac{1}{N_{\text{tst}}} \mathbb{E}_{\mathbf{A}_{\text{tm}}, \mathbf{A}_{\text{tst}}} \left[ \| \mathbf{X}_{\text{tst}} - \mathbf{W}_{c}(\mathbf{X}_{\text{tst}} + \mathbf{A}_{\text{tst}}) \|_{F}^{2} \right] 
= \frac{1}{N_{\text{tst}}} \mathbb{E}_{\mathbf{A}_{\text{tm}}, \mathbf{A}_{\text{tst}}} \left[ \text{Tr}((\mathbf{X}_{\text{tst}} - \mathbf{W}_{c}(\mathbf{X}_{\text{tst}} + \mathbf{A}_{\text{tst}})) (\mathbf{X}_{\text{tst}} - \mathbf{W}_{c}(\mathbf{X}_{\text{tst}} + \mathbf{A}_{\text{tst}})^{\top})) \right] 
= \frac{1}{N_{\text{tst}}} \mathbb{E}_{\mathbf{A}_{\text{tm}}, \mathbf{A}_{\text{tst}}} \left[ \| \mathbf{X}_{\text{tst}} - \mathbf{W}_{c} \mathbf{X}_{\text{tst}} \|_{F}^{2} + \| \mathbf{W}_{c} \mathbf{A}_{\text{tst}} \|_{F}^{2} \right] 
= \frac{1}{N_{\text{tst}}} \mathbb{E}_{\mathbf{A}_{\text{tm}}} \left[ \| \mathbf{X}_{\text{tst}} - \mathbf{W}_{c} \mathbf{X}_{\text{tst}} \|_{F}^{2} \right] + \frac{\eta_{\text{tst}}^{2}}{d} \mathbb{E}_{\mathbf{A}_{\text{tm}}} \left[ \| \mathbf{W}_{c} \|_{F}^{2} \right].$$
(28)

Based on this decomposition, our goal is to prove Lemmas E.3 (the first term) and E.4(the second term), which together establish Theorem 2.4.

We begin by presenting the technical lemmas required for the proof. Recall that critical points satisfy  $\mathbf{W}_c = P_{I^x}(\mathbf{X})(\mathbf{X} + \mathbf{A})^{\dagger}$ . By a slight abuse of notation, we denote  $\mathbf{W}_c = \mathbf{X}_{I^x}(\mathbf{X} + \mathbf{A})^{\dagger}$ .

**Lemma E.1** (Expanding the Pseudo-Inverse Term for Critical Points). In the case of  $d \ge n + r$  and given the solution  $\mathbf{W}_c = \mathbf{X}_{I^x}(\mathbf{X} + \mathbf{A})^{\dagger}$ , it holds that

$$\mathbf{W}_{c} = \tilde{\mathbf{U}} \mathbf{D}_{I^{x}} (\mathbf{P}^{\top} \mathbf{P})^{-1} \mathbf{Z}^{\top} \mathbf{K}_{1}^{-1} \mathbf{H} - \tilde{\mathbf{U}} \mathbf{D}_{I^{x}} \mathbf{Z}^{-1} \mathbf{H} \mathbf{H}^{\top} \mathbf{K}_{1}^{-1} \mathbf{Z} \mathbf{P}^{\dagger}$$
(29)

As  $r \leq d - n$ , the above **P** matrix has full rank. Thus we can invoke Corollary 2.1 from [45], then we have

$$(\mathbf{A} + \mathbf{\tilde{U}}\mathbf{D}\mathbf{\tilde{V}}^\top)^\dagger = \mathbf{A}^\dagger + \mathbf{A}^\dagger\mathbf{\tilde{U}}\mathbf{D}\mathbf{P}^\dagger - (\mathbf{A}^\dagger\mathbf{H}^\top + \mathbf{A}^\dagger\mathbf{\tilde{U}}\mathbf{D}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{Z}^\top)\mathbf{K}_1^{-1}(\mathbf{H} + \mathbf{Z}\mathbf{P}^\dagger).$$

Then, multiplying  $\mathbf{X}_{I^x} = \tilde{\mathbf{U}} \mathbf{D}_{I^x} \tilde{\mathbf{V}}^{\top}$  to the left side, we get

$$\begin{split} &\mathbf{X}_{I^x}(\mathbf{X}+\mathbf{A})^{\dagger} \\ &= \tilde{\mathbf{U}}\mathbf{D}_{I^x}\tilde{\mathbf{V}}^{\top}(\mathbf{A}^{\dagger}+\mathbf{A}^{\dagger}\mathbf{U}\mathbf{D}\mathbf{P}^{\dagger}-(\mathbf{A}^{\dagger}\mathbf{H}^{\top}+\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top})\mathbf{K}_{1}^{-1}(\mathbf{H}+\mathbf{Z}\mathbf{P}^{\dagger})) \\ &= \tilde{\mathbf{U}}\mathbf{D}_{I^x}\tilde{\mathbf{V}}^{\top}\mathbf{A}^{\dagger}+\tilde{\mathbf{U}}\mathbf{D}_{I^x}\tilde{\mathbf{V}}^{\top}\mathbf{A}^{\dagger}\mathbf{U}\mathbf{D}\mathbf{P}^{\dagger}-\tilde{\mathbf{U}}\mathbf{D}_{I^x}\tilde{\mathbf{V}}^{\top}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}(\mathbf{H}+\mathbf{Z}\mathbf{P}^{\dagger})-\\ \tilde{\mathbf{U}}\mathbf{D}_{I^x}\tilde{\mathbf{V}}^{\top}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}(\mathbf{H}+\mathbf{Z}\mathbf{P}^{\dagger}). \end{split}$$

The desired result can be obtained by successively substituting  $\tilde{\mathbf{V}}^{\top}\mathbf{A} = \mathbf{H}, \mathbf{H}\tilde{\mathbf{U}}\mathbf{D} = \mathbf{Z} - \mathbb{I}$ , and  $\mathbf{H}\mathbf{H}^{\top} + \mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top} = \mathbf{K}_{1}.$ 

**Lemma E.2** Let  $\mathbf{T} \in \mathbb{R}^{r \times r}$  be a diagonal matrix which satisfies

$$\mathbf{T}_{ii} = \begin{cases} 1 & \textit{if } i \in I^x \\ 0 & \textit{otherwise}. \end{cases}$$

Then, 
$$\mathbb{E}[\mathbf{D}_{I^x}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}] = \frac{c}{c-1}\mathbf{T} + O(d^{-1}).$$

*Proof.* Observe that

$$\begin{split} \mathbb{E}[\mathbf{D}_{I^x}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{\Sigma}] &= \mathbb{E}[\mathbf{D}_{I^x}\mathbf{D}^{-1}(\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D})] \\ &= \mathbf{D}_{I^x}\mathbf{D}^{-1}\mathbb{E}[\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}] \\ &= \mathbf{T}\mathbb{E}[\mathbf{D}\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}]. \end{split}$$

By applying Lemma 6 of [27], we have

$$\mathbb{E}[\mathbf{D}_{I^x}(\mathbf{P}^{\top}\mathbf{P})^{-1}\boldsymbol{\Sigma}] = \frac{c}{c-1}\mathbf{T} + O(d^{-1}).$$

Because T is a diagonal matrix with 0 or 1 entries on its diagonal part, the element-wise variance is still  $O(d^{-1})$ .

Now we state lemmas that are directly related to the test metric.

**Lemma E.3** (Variance Term). For  $d \ge n + r$ ,

$$\mathbb{E}[\|\mathbf{W}_c\|_F^2] = \frac{c}{c-1} \sum_{i \in I^x} \frac{\sigma_j^2}{\eta_{tm}^2 + \sigma_j^2} + O\left(\frac{\|\mathbf{D}\|^2}{d}\right) + o(1).$$

*Proof.* Note that  $\|\mathbf{W}_c\|_F^2 = \text{Tr}(\mathbf{W}_c^{\mathsf{T}}\mathbf{W}_c)$ . Use Lemma E.1 to expand this.

$$\mathbb{E}[\operatorname{Tr}(\mathbf{W}_{c}^{\top}\mathbf{W}_{c})] = \mathbb{E}[\operatorname{Tr}(\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{T}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}_{Ix}^{2}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H}) -2\operatorname{Tr}(\mathbf{K}_{1}^{-1}\mathbf{Z}^{\top}\mathbf{D}^{-1}(\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}_{Ix}^{\top}\mathbf{Z}\mathbf{Z}^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{Z}\mathbf{P}^{\dagger}\mathbf{H}^{\top})) +\operatorname{Tr}(\mathbf{P}^{\dagger})^{\top}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H}\mathbf{H}^{\top}(\mathbf{Z}^{-1})^{\top}\mathbf{D}_{Ix}^{\top}\tilde{\mathbf{U}}^{\top}\tilde{\mathbf{U}}\mathbf{D}_{Ix}\mathbf{Z}^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger})].$$
(30)

Using the cyclic invariance property of trace, the first term is equivalent to

$$\mathbb{E}[\mathrm{Tr}(\mathbf{D}_{I^x}(\mathbf{P}^{\top}\mathbf{P})^{-1}Z^{\top}\mathbf{K}_1^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_1^{-1}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}_{I^x})]$$

. This is equivalent to

$$\mathbb{E}[\mathrm{Tr}(\mathbf{D}_{I^x}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}(\mathbf{D}^{-1}\mathbf{Z}^{\top})\mathbf{K}_1^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_1^{-1}(\mathbf{Z}\mathbf{D}^{-1})\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}_{I^x})]$$

We invoke Lemma E.2, and 4, 7, 8 from [27] Each of the terms  $\mathbf{D}_{I^x}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}$ ,  $\mathbf{D}^{-1}\mathbf{Z}^{\top}$ ,  $\mathbf{K}_1^{-1}$ ,  $\mathbf{H}\mathbf{H}^{\top}$  has element-wise variance of  $O(d^{-1})$  which vanishes away. Thus, we have concentration around the product of expectations (more details in [27]). Applying the lemmas to each of these terms gives us that

$$\mathbb{E}[\operatorname{Tr}(\mathbf{D}_{I^x}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}(\mathbf{D}^{-1}\mathbf{Z}^{\top})\mathbf{K}_1^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_1^{-1}(\mathbf{Z}\mathbf{D}^{-1})\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}_{I^x})]$$

$$= \frac{\eta_{\text{tm}}^2 c}{(c-1)}\operatorname{Tr}(\mathbf{T}\mathbf{D}^{-1}(\eta_{\text{tm}}^2\mathbf{D}^{-2} + \mathbb{I}_r)^{-2}\mathbf{D}^{-1}\mathbf{T}^{\top}) + o(1).$$

This is equivalent to

$$\frac{\eta_{\text{trn}}^2 c}{(c-1)} \operatorname{Tr}(\mathbf{D}_{Ix}^{-1} (\eta_{\text{trn}}^2 \mathbf{D}_{Ix}^{-2} + \mathbb{I}_{Ix})^{-2} \mathbf{D}_{Ix}^{-1}) + o(1)$$

$$= \frac{\eta_{\text{trn}}^2 c}{(c-1)} \operatorname{Tr}(\mathbf{D}_{Ix}^2 (\eta_{\text{trn}}^2 \mathbb{I}_{Ix} + \mathbf{D}_{Ix}^2)^{-2}) + o(1).$$

Recall that the o(1) term accounts for the error introduced when replacing the expectation of a product with the product of expectations.

The second term of (30) directly follows from the argument of the fact that  $\mathbf{P}^{\dagger}\mathbf{H}^{\top}=0$ . Thus this term is 0. For the third term of (30), , we proceed as follows:

$$\mathbb{E}[\text{Tr}((\mathbf{P}^{\dagger})^{\top}\mathbf{Z}^{\top}(\mathbf{K}_{1}^{-1})^{\top}\mathbf{H}\mathbf{H}^{\top}(\mathbf{Z}^{-1})^{\top}\mathbf{D}_{I^{x}}^{\top}\mathbf{D}_{I^{x}}\mathbf{Z}^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger})]$$

$$= \mathbb{E}[\text{Tr}((\mathbf{K}_{1}^{-1})^{\top}\mathbf{H}\mathbf{H}^{\top}(\mathbf{D}_{I^{x}}\mathbf{Z}^{-1})^{\top}\mathbf{D}_{I^{x}}\mathbf{Z}^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{D}^{1}(\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D})\mathbf{D}^{-1}\mathbf{Z}^{\top})]$$

Applying Lemma 4, 6, 8 of [27], and Lemma E.2, it follows that

$$\frac{c}{(c-1)} \mathbb{E}\left[\operatorname{Tr}((\eta_{\operatorname{trn}}^2 \mathbf{D}^{-2} + \mathbb{I}_r)^{-1} \mathbf{T}(\mathbf{D} \mathbf{Z}^{-1})^{\top} \mathbf{T}(\mathbf{D} \mathbf{Z}^{-1})(\eta_{\operatorname{trn}}^2 \mathbf{D}^{-2} + \mathbb{I}_r)^{-1} \mathbf{Z} \mathbf{D}^{-1} \mathbf{D}^{-1} \mathbf{Z}^{\top})\right] + o(1)$$

for T defined in Lemma E.2. Applying Lemma 7 of [27] for  $\mathbb{Z}^{-1}$  and  $\mathbb{Z}$ , we finally get

$$\frac{c}{c-1} \operatorname{Tr}(\mathbf{D}_{I^x}^2 (\eta_{\text{trn}}^2 \mathbf{D}^{-2} + \mathbb{I}_r)^{-2} \mathbf{D}^{-2}) + O\left(\frac{\|\mathbf{D}\|^2}{d}\right) + o(1)$$

$$= \frac{c}{c-1} \operatorname{Tr}(\mathbf{D}_{I^x}^4 (\eta_{\text{trn}}^2 \mathbb{I}_{I^x} + \mathbf{D}_{I^x}^2)^{-2})) + O\left(\frac{\|\mathbf{D}\|^2}{d}\right) + o(1).$$

Finally, combining all three terms gives us that

$$\mathbb{E}[\|\mathbf{W}_{c}\|_{F}^{2}] = \frac{c}{c-1} \operatorname{Tr}(\mathbf{D}_{Ix}^{2} (\eta_{tm}^{2} \mathbb{I}_{Ix} + \mathbf{D}_{Ix}^{2})^{-1}) + O\left(\frac{\|\mathbf{D}\|^{2}}{d}\right) + o(1)$$
$$= \frac{c}{c-1} \sum_{j \in Ix} \frac{\sigma_{j}^{2}}{\eta_{tm}^{2} + \sigma_{j}^{2}} + O\left(\frac{\|\mathbf{D}\|^{2}}{d}\right) + o(1).$$

**Lemma E.4** (Bias Term). For  $d \ge n + r$ , and given  $\mathbf{X}_{tst} = \tilde{\mathbf{U}} \mathbf{L}$  for some  $\mathbf{L} \in \mathbb{R}^{r \times N_{tst}}$ , we have

$$\begin{split} &\frac{1}{N_{tst}} \mathbb{E}[\|\mathbf{X}_{tst} - \mathbf{W}_{c} \mathbf{X}_{tst}\|_{F}^{2}] \\ &= \frac{1}{N_{tst}} \operatorname{Tr}(\mathbf{J} \mathbf{L} \mathbf{L}^{\top}) + o(1). \end{split}$$

 $\mathbf{J} \in \mathbb{R}^{d \times d}$  is a diagonal matrix defined as

$$\mathbf{J}_{ii} = egin{cases} \left( \left( rac{\sigma_i}{\eta_{im}} 
ight)^2 + 1 
ight)^{-2} & \textit{if } i \in I^x \ 1 & \textit{otherwise} \end{cases}$$

*Proof.* Using Lemma E.1 to replace  $W_c$ , we obtain that

$$\begin{split} \mathbf{X}_{tst} - \mathbf{W}_c \mathbf{X}_{tst} &= \tilde{\mathbf{U}} \mathbf{L} - (\tilde{\mathbf{U}} \mathbf{D}_{I^x} (\mathbf{P}^\top \mathbf{P})^{-1} \mathbf{Z}^\top \mathbf{K}_1^{-1} \mathbf{H} - \tilde{\mathbf{U}} \mathbf{D}_{I^x} \mathbf{Z}^{-1} \mathbf{H} \mathbf{H}^\top \mathbf{K}_1^{-1} \mathbf{Z} \mathbf{P}^\dagger) \tilde{\mathbf{U}} \mathbf{L}. \end{split}$$
 With  $\mathbf{P}^\dagger \tilde{\mathbf{U}} &= -\mathbf{D}^{-1}$  and  $\mathbf{H} \tilde{\mathbf{U}} = (\mathbf{Z} - \mathbb{I}) \mathbf{D}^{-1}$ , we have that 
$$\begin{split} \tilde{\mathbf{U}} \mathbf{L} - \tilde{\mathbf{U}} \mathbf{D}_{I^x} (\mathbf{P}^\top \mathbf{P})^{-1} \mathbf{Z}^\top \mathbf{K}_1^{-1} (\mathbf{Z} - \mathbb{I}) \mathbf{D}^{-1} \mathbf{L} - \tilde{\mathbf{U}} \mathbf{D}_{I^x} \mathbf{Z}^{-1} \mathbf{H} \mathbf{H}^\top \mathbf{K}_1^{-1} \mathbf{Z} \mathbf{D}^{-1} \mathbf{L} \\ &= \tilde{\mathbf{U}} \mathbf{L} - \tilde{\mathbf{U}} \mathbf{D}_{I^x} \mathbf{Z}^{-1} (\mathbf{Z} (\mathbf{P}^\top \mathbf{P})^{-1} \mathbf{Z}^\top \mathbf{K}_1^{-1} (\mathbf{Z} - \mathbb{I}) + \mathbf{H} \mathbf{H}^\top \mathbf{K}_1^{-1} \mathbf{Z}) \mathbf{D}^{-1} \mathbf{L} \\ &= \tilde{\mathbf{U}} \mathbf{L} - \tilde{\mathbf{U}} \mathbf{D}_{I^x} \mathbf{Z}^{-1} ((\mathbf{K}_1 - \mathbf{H} \mathbf{H}^\top) \mathbf{K}_1^{-1} (\mathbf{Z} - \mathbb{I}) + \mathbf{H} \mathbf{H}^\top \mathbf{K}_1^{-1} \mathbf{Z}) \mathbf{D}^{-1} \mathbf{L} \\ &= \tilde{\mathbf{U}} \mathbf{L} - \tilde{\mathbf{U}} \mathbf{D}_{I^x} \mathbf{Z}^{-1} (\mathbf{Z} - \mathbb{I} + \mathbf{H} \mathbf{H}^\top \mathbf{K}_1^{-1}) \mathbf{D}^{-1} \mathbf{L}. \end{split}$$

The third inequality is due to  $\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top} = \mathbf{K}_1 - \mathbf{H}\mathbf{H}^{\top}$ . Expanding the parentheses, for  $\mathbf{T}^c := \mathbb{I} - \mathbf{T}$  (T from Lemma E.2), we obtain that

$$\tilde{\mathbf{U}}\mathbf{L} - \tilde{\mathbf{U}}\mathbf{D}_{Ix}\mathbf{D}^{-1}\mathbf{L} + \tilde{\mathbf{U}}\mathbf{D}_{Ix}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{D}^{-1}\mathbf{L} \\
= \tilde{\mathbf{U}}\mathbf{T}^{c}\mathbf{L} + \tilde{\mathbf{U}}\mathbf{D}_{Ix}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{D}^{-1}\mathbf{L}.$$

We consider the term  $\|\mathbf{X}_{tst} - \mathbf{W}_c \mathbf{X}_{tst}\|_F^2$ . Using the equivalence between the Frobenius norm and the trace, this can be written as  $\text{Tr}((\mathbf{X}_{tst} - \mathbf{W}_c \mathbf{X}_{tst})^{\top}(\mathbf{X}_{tst} - \mathbf{W}_c \mathbf{X}_{tst}))$ . Expanding the expression, and applying the cyclic invariance of the trace along with the identity  $\mathbf{T}^c \mathbf{D}_{I^x} = 0$ , we obtain that,

$$\begin{split} & \mathbb{E}[\mathrm{Tr}((\mathbf{X}_{tst} - \mathbf{W}_c \mathbf{X}_{tst})^{\top} (\mathbf{X}_{tst} - \mathbf{W}_c \mathbf{X}_{tst}))] = \\ & = \mathrm{Tr}(\mathbf{T}^c \mathbf{L} \mathbf{L}^{\top}) + \mathbb{E}[\mathrm{Tr}(\mathbf{D}^{-1} \mathbf{K}_1^{-1} \mathbf{Z} (\mathbf{P}^{\top} \mathbf{P})^{-\top} \mathbf{D}_{I^x}^2 (\mathbf{P}^{\top} \mathbf{P})^{-1} \mathbf{Z}^{\top} \mathbf{K}_1^{-1} \mathbf{D}^{-1} \mathbf{L} \mathbf{L}^{\top})]. \end{split}$$

Using Lemma E.2 and Lemma 7, 8 of [27], we get that the second term is equivalent to

$$\begin{split} & \eta_{\text{trn}}^4 \operatorname{Tr}(\mathbf{D}^{-4}(\eta_{\text{trn}}^2 \mathbf{D}^{-2} + \mathbb{I}_r)^{-2} \mathbf{T} \mathbf{L} \mathbf{L}^\top) + o(1) \\ &= \eta_{\text{trn}}^4 \operatorname{Tr}((\mathbf{D}^2 + \eta_{\text{trn}}^2 \mathbb{I}_r)^{-2} T \mathbf{L} \mathbf{L}^\top) + O(d^{-1}) + o(1). \end{split}$$

The o(1) term refers to the error incurred when approximating the expectation of products by the product of expectations. To summarize the result, we have that

$$\begin{split} &\frac{1}{N_{\text{tst}}} \mathbb{E}[\|\mathbf{X}_{\text{tst}} - \mathbf{W}_c \mathbf{X}_{\text{tst}}\|_F^2] \\ &= \frac{1}{N_{\text{tst}}} \left( \text{Tr}(\mathbf{T}^c \mathbf{L} \mathbf{L}^\top) + \text{Tr}((\eta_{\text{trn}}^{-2} \mathbf{D}^2 + \mathbb{I}_r)^{-2} \mathbf{T} \mathbf{L} \mathbf{L}^\top) + O(d^{-1}) + o(1) \right) \\ &= \frac{1}{N_{\text{tst}}} \operatorname{Tr}(\mathbf{J} \mathbf{L} \mathbf{L}^\top) + O(d^{-1}) + o(1). \end{split}$$

Combining Lemmas E.3 and E.4 within the decomposition in Equation (28) yields the desired result.

#### E.1.2 Proof of Theorem 3.5

We first decompose the test risk (8):

$$\mathbb{E}\left[\frac{1}{N_{\text{tst}}} \| \mathbf{X}_{\text{tst}} - (\mathbf{W}_{c}^{\text{sc}} + \mathbb{I}_{d})(\mathbf{X}_{\text{tst}} + \mathbf{A}_{\text{tst}}) \|_{F}^{2}\right] \\
= \frac{1}{N_{\text{tst}}} \mathbb{E}\left[\text{Tr}(\mathbf{A}_{\text{tst}} \mathbf{A}_{\text{tst}}^{\top} + 2\mathbf{A}_{\text{tst}}(\mathbf{X}_{\text{tst}} + \mathbf{A}_{\text{tst}})^{\top}(\mathbf{W}_{c}^{\text{sc}})^{\top} + \mathbf{W}_{c}^{\text{sc}}(\mathbf{X}_{\text{tst}} + \mathbf{A}_{\text{tst}})(\mathbf{X}_{\text{tst}} + \mathbf{A}_{\text{tst}})^{\top}(\mathbf{W}_{c}^{\text{sc}})^{\top})\right] \\
= \frac{1}{N_{\text{tst}}} \mathbb{E}_{\mathbf{A}_{\text{trn}}} \left[ \frac{N_{\text{tst}} \eta_{\text{tst}}^{2}}{d} \left( \text{Tr}(\mathbb{I}_{d}) + \text{Tr}(\mathbf{W}_{c}^{\text{sc}}) + \|\mathbf{W}_{c}^{\text{sc}}\|_{F}^{2} \right) + \|\mathbf{W}_{c}^{\text{sc}} \mathbf{X}_{\text{tst}}\|_{F}^{2} \right] \\
= \eta_{\text{tst}}^{2} + \mathbb{E}_{\mathbf{A}_{\text{trn}}} \left[ \text{Tr}\left( \frac{2\eta_{\text{tst}}^{2}}{d} \mathbf{W}_{c}^{\text{sc}} \right) + \frac{\eta_{\text{tst}}^{2}}{d} \|\mathbf{W}_{c}^{\text{sc}}\|_{F}^{2} + \frac{1}{N_{\text{tst}}} \|\mathbf{W}_{c}^{\text{sc}} \mathbf{X}_{\text{tst}}\|_{F}^{2} \right]. \tag{31}$$

We aim to prove Lemmas E.12 (the first term), E.11(the second term), and Lemma E.13, in order to establish Theorem 2.4.

We begin with couple of technical lemmas that are necessary for the proof of this theorem. Recall that critical points satisfy  $\mathbf{W}_c^{\text{sc}} = -P_{I^a}(\mathbf{A})(\mathbf{X} + \mathbf{A})^{\dagger}$ . By a slight abuse of notation, we write  $\mathbf{W}_c^{\text{sc}} = -\mathbf{A}_{I^a}(\mathbf{X} + \mathbf{A})^{\dagger}$ .

**Lemma E.5** ([27, Lemma 3]). Consider  $\mathbf{a} \in \mathbb{R}^d$  and  $\mathbf{b} \in \mathbb{R}^d$ , and uniform random orthogonal matrix  $\mathbf{Q} \in \mathbb{R}^{d \times d}$ . If  $\langle \mathbf{a}, \mathbf{b} \rangle = 0$ , then  $\mathbb{E}[(\mathbf{Q}\mathbf{a})_i(\mathbf{Q}\mathbf{b})_i] = 0$ .

*Proof.* Note that 
$$\langle \mathbf{Qa}, \mathbf{Qb} \rangle = 0$$
. Then,  $\sum_{i=1}^{d} \mathbb{E}[(\mathbf{Qa})_i(\mathbf{Qb})_i] = 0$ . Due to symmetry of  $Q$ ,  $\mathbb{E}[(\mathbf{Qa})_i(\mathbf{Qb})_i] = \mathbb{E}[(\mathbf{Qa})_j(\mathbf{Qb})_j] = 0$ , for any  $1 \leq i, j \leq d$ .

**Lemma E.6** Consider an unit vector  $\mathbf{a} \in \mathbb{R}^d$ , and an uniform random orthogonal matrix  $\mathbf{Q} \in \mathbb{R}^{d \times d}$ . Then,  $\mathbb{E}\left[(\mathbf{Q}\mathbf{a})_i(\mathbf{Q}\mathbf{a})_i\right] = \frac{1}{d}$ .

*Proof.* Note that  $\mathbb{E}\left[\mathbf{a}^{\top}\mathbf{Q}\mathbf{Q}^{\top}\mathbf{a}\right] = 1$ . Then,  $\sum_{i=1}^{d} \mathbb{E}\left[\mathbf{a}^{\top}\mathbf{q}_{i}\mathbf{q}_{i}^{\top}\mathbf{a}\right] = 1$ . Then the result follows from the symmetry of  $\mathbf{Q}$ .

**Lemma E.7** For  $c = \frac{d}{n}$ ,  $\mathbb{E}[\mathbf{H}\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}] = \frac{|I^a|}{n}\frac{1}{c-1}\mathbb{I} + o(\frac{|I^a|}{n})$  with element wise variance of  $O(\frac{k}{N^2})$ .

 $\textit{Proof.} \quad \text{For } \mathbf{A} = \mathbf{U}_{\mathbf{A}} \mathbf{\Sigma}_{\mathbf{A}} \mathbf{V}_{\mathbf{A}}^{\top} \text{, we define a diagonal matrix } \mathbf{T}_a \in \mathbb{R}^{d \times d} \text{, such that} \\ (\mathbf{T}_a)_{ii} = \begin{cases} (\sigma_i^{\mathbf{A}})^{-2} & \text{if } i \in I^a \\ 0 & \text{otherwise.} \end{cases}$ 

$$(\mathbf{T}_a)_{ii} = \begin{cases} (\sigma_i^{\mathbf{A}})^{-2} & \text{if } i \in I^a \\ 0 & \text{otherwise.} \end{cases}$$

Then, we can write  $\mathbb{E}[\mathbf{H}\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}] = \mathbb{E}[\tilde{\mathbf{V}}^{\top}\mathbf{A}^{\dagger}\mathbf{A}_{I^a}\mathbf{A}^{\dagger}(\mathbf{A}^{\dagger})^{\top}\tilde{\mathbf{V}}] = \mathbb{E}[\tilde{\mathbf{V}}^{\top}\mathbf{V}_{\mathbf{A}}\mathbf{T}_a\mathbf{V}_{\mathbf{A}}^{\top}\tilde{\mathbf{V}}]$ . Observe  $(\tilde{\mathbf{V}}^{\top}\mathbf{V}_{\mathbf{A}}\mathbf{T}_{a}\mathbf{V}_{\mathbf{A}}^{\top}\tilde{\mathbf{V}})_{ij} = \mathbf{v}_{i}^{\top}\mathbf{V}_{\mathbf{A}}\mathbf{T}_{a}\mathbf{V}_{\mathbf{A}}^{\top}\mathbf{v}_{j}$ , which is equivalent to  $\mathbf{a}_{i}^{\top}\mathbf{T}_{a}\mathbf{a}_{j}$ , for  $\mathbf{a}_{i} := \mathbf{v}_{i}^{\top}\mathbf{V}_{\mathbf{A}}$ . For  $i \neq j$ , this expression evaluates to 0 due to Lemma E.5. On the other hand, if i = j, then  $\mathbb{E}\left[\mathbf{a}_i^{\top}\mathbf{T}_a\mathbf{a}_i\right] = \sum_{l=1}^{l} \mathbb{E}\left[(\mathbf{a}_i)_l^2\right] \mathbb{E}\left[(\sigma_l^{\mathbf{A}})^{-2}\right] = \frac{k}{n}\left(\frac{1}{c-1} + o(1)\right).$  This follows from the fact that  $\mathbf{A}$  is a Gaussian matrix, for which the matrix of singular values is independent of the singular vectors. The final equality results from a direct evaluation of the Stieltjes transform of the Marchenko-Pastur

For the variance part, we consider  $\mathbb{E}\left[\sum_{m=1}^{|I^a|}\sum_{l=1}^{|I^a|}(\mathbf{a}_i)_m(\mathbf{a}_j)_m(\mathbf{a}_j)_l(\sigma_l^{\mathbf{A}})^{-2}(\sigma_m^{\mathbf{A}})^{-2}\right]$ , which matches the variance computation in Lemma 4 of [27], with N replaced by  $|I^a|$ . As a result, the element-wise variance is of order  $O(\frac{|I^a|}{n^2})$ .

**Lemma E.8**  $\mathbf{A}_{I^a}\mathbf{A}^+(\mathbf{P}^\dagger)^\top=0$ ,  $\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^\dagger\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^\dagger)=0$ , and  $\mathbf{P}^\dagger\mathbf{H}^\top=0$ .

For the first term, note that  $\mathbf{A}_{I^a}\mathbf{A}^{\dagger}(\mathbf{P}^{\dagger})^{\top} = \mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{P}(\mathbf{P}^{\top}\mathbf{P})^{-\top}$ . Also, it applies that  $\mathbf{P} = -(\mathbb{I}_d - \mathbf{A}\mathbf{A}^{\dagger})\tilde{\mathbf{U}}\mathbf{D}$ . Then,

$$\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{P} = -\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D} + \mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{A}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}$$
$$= -\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D} + \mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D} = 0.$$

We can prove  $\text{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger})=0$  similarly.  $\mathbf{P}^{\dagger}\mathbf{H}^{\top}=0$  was already proved in Lemma 9 of [27].

**Lemma E.9** For c > 1,  $\mathbb{E}[\tilde{\mathbf{V}}^{\top} \mathbf{A}^{\dagger} \mathbf{A}_{I^a} \mathbf{A}^{\dagger} \tilde{\mathbf{U}}] = \mathbb{E}[\mathbf{H} \mathbf{A}_{I^a} \mathbf{A}^{\dagger} \tilde{\mathbf{U}}] = 0$ , with element-wise variance  $O(\frac{|I^a|}{dn})$ .

*Proof.* For notational convenience, we define

distribution ([38, Lemma 5]).

$$\mathbf{T}_a = \begin{cases} (\sigma_i^{\mathbf{A}})^{-1} & \text{if } i \in I^a \\ 0 & \text{otherwise} \end{cases}$$

Then,  $\mathbb{E}[(\tilde{\mathbf{V}}^{\top}\mathbf{A}^{\dagger}\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}})_{ij}] = \mathbb{E}[(\tilde{\mathbf{V}}^{\top}\mathbf{A}_{I^a}^{\dagger}\tilde{\mathbf{U}})_{ij}] = \mathbb{E}[\mathbf{v}_i^{\top}\mathbf{V}_{\mathbf{A}}\mathbf{T}_a\mathbf{U}_{\mathbf{A}}^{\top}\mathbf{u}_j].$  This is equal to  $\mathbb{E}\left[\mathbf{a}^{\top}\mathbf{T}_{a}\mathbf{b}\right] = \mathbb{E}\left[\sum_{l=1}^{|I^{a}|}(\sigma_{l}^{\mathbf{A}})^{-1}\mathbf{a}_{l}\mathbf{b}_{l}\right] = 0, \text{ for } \mathbf{a} := \mathbf{v}_{i}^{\top}\mathbf{V}_{\mathbf{A}}, \ \mathbf{b} := \mathbf{u}_{j}^{\top}\mathbf{U}_{\mathbf{A}}. \text{ For the variance } \mathbf{v}_{i}^{\top}\mathbf{v}_{\mathbf{A}}$ part, if  $i \neq j$ ,

$$\mathbb{E}\left[(\tilde{\mathbf{V}}^{\top}\mathbf{A}^{\dagger}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}})_{ij}^{2}\right] = \mathbb{E}\left[\sum_{l=1}^{|I^{a}|}(\sigma_{l}^{\mathbf{A}})^{-2}\mathbf{a}_{l}^{2}\mathbf{b}_{l}^{2}\right]$$
$$= \frac{|I^{a}|}{dn}\left(\frac{1}{c-1} + o(1)\right) = O\left(\frac{|I^{a}|}{dn}\right).$$

If i = j,

$$\begin{split} \mathbb{E}\left[(\tilde{\mathbf{V}}^{\top}\mathbf{A}^{\dagger}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}})_{ij}^{2}\right] &= \mathbb{E}\left[\sum_{l=1}^{|I^{a}|}(\sigma_{l}^{\mathbf{A}})^{-2}\mathbf{a}_{l}^{4}\right] \\ &= \frac{3|I^{a}|}{d(d+2)}\left(\frac{1}{c-1} + o(1)\right) = O\left(\frac{|I^{a}|}{d^{2}}\right). \end{split}$$

**Lemma E.10**  $\mathbb{E}[\tilde{\mathbf{U}}^{\top}\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}] = \frac{|I^a|}{d}\mathbb{I}_r$ , with its element-wise variance  $O(\frac{|I^a|^2}{d^2})$ .

For notational convenience, we write

$$\mathbf{T}_a = \begin{cases} 1 & \text{if } i \in I^a \\ 0 & \text{otherwise.} \end{cases}$$

Then,  $\mathbb{E}[(\tilde{\mathbf{U}}^{\top}\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}})_{ij}] = \mathbb{E}[\mathbf{u}_i^{\top}\mathbf{U}_{\mathbf{A}}\mathbf{T}_a\mathbf{U}_{\mathbf{A}}^{\top}\mathbf{u}_j]$ . Thus with Lemma E.5, this is 0 if  $i \neq j$ . On the other hand, if i=j, then this is  $\frac{|I^a|}{d}$ . For the variance part, first assume that  $i\neq j$ . Next, observe that for uniformly distributed random vectors  $\mathbf{a} := \mathbf{U}_{\mathbf{A}}^{\top} \mathbf{u}_i, \mathbf{b} := \mathbf{U}_{\mathbf{A}}^{\top} \mathbf{u}_j$ , we have the following:

$$\mathbb{E}\left[(\tilde{\mathbf{U}}^{\top}\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}})_{ij}^2\right] = \mathbb{E}\left[(\mathbf{a}^{\top}\mathbf{T}_a\mathbf{b})^2\right] = \mathbb{E}\left[\sum_{l=1}^{|I^a|}\sum_{m=1}^{|I^a|}\mathbf{a}_l\mathbf{b}_l\mathbf{a}_m\mathbf{b}_m\right].$$

Note that  $\langle \mathbf{a}, \mathbf{b} \rangle = 0$ . This is because

$$0 \leq \sum_{l=1}^{|I^a|} (a_l b_l) \sum_{m=1}^{|I^a|} (\mathbf{a}_m \mathbf{b}_m) \leq \sum_{l=1}^d (\mathbf{a}_l \mathbf{b}_l) \sum_{m=1}^d (\mathbf{a}_m \mathbf{b}_m)$$
$$= \langle \mathbf{a}, \mathbf{b} \rangle^2$$
$$= 0.$$

The first inequality satisties due to  $\sum\limits_{l=1}^{|I^a|}(\mathbf{a}_l\mathbf{b}_l) = \sum\limits_{h=1}^{|I^a|}\mathbf{u}_i^\top(\mathbf{u}_h^\mathbf{A})(\mathbf{u}_h^\mathbf{A})^\top\mathbf{u}_j$ , and each  $(\mathbf{u}_h^\mathbf{A})(\mathbf{u}_h^\mathbf{A})^\top$  is a positive semi-definite matrix, which means  $\mathbf{u}_i^{\top} \mathbf{u}_h^{\mathbf{A}} (\mathbf{u}_h^{\mathbf{A}})^{\top} \mathbf{u}_j \geq 0$ . Thus, adding extra terms will only increase this. With this, the above term is 0. Now for i=j, this is the case where  $\mathbf{a}=\mathbf{b}$ . Then, we have that

$$\mathbb{E}\left[\sum_{l=1}^{|I^{a}|} \sum_{m=1}^{|I^{a}|} a_{l}^{2} \mathbf{a}_{m}^{2}\right] = \mathbb{E}\left[\sum_{l=1}^{|I^{a}|} \mathbf{a}_{l}^{4}\right] + \mathbb{E}\left[\sum_{l \neq |I^{a}|} \mathbf{a}_{l}^{2} \mathbf{a}_{m}^{2}\right]$$

$$= |I^{a}| \times \frac{3}{d(d+2)} + |I^{a}| \left(|I^{a}| - 1\right) \times \frac{1}{d(d+2)}$$

$$= O\left(\frac{|I^{a}|^{2}}{d^{2}}\right).$$

From the decomposition of the test metric, we have the following lemmas that are directly relevant to Theorem 3.5.

**Lemma E.11**  $(\mathbb{E}[\|\mathbf{W}_c^{\text{sc}}\|_F^2] \text{ Term}).$  For  $c:=\frac{d}{n}$  and  $d\geq n+r$ , we obtain that

$$\mathbb{E}[\|\mathbf{W}_{c}^{\alpha}\|_{F}^{2}] = |I^{a}| + \frac{|I^{a}|}{d} \frac{c}{c-1} \sum_{i=1}^{d} \frac{\sigma_{i}^{2}}{(\eta_{rm}^{2} + \sigma_{i}^{2})} + \frac{|I^{a}|}{n} \frac{1}{c} \sum_{i=1}^{d} \frac{\eta_{rm}^{2} \sigma_{i}^{2}}{(\eta_{rm}^{2} + \sigma_{i}^{2})} + o(1).$$

We evaluate  $\mathbf{E}[\|\mathbf{W}_c^{\text{sc}}\|_F^2] = \mathbf{E}[\text{Tr}(\mathbf{W}_c^{\text{sc}}(\mathbf{W}_c^{\text{sc}})^{\top})]$ . Following E.1.1, we approximate the term using the concentration argument. By using Corollary 2.1 of [45],

$$\begin{split} \mathbf{W}_{c}^{\mathrm{sc}} &= -\mathbf{A}_{I^{a}}(\mathbf{A} + \tilde{\mathbf{U}}\mathbf{D}\tilde{\mathbf{V}}^{\top})^{\dagger} \\ &= -\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger} - \mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger} + \mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H} + \mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger} + \\ \mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H} + \mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger}. \end{split}$$

We now examine how each of the six terms behaves after expanding them by multiplying with  $(\mathbf{W}_c^{\mathrm{sc}})^{\top}$ .

1. Distributing  $-\mathbf{A}_{I^a}\mathbf{A}^{\dagger}$ : For a notational convenience, we define

$$\mathbf{T}_a = \begin{cases} 1 & \text{if } i \in I^a \\ 0 & \text{otherwise.} \end{cases}$$

Due to the fact that  $\mathbf{A}_{I^a}\mathbf{A}^{\dagger}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger})^{\top} = \mathbf{U}_{\mathbf{A}}\mathbf{T}_a\mathbf{U}_{\mathbf{A}}^{\top} = \mathbf{A}_{I^a}\mathbf{A}^{\dagger}$ , we have that

$$-\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}(\mathbf{W}_c^{\operatorname{sc}})^{\top}) = \operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger})^{\top}) - \operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}(\mathbf{K}_1)^{-1}\mathbf{H}) - \operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_1^{\top}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-\top}\mathbf{D}\tilde{\mathbf{U}}^{\top}).$$

Note that, due to Lemma E.8, certain terms evaluate to zero and therefore do not appear in the expression. The first term  $\mathbb{E}\left[\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger})^{\top})\right] = \mathbb{E}\left[\operatorname{Tr}(\mathbf{U}_{\mathbf{A}}\mathbf{T}_a\mathbf{U}_{\mathbf{A}}^{\top})\right] = |I^a|$ . The second term, with Lemma E.7, and Lemma 8 of [27],

$$\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}(\mathbf{K}_1)^{-1}\mathbf{H}) = \operatorname{Tr}(\mathbf{H}\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}(\mathbf{K}_1)^{-1})$$
$$= \frac{|I^a|}{n} \frac{\eta_{\text{trn}}^2}{c} \operatorname{Tr}((\eta_{\text{trn}}^2\mathbf{D}^{-2} + \mathbb{I}_r)^{-1}) + o\left(\frac{|I^a|}{n}\right).$$

Because two terms have element-wise variance of  $O(\frac{|I^a|}{n^2})$ , and  $O(d^{-1})$  respectively, the element-wise estimation error of the whole term would be o(1) ( $O\left(\frac{\sqrt{|I^a|}}{d\sqrt{d}}\right)$  to be more precise).

The third term  $\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_1^{-\top}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-\top}\mathbf{D}\tilde{\mathbf{U}}^{\top})$  has mean of 0 due to Lemma E.9. Thus we only need to take a look at the element-wise variance. Due to the cyclic invariance of trace.

$$Tr(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-\top}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-\top}\mathbf{D}\tilde{\mathbf{U}}^{\top})$$

$$= Tr(\tilde{\mathbf{U}}^{\top}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-\top}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-\top}\mathbf{D})$$

$$= Tr(\tilde{\mathbf{U}}^{\top}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-\top}\mathbf{Z}\mathbf{D}^{-1}(\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-\top}\mathbf{D}))$$

. From [27], the element-wise variances of  $\mathbf{K}^{-1}, \mathbf{Z}\mathbf{D}^{-1}, \mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}$  are all  $O(d^{-1})$ . Thus applying the concentration argument gives us the total element wise variance of  $O(d^{-1})$ .

To summarize the result of distributing  $-\mathbf{A}_{I^a}\mathbf{A}^{\dagger}$ , this can be written down as  $|I^a|+\frac{|I^a|}{r}\frac{\eta_{\text{tra}}^2}{r}\operatorname{Tr}((\eta_{\text{tra}}^2\mathbf{D}^{-2}+\mathbb{I}_r)^{-1})+O(d^{-1})$ .

2. Distributing  $-\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger}$ : This is

$$-\operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger}(\mathbf{W}_{c}^{sc})^{\top}) =$$

$$\operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger}(\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger})^{\top}) - \operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-\top}\mathbf{H}) -$$

$$\operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-\top}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-\top}\mathbf{D}\tilde{\mathbf{U}}^{\top})$$

Note that some terms do not appear here due to Lemma E.8. The second term has also mean 0 due to Lemma E.9, thus only the variance needs to be bounded for this term. Due to the cyclic invariance of trace,

$$-\operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-\top}\mathbf{H})$$

$$=-\operatorname{Tr}(\mathbf{H}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-\top})$$

$$=-\operatorname{Tr}(\mathbf{H}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}(\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D})(\mathbf{D}^{-1}\mathbf{Z}^{\top})\mathbf{K}_{1}^{-\top}).$$

Thus from Lemmas 6, 7, 8 from [27],  $\mathbf{DP}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{D}$ ,  $\mathbf{D}^{-1}\mathbf{Z}^{\top}$ , and  $\mathbf{K}_{1}^{-1}$  have variance of  $O(d^{-1})$ . From the concentration argument, this has an error rate of  $O(d^{-1})$ .

The first term is  $\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger}(\mathbf{U}\mathbf{D}\mathbf{P}^{\dagger})^{\top}) = \operatorname{Tr}(\tilde{\mathbf{U}}^{\top}\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{D}) = \operatorname{Tr}(\tilde{\mathbf{U}}^{\top}\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D})$ . From Lemma E.10, and Lemma 6 from [27], we have that

$$\operatorname{Tr}(\tilde{\mathbf{U}}^{\top} \mathbf{A}_{I^{a}} \mathbf{A}^{\dagger} \tilde{\mathbf{U}} \mathbf{D} (\mathbf{P}^{\top} \mathbf{P})^{-1} \mathbf{D}) = \frac{|I^{a}|}{d} \frac{c}{c-1} \operatorname{Tr}(\mathbb{I}_{r}) + O\left(\frac{|I^{a}|}{dn}\right)$$
(32)

with element-wise variance  $O\left(\frac{|I^a|}{d\sqrt{d}}\right)$ . The final term is

$$Tr(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-\top}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-\top}\mathbf{D}\tilde{\mathbf{U}}^{\top})$$

$$= Tr((\tilde{\mathbf{U}}^{\top}\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}})(\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D})\mathbf{D}^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-\top}(\mathbf{Z}\mathbf{D}^{-1})(\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-\top}\mathbf{D})).$$

. Using Lemma E.10, and Lemmas 6, 7, 8 from [27], we have that

$$-\operatorname{Tr}((\tilde{\mathbf{U}}^{\top}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}})(\mathbf{D}\mathbf{P}^{\top}\mathbf{P}\mathbf{D})^{-1}\mathbf{D}^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-\top}(\mathbf{Z}\mathbf{D}^{-1})(\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-\top}\mathbf{D}))$$

$$= -\frac{|I^{a}|}{d}\frac{\eta_{\text{trn}}^{2}c}{c-1}\operatorname{Tr}(\mathbf{D}^{-2}(\eta_{\text{trn}}^{2}\mathbf{D}^{-2} + \mathbb{I}_{r})^{-1}) + O\left(\frac{|I^{a}|}{d^{2}}\right)$$

with its element-wise variance  $O(d^{-1})$ . Thus, in total, this can be written as

$$\frac{|I^a|}{d}\frac{c}{c-1}\operatorname{Tr}(\mathbf{D}^2(\eta_{\operatorname{trn}}^2\mathbb{I}_r+\mathbf{D}^2)^{-1})+O(\frac{|I^a|}{d^2})$$

3. Distributing  $\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_1^{-1}\mathbf{H}$ : This is

$$Tr(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_1^{-1}\mathbf{H}(\mathbf{W}_c^{sc})^{\top}) = -Tr(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_1^{-1}\mathbf{H}) + Tr(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_1^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_1^{-1}\mathbf{H}) + Tr(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_1^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_1^{-1}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}\tilde{\mathbf{U}}^{\top})$$

Again, using Lemma E.8, some terms are filtered out. The first term is, from Lemma E.7, and Lemma 8 of [27],

$$-\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_1^{-1}\mathbf{H}) = -\frac{|I^a|}{n}\frac{\eta_{\text{trn}}^2}{c}\operatorname{Tr}((\eta_{\text{trn}}^2\mathbf{D}^{-2} + \mathbb{I}_r)^{-1}) + o\left(\frac{|I^a|}{n}\right).$$

, with element-wise variance of  $O\left(\frac{\sqrt{|I^a|}}{N\sqrt{d}}\right)$ . Due to Lemma E.7, and Lemma 4, 8 of [27], the second term is

$$\operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H}) = \operatorname{Tr}((\mathbf{H}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top})\mathbf{K}_{1}^{-1}(\mathbf{H}\mathbf{H}^{\top})\mathbf{K}_{1}^{-1})$$

$$= \frac{|I^{a}|}{n} \frac{\eta_{tm}^{2}}{c} \operatorname{Tr}((\eta_{tm}^{2}\mathbf{D}^{-2} + \mathbb{I}_{r})^{-2}) + o(\frac{|I^{a}|}{n}).$$

with element-wise variance of  $O(d^{-1})$ . The final term is, due to Lemma E.9, has mean of 0, thus only the variance needs to be bounded. Due to the cyclic invariance of trace, this is

$$\begin{split} &\operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}\tilde{\mathbf{U}}^{\top}) \\ &= \operatorname{Tr}(\tilde{\mathbf{U}}^{\top}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}) \\ &= \operatorname{Tr}((\tilde{\mathbf{U}}^{\top}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top})\mathbf{K}_{1}^{-1}(\mathbf{H}\mathbf{H}^{\top})\mathbf{K}_{1}^{-1}(\mathbf{Z}\mathbf{D}^{-1})(\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D})) \end{split}$$

Thus, from Lemma E.9, and Lemma 4, 6, 7, 8 of [27], this term has element-wise variance of  $O(d^{-1})$ . Then we have in total, this term is  $\frac{k}{n}\frac{\eta_{\rm trn}^2}{c}\operatorname{Tr}((\eta_{\rm trn}^2{\bf D}^{-2}+\mathbb{I}_r)^{-2})-\frac{|I^a|}{n}\frac{\eta_{\rm trn}^2}{c}\operatorname{Tr}((\eta_{\rm trn}^2{\bf D}^{-2}+\mathbb{I}_r)^{-1})+O(d^{-1}).$ 

4. Distributing  $\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_1^{-1}\mathbf{Z}\mathbf{P}^{\dagger}$ : This is

$$Tr(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger}(\mathbf{W}_{c}^{sc})^{\top}) = -Tr(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{D}\mathbf{U}^{\top}) + Tr(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H}) + Tr(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}\mathbf{U}^{\top})$$

From this point onward, since the proof follows a similar structure to previous arguments, we provide only a brief sketch. The first term has zero mean, as shown in Lemma E.9. Its variance can be bounded in the usual way by  $O(d^{-1})$ . The second term is, by Lemma E.7 and Lemmas 6, 7, and 8 of [27],

$$\operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H})$$

$$= \frac{|I^{a}|}{n} \frac{\eta_{\text{tm}}^{4}}{c} \operatorname{Tr}(\mathbf{D}^{-2}(\eta_{\text{tm}}^{2}\mathbf{D}^{-2} + \mathbb{I}_{r})^{-2}) + o\left(\frac{|I^{a}|}{n}\right).$$

This has the element-wise variance of  $O(d^{-1})$ . The last term, by Lemma E.9, has zero mean, and applying the standard concentration argument yields an element-wise variance of  $O(d^{-1})$ . Combining all contributions, the total expression is  $\frac{|I^a|}{n} \frac{\eta_{\rm lm}^4}{c} \operatorname{Tr}(\mathbf{D}^{-2}(\eta_{\rm trn}^2\mathbf{D}^{-2} + \mathbb{I}_r)^{-2}) + O(d^{-1})$ .

5. Distributing  $\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_1^{-1}\mathbf{H}$ : This is

$$\begin{split} &\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_1^{-1}\mathbf{H}(\mathbf{W}_c^{sc})^{\top}) = -\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_1^{-1}\mathbf{H}) - \\ &\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_1^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_1^{-1}\mathbf{H}) + \\ &\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_1^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_1^{-1}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}\tilde{\mathbf{U}}^{\top}) \end{split}$$

The first and second terms have zero mean, due to Lemma E.8. For the element-wise variance, Lemma E.3 and Lemmas 4, 6, 7, and 8 from [27] imply that it is of order o(1). The final term, by the cyclic invariance of the trace, Lemma E.10, and again Lemmas 4, 6, 7, and 8 from [27], satisfies that

$$\operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}\tilde{\mathbf{U}}^{\top})$$

$$=\frac{k}{d}\frac{\eta_{\text{tm}}^{2}c}{c-1}\operatorname{Tr}(\mathbf{D}^{-2}(\eta_{\text{tm}}^{2}\mathbf{D}^{-2}+\mathbb{I}_{r})^{-2})+o(\frac{k}{d}).$$

This term has element-wise variance of o(1). Therefore, the total contribution can be summarized as  $\frac{|I^a|}{d} \frac{\eta_{\text{tm}}^2 c}{c^{-1}} \operatorname{Tr}(\mathbf{D}^{-2}(\eta_{\text{tm}}^2 \mathbf{D}^{-2} + \mathbb{I}_r)^{-2}) + o(1)$ .

6. Distributing  $\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{U}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_1^{-1}\mathbf{Z}\mathbf{P}^{\dagger}$ : This is

$$Tr(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger}(\mathbf{W}_{c}^{sc})^{\top}) =$$

$$- Tr(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{D}\tilde{\mathbf{U}}^{\top}) +$$

$$Tr(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H}) +$$

$$Tr(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}\tilde{\mathbf{U}}^{\top})$$

The second term has zero mean due to Lemma E.4. The first term, by Lemma E.10 and Lemmas 6, 7, and 8 of [27], evaluates to

$$\begin{split} &-\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_1^{-1}\mathbf{Z}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{D}\tilde{\mathbf{U}}^{\top})\\ &=-\frac{|I^a|}{d}\frac{\eta_{\mathsf{trn}}^2c}{c-1}\operatorname{Tr}(\mathbf{D}^{-2}(\eta_{\mathsf{trn}}^2\mathbf{D}^{-2}+\mathbb{I}_r)^{-1})+o\left(\frac{|I^a|}{d}\right). \end{split}$$

The last term, again by Lemma E.10 and Lemmas 6, 7, and 8 of [27], becomes

$$\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_1^{-1}\mathbf{Z}\mathbf{P}^{\dagger}(\mathbf{P}^{\dagger})^{\top}\mathbf{Z}^{\top}\mathbf{K}_1^{-1}\mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}\tilde{\mathbf{U}}^{\top})$$

$$= \frac{|I^a|}{d} \frac{\eta_{\text{trn}}^4 c}{c-1} \operatorname{Tr}(\mathbf{D}^{-4}(\eta_{\text{trn}}^2\mathbf{D}^{-2} + \mathbb{I}_r)^{-2}) + o\left(\frac{|I^a|}{d}\right).$$

All of these terms have variance of order o(1). Thus, in total, we have

$$\frac{|I^a|}{d} \frac{\eta_{\text{trn}}^2 c}{c - 1} \operatorname{Tr}(\eta_{\text{trn}}^2 \mathbf{D}^{-4} (\eta_{\text{trn}}^2 \mathbf{D}^{-2} + \mathbb{I}_r)^{-2} - \mathbf{D}^{-2} (\eta_{\text{trn}}^2 \mathbf{D}^{-2} + \mathbb{I}_r)^{-1}) + o(1).$$

Now that all terms have been computed, we organize them to obtain the final expression for  $\mathbb{E}[\|\mathbf{W}_c^{\text{sc}}\|_F^2]$ .

$$\mathbb{E}[\|\mathbf{W}_{c}^{\mathrm{sc}}\|_{F}^{2}] = |I^{a}| + \frac{|I^{a}|}{d} \frac{c}{c-1} \operatorname{Tr}(\mathbf{D}^{2}(\eta_{\mathrm{trn}}^{2} \mathbb{I}_{r} + \mathbf{D}^{2})^{-1}) + \frac{|I^{a}|}{n} \frac{1}{c} \operatorname{Tr}((\mathbf{D}^{-2} + \eta_{\mathrm{trn}}^{-2} \mathbb{I}_{r})^{-1}) + o(1).$$

**Lemma E.12** ( $\mathbb{E}[\operatorname{Tr}(\mathbf{W}_c^{\mathrm{sc}})]$  Term.) For  $c := \frac{d}{r}$  and  $d \ge n + r$ ,

$$\mathbb{E}[\text{Tr}(\mathbf{W}_{c}^{sc})] = -|I^{a}| + \frac{|I^{a}|}{n} \frac{1}{c} \sum_{i=1}^{d} \frac{\eta_{tm}^{2} \sigma_{i}^{2}}{(\eta_{tm}^{2} + \sigma_{i}^{2})} + O(d^{-1}).$$

*Proof.* By using corollary 2.1 of [45] of expanding  $(\mathbf{X} + \mathbf{A})^{\dagger}$ , we have that

$$\begin{aligned} \operatorname{Tr}(\mathbf{W}_{c}^{sc}) &= -\operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}) - \operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger}) + \operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H}) + \\ &\operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger}) + \operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H}) + \\ &\operatorname{Tr}(\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger}) \end{aligned}$$

Note that the first term  $-\operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger})=-|I^a|$ . The second, the fourth, and the sixth terms are 0 due to lemma E.8. The fifth term has mean of 0 due to Lemma E.9. Finally, the third term is

$$\begin{split} \operatorname{Tr}(\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_1^{-1}\mathbf{H}) &= \operatorname{Tr}((\mathbf{H}\mathbf{A}_{I^a}\mathbf{A}^{\dagger}\mathbf{H}^{\top})\mathbf{K}_1^{-1}) \\ &= \frac{|I^a|}{n}\frac{\eta_{\text{tm}}^2}{c}\operatorname{Tr}((\eta_{\text{tm}}^2\mathbf{D}^{-2} + \mathbb{I}_r)^{-1}) + o\left(\frac{|I^a|}{n}\right). \end{split}$$

Using standard concentration arguments, each of these terms exhibits an element-wise variance of order  $O(d^{-1})$ .

**Lemma E.13** ( $\|\mathbf{W}_{c}^{\text{sc}}\mathbf{X}_{\text{tst}}\|_{F}^{2}$  Term).

$$\mathbb{E}\|(\mathbf{W}_{c}^{sc})\mathbf{X}_{tst}\|_{F}^{2}$$

$$=\frac{|I^{a}|}{d}\operatorname{Tr}\left(((c-1)\mathbf{D}^{2}+\mathbb{I}_{d})(\mathbb{I}_{r}+\eta_{tm}^{-2}\mathbf{D}^{2})^{-2}\mathbf{L}\mathbf{L}^{\top}\right)+O(d^{-1}).$$

*Proof.* From Corollary 2.1 of [45], we have that

$$\mathbf{W}_{c}^{sc} = -\mathbf{A}_{I^{a}}(\mathbf{A} + \tilde{\mathbf{U}}\mathbf{D}\tilde{\mathbf{V}}^{\top})^{\dagger}$$

$$= -\mathbf{A}_{I^{a}}(\mathbf{A}^{\dagger} + \mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger} - (\mathbf{A}^{\dagger}\mathbf{H}^{\top} + \mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top})\mathbf{K}_{1}^{-1}(\mathbf{H} + \mathbf{Z}\mathbf{P}^{\dagger}))$$

$$= -\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger} - \mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}\mathbf{P}^{\dagger} + \mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H} + \mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger} +$$

$$\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{H} + \mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top}\mathbf{K}_{1}^{-1}\mathbf{Z}\mathbf{P}^{\dagger}$$
(33)

Using  $X_{tst} = \tilde{U}L$ , with the fact that  $P^{\dagger}\tilde{U} = -D^{-1}$  and  $H\tilde{U} = (Z - I)D^{-1}$ , we have that,

$$\mathbf{W}_c^{sc}\mathbf{X}_{tst} = -(\mathbf{A}_{\mathit{I}^\mathit{a}}\mathbf{A}^\dagger\mathbf{H}^\top\mathbf{K}_1^{-1}\mathbf{D}^{-1} + \mathbf{A}_{\mathit{I}^\mathit{a}}\mathbf{A}^\dagger\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{Z}^\top\mathbf{K}_1^{-1}\mathbf{D}^{-1})\mathbf{L}.$$

Now consider  $\|\mathbf{W}_c^{\text{sc}}\mathbf{X}_{\text{tst}}\|_F^2 = \text{Tr}(\mathbf{X}_{\text{tst}}^{\top}(\mathbf{W}_c^{\text{sc}})^{\top}\mathbf{W}_c^{\text{sc}}\mathbf{X}_{\text{tst}})$ . This is expanded as follows.

$$\begin{aligned} &\operatorname{Tr}(\mathbf{D}^{-1}\mathbf{K}_{1}^{-1}(\mathbf{H}(\mathbf{A}^{\dagger})^{\top}\mathbf{A}_{I^{a}}^{\top}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top} + \mathbf{H}(\mathbf{A}^{\dagger})^{\top}\mathbf{A}_{I^{a}}^{\top}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top} \\ &+ \mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}\tilde{\mathbf{U}}^{\top}(\mathbf{A}^{\dagger})^{\top}\mathbf{A}_{I^{a}}^{\top}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\mathbf{H}^{\top} \\ &+ \mathbf{Z}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{D}\tilde{\mathbf{U}}^{\top}(\mathbf{A}^{\dagger})^{\top}\mathbf{A}_{I^{a}}^{\top}\mathbf{A}_{I^{a}}\mathbf{A}^{\dagger}\tilde{\mathbf{U}}\mathbf{D}(\mathbf{P}^{\top}\mathbf{P})^{-1}\mathbf{Z}^{\top})\mathbf{K}_{1}^{-1}\mathbf{D}^{-1}) \end{aligned}$$

Note that the second and third term have mean of 0 due to Lemma E.9. Thus both terms are 0, and since their element-wise variances are of order  $O(d^{-1})$ , the error introduced by this approximation is also  $O(d^{-1})$ . According to Lemma E.7 and Lemma 8 of [27], the first term has mean  $\eta_{\text{trn}}^4 \frac{|I^a|}{n} \frac{c-1}{c} (\eta_{\text{trn}}^2 \mathbf{D}^{-2} + \mathbb{I}_r)^{-2} \mathbf{D}^{-2}$ , with the element-wise variance of  $O(d^{-1})$ . Similarly, the fourth term has mean  $\eta_{\text{trn}}^4 \frac{|I^a|}{d} \mathbf{D}^{-4} (\eta_{\text{trn}}^2 \mathbf{D}^{-2} + \mathbb{I}_r)^{-2}$  with element-wise variance of  $O(d^{-1})$ . This is due to Lemma E.10, and Lemmas 6, 7, 8 of [27]. Organizing the terms, we have that

$$\mathbb{E}\|\mathbf{W}_{c}^{\text{sc}}\mathbf{X}_{\text{tst}}\|_{F}^{2}$$

$$=\frac{|I^{a}|}{d}\operatorname{Tr}\left(((c-1)\mathbf{D}^{2}+c\mathbb{I}_{d})(c\mathbb{I}_{r}+c\eta_{\text{trn}}^{-2}\mathbf{D}^{2})^{-2}\mathbf{L}\mathbf{L}^{\top}\right)+O(d^{-1}).$$

Substituting Lemmas E.11, E.12, and E.13 into the decomposition in Equation (28) yields the desired result.

#### **E.2** Bias-Variance Decomposition

For the model with skip connection, we defined the term  $\|\mathbf{X}_{tst} - \mathbf{W}_c \mathbf{X}_{tst}\|_F^2$  as the bias term, and  $N_{tst}^{-1} \|\mathbf{W}_c^{sc}\|_F^2$  as the variance term in Section 3. From Lemma E.4, the bias term is indeed asymptotically equal to  $\frac{1}{N_{tst}} \operatorname{Tr}(\mathbf{JLL}^T)$ . On the other hand, Lemma E.3 shows that the variance term satisfies  $\frac{\eta_{tst}^2}{d} \mathbb{E}[\|\mathbf{W}_c\|_F^2] \approx \frac{\eta_{tst}^2}{d} \frac{c}{c-1} \sum_{j \in I^x} \frac{\sigma_j^2}{\eta_{trn}^2 + \sigma_j^2}$ . These bias and variance expressions exhibit a trade-off behavior as the bottleneck dimension varies, closely resembling the classical bias-variance

**Definition E.14** (Bias and Variance in Two-Layer Linear DAEs). The bias term in the under-complete linear DAE is defined as the component of the test error that **decreases** as the model complexity (i.e., the bottleneck dimension k) increases. Conversely, the variance term is defined as the component of the test error that **increases** as the model complexity grows.

relationship. Motivated by this observation, we propose the following definition.

Unlike the model without skip connections, the skip-connected model does not admit a clear decomposition that allows for straightforward interpretation. Nevertheless, in Remark 3.6, we defined  $\|\mathbf{W}_c^{\text{sc}}\|_F^2$  as a variance term, following the definition provided in Definition E.14. Lemma E.11 supports this interpretation by showing that this quantity captures the variance behavior described therein. Especially, among the various variance contributions, the term involving the  $(c-1)^{-1}$  factor becomes dominant as c gets closer to 1, and this definition includes this term.

On the Bias Term of the Model with Skip Connections We have seen that the bias term of the skip connection model includes  $\eta_{\text{tst}}^2$ , which is relatively large compared to the model without a skip connection, unless we have a very high signal-to-noise ratio (low  $\eta_{\text{tst}}$ ). The decomposition early in this subsection (31) shows that the constant  $\eta_{\text{tst}}^2$  originates from

$$rac{1}{N_{ ext{tst}}} \mathbb{E}_{\mathbf{A}_{ ext{tst}}} \left[ \operatorname{Tr}(\mathbf{A}_{ ext{tst}} \mathbf{A}_{ ext{tst}}^T) 
ight].$$

This occurs because incorporating a skip connection in two-layer model makes the target changes from low-rank target  $\mathbf{X}_{tst}$  to full-rank noisy target  $\mathbf{A}_{tst}$ . In contrast, the model we consider has fixed rank budget k. We believe this is due to of the limitation of having skip connection in two-layer linear models. Things could be different, for instance, in *four-layer linear* models, with the skip connection exists between the two hidden layers in the middle. In this case, the target is no longer the full rank noisy matrix  $A_{tst}$ . To illustrate this, let us examine the decomposition of this model. For a skip connection between the middle hidden layers, the network structure is defined as

$$\mathbf{W} := \mathbf{W}_4(\mathbf{W}_3\mathbf{W}_2 + \mathbb{I})\mathbf{W}_1.$$

Then, the decomposition of the test metric for a four-layer linear model is given by

$$\begin{split} &\frac{1}{N_{\text{tst}}} \mathbb{E}\left[\|\mathbf{X}_{\text{tst}} - \mathbf{W}_4(\mathbf{W}_3\mathbf{W}_2 + \mathbb{I})\mathbf{W}_1(\mathbf{X}_{\text{tst}} + \mathbf{A}_{\text{tst}})\|_F^2\right] \\ &= \frac{1}{N_{\text{tst}}} (\|\mathbf{X}_{\text{tst}}\|_F^2 - 2\operatorname{Tr}(\mathbf{W}\mathbf{X}_{\text{tst}}\mathbf{X}_{\text{tst}}^T) + \frac{\eta_{\text{tst}}^2 N_{\text{tst}}}{d}\|\mathbf{W}\|_F^2). \end{split}$$

Now there is no full-rank noisy target  $A_{tst}$  contributing to the constant  $\eta_{tst}^2$ . Thus, there will be no large bias term arising from this. It is left to future work to characterize the exact bias term for this model and compare it with the two-layer linear models with skip connections.

## F Proofs and Supporting Results for Section 4

In this section, we first identify the eigenvalue locations of the model described in Definition 4.1, which implies that its eigenvalue distribution follows the Marchenko–Pastur law (Theorem C.2). We then show that the eigenvalue distribution of *information-plus-noise* model, which is used throughout the paper, follows also the Marchenko–Pastur. These results suggest that the intuition developed from the simplified model (Definition 4.1) may extend to the information-plus-noise setting. This analysis demonstrates that the peak observed near  $c \approx 1$  arises from the accumulation of small eigenvalues near zero, whose number increases as  $c \to 1$ . This supports the argument presented in Section 4. Finally, in Subsection F.2, we provide the proof of Theorem 4.3 in the main text.

#### **F.1** Understanding the Peak Near $c \approx 1$

First, we identify the location of eigenvalues of S that was described in Definition 4.1. Recall that  $\lambda_1$ ,  $\mathbf{u}_1$  denotes the eigenvalue and eigenvector of the rank-1  $\mathbf{X}\mathbf{X}^T$ , respectively.

**Lemma F.1** (Location of Eigenvalues for Rank-1 Additive Model). Let  $\lambda^{\mathbf{S}}$  be any non-zero eigenvalue of  $\mathbf{S}$ . Then, it satisfies that  $\mathbf{u}_1^T\mathbf{Q}_{\mathbf{A}}(\lambda^{\mathbf{S}})\mathbf{u}_1 = -\frac{1}{\lambda_1}$ . Furthermore, let  $\lambda_m^{\mathbf{S}}$  be a m-th eigenvalue of  $\mathbf{S}$ , for  $m \in \{2, \ldots, n\}$ . Then,  $\lambda_m^{\mathbf{S}} \in (\lambda_m^{\mathbf{A}}, \lambda_{m-1}^{\mathbf{A}})$ .

*Proof.* To identify the eigenvalue information, we observe  $det(\mathbf{S} - \alpha \mathbb{I}_d)$ . For  $\lambda_j^{\mathbf{S}}, j \in \{1, \dots, d\}$ , we now derive an equivalent condition for  $det(\mathbf{S} - \lambda_j^{\mathbf{S}} \mathbb{I}_d) = 0$ . Observe that,

$$det(\mathbf{S} - \lambda^{\mathbf{S}} \mathbb{I}_{d})$$

$$= det(\mathbf{X} \mathbf{X}^{\mathbf{T}} + \mathbf{A} \mathbf{A}^{T} - \lambda^{\mathbf{S}} \mathbb{I}_{d})$$

$$= det(\mathbf{Q}_{\mathbf{A}}^{-1}) det(\mathbb{I} + \lambda_{1} \mathbf{Q}_{\mathbf{A}} \mathbf{u}_{1} \mathbf{u}_{1}^{T})$$

$$= det(\mathbf{Q}_{\mathbf{A}}^{-1}) det(1 + \lambda_{1} \mathbf{u}_{1}^{T} \mathbf{Q}_{\mathbf{A}} \mathbf{u}_{1}).$$

The last equality is due to Lemma C.4-(5). Thus we have the equivalent condition that,

$$det(\mathbf{Q}_{\mathbf{A}}^{-1})det(1+\lambda_1\mathbf{u}_1^T\mathbf{Q}_{\mathbf{A}}\mathbf{u}_1)=0.$$
(34)

Thus we have either  $det\left(\mathbf{Q_A^{-1}}(\lambda_j^{\mathbf{S}})\right)=0$ , or  $det\left(1+\lambda_1\mathbf{u}_1^T\mathbf{Q_A}(\lambda^{\mathbf{S}})\mathbf{u}_1\right)=0$ . The former one is not possible, as  $\lambda_j^{\mathbf{S}}$  is not an eigenvalue of  $\mathbf{AA}^T$ . The latter one is equivalent to finding  $\alpha\in\mathbb{R}_+\backslash\{\lambda_i^{\mathbf{A}},\ldots,\lambda_d^{\mathbf{A}}\}$ , such that  $\mathbf{u}_1^T\mathbf{Q_A}(\alpha)\mathbf{u}_1=-\frac{1}{\lambda_1}$ . Note that  $f(\alpha):=\mathbf{u}_1^T\mathbf{Q_A}(\alpha)\mathbf{u}_1$  is increasing in every interval of  $(\lambda_k^{\mathbf{A}},\lambda_{k-1}^{\mathbf{A}})$ , for  $k\in\{2,\ldots,n\}$ , as  $f'(\alpha):=\mathbf{u}_1^T\mathbf{Q_A}(\alpha)\mathbf{u}_1>0$ . In addition to this,  $\lim_{\alpha\downarrow\lambda_k^{\mathbf{A}}}f(\alpha)=-\infty$ ,  $\lim_{\alpha\uparrow\lambda_{k-1}^{\mathbf{A}}}f(\alpha)=\infty$ , thus this  $f(\alpha)$  is monotonically increasing from  $-\infty$  to

 $\infty$ . This means that we have eigenvalue  $\lambda_k^{\mathbf{S}}$  inside every interval  $(\lambda_k^{\mathbf{A}}, \lambda_{k-1}^{\mathbf{A}}), \forall k \in \{2, \dots, n\}$ . Thus we have exactly n eigenvalues, including  $\lambda_1^{\mathbf{S}} \in (\lambda_1, \infty)$ .

This result shows that the empirical spectral distributions of S and  $AA^{\top}$  are essentially identical in the limit. A similar property holds for the *information-plus-noise* model: the empirical eigenvalue distribution converges weakly to the Marchenko–Pastur law. This is formalized in the following theorem.

**Theorem F.2** ( $\mu_N$  converges weakly to  $\mu_{MP}$ ).

For  $\alpha \in \mathbb{C} \setminus \mathbb{R}_+$ , let  $\mu_N$  be the empirical spectral measure (Def C.1) of  $(\mathbf{X} + \mathbf{A})(\mathbf{X} + \mathbf{A})^T$ , for  $\mathbf{X}$  and  $\mathbf{A}$  satisfying Assumption 3.1. In addition to this, assume that for some constant  $C_1 > 0$ , it satisfies that  $\|\mathbf{x}_i\|_2 \leq \frac{C_1}{\sqrt{N}}$ , for  $i \in \{1, \dots, N\}$ . Let  $\mu_{MP}$  be a version of Marchenko-Pastur distribution, where

$$\mu_{MP}(\alpha) = \begin{cases} \frac{\sqrt{(c\alpha - \eta^2(\sqrt{c} - 1)^2)(\eta^2(\sqrt{c} + 1)^2 - c\alpha)}}{2\pi\alpha c\eta^2} & \text{if } \alpha \in \left[\frac{\eta^2}{c}(\sqrt{c} - 1)^2, \frac{\eta^2}{c}(\sqrt{c} + 1)^2\right] \\ 1 - \frac{1}{c} & \text{else.} \end{cases}$$

Then,  $\mu_N$  converges weakly to  $\mu_{MP}$ .

Note that the assumption regarding the norm of the columns of X is natural, given that we have assumed  $\|X\|_2$  scales as  $\Theta(1)$ . This implies the Frobenius norm of X also scales as  $\Theta(1)$ , since X has fixed rank r. If each data point  $x_i$  scales at the same rate, then each individual data point would scale as  $\Theta(N^{-1/2})$ . Thus, this assumption is equivalent to stating that each data point  $x_i$  follows the same scaling behavior.

*Proof.* It is sufficient to show  $m_{\mu_N} \xrightarrow{a.s} m_{\mu_{MP}}$  to reach the conclusion, based on the fact of that  $\mathbb{P}(\mu_N \to \mu_{MP} \text{ weakly}) = 1 \Leftrightarrow m_{\mu_N} \xrightarrow{a.s} m_{\mu_{MP}}$  [39, Exercise 2.4.10]. To establish this convergence, we first show that  $m_{\mu_N} \xrightarrow{a.s} \mathbb{E}[m_{\mu_N}]$  and then show  $\mathbb{E}[m_{\mu_N}] \xrightarrow{a.s} m_{\mu_{MP}}$ , to conclude  $m_{\mu_N} \xrightarrow{a.s} m_{\mu_{MP}}$ .

Firstly, in order to show  $m_{\mu_N} \stackrel{a.s}{\longrightarrow} \mathbb{E}[m_{\mu_N}]$ , we follow the standard approach outlined in Lemma 2.12 and 2.13 of [6]. In essence, we first construct a Martingale Difference Sequence to show that  $\frac{1}{d}\operatorname{Tr}(\mathbf{Q}_{-i})$  converges to  $\frac{1}{d}\operatorname{Tr}(\mathbf{Q})$  almost surely. Then, we use Lemma 2.12 [6] to find an upper bound and finish with the Borel-Cantelli Lemma to establish the almost sure convergence. A key step in the proof is demonstrating that  $\left|\frac{1}{d}\operatorname{Tr}(\mathbf{Q}_{-i})-\frac{1}{d}\operatorname{Tr}(\mathbf{Q})\right|$  is sufficiently small. We cannot directly use Lemma C.4-(4) however, since  $\mathbf{Q}=(\mathbf{Z}\mathbf{Z}^T-\alpha\mathbb{I}_d)^{-1}$ , and  $\mathbf{z}_i=\mathbf{x}_i+\mathbf{a}_i$  is not just a mean-zero gaussian. Nonetheless, leveraging the *low-rank* property of  $\mathbf{X}$ , we can still show that this term remains small, as stated in the following lemma.

**Lemma F.3** Consider the setting of Theorem F.2. Let  $\mathbf{Q} := (\mathbf{Z}\mathbf{Z}^T - \alpha \mathbb{I}_d)^{-1}$ , where  $\mathbf{Z} = \mathbf{X} + \mathbf{A}$ . Then,  $\left|\frac{1}{d}\operatorname{Tr}(\mathbf{Q}) - \frac{1}{d}\operatorname{Tr}(\mathbf{Q}_{-j})\right| = O(n^{-1})$ .

Proof of Lemma F.3. Due to the Sherman-Morrison Lemma, (Lemma C.4-(2)), we have that  $\mathbf{Q} = \mathbf{Q}_{-j} - \frac{\mathbf{Q}_{-j}\mathbf{z}_{j}\mathbf{z}_{j}^{T}\mathbf{Q}_{-j}}{1+\mathbf{z}_{j}^{T}\mathbf{Q}_{-j}\mathbf{z}_{j}}$ . Thus  $\frac{1}{d}\operatorname{Tr}(\mathbf{Q}) - \frac{1}{d}\operatorname{Tr}(\mathbf{Q}_{-j}) = -\frac{1}{d}\operatorname{Tr}\left(\frac{\mathbf{z}_{j}^{T}\mathbf{Q}_{-j}^{2}\mathbf{z}}{1+\mathbf{z}_{j}^{T}\mathbf{Q}_{-j}\mathbf{z}}\right)$ . Note that  $\mathbf{z}_{j}^{T}\mathbf{Q}_{-j}\mathbf{z}_{j} = \mathbf{x}_{j}^{T}\mathbf{Q}_{-j}\mathbf{x}_{j} + \mathbf{x}_{j}^{T}\mathbf{Q}_{-j}\mathbf{x}_{j} + \mathbf{a}_{j}^{T}\mathbf{Q}_{-j}\mathbf{a}_{j}$ . The first term is  $\mathbf{x}^{T}\mathbf{Q}_{-j}\mathbf{x}_{j} \leq \|\mathbf{x}_{j}\|_{2}^{2}\|\mathbf{Q}_{-j}\|_{2} = O(n^{-1})$ , as  $\|\mathbf{Q}_{-j}\|_{2}$  is bounded and  $\|\mathbf{x}_{j}\|_{2} = O(\frac{1}{\sqrt{N}})$ .  $\mathbf{x}_{j}^{T}\mathbf{Q}_{-j}\mathbf{a}_{j}$  has mean of 0, and variance is  $\mathbb{E}\left[\mathbf{x}_{j}^{T}\mathbf{Q}_{-j}\mathbf{a}_{j}\mathbf{a}_{j}^{T}\mathbf{Q}_{-j}\mathbf{x}_{j}\right] \leq \frac{n^{2}}{d}\|\mathbf{x}_{j}\|_{2}^{2}\|\mathbf{Q}_{-j}\|_{2}^{2} = O(n^{-2})$ . Thus applying the Borel-Cantelli lemma will give us that  $\mathbf{x}_{j}^{T}\mathbf{Q}_{-j}\mathbf{a}_{j} \stackrel{a.s}{\longrightarrow} 0$ . Therefore, we have that  $\mathbf{z}_{j}^{T}\mathbf{Q}_{-j}\mathbf{z}_{j} \stackrel{a.s}{\longrightarrow} \mathbf{a}_{j}^{T}\mathbf{Q}_{-j}\mathbf{a}_{j}$ . With this, we conclude that

$$\left| \frac{1}{d} \operatorname{Tr}(\mathbf{Q}) - \frac{1}{d} \operatorname{Tr}(\mathbf{Q}_{-j}) \right| = \left| \frac{1}{d} \operatorname{Tr} \left( \frac{\mathbf{z}_{j}^{T} \mathbf{Q}_{-j}^{2} \mathbf{z}}{1 + \mathbf{z}_{j}^{T} \mathbf{Q}_{-j} \mathbf{z}} \right) \right|$$

$$\simeq \left| \frac{1}{d} \operatorname{Tr} \left( \frac{\mathbf{a}_{j}^{T} \mathbf{Q}_{-j}^{2} \mathbf{a}_{j}}{1 + \mathbf{a}_{j}^{T} \mathbf{Q}_{-j} \mathbf{a}_{j}} \right) \right|$$

$$\stackrel{C.4.(3)}{\simeq} \left| \frac{1}{d} \operatorname{Tr} \left( \frac{\frac{\eta^{2}}{d} \operatorname{Tr}(\mathbf{Q}_{-j}^{2})}{1 + \frac{\eta^{2}}{d} \operatorname{Tr}(\mathbf{Q}_{-j})} \right) \right|$$

$$= O(d^{-1}) = O(n^{-1}).$$

Now back to the original proof, recall that  $m_{\mu_N} = \frac{1}{d} \operatorname{Tr}(\mathbf{Q})$ . Then, we construct the martigale difference sequence, which is

$$m_{\mu_N} - \mathbb{E}[m_{\mu_N}] = \sum_{j=1}^N \left( \mathbb{E}_j \left[ \frac{1}{d} \operatorname{Tr}(\mathbf{Q}) \right] - \mathbb{E}_{j-1} \left[ \frac{1}{d} \operatorname{Tr}(\mathbf{Q}) \right] \right),$$

for  $\mathbb{E}_j\left[\frac{1}{d}\operatorname{Tr}(\mathbf{Q})\right]:=\mathbb{E}\left[\frac{1}{d}\operatorname{Tr}(\mathbf{Q});\mathbf{z}_1,\ldots,\mathbf{z}_j\right]$ , and  $\mathbb{E}_0[m_{\mu_N}]:=\mu_N$ . This is by construction a martingale difference sequence, since

$$\mathbb{E}\left[\left(\mathbb{E}_{j} - \mathbb{E}_{j-1}\right) \left[\frac{1}{d}\operatorname{Tr}(\mathbf{Q})\right]; \mathbf{z}_{1}, \dots, \mathbf{z}_{j-1}\right] = 0.$$

This is due to the fact that  $\mathbb{E}\left[\mathbb{E}_{j}\left[\frac{1}{d}\operatorname{Tr}(\mathbf{Q})\right];\mathbf{z}_{1},\ldots,\mathbf{z}_{j-1}\right]=\mathbb{E}_{j-1}\left[\frac{1}{d}\operatorname{Tr}(\mathbf{Q})\right]$  ([17, Theorem 4.1.13]).

Now, observe that  $\mathbb{E}_j\left[\frac{1}{d}\operatorname{Tr}(\mathbf{Q}_{-j})\right] = \mathbb{E}_{j-1}\left[\frac{1}{d}\operatorname{Tr}(\mathbf{Q}_{-j})\right]$ , then we have

$$\sum_{j=1}^{N} \left( \mathbb{E}_{j} \left[ \frac{1}{d} \operatorname{Tr}(\mathbf{Q}) \right] - \mathbb{E}_{j-1} \left[ \frac{1}{d} \operatorname{Tr}(\mathbf{Q}) \right] \right)$$

$$= \sum_{j=1}^{N} \left( \mathbb{E}_{j} \left[ \frac{1}{d} \operatorname{Tr}(\mathbf{Q}) - \frac{1}{d} \operatorname{Tr}(\mathbf{Q}_{-j}) \right] - \mathbb{E}_{j-1} \left[ \frac{1}{d} \operatorname{Tr}(\mathbf{Q}) - \frac{1}{d} \operatorname{Tr}(\mathbf{Q}_{-j}) \right] \right).$$

With Lemma F.3, we have that  $(\mathbb{E}_j - \mathbb{E}_{j-1}) \left[ \frac{1}{d} \operatorname{Tr}(\mathbf{Q}) \right] = O(n^{-1})$ . Applying [6, Lemma 2.12], for some constant  $K_2 > 0$ , we have that

$$\mathbb{E}\left[\left|m_{\mu_{N}} - \mathbb{E}[m_{\mu_{N}}]\right|^{4}\right] = \mathbb{E}\left[\sum_{j=1}^{N} (\mathbb{E}_{j} - \mathbb{E}_{j-1}) \left[\frac{1}{d}\operatorname{Tr}(\mathbf{Q}) - \frac{1}{d}\operatorname{Tr}(\mathbf{Q}_{-i})\right]\right]$$

$$\leq K_{2}\mathbb{E}\left[\left(\sum_{j=1}^{N} \left| (\mathbb{E}_{j} - \mathbb{E}_{j-1}) \left[\frac{1}{d}\operatorname{Tr}(\mathbf{Q}) - \frac{1}{d}\operatorname{Tr}(\mathbf{Q}_{-i})\right]\right|^{2}\right)^{2}\right]$$

$$= O(n^{-2}).$$

It follows that, for any  $\epsilon > 0$ ,

$$\mathbb{P}(|m_{\mu_N} - \mathbb{E}[m_{\mu_N}]| > \epsilon) \le \frac{\mathbb{E}\left[m_{\mu_N} - \mathbb{E}[m_{\mu_N}]|^4\right]}{\epsilon^4}$$
$$= O(n^{-2}).$$

Applying the standard Borel-Cantelli lemma, we obtain  $m_{\mu_N} \xrightarrow{a.s.} \mathbb{E}[m_{\mu_N}]$ . This completes the first step. From this point onward, we denote  $m := \mathbb{E}[m_{\mu_N}]$ .

Since we have established the convergence of  $m_{\mu_N} \xrightarrow{a.s} m$ , our goal is now to show  $m \xrightarrow{a.s} m_{\mu_{MP}}$ . For this, the key idea is to find a fixed-point equation, leveraging the close asymptotical relationship between  $\mathbf{Q}$  and  $\mathbf{Q}_{-j}$ . From  $\mathbf{Q} = \mathbf{Q}_{-j} - \frac{\mathbf{Q}_{-j}\mathbf{z}_j\mathbf{z}_j^T\mathbf{Q}_{-j}}{1+\mathbf{z}_j^T\mathbf{Q}_{-j}\mathbf{z}_j}$  (Lemma C.4-(2)), it holds that  $\mathbf{z}_j^T\mathbf{Q}\mathbf{z}_j = \frac{\mathbf{z}_j^T\mathbf{Q}_{-j}\mathbf{z}_j}{1+\mathbf{z}_j^T\mathbf{Q}_{-j}\mathbf{z}_j}$ . From the proof of Lemma F.3, we already established that

$$\mathbf{z}_{j}^{T}\mathbf{Q}\mathbf{z}_{j} \xrightarrow{a.s} \mathbf{a}_{j}^{T}\mathbf{Q}_{-j}\mathbf{a}_{j}.$$

Using this result, it satisfies that  $\mathbf{z}_j^T \mathbf{Q} \mathbf{z}_j \simeq \frac{\eta^2}{d} \frac{\mathrm{Tr}(\mathbf{Q}_{-j})}{1 + \frac{\eta^2}{d} \mathrm{Tr}(\mathbf{Q}_{-j})} \simeq \frac{\eta^2}{d} \frac{\mathrm{Tr}(\mathbf{Q})}{1 + \frac{\eta^2}{d} \mathrm{Tr}(\mathbf{Q})}$ . Since this holds for all  $j \in [n]$ , it follows that  $\sum_{j=1}^n \mathbf{z}_j^T \mathbf{Q} \mathbf{z}_j \simeq n \frac{\eta^2}{d} \frac{\mathrm{Tr}(\mathbf{Q})}{1 + \frac{\eta^2}{d} \mathrm{Tr}(\mathbf{Q})}$ . Using the identity  $\sum_{j=1}^n \mathbf{z}_j^T \mathbf{Q} \mathbf{z}_j = \mathrm{Tr}(\mathbf{Z} \mathbf{Z}^T \mathbf{Q})$ , and the fact that  $\mathbf{Z} \mathbf{Z}^T = (\mathbf{Q}^{-1} + \alpha \mathbb{I}_d)$ , we obtain  $\sum_{j=1}^n \mathbf{z}_j^T \mathbf{Q} \mathbf{z}_j = \mathrm{Tr}(\mathbb{I}_d + \alpha \mathbf{Q})$ . Thus, we arrive at  $d + \alpha \mathrm{Tr}(\mathbf{Q}) \simeq n \frac{\eta^2}{d} \frac{\mathrm{Tr}(\mathbf{Q})}{1 + \frac{\eta^2}{d} \mathrm{Tr}(\mathbf{Q})}$ . In the limit case where  $d, n \to \infty$ , this simplifies to the fixed point equation

$$1 + \alpha m = \frac{\eta^2 cm}{1 + \eta^2 m}$$

, for  $c = \frac{d}{n}$ . After rearranging, we obtain the quadratic equaion

$$\alpha c\eta^2 m^2 + (\alpha c + c\eta^2 - \eta^2) m + c = 0.$$

This is precisely the quadratic equation for  $m_{\mu_{MP}}$ , which proves  $m \xrightarrow{a.s} m_{\mu_{MP}}$ . To see this more clearly, note that

$$m(\alpha) = \frac{\eta^2 - c\eta^2 - \alpha c}{2\alpha c\eta^2} \pm \frac{\sqrt{(c\alpha - \eta^2(\sqrt{c} - 1)^2)(c\alpha - \eta^2(\sqrt{c} + 1)^2)}}{2\alpha c\eta^2}.$$

Using the Inverse Stieltjes Transform(See [12, Theorem 2.4]), we find that for all  $\alpha \in \mathbb{C} \setminus \{0\}$ ,

$$\mu_{MP}(\alpha) = \frac{1}{\pi} \lim_{\epsilon \to 0} Im(m(\alpha + i\epsilon))$$

$$= \frac{\sqrt{(c\alpha - \eta^2(\sqrt{c} - 1)^2)(c\alpha - \eta^2(\sqrt{c} + 1)^2)}}{2\pi\alpha c\eta^2}, \text{ for } \alpha \in \left[\frac{\eta^2}{c}(\sqrt{c} - 1)^2, \frac{\eta^2}{c}(\sqrt{c} + 1)^2\right].$$

For  $\alpha = 0$ , we have that

$$\begin{split} \mu_{MP}(\{0\}) &= -\lim_{\epsilon \downarrow 0} i \epsilon m(i \epsilon) \\ &= \begin{cases} 0 & \text{if } c < 1 \\ 1 - \frac{1}{c} & \text{if } c \geq 1. \end{cases}. \end{split}$$

Therefore  $m \xrightarrow{a.s} m_{\mu_{MP}}$ . Then,  $m_{\mu_N} \xrightarrow{a.s} m_{\mu_{MP}}$  follows and this leads to  $\mu_N \xrightarrow{a.s} \mu_{MP}$ , which concludes the proof.

#### F.2 Proof of Theorem 4.3

We begin by establishing the condition for the location of eigenvalues of S. Then, we apply the Cauchy integral formula as introduced in Subsection C.2. We denote the resolvent of S as  $Q_S(\alpha) = (S - \alpha \mathbb{I}_d)^{-1}$  and the resolvent of  $AA^T$  as  $Q_A(\alpha) = (AA^T - \alpha \mathbb{I}_d)^{-1}$ .

We aim to analyze the following quantity, where  $\Gamma_{\lambda_j^{\mathbf{S}}}$  is a closed, positive oriented contour that *only* encompasses  $\lambda_j^{\mathbf{S}}$ .

$$\langle \mathbf{u}_1, \mathbf{u}_j^{\mathbf{S}} \rangle^2 = -\frac{1}{2\pi i} \int_{\Gamma_{\lambda_s^{\mathbf{S}}}} \mathbf{u}_1^T \mathbf{Q}_{\mathbf{S}} \mathbf{u}_1 \, d\alpha. \tag{35}$$

We first pull  $Q_A$  out of  $Q_S$  using the Woodbury Identity (Lemma C.4-(6)).

$$\begin{aligned} \mathbf{Q_S} &= (\mathbf{Q_A}^{-1} + \lambda_1 \mathbf{u}_1 \mathbf{u}_1^T)^{-1} \\ &= \mathbf{Q_A} - \frac{\lambda_1}{1 + \mathbf{u}_1^T \mathbf{Q_A} \mathbf{u}_1} \mathbf{Q_A} \mathbf{u}_1 \mathbf{u}_1^T \mathbf{Q_A}. \end{aligned}$$

With this, we have that

$$\begin{split} \langle \mathbf{u}_1, \mathbf{u}_j^{\mathbf{S}} \rangle^2 &= -\frac{1}{2\pi i} \int_{\Gamma_{\lambda_j^{\mathbf{S}}}} \mathbf{u}_1^T \mathbf{Q}_{\mathbf{S}} \mathbf{u}_1 \, d\alpha \\ &= -\frac{1}{2\pi i} \int_{\Gamma_{\lambda_j^{\mathbf{S}}}} \mathbf{u}_1^T \mathbf{Q}_{\mathbf{A}} \mathbf{u}_1 \, d\alpha + \\ &\frac{1}{2\pi i} \int_{\Gamma_{\lambda_j^{\mathbf{S}}}} \frac{1}{1 + \mathbf{u}_1^T \mathbf{Q}_{\mathbf{A}} \mathbf{u}_1} \mathbf{u}_1^T \mathbf{Q}_{\mathbf{A}} \mathbf{u}_1 \mathbf{u}_1^T \mathbf{Q}_{\mathbf{A}} \mathbf{u}_1 \, d\alpha. \end{split}$$

Note that the first integral is 0, since there is no singularity inside the contour. For the second integral, we have the singularity at  $\lambda_i^{\mathbf{S}}$  from Lemma F.1. Then, using the residue calculus, it follows that

$$\begin{split} &\frac{1}{2\pi i} \int_{\Gamma_{\boldsymbol{\lambda}_{j}^{\mathbf{S}}}} \frac{1}{1 + \mathbf{u}_{1}^{T} \mathbf{Q}_{\mathbf{A}} \mathbf{u}_{1}} \mathbf{u}_{1}^{T} \mathbf{Q}_{\mathbf{A}} \mathbf{u}_{1} \mathbf{u}_{1}^{T} \mathbf{Q}_{\mathbf{A}} \mathbf{u}_{1} d\alpha \\ &= \lim_{\alpha \to \boldsymbol{\lambda}_{j}^{\mathbf{S}}} (\alpha - \boldsymbol{\lambda}_{j}^{\mathbf{S}}) \frac{1}{1 + \mathbf{u}_{1}^{T} \mathbf{Q}_{\mathbf{A}} \mathbf{u}_{1}} \mathbf{u}_{1}^{T} \mathbf{Q}_{\mathbf{A}} \mathbf{u}_{1} \mathbf{u}_{1}^{T} \mathbf{Q}_{\mathbf{A}} \mathbf{u}_{1} \\ &= \frac{1}{\lambda_{1}^{2}} \lim_{\alpha \to \boldsymbol{\lambda}_{j}^{\mathbf{S}}} (\alpha - \boldsymbol{\lambda}_{j}^{\mathbf{S}}) \frac{1}{1 + \mathbf{u}_{1}^{T} \mathbf{Q}_{\mathbf{A}} \mathbf{u}_{1}}. \end{split}$$

The last equality is due to Lemma F.1.

We denote  $f(\alpha) := \mathbf{u}_1^T \mathbf{Q}_{\mathbf{A}}(\alpha) \mathbf{u}_1$ , then we have that

$$\lim_{\alpha \to \lambda_j^{\mathbf{S}}} (\alpha - \lambda_j^{\mathbf{S}}) \frac{1}{1 + \mathbf{u}_1^T \mathbf{Q_A} \mathbf{u}_1} = \frac{1}{f'(\lambda_j^{\mathbf{S}})}.$$

Therefore, we conclude that

$$\langle \mathbf{u}_1, \mathbf{u}_j^{\mathbf{S}} \rangle^2 = \frac{1}{\lambda_1^2 f'(\lambda_j^{\mathbf{S}})}.$$
 (36)

Now we work on  $\langle \mathbf{u}_i^{\mathbf{A}}, \mathbf{u}_j^{\mathbf{S}} \rangle^2$ . We are interested in:

$$\langle \mathbf{u}_i^{\mathbf{A}}, \mathbf{u}_j^{\mathbf{S}} \rangle^2 = -\frac{1}{2\pi i} \int_{\Gamma_{\lambda_i^{\mathbf{S}}}} (\mathbf{u}_i^{\mathbf{A}})^T \mathbf{Q}_{\mathbf{S}} \mathbf{u}_i^{\mathbf{A}} d\alpha.$$
 (37)

Following the same steps as above, we have the non-trivial term:

$$\begin{split} &\frac{1}{2\pi i} \int_{\Gamma_{\boldsymbol{\lambda_{j}^{S}}}} \frac{1}{1+\mathbf{u}_{1}^{T}\mathbf{Q_{A}}\mathbf{u}_{1}} (\mathbf{u}_{i}^{\mathbf{A}})^{T}\mathbf{Q_{A}}\mathbf{u}_{1}\mathbf{u}_{1}^{T}\mathbf{Q_{A}}\mathbf{u}_{i}^{\mathbf{A}} \, d\alpha \\ &= \lim_{\alpha \to \boldsymbol{\lambda_{j}^{S}}} (\alpha - \boldsymbol{\lambda_{j}^{S}}) \frac{1}{1+\mathbf{u}_{1}^{T}\mathbf{Q_{A}}\mathbf{u}_{1}} (\mathbf{u}_{i}^{\mathbf{A}})^{T}\mathbf{Q_{A}}\mathbf{u}_{1}\mathbf{u}_{1}^{T}\mathbf{Q_{A}}\mathbf{u}_{i}^{\mathbf{A}} \\ &= \frac{\langle \mathbf{u}_{i}^{\mathbf{A}}, \mathbf{u}_{1} \rangle^{2}}{(\boldsymbol{\lambda_{i}^{A}} - \boldsymbol{\lambda_{j}^{S}})^{2}} \lim_{\alpha \to \boldsymbol{\lambda_{j}^{S}}} (\alpha - \boldsymbol{\lambda_{j}^{S}}) \frac{1}{1+\mathbf{u}_{1}^{T}\mathbf{Q_{A}}\mathbf{u}_{1}} \\ &= \frac{\langle \mathbf{u}_{i}^{\mathbf{A}}, \mathbf{u}_{1} \rangle^{2}}{(\boldsymbol{\lambda_{i}^{A}} - \boldsymbol{\lambda_{j}^{S}})^{2}} \frac{1}{f'(\boldsymbol{\lambda_{j}^{S}})}. \end{split}$$

Thus, we have the relative proportion of the alignments as:

$$\frac{\langle \mathbf{u}_i^{\mathbf{A}}, \mathbf{u}_j^{\mathbf{S}} \rangle^2}{\langle \mathbf{u}_1, \mathbf{u}_j^{\mathbf{S}} \rangle^2} = \frac{\langle \mathbf{u}_i^{\mathbf{A}}, \mathbf{u}_1 \rangle^2}{(\lambda_1^{\mathbf{A}} - \lambda_j^{\mathbf{S}})^2} \frac{1}{f'(\lambda_j^{\mathbf{S}})} \lambda_1^2 f'(\lambda_j^{\mathbf{S}}) = \lambda_1^2 \frac{\langle \mathbf{u}_i^{\mathbf{A}}, \mathbf{u}_1 \rangle^2}{(\lambda_i^{\mathbf{A}} - \lambda_j^{\mathbf{S}})^2}.$$

Note that from Lemma F.1, for  $j \in [2, n] \setminus \{i - 1, i\}$ ,  $(\lambda_i^{\mathbf{A}} - \lambda_i^{\mathbf{S}})^2 = \Theta\left((\lambda_i^{\mathbf{A}} - \lambda_i^{\mathbf{A}})^2\right)$ . Thus

$$\frac{\langle \mathbf{u}_i^{\mathbf{A}}, \mathbf{u}_j^{\mathbf{S}} \rangle^2}{\langle \mathbf{u}_1, \mathbf{u}_j^{\mathbf{S}} \rangle^2} = \Theta\left(\lambda_1^2 \frac{\langle \mathbf{u}_i^{\mathbf{A}}, \mathbf{u}_1 \rangle^2}{(\lambda_i^{\mathbf{A}} - \lambda_j^{\mathbf{A}})^2}\right)$$

Due to the fact that  $\mathbf{u}_i^{\mathbf{A}}$  is an uniform random vector, it follows from Lemma E.6, that  $\mathbb{E}\left[\langle \mathbf{u}_i^{\mathbf{A}}, \mathbf{u}_1 \rangle^2\right] = d^{-1}$ . Moreover, under our assumptions,  $\lambda_1 = \Theta(1)$ . Using the fact that the eigenvectors of a Gaussian random matrix are independent of its eigenvalues, for  $i \in [k]$  and  $j \in [2, n] \setminus \{i - 1, i\}$ , we obtain

$$\mathbb{E}\left[\frac{\langle \mathbf{u}_i^{\mathbf{A}}, \mathbf{u}_j^{\mathbf{S}} \rangle^2}{\langle \mathbf{u}_1, \mathbf{u}_j^{\mathbf{S}} \rangle^2}\right] = \Theta\left(\frac{1}{d(\lambda_i^{\mathbf{A}} - \lambda_j^{\mathbf{A}})^2}\right) = \Theta\left(\frac{1}{d(\lambda_i^{\mathbf{A}} - \lambda_j^{\mathbf{S}})^2}\right).$$

**Remark F.4** (The case of j=1). Note that the above result also applies for j=1, since  $\lambda_j^{\mathbf{S}}$  converges almost surely to some constant. This can be proven utilizing the tools introduced in [7]. We omit the proof for the brevity.

**Remark F.5** (For j that Corresponds to Small Eigenvalues and Their Influence on Variance). As it was proved in Lemma F.1, the location of eigenvalues of S follows that of  $AA^{\top}$ . According to [36], the smallest eigenvalues of  $AA^{\top}$  scale  $O(n^{-2})$ , and these terms dominate the variance contribution to  $\|\mathbf{W}_c\|_F^2$ . For eigenvectors corresponding to these small eigenvalues, which scale at the same rate, the theorem becomes  $O(d^{-1})$ , since  $(\lambda_i^A - \lambda_j^S)^2 = O(1)$ . Moreover, as  $c \to 1$ , the number of such small eigenvalues increases, further amplifying their influence on the variance.

## **G** Additional Details on Numerical Results

## G.1 Data and Test Setting

We used the CIFAR-10 dataset [29] throughout the main text. Training and test data were sampled from disjoint splits. Each data point was reshaped into a 3072-dimensional vector. The number of test samples was fixed at 4500. Since the dataset has a fixed ambient dimension d, our numerical experiments focused on varying the number of training samples n. To generate Figure 1 and the left and center plots of Figure 3, we set the data rank to r=100 and the bottleneck dimension to k=50. To obtain low-rank representations, we performed singular value decomposition (SVD) and retained the top r components. All figures were produced using appropriately scaled data to ensure a signal-to-noise ratio of approximately  $\frac{\|\mathbf{X}\|_2}{\|\mathbf{A}\|_2} \approx 30$ , where  $\|\cdot\|_2$  denotes the operator norm.

Although computing the test error as described in Eq. (7) ideally requires multiple trials to reduce variance, we conducted only a single trial. The results demonstrated strong agreement with theoretical expectations, likely due to concentration effects. Our numerical experiments were conducted using a T4 GPU on Google Colab. Generating Figure 2 took approximately 5 hours, using a stride of 20 and starting from n=2568. Figure 3 required approximately 2 hours to compute.

## **G.2** Solutions from Existing Methods Used in Plots

To set the clear line from our denoising setting to other settings, in Figure 1 we generated generalization error curves for other models in the overparameterized setting. For this, we used the regularized expressions for the critical points from Section 2, and utilized the minimum-norm solutions. For the underparameterized solution depicted in Figure 3, we directly used the analytical solutions provided by [8], adapting them to our setting as needed.