Fine-grained Analysis of In-context Linear Estimation: Data, Architecture, and Beyond

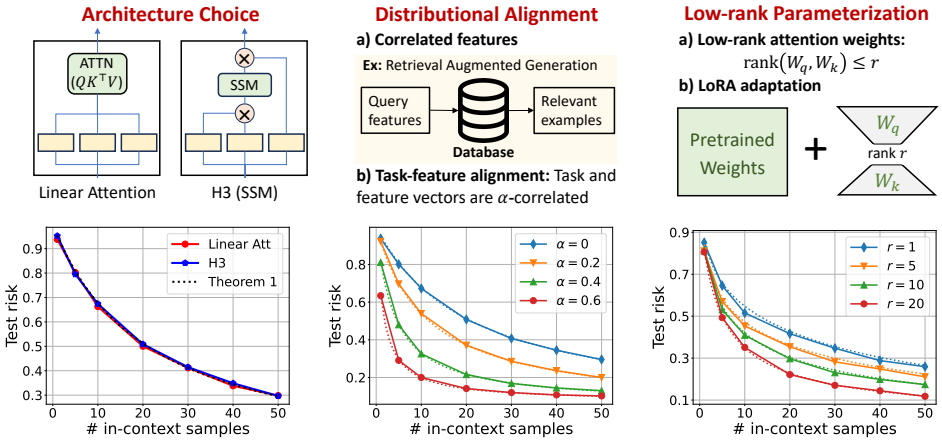
Yingcong Li
University of Michigan
yingcong@umich.edu

Ankit Singh Rawat
Google Research NYC
ankitsrawat@google.com

Samet Oymak
University of Michigan
oymak@umich.edu

Abstract

Recent research has shown that Transformers with linear attention are capable of in-context learning (ICL) by implementing a linear estimator through gradient descent steps. However, the existing results on the optimization landscape apply under stylized settings where task and feature vectors are assumed to be IID and the attention weights are fully parameterized. 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. We show that thanks to its native convolution filters, H3 also has the advantage of implementing sample weighting and outperforming linear attention in suitable settings. (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. Experimental results corroborate our theoretical findings. Overall, this work explores the optimization and risk landscape of ICL in practically meaningful settings and contributes to a more thorough understanding of its mechanics.



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

Figure 1: 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 deferred to Section 4.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1 Introduction

Modern language models exhibit the remarkable ability to learn novel tasks or solve complex problems from the demonstrations provided within their context window [Brown et al., 2020, GeminiTeam et al., 2023, OpenAI, 2023, Touvron et al., 2023]. Such *in-context learning* (ICL) offers a novel and effective alternative to traditional fine-tuning techniques and has become an important feature of LLM with its applications spanning retrieval-augmented generation [Lewis et al., 2020], and reasoning via advanced prompting techniques, such as chain-of-thought [Wei et al., 2022].

ICL ability presents an important research avenue to develop stronger theoretical and mechanistic understanding of large language models. To this aim, there has been significant recent interest in demystifying ICL through the lens of function approximation [Liu et al., 2023a], Bayesian inference [Müller et al., 2021, Xie et al., 2022, Han et al., 2023], and learning and optimization theory [Ahn et al., 2023, Mahankali et al., 2024, Zhang et al., 2024, Duraisamy, 2024]. The latter is concerned with understanding the optimization landscape of ICL, which is also crucial for understanding the generalization properties of the model. A notable result in this direction is the observation that linear attention models [Schlag et al., 2021, Von Oswald et al., 2023, Ahn et al., 2023] implement *preconditioned gradient descent* (PGD) during ICL [Ahn et al., 2023, Mahdavi et al., 2024]. While this line of works provide a fresh perspective to ICL, the existing studies do not address many questions arising from real-life applications nor provide guiding principles for various ICL setups motivated by practical considerations.

To this aim, we revisit the theoretical exploration of ICL with linear data model where we feed an in-context prompt containing n examples $(\boldsymbol{x}_i, y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \xi_i)_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ and a test instance or query $\boldsymbol{x}_{n+1} \in \mathbb{R}^d$ to the model, with d being the feature dimension, $\boldsymbol{\beta} \in \mathbb{R}^d$ being the task weight vector, and $(\xi_i)_{i=1}^n$ denoting the noise in individual labels. Given the in-context prompt, the model is tasked to predict $\hat{y}_{n+1}$ – an estimate for $y_{n+1} = \boldsymbol{x}_{n+1}^\top \boldsymbol{\beta} + \xi_{n+1}$. We aim to provide answers to the following questions by exploring the loss landscape of ICL:

(Q1) Is the ability to implement gradient-based ICL unique to (linear) attention? Can alternative sequence models implement richer algorithms beyond PGD?

(Q2) In language modeling, ICL often works well with few-shot samples whereas standard linear estimation typically requires $O(d)$ samples. How can we reconcile this discrepancy between classical learning and ICL?

(Q3) To our knowledge, existing works assume linear-attention is fully parameterized, i.e., key and query projections $\boldsymbol{W}_k, \boldsymbol{W}_q \in \mathbb{R}^{d \times d}$. What happens when they are low-rank? What happens when there is distribution shift between training and test in-context prompts and we use LoRA [Hu et al., 2022] for adaptation?

In this work, we conduct a careful investigation of these questions. Specifically, we focus on ICL with 1-layer models and make the following contributions:

(A1) We jointly investigate the landscape of linear attention and H3 [Fu et al., 2023], a widely popular state-space model (SSM). We prove that under correlated design, both models implement 1-step PGD (c.f. Proposition 1) and the alignments in Fig. 1a verify that where the dotted curve represents the theoretical PGD result derived from Theorem 1. Our analysis reveals that the gating mechanism in H3 imitates attention. We also empirically show that H3 has the advantage of implementing sample-weighting which allows it to outperform linear attention in temporally-heterogeneous problem settings in Appendix D.

(A2) Proposition 1 allows for task and features to be correlated to each other as long as odd moments are zero. Through this, we can 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 2 and Fig. 1b) and *Task-Feature Alignment* (c.f. Theorem 3), 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.

(A3) 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. (14) and Fig. 1c), and show that, these bounds accurately predict the empirical performance.

2 Problem Setup and Preliminaries

We begin with a short note on notation. Let bold lowercase and uppercase letters (e.g., $\boldsymbol{x}$ and $\boldsymbol{X}$) represent vectors and matrices, respectively. The symbol $\odot$ is defined as the element-wise (Hadamard) product, and $*$ denotes the convolution operator. $\mathbf{1}_d$ and $\mathbf{0}_d$ denote the d-dimensional all-ones and all-zeros vectors, respectively; and $\boldsymbol{I}_d$ denotes the identity matrix of dimension $d \times d$. Additionally, let $\mathtt{tr}(\boldsymbol{W})$ denote the trace of the square matrix $\boldsymbol{W}$.

As mentioned earlier, we study the optimization landscapes of 1-layer linear attention [Katharopoulos et al., 2020, Schlag et al., 2021] and H3 [Fu et al., 2023] models when training with prompts containing in-context data following a linear model. We construct the input in-context prompt similar to Ahn et al. [2023], Mahankali et al. [2024], Zhang et al. [2024] as follows.

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

$$\boldsymbol{Z} = [z_1 \ \dots \ z_n \ z_{n+1}]^\top = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n & \boldsymbol{x}_{n+1} \\ y_1 & \dots & y_n & 0 \end{bmatrix}^\top \in \mathbb{R}^{(n+1)\times(d+1)}. \tag{1}$$

Here, we set $z_i = \begin{bmatrix} \boldsymbol{x}_i \\ y_i \end{bmatrix}$ for $i \le n$ and the last/query token $z_{n+1} = \begin{bmatrix} \boldsymbol{x}_{n+1} \\ 0 \end{bmatrix}$. Then, given $\boldsymbol{Z}$, the goal of the model is to predict the correct label y_{n+1} corresponding to $\boldsymbol{x}_{n+1}$. For cleaner notation, when it is clear from context, we drop the subscript $n+1$ and set $\boldsymbol{x} = \boldsymbol{x}_{n+1}$, $z = z_{n+1}$. Different from the previous work [Ahn et al., 2023, Mahankali et al., 2024, Zhang et al., 2024, Mahdavi et al., 2024] where $(\boldsymbol{x}_i)_{i=1}^{n+1}$ and $\boldsymbol{\beta}$ are assumed to be independent, our analysis focuses on a more general linear setting that captures the dependency between $(\boldsymbol{x}_i)_{i=1}^{n+1}$ and $\boldsymbol{\beta}$.

Model architectures. To start with, we first review the architectures of both Transformer and state-space model (SSM). Similar to the previous work [Von Oswald et al., 2023, Ahn et al., 2023, Mahankali et al., 2024, Zhang et al., 2024] and to simplify the model structure, we focus on single-layer models and omit the nonlinearity, e.g., softmax operation and MLP activation, from the Transformer. Given the input prompt $\boldsymbol{Z} \in \mathbb{R}^{(n+1)\times(d+1)}$ in (1), which can be treated as a sequence of $(d+1)$-dimensional tokens, the single-layer linear attention $\mathtt{ATT}$ and H3-like single-layer SSM $\mathtt{SSM}$ are denoted by

$$\mathtt{ATT}(\boldsymbol{Z}) = (\boldsymbol{Z}\boldsymbol{W}_q\boldsymbol{W}_k^\top\boldsymbol{Z}^\top)\boldsymbol{Z}\boldsymbol{W}_v \tag{2a}$$

$$\mathtt{SSM}(\boldsymbol{Z}) = \Big((\boldsymbol{Z}\boldsymbol{W}_q) \odot ((\boldsymbol{Z}\boldsymbol{W}_k \odot \boldsymbol{Z}\boldsymbol{W}_v) * \boldsymbol{f})\Big) \tag{2b}$$

where $\boldsymbol{W}_k,\ \boldsymbol{W}_q,\ \boldsymbol{W}_v \in \mathbb{R}^{(d+1)\times(d+1)}$ denote the key, query and value weight matrices, respectively. In (2b), the parameter $\boldsymbol{f} \in \mathbb{R}^{n+1}$ is a 1-D convolutional filter that mixes tokens. The Hadamard product $\odot$ is the gating mechanism [Dauphin et al., 2017] between key and query channels, which is crucial for attention-like feature creation. Thus, (2b) is more generally a gated-convolution layer. For $\boldsymbol{f}$ only, we use indexing $\boldsymbol{f} = [f_0 \ \dots \ f_n]^\top \in \mathbb{R}^{n+1}$ and given any vector $\boldsymbol{a}$, denote convolution output $(\boldsymbol{a} * \boldsymbol{f})_i = \sum_{j=1}^{i} f_{i-j} a_j$. Note that our notation slightly differs from the original H3 model [Fu et al., 2023] in two ways:

1. SSMs provide efficient parameterization of $\boldsymbol{f}$ which would otherwise grow with sequence length. In essence, H3 utilizes a linear state-space model $\boldsymbol{s}_i = \boldsymbol{A}\boldsymbol{s}_{i-1} + \boldsymbol{B}u_i$ and $y_i = \boldsymbol{C}\boldsymbol{s}_i$ with parameters $(\boldsymbol{A} \in \mathbb{R}^{d\times d}, \boldsymbol{B} \in \mathbb{R}^{d\times 1}, \boldsymbol{C} \in \mathbb{R}^{1\times d})$ from which the filter $\boldsymbol{f}$ is obtained via the impulse response $f_i = \boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B}$ for $i \ge 0$. Here d is the state dimension and, in practice, $\boldsymbol{A}$ is chosen to be diagonal. Observe that, setting $d = 1$ and $\boldsymbol{A} = \rho, \boldsymbol{C} = \boldsymbol{B} = 1$, SSM reduces to the exponential smoothing $f_i = \rho^i$ for $i \ge 0$. Thus, H3 also captures the all-ones filter as a special instance. As we show in Proposition 1, this simple filter is optimal under independent data model and exactly imitates linear attention. Note that, utilizing a filter $\boldsymbol{f}$ as in (2b) is strictly more expressive than the SSM as it captures all possible impulse responses.
2. H3 also applies a shift SSM to the key embeddings to enable the retrieval of the local context around associative recall hits. We opted not to incorporate this shift operator in our model. This is because unless the features of the neighboring tokens are correlated (which is not the case for

the typical independent data model), the entry-wise products between values and shifted keys will have zero mean and be redundant for the final prediction.

We note that we conduct all empirical evaluations with the original H3 model, which displays exact agreement with our theory formalized for (6b), further validating our modeling choice.

2.1 In-context Linear Estimation

We will next study the algorithms that can be implemented by the single-layer attention and state-space models. Through this, we will show that training ATT and SSM with linear ICL data is equivalent to the prediction obtained from one step of optimally-*preconditioned gradient descent* (PGD) and *sample-weighted preconditioned gradient descent* (WPGD), respectively. We will further show that under mild assumption, the optimal sample weighting for SSM (e.g., $\boldsymbol{f}$) is an all-ones vector and therefore, establishing the equivalence among PGD, 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 $(\boldsymbol{x}_i, y_i)_{i=1}^n$, define

$$\boldsymbol{X} = [\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d} \quad \text{and} \quad \boldsymbol{y} = [y_1 \ \cdots \ y_n]^\top \in \mathbb{R}^n.$$

Starting from $\boldsymbol{\beta}_0 = \boldsymbol{0}_d$ and letting $\eta = 1/2$ be the step size, a single-step GD preconditioned with weights $\boldsymbol{W}$ returns prediction

$$\hat{y} = \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top \boldsymbol{y} := g_{\texttt{PGD}}(\boldsymbol{Z}), \tag{3}$$

and a single-step *sample-weighted* GD given weights $\boldsymbol{\omega} \in \mathbb{R}^n$ and $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ returns prediction

$$\hat{y} = \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top (\boldsymbol{\omega} \odot \boldsymbol{y}) := g_{\texttt{WPGD}}(\boldsymbol{Z}), \tag{4}$$

where $\boldsymbol{Z}$ is defined in (1) consisting of $\boldsymbol{X}, \boldsymbol{y}$ and $\boldsymbol{x}$. Our goal is to find the optimal $\boldsymbol{W}$, as well as $\boldsymbol{\omega}$ in (4) that minimize the population risks defined as follows.

$$\min_{\boldsymbol{W}} \mathcal{L}_{\texttt{PGD}}(\mathcal{W}) \quad \text{where} \quad \mathcal{L}_{\texttt{PGD}}(\mathcal{W}) = \mathbb{E}\left[(y - g_{\texttt{PGD}}(\boldsymbol{Z}))^2\right], \tag{5a}$$

$$\min_{\boldsymbol{W},\boldsymbol{\omega}} \mathcal{L}_{\texttt{WPGD}}(\mathcal{W}) \quad \text{where} \quad \mathcal{L}_{\texttt{WPGD}}(\mathcal{W}) = \mathbb{E}\left[(y - g_{\texttt{WPGD}}(\boldsymbol{Z}))^2\right]. \tag{5b}$$

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

As per (2), given input prompt $\boldsymbol{Z} \in \mathbb{R}^{(n+1)\times(d+1)}$, either of the underlying models outputs a $(n+1)$-length sequence. Note that the label for the query $\boldsymbol{x} = \boldsymbol{x}_{n+1}$ is excluded from the prompt $\boldsymbol{Z}$. Similar to Ahn et al. [2023], Mahankali et al. [2024], we consider a training objective with a causal mask to ensure inputs cannot attend to their own labels and training can be parallelized. Let $\boldsymbol{Z}_0 = [z_1 \ \ldots \ z_n \ 0]^\top$ be the features post-causal masking at time/index $n+1$. Given weights $\boldsymbol{W}_k, \boldsymbol{W}_q, \boldsymbol{W}_v$ and the filter $\boldsymbol{f}$ for SSM, predictions at the query token $z = \begin{bmatrix} \boldsymbol{x} \\ 0 \end{bmatrix}$ take the following forms following sequence-to-sequence mappings in (2):

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

where $\boldsymbol{v} \in \mathbb{R}^{d+1}$ is the linear prediction head and $((\boldsymbol{Z}_0 \boldsymbol{W}_k \odot \boldsymbol{Z}_0 \boldsymbol{W}_v) * \boldsymbol{f})_{n+1}$ returns the last row of the convolution output. Note that SSM can implement the mask by setting $f_0 = 0$. Now consider the meta learning setting and select loss function to be the squared loss, same as in (5). Thus, the objectives for both models take the following forms.

$$\min_{\boldsymbol{W}_k,\boldsymbol{W}_q,\boldsymbol{W}_v,\boldsymbol{v}} \mathcal{L}_{\texttt{ATT}}(\mathcal{W}) \quad \text{where} \quad \mathcal{L}_{\texttt{ATT}}(\mathcal{W}) = \mathbb{E}\left[(y - g_{\texttt{ATT}}(\boldsymbol{Z}))^2\right], \tag{6a}$$

$$\min_{\boldsymbol{W}_k,\boldsymbol{W}_q,\boldsymbol{W}_v,\boldsymbol{v},\boldsymbol{f}} \mathcal{L}_{\texttt{SSM}}(\mathcal{W}) \quad \text{where} \quad \mathcal{L}_{\texttt{SSM}}(\mathcal{W}) = \mathbb{E}\left[(y - g_{\texttt{SSM}}(\boldsymbol{Z}))^2\right]. \tag{6b}$$

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

Note that for all the optimization methods (c.f. (5), (6)), 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. For example, $\mathbb{E}[\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{\beta}]$ is a sum of cross-moments of order 2. Then, it motivates the following data assumptions.

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

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

Note that compared to Ahn et al. [2023], Mahankali et al. [2024], Zhang et al. [2024], Assumption 1 is more general which also subsumes the dependent distribution settings. In this work, we consider the following three linear models (omitting noise) satisfying Assumption 1. Let $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{x}} \in \mathbb{R}^{d\times d}$ represent the task and feature covariance matrices for independent data, and let $0 \le \alpha \le 1$ be the correlation level when considering data dependency. More specific discussions are deferred to Section 3.

- Independent task and data: $\boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),\ \boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{x}}),$ for all $1 \le i \le n+1$.
- Retrieval augmented generation: $\boldsymbol{\beta}, \boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d),\ \boldsymbol{x}_i \,\big|\, \boldsymbol{x} \sim \mathcal{N}(\alpha\boldsymbol{x}, (1-\alpha^2)\boldsymbol{I}_d),$ for all $1 \le i \le n$.
- Task-feature alignment: $\boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{I}_d),\ \boldsymbol{x}_i \,\big|\, \boldsymbol{\beta} \sim \mathcal{N}(\alpha\boldsymbol{\beta}, \boldsymbol{I}_d),$ for all $1 \le i \le n+1$.

Next, we introduce the following result which establishes the equivalence among optimizing 1-layer linear attention (c.f. (6a)), 1-layer H3 (c.f. (6b)), and 1-step gradient descent (c.f. (5)).

Proposition 1 *Suppose Assumptions 1 and 2 hold. Consider the objectives as defined in (5) and (6), and let $\mathcal{L}^\star_{PGD}$, $\mathcal{L}^\star_{WPGD}$, $\mathcal{L}^\star_{ATT}$, and $\mathcal{L}^\star_{SSM}$ be their optimal risks, respectively. Then,*

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

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

We defer the proof to Appendix A.1. Proposition 1 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 $(\boldsymbol{x}_i, y_i)_{i=1}^n$ are independently sampled given $\boldsymbol{x}$ and $\boldsymbol{\beta}$, and we therefore conclude that $\mathcal{L}^\star_{\mathrm{PGD}} = \mathcal{L}^\star_{\mathrm{ATT}} = \mathcal{L}^\star_{\mathrm{SSM}}$. Leveraging this result, the subsequent section of the paper concentrate on addressing (5a), taking into account various linear data distributions.

While Proposition 1 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 $\mathcal{L}_{\mathrm{PGD}}(\mathcal{W})$ and derive the subsequent lemmas.

Lemma 1 *Suppose Assumption 2 holds and let $\boldsymbol{\xi} = [\xi_1\ \xi_2\ \cdots\ \xi_n]^\top$. Then the loss $\mathcal{L}_{PGD}(\mathcal{W})$ in (5a) 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 $g^\star_{PGD}$, $g^\star_{ATT}$ be the optimal prediction functions of (5a) and (6a). Then under the conditions of Assumptions 1 and 2, and the strong convexity, $g^\star_{PGD} = g^\star_{ATT}$.*

Lemma 2 *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 $\mathcal{L}_{PGD}(\mathcal{W})$ in (5a) 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 1, we have $\mathcal{L}^\star_{\texttt{PGD}} = \mathcal{L}^\star_{\texttt{ATT}} = \mathcal{L}^\star_{\texttt{SSM}}$. Additionally, the second claim in Lemma 2 holds, and $\mathcal{L}_{\texttt{PGD}}(\mathcal{W})$ is strongly convex. Therefore, following Lemma 1, we have $g^\star_{\texttt{PGD}} = g^\star_{\texttt{ATT}}$. Thanks to the equivalence among PGD, ATT, and SSM, in the next section, we focus on the solution of objective (5a) under different scenarios, which will reflect the optimization landscapes of ATT and SSM models.

3 Main Results

In light of Proposition 1, optimizing a single layer linear-attention or H3 model is equivalent to solving the objective (5a). Therefore, in this section, we examine the properties of the one-step PGD in (5a). To this end, we consider multiple problem settings, including distinct data distributions and low-rank training. The latter refers to the scenario where the key and query matrices have rank restrictions, e.g., $\boldsymbol{W}_k, \boldsymbol{W}_q \in \mathbb{R}^{(d+1)\times r}$, as well as LoRA-tuning when adapting the model under distribution shift.

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 $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ be the covariance matrices of the input feature and task vectors, respectively, and $\sigma \geq 0$ be the noise level. We assume

$$\boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \quad \boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{x}}), \quad \xi_i \sim \mathcal{N}(0, \sigma^2), \quad 1 \leq i \leq n+1 \tag{7}$$

and the label is obtained via $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \xi_i$. Our following result characterizes the optimal solution of (5a). Note that the data generated from (7) satisfies the conditions in Proposition 1. Therefore, the same results can be applied to both linear-attention and H3 models.

Theorem 1 *Consider independent linear data as defined in (7), and suppose the covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ are full rank. Recap the objective from (5a) 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, let $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}$ and $M = \texttt{tr}(\boldsymbol{\Sigma}) + \sigma^2$. Then $\boldsymbol{W}_\star$ and $\mathcal{L}_\star$ satisfy*

$$\boldsymbol{W}_\star = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1/2}\bar{\boldsymbol{W}}_\star\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1/2} \qquad \text{and} \qquad \mathcal{L}_\star = M - n\texttt{tr}\left(\boldsymbol{\Sigma}\bar{\boldsymbol{W}}_\star\right), \tag{8}$$

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

Corollary 1 *Consider noiseless i.i.d. linear data where $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \boldsymbol{I}_d$ and $\sigma = 0$. Then, the objective in (5a) returns*

$$\boldsymbol{W}_\star = \frac{1}{n+d+1}\boldsymbol{I}_d \qquad \text{and} \qquad \mathcal{L}_\star = d - \frac{nd}{n+d+1}.$$

See Appendix B.2 for proofs. Note that Theorem 1 is consistent with prior work [Ahn et al., 2023, Theorem 1] when specialized to isotropic task covariance, i.e., $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \boldsymbol{I}_d$. 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 RAG-inspired and task-feature alignment models, where Assumptions 1 and 2 continue to hold and Proposition 1 applies.

Retrieval augmented generation. To provide a statistical model of the practical RAG approaches, given the query vector $\boldsymbol{x}_{n+1} = \boldsymbol{x}$, we propose to draw ICL demonstrations that are similar to $\boldsymbol{x}$ with the same shared task vector $\boldsymbol{\beta}$. Modeling feature similarity through the cosine angle, RAG should sample the ICL examples $\boldsymbol{x}_i$, $i \leq n$, from the original feature distribution conditioned on the event $\cos(\boldsymbol{x}_i, \boldsymbol{x}) \geq \alpha$ where α is the similarity threshold. As an approximate proxy, under the Gaussian distribution model, we assume that $\boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and that RAG samples α-correlated demonstrations $(\boldsymbol{x}_i, y_i)_{i=1}^n$ as follows:

$$\boldsymbol{x}_i \,\big|\, \boldsymbol{x} \sim \mathcal{N}(\alpha\boldsymbol{x}, (1-\alpha^2)\boldsymbol{I}_d), \quad \xi_i \sim \mathcal{N}(0, \sigma^2) \quad \text{and} \quad y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + \xi_i, \quad 1 \leq i \leq n. \tag{9}$$

Note that the above normalization ensures that the marginal feature distribution remains $\mathcal{N}(0, \boldsymbol{I}_d)$. The full analysis of RAG is provides in Appendix B.3. Specifically, when we carry out the analysis by assuming $\alpha = \mathcal{O}\left(1/\sqrt{d}\right)$ and $d/n = \mathcal{O}(1)$ where $\mathcal{O}(\cdot)$ denotes proportionality, our derivation leads to the following result:

Theorem 2 *Consider linear model as defined in* (9)*. Recap the objective from* (5a) *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, let* $\kappa = \alpha^2 d + 1$ *and suppose* $\alpha = \mathcal{O}\left(1/\sqrt{d}\right)$*,* $d/n = \mathcal{O}(1)$ *and* d *is sufficiently large. Then* $\boldsymbol{W}_\star$ *and* $\mathcal{L}_\star$ *have approximate forms*

$$\boldsymbol{W}_\star \approx \frac{1}{\kappa n + d + \sigma^2}\boldsymbol{I}_d \qquad \text{and} \qquad \mathcal{L}_\star \approx d + \sigma^2 - \frac{\kappa n d}{\kappa n + d + \sigma^2}. \tag{10}$$

Here, (10) is reminiscent of Corollary 1 and has a surprisingly clean message. Observe that, $\alpha^2 d + 1$ is the dominant multiplier ahead of n in both equations. Thus, we deduce that, RAG model follows the same error bound as the independent data model, however, its sample size is amplified by a factor of $\alpha^2 d + 1$. $\alpha = 0$ reduces to the result of Corollary 1 whereas we need to set $\alpha = \mathcal{O}\left(1/\sqrt{d}\right)$ for constant amplification. When $\alpha = 1$, RAG achieves the approximate risk $\mathcal{L}_\star \approx 2 + \sigma^2$, where the constant bias is due to the higher order moments (e.g., the 4'th and 6'th moments) of the standard Gaussian distribution. As d increases, the normalized loss $\mathcal{L}_\star/d \to 0$. The full analysis of its optimal solution $\boldsymbol{W}_\star$ and loss $\mathcal{L}_\star$ are deferred Theorem 4 in Appendix B.3.

Task-feature alignment. We also consider another dependent data setting where task and feature vectors are assumed to be correlated. This dataset model has the following motivation: In general, an LLM can generate any token within the vocabulary. However, once we specify the task (e.g. domain of the prompt), the LLM output becomes more deterministic and there are much fewer token candidates. For instance, if the task is "Country", "France" is a viable output compared to "Helium" and vice versa when the task is "Chemistry". Formally speaking, this can be formalized as the input $\boldsymbol{x}$ having a diverse distribution whereas it becomes more predictable conditioned on $\boldsymbol{\beta}$. Therefore, it can be captured through a linear model by making the conditional covariance of $\boldsymbol{x} \,\big|\, \boldsymbol{\beta}$ to be approximately low-rank. This formalism can be viewed as a *spectral alignment* between input and task, which is also well-established in deep learning both empirically and theoretically [Li et al., 2020, Arora et al., 2019, Canatar et al., 2021, Cao et al., 2019]. Here, we consider such a setting where the shared task vector is sampled as standard Gaussian distribution $\boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ 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:

$$\boldsymbol{x}_i \,\big|\, \boldsymbol{\beta} \sim \mathcal{N}(\alpha\boldsymbol{\beta}, \boldsymbol{I}_d), \quad \xi_i \sim \mathcal{N}(0, \sigma^2) \quad \text{and} \quad y_i = \kappa^{-1/2}\boldsymbol{x}_i^\top \boldsymbol{\beta} + \xi_i, \quad 1 \le i \le n+1. \tag{11}$$

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 of its optimal solution $\boldsymbol{W}_\star$ and loss $\mathcal{L}_\star$ to Theorem 5 in Appendix B.4. Similar to the RAG setting, by assuming $\alpha = \mathcal{O}\left(1/\sqrt{d}\right)$ and $d/n = \mathcal{O}(1)$, we obtain the following results for the optimal parameter and risk.

Theorem 3 *Consider linear model as defined in* (11)*. Recap the objective from* (5a) *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* $\alpha = \mathcal{O}\left(1/\sqrt{d}\right)$*,* $d/n = \mathcal{O}(1)$ *and* d *is sufficiently large. Then* $\boldsymbol{W}_\star$ *and* $\mathcal{L}_\star$ *have approximate forms*

$$\boldsymbol{W}_\star \approx \frac{1}{\kappa n + (d + \sigma^2)/\kappa}\boldsymbol{I}_d \qquad \text{and} \qquad \mathcal{L}_\star \approx d + \sigma^2 - \frac{\kappa n d}{\kappa n + (d + \sigma^2)/\kappa}. \tag{12}$$

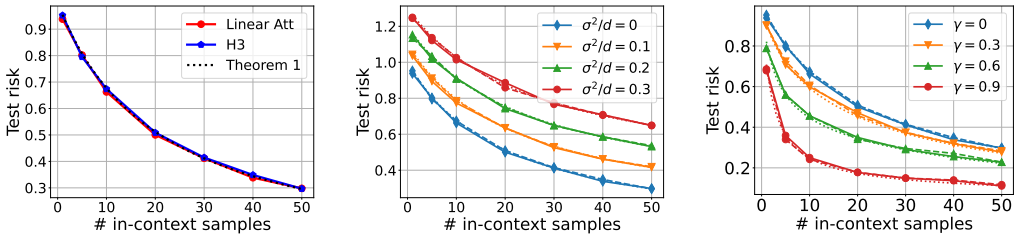
Similar to (10), (12) contains $\kappa = \alpha^2 + 1$ multiplier ahead of n, which reduces the in-context sample complexity and setting $\alpha = 0$ reduces to the results of Corollary 1.

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 (5a) under condition $\text{rank}(\boldsymbol{W}) = r$.

Lemma 3 *Consider independent linear data as defined in* (7)*. Recap the objective from* (5a) *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* $M = \texttt{tr}(\boldsymbol{\Sigma}) + \sigma^2$*. Denoting* λ_i *to be the i'th largest eigenvalue of* $\boldsymbol{\Sigma}$*, we have that*

$$\min_{\text{rank}(\boldsymbol{W})\le r, \boldsymbol{W}=\boldsymbol{W}^\top} \mathcal{L}(\boldsymbol{W}) = M - \sum_{i=1}^{r} \frac{n\lambda_i^2}{(n+1)\lambda_i + M}. \tag{13}$$



(a) $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \boldsymbol{I}_d$ and $\sigma = 0$ (b) Noisy label (c) Non-isotropic task

Figure 2: Empirical evidence validates Theorem 1 and Proposition 1. 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. (8). **(a):** We consider noiseless i.i.d. setting where $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \boldsymbol{I}_d$ 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 $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \gamma \mathbf{1}\mathbf{1}^\top + (1-\gamma)\boldsymbol{I}_d$. 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 1 and Proposition 1. More experimental details are discussed in Section 4.

Note that $\mathtt{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^d \lambda_i$. 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 (5a) with fixed $\boldsymbol{W}$. Suppose $\mathtt{tr}(\boldsymbol{\Sigma}) = \mathtt{tr}(\boldsymbol{\Sigma}^{new}) = M$, $\sigma = 0$ and $\boldsymbol{W}$ is jointly diagonalizable with $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{new}$, then LoRA's risk is upper-bounded by

$$\min_{\mathrm{rank}(\boldsymbol{W}_{lora}) \le r} \tilde{\mathcal{L}}(\boldsymbol{W}_{lora}) \le \min_{|\mathcal{I}| \le r, \mathcal{I} \subset [d]} \left(\sum_{i \notin \mathcal{I}} \frac{\lambda_i + M}{n + 1 + M/\lambda_i} + \sum_{i \in \mathcal{I}} \frac{\lambda_i^{new} + M}{n + 1 + M/\lambda_i^{new}} \right). \tag{14}$$

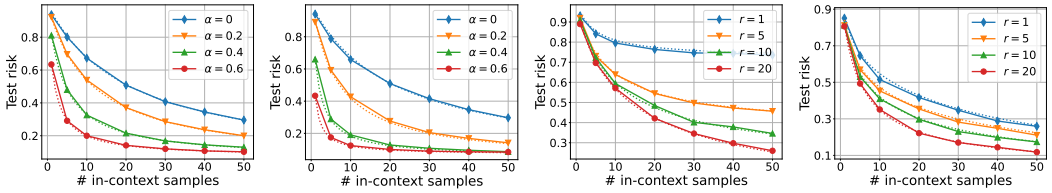
Note that, the right hand side is provided assuming the optimal LoRA-updated model $\boldsymbol{W}_{lora}$ is also jointly diagonalizable with covariances $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{new}$, and $\boldsymbol{W}$.

4 Experiments

We now 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 $\boldsymbol{\beta}$ 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 $\mathcal{L}^\star_{\mathtt{PGD}}$, $\mathcal{L}^\star_{\mathtt{ATT}}$ and $\mathcal{L}^\star_{\mathtt{SSM}}$ (Figure 2).** To verify Proposition 1 as well as Theorem 1, we run random linear regression instances where in-context samples are generated obeying (7). Fig. 2a is identical to Fig. 1a where we set $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \boldsymbol{I}_d$ and $\sigma = 0$. In Fig. 2b, set $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \boldsymbol{I}$ and vary noise level σ^2 from 0 to $0.3 \times d$. In Fig. 2c, we consider noiseless labels, $\sigma = 0$, isotropic feature distribution $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{I}_d$ and set task covariance to be $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \gamma \mathbf{1}\mathbf{1}^\top + (1-\gamma)\boldsymbol{I}_d$ by choosing γ in $\{0, 0.3, 0.6, 0.9\}$. Note that in Fig. 2c, 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



(a) RAG (b) Task-feature alignment (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 (9) and (11), respectively, by changing α from 0 to 0.6. In **(c)**, we train low-rank linear attention models by setting $\boldsymbol{W}_k, \boldsymbol{W}_q \in \mathbb{R}^{(d+1)\times r}$ and in **(d)**, we apply the low-rank LoRA adaptor, $\boldsymbol{W}_{lora} := \boldsymbol{W}_{\text{up}}\boldsymbol{W}_{\text{down}}^\top$ where $\boldsymbol{W}_{\text{up}}, \boldsymbol{W}_{\text{down}} \in \mathbb{R}^{(d+1)\times r}$, 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 4.

ICL results from training 1-layer `ATT` and `SSM` models, respectively, and dotted curves are obtained from (8) in Theorem 1. The alignment of solid, dashed and dotted curves validates our Proposition 1 and Theorem 1.

• **Distributional alignment experiments (Figs. 3a&3b).** In Figs. 3a and 3b, we generate RAG and task-feature alignment data following (9) and (11), respectively, 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 (36) in Appendix B.3 and (42) in Appendix B.4. The empirical alignments corroborate Theorems 4 and 5, further confirming that Proposition 1 is applicable to a broader range of real-world distributional alignment data.

• **Low-rank (Fig. 3c) and LoRA (Fig. 3d) experiments.** We also run simulations to verify our theoretical findings in Section 3.2. Consider the independent data setting as described in (7). In Fig. 3c, we set $\boldsymbol{\Sigma_x} = \boldsymbol{I}_d$, $\sigma = 0$ 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 $\boldsymbol{W}_k, \boldsymbol{W}_q \in \mathbb{R}^{(d+1)\times r}$ and vary r across the set $\{1, 5, 10, 20\}$. Results show that empirical (solid) and theoretical (dotted, c.f. (13)) curves overlap. In Fig. 3d, we implement two phases of training. *Phase 1:* Setting $\boldsymbol{\Sigma_x} = \boldsymbol{\Sigma_\beta} = \boldsymbol{I}_d$ and $\sigma = 0$, we pretrain the model with full rank parameters and obtain weights $\hat{\boldsymbol{W}}_k, \hat{\boldsymbol{W}}_q, \hat{\boldsymbol{W}}_v \in \mathbb{R}^{(d+1)\times(d+1)}$. *Phase 2:* We generate new examples with task covariance $\boldsymbol{\Sigma_\beta}$ 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 $\boldsymbol{W}_{\text{up}}, \boldsymbol{W}_{\text{down}} \in \mathbb{R}^{(d+1)\times r}$ where $\boldsymbol{W}_{lora} := \boldsymbol{W}_{\text{up}}\boldsymbol{W}_{\text{down}}^\top$ and (2a) becomes $\texttt{ATT}(\boldsymbol{Z}) = (\boldsymbol{Z}(\hat{\boldsymbol{W}}_q\hat{\boldsymbol{W}}_k^\top + \boldsymbol{W}_{\text{up}}\boldsymbol{W}_{\text{down}}^\top)\boldsymbol{Z}^\top)\boldsymbol{Z}\hat{\boldsymbol{W}}_v$. Fig. 3d presents the results after two phases of training where dotted curves are drawn from the right hand side of (14) directly. Here, note that since $\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^{new}$ are diagonal, the right hand side of (14) returns the exact optimal risk of LoRA and the alignments verify it.

5 Related Work

There is growing interest in understanding the mechanisms behind ICL [Brown et al., 2020, Liu et al., 2023b, Rae et al., 2021] in LLMs due to its success in continuously enabling novel applications for LLMs [GeminiTeam et al., 2023, OpenAI, 2023, Touvron et al., 2023]. In the previous work, Garg et al. [2022] 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 [von Oswald et al., 2023, Gatmiry et al., Collins et al., 2024, Lin and Lee, 2024, Li et al., 2024, Bai et al., 2024, Akyürek et al., 2023, Zhang et al., 2023, Du et al., 2023]. Akyürek et al. [2023] focus on linear regression and provide a construction of Transformer weights that can enable a single step of GD based on in-context examples. Along the similar line, Von Oswald et al. [2023] provide a construction of weights in linear attention-only Transformers that can emulate GD steps on in-context examples for a linear regression task. Similar to this line of work, Dai et al. [2023] 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.

Building on these primarily empirical studies, Zhang et al. [2024], Mahankali et al. [2024], Ahn et al. [2023], Duraisamy [2024] focus on developing a theoretical understanding of Transformers trained to perform ICL. For single-layer linear attention model trained on independent in-context prompts for random linear regression tasks, Mahankali et al. [2024], Ahn et al. [2023] show that the resulting model implements a single step of PGD on in-context examples in a test prompt, thereby corroborating the findings of [Von Oswald et al., 2023]. Zhang et al. [2024] study the optimization dynamics of gradient flow while training a single-layer linear attention model on in-context prompts for random linear regression tasks. Similar to Mahankali et al. [2024], Ahn et al. [2023], 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.

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 [Lewis et al., 2020, Nakano et al., 2021] 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 [Izacard et al., 2023], which can significantly outperform the typical ICL setup where fixed task-specific demonstrations are provided as in-context examples [Wang et al., 2022, Basu et al., 2023]. 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) [Gu et al., 2021b,a, Fu et al., 2023, Gu and Dao, 2023] 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 [Park et al., 2024, Grazzi et al., 2024] as well as complex NLP tasks [Grazzi et al., 2024]. That said, a rigorous theoretical understanding of ICL for SSMs akin to Zhang et al. [2024], Mahankali et al. [2024], Ahn et al. [2023] is missing from the literature. In this work, we provide the first such theoretical treatment for ICL with SSMs. Focusing on H3 architecture [Fu et al., 2023], we highlight its advantages over linear attention in specific ICL settings.

6 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 [Ahn et al., 2023, Von Oswald et al., 2023]. Finally, we have studied the population risk of ICL training whereas one can also explore the sample complexity of pretraining [Wu et al., 2023, Lu et al., 2024]. 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.

Acknowledgements

This work was supported in part by the National Science Foundation grants CCF-2046816, CCF-2403075, the Office of Naval Research award N000142412289, an Adobe Data Science Research award, and a gift by Google Research.

References

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.

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.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

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.

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.

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.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. *arXiv preprint arXiv:1912.01198*, 2019.

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.

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.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

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.

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

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.

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.

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*.

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.

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

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

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

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, 2021b.

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.

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.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.

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.

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.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International conference on artificial intelligence and statistics*, pages 4313–4324. PMLR, 2020.

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.

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.

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

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*, 2023a. URL https://openreview.net/forum?id=De4FYqjFueZ.

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, 2023b.

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.

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.

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.

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

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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.

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.

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

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.

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.

Appendix

Table of Contents

A Equivalence among Gradient Descent, Attention, and State-Space Models **15**
- A.1 Proof of Proposition 1 . . . 16
- A.2 Proof of Lemma 1 . . . 20
- A.3 Proof of Lemma 2 . . . 21

B Analysis of General Data Distribution **21**
- B.1 Supporting Results . . . 22
- B.2 Independent Data with General Covariance . . . 24
- B.3 Retrieval Augmented Generation with α Correlation . . . 25
- B.4 Task-feature Alignment with α Correlation . . . 28

C Analysis of Low-Rank Parameterization **31**
- C.1 Proof of Lemma 3 . . . 31

D Additional Experiments **32**

E Extended Related Work **33**

A Equivalence among Gradient Descent, Attention, and State-Space Models

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

$$
\begin{aligned}
&\boldsymbol{X} = [\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d},\\
&\boldsymbol{\xi} = [\xi_1 \ \cdots \ \xi_n]^\top \in \mathbb{R}^{n},\\
&\boldsymbol{y} = [y_1 \ \cdots \ y_n]^\top = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi} \in \mathbb{R}^{n},\\
&\boldsymbol{Z}_0 = [\boldsymbol{z}_1 \ \ldots \ \boldsymbol{z}_n \ \boldsymbol{0}_{d+1}]^\top = \begin{bmatrix} \boldsymbol{x}_1 & \ldots & \boldsymbol{x}_n & \boldsymbol{0}_d \\ y_1 & \ldots & y_n & 0 \end{bmatrix}^\top \in \mathbb{R}^{(n+1)\times(d+1)},
\end{aligned}
$$

and corresponding prediction functions

$$
\begin{aligned}
g_{\texttt{PGD}}(\boldsymbol{Z}) &= \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top \boldsymbol{y}, && \text{(15a)}\\
g_{\texttt{WPGD}}(\boldsymbol{Z}) &= \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top (\boldsymbol{\omega} \odot \boldsymbol{y}), && \text{(15b)}\\
g_{\texttt{ATT}}(\boldsymbol{Z}) &= (\boldsymbol{z}^\top \boldsymbol{W}_q \boldsymbol{W}_k^\top \boldsymbol{Z}_0^\top)\boldsymbol{Z}_0 \boldsymbol{W}_v \boldsymbol{v}, && \text{(15c)}\\
g_{\texttt{SSM}}(\boldsymbol{Z}) &= \Big((\boldsymbol{z}^\top \boldsymbol{W}_q)^\top \odot ((\boldsymbol{Z}_0 \boldsymbol{W}_k \odot \boldsymbol{Z}_0 \boldsymbol{W}_v) * \boldsymbol{f})_{n+1}\Big)\boldsymbol{v}, && \text{(15d)}
\end{aligned}
$$

we have objectives

$$
\begin{aligned}
&\min_{\boldsymbol{W}} \mathcal{L}_{\texttt{PGD}}(\mathcal{W}) \quad \text{where} \quad \mathcal{L}_{\texttt{PGD}}(\mathcal{W}) = \mathbb{E}\left[(y - g_{\texttt{PGD}}(\boldsymbol{Z}))^2\right], && \text{(16a)}\\
&\min_{\boldsymbol{W},\boldsymbol{\omega}} \mathcal{L}_{\texttt{WPGD}}(\mathcal{W}) \quad \text{where} \quad \mathcal{L}_{\texttt{WPGD}}(\mathcal{W}) = \mathbb{E}\left[(y - g_{\texttt{WPGD}}(\boldsymbol{Z}))^2\right], && \text{(16b)}\\
&\min_{\boldsymbol{W}_k,\boldsymbol{W}_q,\boldsymbol{W}_v,\boldsymbol{v}} \mathcal{L}_{\texttt{ATT}}(\mathcal{W}) \quad \text{where} \quad \mathcal{L}_{\texttt{ATT}}(\mathcal{W}) = \mathbb{E}\left[(y - g_{\texttt{ATT}}(\boldsymbol{Z}))^2\right], && \text{(16c)}\\
&\min_{\boldsymbol{W}_k,\boldsymbol{W}_q,\boldsymbol{W}_v,\boldsymbol{v},\boldsymbol{f}} \mathcal{L}_{\texttt{SSM}}(\mathcal{W}) \quad \text{where} \quad \mathcal{L}_{\texttt{SSM}}(\mathcal{W}) = \mathbb{E}\left[(y - g_{\texttt{SSM}}(\boldsymbol{Z}))^2\right]. && \text{(16d)}
\end{aligned}
$$

Here, the expectation is over the randomness in $(\boldsymbol{x}_i, \xi_i)_{i=1}^n$ and $\boldsymbol{\beta}$, and the search space for $\boldsymbol{W}$ is $\mathbb{R}^{d\times d}$, for $\boldsymbol{\omega}$ is $\mathbb{R}^n$, for $\boldsymbol{W}_k, \boldsymbol{W}_q, \boldsymbol{W}_v$ is $\mathbb{R}^{(d+1)\times(d+1)}$, for $\boldsymbol{v}$ is $\mathbb{R}^{d+1}$, and for $\boldsymbol{f}$ is $\mathbb{R}^{n+1}$.

A.1 Proof of Proposition 1

Consider the problem setting as discussed in Section 2, Proposition 1 can be proven by the following two lemmas.

Lemma 4 *Suppose Assumptions 1 and 2 hold. Then, given the objectives* (16a) *and* (16c)*, we have*

$$\min_{\boldsymbol{W}_q,\boldsymbol{W}_k,\boldsymbol{W}_v,\boldsymbol{v}} \mathcal{L}_{ATT}(\mathcal{W}) = \min_{\boldsymbol{W}} \mathcal{L}_{PGD}(\mathcal{W}).$$

Proof. Recap the linear attention estimator from (15c) and denote

$$\boldsymbol{W}_q\boldsymbol{W}_k^\top = \begin{bmatrix} \bar{\boldsymbol{W}} & \boldsymbol{w}_1 \\ \boldsymbol{w}_2^\top & w \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{W}_v\boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}_1 \\ v \end{bmatrix},$$

where $\bar{\boldsymbol{W}} \in \mathbb{R}^{d\times d}$, $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{v}_1 \in \mathbb{R}^d$, and $w, v \in \mathbb{R}$. Then we have

$$\begin{aligned}
g_{\mathtt{ATT}}(\boldsymbol{Z}) &= (z^\top \boldsymbol{W}_q\boldsymbol{W}_k^\top \boldsymbol{Z}_0^\top)\boldsymbol{Z}_0\boldsymbol{W}_v\boldsymbol{v} \\
&= [\boldsymbol{x}^\top \ 0] \begin{bmatrix} \bar{\boldsymbol{W}} & \boldsymbol{w}_1 \\ \boldsymbol{w}_2^\top & w \end{bmatrix} \begin{bmatrix} \boldsymbol{X}^\top & \boldsymbol{0}_d \\ \boldsymbol{y}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{X} & \boldsymbol{y} \\ \boldsymbol{0}_d^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 \\ v \end{bmatrix} \\
&= (\boldsymbol{x}^\top \bar{\boldsymbol{W}}\boldsymbol{X}^\top + \boldsymbol{x}^\top \boldsymbol{w}_1\boldsymbol{y}^\top)(\boldsymbol{X}\boldsymbol{v}_1 + \boldsymbol{y}v) \\
&= \boldsymbol{x}^\top(v\bar{\boldsymbol{W}})\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{x}^\top\boldsymbol{w}_1\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{v}_1 + \boldsymbol{x}^\top\left(\bar{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_1 + v\,\|\boldsymbol{y}\|_{\ell_2}^2\,\boldsymbol{w}_1\right) \\
&= \boldsymbol{x}^\top(v\bar{\boldsymbol{W}} + \boldsymbol{w}_1\boldsymbol{v}_1^\top)\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{x}^\top\left(\bar{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_1 + v\,\|\boldsymbol{y}\|_{\ell_2}^2\,\boldsymbol{w}_1\right) \\
&= \underbrace{\boldsymbol{x}^\top\tilde{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{y}}_{\tilde{g}_{\mathtt{ATT}}(\boldsymbol{Z})} + \underbrace{\boldsymbol{x}^\top\left(\bar{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_1 + v\,\|\boldsymbol{y}\|_{\ell_2}^2\,\boldsymbol{w}_1\right)}_{\varepsilon},
\end{aligned} \tag{17}$$

where $\tilde{\boldsymbol{W}} := v\bar{\boldsymbol{W}} + \boldsymbol{w}_1\boldsymbol{v}_1^\top$.

We first show that for any given parameters $\boldsymbol{W}_k, \boldsymbol{W}_q, \boldsymbol{W}_v, \boldsymbol{v}$,

$$\mathbb{E}\left[(g_{\mathtt{ATT}}(\boldsymbol{Z}) - y)^2\right] \geq \mathbb{E}\left[(\tilde{g}_{\mathtt{ATT}}(\boldsymbol{Z}) - y)^2\right]. \tag{18}$$

To this goal, we have

$$\begin{aligned}
\mathbb{E}\left[(g_{\mathtt{ATT}}(\boldsymbol{Z}) - y)^2\right] - \mathbb{E}\left[(\tilde{g}_{\mathtt{ATT}}(\boldsymbol{Z}) - y)^2\right] &= \mathbb{E}\left[(\tilde{g}_{\mathtt{ATT}}(\boldsymbol{Z}) + \varepsilon - y)^2\right] - \mathbb{E}\left[(\tilde{g}_{\mathtt{ATT}}(\boldsymbol{Z}) - y)^2\right] \\
&= \mathbb{E}[\varepsilon^2] + 2\,\mathbb{E}[(\tilde{g}_{\mathtt{ATT}}(\boldsymbol{Z}) - y)\varepsilon]
\end{aligned} \tag{19}$$

where we have decomposition

$$\begin{aligned}
(\tilde{g}_{\mathtt{ATT}}(\boldsymbol{Z}) - y)\varepsilon &= (\boldsymbol{x}^\top\tilde{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{y} - y)\boldsymbol{x}^\top\left(\bar{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_1 + v\,\|\boldsymbol{y}\|_{\ell_2}^2\,\boldsymbol{w}_1\right) \\
&= \boldsymbol{y}^\top\boldsymbol{X}\tilde{\boldsymbol{W}}^\top\boldsymbol{x}\boldsymbol{x}^\top\left(\bar{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_1 + v\,\|\boldsymbol{y}\|_{\ell_2}^2\,\boldsymbol{w}_1\right) - y\boldsymbol{x}^\top\left(\bar{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_1 + v\,\|\boldsymbol{y}\|_{\ell_2}^2\,\boldsymbol{w}_1\right) \\
&= \underbrace{\boldsymbol{y}^\top\boldsymbol{X}\tilde{\boldsymbol{W}}^\top\boldsymbol{x}\boldsymbol{x}^\top\bar{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_1}_{(a)} + \underbrace{v\,\|\boldsymbol{y}\|_{\ell_2}^2\,\boldsymbol{y}^\top\boldsymbol{X}\tilde{\boldsymbol{W}}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{w}_1}_{(b)} - \underbrace{y\boldsymbol{x}^\top\bar{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_1}_{(c)} - \underbrace{vy\,\|\boldsymbol{y}\|_{\ell_2}^2\,\boldsymbol{x}^\top\boldsymbol{w}_1}_{(d)}.
\end{aligned}$$

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 $\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\beta}$ returns zero, i.e., order of $\boldsymbol{x}^\top\boldsymbol{\beta}\boldsymbol{x}$ is 3 and therefore $\mathbb{E}[\boldsymbol{x}^\top\boldsymbol{\beta}\boldsymbol{x}] = \boldsymbol{0}_d$.

$$\begin{aligned}
(a): \quad & \mathbb{E}\left[\boldsymbol{y}^\top\boldsymbol{X}\tilde{\boldsymbol{W}}^\top\boldsymbol{x}\boldsymbol{x}^\top\bar{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_1\right] \\
&= \mathbb{E}\left[(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi})^\top\boldsymbol{X}\tilde{\boldsymbol{W}}^\top\boldsymbol{x}\boldsymbol{x}^\top\bar{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_1\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\tilde{\boldsymbol{W}}^\top\boldsymbol{x}\boldsymbol{x}^\top\bar{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_1\right] + \mathbb{E}\left[\boldsymbol{\xi}^\top\boldsymbol{X}\tilde{\boldsymbol{W}}^\top\boldsymbol{x}\boldsymbol{x}^\top\bar{\boldsymbol{W}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_1\right] \\
&= 0.
\end{aligned}$$

$$\begin{aligned}
(b): \quad & \mathbb{E}\left[v\,\|\boldsymbol{y}\|_{\ell_2}^2\,\boldsymbol{y}^\top\boldsymbol{X}\tilde{\boldsymbol{W}}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{w}_1\right] \\
&= \mathbb{E}\left[v(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi})^\top(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi})^\top\boldsymbol{X}\tilde{\boldsymbol{W}}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{w}_1\right] \\
&= \mathbb{E}\left[v\,\|\boldsymbol{\xi}\|_{\ell_2}^2\,\boldsymbol{\xi}^\top\boldsymbol{X}\tilde{\boldsymbol{W}}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{w}_1\right] \\
&= 0.
\end{aligned}$$

$$
\begin{aligned}
(c): \quad & \mathbb{E}\left[y\boldsymbol{x}^\top \bar{\boldsymbol{W}}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}_1\right] \\
&= \mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{\beta}+\xi)\boldsymbol{x}^\top \bar{\boldsymbol{W}}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}_1\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^\top\boldsymbol{x}\boldsymbol{x}^\top \bar{\boldsymbol{W}}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}_1\right] + \mathbb{E}\left[\xi\boldsymbol{x}^\top \bar{\boldsymbol{W}}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}_1\right] \\
&= 0.
\end{aligned}
$$

$$
\begin{aligned}
(d): \quad & \mathbb{E}\left[vy\,\|\boldsymbol{y}\|_{\ell_2}^2\,\boldsymbol{x}^\top\boldsymbol{w}_1\right] \\
&= v\,\mathbb{E}\left[(\boldsymbol{\beta}^\top\boldsymbol{x}+\xi)(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\xi})^\top(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\xi})\boldsymbol{x}^\top\boldsymbol{w}_1\right] \\
&= v\,\mathbb{E}\left[\xi\,\|\boldsymbol{\xi}\|_{\ell_2}^2\,\boldsymbol{x}^\top\boldsymbol{w}_1\right] \\
&= 0.
\end{aligned}
$$

Combining the results with (19) returns that

$$
\mathbb{E}\left[(g_{\texttt{ATT}}(\boldsymbol{Z})-y)^2\right] - \mathbb{E}\left[(\tilde{g}_{\texttt{ATT}}(\boldsymbol{Z})-y)^2\right] = \mathbb{E}[\varepsilon^2] \geq 0 \tag{20}
$$

which completes the proof of (18). Therefore, we obtain

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

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

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

Then we obtain

$$
g_{\texttt{ATT}}(\boldsymbol{Z}) = \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{y} = g_{\texttt{PGD}}(\boldsymbol{Z}), \tag{21}
$$

which completes the proof. ■

Lemma 5 *Suppose Assumptions 1 and 2 hold. Then, given the objectives in (16), we have*

$$
\min_{\boldsymbol{W}_q,\boldsymbol{W}_k,\boldsymbol{W}_v,\boldsymbol{v},\boldsymbol{f}} \mathcal{L}_{\texttt{SSM}}(\mathcal{W}) = \min_{\boldsymbol{W},\boldsymbol{\omega}} \mathcal{L}_{\texttt{WPGD}}(\mathcal{W}). \tag{22}
$$

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

$$
\min_{\boldsymbol{W},\boldsymbol{\omega}} \mathcal{L}_{\texttt{WPGD}}(\mathcal{W}) = \min_{\boldsymbol{W}} \mathcal{L}_{\texttt{PGD}}(\mathcal{W}). \tag{23}
$$

Proof. Recap the SSM estimator from (15d) and let

$$
\begin{aligned}
\boldsymbol{W}_q &= \begin{bmatrix} \boldsymbol{w}_{q1} & \boldsymbol{w}_{q2} & \cdots & \boldsymbol{w}_{q,d+1} \end{bmatrix}, \\
\boldsymbol{W}_k &= \begin{bmatrix} \boldsymbol{w}_{k1} & \boldsymbol{w}_{k2} & \cdots & \boldsymbol{w}_{k,d+1} \end{bmatrix}, \\
\boldsymbol{W}_v &= \begin{bmatrix} \boldsymbol{w}_{v1} & \boldsymbol{w}_{v2} & \cdots & \boldsymbol{w}_{v,d+1} \end{bmatrix},
\end{aligned}
$$

where $\boldsymbol{w}_{qj}, \boldsymbol{w}_{kj}, \boldsymbol{w}_{vj} \in \mathbb{R}^{d+1}$ for $j \leq d+1$, and let

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

Then we have

$$
\begin{aligned}
g_{\texttt{SSM}}(\boldsymbol{Z}) &= \left((z^\top\boldsymbol{W}_q)^\top \odot ((\boldsymbol{Z}_0\boldsymbol{W}_k \odot \boldsymbol{Z}_0\boldsymbol{W}_v) * \boldsymbol{f})_{n+1}\right)\boldsymbol{v} \\
&= \sum_{i=1}^n f_{n+1-i} \cdot \boldsymbol{v}^\top \left(\begin{bmatrix} \boldsymbol{w}_{q1}^\top z \\ \cdots \\ \boldsymbol{w}_{q,d+1}^\top z \end{bmatrix} \odot \begin{bmatrix} \boldsymbol{w}_{k1}^\top z_i \boldsymbol{w}_{v1}^\top z_i \\ \cdots \\ \boldsymbol{w}_{k,d+1}^\top z_i \boldsymbol{w}_{v,d+1}^\top z_i \end{bmatrix} \right) \\
&= \sum_{i=1}^n f_{n+1-i} \cdot \boldsymbol{v}^\top \begin{bmatrix} \boldsymbol{w}_{q1}^\top z \boldsymbol{w}_{k1}^\top z_i \boldsymbol{w}_{v1}^\top z_i \\ \cdots \\ \boldsymbol{w}_{q,d+1}^\top z \boldsymbol{w}_{k,d+1}^\top z_i \boldsymbol{w}_{v,d+1}^\top z_i \end{bmatrix}.
\end{aligned}
$$

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

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

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

$$\begin{aligned}
\boldsymbol{w}_{qj}^\top \boldsymbol{z} \boldsymbol{w}_{kj}^\top \boldsymbol{z}_i \boldsymbol{w}_{vj}^\top \boldsymbol{z}_i &= \left(\bar{\boldsymbol{w}}_{qj}^\top \boldsymbol{x}\right)\left(\bar{\boldsymbol{w}}_{kj}^\top \boldsymbol{x}_i + w_{kj} y_i\right)\left(\bar{\boldsymbol{w}}_{vj}^\top \boldsymbol{x}_i + w_{vj} y_i\right) \\
&= \boldsymbol{x}^\top \bar{\boldsymbol{w}}_{qj}\left(w_{vj}\bar{\boldsymbol{w}}_{kj}^\top + w_{kj}\bar{\boldsymbol{w}}_{vj}^\top\right)\boldsymbol{x}_i y_i + \left(\bar{\boldsymbol{w}}_{qj}^\top \boldsymbol{x}\right)\left(\bar{\boldsymbol{w}}_{kj}^\top \boldsymbol{x}_i\right)\left(\bar{\boldsymbol{w}}_{vj}^\top \boldsymbol{x}_i\right) + \left(w_{kj} w_{vj} \bar{\boldsymbol{w}}_{qj}^\top \boldsymbol{x} y_i^2\right) \\
&= \boldsymbol{x}^\top \boldsymbol{W}_j' \boldsymbol{x}_i y_i + \delta_j(\boldsymbol{x}, \boldsymbol{x}_i, \boldsymbol{x}_i) + \boldsymbol{w}_j'^\top \boldsymbol{x} y_i^2
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{W}_j' &:= \bar{\boldsymbol{w}}_{qj}\left(w_{vj}\bar{\boldsymbol{w}}_{kj}^\top + w_{kj}\bar{\boldsymbol{w}}_{vj}^\top\right) \in \mathbb{R}^{d\times d}, \\
\boldsymbol{w}_j' &:= w_{kj} w_{vj} \bar{\boldsymbol{w}}_{qj} \in \mathbb{R}^d, \\
\delta_j(\boldsymbol{x}, \boldsymbol{x}_i, \boldsymbol{x}_i) &:= \left(\bar{\boldsymbol{w}}_{qj}^\top \boldsymbol{x}\right)\left(\bar{\boldsymbol{w}}_{kj}^\top \boldsymbol{x}_i\right)\left(\bar{\boldsymbol{w}}_{vj}^\top \boldsymbol{x}_i\right) \in \mathbb{R}.
\end{aligned}$$

Then

$$\begin{aligned}
g_{\texttt{SSM}}(\boldsymbol{Z}) &= \sum_{i=1}^n f_{n+1-i} \cdot \sum_{j=1}^{d+1} v_j \left(\boldsymbol{x}^\top \boldsymbol{W}_j' \boldsymbol{x}_i y_i + \delta_j(\boldsymbol{x}, \boldsymbol{x}_i, \boldsymbol{x}_i) + \boldsymbol{w}_j'^\top \boldsymbol{x} y_i^2\right) \\
&= \boldsymbol{x}^\top \left(\sum_{j=1}^{d+1} v_j \boldsymbol{W}_j'\right) \boldsymbol{X}(\boldsymbol{y} \odot \tilde{\boldsymbol{f}}) + \sum_{i=1}^n f_{n+1-i} \cdot \sum_{j=1}^{d+1} v_j \cdot \delta_j(\boldsymbol{x}, \boldsymbol{x}_i, \boldsymbol{x}_i) + \left(\sum_{j=1}^{d+1} v_j \boldsymbol{w}_j'^\top\right) \boldsymbol{x}\boldsymbol{y}^\top (\boldsymbol{y} \odot \tilde{\boldsymbol{f}}) \\
&= \underbrace{\boldsymbol{x}^\top \tilde{\boldsymbol{W}} \boldsymbol{X} \tilde{\boldsymbol{y}}}_{\tilde{g}_{\texttt{SSM}}(\boldsymbol{Z})} + \underbrace{\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X})}_{\varepsilon_1} + \underbrace{\tilde{\boldsymbol{w}}^\top \boldsymbol{x}\boldsymbol{y}^\top \tilde{\boldsymbol{y}}}_{\varepsilon_2}.
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\boldsymbol{f}} &:= [f_n \ \cdots \ f_1]^\top \in \mathbb{R}^n, \\
\tilde{\boldsymbol{y}} &:= \boldsymbol{y} \odot \tilde{\boldsymbol{f}} \in \mathbb{R}^n, \\
\tilde{\boldsymbol{W}} &:= \sum_{j=1}^{d+1} v_j \boldsymbol{W}_j' \in \mathbb{R}^{d\times d}, \\
\tilde{\boldsymbol{w}} &:= \sum_{j=1}^{d+1} v_j \boldsymbol{w}_j' \in \mathbb{R}^d, \\
\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) &:= \sum_{i=1}^n f_{n+1-i} \cdot \sum_{j=1}^{d+1} v_j \cdot \delta_j(\boldsymbol{x}, \boldsymbol{x}_i, \boldsymbol{x}_i) \in \mathbb{R}.
\end{aligned}$$

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

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

To start with, we obtain

$$\begin{aligned}
\mathbb{E}\left[(g_{\texttt{SSM}}(\boldsymbol{Z}) - y)^2\right] &= \mathbb{E}\left[(\tilde{g}_{\texttt{SSM}}(\boldsymbol{Z}) + \varepsilon_1 + \varepsilon_2 - y)^2\right] \\
&= \mathbb{E}\left[(\tilde{g}_{\texttt{SSM}}(\boldsymbol{Z}) - y)^2\right] + \mathbb{E}\left[(\varepsilon_1 + \varepsilon_2)^2\right] + 2\,\mathbb{E}\left[(\tilde{g}_{\texttt{SSM}}(\boldsymbol{Z}) - y)(\varepsilon_1 + \varepsilon_2)\right] \qquad (24)
\end{aligned}$$

where there is decomposition

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

In the following, similar to the proof of Lemma 4, 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.

$$
\begin{aligned}
(a): \quad & \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}} \boldsymbol{X} \tilde{\boldsymbol{y}}\right] \\
& = \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}} \boldsymbol{X} (\boldsymbol{X}\boldsymbol{\beta} \odot \tilde{\boldsymbol{f}})\right] + \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}} \boldsymbol{X} (\boldsymbol{\xi} \odot \tilde{\boldsymbol{f}})\right] \\
& = \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}} \boldsymbol{X}\right] \mathbb{E}\left[\boldsymbol{\xi} \odot \tilde{\boldsymbol{f}}\right] \\
& = 0.
\end{aligned}
$$

$$
\begin{aligned}
(b): \quad & \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) y\right] \\
& = \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X})(\boldsymbol{x}^\top \boldsymbol{\beta} + \xi)\right] \\
& = \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) \boldsymbol{x}^\top \boldsymbol{\beta}\right] + \mathbb{E}\left[\tilde{\delta}(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{X}) \xi\right] \\
& = 0.
\end{aligned}
$$

$$
\begin{aligned}
(c): \quad & \mathbb{E}\left[\tilde{\boldsymbol{w}}^\top \boldsymbol{x} \boldsymbol{y}^\top \tilde{\boldsymbol{y}} \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}} \boldsymbol{X} \tilde{\boldsymbol{y}}\right] \\
& = \mathbb{E}\left[\tilde{\boldsymbol{w}}^\top \boldsymbol{x} (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi})^\top (\boldsymbol{X}\boldsymbol{\beta} \odot \tilde{\boldsymbol{f}} + \boldsymbol{\xi} \odot \tilde{\boldsymbol{f}}) \cdot \boldsymbol{x}^\top \tilde{\boldsymbol{W}} \boldsymbol{X} (\boldsymbol{X}\boldsymbol{\beta} \odot \tilde{\boldsymbol{f}} + \boldsymbol{\xi} \odot \tilde{\boldsymbol{f}})\right] \\
& = 0.
\end{aligned}
$$

$$
\begin{aligned}
(d): \quad & \mathbb{E}\left[y \cdot \tilde{\boldsymbol{w}}^\top \boldsymbol{x} \boldsymbol{y}^\top \tilde{\boldsymbol{y}}\right] \\
& = \mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{\beta} + \xi) \cdot \tilde{\boldsymbol{w}}^\top \boldsymbol{x} (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi})^\top (\boldsymbol{X}\boldsymbol{\beta} \odot \tilde{\boldsymbol{f}} + \boldsymbol{\xi} \odot \tilde{\boldsymbol{f}})\right] \\
& = 0.
\end{aligned}
$$

Combining the results with (24) results that

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

Therefore we obtain,

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

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

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

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

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

which completes the proof of (22).

Next, to show (23), 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 (23). ■

A.2 Proof of Lemma 1

Proof. Recap the loss $\mathcal{L}_{\mathtt{PGD}}(\mathcal{W})$ in (16a) and prediction $g_{\mathtt{PGD}}(\boldsymbol{Z})$ in (15a), we have

$$\begin{aligned}\mathcal{L}_{\mathtt{PGD}}(\mathcal{W}) &= \mathbb{E}[(y - g_{\mathtt{PGD}}(\boldsymbol{Z}))^2] \\ &= \mathbb{E}\left[\left(\boldsymbol{x}^\top\boldsymbol{\beta} + \xi - \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi})\right)^2\right] \\ &= \mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})^2 + 2(\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] \\ &= \mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})^2 + (\xi - \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi})^2\right] + 2\,\mathbb{E}[(\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})] \\ &= \mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})^2 + (\xi - \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi})^2\right] \qquad (25) \\ &= \underbrace{\mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})^2 + (\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi})^2\right]}_{f_1(\boldsymbol{W})} \underbrace{-2\,\mathbb{E}[\boldsymbol{\beta}^\top\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} + \xi\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi}]}_{f_2(\boldsymbol{W})} + \underbrace{\mathbb{E}[(\boldsymbol{x}^\top\boldsymbol{\beta})^2 + \xi^2]}_{\text{constant}}\end{aligned}$$

where (25) follows Assumption 2. Since $f_2(\boldsymbol{W})$ is convex, $\mathcal{L}_{\mathtt{PGD}}(\mathcal{W})$ is strongly-convex if and only if $f_1(\boldsymbol{W})$ is strongly-convex, which completes the proof of strong convexity.

Next, (20) and (21) in the proof of Lemma 4 demonstrate that the optimal loss is achievable and is achieved at $\varepsilon = 0$. Subsequently, (17) indicates that $g^\star_{\mathtt{ATT}}$ has the same form as $g^\star_{\mathtt{PGD}}$. Under the strong convexity assumption, $g^\star_{\mathtt{PGD}}$ is unique, which leads to the conclusion that $g^\star_{\mathtt{PGD}} = g^\star_{\mathtt{ATT}}$. ■

A.3 Proof of Lemma 2

Proof. According to Lemma 1, $\mathcal{L}_{\text{PGD}}(\mathcal{W})$ is strongly-convex as long as either $\mathbb{E}[(\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta})^2]$ or $\mathbb{E}[(\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top \boldsymbol{\xi})^2]$ is strongly-convex. Therefore, in this lemma, the two claims correspond to the strong convexity of $\mathbb{E}[(\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top \boldsymbol{\xi})^2]$ and $\mathbb{E}[(\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta})^2]$ terms, respectively.

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

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

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

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

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

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

• **Claim 2:** Now recall that $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = \mathbb{E}[\boldsymbol{X}_1^\top \boldsymbol{X}_1]$ and set $\boldsymbol{A} = \boldsymbol{X}_1^\top \boldsymbol{X}_1 - \bar{\boldsymbol{\Sigma}}_{\boldsymbol{X}}$ and $\boldsymbol{B} = \boldsymbol{X}_2^\top \boldsymbol{X}_2 + \bar{\boldsymbol{\Sigma}}_{\boldsymbol{X}}$. Observe that $\mathbb{E}[\boldsymbol{A}] = 0$. If the second claim holds, $\mathbb{E}[\boldsymbol{X}^\top \boldsymbol{X}] = \mathbb{E}[\boldsymbol{A} + \boldsymbol{B}]$. Note that $(\boldsymbol{A}, \boldsymbol{\beta}_1, \boldsymbol{x}_1)$ are independent of each other and $(\boldsymbol{B}, \boldsymbol{\beta}_2, \boldsymbol{x}_2)$. Using independence and $\mathbb{E}[\boldsymbol{A}] = 0$, similarly write

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

Now using $\mathbb{E}[\boldsymbol{\beta}_1] = \mathbb{E}[\boldsymbol{x}_1] = 0$ and their independence from rest, these terms obeys

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{A}\boldsymbol{\beta})^2] &= \mathbb{E}[(\boldsymbol{x}_1^\top \boldsymbol{W}\boldsymbol{A}\boldsymbol{\beta}_1)^2] + \mathbb{E}[(\boldsymbol{x}_1^\top \boldsymbol{W}\boldsymbol{A}\boldsymbol{\beta}_2)^2] + \mathbb{E}[(\boldsymbol{x}_2^\top \boldsymbol{W}\boldsymbol{A}\boldsymbol{\beta}_1)^2] + \mathbb{E}[(\boldsymbol{x}_2^\top \boldsymbol{W}\boldsymbol{A}\boldsymbol{\beta}_2)^2] \\
\mathbb{E}[(\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{B}\boldsymbol{\beta})^2] &= \mathbb{E}[(\boldsymbol{x}_1^\top \boldsymbol{W}\boldsymbol{B}\boldsymbol{\beta}_1)^2] + \mathbb{E}[(\boldsymbol{x}_1^\top \boldsymbol{W}\boldsymbol{B}\boldsymbol{\beta}_2)^2] + \mathbb{E}[(\boldsymbol{x}_2^\top \boldsymbol{W}\boldsymbol{B}\boldsymbol{\beta}_1)^2] + \mathbb{E}[(\boldsymbol{x}_2^\top \boldsymbol{W}\boldsymbol{B}\boldsymbol{\beta}_2)^2].
\end{aligned}$$

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 $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathbb{E}[\boldsymbol{X}^\top \boldsymbol{X}] \succeq \bar{\boldsymbol{\Sigma}}_{\boldsymbol{X}} \succ 0$, we note that

$$\mathbb{E}[(\boldsymbol{x}_1^\top \boldsymbol{W}\boldsymbol{A}\boldsymbol{\beta}_1)^2] + \mathbb{E}[(\boldsymbol{x}_1^\top \boldsymbol{W}\boldsymbol{B}\boldsymbol{\beta}_1)^2] = \mathbb{E}[(\boldsymbol{x}_1^\top \boldsymbol{W}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}_1)^2]$$

where $\boldsymbol{x}_1, \boldsymbol{\beta}_1, \boldsymbol{X}$ are independent and full-rank covariance. To proceed, note that

$$\mathbb{E}[(\boldsymbol{x}_1^\top \boldsymbol{W}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}_1)^2] = \mathbb{E}[(\boldsymbol{x}_1^\top \boldsymbol{W}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\beta}_1)^2] + \mathbb{E}[(\boldsymbol{x}_1^\top \boldsymbol{W}(\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{\Sigma}_{\boldsymbol{X}})\boldsymbol{\beta}_1)^2].$$

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

$$\mathbb{E}[(\boldsymbol{x}_1^\top \boldsymbol{W}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\beta}_1)^2] = \mathtt{tr}\left(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \boldsymbol{W}\boldsymbol{\Sigma}_{\boldsymbol{X}} \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{W}^\top\right) = \left\| \sqrt{\bar{\boldsymbol{\Sigma}}_{x}}\, \boldsymbol{W}\boldsymbol{\Sigma}_{\boldsymbol{X}} \sqrt{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}} \right\|_F^2 .$$

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

B Analysis of General Data Distribution

In this section, we provide the proofs in Section 3, which focuses on solving Objective (5a). For the sake of clean notation, let $\mathcal{L}(\boldsymbol{W}) := \mathcal{L}_{\text{PGD}}(\mathcal{W})$ and $g := g_{\text{PGD}}$ 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

$$\mathbb{E}[u^{2n}] = \sigma^{2n}(2n-1)!!. \tag{26}$$

• **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

$$\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{u}^\top \boldsymbol{W}\boldsymbol{u})(\boldsymbol{u}^\top \boldsymbol{W}'\boldsymbol{u})\right] \\
&= \mathbb{E}\left[\left(\sum_{i,j=1}^{d} W_{ij}u_iu_j\right)\left(\sum_{i,j=1}^{d} W'_{ij}u_iu_j\right)\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{d} W_{ii}u_i^2\right)\left(\sum_{i=1}^{d} W'_{ii}u_i^2\right)\right] + \mathbb{E}\left[\left(\sum_{i\neq j} W_{ij}u_iu_j\right)\left(\sum_{i\neq j} W'_{ij}u_iu_j\right)\right] \\
&= \sum_{i=1}^{d} W_{ii}W'_{ii}\,\mathbb{E}\left[u_i^4\right] + \sum_{i\neq j} W_{ii}W'_{jj}\,\mathbb{E}[u_i^2]\,\mathbb{E}[u_j^2] + \sum_{i\neq j} W_{ij}W'_{ij}\,\mathbb{E}[u_i^2]\,\mathbb{E}[u_j^2] + \sum_{i\neq j} W_{ij}W'_{ji}\,\mathbb{E}[u_i^2]\,\mathbb{E}[u_j^2] \\
&= 3\sum_{i=1}^{d} W_{ii}W'_{ii} + \sum_{i\neq j} W_{ii}W'_{jj} + \sum_{i\neq j} W_{ij}W'_{ij} + \sum_{i\neq j} W_{ij}W'_{ji} \\
&= \sum_{i,j=1}^{d} W_{ii}W'_{jj} + \sum_{i,j=1}^{d} W_{ij}W'_{ij} + \sum_{i,j=1}^{d} W_{ij}W'_{ji} \\
&= \mathtt{tr}\left(\boldsymbol{W}\right)\mathtt{tr}\left(\boldsymbol{W}'\right) + \mathtt{tr}\left(\boldsymbol{W}'\boldsymbol{W}^\top\right) + \mathtt{tr}\left(\boldsymbol{W}\boldsymbol{W}'\right).
\end{aligned} \tag{27}$$

• **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

$$\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{u}^\top \boldsymbol{W}\boldsymbol{v}\boldsymbol{v}^\top \boldsymbol{u})^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i,j=1}^{d} W_{ij}u_iv_j\right)^2\left(\sum_{i=1}^{d} u_iv_i\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i,j=1}^{d} W_{ij}^2u_i^2v_j^2 + \sum_{i\neq i'} W_{ij}W_{i'j}u_iu_{i'}v_j^2 + \sum_{j\neq j'} W_{ij}W_{ij'}u_i^2v_jv_{j'} + \sum_{i'\neq i, j'\neq j} W_{ij}W_{i'j'}u_iu_{i'}v_jv_{j'}\right)\left(\sum_{i=1}^{d} u_i^2v_i^2 + \sum_{i\neq j} u_iu_jv_iv_j\right)\right] \\
&= \mathbb{E}\left[\left(\sum_{i,j=1}^{d} W_{ij}^2u_i^2v_j^2\right)\left(\sum_{i=1}^{d} u_i^2v_i^2\right) + \left(\sum_{i\neq j} W_{ij}W_{ji}u_i^2u_j^2v_i^2v_j^2\right)\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{d} W_{ii}^2u_i^2v_i^2 + \sum_{i\neq j} W_{ij}^2u_i^2v_j^2\right)\left(\sum_{i=1}^{d} u_i^2v_i^2\right)\right] + \sum_{i\neq j} W_{ij}W_{ji} \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{d} W_{ii}^2u_i^4v_i^4 + \sum_{i\neq j} W_{ii}^2u_i^2v_i^2u_j^2v_j^2\right)\right] + \mathbb{E}\left[\left(\sum_{i\neq j} W_{ij}^2u_i^4v_j^2v_i^2 + \sum_{i\neq j} W_{ij}^2u_i^2v_j^4u_j^2 + \sum_{i\neq j\neq k} W_{ij}^2u_i^2v_j^2u_k^2v_k^2\right)\right] + \sum_{i\neq j} W_{ij}W_{ji} \\
&= 9\sum_{i=1}^{d} W_{ii}^2 + (d-1)\sum_{i=1}^{d} W_{ii}^2 + 6\sum_{i\neq j} W_{ij}^2 + (d-2)\sum_{i\neq j} W_{ij}^2 + \sum_{i\neq j} W_{ij}W_{ji} \\
&= 3\sum_{i=1}^{d} W_{ii}^2 + (d+4)\sum_{i,j=1}^{d} W_{ij}^2 + \sum_{i,j=1}^{d} W_{ij}W_{ji} \\
&= 3\mathtt{tr}\left(\boldsymbol{\Lambda}_{\boldsymbol{W}}^2\right) + (d+4)\mathtt{tr}\left(\boldsymbol{W}\boldsymbol{W}^\top\right) + \mathtt{tr}\left(\boldsymbol{W}^2\right).
\end{aligned} \tag{28}$$

• **6'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

$$
\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{u}^\top \boldsymbol{W}\boldsymbol{u})(\boldsymbol{u}^\top \boldsymbol{W}'\boldsymbol{u})\,\|\boldsymbol{u}\|_{\ell_2}^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i,j=1}^d W_{ij}u_iu_j\right)\left(\sum_{i,j=1}^d W'_{ij}u_iu_j\right)\left(\sum_{i=1}^d u_i^2\right)\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^d W_{ii}u_i^2\right)\left(\sum_{i=1}^d W'_{ii}u_i^2\right)\left(\sum_{i=1}^d u_i^2\right)\right] + \mathbb{E}\left[\left(\sum_{i\neq j} W_{ij}u_iu_j\right)\left(\sum_{i\neq j} W'_{ij}u_iu_j\right)\left(\sum_{i=1}^d u_i^2\right)\right] \\
&= \sum_{i=1}^d W_{ii}W'_{ii}\,\mathbb{E}\left[u_i^4\left(\sum_{i'=1}^d u_{i'}^2\right)\right] + \sum_{i\neq j} W_{ii}W'_{jj}\,\mathbb{E}\left[u_i^2u_j^2\left(\sum_{i'=1}^d u_{i'}^2\right)\right] \\
&\quad + \sum_{i\neq j} W_{ij}W'_{ij}\,\mathbb{E}\left[u_i^2u_j^2\left(\sum_{i'=1}^d u_{i'}^2\right)\right] + \sum_{i\neq j} W_{ij}W'_{ji}\,\mathbb{E}\left[u_i^2u_j^2\left(\sum_{i'=1}^d u_{i'}^2\right)\right] \\
&= (d+4)\left(3\sum_{i=1}^d W_{ii}W'_{ii} + \sum_{i\neq j} W_{ii}W'_{jj} + \sum_{i\neq j} W_{ij}W'_{ij} + \sum_{i\neq j} W_{ij}W'_{ji}\right) \qquad (29) \\
&= (d+4)\left(\sum_{i,j=1}^d W_{ii}W'_{jj} + \sum_{i,j=1}^d W_{ij}W'_{ij} + \sum_{i,j=1}^d W_{ij}W'_{ji}\right) \\
&= (d+4)\left(\mathtt{tr}\,(\boldsymbol{W})\,\mathtt{tr}\,(\boldsymbol{W}') + \mathtt{tr}\left(\boldsymbol{W}'\boldsymbol{W}^\top\right) + \mathtt{tr}\,(\boldsymbol{W}\boldsymbol{W}')\right), \qquad (30)
\end{aligned}
$$

where (29) is obtained by following

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

• **8'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

$$
\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{u}^\top \boldsymbol{W}\boldsymbol{u})(\boldsymbol{u}^\top \boldsymbol{W}'\boldsymbol{u})\,\|\boldsymbol{u}\|_{\ell_2}^4\right] \\
&= \mathbb{E}\left[\left(\sum_{i,j=1}^d W_{ij}u_iu_j\right)\left(\sum_{i,j=1}^d W'_{ij}u_iu_j\right)\left(\sum_{i,j=1}^d u_i^2u_j^2\right)\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^d W_{ii}u_i^2\right)\left(\sum_{i=1}^d W'_{ii}u_i^2\right)\left(\sum_{i=1}^d u_i^4 + \sum_{i\neq j} u_i^2u_j^2\right)\right] + \mathbb{E}\left[\left(\sum_{i\neq j} W_{ij}u_iu_j\right)\left(\sum_{i\neq j} W'_{ij}u_iu_j\right)\left(\sum_{i=1}^d u_i^4 + \sum_{i\neq j} u_i^2u_j^2\right)\right] \\
&= \sum_{i=1}^d W_{ii}W'_{ii}\,\mathbb{E}\left[u_i^4\left(\sum_{i'=1}^d u_{i'}^4 + \sum_{i'\neq j'} u_{i'}^2u_{j'}^2\right)\right] + \sum_{i\neq j} W_{ii}W'_{jj}\,\mathbb{E}\left[u_i^2u_j^2\left(\sum_{i'=1}^d u_{i'}^4 + \sum_{i'\neq j'} u_{i'}^2u_{j'}^2\right)\right] \\
&\quad + \sum_{i\neq j} W_{ij}W'_{ij}\,\mathbb{E}\left[u_i^2u_j^2\left(\sum_{i'=1}^d u_{i'}^4 + \sum_{i'\neq j'} u_{i'}^2u_{j'}^2\right)\right] + \sum_{i\neq j} W_{ij}W'_{ji}\,\mathbb{E}\left[u_i^2u_j^2\left(\sum_{i'=1}^d u_{i'}^4 + \sum_{i'\neq j'} u_{i'}^2u_{j'}^2\right)\right] \\
&= (d+4)(d+6)\left(3\sum_{i=1}^d W_{ii}W'_{ii} + \sum_{i\neq j} W_{ii}W'_{jj} + \sum_{i\neq j} W_{ij}W'_{ij} + \sum_{i\neq j} W_{ij}W'_{ji}\right) \qquad (31) \\
&= (d+4)(d+6)\left(\sum_{i,j=1}^d W_{ii}W'_{jj} + \sum_{i,j=1}^d W_{ij}W'_{ij} + \sum_{i,j=1}^d W_{ij}W'_{ji}\right) \\
&= (d+4)(d+6)\left(\mathtt{tr}\,(\boldsymbol{W})\,\mathtt{tr}\,(\boldsymbol{W}') + \mathtt{tr}\left(\boldsymbol{W}'\boldsymbol{W}^\top\right) + \mathtt{tr}\,(\boldsymbol{W}\boldsymbol{W}')\right). \qquad (32)
\end{aligned}
$$

where (31) is obtained by following

$$
\begin{aligned}
&\mathbb{E}\left[u_i^4\left(\sum_{i'=1}^{d} u_{i'}^4 + \sum_{i'\neq j'} u_{i'}^2 u_{j'}^2\right)\right] \\
&= \mathbb{E}[u^8] + (d-1)\,\mathbb{E}[u^4]\,\mathbb{E}[u^4] + 2(d-1)\,\mathbb{E}[u^6]\,\mathbb{E}[u^2] + (d-1)(d-2)\,\mathbb{E}[u^4]\,\mathbb{E}[u^2]\,\mathbb{E}[u^2] \\
&= 105 + 9(d-1) + 30(d-1) + 3(d-1)(d-2) \\
&= 3(d+4)(d+6), \\
&\mathbb{E}\left[u_i^2 u_j^2\left(\sum_{i'=1}^{d} u_{i'}^4 + \sum_{i'\neq j'} u_{i'}^2 u_{j'}^2\right)\right] \\
&= 2\,\mathbb{E}[u^6]\,\mathbb{E}[u^2] + (d-2)\,\mathbb{E}[u^4](\mathbb{E}[u^2])^2 + 2\,\mathbb{E}[u^4]\,\mathbb{E}[u^4] + 4(d-2)\,\mathbb{E}[u^4](\mathbb{E}[u^2])^2 + (d-2)(d-3)(\mathbb{E}[u^2])^4 \\
&= 30 + 3(d-2) + 18 + 12(d-2) + (d-2)(d-3) \\
&= (d+4)(d+6).
\end{aligned}
$$

B.2 Independent Data with General Covariance

Proof of Theorem 1. Consider a general independent linear model as defined in (7) where $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ are full-rank feature and task convariance matrices and

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

Let

$$
\boldsymbol{X} = [\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_n]^\top, \quad \boldsymbol{\xi} = [\xi_1 \ \cdots \ \xi_n]^\top, \quad \text{and} \quad \boldsymbol{y} = [y_1 \ \cdots \ y_n]^\top = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi}.
$$

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

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

and

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

Then recap the loss from (5a), and we obtain

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{W}) &= \mathbb{E}\left[(y - g(\boldsymbol{Z}))^2\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{x}^\top\boldsymbol{\beta} + \xi - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi})\right)^2\right] \\
&= \mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})^2 + 2(\boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})(\xi - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi}) + (\xi - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi})^2\right] \\
&= \mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})^2\right] + \mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi})^2\right] + \sigma^2,
\end{aligned} \tag{33}
$$

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

We first consider the following term

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

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

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

where $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}$.

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

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

Then we get

$$\begin{aligned}
i \neq j: \quad \mathbb{E}\left[\left(\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}} \boldsymbol{\Sigma} \bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}}\right)_{ij}\right] &= \boldsymbol{\Sigma}_{ij}\, \mathbb{E}[\bar{\boldsymbol{x}}_i^\top \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_j^\top \bar{\boldsymbol{x}}_j] + \boldsymbol{\Sigma}_{ji}\, \mathbb{E}[\bar{\boldsymbol{x}}_i^\top \bar{\boldsymbol{x}}_j \bar{\boldsymbol{x}}_i^\top \bar{\boldsymbol{x}}_j] = n^2 \boldsymbol{\Sigma}_{ij} + n \boldsymbol{\Sigma}_{ji} \\
i = j: \quad \mathbb{E}\left[\left(\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}} \boldsymbol{\Sigma} \bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}}\right)_{ii}\right] &= \boldsymbol{\Sigma}_{ii}\, \mathbb{E}\left[\bar{\boldsymbol{x}}_i^\top \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^\top \bar{\boldsymbol{x}}_i\right] + \sum_{j \neq i} \boldsymbol{\Sigma}_{jj}\, \mathbb{E}\left[\bar{\boldsymbol{x}}_i^\top \bar{\boldsymbol{x}}_j \bar{\boldsymbol{x}}_j^\top \bar{\boldsymbol{x}}_i\right] \\
&= \boldsymbol{\Sigma}_{ii}\, \mathbb{E}\left[(x_{i1}^2 + \cdots + x_{in}^2)^2\right] + n \sum_{j \neq i} \boldsymbol{\Sigma}_{jj} \\
&= \boldsymbol{\Sigma}_{ii}(3n + n(n-1)) + n \sum_{j \neq i} \boldsymbol{\Sigma}_{jj} \\
&= n\left(\boldsymbol{\Sigma}_{ii}(n+1) + \sum_{j=1}^{d} \boldsymbol{\Sigma}_{jj}\right) \\
&= n\left(\boldsymbol{\Sigma}_{ii}(n+1) + \mathtt{tr}\left(\boldsymbol{\Sigma}\right)\right).
\end{aligned}$$

Therefore

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

Combining all together results in

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W}) &= \mathtt{tr}\left(\boldsymbol{\Sigma}\right) - 2n\mathtt{tr}\left(\boldsymbol{\Sigma}\bar{\boldsymbol{W}}\right) + n(n+1)\mathtt{tr}\left(\boldsymbol{\Sigma}\bar{\boldsymbol{W}}^\top \bar{\boldsymbol{W}}\right) + n(\mathtt{tr}\left(\boldsymbol{\Sigma}\right) + \sigma^2)\mathtt{tr}\left(\bar{\boldsymbol{W}}\bar{\boldsymbol{W}}^\top\right) + \sigma^2, \\
&= M - 2n\mathtt{tr}\left(\boldsymbol{\Sigma}\bar{\boldsymbol{W}}\right) + n(n+1)\mathtt{tr}\left(\boldsymbol{\Sigma}\bar{\boldsymbol{W}}^\top \bar{\boldsymbol{W}}\right) + nM\mathtt{tr}\left(\bar{\boldsymbol{W}}\bar{\boldsymbol{W}}^\top\right),
\end{aligned} \tag{34}$$

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

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

Then we have

$$\boldsymbol{W}_\star = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1/2} \left((n+1)\boldsymbol{I} + M\boldsymbol{\Sigma}^{-1}\right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1/2}$$

and

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

■

B.3 Retrieval Augmented Generation with α Correlation

In this section, we consider the retrieval augmented generation (RAG) linear model similar to (9), where we first draw the query vector $\boldsymbol{x}$ and task vector $\boldsymbol{\beta}$ via

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

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

$$\boldsymbol{x}_i \,\Big|\, \boldsymbol{x} \sim \mathcal{N}(\alpha \boldsymbol{x}, \gamma^2 \boldsymbol{I}), \quad \xi_i \sim \mathcal{N}(0, \sigma^2) \quad \text{and} \quad y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \xi_i, \tag{35}$$

which results in (9) by setting $\gamma^2 = 1 - \alpha^2$.

Theorem 4 (Extended version of Theorem 2) *Consider linear model as defined in (35). Recap the objective from (5a) 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*

$$\boldsymbol{W}_\star = c\boldsymbol{I} \qquad \textit{and} \qquad \mathcal{L}_\star = d + \sigma^2 - cnd(\alpha^2(d+2) + \gamma^2) \tag{36}$$

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

$$\boldsymbol{W}_\star \approx \frac{1}{\kappa n + d + \sigma^2}\boldsymbol{I} \qquad \text{and} \qquad \mathcal{L}_\star \approx d + \sigma^2 - \frac{\kappa n d}{\kappa n + d + \sigma^2}. \tag{37}$$

Proof. Here, for clean notation and without loss of generality, we define and rewrite (35) via

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

Then we obtain

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W}) &= \mathbb{E}\left[(y - g(\boldsymbol{Z}))^2\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{x}^\top \boldsymbol{\beta} + \xi - \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi})\right)^2\right] \\
&= \mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{\beta} - \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta})^2 + 2(\boldsymbol{x}^\top \boldsymbol{\beta} - \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta})(\xi - \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{\xi}) + (\xi - \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{\xi})^2\right] \\
&= \mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{\beta} - \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta})^2\right] + \mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{\xi})^2\right] + \sigma^2.
\end{aligned} \tag{38}$$

To begin with, let

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

We first focus on the second term in (38)

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{X}^\top \boldsymbol{\xi})^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^n \xi_i \boldsymbol{x}^\top \boldsymbol{W}(\alpha \boldsymbol{x} + \gamma \boldsymbol{g}_i)\right)^2\right] \\
&= n\sigma^2\, \mathbb{E}\left[\boldsymbol{x}^\top \boldsymbol{W}(\alpha \boldsymbol{x} + \gamma \boldsymbol{g})(\alpha \boldsymbol{x} + \gamma \boldsymbol{g})^\top \boldsymbol{W}^\top \boldsymbol{x}\right] \\
&= n\sigma^2 \left(\alpha^2\, \mathbb{E}[\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{W}^\top \boldsymbol{x}] + \gamma^2\, \mathbb{E}[\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{g}\boldsymbol{g}^\top \boldsymbol{W}^\top \boldsymbol{x}]\right) \\
&= n\sigma^2 \left(\alpha^2 N_1 + \gamma^2 N_2\right). \qquad \text{(It follows (27) and independence of } \boldsymbol{x}, \boldsymbol{g}.)
\end{aligned}$$

Next, the first term in (38) can be decomposed into

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

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

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

$$\begin{aligned}
(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] \\
&\quad + 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 n d N_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 n N_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.
\end{aligned}$$

$$
\begin{aligned}
(c): \quad \mathbb{E}\left[\boldsymbol{x}^\top\boldsymbol{\beta}\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}\right] &= \mathbb{E}\left[\sum_{i=1}^n \boldsymbol{x}^\top\boldsymbol{\beta}\boldsymbol{x}^\top\boldsymbol{W}(\alpha\boldsymbol{x}+\gamma\boldsymbol{g}_i)(\alpha\boldsymbol{x}+\gamma\boldsymbol{g}_i)^\top\boldsymbol{\beta}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n \boldsymbol{x}^\top\boldsymbol{\beta}\boldsymbol{x}^\top\boldsymbol{W}(\alpha^2\boldsymbol{x}\boldsymbol{x}^\top+\gamma^2\boldsymbol{g}_i\boldsymbol{g}_i^\top+\alpha\gamma\boldsymbol{x}\boldsymbol{g}_i^\top+\alpha\gamma\boldsymbol{g}_i\boldsymbol{x}^\top)\boldsymbol{\beta}\right] \\
&= \alpha^2 n\,\mathbb{E}\left[\boldsymbol{x}^\top\boldsymbol{\beta}\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\beta}\right] + \gamma^2 n\,\mathbb{E}\left[\boldsymbol{x}^\top\boldsymbol{\beta}\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{g}\boldsymbol{g}^\top\boldsymbol{\beta}\right] \\
&= \alpha^2 n(d+2)\mathtt{tr}\,(\boldsymbol{W}) + \gamma^2 n\mathtt{tr}\,(\boldsymbol{W}) \\
&= \left(\alpha^2 n(d+2)+\gamma^2 n\right)N_3 \\
&= A_3 N_3.
\end{aligned}
$$

Here, (b) utilizes the 4'th and 6'th moment results (27) and (30) and we define

$$
\begin{aligned}
A_1 &= \alpha^4 n^2(d+4) + \alpha^2\gamma^2 n(2n+d+2) \\
A_2 &= \alpha^2\gamma^2 n(d+2) + \gamma^4 n(d+n+1) \\
A_3 &= \alpha^2 n(d+2) + \gamma^2 n.
\end{aligned}
$$

Then combining all together results in

$$
\mathcal{L}(\boldsymbol{W}) = A_1N_1 + A_2N_2 - 2A_3N_3 + n\sigma^2(\alpha^2N_1+\gamma^2N_2) + d + \sigma^2.
$$

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

$$
A_1\nabla N_1 + A_2\nabla N_2 - 2A_3\nabla N_3 + n\sigma^2(\alpha^2\nabla N_1 + \gamma^2\nabla N_2) = 0. \tag{39}
$$

Note that we have

$$
\begin{aligned}
\nabla N_1 &= \nabla\left(\mathtt{tr}\,(\boldsymbol{W})^2 + \mathtt{tr}\left(\boldsymbol{W}\boldsymbol{W}^\top\right) + \mathtt{tr}\left(\boldsymbol{W}^2\right)\right) = 2\mathtt{tr}\,(\boldsymbol{W})\,\boldsymbol{I} + 2\boldsymbol{W} + 2\boldsymbol{W}^\top \\
\nabla N_2 &= \nabla\mathtt{tr}\left(\boldsymbol{W}\boldsymbol{W}^\top\right) = 2\boldsymbol{W} \\
\nabla N_3 &= \nabla\mathtt{tr}\,(\boldsymbol{W}) = \boldsymbol{I}.
\end{aligned}
$$

Therefore, (39) returns

$$
2A_1\left(\mathtt{tr}\,(\boldsymbol{W})\,\boldsymbol{I} + \boldsymbol{W} + \boldsymbol{W}^\top\right) + 2A_2\boldsymbol{W} - 2A_3 + 2n\sigma^2(\alpha^2(\mathtt{tr}\,(\boldsymbol{W})\,\boldsymbol{I} + \boldsymbol{W} + \boldsymbol{W}^\top) + \gamma^2\boldsymbol{W})\boldsymbol{I} = 0, \tag{40}
$$

which implies that the optimal solution $\boldsymbol{W}_\star$ has the form of $c\boldsymbol{I}$ for some constant c. Then suppose $\boldsymbol{W}_\star = c\boldsymbol{I}$, we have $\mathtt{tr}\,(\boldsymbol{W}) = cd$ and (40) returns

$$
\begin{aligned}
&2A_1(d+2)c\boldsymbol{I} + 2A_2c\boldsymbol{I} - 2A_3\boldsymbol{I} + 2n\sigma^2(\alpha^2(d+2)c\boldsymbol{I} + \gamma^2c\boldsymbol{I}) = 0 \\
\implies c &= \frac{A_3}{A_1(d+2) + A_2 + n\sigma^2(\alpha^2(d+2)+\gamma^2)} \\
&= \frac{\alpha^2(d+2)+\gamma^2}{\alpha^4n(d+2)(d+4) + \alpha^2\gamma^2(d+2)(d+2n+3) + \gamma^4(d+n+1) + \sigma^2(\alpha^2(d+2)+\gamma^2)}.
\end{aligned}
$$

Then the optimal loss is obtained by setting $\boldsymbol{W}_\star = c\boldsymbol{I}$ and

$$
\begin{aligned}
\mathcal{L}_\star = \mathcal{L}(\boldsymbol{W}_\star) &= A_1c^2d(d+2) + A_2c^2d - 2A_3cd + n\sigma^2c^2d(\alpha^2(d+2)+\gamma^2) + d + \sigma^2 \\
&= c^2d\left(A_1(d+2) + A_2 + n\sigma^2(\alpha^2(d+2)+\gamma^2)\right) - 2A_3cd + d + \sigma^2 \\
&= d + \sigma^2 - A_3cd.
\end{aligned}
$$

It completes the proof of (36). Now if assuming $\alpha = O\left(1/\sqrt{d}\right)$, $d/n = O(1)$ and sufficiently large dimension d, we have the approximate

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

and

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

■

B.4 Task-feature Alignment with α Correlation

In this section, we consider the task-feature alignment data model similar to (11), 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 $\text{corr_coef}(\boldsymbol{x}_i, \boldsymbol{\beta}) \geq \alpha \geq 0$ via

$$\boldsymbol{x}_i \,\Big|\, \boldsymbol{\beta} \sim \mathcal{N}(\alpha\boldsymbol{\beta}, \boldsymbol{I}), \quad \xi_i \sim \mathcal{N}(0, \sigma^2) \quad \text{and} \quad y_i = \gamma \cdot \boldsymbol{x}_i^\top \boldsymbol{\beta} + \xi_i, \tag{41}$$

which results in (11) by setting $\gamma^2 = 1/(\alpha^2 d + 1)$.

Theorem 5 (Extended version of Theorem 3) *Consider linear model as defined in (41). Recap the objective from (5a) 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*

$$\boldsymbol{W}_\star = c\boldsymbol{I} \qquad \text{and} \qquad \mathcal{L}_\star = d\gamma^2(\Delta_0\alpha^2 + 1) + \sigma^2 - cnd\gamma^2(\Delta_1\alpha^4 + 2\Delta_0\alpha^2 + 1) \tag{42}$$

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

$$\boldsymbol{W}_\star \approx \frac{1}{\kappa n + (d+\sigma^2)/\kappa} \qquad \text{and} \qquad \mathcal{L}_\star \approx d + \sigma^2 - \frac{\kappa n d}{\kappa n + (d+\sigma^2)/\kappa}. \tag{43}$$

Proof. Here, for clean notation and without loss of generality, we define and rewrite (41) 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 (5a), we obtain

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W}) &= \mathbb{E}\left[(y - g(\boldsymbol{Z}))^2\right] \\
&= \mathbb{E}\left[\left(\gamma\boldsymbol{x}^\top\boldsymbol{\beta} + \xi - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top(\gamma\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi})\right)^2\right] \\
&= \mathbb{E}\left[\gamma^2(\boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})^2 + 2\gamma(\boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})(\xi - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi}) + (\xi - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi})^2\right] \\
&= \gamma^2\,\mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})^2\right] + \mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi})^2\right] + \sigma^2.
\end{aligned} \tag{44}$$

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

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

and additionally, given $\boldsymbol{\Lambda}_{\boldsymbol{W}} = \boldsymbol{W} \odot \boldsymbol{I}$, let

$$N_4 = 3\mathtt{tr}\left(\boldsymbol{\Lambda}_{\boldsymbol{W}}^2\right) + (d+4)\mathtt{tr}\left(\boldsymbol{W}\boldsymbol{W}^\top\right) + \mathtt{tr}\left(\boldsymbol{W}^2\right).$$

We first focus on the second term in (44)

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\xi})^2\right] &= \mathbb{E}\left[\left((\alpha\boldsymbol{\beta} + \boldsymbol{g})^\top\boldsymbol{W}\sum_{i=1}^n \xi_i(\alpha\boldsymbol{\beta} + \boldsymbol{g}_i)\right)^2\right] \\
&= n\sigma^2\,\mathbb{E}\left[\left((\alpha\boldsymbol{\beta} + \boldsymbol{g})^\top\boldsymbol{W}(\alpha\boldsymbol{\beta} + \boldsymbol{g}')\right)^2\right] \\
&= n\sigma^2\left(\alpha^4\,\mathbb{E}\left[(\boldsymbol{\beta}^\top\boldsymbol{W}\boldsymbol{\beta})^2\right] + 2\alpha^2\,\mathbb{E}\left[(\boldsymbol{\beta}^\top\boldsymbol{W}\boldsymbol{g}')^2\right] + \mathbb{E}\left[(\boldsymbol{g}^\top\boldsymbol{W}\boldsymbol{g}')^2\right]\right) \\
&= n\sigma^2\left(\alpha^4\left(\mathtt{tr}(\boldsymbol{W})^2 + \mathtt{tr}\left(\boldsymbol{W}^2\right) + \mathtt{tr}\left(\boldsymbol{W}\boldsymbol{W}^\top\right)\right) + (2\alpha^2 + 1)\mathtt{tr}\left(\boldsymbol{W}\boldsymbol{W}^\top\right)\right) \\
&= n\sigma^2\left(\alpha^4 N_1 + (2\alpha^2 + 1)N_2\right). \qquad \text{(It follows (27) and independence of } \boldsymbol{\beta}, \boldsymbol{g}, \boldsymbol{g}'.)
\end{aligned}$$

Next, the first term of (44) (omitting γ^2) returns the following decomposition:

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

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

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

$$
\begin{aligned}
\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} &= \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{\beta} \\
&= \sum_{i=1}^n (\alpha\boldsymbol{\beta} + \boldsymbol{g}_i)(\alpha\boldsymbol{\beta} + \boldsymbol{g}_i)^\top\boldsymbol{\beta} \\
&= \sum_{i=1}^n \alpha^2\boldsymbol{\beta}\boldsymbol{\beta}^\top\boldsymbol{\beta} + \alpha\boldsymbol{\beta}\boldsymbol{g}_i^\top\boldsymbol{\beta} + \alpha\boldsymbol{g}_i\boldsymbol{\beta}^\top\boldsymbol{\beta} + \boldsymbol{g}_i\boldsymbol{g}_i^\top\boldsymbol{\beta}.
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
(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] \\
&\quad + 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] \\
&\quad + 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 && (45) \\
&\quad + n(n-1)N_1 + nN_4 && (46) \\
&\quad + 2\alpha^2 n^2(d+4)N_1 + 2\alpha^2 n(d+4)N_1 && (47) \\
&= \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 + nN_4 && (48) \\
&= B_1N_1 + B_2N_2 + nN_4,
\end{aligned}
$$

where (45) and (47) utilize (30) and (32), and (46) is obtained via

$$
\begin{aligned}
\mathbb{E}\left[\left(\sum_{i=1}^n \boldsymbol{\beta}^\top\boldsymbol{W}\boldsymbol{g}_i\boldsymbol{g}_i^\top\boldsymbol{\beta}\right)^2\right] &= n\,\mathbb{E}\left[\left(\boldsymbol{\beta}^\top\boldsymbol{W}\boldsymbol{g}'\boldsymbol{g}'^\top\boldsymbol{\beta}\right)^2\right] + n(n-1)\,\mathbb{E}\left[\boldsymbol{\beta}^\top\boldsymbol{W}\boldsymbol{g}'\boldsymbol{g}'^\top\boldsymbol{\beta}\boldsymbol{\beta}^\top\boldsymbol{W}\boldsymbol{g}''\boldsymbol{g}''^\top\boldsymbol{\beta}\right] \\
&= nN_4 + n(n-1)N_1,
\end{aligned}
$$

which follows (27) and (28).

$$
\begin{aligned}
(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] \\
&\quad + 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] \\
&\quad + 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 \qquad (49) \\
&\quad + 2\alpha^2 n^2 (d+2) N_2 + 2\alpha^2 n (d+2) N_2 \qquad (50) \\
&= \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,
\end{aligned}
$$

where (49) and (50) are obtained using (27), (30) and

$$
\begin{aligned}
\mathbb{E}\left[\left(\sum_{i=1}^n \boldsymbol{g}^\top \boldsymbol{W}\boldsymbol{g}_i\boldsymbol{g}_i^\top\boldsymbol{\beta}\right)^2\right] &= n\, \mathbb{E}\left[\left(\boldsymbol{g}^\top \boldsymbol{W}\boldsymbol{g}'\boldsymbol{g}'^\top\boldsymbol{\beta}\right)^2\right] + n(n-1)\, \mathbb{E}\left[\boldsymbol{g}^\top \boldsymbol{W}\boldsymbol{g}'\boldsymbol{g}'^\top\boldsymbol{\beta}\boldsymbol{g}^\top \boldsymbol{W}\boldsymbol{g}''\boldsymbol{g}''^\top\boldsymbol{\beta}\right] \\
&= n(d+2)N_2 + n(n-1)N_2 = n(n+d+1)N_2.
\end{aligned}
$$

$$
\begin{aligned}
(c): \quad & \mathbb{E}\left[\boldsymbol{\beta}^\top\boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}\right] \\
&= n\, \mathbb{E}\left[\boldsymbol{\beta}^\top\boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{W}(\alpha\boldsymbol{\beta} + \boldsymbol{g}')(\alpha\boldsymbol{\beta} + \boldsymbol{g}')^\top\boldsymbol{\beta}\right] \\
&= \alpha^2 n\, \mathbb{E}\left[\boldsymbol{\beta}^\top\boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{W}\boldsymbol{\beta}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right] + n\, \mathbb{E}\left[\boldsymbol{\beta}^\top\boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{W}\boldsymbol{g}'\boldsymbol{g}'^\top\boldsymbol{\beta}\right] \\
&= \alpha^2 n (d+2)(d+4)\mathtt{tr}(\boldsymbol{W}) + n(d+2)\mathtt{tr}(\boldsymbol{W}) \\
&= \left(\alpha^2 n (d+2)(d+4) + n(d+2)\right) N_3 \\
&= B_4 N_3.
\end{aligned}
$$

$$
\begin{aligned}
(d): \quad & \mathbb{E}\left[\boldsymbol{\beta}^\top\boldsymbol{g}\boldsymbol{g}^\top \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}\right] \\
&= n\, \mathbb{E}\left[\boldsymbol{\beta}^\top\boldsymbol{g}\boldsymbol{g}^\top \boldsymbol{W}(\alpha\boldsymbol{\beta} + \boldsymbol{g}')(\alpha\boldsymbol{\beta} + \boldsymbol{g}')^\top\boldsymbol{\beta}\right] \\
&= \alpha^2 n\, \mathbb{E}\left[\boldsymbol{\beta}^\top\boldsymbol{g}\boldsymbol{g}^\top \boldsymbol{W}\boldsymbol{\beta}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right] + n\, \mathbb{E}\left[\boldsymbol{\beta}^\top\boldsymbol{g}\boldsymbol{g}^\top \boldsymbol{W}\boldsymbol{g}'\boldsymbol{g}'^\top\boldsymbol{\beta}\right] \\
&= \alpha^2 n (d+2)\mathtt{tr}(\boldsymbol{W}) + n\mathtt{tr}(\boldsymbol{W}) \\
&= \left(\alpha^2 n (d+2) + n\right) N_3 \\
&= B_5 N_3.
\end{aligned}
$$

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

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

and differentiating it results in

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

Similar to the proof in Appendix B.3, $\boldsymbol{W}_\star$ has the form of $\boldsymbol{W}_\star = c\boldsymbol{I}$ and we have

$$\begin{aligned}
\nabla N_1 &= \nabla\left(\mathtt{tr}\,(\boldsymbol{W})^2 + \mathtt{tr}\left(\boldsymbol{W}\boldsymbol{W}^\top\right) + \mathtt{tr}\left(\boldsymbol{W}^2\right)\right) = 2\mathtt{tr}\,(\boldsymbol{W})\,\boldsymbol{I} + 2\boldsymbol{W} + 2\boldsymbol{W}^\top = 2c(d+2)\boldsymbol{I} \\
\nabla N_2 &= \nabla\mathtt{tr}\left(\boldsymbol{W}\boldsymbol{W}^\top\right) = 2\boldsymbol{W} = 2c\boldsymbol{I} \\
\nabla N_3 &= \nabla\mathtt{tr}\,(\boldsymbol{W}) = \boldsymbol{I} \\
\nabla N_4 &= \nabla\left(3\mathtt{tr}\left(\boldsymbol{\Lambda}_{\boldsymbol{W}}^2\right) + (d+4)\mathtt{tr}\left(\boldsymbol{W}\boldsymbol{W}^\top\right) + \mathtt{tr}\left(\boldsymbol{W}^2\right)\right) \\
&= 6\cdot\mathrm{diag}\,(\boldsymbol{\Lambda}_{\boldsymbol{W}}) + 2(d+4)\boldsymbol{W} + 2\boldsymbol{W}^\top \\
&= 2c(d+8)\boldsymbol{I}.
\end{aligned}$$

Therefore, setting $\nabla\mathcal{L}(\boldsymbol{W}) = 0$ 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 $\boldsymbol{W}_\star = c\boldsymbol{I}$ and

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

It completes the proof of (42). 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

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

■

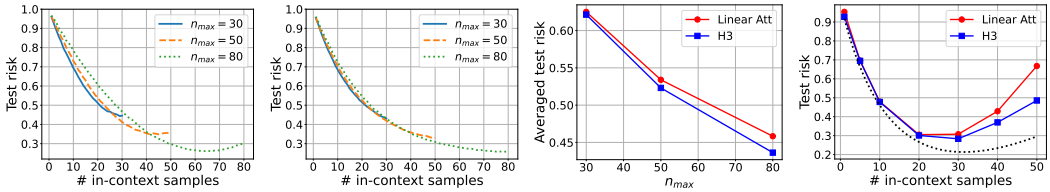
C Analysis of Low-Rank Parameterization

C.1 Proof of Lemma 3

Proof. Recall the loss function from (34)

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

where $\bar{\boldsymbol{W}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}\boldsymbol{W}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}$ and $M = \mathtt{tr}\,(\boldsymbol{\Sigma}) + \sigma^2$. For any $\bar{\boldsymbol{W}}$, let us parameterize $\bar{\boldsymbol{W}} = \boldsymbol{U}\boldsymbol{E}\boldsymbol{U}^\top$ where $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ denotes the eigenvectors of $\bar{\boldsymbol{W}}$ and $\boldsymbol{E} \in \mathbb{R}^{r\times r}$ is a symmetric square



(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)). Implementation details are discussed in Section 4.

matrix. We will first treat $\boldsymbol{U}$ as fixed and optimize $\boldsymbol{E}$. We will then optimize $\boldsymbol{U}$. Fixing $\boldsymbol{U}$, setting $\bar{\boldsymbol{\Sigma}} = \boldsymbol{U}^\top \boldsymbol{\Sigma} \boldsymbol{U}$, we obtain

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

Differentiating, we obtain

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

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

$$\boldsymbol{E}_\star = (M\boldsymbol{I} + (n+1)\bar{\boldsymbol{\Sigma}})^{-1}\bar{\boldsymbol{\Sigma}}. \tag{51}$$

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

$$\mathcal{L}_\star(\boldsymbol{U}) = M - n \cdot \mathtt{tr}\left(\bar{\boldsymbol{\Sigma}}\boldsymbol{E}_\star\right) = M - n \cdot \mathtt{tr}\left((M\boldsymbol{I} + (n+1)\bar{\boldsymbol{\Sigma}})^{-1}\bar{\boldsymbol{\Sigma}}^2\right) \tag{52}$$

$$= M - n\sum_{i=1}^{r} \frac{\bar{\lambda}_i^2}{(n+1)\bar{\lambda}_i + M} = M - n\sum_{i=1}^{r} \frac{\bar{\lambda}_i}{n + 1 + M\bar{\lambda}_i^{-1}}. \tag{53}$$

Now observe that, the right hand side is strictly decreasing function of the eigenvalues $\bar{\lambda}_i$ of $\bar{\boldsymbol{\Sigma}} = \boldsymbol{U}^\top\boldsymbol{\Sigma}\boldsymbol{U}$. Thus, to minimize $\mathcal{L}_\star(\boldsymbol{U})$, 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 $\boldsymbol{\Sigma}$ since $\bar{\boldsymbol{\Sigma}}$ is an orthogonal projection of $\boldsymbol{\Sigma}$ on $\boldsymbol{U}$. 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 $\boldsymbol{U}$ to be the top-r eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{E} = \boldsymbol{E}_\star(\boldsymbol{U})$ to be the diagonal matrix according to (51). ■

D Additional Experiments

In this section, we present additional experiments demonstrating that the H3 model can outperform the linear attention model under different training or data settings. The implementation details are consistent with those outlined in Section 4.

• **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. 4a and 4b, 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 [Garg et al., 2022,

Akyürek et al., 2023, Li et al., 2023]. We generate data according to (7) with $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \boldsymbol{I}_d$ and $\sigma = 0$, and train 1-layer linear attention (Fig. 4a) and H3 (Fig. 4b) models with different max-lengths $n_{\max} = 30, 50, 80$. Comparison between Fig. 4a and 4b shows that 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. 4c, we plot the averaged risks for each model and H3 outperforms linear attention. Furthermore, in Fig. 4d, 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. 4d 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 1.

E Extended Related Work

There is growing interest in understanding the mechanisms behind ICL [Brown et al., 2020, Liu et al., 2023b, Rae et al., 2021] in large language models (LLMs) due to its success in continuously enabling novel applications for LLMs [GeminiTeam et al., 2023, OpenAI, 2023, Touvron et al., 2023]. Towards this, Xie et al. [2022] 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. Müller et al. [2021], Hollmann et al. [2022], Müller et al. [2023] 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. [2022] posit induction heads as the key mechanism enabling ICL in Transformers. Park et al. [2024] study how various distributional properties of training data aid in the emergence of ICL in Transformers.

In the previous work, Garg et al. [2022] 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 [von Oswald et al., 2023, Gatmiry et al., Collins et al., 2024, Lin and Lee, 2024, Li et al., 2024, Bai et al., 2024, Akyürek et al., 2023, Zhang et al., 2023, Du et al., 2023]. Akyürek et al. [2023] 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. [2023] 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. [2023] 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 [Lee et al., 2023, Zucchet et al., 2023, Grazzi et al., 2024] demonstrate that other model architectures, such as SSM and RNNs, are also capable of in-context learning (ICL).

Building on these primarily empirical studies, Zhang et al. [2024], Mahankali et al. [2024], Ahn et al. [2023], Duraisamy [2024] 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, Mahankali et al. [2024], Ahn et al. [2023] show that the resulting model implements a single step of GD on in-context examples in a test prompt, thereby corroborating the findings of [Von Oswald et al., 2023]. They also show that the learned model implements a PGD step, when faced with anisotropic Gaussian features, with Mahankali et al. [2024] also considering anisotropic Gaussian weight vectors. Ahn et al. [2023] further study multi-layer model and show that the trained model can implement a generalization of GD++ algorithm, supporting an empirical observation in Von Oswald et al. [2023]. On the other hand, Mahankali et al. [2024] 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. [2024] 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 Mahankali et al. [2024], Ahn et al. [2023], 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. [2024] 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 [Lewis et al., 2020, Nakano et al., 2021] 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 [Izacard et al., 2023], which can significantly outperform the typical ICL setup where fixed task-specific demonstrations are provided as in-context examples [Wang et al., 2022, Basu et al., 2023]. 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) [Gu et al., 2021b,a, Fu et al., 2023, Gu and Dao, 2023] 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 [Park et al., 2024, Grazzi et al., 2024] as well as complex NLP tasks [Grazzi et al., 2024]. That said, a rigorous theoretical understanding of ICL for SSMs akin to Zhang et al. [2024], Mahankali et al. [2024], Ahn et al. [2023] is missing from the literature. In this work, we provide the first such theoretical treatment for ICL with SSMs. Focusing on H3 architecture [Fu et al., 2023], we highlight its advantages over linear attention in specific ICL settings.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: **All the theoretical contributions claimed in the abstract and introduction along with the underlying data model are presented in Section 2 and Section 3.**

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: **This is a theoretical study which (similar to prior studies in the field) relies on a precise but simplified data model to draw quantitatively precise conclusions. All the assumptions on the data model are clearly stated in Section 2 and Section 3. We have also added a paragraph after the conclusion to specifically highlight various limitations of our work.**

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

Answer: [Yes]

Justification: **As discussed above, for all of our theoretical results and proofs we state the precise setup and assumptions in Section 2 and Section 3. Due to page limit, proofs are deferred to the supplemental material.**

Guidelines:

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

4. **Experimental Result Reproducibility**

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

Answer: [Yes]

Justification: **This is primarily a theoretical work where detailed synthetic experiments (on the same data model studied in our theoretical analysis) have been conducted to corroborate our theoretical findings. We provide sufficient details in Section 4 for reproducing these experiments.**

Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: **As discussed above, this paper conducts small scale synthetic experiments to corroborate our theoretical findings. We have provided sufficient details to reproduce these experiments in Section 4.**

   Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: **All the relevant details for our small scale experiments are provided in Section 4.**

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [No]

Justification: **This work is a theoretical work studying the optimization landscape of linear attention/H3 under population risk. Then our goal of simulations is to find the optimal solution corresponding to the minimal risks. Therefore, we do not report the error bars and we have included the discussion in the experiment section.**

Guidelines:

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

8. **Experiments Compute Resources**

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

Answer: [No]

Justification: **Our work only focuses on 1-layer attention/H3 model training with hidden dimension** 21 **and maximal context length** < 100**, which can be implemented easily.**

Guidelines:

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

9. **Code Of Ethics**

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

Answer: [Yes]

Justification: **Yes, the authors confirm that the research conducted in the paper conform wiht the NeurIPS Code of Ethics.**

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

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

    Answer: [Yes]

    Justification: **In its current form, we don't see any specific negative impacts of our theoretical study. However, we have discussed potential broader impacts of the future extensions of this work, e.g., the ones tied to eliciting undesirable behavior of LLMs with in-context learning, while discussing the limitations of the work after the conclusion section.**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: **The synthetic setup studied in the paper does not pose such risks.**

    Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [NA]

Justification: **The paper does not rely on existing assets.**

Guidelines:

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

13. **New Assets**

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

Answer:[NA]

Justification: **The paper does not release new assets such as code, data, or models.**

Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

Answer: [NA]

Justification: **The paper does not involve crowdsourcing nor research with human subjects.**

Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: **The paper does not involve crowdsourcing nor research with human subjects.**

    Guidelines:

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