Mamba Can Learn Low-Dimensional Targets In-Context via Test-Time Feature Learning

Junsoo Oh*
KAIST AI
junsoo.oh@kaist.ac.kr

Wei Huang
RIKEN AIP/The Institute of Statistical Mathematics
wei.huang.vr@riken.jp

Taiji Suzuki
The University of Tokyo/RIKEN AIP
taiji@mist.i.u-tokyo.ac.jp

Abstract

Mamba, a recently proposed linear-time sequence model, has attracted significant attention for its computational efficiency and strong empirical performance. However, a rigorous theoretical understanding of its underlying mechanisms remains limited. In this work, we provide a theoretical analysis of Mamba's in-context learning (ICL) capability by focusing on tasks defined by low-dimensional nonlinear target functions. Specifically, we study in-context learning of a single-index model $y \approx g_*(\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle)$, which depends on only a single relevant direction $\boldsymbol{\beta}$, referred to as *feature*. We prove that Mamba, pretrained by gradient-based methods, can achieve efficient ICL via *test-time feature learning*, extracting the relevant direction directly from context examples. Consequently, we establish a test-time sample complexity that improves upon linear Transformers—analyzed to behave like kernel methods—and is comparable to nonlinear Transformers, which have been shown to surpass the Correlational Statistical Query (CSQ) lower bound and achieve near information-theoretically optimal rate in previous works. Our analysis reveals the crucial role of the *nonlinear gating* mechanism in Mamba for feature extraction, highlighting it as the fundamental driver behind Mamba's ability to achieve both computational efficiency and high performance.

1 Introduction

Mamba (Gu and Dao, 2024), a recently proposed state space model, has rapidly gained attention for its remarkable balance of computational efficiency and empirical performance. By replacing the quadratic-time attention mechanism of Transformers (Vaswani et al., 2017) with a selective state-space recurrence with nonlinear gating, Mamba enables scalable modeling of long sequences while maintaining competitive accuracy across a variety of tasks (Dao and Gu, 2024; Waleffe et al., 2024; Wang et al., 2024; Patro and Agneeswaran, 2025). Despite Mamba's remarkable computational efficiency, it remains unknown whether it can exhibit strong adaptability (often referred to as *feature learning*), a property widely recognized as critical to the success of deep learning neural networks (Girshick et al., 2014; Suzuki, 2019; Damian et al., 2022).

A key benchmark for test-time adaptability is in-context learning (ICL) (Brown et al., 2020), which has emerged as a canonical paradigm for understanding the adaptability of large language models and sequence architectures. By conditioning on context examples provided in the input prompt, a model can achieve strong performance on new tasks at test time without explicit parameter updates. While the empirical effectiveness of ICL is well documented, theoretical understanding of when

*Based on work performed during internship at the University of Tokyo

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and how different architectures exhibit this behavior remains limited (Xie et al., 2022; Garg et al., 2022; Zhou et al., 2024). In particular, most existing theoretical analyses focus on Transformers (Ahn et al., 2023; Zhang et al., 2024; Mahankali et al., 2024; Huang et al., 2024; Kim and Suzuki, 2024), whose quadratic attention mechanisms make them both powerful and computationally demanding. It remains unclear whether alternative architectures such as Mamba can offer comparable adaptability.

Recent works have investigated Mamba's ICL capabilities, empirically demonstrating that Mamba performs competitively across various ICL benchmarks (Grazzi et al., 2024; Park et al., 2024; Li et al., 2024c). However, our understanding of Mamba's ICL capabilities remains lacking. This is due to its distinct inductive bias from the Transformer. The recurrent state-space model with nonlinear gating processes inputs through recurrent updates that maintain and transform hidden states over time, rather than relying on global attention over the entire context. This distinction motivates new theoretical questions:

Can Mamba provably achieve strong test-time adaptability like Transformers with its recurrent state-space updates and nonlinear gating?

1.1 Summary of Contributions

In this paper, we study the ICL capabilities of Mamba, focusing on a single-index model—a widely adopted theoretical tool for studying adaptability. We summarize our contributions as follows:

- We introduce a theoretical framework for analyzing Mamba's ICL of single-index models, including input embeddings, the Mamba architecture, and a gradient-based pretraining algorithm. Under this framework, we characterize the optimization dynamics and establish the sample complexity in terms of the number of pretraining tasks and the number of context examples at pretraining and test time required to achieve strong performance (Theorem 3.3).
- Our analysis reveals that pretrained Mamba is capable of *test-time feature learning*, enabling it to extract task-relevant features directly from context examples (Proposition 4.1). This result implies that Mamba can surpass the performance of kernel regression baselines and achieve adaptation at test time. Specifically, the gating mechanism enables Mamba to achieve test-time feature learning, thereby overcoming the limitations inherent to purely linear recurrent updates.
- We provide a comparative analysis between Mamba and Transformer, highlighting similarities and differences in their ICL mechanisms. Our results reveal that Mamba can achieve test-time feature learning via a qualitatively different mechanism—recurrent state-space updates with nonlinear gating—thus extending the theoretical landscape of ICL beyond attention-based models.

1.2 Related Works

Theory of In-Context Learning. Theoretical investigations of ICL have predominantly centered on Transformers. Beyond initial results showing that Transformers trained on regression tasks can reproduce ordinary least squares solutions in-context (Akyürek et al., 2023; Zhang et al., 2024; Mahankali et al., 2024; Han et al., 2025), subsequent analyses reveal their ability to emulate more complex procedures such as multi-step gradient descent (Ahn et al., 2023; Saunshi et al., 2025), functional gradient descent (Cheng et al., 2024), and sparse regression (Bai et al., 2023). Parallel works extend this line of inquiry to classification, where recent studies provide provable insights into how Transformers implement in-context classification (Li et al., 2024a; Bu et al., 2025).

While the theoretical literature on ICL has dominantly focused on Transformers, a growing body of work is extending this theoretical analysis to linear-time sequence models. Recent works (Li et al., 2024b, 2025b) prove that H3-like model (Fu et al., 2023) and gated linear attention (Yang et al., 2024) can implement weighted preconditioned gradient descent based on loss landscape analysis. Bondaschi et al. (2025) study ICL of Mamba on Markov chains and show that it learns a Laplacian smoothing estimator in-context. However, these works do not provide optimization or generalization guarantees. Recent works by Jiang et al. (2025) and Li et al. (2025a) provide such guarantees for ICL of Mamba on linear regression and classification tasks with outliers, as we also do in this work.

Learning Low-Dimensional Target Function. Low-dimensional target function classes, such as sparse parities (Barak et al., 2022; Suzuki et al., 2023; Glasgow, 2024), signal-noise models (Allen-Zhu and Li, 2020; Cao et al., 2022), are widely adopted as theoretical benchmarks for studying a

neural network's ability to perform feature learning. This work specifically focuses on the single-index model. A line of theoretical work has analyzed the learning of these models and has established key results on sample complexity. The required sample complexity is governed by either the *information exponent* (for algorithms utilizing correlational information (Ben Arous et al., 2021; Bietti et al., 2022; Damian et al., 2023; Mousavi-Hosseini et al., 2023)) or the *generative exponent* (for algorithms that employ suitable label transformations (Damian et al., 2024; Lee et al., 2024; Arnaboldi et al., 2024; Joshi et al., 2024)). We discuss these sample complexity results in more detail in Section 3.1.

Our work is most closely related to Oko et al. (2024); Nishikawa et al. (2025), which lie at the intersection of ICL and the single-index model. Specifically, Oko et al. (2024) show that a pretrained linear Transformer can effectively learn a single-index model in-context. More recent work by Nishikawa et al. (2025) establish an even smaller sample complexity and reveal that the nonlinear Transformer can perform test-time feature learning. A detailed comparison is provided in Section 3.2.

2 Problem Setting

In this section, we provide a formal description of the key components we focus on: the ICL data distribution, the Mamba model, and the gradient-based pretraining algorithm.

Notation. We denote the i-th coordinate of a vector $\boldsymbol{v}$ as $\boldsymbol{v}[i]$, and the (i, j)-th coordinate of a matrix $\boldsymbol{M}$ as $\boldsymbol{M}[i, j]$. The matrix $\mathrm{diag}(\boldsymbol{v})$ represents a diagonal matrix with a vector $\boldsymbol{v}$ on its main diagonal. We use $\odot$ for the element-wise product. For any $k \in \mathbb{N}$, we denote the vectors with all entries equal to one and zero as $\mathbf{1}_k$ and $\mathbf{0}_k$, respectively. We omit the subscript k when it is clear.

2.1 Data Distribution for In-Context Learning

ICL aims to solve the *task* of predicting the *label* y of a *query* $\boldsymbol{x}$ using a sequence of input-label pairs $\{(\boldsymbol{x}_i, y_i)\}_{i\in[N]}$, which are referred to as *context examples*. The model then utilizes a prompt, which is a sequence $(\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_N, y_N, \boldsymbol{x})$ consisting of the context examples and the query, as its input. We focus on prompts constructed from the Gaussian single-index model, defined as follows.

Definition 2.1 (Gaussian Single-Index Model)**.** Given a feature vector $\boldsymbol{\beta} \in \mathbb{R}^d$, we draw input-label pairs $(\boldsymbol{x}, y) \sim \mathcal{D}_{\boldsymbol{\beta}}$ as

$$\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d), \quad y = g_*(\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle) + \zeta, \quad \zeta \sim \mathrm{Unif}(\{-\tau, \tau\}),$$

where g_* is a polynomial *link function* and $\tau > 0$ represents the noise level. For simplicity, we assume that $\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,1)}[g_*(\mathbf{z})] = 0, \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,1)}[(g_*(\mathbf{z}))^2] = 1$ and τ is a small enough constant.

For each task, a prompt is constructed with a random choice of feature vectors.

Definition 2.2 (ICL Data Distribution)**.** For given a context length N, we define a data distribution $\mathcal{D}(N)$ such that $(\boldsymbol{\beta}, \{(\boldsymbol{x}_i, y_i)\}_{i\in[N]}, \boldsymbol{x}, y) \sim \mathcal{D}(N)$ is constructed as follows.

1. We draw the feature vector $\boldsymbol{\beta} \in \mathbb{R}^d$ uniformly from the unit sphere S_r of a low-dimensional *intrinsic feature space* with dimension r, defined as:

$$S_r := \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\| = 1, \boldsymbol{\theta}[j] = 0 \text{ for all } j \notin \mathcal{I} \right\},$$

   for some unknown feature index set $\mathcal{I}$ with $|\mathcal{I}| = r$.

2. We sample N context examples $\{(\boldsymbol{x}_i, y_i)\}_{i\in[N]}$ and a query-label pair $(\boldsymbol{x}, y)$ from $\mathcal{D}_{\boldsymbol{\beta}}$.

Our task distribution exhibits a low-dimensional structure in two key aspects: (1) the label depends solely on the projection of the input onto the feature vector, and (2) feature vectors are supported on an r-dimensional subspace. We note that to achieve low prediction errors, it is crucial to extract both of these structures and estimate the link function g_*.

2.2 Prediction Model Architecture

We consider a simplified prediction model for ICL consisting of three components: an input embedding, a single Mamba layer, and a multi-layer perceptron (MLP).

Input Embedding. Given a prompt $(\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_N, y_N, \boldsymbol{x})$ with context length N and label y, we construct an input embedding $\boldsymbol{Z} \in \mathbb{R}^{(\tilde{d}+1)\times(N+1)}$ as

$$\boldsymbol{Z} = \begin{bmatrix} \phi(\boldsymbol{x}_1) & \phi(\boldsymbol{x}_2) & \ldots & \phi(\boldsymbol{x}_N) & \phi(\boldsymbol{x}) \\ y_1 & y_2 & \ldots & y_N & 0 \end{bmatrix} = [\boldsymbol{z}_1, \ldots \boldsymbol{z}_N, \boldsymbol{z}_{N+1}] \in \mathbb{R}^{(\tilde{d}+1)\times(N+1)},$$

where $\tilde{d} = 1 + d + \frac{d(d+1)}{2}$ and $\phi : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ is defined as

$$\phi(\boldsymbol{\theta}) = \left[1, \boldsymbol{\theta}, \frac{\boldsymbol{\theta} \odot \boldsymbol{\theta} - \mathbf{1}_d}{\sqrt{2}}, \boldsymbol{\theta}[1]\boldsymbol{\theta}[2], \ldots, \boldsymbol{\theta}[d-1]\boldsymbol{\theta}[d]\right].$$

An input embedding similar to ours was also considered in the recent work by Sun et al. (2025), who studied the in-context learning of high-order polynomial target functions. They showed this embedding can be implemented with a simple version of Gated Linear Unit (GLU) and demonstrated its efficacy for enabling linear Transformers to learn these functions in-context. Unlike Sun et al. (2025), who repeatedly stacked a linear Transformer and a GLU layer, in our work, a single GLU-based embedding is sufficient due to the nonlinearity in Mamba and MLP layers. We discuss the efficacy of this input embedding in more detail in Section 4.1.

Remark 2.3. Our input embedding is based on a basis for degree-2 polynomials in $\mathbb{R}^d$. Specifically, we use the standard basis of $\mathbb{R}^d$ for the construction of both the input embedding and the intrinsic feature space S_r. While extending our results to a general choice of S_r with an arbitrary basis may require additional techniques, our setting remains valuable for studying Mamba's ability to learn low-dimensional structure. For non-even link functions, a similar approach may also allow an extension to the standard input embedding $\phi(\boldsymbol{x}) = \boldsymbol{x}$ used in prior works (Von Oswald et al., 2023).

One-Layer Mamba. Given an input embedding $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{N+1}) \in \mathbb{R}^{(\tilde{d}+1)\times(N+1)}$, a one-layer Mamba model $\mathsf{Mamba}(\cdot; \boldsymbol{\Theta})$ with parameters $\boldsymbol{\Theta}$ has sequential outputs $\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{N+1} \in \mathbb{R}^{\tilde{d}+1}$ and hidden states for i-th channel $\boldsymbol{h}_1^{(i)}, \ldots, \boldsymbol{h}_{N+1}^{(i)} \in \mathbb{R}^{\tilde{d}+1}$ defined as below:

$$\boldsymbol{h}_l^{(i)} = \overline{\boldsymbol{A}}_l \boldsymbol{h}_{l-1}^{(i)} + \overline{\boldsymbol{B}}_l \boldsymbol{z}_l[i] \in \mathbb{R}^{\tilde{d}+1}, \quad \boldsymbol{o}_l[i] = \boldsymbol{C}_l^\top \boldsymbol{h}_l^{(i)} \in \mathbb{R},$$
$$\overline{\boldsymbol{A}}_l = \exp(\Delta_l \boldsymbol{A}) \in \mathbb{R}^{(\tilde{d}+1)\times(\tilde{d}+1)}, \quad \overline{\boldsymbol{B}}_l = (\Delta_l \boldsymbol{A})^{-1} (\exp(\Delta_l \boldsymbol{A}) - \boldsymbol{I}_{\tilde{d}+1}) \Delta_l \boldsymbol{B}_l \in \mathbb{R}^{\tilde{d}+1},$$

where $\boldsymbol{h}_0^{(i)} = \mathbf{0}_{\tilde{d}+1}$. Here, the components of the selection algorithm $\boldsymbol{A}, \boldsymbol{B}_l, \boldsymbol{C}_l, \Delta_l$ are chosen as

$$\boldsymbol{A} = -\boldsymbol{I}_{\tilde{d}+1}, \quad \boldsymbol{B}_l = \boldsymbol{W}_B \boldsymbol{z}_l, \quad \boldsymbol{C}_l = \boldsymbol{W}_C \boldsymbol{z}_l, \quad \Delta_l = \mathrm{softplus}\left(\boldsymbol{w}^\top \boldsymbol{z}_l + b\right),$$

with $\boldsymbol{W}_B, \boldsymbol{W}_C \in \mathbb{R}^{(\tilde{d}+1)\times(\tilde{d}+1)}, \boldsymbol{w} \in \mathbb{R}^{\tilde{d}+1}, b \in \mathbb{R}$. Then, the l-th output can be expressed as

$$\boldsymbol{o}_l = \sum_{j=1}^{l} G_{j,l}(\boldsymbol{Z}) \boldsymbol{z}_j \boldsymbol{z}_j^\top \boldsymbol{W}_B^\top \boldsymbol{W}_C \boldsymbol{z}_l, \tag{1}$$

where $G_{j,l}(\boldsymbol{Z}) = \sigma\left(\boldsymbol{w}^\top \boldsymbol{z}_j + b\right) \prod_{k=j+1}^{l} \left(1 - \sigma\left(\boldsymbol{w}^\top \boldsymbol{z}_k + b\right)\right)$ with sigmoid function $\sigma(\cdot)$. It implies that Mamba involves two key mechanisms: *nonlinear gating* $G_{j,l}(\boldsymbol{Z})$ and *linear attention* with projection matrices $\boldsymbol{W}_B$ and $\boldsymbol{W}_C$. Yang et al. (2024) refer the combination of these mechanisms as *gated linear attention* and recent recurrent models including Mamba, mLSTM (Beck et al., 2024), and RWKV-6 (Peng et al., 2024) can be viewed within this framework.

To ensure a tractable optimization guarantee, we further introduce the following simplifications to our model, while preserving the essential mechanism:

$$\boldsymbol{W}_B^\top \boldsymbol{W}_C = \mathrm{diag}(\boldsymbol{\gamma}, 0), \quad \boldsymbol{w} = \begin{bmatrix} \mathbf{0}_{\tilde{d}} \\ \rho^{-1} \end{bmatrix},$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{\tilde{d}}$ is a learnable parameter, and $\rho, b \in \mathbb{R}$ are fixed (as detailed in Theorem 3.3). Our approach of merging the product of two learnable matrices into a single matrix and using sparse learnable parameters is a technique also adopted in the theoretical literature on optimization of attention mechanisms (Ahn et al., 2023; Zhang et al., 2024; Mahankali et al., 2024; Kim and Suzuki, 2024). We fix the essential last coordinate of $\boldsymbol{w}$ to a sufficiently small constant, closely approximating a random initialization, and set the remaining coordinates to zero to avoid unnecessarily complicated

calculations. Additionally, we fix b to a large negative value so that the product factor in $G_{j,l}$ is close to one. The remaining factor $\sigma(y_j/\rho + b)$ still transforms the context labels nonlinearly. Under this simplification, the last coordinate of the final output which serves as the input to the MLP can be expressed as follows:

$$\mathsf{Mamba}(\boldsymbol{Z};\boldsymbol{\gamma})[\tilde{d}+1, N+1] = \sum_{j=1}^{N} G_{j,N+1}(\boldsymbol{Z}) y_j \phi(\boldsymbol{x}_j)^\top (\boldsymbol{\gamma} \odot \phi(\boldsymbol{x})) .$$

Multi-Layer Perceptron. We use a two-layer MLP with ReLU activation, width m and parameters $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{a} \in \mathbb{R}^m$ defined as follows:

$$\mathsf{MLP}(z; \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{a}) := \sum_{k=1}^{m} \boldsymbol{u}[k] \mathrm{ReLU}\left(\boldsymbol{v}[k] z + \boldsymbol{a}[k]\right) .$$

We apply this MLP to the output of the Mamba layer, after normalizing it by its context length N. Then, the final output is given by

$$f(\boldsymbol{Z}; \boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{a}) := \mathsf{MLP}\left(N^{-1}\mathsf{Mamba}(\boldsymbol{Z};\boldsymbol{\gamma})[\tilde{d}+1, N+1]; \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{a}\right)$$

Remark 2.4. A similar structure to our models, which combines a sequence model with a MLP, has also been utilized in two closely related prior works. Nishikawa et al. (2025) follow a similar structure but use a softmax Transformer in place of Mamba. In contrast, Oko et al. (2024) apply the MLP to the input embedding before a linear Transformer, rather than after the sequence model.

Our goal for ICL is to find parameters $\boldsymbol{\gamma} \in \mathbb{R}^{\tilde{d}}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{a} \in \mathbb{R}^m$, and context length N, achieving a small ICL test error, which is defined as

$$\mathcal{R}_N(\boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{a}) := \mathbb{E}_{(\boldsymbol{Z},y)\sim\mathcal{D}(N)}[|f(\boldsymbol{Z}, \boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{a}) - y|].$$

Here, we abuse notation and use $(\boldsymbol{Z}, y)$ to denote the input embedding and label for a prompt sampled from the ICL data distribution $\mathcal{D}(N)$. More precisely, we are interested in the sample complexity of context examples required for the parameters learned from pretraining to achieve a low ICL test error.

2.3 Pretraining Algorithm

Our prediction model is pretrained on a set of $T_{\mathrm{pt}} = T_1 + T_2$ tasks with context length N_{pt} drawn from $\mathcal{D}(N_{\mathrm{pt}})$. For each task $t \in [T_{\mathrm{pt}}]$, we have an input embedding $\boldsymbol{Z}^t$ constructed from context examples $\{(\boldsymbol{x}_i^t, y_i^t)\}_{i\in[N_{\mathrm{pt}}]}$ and a query-label pair $(\boldsymbol{x}^t, y^t)$ with a feature vector $\boldsymbol{\beta}^t$. Then, our training losses can be written as

$$L_l(\boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{a}) := \frac{1}{T_l} \sum_{t=T_{l-1}+1}^{T_{l-1}+T_l} \left(f\left(\boldsymbol{Z}^t; \boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{a}\right) - y^t\right)^2 ,$$

for $l = 1, 2$ and $T_0 = 0$. Using these objectives, we employ Algorithm 1 which proceeds in two stages: Stage I recovers intrinsic feature structures, enabling Stage II to perform in-context link estimation conditioned on these features. This approach enables tractable analysis and has been widely adopted in prior works (Ba et al., 2022; Damian et al., 2022).

1. In Stage I, we only train the Mamba layer parameter $\boldsymbol{\gamma}$, starting from proper initialization. Our training objective is ℓ_2-regularized loss $L_1(\boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{a}) + \frac{\lambda_1}{2}\|\boldsymbol{\gamma}\|^2$. Due to the non-linearity introduced by the MLP, this objective is non-convex. To make the training dynamics tractable, we apply one-step gradient descent, following the approaches studied in the literature of feature learning (Ba et al., 2022; Damian et al., 2022). As we describe in Section 4.1, a single step update is sufficient to capture the low-dimensional structure of the feature vectors.
2. In Stage II, we fix the Mamba layer parameter $\boldsymbol{\gamma}^*$ obtained from Stage I and optimize the outer layer $\boldsymbol{u}$ of MLP on ℓ_2-regularized loss $L_2(\boldsymbol{\gamma}^*, \boldsymbol{u}, \boldsymbol{v}^*, \boldsymbol{a}^*) + \frac{\lambda_2}{2}\|\boldsymbol{u}\|^2$ with reinitialized inner layer parameters $\boldsymbol{v}^*, \boldsymbol{a}^*$. This induces a convex problem that gradient-based methods can solve. As we show in Section 4.2, the optimized MLP is capable of estimating the link function g_*.

Algorithm 1: Gradient-based Pretraining of the Mamba Model

Input : Learning rate η, weight decay λ_1, λ_2, context length N_{pt}, the number of tasks T_1, T_2, initialization scale γ, ρ, b.

Stage I: Gradient descent on Mamba layer

  Initialize $\boldsymbol{\gamma}^{(0)} = (\gamma^2, 1, \ldots, 1, \gamma, \cdots, \gamma), \boldsymbol{u}^{(0)} = m^{-1}\mathbf{1}_m, \boldsymbol{v}^{(0)} = \mathbf{1}_m, \boldsymbol{a}^{(0)} = \mathbf{0}_m$.

  $\boldsymbol{\gamma}^* \leftarrow \boldsymbol{\gamma}^{(0)} - \eta \nabla_{\boldsymbol{\gamma}} \left(L_1(\boldsymbol{\gamma}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{a}^{(0)}) + \frac{\lambda_1}{2} \|\boldsymbol{\gamma}\|^2 \right)$.

Stage II: Optimization of MLP Layer

  Initialize $\boldsymbol{v}^* \sim \mathrm{Unif}\left(\{\pm 1\}^m\right)$, $\boldsymbol{a}^* \sim \mathrm{Unif}([-1, 1]^m)$.

  $\boldsymbol{u}^* \leftarrow \arg\min_{\boldsymbol{u}} \left(L_2(\boldsymbol{\gamma}^*, \boldsymbol{u}, \boldsymbol{v}^*, \boldsymbol{a}^*) + \frac{\lambda_2}{2} \|\boldsymbol{u}\|^2 \right)$.

Output : Prediction function $f(\cdot; \boldsymbol{\gamma}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, \boldsymbol{a}^*)$.

3 Mamba Efficiently Learns Single-Index Models In-Context

In this section, we present our theoretical results. Our analysis focuses on the asymptotic dependencies on the input dimension d, while the link function g_* is fixed. For our analysis, we assume that the intrinsic dimension satisfies $r \geq (\log d)^{C_r}$ for a sufficiently large constant $C_r > 0$. This technical condition controls residual terms in our analysis, and removing it appears to require additional techniques. This condition still allows the low-dimensional regime $r \ll d$. We let N^* and T^* be the maximum admissible context length and the number of pretraining tasks, respectively. We assume that $N^*, T^* \leq d^{C^*}$ for some large constant $C^* > 0$. We use the standard asymptotic notation $\mathcal{O}(\cdot), \Omega(\cdot), \Theta(\cdot), o(\cdot)$ to express dependencies on d, and $\tilde{\mathcal{O}}(\cdot), \tilde{\Omega}(\cdot), \tilde{\Theta}(\cdot)$ to hide logarithmic factors of d.

3.1 Preliminaries

We first provide backgrounds on learning Gaussian single-index models, which are essential for understanding our main result. Let $\mathrm{He}_i(z) = (-1)^i e^{\frac{z^2}{2}} \frac{\mathrm{d}^i}{\mathrm{d}z^i} e^{-\frac{z^2}{2}}$ denote the probabilist's Hermite polynomials. Then, the set $\{\mathrm{He}_i(z)/\sqrt{i!}\}_{i \in \mathbb{N} \cup \{0\}}$ forms an orthonormal basis of the L^2 space with respect to the Gaussian measure and serves as a key technical tool for the analysis of Gaussian single-index models. We now introduce two key terms relevant to the sample complexity of learning.

Definition 3.1. For any function $h : \mathbb{R} \to \mathbb{R}$ which is L^2-integrable with respect to the Gaussian measure, we express its Hermite expansion as

$$h(z) = \sum_{i=0}^{\infty} \frac{H(h, i)}{i!} \mathrm{He}_i(z), \quad H(h, i) := \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(0,1)}[h(\mathbf{z}) \mathrm{He}_i(\mathbf{z})].$$

We also define the following quantities:

- We define $\deg(h)$ as the degree of h, if it is a polynomial.
- The *information exponent* (Ben Arous et al., 2021; Damian et al., 2023) of h is defined as

  $$\mathrm{ie}(h) := \min\{i \in \mathbb{N} : H(h, i) \neq 0\}.$$

  It implies that $\mathbb{E}_{\mathbf{z} \sim \mathcal{N}(0,1)}[h(\mathbf{z}) \mathrm{He}_k(\mathbf{z})] = 0$ for any $k \in \mathbb{N}$ with $k < \mathrm{ie}(h)$.
- The *generative exponent* (Damian et al., 2024) of h is defined as the lowest possible information exponent after an L^2 transformation. It is formally defined as:

  $$\mathrm{ge}(h) := \min_{\mathcal{T} \in L^2(\mathbb{P}_h)} \min\{i \in \mathbb{N} : H(\mathcal{T} \circ h, i) \neq 0\},$$

  where $L^2(\mathbb{P}_h)$ is the set of L^2-integrable functions with respect to $\mathbb{P}_h$. Here, $\mathbb{P}_h$ is the push-forward measure of the Gaussian measure by h.

While the definition of the generative exponent may seem difficult to apply at first glance, Lee et al. (2024) provides a characterization of the generative exponent for polynomials.

Lemma 3.2 (Proposition 6 in Lee et al. (2024))**.** *For a polynomial link function g_*, the generative exponent is characterized as $\mathrm{ge}(g_*) = 2$ if g_* is an even function, and $\mathrm{ge}(g_*) = 1$ otherwise.*

From the definition, $\mathrm{ge}(g_*) \le \mathrm{ie}(g_*) \le \deg(g_*)$ and Lemma 3.2 implies that the gap between these three terms can be arbitrarily large depending on the choice of g_*[2]. With a slight abuse of notation, we use $\Theta(\deg(g_*))$ to denote a quantity that is bounded by a universal constant multiple of $\deg(g_*)$. We also use $\Theta(\mathrm{ie}(g_*))$ and $\Theta(\mathrm{ge}(g_*))$, in similar manners.

Sample Complexity of Learning Single-Index Models. Previous works have established the sample complexity of various methods for learning a Gaussian single-index model. For example, kernel methods, which lack an adaptive basis, require at least $d^{\deg(g_*)}$ samples (Ghorbani et al., 2021; Donhauser et al., 2021). In contrast, adaptive methods such as gradient-based methods on two-layer neural networks can achieve a sample complexity of $\tilde{\mathcal{O}}\left(d^{\Theta(\mathrm{ie}(g_*))}\right)$ by learning an adaptive feature map (Ben Arous et al., 2021; Ba et al., 2022; Damian et al., 2022, 2023; Dandi et al., 2024). These approaches fall under the category of CSQ algorithms, and in this category, a sample complexity that depends on the information exponent is inevitable (Damian et al., 2022). However, recent works show that a nonlinear transformation introduced by data reuse (Arnaboldi et al., 2024; Lee et al., 2024) enables the algorithm to move into the broader class of Statistical Query (SQ) algorithms. This transformation allows the "effective" information exponent to be lowered to the generative exponent, thereby achieving a sample complexity of $\tilde{\mathcal{O}}\left(d^{\Theta(\mathrm{ge}(g_*))}\right)$.

3.2 Main Result

We now present our main result, which provides a theoretical characterization of the pretraining and test-time sample complexities for achieving low ICL errors.

Theorem 3.3. *Let $f(\cdot; \boldsymbol{\gamma}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, \boldsymbol{a}^*)$ be the Mamba model pretrained using Algorithm 1. We assume the following conditions hold for its hyperparameters:*

- *The context length is $N_{\mathrm{pt}} = \tilde{\Omega}\left(r^{3\mathrm{ge}(g_*)} d^8 \vee T_1^2 d^4\right)$.*
- *The number of pretraining tasks are $T_1 = \tilde{\Omega}\left(r^{3\mathrm{ge}(g_*)} d^6\right)$ and $T_2 = \tilde{\Omega}\left(r^{3\mathrm{ge}(g_*)}\right)$.*
- *The MLP width is $m = \tilde{\Omega}\left(r^{4\mathrm{ge}(g_*)}\right)$.*
- *The fixed weights are $\rho = \Theta\left((\log d)^{C_\rho}\right)$ and $b = -C_b \log d$, and the initialization scale is $\gamma = \Theta((\log d)^{-C_\gamma})$ for sufficiently large constants $C_\gamma, C_\rho, C_b > 0$.*

Then, there exist hyperparameters λ_1, λ_2, and η such that with probability at least 0.99 over the training data and random initialization, the following holds: If the test prompt length satisfies $N_{\mathrm{test}} = \tilde{\Omega}\left(r^{3\mathrm{ge}(g_*)}\right)$, then the test error $\mathcal{R}_{N_{\mathrm{test}}}(\boldsymbol{\gamma}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, \boldsymbol{a}^*)$ is bounded by $\tau + o(1)$.

We discuss our sample complexity results in comparison with other methods, including regression on test prompts and prior theoretical works (Oko et al., 2024; Nishikawa et al., 2025). We summarize these results in Table 1 and highlight the following key points:

Adaptation to Low-Dimensional Structure. Our test-time sample complexity depends on the intrinsic dimension of the feature vectors r, rather than the ambient dimension d. This is consistent with prior works (Oko et al., 2024; Nishikawa et al., 2025) that also demonstrate a dependence on intrinsic dimensionality. In contrast, the sample complexities of various regression algorithms we have discussed depend on the full input dimension d. This difference arises because pretrained models can learn the low-dimensional structure of the intrinsic feature space during pretraining.

Test-Time Feature Learning. The dependence of the sample complexity on the intrinsic dimension r in the work of Oko et al. (2024) is controlled by the degree of the link function g_*. This means that while their approach is more efficient than simple regression on full dimensions, its performance remains close to that of kernel methods on intrinsic dimensions. In contrast, our result depends on the generative exponent $\mathrm{ge}(g_*)$, instead of its degree. This implies that Mamba's efficient in-context

[2]For example, consider $g_*(z) = \mathrm{He}_q(z) + \mathrm{He}_p(z)$ with $1 \ll q \ll p$.

learning is enabled not just by its ability to learn an intrinsic feature space, but also by a process called *test-time feature learning*, which allows the model to extract features directly from the context. The same process also works for the softmax Transformers considered in Nishikawa et al. (2025) and thus achieved a similar sample complexity. However, they perform test-time feature learning through different mechanisms: Mamba relies on nonlinear gating, while Transformer uses softmax attention.

Improved Pretraining Sample Complexity. The conditions for the pretraining in our theorem can be satisfied with a pretraining sample complexity of $N_{\mathrm{pt}} = \Theta\left(d^{\Theta(\mathrm{ge}(g_*))}\right)$. In contrast, the pretraining sample complexities in previous works (Oko et al., 2024; Nishikawa et al., 2025) are governed by the information exponent, which can lead to a suboptimal rate in the worst case. This improvement is due to the nonlinearity of the MLP, as we discuss in detail in Section 4. Note that since these sample complexity bounds are sufficient conditions, it remains open whether Mamba is theoretically more efficient than Transformers in pretraining. While we leave this theoretical investigation as a future direction, our experiments in Section 5 suggest this is indeed the case.

Table 1: Summary of sample complexity results for regression algorithms on test prompt and prior works on ICL (Oko et al., 2024; Nishikawa et al., 2025).

Regression on Test Prompt		
Kernel	CSQ	SQ
$d^{\Theta(\deg(g_*))}$	$d^{\Theta(\mathrm{ie}(g_*))}$	$d^{\Theta(\mathrm{ge}(g_*))}$
In-Context Learning		
Linear Transformer	Softmax Transformer	Mamba
Oko et al. (2024)	Nishikawa et al. (2025)	This Work
Pretrain: $d^{\Theta(\mathrm{ie}(g_*))}$ *Test:* $r^{\Theta(\deg(g_*))}$	*Pretrain:* $d^{\Theta(\mathrm{ie}(g_*))}$ *Test:* $r^{\Theta(\mathrm{ge}(g_*))}$	*Pretrain:* $d^{\Theta(\mathrm{ge}(g_*))}$ *Test:* $r^{\Theta(\mathrm{ge}(g_*))}$

4 Proof Overview

In this section, we provide an overview of the proof for our theorem. The proof consists of three main parts: an analysis of one-step gradient descent on the Mamba layer, the optimization of the MLP, and a test error analysis. The formal proofs for each of these steps are provided in Appendices B, C, and D, respectively. In the following, we introduce the key ideas behind each step, providing a detailed discussion of our theoretical components.

4.1 One-Step Gradient Descent on the Mamba Layer

We show that the pretrained parameter $\boldsymbol{\gamma}^*$ recovers the intrinsic feature space by assigning significantly larger weights to embedding coordinates supported on the feature index set $\mathcal{I}$ than to those involving irrelevant coordinates. This identifies the structure shared across tasks; the context examples then provide the information about the particular task's feature vector $\boldsymbol{\beta}$. The resulting test-time feature learning is summarized in the following proposition (formally stated in Proposition B.5):

Proposition 4.1 (Informal)**.** *For a sampled ICL input embedding $\boldsymbol{Z}$ with context length $N = \tilde{\Omega}\left(r^{3\mathrm{ge}(g_*)}\right)$, query $\boldsymbol{x}$, and feature vector $\boldsymbol{\beta}$, the following holds with high probability:*

$$N^{-1}\mathsf{Mamba}(\boldsymbol{Z};\boldsymbol{\gamma}^*) \approx P_1 + P_2\left(\frac{\langle \boldsymbol{\beta}, \boldsymbol{x}\rangle}{r}\right)^{\mathrm{ge}(g_*)},$$

where P_1 and P_2 are independent of the data.

Assuming a negative bias b with sufficiently large absolute value, and a large enough number of tasks T_1 and context length N_{pt}, the updated parameter $\boldsymbol{\gamma}^*$ can be approximated as follows:

$$\begin{aligned}\boldsymbol{\gamma}^* &\approx 2\eta\mathbb{E}_{(\boldsymbol{Z},y)\sim\mathcal{D}(N_{\mathrm{pt}})}[y\nabla_{\boldsymbol{\gamma}} f(\boldsymbol{Z};\boldsymbol{\gamma}^{(0)},\boldsymbol{u}^{(0)},\boldsymbol{v}^{(0)},\boldsymbol{a}^{(0)})]\\ &\approx 2\eta\mathbb{E}_{\substack{\boldsymbol{\beta}\sim\mathrm{Unif}(S_r)\\(\boldsymbol{x},y)\sim\mathcal{D}_{\boldsymbol{\beta}}}}\left[y\,\mathbb{1}\left[\left\langle \boldsymbol{c}_{\boldsymbol{\beta}}, \boldsymbol{\gamma}^{(0)}\odot\phi(\boldsymbol{x})\right\rangle > 0\right]\boldsymbol{c}_{\boldsymbol{\beta}}\odot\phi(\boldsymbol{x})\right],\end{aligned}$$

where $\boldsymbol{c}_{\boldsymbol{\beta}} := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\boldsymbol{\beta}}}[y\sigma(y/\rho+b)\phi(\boldsymbol{x})]$ corresponds to a simplified expectation over context examples, neglecting the effect of "forgetting" in the gating mechanism. The negative bias makes cumulative forgetting negligible over the admissible context lengths, while concentration over context examples and pretraining tasks justifies the population approximation. Appendix B bounds these errors together with the contribution of the initial model output.

The Role of Gating and Input Embedding. For the proof of Proposition 4.1 and our test-time sample complexity, nonlinear transformation introduced by gating mechanism plays a crucial role. In the absence of a gating mechanism and with only a linear attention, the term $\boldsymbol{c}_{\boldsymbol{\beta}}$ is replaced by $\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\boldsymbol{\beta}}}[y\phi(\boldsymbol{x})]$ and this term vanishes when $\mathrm{ie}(g_*) > 2$. This is a consequence of our input embedding using Hermite polynomials only up to the second degree, in combination with Stein's lemma. As a result, pretraining in this case is unable to learn useful information. However, we prove that the gating mechanism ensures $c_{\boldsymbol{\beta}}$ is non-zero, thereby enabling the model to extract information. This is because the nonlinear transformation introduced by the gating mechanism reduces the information exponent to the generative exponent: $\mathrm{ie}(g_*\sigma(g_*/\rho + b)) = \mathrm{ge}(g_*)$. This reduction, combined with Lemma 3.2 and our input embedding, makes information extraction possible. A similar approach may also allow an extension to the standard input embedding $\phi(\boldsymbol{x}) = \boldsymbol{x}$ for non-even link functions. Reducing the information exponent to the generative exponent crucially affects the achievement of a test-time sample complexity below the CSQ lower bounds. Nishikawa et al. (2025) shows that the softmax operator in Transformer can also perform a nonlinear transformation on the label, which reduces the information exponent. This highlights a key difference in the mechanisms used by Mamba and Transformer to achieve this result.

Improved Pretraining Sample Complexity. Our pretraining sample complexity scales as $d^{\Theta(\mathrm{ge}(g_*))}$. This improvement relies on the nonlinearity of the MLP, whose ReLU activation introduces the indicator in the gradient. If the indicator $\mathbb{1}[\cdot]$ inside the expectation is 1 with high probability, then the updated parameter $\boldsymbol{\gamma}^*$ can be approximated as $\boldsymbol{\gamma}^* \approx 2\eta\mathbb{E}_{\boldsymbol{\beta}\sim\mathrm{Unif}(S_r)}\left[\mathbb{E}_{(\boldsymbol{x},y\sim\mathcal{D}_{\boldsymbol{\beta}})}[y\boldsymbol{c}_{\boldsymbol{\beta}} \odot \phi(\boldsymbol{x})]\right]$ and this close to zero when $\mathrm{ie}(g_*) > 2$, leading to less information gain. However, we show that the indicator deviates significantly from a constant value. We formally prove that this deviation allows the indicator to reduce the information exponent when multiplied by the label y, thereby inducing a pretraining sample complexity not governed by $\mathrm{ie}(g_*)$. While Nishikawa et al. (2025) employ an architecture with a similar structure to ours—an MLP following a softmax Transformer—they do not achieve the same improvement. This is because their use of softmax places the model in a regime where a key indicator function is 1 with high probability, which is sufficient in their case as softmax generates the necessary higher-order functions of the input, while our analysis only uses up to second-order functions. In addition, our observation for this improvement cannot be directly applied to the work of Oko et al. (2024) due to a key architectural difference: applying the MLP layer in the input embedding rather than at the output layer.

4.2 Optimization of the MLP and Test Error Analysis

In our analysis of Stage II pretraining, we show that the MLP can fit the link function g_*. We first construct an outer layer parameter $\boldsymbol{u}'$ such that the loss $L_2(\boldsymbol{\gamma}^*, \boldsymbol{u}', \boldsymbol{v}, \boldsymbol{a})$ is sufficiently small and the norm $\|\boldsymbol{u}'\|$ is well-bounded. Our construction is based on the techniques in Damian et al. (2022), which constructed a ReLU network approximating monomials. This allows our model to learn high-order polynomials with a few layers, in contrast to the multi-layer approach in Sun et al. (2025). This is possible because Lemma 3.2 implies that $g_*(z)$ is a polynomial of $z^{\mathrm{ge}(g_*)}$, allowing us to construct an MLP that approximates $g_*(\langle\boldsymbol{\beta}, \boldsymbol{x}\rangle)$ when the input is provided in the form given in Proposition 4.1.

For a suitable $\lambda_2 > 0$, we show that the optimized output weights $\boldsymbol{u}^*$ attain empirical absolute loss $\tau + o(1)$ with a well-bounded norm. Next, we show that the trained model achieves a small test error with context length N_{pt} by applying a standard generalization bound based on Rademacher complexity, which is applicable due to a well-bounded norm $\|\boldsymbol{u}^*\|$. Lastly, we extend this error bound to a general context length $N_{\mathrm{test}} = \tilde{\Omega}\left(r^{3\mathrm{ge}(g_*)}\right)$. This is possible because Proposition 4.1 implies that prompts with N_{test} context examples and N_{pt} context examples give similar normalized Mamba outputs, given the same feature vector $\boldsymbol{\beta}$ and query.

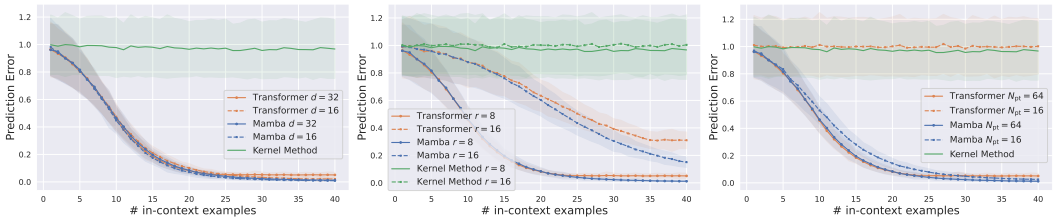
5 Experiments

To support our theoretical findings, we conduct experiments beyond our simplified setting using standard architectures and optimization algorithms on NVIDIA RTX A6000 GPUs.

We pretrain and evaluate both Transformer and Mamba on our data distribution. Our base configuration uses a link function $g_*(z) = \mathrm{He}_3(z)/\sqrt{6}$, an intrinsic dimension of $r = 8$, and an ambient

dimension $d = 32$. We employ a 6-layer GPT-2 model (Radford et al., 2019) with 8 attention heads and a 12-layer Mamba model. To ensure a fair comparison, both models have an embedding dimension of 256 and a similar number of parameters. The overall experimental settings for pretraining follow those of prior works (Garg et al., 2022; Park et al., 2024),[3] including the optimization algorithm and a pretraining context length $N_{\mathrm{pt}} = 64$. We also conduct kernel ridge regression on the intrinsic feature space to serve as a baseline for understanding the effect of feature learning. For this, we use a Gaussian RBF kernel with a bandwidth of 1 and a ridge parameter of 1. For evaluation, we measure the prediction error using squared error, with the number of context examples ranging from 1 to 40. We estimate the test error on 1024 randomly sampled tasks, using 2048 independent prompts for each task, and represent the results with the mean and standard deviation over these tasks.

To validate our results on the influence of problem parameters, we analyze trends by varying parameters from our base configuration: the ambient dimension to $d = 16$, the intrinsic dimension to $r = 16$, and the pretraining context length to $N_{\mathrm{pt}} = 16$. Figure 1a demonstrates the influence of the ambient dimension d. Both Transformer and Mamba models exhibit comparable performance that is rarely affected by the ambient dimension d. In contrast, when the intrinsic dimension r is increased, both models exhibit performance degradation, while their performance remains comparable (Figure 1b). This suggests that both models mainly utilize information from the intrinsic feature space. In addition, these methods outperform kernel methods, even when we restrict the input of the kernel method to the intrinsic feature space. This observation aligns with our finding that Mamba, similar to Transformers, not only benefits from its adaptation to the intrinsic feature space but also performs test-time feature learning. Lastly, when using a small pretraining context length $N_{\mathrm{pt}} = 16$, the Transformer's performance deteriorates significantly, while that of Mamba remains comparable to its performance with $N_{\mathrm{pt}} = 64$ (Figure 1c). This observation aligns with our pretraining sample complexity result, which is lower than that of the Transformer established by Nishikawa et al. (2025).



(a) $d = 32$ vs. $d = 16$ (b) $r = 8$ vs. $r = 16$ (c) $N_{\mathrm{pt}} = 64$ vs. $N_{\mathrm{pt}} = 16$

Figure 1: Comparison of prediction error for in-context learning with Transformer and Mamba models, and kernel regression across different problem parameter settings.

6 Conclusion

We investigated Mamba's in-context learning capability by focusing on a Gaussian single-index model. We proved that Mamba, when pretrained with gradient-based optimization, can efficiently learn in-context through a mechanism we termed test-time feature learning. Our derived test-time sample complexity is comparable to that of the softmax Transformer, a result established by Nishikawa et al. (2025) and also surpasses the CSQ lower bound. Our analysis reveals that Mamba's gating mechanism is a key factor in enabling feature learning and strong performance. We also presented experimental results to support our findings.

We discuss the limitations of our work and future directions. First, our analysis uses a simplified data distribution, model architecture, and training procedure; extending our results beyond this setting is an important direction. Second, our experiments suggest that Mamba is more sample-efficient than Transformers in pretraining, but a theoretical explanation for this difference remains open. Third, our analysis assumes a negligible effect of forgetting, which can benefit tasks with outliers (Li et al., 2025a). Understanding the interplay between forgetting and our feature-learning mechanism is an interesting direction.

[3]We build upon the official codebases of these works (MIT and Apache 2.0 licenses).

Acknowledgement

JO acknowledges partial support by the Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)). WH was supported by JSPS KAKENHI (24K20848) and JST BOOST (JPMJBY24G6). TS was partially supported by JSPS KAKENHI (24K02905) and JST CREST (JPMJCR2115). This work was partially supported by JST ERATO (JPMJER2601). This research is supported by the National Research Foundation, Singapore, Infocomm Media Development Authority under its Trust Tech Funding Initiative, and the Ministry of Digital Development and Information under the AI Visiting Professorship Programme (award number AIVP-2024-004). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, Infocomm Media Development Authority, and the Ministry of Digital Development and Information.

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*, 36:45614–45650, 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.

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Luca Pesce, and Ludovic Stephan. Repetita iuvant: Data repetition allows SGD to learn high-dimensional multi-index functions. *arXiv preprint arXiv:2405.15459*, 2024.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

Francis Bach. *Learning theory from first principles*. MIT press, 2024.

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*, 36:57125–57211, 2023.

Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. *Advances in Neural Information Processing Systems*, 35:21750–21764, 2022.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *Advances in Neural Information Processing Systems*, 37:107547–107603, 2024.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22 (106):1–51, 2021.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in neural information processing systems*, 35:9768–9783, 2022.

Marco Bondaschi, Nived Rajaraman, Xiuying Wei, Kannan Ramchandran, Razvan Pascanu, Caglar Gulcehre, Michael Gastpar, and Ashok Vardhan Makkuva. From markov to laplace: How mamba in-context learns markov chains. *arXiv preprint arXiv:2502.10178*, 2025.

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.

Dake Bu, Wei Huang, Andi Han, Atsushi Nitanda, Qingfu Zhang, Hau-San Wong, and Taiji Suzuki. Provable in-context vector arithmetic via retrieving task concepts. In *Forty-second International Conference on Machine Learning*, 2025.

Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.

Xiang Cheng, Yuxin Chen, and Suvrit Sra. Transformers implement functional gradient descent to learn non-linear functions in context. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ah1BlQcLv4.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems*, 36:752–784, 2023.

Alex Damian, Loucas Pillaud-Vivien, Jason Lee, and Joan Bruna. Computational-statistical gaps in gaussian single-index models. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1262–1262. PMLR, 2024.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349): 1–65, 2024. URL http://jmlr.org/papers/v25/23-1543.html.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ztn8FCR1td.

Konstantin Donhauser, Mingqi Wu, and Fanny Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR, 2021.

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.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2), 2021.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Margalit Glasgow. SGD finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the XOR problem. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=HgOJlxzB16.

Riccardo Grazzi, Julien Niklas Siems, Simon Schrodi, Thomas Brox, and Frank Hutter. Is mamba capable of in-context learning? In *International Conference on Automated Machine Learning*, pages 1–1. PMLR, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. Understanding emergent in-context learning from a kernel regression perspective. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=6rD50Q6yYz.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=9GLvXGkUE2.

Jiarui Jiang, Wei Huang, Miao Zhang, Taiji Suzuki, and Liqiang Nie. Trained mamba emulates online gradient descent in in-context linear regression. *arXiv preprint arXiv:2509.23779*, 2025.

Nirmit Joshi, Theodor Misiakiewicz, and Nati Srebro. On the complexity of learning sparse functions with statistical and gradient queries. *Advances in Neural Information Processing Systems*, 37: 103198–103241, 2024.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. In *International Conference on Machine Learning*, pages 24527–24561. PMLR, 2024.

Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit. *Advances in Neural Information Processing Systems*, 37:58716–58756, 2024.

Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning? In *Forty-first International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=I4HTPws9P6.

Hongkang Li, Songtao Lu, Xiaodong Cui, Pin-Yu Chen, and Meng Wang. Can mamba learn in context with outliers? a theoretical generalization analysis. *arXiv preprint arXiv:2510.00399*, 2025a.

Yingcong Li, Ankit S Rawat, and Samet Oymak. Fine-grained analysis of in-context linear estimation: Data, architecture, and beyond. *Advances in Neural Information Processing Systems*, 37:138324–138364, 2024b.

Yingcong Li, Xupeng Wei, Haonan Zhao, and Taigao Ma. Can mamba in-context learn task mixtures? In *ICML 2024 Workshop on In-Context Learning*, 2024c.

Yingcong Li, Davoud Ataee Tarzanagh, Ankit Singh Rawat, Maryam Fazel, and Samet Oymak. Gating is weighting: Understanding gated linear attention through in-context learning. *arXiv preprint arXiv:2504.04308*, 2025b.

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.

Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. Gradient-based feature learning under structured data. *Advances in Neural Information Processing Systems*, 36:71449–71485, 2023.

Naoki Nishikawa, Yujin Song, Kazusato Oko, Denny Wu, and Taiji Suzuki. Nonlinear transformers can perform inference-time feature learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=xQTSvP57C3.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context. *Advances in Neural Information Processing Systems*, 37:77316–77365, 2024.

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. In *International Conference on Machine Learning*, pages 39793–39812. PMLR, 2024.

Badri Narayana Patro and Vijay Srinivas Agneeswaran. Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges. *Engineering Applications of Artificial Intelligence*, 159:111279, 2025.

Bo Peng, Daniel Goldstein, Quentin Gregory Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Teddy Ferdinan, Kranthi Kiran GV, Haowen Hou, Satyapriya Krishna, Ronald McClelland Jr., Niklas Muennighoff, Fares Obeid, Atsushi Saito, Guangyu Song, Haoqin Tu, Ruichong Zhang, Bingchen Zhao, Qihang Zhao, Jian Zhu, and Rui-Jie Zhu. Eagle and finch: RWKV with matrix-valued states and dynamic recurrence. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=soz1SEiPeq.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. Reasoning with latent thoughts: On the power of looped transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=din0lGfZFd.

Haoyuan Sun, Ali Jadbabaie, and Navid Azizan. On the role of transformer feed-forward layers in nonlinear in-context learning. *arXiv preprint arXiv:2501.18187*, 2025.

Taiji Suzuki. Adaptivity of deep reLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1ebTsActm.

Taiji Suzuki, Denny Wu, Kazusato Oko, and Atsushi Nitanda. Feature learning via mean-field langevin dynamics: classifying sparse parities and beyond. *Advances in Neural Information Processing Systems*, 36:34536–34556, 2023.

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

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.

Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, Tri Dao, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, et al. An empirical study of mamba-based language models. *arXiv preprint arXiv:2406.07887*, 2024.

Junxiong Wang, Daniele Paliotta, Avner May, Alexander Rush, and Tri Dao. The mamba in the llama: Distilling and accelerating hybrid models. *Advances in Neural Information Processing Systems*, 37:62432–62457, 2024.

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.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *International Conference on Machine Learning*, pages 56501–56523. PMLR, 2024.

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.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Joshua M. Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? a study in length generalization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AssIuHnmHX.

Contents

1 Introduction **1**

1.1 Summary of Contributions 2

1.2 Related Works 2

2 Problem Setting **3**

2.1 Data Distribution for In-Context Learning 3

2.2 Prediction Model Architecture 3

2.3 Pretraining Algorithm 5

3 Mamba Efficiently Learns Single-Index Models In-Context **6**

3.1 Preliminaries 6

3.2 Main Result 7

4 Proof Overview **8**

4.1 One-Step Gradient Descent on the Mamba Layer 8

4.2 Optimization of the MLP and Test Error Analysis 9

5 Experiments **9**

6 Conclusion **10**

A Proof Preliminaries **16**

A.1 Simplification of Mamba Output 16

A.2 High Probability Events 17

A.3 Reducing the Information Exponent with Label Transformation 17

B One-Step Gradient Descent on the Mamba Layer **19**

B.1 Estimation of Label-Gradient Correlation 20

B.2 Characterization of Updated Parameter 23

B.3 Output of Updated Mamba Layer 25

C Optimizing MLP Layer **27**

C.1 Construction of Approximating MLP Layer 27

C.2 Characterization of Estimation Error on the Training Set 28

D Test Error Analysis **30**

D.1 Test Error for Prompts with Pretraining Context Length 30

D.2 Test Error for Prompts with General Length 31

A Proof Preliminaries

Notation. We introduce the following additional notation for ease of presentation. We use $\mathbb{1}[\cdot]$ to represent indicator function. We write $a \lesssim b$, $a \gtrsim b$, and $a \asymp b$ to denote that $a = \mathcal{O}(b)$, $a = \Omega(b)$, and $a = \Theta(b)$, respectively. We also use $\mathrm{poly}(d)$ and $\mathrm{polylog}(d)$ to represent a sufficiently large polynomial in d and $\log d$, respectively. Lastly, we use $o(1/\mathrm{polylog}(d))$ to represent a quantity that decreases faster than $(\log d)^{-C}$ for any constant $C > 0$. Lastly, with a slight abuse of notation, we use the asymptotic notation we introduced to represent a vector when its norm satisfies the corresponding bound.

A.1 Simplification of Mamba Output

The following lemma immediately implies (1).

Lemma A.1. *Given a prompt $(\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_N, y_N, \boldsymbol{x})$ and its input embedding $\boldsymbol{Z} \in \mathbb{R}^{(\tilde{d}+1)\times(N+1)}$, let $\mathsf{Mamba}(\boldsymbol{Z}; \boldsymbol{\Theta}) = (\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{N+1}) \in \mathbb{R}^{(\tilde{d}+1)\times(N+1)}$ be outputs and $\boldsymbol{h}_l^{(i)}$'s be hidden states. For each $i \in [\tilde{d}+1]$ and $l \in [N+1]$, we have*

$$\boldsymbol{h}_l^{(i)} = \sum_{j=1}^{l} G_{j,l}(\boldsymbol{Z}) \boldsymbol{z}_j[i] \boldsymbol{W}_B \boldsymbol{z}_j, \quad \boldsymbol{o}_l[i] = \sum_{j=1}^{l} G_{j,l}(\boldsymbol{Z}) \boldsymbol{z}_j[i] \boldsymbol{z}_j^\top \boldsymbol{W}_B^\top \boldsymbol{W}_C \boldsymbol{z}_l,$$

where $G_{j,l}(\boldsymbol{Z}) = \sigma\left(\boldsymbol{w}^\top \boldsymbol{z}_j + b\right) \prod_{k=j+1}^{l} \left(1 - \sigma\left(\boldsymbol{w}^\top \boldsymbol{z}_k + b\right)\right)$.

Proof of Lemma A.1. For each $l \in [N+1]$, we have

$$\overline{\boldsymbol{A}}_l = \exp\left(-\mathrm{softplus}(\boldsymbol{w}^\top \boldsymbol{z}_l + b) \boldsymbol{I}_{\tilde{d}+1}\right) = \frac{1}{1 + \exp(\boldsymbol{w}^\top \boldsymbol{z}_l + b)} \boldsymbol{I}_{\tilde{d}+1} = \left(1 - \sigma\left(\boldsymbol{w}^\top \boldsymbol{z}_l + b\right)\right) \boldsymbol{I}_{\tilde{d}+1}$$

and

$$\overline{\boldsymbol{B}}_l = -\left(\overline{\boldsymbol{A}}_l - \boldsymbol{I}_{\tilde{d}+1}\right) \boldsymbol{W}_B \boldsymbol{z}_l = \sigma\left(\boldsymbol{w}^\top \boldsymbol{z}_l + b\right) \boldsymbol{W}_B \boldsymbol{z}_l.$$

We fix any $i \in [\tilde{d}+1]$ and we will prove by applying induction on $l \in [N+1]$. Let us first consider the case $l = 1$. We have

$$\begin{aligned}
\boldsymbol{h}_1^{(i)} &= \left(1 - \sigma\left(\boldsymbol{w}^\top \boldsymbol{z}_1 + b\right)\right) \boldsymbol{h}_0^{(i)} + \sigma\left(\boldsymbol{w}^\top \boldsymbol{z}_1 + b\right) \boldsymbol{W}_B \boldsymbol{z}_1 \boldsymbol{z}_1[i] \\
&= \sigma\left(\boldsymbol{w}^\top \boldsymbol{z}_1 + b\right) \boldsymbol{W}_B \boldsymbol{z}_1 \boldsymbol{z}_1[i] \\
&= G_{1,1}(\boldsymbol{Z}) \boldsymbol{z}_1[i] \boldsymbol{W}_B \boldsymbol{z}_1
\end{aligned}$$

and

$$\boldsymbol{o}_1[i] = \left(\boldsymbol{W}_C \boldsymbol{z}_1\right)^\top \boldsymbol{h}_1^{(i)} = \sigma\left(\boldsymbol{w}^\top \boldsymbol{z}_1 + b\right) \boldsymbol{z}_1[i] \boldsymbol{z}_1^\top \boldsymbol{W}_B^\top \boldsymbol{W}_C \boldsymbol{z}_1 = G_{1,1}(\boldsymbol{Z}) \boldsymbol{z}_1[i] \boldsymbol{z}_1^\top \boldsymbol{W}_B^\top \boldsymbol{W}_C \boldsymbol{z}_1.$$

Therefore, desired conclusions hold for the case $l = 1$.

Next, we assume that our conclusion holds for $l < N + 1$.

$$\begin{aligned}
\boldsymbol{h}_{l+1}^{(i)} &= \left(1 - \sigma\left(\boldsymbol{w}^\top \boldsymbol{z}_{l+1} + b\right)\right) \boldsymbol{h}_l^{(i)} + \sigma\left(\boldsymbol{w}^\top \boldsymbol{z}_{l+1} + b\right) \boldsymbol{W}_B \boldsymbol{z}_{l+1} (\boldsymbol{z}_{l+1})_i \\
&= \left(1 - \sigma\left(\boldsymbol{w}^\top \boldsymbol{z}_{l+1} + b\right)\right) \sum_{j=1}^{l} G_{j,l}(\boldsymbol{Z}) \boldsymbol{z}_j[i] \boldsymbol{W}_B \boldsymbol{z}_j + G_{l+1,l+1}(\boldsymbol{Z}) \boldsymbol{z}_{l+1}[i] \boldsymbol{W}_B \boldsymbol{z}_{l+1} \\
&= \sum_{j=1}^{l+1} G_{j,l+1}(\boldsymbol{Z}) \boldsymbol{z}_j[i] \boldsymbol{W}_B \boldsymbol{z}_j
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{o}_{l+1}[i] &= \left(\boldsymbol{W}_C \boldsymbol{z}_{l+1}\right)^\top \boldsymbol{h}_{l+1}^{(i)} \\
&= \left(\boldsymbol{W}_C \boldsymbol{z}_{l+1}\right)^\top \sum_{j=1}^{l+1} G_{j,l+1}(\boldsymbol{Z}) \boldsymbol{z}_j[i] \boldsymbol{W}_B \boldsymbol{z}_j \\
&= \sum_{j=1}^{l+1} G_{j,l+1}(\boldsymbol{Z}) \boldsymbol{z}_j[i] \boldsymbol{z}_j^\top \boldsymbol{W}_B^\top \boldsymbol{W}_C \boldsymbol{z}_{l+1}.
\end{aligned}$$

Therefore, we have the desired conclusions. □

A.2 High Probability Events

Throughout the proof, we use the term "with high probability" which is defined as follows.

Definition A.2. We call that an event E occurs *with high probability*, when

$$\mathbb{P}[E] \geq 1 - d^{-C_{\mathrm{whp}}}$$

with a large enough $C_{\mathrm{whp}} > 0$.

For example, $\mathbf{z} = \mathcal{O}(\sqrt{\log d})$ for $\mathbf{z} \sim \mathcal{N}(0, 1)$, with high probability, which is a direct consequence of Gaussian tail bound. In addition, the intersection of a $\mathrm{poly}(d)$ events also occurs with high probability. We use these property frequently throughout our proof.

The following lemma is useful when we bound some quantities with high probability.

Lemma A.3 (Corollary 17 in Oko et al. (2024), adapted)**.** *Let P be a polynomial with degree $\deg(P)$. If $\|\boldsymbol{\beta}\| = 1$ and $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, then $|P(\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle)| \lesssim (\log d)^{\deg(P)/2}$ holds with high probability.*

This lemma implies that $y_i^t = \tilde{\mathcal{O}}(1), \|\phi(\boldsymbol{x}_i^t)\|, \|\phi(\boldsymbol{x}^t)\| = \tilde{\mathcal{O}}(d)$ holds for any $i \in [N_{\mathrm{pt}}], t \in [T_{\mathrm{pt}}]$, with high probability. We utilize these properties frequently in our proof.

The following lemma provides a high-probability guarantee regarding our input embedding, which is crucial for our analysis.

Lemma A.4. *Let $(\mathbf{x}_i, \mathbf{x}_i', \mathbf{z}_i)_{i=1}^N$ be i.i.d. tuples whose first two coordinates have standard Gaussian marginals; $\mathbf{z}_i$ may depend on these coordinates. Suppose $\mathbb{E}[\mathbf{z}_i^4] = O(1)$ and $|\mathbf{z}_i| \leq C$ with high probability, where C is bounded above and below by fixed powers of d. If $1 \leq N \leq N^*$, then for $k = 0, 1, 2$,*

$$\left| \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \mathrm{He}_k(\mathbf{x}_i) - \mathbb{E}[\mathbf{z}_1 \mathrm{He}_k(\mathbf{x}_1)] \right| = \tilde{\mathcal{O}}(CN^{-1/2})$$

with high probability. The same bound holds with $\mathrm{He}_k(\mathbf{x}_i)$ replaced by $\mathbf{x}_i \mathbf{x}_i'$.

Proof of Lemma A.4. We define $\widehat{z}_i = \mathbf{z}_i \mathbb{1}[|\mathbf{z}_i| \leq C]$. The products $\widehat{z}_i \mathrm{He}_k(\mathbf{x}_i)$ and $\widehat{z}_i \mathbf{x}_i \mathbf{x}_i'$ have subexponential norms $O(C)$, since the multiplier is bounded and Gaussian coordinates and their quadratic products have bounded subexponential norms. By applying Bernstein's inequality and using $N \leq d^{C^*}$, we have

$$\left| \frac{1}{N} \sum_{i=1}^{N} \widehat{z}_i \mathrm{He}_k(\mathbf{x}_i) - \mathbb{E}[\widehat{z}_1 \mathrm{He}_k(\mathbf{x}_1)] \right| = O\left(C\sqrt{\frac{\log d}{N}} + \frac{C \log d}{N} \right) = \tilde{\mathcal{O}}(CN^{-1/2}).$$

By a union bound, we can replace all $\widehat{z}_i$ in the sample average by $\mathbf{z}_i$ with high probability. For the expectation, by applying the Cauchy–Schwarz inequality twice, we have

$$\begin{aligned} |\mathbb{E}[(\mathbf{z}_1 - \widehat{z}_1)\mathrm{He}_k(\mathbf{x}_1)]| &\leq \left(\mathbb{E}[\mathbf{z}_1^2 \mathbb{1}[|\mathbf{z}_1| > C]]\right)^{1/2} \left(\mathbb{E}[\mathrm{He}_k(\mathbf{x}_1)^2]\right)^{1/2} \\ &\leq (\mathbb{E}[\mathbf{z}_1^4])^{1/4} \mathbb{P}[|\mathbf{z}_1| > C]^{1/4} \left(\mathbb{E}[\mathrm{He}_k(\mathbf{x}_1)^2]\right)^{1/2}. \end{aligned}$$

By choosing the high-probability exponent sufficiently large, we make this term negligible relative to $CN^{-1/2}$, since C^{-1} and N are polynomially bounded. We apply the same argument to $\mathbf{x}_1 \mathbf{x}_1'$. Its second moment is bounded by the Cauchy–Schwarz inequality, which does not require independence of the two coordinates within a tuple. □

A.3 Reducing the Information Exponent with Label Transformation

For any polynomial h and $p \in \mathbb{N} \cup \{0\}$, we define $e_p(h) := \min\{i \in \mathbb{N} : H(h^i, p) \neq 0\}$. If this minimum does not exist (i.e., the set is empty), we set $e_p(h) = \infty$. In addition, we define

$$\bar{g}_*(z) := \begin{cases} \frac{g_*(z)}{\rho} & \text{if } \left|\frac{g_*(z)}{\rho}\right| \leq \frac{1}{\log d} \\ 0 & \text{otherwise} \end{cases}.$$

By our choice of ρ, we have $\bar{g}_*(\langle \boldsymbol{\beta}^t, \boldsymbol{x}_i^t \rangle) = g_*(\langle \boldsymbol{\beta}^t, \boldsymbol{x}_i^t \rangle)/\rho$ for all pretraining context examples with high probability. We use this identity in the following analysis and treat the independent noise ζ_i^t separately.

We also define the following function, which naturally appears in our analysis:

$$A(z) := \frac{1}{2}\big[(\rho\bar{g}_*(z)+\tau)\sigma\left(\bar{g}_*(z)+\tau/\rho+b\right) + (\rho\bar{g}_*(z)-\tau)\sigma\left(\bar{g}_*(z)-\tau/\rho+b\right)\big].$$

Let $A(z) = \sum_{k=0}^{\infty} \frac{a_k}{k!}\mathrm{He}_k(z)$ be the Hermite expansion. The following lemma characterizes its Hermite coefficients.

Lemma A.5. *Let $k = \deg(g_*)$ and choose $C_\rho > 1 + k/2$. For any fixed $p \geq 1$ such that $e = e_p(g_*) < \infty$,*

$$a_p = \frac{\sigma^{(e-1)}(b)}{(e-1)!\rho^{e-1}} H(g_*^e, p) + O(d^{-C_b}\rho^{-e}). \tag{2}$$

Consequently, $d^{C_b}|a_p| = \Theta(\rho^{-(e-1)})$. The constant coefficient satisfies

$$a_0 = \frac{\sigma'(b)}{\rho}(1+\tau^2) + O(d^{-C_b}\rho^{-2}), \qquad d^{C_b}a_0 = \Theta(\rho^{-1}).$$

The constants may depend on the fixed link, p, and τ. If g_* is even, then A is even and $a_1 = 0$ exactly. The indices needed below satisfy $e_0(g_*) = 2$, $e_2(g_*) < \infty$, and $e_1(g_*) < \infty$ whenever g_* is not even.

The following two lemmas are crucial for our proof of Lemma A.5.

Lemma A.6 (Proposition 6 in Lee et al. (2024))**.** *For any nonconstant polynomial P, there exist $C_P, D_P > 0$ depending only on P such that the following holds.*

- *If P is not an even function, then there exists $i \leq C_P$ such that $\left|H(P^i, 1)\right| \geq D_P$.*
- *If P is an even function, then there exists $i \leq C_P$ such that $\left|H(P^i, 2)\right| \geq D_P$.*

Lemma A.7. *For any $k \in \mathbb{N} \cup \{0\}$ and $z < -k-2$, we have $\frac{e^z}{2} \leq \sigma^{(k)}(z) \leq 2e^z$.*

Proof of Lemma A.7. For any $x < 0$, we have

$$\sigma(x) = \frac{1}{1+\exp(-x)} = 1 - \frac{1}{1+\exp(x)} = \sum_{j=1}^{\infty}(-1)^{j-1}e^{jx}.$$

Therefore, we have

$$\sigma^{(k)}(z) = \sum_{j=1}^{\infty}(-1)^{j-1}j^k e^{jz} = e^z + \sum_{j=2}^{\infty}(-1)^{j-1}j^k e^{jz}.$$

For each $j \geq 2$, since $\frac{(j+1)^k e^{(j+1)z}}{j^k e^{jz}} \leq 2^k e^z$ and $2^k e^z < \frac{1}{3}$, we have

$$\left|\sigma^{(k)}(z) - e^z\right| \leq \sum_{j=2}^{\infty} j^k e^{jz} \leq 2^k e^{2z}\sum_{j=0}^{\infty}\left(2^k e^z\right)^j = \frac{2^k e^{2z}}{1-2^k e^z} \leq \frac{e^z}{2}.$$

Hence, we have a desired conclusion. □

We now prove Lemma A.5.

Proof of Lemma A.5. Let $Z \sim \mathcal{N}(0,1)$ and let ζ be independent and uniform on $\{\pm\tau\}$. Let

$$E_\rho := \{|g_*(Z)| > \rho/\log d\}.$$

Then $A(z) = \mathbb{E}_\zeta[(\rho\bar{g}_*(z)+\zeta)\sigma(b+(\rho\bar{g}_*(z)+\zeta)/\rho)]$. Since g_* has fixed degree k and $C_\rho > 1+k/2$, the Gaussian tail bound gives, for every fixed j, p, M,

$$\mathbb{E}[(1+|g_*(Z)|^j)|\mathrm{He}_p(Z)|\mathbb{1}[E_\rho]] = O(d^{-M}). \tag{3}$$

To show this, note that E_ρ implies $|Z| \geq c(\log d)^{(C_\rho - 1)/k}$ for sufficiently large d. By applying the Cauchy–Schwarz inequality, we bound this expectation by a fixed Gaussian polynomial moment times $\mathbb{P}[E_\rho]^{1/2}$, which decays faster than every power of d^{-1}.

For any fixed integer $J \geq 0$, by applying Taylor's theorem to the sigmoid at b, we have

$$A(z) = \sum_{j=0}^{J} \frac{\sigma^{(j)}(b)}{j!\rho^j} \mathbb{E}_\zeta[(\rho \bar{g}_*(z) + \zeta)^{j+1}] + R_J(z).$$

We have $|(\rho \bar{g}_*(z) + \zeta)/\rho| \leq 2/\log d$ for large d. By Lemma A.7, we obtain

$$|R_J(z)| \leq C_J d^{-C_b} \rho^{-(J+1)} \mathbb{E}_\zeta[|\rho \bar{g}_*(z) + \zeta|^{J+2}], \qquad |H(R_J, p)| = O(d^{-C_b} \rho^{-(J+1)}).$$

The moments in this bound are finite and uniformly bounded in d. By (3), we can replace each $\rho \bar{g}_*$ in the finite sum by g_* with an error $O(d^{-C_b - M})$ for arbitrary fixed M. Therefore, the truncation error also contains the factor d^{-C_b}.

For $p \geq 1$ and $e = e_p(g_*) < \infty$, we choose $J = e - 1$. By expanding $(g_* + \zeta)^{j+1}$ and using independence, we find that its pth Hermite coefficient vanishes for $j + 1 < e$, since all positive powers of g_* below e have zero coefficient and $H(1, p) = 0$. At $j = e - 1$, the only nonzero term is $H(g_*^e, p)$. Thus, we obtain (2). By the series expansion in the proof of Lemma A.7, we also have $\sigma^{(j)}(b) = d^{-C_b}(1 + O(d^{-C_b}))$ for fixed j, which gives the stated magnitude bound without a sign assumption on a_p.

For $p = 0$, take $J = 1$ and use $\mathbb{E}[g_*(Z) + \zeta] = 0$ and $\mathbb{E}[(g_*(Z) + \zeta)^2] = 1 + \tau^2$. This gives the claimed expression for a_0, including the noise contribution.

Finally, $e_0 = 2$ follows from the normalization of g_*. By Lemma A.6, we have finite e_1 for non-even links. For even links, $a_1 = 0$ follows directly from evenness and does not use a truncation approximation. To show that $e_2 < \infty$ for every nonconstant polynomial, we let $M_0 = \max_{|z| \leq 1} |g_*(z)|$. We choose a fixed interval I outside $[-1, 1]$ on which $|g_*(z)| > 2M_0$ and $\mathrm{He}_2(z) > 0$. For sufficiently large integers j, the positive contribution of I to $H(g_*^{2j}, 2)$ exceeds the negative contribution from $[-1, 1]$, and all other contributions outside $[-1, 1]$ are nonnegative. Therefore, $H(g_*^{2j}, 2) > 0$ for some finite j.

□

B One-Step Gradient Descent on the Mamba Layer

Let us define the function $\psi : \mathbb{R}^{d+3} \to \mathbb{R}^{\tilde{d}}$, which we use repeatedly in our proof. It is defined as

$$\psi(\boldsymbol{\theta}, c_0, c_1, c_2) := \left[c_0, c_1 \boldsymbol{\theta}^\top, \frac{c_2 (\boldsymbol{\theta} \odot \boldsymbol{\theta})^\top}{\sqrt{2}}, c_2 \boldsymbol{\theta}[1]\boldsymbol{\theta}[2], \ldots, c_2 \boldsymbol{\theta}[d-1]\boldsymbol{\theta}[d] \right]^\top.$$

Note that for any vector $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\|\psi(\boldsymbol{\theta}, c_0, c_1, c_2)\|^2 = c_0^2 + c_1^2 \|\boldsymbol{\theta}\|^2 + \frac{c_2^2 \|\boldsymbol{\theta}\|^4}{2}. \tag{4}$$

If we choose $\lambda_1 = \eta^{-1}$, the updated parameter $\boldsymbol{\gamma}^*$ can be expressed as

$$\begin{aligned}
\boldsymbol{\gamma}^* &= \frac{2\eta}{T_1} \sum_{t \in [T_1]} \left[\left(y^t - f\left(\boldsymbol{Z}^t; \boldsymbol{\gamma}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{a}^{(0)} \right) \right) \nabla_{\boldsymbol{\gamma}} f\left(\boldsymbol{Z}^t; \boldsymbol{\gamma}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{a}^{(0)} \right) \right] \\
&= \frac{2\eta}{T_1} \sum_{t \in [T_1]} y^t \nabla_{\boldsymbol{\gamma}} f\left(\boldsymbol{Z}^t; \boldsymbol{\gamma}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{a}^{(0)} \right) \\
&\quad - \frac{2\eta}{T_1} \sum_{t \in [T_1]} f\left(\boldsymbol{Z}^t; \boldsymbol{\gamma}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{a}^{(0)} \right) \nabla_{\boldsymbol{\gamma}} f\left(\boldsymbol{Z}^t; \boldsymbol{\gamma}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{a}^{(0)} \right).
\end{aligned}$$

The initial output evaluated at $\boldsymbol{Z}^t$ can be bounded as

$$\begin{aligned}\left|f\left(\boldsymbol{Z}^t;\boldsymbol{\gamma}^{(0)},\boldsymbol{u}^{(0)},\boldsymbol{v}^{(0)},\boldsymbol{a}^{(0)}\right)\right| &= \mathrm{ReLU}\left(N_{\mathrm{pt}}^{-1}\sum_{j=1}^{N_{\mathrm{pt}}} G_{j,N_{\mathrm{pt}}+1}(\boldsymbol{Z}^t) y_j^t \phi\left(\boldsymbol{x}_j^t\right)^\top \left(\boldsymbol{\gamma}^{(0)}\odot\phi\left(\boldsymbol{x}^t\right)\right)\right)\\ &\le N_{\mathrm{pt}}^{-1}\sum_{j=1}^{N_{\mathrm{pt}}} G_{j,N_{\mathrm{pt}}+1}(\boldsymbol{Z}^t)\left|y_j^t\right|\left\|\phi\left(\boldsymbol{x}_j^t\right)\right\|\left\|\phi\left(\boldsymbol{x}^t\right)\right\|\\ &= \tilde{\mathcal{O}}\left(d^{-C_b+2}\right),\end{aligned}$$

with high probability.

The gradient of $\boldsymbol{\gamma}$ of output evaluated at $\boldsymbol{Z}^t$ can be calculated as

$$\begin{aligned}&\nabla_{\boldsymbol{\gamma}} f\left(\boldsymbol{Z}^t;\boldsymbol{\gamma}^{(0)},\boldsymbol{u}^{(0)},\boldsymbol{v}^{(0)},\boldsymbol{a}^{(0)}\right)\\ =\ &\mathbb{1}\left[\sum_{j=1}^{N_{\mathrm{pt}}} G_{j,N_{\mathrm{pt}}+1}\left(\boldsymbol{Z}^t\right) y_j^t\phi\left(\boldsymbol{x}_j^t\right)^\top\left(\boldsymbol{\gamma}^{(0)}\odot\phi\left(\boldsymbol{x}^t\right)\right)>0\right]\\ &\times\left(N_{\mathrm{pt}}^{-1}\sum_{j=1}^{N_{\mathrm{pt}}} G_{j,N_{\mathrm{pt}}+1}\left(\boldsymbol{Z}^t\right) y_j^t\phi\left(\boldsymbol{x}_j^t\right)\odot\phi\left(\boldsymbol{x}^t\right)\right)\end{aligned}$$

and its norm can be bounded as

$$\begin{aligned}\left\|\nabla_{\boldsymbol{\gamma}} f\left(\boldsymbol{Z}^t;\boldsymbol{\gamma}^{(0)},\boldsymbol{u}^{(0)},\boldsymbol{v}^{(0)},\boldsymbol{a}^{(0)}\right)\right\| &\le N_{\mathrm{pt}}^{-1}\sum_{j=1}^{N_{\mathrm{pt}}} G_{j,N_{\mathrm{pt}}+1}(\boldsymbol{Z}^t)\left|y_j^t\right|\left\|\phi\left(\boldsymbol{x}_j^t\right)\right\|\left\|\phi\left(\boldsymbol{x}^t\right)\right\|\\ &= \tilde{\mathcal{O}}\left(d^{-C_b+2}\right),\end{aligned}$$

with high probability. Therefore, with high probability, we have

$$\boldsymbol{\gamma}^* = \frac{2\eta}{T_1}\sum_{t\in[T_1]} y^t\nabla_{\boldsymbol{\gamma}} f\left(\boldsymbol{Z}^t;\boldsymbol{\gamma}^{(0)},\boldsymbol{u}^{(0)},\boldsymbol{v}^{(0)},\boldsymbol{a}^{(0)}\right)+\tilde{\mathcal{O}}\left(\eta d^{-2C_b+4}\right).$$

Hence, we will focus on estimating the first term.

B.1 Estimation of Label-Gradient Correlation

We first establish a high-probability guarantee for the term containing context examples.

Lemma B.1. *Let $(\boldsymbol{x}_1,y_1,\ldots,\boldsymbol{x}_N,y_N,\boldsymbol{x})$ be a prompt with context length $N\le N^*$ and feature vector $\boldsymbol{\beta}\in S_r$ and its embedding $\boldsymbol{Z}\in\mathbb{R}^{(\tilde{d}+1)\times(N+1)}$. Then, the following holds with high probability:*

$$N^{-1}\sum_{j=1}^{N} G_{j,N+1}\left(\boldsymbol{Z}\right) y_j\phi\left(\boldsymbol{x}_j\right) = \psi\left(\boldsymbol{\beta},a_0,a_1,a_2\right)+\tilde{\mathcal{O}}\left(d^{-C_b+1}N^{-1/2}\right).$$

Proof of Lemma B.1. Recall that, with high probability, $y_j/\rho=\bar{g}_*(\langle\boldsymbol{\beta},\boldsymbol{x}_j\rangle)+\zeta_j/\rho$ simultaneously for all $j\in[N]$. On this event,

$$\begin{aligned}&\frac{1}{N}\sum_{j=1}^{N} y_j\sigma(y_j/\rho+b)\phi(\boldsymbol{x}_j)\\ =\ &\frac{1}{N}\sum_{j=1}^{N}\left(\rho\bar{g}_*(\langle\boldsymbol{\beta},\boldsymbol{x}_j\rangle)+\zeta_j\right)\sigma\left(\bar{g}_*(\langle\boldsymbol{\beta},\boldsymbol{x}_j\rangle)+\zeta_j/\rho+b\right)\phi(\boldsymbol{x}_j).\end{aligned}$$

For fixed $\boldsymbol{\beta}$, the truncated summands on the right are independent. After multiplication by d^{C_b}, their scalar multipliers have bounded fourth moments and are bounded by a polylogarithm with high

probability. By the definition of A and Gaussian conditional expectation for the first two Hermite degrees,

$$\begin{aligned}&\mathbb{E}\left[(\rho\bar{g}_*(\langle\boldsymbol{\beta},\boldsymbol{x}_j\rangle)+\zeta_j)\,\sigma\left(\bar{g}_*(\langle\boldsymbol{\beta},\boldsymbol{x}_j\rangle)+\zeta_j/\rho+b\right)\phi(\boldsymbol{x}_j)\right]\\&=\mathbb{E}\left[A(\langle\boldsymbol{\beta},\boldsymbol{x}_j\rangle)\phi(\boldsymbol{x}_j)\right]=\psi(\boldsymbol{\beta},a_0,a_1,a_2).\end{aligned}$$

Applying Lemma A.4 to these truncated summands coordinatewise, and taking a union bound over the $O(d^2)$ embedding coordinates, therefore gives

$$\left\|\frac{1}{N}\sum_{j=1}^{N}y_j\sigma(y_j/\rho+b)\phi(\boldsymbol{x}_j)-\psi(\boldsymbol{\beta},a_0,a_1,a_2)\right\|=\tilde{\mathcal{O}}(d^{-C_b+1}N^{-1/2})$$

with high probability.

In addition, all gate probabilities are $O(d^{-C_b})$ with high probability, and

$$0\le 1-(1-\sigma(b))\prod_{i=j+1}^{N}(1-\sigma(y_i/\rho+b))\le C(N+1)d^{-C_b}.$$

Using $|y_j|=\tilde{\mathcal{O}}(1)$ and $\|\phi(\boldsymbol{x}_j)\|=\tilde{\mathcal{O}}(d)$, we obtain

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{j=1}^{N}G_{j,N+1}(\boldsymbol{Z})y_j\phi(\boldsymbol{x}_j)-\frac{1}{N}\sum_{j=1}^{N}y_j\sigma(y_j/\rho+b)\phi(\boldsymbol{x}_j)\right\|\\&\le\frac{1}{N}\sum_{j=1}^{N}|y_j\sigma(y_j/\rho+b)|\left(1-(1-\sigma(b))\prod_{i=j+1}^{N}(1-\sigma(y_i/\rho+b))\right)\|\phi(\boldsymbol{x}_j)\|\\&=\tilde{\mathcal{O}}(Nd^{-2C_b+1}).\end{aligned}$$

For $N\le d^{C^*}$ and $C_b>3C^*/2$, this is absorbed into $\tilde{\mathcal{O}}(d^{-C_b+1}N^{-1/2})$. The triangle inequality completes the proof.

□

Corollary B.2. *For each $t\in[T_1]$, the following holds with high probability:*

$$\begin{aligned}&N_{\mathrm{pt}}^{-1}\sum_{j=1}^{N_{\mathrm{pt}}}G_{j,N_{\mathrm{pt}}+1}\left(\boldsymbol{Z}^t\right)y_j^t\phi\left(\boldsymbol{x}_j^t\right)^\top\left(\boldsymbol{\gamma}^{(0)}\odot\phi\left(\boldsymbol{x}^t\right)\right)\\&=a_0\gamma^2+a_1\left\langle\boldsymbol{\beta}^t,\boldsymbol{x}^t\right\rangle+\frac{1}{2}a_2\gamma\mathrm{He}_2\left(\left\langle\boldsymbol{\beta}^t,\boldsymbol{x}^t\right\rangle\right)+\tilde{\mathcal{O}}\left(d^{-C_b+2}N_{\mathrm{pt}}^{-1/2}\right).\end{aligned}$$

Proof of Corollary B.2. From Lemma B.1, for each $t\in[T_1]$, with high probability, we have

$$\begin{aligned}&N_{\mathrm{pt}}^{-1}\sum_{j=1}^{N_{\mathrm{pt}}}G_{j,N_{\mathrm{pt}}+1}\left(\boldsymbol{Z}^t\right)y_j^t\phi\left(\boldsymbol{x}_j^t\right)^\top\left(\boldsymbol{\gamma}^{(0)}\odot\phi\left(\boldsymbol{x}^t\right)\right)\\&=a_0\gamma^2+a_1\left\langle\boldsymbol{\beta}^t,\boldsymbol{x}^t\right\rangle+\frac{1}{2}a_2\gamma\mathrm{He}_2\left(\left\langle\boldsymbol{\beta}^t,\boldsymbol{x}^t\right\rangle\right)+\tilde{\mathcal{O}}\left(d^{-C_b+1}N_{\mathrm{pt}}^{-1/2}\right)\left\|\phi\left(\boldsymbol{x}^t\right)\right\|\\&=a_0\gamma^2+a_1\left\langle\boldsymbol{\beta}^t,\boldsymbol{x}^t\right\rangle+\frac{1}{2}a_2\gamma\mathrm{He}_2\left(\left\langle\boldsymbol{\beta}^t,\boldsymbol{x}^t\right\rangle\right)+\tilde{\mathcal{O}}\left(d^{-C_b+2}N_{\mathrm{pt}}^{-1/2}\right),\end{aligned}$$

where we use Lemma A.3 and (4) for the last equality. □

Next, let us estimate the expectation of the first term that appears in the label-gradient correlation. The following lemma is useful for this purpose.

Lemma B.3. *For any $\delta>0$ with $\delta=\tilde{\mathcal{O}}(d^{-C_\delta})$ for some constant $C_\delta>0$, the following holds:*

$$\mathbb{P}_{\mathbf{z}\sim\mathcal{N}(0,1)}\left[\left|a_0\gamma^2+a_1\mathbf{z}+\frac{1}{2}a_2\gamma\mathrm{He}_2\left(\mathbf{z}\right)\right|<d^{-C_b}\delta\right]\le\tilde{\mathcal{O}}\left(d^{-C_\delta}\right).$$

Proof of Lemma B.3. Let $\alpha_i = d^{C_b} a_i$ and $h(z) = \alpha_0\gamma^2 + \alpha_1 z + \alpha_2\gamma(z^2-1)/2$. Lemma A.5 gives $\alpha_0 = \Theta(\rho^{-1})$ and $|\alpha_p| = \Theta(\rho^{-(e_p(g_*)-1)})$ whenever $p = 1, 2$ and $e_p(g_*) < \infty$. For even links, $\alpha_1 = 0$ exactly.

If $\alpha_2 = 0$, then $\alpha_1 \neq 0$, and the event $|h(Z)| < \delta$ is an interval of length $2\delta/|\alpha_1|$. Its Gaussian probability is $\tilde{\mathcal{O}}(\delta)$. Otherwise, we define

$$D := \alpha_1^2 + \alpha_2^2\gamma^2 - 2\alpha_0\alpha_2\gamma^3.$$

For non-even links, we choose C_γ sufficiently large that $|\alpha_2|\gamma = o(|\alpha_1|)$ and $|\alpha_0\alpha_2|\gamma^3 = o(\alpha_1^2)$; then $D \asymp \alpha_1^2$. For even links, we choose $|\alpha_0|\gamma = o(|\alpha_2|)$; then $D \asymp \alpha_2^2\gamma^2$. In both cases $D > 0$ and $\sqrt{D}$ is bounded below by an inverse power of $\log d$.

We obtain the endpoints of $\{|h(z)| < \delta\}$ from the roots of $h(z) = \pm\delta$, whose discriminants are $D \pm 2\alpha_2\gamma\delta$. Since δ is polynomially small, $|\alpha_2|\gamma\delta = o(D)$. By rationalizing the differences of these roots, we bound the total length of the intervals by

$$\frac{C\delta}{\sqrt{D + 2\alpha_2\gamma\delta} + \sqrt{D - 2\alpha_2\gamma\delta}} = O(\delta/\sqrt{D}) = \tilde{\mathcal{O}}(\delta).$$

We use the absolute lengths of the intervals, so the bound holds for either sign of α_2. Since the standard Gaussian density is bounded, the same bound holds for the probability. □

For each $t \in [T_1]$, let E_t be the event that

$$\mathbb{1}\left[N_{\text{pt}}^{-1}\sum_{j=1}^{N_{\text{pt}}} G_{j,N_{\text{pt}}+1}(\boldsymbol{Z}^t) y_j^t \phi(\boldsymbol{x}_j^t)^\top(\boldsymbol{\gamma}^{(0)} \odot \phi(\boldsymbol{x}^t)) > 0\right]$$
$$\neq \mathbb{1}\left[a_0\gamma^2 + a_1\langle\boldsymbol{\beta}^t, \boldsymbol{x}^t\rangle + \frac{1}{2}a_2\gamma\text{He}_2(\langle\boldsymbol{\beta}^t, \boldsymbol{x}^t\rangle) > 0\right].$$

From Corollary B.2 and Lemma B.3, we have

$$\mathbb{P}_{\boldsymbol{Z}^t}[E_t] \le \mathbb{P}_{\boldsymbol{Z}^t}\left[\left|a_0\gamma^2 + a_1\left\langle\boldsymbol{\beta}^t, \boldsymbol{x}^t\right\rangle + \frac{1}{2}a_2\gamma\text{He}_2\left(\left\langle\boldsymbol{\beta}^t, \boldsymbol{x}^t\right\rangle\right)\right| < \tilde{\mathcal{O}}\left(d^{-C_b+2}N_{\text{pt}}^{-1/2}\right)\right] + \frac{1}{\text{poly}(d)}$$
$$= \tilde{\mathcal{O}}\left(d^2 N_{\text{pt}}^{-1/2}\right).$$

Combining with Corollary B.2, with probability at least $1 - \tilde{\mathcal{O}}\left(d^2 T_1 N_{\text{pt}}^{-1/2}\right)$ the following holds: For any $t \in [T_1]$, we have

$$y^t\nabla_{\boldsymbol{\gamma}} f\left(\boldsymbol{Z}^t; \boldsymbol{\gamma}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{a}^{(0)}\right)$$
$$= y^t\mathbb{1}\left[N_{\text{pt}}^{-1}\sum_{j=1}^{N_{\text{pt}}} G_{j,N_{\text{pt}}+1}\left(\boldsymbol{Z}^t\right) y_j^t\phi\left(\boldsymbol{x}_j^t\right)^\top\left(\boldsymbol{\gamma}^{(0)} \odot \phi\left(\boldsymbol{x}^t\right)\right) > 0\right]$$
$$\times\left(N_{\text{pt}}^{-1}\sum_{j=1}^{N_{\text{pt}}} G_{j,N_{\text{pt}}+1}\left(\boldsymbol{Z}^t\right) y_j^t\phi\left(\boldsymbol{x}_j^t\right) \odot \phi\left(\boldsymbol{x}^t\right)\right)$$
$$= y^t\mathbb{1}\left[a_0\gamma^2 + a_1\left\langle\boldsymbol{\beta}^t, \boldsymbol{x}^t\right\rangle + \frac{1}{2}a_2\gamma\text{He}_2\left(\left\langle\boldsymbol{\beta}^t, \boldsymbol{x}^t\right\rangle\right) > 0\right]\phi(\boldsymbol{x}^t) \odot \psi\left(\boldsymbol{\beta}^t, a_0, a_1, a_2\right)$$
$$+ y^t\mathbb{1}\left[a_0\gamma^2 + a_1\left\langle\boldsymbol{\beta}^t, \boldsymbol{x}^t\right\rangle + \frac{1}{2}a_2\gamma\text{He}_2\left(\left\langle\boldsymbol{\beta}^t, \boldsymbol{x}^t\right\rangle\right) > 0\right]\phi(\boldsymbol{x}^t) \odot \boldsymbol{n}\left(\boldsymbol{Z}^t\right),$$

where $\|\boldsymbol{n}(\boldsymbol{Z}^t)\| = \tilde{\mathcal{O}}\left(d^{-C_b+1}N_{\text{pt}}^{-1/2}\right)$ with high probability. With high probability, the following holds for all $t \in [T_1]$:

$$\left\|y^t\mathbb{1}\left[a_0\gamma^2 + a_1\left\langle\boldsymbol{\beta}^t, \boldsymbol{x}^t\right\rangle + \frac{1}{2}a_2\gamma\text{He}_2\left(\left\langle\boldsymbol{\beta}^t, \boldsymbol{x}^t\right\rangle\right) > 0\right]\phi\left(\boldsymbol{x}^t\right) \odot \boldsymbol{n}(\boldsymbol{Z}^t)\right\|$$

$$\leq |y^t| \left\|\boldsymbol{n}\left(\boldsymbol{Z}^t\right)\right\| \left\|\phi\left(\boldsymbol{x}^t\right)\right\|$$
$$= \tilde{\mathcal{O}}\left(d^{-C_b+2} N_{\mathrm{pt}}^{-1/2}\right).$$

Estimation of label-gradient correlation. With probability at least $1 - \tilde{\mathcal{O}}\left(d^2 T_1 N_{\mathrm{pt}}^{-\frac{1}{2}}\right)$, the following holds for all $t \in [T_1]$:

$$y^t \nabla_{\boldsymbol{\gamma}} f\left(\boldsymbol{Z}^t; \boldsymbol{\gamma}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{a}^{(0)}\right)$$
$$= y^t \mathbb{1}\left[a_0\gamma^2 + a_1\left\langle \boldsymbol{\beta}^t, \boldsymbol{x}^t\right\rangle + \frac{1}{2}a_2\gamma \mathrm{He}_2\left(\left\langle \boldsymbol{\beta}^t, \boldsymbol{x}^t\right\rangle\right) > 0\right] \phi(\boldsymbol{x}^t) \odot \psi\left(\boldsymbol{\beta}^t, a_0, a_1, a_2\right)$$
$$+ \tilde{\mathcal{O}}\left(d^{-C_b+2} N_{\mathrm{pt}}^{-1/2}\right).$$

B.2 Characterization of Updated Parameter

In this step, we characterize the updated parameter by establishing concentration results.

Every entry of $\psi(\boldsymbol{\beta}^t, a_0, a_1, a_2)$ is $\tilde{\mathcal{O}}(d^{-C_b})$. By applying Lemma A.4 coordinatewise to the independent tasks, we obtain, with high probability,

$$\left\| \frac{1}{T_1}\sum_{t=1}^{T_1} y^t \mathbb{1}\left[a_0\gamma^2 + a_1\langle \boldsymbol{\beta}^t, \boldsymbol{x}^t\rangle + \frac{1}{2}a_2\gamma \mathrm{He}_2(\langle \boldsymbol{\beta}^t, \boldsymbol{x}^t\rangle) > 0\right] \phi(\boldsymbol{x}^t) \odot \psi(\boldsymbol{\beta}^t, a_0, a_1, a_2) - \boldsymbol{c} \right\|$$
$$= \tilde{\mathcal{O}}(d^{-C_b+1} T_1^{-1/2}),$$

where $\boldsymbol{c}$ is the expectation of a summand. We define

$$B(z) = g_*(z)\mathbb{1}[a_0\gamma^2 + a_1 z + \tfrac{1}{2}a_2\gamma \mathrm{He}_2(z) > 0], \qquad B(z) = \sum_{j=0}^{\infty} \frac{b_j}{j!}\mathrm{He}_j(z).$$

The query noise is centered and independent of the activation indicator. Thus, by Gaussian conditional expectation, we have

$$\boldsymbol{c} = \mathbb{E}_{\boldsymbol{\beta}\sim \mathrm{Unif}(S_r)}\left[\psi(\boldsymbol{\beta}, a_0, a_1, a_2) \odot \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[B(\langle \boldsymbol{\beta}, \boldsymbol{x}\rangle)\phi(\boldsymbol{x})]\right]$$
$$= \mathbb{E}_{\boldsymbol{\beta}\sim \mathrm{Unif}(S_r)}[\psi(\boldsymbol{\beta}, a_0, a_1, a_2) \odot \psi(\boldsymbol{\beta}, b_0, b_1, b_2)].$$

Therefore, we conclude that

$$\frac{1}{T_1}\sum_{t=1}^{T_1} y^t \nabla_{\boldsymbol{\gamma}} f(\boldsymbol{Z}^t; \boldsymbol{\gamma}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{a}^{(0)})$$
$$= \mathbb{E}_{\boldsymbol{\beta}\sim \mathrm{Unif}(S_r)}[\psi(\boldsymbol{\beta}, a_0, a_1, a_2) \odot \psi(\boldsymbol{\beta}, b_0, b_1, b_2)] + \tilde{\mathcal{O}}(d^{-C_b+2} N_{\mathrm{pt}}^{-1/2}) + \tilde{\mathcal{O}}(d^{-C_b+1} T_1^{-1/2}),$$

with probability at least $1 - \tilde{\mathcal{O}}\left(d^2 T_1 N_{\mathrm{pt}}^{-1/2}\right)$.

The remaining step is to characterize the Hermite coefficients b_0, b_1, b_2, and the following lemma is useful.

Lemma B.4. *For any non zero polynomial P independent of r and d, the following holds except for the cases g_* is even function and P is an odd function:*

$$\left|\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,1)}\left[P(\mathbf{z})\mathbb{1}\left[a_0\gamma^2 + a_1\mathbf{z} + \frac{1}{2}a_2\gamma \mathrm{He}_2(\mathbf{z}) > 0\right]\right]\right| \gtrsim \frac{1}{\mathrm{polylog}(d)}.$$

Here, dependency of g_* appears in a_0, a_1, a_2.

Proof of Lemma B.4. Let $\varphi(z) = (2\pi)^{-1/2} e^{-z^2/2}$ and $h(z) = a_0\gamma^2 + a_1 z + a_2\gamma(z^2-1)/2$. By Lemma A.5, all nonzero coefficients used below, after multiplication by d^{C_b}, are bounded above

and below in magnitude by powers of $\log d$. In particular, $a_0 > 0$ and $e_2(g_*) < \infty$. We choose C_γ sufficiently large relative to these exponents.

Case 1: g_* is not an even function. Here $a_1 \neq 0$. By substituting $z \mapsto -z$ and replacing P accordingly, we may assume $a_1 > 0$. If $a_2 = 0$, the event $h(z) > 0$ is $z > t$, where $t = -a_0\gamma^2/a_1$. This threshold tends to zero and its magnitude is bounded below by an inverse power of $\log d$.

If $a_2 \neq 0$, set

$$D = a_1^2 + a_2^2\gamma^2 - 2a_0a_2\gamma^3, \qquad t = \frac{-a_1 + \sqrt{D}}{a_2\gamma} = \frac{(a_2 - 2a_0\gamma)\gamma}{a_1 + \sqrt{D}}.$$

We choose γ so that $|a_0|\gamma = o(|a_2|)$ and $|a_2|\gamma = o(|a_1|)$. Then

$$D \asymp a_1^2, \qquad |t| \asymp |a_2|\gamma/|a_1| = o(1).$$

By the coefficient bounds, $|t|$ is bounded below by an inverse power of $\log d$, in addition to the upper bound on $|t|$. The other root is $t_{\mathrm{far}} = (-a_1 - \sqrt{D})/(a_2\gamma)$. For sufficiently large C_γ, we have $|t_{\mathrm{far}}| \geq (\log d)^2$. Since $h'(t) = \sqrt{D} > 0$, the indicators of $h(z) > 0$ and $z > t$ differ only beyond this root for either sign of a_2. Hence, for every fixed $M > 0$,

$$\mathbb{E}[P(Z)\mathbb{1}[h(Z) > 0]] = F(t) + O(d^{-M}), \qquad F(t) := \int_t^\infty P(z)\varphi(z)\,\mathrm{d}z.$$

The same identity holds without the tail error when $a_2 = 0$. If $F(0) \neq 0$, its magnitude stays bounded away from zero. If $F(0) = 0$, write $P(z) = c_\nu z^\nu + O(z^{\nu+1})$ with $c_\nu \neq 0$. By integrating near the boundary, we have

$$F(t) = -\frac{c_\nu\varphi(0)}{\nu+1}t^{\nu+1} + O(|t|^{\nu+2}).$$

The lower bound on $|t|$ and the Gaussian tail bound give the desired inverse-polylogarithmic bound.

Case 2: g_* is an even function. Here $a_1 = 0$ exactly, $a_2 \neq 0$, and only the even part P_{e} of P contributes. By the assumption on P, we have $P_{\mathrm{e}} \neq 0$. Set $t = (1 - 2a_0\gamma/a_2)^{1/2}$. With $|a_0|\gamma = o(|a_2|)$,

$$t \longrightarrow 1, \qquad |t - 1| \asymp |a_0|\gamma/|a_2|.$$

The last quantity is bounded below by an inverse power of $\log d$. The required expectation is

$$F(t) = \begin{cases} 2\int_t^\infty P_{\mathrm{e}}(z)\varphi(z)\,\mathrm{d}z, & a_2 > 0, \\ 2\int_0^t P_{\mathrm{e}}(z)\varphi(z)\,\mathrm{d}z, & a_2 < 0. \end{cases}$$

If $F(1) \neq 0$, the conclusion follows by continuity. Otherwise, let ν be the multiplicity of the zero of P_{e} at 1 (with $\nu = 0$ if it is nonzero there). By Taylor expansion of the integrand, we have $|F(t)| \asymp |t-1|^{\nu+1}$, which gives the desired lower bound for both signs. □

By applying Lemma B.4 to $P(z) = g_*(z)\mathrm{He}_j(z)$ for $j = 0, 1, 2$, we obtain $|b_0|, |b_2| = \tilde{\Theta}(1)$ and $|b_1| = \tilde{\Theta}(1)$ for non-even links. For even links, B is even and $b_1 = 0$ exactly. These bounds hold for either sign of the nonzero coefficients.

Updated Parameter $\boldsymbol{\gamma}^*$. With probability at least $1 - \tilde{\mathcal{O}}\left(d^2T_1N_{\mathrm{pt}}^{-1/2}\right)$, the updated parameter $\boldsymbol{\gamma}^*$ is given by:

$$\begin{aligned} \boldsymbol{\gamma}^* = {} & 2\eta\mathbb{E}_{\boldsymbol{\beta}\sim\mathrm{Unif}(S_r)}\left[\psi(\boldsymbol{\beta}, a_0, a_1, a_2) \odot \psi(\boldsymbol{\beta}, b_0, b_1, b_2)\right] \\ & + \eta\tilde{\mathcal{O}}\left(\max\left\{d^{-2C_b+4}, d^{-C_b+2}N_{\mathrm{pt}}^{-1/2}, d^{-C_b+1}T_1^{-1/2}\right\}\right), \end{aligned} \tag{5}$$

Here, we have $d^{C_b}a_0, d^{C_b}|a_{\mathrm{ge}(g_*)}|, |b_0|, |b_2| = \tilde{\Theta}(1)$ and $d^{C_b}|a_{3-\mathrm{ge}(g_*)}| = \tilde{\mathcal{O}}(1)$. In addition, $|b_1| = \tilde{\Theta}(1)$ if $\mathrm{ge}(g_*) = 1$. If g_* is even, we have $a_1 = b_1 = 0$ by symmetry.

B.3 Output of Updated Mamba Layer

Lastly, we characterize the output of the Mamba layer with the updated parameter $\boldsymbol{\gamma}^*$, which serves as the input to the MLP layer. This characterization is given in the following proposition, which is a formal statement of Proposition 4.1.

Proposition B.5. *Let $(\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_N, y_N, \boldsymbol{x})$ be a prompt from the task distribution with $\boldsymbol{\beta} \sim \mathrm{Unif}(S_r)$ and $N \le N^*$. Suppose (5) holds, $N = \tilde{\Omega}(r^{3\mathrm{ge}(g_*)})$, and $\eta = \Theta(d^{2C_b}(\log d)^{-C_\eta})$ with sufficiently large C_η. Let $q = \mathrm{ge}(g_*)$, $k = \deg(g_*)$, and $s = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle$. Under the hyperparameter conditions of Theorem 3.3, the following holds with high probability:*

$$X_N := N^{-1}\mathsf{Mamba}(\boldsymbol{Z}; \boldsymbol{\gamma}^*)[\tilde{d}+1, N+1] = P_1 + P_2(s/r)^q + Q(\boldsymbol{\beta}, \boldsymbol{x}) + R_{\mathrm{mix}}(\boldsymbol{\beta}, \boldsymbol{x}) + \epsilon_N, \quad (6)$$

where

$$P_1 = 2\eta\left(a_0^2 b_0 - \frac{a_2^2 b_2}{2r(r+2)}\right), \qquad B_2 = \frac{\eta r}{r+2} a_2^2 b_2,$$

$$P_2 = \begin{cases} 2\eta a_1^2 b_1, & q = 1, \\ B_2, & q = 2, \end{cases} \qquad Q(\boldsymbol{\beta}, \boldsymbol{x}) = \frac{B_2}{2r^2} \sum_{i \in \mathcal{I}} \boldsymbol{\beta}[i]^2 (\boldsymbol{x}[i]^2 - 1),$$

$$R_{\mathrm{mix}}(\boldsymbol{\beta}, \boldsymbol{x}) = \mathbb{1}[q = 1] B_2 (s/r)^2.$$

These coefficients are independent of the prompt data, $P_1 = o(1)$, and $|P_2| = \Theta((\log d)^{-C_{P_2}})$ for some $C_{P_2} > 0$. The error satisfies

$$|\epsilon_N| \le \tilde{\mathcal{O}}\Big(N^{-1/2} + d^{-C_b+6} + d^{-C_b+C^*} + d^4 N_{\mathrm{pt}}^{-1/2} + d^3 T_1^{-1/2}\Big). \quad (7)$$

For $q = 1$, we choose a fixed $L_{\mathrm{mix}} \ge 0$ such that $|B_2/P_2| = O((\log d)^{L_{\mathrm{mix}}})$; for $q = 2$, we set $L_{\mathrm{mix}} = 0$. A sufficient growth condition is

$$r \ge (\log d)^{C_r}, \qquad C_r \ge \max\{4k + 2, L_{\mathrm{mix}} + 2k + 2\}. \quad (8)$$

With sufficiently large logarithmic factors in the sample sizes and sufficiently large C_b, we have

$$|Q + R_{\mathrm{mix}} + \epsilon_N| = o(|P_2| r^{-q} (\log d)^{-2k}) \quad (9)$$

with high probability. For even links, $R_{\mathrm{mix}} = 0$ and $C_r \ge 4k + 2$ suffices.

We fix C^* first and successively choose $C_b, C_\rho, C_\gamma, C_\eta, C_r$ sufficiently large. The logarithmic factors in the sample sizes and the high-probability constants are chosen last.

Proof of Proposition B.5. Let $\boldsymbol{c} = \mathbb{E}_{\boldsymbol{\beta}' \sim \mathrm{Unif}(S_r)}[\psi(\boldsymbol{\beta}', a_0, a_1, a_2) \odot \psi(\boldsymbol{\beta}', b_0, b_1, b_2)]$. By the Cauchy–Schwarz inequality and (4), we have

$$\|\boldsymbol{c}\|_1 \le \mathbb{E}_{\boldsymbol{\beta}'}[\|\psi(\boldsymbol{\beta}', a_0, a_1, a_2)\| \|\psi(\boldsymbol{\beta}', b_0, b_1, b_2)\|] = \tilde{\mathcal{O}}(d^{-C_b}).$$

From (5), we can write $\boldsymbol{\gamma}^* = 2\eta \boldsymbol{c} + \mathbf{e}_\gamma$ with

$$\|\mathbf{e}_\gamma\| = \eta \tilde{\mathcal{O}}(d^{-2C_b+4} + d^{-C_b+2} N_{\mathrm{pt}}^{-1/2} + d^{-C_b+1} T_1^{-1/2}).$$

On the Gaussian high-probability event, $|y_j| = \tilde{\mathcal{O}}(1)$, all gates are $O(d^{-C_b})$, and $\|\phi(\boldsymbol{x}_j)\|, \|\phi(\boldsymbol{x})\| = \tilde{\mathcal{O}}(d)$. Thus, the contribution of $\mathbf{e}_\gamma$ to X_N is bounded by

$$\tilde{\mathcal{O}}(d^{-C_b+2}) \|\mathbf{e}_\gamma\| = \tilde{\mathcal{O}}(d^{-C_b+6} + d^4 N_{\mathrm{pt}}^{-1/2} + d^3 T_1^{-1/2}).$$

For the contribution of $2\eta \boldsymbol{c}$, the error from replacing $G_{j,N+1}$ by the single gate $\sigma(y_j/\rho + b)$ is bounded by

$$2\eta\, \tilde{\mathcal{O}}(d^{-2C_b})(N+1) d^{-C_b} = \tilde{\mathcal{O}}(d^{-C_b+C^*}).$$

Here, we used $N \le d^{C^*}$, $\|\phi(\boldsymbol{x}_j)\|_\infty, \|\phi(\boldsymbol{x})\|_\infty = \tilde{\mathcal{O}}(1)$, and the bound on $\|\boldsymbol{c}\|_1$.

We use the same truncation as in Lemma B.1. Conditional on $\boldsymbol{\beta}, \boldsymbol{x}$, the variables

$$V_j = 2\eta\Big(\rho \bar{g}_*\big(\langle \boldsymbol{\beta}, \boldsymbol{x}_j \rangle\big) + \zeta_j\Big) \sigma\Big(b + \bar{g}_*\big(\langle \boldsymbol{\beta}, \boldsymbol{x}_j \rangle\big) + \zeta_j/\rho\Big) \phi(\boldsymbol{x}_j)^\top \big(\boldsymbol{c} \odot \phi(\boldsymbol{x})\big)$$

are independent and equal $2\eta y_j \sigma(y_j/\rho + b)\phi(\boldsymbol{x}_j)^\top(\boldsymbol{c} \odot \phi(\boldsymbol{x}))$ simultaneously with high probability. For sufficiently large C_η, $|V_j| \le 1$ with high probability. The truncated gate is uniformly $O(d^{-C_b})$, and the magnitude of the truncated label is at most $|g_*(\langle \boldsymbol{\beta}, \boldsymbol{x}_j\rangle)| + \tau$. Since each embedding coordinate has bounded fixed moments and $2\eta d^{-C_b}\|\boldsymbol{c}\|_1\|\phi(\boldsymbol{x})\|_\infty = \tilde{\mathcal{O}}(1)$, the conditional second moments of V_j are $\tilde{\mathcal{O}}(1)$ on the query event.

We apply Höeffding's inequality to $V_j \mathbb{1}[|V_j| \le 1]$. By Cauchy–Schwarz, the difference of its expectation from $\mathbb{E}[V_j \mid \boldsymbol{\beta}, \boldsymbol{x}]$ is bounded by

$$\left(\mathbb{E}\left[V_j^2 \mid \boldsymbol{\beta}, \boldsymbol{x}\right]\right)^{1/2} \mathbb{P}\left[|V_j| > 1 \mid \boldsymbol{\beta}, \boldsymbol{x}\right]^{1/2},$$

which is negligible compared with $N^{-1/2}$ for sufficiently large high-probability constants. By the expectation calculation in Lemma B.1, we obtain

$$X_N = 2\eta\psi(\boldsymbol{\beta}, a_0, a_1, a_2)^\top(\boldsymbol{c} \odot \phi(\boldsymbol{x})) + \epsilon_N, \tag{10}$$

with (7).

We now compute the population term. For distinct $i, j \in \mathcal{I}$,

$$\mathbb{E}[\boldsymbol{\beta}'[i]^2] = \frac{1}{r}, \qquad \mathbb{E}[\boldsymbol{\beta}'[i]^4] = \frac{3}{r(r+2)}, \qquad \mathbb{E}[\boldsymbol{\beta}'[i]^2\boldsymbol{\beta}'[j]^2] = \frac{1}{r(r+2)}.$$

Coordinates outside $\mathcal{I}$ vanish. By substituting these moments into the coordinatewise product, we have

$$\begin{aligned}
&\psi(\boldsymbol{\beta}, a_0, a_1, a_2)^\top(\boldsymbol{c} \odot \phi(\boldsymbol{x})) \\
&= a_0^2 b_0 + \frac{a_1^2 b_1}{r} s + \frac{3a_2^2 b_2}{4r(r+2)} \sum_{i\in\mathcal{I}} \boldsymbol{\beta}[i]^2(\boldsymbol{x}[i]^2 - 1) + \frac{a_2^2 b_2}{r(r+2)} \sum_{\substack{i,j\in\mathcal{I} \\ i<j}} \boldsymbol{\beta}[i]\boldsymbol{\beta}[j]\boldsymbol{x}[i]\boldsymbol{x}[j] \\
&= a_0^2 b_0 - \frac{a_2^2 b_2}{2r(r+2)} + a_1^2 b_1 \frac{s}{r} + \frac{a_2^2 b_2 r}{2(r+2)}(s/r)^2 + \frac{a_2^2 b_2}{4r(r+2)} \sum_{i\in\mathcal{I}} \boldsymbol{\beta}[i]^2(\boldsymbol{x}[i]^2 - 1).
\end{aligned}$$

Multiplying by 2η gives (6). For even links, $a_1 = b_1 = 0$. For non-even links, both degrees can occur. Using the coefficient magnitude bounds with a sufficiently large C_η, we obtain $P_1 = o(1)$ and the stated inverse-polylogarithmic magnitude of P_2.

It remains to bound the residuals. Write the nonzero coordinates of $\boldsymbol{\beta}$ as $\mathbf{g}/\|\mathbf{g}\|$, where $\mathbf{g} \sim \mathcal{N}(0, \boldsymbol{I}_r)$. Gaussian concentration for the 1-Lipschitz function $\|\mathbf{g}\|_4$, together with $\mathbb{E}[\|\mathbf{g}\|_4] \le (3r)^{1/4}$ and the chi-square lower-tail bound, gives

$$\sum_{i\in\mathcal{I}} \boldsymbol{\beta}[i]^4 = O(r^{-1}), \qquad \max_{i\in\mathcal{I}} \boldsymbol{\beta}[i]^2 = O((\log d)/r)$$

with high probability under (8), since $k \ge 1$ implies $C_r \ge 6$; the second bound follows by a coordinatewise Gaussian tail bound. Conditional on such a direction, we apply Bernstein's inequality to obtain

$$\begin{aligned}
\left|\sum_{i\in\mathcal{I}} \boldsymbol{\beta}[i]^2(\boldsymbol{x}[i]^2 - 1)\right| &\le O\left(\sqrt{(\log d)\sum_i \boldsymbol{\beta}[i]^4} + (\log d)\max_i \boldsymbol{\beta}[i]^2\right) \\
&= O\left(\sqrt{\frac{\log d}{r}} + \frac{(\log d)^2}{r}\right) = O\left(\sqrt{\frac{\log d}{r}}\right).
\end{aligned}$$

The last equality follows from (8). Thus, $|Q| = O(|B_2| r^{-5/2}(\log d)^{1/2})$.

For $q = 2$, $B_2 = P_2$ and $R_{\mathrm{mix}} = 0$. The condition $C_r \ge 4k + 2$ implies $r^{-1/2}(\log d)^{2k+1/2} \le (\log d)^{-1/2} = o(1)$, so $|Q| = o(|P_2| r^{-2}(\log d)^{-2k})$. For $q = 1$, $s^2 = O(\log d)$ with high probability, and hence

$$|R_{\mathrm{mix}}| = O(|P_2| r^{-2}(\log d)^{L_{\mathrm{mix}}+1}), \qquad |Q| = O(|P_2| r^{-5/2}(\log d)^{L_{\mathrm{mix}}+1/2}).$$

Both are $o(|P_2|r^{-1}(\log d)^{-2k})$ when $C_r \geq L_{\mathrm{mix}} + 2k + 2$, since

$$r^{-1}(\log d)^{L_{\mathrm{mix}}+2k+1} \leq (\log d)^{-1} = o(1),$$
$$r^{-3/2}(\log d)^{L_{\mathrm{mix}}+2k+1/2} \leq (\log d)^{-(C_r+3)/2} = o(1).$$

Finally, (7) is $o(|P_2|r^{-q}(\log d)^{-2k})$ under the sample-size assumptions with sufficiently large logarithmic factors. The terms containing d^{-C_b} are absorbed when C_b is sufficiently large relative to C^* and the fixed exponents. This proves (9). □

C Optimizing MLP Layer

In this section, we analyze the second stage of pretraining, which focuses on the MLP layer. All probability statements below are conditional on the Stage-I event on which (5) holds.

C.1 Construction of Approximating MLP Layer

First, we construct an infinite-width MLP approximating the link function g_*.

Lemma C.1. *Assume (8). For a given $\boldsymbol{\beta} \in \mathbb{R}^d$ with $\|\boldsymbol{\beta}\| = 1$, suppose*

$$h(\boldsymbol{x}) = P_1 + P_2 \left(\frac{\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle}{r} \right)^{\mathrm{ge}(g_*)} + n(\boldsymbol{x}),$$

where $P_1 = o(1)$, $|P_2| = \Theta((\log d)^{-C_{P_2}})$, and $|n(\boldsymbol{x})| = o(|P_2|r^{-\mathrm{ge}(g_*)}(\log d)^{-2\mathrm{deg}(g_*)+2})$ with high probability. Then there exists a function $\pi : \{\pm 1\} \times [-1, 1] \to \mathbb{R}$, depending on g_*, P_1, P_2, r but not on $\boldsymbol{\beta}$, such that

$$\left| \mathbb{E}_{v \sim \mathrm{Unif}(\{\pm 1\}),\, a \sim \mathrm{Unif}([-1,1])} [\pi(v, a)\mathrm{ReLU}(vh(\boldsymbol{x}) + a)] - g_*(\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle) \right| = o(1)$$

with high probability, and $\sup_{v,a} |\pi(v, a)| = \tilde{\mathcal{O}}(r^{2\mathrm{ge}(g_*)})$. The conclusion also applies when n depends on the context, with the probability statement including that randomness.

Proof of Lemma C.1. By Lemma 3.2, there exists a polynomial $\tilde{g}_*$ such that $g_*(z) = \tilde{g}_*(\mathrm{sign}(P_2)z^{\mathrm{ge}(g_*)})$. Let $\tilde{g}_*(z) = \sum_{k=0}^{\deg(\tilde{g}_*)} s_k z^k$ and $p = |P_2|r^{-\mathrm{ge}(g_*)} > 0$. By Lemma 17 in Damian et al. (2022), for each k there exists a bounded function π'_k such that

$$\mathbb{E}_{v,a}[\pi'_k(v, a)\mathrm{ReLU}(vz + a)] = z^k \quad (|z| \leq 1), \qquad \sup_{v,a} |\pi'_k(v, a)| = O(1),$$

where v and a are uniform on $\{\pm 1\}$ and $[-1, 1]$, respectively. Define

$$\pi'(v, a) = \sum_{k=0}^{\deg(\tilde{g}_*)} s_k \frac{(\log d)^{2k}}{p(\log d)^2} \pi'_k \left(v, \frac{a}{p(\log d)^2} \right) \qquad (|a| \leq p(\log d)^2).$$

By positive homogeneity of ReLU and a change of variables, for $|z| \leq (\log d)^2$ we have

$$\mathbb{E}_{\substack{v \sim \mathrm{Unif}(\{\pm 1\}) \\ a \sim \mathrm{Unif}([-p(\log d)^2, p(\log d)^2])}} [\pi'(v, a)\mathrm{ReLU}(vpz + a)]$$
$$= \sum_{k=0}^{\deg(\tilde{g}_*)} s_k (\log d)^{2k} \mathbb{E}_{v,a} \left[\pi'_k(v, a)\mathrm{ReLU} \left(\frac{vz}{(\log d)^2} + a \right) \right] = \tilde{g}_*(z).$$

Moreover, $\sup_{v,a} |\pi'(v, a)| = O(p^{-1}(\log d)^{2\mathrm{deg}(g_*)-2})$.

To account for the offset and the original bias distribution, define

$$\pi(v, a) = \frac{\mathbb{1}[|a + vP_1| \leq p(\log d)^2]}{p(\log d)^2} \pi'(v, a + vP_1). \tag{11}$$

For large d, $|P_1| + p(\log d)^2 < 1$ by (8). For each v, the selected interval has probability $p(\log d)^2$ under the uniform bias distribution, which accounts for the normalization in (11). Consequently,

$$\mathbb{E}_{v,a} \left[\pi(v, a)\mathrm{ReLU} \left(v(P_1 + pz) + a \right) \right]$$

$$= \mathbb{E}_{v\sim\mathrm{Unif}(\{\pm1\})a\sim\mathrm{Unif}([-p(\log d)^2,p(\log d)^2])}[\pi'(v,a)\mathrm{ReLU}(vpz+a)] = \tilde{g}_*(z).$$

With high probability, $|\langle\boldsymbol{\beta},\boldsymbol{x}\rangle|^{\mathrm{ge}(g_*)} \le (\log d)^2$. Substituting $z = \mathrm{sign}(P_2)\langle\boldsymbol{\beta},\boldsymbol{x}\rangle^{\mathrm{ge}(g_*)}$ gives the target $g_*(\langle\boldsymbol{\beta},\boldsymbol{x}\rangle)$. By the Lipschitz continuity of ReLU, adding $n(\boldsymbol{x})$ changes this expectation by at most

$$\sup_{v,a}|\pi'(v,a)|\,|n(\boldsymbol{x})| = O(p^{-1}(\log d)^{2\deg(g_*)-2})\,|n(\boldsymbol{x})| = o(1).$$

Finally, we have

$$\sup_{v,a}|\pi(v,a)| \le \frac{\sup_{v,a}|\pi'(v,a)|}{p(\log d)^2} = \tilde{\mathcal{O}}(r^{2\mathrm{ge}(g_*)}).$$

Thus, the construction applies to every task direction and either sign of P_2.

□

Next, we approximate the link function with a finite-width MLP.

Lemma C.2. *Let $(\boldsymbol{v}[j],\boldsymbol{a}[j])_{j=1}^m$ be independent and uniform on $\{\pm1\}\times[-1,1]$. Under the conditions of Lemma C.1 and $m = \tilde{\Omega}(r^{4\mathrm{ge}(g_*)})$ with sufficiently large logarithmic factors, there exists a common vector $\boldsymbol{u}' \in \mathbb{R}^m$ such that*

$$\left|\sum_{j=1}^m \boldsymbol{u}'[j]\mathrm{ReLU}(\boldsymbol{v}[j]h(\boldsymbol{x})+\boldsymbol{a}[j]) - g_*(\langle\boldsymbol{\beta},\boldsymbol{x}\rangle)\right| = o(1), \qquad \|\boldsymbol{u}'\|^2 = \tilde{\mathcal{O}}(r^{3\mathrm{ge}(g_*)}m^{-1})$$

with high probability. The approximation holds simultaneously on polynomially many independent prompts.

Proof of Lemma C.2. We choose $\boldsymbol{u}'[j] = \pi(\boldsymbol{v}[j],\boldsymbol{a}[j])/m$ with π from (11). Since $|h(\boldsymbol{x})| \le 1$ with high probability, each activation is at most 2. By Hoeffding's inequality conditional on a prompt,

$$\begin{aligned}\sum_{j=1}^m \boldsymbol{u}'[j]\mathrm{ReLU}(\boldsymbol{v}[j]h(\boldsymbol{x})+\boldsymbol{a}[j]) &= \mathbb{E}_{v,a}[\pi(v,a)\mathrm{ReLU}(vh(\boldsymbol{x})+a)] + \tilde{\mathcal{O}}(r^{2\mathrm{ge}(g_*)}m^{-1/2}) \\ &= g_*(\langle\boldsymbol{\beta},\boldsymbol{x}\rangle) + o(1).\end{aligned}$$

A union bound gives the approximation simultaneously on polynomially many prompts, even though they share the same random features.

By (11), $\pi(v,a)$ is nonzero with probability $\tilde{\mathcal{O}}(r^{-\mathrm{ge}(g_*)})$. Combining this with $\sup_{v,a}|\pi(v,a)| = \tilde{\mathcal{O}}(r^{2\mathrm{ge}(g_*)})$ gives $\mathbb{E}_{v,a}[\pi(v,a)^2] = \tilde{\mathcal{O}}(r^{3\mathrm{ge}(g_*)})$. Another application of Hoeffding's inequality yields

$$\begin{aligned}\|\boldsymbol{u}'\|^2 &= m^{-2}\sum_{j=1}^m \pi(\boldsymbol{v}[j],\boldsymbol{a}[j])^2 \\ &\le \tilde{\mathcal{O}}(r^{3\mathrm{ge}(g_*)}m^{-1} + r^{4\mathrm{ge}(g_*)}m^{-3/2}) = \tilde{\mathcal{O}}(r^{3\mathrm{ge}(g_*)}m^{-1}).\end{aligned}$$

This vector is independent of the Stage-II query-label noises.

□

C.2 Characterization of Estimation Error on the Training Set

The following lemma characterizes the training error after pretraining.

Lemma C.3. *There is a deterministic choice $\lambda_2 > 0$ such that, with probability at least 0.999,*

$$\frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}|y^t - f(\boldsymbol{Z}^t;\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*)| = \tau + o(1), \qquad \|\boldsymbol{u}^*\| = \tilde{\mathcal{O}}(r^{3\mathrm{ge}(g_*)/2}m^{-1/2}).$$

Proof of Lemma C.3. By (6)–(9), the Mamba output satisfies the conditions of Lemma C.1. By Lemma C.2 and a union bound over the Stage-II prompts, there exists $\boldsymbol{u}'$ such that

$$\max_{T_1<t\le T_1+T_2} |f(\boldsymbol{Z}^t;\boldsymbol{\gamma}^*,\boldsymbol{u}',\boldsymbol{v}^*,\boldsymbol{a}^*) - g_*(\langle\boldsymbol{\beta}^t,\boldsymbol{x}^t\rangle)| = o(1),$$
$$\|\boldsymbol{u}'\|^2 = \tilde{\mathcal{O}}(r^{3\mathrm{ge}(g_*)}m^{-1})$$

on an event of probability at least $1-5\cdot10^{-4}$. On this event, the normalized Mamba outputs have magnitude at most 1, so the MLP activations are at most 2.

We condition on Stage I, the random features, and the Stage-II prompts except their query-label noises. The noises $\zeta^t = y^t - g_*(\langle\boldsymbol{\beta}^t,\boldsymbol{x}^t\rangle)$ are then independent, centered, and uniform on $\{\pm\tau\}$. Using this independence and the activation bound gives

$$\mathbb{E}\left[\sum_{j=1}^{m}\left(\frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}\zeta^t\mathrm{ReLU}\left(\boldsymbol{v}^*[j]N_{\mathrm{pt}}^{-1}\mathsf{Mamba}(\boldsymbol{Z}^t;\boldsymbol{\gamma}^*)+\boldsymbol{a}^*[j]\right)\right)^2\right]$$
$$\le \frac{4\tau^2 m}{T_2},$$

where the expectation is over the remaining noises. Here and below, $\mathsf{Mamba}(\boldsymbol{Z}^t;\boldsymbol{\gamma}^*)$ denotes the scalar output at $[\tilde{d}+1, N_{\mathrm{pt}}+1]$, as in the model definition. By Markov's inequality and the Cauchy–Schwarz inequality, with conditional probability at least $1-5\cdot10^{-4}$,

$$\left|\frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}\zeta^t f(\boldsymbol{Z}^t;\boldsymbol{\gamma}^*,\boldsymbol{u},\boldsymbol{v}^*,\boldsymbol{a}^*)\right| \le O\left(\sqrt{\frac{m}{T_2}}\right)\|\boldsymbol{u}\| \qquad \text{for every } \boldsymbol{u}\in\mathbb{R}^m.$$

Choose a deterministic regularization coefficient

$$\lambda_2 = mr^{-3\mathrm{ge}(g_*)}(\log d)^{-C} \tag{12}$$

for a sufficiently large fixed constant $C>0$. The norm bound on $\boldsymbol{u}'$ gives $\lambda_2\|\boldsymbol{u}'\|^2 = o(1)$. With sufficiently large logarithmic factors in the assumed lower bound on T_2, we also have $m/(\lambda_2 T_2) = o(1)$. By optimality of $\boldsymbol{u}^*$ for the empirical squared loss plus $\lambda_2\|\boldsymbol{u}\|^2/2$, expanding around the noiseless targets gives

$$\frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}\left(f(\boldsymbol{Z}^t;\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*) - g_*(\langle\boldsymbol{\beta}^t,\boldsymbol{x}^t\rangle)\right)^2 + \frac{\lambda_2}{2}\|\boldsymbol{u}^*\|^2$$
$$\le o(1) + \frac{\lambda_2}{2}\|\boldsymbol{u}'\|^2 + \frac{2}{T_2}\sum_{t=T_1+1}^{T_1+T_2}\zeta^t f(\boldsymbol{Z}^t;\boldsymbol{\gamma}^*,\boldsymbol{u}^*-\boldsymbol{u}',\boldsymbol{v}^*,\boldsymbol{a}^*)$$
$$\le o(1) + \frac{\lambda_2}{2}\|\boldsymbol{u}'\|^2 + O\left(\sqrt{\frac{m}{T_2}}\right)(\|\boldsymbol{u}^*\| + \|\boldsymbol{u}'\|)$$
$$\le o(1) + \lambda_2\|\boldsymbol{u}'\|^2 + \frac{\lambda_2}{4}\|\boldsymbol{u}^*\|^2 + O\left(\frac{m}{\lambda_2 T_2}\right),$$

where the last step uses Young's inequality. Therefore,

$$\frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}\left(f(\boldsymbol{Z}^t;\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*) - g_*(\langle\boldsymbol{\beta}^t,\boldsymbol{x}^t\rangle)\right)^2 + \frac{\lambda_2}{4}\|\boldsymbol{u}^*\|^2 = o(1).$$

In particular, $\|\boldsymbol{u}^*\|^2 = o(\lambda_2^{-1}) = \tilde{\mathcal{O}}(r^{3\mathrm{ge}(g_*)}m^{-1})$. Finally, since $|\zeta^t| = \tau$, the triangle and Cauchy–Schwarz inequalities imply

$$\left|\frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}|y^t - f(\boldsymbol{Z}^t;\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*)| - \tau\right|$$
$$\le \frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}|f(\boldsymbol{Z}^t;\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*) - g_*(\langle\boldsymbol{\beta}^t,\boldsymbol{x}^t\rangle)| = o(1).$$

A union bound gives the desired conclusions with probability at least 0.999.

□

D Test Error Analysis

We condition throughout this section on the Stage-I event on which (5) holds. Its failure probability is included in the final union bound. We analyze the test-time estimation error

$$\mathcal{R}_{N_{\mathrm{test}}}(\boldsymbol{\gamma}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, \boldsymbol{a}^*) = \mathbb{E}_{(\boldsymbol{Z},y)\sim\mathcal{D}(N_{\mathrm{test}})}[|f(\boldsymbol{Z}; \boldsymbol{\gamma}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, \boldsymbol{a}^*) - y|].$$

As in the model definition, $\mathsf{Mamba}(\boldsymbol{Z}; \boldsymbol{\gamma}^*)$ denotes the scalar output at $[\tilde{d}+1, N+1]$ for a prompt of context length N.

D.1 Test Error for Prompts with Pretraining Context Length

We first prove the conclusion for $N_{\mathrm{test}} = N_{\mathrm{pt}}$ by establishing a generalization bound using Rademacher complexity.

We define a family of functions on inputs with context length N_{pt}:

$$\mathcal{F}_U = \left\{ (\boldsymbol{Z}, y) \mapsto \sum_{j=1}^{m} \boldsymbol{u}[j]\mathrm{ReLU}\left(\boldsymbol{v}^*[j] N_{\mathrm{pt}}^{-1} \mathsf{Mamba}(\boldsymbol{Z}; \boldsymbol{\gamma}^*) + \boldsymbol{a}^*[j]\right) \;\middle|\; \|\boldsymbol{u}\| \le U \right\}.$$

Its Rademacher complexity for sample size T_2 is

$$\mathrm{Rad}_{T_2}(\mathcal{F}_U) = \mathbb{E}_{\substack{(\boldsymbol{Z}^t, y^t)\sim\mathcal{D}(N_{\mathrm{pt}}) \\ \boldsymbol{\epsilon}\sim\mathrm{Unif}(\{\pm1\}^{T_2})}} \left[\sup_{\tilde{f}\in\mathcal{F}_U} \frac{1}{T_2} \sum_{t=1}^{T_2} \boldsymbol{\epsilon}[t] \tilde{f}(\boldsymbol{Z}^t, y^t) \right].$$

Lemma D.1. *It holds that*

$$\mathrm{Rad}_{T_2}(\mathcal{F}_U) = \tilde{\mathcal{O}}(U m^{1/2} T_2^{-1/2}).$$

Proof of Lemma D.1. We condition on the Stage-I data and Stage-II initialization satisfying the preceding bounds. These parameters are independent of the Stage-II training prompts. By the Cauchy–Schwarz and Jensen inequalities and the independence of the Rademacher signs,

$$\begin{aligned}
\mathrm{Rad}_{T_2}(\mathcal{F}_U) &\le \frac{U}{T_2} \left(\sum_{t=1}^{T_2} \sum_{j=1}^{m} \mathbb{E}\left[\mathrm{ReLU}\left(\boldsymbol{v}^*[j] N_{\mathrm{pt}}^{-1} \mathsf{Mamba}(\boldsymbol{Z}^t; \boldsymbol{\gamma}^*) + \boldsymbol{a}^*[j]\right)^2 \right] \right)^{1/2} \\
&\le U \left(\frac{2m}{T_2} \left(1 + \mathbb{E}_{\boldsymbol{Z}}\left[N_{\mathrm{pt}}^{-2} \mathsf{Mamba}(\boldsymbol{Z}; \boldsymbol{\gamma}^*)^2 \right]\right) \right)^{1/2},
\end{aligned}$$

since $|\boldsymbol{v}^*[j]| = 1$ and $|\boldsymbol{a}^*[j]| \le 1$.

We bound the second moment for every context length satisfying Proposition B.5. By Lemma A.5, $a_i = \tilde{\mathcal{O}}(d^{-C_b})$ for $i = 0, 1, 2$, while $b_i = \tilde{\mathcal{O}}(1)$. Thus, (5), our choice of η, and the sample-size bounds give $\|\boldsymbol{\gamma}^*\|_\infty = \tilde{\mathcal{O}}(d^{C_b})$. Since the gating factors lie in $[0, 1]$,

$$|N^{-1}\mathsf{Mamba}(\boldsymbol{Z}; \boldsymbol{\gamma}^*)| \le \|\boldsymbol{\gamma}^*\|_\infty \|\phi(\boldsymbol{x})\| \frac{1}{N} \sum_{j=1}^{N} |y_j| \|\phi(\boldsymbol{x}_j)\|.$$

Each of the $O(d^2)$ embedding coordinates is a fixed-degree Hermite polynomial. Consequently,

$$\mathbb{E}[\|\phi(\boldsymbol{x})\|^4] = O(d^4), \qquad \mathbb{E}[\|\phi(\boldsymbol{x})\|^8] = O(d^8), \qquad \sup_{\boldsymbol{\beta}\in S_r} \mathbb{E}[|y_1|^8 \mid \boldsymbol{\beta}] = O(1).$$

Conditional on $\boldsymbol{\beta}$, the query is independent of the context. Applying Jensen's inequality to the context average and then the Cauchy–Schwarz inequality gives

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Z}\sim\mathcal{D}(N)}[N^{-4}\mathsf{Mamba}(\boldsymbol{Z}; \boldsymbol{\gamma}^*)^4] &\le \|\boldsymbol{\gamma}^*\|_\infty^4 \mathbb{E}[\|\phi(\boldsymbol{x})\|^4] \sup_{\boldsymbol{\beta}\in S_r} \mathbb{E}[|y_1|^4 \|\phi(\boldsymbol{x}_1)\|^4 \mid \boldsymbol{\beta}] \\
&\le \|\boldsymbol{\gamma}^*\|_\infty^4 O(d^4) \sup_{\boldsymbol{\beta}\in S_r} \left(\mathbb{E}[|y_1|^8 \mid \boldsymbol{\beta}]\right)^{1/2} \left(\mathbb{E}[\|\phi(\boldsymbol{x}_1)\|^8]\right)^{1/2} \\
&= \tilde{\mathcal{O}}(d^{4C_b+8}). \qquad (13)
\end{aligned}$$

Let $\mathcal{E}$ be the event that the conclusion of Proposition B.5 holds for this context length. On $\mathcal{E}$, $|N^{-1}\mathsf{Mamba}(\boldsymbol{Z};\boldsymbol{\gamma}^*)| = \tilde{\mathcal{O}}(1)$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z}\sim\mathcal{D}(N)}[N^{-2}\mathsf{Mamba}(\boldsymbol{Z};\boldsymbol{\gamma}^*)^2] \\
&\le \tilde{\mathcal{O}}(1) + \left(\mathbb{E}_{\boldsymbol{Z}\sim\mathcal{D}(N)}[N^{-4}\mathsf{Mamba}(\boldsymbol{Z};\boldsymbol{\gamma}^*)^4]\right)^{1/2}\mathbb{P}[\mathcal{E}^c]^{1/2} \\
&\le \tilde{\mathcal{O}}(1) + \tilde{\mathcal{O}}(d^{2C_b+4-C_{\mathrm{whp}}/2}) = \tilde{\mathcal{O}}(1) \qquad (14)
\end{aligned}$$

for $C_{\mathrm{whp}} > 4C_b + 8$. This bound is uniform over the admissible context lengths. Taking $N = N_{\mathrm{pt}}$ in the Rademacher bound proves the lemma.

□

Next, we obtain the following result on test error.

Lemma D.2. *With probability at least* 0.995*, it holds that* $\mathcal{R}_{N_{\mathrm{pt}}}(\boldsymbol{\gamma}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, \boldsymbol{a}^*) - \tau = o(1)$.

Proof of Lemma D.2. By Lemma C.3, there is a deterministic $U = \tilde{\mathcal{O}}(r^{3\mathrm{ge}(g_*)/2}m^{-1/2})$ such that, with probability at least 0.999, the trained predictor belongs to $\mathcal{F}_U$ and its empirical absolute loss is $\tau + o(1)$.

The map $z \mapsto |y-z| - |y|$ is 1-Lipschitz and vanishes at zero. Applying symmetrization (Proposition 4.2 in Bach (2024)) and the Ledoux–Talagrand contraction inequality to these centered losses gives

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\tilde{f}\in\mathcal{F}_U}\left|\mathbb{E}_{(\boldsymbol{Z},y)\sim\mathcal{D}(N_{\mathrm{pt}})}[|y-\tilde{f}(\boldsymbol{Z},y)| - |y|] - \frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}(|y^t - \tilde{f}(\boldsymbol{Z}^t, y^t)| - |y^t|)\right|\right] \\
&\le 4\,\mathrm{Rad}_{T_2}(\mathcal{F}_U).
\end{aligned}$$

Here, the outer expectation is over the independent Stage-II training prompts. The remaining label-only term satisfies

$$\mathbb{E}\left[\left|\mathbb{E}[|y|] - \frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}|y^t|\right|\right] \le \sqrt{\frac{\mathrm{Var}(|y|)}{T_2}} \le \sqrt{\frac{\mathbb{E}[y^2]}{T_2}} = O(T_2^{-1/2}).$$

Combining these bounds with Lemma D.1,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\tilde{f}\in\mathcal{F}_U}\left|\mathbb{E}_{(\boldsymbol{Z},y)\sim\mathcal{D}(N_{\mathrm{pt}})}[|y-\tilde{f}(\boldsymbol{Z},y)|] - \frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}|y^t - \tilde{f}(\boldsymbol{Z}^t, y^t)|\right|\right] \\
&\le 4\,\mathrm{Rad}_{T_2}(\mathcal{F}_U) + O(T_2^{-1/2}) = \tilde{\mathcal{O}}(r^{3\mathrm{ge}(g_*)/2}T_2^{-1/2}) = o(1).
\end{aligned}$$

By Markov's inequality, the displayed supremum is $o(1)$ with probability at least 0.996. Together with the event of Lemma C.3, this yields

$$\mathcal{R}_{N_{\mathrm{pt}}}(\boldsymbol{\gamma}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, \boldsymbol{a}^*) \le \frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}|y^t - f(\boldsymbol{Z}^t;\boldsymbol{\gamma}^*, \boldsymbol{u}^*, \boldsymbol{v}^*, \boldsymbol{a}^*)| + o(1) = \tau + o(1)$$

with probability at least 0.995. The independent query noise is uniform on $\{\pm\tau\}$, and $(|a-\tau| + |a+\tau|)/2 \ge \tau$ for every $a \in \mathbb{R}$. Thus the risk is at least τ, proving the desired equality.

□

D.2 Test Error for Prompts with General Length

For the last step, we extend the result to a general test-time context length $N_{\mathrm{test}} = \tilde{\Omega}(r^{3\mathrm{ge}(g_*)})$.

Proof of Theorem 3.3. We condition on the trained parameters satisfying the preceding bounds. We compare test prompts with the same feature vector $\boldsymbol{\beta}$ and query-label pair $(\boldsymbol{x}, y)$. Their contexts consist of the first N_{pt} or N_{test} examples of a common context of length $\max\{N_{\mathrm{pt}}, N_{\mathrm{test}}\}$, and we denote the corresponding embeddings by $\boldsymbol{Z}_{N_{\mathrm{pt}}}$ and $\boldsymbol{Z}_{N_{\mathrm{test}}}$.

In (6), all terms except the context error depend only on $\boldsymbol{\beta}, \boldsymbol{x}$ and the trained parameters. They therefore cancel when we subtract the two normalized Mamba outputs. By (7), with high probability,

$$\begin{aligned}&\left|N_{\mathrm{pt}}^{-1}\mathsf{Mamba}(\boldsymbol{Z}_{N_{\mathrm{pt}}};\boldsymbol{\gamma}^*) - N_{\mathrm{test}}^{-1}\mathsf{Mamba}(\boldsymbol{Z}_{N_{\mathrm{test}}};\boldsymbol{\gamma}^*)\right| \\ &\le \tilde{\mathcal{O}}\left(N_{\mathrm{pt}}^{-1/2} + N_{\mathrm{test}}^{-1/2} + d^{-C_b+6} + d^{-C_b+C^*} + d^4 N_{\mathrm{pt}}^{-1/2} + d^3 T_1^{-1/2}\right).\end{aligned}$$

With sufficiently large logarithmic factors in the sample sizes and sufficiently large C_b, multiplying the right-hand side by $\|\boldsymbol{u}^*\|\sqrt{m} = \tilde{\mathcal{O}}(r^{3\mathrm{ge}(g_*)/2})$ gives $o(1)$. By the Lipschitz continuity of ReLU and $|\boldsymbol{v}^*[j]| = 1$, there is an event $\mathcal{E}$ with $\mathbb{P}[\mathcal{E}^c] \le d^{-C_{\mathrm{whp}}}$ on which

$$\begin{aligned}&|f(\boldsymbol{Z}_{N_{\mathrm{pt}}};\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*) - f(\boldsymbol{Z}_{N_{\mathrm{test}}};\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*)| \\ &\le \|\boldsymbol{u}^*\|\sqrt{m}\left|N_{\mathrm{pt}}^{-1}\mathsf{Mamba}(\boldsymbol{Z}_{N_{\mathrm{pt}}};\boldsymbol{\gamma}^*) - N_{\mathrm{test}}^{-1}\mathsf{Mamba}(\boldsymbol{Z}_{N_{\mathrm{test}}};\boldsymbol{\gamma}^*)\right| = o(1).\end{aligned}$$

To pass to the expected error, we also control the complementary event. By the same Lipschitz bound and (14),

$$\begin{aligned}&\mathbb{E}\left[|f(\boldsymbol{Z}_{N_{\mathrm{pt}}};\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*) - f(\boldsymbol{Z}_{N_{\mathrm{test}}};\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*)|^2\right] \\ &\le 2m\|\boldsymbol{u}^*\|^2\left(\mathbb{E}[N_{\mathrm{pt}}^{-2}\mathsf{Mamba}(\boldsymbol{Z}_{N_{\mathrm{pt}}};\boldsymbol{\gamma}^*)^2] + \mathbb{E}[N_{\mathrm{test}}^{-2}\mathsf{Mamba}(\boldsymbol{Z}_{N_{\mathrm{test}}};\boldsymbol{\gamma}^*)^2]\right) \\ &= \tilde{\mathcal{O}}(r^{3\mathrm{ge}(g_*)}).\end{aligned}$$

Splitting the expectation over $\mathcal{E}$ and $\mathcal{E}^c$ and applying the Cauchy–Schwarz inequality therefore yields

$$\begin{aligned}&\mathbb{E}\left[|f(\boldsymbol{Z}_{N_{\mathrm{pt}}};\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*) - f(\boldsymbol{Z}_{N_{\mathrm{test}}};\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*)|\right] \\ &\le o(1) + \tilde{\mathcal{O}}(r^{3\mathrm{ge}(g_*)/2})\mathbb{P}[\mathcal{E}^c]^{1/2} = o(1),\end{aligned}$$

where the last equality uses $r \le d$ and $C_{\mathrm{whp}} > 3\mathrm{ge}(g_*)$. Finally, by the triangle inequality,

$$\begin{aligned}&|\mathcal{R}_{N_{\mathrm{test}}}(\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*) - \mathcal{R}_{N_{\mathrm{pt}}}(\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*)| \\ &\le \mathbb{E}\left[|f(\boldsymbol{Z}_{N_{\mathrm{pt}}};\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*) - f(\boldsymbol{Z}_{N_{\mathrm{test}}};\boldsymbol{\gamma}^*,\boldsymbol{u}^*,\boldsymbol{v}^*,\boldsymbol{a}^*)|\right] = o(1).\end{aligned}$$

Combining this with Lemma D.2 and taking a union bound over the preceding Stage-I and initialization events proves the conclusion with probability at least 0.99.

□

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions of the paper, including limitations of prior works, a brief description of the problem setting, and a summary of our theoretical findings.

   Guidelines:

   - The answer [N/A] 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 [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations of our theoretical results in Section 6.

   Guidelines:

   - The answer [N/A] 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: We introduce formal theoretical settings in Section 2 and provide full set of assumptions in the statement of Theorem 3.3. In addition, we provide complete proof of our theoretical findings in the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide details for our main theoretical results in Section 5.

   Guidelines:

   - The answer [N/A] 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: We do not provide code in supplemental material. However, our experiments on synthetic data can be easily reproduced since we opened all details.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all details for our main experimental results in Section 5.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We represent the experimental results with the mean and standard deviation in Figure 1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide computer resource information in Section 5.

   Guidelines:

   - The answer [N/A] 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: This paper does not violate the NeurIPS Code of Ethics in terms of data privacy, potential for misuse, or other ethical issues.

   Guidelines:

   - The answer [N/A] 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: [N/A]

    Justification: This work is primarily theoretical.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work is primarily theoretical.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We clarify the reference of our code base in Section 5.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: Our work does not release new assets.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This research does not involve any human subjects or crowdsourcing.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This research does not involve any human subjects. Thus, IRB approval is notapplicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: We used LLMs only for writing and editing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.