Towards Understanding Out-of-Distribution Generalization for In-Context Learning via Low-Dimensional Subspaces

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.

1 INTRODUCTION

The remarkable capability of ICL in transformer-based large language models (LLMs) (Vaswani et al., 2017) has sparked a wide range of both empirical and theoretical research dedicated to understanding its foundations (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; 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). However, its generalization capabilities, particularly whether ICL can generalize beyond its pre-training data, remain unclear. There are conflicting views in the literature: Garg et al. (2022); Zhang et al. (2024) showed that ICL is relatively robust to distribution shifts in several settings, while Wang et al. (2025) empirically demonstrated that ICL can only solve in-distribution (language) tasks in general. To resolve these contrasting perspectives, Goddard et al. (2025) presented yet another setting in which ICL can generalize OOD, attributing this capability to pre-training task diversity. However, the complex nature of their definition of the task vector made theoretical analysis difficult, restricting their study to empirical results. Overall, these observations 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. 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.

2 PROBLEM SETUP

2.1 Preliminaries

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 via the following input prompt $\mathbf{Z} \in \mathbb{R}^{(n+1)\times(d+1)}$:

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

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, the transformer $g_{\texttt{ATT}}$, parameterized by weights $\mathcal{W}$, is trained by minimizing the following expected squared loss with respect to $\mathcal{W}$:

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

At inference, we test the trained model $g^\star_{\texttt{ATT}}$ using $m+1$ paired examples $\{\mathbf{x}_j, \widetilde{y}_j\}_{j=1}^{m+1}$, where the prompts are constructed in the same manner:

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

Since we are interested in the OOD capabilities of transformers, the labels are generated via $\widetilde{y}_j = \widetilde{f}(\mathbf{x}_j)$ for some function $\widetilde{f} \neq f$, i.e., the task function between training and testing are different.

2.2 Single-Layer Linear Attention

We now introduce the linear attention architecture. First, define the masked input prompts as

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

Given the input sequence $\mathbf{Z}_{\mathcal{M}}$ and query $\mathbf{z}_{n+1}$, the predicted label $\widehat{y}_{n+1}$ is the output of a linear attention model $g_{\texttt{ATT}}$, where

$$g_{\texttt{ATT}}(\mathbf{z}_{n+1}, \mathbf{Z}_{\mathcal{M}}) = \frac{\left(\mathbf{z}_{n+1}^\top \mathbf{W}_Q \mathbf{W}_K^\top \mathbf{Z}_{\mathcal{M}}^\top\right) \mathbf{Z}_{\mathcal{M}} \mathbf{W}_V \mathbf{p}}{n}, \quad (2)$$

$\mathbf{p} = [\mathbf{0}_d \ 1]^\top$, 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, meaning $\mathcal{W} = \{\mathbf{W}_K, \mathbf{W}_Q, \mathbf{W}_V\}$. Then, letting $\mathcal{W}^\star = \{\mathbf{W}_K^\star, \mathbf{W}_Q^\star, \mathbf{W}_V^\star\}$ denote the optimal weights obtained via Equation (1). At inference, the predicted test-time label $\widehat{y}_{m+1}$ is the output of the trained linear attention model $g^\star_{\texttt{att}}\left(\widetilde{\mathbf{z}}_{m+1}, \widetilde{\mathbf{Z}}_{\mathcal{M}}\right)$, where

$$g^\star_{\texttt{att}}\left(\widetilde{\mathbf{z}}_{m+1}, \widetilde{\mathbf{Z}}_{\mathcal{M}}\right) = \frac{\left(\mathbf{z}_{n+1}^\top \mathbf{W}_Q^\star \mathbf{W}_K^{\star\top} \mathbf{Z}_{\mathcal{M}}^\top\right) \mathbf{Z}_{\mathcal{M}} \mathbf{W}_V^\star \mathbf{p}}{m}. \quad (3)$$

2.3 ICL with Linear Regression

We consider linear regression tasks, i.e., $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ for some input $\mathbf{x} \in \mathbb{R}^d$ and task vector $\mathbf{w} \in \mathbb{R}^d$.

Training Distribution. 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:

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

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

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

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

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

or

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

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 in Section A.

Testing Distribution. We 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):

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

where $\theta \in \left[0, \frac{\pi}{2}\right]$ denotes the (r identical) principal angles between $\mathbf{U}_s$ and $\mathbf{U}_t$. Finally, we define the testing covariance $\mathbf{\Sigma}_t$ as

$$\mathbf{\Sigma}_t = \mathbf{U}_t \mathbf{U}_t^\top + \epsilon \cdot \mathbf{I}_d. \quad (6)$$

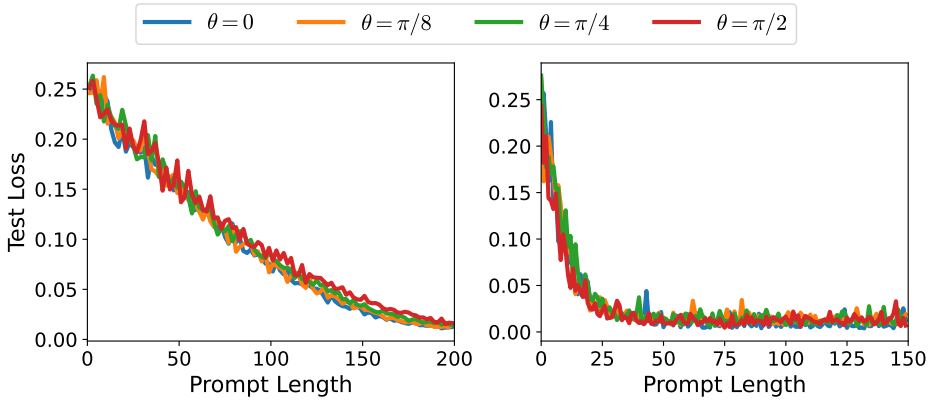

Figure 1: Plot of the test risk as a function of the prompt length for a linear transformer (left) and a GPT-2 model (right) under the data setting of Theorem 1. When the prompt length at test time is large enough, the test risk goes nearly to zero for all $\theta \in \left[0, \frac{\pi}{2}\right]$, corroborating Theorem 1.

We 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 show that when the linear attention model is trained on prompts whose task vectors are sampled via Equation (Union of Subs.), and tested on a task vector sampled via Equation (7) at any $\theta \in [0, \frac{\pi}{2}]$, the test risk can be arbitrarily close to the optimal test risk (the label noise variance). Notably, this test risk is *independent* of θ, implying 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

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

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. We empirically corroborate Theorem 1 in Figure 1 on both linear attention and GPT-2, deferring experimental details to Section B, and the proof to Section C.2.

3.2 Transformers Cannot Generalize When Trained on a Single Subspace

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.

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 vec-*

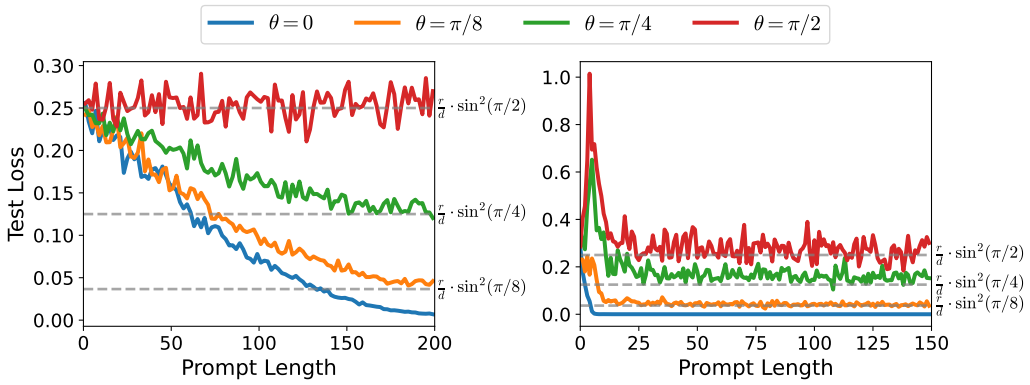

Figure 2: Plot of the normalized test risk as a function of the prompt length for a linear transformer (left) and a GPT-2 model (right) under the data setting of Proposition 1. The test risk exactly matches the predicted risk from Proposition 1.

tors 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 $\mathbf{U}_t \in \mathbb{R}^{d\times r}$.

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, which is the largest possible error in this setting (Garg et al., 2022; Oko et al., 2025). We empirically corroborate Proposition 1 in Figure 2, again deferring experimental details to Section B, and the proof to Section C.3.

4 EXPERIMENTS

In this section, we provide experimental results showing our theoretical results hold beyond linear function classes. We defer specific experimental details to Section B. We look at square-integrable functions under the uniform measure. We construct an orthonormal basis via cosines, i.e., $\psi_n(x) = (1/\sqrt{2})\cos(n\pi x)$ for $n \in \mathbb{N}$. As in previous sections, we consider two settings: observing instances of a single (one-dimensional) subspace, as well as for a union of three (one-dimensional) subspaces. We draw the inputs x via $x \sim \mathcal{U}([0, 1])$. The results are shown in Figure 3: as seen on the left, transformers are not robust to subspace shifts, as the test risk increases with the subspace angle from the training subspace, in accordance with Proposition 1. However, in Figure 3 (right), we have

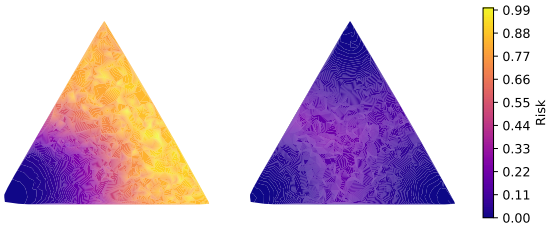

Figure 3: Visualization of the generalization behavior of transformers for learning cosines in-context. Each corner of a triangle represents a one-dimensional subspace spanned by ψ_1 (bottom left), ψ_2 (bottom right), or ψ_3 (top), with all possible convex combinations given by the interior. We show the risk when evaluated at different points in $\text{span}(\{\psi_1, \psi_2, \psi_3\})$ for the appropriate function space. **Left:** train on prompts drawn from $\text{span}(\{\psi_1\})$. **Right:** train on prompts drawn from $\text{span}(\{\psi_1\}) \cup \text{span}(\{\psi_2\}) \cup \text{span}(\{\psi_3\})$.

the generalization behavior described by Theorem 1, where training on the mixture of subspaces results in low risk in the space spanned by the basis vectors.

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.

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.

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.

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*.

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*.

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*.

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

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, 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*.

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

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.

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, 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*.

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.

Supplementary Materials

A Beyond a Mixture of Two Gaussians

We now generalize Theorem 1 to consider a mixture of $K > 2$ Gaussians. 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 $g_{ATT}^\star$ 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

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

The proof is deferred to Section C.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.

B Experimental Details

In this section, we provide specific experimental details for Figures 1 to 3. To generate linear regression data, we set $d = 20$, $r = 5$, $\sigma^2 = 0$, 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 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.

C Deferred Proofs

This section presents all deferred proofs. First, Section C.1 provides auxiliary results used to support both the task and feature shift proofs. Next, Sections C.2 and C.3 provide proofs of Theorems 1 and 2 and proposition 1.

C.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,

$$\widetilde{\mathbf{Z}}_{\mathcal{M}} = \begin{bmatrix}\tilde{\mathbf{z}}_1 & \dots & \tilde{\mathbf{z}}_m & \mathbf{0}\end{bmatrix}^\top = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_m & \mathbf{0} \\ \widetilde{y}_1 & \dots & \widetilde{y}_m & 0\end{bmatrix}^\top \quad \text{and} \quad \tilde{\mathbf{z}}_q = \begin{bmatrix}\mathbf{x}_{m+1} \\ 0\end{bmatrix} := \begin{bmatrix}\mathbf{x}_q \\ 0\end{bmatrix}. \tag{10}$$

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 2, we have

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

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,

$$\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q + \eta_q - \widehat{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right] = \underbrace{\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q + \eta_q\right)^2\right]}_{(a)} - 2\underbrace{\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q + \eta_q\right)\left(\mathbf{x}_q^\top \widehat{\mathbf{w}}\right)\right]}_{(b)} + \underbrace{\mathbb{E}\left[\left(\widehat{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right]}_{(c)}. \tag{11}$$

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):

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

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:

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

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):

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

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 $\mathrm{GAL}_p(\mathbf{\Sigma}, \boldsymbol{\mu}, s)$ 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,

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

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:

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

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 (11). This completes the proof. □

C.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 3, we have

$$\mathbb{E}\left[\left(\widetilde{y} - g^\star_{\texttt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = \left(\frac{1}{m}\operatorname{Tr}\left(\mathbf{A}^\top\mathbf{A}\right) + 1\right)\left(\operatorname{Tr}\left(\overline{\boldsymbol{\Sigma}}_t\right) + \sigma^2\right) - 2\operatorname{Tr}\left(\overline{\boldsymbol{\Sigma}}_t\mathbf{A}\right) + \frac{m+1}{m}\operatorname{Tr}\left(\mathbf{A}\overline{\boldsymbol{\Sigma}}_t\mathbf{A}^\top\right), \quad (12)$$

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 4,

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

where

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

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 $\operatorname{Tr}(\overline{\boldsymbol{\Sigma}}_t)$. We can write $\operatorname{Tr}(\overline{\boldsymbol{\Sigma}}_t)$ as such:

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

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

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

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 $\mathrm{Tr}(\overline{\mathbf{\Sigma}}_t\mathbf{A})$:

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

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 (12) yields

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

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

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

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. □

C.3 Proof of Proposition 1

We now provide the proof of Proposition 1.

Proof. For simplicity, we denote $\widetilde{y} \coloneqq \widetilde{y}_{m+1}$. Recall $\mathbf{U} \coloneqq \begin{bmatrix}\mathbf{U}_s & \mathbf{U}_{s,\perp} & \mathbf{U}_{2r,\perp}\end{bmatrix} \in \mathbb{R}^{d\times d}$, 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 4, we have

$$\mathbf{A} = \left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\mathbf{\Sigma}_s^{-1}\right)^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top, \tag{13}$$

where

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

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

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

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

We simplify the remaining $\operatorname{Tr}(\cdot)$ terms using Equation (13).

Simplifying $\operatorname{Tr}(\mathbf{A})$ and $\operatorname{Tr}(\mathbf{A}^\top\mathbf{A})$. Directly from Equation (13):

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

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

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

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

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,

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

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

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

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

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

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. □

C.4 Auxiliary Results

Here, we provide auxiliary results to support the proofs in Sections C.1 to C.3.

C.4.1 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 2 (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*

$$\mathbf{W}_K^{\star}=\mathbf{W}_V^{\star}=\mathbf{I}_{d+1},\quad \textit{and}\quad \mathbf{W}_Q^{\star}=\begin{bmatrix}\mathbf{A} & \mathbf{0}_d\\ \mathbf{0}_d^{\top} & 0\end{bmatrix} \tag{15}$$

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 3 (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*

$$\mathbf{W}_K^{\star}=\mathbf{W}_V^{\star}=\mathbf{I}_{d+1},\quad \textit{and}\quad \mathbf{W}_Q^{\star}=\begin{bmatrix}\mathbf{A} & \mathbf{0}_d\\ \mathbf{0}_d^{\top} & 0\end{bmatrix}, \tag{16}$$

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 2 under the new form of $\mathbf{\Sigma}_s$. □

C.4.2 Miscellaneous Results

Lemma 4. *Let $0 \prec \mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ and $c, k > 0$ be constants. Then,*

$$\left(c \cdot \mathbf{I}_d + k \cdot \mathbf{\Sigma}^{-1}\right)^{-1} = \mathbf{V} \begin{bmatrix} \frac{\lambda_1}{c\cdot\lambda_1 + k} & 0 & \dots & 0 \\ 0 & \frac{\lambda_2}{c\cdot\lambda_2 + k} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\lambda_d}{c\cdot\lambda_d + k} \end{bmatrix} \mathbf{V}^\top, \tag{17}$$

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. □