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

Soo Min Kwon*, Alec S. Xu*, Can Yaras, Laura Balzano, Qing Qu

Electrical Engineering and Computer Science Department
University of Michigan
Ann Arbor, MI, 48109
kwonsm@umich.edu

1 Introduction

The remarkable capability of ICL in Transformer-based large language models (LLMs) [1] has sparked both empirical [2–10] and theoretical research [7, 8, 11–15]. However, the generalization capabilities of ICL, particularly whether it can extend beyond its pre-training distribution, remain unclear. For example, Garg et al. [2] empirically showed that ICL is relatively robust to distribution shifts in several settings, as the performance of the Transformer closely matched that of the least-squares estimator on linear regression tasks. Zhang et al. [7] shared a similar conclusion, showing that while shifts in the features cannot be tolerated for a one-layer linear attention model, shifts in the regression weights can be handled well. However, Wang et al. [5] challenged these views, empirically demonstrating that ICL can only solve in-distribution tasks in general. These contrasting views, combined with the lack of a theoretical foundation, highlight the need for a rigorous characterization of the OOD generalization capabilities of ICL.

This work proposes a mathematical framework to demystify and quantify the OOD generalization capabilities of ICL. We theoretically study ICL on a single-layer linear attention model with linear regression, where the weight (or task) vectors are sampled from low-dimensional subspaces. This setup enables us to quantify the distribution shift in the task vectors via the principal angles between subspaces and to characterize the OOD test risk as a function of these angles. Then, we precisely identify the conditions on the pre-training task vectors under which the OOD test risk is either sensitive to or independent of these angles, thereby explaining both the limitations and capabilities of ICL. Specifically, we prove that when the training task vectors are drawn from a single r-dimensional subspace, ICL inevitably incurs test error as a function of the principal angle. On the other hand, when the training task vectors are drawn from a union of subspaces, we show that ICL incurs a test risk that is independent of the principal angles. Unlike the single-subspace setting, this result implies that ICL can generalize to any subspace within the span of the training subspaces, even regions with zero probability density under the training distribution. We hypothesize that this explains when ICL exhibits OOD generalization: the testing task vector lies within the span of the training task vectors.

2 Problem Setup and Theoretical Results

Problem Setup. We study a standard ICL task of predicting the next token. For training, we draw a feature and label pair $(\mathbf{x}_i, y_i)$ as follows: let each feature vector be $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. For all $i \in [n+1]$, we generate each label $y_i \in \mathbb{R}$ as such:

$$y_i = \mathbf{w}^\top \mathbf{x}_i + \eta_i \quad \text{where} \quad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_s), \quad \eta_i \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

*Equal Contribution. Correspondence to {kwonsm, alecx}@umich.edu. LB, SK, CY were supported in part by NSF award CCF-1845076 and CCF-2331590.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

$\sigma \geq 0$ is the noise level, and $\mathbf{\Sigma}_s \in \mathbb{R}^{d\times d}$ is the source task covariance matrix, i.e., the covariance of the training weight $\mathbf{w}$, which we often refer to as the "task vector". Then, given $n+1$ paired examples, we train the single-layer linear attention model in Equation (9) by solving Equation (10). We use $g^\star_{\texttt{ATT}}$ and $\mathcal{W}^\star$ to respectively denote the optimal model and weights according to this setup. Then, our main goal is to investigate how distribution shifts in the task vector affect the test risk of the optimal model $g^\star_{\texttt{ATT}}$. To this end, at test time, we draw a feature and label pair $(\mathbf{x}_j, \widetilde{y}_j)$ independent of the training data in a similar fashion: let each feature vector be $\mathbf{x}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. For all $j \in [m+1]$, we generate the label $\widetilde{y}_j \in \mathbb{R}$ according to each $\mathbf{x}_j \in \mathbb{R}^d$ as

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

and $\mathbf{\Sigma}_t \in \mathbb{R}^{d\times d}$ is the target covariance matrix, i.e., the covariance for the task vector at test time. Next, we give forms to $\mathbf{\Sigma}_s$ and $\mathbf{\Sigma}_t$ to quantify the distribution shift from training to test time.

Suppose $d \gg r$, and let $\mathbf{U}_s \in \mathbb{R}^{d\times r}$ be an orthonormal basis for an r-dimensional subspace in $\mathbb{R}^d$. We parameterize $\mathbf{\Sigma}_s$ and $\mathbf{\Sigma}_t$ as follows:

$$\mathbf{\Sigma}_s = \mathbf{U}_s\mathbf{U}_s^\top + \epsilon \cdot \mathbf{I}_d \quad \text{and} \quad \mathbf{\Sigma}_t = \mathbf{U}_t\mathbf{U}_t^\top + \epsilon \cdot \mathbf{I}_d, \tag{3}$$

where $\epsilon > 0$ is a small constant to ensure invertibility, and $\mathbf{U}_t$ is parameterized as [16, Section 3.8]:

$$\mathbf{U}_t = \mathbf{U}_s \cdot \cos(\mathbf{\Theta}) + \mathbf{U}_{s,\perp} \cdot \sin(\mathbf{\Theta}), \tag{4}$$

and $\mathbf{U}_{s,\perp} \in \mathbb{R}^{d\times r}$ is an r-dimensional orthonormal basis is that is *orthogonal* to $\mathbf{U}_s$. 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, by parameterizing $\mathbf{\Sigma}_s$ and $\mathbf{\Sigma}_t$ using $\mathbf{U}_s$ and $\mathbf{U}_t$, changing the value of θ allows us to control how aligned the testing covariance $\mathbf{\Sigma}_t$ is with the training covariance $\mathbf{\Sigma}_s$. In other words, θ measures the distribution shift from training to testing. Our goal is to quantify the OOD test risk of $g^\star_{\texttt{ATT}}$ in terms of θ.

Main Results. In Proposition 1 (available in Appendix B), we prove that even with infinitely many samples, ICL with a single-layer linear attention model exhibits test risk with a non-negligible dependence on the shift between the covariance matrices $\mathbf{\Sigma}_t$ and $\mathbf{\Sigma}_s$, as measured by θ. This also empirically holds for nonlinear models such as GPT-2 (see Figure 2), which demonstrates that ICL is not inherently robust to subspace shifts. However, consider the following covariance matrices:

$$\mathbf{\Sigma}_s = \mathbf{U}_s\mathbf{U}_s^\top + \epsilon \cdot \mathbf{I}_d \quad \text{and} \quad \mathbf{\Sigma}_{s,\perp} = \mathbf{U}_{s,\perp}\mathbf{U}_{s,\perp}^\top + \epsilon \cdot \mathbf{I}_d. \tag{5}$$

Then, instead of the training task vector in Equation (1), consider training $g_{\texttt{ATT}}$ on prompts with labels $y_i = \mathbf{w}^\top \mathbf{x}_i + \eta_i$ whose task vector is drawn from a mixture of two Gaussians:

$$\mathbf{w} \sim \gamma \cdot \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_s) + (1-\gamma) \cdot \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{s,\perp}), \tag{6}$$

where γ is the mixture probability. The following result states that this training procedure mitigates dependence on θ, given that the prompt lengths are sufficiently large.

Theorem 1 (Test Risk under the Span of Covariance Matrices)**.** *Let $g^\star_{\texttt{ATT}}$ denote the optimal linear attention model corresponding to the independent data setting in Equation (1), where the task vector now follows Equation (6) 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*

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

and $\mathbf{\Sigma}_t \in \mathbb{R}^{d\times d}$ is from Equation (3). For any $\theta \in \left[0, \frac{\pi}{2}\right]$ and $\delta \in (0, r)$, we have

$$\lim_{m,n\to\infty} \lim_{\epsilon\to 0} \mathbb{E}\left[\left(\tilde{y}_{m+1} - g^\star_{\mathit{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = \sigma^2. \tag{7}$$

This highlights an interesting property of Transformers: if the pre-training task vectors are drawn from a union of subspaces, then ICL can interpolate to the space between the subspaces. In other words, even if certain regions have zero probability density in the distribution over the training task vector, ICL can still generalize to those regions at test time, as long as they lie within the overall span of the training task vectors. We hypothesize this can explain why ICL can seemingly achieve OOD generalization: the test data actually lies within the span of the training data. Due to space limitations, we only present the main ideas and defer all other details to the Appendix.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems*, 2022.

[3] Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in neural information processing systems*, 36:14228–14246, 2023.

[4] Steve Yadlowsky, Lyric Doshi, and Nilesh Tripuraneni. Pretraining data mixtures enable narrow model selection capabilities in transformer models. *arXiv preprint arXiv:2311.00871*, 2023.

[5] Qixun Wang, Yifei Wang, Yisen Wang, and Xianghua Ying. Can in-context learning really generalize to out-of-distribution tasks? In *The Thirteenth International Conference on Learning Representations*, 2025.

[6] Kartik Ahuja and David Lopez-Paz. A closer look at in-context learning under distribution shifts. *arXiv preprint arXiv:2305.16704*, 2023.

[7] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

[8] Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning? In *International Conference on Machine Learning*, pages 28734–28783. PMLR, 2024.

[9] Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[10] Jannik Kossen, Yarin Gal, and Tom Rainforth. In-context learning learns label relationships but is not conventional learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[11] Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In *International Conference on Machine Learning*, pages 19660–19722. PMLR, 2024.

[12] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *The Eleventh International Conference on Learning Representations*, 2023.

[13] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

[14] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650, 2023.

[15] Yingcong Li, Ankit Singh Rawat, and Samet Oymak. Fine-grained analysis of in-context linear estimation: Data, architecture, and beyond. *arXiv preprint arXiv:2407.10005*, 2024.

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

[17] Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *The Twelfth International Conference on Learning Representations*, 2024.

[18] Arvind V. Mahankali, Tatsunori Hashimoto, and Tengyu Ma. 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*, 2024.

[19] Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context. *Advances in Neural Information Processing Systems*, 37:77316–77365, 2025.

[20] Khashayar Gatmiry, Nikunj Saunshi, Sashank J. Reddi, Stefanie Jegelka, and Sanjiv Kumar. Can looped transformers learn to implement multi-step gradient descent for in-context learning? In *International Conference on Machine Learning*, 2024.

[21] Ruiqi Zhang, Jingfeng Wu, and Peter Bartlett. In-context learning of a linear transformer block: benefits of the mlp component and one-step gd initialization. *Advances in Neural Information Processing Systems*, 37:18310–18361, 2024.

[22] Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. Revisiting the equivalence of in-context learning and gradient descent: The impact of data distribution. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7410–7414. IEEE, 2024.

[23] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19565–19594. PMLR, 23–29 Jul 2023.

[24] Ying Fan, Steve Yadlowsky, Dimitris Papailiopoulos, and Kangwook Lee. Transformers can learn meta-skills for task generalization in in-context learning. In *NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward*, 2024.

[25] Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. In-context learning with representations: Contextual generalization of trained transformers. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 85867–85898. Curran Associates, Inc., 2024.

[26] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[27] Pierre-Alexandre Mattei. Multiplying a gaussian matrix by a gaussian vector. *Statistics & Probability Letters*, 128:67–70, 2017.

Appendix

Contents

1 Introduction 1

2 Problem Setup and Theoretical Results 1

A Background: Single-Layer Linear Attention Model 5

B Main Results 7
- B.1 Transformers Are Not Robust To Subspace Shifts . . . 7
- B.2 Transformers Can Generalize to the Span When Trained on a Union of Subspaces . 7

C Experimental Results 9
- C.1 More Results on Linear Function Classes . . . 9
- C.2 Beyond Linear Function Classes . . . 10

D Discussion 10

E Related Work 11

F Additional Results 12
- F.1 Result with Different Principal Angles . . . 12
- F.2 Generalization Beyond a Mixture of Two Gaussians . . . 13

G Deferred Proofs 13
- G.1 Proofs for Task Shifts . . . 13
  - G.1.1 Supporting Results . . . 13
  - G.1.2 Proof of Proposition 1 . . . 17
  - G.1.3 Proof of Theorem 2 . . . 18
  - G.1.4 Proof of Theorem 3 . . . 20
- G.2 Auxiliary Results . . . 21
  - G.2.1 Optimal Linear Attention Weights . . . 21
  - G.2.2 Miscellaneous Results . . . 22

A Background: Single-Layer Linear Attention Model

The work by Ahn et al. [17] empirically showed many phenomena observed in vanilla Transformers can be replicated in Transformers with linear attention. These findings motivated other works [7, 14, 15] to use linear attention as a test-bed for studying ICL. Following these works, we consider a single-layer linear attention model for analysis. Let $\{\mathbf{x}, y\} \in \mathbb{R}^d \times \mathbb{R}$ denote a feature and label pair.

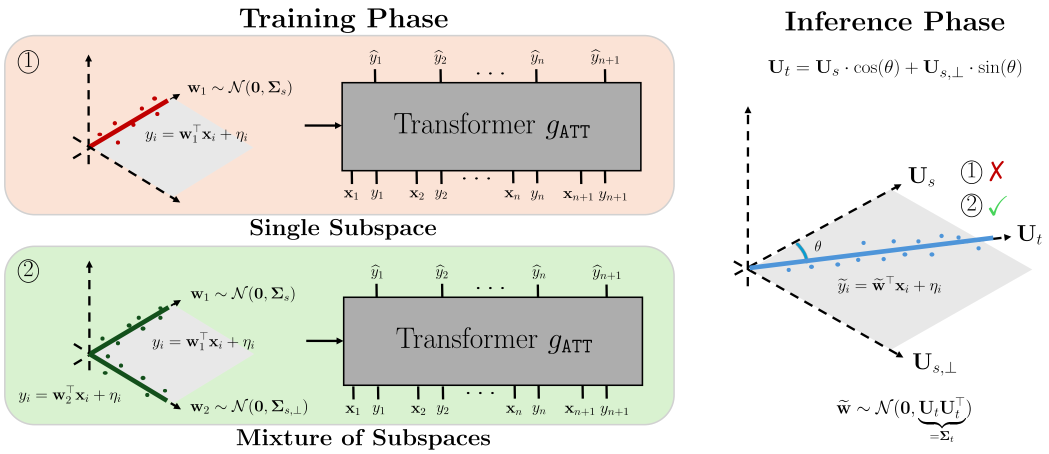

Figure 1: **Overview of this paper.** We consider two models: one trained with task vectors drawn from a single subspace, and one with task vectors drawn from a union of subspaces. At inference, we test both models using a task vector at an angle between two subspaces. The single subspace model fails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

Given $n+1$ paired examples $\{\mathbf{x}_i, y_i\}_{i=1}^{n+1}$, we construct the training-time input prompt as such:

$$\mathbf{Z} = \begin{bmatrix}\mathbf{z}_1 & \dots & \mathbf{z}_n & \mathbf{z}_{n+1}\end{bmatrix}^\top = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_n & \mathbf{x}_{n+1} \\ y_1 & \dots & y_n & 0\end{bmatrix}^\top \in \mathbb{R}^{(n+1)\times(d+1)},$$

Following Ahn et al. [14] and Mahankali et al. [18], we employ a causal mask to the prompt to ensure inputs cannot attend to their own labels:

$$\mathbf{Z}_{\mathcal{M}} = \begin{bmatrix}\mathbf{z}_1 & \dots & \mathbf{z}_n & \mathbf{0}\end{bmatrix}^\top, \quad \text{where} \quad \mathbf{z}_i = \begin{bmatrix}\mathbf{x}_i \\ y_i\end{bmatrix} \quad \text{and} \quad \mathbf{z}_q = \begin{bmatrix}\mathbf{x}_{n+1} \\ 0\end{bmatrix}. \tag{8}$$

The goal of ICL is to leverage the in-context examples $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ in the prompt $\mathbf{Z}_{\mathcal{M}}$ to predict the correct label y_{n+1} according to the query $\mathbf{x}_{n+1}$ (equivalently $\mathbf{z}_q$). We input the prompt $\mathbf{Z}_{\mathcal{M}}$ and query $\mathbf{z}_q$ into a (normalized) single head linear attention model to make the prediction $\hat{y}_{n+1}$:

$$\widehat{y}_{n+1} = g_{\texttt{ATT}}(\mathbf{z}_q, \mathbf{Z}_{\mathcal{M}}) = \frac{1}{n}\left(\mathbf{z}_q^\top \mathbf{W}_Q \mathbf{W}_K^\top \mathbf{Z}_{\mathcal{M}}^\top\right) \mathbf{Z}_{\mathcal{M}} \mathbf{W}_V \mathbf{p}, \tag{9}$$

where $\mathbf{W}_K, \mathbf{W}_Q, \mathbf{W}_V \in \mathbb{R}^{(d+1)\times(d+1)}$ are the key, query, and value weight matrices, respectively, and $\mathbf{p} \in \mathbb{R}^{d+1}$ is the linear prediction head. We denote $\mathcal{W} = \{\mathbf{W}_K, \mathbf{W}_Q, \mathbf{W}_V, \mathbf{p}\}$ as the collection of trainable weights corresponding to the linear attention model. We train the model $g_{\texttt{ATT}}$ by minimizing the following expected squared loss with respect to the parameters $\mathcal{W}$:

$$\min_{\mathcal{W}} \mathcal{L}_{\texttt{ATT}}(\mathcal{W}), \quad \text{where} \quad \mathcal{L}_{\texttt{ATT}}(\mathcal{W}) = \mathbb{E}\left[\left(y_{n+1} - g_{\texttt{ATT}}(\mathbf{z}_q, \mathbf{Z}_{\mathcal{M}})\right)^2\right]. \tag{10}$$

For inference, given $m+1$ paired examples $\{\mathbf{x}_j, \tilde{y}_j\}_{j=1}^{m+1}$, we construct the input prompts as such:

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

Then, the inputs $\widetilde{\mathbf{Z}}_{\mathcal{M}}$ and $\tilde{\mathbf{z}}_q$ are fed into the trained linear attention model to obtain a prediction for $\tilde{y}_{m+1}$. Specifically, let $\mathcal{W}^\star = \{\mathbf{W}_K^\star, \mathbf{W}_Q^\star, \mathbf{W}_V^\star, \mathbf{p}^\star\}$ denote the optimally trained linear attention model for minimizing the loss in Equation (10). We compute

$$\widehat{y}_{m+1} = g_{\texttt{ATT}}^\star(\widetilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}}) = \frac{1}{m}\left(\widetilde{\mathbf{z}}_q^\top \mathbf{W}_Q^\star \mathbf{W}_K^{\star\top} \widetilde{\mathbf{Z}}_{\mathcal{M}}^\top\right) \widetilde{\mathbf{Z}}_{\mathcal{M}} \mathbf{W}_V^\star \mathbf{p}^\star,$$

where we normalize by a factor of m instead of n. Doing so decouples the training and testing prompt lengths, which allows us to analyze the behavior of ICL under different conditions.

B Main Results

This section presents our main results in detail to support those discussed in the main text. We illustrate an overview of the setup in Figure 1.

B.1 Transformers Are Not Robust To Subspace Shifts

In this section, we consider the setup in Section 2, where we train a single-layer linear attention model according to Equation (1), and test the (optimal) model with Equation (2). We prove that even with infinitely many samples, ICL exhibits test risk with a non-negligible dependence on the shift between the covariance matrices $\mathbf{\Sigma}_t$ and $\mathbf{\Sigma}_s$, as measured by θ. This result demonstrates that ICL is not inherently robust to subspace shifts.

Proposition 1 (Task Distribution Shift)**.** *Let $g^\star_{ATT}$ denote the optimal linear attention model corresponding to the independent data setting in Equation (1). 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*

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

and $\mathbf{\Sigma}_t \in \mathbb{R}^{d \times d}$ is from Equation (3). Then, we have

$$\lim_{m\to\infty} \lim_{n\to\infty} \lim_{\epsilon\to 0} \mathbb{E}\left[\left(\tilde{y}_{m+1} - g^\star_{ATT}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = r\sin^2(\theta) + \sigma^2, \tag{11}$$

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

The proof is provided in Appendix G.1.2. 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) to ensure that the covariance matrices are exactly low-rank. Then, 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 [2, 19].

The analysis involves deriving the test risk under an arbitrary distribution shift, assuming the linear attention model is parameterized by the optimal weights according to Equation (10). At the optimal weights, the model reduces to a single step of projected gradient descent (PGD) [13–15, 18]. 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 arises (assuming $\sigma = 0$ for simplicity):

$$\begin{aligned}
\widehat{y}_{m+1} = g^\star_{\mathtt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}}) &= \frac{1}{m}\mathbf{x}_{m+1}^\top \mathbf{A}\mathbf{X}^\top \mathbf{y} = \frac{1}{m}\mathbf{x}_{m+1}^\top \mathbf{A}\mathbf{X}^\top \mathbf{X}\tilde{\mathbf{w}} && (\text{Substitute } \mathbf{y} = \mathbf{X}\tilde{\mathbf{w}})\\
&\to \mathbf{x}_{m+1}^\top \mathbf{U}_s\mathbf{U}_s^\top \tilde{\mathbf{w}} && (\text{Take } m, n \to \infty \text{ and } \epsilon \to 0)\\
&= \mathbf{x}_{m+1}^\top \mathbf{U}_s\mathbf{U}_s^\top \mathbf{U}_t\mathbf{g}, && (\tilde{\mathbf{w}} = \mathbf{U}_t\mathbf{g} \text{ for } \mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r))
\end{aligned}$$

where $\mathbf{X} \coloneqq [\mathbf{x}_1 \quad \dots \quad \mathbf{x}_m]^\top \in \mathbb{R}^{m\times d}$ and $\mathbf{y} \coloneqq [\tilde{y}_1 \quad \dots \quad \tilde{y}_m]^\top \in \mathbb{R}^m$. 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. Put differently, in cases in which $\mathbf{\Sigma}_t \neq \mathbf{\Sigma}_s$, ICL can generalize only if $\mathcal{R}(\mathbf{\Sigma}_t) \subset \mathcal{R}(\mathbf{\Sigma}_s)$.

In Figure 2, we present experiments corroborating Proposition 1 on both linear and nonlinear Transformers. 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 F.1.

B.2 Transformers Can Generalize to the Span When Trained on a Union of Subspaces

Previously, we observed that shifting the covariance matrix induces a dependence on θ in the test risk due to projection onto a misaligned subspace, implying that the training and testing data must

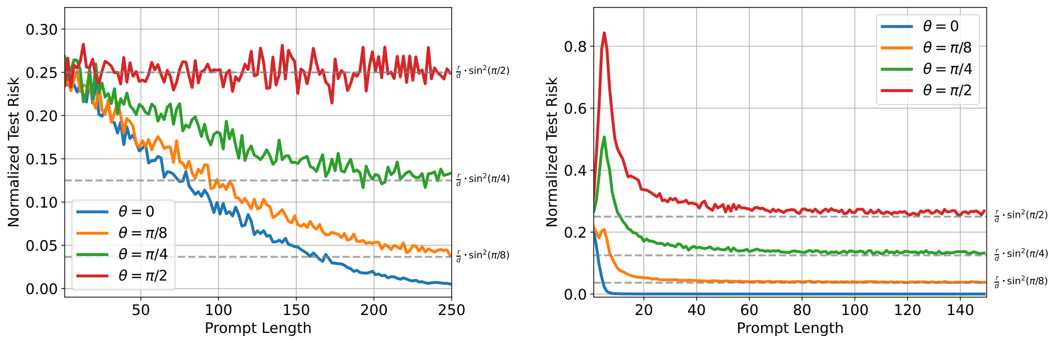

Figure 2: Plot of the normalized test risk for OOD linear regression as a function of the prompt length for a linear Transformer (left) and a nonlinear Transformer (right) under covariance shifts. As the covariance at test time shifts away from the covariance used at training time as a function of θ, 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.

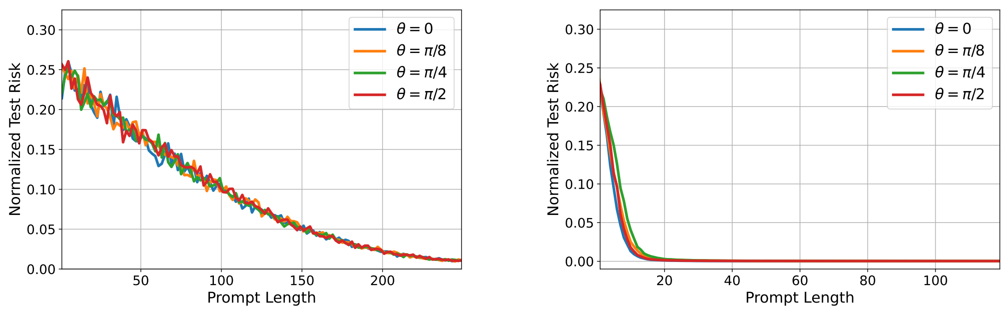

Figure 3: Plot of the test risk for OOD linear regression as a function of the prompt length for a linear Transformer (left) and a nonlinear Transformer (right). When the prompt length at test time is large enough, the test risk goes nearly to zero for all $\theta \in \left[0, \frac{\pi}{2}\right]$,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

span the same r-dimensional subspace. This raises the question: are there settings in ICL where the dependence on θ can be mitigated? In the main text, we showed that this dependence can be mitigated, roughly speaking, by introducing diversity into the training prompts. Specifically, we showed that by drawing task vectors from a union of subspaces, the projection matrix can better capture shifts in θ, allowing OOD generalization. In the following, we re-phrase Theorem 1 in the same format as Proposition 1.

Theorem 2 (Test Risk under the Span of Covariance Matrices)**.** *Let $g^{\star}_{\texttt{ATT}}$ denote the optimal linear attention model corresponding to the independent data setting in Equation (1), where the task vector now follows Equation (6) 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*

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

and $\mathbf{\Sigma}_t \in \mathbb{R}^{d \times d}$ is from Equation (3). For any $\theta \in \left[0, \frac{\pi}{2}\right]$ and $\delta \in (0, r)$, if

$$m \geq n > \frac{\left(2(r+\sigma^2)+1\right) r}{\delta} - \left(2(r+\sigma^2)+1\right), \tag{12}$$

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

The proof technique is similar to that of Proposition 1 and is available in Appendix G.1.3. Moreover, we can generalize this result to a mixture of $K > 2$ subspaces; see Appendix F.2. For Theorem 2,

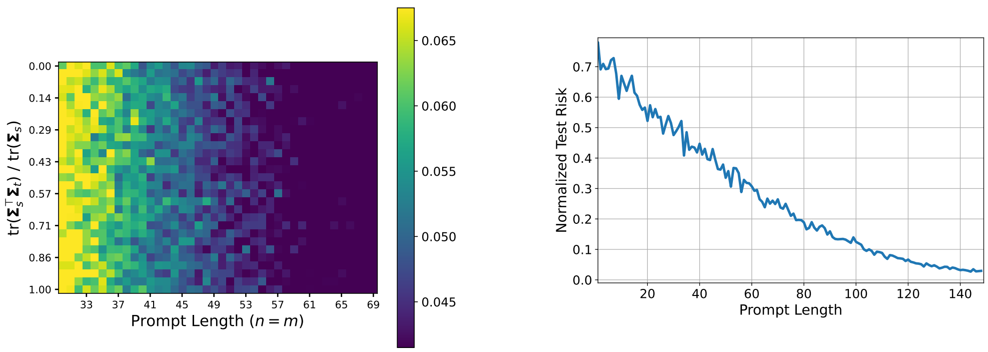

Figure 4: Left: Phase plot of the test risk as we vary the angle between $\mathbf{\Sigma}_s$ and $\mathbf{\Sigma}_t$ and the prompt length with $m = n$ for a linear attention model trained with a mixture of Gaussians. The test risk is low across all angle shifts, and decreases further as the prompt length 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 at al [20]. This serves to explain why ICL can seemingly do OOD generalization as observed in the literature.

we similarly sketch how θ becomes mitigated in the test risk. Consider the case where $\delta \to 0$, i.e., $m, n \to \infty$. Then, we can simplify the linear attention model as such (again assuming $\sigma = 0$):

$$\widehat{y}_{m+1} = g^{\star}_{\mathtt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}}) \to \mathbf{x}_{m+1}^\top \mathbf{U}_{2r} \mathbf{U}_{2r}^\top \widetilde{\mathbf{w}} = \mathbf{x}_{m+1}^\top \mathbf{U}_{2r} \mathbf{U}_{2r}^\top \mathbf{U}_t \mathbf{g},$$

where again $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r)$ and $\mathbf{U}_{2r} = [\mathbf{U}_s \quad \mathbf{U}_{s,\perp}]$. 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 $\widetilde{y}_{m+1}$. In Figure 3, we present results on linear Transformers and GPT-2 that corroborate our theory. In both models, the test risk approaches zero for all $\theta \in \left[0, \frac{\pi}{2}\right]$, meaning there is no dependence on θ. The only noticeable difference is the linear attention model requires a longer prompt length to reach near-zero risk, which is also highlighted by our theory.

Overall, this highlights an interesting property of Transformers: if the training task vectors are drawn from a union of subspaces, then ICL can interpolate to the space between the subspaces. In other words, even if certain regions have zero probability density in the distribution over the training task vector, ICL can still generalize to those regions at test time, as long as they lie within the overall span of the training task vectors. We hypothesize this can explain why ICL can seemingly achieve OOD generalization: the test data actually remains within the span of the training distribution.

C Experimental Results

Experimental Setup. Unless otherwise stated, the experimental setup is as follows: for both the linear and nonlinear Transformer, we consider noiseless linear regression, and set $d = 20$, $r = 5$, and $\epsilon = 10^{-6}$. 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 $m = n = 250$. For the nonlinear Transformer, following Garg et al. [2], 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 64, prompt lengths $m = n = 120$, and train for 100K iterations.

C.1 More Results on Linear Function Classes

Linear Regression. Previously, we presented results on the test risk as a function of the prompt length. In Figure 4 (left), we present a phase plot of the test risk as a function of both $\mathrm{Tr}(\mathbf{\Sigma}_s^\top \mathbf{\Sigma}_t) \,/\, \mathrm{Tr}(\mathbf{\Sigma}_s)$ (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

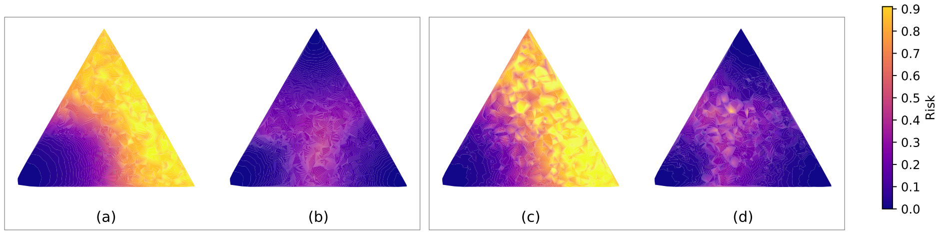

Figure 5: Visualization of the generalization behavior of Transformers for learning nonlinear function classes in-context. Each corner of a triangle represents a one-dimensional subspace spanned by ψ_1 (bottom left), ψ_2 (bottom right), or ψ_3 (top), with all possible convex combinations given by the interior. In all cases, we show the risk when evaluated at different points in $\text{span}(\{\psi_1, \psi_2, \psi_3\})$ for the appropriate function space. (a) Train on prompts drawn from $\text{span}(\{\psi_1^C\})$. (b) Train on prompts drawn from $\text{span}(\{\psi_1^C\}) \cup \text{span}(\{\psi_2^C\}) \cup \text{span}(\{\psi_3^C\})$. (c) Train on prompts drawn from $\text{span}(\{\psi_1^H\})$. (d) Train on prompts drawn from $\text{span}(\{\psi_1^H\}) \cup \text{span}(\{\psi_2^H\}) \cup \text{span}(\{\psi_3^H\})$.

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 B.2, 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. [20], with $d = 5$, $\mathbf{\Sigma}_s = \mathbf{I}_5$ 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 $\mathbf{\Lambda}_t = \text{Diag}(1, 1, 1/2, 1/4, 1)$. In Figure 4 (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.

C.2 Beyond Linear Function Classes

Finally, we demonstrate that the theoretical findings in Appendix B extend to *nonlinear* function classes. Specifically, we look at two function spaces, namely $L^2([0,1])$ and $L^2(\mathbb{R}, e^{-x^2/2}/\sqrt{2\pi}\,dx)$, i.e., square-integrable functions under the uniform and Gaussian measures respectively, which model rich sets of signals observed in real-world data. For the former, we construct an orthonormal basis via cosines, i.e., $\psi_n^C(x) = (1/\sqrt{2})\cos(n\pi x)$ for $n \in \mathbb{N}$. For the latter, we construct an orthonormal basis via Hermite polynomials:

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

As described 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. As before, we draw the function coefficients from standard multivariate Gaussian. We draw the inputs from the distribution appropriate to the function space measure, i.e., $x \sim \mathcal{U}([0,1])$ for $L^2([0,1])$ and $x \sim \mathcal{N}(0,1)$ for $L^2(\mathbb{R}, e^{-x^2/2}/\sqrt{2\pi}\,dx)$. All other details are identical to previous (nonlinear) Transformer experiments. The results are shown in Figure 5. As shown in panels (a) and (c) of Figure 5, we see that Transformers are not robust to subspace shifts for either function class, with increasing test risk with respect to the subspace angle from the train subspace, in accordance with Proposition 1. On the other hand, as shown in panels (b) and (d) of Figure 5, we have the generalization behavior described by Theorem 3, where training on the mixture of subspaces results in low risk in the space spanned by the basis vectors.

D Discussion

In this work, we analyzed the OOD generalization capabilities of ICL by studying a single-layer linear attention model with linear regression, where the task vector was parameterized by low-dimensional subspaces. We uncovered two key properties of ICL: (i) it is not inherently robust to subspace shifts, and (ii) it can generalize to the span of covariance matrices if trained on a union of subspaces. We also provided insights into how LoRA can be used to model distribution shifts, and showed how our findings extend to nonlinear function classes. One limitation of this work is that the analysis focuses on single-layer linear attention, as in prior studies; a promising direction for future research is to extend the analysis to multi-layer nonlinear Transformers.

E Related Work

ICL on Transformers with Linear Attention. There is abundant research on ICL that analyzes single-layer linear attention models. Below, we survey several works most relevant to our work; like ours, many of them focus on linear regression settings, where for all $i \in [n+1]$:

$$y_i = f(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + \eta, \quad \text{where} \quad \mathbf{w} \sim \mathcal{N}\left(\mathbf{0}_d, \boldsymbol{\Sigma}_{\mathbf{w}}\right), \quad \mathbf{x}_i \sim \mathcal{N}\left(\mathbf{0}_d, \boldsymbol{\Sigma}_{\mathbf{x}}\right),$$

and η is additive Gaussian noise. As previously mentioned, Zhang et al. [7] studied the training dynamics of a single-layer linear attention model on the population loss for a linear regression ICL task. Specifically, assuming $\boldsymbol{\Sigma}_{\mathbf{w}} = \mathbf{I}_d$ and an arbitrary $\boldsymbol{\Sigma}_{\mathbf{x}}$, they showed the model weights converge to a globally optimal solution under gradient flow, despite the non-convex objective. They also provide closed-form expressions for the model weights at the global minima. A follow-up work [21] considered a linear regression task with $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}})$ and a linear Transformer model (a linear attention layer followed by a two-layer linear network). They showed a single linear attention layer incurs a sub-optimal risk that depends on $\boldsymbol{\mu}_{\mathbf{w}}$, but adding a linear network allows the model to achieve the Bayes optimal risk.

Other works [13–15, 21–23] study the underlying learning algorithms that linear attention models implement when learning linear functions in-context. Specifically, for a single linear attention layer, Von Oswald et al. [13] demonstrated the existence of model weights that implement a single step of GD on a mean-squared error loss. They further showed empirically that the weights of a trained linear attention layer closely align with those that implement a GD step. Follow-up works [14,15,22] rigorously proved the equivalence between a single step of preconditioned gradient descent (PGD) with zero initialization and the weights of a single-layer linear attention model under the population loss. Specifically, Ahn et al. [14] theoretically showed when $\boldsymbol{\Sigma}_{\mathbf{w}} = \mathbf{I}_d$ and $\boldsymbol{\Sigma}_{\mathbf{x}}$ is arbitrary, the single-layer linear attention model learns a preconditioning matrix that is dependent on $\boldsymbol{\Sigma}_{\mathbf{x}}$. Li et al. [15] generalized this result by considering an arbitrary $\boldsymbol{\Sigma}_{\mathbf{w}}$ in addition to $\boldsymbol{\Sigma}_{\mathbf{x}}$ — they showed the learned preconditioning matrix depends on both $\boldsymbol{\Sigma}_{\mathbf{x}}$ and $\boldsymbol{\Sigma}_{\mathbf{w}}$. Finally, [21] showed when $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}})$, a linear attention layer followed by a linear network implements a PGD step while *learning* the initialization. While our work builds on the fact that a single-layer linear attention model implements PGD, our goal differs from these prior works: we study how ICL under this model can generalize out-of-distribution.

Empirical Observations on ICL's OOD Generalization. As part of their study, Garg et al. [2] empirically observed Transformer-based ICL is robust to a number of distribution shifts, such as between the train and test distributions of the features $\mathbf{x}_i$, as well as between the features $\mathbf{x}_i$ and query $\mathbf{x}_q$. These observations inspired an extensive line of empirical work studying ICL's ability to generalize to OOD tasks [3–6, 9, 10, 24]. To our knowledge, [4, 5] are the most closely related with our setting. Specifically, these works consider sampling tasks from a mixture of *function class* distributions, e.g., f is sampled from the class of dense linear functions with probability $\gamma \in (0, 1)$, or from the class of sparse linear functions with probability $1 - \gamma$. Yadlowskey et al. [4] showed when Transformers are trained for ICL on a mixture of function classes, ICL cannot generalize well to function classes not present in the training mixture. Wang et al. [5] argue if the test task is not in the training mixture, Transformers select a task from the training mixture that minimizes the test error. In contrast, our work assumes that the target function is sampled from a mixture of low-dimensional subspaces in a fixed function space. In other words, the mixture distribution from which we sample is always within a *single* function class. We emphasize this is different from sampling from a mixture of *multiple* function class distributions.

Theoretical Studies on ICL's OOD Generalization. The above empirical observations motivated theoretical studies on ICL's OOD generalization ability. Under their setting, Zhang et al. [7] studied how a trained single linear attention layer handles various distribution shifts. Assuming the model weights were at the global minima of Equation (10), they derived a closed-form expression for the prediction $\widehat{y}_q$ for a given query $\mathbf{x}_q$ and in-context examples $\left(\mathbf{x}_1, \mathbf{w}^\top \mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{w}^\top \mathbf{x}_m\right)$. Using this expression for $\widehat{y}_q$, they concluded a trained linear attention model is robust to task and query shifts, but cannot tolerate feature shifts well.

Other works have studied *nonlinear* models and function classes. For instance, [8] considered a binary classification ICL task. They showed a sufficiently trained single-layer, single-head Transformer model (one softmax attention layer followed by a two-layer perceptron) can achieve arbitrarily small generalization error when the inference-time features are *linear combinations* of the training features. Another work [25] assumed the function to learn in-context was $f(\mathbf{x}) = \mathbf{w}^\top g(\mathbf{x}) + \eta$, where $g(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_\ell(\mathbf{x}))$ is an arbitrary feature mapping. They showed if $\mathbf{w}$ has iid, zero mean, unit variance entries at train time, and $\|\mathbf{w}\|_2$ is bounded at inference time, a trained single-layer, multi-head softmax attention model generalizes well under *any* shift in $\mathbf{w}$. Again, our paper differs from these works by studying when ICL can and cannot perform OOD generalization, particularly by using low-dimensional subspaces to parameterize the covariance matrices.

Learning Functions with Low-Dimensional Structure In-Context. To the best of our knowledge, the work by [19] is the only most related work that also considers learning functions with low-dimensional structures. In their setting, the function to learn in-context is a single-index model $f(\mathbf{x}) = \sigma\left(\mathbf{w}^\top \mathbf{x}\right) + \eta$, where $\sigma(\cdot)$ is a nonlinear link function, $\mathbf{w}$ is drawn from a low-dimensional subspace, and η is additive noise. We only consider linear functions $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + \eta$ in our analysis, but also assume $\mathbf{w}$ is sampled from a low-dimensional distribution. In our experiments, we sample nonlinear functions from subspaces of the *function space*, which differs from sampling the function *parameters* from a subspace of Euclidean space. Furthermore, our goal is to use such a parameterization to study OOD generalization, whereas the main focus of [19] is to examine whether ICL can solve such functions at all.

F Additional Results

In this section, we present additional results to supplement those presented in the main text. All experiments were run using either a Macbook Pro with an Apple M2 Pro Chip or a NVIDIA A100 GPU.

Appendix F.1 presents an additional theoretical result for Proposition 1 for when the principal angles are different. Additionally, in Appendix F.2, we present another result where we generalize the mixture of two Gaussians from Theorem 2 to a mixture of $K \geq 2$ Gaussians.

F.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 $g^\star_{\texttt{ATT}}$ denote the optimal linear attention model corresponding to the independent data setting in Equation (1). 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 (3). Then, we have

$$\lim_{m\to\infty} \lim_{n\to\infty} \lim_{\epsilon\to 0} \mathbb{E}\left[\left(\tilde{y}_{m+1} - g^\star_{\texttt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = \sum_{i=1}^{r} \sin^2(\theta_i) + \sigma^2, \tag{13}$$

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

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., $\sum_{i=1}^r \sin^2(\theta_i) = r\sin^2(\theta)$.

F.2 Generalization Beyond a Mixture of Two Gaussians

We now discuss how ICL can achieve OOD generalization when $\mathbf{w} \in \mathbb{R}^d$ is sampled from a mixture of K low-rank Gaussians for any $K \geq 2$. Let $\mathbf{U} = [\mathbf{u}_1 \quad \dots \quad \mathbf{u}_d] \in \mathbb{R}^{d\times d}$ be an orthonormal basis for $\mathbb{R}^d$. Assuming $d > Kr$, let $\mathbf{U}_{s,k} = \left[\mathbf{u}_{(k-1)\cdot r+1} \quad \dots \quad \mathbf{u}_{kr}\right] \in \mathbb{R}^{d\times r}$ for all $k \in [K]$. Note $\mathbf{U}_{s,k}^\top \mathbf{U}_{s,l} = \mathbf{0}_{r\times r}$ for all $k \neq l$.

We assume the training task $\mathbf{w} \in \mathbb{R}^d$ is sampled as such:

$$\mathbf{w} \sim \sum_{k=1}^{K} \gamma_k \cdot \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{s,k}), \quad \text{where} \quad \mathbf{\Sigma}_{s,k} = \mathbf{U}_{s,k}\mathbf{U}_{s,k}^\top + \epsilon \cdot \mathbf{I}_d \quad \text{and} \quad \sum_{k=1}^{K} \gamma_k = 1. \tag{14}$$

Then, let $\overline{\mathbf{U}}_t$ be an arbitrary orthonormal basis for an r-dimensional subspace that lies in the span of $[\mathbf{U}_{s,1} \quad \dots \quad \mathbf{U}_{s,K}]$, i.e.,

$$\overline{\mathbf{U}}_t = \sum_{k=1}^{K} \alpha_k \mathbf{U}_{s,k} \quad \text{for some } \{\alpha_k\}_{k=1}^K \text{ s.t. } \sum_{k=1}^{K} \alpha_k^2 = 1, \tag{15}$$

where the last constraint on $\{\alpha_k\}_{k=1}^K$ ensures $\overline{\mathbf{U}}_t$ is an orthonormal basis. Similar to the $K = 2$ case, we show when tested on $\widetilde{\mathbf{w}} \sim \mathcal{N}\left(\mathbf{0}, \overline{\mathbf{\Sigma}}_t\right)$ with $\overline{\mathbf{\Sigma}}_t = \overline{\mathbf{U}}_t\overline{\mathbf{U}}_t^\top + \epsilon \cdot \mathbf{I}_d$, the trained model can generalize to this previously unseen subspace $\overline{\mathbf{U}}_t$.

Theorem 3. *Let $g^\star_{\mathtt{ATT}}$ denote the optimal linear attention model corresponding to the independent data setting in Equation (1), where the task vector is drawn from Equation (14) with $\gamma_k = \frac{1}{K}$ for all $k \in [K]$. 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*

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

where $\overline{\mathbf{\Sigma}}_t = \overline{\mathbf{U}}_t\overline{\mathbf{U}}_t^\top + \epsilon \cdot \mathbf{I}_d$ and $\overline{\mathbf{U}}_t$ is defined in Equation (15). For any $\{\alpha_k\}_{k=1}^K$ s.t. $\sum_{k=1}^{K} \alpha_k^2 = 1$ and $\delta \in (0, r)$, if

$$m \geq n > \frac{(K(r+\sigma^2)+1)r}{\delta} - (K(r+\sigma^2)+1), \tag{16}$$

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

The proof is deferred to Appendix G.1.4. Similar to Theorem 2, 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. We note setting $K = 2$, $\alpha_1 = \cos(\theta)$, and $\alpha_2 = \sin(\theta)$ perfectly recovers Theorem 2.

G Deferred Proofs

This section presents all deferred proofs and is organized as follows: Section G.1 contains all proofs related to shifts in the task vector $\mathbf{w} \in \mathbb{R}^d$, and Appendix G.2 provides auxiliary results used to support both the task and feature shift proofs.

G.1 Proofs for Task Shifts

G.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 $g^\star_{\mathtt{ATT}}$ denote the optimal linear attention model corresponding to the independent data setting in Equation (1). 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,

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

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

Proof. Recall at inference time,

$$\widetilde{\mathbf{Z}}_{\mathcal{M}} = \begin{bmatrix}\tilde{\mathbf{z}}_1 & \dots & \tilde{\mathbf{z}}_m & \mathbf{0}\end{bmatrix}^\top = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_m & \mathbf{0} \\ \widetilde{y}_1 & \dots & \widetilde{y}_m & 0\end{bmatrix}^\top \quad \text{and} \quad \tilde{\mathbf{z}}_q = \begin{bmatrix}\mathbf{x}_{m+1} \\ 0\end{bmatrix} := \begin{bmatrix}\mathbf{x}_q \\ 0\end{bmatrix}. \tag{17}$$

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

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

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 $\mathbb{E}\left[\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q\right)^2 \mid \widetilde{\mathbf{w}}\right] = \operatorname{Var}\left(\widetilde{\mathbf{w}}^\top \mathbf{x}_q \mid \widetilde{\mathbf{w}}\right) = \|\widetilde{\mathbf{w}}\|^2$. Therefore,

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

Therefore,

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

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

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

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

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):

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

where $\widetilde{\mathbf{X}}_{te}^\top := \mathbf{A}\mathbf{X}_{te}^\top$ and $\overline{\mathbf{X}}_{te}^\top := \boldsymbol{\Sigma}_t^{1/2}\mathbf{X}_{te}^\top$. Note $\widetilde{\mathbf{X}}_{te}^\top = [\mathbf{A}\mathbf{x}_1 \quad \dots \quad \mathbf{A}\mathbf{x}_m] := [\widetilde{\mathbf{x}}_1 \quad \dots \quad \widetilde{\mathbf{x}}_m]$ where $\widetilde{\mathbf{x}}_i := \mathbf{A}\mathbf{x}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}_d, \mathbf{A}\mathbf{A}^\top)$, and $\overline{\mathbf{X}}_{te}^\top = \left[\boldsymbol{\Sigma}_t^{1/2}\mathbf{x}_1 \quad \dots \quad \boldsymbol{\Sigma}_t^{1/2}\mathbf{x}_m\right] := [\overline{\mathbf{x}}_1 \quad \dots \quad \overline{\mathbf{x}}_m]$ where $\overline{\mathbf{x}}_i := \boldsymbol{\Sigma}_t^{1/2}\mathbf{x}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}_d, \boldsymbol{\Sigma}_t)$. 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,

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

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

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

Therefore,

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

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

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

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 [26]).

We now focus on (e):

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

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 [27], $\widetilde{\boldsymbol{\eta}}_{te} \sim \mathrm{GAL}_d(2\sigma^2\mathbf{A}\mathbf{A}^{\top}, \mathbf{0}_d, m/2)$, 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 [27]). 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:

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

Combining like terms yields

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

which is exactly Equation (18). This completes the proof. □

G.1.2 Proof of Proposition 1

Proof. For simplicity, we denote $\widetilde{y} := \widetilde{y}_{m+1}$. Recall $\mathbf{U} := [\mathbf{U}_s \quad \mathbf{U}_{s,\perp} \quad \mathbf{U}_{2r,\perp}] \in \mathbb{R}^{d\times d}$, 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 5, we have

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

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

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

$$\mathbb{E}\left[\left(\widetilde{y} - g^\star_{\mathtt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = \left(\frac{1}{m}\operatorname{Tr}\left(\mathbf{A}^\top\mathbf{A}\right)+1\right)\left(\operatorname{Tr}\left(\mathbf{\Sigma}_t\right)+\sigma^2\right) - 2\operatorname{Tr}\left(\mathbf{\Sigma}_t\mathbf{A}\right) + \frac{m+1}{m}\operatorname{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t\mathbf{A}^\top\right). \tag{20}$$

We simplify the remaining $\operatorname{Tr}(\cdot)$ terms using Equation (19).

Simplifying $\operatorname{Tr}(\mathbf{A})$ and $\operatorname{Tr}(\mathbf{A}^\top\mathbf{A})$. Directly from Equation (19):

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

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

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

Recall we defined $\mathbf{U}_t$ in Equation (4) as follows:

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

Therefore:

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

and thus,

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

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 (20) yields

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

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

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

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

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

which completes the proof. □

G.1.3 Proof of Theorem 2

Proof. For simplicity, we denote $\widetilde{y} := \widetilde{y}_{m+1}$. Recall that by Lemma 1 and Lemma 3, we have

$$\begin{aligned}\mathbb{E}\left[\left(\widetilde{y} - g^{\star}_{\mathtt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] &= \left(\frac{1}{m}\mathrm{Tr}\left(\mathbf{A}^\top\mathbf{A}\right) + 1\right)\left(\mathrm{Tr}\left(\mathbf{\Sigma}_t\right) + \sigma^2\right) \\ &\quad - 2\,\mathrm{Tr}\left(\mathbf{\Sigma}_t\mathbf{A}\right) + \frac{m+1}{m}\mathrm{Tr}\left(\mathbf{A}\mathbf{\Sigma}_t\mathbf{A}^\top\right),\end{aligned} \tag{21}$$

where $\mathbf{A} = \left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\mathbf{\Sigma}^{-1}\right)^{-1}$, $M_s = \mathrm{Tr}(\mathbf{\Sigma}) + \sigma^2$ with $\mathbf{\Sigma} = \gamma\mathbf{\Sigma}_s + (1-\gamma)\mathbf{\Sigma}_{s,\perp}$. First, we simplify $\mathbf{\Sigma}$ as such:

$$\begin{aligned}\mathbf{\Sigma} &= \gamma\mathbf{\Sigma}_s + (1-\gamma)\mathbf{\Sigma}_{s,\perp} \\ &= \gamma\big(\mathbf{U}_s\mathbf{U}_s^\top + \epsilon\cdot\mathbf{I}_d\big) + (1-\gamma)\big(\mathbf{U}_{s,\perp}\mathbf{U}_{s,\perp}^\top + \epsilon\cdot\mathbf{I}_d\big) \\ &= \mathbf{U}\begin{bmatrix}\gamma(1+\epsilon)\cdot\mathbf{I}_r & \\ & \gamma\epsilon\cdot\mathbf{I}_{d-r}\end{bmatrix}\mathbf{U}^\top + \mathbf{U}\begin{bmatrix}(1-\gamma)\epsilon\cdot\mathbf{I}_r & & \\ & (1-\gamma)(1+\epsilon)\cdot\mathbf{I}_r & \\ & & (1-\gamma)\epsilon\cdot\mathbf{I}_{d-2r}\end{bmatrix}\mathbf{U}^\top \\ &= \mathbf{U}\begin{bmatrix}(\gamma+\epsilon)\cdot\mathbf{I}_r & & \\ & (\epsilon-\gamma+1)\cdot\mathbf{I}_r & \\ & & \epsilon\cdot\mathbf{I}_{d-2r}\end{bmatrix}\mathbf{U}^\top,\end{aligned}$$

and so we have

$$M_s = \mathrm{Tr}(\mathbf{\Sigma}) + \sigma^2 = r + \epsilon d + \sigma^2 \quad \text{and} \quad \mathbf{\Sigma}^{-1} = \mathbf{U}\begin{bmatrix}\frac{1}{\gamma+\epsilon}\cdot\mathbf{I}_r & & \\ & \frac{1}{\epsilon-\gamma+1}\cdot\mathbf{I}_r & \\ & & \frac{1}{\epsilon}\cdot\mathbf{I}_{d-2r}\end{bmatrix}\mathbf{U}^\top.$$

Then, by Lemma 5, we have

$$\mathbf{A} = \left(\frac{n+1}{n}\mathbf{I}_d + \frac{M_s}{n}\mathbf{\Sigma}^{-1}\right)^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top, \tag{22}$$

where

$$\mathbf{\Lambda} = \begin{bmatrix}\frac{n(\gamma+\epsilon)}{(n+1)(\gamma+\epsilon)+M_s}\cdot\mathbf{I}_r & & \\ & \frac{n(\epsilon-\gamma+1)}{(n+1)(\epsilon-\gamma+1)+M_s}\cdot\mathbf{I}_r & \\ & & \frac{n\epsilon}{(n+1)\epsilon+M_s}\cdot\mathbf{I}_{d-2r}\end{bmatrix} := \begin{bmatrix}\nu_1\cdot\mathbf{I}_r & & \\ & \nu_2\cdot\mathbf{I}_r & \\ & & \nu_3\cdot\mathbf{I}_{d-2r}\end{bmatrix}.$$

We simplify the $\mathrm{Tr}(\cdot)$ terms using Equation (22).

Simplifying $\mathrm{Tr}(\mathbf{A})$ **and** $\mathrm{Tr}(\mathbf{A}^\top\mathbf{A})$**.** Directly from Equation (22):

$$\mathrm{Tr}(\mathbf{A}) = r\nu_1 + r\nu_2 + (d-2r)\nu_3, \quad \text{and} \quad \mathrm{Tr}(\mathbf{A}^\top\mathbf{A}) = \mathrm{Tr}(\mathbf{A}^2) = r\nu_1^2 + r\nu_2^2 + (d-2r)\nu_3^2.$$

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 $\mathbf{U}_t = \mathbf{U}_s\cos(\mathbf{\Theta}) + \mathbf{U}_{s,\perp}\sin(\mathbf{\Theta})$, and so we have

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

$$\begin{aligned}\mathrm{Tr}\left(\mathbf{U}_t^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\mathbf{U}_t\right) &= \mathrm{Tr}\left(\left[\cos(\mathbf{\Theta}) \quad \sin(\mathbf{\Theta}) \quad \mathbf{0}_{d\times(d-2r)}\right]\begin{bmatrix}\nu_1\mathbf{I}_r & & \\ & \nu_2\mathbf{I}_r & \\ & & \nu_3\mathbf{I}_{d-2r}\end{bmatrix}\begin{bmatrix}\cos(\mathbf{\Theta}) \\ \sin(\mathbf{\Theta}) \\ \mathbf{0}_{(d-2r)\times d}\end{bmatrix}\right) \\ &= \mathrm{Tr}\left(\begin{bmatrix}\nu_1\cos^2(\mathbf{\Theta}) & & \\ & \nu_2\sin^2(\mathbf{\Theta}) & \\ & & \mathbf{0}_{(d-2r)\times(d-2r)}\end{bmatrix}\right) = r\cdot\nu_1\cdot\cos^2(\theta) + r\cdot\nu_2\cdot\sin^2(\theta),\end{aligned}$$

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 (21) yields

$$\begin{aligned}\mathbb{E}\left[\left(\widetilde{y} - g^\star_{\mathtt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = &\left(\frac{1}{m}\left(r\nu_1^2 + r\nu_2^2 + (d-2r)\nu_3^2\right) + 1\right)\left(r + \epsilon d + \sigma^2\right) \\ &- 2\left(r\nu_1\cos^2(\theta) + r\nu_2\sin^2(\theta) + \left(r\nu_1 + r\nu_2 + (d-2r)\nu_3\right)\epsilon\right) \\ &+ \frac{m+1}{m}\left(r\nu_1^2\cos^2(\theta) + r\nu_2^2\sin^2(\theta) + \left(r\nu_1^2 + r\nu_2^2 + (d-2r)\nu_3^2\right)\epsilon\right).\end{aligned}$$

Then, taking $\epsilon \to 0$:

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

Substituting $\gamma = 0.5$ and combining like terms yields

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

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

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

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

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

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

$\square$

G.1.4 Proof of Theorem 3

Proof. The proof is similar to that of Theorem 2. Again let $\widetilde{y} := \widetilde{y}_{m+1}$. By Lemma 1 and Lemma 4, we have

$$\mathbb{E}\left[\left(\widetilde{y} - g^{\star}_{\mathtt{ATT}}(\tilde{\mathbf{z}}_q, \widetilde{\mathbf{Z}}_{\mathcal{M}})\right)^2\right] = \left(\frac{1}{m}\operatorname{Tr}\left(\mathbf{A}^\top\mathbf{A}\right)+1\right)\left(\operatorname{Tr}\left(\overline{\mathbf{\Sigma}}_t\right)+\sigma^2\right) - 2\operatorname{Tr}\left(\overline{\mathbf{\Sigma}}_t\mathbf{A}\right) + \frac{m+1}{m}\operatorname{Tr}\left(\mathbf{A}\overline{\mathbf{\Sigma}}_t\mathbf{A}^\top\right), \tag{23}$$

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

Let $\mathbf{U} := [\mathbf{U}_{s,1} \quad \mathbf{U}_{s,2} \quad \dots \quad \mathbf{U}_{s,K} \quad \mathbf{U}_\perp]$, where $\mathbf{U}_\perp \in \mathbb{R}^{d\times(d-Kr)}$ completes the orthonormal basis for $\mathbb{R}^d$. By Lemma 5,

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

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

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

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

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

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

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

Thus,

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

Using a similar argument,

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

Simplifying the test risk. Substituting the expressions for the $\operatorname{Tr}(\cdot)$ terms into Equation (23) yields

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

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

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

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

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

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

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

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

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

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

□

G.2 Auxiliary Results

G.2.1 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 loss in Equation (10). The following results are largely inspired by Theorem 1 in [15], but are slightly different since we consider a normalization factor of $1/n$ in our linear attention model.

Lemma 2 (Optimal Attention Weights [15])**.** *Consider the independent data model in Equation (1) where the task vector is drawn from $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_s)$, 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 (10) are given by*

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

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

Proof. The proof is the same as that of Theorem 1 in [15] by absorbing the $1/n$ factor into $\mathbf{W}_Q$. □

Lemma 3 (Optimal Attention Weights for Mixture of 2 Gaussians)**.** *Consider the independent data model in Equation (1) where the task vector is drawn from* $\mathbf{w} \sim \gamma \cdot \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_s) + (1-\gamma) \cdot \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{s,\perp})$ *for some* $\gamma \in (0,1)$*. Let* $n \in \mathbb{N}$ *denote the in-context prompt length used at training. Define* $\mathbf{\Sigma} = \gamma \cdot \mathbf{\Sigma}_s + (1-\gamma) \cdot \mathbf{\Sigma}_{\mathbf{s},\perp}$*. Then, the optimal linear attention weights obtained by minimizing the loss in Equation (10) are given by*

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

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 $\mathbf{w} \sim \gamma \cdot \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_s) + (1-\gamma) \cdot \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{s,\perp})$, then

$$\mathbf{\Sigma} := \mathrm{Cov}(\mathbf{w}) = \gamma \cdot \mathbf{\Sigma}_s + (1-\gamma) \cdot \mathbf{\Sigma}_{s,\perp}.$$

Then, the proof from Lemma 2 follows verbatim by using $\mathbf{\Sigma}$ instead of $\mathbf{\Sigma}_s$. ☐

Lemma 4 (Optimal Attention Weights for Mixture of K Gaussians)**.** *Consider the independent data model in Equation (1) where the task vector is drawn from* $\mathbf{w} \sim \sum_{k=1}^{K} \gamma_k \cdot \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{s,k}\right)$ *with* $\gamma_k \in (0,1)$ *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 (10) are given by*

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

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. The proof is equivalent to that of Lemma 3 by letting $\mathbf{\Sigma} = \sum_{k=1}^{K} \gamma_k \mathbf{\Sigma}_{s,k}$ instead. ☐

G.2.2 Miscellaneous Results

Lemma 5. *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. ☐

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper is about in-context learning, which the main body of the paper discusses.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation of the paper in the conclusion, which is simplification of the model to linear attention for analysis.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The proofs available in the Appendix clearly outlines the assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We discuss the experimental setup in the main body, as well as the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release the code upon completion of the double-blind review process.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We clearly discuss the experimental setup in the main body, as well as the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We include error bars for the LoRA experiments. Error bars were omitted for the experiments supporting the theory, as the theoretical results describe the risk in expectation; thus, reporting the average aligns naturally with the theory.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the guidelines and our paper reflects this.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There are no societal impacts of this paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not pose such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the appropriate papers where necessary.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release our code upon completion of the review process.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper is about emergent capabilities of LLMs. LLMs are the backbone for ICL, which we use for experiments (e.g., GPT-2). LLMs were not used for any other components of this paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.