Fine-grained Analysis of In-context Linear Estimation

Yingcong Li   YINGCONG@UMICH.EDU
University of Michigan, Ann Arbor

Ankit Singh Rawat   ANKITSRAWAT@GOOGLE.COM
Google Research NYC

Samet Oymak   OYMAK@UMICH.EDU
University of Michigan, Ann Arbor

Abstract

In this work, we develop a stronger characterization of the optimization and generalization landscape of ICL through contributions on architectures, low-rank parameterization, and correlated designs: (1) We study the landscape of 1-layer linear attention and 1-layer H3, a state-space model. Under a suitable correlated design assumption, we prove that both implement 1-step preconditioned gradient descent. (2) By studying correlated designs, we provide new risk bounds for retrieval augmented generation (RAG) and task-feature alignment which reveal how ICL sample complexity benefits from distributional alignment. (3) We derive the optimal risk for low-rank parameterized attention weights in terms of covariance spectrum. Through this, we also shed light on how LoRA can adapt to a new distribution by capturing the shift between task covariances.

1. Introduction

ICL capability has become an important feature of LLM and its ability presents an important research avenue to develop stronger theoretical and mechanistic understanding of LLMs. To this aim, there has been significant recent interest in demystifying ICL [1, 18, 26, 29, 44, 46]. A notable result in this direction is the observation that multilayer linear attention models [38, 40] implement *preconditioned gradient descent* (PGD) during ICL [1]. While this line of works provide a fresh perspective to ICL, the existing studies do not address many questions arising from real-life applications.

To this aim, we revisit the theoretical exploration of ICL with linear data model and make the following contributions:

1. We investigate the landscape of linear attention and H3 [10], a widely popular state-space model (SSM), and prove that under correlated design, both implement 1-step PGD (c.f. Proposition 3). Our analysis reveals that the gating mechanism in H3 imitates attention. We also empirically show that H3 has the advantage of implementing sample-weighting allowing it to outperform linear attention in temporally-heterogeneous settings in Appendix D and Figure 4.

2. We assess the impact of distributional alignment on the sample complexity of ICL. Specifically, we characterize the performance of *Retrieval Augmented Generation* (RAG) (c.f. Theorem 6) and *Task-Feature Alignment* (c.f. Theorem 7), where the in-context examples are α-correlated with either the query or the task vector. For both settings, we prove that alignment amplifies the *effective sample size* of ICL by a factor of $\alpha^2 d + 1$, highlighting that aligned data are crucial for the success of ICL in few-shot settings.

3. We show that, under low-rank parameterization, optimal attention-weights still implements PGD according to the truncated eigenspectrum of the fused task-feature covariance (see Section 3.2). We similarly derive risk upper bounds for LoRA adaptation (c.f. Eq. (13)), and show that, these bounds accurately predict the empirical performance.

2. Problem Setup and Preliminaries

We construct the input in-context prompt similar to [1, 29, 46] as follows.

Linear data distribution. Let $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ be a (feature, label) pair generated by a linear model parameterized by $\beta \in \mathbb{R}^d$, i.e., $y = x^\top \beta + \xi$, where x and β are feature and task vectors, and ξ is the label noise. Given demonstrations $(x_i, y_i)_{i=1}^{n+1}$ sampled from a single β, define the input prompt

$$Z = [z_1 \ \ldots \ z_n \ z_{n+1}]^\top = \begin{bmatrix} x_1 & \ldots & x_n & x_{n+1} \\ y_1 & \ldots & y_n & 0 \end{bmatrix}^\top \in \mathbb{R}^{(n+1)\times(d+1)}. \tag{1}$$

Then, given Z, the goal of the model is to predict the correct label y_{n+1} corresponding to x_{n+1}. For cleaner notation, when it is clear from context, we drop the subscript $n+1$ and set $x = x_{n+1}$, $z = z_{n+1}$. Different from the previous work [1, 29, 30, 46] where $(x_i)_{i=1}^{n+1}$ and β are assumed to be independent, our analysis focuses on a more general setting that captures the dependency between $(x_i)_{i=1}^{n+1}$ and β.

Model architectures. Similar to the previous work [1, 29, 40, 46] and to simplify the model structure, we focus on single-layer models and omit the nonlinearity from the Transformer. Given the input prompt $Z \in \mathbb{R}^{(n+1)\times(d+1)}$ in (1), which can be treated as a sequence of $(d+1)$-dimensional tokens, the single-layer linear attention ATT and H3-like single-layer SSM SSM are denoted by

$$\mathtt{ATT}(Z) = (ZW_qW_k^\top Z^\top)ZW_v \tag{2a}$$

$$\mathtt{SSM}(Z) = \left((ZW_q) \odot ((ZW_k \odot ZW_v) * f) \right) \tag{2b}$$

where W_k, W_q, $W_v \in \mathbb{R}^{(d+1)\times(d+1)}$ denote the key, query and value weight matrices, respectively. In (2b), the parameter $f \in \mathbb{R}^{n+1}$ is a 1-D convolutional filter that mixes tokens.

2.1. In-context Linear Estimation

We will next study the algorithms that can be implemented by the single-layer models ATT and SSM.

Background: 1-step gradient descent. Consider minimizing squared loss and solving linear regression using one step of PGD and WPGD. Given n samples $(x_i, y_i)_{i=1}^n$, define

$$X = [x_1 \ \cdots \ x_n]^\top \in \mathbb{R}^{n\times d} \quad \text{and} \quad y = [y_1 \ \cdots \ y_n]^\top \in \mathbb{R}^n.$$

Starting from $\beta_0 = 0_d$ and letting $\eta = 1/2$ be the step size, a single-step (sample-weighted) GD preconditioned with weights $W \in \mathbb{R}^{d\times d}$ and $\omega \in \mathbb{R}^n$ return predictions

$$\hat{y} = x^\top W X^\top y := g_{\mathtt{PGD}}(Z), \quad \text{and} \quad \hat{y} = x^\top W X^\top (\omega \odot y) := g_{\mathtt{WPGD}}(Z), \tag{3}$$

where Z is defined in (1) consisting of X, y and x. Our goal is to find the optimal W, as well as ω that minimize the population risks defined as follows.

$$\min_W \mathcal{L}_{\mathtt{PGD}}(\mathcal{W}) \quad \text{where} \quad \mathcal{L}_{\mathtt{PGD}}(\mathcal{W}) = \mathbb{E}\left[(y - g_{\mathtt{PGD}}(Z))^2 \right], \tag{4a}$$

$$\min_{W,\omega} \mathcal{L}_{\mathtt{WPGD}}(\mathcal{W}) \quad \text{where} \quad \mathcal{L}_{\mathtt{WPGD}}(\mathcal{W}) = \mathbb{E}\left[(y - g_{\mathtt{WPGD}}(Z))^2 \right]. \tag{4b}$$

Here, the expectation is over the randomness in $(\mathbf{x}_i, \xi_i)_{i=1}^{n+1}$ and $\boldsymbol{\beta}$, and we use $\mathcal{W}$ to represent the set of corresponding trainable parameters. The search spaces for $\boldsymbol{\omega}$ and $\mathbf{W}$ are $\mathbb{R}^n$ and $\mathbb{R}^{d \times d}$, respectively.

Note that similar to [1, 29], we consider a training objective with a causal mask to ensure inputs cannot attend to their own labels and training can be parallelized. Let $\mathbf{Z}_0 = [\mathbf{z}_1 \ \dots \ \mathbf{z}_n \ \mathbf{0}]^\top$ be the features post-causal masking at time/index $n + 1$. Given weights $\mathbf{W}_k, \mathbf{W}_q, \mathbf{W}_v$ and the filter $\boldsymbol{f}$ for SSM, predictions at the query token $\mathbf{z}$ take the forms following sequence-to-sequence mappings in (2):

$$g_{\text{ATT}}(\mathbf{Z}) = (\mathbf{z}^\top \mathbf{W}_q \mathbf{W}_k^\top \mathbf{Z}_0^\top) \mathbf{Z}_0 \mathbf{W}_v \mathbf{v}, \quad \text{and} \quad g_{\text{SSM}}(\mathbf{Z}) = \left((\mathbf{z}^\top \mathbf{W}_q)^\top \odot ((\mathbf{Z}_0 \mathbf{W}_k \odot \mathbf{Z}_0 \mathbf{W}_v) * \boldsymbol{f})_{n+1} \right) \mathbf{v},$$

where $\mathbf{v} \in \mathbb{R}^{d+1}$ is the linear prediction head and $((\mathbf{Z}_0 \mathbf{W}_k \odot \mathbf{Z}_0 \mathbf{W}_v) * \boldsymbol{f})_{n+1}$ returns the last row of the convolution output. Note that SSM can implement the mask by setting $f_0 = 0$. Thus, the objectives for both models take the following forms.

$$\min_{\mathbf{W}_k, \mathbf{W}_q, \mathbf{W}_v, \mathbf{v}} \mathcal{L}_{\text{ATT}}(\mathcal{W}) \quad \text{where} \quad \mathcal{L}_{\text{ATT}}(\mathcal{W}) = \mathbb{E}\left[(y - g_{\text{ATT}}(\mathbf{Z}))^2 \right], \tag{5a}$$

$$\min_{\mathbf{W}_k, \mathbf{W}_q, \mathbf{W}_v, \mathbf{v}, \boldsymbol{f}} \mathcal{L}_{\text{SSM}}(\mathcal{W}) \quad \text{where} \quad \mathcal{L}_{\text{SSM}}(\mathcal{W}) = \mathbb{E}\left[(y - g_{\text{SSM}}(\mathbf{Z}))^2 \right]. \tag{5b}$$

Here, similarly, the expectation subsumes the randomness of $(\mathbf{x}_i, \xi_i)_{i=1}^{n+1}$ and $\boldsymbol{\beta}$ and $\mathcal{W}$ represents the set of trainable parameters. The search space for matrices $\mathbf{W}_k, \mathbf{W}_q, \mathbf{W}_v$ is $\mathbb{R}^{(d+1) \times (d+1)}$, for head $\mathbf{v}$ is $\mathbb{R}^{d+1}$, and for $\boldsymbol{f}$ is $\mathbb{R}^{n+1}$.

Note that for all the optimization methods (c.f. (4), (5)), to simplify the analysis, we train the models without capturing additional bias terms. Therefore, in the following, we introduce the centralized data assumptions such that the models are trained to make unbiased predictions.

To begin with, a cross moment of random variables is defined as the expectation of a monomial of these variables, with the order of the cross moment being the same as order of the monomial. Then, it motivates the following data assumptions.

Assumption 1 *All cross moments of the entries of $(\mathbf{x}_i)_{i=1}^{n+1}$ and $\boldsymbol{\beta}$ with odd orders are zero.*

Assumption 2 *$(\xi_i)_{i=1}^{n+1}$ are independent of $(\mathbf{x}_i)_{i=1}^{n+1}, \boldsymbol{\beta}$, and their odd-order cross moments are zero.*

Note that compared to [1, 29, 46], Assumption 1 is more general which also subsumes the dependent data setting. Next, we introduce the following result establishing the equivalence among optimizing 1-layer linear attention (c.f. (5a)), 1-layer H3 (c.f. (5b)), and 1-step gradient descent (c.f. (4)).

Proposition 3 *Suppose Assumptions 1 and 2 hold. Consider the objectives as defined in (4) and (5), and let $\mathcal{L}_{\text{PGD}}^\star, \mathcal{L}_{\text{WPGD}}^\star, \mathcal{L}_{\text{ATT}}^\star,$ and $\mathcal{L}_{\text{SSM}}^\star$ be their optimal risks, respectively. Then,*

$$\mathcal{L}_{\text{PGD}}^\star = \mathcal{L}_{\text{ATT}}^\star \quad \text{and} \quad \mathcal{L}_{\text{WPGD}}^\star = \mathcal{L}_{\text{SSM}}^\star.$$

Additionally, if the examples $(\mathbf{x}_i, y_i)_{i=1}^n$ follow the same distribution and are conditionally independent given $\mathbf{x}, \boldsymbol{\beta}$, then SSM/H3 can achieve the optimal loss using the all-ones filter and $\mathcal{L}_{\text{PGD}}^\star = \mathcal{L}_{\text{SSM}}^\star$.

We defer the proof to Appendix A.1. Proposition 3 establishes that analyzing the optimization landscape of ICL for both single-layer linear attention and the H3 model can be effectively reduced to examining the behavior of a one-step PGD algorithm. Notably, under the independent, RAG and task-feature alignment data settings discussed above, examples $(\mathbf{x}_i, y_i)_{i=1}^n$ are independently sampled given $\mathbf{x}$ and $\boldsymbol{\beta}$, and we therefore conclude that $\mathcal{L}_{\text{PGD}}^\star = \mathcal{L}_{\text{ATT}}^\star = \mathcal{L}_{\text{SSM}}^\star$. Leveraging this result, the subsequent section of the paper concentrate on addressing (4a), taking into account various linear data distributions.

3. Main Results

In light of Proposition 3, optimizing a 1-layer linear-attention or H3 model is equivalent to solving the objective (4a). Therefore, in this section, we examine the properties of the one-step PGD in (4a) considering multiple problem settings, including distinct data distributions and low-rank training.

3.1. Analysis of Linear Data Models

We first consider the standard independent data setting. We will then examine correlated designs.

Independent data model. Let Σ_x and Σ_β be the covariance matrices of the input feature and task vectors, respectively, and $\sigma \geq 0$ be the noise level. We assume

$$\beta \sim \mathcal{N}(0, \Sigma_\beta), \quad x_i \sim \mathcal{N}(0, \Sigma_x), \quad \xi_i \sim \mathcal{N}(0, \sigma^2), \quad 1 \leq i \leq n+1 \tag{6}$$

and the label is obtained via $y_i = x_i^\top \beta + \xi_i$. Our following result characterizes the optimal solution of (4a). Note that the data generated from (6) satisfies the conditions in Proposition 3. Therefore, the same results can be applied to both linear-attention and H3 models.

Theorem 4 *Consider independent linear data as defined in (6), and suppose the covariance matrices Σ_x, Σ_β are full rank. Recap the objective from (4a) and let $W_\star := \arg\min_W \mathcal{L}_{PGD}(W)$, and $\mathcal{L}_\star = \mathcal{L}_{PGD}(W_\star)$. Additionally, let $\Sigma = \Sigma_x^{1/2} \Sigma_\beta \Sigma_x^{1/2}$ and $M = \mathtt{tr}(\Sigma) + \sigma^2$. Then $W_\star$ and $\mathcal{L}_\star$ satisfy*

$$W_\star = \Sigma_x^{-1/2} \bar{W}_\star \Sigma_x^{-1/2} \quad and \quad \mathcal{L}_\star = M - n\,\mathtt{tr}\left(\Sigma \bar{W}_\star\right), \tag{7}$$

where we define $\bar{W}_\star = \left((n+1)I_d + M\Sigma^{-1}\right)^{-1}$.

Corollary 5 *Consider noiseless i.i.d. linear data where $\Sigma_x = \Sigma_\beta = I_d$ and $\sigma = 0$. Then, the objective in (4a) returns*

$$W_\star = \frac{1}{n+d+1} I_d \quad and \quad \mathcal{L}_\star = d - \frac{nd}{n+d+1}.$$

See Appendix B.2 for proofs. Note that Theorem 4 is consistent with prior work [1, Theorem 1] when specialized to isotropic task covariance. However, their result is limited as the features and task are assumed to be independent. This prompts us to ask: *What is the optimization landscape with correlated in-context samples?* Toward this, we consider the following correlated linear models.

Retrieval augmented generation. To provide a statistical model of the practical RAG approaches, given the query vector $x_{n+1} = x$, we propose to draw ICL demonstrations that are similar to x with the same shared task vector β. As an approximate proxy, we assume that $\beta \sim \mathcal{N}(0, I_d)$, $x \sim \mathcal{N}(0, I_d)$ and that RAG samples α-correlated demonstrations $(x_i, y_i)_{i=1}^n$ as follows:

$$x_i \mid x \sim \mathcal{N}(\alpha x, (1-\alpha^2) I_d), \quad \xi_i \sim \mathcal{N}(0, \sigma^2) \quad and \quad y_i = x_i^\top \beta + \xi_i, \quad 1 \leq i \leq n. \tag{8}$$

The full analysis of RAG is provides in Appendix B.3. Specifically, when we carry out the analysis by assuming $\alpha = O\left(1/\sqrt{d}\right)$ and $d/n = O(1)$, our derivation leads to the following result:

Theorem 6 *Consider linear model as defined in (8). Recap the objective from (4a) and let $W_\star := \arg\min_W \mathcal{L}_{PGD}(W)$, and $\mathcal{L}_\star = \mathcal{L}_{PGD}(W_\star)$. Additionally, let $\kappa = \alpha^2 d + 1$ and suppose $\alpha = O\left(1/\sqrt{d}\right)$, $d/n = O(1)$ and d is sufficiently large. Then $W_\star$ and $\mathcal{L}_\star$ have approximate forms*

$$W_\star \approx \frac{1}{\kappa n + d + \sigma^2} I_d \quad and \quad \mathcal{L}_\star \approx d + \sigma^2 - \frac{\kappa nd}{\kappa n + d + \sigma^2}. \tag{9}$$

4

Task-feature alignment. We also consider another dependent data setting where task and feature vectors are assumed to be correlated. Assume that the shared task vector is sampled by $\boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and letting $\kappa = \alpha^2 d + 1$, we sample the α-correlated ICL demonstrations $(\boldsymbol{x}_i, y_i)_{i=1}^{n+1}$ as follows:

$$\boldsymbol{x}_i \,\big|\, \boldsymbol{\beta} \sim \mathcal{N}(\alpha\boldsymbol{\beta}, \boldsymbol{I}_d), \quad \xi_i \sim \mathcal{N}(0, \sigma^2) \quad \text{and} \quad y_i = \kappa^{-1/2} \boldsymbol{x}_i^\top \boldsymbol{\beta} + \xi_i, \quad 1 \le i \le n+1. \tag{10}$$

Above, $\kappa^{-1/2}$ is a normalization factor to ensure that label variance remains invariant to α. To keep the exposition cleaner, we defer the full analysis to Theorem 14 in Appendix B.4. Similar to the RAG setting, we obtain the following results for the optimal parameter and risk.

Theorem 7 *Consider linear model as defined in* (10). *Recap the objective from* (4a) *and let* $\boldsymbol{W}_\star := \arg\min_{\boldsymbol{W}} \mathcal{L}_{PGD}(\boldsymbol{W})$, *and* $\mathcal{L}_\star = \mathcal{L}_{PGD}(\boldsymbol{W}_\star)$. *Additionally, given* $\kappa = \alpha^2 d + 1$ *and suppose* $\alpha = O\left(1/\sqrt{d}\right)$, $d/n = O(1)$ *and* d *is sufficiently large. Then* $\boldsymbol{W}_\star$ *and* $\mathcal{L}_\star$ *have approximate forms*

$$\boldsymbol{W}_\star \approx \frac{1}{\kappa n + (d + \sigma^2)/\kappa} \boldsymbol{I}_d \qquad \text{and} \qquad \mathcal{L}_\star \approx d + \sigma^2 - \frac{\kappa n d}{\kappa n + (d + \sigma^2)/\kappa}. \tag{11}$$

Here, (9) and (11) are reminiscent of Corollary 5 and setting $\alpha = 0$ reduces to the results of Corollary 5. Observe that, $\alpha^2 d + 1$ is the dominant multiplier ahead of n in both results. Thus, we deduce that, α-correlated models reduce the in-context sample complexity by a factor of $\alpha^2 d + 1$.

3.2. Low-rank Parameterization and LoRA

In this section, we investigate training low-rank models, which assume $\boldsymbol{W}_k, \boldsymbol{W}_q \in \mathbb{R}^{(d+1)\times r}$ where r is the rank restriction. Equivalently, we consider objective (4a) under condition $\text{rank}(\boldsymbol{W}) = r$.

Lemma 8 *Consider independent linear data as defined in* (6). *Recap the objective from* (4a) *and enforce* $\text{rank}(\boldsymbol{W}) \le r$ *and* $\boldsymbol{W}^\top = \boldsymbol{W}$. *Let* $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_x^{1/2} \boldsymbol{\Sigma}_\beta \boldsymbol{\Sigma}_x^{1/2}$ *and* $M = \text{tr}(\boldsymbol{\Sigma}) + \sigma^2$. *Denoting* λ_i *to be the* i'*th largest eigenvalue of* $\boldsymbol{\Sigma}$, *we have that*

$$\min_{\text{rank}(\boldsymbol{W}) \le r, \boldsymbol{W} = \boldsymbol{W}^\top} \mathcal{L}(\boldsymbol{W}) = M - \sum_{i=1}^{r} \frac{n \lambda_i^2}{(n+1)\lambda_i + M}. \tag{12}$$

Note that $\text{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^{d} \lambda_i$. Removing the rank constraint and considering noiseless data setting, this reduces to the following optimal risk $\mathcal{L}_\star = \sum_{i=1}^{d} \frac{\lambda_i + M}{n + 1 + M/\lambda_i}$. See Appendix C.1 for more details.

Impact of LoRA: Based on the above lemma, we consider the impact of LoRA for adapting the pretrained model to a new task distribution under jointly-diagonalizable old and new eigenvalues of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{new}$, $(\lambda_i)_{i=1}^{d}$, $(\lambda_i^{new})_{i=1}^{d}$. Consider adapting LoRA matrix to the combined key and value weights in attention, which reflects minimizing the population loss $\tilde{\mathcal{L}}(\boldsymbol{W}_{lora}) := \mathcal{L}(\boldsymbol{W} + \boldsymbol{W}_{lora})$ in (4a) with fixed $\boldsymbol{W}$. Suppose $\text{tr}(\boldsymbol{\Sigma}) = \text{tr}(\boldsymbol{\Sigma}^{new}) = M$, $\sigma = 0$ and $\boldsymbol{W}$ is jointly diagonalizable with $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{new}$, then LoRA's risk is upper-bounded by

$$\min_{\text{rank}(\boldsymbol{W}_{lora}) \le r} \tilde{\mathcal{L}}(\boldsymbol{W}_{lora}) \le \min_{|\mathcal{I}| \le r, \mathcal{I} \subset [d]} \left(\sum_{i \notin \mathcal{I}} \frac{\lambda_i + M}{n + 1 + M/\lambda_i} + \sum_{i \in \mathcal{I}} \frac{\lambda_i^{new} + M}{n + 1 + M/\lambda_i^{new}} \right). \tag{13}$$

Note that, the right hand side is provided assuming the optimal LoRA-updated model $\boldsymbol{W}_{lora}$ is also jointly diagonalizable with covariances $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{new}$, and $\boldsymbol{W}$.

References

[1] 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, 2023.

[2] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

[3] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36, 2024.

[4] Soumya Basu, Ankit Singh Rawat, and Manzil Zaheer. A statistical perspective on retrieval-based models. In *International Conference on Machine Learning*, pages 1852–1886. PMLR, 2023.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Liam Collins, Advait Parulekar, Aryan Mokhtari, Sujay Sanghavi, and Sanjay Shakkottai. In-context learning with transformers: Softmax attention adapts to function lipschitzness. *arXiv preprint arXiv:2402.11639*, 2024.

[7] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.247. URL https://aclanthology.org/2023.findings-acl.247.

[8] Zhe Du, Haldun Balim, Samet Oymak, and Necmiye Ozay. Can transformers learn optimal filtering for unknown systems? *IEEE Control Systems Letters*, 7:3525–3530, 2023.

[9] Karthik Duraisamy. Finite sample analysis and bounds of generalization error of gradient descent in in-context linear regression. *arXiv preprint arXiv:2405.02462*, 2024.

[10] Daniel Y Fu, Tri Dao, Khaled Kamal Saab, Armin W Thomas, Atri Rudra, and Christopher Re. Hungry hungry hippos: Towards language modeling with state space models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=COZDy0WYGg.

[11] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

[12] 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 *Forty-first International Conference on Machine Learning*.

[13] GeminiTeam, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[14] Riccardo Grazzi, Julien Siems, Simon Schrodi, Thomas Brox, and Frank Hutter. Is mamba capable of in-context learning? *arXiv preprint arXiv:2402.03170*, 2024.

[15] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[16] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2021.

[17] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021.

[18] Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. In-context learning of large language models explained as kernel regression. *arXiv preprint arXiv:2305.12766*, 2023.

[19] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

[20] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023.

[21] Ivan Lee, Nan Jiang, and Taylor Berg-Kirkpatrick. Exploring the relationship between model architecture and in-context learning ability. *arXiv preprint arXiv:2310.08049*, 2023.

[22] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[23] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pages 19565–19594. PMLR, 2023.

[24] Yingcong Li, Kartik Sreenivasan, Angeliki Giannou, Dimitris Papailiopoulos, and Samet Oymak. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[25] Ziqian Lin and Kangwook Lee. Dual operating modes of in-context learning. *arXiv preprint arXiv:2402.18819*, 2024.

[26] Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=De4FYqjFueZ.

[27] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[28] Yue Lu, Mary I. Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *arXiv preprint arXiv:2310.08391*, 2024.

[29] 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. URL https://openreview.net/forum?id=8p3fu56lKc.

[30] 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.

[31] Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. *arXiv preprint arXiv:2112.10510*, 2021.

[32] Samuel Müller, Matthias Feurer, Noah Hollmann, and Frank Hutter. Pfns4bo: In-context learning for bayesian optimization. In *International Conference on Machine Learning*, pages 25444–25470. PMLR, 2023.

[33] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[34] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[35] OpenAI. Gpt-4 technical report. *arXiv preprintarXiv:2303.08774*, 2023.

[36] Jongho Park, Jaeseung Park, Zheyang Xiong, Nayoung Lee, Jaewoong Cho, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. Can mamba learn how to learn? a comparative study on in-context learning tasks. *International Conference on Machine Learning*, 2024.

[37] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

[38] Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, pages 9355–9366. PMLR, 2021.

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[40] 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.

[41] Johannes von Oswald, Eyvind Niklasson, Maximilian Schlegel, Seijin Kobayashi, Nicolas Zucchet, Nino Scherrer, Nolan Miller, Mark Sandler, Max Vladymyrov, Razvan Pascanu, et al. Uncovering mesa-optimization algorithms in transformers. *arXiv preprint arXiv:2309.05858*, 2023.

[42] Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. Training data is more valuable than you think: A simple and effective method by retrieving from training data. *arXiv preprint arXiv:2203.08773*, 2022.

[43] Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter L Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? *arXiv preprint arXiv:2310.08391*, 2023.

[44] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

[45] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

[46] 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.

[47] Nicolas Zucchet, Seijin Kobayashi, Yassir Akram, Johannes Von Oswald, Maxime Larcher, Angelika Steger, and Joao Sacramento. Gated recurrent neural networks discover attention. *arXiv preprint arXiv:2309.01775*, 2023.

Table of Contents

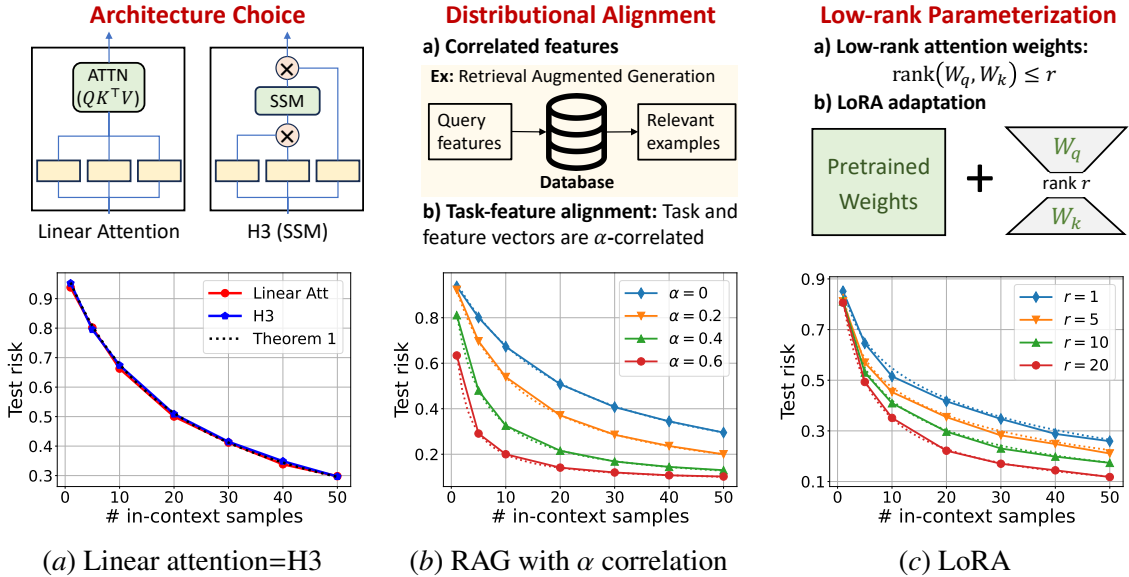(a) Linear attention=H3   (b) RAG with α correlation   (c) LoRA

Figure 1: In this work, we investigate the optimization landscape of in-context learning from the lens of architecture choice, the role of distributional alignment, and low-rank parameterization. The empirical performance (solid curves) are aligned with our theoretical results (dotted curves) from Section 3. More experimental details and discussion are discussed in Appendix D.

Appendix A. Equivalence among Gradient Descent, Attention, and SSMs

In this section, we present the proofs related to Section 2. Recap that given data

$$
\begin{aligned}
X &= [x_1 \ \cdots \ x_n]^\top \in \mathbb{R}^{n \times d}, \\
\xi &= [\xi_1 \ \cdots \ \xi_n]^\top \in \mathbb{R}^n, \\
y &= [y_1 \ \cdots \ y_n]^\top = X\beta + \xi \in \mathbb{R}^n, \\
Z_0 &= [z_1 \ \ldots \ z_n \ \mathbf{0}_{d+1}]^\top = \begin{bmatrix} x_1 & \ldots & x_n & \mathbf{0}_d \\ y_1 & \ldots & y_n & 0 \end{bmatrix}^\top \in \mathbb{R}^{(n+1) \times (d+1)},
\end{aligned}
$$

and corresponding prediction functions

$$
\begin{aligned}
g_{\text{PGD}}(Z) &= x^\top W X^\top y, &&\text{(14a)} \\
g_{\text{WPGD}}(Z) &= x^\top W X^\top (\omega \odot y), &&\text{(14b)} \\
g_{\text{ATT}}(Z) &= (z^\top W_q W_k^\top Z_0^\top) Z_0 W_v v, &&\text{(14c)} \\
g_{\text{SSM}}(Z) &= \left((z^\top W_q)^\top \odot ((Z_0 W_k \odot Z_0 W_v) * f)_{n+1} \right) v, &&\text{(14d)}
\end{aligned}
$$

we have objectives

$$
\begin{aligned}
\min_{W} \mathcal{L}_{\text{PGD}}(W) \quad &\text{where} \quad \mathcal{L}_{\text{PGD}}(W) = \mathbb{E}\left[(y - g_{\text{PGD}}(Z))^2 \right], &&\text{(15a)} \\[4pt]
\min_{W, \omega} \mathcal{L}_{\text{WPGD}}(W) \quad &\text{where} \quad \mathcal{L}_{\text{WPGD}}(W) = \mathbb{E}\left[(y - g_{\text{WPGD}}(Z))^2 \right], &&\text{(15b)} \\[4pt]
\min_{W_k, W_q, W_v, v} \mathcal{L}_{\text{ATT}}(W) \quad &\text{where} \quad \mathcal{L}_{\text{ATT}}(W) = \mathbb{E}\left[(y - g_{\text{ATT}}(Z))^2 \right], &&\text{(15c)} \\[4pt]
\min_{W_k, W_q, W_v, v, f} \mathcal{L}_{\text{SSM}}(W) \quad &\text{where} \quad \mathcal{L}_{\text{SSM}}(W) = \mathbb{E}\left[(y - g_{\text{SSM}}(Z))^2 \right]. &&\text{(15d)}
\end{aligned}
$$

Here, the expectation is over the randomness in $(x_i, \xi_i)_{i=1}^n$ and β, and the search space for W is $\mathbb{R}^{d \times d}$, for ω is $\mathbb{R}^n$, for W_k, W_q, W_v is $\mathbb{R}^{(d+1) \times (d+1)}$, for v is $\mathbb{R}^{d+1}$, and for f is $\mathbb{R}^{n+1}$.

A.1. Proof of Proposition 3

Consider the problem setting as discussed in Section 2, Proposition 3 can be proven by the following two lemmas.

Lemma 9 *Suppose Assumptions 1 and 2 hold. Then, given the objectives* (15a) *and* (15c)*, we have*

$$
\min_{W_q, W_k, W_v, v} \mathcal{L}_{ATT}(W) = \min_{W} \mathcal{L}_{PGD}(W).
$$

Proof Recap the linear attention estimator from (14c) and denote

$$
W_q W_k^\top = \begin{bmatrix} \bar{W} & w_1 \\ w_2^\top & w \end{bmatrix} \qquad \text{and} \qquad W_v v = \begin{bmatrix} v_1 \\ v \end{bmatrix},
$$

where $\bar{W} \in \mathbb{R}^{d \times d}$, $w_1, w_2, v_1 \in \mathbb{R}^d$, and $w, v \in \mathbb{R}$. Then we have

$$
\begin{aligned}
g_{\text{ATT}}(Z) &= (z^\top W_q W_k^\top Z_0^\top) Z_0 W_v v \\
&= [x^\top \ 0] \begin{bmatrix} \bar{W} & w_1 \\ w_2^\top & w \end{bmatrix} \begin{bmatrix} X^\top & 0_d \\ y^\top & 0 \end{bmatrix} \begin{bmatrix} X & y \\ 0_d^\top & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v \end{bmatrix} \\
&= (x^\top \bar{W} X^\top + x^\top w_1 y^\top)(X v_1 + y v) \\
&= x^\top (v\bar{W}) X^\top y + x^\top w_1 y^\top X v_1 + x^\top \left(\bar{W} X^\top X v_1 + v \|y\|_{\ell_2}^2 w_1 \right) \\
&= x^\top (v\bar{W} + w_1 v_1^\top) X^\top y + x^\top \left(\bar{W} X^\top X v_1 + v \|y\|_{\ell_2}^2 w_1 \right) \\
&= \underbrace{x^\top \tilde{W} X^\top y}_{\tilde{g}_{\text{ATT}}(Z)} + \underbrace{x^\top \left(\bar{W} X^\top X v_1 + v \|y\|_{\ell_2}^2 w_1 \right)}_{\varepsilon},
\end{aligned}
\tag{16}
$$

where $\tilde{W} := v\bar{W} + w_1 v_1^\top$.

We first show that for any given parameters W_k, W_q, W_v, v,

$$
\mathbb{E}\left[(g_{\text{ATT}}(Z) - y)^2 \right] \geq \mathbb{E}\left[(\tilde{g}_{\text{ATT}}(Z) - y)^2 \right].
\tag{17}
$$

To this goal, we have

$$
\begin{aligned}
\mathbb{E}\left[(g_{\text{ATT}}(Z) - y)^2 \right] - \mathbb{E}\left[(\tilde{g}_{\text{ATT}}(Z) - y)^2 \right] &= \mathbb{E}\left[(\tilde{g}_{\text{ATT}}(Z) + \varepsilon - y)^2 \right] - \mathbb{E}\left[(\tilde{g}_{\text{ATT}}(Z) - y)^2 \right] \\
&= \mathbb{E}[\varepsilon^2] + 2\mathbb{E}[(\tilde{g}_{\text{ATT}}(Z) - y)\varepsilon]
\end{aligned}
\tag{18}
$$

where we have decomposition

$$
\begin{aligned}
(\tilde{g}_{\text{ATT}}(Z) - y)\varepsilon &= (x^\top \tilde{W} X^\top y - y) x^\top \left(\bar{W} X^\top X v_1 + v \|y\|_{\ell_2}^2 w_1 \right) \\
&= y^\top X \tilde{W}^\top x x^\top \left(\bar{W} X^\top X v_1 + v \|y\|_{\ell_2}^2 w_1 \right) - y x^\top \left(\bar{W} X^\top X v_1 + v \|y\|_{\ell_2}^2 w_1 \right) \\
&= \underbrace{y^\top X \tilde{W}^\top x x^\top \bar{W} X^\top X v_1}_{(a)} + \underbrace{v \|y\|_{\ell_2}^2 y^\top X \tilde{W}^\top x x^\top w_1}_{(b)} - \underbrace{y x^\top \bar{W} X^\top X v_1}_{(c)} - \underbrace{v y \|y\|_{\ell_2}^2 x^\top w_1}_{(d)}.
\end{aligned}
$$

In the following, we consider the expectations of $(a), (b), (c), (d)$ sequentially, which return zeros under Assumptions 1 and 2. Note that since Assumption 1 holds, expectation of any odd *order* of monomial of the entries of X, x, β returns zero, i.e., order of $x^\top \beta x$ is 3 and therefore $\mathbb{E}[x^\top \beta x] = 0_d$.

$(a):$ $\quad \mathbb{E}\left[y^\top X \tilde{W}^\top x x^\top \bar{W} X^\top X v_1 \right]$

$\qquad = \mathbb{E}\left[(X\beta + \xi)^\top X \tilde{W}^\top x x^\top \bar{W} X^\top X v_1 \right]$

$\qquad = \mathbb{E}\left[\beta^\top X^\top X \tilde{W}^\top x x^\top \bar{W} X^\top X v_1 \right] + \mathbb{E}\left[\xi^\top X \tilde{W}^\top x x^\top \bar{W} X^\top X v_1 \right]$

$\qquad = 0.$

$(b):$ $\quad \mathbb{E}\left[v \|y\|_{\ell_2}^2 y^\top X \tilde{W}^\top x x^\top w_1 \right]$

$\qquad = \mathbb{E}\left[v(X\beta + \xi)^\top (X\beta + \xi)(X\beta + \xi)^\top X \tilde{W}^\top x x^\top w_1 \right]$

$\qquad = \mathbb{E}\left[v \|\xi\|_{\ell_2}^2 \xi^\top X \tilde{W}^\top x x^\top w_1 \right]$

$\qquad = 0.$

12

$(c):\quad \mathbb{E}\left[y\mathbf{x}^\top \bar{\mathbf{W}} \mathbf{X}^\top \mathbf{X} \mathbf{v}_1\right]$

$\qquad = \mathbb{E}\left[(\mathbf{x}^\top \boldsymbol{\beta} + \xi)\mathbf{x}^\top \bar{\mathbf{W}} \mathbf{X}^\top \mathbf{X} \mathbf{v}_1\right]$

$\qquad = \mathbb{E}\left[\boldsymbol{\beta}^\top \mathbf{x}\mathbf{x}^\top \bar{\mathbf{W}} \mathbf{X}^\top \mathbf{X} \mathbf{v}_1\right] + \mathbb{E}\left[\xi\mathbf{x}^\top \bar{\mathbf{W}} \mathbf{X}^\top \mathbf{X} \mathbf{v}_1\right]$

$\qquad = 0.$

$(d):\quad \mathbb{E}\left[vy\,\|\mathbf{y}\|_{\ell_2}^2\,\mathbf{x}^\top \mathbf{w}_1\right]$

$\qquad = v\,\mathbb{E}\left[(\boldsymbol{\beta}^\top \mathbf{x} + \xi)(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi})^\top (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi})\mathbf{x}^\top \mathbf{w}_1\right]$

$\qquad = v\,\mathbb{E}\left[\xi\,\|\boldsymbol{\xi}\|_{\ell_2}^2\,\mathbf{x}^\top \mathbf{w}_1\right]$

$\qquad = 0.$

Combining the results with (18) returns that

$$\mathbb{E}\left[(g_{\text{ATT}}(\mathbf{Z}) - y)^2\right] - \mathbb{E}\left[(\tilde{g}_{\text{ATT}}(\mathbf{Z}) - y)^2\right] = \mathbb{E}[\varepsilon^2] \geq 0 \tag{19}$$

which completes the proof of (17). Therefore, we obtain

$$\min_{\mathbf{W}_q,\mathbf{W}_k,\mathbf{W}_v,\mathbf{v}} \mathbb{E}\left[(g_{\text{ATT}}(\mathbf{Z}) - y)^2\right] \geq \min_{\tilde{\mathbf{W}}} \mathbb{E}\left[(\tilde{g}_{\text{ATT}}(\mathbf{Z}) - y)^2\right] = \min_{\mathbf{W}} \mathbb{E}\left[(g_{\text{PGD}}(\mathbf{Z}) - y)^2\right].$$

We conclude the proof of this lemma by showing that for any $\mathbf{W} \in \mathbb{R}^{d\times d}$ in g_{PGD}, there exist $\mathbf{W}_k, \mathbf{W}_q, \mathbf{W}_v, \mathbf{v}$ such that $g_{\text{ATT}}(\mathbf{Z}) = g_{\text{PGD}}(\mathbf{Z})$. Let

$$\mathbf{W}_k = \mathbf{W}_v = \mathbf{I}_{d+1}, \qquad \mathbf{W}_q = \begin{bmatrix} \mathbf{W} & \mathbf{0}_d \\ \mathbf{0}_d^\top & 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{v} = \begin{bmatrix} \mathbf{0}_d \\ 1 \end{bmatrix}.$$

Then we obtain

$$g_{\text{ATT}}(\mathbf{Z}) = \mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \mathbf{y} = g_{\text{PGD}}(\mathbf{Z}), \tag{20}$$

which completes the proof. ∎

Lemma 10 *Suppose Assumptions 1 and 2 hold. Then, given the objectives in (15), we have*

$$\min_{\mathbf{W}_q,\mathbf{W}_k,\mathbf{W}_v,\mathbf{v},f} \mathcal{L}_{\text{SSM}}(\mathcal{W}) = \min_{\mathbf{W},\omega} \mathcal{L}_{\text{WPGD}}(\mathcal{W}). \tag{21}$$

Additionally, if the examples $(\mathbf{x}_i, y_i)_{i=1}^n$ follow the same distribution and are conditionally independent given $\mathbf{x}$ and $\boldsymbol{\beta}$, then SSM/H3 can achieve the optimal loss using the all-ones filter and

$$\min_{\mathbf{W},\omega} \mathcal{L}_{\text{WPGD}}(\mathcal{W}) = \min_{\mathbf{W}} \mathcal{L}_{\text{PGD}}(\mathcal{W}). \tag{22}$$

Proof Recap the SSM estimator from (14d) and let

$$\mathbf{W}_q = \begin{bmatrix} \mathbf{w}_{q1} & \mathbf{w}_{q2} & \cdots & \mathbf{w}_{q,d+1} \end{bmatrix},$$

$$\mathbf{W}_k = \begin{bmatrix} \mathbf{w}_{k1} & \mathbf{w}_{k2} & \cdots & \mathbf{w}_{k,d+1} \end{bmatrix},$$

$$\mathbf{W}_v = \begin{bmatrix} \mathbf{w}_{v1} & \mathbf{w}_{v2} & \cdots & \mathbf{w}_{v,d+1} \end{bmatrix},$$

where $w_{qj}, w_{kj}, w_{vj} \in \mathbb{R}^{d+1}$ for $j \le d+1$, and let

$$
v = \begin{bmatrix} v_1 \\ v_2 \\ \cdots \\ v_{d+1} \end{bmatrix}, \quad \text{and} \quad f = \begin{bmatrix} f_0 \\ f_1 \\ \cdots \\ f_n \end{bmatrix}.
$$

Then we have

$$
g_{\text{SSM}}(Z) = \left((z^\top W_q)^\top \odot ((Z_0 W_k \odot Z_0 W_v) * f)_{n+1} \right) v
$$

$$
= \sum_{i=1}^{n} f_{n+1-i} \cdot v^\top \left(\begin{bmatrix} w_{q1}^\top z \\ \cdots \\ w_{q,d+1}^\top z \end{bmatrix} \odot \begin{bmatrix} w_{k1}^\top z_i w_{v1}^\top z_i \\ \cdots \\ w_{k,d+1}^\top z_i w_{v,d+1}^\top z_i \end{bmatrix} \right)
$$

$$
= \sum_{i=1}^{n} f_{n+1-i} \cdot v^\top \begin{bmatrix} w_{q1}^\top z w_{k1}^\top z_i w_{v1}^\top z_i \\ \cdots \\ w_{q,d+1}^\top z w_{k,d+1}^\top z_i w_{v,d+1}^\top z_i \end{bmatrix}.
$$

Next for all $j \le d+1$, let

$$
w_{qj} = \begin{bmatrix} \bar{w}_{qj} \\ w_{qj} \end{bmatrix}, \quad w_{kj} = \begin{bmatrix} \bar{w}_{kj} \\ w_{kj} \end{bmatrix}, \quad w_{vj} = \begin{bmatrix} \bar{w}_{vj} \\ w_{vj} \end{bmatrix}
$$

where $\bar{w}_{qj}, \bar{w}_{kj}, \bar{w}_{vj} \in \mathbb{R}^d$ and $w_{qj}, w_{kj}, w_{vj} \in \mathbb{R}$. Then we have

$$
w_{qj}^\top z w_{kj}^\top z_i w_{vj}^\top z_i = \left(\bar{w}_{qj}^\top x \right) \left(\bar{w}_{kj}^\top x_i + w_{kj} y_i \right) \left(\bar{w}_{vj}^\top x_i + w_{vj} y_i \right)
$$

$$
= x^\top \bar{w}_{qj} \left(w_{vj} \bar{w}_{kj}^\top + w_{kj} \bar{w}_{vj}^\top \right) x_i y_i + \left(\bar{w}_{qj}^\top x \right) \left(\bar{w}_{kj}^\top x_i \right) \left(\bar{w}_{vj}^\top x_i \right) + \left(w_{kj} w_{vj} \bar{w}_{qj}^\top x y_i^2 \right)
$$

$$
= x^\top W_j' x_i y_i + \delta_j(x, x_i, x_i) + w_j'^\top x y_i^2
$$

where

$$
W_j' := \bar{w}_{qj} \left(w_{vj} \bar{w}_{kj}^\top + w_{kj} \bar{w}_{vj}^\top \right) \in \mathbb{R}^{d \times d},
$$

$$
w_j' := w_{kj} w_{vj} \bar{w}_{qj} \in \mathbb{R}^d,
$$

$$
\delta_j(x, x_i, x_i) := \left(\bar{w}_{qj}^\top x \right) \left(\bar{w}_{kj}^\top x_i \right) \left(\bar{w}_{vj}^\top x_i \right) \in \mathbb{R}.
$$

Then

$$
g_{\text{SSM}}(Z) = \sum_{i=1}^{n} f_{n+1-i} \cdot \sum_{j=1}^{d+1} v_j \left(x^\top W_j' x_i y_i + \delta_j(x, x_i, x_i) + w_j'^\top x y_i^2 \right)
$$

$$
= x^\top \left(\sum_{j=1}^{d+1} v_j W_j' \right) X(y \odot \tilde{f}) + \sum_{i=1}^{n} f_{n+1-i} \cdot \sum_{j=1}^{d+1} v_j \cdot \delta_j(x, x_i, x_i) + \left(\sum_{j=1}^{d+1} v_j w_j'^\top \right) x y^\top (y \odot \tilde{f})
$$

$$
= \underbrace{x^\top \tilde{W} X \tilde{y}}_{\tilde{g}_{\text{SSM}}(Z)} + \underbrace{\tilde{\delta}(x, X, X)}_{\varepsilon_1} + \underbrace{\tilde{w}^\top x y^\top \tilde{y}}_{\varepsilon_2}.
$$

14

where

$$\tilde{\boldsymbol{f}} := [f_n \ \cdots \ f_1]^\top \in \mathbb{R}^n,$$

$$\tilde{\boldsymbol{y}} := \boldsymbol{y} \odot \tilde{\boldsymbol{f}} \in \mathbb{R}^n,$$

$$\tilde{\boldsymbol{W}} := \sum_{j=1}^{d+1} v_j \boldsymbol{W}'_j \in \mathbb{R}^{d \times d},$$

$$\tilde{\boldsymbol{w}} := \sum_{j=1}^{d+1} v_j \boldsymbol{w}'_j \in \mathbb{R}^d,$$

$$\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) := \sum_{i=1}^{n} f_{n+1-i} \cdot \sum_{j=1}^{d+1} v_j \cdot \delta_j(\boldsymbol{x}, \boldsymbol{x}_i, \boldsymbol{x}_i) \in \mathbb{R}.$$

Next we will show that for any $\boldsymbol{W}_k, \boldsymbol{W}_q, \boldsymbol{W}_v, \boldsymbol{v}$,

$$\mathbb{E}\left[(g_{\mathsf{SSM}}(\boldsymbol{Z}) - y)^2\right] \geq \mathbb{E}\left[(\tilde{g}_{\mathsf{SSM}}(\boldsymbol{Z}) - y)^2\right].$$

To start with, we obtain

$$\mathbb{E}\left[(g_{\mathsf{SSM}}(\boldsymbol{Z}) - y)^2\right] = \mathbb{E}\left[(\tilde{g}_{\mathsf{SSM}}(\boldsymbol{Z}) + \varepsilon_1 + \varepsilon_2 - y)^2\right]$$
$$= \mathbb{E}\left[(\tilde{g}_{\mathsf{SSM}}(\boldsymbol{Z}) - y)^2\right] + \mathbb{E}\left[(\varepsilon_1 + \varepsilon_2)^2\right] + 2\,\mathbb{E}\left[(\tilde{g}_{\mathsf{SSM}}(\boldsymbol{Z}) - y)(\varepsilon_1 + \varepsilon_2)\right] \quad (23)$$

where there is decomposition

$$(\tilde{g}_{\mathsf{SSM}}(\boldsymbol{Z}) - y)(\varepsilon_1 + \varepsilon_2) = \underbrace{\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}} \boldsymbol{X} \tilde{\boldsymbol{y}}}_{(a)} - \underbrace{\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) y}_{(b)} + \underbrace{\tilde{\boldsymbol{w}}^\top \boldsymbol{x} \boldsymbol{y}^\top \tilde{\boldsymbol{y}} \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}} \boldsymbol{X} \tilde{\boldsymbol{y}}}_{(c)} - \underbrace{y \cdot \tilde{\boldsymbol{w}}^\top \boldsymbol{x} \boldsymbol{y}^\top \tilde{\boldsymbol{y}}}_{(d)}.$$

In the following, similar to the proof of Lemma 9, we consider the expectations of $(a), (b), (c), (d)$ sequentially, which return zeros under Assumptions 1 and 2. Note that $\delta_j(\boldsymbol{x}, \boldsymbol{x}_i, \boldsymbol{x}_i)$'s and $\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X})$ are summation of monomials of entries of $(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{\beta})$ with order 3, and entries of $\boldsymbol{y}$ and y are summation of monomials of entries of $(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{\beta})$ with even orders: e.g., $y = \boldsymbol{x}^\top \boldsymbol{\beta} + \xi$ where ξ is of oder 0 and $\boldsymbol{x}^\top \boldsymbol{\beta}$ is of order 2.

$(a):$ $\mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}} \boldsymbol{X} \tilde{\boldsymbol{y}}\right]$

$$= \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}} \boldsymbol{X}(\boldsymbol{X}\boldsymbol{\beta} \odot \tilde{\boldsymbol{f}})\right] + \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}} \boldsymbol{X}(\boldsymbol{\xi} \odot \tilde{\boldsymbol{f}})\right]$$
$$= \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}} \boldsymbol{X}\right] \mathbb{E}\left[\boldsymbol{\xi} \odot \tilde{\boldsymbol{f}}\right]$$
$$= 0.$$

$(b):$ $\mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) y\right]$

$$= \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X})(\boldsymbol{x}^\top \boldsymbol{\beta} + \xi)\right]$$
$$= \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X})\boldsymbol{x}^\top \boldsymbol{\beta}\right] + \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X})\xi\right]$$
$$= 0.$$

$(c):\quad \mathbb{E}\left[\tilde{\boldsymbol{w}}^\top \boldsymbol{x}\boldsymbol{y}^\top \tilde{\boldsymbol{y}} \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}}\boldsymbol{X}\tilde{\boldsymbol{y}}\right]$

$\qquad = \mathbb{E}\left[\tilde{\boldsymbol{w}}^\top \boldsymbol{x}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi})^\top(\boldsymbol{X}\boldsymbol{\beta}\odot\tilde{\boldsymbol{f}} + \boldsymbol{\xi}\odot\tilde{\boldsymbol{f}}) \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}}\boldsymbol{X}(\boldsymbol{X}\boldsymbol{\beta}\odot\tilde{\boldsymbol{f}} + \boldsymbol{\xi}\odot\tilde{\boldsymbol{f}})\right]$

$\qquad = 0.$

$(d):\quad \mathbb{E}\left[y \cdot \tilde{\boldsymbol{w}}^\top \boldsymbol{x}\boldsymbol{y}^\top \tilde{\boldsymbol{y}}\right]$

$\qquad = \mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{\beta} + \xi) \cdot \tilde{\boldsymbol{w}}^\top \boldsymbol{x}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi})^\top(\boldsymbol{X}\boldsymbol{\beta}\odot\tilde{\boldsymbol{f}} + \boldsymbol{\xi}\odot\tilde{\boldsymbol{f}})\right]$

$\qquad = 0.$

Combining the results with (23) results that

$$\mathbb{E}\left[(g_{\text{SSM}}(\boldsymbol{Z}) - y)^2\right] - \mathbb{E}\left[(\tilde{g}_{\text{SSM}}(\boldsymbol{Z}) - y)^2\right] = \mathbb{E}\left[(\varepsilon_1 + \varepsilon_2)^2\right] \geq 0.$$

Therefore we obtain,

$$\min_{\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v, \boldsymbol{v}, \boldsymbol{f}} \mathbb{E}\left[(g_{\text{SSM}}(\boldsymbol{Z}) - y)^2\right] \geq \min_{\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{f}}} \mathbb{E}\left[(\tilde{g}_{\text{SSM}}(\boldsymbol{Z}) - y)^2\right] = \min_{\boldsymbol{W}, \boldsymbol{\omega}} \mathbb{E}\left[(g_{\text{WPGD}}(\boldsymbol{Z}) - y)^2\right].$$

Next we show that for any choices of $\boldsymbol{W}$ and $\boldsymbol{\omega}$ in g_{WPGD}, there are $\boldsymbol{W}_{q,k,v}, \boldsymbol{v}, \boldsymbol{f}$ such that $g_{\text{SSM}} \equiv g_{\text{WPGD}}$. To this end, given $\boldsymbol{\omega} = [\omega_1 \ \ldots \ \omega_n]^\top$, let

$$\boldsymbol{W}_q = \boldsymbol{I}_{d+1}, \quad \boldsymbol{W}_k = \begin{bmatrix} \boldsymbol{W}^\top & \boldsymbol{0}_d \\ \boldsymbol{0}_d^\top & 0 \end{bmatrix}, \quad \boldsymbol{W}_v = \begin{bmatrix} \boldsymbol{0}_{d\times d} & \boldsymbol{0}_d \\ \boldsymbol{1}_d^\top & 0 \end{bmatrix}, \quad \boldsymbol{v} = \begin{bmatrix} \boldsymbol{1}_d \\ 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{f} = \begin{bmatrix} 0 \\ \omega_n \\ \ldots \\ \omega_1 \end{bmatrix}.$$

Then we get

$$\begin{aligned}((\boldsymbol{Z}_0\boldsymbol{W}_k \odot \boldsymbol{Z}_0\boldsymbol{W}_v) * \boldsymbol{f})_{n+1} &= \left(\left(\begin{bmatrix} \boldsymbol{X}\boldsymbol{W}^\top & \boldsymbol{0}_n \\ \boldsymbol{0}_d & 0 \end{bmatrix} \odot \begin{bmatrix} \boldsymbol{y}\boldsymbol{1}_d^\top & \boldsymbol{0}_n \\ \boldsymbol{0}_d & 0 \end{bmatrix}\right) * \boldsymbol{f}\right)_{n+1} \\ &= \begin{bmatrix} \sum_{i=1}^n \omega_i \cdot y_i \boldsymbol{W}\boldsymbol{x}_i \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{W}\boldsymbol{X}^\top(\boldsymbol{y}\odot\boldsymbol{\omega}) \\ 0 \end{bmatrix},\end{aligned}$$

and therefore

$$g_{\text{SSM}}(\boldsymbol{Z}) = \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top(\boldsymbol{y}\odot\boldsymbol{\omega}) = g_{\text{WPGD}}(\boldsymbol{Z}),$$

which completes the proof of (21).

Next, to show (22), for any $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, let $\mathcal{L}(\boldsymbol{\omega}) = \mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top(\boldsymbol{y}\odot\boldsymbol{\omega}) - y)^2\right]$. Then we have

$$\begin{aligned}\frac{\partial\mathcal{L}(\boldsymbol{\omega})}{\partial\omega_i} &= \mathbb{E}\left[2\left(\boldsymbol{x}^\top \boldsymbol{W}\sum_{j=1}^n \omega_j y_j \boldsymbol{x}_j - y\right)\left(\boldsymbol{x}^\top \boldsymbol{W}y_i\boldsymbol{x}_i\right)\right] \\ &= 2\sum_{j=1}^n \omega_j \mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{W}y_j\boldsymbol{x}_j)(\boldsymbol{x}^\top \boldsymbol{W}y_i\boldsymbol{x}_i)\right] - 2\mathbb{E}\left[y\boldsymbol{x}^\top \boldsymbol{W}y_i\boldsymbol{x}_i\right].\end{aligned}$$

Here since $(\boldsymbol{x}_i, y_i)_{i=1}^n$ follow the same distribution and are conditionally independent given $\boldsymbol{x}$ and $\boldsymbol{\beta}$, for any $i \neq j \neq j'$, $\mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{W} y_i \boldsymbol{x}_i)^2\right] = \mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{W} y_j \boldsymbol{x}_j)^2\right]$ and $\mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{W} y_j \boldsymbol{x}_j)(\boldsymbol{x}^\top \boldsymbol{W} y_i \boldsymbol{x}_i)\right] = \mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{W} y_{j'} \boldsymbol{x}_{j'})(\boldsymbol{x}^\top \boldsymbol{W} y_i \boldsymbol{x}_i)\right]$. Then let

$$\mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{W} y_j \boldsymbol{x}_j)(\boldsymbol{x}^\top \boldsymbol{W} y_i \boldsymbol{x}_i)\right] = \begin{cases} c_1, & i \neq j \\ c_2, & i = j \end{cases} \quad \text{and} \quad \mathbb{E}\left[y \boldsymbol{x}^\top \boldsymbol{W} y_i \boldsymbol{x}_i\right] = c_3,$$

where $(c_1, c_2, c_3) := (c_1(\boldsymbol{W}), c_2(\boldsymbol{W}), c_3(\boldsymbol{W}))$. We get

$$\frac{\partial \mathcal{L}(\boldsymbol{\omega})}{\partial \omega_i} = 2c_1 \boldsymbol{\omega}^\top \mathbf{1}_n + 2(c_2 - c_1)\omega_i - 2c_3.$$

If $c_2 - c_1 = 0$, then $\frac{\partial \mathcal{L}(\boldsymbol{\omega})}{\partial \omega_i} \equiv 2c_1 \boldsymbol{\omega}^\top \mathbf{1}_n - 2c_3$ for all $i \leq n$ and any $\boldsymbol{\omega} \in \mathbb{R}^n$ achieves the same performance.

If $c_2 - c_1 \neq 0$, setting $\frac{\partial \mathcal{L}(\boldsymbol{\omega})}{\partial \omega_i} = 0$ returns

$$\omega_i = \frac{c_3 - c_1 \sum_{j=1}^n \omega_j}{c_2 - c_1} := C \quad \text{for all } i \leq n.$$

Therefore the optimal loss is achieved via setting $\boldsymbol{\omega} = C\mathbf{1}_n$. Without loss of generality, we can update $\boldsymbol{W} \to C\boldsymbol{W}$. Then $\boldsymbol{\omega} = \mathbf{1}_n$, and we obtain

$$\min_{\boldsymbol{W}, \boldsymbol{\omega}} \mathbb{E}\left[\left(\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{X}^\top (\boldsymbol{y} \odot \boldsymbol{\omega}) - y\right)^2\right] = \min_{\boldsymbol{W}} \mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{y} - y)^2\right]$$

which completes the proof of (22). ∎

A.2. Equivalence between $g_{\mathrm{ATT}}^\star$ and $g_{\mathrm{PGD}}^\star$

While Proposition 3 demonstrates the equivalence of optimal losses, we also study the uniqueness and equivalence of optimal prediction functions. To this end, we analyze the strong convexity of $\mathcal{L}_{\mathrm{PGD}}(\boldsymbol{W})$ and derive the subsequent lemmas.

Lemma 11 *Suppose Assumption 2 holds and let $\boldsymbol{\xi} = [\xi_1 \; \xi_2 \; \cdots \; \xi_n]^\top$. Then the loss $\mathcal{L}_{\mathrm{PGD}}(\boldsymbol{W})$ in (4a) is strongly-convex if and only if $\mathbb{E}[(\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta})^2] + \mathbb{E}[(\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{\xi})^2]$ is strongly-convex. Additionally, let $g_{\mathrm{PGD}}^\star$, $g_{\mathrm{ATT}}^\star$ be the optimal prediction functions of (4a) and (5a). Then under the conditions of Assumptions 1 and 2, and the strong convexity, $g_{\mathrm{PGD}}^\star = g_{\mathrm{ATT}}^\star$.*

Lemma 12 *Suppose that the label noise $(\xi_i)_{i=1}^n$ are i.i.d., zero-mean, variance σ^2 and independent of everything else, and that there is a decomposition $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$, $\boldsymbol{X} = \boldsymbol{X}_1 + \boldsymbol{X}_2$, and $\boldsymbol{\beta} = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2$ such that either of the following holds*

- *$\sigma > 0$, and $(\boldsymbol{x}_1, \boldsymbol{X}_1)$ have full rank covariance and are independent of each other and $(\boldsymbol{x}_2, \boldsymbol{X}_2)$.*

- *$(\boldsymbol{x}_1, \boldsymbol{\beta}_1, \boldsymbol{X}_1)$ have full rank covariance and are independent of each other and $(\boldsymbol{x}_2, \boldsymbol{\beta}_2, \boldsymbol{X}_2)$.*

Then, the loss $\mathcal{L}_{\mathrm{PGD}}(\boldsymbol{W})$ in (4a) is strongly-convex.

As mentioned above, in this work, we study three specific linear models: with general independent, RAG-related, and task-feature alignment data. Note that for all the three cases, according to Proposition 3, we have $\mathcal{L}_{\mathrm{PGD}}^\star = \mathcal{L}_{\mathrm{ATT}}^\star = \mathcal{L}_{\mathrm{SSM}}^\star$. Additionally, the second claim in Lemma 12 holds, and $\mathcal{L}_{\mathrm{PGD}}(\boldsymbol{W})$ is strongly convex. Therefore, following Lemma 11, we have $g_{\mathrm{PGD}}^\star = g_{\mathrm{ATT}}^\star$.

A.3. Proof of Lemma 11

Proof Recap the loss $\mathcal{L}_{\text{PGD}}(\mathcal{W})$ in (15a) and prediction $g_{\text{PGD}}(\mathbf{Z})$ in (14a), we have

$$
\begin{aligned}
\mathcal{L}_{\text{PGD}}(\mathcal{W}) &= \mathbb{E}[(y - g_{\text{PGD}}(\mathbf{Z}))^2] \\
&= \mathbb{E}\left[\left(\mathbf{x}^\top \boldsymbol{\beta} + \xi - \mathbf{x}^\top \mathbf{W} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi})\right)^2\right] \\
&= \mathbb{E}\left[(\mathbf{x}^\top \boldsymbol{\beta} - \mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta})^2 + 2(\mathbf{x}^\top \boldsymbol{\beta} - \mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta})(\xi - \mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \boldsymbol{\xi}) + (\xi - \mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \boldsymbol{\xi})^2\right] \\
&= \mathbb{E}\left[(\mathbf{x}^\top \boldsymbol{\beta} - \mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta})^2 + (\xi - \mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \boldsymbol{\xi})^2\right] + 2\,\mathbb{E}[(\mathbf{x}^\top \boldsymbol{\beta} - \mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta})(\xi - \mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \boldsymbol{\xi})] \\
&= \mathbb{E}\left[(\mathbf{x}^\top \boldsymbol{\beta} - \mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta})^2 + (\xi - \mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \boldsymbol{\xi})^2\right] \qquad\qquad (24) \\
&= \underbrace{\mathbb{E}\left[(\mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta})^2 + (\mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \boldsymbol{\xi})^2\right]}_{f_1(\mathbf{W})} \underbrace{- 2\,\mathbb{E}[\boldsymbol{\beta}^\top \mathbf{x}\mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + \xi \mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \boldsymbol{\xi}]}_{f_2(\mathbf{W})} + \underbrace{\mathbb{E}[(\mathbf{x}^\top \boldsymbol{\beta})^2 + \xi^2]}_{\text{constant}}
\end{aligned}
$$

where (24) follows Assumption 2. Since $f_2(\mathbf{W})$ is convex, $\mathcal{L}_{\text{PGD}}(\mathcal{W})$ is strongly-convex if and only if $f_1(\mathbf{W})$ is strongly-convex, which completes the proof of strong convexity.

Next, (19) and (20) in the proof of Lemma 9 demonstrate that the optimal loss is achievable and is achieved at $\varepsilon = 0$. Subsequently, (16) indicates that $g_{\text{ATT}}^\star$ has the same form as $g_{\text{PGD}}^\star$. Under the strong convexity assumption, $g_{\text{PGD}}^\star$ is unique, which leads to the conclusion that $g_{\text{PGD}}^\star = g_{\text{ATT}}^\star$. ∎

A.4. Proof of Lemma 12

Proof According to Lemma 11, $\mathcal{L}_{\text{PGD}}(\mathcal{W})$ is strongly-convex as long as either $\mathbb{E}[(\mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta})^2]$ or $\mathbb{E}[(\mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \boldsymbol{\xi})^2]$ is strongly-convex. Therefore, in this lemma, the two claims correspond to the strong convexity of $\mathbb{E}[(\mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \boldsymbol{\xi})^2]$ and $\mathbb{E}[(\mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta})^2]$ terms, respectively.

Suppose the decomposition claim holds. Without losing generality, we may assume $(\mathbf{x}_1, \boldsymbol{\beta}_1, \mathbf{X}_1)$ are zero-mean because we can allocate the mean component to $(\mathbf{x}_2, \boldsymbol{\beta}_2, \mathbf{X}_2)$ without changing the covariance.

• **Claim 1:** Let $\bar{\boldsymbol{\Sigma}}_x = \mathbb{E}[\mathbf{x}_1 \mathbf{x}_1^\top]$, $\bar{\boldsymbol{\Sigma}}_\beta = \mathbb{E}[\boldsymbol{\beta}_1 \boldsymbol{\beta}_1^\top]$, and $\bar{\boldsymbol{\Sigma}}_X = \mathbb{E}[\mathbf{X}_1^\top \mathbf{X}_1]$. If the first claim holds, using independence, observe that we can write

$$
\mathbb{E}[(\mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \boldsymbol{\xi})^2] = \mathbb{E}[(\mathbf{x}_1^\top \mathbf{W} \mathbf{X}_1^\top \boldsymbol{\xi})^2] + \mathbb{E}[(\mathbf{x}_1^\top \mathbf{W} \mathbf{X}_2^\top \boldsymbol{\xi})^2] + \mathbb{E}[(\mathbf{x}_2^\top \mathbf{W} \mathbf{X}_1^\top \boldsymbol{\xi})^2] + \mathbb{E}[(\mathbf{x}_2^\top \mathbf{W} \mathbf{X}_2^\top \boldsymbol{\xi})^2],
$$

where the last three terms of the right hand side are convex and the first term obeys

$$
\begin{aligned}
\mathbb{E}[(\mathbf{x}_1^\top \mathbf{W} \mathbf{X}_1^\top \boldsymbol{\xi})^2] &= \sigma^2 \,\mathbb{E}[\mathbf{x}_1^\top \mathbf{W} \mathbf{X}_1^\top \mathbf{X}_1 \mathbf{W}^\top \mathbf{x}_1] \\
&= \sigma^2 \text{tr}\left(\mathbb{E}[\mathbf{x}_1 \mathbf{x}_1^\top \mathbf{W} \mathbf{X}_1^\top \mathbf{X}_1 \mathbf{W}^\top]\right) \\
&= \sigma^2 \text{tr}\left(\bar{\boldsymbol{\Sigma}}_x \mathbf{W} \bar{\boldsymbol{\Sigma}}_X \mathbf{W}^\top\right) \\
&= \sigma^2 \left\| \sqrt{\bar{\boldsymbol{\Sigma}}_x} \mathbf{W} \sqrt{\bar{\boldsymbol{\Sigma}}_X} \right\|_F^2.
\end{aligned}
$$

Since noise level $\sigma > 0$, using the full-rankness of covariance matrices $\bar{\boldsymbol{\Sigma}}_x$ and $\bar{\boldsymbol{\Sigma}}_X$, we conclude with strong convexity of $\mathbb{E}[(\mathbf{x}^\top \mathbf{W} \mathbf{X}^\top \boldsymbol{\xi})^2]$.

• **Claim 2:** Now recall that $\bar{\Sigma}_X = \mathbb{E}[X_1^\top X_1]$ and set $A = X_1^\top X_1 - \bar{\Sigma}_X$ and $B = X_2^\top X_2 + \bar{\Sigma}_X$. Observe that $\mathbb{E}[A] = 0$. If the second claim holds, $\mathbb{E}[X^\top X] = \mathbb{E}[A + B]$. Note that (A, β_1, x_1) are independent of each other and (B, β_2, x_2). Using independence and $\mathbb{E}[A] = 0$, similarly write

$$\mathbb{E}[(x^\top W X^\top X \beta)^2] = \mathbb{E}[(x^\top W A \beta)^2] + \mathbb{E}[(x^\top W B \beta)^2].$$

Now using $\mathbb{E}[\beta_1] = \mathbb{E}[x_1] = 0$ and their independence from rest, these terms obeys

$$\mathbb{E}[(x^\top W A \beta)^2] = \mathbb{E}[(x_1^\top W A \beta_1)^2] + \mathbb{E}[(x_1^\top W A \beta_2)^2] + \mathbb{E}[(x_2^\top W A \beta_1)^2] + \mathbb{E}[(x_2^\top W A \beta_2)^2]$$
$$\mathbb{E}[(x^\top W B \beta)^2] = \mathbb{E}[(x_1^\top W B \beta_1)^2] + \mathbb{E}[(x_1^\top W B \beta_2)^2] + \mathbb{E}[(x_2^\top W B \beta_1)^2] + \mathbb{E}[(x_2^\top W B \beta_2)^2].$$

In both equations, the last three terms of the right hand side are convex. To proceed, we focus on the first terms. Using independence and setting $\Sigma_X = \mathbb{E}[X^\top X] \geq \bar{\Sigma}_X > 0$, we note that

$$\mathbb{E}[(x_1^\top W A \beta_1)^2] + \mathbb{E}[(x_1^\top W B \beta_1)^2] = \mathbb{E}[(x_1^\top W X^\top X \beta_1)^2]$$

where x_1, β_1, X are independent and full-rank covariance. To proceed, note that

$$\mathbb{E}[(x_1^\top W X^\top X \beta_1)^2] = \mathbb{E}[(x_1^\top W \Sigma_X \beta_1)^2] + \mathbb{E}[(x_1^\top W (X^\top X - \Sigma_X) \beta_1)^2].$$

Observing the convexity of the right hand side and focusing on the first term, we get

$$\mathbb{E}[(x_1^\top W \Sigma_X \beta_1)^2] = \mathrm{tr}\left(\bar{\Sigma}_x W \Sigma_X \bar{\Sigma}_\beta \Sigma_X W^\top\right) = \left\| \sqrt{\bar{\Sigma}_x} W \Sigma_X \sqrt{\bar{\Sigma}_\beta} \right\|_F^2.$$

Using the fact that covariance matrices, $\bar{\Sigma}_x, \Sigma_X, \bar{\Sigma}_\beta$, are full rank concludes the strong convexity proof of $\mathbb{E}[(x^\top W X^\top X \beta)^2]$. ∎

Appendix B. Analysis of General Data Distribution

In this section, we provide the proofs in Section 3, which focuses on solving Objective (4a). For the sake of clean notation, let $\mathcal{L}(W) := \mathcal{L}_{\mathrm{PGD}}(W)$ and $g := g_{\mathrm{PGD}}$ in this section.

B.1. Supporting Results

We begin by deriving the even moments of random variables.
• **$2n$'th moment of a normally distributed variable:** Let $u \sim \mathcal{N}(0, \sigma^2)$. Then we have

$$\mathbb{E}[u^{2n}] = \sigma^{2n}(2n - 1)!!. \tag{25}$$

- **4'th moment:** Let $\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Then for any $\boldsymbol{W}, \boldsymbol{W}' \in \mathbb{R}^{d \times d}$, we have

$$
\mathbb{E}\left[(\boldsymbol{u}^\top \boldsymbol{W} \boldsymbol{u})(\boldsymbol{u}^\top \boldsymbol{W}' \boldsymbol{u})\right]
$$

$$
= \mathbb{E}\left[\left(\sum_{i,j=1}^d W_{ij} u_i u_j\right)\left(\sum_{i,j=1}^d W'_{ij} u_i u_j\right)\right]
$$

$$
= \mathbb{E}\left[\left(\sum_{i=1}^d W_{ii} u_i^2\right)\left(\sum_{i=1}^d W'_{ii} u_i^2\right)\right] + \mathbb{E}\left[\left(\sum_{i \neq j} W_{ij} u_i u_j\right)\left(\sum_{i \neq j} W'_{ij} u_i u_j\right)\right]
$$

$$
= \sum_{i=1}^d W_{ii} W'_{ii} \mathbb{E}\left[u_i^4\right] + \sum_{i \neq j} W_{ii} W'_{jj} \mathbb{E}[u_i^2]\mathbb{E}[u_j^2] + \sum_{i \neq j} W_{ij} W'_{ij} \mathbb{E}[u_i^2]\mathbb{E}[u_j^2] + \sum_{i \neq j} W_{ij} W'_{ji} \mathbb{E}[u_i^2]\mathbb{E}[u_j^2]
$$

$$
= 3\sum_{i=1}^d W_{ii} W'_{ii} + \sum_{i \neq j} W_{ii} W'_{jj} + \sum_{i \neq j} W_{ij} W'_{ij} + \sum_{i \neq j} W_{ij} W'_{ji}
$$

$$
= \sum_{i,j=1}^d W_{ii} W'_{jj} + \sum_{i,j=1}^d W_{ij} W'_{ij} + \sum_{i,j=1}^d W_{ij} W'_{ji}
$$

$$
= \mathtt{tr}\,(\boldsymbol{W})\,\mathtt{tr}\,(\boldsymbol{W}') + \mathtt{tr}\left(\boldsymbol{W}' \boldsymbol{W}^\top\right) + \mathtt{tr}\,(\boldsymbol{W} \boldsymbol{W}'). \tag{26}
$$

- **4'th cross-moment:** Let $\boldsymbol{u}, \boldsymbol{v} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and for any $\boldsymbol{W} \in \mathbb{R}^{d \times d}$, let $\boldsymbol{\Lambda}_{\boldsymbol{W}} = \boldsymbol{W} \odot \boldsymbol{I}_d$. Then we have

$$
\mathbb{E}\left[(\boldsymbol{u}^\top \boldsymbol{W} \boldsymbol{v} \boldsymbol{v}^\top \boldsymbol{u})^2\right]
$$

$$
= \mathbb{E}\left[\left(\sum_{i,j=1}^d W_{ij} u_i v_j\right)^2 \left(\sum_{i=1}^d u_i v_i\right)^2\right]
$$

$$
= \mathbb{E}\left[\left(\sum_{i,j=1}^d W_{ij}^2 u_i^2 v_j^2 + \sum_{i \neq i'} W_{ij} W_{i'j} u_i u_{i'} v_j^2 + \sum_{j \neq j'} W_{ij} W_{ij'} u_i^2 v_j v_{j'} + \sum_{i' \neq i, j' \neq j} W_{ij} W_{i'j'} u_i u_{i'} v_j v_{j'}\right)\left(\sum_{i=1}^d u_i^2 v_i^2 + \sum_{i \neq j} u_i u_j v_i v_j\right)\right]
$$

$$
= \mathbb{E}\left[\left(\sum_{i,j=1}^d W_{ij}^2 u_i^2 v_j^2\right)\left(\sum_{i=1}^d u_i^2 v_i^2\right) + \left(\sum_{i \neq j} W_{ij} W_{ji} u_i^2 u_j^2 v_i^2 v_j^2\right)\right]
$$

$$
= \mathbb{E}\left[\left(\sum_{i=1}^d W_{ii}^2 u_i^2 v_i^2 + \sum_{i \neq j} W_{ij}^2 u_i^2 v_j^2\right)\left(\sum_{i=1}^d u_i^2 v_i^2\right)\right] + \sum_{i \neq j} W_{ij} W_{ji}
$$

$$
= \mathbb{E}\left[\left(\sum_{i=1}^d W_{ii}^2 u_i^4 v_i^4 + \sum_{i \neq j} W_{ii}^2 u_i^2 v_i^2 u_j^2 v_j^2\right)\right] + \mathbb{E}\left[\left(\sum_{i \neq j} W_{ij}^2 u_i^4 v_j^2 v_i^2 + \sum_{i \neq j} W_{ij}^2 u_i^2 v_j^4 u_j^2 + \sum_{i \neq j \neq k} W_{ij}^2 u_i^2 v_j^2 u_k^2 v_k^2\right)\right] + \sum_{i \neq j} W_{ij} W_{ji}
$$

$$
= 9\sum_{i=1}^d W_{ii}^2 + (d-1)\sum_{i=1}^d W_{ii}^2 + 6\sum_{i \neq j} W_{ij}^2 + (d-2)\sum_{i \neq j} W_{ij}^2 + \sum_{i \neq j} W_{ij} W_{ji}
$$

$$
= 3\sum_{i=1}^d W_{ii}^2 + (d+4)\sum_{i,j=1}^d W_{ij}^2 + \sum_{i,j=1}^d W_{ij} W_{ji}
$$

$$
= 3\mathtt{tr}\left(\boldsymbol{\Lambda}_{\boldsymbol{W}}^2\right) + (d+4)\mathtt{tr}\left(\boldsymbol{W} \boldsymbol{W}^\top\right) + \mathtt{tr}\left(\boldsymbol{W}^2\right). \tag{27}
$$

- **6'th moment:** Let $u \sim \mathcal{N}(0, I_d)$. Then for any $W, W' \in \mathbb{R}^{d \times d}$, we have

$$\mathbb{E}\left[(u^\top W u)(u^\top W' u) \|u\|_{\ell_2}^2\right]$$

$$= \mathbb{E}\left[\left(\sum_{i,j=1}^d W_{ij} u_i u_j\right)\left(\sum_{i,j=1}^d W'_{ij} u_i u_j\right)\left(\sum_{i=1}^d u_i^2\right)\right]$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^d W_{ii} u_i^2\right)\left(\sum_{i=1}^d W'_{ii} u_i^2\right)\left(\sum_{i=1}^d u_i^2\right)\right] + \mathbb{E}\left[\left(\sum_{i \neq j} W_{ij} u_i u_j\right)\left(\sum_{i \neq j} W'_{ij} u_i u_j\right)\left(\sum_{i=1}^d u_i^2\right)\right]$$

$$= \sum_{i=1}^d W_{ii} W'_{ii} \mathbb{E}\left[u_i^4\left(\sum_{i'=1}^d u_{i'}^2\right)\right] + \sum_{i \neq j} W_{ii} W'_{jj} \mathbb{E}\left[u_i^2 u_j^2\left(\sum_{i'=1}^d u_{i'}^2\right)\right]$$

$$+ \sum_{i \neq j} W_{ij} W'_{ij} \mathbb{E}\left[u_i^2 u_j^2\left(\sum_{i'=1}^d u_{i'}^2\right)\right] + \sum_{i \neq j} W_{ij} W'_{ji} \mathbb{E}\left[u_i^2 u_j^2\left(\sum_{i'=1}^d u_{i'}^2\right)\right]$$

$$= (d+4)\left(3\sum_{i=1}^d W_{ii} W'_{ii} + \sum_{i \neq j} W_{ii} W'_{jj} + \sum_{i \neq j} W_{ij} W'_{ij} + \sum_{i \neq j} W_{ij} W'_{ji}\right) \tag{28}$$

$$= (d+4)\left(\sum_{i,j=1}^d W_{ii} W'_{jj} + \sum_{i,j=1}^d W_{ij} W'_{ij} + \sum_{i,j=1}^d W_{ij} W'_{ji}\right)$$

$$= (d+4)\left(\mathrm{tr}\,(W)\,\mathrm{tr}\,(W') + \mathrm{tr}\,\left(W' W^\top\right) + \mathrm{tr}\,(W W')\right), \tag{29}$$

where (28) is obtained by following

$$\mathbb{E}\left[u_i^4\left(\sum_{i'=1}^d u_{i'}^2\right)\right] = \mathbb{E}[u^6] + (d-1)\,\mathbb{E}[u^4]\,\mathbb{E}[u^2] = 3(d+4),$$

$$\mathbb{E}\left[u_i^2 u_j^2\left(\sum_{i'=1}^d u_{i'}^2\right)\right] = 2\,\mathbb{E}[u^4]\,\mathbb{E}[u^2] + (d-2)\,\mathbb{E}[u^2]\,\mathbb{E}[u^2]\,\mathbb{E}[u^2] = d+4.$$

- **8'th moment:** Let $u \sim \mathcal{N}(0, I_d)$. Then for any $W, W' \in \mathbb{R}^{d \times d}$, we have

$$\mathbb{E}\left[(u^\top W u)(u^\top W' u) \|u\|_{\ell_2}^4\right]$$

$$= \mathbb{E}\left[\left(\sum_{i,j=1}^d W_{ij} u_i u_j\right)\left(\sum_{i,j=1}^d W'_{ij} u_i u_j\right)\left(\sum_{i,j=1}^d u_i^2 u_j^2\right)\right]$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^d W_{ii} u_i^2\right)\left(\sum_{i=1}^d W'_{ii} u_i^2\right)\left(\sum_{i=1}^d u_i^4 + \sum_{i \neq j} u_i^2 u_j^2\right)\right] + \mathbb{E}\left[\left(\sum_{i \neq j} W_{ij} u_i u_j\right)\left(\sum_{i \neq j} W'_{ij} u_i u_j\right)\left(\sum_{i=1}^d u_i^4 + \sum_{i \neq j} u_i^2 u_j^2\right)\right]$$

$$= \sum_{i=1}^d W_{ii} W'_{ii} \mathbb{E}\left[u_i^4\left(\sum_{i'=1}^d u_{i'}^4 + \sum_{i' \neq j'} u_{i'}^2 u_{j'}^2\right)\right] + \sum_{i \neq j} W_{ii} W'_{jj} \mathbb{E}\left[u_i^2 u_j^2\left(\sum_{i'=1}^d u_{i'}^4 + \sum_{i' \neq j'} u_{i'}^2 u_{j'}^2\right)\right]$$

$$+ \sum_{i \neq j} W_{ij} W'_{ij} \mathbb{E}\left[u_i^2 u_j^2\left(\sum_{i'=1}^d u_{i'}^4 + \sum_{i' \neq j'} u_{i'}^2 u_{j'}^2\right)\right] + \sum_{i \neq j} W_{ij} W'_{ji} \mathbb{E}\left[u_i^2 u_j^2\left(\sum_{i'=1}^d u_{i'}^4 + \sum_{i' \neq j'} u_{i'}^2 u_{j'}^2\right)\right]$$

$$= (d+4)(d+6)\left(3\sum_{i=1}^d W_{ii} W'_{ii} + \sum_{i \neq j} W_{ii} W'_{jj} + \sum_{i \neq j} W_{ij} W'_{ij} + \sum_{i \neq j} W_{ij} W'_{ji}\right) \tag{30}$$

$$= (d+4)(d+6)\left(\sum_{i,j=1}^d W_{ii} W'_{jj} + \sum_{i,j=1}^d W_{ij} W'_{ij} + \sum_{i,j=1}^d W_{ij} W'_{ji}\right)$$

$$= (d+4)(d+6)\left(\text{tr}(W)\,\text{tr}(W') + \text{tr}\left(W' W^\top\right) + \text{tr}(WW')\right). \tag{31}$$

where (30) is obtained by following

$$\mathbb{E}\left[u_i^4\left(\sum_{i'=1}^d u_{i'}^4 + \sum_{i' \neq j'} u_{i'}^2 u_{j'}^2\right)\right]$$

$$= \mathbb{E}[u^8] + (d-1)\mathbb{E}[u^4]\mathbb{E}[u^4] + 2(d-1)\mathbb{E}[u^6]\mathbb{E}[u^2] + (d-1)(d-2)\mathbb{E}[u^4]\mathbb{E}[u^2]\mathbb{E}[u^2]$$

$$= 105 + 9(d-1) + 30(d-1) + 3(d-1)(d-2)$$

$$= 3(d+4)(d+6),$$

$$\mathbb{E}\left[u_i^2 u_j^2\left(\sum_{i'=1}^d u_{i'}^4 + \sum_{i' \neq j'} u_{i'}^2 u_{j'}^2\right)\right]$$

$$= 2\mathbb{E}[u^6]\mathbb{E}[u^2] + (d-2)\mathbb{E}[u^4](\mathbb{E}[u^2])^2 + 2\mathbb{E}[u^4]\mathbb{E}[u^4] + 4(d-2)\mathbb{E}[u^4](\mathbb{E}[u^2])^2 + (d-2)(d-3)(\mathbb{E}[u^2])^4$$

$$= 30 + 3(d-2) + 18 + 12(d-2) + (d-2)(d-3)$$

$$= (d+4)(d+6).$$

B.2. Independent Data with General Covariance

Proof of Theorem 4. Consider a general independent linear model as defined in (6) where Σ_x and Σ_β are full-rank feature and task convariance matrices and

$$x \sim \mathcal{N}(0, \Sigma_x), \quad \beta \sim \mathcal{N}(0, \Sigma_\beta), \quad \xi \sim \mathcal{N}(0, \sigma^2), \quad \text{and} \quad y = x^\top \beta + \xi.$$

Let

$$X = [x_1 \ \cdots \ x_n]^\top, \quad \xi = [\xi_1 \ \cdots \ \xi_n]^\top, \quad \text{and} \quad y = [y_1 \ \cdots \ y_n]^\top = X\beta + \xi.$$

To simplify and without loss of generality, let $\bar{x} = \Sigma_x^{-1/2} x$, $\bar{X} = X\Sigma_x^{-1/2}$, $\bar{\beta} = \Sigma_x^{1/2}\beta$ where we have

$$\bar{x} \sim \mathcal{N}(0, I), \qquad \bar{\beta} \sim \mathcal{N}(0, \Sigma_x^{1/2}\Sigma_\beta\Sigma_x^{1/2})$$

and

$$y = \bar{x}^\top \bar{\beta} + \xi, \qquad y = \bar{X}\bar{\beta} + \boldsymbol{\xi}.$$

Then recap the loss from (4a), and we obtain

$$
\begin{aligned}
\mathcal{L}(W) &= \mathbb{E}\left[(y - g(Z))^2\right] \\
&= \mathbb{E}\left[\left(x^\top\beta + \xi - x^\top W X^\top (X\beta + \boldsymbol{\xi})\right)^2\right] \\
&= \mathbb{E}\left[(x^\top\beta - x^\top W X^\top X\beta)^2 + 2(x^\top\beta - x^\top W X^\top X\beta)(\xi - x^\top W X^\top \boldsymbol{\xi}) + (\xi - x^\top W X^\top \boldsymbol{\xi})^2\right] \\
&= \mathbb{E}\left[(x^\top\beta - x^\top W X^\top X\beta)^2\right] + \mathbb{E}\left[(x^\top W X^\top \boldsymbol{\xi})^2\right] + \sigma^2,
\end{aligned}
\tag{32}
$$

where the last equality comes from the independence of label noise $\xi, \boldsymbol{\xi}$.

We first consider the following term

$$\mathbb{E}\left[(x^\top W X^\top \boldsymbol{\xi})^2\right] = \mathbb{E}\left[(\bar{x}^\top (\Sigma_x^{1/2} W \Sigma_x^{1/2})\bar{X}^\top \boldsymbol{\xi})^2\right] = n\sigma^2 \cdot \mathrm{tr}\left(\bar{W}\bar{W}^\top\right)$$

where we define $\bar{W} = \Sigma_x^{1/2} W \Sigma_x^{1/2}$. Next, focus on the following

$$
\begin{aligned}
\mathbb{E}\left[(x^\top\beta - x^\top W X^\top X\beta)^2\right] &= \mathbb{E}\left[(\bar{x}^\top\bar{\beta} - \bar{x}^\top \bar{W}\bar{X}^\top \bar{X}\bar{\beta})^2\right] \\
&= \mathbb{E}\left[\left(\bar{x}^\top \left(I - \bar{W}\bar{X}^\top \bar{X}\right)\bar{\beta}\right)^2\right] \\
&= \mathrm{tr}\left(\mathbb{E}\left[\left(I - \bar{W}\bar{X}^\top \bar{X}\right)\Sigma\left(I - \bar{W}\bar{X}^\top \bar{X}\right)^\top\right]\right) \\
&= \mathrm{tr}(\Sigma) - \mathrm{tr}\left(\Sigma(\bar{W} + \bar{W}^\top)\mathbb{E}[\bar{X}^\top \bar{X}]\right) + \mathrm{tr}\left(\bar{W}^\top \bar{W}\,\mathbb{E}[\bar{X}^\top \bar{X}\Sigma\bar{X}^\top \bar{X}]\right) \\
&= \mathrm{tr}(\Sigma) - 2n \cdot \mathrm{tr}\left(\Sigma\bar{W}\right) + \mathrm{tr}\left(\bar{W}^\top \bar{W}\,\mathbb{E}[\bar{X}^\top \bar{X}\Sigma\bar{X}^\top \bar{X}]\right),
\end{aligned}
$$

where $\Sigma := \Sigma_x^{1/2}\Sigma_\beta\Sigma_x^{1/2}$.

Let $\bar{x}_i \in \mathbb{R}^n$ be the i'th column of $\bar{X}$ and Σ_{ij} be the (i, j)'th entry of Σ. Then the (i, j) entry of matrix $\bar{X}^\top \bar{X}\Sigma\bar{X}^\top \bar{X}$ is

$$(\bar{X}^\top \bar{X}\Sigma\bar{X}^\top \bar{X})_{ij} = \sum_{k=1}^{d}\sum_{p=1}^{d} \Sigma_{kp}\bar{x}_i^\top \bar{x}_k \bar{x}_p^\top \bar{x}_j.$$

Then we get

$$i \neq j: \quad \mathbb{E}\left[\left(\bar{X}^\top \bar{X} \Sigma \bar{X}^\top \bar{X}\right)_{ij}\right] = \Sigma_{ij} \mathbb{E}[\bar{x}_i^\top \bar{x}_i \bar{x}_j^\top \bar{x}_j] + \Sigma_{ji} \mathbb{E}[\bar{x}_i^\top \bar{x}_j \bar{x}_i^\top \bar{x}_j] = n^2 \Sigma_{ij} + n\Sigma_{ji}$$

$$i = j: \quad \mathbb{E}\left[\left(\bar{X}^\top \bar{X} \Sigma \bar{X}^\top \bar{X}\right)_{ii}\right] = \Sigma_{ii} \mathbb{E}\left[\bar{x}_i^\top \bar{x}_i \bar{x}_i^\top \bar{x}_i\right] + \sum_{j \neq i} \Sigma_{jj} \mathbb{E}\left[\bar{x}_i^\top \bar{x}_j \bar{x}_j^\top \bar{x}_i\right]$$

$$= \Sigma_{ii} \mathbb{E}\left[(x_{i1}^2 + \cdots + x_{in}^2)^2\right] + n \sum_{j \neq i} \Sigma_{jj}$$

$$= \Sigma_{ii}(3n + n(n-1)) + n \sum_{j \neq i} \Sigma_{jj}$$

$$= n\left(\Sigma_{ii}(n+1) + \sum_{j=1}^{d} \Sigma_{jj}\right)$$

$$= n\left(\Sigma_{ii}(n+1) + \mathtt{tr}\left(\Sigma\right)\right).$$

Therefore

$$\mathbb{E}[\bar{X}^\top \bar{X} \Sigma \bar{X}^\top \bar{X}] = n(n+1)\Sigma + n \cdot \mathtt{tr}\left(\Sigma\right) I.$$

Combining all together results in

$$\mathcal{L}(W) = \mathtt{tr}\left(\Sigma\right) - 2n\mathtt{tr}\left(\Sigma \bar{W}\right) + n(n+1)\mathtt{tr}\left(\Sigma \bar{W}^\top \bar{W}\right) + n(\mathtt{tr}\left(\Sigma\right) + \sigma^2)\mathtt{tr}\left(\bar{W}\bar{W}^\top\right) + \sigma^2,$$

$$= M - 2n\mathtt{tr}\left(\Sigma \bar{W}\right) + n(n+1)\mathtt{tr}\left(\Sigma \bar{W}^\top \bar{W}\right) + nM\mathtt{tr}\left(\bar{W}\bar{W}^\top\right), \tag{33}$$

where $M := \mathtt{tr}\left(\Sigma\right) + \sigma^2$. Setting $\nabla_{\bar{W}} \mathcal{L}(W) = 0$ returns

$$-2n \cdot \Sigma + 2n(n+1) \cdot \Sigma \bar{W} + 2nM\bar{W} = 0 \implies \bar{W}_\star = \left((n+1)I + M\Sigma^{-1}\right)^{-1}.$$

Then we have

$$W_\star = \Sigma_x^{-1/2}\left((n+1)I + M\Sigma^{-1}\right)^{-1} \Sigma_x^{-1/2}$$

and

$$\mathcal{L}_\star = \mathcal{L}(W_\star) = M - n\mathtt{tr}\left(((n+1)\Sigma^{-1} + M\Sigma^{-2})^{-1}\right).$$

∎

B.3. Retrieval Augmented Generation with α Correlation

In this section, we consider the retrieval augmented generation (RAG) linear model similar to (8), where we first draw the query vector x and task vector β via

$$x \sim \mathcal{N}(0, I) \quad \text{and} \quad \beta \sim \mathcal{N}(0, I).$$

We then draw data $(x_i)_{i=1}^n$ to be used in-context according to the rule corr_coef$(x, x_i) \geq \alpha \geq 0$. Hence, for $i \leq n$ we sample

$$x_i \mid x \sim \mathcal{N}(\alpha x, \gamma^2 I), \quad \xi_i \sim \mathcal{N}(0, \sigma^2) \quad \text{and} \quad y_i = x_i^\top \beta + \xi_i, \tag{34}$$

which results in (8) by setting $\gamma^2 = 1 - \alpha^2$.

Theorem 13 (Extended version of Theorem 6) *Consider linear model as defined in* (34). *Recap the objective from* (4a) *and let* $W_\star := \arg\min_W \mathcal{L}_{PGD}(W)$, *and* $\mathcal{L}_\star = \mathcal{L}_{PGD}(W_\star)$. *Then* $W_\star$ *and* $\mathcal{L}_\star$ *satisfy*

$$W_\star = cI \qquad and \qquad \mathcal{L}_\star = d + \sigma^2 - cnd(\alpha^2(d+2) + \gamma^2) \tag{35}$$

where

$$c = \frac{\alpha^2(d+2) + \gamma^2}{\alpha^4 n(d+2)(d+4) + \alpha^2\gamma^2(d+2)(d+2n+3) + \gamma^4(d+n+1) + \sigma^2(\alpha^2(d+2) + \gamma^2)}.$$

Suppose $\alpha = O(1/\sqrt{d})$, $d/n = O(1)$ *and* d *is sufficiently large. Let* $\kappa = \alpha^2 d + 1$ *and* $\gamma^2 = 1 - \alpha^2$. *Then* $W_\star$ *and* $\mathcal{L}_\star$ *have approximate forms*

$$W_\star \approx \frac{1}{\kappa n + d + \sigma^2}I \qquad and \qquad \mathcal{L}_\star \approx d + \sigma^2 - \frac{\kappa nd}{\kappa n + d + \sigma^2}. \tag{36}$$

Proof Here, for clean notation and without loss of generality, we define and rewrite (34) via

$$g_i \sim \mathcal{N}(0, I), \quad \xi_i \sim \mathcal{N}(0, \sigma^2) \quad \text{and} \quad x_i = \alpha x + \gamma g_i, \quad y_i = (\alpha x + \gamma g_i)^\top \beta + \xi_i.$$

Then we obtain

$$\begin{aligned}
\mathcal{L}(W) &= \mathbb{E}\left[(y - g(Z))^2\right] \\
&= \mathbb{E}\left[\left(x^\top \beta + \xi - x^\top W X^\top (X\beta + \xi)\right)^2\right] \\
&= \mathbb{E}\left[(x^\top\beta - x^\top W X^\top X\beta)^2 + 2(x^\top\beta - x^\top W X^\top X\beta)(\xi - x^\top W X^\top \xi) + (\xi - x^\top W X^\top \xi)^2\right] \\
&= \mathbb{E}\left[(x^\top\beta - x^\top W X^\top X\beta)^2\right] + \mathbb{E}\left[(x^\top W X^\top \xi)^2\right] + \sigma^2. \tag{37}
\end{aligned}$$

To begin with, let

$$N_1 = \operatorname{tr}(W)^2 + \operatorname{tr}\left(WW^\top\right) + \operatorname{tr}\left(W^2\right), \quad N_2 = \operatorname{tr}\left(WW^\top\right), \quad \text{and} \quad N_3 = \operatorname{tr}(W).$$

We first focus on the second term in (37)

$$\begin{aligned}
\mathbb{E}\left[(x^\top W X^\top \xi)^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^n \xi_i x^\top W (\alpha x + \gamma g_i)\right)^2\right] \\
&= n\sigma^2 \mathbb{E}\left[x^\top W(\alpha x + \gamma g)(\alpha x + \gamma g)^\top W^\top x\right] \\
&= n\sigma^2\left(\alpha^2 \mathbb{E}[x^\top W x x^\top W^\top x] + \gamma^2 \mathbb{E}[x^\top W g g^\top W^\top x]\right) \\
&= n\sigma^2\left(\alpha^2 N_1 + \gamma^2 N_2\right). \qquad \text{(It follows (26) and independence of } x, g.)
\end{aligned}$$

Next, the first term in (37) can be decomposed into

$$\mathbb{E}\left[(x^\top\beta - x^\top W X^\top X\beta)^2\right] = \underbrace{\mathbb{E}\left[(x^\top\beta)^2\right]}_{(a)} + \underbrace{\mathbb{E}\left[(x^\top W X^\top X\beta)^2\right]}_{(b)} - \underbrace{2\mathbb{E}\left[x^\top\beta x^\top W X^\top X\beta\right]}_{(c)}.$$

25

In the following, we consider solving (a)-(c) sequentially.

$(a):\quad \mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{\beta})^2\right] = d.$

$(b):\quad \mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})^2\right]$

$$= \mathbb{E}\left[\left(\boldsymbol{x}^\top\boldsymbol{W}\sum_{i=1}^n(\alpha\boldsymbol{x}+\gamma\boldsymbol{g}_i)(\alpha\boldsymbol{x}+\gamma\boldsymbol{g}_i)^\top\boldsymbol{\beta}\right)^2\right]$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^n\boldsymbol{x}^\top\boldsymbol{W}(\alpha^2\boldsymbol{x}\boldsymbol{x}^\top+\gamma^2\boldsymbol{g}_i\boldsymbol{g}_i^\top+\alpha\gamma\boldsymbol{x}\boldsymbol{g}_i^\top+\alpha\gamma\boldsymbol{g}_i\boldsymbol{x}^\top)\boldsymbol{\beta}\right)^2\right]$$

$$= \alpha^4 n^2\,\mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta})^2\right]+\gamma^4\,\mathbb{E}\left[\left(\sum_{i=1}^n\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{g}_i\boldsymbol{g}_i^\top\boldsymbol{\beta}\right)^2\right]+\alpha^2\gamma^2\,\mathbb{E}\left[\left(\sum_{i=1}^n\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x}\boldsymbol{g}_i^\top\boldsymbol{\beta}\right)^2\right]+\alpha^2\gamma^2\,\mathbb{E}\left[\left(\sum_{i=1}^n\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{g}_i\boldsymbol{x}^\top\boldsymbol{\beta}\right)^2\right]$$

$$+2\alpha^2\gamma^2 n^2\,\mathbb{E}\left[\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}\boldsymbol{\beta}^\top\boldsymbol{g}\boldsymbol{g}^\top\boldsymbol{W}^\top\boldsymbol{x}\right]+2\alpha^2\gamma^2 n\,\mathbb{E}\left[\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x}\boldsymbol{g}^\top\boldsymbol{\beta}\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{g}\boldsymbol{x}^\top\boldsymbol{\beta}\right]$$

$$= \left(\alpha^4 n^2(d+4)N_1+\gamma^4 n(d+n+1)N_2\right)+\left(\alpha^2\gamma^2 ndN_1+\alpha^2\gamma^2 n(d+2)N_2\right)+\left(2\alpha^2\gamma^2 n^2 N_1+2\alpha^2\gamma^2 nN_1\right)$$

$$= \left(\alpha^4 n^2(d+4)+\alpha^2\gamma^2 n(2n+d+2)\right)N_1+\left(\alpha^2\gamma^2 n(d+2)+\gamma^4 n(d+n+1)\right)N_2$$

$$= A_1 N_1+A_2 N_2.$$

$(c):\quad \mathbb{E}\left[\boldsymbol{x}^\top\boldsymbol{\beta}\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}\right]=\mathbb{E}\left[\sum_{i=1}^n\boldsymbol{x}^\top\boldsymbol{\beta}\boldsymbol{x}^\top\boldsymbol{W}(\alpha\boldsymbol{x}+\gamma\boldsymbol{g}_i)(\alpha\boldsymbol{x}+\gamma\boldsymbol{g}_i)^\top\boldsymbol{\beta}\right]$

$$= \mathbb{E}\left[\sum_{i=1}^n\boldsymbol{x}^\top\boldsymbol{\beta}\boldsymbol{x}^\top\boldsymbol{W}(\alpha^2\boldsymbol{x}\boldsymbol{x}^\top+\gamma^2\boldsymbol{g}_i\boldsymbol{g}_i^\top+\alpha\gamma\boldsymbol{x}\boldsymbol{g}_i^\top+\alpha\gamma\boldsymbol{g}_i\boldsymbol{x}^\top)\boldsymbol{\beta}\right]$$

$$= \alpha^2 n\,\mathbb{E}\left[\boldsymbol{x}^\top\boldsymbol{\beta}\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}\right]+\gamma^2 n\,\mathbb{E}\left[\boldsymbol{x}^\top\boldsymbol{\beta}\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{g}\boldsymbol{g}^\top\boldsymbol{\beta}\right]$$

$$= \alpha^2 n(d+2)\mathrm{tr}\,(\boldsymbol{W})+\gamma^2 n\mathrm{tr}\,(\boldsymbol{W})$$

$$= \left(\alpha^2 n(d+2)+\gamma^2 n\right)N_3$$

$$= A_3 N_3.$$

Here, (b) utilizes the 4'th and 6'th moment results (26) and (29) and we define

$$A_1 = \alpha^4 n^2(d+4)+\alpha^2\gamma^2 n(2n+d+2)$$
$$A_2 = \alpha^2\gamma^2 n(d+2)+\gamma^4 n(d+n+1)$$
$$A_3 = \alpha^2 n(d+2)+\gamma^2 n.$$

Then combining all together results in

$$\mathcal{L}(\boldsymbol{W}) = A_1 N_1+A_2 N_2-2A_3 N_3+n\sigma^2(\alpha^2 N_1+\gamma^2 N_2)+d+\sigma^2.$$

To find the optimal solution, set $\nabla\mathcal{L}(\boldsymbol{W}) = 0$ and we obtain

$$A_1\nabla N_1+A_2\nabla N_2-2A_3\nabla N_3+n\sigma^2(\alpha^2\nabla N_1+\gamma^2\nabla N_2) = 0. \tag{38}$$

Note that we have

$$\nabla N_1 = \nabla \left(\text{tr}\,(W)^2 + \text{tr}\left(WW^\top\right) + \text{tr}\left(W^2\right)\right) = 2\text{tr}\,(W)\,I + 2W + 2W^\top$$
$$\nabla N_2 = \nabla \text{tr}\left(WW^\top\right) = 2W$$
$$\nabla N_3 = \nabla \text{tr}\,(W) = I.$$

Therefore, (38) returns

$$2A_1\left(\text{tr}\,(W)\,I + W + W^\top\right) + 2A_2 W - 2A_3 + 2n\sigma^2(\alpha^2(\text{tr}\,(W)\,I + W + W^\top) + \gamma^2 W)I = 0, \quad (39)$$

which implies that the optimal solution $W_\star$ has the form of cI for some constant c. Then suppose $W_\star = cI$, we have $\text{tr}\,(W) = cd$ and (39) returns

$$2A_1(d+2)cI + 2A_2 cI - 2A_3 I + 2n\sigma^2(\alpha^2(d+2)cI + \gamma^2 cI) = 0$$

$$\implies c = \frac{A_3}{A_1(d+2) + A_2 + n\sigma^2(\alpha^2(d+2) + \gamma^2)}$$
$$= \frac{\alpha^2(d+2) + \gamma^2}{\alpha^4 n(d+2)(d+4) + \alpha^2\gamma^2(d+2)(d+2n+3) + \gamma^4(d+n+1) + \sigma^2(\alpha^2(d+2) + \gamma^2)}.$$

Then the optimal loss is obtained by setting $W_\star = cI$ and

$$\mathcal{L}_\star = \mathcal{L}(W_\star) = A_1 c^2 d(d+2) + A_2 c^2 d - 2A_3 cd + n\sigma^2 c^2 d(\alpha^2(d+2) + \gamma^2) + d + \sigma^2$$
$$= c^2 d\left(A_1(d+2) + A_2 + n\sigma^2(\alpha^2(d+2) + \gamma^2)\right) - 2A_3 cd + d + \sigma^2$$
$$= d + \sigma^2 - A_3 cd.$$

It completes the proof of (35). Now if assuming $\alpha = O\left(1/\sqrt{d}\right)$, $d/n = O(1)$ and sufficiently large dimension d, we have the approximate

$$c \approx \frac{\alpha^2 d + 1}{\alpha^4 d^2 n + \alpha^2 d(d+2n) + (d+n) + \sigma^2(\alpha^2 d + 1)}$$
$$= \frac{\alpha^2 d + 1}{(\alpha^2 d + 1)^2 n + (\alpha^2 d + 1)d + \sigma^2(\alpha^2 d + 1)}$$
$$= \frac{1}{(\alpha^2 d + 1)n + d + \sigma^2}$$

and

$$\mathcal{L}_\star \approx d + \sigma^2 - \frac{(\alpha^2 d + 1)nd}{(\alpha^2 d + 1)n + d + \sigma^2}.$$

■

B.4. Task-feature Alignment with α Correlation

In this section, we consider the task-feature alignment data model similar to (10), where we first draw task vector $\boldsymbol{\beta}$ via

$$\boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{I}).$$

Then we generate examples $(\boldsymbol{x}_i, y_i)_{i=1}^{n+1}$ according to the rule corr_coef$(\boldsymbol{x}_i, \boldsymbol{\beta}) \geq \alpha \geq 0$ via

$$\boldsymbol{x}_i \mid \boldsymbol{\beta} \sim \mathcal{N}(\alpha\boldsymbol{\beta}, \boldsymbol{I}), \quad \xi_i \sim \mathcal{N}(0, \sigma^2) \quad \text{and} \quad y_i = \gamma \cdot \boldsymbol{x}_i^\top \boldsymbol{\beta} + \xi_i, \tag{40}$$

which results in (10) by setting $\gamma^2 = 1/(\alpha^2 d + 1)$.

Theorem 14 (Extended version of Theorem 7) *Consider linear model as defined in* (40). *Recap the objective from* (4a) *and let* $\boldsymbol{W}_\star := \arg\min_{\boldsymbol{W}} \mathcal{L}_{PGD}(\boldsymbol{W})$, *and* $\mathcal{L}_\star = \mathcal{L}_{PGD}(\boldsymbol{W}_\star)$. *Then* $\boldsymbol{W}_\star$ *and* $\mathcal{L}_\star$ *satisfy*

$$\boldsymbol{W}_\star = c\boldsymbol{I} \quad \text{and} \quad \mathcal{L}_\star = d\gamma^2(\Delta_0\alpha^2 + 1) + \sigma^2 - cnd\gamma^2(\Delta_1\alpha^4 + 2\Delta_0\alpha^2 + 1) \tag{41}$$

where

$$c = \frac{\Delta_1\alpha^4 + 2\Delta_0\alpha^2 + 1}{\Delta_2\alpha^6 + \Delta_3\alpha^4 + \Delta_4\alpha^2 + (d + n + 1) + \sigma^2(\Delta_0\alpha^4 + 2\alpha^2 + 1)/\gamma^2}$$

and

$$\begin{cases} \Delta_0 = d + 2 \\ \Delta_1 = (d + 2)(d + 4) \\ \Delta_2 = (d + 2)(d + 4)(d + 6)n \\ \Delta_3 = (d + 2)(d + 4)(3n + 4) \\ \Delta_4 = (d + 2)(3n + d + 3) + (d + 8). \end{cases}$$

Suppose $\alpha = O\left(1/\sqrt{d}\right)$, $d/n = O(1)$ *and* d *is sufficiently large. Let* $\kappa = \alpha^2 d + 1$ *and* $\gamma^2 = 1/\kappa$. *Then* $\boldsymbol{W}_\star$ *and* $\mathcal{L}_\star$ *have approximate forms*

$$\boldsymbol{W}_\star \approx \frac{1}{\kappa n + (d + \sigma^2)/\kappa} \quad \text{and} \quad \mathcal{L}_\star \approx d + \sigma^2 - \frac{\kappa nd}{\kappa n + (d + \sigma^2)/\kappa}. \tag{42}$$

Proof Here, for clean notation and without loss of generality, we define and rewrite (40) via

$$\boldsymbol{g}_i \sim \mathcal{N}(0, \boldsymbol{I}), \quad \xi_i \sim \mathcal{N}(0, \sigma^2) \quad \text{and} \quad \boldsymbol{x}_i = \alpha\boldsymbol{\beta} + \boldsymbol{g}_i, \quad y_i = \gamma\boldsymbol{x}_i^\top\boldsymbol{\beta} + \xi_i = \gamma \cdot (\alpha\boldsymbol{\beta} + \boldsymbol{g}_i)^\top\boldsymbol{\beta} + \xi_i.$$

Recap the loss function from (4a), we obtain

$$\begin{aligned} \mathcal{L}(\boldsymbol{W}) &= \mathbb{E}\left[(y - g(\boldsymbol{Z}))^2\right] \\ &= \mathbb{E}\left[\left(\gamma\boldsymbol{x}^\top\boldsymbol{\beta} + \xi - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top(\gamma\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi})\right)^2\right] \\ &= \mathbb{E}\left[\gamma^2(\boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})^2 + 2\gamma(\boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})(\xi - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi}) + (\xi - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi})^2\right] \\ &= \gamma^2\mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})^2\right] + \mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi})^2\right] + \sigma^2. \end{aligned} \tag{43}$$

Similar to Appendix B.3, to begin with, let

$$N_1 = \text{tr}\left(\boldsymbol{W}\right)^2 + \text{tr}\left(\boldsymbol{W}\boldsymbol{W}^\top\right) + \text{tr}\left(\boldsymbol{W}^2\right), \quad N_2 = \text{tr}\left(\boldsymbol{W}\boldsymbol{W}^\top\right), \quad \text{and} \quad N_3 = \text{tr}\left(\boldsymbol{W}\right),$$

and additionally, given $\Lambda_W = W \odot I$, let

$$N_4 = 3\mathrm{tr}\left(\Lambda_W^2\right) + (d+4)\mathrm{tr}\left(WW^\top\right) + \mathrm{tr}\left(W^2\right).$$

We first focus on the second term in (43)

$$
\begin{aligned}
\mathbb{E}\left[(x^\top W X^\top \xi)^2\right] &= \mathbb{E}\left[\left((\alpha\beta + g)^\top W \sum_{i=1}^n \xi_i(\alpha\beta + g_i)\right)^2\right] \\
&= n\sigma^2 \mathbb{E}\left[\left((\alpha\beta + g)^\top W(\alpha\beta + g')\right)^2\right] \\
&= n\sigma^2 \left(\alpha^4 \mathbb{E}\left[(\beta^\top W\beta)^2\right] + 2\alpha^2 \mathbb{E}\left[(\beta^\top W g')^2\right] + \mathbb{E}\left[(g^\top W g')^2\right]\right) \\
&= n\sigma^2 \left(\alpha^4 \left(\mathrm{tr}(W)^2 + \mathrm{tr}\left(W^2\right) + \mathrm{tr}\left(WW^\top\right)\right) + (2\alpha^2 + 1)\mathrm{tr}\left(WW^\top\right)\right) \\
&= n\sigma^2 \left(\alpha^4 N_1 + (2\alpha^2 + 1)N_2\right). \quad \text{(It follows (26) and independence of } \beta, g, g'.\text{)}
\end{aligned}
$$

Next, the first term of (43) (omitting γ^2) returns the following decomposition:

$$
\begin{aligned}
\mathbb{E}\left[(x^\top \beta - x^\top W X^\top X\beta)^2\right] &= \mathbb{E}\left[((\alpha\beta + g)^\top(\beta - W X^\top X\beta))^2\right] \\
&= \mathbb{E}\left[\left(\alpha\beta^\top \beta - \alpha\beta^\top W X^\top X\beta + g^\top \beta - g^\top W X^\top X\beta\right)^2\right] \\
&= \alpha^2 \mathbb{E}[(\beta^\top \beta)^2] + \alpha^2 \mathbb{E}[(\beta^\top W X^\top X\beta)^2] + \mathbb{E}[(g^\top \beta)^2] + \mathbb{E}[(g^\top W X^\top X\beta)^2] \\
&\quad - 2\alpha^2 \mathbb{E}[\beta^\top \beta\beta^\top W X^\top X\beta] - 2\mathbb{E}[\beta^\top g g^\top W X^\top X\beta] \\
&= \alpha^2 d(d+2) + \alpha^2 \underbrace{\mathbb{E}[(\beta^\top W X^\top X\beta)^2]}_{(a)} + d + \underbrace{\mathbb{E}[(g^\top W X^\top X\beta)^2]}_{(b)} \\
&\quad - 2\alpha^2 \underbrace{\mathbb{E}[\beta^\top \beta\beta^\top W X^\top X\beta]}_{(c)} - 2\underbrace{\mathbb{E}[\beta^\top g g^\top W X^\top X\beta]}_{(d)}.
\end{aligned}
$$

Consider solving (a)-(d) sequentially as follows:

To begin with, we use the following decomposition for all (a)-(d):

$$
\begin{aligned}
X^\top X\beta &= \sum_{i=1}^n x_i x_i^\top \beta \\
&= \sum_{i=1}^n (\alpha\beta + g_i)(\alpha\beta + g_i)^\top \beta \\
&= \sum_{i=1}^n \alpha^2 \beta\beta^\top \beta + \alpha\beta g_i^\top \beta + \alpha g_i \beta^\top \beta + g_i g_i^\top \beta.
\end{aligned}
$$

Then, we have

$$(a): \quad \mathbb{E}[(\boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta})^2]$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^n \alpha^2 \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{\beta} + \alpha \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{g}_i^\top \boldsymbol{\beta} + \alpha \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{\beta}^\top \boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{g}_i^\top \boldsymbol{\beta}\right)^2\right]$$

$$= \alpha^4 n^2 \, \mathbb{E}\left[\left(\boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{\beta}\right)^2\right] + \alpha^2 \, \mathbb{E}\left[\left(\sum_{i=1}^n \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{g}_i^\top \boldsymbol{\beta}\right)^2\right] + \alpha^2 \, \mathbb{E}\left[\left(\sum_{i=1}^n \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{\beta}^\top \boldsymbol{\beta}\right)^2\right] + \mathbb{E}\left[\left(\sum_{i=1}^n \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{g}_i^\top \boldsymbol{\beta}\right)^2\right]$$

$$+ 2\alpha^2 n \, \mathbb{E}\left[\sum_{i=1}^n \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{g}_i^\top \boldsymbol{\beta}\right] + 2\alpha^2 \, \mathbb{E}\left[\sum_{i=1}^n \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{g}_i^\top \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{\beta}^\top \boldsymbol{\beta}\right]$$

$$= \alpha^4 n^2 \, \mathbb{E}\left[\left(\boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{\beta}\right)^2\right] + \alpha^2 n \, \mathbb{E}\left[\left(\boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{g}'^\top \boldsymbol{\beta}\right)^2\right] + \alpha^2 n \, \mathbb{E}\left[\left(\boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}' \boldsymbol{\beta}^\top \boldsymbol{\beta}\right)^2\right] + \mathbb{E}\left[\left(\sum_{i=1}^n \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{g}_i^\top \boldsymbol{\beta}\right)^2\right]$$

$$+ 2\alpha^2 n^2 \, \mathbb{E}\left[\boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}' \boldsymbol{g}'^\top \boldsymbol{\beta}\right] + 2\alpha^2 n \, \mathbb{E}\left[\boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{g}_i^\top \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{\beta}^\top \boldsymbol{\beta}\right]$$

$$= \alpha^4 n^2 (d+4)(d+6) N_1 + \alpha^2 n (d+4) N_1 + \alpha^2 n (d+2)(d+4) N_2 \tag{44}$$

$$+ n(n-1) N_1 + n N_4 \tag{45}$$

$$+ 2\alpha^2 n^2 (d+4) N_1 + 2\alpha^2 n (d+4) N_1 \tag{46}$$

$$= \left(\alpha^2 n(d+4)(\alpha^2 n(d+6) + 2n + 3) + n(n-1)\right) N_1 + \alpha^2 n(d+2)(d+4) N_2 + n N_4 \tag{47}$$

$$= B_1 N_1 + B_2 N_2 + n N_4,$$

where (44) and (46) utilize (29) and (31), and (45) is obtained via

$$\mathbb{E}\left[\left(\sum_{i=1}^n \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{g}_i^\top \boldsymbol{\beta}\right)^2\right] = n \, \mathbb{E}\left[\left(\boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}' \boldsymbol{g}'^\top \boldsymbol{\beta}\right)^2\right] + n(n-1) \, \mathbb{E}\left[\boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}' \boldsymbol{g}'^\top \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}'' \boldsymbol{g}''^\top \boldsymbol{\beta}\right]$$

$$= n N_4 + n(n-1) N_1,$$

which follows (26) and (27).

$(b):\quad \mathbb{E}\left[(\boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta})^2\right]$

$$= \mathbb{E}\left[\left(\sum_{i=1}^{n} \alpha^2 \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{\beta} + \alpha \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{g}_i^\top \boldsymbol{\beta} + \alpha \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{\beta}^\top \boldsymbol{\beta} + \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{g}_i^\top \boldsymbol{\beta}\right)^2\right]$$

$$= \alpha^4 n^2 \mathbb{E}\left[\left(\boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{\beta}\right)^2\right] + \alpha^2 \mathbb{E}\left[\left(\sum_{i=1}^{n} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{g}_i^\top \boldsymbol{\beta}\right)^2\right] + \alpha^2 \mathbb{E}\left[\left(\sum_{i=1}^{n} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{\beta}^\top \boldsymbol{\beta}\right)^2\right] + \mathbb{E}\left[\left(\sum_{i=1}^{n} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{g}_i^\top \boldsymbol{\beta}\right)^2\right]$$

$$+ 2\alpha^2 n \mathbb{E}\left[\sum_{i=1}^{n} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{\beta} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{g}_i^\top \boldsymbol{\beta}\right] + 2\alpha^2 \mathbb{E}\left[\sum_{i=1}^{n} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{g}_i^\top \boldsymbol{\beta} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{\beta}^\top \boldsymbol{\beta}\right]$$

$$= \alpha^4 n^2 \mathbb{E}\left[\left(\boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{\beta}\right)^2\right] + \alpha^2 n \mathbb{E}\left[\left(\boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{g}'^\top \boldsymbol{\beta}\right)^2\right] + \alpha^2 n \mathbb{E}\left[\left(\boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}' \boldsymbol{\beta}^\top \boldsymbol{\beta}\right)^2\right] + \mathbb{E}\left[\left(\sum_{i=1}^{n} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{g}_i^\top \boldsymbol{\beta}\right)^2\right]$$

$$+ 2\alpha^2 n^2 \mathbb{E}\left[\boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{\beta} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}' \boldsymbol{g}'^\top \boldsymbol{\beta}\right] + 2\alpha^2 n \mathbb{E}\left[\boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{g}_i^\top \boldsymbol{\beta} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{\beta}^\top \boldsymbol{\beta}\right]$$

$$= \alpha^4 n^2 (d+2)(d+4) N_2 + \alpha^2 n(d+2) N_2 + \alpha^2 n d(d+2) N_2 + n(d+n+1) N_2 \tag{48}$$

$$+ 2\alpha^2 n^2 (d+2) N_2 + 2\alpha^2 n(d+2) N_2 \tag{49}$$

$$= \left(\alpha^2 n(d+2)(\alpha^2 n(d+4) + 2n + d + 3) + n(d+n-1)\right) N_2$$

$$= B_3 N_2,$$

where (48) and (49) are obtained using (26), (29) and

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}_i \boldsymbol{g}_i^\top \boldsymbol{\beta}\right)^2\right] = n \mathbb{E}\left[\left(\boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}' \boldsymbol{g}'^\top \boldsymbol{\beta}\right)^2\right] + n(n-1) \mathbb{E}\left[\boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}' \boldsymbol{g}'^\top \boldsymbol{\beta} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}'' \boldsymbol{g}''^\top \boldsymbol{\beta}\right]$$

$$= n(d+2) N_2 + n(n-1) N_2 = n(n+d+1) N_2.$$

$(c):\quad \mathbb{E}\left[\boldsymbol{\beta}^\top \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta}\right]$

$$= n \mathbb{E}\left[\boldsymbol{\beta}^\top \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{W} (\alpha \boldsymbol{\beta} + \boldsymbol{g}')(\alpha \boldsymbol{\beta} + \boldsymbol{g}')^\top \boldsymbol{\beta}\right]$$

$$= \alpha^2 n \mathbb{E}\left[\boldsymbol{\beta}^\top \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{\beta}\right] + n \mathbb{E}\left[\boldsymbol{\beta}^\top \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{W} \boldsymbol{g}' \boldsymbol{g}'^\top \boldsymbol{\beta}\right]$$

$$= \alpha^2 n(d+2)(d+4) \mathrm{tr}(\boldsymbol{W}) + n(d+2) \mathrm{tr}(\boldsymbol{W})$$

$$= \left(\alpha^2 n(d+2)(d+4) + n(d+2)\right) N_3$$

$$= B_4 N_3.$$

$(d):\quad \mathbb{E}\left[\boldsymbol{\beta}^\top \boldsymbol{g} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta}\right]$

$$= n \mathbb{E}\left[\boldsymbol{\beta}^\top \boldsymbol{g} \boldsymbol{g}^\top \boldsymbol{W} (\alpha \boldsymbol{\beta} + \boldsymbol{g}')(\alpha \boldsymbol{\beta} + \boldsymbol{g}')^\top \boldsymbol{\beta}\right]$$

$$= \alpha^2 n \mathbb{E}\left[\boldsymbol{\beta}^\top \boldsymbol{g} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{\beta} \boldsymbol{\beta}^\top \boldsymbol{\beta}\right] + n \mathbb{E}\left[\boldsymbol{\beta}^\top \boldsymbol{g} \boldsymbol{g}^\top \boldsymbol{W} \boldsymbol{g}' \boldsymbol{g}'^\top \boldsymbol{\beta}\right]$$

$$= \alpha^2 n(d+2) \mathrm{tr}(\boldsymbol{W}) + n \mathrm{tr}(\boldsymbol{W})$$

$$= \left(\alpha^2 n(d+2) + n\right) N_3$$

$$= B_5 N_3.$$

Here we define

$$
\begin{aligned}
B_1 &= \alpha^2 n(d+4)(\alpha^2 n(d+6) + 2n + 3) + n(n-1) \\
B_2 &= \alpha^2 n(d+2)(d+4) \\
B_3 &= \alpha^2 n(d+2)(\alpha^2 n(d+4) + 2n + d + 3) + n(d+n-1) \\
B_4 &= \alpha^2 n(d+2)(d+4) + n(d+2) \\
B_5 &= \alpha^2 n(d+2) + n.
\end{aligned}
$$

Then combining all together results in

$$
\begin{aligned}
\mathcal{L}(W) &= \gamma^2 \left(\alpha^2 d(d+2) + d + \alpha^2 (B_1 N_1 + B_2 N_2 + nN_4) + B_3 N_2 - 2\alpha^2 B_4 N_3 - 2B_5 N_3 \right) + n\sigma^2(\alpha^4 N_1 + (2\alpha^2 + 1)N_2) + \sigma^2 \\
&= \gamma^2 \left(\alpha^2 B_1 N_1 + (\alpha^2 B_2 + B_3)N_2 - 2(\alpha^2 B_4 + B_5)N_3 + \alpha^2 n N_4 \right) + n\sigma^2(\alpha^4 N_1 + (2\alpha^2 + 1)N_2) + \gamma^2 d\left(\alpha^2(d+2) + 1\right) + \sigma^2
\end{aligned}
$$

and differentiating it results in

$$
\nabla\mathcal{L}(W) = \gamma^2 \left(\alpha^2 B_1 \nabla N_1 + (\alpha^2 B_2 + B_3)\nabla N_2 - 2(\alpha^2 B_4 + B_5)\nabla N_3 + \alpha^2 n \nabla N_4 \right) + n\sigma^2(\alpha^4 \nabla N_1 + (2\alpha^2 + 1)\nabla N_2).
$$

Similar to the proof in Appendix B.3, $W_\star$ has the form of $W_\star = cI$ and we have

$$
\begin{aligned}
\nabla N_1 &= \nabla\left(\operatorname{tr}(W)^2 + \operatorname{tr}\left(WW^\top\right) + \operatorname{tr}\left(W^2\right)\right) = 2\operatorname{tr}(W)I + 2W + 2W^\top = 2c(d+2)I \\
\nabla N_2 &= \nabla\operatorname{tr}\left(WW^\top\right) = 2W = 2cI \\
\nabla N_3 &= \nabla\operatorname{tr}(W) = I \\
\nabla N_4 &= \nabla\left(3\operatorname{tr}\left(\Lambda_W^2\right) + (d+4)\operatorname{tr}\left(WW^\top\right) + \operatorname{tr}\left(W^2\right)\right) \\
&= 6 \cdot \operatorname{diag}(\Lambda_W) + 2(d+4)W + 2W^\top \\
&= 2c(d+8)I.
\end{aligned}
$$

Therefore, setting $\nabla\mathcal{L}(W) = 0$ returns

$$
\gamma^2 \left(2c(d+2)\alpha^2 B_1 + 2c(\alpha^2 B_2 + B_3) - 2(\alpha^2 B_4 + B_5) + 2c(d+8)\alpha^2 n \right) + 2cn\sigma^2(\alpha^4(d+2) + 2\alpha^2 + 1) = 0
$$

$$
\begin{aligned}
\Longrightarrow c &= \frac{\alpha^2 B_4 + B_5}{(d+2)\alpha^2 B_1 + (\alpha^2 B_2 + B_3) + (d+8)\alpha^2 n + n\sigma^2(\alpha^4(d+2) + 2\alpha^2 + 1)/\gamma^2} \\
&= \frac{\alpha^4 n(d+2)(d+4) + 2\alpha^2 n(d+2) + n}{\alpha^6 n^2(d+2)(d+4)(d+6) + \alpha^4 n(d+2)(d+4)(3n+4) + \alpha^2 n((d+2)(3n+d+3) + (d+8)) + n(d+n+1) + n\sigma^2(\alpha^4(d+2) + 2\alpha^2 + 1)/\gamma^2} \\
&= \frac{\alpha^4(d+2)(d+4) + 2\alpha^2(d+2) + 1}{\alpha^6 n(d+2)(d+4)(d+6) + \alpha^4(d+2)(d+4)(3n+4) + \alpha^2((d+2)(3n+d+3) + (d+8)) + (d+n+1) + \sigma^2(\alpha^4(d+2) + 2\alpha^2 + 1)/\gamma^2}.
\end{aligned}
$$

Then the optimal loss is obtained by setting $W_\star = cI$ and

$$
\mathcal{L}_\star = \mathcal{L}(W_\star) = \gamma^2 d(\alpha^2(d+2) + 1) + \sigma^2 - \gamma^2(\alpha^2 B_4 + B_5)cd.
$$

It completes the proof of (41). Now if assuming $\alpha = O\left(1/\sqrt{d}\right), d/n = O(1), \gamma^2 = 1/(\alpha^2 d + 1)$ and sufficiently large dimension d, we have the approximate

$$
\begin{aligned}
c &\approx \frac{\alpha^4 d^2 + 2\alpha^2 d + 1}{n\alpha^6 d^3 + 3n\alpha^4 d^2 + (3n+d)\alpha^2 d + d + n + \sigma^2(\alpha^4 d + 2\alpha^2 + 1)/\gamma^2} \\
&\approx \frac{(\alpha^2 d + 1)^2}{n(\alpha^2 d + 1)^3 + d(\alpha^2 d + 1) + \sigma^2(\alpha^2 d + 1)} \\
&\approx \frac{1}{(\alpha^2 d + 1)n + (d + \sigma^2)/(\alpha^2 d + 1)}
\end{aligned}
$$

and

$$\mathcal{L}_\star \approx \gamma^2 d(\alpha^2 d + 1) + \sigma^2 - \frac{\gamma^2(\alpha^2 d + 1)^2 nd}{(\alpha^2 d + 1)n + (d + \sigma^2)/(\alpha^2 d + 1)}$$

$$= d + \sigma^2 - \frac{(\alpha^2 d + 1)nd}{(\alpha^2 d + 1)n + (d + \sigma^2)/(\alpha^2 d + 1)}.$$

∎

Appendix C. Analysis of Low-Rank Parameterization

C.1. Proof of Lemma 8

Proof Recall the loss function from (33)

$$\mathcal{L}(W) = M - 2n\text{tr}\left(\Sigma\bar{W}\right) + n(n + 1)\text{tr}\left(\Sigma\bar{W}^\top\bar{W}\right) + nM\text{tr}\left(\bar{W}\bar{W}^\top\right)$$

where $\bar{W} = \Sigma_x^{1/2}W\Sigma_x^{1/2}$, $\Sigma = \Sigma_x^{1/2}\Sigma_\beta\Sigma_x^{1/2}$ and $M = \text{tr}(\Sigma) + \sigma^2$. For any $\bar{W}$, let us parameterize $\bar{W} = UEU^\top$ where $U \in \mathbb{R}^{d\times r}$ denotes the eigenvectors of $\bar{W}$ and $E \in \mathbb{R}^{r\times r}$ is a symmetric square matrix. We will first treat U as fixed and optimize E. We will then optimize U. Fixing U, setting $\bar{\Sigma} = U^\top\Sigma U$, we obtain

$$\mathcal{L}(E) = M - 2n\text{tr}\left(\bar{\Sigma}E\right) + n(n + 1)\text{tr}\left(\bar{\Sigma}E^2\right) + nM\text{tr}\left(E^2\right).$$

Differentiating, we obtain

$$0.5n^{-1}\nabla\mathcal{L}(E) = -\bar{\Sigma} + (n + 1)\bar{\Sigma}E + ME.$$

Setting $\nabla\mathcal{L}(E) = 0$ returns

$$E_\star = (MI + (n + 1)\bar{\Sigma})^{-1}\bar{\Sigma}. \tag{50}$$

Let $\bar{\lambda}_i$ denote the i'th largest eigenvalue of $\bar{\Sigma}$. Plugging in this value, we obtain the optimal risk as a function of U is given by

$$\mathcal{L}_\star(U) = M - n \cdot \text{tr}\left(\bar{\Sigma}E_\star\right) = M - n \cdot \text{tr}\left((MI + (n + 1)\bar{\Sigma})^{-1}\bar{\Sigma}^2\right) \tag{51}$$

$$= M - n \sum_{i=1}^r \frac{\bar{\lambda}_i^2}{(n + 1)\bar{\lambda}_i + M} = M - n \sum_{i=1}^r \frac{\bar{\lambda}_i}{n + 1 + M\bar{\lambda}_i^{-1}}. \tag{52}$$

Now observe that, the right hand side is strictly decreasing function of the eigenvalues $\bar{\lambda}_i$ of $\bar{\Sigma} = U^\top\Sigma U$. Thus, to minimize $\mathcal{L}_\star(U)$, we need to maximize $\sum_{i=1}^r \frac{\bar{\lambda}_i}{n+1+M\bar{\lambda}_i^{-1}}$. It follows from Cauchy interlacing theorem that $\bar{\lambda}_j \leq \lambda_i$ where λ_i is the i'th largest eigenvalue of Σ since $\bar{\Sigma}$ is an orthogonal projection of Σ on U. Consequently, we find the desired bound where

$$\mathcal{L}_\star = M - n \sum_{i=1}^r \frac{\lambda_i}{n + 1 + M\lambda_i^{-1}}.$$

The equality holds by setting U to be the top-r eigenvectors of Σ and $E = E_\star(U)$ to be the diagonal matrix according to (50). ∎

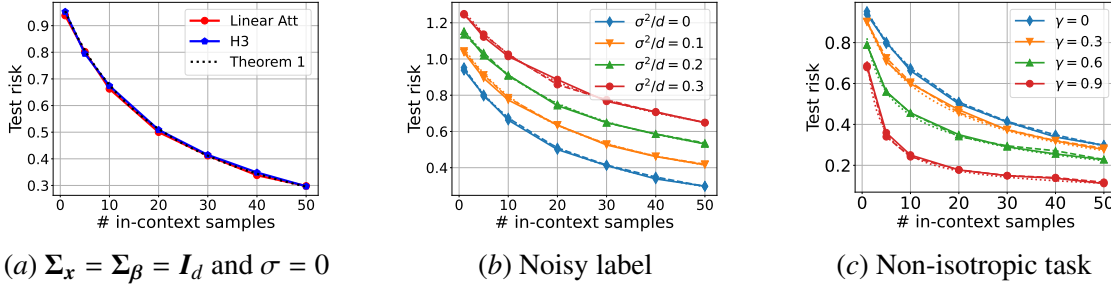(a) $\Sigma_x = \Sigma_\beta = I_d$ and $\sigma = 0$ (b) Noisy label (c) Non-isotropic task

Figure 2: Empirical evidence validates Theorem 4 and Proposition 3. We train 1-layer linear attention and H3 models with prompts containing independent demonstrations following a linear model, and dotted curves are the theory curves following Eq. (7). **(a):** We consider noiseless i.i.d. setting where $\Sigma_x = \Sigma_\beta = I_d$ and $\sigma = 0$, with results presented in red (attention) and blue (H3) solid curves. **(b):** We conduct noisy label experiments by choosing $\sigma \neq 0$. **(c):** Consider non-isotropic task by setting $\Sigma_\beta = \gamma \mathbf{1}\mathbf{1}^\top + (1 - \gamma)I_d$. Solid and dashed curves in (b) and (c) represent attention and H3 results, respectively. The alignments in (a), (b) and (c) show the equivalence between attention and H3, validating Theorem 4 and Proposition 3. More experimental details are discussed in Section D.

Appendix D. Experiments

In this section, we conduct synthetic experiments to support our theoretical findings and further explore the behavior of different models of interest under different conditions. The experiments are designed to investigate various scenarios, including independent data, retrieval-augmented generation (RAG), task-feature alignment, low-rank parameterization, and LoRA adaption.

Experimental setting. We train 1-layer attention and H3 models for solving the linear regression ICL. As described in Section 2, we consider meta-learning setting where task parameter β is randomly generated for each training sequence. In all experiments, we set the dimension $d = 20$. Depending on the in-context length (n), different models are trained to make in-context predictions. We train each model for 10000 iterations with batch size 128 and Adam optimizer with learning rate 10^{-3}. Since our study focuses on the optimization landscape, and experiments are implemented via gradient descent, we repeat 20 model trainings from different initialization and results are presented as the minimal test risk among those 20 trails. In all the plots, theoretical predictions are obtained via the corresponding formulae presented in Section 3 and the test risks are normalized by the dimension d.

• **Equivalence among $\mathcal{L}_{\text{PGD}}^\star$, $\mathcal{L}_{\text{ATT}}^\star$ and $\mathcal{L}_{\text{SSM}}^\star$ (Figure 2).** To verify Proposition 3 as well as Theorem 4, we run random linear regression instances where in-context samples are generated obeying (6). Fig. 2(a)subfigure is identical to Fig. 1(a)subfigure where we set $\Sigma_x = \Sigma_\beta = I_d$ and $\sigma = 0$. In Fig. 2(b)subfigure, set $\Sigma_x = \Sigma_\beta = I$ and vary noise level σ^2 from 0 to $0.3 \times d$. In Fig. 2(c)subfigure, we consider noiseless labels, $\sigma = 0$, isotropic feature distribution $\Sigma_x = I_d$ and set task covariance to be $\Sigma_\beta = \gamma \mathbf{1}\mathbf{1}^\top + (1 - \gamma)I_d$ by choosing γ in $\{0, 0.3, 0.6, 0.9\}$. Note that in Fig. 2(c)subfigure, we train a sufficient number of models (greater than 20) to ensure the optimal model is obtained. In all the figures, solid and dashed curves correspond to the ICL results from training 1-layer ATT and SSM models, respectively, and dotted curves are obtained from (7) in Theorem 4. The alignment of solid, dashed and dotted curves validates our Proposition 3 and Theorem 4.

• **Distributional alignment experiments (Figs. 3(a)subfigure&3(b)subfigure).** In Figs. 3(a)subfigure and 3(b)subfigure, we generate RAG and task-feature alignment data following (8) and (10), re-

(a) RAG   (b) Task-feature align.   (c) Low-rank attention   (d) LoRA adaptation
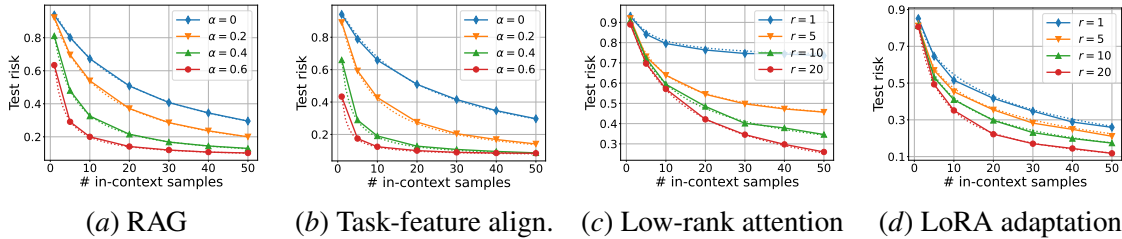
Figure 3: Distributional alignment and low-rank parameterization experiments. **(a)** and **(b)** show the ICL results using data generated via (8) and (10), respectively, by changing α from 0 to 0.6. In **(c)**, we train low-rank linear attention models by setting $W_k, W_q \in \mathbb{R}^{(d+1)\times r}$ and in **(d)**, we apply the low-rank LoRA adaptor, $W_{lora} := W_{up}W_{down}^\top$ where $W_{up}, W_{down} \in \mathbb{R}^{(d+1)\times r}$, to pretrained linear attention models and adjust the LoRA parameters under different task distribution. Solid and dotted curves correspond to the linear attention and theoretical results (c.f. Section 3), respectively, and the alignments validate our theorems in Section 3. More experimental details are discussed in Section D.

spectively, by setting $\sigma = 0$ and varying α from 0 to 0.6. Attention training results are displayed in solid curves, and we generate theory curve (dotted) via the $\mathcal{L}_\star$ formula as described in (35) in Appendix B.3 and (41) in Appendix B.4. The empirical alignments corroborate Theorems 13 and 14, further confirming that Proposition 3 is applicable to a broader range of real-world distributional alignment data.

• **Low-rank (Fig. 3(c)subfigure) and LoRA (Fig. 3(d)subfigure) experiments.** We also run simulations to verify our theoretical findings in Section 3.2. Consider the independent data setting as described in (6). In Fig. 3(c)subfigure, we set $\Sigma_x = I_d$, $\sigma = 0$ and task covariance to be diagonal with diagonal entries $c[1 \ 2^{-1} \ \cdots \ d^{-1}]^\top$ for some normalization constant $c = d/\sum_{i=1}^d i^{-1}$, and parameterize the attention model using matrices $W_k, W_q \in \mathbb{R}^{(d+1)\times r}$ and vary r across the set $\{1, 5, 10, 20\}$. Results show that empirical (solid) and theoretical (dotted, c.f. (12)) curves overlap. In Fig. 3(d)subfigure, we implement two phases of training. *Phase 1:* Setting $\Sigma_x = \Sigma_\beta = I_d$ and $\sigma = 0$, we pretrain the model with full rank parameters and obtain weights $\hat{W}_k, \hat{W}_q, \hat{W}_v \in \mathbb{R}^{(d+1)\times(d+1)}$. *Phase 2:* We generate new examples with task covariance Σ_β being a diagonal matrix with diagonal entries $c'[2^{-1} \ 2^{-2} \ \cdots \ 2^{-d}]^\top$ for some normalization constant $c' = d/\sum_{i=1}^d 2^{-i}$. Given the rank restriction r, we train additional LoRA parameters $W_{up}, W_{down} \in \mathbb{R}^{(d+1)\times r}$ where $W_{lora} := W_{up}W_{down}^\top$ and (2a) becomes $\mathtt{ATT}(Z) = (Z(\hat{W}_q\hat{W}_k^\top + W_{up}W_{down}^\top)Z^\top)Z\hat{W}_v$. Fig. 3(d)subfigure presents the results after two phases of training where dotted curves are drawn from the right hand side of (13) directly. Here, note that since Σ, Σ^{new} are diagonal, the right hand side of (13) returns the exact optimal risk of LoRA and the alignments verify it.

• **H3 outperforms linear attention (Figure 4).** Until now, our analysis has established the equivalence between linear attention and H3 models in solving linear ICL problem. Furthermore, we also investigate settings where H3 could outperform linear attention due to its sample weighting ability. In Figs. 4(a)subfigure and 4(b)subfigure, instead of training separate models to fit the different context lengths, we train a single model with fixed max-length n_{max} and loss is evaluated as the average loss given samples from 1 to n_{max}. Such setting has been wildly studied in the previous ICL work [2, 11, 23]. We generate data according to (6) with $\Sigma_x = \Sigma_\beta = I_d$ and $\sigma = 0$, and train 1-layer linear attention (Fig. 4(a)subfigure) and H3 (Fig. 4(b)subfigure) models with different max-lengths $n_{max} = 30, 50, 80$. Comparison between Fig. 4(a)subfigure and 4(b)subfigure shows that
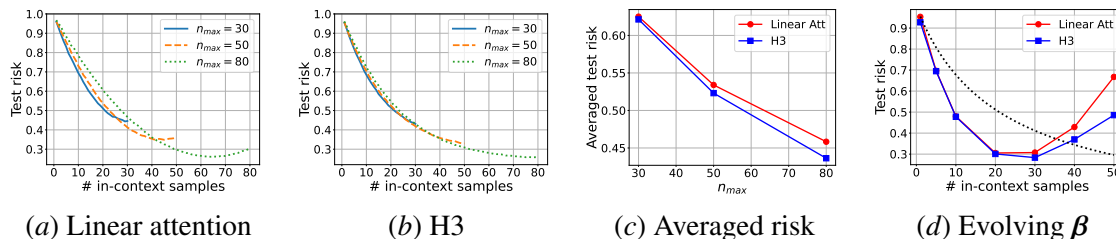
(a) Linear attention   (b) H3   (c) Averaged risk   (d) Evolving $\boldsymbol{\beta}$

Figure 4: Further comparison for linear attention and H3. In **(a)** and **(b)**, given maximum context lengths $n_{\max}$, we train linear attention and H3 models to minimize the average loss across all positions n from 1 to $n_{\max}$. Averaged test risks are presented in **(c)**. In **(d)**, the task vector $\boldsymbol{\beta}$ evolves gradually over the context positions $i \leq n$ via $\boldsymbol{\beta}_i = (i/n)\boldsymbol{\beta}_1 + (1 - i/n)\boldsymbol{\beta}_2$. In both scenarios, H3 outperforms linear attention benefiting from its additional convolutional filter (c.f. $\boldsymbol{f}$ in (2b)). More experimental details are discussed in Section D.

1-layer attention and H3 implement different algorithms in solving the averaged linear regression problem and H3 is more consistent in generalizing to longer context lengths. In Fig. 4(c)subfigure, we plot the averaged risks for each model and H3 outperforms linear attention. Furthermore, in Fig. 4(d)subfigure, we focus on the setting where in-context examples are generated using evolving task vector $\boldsymbol{\beta}$. Specifically, consider that each sequence corresponds to two individual task parameters $\boldsymbol{\beta}_1 \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $\boldsymbol{\beta}_2 \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Then the i'th sample is generated via $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $y_i = \boldsymbol{\beta}_i^\top \boldsymbol{x}_i$ where $\boldsymbol{\beta}_i = \lambda_i \boldsymbol{\beta}_1 + (1 - \lambda_i)\boldsymbol{\beta}_2$ and $\lambda_i = i/n$. The results are reported in Fig. 4(d)subfigure which again shows that H3 achieves better performance compared to linear attention, as H3 may benefit from the additional convolutional filter (c.f. $\boldsymbol{f}$ in (2b)). Here, dotted curve represent the theoretical results under i.i.d. and noiseless setting, derived from Corollary 5.

Appendix E. Related Work

There is growing interest in understanding the mechanisms behind ICL [5, 27, 37] in large language models (LLMs) due to its success in continuously enabling novel applications for LLMs [13, 35, 39]. Towards this, Xie et al. [44] explain ICL by language model's ability to perform implicit Bayesian inference where, under specific assumptions on the pre-training data distribution, the model infers a shared latent concept among the in-context examples and leverages the concept to make a prediction. Hollmann et al. [19], Müller et al. [31, 32] introduce prior-data fitted network (PFN) to approximate Bayesian inference on synthetic datasets and use it to perform downstream tasks such as tabular dataset classification. On the other hand, Olsson et al. [34] posit induction heads as the key mechanism enabling ICL in Transformers. Park et al. [36] study how various distributional properties of training data aid in the emergence of ICL in Transformers.

In the previous work, Garg et al. [11] explored ICL ability of Transformers. In particular, they considered in-context prompts where each in-context example is labeled by a target function from a given function class, including linear models. A number of works have studied this and related settings to develop a theoretical understanding of ICL [2, 3, 6, 8, 12, 24, 25, 41, 45]. Akyürek et al. [2] focus on linear regression and provide a construction of Transformer weights that can enable a single step of GD based on in-context examples. They further show that Transformers trained on in-context prompts exhibit behaviors similar to the models recovered via explicit learning algorithm on the in-context examples in a prompt. Along the similar line, Von Oswald et al. [40]

provide a construction of weights in linear attention-only Transformers that can emulate GD steps on in-context examples for a linear regression task. Interestingly, they find similarity between their constructed networks and the networks resulting from training on in-context prompts corresponding to linear regression tasks. Similar to this line of work, Dai et al. [7] argue that pre-trained language models act as meta-optimizer which utilize attention to apply meta-gradients to the original language model based on the in-context examples. Focusing on various NLP tasks, they further connect it to a specific form of explicit fine-tuning that performs gradient updates to the attention-related parameters. Inspired by the connection between linear attention and GD, they developed a novel attention mechanism that mirrors the behavior of GD with momentum. Beyond Transformers, existing work [14, 21, 47] demonstrate that other model architectures, such as SSM and RNNs, are also capable of in-context learning (ICL).

Building on these primarily empirical studies, Ahn et al. [1], Duraisamy [9], Mahankali et al. [29], Zhang et al. [46] focus on developing a theoretical understanding of Transformers trained to perform ICL. For single-layer linear attention model trained on in-context prompts for random linear regression tasks with isotropic Gaussian features and isotropic Gaussian weight vectors, Ahn et al. [1], Mahankali et al. [29] show that the resulting model implements a single step of GD on in-context examples in a test prompt, thereby corroborating the findings of [40]. They also show that the learned model implements a PGD step, when faced with anisotropic Gaussian features, with Mahankali et al. [29] also considering anisotropic Gaussian weight vectors. Ahn et al. [1] further study multi-layer model and show that the trained model can implement a generalization of GD++ algorithm, supporting an empirical observation in [40]. On the other hand, Mahankali et al. [29] extend their single-layer setup to consider suitable non-linear target functions, showing that learned Transformer again implements a single step of GD on lineare regression objective. For a single-layer linear attention model, Zhang et al. [46] study the optimization dynamics of gradient flow while training such a model on in-context prompts for random linear regression tasks. Despite the non-convexity of the underlying problem, they show the convergence to the global minimum of the population objective. Similar to Ahn et al. [1], Mahankali et al. [29], they show that the trained model implements a single step of GD and PGD for isotropic and anisotropic Gaussian features, respectively. In addition, they also characterize the test-time prediction error for the trained model while highlighting its dependence on train and test prompt lengths. Interestingly, Zhang et al. [46] further explore the effect of various distributional shifts, including the shift in task weight vector distributions between train and test time as well as the covariate shifts among train and test in-context prompts. Interestingly, they find that while linear-attention models are robust to most shifts, they exhibit brittleness to the covariate shifts.

While our work shares similarities with this line of works, as discussed in our contributions in the introduction, we expand the theoretical understanding of ICL along multiple novel dimensions, which includes the first study of LoRA adaptation for ICL in the presence of a distributional shift. Furthermore, we strive to capture the effect of retrieval augmentation [22, 33] on ICL through our analysis. Retrieval augmentation allows for selecting most relevant demonstration out of a large collection for a test instance, e.g., via a dense retrieval model [20], which can significantly outperform the typical ICL setup where fixed task-specific demonstrations are provided as in-context examples [4, 42]. Through a careful modeling of retrieval augmentation via correlated design, we show that it indeed has a desirable amplification effect where the effective number in-context examples becomes larger with higher correlation which corresponds to preforming a successful retrieval of query-relevant demonstrations in a practical retrieval augmented setup.

Recently, state space models (SSMs) [10, 15–17] have appeared as potential alternatives to Transformer architecture, with more efficient scaling to input sequence length. Recent studies demonstrate that such SSMs can also perform ICL for simple non-language tasks [14, 36] as well as complex NLP tasks [14]. That said, a rigorous theoretical understanding of ICL for SSMs akin to [1, 29, 46] is missing from the literature. In this work, we provide the first such theoretical treatment for ICL with SSMs. Focusing on H3 architecture [10], we highlight its advantages over linear attention in specific ICL settings.

Appendix F. Discussion

In this work, we revisited the loss landscape of in-context learning with 1-layer sequence models. We have established a general connection between ICL and gradient methods that accounts for correlated data, non-attention architectures (specifically SSMs), and the impact of low-rank parameterization including LoRA adaptation. Our results elucidate two central findings: (1) The functions learned by different sequence model architectures exhibit a strong degree of *universality* and (ii) *Dataset and prompt design*, such as RAG, can substantially benefit ICL performance.

Future directions and limitations. The results of this work fall short of being a comprehensive theory for ICL in LLMs and can be augmented in multiple directions. First, while the exact equivalence between H3 and linear attention is remarkable, we should examine whether it extends to other SSMs. Secondly, while empirically predictive, our RAG and LoRA analyses are not precise and fully formal. Thirdly, it is desirable to develop a deeper understanding of multilayer architectures and connect to iterative GD methods as in [1, 40]. Finally, we have studied the population risk of ICL training whereas one can also explore the sample complexity of pretraining [28, 43]. Moving beyond the theoretically tractable setup of this work, our simplified models are trained on in-context prompts from random initialization. Therefore, this theoretical study doesn't address more challenging in-context learning tasks, such as question answering, where both in-context demonstration and general knowledge from pretraining are required. Future work in this area could also shed light on how certain contexts might elicit undesirable behaviors acquired by an LLM during pretraining, an aspect not covered in our current analysis. This work also studies a theoretical model for retrieval augmentation-based ICL. In a real-life retrieval augmentation-based ICL, one needs to account for the quality of the collection of the retrievable demonstrations and its (negative) impacts on the final predictions.