Out-of-Distribution Generalization of In-Context Learning: A Low-Dimensional Subspace Perspective

Soo Min Kwon* University of Michigan

Alec S. Xu* University of Michigan

Can Yaras University of Michigan

Laura Balzano University of Michigan

Qing Qu University of Michigan

Abstract

The transformer's remarkable ability to perform in-context learning (ICL) has sparked a wide range of studies designed to understand its strengths and limitations. However, a theoretical understanding of when ICL can and cannot generalize beyond its pre-training data still remains unclear. This paper puts forth a minimal mathematical model that provably identifies when ICL can generalize out-of-distribution (OOD). By studying linear regression tasks parameterized with low-rank covariance matrices, we model distribution shifts as varying angles between subspaces and derive conditions under which a single-layer linear attention model interpolates across all angles. We show that if pre-training task vectors are drawn from a union of subspaces, transformers can generalize to all angle shifts—enabling ICL even in regions with zero probability mass in the training distribution. On the other hand, if the pre-training tasks are drawn from a single Gaussian, the test risk shows a non-negligible dependence on the angle, implying that ICL cannot generalize OOD. We empirically show that our results also hold for models such as GPT-2, and present experiments on how our results extend to nonlinear function classes.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Transformer-based large language models (LLMs) (Vaswani et al., 2017) have revolutionized natural language processing and driven significant progress across a wide range of domains, including logical reasoning (Wei et al., 2022b), sentiment classification (Chen et al., 2024; Wang et al., 2024c; Xu et al., 2024), machine translation (Vilar et al., 2022; Agrawal et al., 2023), and code generation (Li et al., 2023; Patel et al., 2024). Their success is largely attributed to scaling up model size, which has been shown to improve both performance and sample efficiency (Kaplan et al., 2020). Interestingly, large-scale transformers also exhibit emergent capabilities—abilities that arise only beyond a certain scale (Wei et al., 2022a). One striking example is in-context learning (ICL), where a model can perform a task simply by being prompted with a few input–output examples, without any gradient-based updates. This has sparked a variety of research aimed at understanding the underlying mechanism behind ICL, along with its strengths and limitations.

While there is an abundance of work on ICL, including studies on test-time training (Gozeten et al., 2025) and how specific prompts affect performance (Chang et al., 2025; Fang et al., 2025), the literature can be broadly divided into two categories: (i) empirical investigations into the extent to which ICL can solve novel tasks (Garg et al., 2022; Raventós et al., 2023; Yadlowsky et al., 2023; Wang et al., 2025; Ahuja and Lopez-Paz, 2023; Zhang et al., 2024; Li et al., 2024a; Pan et al., 2023; Kossen et al., 2024), and (ii) theoretical analyses of the mechanisms underlying ICL (Zhang et al., 2024; Huang et al., 2024; Li et al., 2024a; Akyürek et al., 2023; Von Oswald et al., 2023; Ahn et al., 2023; Li et al., 2024b). Interestingly, both lines of work share a common observation: ICL often generalizes out of distribution (OOD) beyond its training

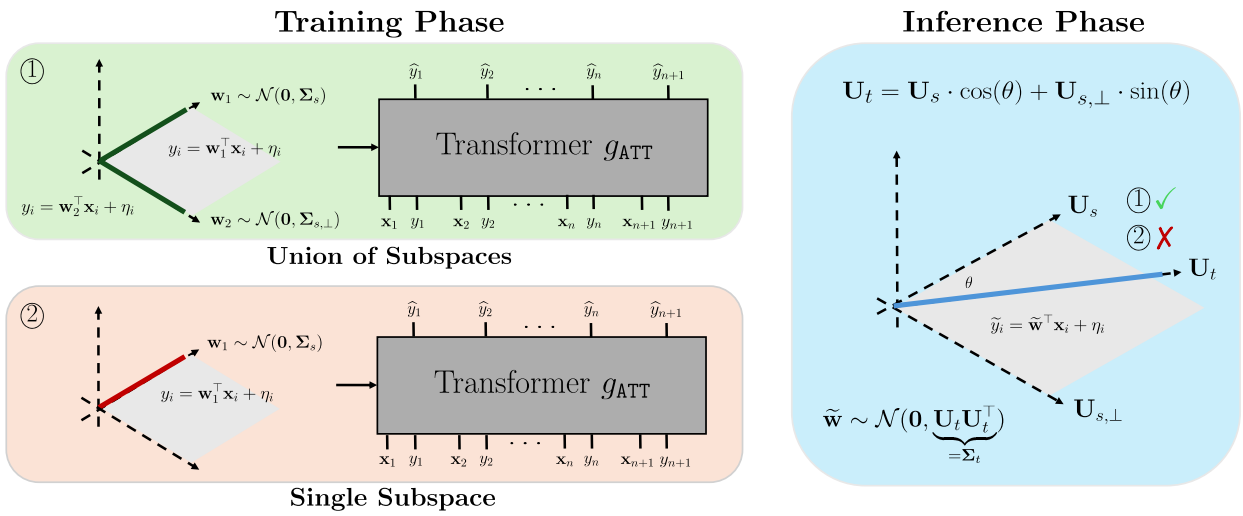

Figure 1: Overview of the setup for the pre-training task vectors in this paper. We consider two models: one with task vectors drawn from a union of subspaces and one trained with task vectors drawn from a single subspace. At inference, we test both models using a task vector at an angle between two subspaces. The union of subspace model generalizes across all angles, while the latter fails to generalize under distribution shifts.

data. For ICL, there are a few types of distribution shift: covariate shifts (i.e., shifts in the inputs), task-function shifts (i.e., shifts in the function that generates outputs), and query shifts (i.e., the query input differs from the inputs in the test prompts). Zhang et al. (2024) showed that while a single-layer linear attention model cannot tolerate input shifts, it can handle shifts in the regression weights (task vectors) well. Garg et al. (2022) reached a similar conclusion, empirically showing that ICL is relatively robust to distribution shifts in several settings, as transformer performance closely matched that of the least-squares estimator on linear regression tasks. However, Wang et al. (2025) recently challenged these views on language data, empirically demonstrating that ICL can generally solve only in-distribution tasks. To resolve these contrasting perspectives, the concurrent work of Goddard et al. (2025) investigates task-function shifts and attributes ICL OOD capabilities to pre-training task diversity. Yet, the complex nature of their definition of the task vector made theoretical analysis difficult, restricting their study to empirical results. Taken together, these works highlight the lack of a theoretical framework that clearly explains when ICL can and cannot generalize OOD.

In this work, we present a mathematical model to demystify and quantify the OOD generalization capabilities of ICL. We primarily focus on task distribution shifts by studying ICL in a single-layer linear attention model performing linear regression, where the weight (or task) vectors are sampled from low-dimensional subspaces. This allows us to quantify the distribution shift in the task vector via the principal angles between subspaces, and characterize the OOD test risk as a function of these angles (see Figure 1 for an illustration). Unlike Goddard et al. (2025), we then provide theoretical guarantees by proving conditions on the pre-training task vectors under which OOD generalization is possible. Furthermore, we empirically show that our findings extend beyond linear settings, in that (i) our results hold for nonlinear transformers such as GPT-2; and (ii) our results apply to nonlinear function classes. Overall, our contributions can be summarized as follows:

- **OOD Generalization When Trained on a Union of Subspaces.** We prove that when the pre-training task vectors are sampled from a union of subspaces, a linear attention model can generalize across all principal angles between the training subspaces. This result implies that ICL can generalize to subspaces within the span of the training subspaces, including regions with zero probability density under the training distribution. We hypothesize that this explains the apparent OOD capabilities of ICL observed in the literature: the testing task vector lies within the span of the training tasks.

- **No OOD Generalization When Trained on a Single Subspace.** On the other hand, we prove that when the training task vectors are drawn from a single r-dimensional subspace, ICL applied to a task vector whose subspace is shifted away from the training subspace by an angle θ incurs a test risk that depends on θ. This implies that ICL cannot

generalize beyond the span of the training subspace.

Consequently, our results complement those of Goddard et al. (2025) and advocate for task diversity in the pre-training data, since training on a union of subspaces can be viewed as a form of task diversity, with each subspace spanning different directions. Given that large-scale transformers are typically trained on vast corpora and thus exposed to diverse pre-training data, we attribute their OOD generalization in ICL to this diversity.

2 PROBLEM SETUP

In this section, we present the basic setup for ICL, together with a motivating example that illustrates how OOD generalization arises depending on the pre-training data. We then describe the analysis setup, which involves both linear attention and linear regression task vectors.

2.1 Preliminaries

Notation. We denote scalars with unbolded letters (e.g., m, M), vectors with bold lower-case letters (e.g., $\mathbf{x}$) and matrices with bold upper-case letters (e.g., $\mathbf{X}$). We use $\mathbf{I}_n$ to denote an identity matrix of size $n \in \mathbb{N}$. We use $\mathcal{R}(\mathbf{X})$ to denote the range or the column space of the matrix $\mathbf{X}$. Lastly, given any $n \in \mathbb{N}$, we use $[n]$ to denote the index set $\{1, \dots, n\}$.

ICL Setup. Given a sequence of n input-output example pairs $\{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, the objective of ICL is to predict the output $y_{n+1} \in \mathbb{R}$ corresponding to an unseen query $\mathbf{x}_{n+1} \in \mathbb{R}^d$. Following prior works (Garg et al., 2022), we assume each output is generated via $y_i = f(\mathbf{x}_i)$ for some function $f(\cdot)$, where $f \in \mathcal{F}$ is sampled from a distribution over a function class $\mathcal{F}$. By convention, a transformer takes in these $n+1$ pairs as an input prompt $\mathbf{Z} \in \mathbb{R}^{(n+1)\times(d+1)}$ constructed in the following form:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \dots & \mathbf{z}_n & \mathbf{z}_{n+1} \end{bmatrix}^\top = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_n & \mathbf{x}_{n+1} \\ y_1 & \dots & y_n & 0 \end{bmatrix}^\top,$$

where $\mathbf{z}_i := \begin{bmatrix} \mathbf{x}_i^\top & y_i \end{bmatrix}^\top$ and $\mathbf{z}_{n+1} := \begin{bmatrix} \mathbf{x}_{n+1}^\top & 0 \end{bmatrix}^\top$.

Then, a transformer $g_{\texttt{ATT}}$, parameterized by weights $\mathcal{W}$, takes these prompts as input and is trained by minimizing the following expected squared loss with respect to $\mathcal{W}$:

$$\min_{\mathcal{W}} \mathcal{L}_{\texttt{ATT}}(\mathcal{W}) := \mathbb{E}\left[(y_{n+1} - g_{\texttt{ATT}}(\mathbf{z}_{n+1}, \mathbf{Z}))^2\right]. \quad (1)$$

During inference time, we test the trained model, denoted as $g^\star_{\texttt{ATT}}$, using $m+1$ paired examples $\{\mathbf{x}_j, \widetilde{y}_j\}_{j=1}^{m+1}$.

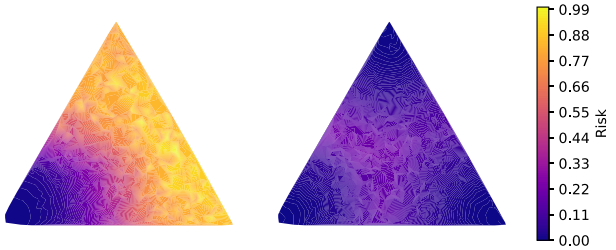

Figure 2: Visualization of the generalization behavior of transformers for learning cosine functions in-context. Each corner of a triangle represents a one-dimensional subspace spanned by ψ_1^C (bottom left), ψ_2^C (bottom right), or ψ_3^C (top), with all possible convex combinations given by the interior. Left: Test risk when trained on $\text{span}(\{\psi_1^C\})$. Right: Test risk when trained on $\text{span}(\{\psi_1^C\}) \cup \text{span}(\{\psi_2^C\}) \cup \text{span}(\{\psi_3^C\})$.

The input prompts are constructed in the same manner:

$$\widetilde{\mathbf{Z}} = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_n & \mathbf{x}_{n+1} \\ \widetilde{y}_1 & \dots & \widetilde{y}_n & 0 \end{bmatrix}^\top \quad \text{and} \quad \widetilde{\mathbf{z}}_{m+1} = \begin{bmatrix} \mathbf{x}_{m+1} \\ 0 \end{bmatrix}.$$

In this work, we are interested in ICL's OOD generalization abilities under distribution shifts in the underlying task function, i.e., the labels are generated via $\widetilde{y}_j = \widetilde{f}(\mathbf{x}_j)$, where $\widetilde{f} \in \widetilde{\mathcal{F}} \neq \mathcal{F}$.

2.2 Case Study: When Can In-Context Learning Generalize Out-of-Distribution?

Now, we present a case study on how the pre-training distribution affects OOD generalization of ICL. Specifically, we train a GPT-2 model[1] by constructing input prompts with inputs $x \sim \mathcal{U}([0,1])$ (and hence $d = 1$) and generate labels via cosine bases, which model rich sets of signals in real-world data:

$$\psi_k^C(x) = \frac{1}{\sqrt{2}} \cos(k\pi x) \quad \text{for } k \in \mathbb{N}.$$

In Figure 2 (left), we show the test risk when training on prompts drawn from $\text{span}(\{\psi_1^C\})$ and testing on prompts drawn from $\text{span}(\{\psi_1^C, \psi_2^C, \psi_3^C\})$. As the test prompts shift away from the pre-training distribution $\text{span}(\{\psi_1^C\})$, the test risk increases, indicating that transformers are not robust to distribution shifts in this case. This raises the question: under what conditions on the pre-training data does generalization to such regions become possible?

In Figure 2 (right), we present the test risk when training on the union of rank-one subspaces $\text{span}(\{\psi_1^C\}) \cup \text{span}(\{\psi_2^C\}) \cup \text{span}(\{\psi_3^C\})$, i.e., we sample $k \in \{1, 2, 3\}$

[1] We defer the experimental details to Section 4.

uniformly and draw the prompt from span($\{\psi_k^C\}$). In this case, the test risk is negligible in all regions, even in those with zero probability density under the pretraining distribution. This implies transformers generalizes to OOD tasks in this setting. We hypothesize that this may explain why ICL achieves OOD generalization: the test data actually lies within the span of the training data. These observations motivate us to formalize this notion with a tractable analytical setting. To facilitate such analysis, we study a single-layer linear attention model and linear regression data.

2.3 Single-Layer Linear Attention

We now introduce the linear attention architecture. Ahn et al. (2024) empirically demonstrated that many phenomena observed in vanilla transformers can also be replicated in transformers with linear attention. These results motivated subsequent works (Ahn et al., 2023; Zhang et al., 2024; Li et al., 2024b) to adopt linear attention as a testbed for studying ICL, which we also adopt for analysis.

For linear attention, we employ a causal mask to the input prompts to ensure that they cannot attend to their own labels (Ahn et al., 2023; Mahankali et al., 2024):

$$\mathbf{Z}_{\mathcal{M}} = \begin{bmatrix}\mathbf{z}_1 & \dots & \mathbf{z}_n & \mathbf{0}\end{bmatrix}^\top \text{ and } \mathbf{z}_{n+1} = \begin{bmatrix}\mathbf{x}_{n+1} \\ 0\end{bmatrix}.$$

Given the input sequence $\mathbf{Z}_{\mathcal{M}}$ and query $\mathbf{z}_{n+1}$, a single-layer linear attention model sets $g_{\texttt{ATT}}$ as follows to make the prediction $\widehat{y}_{n+1}$:

$$\widehat{y}_{n+1} = g_{\texttt{ATT}}(\mathbf{z}_{n+1}, \mathbf{Z}_{\mathcal{M}}) \tag{2}$$

$$= \frac{1}{n}\left(\mathbf{z}_{n+1}^\top \mathbf{W}_Q \mathbf{W}_K^\top \mathbf{Z}_{\mathcal{M}}^\top\right) \mathbf{Z}_{\mathcal{M}} \mathbf{W}_V \mathbf{p}, \tag{3}$$

where $\mathbf{p} = [\mathbf{0}_d \ \ 1]^\top$ and $\mathbf{W}_K, \mathbf{W}_Q, \mathbf{W}_V \in \mathbb{R}^{(d+1)\times(d+1)}$ are the key, query, and value weight matrices, respectively. This sets $\mathcal{W} = \{\mathbf{W}_K, \mathbf{W}_Q, \mathbf{W}_V\}$ as the collection of trainable weights corresponding to the linear attention model. Then, let $\mathcal{W}^\star = \{\mathbf{W}_K^\star, \mathbf{W}_Q^\star, \mathbf{W}_V^\star\}$ be the optimal weights obtained by minimizing the loss in Equation (1). During inference time, we test the optimal linear attention model $g_{\texttt{ATT}}^\star$ using the $m+1$ paired examples:

$$\widehat{y}_{m+1} = g_{\texttt{ATT}}^\star(\widetilde{\mathbf{z}}_{m+1}, \widetilde{\mathbf{Z}}_{\mathcal{M}})$$

$$= \frac{1}{m}\left(\widetilde{\mathbf{z}}_{m+1}^\top \mathbf{W}_Q^\star \mathbf{W}_K^{\star\top} \widetilde{\mathbf{Z}}_{\mathcal{M}}^\top\right) \widetilde{\mathbf{Z}}_{\mathcal{M}} \mathbf{W}_V^\star \mathbf{p}.$$

Notice at inference time, we normalize by a factor of m instead of n.

2.4 ICL with Linear Regression

The most commonly studied ICL task is linear regression, which specifies $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ for some input $\mathbf{x} \in \mathbb{R}^d$ and weight (or task) vector $\mathbf{w} \in \mathbb{R}^d$. In this work, we also adopt the linear regression setting, but assign particular forms to the task vectors to mimic the setup of Section 2.2 using low-dimensional subspaces. Below, we specify the corresponding training and testing distributions.

Training Distribution. Recall that in Figure 2 (left), we trained on prompts drawn from a single cosine basis, while in Figure 2 (right), we trained on prompts drawn from a union of cosine bases. To this end, suppose that $d \geq 2r$ and let $\mathbf{U}_s \in \mathbb{R}^{d\times r}$ and $\mathbf{U}_{s,\perp} \in \mathbb{R}^{d\times r}$ be two r-dimensional orthonormal bases in $\mathbb{R}^d$. Consider the following two covariance matrices:

$$\mathbf{\Sigma}_s = \mathbf{U}_s \mathbf{U}_s^\top + \epsilon \cdot \mathbf{I}_d$$
$$\mathbf{\Sigma}_{s,\perp} = \mathbf{U}_{s,\perp} \mathbf{U}_{s,\perp}^\top + \epsilon \cdot \mathbf{I}_d,$$

where $\epsilon > 0$ is a small constant included to ensure a non-degenerate distribution. For training, we consider two separate models trained on two different task vector distributions. Specifically, each feature and label pair $(\mathbf{x}_i, y_i)$ is generated as follows. For all $i \in [n+1]$, let $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and

$$y_i = \mathbf{w}^\top \mathbf{x}_i + \eta_i, \tag{4}$$

where $\eta_i \sim \mathcal{N}(0, \sigma^2)$ is iid noise with variance σ^2. Finally, we consider the following distributions for the task vector $\mathbf{w}$:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_s), \qquad \text{(Single Subspace)}$$

or

$$\mathbf{w} \sim \begin{cases}\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_s) & \text{w.p. } \gamma, \\ \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{s,\perp}) & \text{w.p. } 1-\gamma,\end{cases} \qquad \text{(Union of Subs.)}$$

where $0 < \gamma < 1$ denotes the mixture probability. Under these two distributions, we show that the trained models exhibit different OOD generalization behaviors with respect to the testing distribution defined in the following section. Finally, in the union of subspaces distribution, while we focus on $K = 2$ bases for ease of exposition, we also generalize our results to consider a union of $K > 2$ bases.

Testing Distribution. Recall that in Section 2.2, we tested the transformer with all possible convex combinations of the cosine bases. Similarly, we aim to define a testing subspace $\mathbf{U}_t \in \mathbb{R}^{d\times r}$ such that it represents a region between the training subspaces. To this end, we parameterize $\mathbf{U}_t$ as such (Absil et al., 2004, Section 3.8):

$$\mathbf{U}_t = \mathbf{U}_s \cdot \cos(\mathbf{\Theta}) + \mathbf{U}_{s,\perp} \cdot \sin(\mathbf{\Theta}), \tag{5}$$

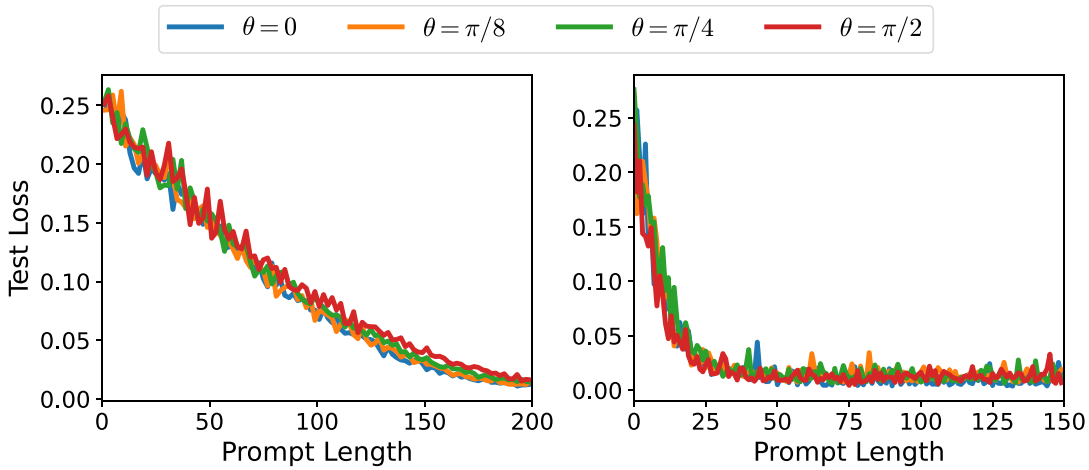

Figure 3: Plot of the test risk as a function of the prompt length for a linear transformer (left) and a GPT-2 model (right). When the prompt length at test time is large enough, the test risk goes nearly to zero for all $\theta \in \left[0, \frac{\pi}{2}\right]$, corroborating Theorem 1 in that both linear and nonlinear transformers can generalize to the span of the training task vectors at test-time.

where $\mathbf{\Theta} \in \mathbb{R}^{r \times r}$ is a diagonal matrix with elements $\theta_i \in \left[0, \frac{\pi}{2}\right]$ for all $i \in [r]$, while $\cos(\cdot)$ and $\sin(\cdot)$ are only applied to the diagonal elements of $\mathbf{\Theta}$. Here, θ_i represents the i-th principal angle between $\mathbf{U}_s$ and $\mathbf{U}_t$. For simplicity, we will assume all principal angles are equal, i.e., for all $i \in [r]$, $\theta_i = \theta$ for some $\theta \in \left[0, \frac{\pi}{2}\right]$ so that $\mathbf{\Theta} = \theta \cdot \mathbf{I}_r$. Notice when $\theta = 0$, $\mathbf{U}_t = \mathbf{U}_s$, and when $\theta = \frac{\pi}{2}$, $\mathbf{U}_t = \mathbf{U}_{s,\perp}$. Hence, for varying values of θ, $\mathbf{U}_t$ can be viewed as a "slice" between $\mathbf{U}_s$ and $\mathbf{U}_{s,\perp}$. Finally, we parameterize $\mathbf{\Sigma}_t$ as

$$\mathbf{\Sigma}_t = \mathbf{U}_t \mathbf{U}_t^\top + \epsilon \cdot \mathbf{I}_d, \tag{6}$$

and generate each testing pair $(\mathbf{x}_j, \widetilde{y}_j)$ independent of the training data in a similar fashion: for all $j \in [m+1]$, let $\mathbf{x}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and

$$\widetilde{y}_j = \widetilde{\mathbf{w}}^\top \mathbf{x}_j + \eta_j, \quad \text{where} \quad \widetilde{\mathbf{w}} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t) \tag{7}$$

and $\eta_j \sim \mathcal{N}(0, \sigma^2)$ is again iid noise.

3 MAIN RESULTS

3.1 Transformers Can Generalize to the Span When Trained on a Union of Subspaces

In this section, we derive the test risk of the optimal linear attention model trained with prompts whose task vectors are drawn from a union of two subspaces, and tested on a task vector from subspace $\mathbf{U}_t$ at any angle $\theta \in [0, \frac{\pi}{2}]$. The following result shows that for sufficiently large prompt lengths, the test risk can be arbitrarily close to the optimal test risk (the label noise variance) independent of θ. This implies that linear attention is robust to such subspace shifts in this setting.

Theorem 1 (Test Risk Under a Union of Subspaces). *Let $g^\star_{ATT}$ denote the optimal linear attention model corresponding to the independent data setting in Equation (4), where the task vector is drawn from Equation (Union of Subs.) with $\gamma = 0.5$. For all $j \in [m+1]$, suppose that the prompts at test time are constructed with features $\mathbf{x}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and labels whose task vectors are drawn from Equation (6). For any $\theta \in \left[0, \frac{\pi}{2}\right]$ and $\delta \in (0, r)$, if*

$$m \geq n > \frac{\left(2(r+\sigma^2)+1\right) r}{\delta} - \left(2(r+\sigma^2)+1\right),$$

then we have

$$\lim_{\epsilon \to 0} \mathbb{E}\left[\left(\widetilde{y}_{m+1} - g^\star_{ATT}(\widetilde{\mathbf{z}}_{m+1}, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] < \sigma^2 + \delta.$$

In Theorem 1, we take $\epsilon \to 0$ for two reasons: (i) to eliminate any dependence on ϵ and isolate its effect on test risk as it is assumed to be a small constant, and (ii) analyze the test risk when the covariance matrices are exactly low-rank. In this limit, task vectors sampled from the test subspace $\mathbf{U}_t$ have zero probability density in the training task distribution. Hence, Theorem 1 implies ICL generalizes well to this OOD task. We hypothesize this can explain why ICL achieves OOD generalization: the test data actually lies within the span of the training data. In Figure 3, we corroborate this theory on both linear attention and a GPT-2 model, showing that for sufficiently large prompt lengths, the test risk indeed tends to zero. Experimental details are deferred to Section 4.

We provide some intuition behind the proof of Theorem 1, deferring the full proof to Section B.1.2. At the optimal linear attention weights under the loss in Equation (1), the model reduces to a single step of projected gradient descent (PGD) (Li et al., 2024b; Ahn et al., 2023; Von Oswald et al., 2023; Mahankali et al., 2024). Denoting $\mathbf{A} \in \mathbb{R}^{d\times d}$ as the PGD projection matrix that arises from the optimal weights, we sketch how the dependence on θ is mitigated (assuming $\sigma = 0$ and $\delta \to 0$ for simplicity):

$$\begin{aligned}\widehat{y}_{m+1} = g^{\star}_{\texttt{ATT}}(\widetilde{\mathbf{z}}_{m+1}, \widetilde{\mathbf{Z}}_{\mathcal{M}}) &= \frac{1}{m}\mathbf{x}_{m+1}^{\top}\mathbf{A}\mathbf{X}^{\top}\underbrace{\mathbf{X}\widetilde{\mathbf{w}}}_{=\mathbf{y}} \\ &\to \mathbf{x}_{m+1}^{\top}\mathbf{U}_{2r}\mathbf{U}_{2r}^{\top}\widetilde{\mathbf{w}} \qquad (m, n \to \infty \text{ and } \epsilon \to 0) \\ &= \mathbf{x}_{m+1}^{\top}\mathbf{U}_{2r}\mathbf{U}_{2r}^{\top}\mathbf{U}_t\mathbf{g}, \qquad (\widetilde{\mathbf{w}} = \mathbf{U}_t\mathbf{g})\end{aligned}$$

where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r)$, and we have defined the following:

$$\begin{aligned}\mathbf{X} &\coloneqq \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_m\end{bmatrix}^{\top} \in \mathbb{R}^{m\times d} \\ \mathbf{y} &\coloneqq \begin{bmatrix}\widetilde{y}_1 & \dots & \widetilde{y}_m\end{bmatrix}^{\top} \in \mathbb{R}^{m} \\ \mathbf{U}_{2r} &\coloneqq \begin{bmatrix}\mathbf{U}_s & \mathbf{U}_{s,\perp}\end{bmatrix} \in \mathbb{R}^{d\times 2r}.\end{aligned}$$

Since $\mathcal{R}(\mathbf{U}_t) \subset \mathcal{R}(\mathbf{U}_{2r})$ for all $\theta \in \left[0, \frac{\pi}{2}\right]$, the trained model perfectly recovers $\widetilde{y}_{m+1}$. Lastly, we set $\gamma = 0.5$ to simplify constants in the test risk, though our results hold for any $\gamma \in (0, 1)$.

Beyond a Mixture of Two Gaussians. We now generalize the above result to consider a mixture of $K > 2$ Gaussians. This model has been extensively studied in machine learning (Vidal, 2011) and has recently gained attention in deep learning (Wang et al., 2024b; Xu et al., 2025; Wang et al., 2024a). Specifically, assume $d \geq Kr$, and let $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_d\end{bmatrix} \in \mathbb{R}^{d\times d}$ be an orthonormal basis for $\mathbb{R}^d$. For all $k \in [K]$, we define $\mathbf{U}_{s,k} = \begin{bmatrix}\mathbf{u}_{(k-1)\cdot r+1} & \dots & \mathbf{u}_{kr}\end{bmatrix} \in \mathbb{R}^{d\times r}$. Note that $\mathbf{U}_{s,k}^{\top}\mathbf{U}_{s,l} = \mathbf{0}_{r\times r}$ for all $k \neq l$. Then, we assume the training task $\mathbf{w} \in \mathbb{R}^d$ is sampled as such:

$$\mathbf{w} \sim \sum_{k=1}^{K} \gamma_k \cdot \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{s,k}), \tag{8}$$

where $\mathbf{\Sigma}_{s,k} = \mathbf{U}_{s,k}\mathbf{U}_{s,k}^{\top} + \epsilon \cdot \mathbf{I}_d$ and $\sum_{k=1}^{K} \gamma_k = 1$. At inference time, we define an orthonormal basis $\overline{\mathbf{U}}_t \in \mathbb{R}^{d\times r}$ that lies within the span of $\{\mathbf{U}_{s,k}\}_{k=1}^{K}$:

$$\overline{\mathbf{U}}_t = \sum_{k=1}^{K} \alpha_k \mathbf{U}_{s,k}, \quad \text{for } \{\alpha_k\}_{k=1}^{K} \text{ s.t. } \sum_{k=1}^{K} \alpha_k^2 = 1, \tag{9}$$

where the constraint on $\{\alpha_k\}_{k=1}^{K}$ ensures $\overline{\mathbf{U}}_t \in \mathbb{R}^{d\times r}$ is an orthonormal basis. Then, similar to Theorem 1, we consider testing on task vectors $\widetilde{\mathbf{w}} \sim \mathcal{N}\left(\mathbf{0}, \overline{\mathbf{\Sigma}}_t\right)$ with $\overline{\mathbf{\Sigma}}_t = \overline{\mathbf{U}}_t\overline{\mathbf{U}}_t^{\top} + \epsilon \cdot \mathbf{I}_d$. We again emphasize $\overline{\mathbf{U}}_t$ is unseen during training, but lies within the span of the training subspaces.

Theorem 2. *Let $g^{\star}_{ATT}$ denote the optimal linear attention model corresponding to the independent data setting in Equation (4), where the task vector is drawn from Equation (8) with $\gamma_k = \frac{1}{K}$ for all $k \in [K]$. For all $j \in [m+1]$, suppose that the test prompts are constructed with features $\mathbf{x}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and labels whose task vectors are constructed with subspaces defined in Equation (9). For any $\{\alpha_k\}_{k=1}^{K}$ s.t. $\sum_{k=1}^{K} \alpha_k^2 = 1$ and $\delta \in (0, r)$, if*

$$m \geq n > \frac{(K(r+\sigma^2)+1)r}{\delta} - (K(r+\sigma^2)+1),$$

then we have

$$\lim_{\epsilon\to 0} \mathbb{E}\left[\left(\widetilde{y} - g^{\star}_{ATT}(\widetilde{\mathbf{z}}_{m+1}, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] < \sigma^2 + \delta.$$

The proof is deferred to Section B.1.2. Similar to Theorem 1, if the linear attention model is trained on task vectors that lie in a union of K subspaces, it can generalize well to any region within the span of the K subspaces, even if those regions have zero probability density during training. Lastly, note that by setting $K = 2$, $\alpha_1 = \cos(\theta)$, and $\alpha_2 = \sin(\theta)$, we exactly recover Theorem 1.

3.2 Transformers Cannot Generalize When Trained on a Single Subspace

Previously, we saw that task vectors drawn from a union of subspaces—which can be viewed as a form of task diversity—can enable OOD generalization. Similar to Section 2.2, we now aim to identify a setting in which generalization beyond the pre-training data is not possible. The following result shows that if the task vectors are drawn from a Gaussian whose covariance matrix spans only a single subspace $\mathbf{U}_s \in \mathbb{R}^{d\times r}$, then testing a linear attention model on a task vector shifted away from the training subspace by an angle θ yields a test risk with a non-negligible dependence on θ, even as the prompt length tends to infinity.

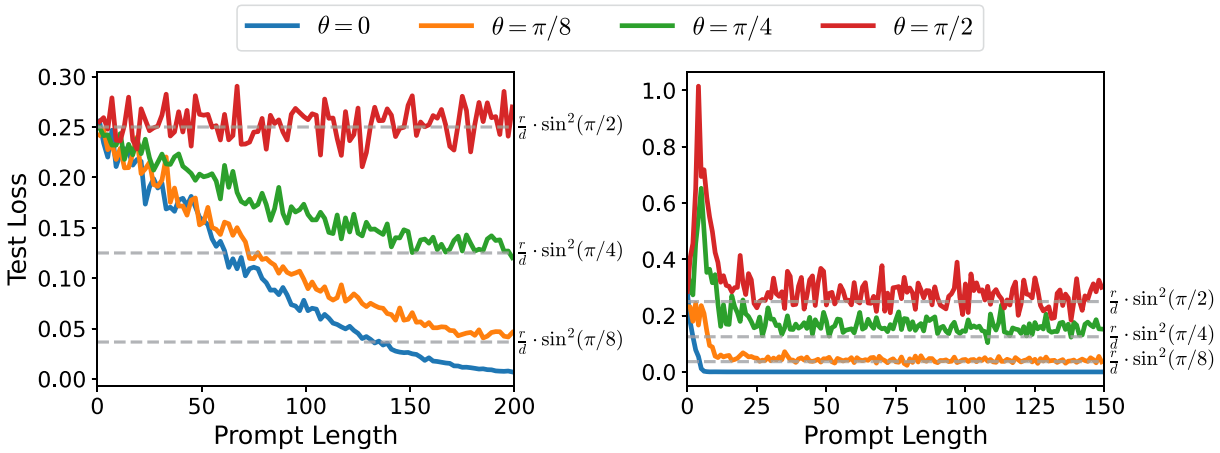

Figure 4: Plot of the normalized test risk as a function of the prompt length for a linear transformer (left) and a GPT-2 model (right) under covariance shifts. As the covariance at test time shifts away from the covariance used at training time as a function of θ, the test risk exhibits a non-negligible dependence on θ for both the linear and nonlinear transformer. Moreover, for both models, the test risk exactly matches the predicted risk from Proposition 1.

Proposition 1 (Test Risk Under a Single Subspace). *Let $g^\star_{ATT}$ denote the optimal linear attention model corresponding to the independent data setting in Equation (4), where the task vector is drawn from Equation (Single Subspace). For all $j \in [m+1]$, suppose that the prompts at test time are constructed with features $\mathbf{x}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and labels whose task vectors are drawn from Equation (6). Then, we have*

$$\lim_{m\to\infty} \lim_{n\to\infty} \lim_{\epsilon\to 0} \mathbb{E}\left[\left(\widetilde{y}_{m+1} - g^\star_{ATT}(\widetilde{\mathbf{z}}_{m+1}, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = r\sin^2(\theta) + \sigma^2,$$

where $\theta \in [0, \frac{\pi}{2}]$ are the r principal angles between $\mathbf{U}_s \in \mathbb{R}^{d\times r}$ and $\mathbf{U}_t \in \mathbb{R}^{d\times r}$.

The proof is provided in Appendix B.1.3. In the asymptotic regime, our result reveals the following: when $\theta = 0$, the $\sin(\cdot)$ term vanishes, allowing perfect recovery up to the label noise variance. However, as θ increases from 0 to $\frac{\pi}{2}$, the test risk increases with respect to θ. At $\theta = \frac{\pi}{2}$, the test risk becomes exactly the rank of the covariance matrix. Notably, this represents the largest possible error in this setting, as a low-rank covariance matrix induces an error dependent on the rank rather than the ambient dimension, as observed in related work (Garg et al., 2022; Oko et al., 2025).

The proof technique is similar to that of Theorem 1:

$$\begin{aligned}\widehat{y}_{m+1} &= g^\star_{\mathtt{ATT}}(\widetilde{\mathbf{z}}_{m+1}, \widetilde{\mathbf{Z}}_{\mathcal{M}}) \\ &\to \mathbf{x}_{m+1}^\top \mathbf{U}_s \mathbf{U}_s^\top \widetilde{\mathbf{w}} \qquad (m, n \to \infty \text{ and } \epsilon \to 0) \\ &= \mathbf{x}_{m+1}^\top \mathbf{U}_s \mathbf{U}_s^\top \mathbf{U}_t \mathbf{g}, \qquad (\widetilde{\mathbf{w}} = \mathbf{U}_t \mathbf{g})\end{aligned}$$

By taking appropriate limits, it is easy to see that the dependence on θ arises from $\mathbf{U}_s^\top \mathbf{U}_t$, which reflects a rotation by an angle θ between the subspaces. Since $\mathbf{A} \to \mathbf{U}_s \mathbf{U}_s^\top$ in the asymptotic regime, PGD projects the data onto an "incorrect" subspace, thereby inducing an error proportional to θ in the test risk.

In Figure 4, we present experiments corroborating Proposition 1 on linear attention and a GPT-2 model. Interestingly, our experiments show that both models incur the same test risk under the distribution shift when given enough in-context examples. This implies that the linear attention model can adequately capture the behavior in this setting, and that the observed error is not merely an artifact of using a linear model. Lastly, we assumed equal principal angles between the subspaces for simplicity, and defer the more general result to Proposition 2 in Appendix A.1.

4 EXPERIMENTAL RESULTS

Experimental Setup. Unless otherwise stated, the experimental setup is as follows: for both the linear and nonlinear Transformer, we consider the noiseless case, and set $d = 20$, $r = 5$, and $\epsilon = 10^{-5}$. To construct the train and test subspaces, we sample an orthogonal matrix $\mathbf{U} \in \mathbb{R}^{d\times d}$ uniformly at random, set $\mathbf{U}_s$ to be the first r columns of $\mathbf{U}$, and set $\mathbf{U}_{s,\perp}$ to

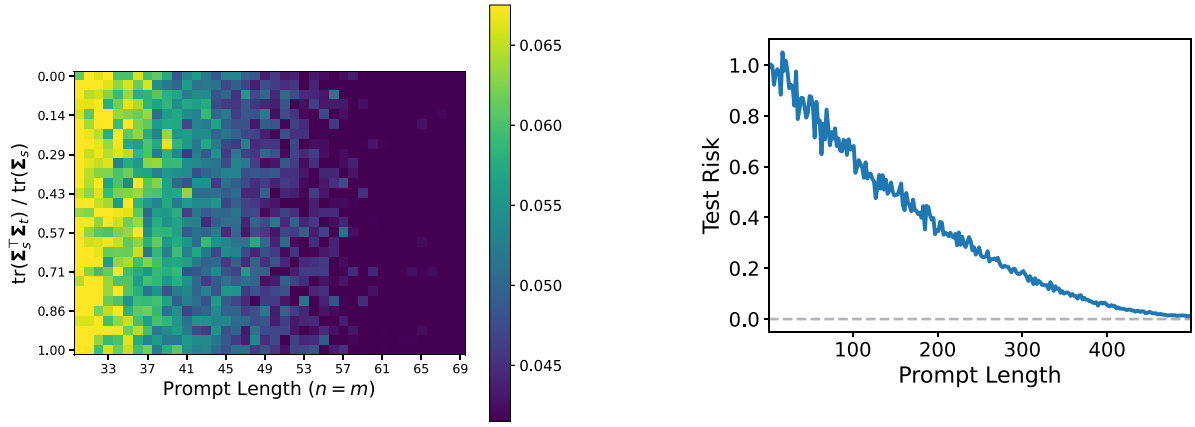

Figure 5: Left: Phase plot of the test risk as we vary the angle between $\mathbf{\Sigma}_s$ and $\mathbf{\Sigma}_t$ and the prompt length with $m = n$ for a linear attention model trained with a mixture of Gaussians. The test risk is low across all angle shifts, and decreases further as the prompt length increases. Right: Plot of the test risk as a function of the prompt length for a case in which $\mathbf{\Sigma}_s \neq \mathbf{\Sigma}_t$ but with $\theta = 0$, following the OOD example in Gatmiry et al. (2024). This serves to explain why ICL can seemingly do OOD generalization as observed in the literature.

be the second r columns. Given this setup, we typically consider a mixture of $K = 2$ subspaces for the experiments.

For the experiments with the linear transformer, we plug in the optimal weights according to their respective settings (e.g., optimal weights using a single subspace or a mixture of subspaces) and set $m = n = 200$. For the nonlinear transformer, following Garg et al. (2022), we use a small GPT-2 model with 6 layers, 4 heads, and a 128-dimensional embedding space. We append a learnable linear transformation to map the vector predicted by the model to a scalar. We use a learning rate of $\eta = 10^{-4}$, batch size 128, prompt lengths $m = n = 150$, and train for 100K iterations. We run all experiments on a single A100 GPU.

4.1 More Results on Linear Function Classes

Previously, we presented results on the test risk as a function of the prompt length. In Figure 5 (left), we present a phase plot of the test risk as a function of $\mathrm{Tr}(\mathbf{\Sigma}_s^\top \mathbf{\Sigma}_t) / \mathrm{Tr}(\mathbf{\Sigma}_s)$ (which measures the angle between two covariance matrices) and the prompt length on linear attention with task vectors drawn from a mixture of two Gaussians. Similar to Figure 3, the test risk is low for all values of $m = n$, and it decreases further as the prompt length increases. Note that the largest possible normalized test risk in this setting is $r/d = 0.25$, so the test risk is still considered low even when the prompt length is small.

In Section 3, we discussed how apparent abilities of ICL to perform OOD generalization arises when the test task lies within the span of the training task vectors. Here, we present an extra experiment to support this claim, using the example from Gatmiry et al. (2024), with $\mathbf{\Sigma}_s = \mathbf{I}_5$ and $\mathbf{\Sigma}_t = \mathbf{V}\mathbf{\Lambda}_t\mathbf{V}^\top$, where $\mathbf{V} \in \mathbb{R}^{5\times 5}$ is a random orthogonal matrix and $\mathbf{\Lambda}_t = \mathrm{Diag}(1, 1, 1/2, 1/4, 1)$. In Figure 5 (right), we observe that the test risk approaches zero given enough samples. This implies that our result may help explain many observations of OOD generalization in ICL and offers a unifying perspective on findings reported in the literature.

4.2 More Results on Nonlinear Function Classes

In this section, we present an additional result using a nonlinear function class to complement both our theory and Section 2.2. We look at the function space $L^2(\mathbb{R}, e^{-x^2/2}/\sqrt{2\pi}\, dx)$, i.e., square-integrable functions under the Gaussian measure, and construct an orthonormal basis via Hermite polynomials:

$$\psi_k^H(x) = \frac{(-1)^k}{\sqrt{k!}} e^{x^2/2} \frac{d^k(e^{-x^2/2})}{dx^k} \quad \text{for } k \in \mathbb{N}.$$

In Figure 6, we observe the same pattern as in Figure 2: training on a single polynomial does not allow OOD generalization beyond that basis, whereas training on a union of bases enables generalization to all points in the interior. This demonstrates that our theory and setup extend beyond linear settings and hold with greater generality.

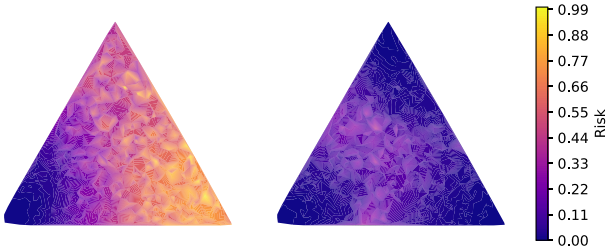

Figure 6: Visualization of the generalization behavior of transformers for learning Hermite polynomials functions in-context. Similar to Figure 2, each corner of a triangle represents a one-dimensional subspace spanned by ψ_k^H. Left: Test risk when trained on $\text{span}(\{\psi_1^H\})$. Right: Test risk when trained on $\text{span}(\{\psi_1^H\}) \cup \text{span}(\{\psi_2^H\}) \cup \text{span}(\{\psi_3^H\})$.

5 CONCLUSION

In this work, we proposed a mathematical framework to analyze the OOD capabilities of ICL. Unlike prior studies, our framework provides theoretical guarantees for when linear transformers can or cannot generalize OOD, based on their pre-training data, which we parameterize using low-dimensional subspaces. We show that pre-training task diversity, defined as a union of subspaces, enables generalization to regions with zero probability density in the training distribution, whereas tasks drawn from a single subspace do not have this property. We further demonstrate that our theoretical results extend empirically to transformer models such as GPT-2, and we present experiments showing that the findings also apply to nonlinear function classes.

Acknowledgments

QQ, SK, AX, and CY acknowledge NSF CAREER CCF-2143904, NSF IIS 2312842, NSF IIS 2402950, and ONR N00014-22-1-2529. LB, CY, and SK acknowledge NSF CAREER CCF-1845076 and NSF CCF-2331590. We would like to thank Emrullah Ildiz (University of Michigan), Samet Oymak (University of Michigan), and Daniel Hsu (Columbia University) for fruitful discussions.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. (2004). Riemannian geometry of grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematica*, 80(2):199–220.

Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., and Ghazvininejad, M. (2023). In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873.

Ahn, K., Cheng, X., Daneshmand, H., and Sra, S. (2023). Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650.

Ahn, K., Cheng, X., Song, M., Yun, C., Jadbabaie, A., and Sra, S. (2024). Linear attention is (maybe) all you need (to understand transformer optimization). In *The Twelfth International Conference on Learning Representations*.

Ahuja, K. and Lopez-Paz, D. (2023). A closer look at in-context learning under distribution shifts. *arXiv preprint arXiv:2305.16704*.

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. (2023). What learning algorithm is in-context learning? investigations with linear models. *The Eleventh International Conference on Learning Representations*.

Chang, X., Li, Y., Kara, M., Oymak, S., and Roy-Chowdhury, A. (2025). Provable benefits of task-specific prompts for in-context learning. In Li, Y., Mandt, S., Agrawal, S., and Khan, E., editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1558–1566. PMLR.

Chen, H., Zhao, Y., Chen, Z., Wang, M., Li, L., Zhang, M., and Zhang, M. (2024). Retrieval-style in-context learning for few-shot hierarchical text classification. *Transactions of the Association for Computational Linguistics*, 12:1214–1231.

Fang, L., Wang, Y., Gatmiry, K., Fang, L., and Wang, Y. (2025). Rethinking invariance in in-context learning. In *The Thirteenth International Conference on Learning Representations*.

Garg, S., Tsipras, D., Liang, P., and Valiant, G. (2022). What can transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems*.

Gatmiry, K., Saunshi, N., Reddi, S. J., Jegelka, S., and Kumar, S. (2024). Can looped transformers learn to implement multi-step gradient descent for in-context learning? In *International Conference on Machine Learning*.

Goddard, C., Smith, L. M., Ngampruetikorn, V., and Schwab, D. J. (2025). When can in-context learning generalize out of task distribution? In *Forty-second International Conference on Machine Learning*.

Gozeten, H. A., Ildiz, M. E., Zhang, X., Soltanolkotabi, M., Mondelli, M., and Oymak, S. (2025). Test-time training provably improves transformers as in-context learners. In *Forty-second International Conference on Machine Learning*.

Huang, Y., Cheng, Y., and Liang, Y. (2024). In-context convergence of transformers. In *International Conference on Machine Learning*, pages 19660–19722. PMLR.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kossen, J., Gal, Y., and Rainforth, T. (2024). In-context learning learns label relationships but is not conventional learning. In *The Twelfth International Conference on Learning Representations*.

Li, H., Wang, M., Lu, S., Cui, X., and Chen, P.-Y. (2024a). How do nonlinear transformers learn and generalize in in-context learning? In *International Conference on Machine Learning*, pages 28734–28783. PMLR.

Li, J., Tao, C., Li, J., Li, G., Jin, Z., Zhang, H., Fang, Z., and Liu, F. (2023). Large language model-aware in-context learning for code generation. *ACM Transactions on Software Engineering and Methodology*.

Li, Y., Rawat, A. S., and Oymak, S. (2024b). Fine-grained analysis of in-context linear estimation: Data, architecture, and beyond. *arXiv preprint arXiv:2407.10005*.

Mahankali, A. V., Hashimoto, T., and Ma, T. (2024). One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The Twelfth International Conference on Learning Representations*.

Mattei, P.-A. (2017). Multiplying a gaussian matrix by a gaussian vector. *Statistics & Probability Letters*, 128:67–70.

Oko, K., Song, Y., Suzuki, T., and Wu, D. (2025). Pretrained transformer efficiently learns low-dimensional target functions in-context. *Advances in Neural Information Processing Systems*, 37:77316–77365.

Pan, J., Gao, T., Chen, H., and Chen, D. (2023). What in-context learning "learns" in-context: Disentan-

gling task recognition and task learning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Patel, A., Reddy, S., Bahdanau, D., and Dasigi, P. (2024). Evaluating in-context learning of libraries for code generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2908–2926.

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, 7(15):510.

Raventós, A., Paul, M., Chen, F., and Ganguli, S. (2023). Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in neural information processing systems*, 36:14228–14246.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vidal, R. (2011). Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68.

Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., and Foster, G. (2022). Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. (2023). Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.

Wang, P., Liu, H., Pai, D., Yu, Y., Zhu, Z., Qu, Q., and Ma, Y. (2024a). A global geometric analysis of maximal coding rate reduction. In *Forty-first International Conference on Machine Learning*.

Wang, P., Zhang, H., Zhang, Z., Ma, Y., and Qu, Q. (2024b). Diffusion models learn low-dimensional distributions via subspace clustering. *arXiv preprint arXiv:2409.02426*.

Wang, Q., Wang, Y., Wang, Y., and Ying, X. (2025). Can in-context learning really generalize to out-of-distribution tasks? In *The Thirteenth International Conference on Learning Representations*.

Wang, Z., Xie, Q., Feng, Y., Ding, Z., Yang, Z., and Xia, R. (2024c). Is ChatGPT a good sentiment analyzer? In *First Conference on Language Modeling*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022a). Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022b). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xu, A. S., Yaras, C., Wang, P., and Qu, Q. (2025). Understanding how nonlinear layers create linearly separable features for low-dimensional data. *arXiv preprint arXiv:2501.02364*.

Xu, H., Wang, Q., Zhang, Y., Yang, M., Zeng, X., Qin, B., and Xu, R. (2024). Improving in-context learning with prediction feedback for sentiment analysis. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3879–3890.

Yadlowsky, S., Doshi, L., and Tripuraneni, N. (2023). Pretraining data mixtures enable narrow model selection capabilities in transformer models. *arXiv preprint arXiv:2311.00871*.

Zhang, R., Frei, S., and Bartlett, P. L. (2024). Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/**No**/Not Applicable] **We will release the code upon the completion of the review process for anonymity.**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/**Not Applicable**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/No/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

Supplementary Materials

Contents

A Additional Results

In this section, we present additional theoretical and experimental results to supplement those in the main text. In Section A.1, we generalize the result in Proposition 1 to the case of r distinct principal angles. In Section A.2, we explore how our framework can be used to analyze the effect of feature shifts between training and testing prompts.

A.1 Result with Different Principal Angles

In Proposition 1, we assumed that all of the r principal angles between the subspaces $\mathbf{U}_s \in \mathbb{R}^{d\times r}$ and $\mathbf{U}_t \in \mathbb{R}^{d\times r}$ were all the same, i.e., $\theta_i = \theta \in [0, \frac{\pi}{2}]$, for simplicity. In Proposition 2, we relax this requirement and present a result where the angles are not necessarily the same.

Proposition 2 (Task Distribution Shift with Different Angles)**.** *Let $g^{\star}_{ATT}$ denote the optimal linear attention model corresponding to the independent data setting in Equation (4) with where the task vector is drawn from a single subspace. For all $j \in [m+1]$, suppose that the prompts at test time are constructed with features $\mathbf{x}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and labels*

$$\widetilde{y}_j = \widetilde{\mathbf{w}}^\top \mathbf{x}_j + \eta_j, \quad \textit{where} \quad \widetilde{\mathbf{w}} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t) \quad \textit{and} \quad \eta_j \sim \mathcal{N}(0, \sigma^2),$$

with covariance matrix $\mathbf{\Sigma}_t \in \mathbb{R}^{d\times d}$ from Equation (6), but now with $\theta_1 \neq \theta_2 \neq \ldots \neq \theta_r$. Then, we have

$$\lim_{m\to\infty} \lim_{n\to\infty} \lim_{\epsilon\to 0} \mathbb{E}\left[\left(\tilde{y}_{m+1} - g^{\star}_{ATT}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = \sum_{i=1}^{r} \sin^2(\theta_i) + \sigma^2, \tag{10}$$

where $\theta_i \in [0, \frac{\pi}{2}]$ is the i-th principal angle between the train subspace $\mathbf{U}_s \in \mathbb{R}^{d\times r}$ and the test subspace $\mathbf{U}_t \in \mathbb{R}^{d\times r}$.

The proof is a direct extension of the proof for Proposition 1 and is left in Section B.1.3. Recall that the test risk presented in Proposition 1 was $r\sin^2(\theta) + \sigma^2$. It is easy to see that if we set $\theta_i = \theta$, then the test risk in Proposition 2 recovers the risk in Proposition 1, i.e., $\sum_{i=1}^{r} \sin^2(\theta_i) = r\sin^2(\theta)$.

A.2 Can Transformers Tolerate Feature Shifts?

Thus far, we studied how changes in the task vector affect the test risk of ICL. In this section, we explore how distribution shifts in the features (in both the prompts and the query) affect the test risk. Recent work by Zhang et al. (2024) demonstrated that Transformers are not robust to feature shifts, showing that ICL fails to remain robust when features are scaled by a constant. We share a similar observation that feature shifts can be much more detrimental to the test risk than task shifts, by considering the mathematical framework used thus far. Consider training a Transformer using prompts whose labels are constructed by a task vector $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$:

$$y_i = \mathbf{w}^\top \mathbf{x}_i + \eta_i, \quad \text{where} \quad \mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_s) \quad \text{and} \quad \eta_i \sim \mathcal{N}(0, \sigma^2). \tag{11}$$

During testing, we construct labels using the task vectors drawn from the same distribution but with shifted features: $\widetilde{y}_j = \mathbf{w}^\top \widetilde{\mathbf{x}}_j + \eta_j$, where $\widetilde{\mathbf{x}}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t)$. Before presenting our theoretical result, we illustrate in Figure 7 how shifts in the features affect the test risk, in contrast to shifts in the task vector. Interestingly, for both linear and nonlinear Transformer models, when the feature shift is small (i.e., $\theta > \frac{\pi}{4}$), the test risk changes only minimally, whereas it increases sharply for $\theta < \frac{\pi}{4}$. The following result captures this behavior, which arises as an artifact of using a degenerate distribution for the features.

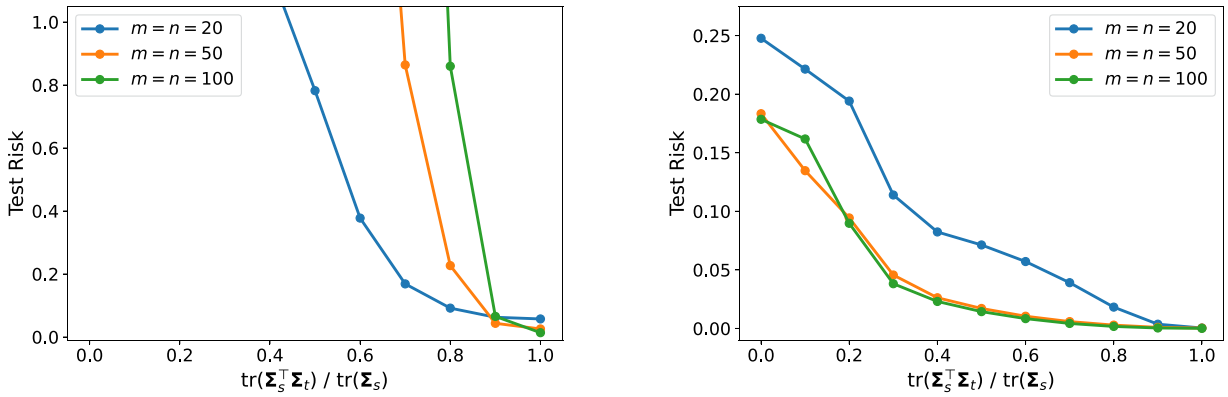

Figure 7: Plot of the test risk as the covariance matrix of the features shift as a function of t on a linear (left) and a nonlinear Transformer (right). When $t > 0.5$, the test risk does not increase as much as it does for $t < 0.5$, which contrasts with the behavior observed under covariance shifts in the task vector.

Proposition 3 (Feature Distribution Shift)**.** *Let $g^\star_{ATT}$ denote the optimal linear attention model corresponding to the independent data setting in Equation (11). For all $j \in [m+1]$, suppose that the prompts at test time are constructed with task vectors $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and labels*

$$\widetilde{y}_j = \mathbf{w}^\top \widetilde{\mathbf{x}}_j + \eta_j, \quad \text{where} \quad \widetilde{\mathbf{x}}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t), \quad \eta_j \sim \mathcal{N}(0, \sigma^2),$$

and $\mathbf{\Sigma}_t \in \mathbb{R}^{d\times d}$ is from Equation (6). Then, we have

$$\lim_{m\to\infty} \lim_{n\to\infty} \mathbb{E}\left[\left(\widetilde{y}_{m+1} - g^\star_{ATT}(\widetilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = \left(\frac{c_1 - (1+\epsilon)c_2}{(1+\epsilon)^2}\right) \cdot r\cos^2(\theta) + \left(\frac{c_1 - \epsilon c_2}{\epsilon^2}\right) \cdot r\sin^2(\theta) + r + \sigma^2 + \mathcal{O}(\epsilon),$$

where $c_1 = \left((1+2\epsilon)(1+\epsilon) + \epsilon^2\right)$ and $c_2 = 2(1+\epsilon)$.

The proof is available in Section B.2.2. Notice that we cannot immediately take $\epsilon \to 0$ in this setting, as that would result in an underdetermined system. The problematic term is the $\sin(\cdot)$, which reveals a few interesting insights: when $\theta = 0$, the $\sin(\cdot)$ term vanishes, allowing us to take $\epsilon \to 0$ and recover the same test risk as in Proposition 1. However, as θ moves away from zero, the risk contribution from the $\sin(\cdot)$ term starts to dominate, scaling with a factor of $\mathcal{O}(1/\epsilon^2)$ and causing the test risk to diverge since ϵ is assumed to be small. This result shows that testing with prompts whose features are shifted toward an almost independent covariance matrix (since ϵ is not exactly zero) can degrade test risk more than shifting the distribution of the task vector, roughly echoing the results of Zhang et al. (2024).

A.3 Experimental Results with Noisy Data

In the main text, we presented results for the noiseless setting ($\sigma^2 = 0$). Recall that our theory in Theorem 1 states that the test risk should converge to the variance of the noise given a sufficiently large prompt length. In Figure 8, we present results for the single-layer linear attention model, showing that when the training and testing prompt lengths are set to $m = n = 500$, the test risk converges roughly to the variance of the noise in two settings, $\sigma^2 = 0.01$ and 0.25, corroborating our theory. Here, we set $d = 20$ and $r = 6$.

B Deferred Proofs

This section presents all deferred proofs and is organized as follows: Section B.1 contains all proofs related to shifts in the task, Section B.2 includes all proofs related to shifts in the features, and Section B.3 provides

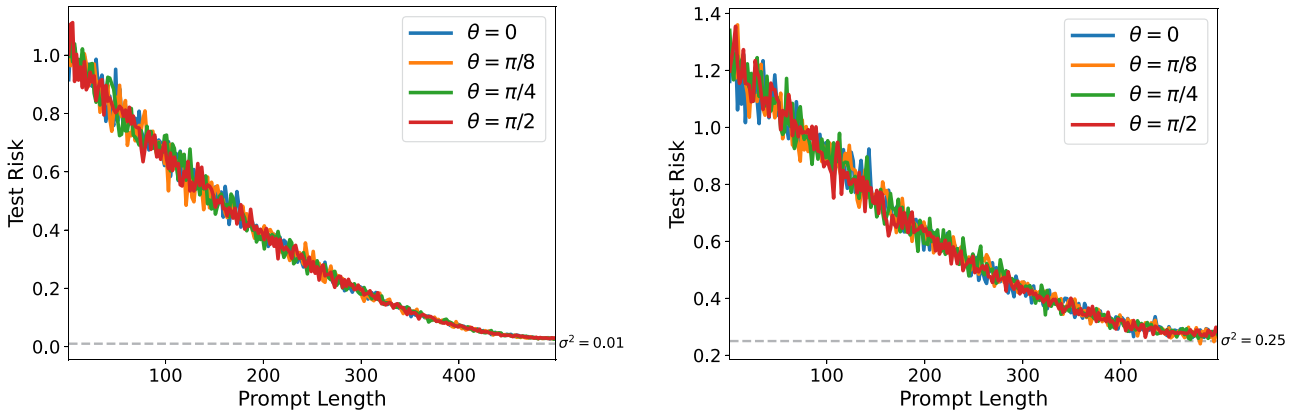

Figure 8: Experimental results for the single-layer linear attention model in the noisy setting, where the model is trained on a mixture of two low-dimensional Gaussians. Left: Test risk as a function of the prompt length with $\sigma^2 = 0.01$. Right: Test risk as a function of the prompt length with $\sigma^2 = 0.25$. Both cases converge to the variance of the noise, corroborating our theory in Theorem 1.

auxiliary results used to support both the task and feature shift proofs.

B.1 Proofs for Task Shifts

Here, we provide proofs of the results related to subspace shifts in the task vector.

B.1.1 Supporting Results

We first derive an expression for the test risk under a general distribution shift for the task vector.

Lemma 1 (Test Risk under General Task Distribution Shift)**.** *Let $g^\star_{ATT}$ denote the optimal linear attention model corresponding to the independent data setting in Equation (4), where $\mathbf{w}$ follows a (mixture of) Gaussian distribution(s) with zero mean and covariance $\mathbf{\Sigma}_s$. For all $j \in [m+1]$, suppose that the prompts at test time are constructed with features $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and labels*

$$\widetilde{y}_j = \widetilde{\mathbf{w}}^\top \mathbf{x}_j + \eta_i, \quad \textit{where} \quad \widetilde{\mathbf{w}} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t) \quad \textit{and} \quad \eta_j \sim \mathcal{N}(0, \sigma^2).$$

Then,

$$\mathbb{E}\left[\left(\widetilde{y}_{m+1} - g^\star_{ATT}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = M_t - \operatorname{Tr}\left(\mathbf{\Sigma}_t \mathbf{A}\right) + \frac{M_t}{m} \operatorname{Tr}\left(\mathbf{A}^\top \mathbf{A}\right) - \operatorname{Tr}\left(\mathbf{\Sigma}_t \mathbf{A}\right) + \frac{m+1}{m} \operatorname{Tr}\left(\mathbf{A} \mathbf{\Sigma}_t \mathbf{A}\right),$$

where $M_t = \operatorname{Tr}(\mathbf{\Sigma}_t) + \sigma^2$.

Proof. Recall at inference time,

$$\widetilde{\mathbf{Z}}_{\mathcal{M}} = \begin{bmatrix} \tilde{\mathbf{z}}_1 & \dots & \tilde{\mathbf{z}}_m & \mathbf{0} \end{bmatrix}^\top = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_m & \mathbf{0} \\ \widetilde{y}_1 & \dots & \widetilde{y}_m & 0 \end{bmatrix}^\top \quad \text{and} \quad \tilde{\mathbf{z}}_q = \begin{bmatrix} \mathbf{x}_{m+1} \\ 0 \end{bmatrix} := \begin{bmatrix} \mathbf{x}_q \\ 0 \end{bmatrix}. \tag{12}$$

Then, let us define

$$\mathbf{X}_{te} := \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_m \end{bmatrix}^\top, \quad \mathbf{y}_{te} := \begin{bmatrix} \widetilde{y}_1 & \widetilde{y}_2 & \dots & \widetilde{y}_m \end{bmatrix}^\top, \quad \boldsymbol{\eta}_{te} := \begin{bmatrix} \eta_1 & \eta_2 & \dots \eta_m \end{bmatrix}^\top,$$

and $\eta_q := \eta_{m+1}$. Note $\mathbf{y}_{te} = \mathbf{X}_{te} \widetilde{\mathbf{w}} + \boldsymbol{\eta}_{te}$. By Lemma 3, we have

$$g^\star_{\mathtt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}}) = \frac{1}{m} \mathbf{x}_q^\top \mathbf{A} \mathbf{X}_{te}^\top \mathbf{y}_{te} = \mathbf{x}_q^\top \underbrace{\left(\frac{1}{m} \mathbf{A} \mathbf{X}_{te}^\top \mathbf{y}_{te}\right)}_{:= \widehat{\mathbf{w}}},$$

where $\mathbf{A} = \left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\mathbf{\Sigma}_s\right)^{-1}$. By plugging this into the risk and linearity of expectation,

$$\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q + \eta_q - \widehat{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right] = \underbrace{\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q + \eta_q\right)^2\right]}_{(a)} - 2\underbrace{\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q + \eta_q\right)\left(\mathbf{x}_q^\top \widehat{\mathbf{w}}\right)\right]}_{(b)} + \underbrace{\mathbb{E}\left[\left(\widehat{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right]}_{(c)}. \tag{13}$$

It suffices to analyze each individual term.

Analyzing (a)**.** We first evaluate $\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q + \eta_q\right)^2\right]$. First, we note

$$\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q + \eta_q\right)^2\right] = \mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right] + 2\mathbb{E}\left[\eta_q \widetilde{\mathbf{w}}^\top \mathbf{x}_q\right] + \mathbb{E}\left[\eta_q^2\right] = \mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right] + \sigma^2,$$

so it suffices to analyze $\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right]$. By law of total expectation and the fact that $\widetilde{\mathbf{w}}, \mathbf{x}_q$ are independent,

$$\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right] = \mathbb{E}_{\widetilde{\mathbf{w}}}\left[\mathbb{E}_{\mathbf{x}_q}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q\right)^2 \mid \widetilde{\mathbf{w}}\right]\right].$$

Conditioned on $\widetilde{\mathbf{w}}$, $\widetilde{\mathbf{w}}^\top \mathbf{x}_q \sim \mathcal{N}(0, \|\widetilde{\mathbf{w}}\|^2)$, so $\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q\right)^2 \mid \widetilde{\mathbf{w}}\right] = \mathrm{Var}\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q \mid \widetilde{\mathbf{w}}\right) = \|\widetilde{\mathbf{w}}\|^2$. Therefore,

$$\mathbb{E}_{\widetilde{\mathbf{w}}}\left[\mathbb{E}_{\mathbf{x}}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q\right)^2 \mid \widetilde{\mathbf{w}}\right]\right] = \mathbb{E}\left[\|\widetilde{\mathbf{w}}\|^2\right] = \mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top\right]\right) = \mathrm{Tr}\left(\mathbf{\Sigma}_t\right).$$

Therefore,

$$\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q + \eta_q\right)^2\right] = \mathrm{Tr}(\mathbf{\Sigma}_t) + \sigma^2.$$

Analyzing (b)**.** Next, we analyze $\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q + \eta_q\right)\left(\mathbf{x}_q^\top \widehat{\mathbf{w}}\right)\right]$. We first note

$$\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q + \eta_q\right)\left(\mathbf{x}_q^\top \widehat{\mathbf{w}}\right)\right] = \mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q\right)\left(\mathbf{x}_q^\top \widehat{\mathbf{w}}\right)\right] + \underbrace{\mathbb{E}\left[\eta_q \mathbf{x}_q^\top \widehat{\mathbf{w}}\right]}_{=0} = \mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q\right)\left(\mathbf{x}_q^\top \widehat{\mathbf{w}}\right)\right],$$

so it suffices to analyze $\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q\right)\left(\mathbf{x}_q^\top \widehat{\mathbf{w}}\right)\right]$. Substituting $\widehat{\mathbf{w}} := \frac{1}{m}\mathbf{A}\mathbf{X}_{te}^\top \mathbf{y}_{te} = \frac{1}{m}\mathbf{A}\mathbf{X}_{te}^\top(\mathbf{X}_{te}\widetilde{\mathbf{w}} + \boldsymbol{\eta}_{te})$ yields

$$\begin{aligned}
\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q\right)\left(\mathbf{x}_q^\top \widehat{\mathbf{w}}\right)\right] &= \frac{1}{m}\mathbb{E}\left[\widetilde{\mathbf{w}}^\top \mathbf{x}_q \mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top(\mathbf{X}_{te}\widetilde{\mathbf{w}} + \boldsymbol{\eta}_{te})\right] \\
&= \frac{1}{m}\Big(\mathbb{E}\left[\widetilde{\mathbf{w}}^\top \mathbf{x}_q \mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\widetilde{\mathbf{w}}\right] + \mathbb{E}\left[\widetilde{\mathbf{w}}^\top \mathbf{x}_q \mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te}\right]\Big) \\
&= \frac{1}{m}\Big(\mathbb{E}\left[\widetilde{\mathbf{w}}^\top \mathbf{x}_q \mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\widetilde{\mathbf{w}}\right] + \underbrace{\mathbb{E}\left[\widetilde{\mathbf{w}}^\top \mathbf{x}_q \mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top\right]\mathbb{E}\left[\boldsymbol{\eta}_{te}\right]}_{=\mathbf{0}}\Big) \\
&= \frac{1}{m}\mathbb{E}\left[\widetilde{\mathbf{w}}^\top \mathbf{x}_q \mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\widetilde{\mathbf{w}}\right] = \frac{1}{m}\mathbb{E}\Big[\mathrm{Tr}\left(\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top \mathbf{x}_q \mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\right)\Big] \\
&= \frac{1}{m}\,\mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top \mathbf{x}_q \mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\right]\right) \\
&= \frac{1}{m}\,\mathrm{Tr}\Big(\underbrace{\mathbb{E}\left[\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top\right]}_{\mathbf{\Sigma}_t}\underbrace{\mathbb{E}\left[\mathbf{x}_q \mathbf{x}_q^\top\right]}_{\mathbf{I}_d}\mathbf{A}\underbrace{\mathbb{E}\left[\mathbf{X}_{te}^\top \mathbf{X}_{te}\right]}_{m\cdot\mathbf{I}_d}\Big) = \mathrm{Tr}\left(\mathbf{\Sigma}_t\mathbf{A}\right).
\end{aligned}$$

Analyzing (c)**.** Finally, we analyze $\mathbb{E}\left[(\mathbf{x}_q^\top \widehat{\mathbf{w}})^2\right]$:

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{x}_q^\top \widehat{\mathbf{w}})^2\right] &= \frac{1}{m^2}\mathbb{E}\left[\left(\mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top(\mathbf{X}_{te}\widetilde{\mathbf{w}} + \boldsymbol{\eta}_{te})\right)^2\right] = \frac{1}{m^2}\mathbb{E}\left[\left(\mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\widetilde{\mathbf{w}} + \mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te}\right)^2\right] \\
&= \frac{1}{m^2}\Big(\mathbb{E}\left[(\mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\widetilde{\mathbf{w}})^2\right] + 2\underbrace{\mathbb{E}\left[(\mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\widetilde{\mathbf{w}})(\mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te})\right]}_{=0} + \mathbb{E}\left[(\mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te})^2\right]\Big) \\
&= \frac{1}{m^2}\Big(\underbrace{\mathbb{E}\left[\mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top \mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \mathbf{x}_q\right]}_{(d)} + \underbrace{\mathbb{E}\left[\mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te}\boldsymbol{\eta}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \mathbf{x}_q\right]}_{(e)}\Big).
\end{aligned}$$

We first focus on (d):

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top \mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \mathbf{x}_q\right] = \mathbb{E}\Big[\operatorname{Tr}\left(\mathbf{x}_q\mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top \mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top\right)\Big] \\
&= \operatorname{Tr}\Big(\underbrace{\mathbb{E}\left[\mathbf{x}_q\mathbf{x}_q^\top\right]}_{\mathbf{I}_d}\mathbf{A}\mathbb{E}\left[\mathbf{X}_{te}^\top \mathbf{X}_{te}\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top \mathbf{X}_{te}^\top \mathbf{X}_{te}\right]\mathbf{A}^\top\Big) \\
&= \mathbb{E}\Big[\operatorname{Tr}\left(\mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top \mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top\right)\Big] = \mathbb{E}\Big[\operatorname{Tr}\left(\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top \mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\right)\Big] \\
&= \operatorname{Tr}\Big(\underbrace{\mathbb{E}\left[\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top\right]}_{\mathbf{\Sigma}_t}\mathbb{E}\left[\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\right]\Big) = \mathbb{E}\Big[\operatorname{Tr}\left(\mathbf{\Sigma}_t\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\right)\Big] \\
&= \mathbb{E}\Big[\operatorname{Tr}\left(\mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{\Sigma}_t^{1/2}\mathbf{\Sigma}_t^{1/2}\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top\right)\Big] := \mathbb{E}\Big[\operatorname{Tr}\left(\widetilde{\mathbf{X}}_{te}^\top \overline{\mathbf{X}}_{te}\overline{\mathbf{X}}_{te}^\top \widetilde{\mathbf{X}}_{te}\right)\Big],
\end{aligned}$$

where $\widetilde{\mathbf{X}}_{te}^\top := \mathbf{A}\mathbf{X}_{te}^\top$ and $\overline{\mathbf{X}}_{te}^\top := \mathbf{\Sigma}_t^{1/2}\mathbf{X}_{te}^\top$. Note $\widetilde{\mathbf{X}}_{te}^\top = \begin{bmatrix}\mathbf{A}\mathbf{x}_1 & \dots & \mathbf{A}\mathbf{x}_m\end{bmatrix} := \begin{bmatrix}\widetilde{\mathbf{x}}_1 & \dots & \widetilde{\mathbf{x}}_m\end{bmatrix}$ where $\widetilde{\mathbf{x}}_i := \mathbf{A}\mathbf{x}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}_d, \mathbf{A}\mathbf{A}^\top)$, and $\overline{\mathbf{X}}_{te}^\top = \begin{bmatrix}\mathbf{\Sigma}_t^{1/2}\mathbf{x}_1 & \dots & \mathbf{\Sigma}_t^{1/2}\mathbf{x}_m\end{bmatrix} := \begin{bmatrix}\overline{\mathbf{x}}_1 & \dots & \overline{\mathbf{x}}_m\end{bmatrix}$ where $\overline{\mathbf{x}}_i := \mathbf{\Sigma}_t^{1/2}\mathbf{x}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}_d, \mathbf{\Sigma}_t)$. We can express $\widetilde{\mathbf{X}}_{te}^\top\overline{\mathbf{X}}_{te}$ and $\overline{\mathbf{X}}_{te}^\top\widetilde{\mathbf{X}}_{te}$ as such:

$$\widetilde{\mathbf{X}}_{te}^\top\overline{\mathbf{X}}_{te} = \sum_{i=1}^m \widetilde{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top \quad \text{and} \quad \overline{\mathbf{X}}_{te}^\top\widetilde{\mathbf{X}}_{te} = \sum_{j=1}^m \overline{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^\top.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\Big[\operatorname{Tr}\left(\widetilde{\mathbf{X}}_{te}^\top\overline{\mathbf{X}}_{te}\overline{\mathbf{X}}_{te}^\top\widetilde{\mathbf{X}}_{te}\right)\Big] = \operatorname{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{X}}_{te}^\top\overline{\mathbf{X}}_{te}\overline{\mathbf{X}}_{te}^\top\widetilde{\mathbf{X}}_{te}\right]\right) \\
&= \operatorname{Tr}\left(\sum_{i=1}^m\sum_{j=1}^m \mathbb{E}\left[\widetilde{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^\top\right]\right) \\
&= \operatorname{Tr}\left(\sum_{i=1}^m\sum_{j\neq i} \mathbb{E}\left[\widetilde{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^\top\right]\right) + \operatorname{Tr}\left(\sum_{i=1}^m \mathbb{E}\left[\widetilde{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\right]\right)
\end{aligned}$$

We first consider the case when $i \neq j$. In this setting, $\mathbf{x}_i$ and $\mathbf{x}_j$ are independent, so

$$\mathbb{E}\left[\widetilde{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^\top\right] = \mathbb{E}\left[\widetilde{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\right]\mathbb{E}\left[\overline{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^\top\right] = \mathbf{A}\underbrace{\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^\top\right]}_{\mathbf{I}_d}\mathbf{\Sigma}_t\underbrace{\mathbb{E}\left[\mathbf{x}_j\mathbf{x}_j^\top\right]}_{\mathbf{I}_d}\mathbf{A}^\top = \mathbf{A}\mathbf{\Sigma}_t\mathbf{A}^\top.$$

Therefore,

$$\operatorname{Tr}\left(\sum_{i=1}^m\sum_{j\neq i} \mathbb{E}\left[\widetilde{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^\top\right]\right) = m\cdot(m-1)\cdot\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t\mathbf{A}^\top\right).$$

We now consider the case where $i = j$:

$$\begin{aligned}
&\operatorname{Tr}\left(\sum_{i=1}^m \mathbb{E}\left[\widetilde{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\right]\right) = \sum_{i=1}^m \mathbb{E}\Big[\operatorname{Tr}\left(\widetilde{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\right)\Big] \\
&= \sum_{i=1}^m \mathbb{E}\left[\widetilde{\mathbf{x}}_i^\top\widetilde{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_i\right] = \sum_{i=1}^m \mathbb{E}\left[(\mathbf{x}_i^\top\mathbf{A}^\top\mathbf{A}\mathbf{x}_i)(\mathbf{x}_i^\top\mathbf{\Sigma}_t\mathbf{x}_i)\right] \\
&\overset{(i)}{=} m\cdot\left(2\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t\mathbf{A}^\top\right) + \operatorname{Tr}\left(\mathbf{A}^\top\mathbf{A}\right)\operatorname{Tr}\left(\mathbf{\Sigma}_t\right)\right),
\end{aligned}$$

where (i) is because for $\mathbf{a} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and fixed $\mathbf{Q}, \mathbf{R} \in \mathbb{R}^{d\times d}$, $\mathbb{E}\left[(\mathbf{a}^\top\mathbf{Q}\mathbf{a})(\mathbf{a}^\top\mathbf{R}\mathbf{a})\right] = \operatorname{Tr}\left(\mathbf{Q}(\mathbf{R}+\mathbf{R}^\top)\right) + \operatorname{Tr}(\mathbf{Q})\operatorname{Tr}(\mathbf{R})$ (see Section 8.2.4 in Petersen et al. (2008)).

We now focus on (e):

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te}\boldsymbol{\eta}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \mathbf{x}_q\right] = \mathbb{E}\Big[\operatorname{Tr}\left(\mathbf{x}_q\mathbf{x}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te}\boldsymbol{\eta}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top\right)\Big] \\
&= \operatorname{Tr}\Big(\underbrace{\mathbb{E}\left[\mathbf{x}_q\mathbf{x}_q^\top\right]}_{\mathbf{I}_d}\mathbf{A}\mathbb{E}\left[\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te}\boldsymbol{\eta}_{te}^\top \mathbf{X}_{te}\right]\mathbf{A}^\top\Big) = \operatorname{Tr}\Big(\mathbb{E}\left[\mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te}\boldsymbol{\eta}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top\right]\Big) \\
&:= \operatorname{Tr}\Big(\mathbb{E}\left[\widetilde{\boldsymbol{\eta}}_{te}\widetilde{\boldsymbol{\eta}}_{te}^\top\right]\Big),
\end{aligned}$$

where $\widetilde{\boldsymbol{\eta}}_{te} := \mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te} = \widetilde{\mathbf{X}}_{te}^\top \boldsymbol{\eta}_{te}$. Note the columns of $\widetilde{\mathbf{X}}_{te}^\top$ are iid Gaussian with covariance $\mathbf{A}\mathbf{A}^\top$. By Corollary 6 in Mattei (2017), $\widetilde{\boldsymbol{\eta}}_{te} \sim \mathrm{GAL}_d(2\sigma^2\mathbf{A}\mathbf{A}^\top, \mathbf{0}_d, m/2)$, where $\mathrm{GAL}_p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, s)$ denotes a p-dimensional *multivariate generalized asymmetric Laplace distribution* with mean $s\boldsymbol{\mu}$ and covariance $s(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top)$ (Definition 1 and Proposition 2 in Mattei (2017)). Therefore,

$$\operatorname{Tr}\Big(\mathbb{E}\left[\widetilde{\boldsymbol{\eta}}_{te}\widetilde{\boldsymbol{\eta}}_{te}^\top\right]\Big) = \operatorname{Tr}\Big(\operatorname{Cov}(\widetilde{\boldsymbol{\eta}}_{te})\Big) = m\sigma^2 \operatorname{Tr}\Big(\mathbf{A}\mathbf{A}^\top\Big).$$

Adding (a), (b), and (c). Adding the expressions for (a), (b), and (c), where $(c) = (d) + (e)$, yields and combining like terms yields the following expression:

$$\begin{aligned}
\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q + \eta_q - \widehat{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right] &= \underbrace{\operatorname{Tr}(\boldsymbol{\Sigma}_t) + \sigma^2}_{=(a)} - 2\underbrace{\operatorname{Tr}\left(\boldsymbol{\Sigma}_t\mathbf{A}\right)}_{=(b)} \\
&+ \underbrace{\frac{1}{m^2}\Big(\underbrace{m(m-1)\operatorname{Tr}\left(\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^\top\right) + 2m\operatorname{Tr}\left(\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^\top\right) + m\operatorname{Tr}(\boldsymbol{\Sigma}_t)\operatorname{Tr}\left(\mathbf{A}^\top\mathbf{A}\right)}_{=(d)} + \underbrace{m\sigma^2\operatorname{Tr}\left(\mathbf{A}^\top\mathbf{A}\right)}_{=(e)}\Big)}_{=(c)}.
\end{aligned}$$

Combining like terms yields

$$\begin{aligned}
\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q + \eta_q - \widehat{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right] &= \left(\frac{1}{m}\operatorname{Tr}\left(\mathbf{A}^\top\mathbf{A}\right) + 1\right)\left(\operatorname{Tr}\left(\boldsymbol{\Sigma}_t\right) + \sigma^2\right) - 2\operatorname{Tr}\left(\boldsymbol{\Sigma}_t\mathbf{A}\right) + \frac{m+1}{m}\operatorname{Tr}\left(\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^\top\right) \\
&= M_t - \operatorname{Tr}\left(\boldsymbol{\Sigma}_t\mathbf{A}\right) + \frac{M_t}{m}\operatorname{Tr}\left(\mathbf{A}^\top\mathbf{A}\right) - \operatorname{Tr}\left(\boldsymbol{\Sigma}_t\mathbf{A}\right) + \frac{m+1}{m}\operatorname{Tr}\left(\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}\right),
\end{aligned}$$

which is exactly Equation (13). This completes the proof. □

B.1.2 Proof of Theorems 1 and 2

We now provide a proof of Theorems 1 and 2. We specifically focus on the setting of Theorem 2. However, we also emphasize that the proof reduces down to a proof of Theorem 1 when $K = 2$, $\alpha_1 = \sin(\theta)$, and $\alpha_2 = \cos(\theta)$ for any $\theta \in [0, \pi/2]$.

Proof. Let $\widetilde{y} := \widetilde{y}_{m+1}$. By Lemmas 1 and 4, we have

$$\mathbb{E}\left[\left(\widetilde{y} - g^\star_{\mathtt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = \left(\frac{1}{m}\operatorname{Tr}\left(\mathbf{A}^\top\mathbf{A}\right) + 1\right)\left(\operatorname{Tr}\left(\overline{\boldsymbol{\Sigma}}_t\right) + \sigma^2\right) - 2\operatorname{Tr}\left(\overline{\boldsymbol{\Sigma}}_t\mathbf{A}\right) + \frac{m+1}{m}\operatorname{Tr}\left(\mathbf{A}\overline{\boldsymbol{\Sigma}}_t\mathbf{A}^\top\right), \quad (14)$$

where $\mathbf{A} = \left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\boldsymbol{\Sigma}^{-1}\right)^{-1}$, $M_s = \operatorname{Tr}(\boldsymbol{\Sigma}) + \sigma^2$, and $\boldsymbol{\Sigma} = \sum_{k=1}^{K}\gamma_k \cdot \boldsymbol{\Sigma}_{s,k}$.

Let $\mathbf{U} := \begin{bmatrix}\mathbf{U}_{s,1} & \mathbf{U}_{s,2} & \dots & \mathbf{U}_{s,K} & \mathbf{U}_\perp\end{bmatrix}$, where $\mathbf{U}_\perp \in \mathbb{R}^{d\times(d-Kr)}$ completes the orthonormal basis for $\mathbb{R}^d$. By Lemma 6,

$$\mathbf{A} = \left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\boldsymbol{\Sigma}^{-1}\right)^{-1} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top,$$

where

$$\mathbf{\Lambda} = \begin{bmatrix} \nu_1 \mathbf{I}_r & & & \\ & \ddots & & \\ & & \nu_K \mathbf{I}_r & \\ & & & \nu_{K+1}\mathbf{I}_{d-Kr} \end{bmatrix}$$

with $\nu_k = \frac{n(\gamma_k+\epsilon)}{(n+1)(\gamma_k+\epsilon)+M_s}$ for all $k \in [K]$, and $\nu_{K+1} = \frac{n\epsilon}{(n+1)\epsilon+r+\epsilon d+\sigma^2}$.

Simplifying $\mathrm{Tr}(\overline{\mathbf{\Sigma}}_t)$. We can write $\mathrm{Tr}(\overline{\mathbf{\Sigma}}_t)$ as such:

$$\mathrm{Tr}(\overline{\mathbf{\Sigma}}_t) = \mathrm{Tr}(\overline{\mathbf{U}}_t \overline{\mathbf{U}}_t^\top) + \epsilon \,\mathrm{Tr}(\mathbf{I}_d) = r + \epsilon d.$$

Simplifying $\mathrm{Tr}(\mathbf{A})$ and $\mathrm{Tr}(\mathbf{A}^\top\mathbf{A})$. We can write $\mathrm{Tr}(\mathbf{A})$ and $\mathrm{Tr}(\mathbf{A}^\top\mathbf{A})$ as such:

$$\mathrm{Tr}(\mathbf{A}) = r\sum_{k=1}^{K}\nu_k + (d-Kr)\nu_{K+1} \quad \text{and} \quad \mathrm{Tr}(\mathbf{A}^\top\mathbf{A}) = \mathrm{Tr}(\mathbf{A}^2) = r\sum_{k=1}^{K}\nu_k^2 + (d-Kr)\nu_{K+1}^2.$$

Simplifying $\mathrm{Tr}(\overline{\mathbf{\Sigma}}_t\mathbf{A})$ and $\mathrm{Tr}(\mathbf{A}\overline{\mathbf{\Sigma}}_t\mathbf{A}^\top)$. Note $\mathrm{Tr}(\mathbf{A}\overline{\mathbf{\Sigma}}_t\mathbf{A}^\top) = \mathrm{Tr}(\overline{\mathbf{\Sigma}}_t\mathbf{A}^2)$. We first focus on $\mathrm{Tr}(\overline{\mathbf{\Sigma}}_t\mathbf{A})$:

$$\begin{aligned} \overline{\mathbf{\Sigma}}_t\mathbf{A} &= \left(\overline{\mathbf{U}}_t\overline{\mathbf{U}}_t^\top + \epsilon\mathbf{I}_d\right)\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top = \overline{\mathbf{U}}_t\overline{\mathbf{U}}_t^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top + \epsilon\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \\ \implies \mathrm{Tr}\left(\overline{\mathbf{\Sigma}}_t\mathbf{A}\right) &= \mathrm{Tr}\left(\overline{\mathbf{U}}_t^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\overline{\mathbf{U}}_t\right) + \epsilon\,\mathrm{Tr}(\mathbf{A}). \end{aligned}$$

Recall $\overline{\mathbf{U}}_t = \sum_{k=1}^{K}\alpha_k\mathbf{U}_{s,k}$ where $\sum_{k=1}^{K}\alpha_k^2 = 1$, and so we have

$$\overline{\mathbf{U}}_t^\top\mathbf{U} = \left(\sum_{k=1}^{K}\alpha_k\mathbf{U}_k\right)^\top \begin{bmatrix}\mathbf{U}_{s,1} & \dots & \mathbf{U}_{s,K} & \mathbf{U}_\perp\end{bmatrix} = \begin{bmatrix} \alpha_1\mathbf{I}_r & & & \\ & \ddots & & \\ & & \alpha_K\mathbf{I}_r & \\ & & & \mathbf{0}_{(d-Kr)\times(d-Kr)} \end{bmatrix}$$

Thus,

$$\mathrm{Tr}\left(\overline{\mathbf{U}}_t^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\overline{\mathbf{U}}_t\right) = \mathrm{Tr}\left(\begin{bmatrix} \alpha_1^2\nu_1\mathbf{I}_r & & & \\ & \ddots & & \\ & & \alpha_K^2\nu_K\mathbf{I}_r & \\ & & & \mathbf{0}_{(d-Kr)\times(d-Kr)} \end{bmatrix}\right) = r\sum_{k=1}^{K}\alpha_k^2\nu_k$$

Using a similar argument,

$$\mathrm{Tr}\left(\overline{\mathbf{\Sigma}}_t^\top\mathbf{A}^2\right) = r\sum_{k=1}^{K}\alpha_k^2\nu_k^2 + \epsilon\,\mathrm{Tr}\left(\mathbf{A}^2\right).$$

Simplifying the test risk. Substituting the expressions for the $\mathrm{Tr}(\cdot)$ terms into Equation (14) yields

$$\begin{aligned} \mathbb{E}\left[\left(\widetilde{y} - g^\star_{\texttt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = &\left(\frac{1}{m}\left(r\sum_{k=1}^{K}\nu_k^2 + (d-Kr)\nu_{K+1}^2\right) + 1\right)\left(r + \epsilon d + \sigma^2\right) \\ &- 2\left(r\sum_{k=1}^{K}\alpha_k^2\nu_k + \left(r\sum_{k=1}^{K}\nu_k + (d-Kr)\nu_{K+1}\right)\epsilon\right) \\ &+ \frac{m+1}{m}\left(r\sum_{k=1}^{K}\alpha_k^2\nu_k^2 + \left(r\sum_{k=1}^{K}\nu_k^2 + (d-Kr)\nu_{K+1}^2\right)\epsilon\right). \end{aligned}$$

Taking $\epsilon \to 0$ results in the following expression for the test risk:

$$\lim_{\epsilon\to 0} \mathbb{E}\left[\left(\widetilde{y} - g^{\star}_{\texttt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = r + \sigma^2 + \frac{(r+\sigma^2)r}{m}\sum_{k=1}^{K}\left(\frac{\gamma_k n}{\gamma_k(n+1)+r+\sigma^2}\right)^2 - 2r\sum_{k=1}^{K}\frac{\alpha_k^2\gamma_k n}{\gamma_k(n+1)+r+\sigma^2} + \frac{(m+1)r}{m}\sum_{k=1}^{K}\left(\frac{\alpha_k\gamma_k n}{\gamma_k(n+1)+r+\sigma^2}\right)^2$$

Substituting $\gamma_k = \frac{1}{K}$ for all $k \in [K]$ and combining like terms yields

$$\lim_{\epsilon\to 0} \mathbb{E}\left[\left(\widetilde{y} - g^{\star}_{\texttt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = r + \sigma^2 + \frac{m+1+K(r+\sigma^2)}{m}\cdot\frac{rn^2}{(n+1+K(r+\sigma^2))^2} - \frac{2rn}{n+1+K(r+\sigma^2)}.$$

Now suppose $n \le m$. Then, we have

$$\lim_{\epsilon\to 0} \mathbb{E}\left[\left(\widetilde{y} - g^{\star}_{\texttt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] \le r + \sigma^2 - \frac{rn^2}{n+1+K(r+\sigma^2)}.$$

Upper bounding this by $\sigma^2 + \delta$ for some $\delta \in (0, r)$, then solving for n, yields the following result. For any $\delta \in (0, r)$, if

$$m \ge n > \frac{(K(r+\sigma^2)+1)r}{\delta} - (K(r+\sigma^2)+1),$$

then $\lim_{\epsilon\to 0} \mathbb{E}\left[\left(\widetilde{y} - g^{\star}_{\texttt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] < \sigma^2 + \delta$, which completes the proof. □

B.1.3 Proof of Propositions 1 and 2

We now provide the proof of Proposition 1, which discusses how the proof can be generalized to the setting of Proposition 2.

Proof. For simplicity, we denote $\widetilde{y} \coloneqq \widetilde{y}_{m+1}$. Recall $\mathbf{U} := \begin{bmatrix}\mathbf{U}_s & \mathbf{U}_{s,\perp} & \mathbf{U}_{2r,\perp}\end{bmatrix} \in \mathbb{R}^{d\times d}$, where $\mathbf{U}_s, \mathbf{U}_{s,\perp} \in \mathbb{R}^{d\times r}$ and $\mathbf{U}_{2r,\perp} \in \mathbb{R}^{d\times(d-2r)}$ all have orthonormal columns, while $\mathbf{U}_s^\top\mathbf{U}_{\perp,s} = \mathbf{0}_{r\times r}$ and $\mathbf{U}_s^\top\mathbf{U}_\perp = \mathbf{U}_{s,\perp}^\top\mathbf{U}_{2r,\perp} = \mathbf{0}_{r\times(d-2r)}$. We re-write $\mathbf{\Sigma}_s$ as such:

$$\mathbf{\Sigma}_s = \mathbf{U}_s\mathbf{U}_s^\top + \epsilon\mathbf{I}_d = \mathbf{U}\begin{bmatrix}\mathbf{I}_r & \\ & \mathbf{0}_{(d-r)\times(d-r)}\end{bmatrix}\mathbf{U}^\top + \epsilon\mathbf{I} = \mathbf{U}\begin{bmatrix}(1+\epsilon)\mathbf{I}_r & \\ & \epsilon\mathbf{I}_{d-r}\end{bmatrix}\mathbf{U}^\top.$$

Note this is a valid eigendecomposition of $\mathbf{\Sigma}_s$. Thus, by Lemma 6, we have

$$\mathbf{A} = \left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\mathbf{\Sigma}_s^{-1}\right)^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top, \tag{15}$$

where

$$\mathbf{\Lambda} = \begin{bmatrix}\frac{n(1+\epsilon)}{(n+1)\epsilon+M_s}\cdot\mathbf{I}_r & \\ & \frac{n\epsilon}{(n+1)\epsilon+M_s}\cdot\mathbf{I}_{d-r}\end{bmatrix} := \begin{bmatrix}\nu_1\mathbf{I}_r & \\ & \nu_2\mathbf{I}_{d-r}\end{bmatrix}.$$

and $M_s = \mathrm{Tr}(\mathbf{\Sigma}_s) + \sigma^2$.

By Lemma 1 (and omitting the subscripts in the expectation),

$$\mathbb{E}\left[\left(\widetilde{y} - g^{\star}_{\texttt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = \left(\frac{1}{m}\mathrm{Tr}\left(\mathbf{A}^\top\mathbf{A}\right)+1\right)\left(\mathrm{Tr}\left(\mathbf{\Sigma}_t\right)+\sigma^2\right) - 2\,\mathrm{Tr}\left(\mathbf{\Sigma}_t\mathbf{A}\right) + \frac{m+1}{m}\mathrm{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t\mathbf{A}^\top\right). \tag{16}$$

We simplify the remaining $\mathrm{Tr}(\cdot)$ terms using Equation (15).

Simplifying $\mathrm{Tr}(\mathbf{A})$ and $\mathrm{Tr}(\mathbf{A}^\top\mathbf{A})$. Directly from Equation (15):

$$\mathrm{Tr}(\mathbf{A}) = r\cdot\nu_1 + (d-r)\cdot\nu_2 \quad\text{and}\quad \mathrm{Tr}(\mathbf{A}^\top\mathbf{A}) = \mathrm{Tr}(\mathbf{A}^2) = r\cdot\nu_1^2 + (d-r)\cdot\nu_2^2,$$

where $\mathbf{A}^2 = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^\top$.

Simplifying $\mathrm{Tr}(\mathbf{\Sigma}_t\mathbf{A})$ and $\mathrm{Tr}(\mathbf{A}\mathbf{\Sigma}_t\mathbf{A}^\top)$. First note $\mathrm{Tr}(\mathbf{A}\mathbf{\Sigma}_t\mathbf{A}^\top) = \mathrm{Tr}(\mathbf{\Sigma}_t\mathbf{A}^2)$. We first focus on $\mathrm{Tr}(\mathbf{\Sigma}_t\mathbf{A})$:

$$\begin{aligned}\mathbf{\Sigma}_t\mathbf{A} &= \left(\mathbf{U}_t\mathbf{U}_t^\top + \epsilon\mathbf{I}_d\right)\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top = \mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top + \epsilon\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \\ \implies \mathrm{Tr}\left(\mathbf{\Sigma}_t\mathbf{A}\right) &= \mathrm{Tr}\left(\mathbf{U}_t^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\mathbf{U}_t\right) + \epsilon\,\mathrm{Tr}(\mathbf{A}).\end{aligned}$$

Recall we defined $\mathbf{U}_t$ in Equation (5) as follows:

$$\mathbf{U}_t = \mathbf{U}_s\cos(\mathbf{\Theta}) + \mathbf{U}_{s,\perp}\sin(\mathbf{\Theta}).$$

Therefore:

$$\mathbf{U}_t^\top\mathbf{U} = \left(\mathbf{U}_s\cos(\mathbf{\Theta}) + \mathbf{U}_{s,\perp}\sin(\mathbf{\Theta})\right)^\top\begin{bmatrix}\mathbf{U}_s & \mathbf{U}_{s,\perp} & \mathbf{U}_\perp\end{bmatrix} = \begin{bmatrix}\cos(\mathbf{\Theta}) & \sin(\mathbf{\Theta}) & \mathbf{0}_{d\times d-2r}\end{bmatrix},$$

and thus,

$$\begin{aligned}\mathrm{Tr}\left(\mathbf{U}_t^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\mathbf{U}_t\right) &= \mathrm{Tr}\left(\begin{bmatrix}\cos(\mathbf{\Theta}) & \sin(\mathbf{\Theta}) & \mathbf{0}_{d\times(d-2r)}\end{bmatrix}\begin{bmatrix}\nu_1\mathbf{I}_r & & \\ & \nu_2\mathbf{I}_r & \\ & & \nu_2\mathbf{I}_{d-2r}\end{bmatrix}\begin{bmatrix}\cos(\mathbf{\Theta}) \\ \sin(\mathbf{\Theta}) \\ \mathbf{0}_{(d-2r)\times d}\end{bmatrix}\right) \\ &= \mathrm{Tr}\left(\begin{bmatrix}\nu_1\cos^2(\mathbf{\Theta}) & & \\ & \nu_2\sin^2(\mathbf{\Theta}) & \\ & & \mathbf{0}_{(d-2r)\times(d-2r)}\end{bmatrix}\right) = r\cdot\nu_1\cdot\cos^2(\theta) + r\cdot\nu_2\cdot\sin^2(\theta),\end{aligned}$$

where we used the fact that the principal angles are all equal to θ. Using a similar argument,

$$\mathrm{Tr}\left(\mathbf{\Sigma}_t^\top\mathbf{A}^2\right) = r\cdot\nu_1^2\cdot\cos^2(\theta) + r\cdot\nu_2^2\cdot\sin^2(\theta) + \epsilon\,\mathrm{Tr}(\mathbf{A}^2)$$

Simplifying the Test Risk. Substituting the expressions for the $\mathrm{Tr}(\cdot)$ terms into Equation (16) yields

$$\begin{aligned}\mathbb{E}\left[\left(\widetilde{y} - g^\star_{\texttt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] =& \left(\frac{1}{m}\left(r\nu_1^2 + (d-r)\nu_2^2\right) + 1\right)\left(r + \epsilon d + \sigma^2\right) \\ &- 2\left(r\nu_1\cos^2(\theta) + r\nu_2\sin^2(\theta) + \left(r\nu_1 + (d-r)\nu_2\right)\epsilon\right) \\ &+ \frac{m+1}{m}\left(r\nu_1^2\cos^2(\theta) + r\nu_2^2\sin^2(\theta) + \left(r\nu_1^2 + (d-r)\nu_2^2\right)\epsilon\right)\end{aligned}$$

Substituting the expressions for ν_1 and ν_2 and taking $\epsilon\to 0$ results in the following:

$$\begin{aligned}\lim_{\epsilon\to 0}\mathbb{E}\left[\left(\widetilde{y} - g^\star_{\texttt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] =& \left(\frac{rn^2}{m(n+1+r+\sigma^2)^2} + 1\right)(r+\sigma^2) \\ &- \frac{2rn\cos^2(\theta)}{n+1+r+\sigma^2} + \frac{(m+1)rn^2\cos^2(\theta)}{m(n+1+r+\sigma^2)^2}\end{aligned}$$

Subsequently taking $m, n\to\infty$ yields

$$\lim_{m\to\infty}\lim_{n\to\infty}\lim_{\epsilon\to 0}\mathbb{E}\left[\left(\widetilde{y} - g^\star_{\texttt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = r + \sigma^2 - r\cos^2(\theta) = r\sin^2(\theta) + \sigma^2,$$

which completes the proof. To generalize this result to the case in which all of the principal angles are not the same, i.e., $\theta_i \neq \theta$ for all $i\in[r]$, we can replace all terms with $r\cos^2(\theta)$ and $r\sin^2(\theta)$ with $\sum_{i=1}^r\cos^2(\theta_i)$ and $\sum_{i=1}^r\sin^2(\theta_i)$, respectively, due to the trace terms. This gives us an overall test risk of

$$\lim_{m\to\infty}\lim_{n\to\infty}\lim_{\epsilon\to 0}\mathbb{E}\left[\left(\widetilde{y} - g^\star_{\texttt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = r + \sigma^2 - \sum_{i=1}^r\cos^2(\theta_i) = \sum_{i=1}^r\sin^2(\theta_i) + \sigma^2,$$

which gives the result in Proposition 2. □

B.2 Proofs for Feature Shifts

In this section, we provide proofs of our theoretical results on the generalization abilities of a linear attention model when the features undergo a distribution shift.

B.2.1 Supporting Results

We first derive an expression for the test risk under a general distribution shift for the features.

Lemma 2 (Test Risk Under General Feature Distribution Shift)**.** *Let $g^{\star}_{ATT}$ denote the optimal linear attention model corresponding to the independent data setting in Equation (11). For all $j \in [m+1]$, suppose that the prompts at test time are constructed with task vectors $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and labels*

$$\widetilde{y}_j = \mathbf{w}^\top \widetilde{\mathbf{x}}_j + \eta_j, \quad \text{where} \quad \widetilde{\mathbf{x}}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t) \quad \text{and} \quad \eta_j \sim \mathcal{N}(0, \sigma^2),$$

Then, we have

$$\begin{aligned}\mathbb{E}\left[\left(\widetilde{y}_{m+1} - g^{\star}_{ATT}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] &= M_t - 2\operatorname{Tr}\left(\mathbf{\Sigma}_t^2 \mathbf{A}\right) + \frac{m+1}{m}\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^3\mathbf{A}\right) \\ &\quad + \frac{1}{m}\operatorname{Tr}\left(\mathbf{\Sigma}_t^2\right)\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t\mathbf{A}\right) + \frac{\sigma^2}{m}\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^2\mathbf{A}\right), \end{aligned} \tag{17}$$

where $M_t = \operatorname{Tr}(\mathbf{\Sigma}_t) + \sigma^2$.

Proof. The proof is similar to that of Lemma 1 with $\widetilde{y}_{m+1} = \mathbf{w}^\top \widetilde{\mathbf{x}}_{m+1} + \eta_q$, where $\widetilde{\mathbf{x}}_{m+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t)$. Recall at inference time,

$$\widetilde{\mathbf{Z}}_{\mathcal{M}} = \begin{bmatrix}\mathbf{z}_1 & \dots & \mathbf{z}_m & \mathbf{0}\end{bmatrix}^\top = \begin{bmatrix}\widetilde{\mathbf{x}}_1 & \dots & \widetilde{\mathbf{x}}_m & \mathbf{0} \\ \widetilde{y}_1 & \dots & \widetilde{y}_m & 0\end{bmatrix}^\top \quad \text{and} \quad \tilde{\mathbf{z}}_q = \begin{bmatrix}\widetilde{\mathbf{x}}_{m+1} \\ 0\end{bmatrix} := \begin{bmatrix}\widetilde{\mathbf{x}}_q \\ 0\end{bmatrix}.$$

Again we define

$$\mathbf{X}_{te} = \begin{bmatrix}\widetilde{\mathbf{x}}_1 & \widetilde{\mathbf{x}}_2 & \dots & \widetilde{\mathbf{x}}_m\end{bmatrix}^\top, \quad \mathbf{y}_{te} = \begin{bmatrix}\widetilde{y}_1 & \widetilde{y}_2 & \dots & \widetilde{y}_m\end{bmatrix}^\top, \quad \boldsymbol{\eta}_{te} = \begin{bmatrix}\eta_1 & \eta_2 & \dots & \eta_m\end{bmatrix}^\top,$$

and $\eta_q := \eta_{m+1}$. By Lemma 5, we have

$$g^{\star}_{\mathtt{ATT}}\left(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}}\right) = \frac{1}{m}\widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{y}_{te} = \widetilde{\mathbf{x}}_q^\top \underbrace{\left(\frac{1}{m}\mathbf{A}\mathbf{X}_{te}^\top \mathbf{y}_{te}\right)}_{:=\widehat{\mathbf{w}}},$$

where $\mathbf{A} = \mathbf{\Sigma}_s^{-1/2}\bar{\mathbf{A}}\mathbf{\Sigma}_s^{-1/2}$ and $\bar{\mathbf{A}} = \left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\mathbf{\Sigma}_s^{-1/2}\right)^{-1}$. By plugging this into the risk and linearity of expectation,

$$\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q + \eta_q - \widehat{\mathbf{w}}^\top \widetilde{\mathbf{x}}_q\right)^2\right] = \underbrace{\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q + \eta_q\right)^2\right]}_{(a)} - 2\underbrace{\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q + \eta_q\right)\left(\widetilde{\mathbf{x}}_q^\top \widehat{\mathbf{w}}\right)\right]}_{(b)} + \underbrace{\mathbb{E}\left[\left(\widehat{\mathbf{w}}^\top \widetilde{\mathbf{x}}_q\right)^2\right]}_{(c)}, \tag{18}$$

so it suffices to analyze each individual term.

Analyzing (a)**.** We first evaluate $\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q + \eta_q\right)^2\right]$. First, we note

$$\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q + \eta_q\right)^2\right] = \mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q\right)^2\right] + 2\mathbb{E}\left[\eta_q \mathbf{w}^\top \widetilde{\mathbf{x}}_q\right] + \mathbb{E}\left[\eta_q^2\right] = \mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q\right)^2\right] + \sigma^2,$$

so it suffices to analyze $\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q\right)^2\right]$. By law of total expectation and the fact that $\mathbf{w}, \widetilde{\mathbf{x}}_q$ are independent,

$$\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q\right)^2\right] = \mathbb{E}_{\widetilde{\mathbf{x}}_q}\left[\mathbb{E}_{\mathbf{w}}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q\right)^2 \mid \widetilde{\mathbf{x}}_q\right]\right].$$

Conditioned on $\widetilde{\mathbf{x}}_q$, $\mathbf{w}^\top \widetilde{\mathbf{x}}_q \sim \mathcal{N}(0, \|\widetilde{\mathbf{x}}_q\|^2)$, so $\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q\right)^2 \mid \widetilde{\mathbf{x}}_q\right] = \mathrm{Var}\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q \mid \mathbf{w}\right) = \|\widetilde{\mathbf{x}}_q\|^2$. Therefore,

$$\mathbb{E}_{\mathbf{w}}\left[\mathbb{E}_{\mathbf{x}}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q\right)^2 \mid \mathbf{w}\right]\right] = \mathbb{E}\left[\|\widetilde{\mathbf{x}}_q\|^2\right] = \mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top\right]\right) = \mathrm{Tr}\left(\mathbf{\Sigma}_t\right).$$

Therefore,

$$\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q + \eta_q\right)^2\right] = \mathrm{Tr}\left(\mathbf{\Sigma}_t\right) + \sigma^2.$$

Analyzing (b)**.** Next, we analyze $\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q + \eta_q\right)\left(\widetilde{\mathbf{x}}_q^\top \widehat{\mathbf{w}}\right)\right]$. We first note

$$\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q + \eta_q\right)\left(\widetilde{\mathbf{x}}_q^\top \widehat{\mathbf{w}}\right)\right] = \mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q\right)\left(\widetilde{\mathbf{x}}_q^\top \widehat{\mathbf{w}}\right)\right] + \underbrace{\mathbb{E}\left[\eta_q \widetilde{\mathbf{x}}_q^\top \widehat{\mathbf{w}}\right]}_{=0} = \mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q\right)\left(\widetilde{\mathbf{x}}_q^\top \widehat{\mathbf{w}}\right)\right],$$

so it suffices to analyze $\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q\right)\left(\widetilde{\mathbf{x}}_q^\top \widehat{\mathbf{w}}\right)\right]$. Note $\widehat{\mathbf{w}} := \frac{1}{m}\mathbf{A}\mathbf{X}_{te}^\top \mathbf{y}_{te} = \frac{1}{m}\mathbf{A}\mathbf{X}_{te}^\top(\mathbf{X}_{te}\mathbf{w} + \boldsymbol{\eta}_{te})$, so

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q\right)\left(\widetilde{\mathbf{x}}_q^\top \widehat{\mathbf{w}}\right)\right] = \frac{1}{m}\mathbb{E}\left[\mathbf{w}^\top \widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top(\mathbf{X}_{te}\mathbf{w} + \boldsymbol{\eta}_{te})\right] \\
&= \frac{1}{m}\Big(\mathbb{E}\left[\mathbf{w}^\top \widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{w}\right] + \mathbb{E}\left[\mathbf{w}^\top \widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te}\right]\Big) \\
&= \frac{1}{m}\Big(\mathbb{E}\left[\mathbf{w}^\top \widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{w}\right] + \underbrace{\mathbb{E}\left[\mathbf{w}^\top \widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top\right]\mathbb{E}\left[\boldsymbol{\eta}_{te}\right]}_{=0}\Big) \\
&= \frac{1}{m}\mathbb{E}\left[\mathbf{w}^\top \widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{w}\right] = \frac{1}{m}\mathbb{E}\Big[\mathrm{Tr}\left(\mathbf{w}\mathbf{w}^\top \widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\right)\Big] \\
&= \frac{1}{m}\mathrm{Tr}\left(\mathbb{E}\left[\mathbf{w}\mathbf{w}^\top \widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\right]\right) \\
&= \frac{1}{m}\mathrm{Tr}\Big(\underbrace{\mathbb{E}\left[\mathbf{w}\mathbf{w}^\top\right]}_{\mathbf{I}_d}\underbrace{\mathbb{E}\left[\widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top\right]}_{\mathbf{\Sigma}_t}\mathbf{A}\underbrace{\mathbb{E}\left[\mathbf{X}_{te}^\top \mathbf{X}_{te}\right]}_{m\cdot\mathbf{\Sigma}_t}\Big) = \mathrm{Tr}\Big(\mathbf{\Sigma}_t^2 \mathbf{A}\Big),
\end{aligned}$$

where again $\mathbf{A} = \mathbf{\Sigma}_s^{-1/2}\bar{\mathbf{A}}\mathbf{\Sigma}_s^{-1/2}$ and $\bar{\mathbf{A}} = \left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\mathbf{\Sigma}_s^{-1}\right)^{-1}$.

Analyzing (c)**.** Finally, we analyze $\mathbb{E}\left[(\widetilde{\mathbf{x}}_q^\top \widehat{\mathbf{w}})^2\right]$:

$$\begin{aligned}
&\mathbb{E}\left[(\widetilde{\mathbf{x}}_q^\top \widehat{\mathbf{w}})^2\right] = \frac{1}{m^2}\mathbb{E}\left[\left(\widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top(\mathbf{X}_{te}\mathbf{w} + \boldsymbol{\eta}_{te})\right)^2\right] = \mathbb{E}\left[\left(\widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{w} + \widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te}\right)^2\right] \\
&= \frac{1}{m^2}\Big(\mathbb{E}\left[(\widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{w})^2\right] + 2\underbrace{\mathbb{E}\left[(\widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{w})(\widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te})\right]}_{=0} + \mathbb{E}\left[(\widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te})^2\right]\Big) \\
&= \frac{1}{m^2}\Big(\underbrace{\mathbb{E}\left[\widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{w}\mathbf{w}^\top \mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \widetilde{\mathbf{x}}_q\right]}_{(d)} + \underbrace{\mathbb{E}\left[\widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \boldsymbol{\eta}_{te}\boldsymbol{\eta}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \widetilde{\mathbf{x}}_q\right]}_{(e)}\Big).
\end{aligned}$$

We first focus on (d):

$$\begin{aligned}
&\mathbb{E}\left[\widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{w}\mathbf{w}^\top \mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \widetilde{\mathbf{x}}_q\right] = \frac{1}{m^2}\mathbb{E}\Big[\mathrm{Tr}\left(\mathbf{w}\mathbf{w}^\top \mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\right)\Big] \\
&= \mathrm{Tr}\Big(\underbrace{\mathbb{E}\left[\mathbf{w}\mathbf{w}^\top\right]}_{\mathbf{I}_d}\mathbb{E}\left[\mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top \mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top\right]\Big) \\
&= \mathbb{E}\Big[\mathrm{Tr}\left(\mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top \mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top\right)\Big] = \mathbb{E}\Big[\mathrm{Tr}\left(\widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top \mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\right)\Big] \\
&= \mathrm{Tr}\Big(\underbrace{\mathbb{E}\left[\widetilde{\mathbf{x}}_q \widetilde{\mathbf{x}}_q^\top\right]}_{\mathbf{\Sigma}_t}\mathbb{E}\left[\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\right]\Big) = \mathbb{E}\Big[\mathrm{Tr}\left(\mathbf{\Sigma}_t \mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top \mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\right)\Big] \\
&= \mathbb{E}\Big[\mathrm{Tr}\left(\mathbf{A}\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{\Sigma}_t^{1/2}\mathbf{\Sigma}_t^{1/2}\mathbf{X}_{te}^\top \mathbf{X}_{te}\mathbf{A}^\top\right)\Big] := \mathbb{E}\Big[\mathrm{Tr}\left(\widehat{\mathbf{X}}_{te}^\top \overline{\mathbf{X}}_{te}\overline{\mathbf{X}}_{te}^\top \widehat{\mathbf{X}}_{te}\right)\Big],
\end{aligned}$$

where $\widehat{\mathbf{X}}_{te}^\top := \mathbf{A}\mathbf{X}_{te}^\top$ and $\overline{\mathbf{X}}_{te}^\top := \mathbf{\Sigma}_t^{1/2}\mathbf{X}_{te}^\top$. Note $\widehat{\mathbf{X}}_{te}^\top = \begin{bmatrix}\mathbf{A}\widetilde{\mathbf{x}}_1 & \dots & \mathbf{A}\widetilde{\mathbf{x}}_m\end{bmatrix} := \begin{bmatrix}\widehat{\mathbf{x}}_1 & \dots & \widehat{\mathbf{x}}_m\end{bmatrix}$ where $\widehat{\mathbf{x}}_i := \mathbf{A}\widetilde{\mathbf{x}}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}_d, \mathbf{A}\mathbf{\Sigma}_t\mathbf{A}^\top)$, and $\overline{\mathbf{X}}_{te}^\top = \begin{bmatrix}\mathbf{\Sigma}_t^{1/2}\widetilde{\mathbf{x}}_1 & \dots & \mathbf{\Sigma}_t^{1/2}\widetilde{\mathbf{x}}_m\end{bmatrix} := \begin{bmatrix}\overline{\mathbf{x}}_1 & \dots & \overline{\mathbf{x}}_m\end{bmatrix}$ where $\overline{\mathbf{x}}_i := \mathbf{\Sigma}_t^{1/2}\widetilde{\mathbf{x}}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}_d, \mathbf{\Sigma}_t^2)$. We can express $\widehat{\mathbf{X}}_{te}^\top\overline{\mathbf{X}}_{te}$ and $\overline{\mathbf{X}}_{te}^\top\widehat{\mathbf{X}}_{te}$ as such:

$$\widehat{\mathbf{X}}_{te}^\top\overline{\mathbf{X}}_{te} = \sum_{i=1}^m \widehat{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top \text{ and } \overline{\mathbf{X}}_{te}^\top\widehat{\mathbf{X}}_{te} = \sum_{j=1}^m \overline{\mathbf{x}}_j\widehat{\mathbf{x}}_j^\top.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\Big[\operatorname{Tr}\left(\widehat{\mathbf{X}}_{te}^\top\overline{\mathbf{X}}_{te}\overline{\mathbf{X}}_{te}^\top\widehat{\mathbf{X}}_{te}\right)\Big] = \operatorname{Tr}\left(\mathbb{E}\left[\widehat{\mathbf{X}}_{te}^\top\overline{\mathbf{X}}_{te}\overline{\mathbf{X}}_{te}^\top\widehat{\mathbf{X}}_{te}\right]\right)\\
&= \operatorname{Tr}\left(\sum_{i=1}^m\sum_{j=1}^m \mathbb{E}\left[\widehat{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_j\widehat{\mathbf{x}}_j^\top\right]\right)\\
&= \operatorname{Tr}\left(\sum_{i=1}^m\sum_{j\neq i} \mathbb{E}\left[\widehat{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_j\widehat{\mathbf{x}}_j^\top\right]\right) + \operatorname{Tr}\left(\sum_{i=1}^m \mathbb{E}\left[\widehat{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_i\widehat{\mathbf{x}}_i^\top\right]\right)
\end{aligned}$$

We first consider the case when $i \neq j$. In this setting, $\mathbf{x}_i$ and $\mathbf{x}_j$ are independent, so

$$\mathbb{E}\left[\widehat{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_j\widehat{\mathbf{x}}_j^\top\right] = \mathbb{E}\left[\widehat{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\right]\mathbb{E}\left[\overline{\mathbf{x}}_j\widehat{\mathbf{x}}_j^\top\right] = \mathbf{A}\underbrace{\mathbb{E}\left[\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\right]}_{\mathbf{\Sigma}_t}\mathbf{\Sigma}_t\underbrace{\mathbb{E}\left[\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^\top\right]}_{\mathbf{\Sigma}_t}\mathbf{A}^\top = \mathbf{A}\mathbf{\Sigma}_t^3\mathbf{A}^\top.$$

Therefore,

$$\operatorname{Tr}\left(\sum_{i=1}^m\sum_{j\neq i} \mathbb{E}\left[\widehat{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_j\widehat{\mathbf{x}}_j^\top\right]\right) = m\cdot(m-1)\cdot\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^3\mathbf{A}^\top\right).$$

We now consider the case where $i = j$:

$$\begin{aligned}
&\operatorname{Tr}\left(\sum_{i=1}^m \mathbb{E}\left[\widehat{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_i\widehat{\mathbf{x}}_i^\top\right]\right) = \sum_{i=1}^m \mathbb{E}\Big[\operatorname{Tr}\left(\widehat{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_i\widehat{\mathbf{x}}_i^\top\right)\Big]\\
&= \sum_{i=1}^m \mathbb{E}\left[\widehat{\mathbf{x}}_i^\top\widehat{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_i\right] = \sum_{i=1}^m \mathbb{E}\left[(\widetilde{\mathbf{x}}_i^\top\mathbf{A}^\top\mathbf{A}\widetilde{\mathbf{x}}_i)(\widetilde{\mathbf{x}}_i^\top\mathbf{\Sigma}_t\widetilde{\mathbf{x}}_i)\right]\\
&\overset{(i)}{=} m\cdot\left(2\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^3\mathbf{A}^\top\right) + \operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t\mathbf{A}^\top\right)\operatorname{Tr}\left(\mathbf{\Sigma}_t^2\right)\right),
\end{aligned}$$

where (i) is because for $\mathbf{a}\sim\mathcal{N}(\mathbf{0}_d, \mathbf{\Sigma}_t)$ and fixed $\mathbf{Q}, \mathbf{R}\in\mathbb{R}^{d\times d}$, $\mathbb{E}\left[(\mathbf{a}^\top\mathbf{Q}\mathbf{a})(\mathbf{a}^\top\mathbf{R}\mathbf{a})\right] = \operatorname{Tr}\left(\mathbf{Q}\mathbf{\Sigma}_t(\mathbf{R}+\mathbf{R}^\top)\mathbf{\Sigma}_t\right) + \operatorname{Tr}\left(\mathbf{Q}\mathbf{\Sigma}_t\right)\operatorname{Tr}\left(\mathbf{R}\mathbf{\Sigma}_t\right)$ (see Section 8.2.4 in Petersen et al. (2008)).

We now focus on (e):

$$\begin{aligned}
&\mathbb{E}\left[\widetilde{\mathbf{x}}_q^\top\mathbf{A}\mathbf{X}_{te}^\top\boldsymbol{\eta}_{te}\boldsymbol{\eta}_{te}^\top\mathbf{X}_{te}\mathbf{A}^\top\widetilde{\mathbf{x}}_q\right] = \mathbb{E}\Big[\operatorname{Tr}\left(\widetilde{\mathbf{x}}_q\widetilde{\mathbf{x}}_q^\top\mathbf{A}\mathbf{X}_{te}^\top\boldsymbol{\eta}_{te}\boldsymbol{\eta}_{te}^\top\mathbf{X}_{te}\mathbf{A}^\top\right)\Big]\\
&= \operatorname{Tr}\left(\underbrace{\mathbb{E}\left[\widetilde{\mathbf{x}}_q\widetilde{\mathbf{x}}_q^\top\right]}_{\mathbf{\Sigma}_t}\mathbb{E}\left[\mathbf{A}\mathbf{X}_{te}^\top\boldsymbol{\eta}_{te}\boldsymbol{\eta}_{te}^\top\mathbf{X}_{te}\mathbf{A}^\top\right]\right) = \operatorname{Tr}\left(\mathbb{E}\left[\mathbf{\Sigma}_t^{1/2}\mathbf{A}\mathbf{X}_{te}^\top\boldsymbol{\eta}_{te}\boldsymbol{\eta}_{te}^\top\mathbf{X}_{te}\mathbf{A}^\top\mathbf{\Sigma}_t^{1/2}\right]\right)\\
&:= \operatorname{Tr}\left(\mathbb{E}\left[\widetilde{\boldsymbol{\eta}}_{te}\widetilde{\boldsymbol{\eta}}_{te}^\top\right]\right),
\end{aligned}$$

where $\widetilde{\boldsymbol{\eta}}_{te} := \mathbf{\Sigma}_t^{1/2}\mathbf{A}\mathbf{X}_{te}^\top\boldsymbol{\eta}_{te} := \widetilde{\mathbf{X}}_{te}^\top\boldsymbol{\eta}_{te}$. Note the columns of $\widetilde{\mathbf{X}}_{te}^\top$ are iid Gaussian with covariance $\mathbf{A}\mathbf{\Sigma}_t^2\mathbf{A}^\top$. By Corollary 1 in Mattei (2017), $\widetilde{\boldsymbol{\eta}}_{te}\sim\mathrm{GAL}_d(2\sigma^2\mathbf{A}\mathbf{\Sigma}_t\mathbf{A}^\top, \mathbf{0}_d, m/2)$, where $\mathrm{GAL}_p(\mathbf{\Sigma}, \boldsymbol{\mu}, s)$ denotes a p-dimensional *multivariate generalized asymmetric Laplace distribution* with mean $s\boldsymbol{\mu}$ and covariance $s(\mathbf{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^\top)$ (Definition 1 and Proposition 2 in Mattei (2017)). Therefore,

$$\operatorname{Tr}\left(\mathbb{E}\left[\widetilde{\boldsymbol{\eta}}_{te}\widetilde{\boldsymbol{\eta}}_{te}^\top\right]\right) = m\sigma^2\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^2\mathbf{A}^\top\right).$$

Adding (a), (b), and (c). Adding the expressions for (a), (b), and (c), where $(c) = (d) + (e)$, yields and combining like terms yields the following expression:

$$\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q + \eta_q - \widehat{\mathbf{w}}^\top \widetilde{\mathbf{x}}_q\right)^2\right] = \underbrace{\operatorname{Tr}\left(\mathbf{\Sigma}_t\right) + \sigma^2}_{=(a)} - 2\underbrace{\operatorname{Tr}\left(\mathbf{\Sigma}_t^2 \mathbf{A}\right)}_{=(b)}$$
$$+ \underbrace{\underbrace{\frac{m(m-1)}{m^2}\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^3\mathbf{A}^\top\right) + \frac{2}{m}\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^3\mathbf{A}^\top\right) + \frac{1}{m}\operatorname{Tr}(\mathbf{\Sigma}_t^2)\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t\mathbf{A}^\top\right)}_{=(d)} + \underbrace{\frac{\sigma^2}{m}\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^2\mathbf{A}^\top\right)}_{=(e)}}_{=(c)}.$$

Combining like terms and using the fact $M_t = \operatorname{Tr}\left(\mathbf{\Sigma}_t\right) + \sigma^2$ yields

$$\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q + \eta_q - \widehat{\mathbf{w}}^\top \widetilde{\mathbf{x}}_q\right)^2\right] = M_t - 2\operatorname{Tr}\left(\mathbf{\Sigma}_t^2\mathbf{A}\right) + \frac{m+1}{m}\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^3\mathbf{A}\right)$$
$$+ \frac{1}{m}\operatorname{Tr}\left(\mathbf{\Sigma}_t^2\right)\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t\mathbf{A}\right) + \frac{\sigma^2}{m}\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^2\mathbf{A}\right),$$

which is exactly Equation (17). This completes the proof. □

B.2.2 Proof of Proposition 3

Proof. From Lemma 2, recall that the test risk is given by

$$\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q + \eta_q - \widehat{\mathbf{w}}^\top \widetilde{\mathbf{x}}_q\right)^2\right] = \operatorname{Tr}\left(\mathbf{\Sigma}_t\right) + \sigma^2 - 2\operatorname{Tr}\left(\mathbf{\Sigma}_t^2\mathbf{A}\right) + \frac{m+1}{m}\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^3\mathbf{A}\right)$$
$$+ \frac{1}{m}\operatorname{Tr}(\mathbf{\Sigma}_t^2)\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t\mathbf{A}\right) + \frac{\sigma^2}{m}\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^2\mathbf{A}\right)$$

where $M_t = \operatorname{Tr}\left(\mathbf{\Sigma}_t\right) + \sigma^2$, $\bar{\mathbf{A}} = \left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\mathbf{\Sigma}_s^{-1}\right)^{-1}$, and $\mathbf{A} = \mathbf{\Sigma}_s^{-1/2}\bar{\mathbf{A}}\mathbf{\Sigma}_s^{-1/2}$. Since $\mathbf{\Sigma}_s = \mathbf{U}_s\mathbf{U}_s^\top + \epsilon\mathbf{I}_d$, we can write its eigendecomposition as such:

$$\mathbf{\Sigma}_s = \mathbf{U}_s\mathbf{U}_s^\top + \epsilon\mathbf{I}_d = \mathbf{U}\begin{bmatrix}\mathbf{I}_r & \\ & \mathbf{0}_{(d-r)\times(d-r)}\end{bmatrix}\mathbf{U}^\top + \epsilon\mathbf{I} = \mathbf{U}\begin{bmatrix}(1+\epsilon)\mathbf{I}_r & \\ & \epsilon\mathbf{I}_{d-r}\end{bmatrix}\mathbf{U}^\top,$$

where $\mathbf{U} = \begin{bmatrix}\mathbf{U}_s & \mathbf{U}_{s,\perp} & \mathbf{U}_{2r,\perp}\end{bmatrix}$. Then, we can write $\mathbf{A}$ as follows:

$$\mathbf{A} = \mathbf{\Sigma}_s^{-1/2}\left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\mathbf{\Sigma}_s^{-1}\right)^{-1}\mathbf{\Sigma}_s^{-1/2} = \left(\frac{n+1}{n}\mathbf{\Sigma}_s + \frac{M_s}{n}\mathbf{I}_d\right)^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \tag{19}$$

where

$$\mathbf{\Lambda} = \begin{bmatrix}\frac{n}{(n+1)(1+\epsilon)+M_s}\cdot\mathbf{I}_r & \\ & \frac{n}{(n+1)\epsilon+M_s}\cdot\mathbf{I}_{d-r}\end{bmatrix} := \begin{bmatrix}\nu_1\mathbf{I}_r & \\ & \nu_2\mathbf{I}_{d-r}\end{bmatrix}.$$

We simplify the $\operatorname{Tr}(\cdot)$ terms using Equation (19). Before proceeding, note

$$\lim_{n\to\infty}\nu_1 = \frac{1}{1+\epsilon} \quad \text{and} \quad \lim_{n\to\infty}\nu_2 = \frac{1}{\epsilon}, \tag{20}$$

as well as

$$\lim_{m\to\infty}\mathbb{E}\left[\left(\mathbf{w}^\top \widetilde{\mathbf{x}}_q + \eta_q - \widehat{\mathbf{w}}^\top \widetilde{\mathbf{x}}_q\right)^2\right] = \operatorname{Tr}\left(\mathbf{\Sigma}_t\right) + \sigma^2 - 2\operatorname{Tr}\left(\mathbf{\Sigma}_t^2\mathbf{A}\right) + \operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^3\mathbf{A}\right), \tag{21}$$

so we only focus on the $\operatorname{Tr}(\mathbf{\Sigma}_t)$, $\operatorname{Tr}(\mathbf{\Sigma}_t^2\mathbf{A})$, and $\operatorname{Tr}(\mathbf{A}\mathbf{\Sigma}_t^3\mathbf{A})$ terms. Also, recall $\mathbf{\Sigma}_t = \mathbf{U}_t\mathbf{U}_t^\top + \epsilon\mathbf{I}_d$, so

$$\mathbf{\Sigma}_t^2 = \left(\mathbf{U}_t\mathbf{U}_t^\top + \epsilon\mathbf{I}_d\right)\left(\mathbf{U}_t\mathbf{U}_t^\top + \epsilon\mathbf{I}_d\right) = (1+2\epsilon)\mathbf{U}_t\mathbf{U}_t^\top + \epsilon^2\mathbf{I}_d, \quad \text{and} \tag{22}$$
$$\mathbf{\Sigma}_t^3 = \left(\mathbf{U}_t\mathbf{U}_t^\top + \epsilon\mathbf{I}_d\right)\left((1+2\epsilon)\mathbf{U}_t\mathbf{U}_t^\top + \epsilon^2\mathbf{I}_d\right) = \left((1+2\epsilon)(1+\epsilon)+\epsilon^2\right)\mathbf{U}_t\mathbf{U}_t^\top + \epsilon^3\mathbf{I}_d. \tag{23}$$

Simplifying $\mathrm{Tr}(\mathbf{\Sigma}_t^2\mathbf{A})$. Using Equation (19) and Equation (22) yields

$$\begin{aligned}\mathrm{Tr}\left(\mathbf{\Sigma}_t^2\mathbf{A}\right) &= (1+2\epsilon)\,\mathrm{Tr}\left(\mathbf{U}_t^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\mathbf{U}_t\right) + \epsilon^2\,\mathrm{Tr}(\mathbf{A}) \\ &= (1+2\epsilon)\left(r\nu_1\cos^2(\theta) + r\nu_2\sin^2(\theta)\right) + \epsilon^2\left(r\nu_1 + (d-r)\nu_2\right).\end{aligned}$$

Simplifying $\mathrm{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^3\mathbf{A}\right)$. Using Equation (19) and Equation (23) yields

$$\begin{aligned}\mathrm{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t^3\mathbf{A}\right) = \mathrm{Tr}\left(\mathbf{\Sigma}_t^3\mathbf{A}^2\right) &= \left((1+2\epsilon)(1+\epsilon)+\epsilon^2\right)\mathrm{Tr}\left(\mathbf{U}_t\mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^\top\mathbf{U}_t\right) + \epsilon^3\,\mathrm{Tr}\left(\mathbf{A}^2\right) \\ &= \left((1+2\epsilon)(1+\epsilon)+\epsilon^2\right)\left(r\nu_1^2\cos^2(\theta) + r\nu_2^2\sin^2(\theta)\right) + \epsilon^2\left(r\nu_1^2 + (d-r)\nu_2^2\right).\end{aligned}$$

Simplifying the Test Risk. Combining Equation (20) and Equation (21), and then substituting the expressions for the $\mathrm{Tr}(\cdot)$ terms yields

$$\begin{aligned}\lim_{m\to\infty}\lim_{n\to\infty}\mathbb{E}\left[\left(\mathbf{w}^\top\widetilde{\mathbf{x}}_q + \eta_q - \widehat{\mathbf{w}}^\top\widetilde{\mathbf{x}}_q\right)^2\right] &= r + \sigma^2 - 2(1+2\epsilon)\left(\frac{r\cos^2(\theta)}{1+\epsilon} + \frac{r\sin^2(\theta)}{\epsilon}\right) \\ &\quad + \left((1+2\epsilon)(1+\epsilon)+\epsilon^2\right)\left(\frac{r\cos^2(\theta)}{(1+\epsilon)^2} + \frac{r\sin^2(\theta)}{\epsilon^2}\right) + \mathcal{O}(\epsilon).\end{aligned}$$

Letting $c_1 := \left((1+2\epsilon)(1+\epsilon)+\epsilon^2\right)$ and $c_2 := 2(1+\epsilon)$ yields

$$\lim_{m\to\infty}\lim_{n\to\infty}\mathbb{E}\left[\left(\mathbf{w}^\top\widetilde{\mathbf{x}}_q + \eta_q - \widehat{\mathbf{w}}^\top\widetilde{\mathbf{x}}_q\right)^2\right] = r + \sigma^2 + \left(\frac{c_1 - (1+\epsilon)c_2}{(1+\epsilon)^2}\right)r\cos^2(\theta) + \left(\frac{c_1 - \epsilon c_2}{\epsilon^2}\right)r\sin^2(\theta) + \mathcal{O}(\epsilon),$$

which completes the proof.

□

B.3 Auxiliary Results

Here, we provide auxiliary results to support the proofs in Sections B.1 and B.2.

B.3.1 Optimal Linear Attention Weights

We first provide results on the form of the weights matrices after training a single-layer linear attention model on the objective Equation (1). The following results are largely inspired by Theorem 1 in Li et al. (2024b), but are slightly different since we consider a normalization factor of $1/n$ in our linear attention model.

Lemma 3 (Optimal Attention Weights (Li et al., 2024b))**.** *Consider the independent data model in Equation (4) with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_s)$, and let $n \in \mathbb{N}$ denote the in-context prompt length used at training. Then, the optimal linear attention weights obtained by minimizing the loss in Equation (1) are given by*

$$\mathbf{W}_K^\star = \mathbf{W}_V^\star = \mathbf{I}_{d+1}, \quad \text{and} \quad \mathbf{W}_Q^\star = \begin{bmatrix}\mathbf{A} & \mathbf{0}_d \\ \mathbf{0}_d^\top & 0\end{bmatrix} \tag{24}$$

where $\mathbf{A} = \left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\mathbf{\Sigma}_s^{-1}\right)^{-1}$ and $M_s = \mathrm{Tr}(\mathbf{\Sigma}_s) + \sigma^2$, with empirical risk $\mathcal{L}_s^\star = M_s - \mathrm{Tr}\left(\mathbf{\Sigma}_s\mathbf{A}\right)$.

Proof. The proof is the same as that of Theorem 1 in Li et al. (2024b) by absorbing the $1/n$ factor into $\mathbf{W}_Q$. □

Lemma 4 (Optimal Attention Weights for Mixture of K Gaussians)**.** *Consider the independent data model in Equation (4) with $\mathbf{w} \sim \sum_{k=1}^K \gamma_k \cdot \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{s,k}\right)$ for $\gamma_k \in (0,1)$ for all $k \in [K]$ and $\sum_{k=1}^K \gamma_k = 1$. Let $n \in \mathbb{N}$ denote the in-context prompt length used at training. Define $\mathbf{\Sigma} = \sum_{k=1}^K \gamma_k \cdot \mathbf{\Sigma}_{s,k}$. Then, the optimal linear attention weights obtained by minimizing the loss in Equation (1) are given by*

$$\mathbf{W}_K^\star = \mathbf{W}_V^\star = \mathbf{I}_{d+1}, \quad \text{and} \quad \mathbf{W}_Q^\star = \begin{bmatrix}\mathbf{A} & \mathbf{0}_d \\ \mathbf{0}_d^\top & 0\end{bmatrix}, \tag{25}$$

where $\mathbf{A} = \left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\mathbf{\Sigma}^{-1}\right)^{-1}$ and $M_s = \mathrm{Tr}(\mathbf{\Sigma}) + \sigma^2$, with empirical risk $\mathcal{L}_s^\star = M_s - \mathrm{Tr}\left(\mathbf{\Sigma}\mathbf{A}\right)$.

Proof. It is straightforward to see that if $\mathbf{w} \sim \sum_{k=1}^{K} \gamma_k \gamma \cdot \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{s,k})$, then

$$\mathbf{\Sigma}_s \coloneqq \mathrm{Cov}(\mathbf{w}) = \sum_{k=1}^{K} \gamma_k \cdot \mathbf{\Sigma}_{s,k}.$$

Then, the proof is equivalent to that of Lemma 3 under the new form of $\mathbf{\Sigma}_s$. □

Lemma 5 (Optimal Attention Weights for Non-Isotropic Features Li et al. (2024b))**.** *Consider the independent data model in Equation (11) where the features are drawn from $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_s)$ for all $i \in [n]$, and $n \in \mathbb{N}$ denotes the in-context prompt length used at training. The optimal linear attention weights obtained by minimizing the loss in Equation (1) are given by*

$$\mathbf{W}_K^\star = \mathbf{W}_V^\star = \mathbf{I}_{d+1}, \quad \textit{and} \quad \mathbf{W}_Q^\star = \begin{bmatrix} \mathbf{A} & \mathbf{0}_d \\ \mathbf{0}_d^\top & 0 \end{bmatrix}, \tag{26}$$

where $\mathbf{A} = \mathbf{\Sigma}_s^{-1/2} \bar{\mathbf{A}} \mathbf{\Sigma}_s^{-1/2}$, $\bar{\mathbf{A}} = \left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\mathbf{\Sigma}_s^{-1}\right)^{-1}$ and $M_s = \mathrm{Tr}(\mathbf{\Sigma}_s) + \sigma^2$, with empirical risk $\mathcal{L}_s^\star = M_s - \mathrm{Tr}(\mathbf{\Sigma}_s^\top \bar{\mathbf{A}})$.

Proof. The proof is the same as that of Lemma 3 and Theorem 1 in Li et al. (2024b). □

B.3.2 Miscellaneous Results

Lemma 6. *Let $0 \prec \mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ and $c, k > 0$ be constants. Then,*

$$\left(c \cdot \mathbf{I}_d + k \cdot \mathbf{\Sigma}^{-1}\right)^{-1} = \mathbf{V} \begin{bmatrix} \frac{\lambda_1}{c \cdot \lambda_1 + k} & 0 & \dots & 0 \\ 0 & \frac{\lambda_2}{c \cdot \lambda_2 + k} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\lambda_d}{c \cdot \lambda_d + k} \end{bmatrix} \mathbf{V}^\top, \tag{27}$$

where $\mathbf{V} \in \mathbb{R}^{d \times d}$ is an orthonormal matrix whose columns are eigenvectors of $\mathbf{\Sigma}$, and λ_i is the i^{th} largest eigenvalue of $\mathbf{\Sigma}$.

Proof. Since $\mathbf{\Sigma} \succ 0$, there exists an eigendecomposition $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ such that $\mathbf{V}$ is an orthonormal matrix and $\mathbf{\Lambda}$ is a diagonal matrix consisting of the real, positive eigenvalues of $\mathbf{\Sigma}$, denoted as $\lambda_1, \lambda_2, \dots, \lambda_d$. Thus,

$$\mathbf{\Sigma}^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^\top \implies c \cdot \mathbf{I}_d + k \cdot \mathbf{\Sigma}^{-1} = \mathbf{V} \underbrace{\begin{bmatrix} c + \frac{k}{\lambda_1} & 0 & \dots & 0 \\ 0 & c + \frac{k}{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & c + \frac{k}{\lambda_d} \end{bmatrix}}_{\widetilde{\mathbf{\Lambda}}} \mathbf{V}^\top$$

$$\implies \left(c \cdot \mathbf{I}_d + k \cdot \mathbf{\Sigma}^{-1}\right)^{-1} = \mathbf{V}\widetilde{\mathbf{\Lambda}}^{-1}\mathbf{V}^\top,$$

which completes the proof. □