Trained Mamba Emulates Online Gradient Descent in In-Context Linear Regression

Jiarui Jiang*[1], Wei Huang*[2], Miao Zhang†[1], Taiji Suzuki[3 2], Liqiang Nie[1]
[1]Harbin Institute of Technology, Shenzhen
[2]RIKEN AIP
[3]University of Tokyo
jiaruij@outlook.com, wei.huang.vr@riken.jp, zhangmiao@hit.edu.cn
taiji@mist.i.u-tokyo.ac.jp, nieliqiang@gmail.com

Abstract

State-space models (SSMs), particularly Mamba, emerge as an efficient Transformer alternative with linear complexity for long-sequence modeling. Recent empirical works demonstrate Mamba's in-context learning (ICL) capabilities competitive with Transformers, a critical capacity for large foundation models. However, theoretical understanding of Mamba's ICL remains limited, restricting deeper insights into its underlying mechanisms. Even fundamental tasks such as linear regression ICL, widely studied as a standard theoretical benchmark for Transformers, have not been thoroughly analyzed in the context of Mamba. To address this gap, we study the training dynamics of Mamba on the linear regression ICL task. By developing novel techniques tackling non-convex optimization with gradient descent related to Mamba's structure, we establish an exponential convergence rate to ICL solution, and derive a loss bound that is comparable to Transformer's. Importantly, our results reveal that Mamba can perform a variant of *online gradient descent* to learn the latent function in context. This mechanism is different from that of Transformer, which is typically understood to achieve ICL through gradient descent emulation. The theoretical results are verified by experimental simulation.

1 Introduction

State-space models (SSMs), notably Mamba (Gu and Dao, 2024), have recently emerged as compelling alternatives to Transformer-based architectures (Vaswani et al., 2017). Mamba integrates gating, convolutions, and state-space modeling with selection mechanisms, enabling linear-time complexity. This effectively addresses the quadratic computational costs typically associated with self-attention mechanisms in Transformers. Consequently, Mamba demonstrates superior efficiency in processing long sequences while maintaining or even surpassing Transformer performance across diverse benchmarks (Gu and Dao, 2024; Dao and Gu, 2024; Patro and Agneeswaran, 2024; Liu et al., 2024; Ahamed and Cheng, 2024; Li et al., 2024a,b).

In-context learning (ICL) (Brown et al., 2020) is a powerful paradigm that enables models to generalize to unseen tasks by dynamically leveraging contextual examples (such as input-output pairs) without task-specific fine-tuning. This capability has become a defining characteristic of large foundation models, significantly enhancing their flexibility and adaptability. While extensive research has provided substantial insights into Transformer-based ICL mechanisms (Garg et al., 2022; Gatmiry et al., 2024; Sander et al., 2024; Zheng et al., 2024; Zhang et al., 2025), the principles underlying

*Equal contribution
†Corresponding author

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

Mamba's ability to perform in-context learning remain largely unexplored, highlighting a compelling research gap.

Recent empirical studies have examined Mamba's (ICL) capabilities, showing it matches Transformers on many standard ICL tasks, while surpassing them in specialized scenarios like sparse parity (Park et al., 2024; Grazzi et al., 2024). Bondaschi et al. (2025) theoretically analyzed its representational capacity for in-context learning of Markov chains, and Li et al. (2025a) investigated binary classification tasks with outliers. (Yang et al., 2024, 2025; Behrouz et al., 2025b,a) leverage the connection between SSMs and online learning to design new architectures. However, even the linear regression model, a canonical setting widely used for theoretical analysis of Transformer-based ICL mechanisms, remains theoretically underexplored in the context of Mamba. To fill this gap, we analyze Mamba's training dynamics on in-context linear regression tasks. More precisely, following the previous ICL analysis in Transformers (Garg et al., 2022; Zhang et al., 2024; Ahn et al., 2023), this paper focuses on a data generative model with N input-output pairs ($\{\boldsymbol{x}_i, y_i\}_{i=1}^N$) and a query input ($\boldsymbol{x}_q$) satisfying $y = f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, where $\boldsymbol{x}$ denotes the input and y denotes the output, and $\boldsymbol{w}$ is randomly sampled from Gaussian distribution, termed the *context*. In this work, we develop a rigorous theoretical framework to analyze how randomly initialized Mamba models, when trained through gradient descent, evolve to implement in-context learning. We demonstrate that the trained Mamba architecture dynamically leverages the input context to perform implicit estimation of the vector $\boldsymbol{w}$. This estimation is achieved through hidden state updates that mimic online gradient descent steps, finally implementing prediction for $y_q = f(\boldsymbol{x}_q) = \boldsymbol{w}^\top \boldsymbol{x}_q$. We also provide a loss bound that is comparable to Transformers'. Our contributions are summarized as follows:

- We construct a Mamba architecture (S6: S4 with selection) capable of ICL, establishing its exponential convergence rate to ICL solution, and further derive the loss bound after convergence. The loss matches that of Transformers.
- Technically, we develop novel techniques to address optimization challenges induced by random initialization and gradient descent, rigorously characterizing Mamba's training dynamics when trained from scratch.
- We reveal how trained Mamba achieves in-context linear regression by progressively aligning its hidden states with the *context* through sequential token processing. This finding provides a new perspective for understanding Mamba's ICL mechanism, distinct from Transformer-based approaches. All the above results are verified by experiments.

2 Related Work

In-Context Learning The seminal work of Brown et al. (2020) demonstrated the in-context learning capability in Transformers, showing their ability to infer functional mappings from input-output exemplars without weight updates. Garg et al. (2022) initiated the investigation of ICL from the perspective of learning particular function classes. Following these, a line of research analyze this phenomenon through the lens of algorithm imitation: Transformers can be trained to implement various learning algorithms that can mimic the latent functions in context, including: a single step of gradient descent (Von Oswald et al., 2023; Akyürek et al., 2023), statistical algorithms (Bai et al., 2023), reinforcement learning algorithm (Lin et al., 2024), multi-step gradient descent (Gatmiry et al., 2024), mesa-optimization (Zheng et al., 2024), Newton's method (Giannou et al., 2025), weighted preconditioned gradient descent (Li et al., 2025b), in context classification (Bu et al., 2024; Shen et al., 2024; Bu et al., 2025) among others.

Recent work extends ICL analysis beyond Transformers: (Lee et al., 2024; Park et al., 2024) empirically compared popular architectures (e.g., RNNs, CNNs, SSMs, Transformers) on synthetic ICL tasks, identifying capability variations across model types and task demands. Tong and Pehlevan (2024) demonstrate that MLPs can learn in-context a series of classical tasks such as regression and classification with less computation than Transformers. Sushma et al. (2024) show that state space models augmented with *local self-attention* can learn linear regression in-context. Unlike existing research on ICL, this work focuses on the ICL mechanism of Mamba (specifically S4 with selection) and its training dynamics.

Theoretical Understanding of SSMs As Gu et al. (2022) introduce structured state spaces models in modeling long sequence and further be extended to Mamba (Gu and Dao, 2024), which gained

significant attention as alternatives to Transformers, extensive research has sought to theoretically understand the mechanisms and capabilities of state-spaces models (SSMs). Dao and Gu (2024) propose the framework of state space duality, which establishes a connection between SSMs and attention variants through the lens of structured matrices. Vankadara et al. (2024) provide a scaling analysis of signal propagation in SSMs through the lens of feature learning. Cirone et al. (2024) draw the link of SSMs to linear CDEs (controlled differential equations) and use tools from rough path theory to study their expressivity. Chen et al. (2025) establish the computational limits of SSMs and Mamba via circuit complexity analysis, questioning the prevailing belief that Mamba possesses superior computational expressivity compared to Transformers. Nishikawa and Suzuki (2025) demonstrate that state space models integrated with nonlinear layers achieve dynamic token selection capabilities comparable to Transformers. Different from the above, we provide theoretical understanding of Mamba from the perspective of ICL.

3 Problem Setup

In this section, we outline the ICL data model, the Mamba model, the prediction strategy, and the gradient descent training algorithm.

Data Model. We consider an in-context linear regression task where each prompt corresponds to a new function $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ with weights $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $d > 1$. For each task, we generate N i.i.d. input-output pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ and a query $\boldsymbol{x}_q$, where all inputs $\boldsymbol{x}_i, \boldsymbol{x}_q \sim \mathcal{N}(0, \boldsymbol{I}_d)$ are independent Gaussian vectors, and the outputs satisfy $y_i = f(\boldsymbol{x}_i)$. The goal is to predict $y_q = f(\boldsymbol{x}_q)$ for the query.

To enable sequential processing of prompts in the Mamba model, we implement an embedding strategy where:

1. The i-th context token is encoded as $\boldsymbol{e}_i = (\boldsymbol{x}_i^\top, y_i)^\top$, formed by concatenating input $\boldsymbol{x}_i$ with its corresponding label y_i.
2. The query token is represented as $\boldsymbol{e}_q = (\boldsymbol{x}_q^\top, 0)^\top$, masking the unknown target value with a zero placeholder.

In many theoretical analyses of Transformer-based in-context learning, token embeddings are conventionally concatenated into a single matrix to enable parallel computation of global attention (Zhang et al., 2024; Ahn et al., 2023; Huang et al., 2023; Mahankali et al., 2024; Wu et al., 2024). In contrast, since Mamba operates as a sequential model, we feed the embeddings of context tokens one by one, and finally the query token ($\boldsymbol{e}_1 \to \boldsymbol{e}_2 \to \cdots \to \boldsymbol{e}_N \to \boldsymbol{e}_q$).

Mamba Model. We consider a S6 layer of Mamba $\boldsymbol{o}_{1:L} = \textbf{Mamba}(\boldsymbol{\theta}; \boldsymbol{u}_{1:L})$ with selection, discretization, and linear recurrence components, where $\boldsymbol{u}_l, \boldsymbol{o}_l \in \mathbb{R}^{d_e}$. It can be described as follows:

$$\boldsymbol{h}_l^{(i)} = \overline{\boldsymbol{A}}_l \boldsymbol{h}_{l-1}^{(i)} + \overline{\boldsymbol{B}}_l u_l^{(i)} \in \mathbb{R}^{d_h \times 1}, \quad \text{(1a)} \qquad \overline{\boldsymbol{A}}_l = \exp(\Delta_l \boldsymbol{A}) \in \mathbb{R}^{d_h \times d_h}, \tag{2a}$$

$$o_l^{(i)} = \boldsymbol{C}_l^\top \boldsymbol{h}_l^{(i)}, \quad \boldsymbol{C}_l \in \mathbb{R}^{d_h \times 1}, \qquad \text{(1b)} \qquad \overline{\boldsymbol{B}}_l = (\Delta_l \boldsymbol{A})^{-1}(\exp(\Delta_l \boldsymbol{A}) - \boldsymbol{I})\Delta_l \boldsymbol{B}_l \in \mathbb{R}^{d_h \times 1} \tag{2b}$$

for $i \in [d_e]$. Here, the superscript (i) denotes the i-th independent processing channel, where each channel operates on a unique feature dimension of the input $\boldsymbol{u}_l$ and output $\boldsymbol{o}_l$ vectors (i.e., $u_l^{(i)}$ and $o_l^{(i)}$ correspond to the i-th elements of $\boldsymbol{u}_l$ and $\boldsymbol{o}_l$, respectively). The hidden state $\boldsymbol{h}_l^{(i)}$ is initialized as $\boldsymbol{h}_0^{(i)} = \boldsymbol{0}$ and evolves according to $\overline{\boldsymbol{A}}_l \in \mathbb{R}^{d_h \times d_h}, \overline{\boldsymbol{B}}_l \in \mathbb{R}^{d_h \times 1}$ and the input $u_l^{(i)}$. $\boldsymbol{C}_l \in \mathbb{R}^{d_h \times 1}$ maps the hidden state $\boldsymbol{h}_l^{(i)}$ to the output $o_l^{(i)}$. As shown in (2), $\overline{\boldsymbol{A}}_l$ and $\overline{\boldsymbol{B}}_l$ are computed using the zero-order hold (ZOH) discretization method applied to $\boldsymbol{A} \in \mathbb{R}^{d_h \times d_h}, \boldsymbol{B}_l \in \mathbb{R}^{d_h \times 1}$ and the timestep $\Delta_l \in \mathbb{R}$. Next, we describe the selection mechanism.

$$\boldsymbol{B}_l = \boldsymbol{W}_B \boldsymbol{u}_l + \boldsymbol{b}_B, \quad (3) \qquad \boldsymbol{C}_l = \boldsymbol{W}_C \boldsymbol{u}_l + \boldsymbol{b}_C, \quad (4) \qquad \Delta_l = \text{softplus}(\boldsymbol{w}_\Delta^\top \boldsymbol{u}_l + b_\Delta), \quad (5)$$

Here, $\text{softplus(x)} = \log(1 + \exp(x))$. $\boldsymbol{W}_B, \boldsymbol{W}_C \in \mathbb{R}^{d_h \times d_e}$, $\boldsymbol{b}_B, \boldsymbol{b}_C \in \mathbb{R}^{d_h \times 1}$, $\boldsymbol{w}_\Delta \in \mathbb{R}^{d_e \times 1}$, $b_\Delta \in \mathbb{R}$, along with $\boldsymbol{A} \in \mathbb{R}^{d_h \times d_h}$ are the parameters of the Mamba model. We use $\boldsymbol{\theta}$ to denote the collection of all the parameters.

Unlike previous work (Sushma et al., 2024) that introduce *local self-attention* component to augment SSMs, which may inherit the Transformer's ICL ability, our model adheres to Mamba's original selective state-space framework (Gu and Dao, 2024). This alignment ensures us to mechanistically analyze how Mamba's architecture enables in-context learning (ICL).

Linear Regression Prediction. In this work, we set $d_e = d + 1$, enabling the Mamba model to process the embeddings $\boldsymbol{e}_{1:N}, \boldsymbol{e}_q$. Given the prompt $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N, \boldsymbol{e}_q)$, the Mamba model will output a sequence $\boldsymbol{o}_{1:N+1} = \textbf{Mamba}(\boldsymbol{\theta}; \boldsymbol{e}_1, \ldots, \boldsymbol{e}_N, \boldsymbol{e}_q)$. The prediction for the linear regression target $y_q = \boldsymbol{w}^\top \boldsymbol{x}_q$ is extracted from the terminal position of the output matrix (corresponding to the zero placeholder in the query token $\boldsymbol{e}_q = (\boldsymbol{x}_q^\top, 0)^\top$). Concretely, $\hat{y}_q = \boldsymbol{o}_{N+1}^{(d+1)}$.

Training Algorithm. To train a Mamba model over the in-context linear regression task, we consider minimizing the following population loss:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}_{1:N}, \boldsymbol{x}_q, \boldsymbol{w}} \Big[\frac{1}{2} (\hat{y}_q - y_q)^2 \Big]. \tag{6}$$

Given a Mamba model, we use gradient descent to minimize population loss $\mathcal{L}(\boldsymbol{\theta})$, and the update of trainable parameters $\boldsymbol{\theta}' = \{\boldsymbol{W}_B, \boldsymbol{W}_C, \boldsymbol{b}_B, \boldsymbol{b}_C\}$ can be written as follows:

$$\boldsymbol{\theta}'(t+1) = \boldsymbol{\theta}'(t) - \eta \nabla_{\boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}(t)). \tag{7}$$

4 Main Results

This section presents our main theoretical results that characterize the convergence state of Mamba and its final loss. We also compare the results with other models.

Assumption 4.1 *(1) Matrix $\boldsymbol{A} = -\boldsymbol{I}_{d_h}$. (2) The vector $\boldsymbol{w}_\Delta$ is fixed as zero $\mathbf{0}$, and b_Δ is fixed as $\ln(\exp((\ln 2)/N) - 1)$. (3) Matrices $\boldsymbol{W}_B$, $\boldsymbol{W}_C$ are initialized with entries drawn i.i.d. from the standard Gaussian distribution $\mathcal{N}(0, 1)$. (4) The hidden state dimension satisfies: $d_h = \widetilde{\Omega}(d^2)$. (5) The learning rate satisfies: $\eta = O(d^{-2} d_h^{-1})$. (6) Bias vectors $\boldsymbol{b}_B$, $\boldsymbol{b}_C$ are initialized as zero $\mathbf{0}$. (7) Token length $N = \Omega(d)$.*

(1) The negative-definite matrix $\boldsymbol{A} = -\boldsymbol{I}_{d_h}$ guarantees the stable convergence of hidden states $\boldsymbol{h}_l^{(i)}$. (2) Given the zero-mean and symmetric distribution of embeddings, $\boldsymbol{w}_\Delta$ can naturally converge to $\mathbf{0}$ during gradient descent, and we fix it as $\mathbf{0}$ for simplicity. We further fix b_Δ to an appropriate constant to maintain a suitable timestep Δ_l, enabling us to concentrate our theoretical analysis on $\boldsymbol{W}_B$, $\boldsymbol{W}_C$, $\boldsymbol{b}_B$, and $\boldsymbol{b}_C$. In prior works on Transformer-based in-context learning, merging key-query weights (e.g., $\boldsymbol{W} := \boldsymbol{W}_Q \boldsymbol{W}_K$) and specific initializations (e.g., $\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{I}$) are often adopted to simplify optimization analysis (Zhang et al., 2024; Ahn et al., 2023; Huang et al., 2023; Mahankali et al., 2024; Wu et al., 2024). (3, 4, 5) In contrast, our Gaussian initialization of $\boldsymbol{W}_B$ and $\boldsymbol{W}_C$ demonstrates more practicality, which requires a sufficiently large hidden state dimension d_h and a sufficiently small learning rate η to ensure favorable loss landscape properties. Assumption (6) is intended to simplify the analysis. (7) Token length should be larger enough than the dimension of $\boldsymbol{w}$ to capture sufficient contextual information.

Theorem 4.1 *Under Assumption 4.1, if the Mamba is trained with gradient descent, and given a new prompt $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N, \boldsymbol{e}_q)$, then with probability at least $1 - \delta$ for some $\delta \in (0, 1)$, the trainable parameters $\boldsymbol{\theta}'(t) = \{\boldsymbol{W}_B(t), \boldsymbol{W}_C(t), \boldsymbol{b}_B(t), \boldsymbol{b}_C(t)\}$ converge as $t \to \infty$ to parameters that satisfies:*

(a) Projected hidden state: $(\boldsymbol{W}_C^\top)_{[1:d,:]}(t) \boldsymbol{h}_l^{(d+1)} = \alpha (\boldsymbol{W}_C^\top(t))_{[1:d,:]} \boldsymbol{h}_{l-1}^{(d+1)} + (1-\alpha)\beta y_l \boldsymbol{x}_l$,

(b) Prediction for target: $\hat{y}_q = \boldsymbol{x}_q^\top \sum_{i=0}^{N-1} (1-\alpha)\alpha^{i+1} \beta y_{N-i} \boldsymbol{x}_{N-i}$,

(c) Population loss: $\mathcal{L}(\boldsymbol{\theta}(t)) \leq \frac{3d(d+1)}{2N}$,

where $\alpha = \exp((-\ln 2)/N)$, $\beta = \frac{2(1+\alpha)}{\alpha\big(3(1-\alpha)d + 4 - 2\alpha\big)}$.

Theorem 4.1 characterizes the in-context learning (ICL) mechanism of Mamba and establishes an upper bound on its population loss. Specifically, (Thm 4.1 (a)) shows how the hidden state is updated according the given prompt $\boldsymbol{e}_l = (\boldsymbol{x}_l^\top, y_l)^\top$. (Thm 4.1 (b)) presents the final prediction given prompt $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N, \boldsymbol{e}_q)$. (Thm 4.1 (c)) provides the upper bound for the population loss, which is comparable to that of the Transformer (Zhang et al., 2024). Next, we'll discuss it in more detail.

Update of Hidden State. If we define $\tilde{\boldsymbol{h}}_l := (\boldsymbol{W}_C^\top)_{[1:d,:]} \boldsymbol{h}_l^{(d+1)}$, then (Thm 4.1 (a)) can be rewritten as follows:

$$\tilde{\boldsymbol{h}}_l = \alpha \tilde{\boldsymbol{h}}_{l-1} + (1-\alpha)\beta y_l \boldsymbol{x}_l = \tilde{\boldsymbol{h}}_{l-1} + (1-\alpha)(\beta y_l \boldsymbol{x}_l - \tilde{\boldsymbol{h}}_{l-1}). \tag{8}$$

We observe its intrinsic connection to **online gradient descent**, which updates the model parameters ($\tilde{\boldsymbol{h}}_l$) with only one currently arriving sample ($\boldsymbol{e}_l = (\boldsymbol{x}_l^\top, y_l)^\top$) at each step. Specifically, the system gradually updates $\tilde{\boldsymbol{h}}_l$ along the pseudo-gradient direction $\beta y_l \boldsymbol{x}_l$, with a fixed step size $(1-\alpha)$.

For a newly defined task $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, given that $\mathbb{E}[y_l \boldsymbol{x}_l] = \boldsymbol{w}$, the direction of $\tilde{\boldsymbol{h}}_l$ converges toward $\boldsymbol{w}$ as mamba processes multiple prompts. This demonstrates mamba's ability to internalize $f(\boldsymbol{x})$ through prompt processing, which ultimately ensures that predictions for query token $\boldsymbol{e}_q = (\boldsymbol{x}_q^\top, 0)^\top$ closely approximate $f(\boldsymbol{x}_q)$.

Previous works have shown that Transformer can mimic a single step of gradient descent to achieve in-context learning ability (Zhang et al., 2024; Mahankali et al., 2024). Concretely, a trained Transformer can be described as follows

$$\textbf{Transformer}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N, \boldsymbol{e}_q) \approx \boldsymbol{x}_q^\top \Big(\frac{1}{N} \sum_{i=1}^{N} y_i \boldsymbol{x}_i \Big) \approx \boldsymbol{x}_q^\top \boldsymbol{w}. \tag{9}$$

Our theoretical analysis reveals that Mamba and Transformer have different in-context learning mechanisms. This divergence stems from their inherent architectural biases: Transformers process contexts globally through self-attention, while Mamba enforces local sequential dependencies via recurrent state transitions. These findings provide fundamental insights into the contrasting capabilities of Transformer-based and Mamba-based models for in-context learning. As experimental work shows, transformers can learn vector-valued MQAR tasks in the context which Mamba cannot, while Mamba succeeds in sparse-parity in-context learning tasks where Transformers fail(Park et al., 2024).

Prediction Outcome. Comparing equations Thm 4.1 (b) and (9), we found both similarities and distinctions in how Transformer and Mamba implement in-context learning (ICL). Both models leverage a weighted aggregation of $y_i \boldsymbol{x}_i$, aligning with the intuition that learning $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{w}$ from context reduces to estimating the latent parameter $\boldsymbol{w}$, since $\mathbb{E}[y_i \boldsymbol{x}_i] = \boldsymbol{w}$. Notably, their token weighting strategies diverge: Transformer's global attention mechanism implicitly assigns nearly uniform weights ($\sim \frac{1}{N}$, where N is the token length) to all $y_i \boldsymbol{x}_i$, while Mamba's linear recurrence imposes position-dependent weight variations. This difference arises from Mamba's iterative state update rule, where the influence of prompt tokens $\boldsymbol{e}_i$ on the hidden state $\boldsymbol{h}_l$ depends on their sequential placement, governed by the model's linear recurrence dynamics.

The derived upper bound (Thm 4.1 (c)) establishes an $O(1/N)$ convergence rate for the loss (ignoring dimension factor), demonstrating that Mamba matches the sample complexity scaling of Transformers in linear regression ICL tasks (Zhang et al., 2024).

Compare with S4. Mamba extends the structured state space model (S4) (Gu et al., 2022) by integrating a selection mechanism, which is critical for enabling ICL. In S4 model, the matrices $\boldsymbol{A} \in \mathbb{R}^{d_h \times d_h}$, and $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{d_h \times 1}$ are static, leading to a fixed linear combination of inputs:

$$o_l^{(i)} = \sum_{j=1}^{l} \boldsymbol{C}^\top \overline{\boldsymbol{A}}^{l-j} \overline{\boldsymbol{B}}\, u_j^{(i)}, \tag{10}$$

where the coefficients $\boldsymbol{C}^\top \overline{\boldsymbol{A}}^{l-j} \overline{\boldsymbol{B}}$ are *task-agnostic*. This formulation inherently limits S4's ability to adapt to *task-specific* parameters $\boldsymbol{w}$ in ICL scenarios, as the model cannot adjust its inductive bias to match distinct $\boldsymbol{w}$ across different tasks. Therefore, the S4 model cannot truly learn in-context.

In contrast, Mamba's selection mechanism dynamically adjusts $\boldsymbol{B}_l$ and $\boldsymbol{C}_l$ (and optionally $\boldsymbol{A}_l$) based on the input tokens $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N)$. This allows the model to implicitly adapt its hidden state to

align with the latent $\boldsymbol{w}$ of each task, effectively transforming the linear combination weights into context-dependent functions $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{w}$. Such adaptability is essential for ICL, as it enables Mamba to reconstruct diverse $\boldsymbol{w}$ from input prompts without task-specific fine-tuning.

5 Proof Sketch

This section outlines the main technical ideas to prove Theorem 4.1. The complete proofs are given in the appendix.

Linear Recurrence. To start with, we show how the hidden states update when receiving token $\boldsymbol{e}_l = (\boldsymbol{x}_l^\top, y_l)^\top$. By (Eq. (5)) and Assumption 4.1(2), we have $\Delta_l = (\ln 2)/N$. Combining it with (Eq. (1)(2)) and get:

$$\boldsymbol{h}_l^{(d+1)} = \alpha \boldsymbol{h}_{l-1}^{(d+1)} + (1-\alpha) y_l \boldsymbol{B}_l, \tag{11}$$

where $\alpha := \exp(-\Delta_l) = \exp((-\ln 2)/N)$, the second equality is by discretization rule (2), the third equality is by Assumption 4.1(2) and $\exp(-\Delta_l \boldsymbol{I}) = \exp(-\Delta_l)\boldsymbol{I}$.

Prediction Output. We next derive the expression of $\hat{y}_q$. By recurring (Eq.(11)), the hidden state after receiving the first l context prompts $\boldsymbol{e}_{1:l}$ is given by $\boldsymbol{h}_l^{(d+1)} = (1-\alpha)\sum_{i=0}^{l-1} \alpha^i y_{l-i} \boldsymbol{B}_{l-i}$. Receiving all the prompt tokens $\boldsymbol{e}_{1:N}$ and the query token $\boldsymbol{e}_q = (\boldsymbol{x}_q^\top, 0)^\top$, we have:

$$\boldsymbol{h}_{N+1}^{(d+1)} = \alpha \boldsymbol{h}_N^{(d+1)} + (1-\alpha)\cdot 0 \cdot \boldsymbol{B}_N = (1-\alpha)\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i} \boldsymbol{B}_{N-i}. \tag{12}$$

Finally, the prediction output is as follows

$$\hat{y}_q = \boldsymbol{C}_{N+1}^\top \boldsymbol{h}_{N+1}^{(d+1)} = (1-\alpha)(\boldsymbol{W}_C \boldsymbol{e}_q + \boldsymbol{b}_C)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i} (\boldsymbol{W}_B \boldsymbol{e}_{N-i} + \boldsymbol{b}_B). \tag{13}$$

To handle $\boldsymbol{W}_C \boldsymbol{e}_q$ and $\boldsymbol{W}_B \boldsymbol{e}_{N-i}$, we further decompose $\boldsymbol{W}_B = [\boldsymbol{B}\, \boldsymbol{b}]$ and $\boldsymbol{W}_C = [\boldsymbol{C}\, \boldsymbol{c}]$, where $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{d_h \times d}$, $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^{d_h \times 1}$. Then we write another form of (Eq. (13)):

$$\hat{y}_q = (1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q + \boldsymbol{b}_C)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i} (\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b} + \boldsymbol{b}_B). \tag{14}$$

The loss becomes:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}\Big[\Big((1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q + \boldsymbol{b}_C)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i} (\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b} + \boldsymbol{b}_B) - \boldsymbol{w}^\top \boldsymbol{x}_q\Big)^2\Big]. \tag{15}$$

By computing the gradient of $\boldsymbol{C}$, $\boldsymbol{b}_C$, $\boldsymbol{B}$, $\boldsymbol{b}$ and $\boldsymbol{b}_B$ with respect to $\mathcal{L}(\boldsymbol{\theta}(t))$, we derive the following update rule according to Eq. (7).

Lemma 5.1 (Update Rule) *Let η be the learning rate and we use gradient descent to update the weights $\boldsymbol{W}_B, \boldsymbol{W}_C, \boldsymbol{b}_B, \boldsymbol{b}_C$, for $t \geq 0$ we have*

$$\boldsymbol{B}(t+1) = \boldsymbol{B}(t) + \eta\beta_3 \boldsymbol{C}(t) - \eta\beta_1 \boldsymbol{C}(t)\boldsymbol{C}(t)^\top \boldsymbol{B}(t),$$

$$\boldsymbol{C}(t+1) = \boldsymbol{C}(t) + \eta\beta_3 \boldsymbol{B}(t) - \eta\beta_1 \boldsymbol{B}(t)\boldsymbol{B}(t)^\top \boldsymbol{C}(t) - \eta\beta_2 \boldsymbol{b}(t)\boldsymbol{b}(t)^\top \boldsymbol{C}(t),$$

$$\boldsymbol{b}(t+1) = \boldsymbol{b}(t) - \eta\beta_2 \boldsymbol{C}(t)\boldsymbol{C}(t)^\top \boldsymbol{b}(t), \quad \boldsymbol{b}_B(t) = \boldsymbol{b}_C(t) = \boldsymbol{0},$$

where $\beta_1 = \mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(1-\alpha)^2\alpha^{i+j+2} y_{N-i} y_{N-j} \boldsymbol{x}_{N-i}\boldsymbol{x}_{N-j}^\top\Big]$, $\beta_2 = \mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(1-\alpha)^2\alpha^{i+j+2} y_{N-i}^2 y_{N-j}^2\Big]$, $\beta_3 = \mathbb{E}\Big[\sum_{i=0}^{N-1}(1-\alpha)\alpha^{i+1} y_{N-i}\boldsymbol{x}_{N-i}\boldsymbol{w}^\top\Big]$.

Technical Challenges. Unlike many prior Transformer-based ICL analyses that simplify dynamics via merged weights or special initializations, our Gaussian-initialized $\boldsymbol{W}_B$, $\boldsymbol{W}_C$ and discrete-time gradient descent introduces more complexity (cf. assumption 4.1). To solve the optimization problem described in Lemma 5.1, we have the following three questions to answer: (1) Convergence Target: Where do the parameters converge? (2) Convergence Proof: How to rigorously establish convergence? (3) Saddle Point Avoidance: How to avoid saddle points? To answer these three questions, we propose two key techniques: *Vector-coupled Dynamic*, *Negative Feedback Convergence*, and apply them with a *Fine-grained Induction*. We next describe them in detail.

5.1 Vector-coupled Dynamics

We can verify by Lemma 5.1 that $\boldsymbol{C}^\top \boldsymbol{B} = \mathrm{Diag}(a_1, \ldots, a_d)$ with $a_i \in \{0, \frac{\beta_3}{\beta_1}\}$, $\boldsymbol{C}^\top \boldsymbol{b} = \boldsymbol{0}$ are the fixed points for the parameters $\boldsymbol{W}_B$, $\boldsymbol{W}_C$.

Combining the loss function Eq. 15 and $\boldsymbol{b}_B(t) = \boldsymbol{b}_C(t) = \boldsymbol{0}$ in Lemma 5.1, the loss function can be rewritten as

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}\Big[\Big((1-\alpha)\sum_{i=0}^{N-1}\alpha^{i+1}(\boldsymbol{x}_q^\top \boldsymbol{C}^\top \boldsymbol{B} y_{N-i}\boldsymbol{x}_{N-i} + y_{N-i}^2 \boldsymbol{x}_q^\top \boldsymbol{C}^\top \boldsymbol{b}) - \boldsymbol{w}^\top \boldsymbol{x}_q\Big)^2\Big].$$

To minimize loss, the term $(1-\alpha)\sum_{i=0}^{N-1}\alpha^{i+1}(\boldsymbol{x}_q^\top \boldsymbol{C}^\top \boldsymbol{B} y_{N-i}\boldsymbol{x}_{N-i} + y_{N-i}^2 \boldsymbol{x}_q^\top \boldsymbol{C}^\top \boldsymbol{b})$ should approximate $\boldsymbol{w}^\top \boldsymbol{x}_q$. Given $\mathbb{E}[y_{N-i}\boldsymbol{x}_{N-i}] = \boldsymbol{w}$ and $\mathbb{E}[y_{N-i}^2] > 0$, we derive that $\boldsymbol{C}^\top \boldsymbol{B}$ should converge to $\frac{\beta_3}{\beta_1}\boldsymbol{I}$, while $\boldsymbol{C}^\top \boldsymbol{b}$ converges to $\boldsymbol{0}$ to minimize the loss. However, as mentioned above, $\boldsymbol{C}^\top \boldsymbol{B} = \mathrm{Diag}(a_1, \ldots, a_d)$ with partial $a_i = 0$ can also enable convergence, which is an undesirable scenario.

To analyze the convergence behavior of $\boldsymbol{C}^\top \boldsymbol{B}$ and $\boldsymbol{C}^\top \boldsymbol{b}$, we introduce the *Vector-coupled Dynamics* technique, which studies the inner product dynamics between decomposed column vectors of $\boldsymbol{B}$ and $\boldsymbol{C}$. Specifically, we decompose $\boldsymbol{B}$ and $\boldsymbol{C}$ into $\boldsymbol{B} = [\boldsymbol{b}_1 \ldots \boldsymbol{b}_d]$, $\boldsymbol{C} = [\boldsymbol{c}_1 \ldots \boldsymbol{c}_d]$. Then we have another form of Lemma 5.1 for $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{b}$ as the following lemma.

Lemma 5.2 (Vectors Update Rule) *Let η be the learning rate and we use gradient descent to update the weights $\boldsymbol{W}_B, \boldsymbol{W}_C, \boldsymbol{b}_B, \boldsymbol{b}_C$, for $i \in [d]$, $t \geq 0$ we have*

$$\boldsymbol{b}_i(t+1) = \boldsymbol{b}_i(t) + \eta\Big(\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_i(t) - \beta_1 \sum_{k\neq i}^{d} \boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\cdot \boldsymbol{c}_k(t)\Big),$$

$$\boldsymbol{c}_i(t+1) = \boldsymbol{c}_i(t) + \eta\Big(\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{b}_i(t) - \beta_1 \sum_{k\neq i}^{d} \boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\cdot \boldsymbol{b}_k(t) - \beta_2 \boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\cdot \boldsymbol{b}(t)\Big),$$

$$\boldsymbol{b}(t+1) = \boldsymbol{b}(t) - \eta\Big(\beta_2 \sum_{k=1}^{d} \boldsymbol{c}_k^\top(t)\boldsymbol{b}(t)\cdot \boldsymbol{c}_k(t)\Big).$$

With Lemma 5.2, we can further analyze the dynamics of the inner products $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)$, $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)$ and $\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)$, precisely characterizing the behavior of $\boldsymbol{C}^\top \boldsymbol{B}$ and $\boldsymbol{C}^\top \boldsymbol{b}$. This technique helps answer the question "*Where do the parameters converge?*"

5.2 Negative Feedback Convergence

As we discuss in Section 5.1, to minimize loss, the following conditions must be satisfied for all $i, j \in [d]$ with $i \neq j$: $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t) \to \frac{\beta_3}{\beta_1}$, $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t) \to 0$, $\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t) \to 0$. To establish the convergence, we introduce the *Negative Feedback Convergence* technique. This technique leverages the negative feedback terms in the dynamical equations of $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)$, $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)$, and $\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)$ to derive an exponential convergence rate. Taking $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)$ as an example, we derive the following

update rule by Lemma 5.2.

$$\begin{aligned}
&\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(t+1)\boldsymbol{b}_i(t+1)\big) = \underline{\beta_3 - \beta_1 \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)} \\
&\underbrace{-\eta\beta_1\underline{\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)}\boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t) - \eta\beta_1\underline{\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)}\boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t)}_{negative\ feedback\ term} \\
&+ \eta\beta_1^2 \sum_{k\neq i}^{d} \boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\cdot\boldsymbol{b}_k^\top(t)\boldsymbol{b}_i(t) + \eta\beta_1^2 \sum_{k\neq i}^{d} \boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\cdot\boldsymbol{c}_i^\top(t)\boldsymbol{c}_k(t) \\
&+ \eta\beta_1\beta_2 \boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{b}_i^\top(t)\boldsymbol{b}(t) - \beta_1\big(\boldsymbol{c}(t+1) - \boldsymbol{c}(t)\big)^\top\big(\boldsymbol{b}(t+1) - \boldsymbol{b}(t)\big).
\end{aligned} \tag{16}$$

The term $\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(t+1)\boldsymbol{b}_i(t+1)\big)$ decomposes into its previous state $\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)$ (marked with underline) plus the remaining terms (increment terms). The increment terms includes a *negative feedback term*, which induces a tendency to drive $\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)$ to 0 ($\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t) \to \frac{\beta_3}{\beta_1}$).

Intuitively, $\boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t)$ and $\boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t)$ are much larger than $\boldsymbol{b}_k^\top(t)\boldsymbol{b}_i(t)$, $\boldsymbol{c}_i^\top(t)\boldsymbol{c}_k(t)$ and $\boldsymbol{b}_i^\top(t)\boldsymbol{b}(t)$ at Gaussian initialization with high probability. Also, as $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t), \boldsymbol{b}_i^\top(t)\boldsymbol{b}(t) \to 0$ with $i \neq j$ and η is small enough, the effect of *negative feedback term* is the dominant term in the increment terms. Therefore, denoting $y(t) = \beta_3 - \beta_1 \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)$ and $\xi(t) = y(t+1) - y(t) -$ *negative feedback term* we can model the update rule of (Eq. 16) as follows:

$$y(t+1) = \big(1 - \eta\beta_1(\boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t) + \boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t))\big)y(t) + \xi(t).$$

Recur this formula from 0 to t, we have:

$$\begin{aligned}
y(t+1) = &\prod_{s=0}^{t}\Big(1 - \eta\beta_1\big(\boldsymbol{b}_i^\top(s)\boldsymbol{b}_i(s) + \boldsymbol{c}_i^\top(s)\boldsymbol{c}_i(s)\big)\Big)y(0) \\
&+ \sum_{s=0}^{t}\prod_{s'=s+1}^{t}\Big(1 - \eta\beta_1\big(\boldsymbol{b}_i^\top(s')\boldsymbol{b}_i(s') + \boldsymbol{c}_i^\top(s')\boldsymbol{c}_i(s')\big)\Big)\xi(s').
\end{aligned} \tag{17}$$

Denoting $\gamma = \min\{\boldsymbol{b}_i^\top(s)\boldsymbol{b}_i(s), \boldsymbol{c}_i^\top(s)\boldsymbol{c}_i(s)\}$ for $s \in [0,t]$, the first term on the RHS of (Eq. (17)) can be upper bounded by $(1 - 2\eta\beta_1\gamma)^{t+1}y(0)$. if $\xi(s')$ has an exponentially decaying upper bound (it can be proved when $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t) \to 0, \boldsymbol{c}_i^\top(t)\boldsymbol{b}(t) \to 0$ with an exponential rate), the second term on the RHS of (Eq. (17)) has an exponentially decaying upper bound. Therefore, we can establish an exponential convergence rate for $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t) \to \frac{\beta_3}{\beta_1}$. The similar method can be used on $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t) \to 0, \boldsymbol{c}_i^\top(t)\boldsymbol{b}(t) \to 0$. This technique helps answer the question "*How to rigorously establish convergence?*"

5.3 Fine-grained Induction

The exponential convergence of $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t) \to \frac{\beta_3}{\beta_1}$ under the *Negative Feedback Convergence* framework requires the following two conditions for all $i, j \in [d]$ with $i \neq j$:

(1) $\boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t)$ and $\boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t)$ dominate $\boldsymbol{b}_i^\top(t)\boldsymbol{b}_j(t)$, $\boldsymbol{c}_i^\top(t)\boldsymbol{c}_j(t)$ and $\boldsymbol{b}_i^\top(t)\boldsymbol{b}(t)$ in magnitude.

(2) $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t) \to 0, \boldsymbol{c}_i^\top(t)\boldsymbol{b}(t) \to 0$ at an exponentially decaying rate.

On the one hand, condition (1) at initialization ($t = 0$) can be established via concentration inequalities, and critically, the preservation of Condition (1) for $t > 0$ relies on the rapid decay of $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t), \boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)$, and $\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)$ (condition (2)). On the other hand, under the framework of *Negative Feedback Convergence*, $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t) \to 0$ in Condition (2) also relies on Condition (1) and the rapid decay of $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t) \to \frac{\beta_3}{\beta_1}$, $\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t) \to 0$. This implies mutual dependencies among the bounds of these *Vector-coupled* inner products.

To handle these dependencies and establish stable bounds, we introduce the technique *Fine-grained Induction*: Divide the inner products into three groups: (1) Squared norms: $\boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t), \boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t), \boldsymbol{b}^\top(t)\boldsymbol{b}(t)$. (2) Target terms: $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t), \boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t), \boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)$. (3) Cross-interactions: $\boldsymbol{b}_i^\top(t)\boldsymbol{b}_j(t), \boldsymbol{c}_i^\top(t)\boldsymbol{c}_j(t), \boldsymbol{b}_i^\top(t)\boldsymbol{b}(t)$. And then carefully give bounds for them with an induction.

Specifically, denoting $\delta(t) = \max_{s\in[0,t]}\{2\sqrt{d_h\log(4d(2d+1)/\delta)}, |\boldsymbol{b}_i^\top(s)\boldsymbol{b}_j(s)|, |\boldsymbol{c}_i^\top(s)\boldsymbol{c}_j(s)|, |\boldsymbol{b}_i^\top(s)\boldsymbol{b}(s)|\}$ and $\gamma = \frac{1}{2}d_h \le \min_{t\ge 0}\{\boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t), \boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t), \boldsymbol{b}^\top(t)\boldsymbol{b}(t)\}$, we establish the following three properties $\mathcal{A}(t)$, $\mathcal{B}(t)$, and $\mathcal{C}(t)$ simultaneously for $t \ge 0$:

$$\mathcal{A}(t): \qquad d_h/2 \le \boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t), \boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t), \boldsymbol{b}^\top(t)\boldsymbol{b}(t) \le 2d_h.$$

$$\mathcal{B}(t): \qquad |\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)| \le \delta(t)\exp(-\eta\beta_1\gamma t), \quad |\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)| \le 2\delta(t)\exp(-\eta\beta_1\gamma t),$$
$$|\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)| \le 2\delta(t)\exp(-\eta\beta_2\gamma t) + \frac{\delta(t)}{\beta_2}\exp(-\eta\beta_1\gamma t).$$

$$\mathcal{C}(t): \qquad |\boldsymbol{b}_i^\top(t)\boldsymbol{b}_j(t)|, |\boldsymbol{c}_i^\top(t)\boldsymbol{c}_j(t)|, |\boldsymbol{b}_i^\top(t)\boldsymbol{b}(t)| \le \delta(t) \le 3\sqrt{d_h\log(4d(2d+1)/\delta)}.$$

The initial conditions $\mathcal{A}(0)$, $\mathcal{B}(0)$, and $\mathcal{C}(0)$ are established with high probability by concentration inequalities. We also provide the following claims to establish $\mathcal{A}(t)$, $\mathcal{B}(t)$, and $\mathcal{C}(t)$ for $t \ge 0$:

Claim 5.1 $\mathcal{A}(0), \ldots, \mathcal{A}(T), \mathcal{B}(0), \ldots, \mathcal{B}(T), \mathcal{C}(0), \ldots, \mathcal{C}(T) \Longrightarrow \mathcal{A}(T+1)$.

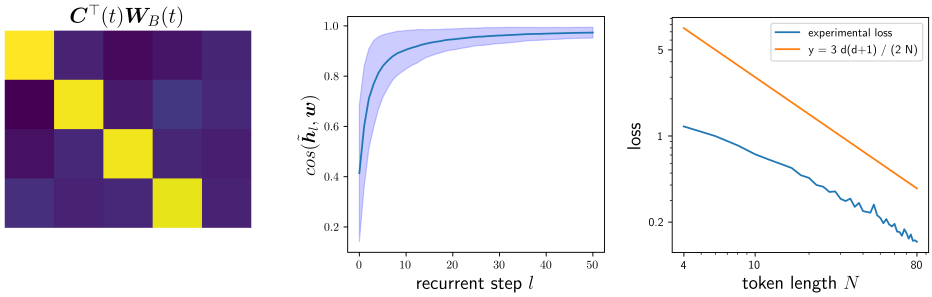
Claim 5.2 $\mathcal{A}(0), \ldots, \mathcal{A}(T), \mathcal{B}(0), \ldots, \mathcal{B}(T), \mathcal{C}(0), \ldots, \mathcal{C}(T) \Longrightarrow \mathcal{B}(T+1)$.

Claim 5.3 $\mathcal{A}(0), \ldots, \mathcal{A}(T), \mathcal{B}(0), \ldots, \mathcal{B}(T), \mathcal{C}(0), \ldots, \mathcal{C}(T) \Longrightarrow \mathcal{C}(T+1)$.

This induction answers the question "*How to avoid saddle points?*" because $\mathcal{B}(t)$ guarantees that $\boldsymbol{C}^\top\boldsymbol{B} \to \frac{\beta_3}{\beta_1}\boldsymbol{I}$ and $\boldsymbol{C}^\top\boldsymbol{b} \to \boldsymbol{0}$, preventing stagnation of partial diagonal entries of $\boldsymbol{C}^\top\boldsymbol{B}$ at zero. Theorem 4.1 can be proved by substituting $\boldsymbol{C}^\top\boldsymbol{B} = \frac{\beta_3}{\beta_1}\boldsymbol{I}, \boldsymbol{C}^\top\boldsymbol{b} = \boldsymbol{0}, \boldsymbol{b}_B = \boldsymbol{b}_C = \boldsymbol{0}$ into (Eq. (11), (14), (15))

6 Experimental Results

We present simulation results on synthetic data to verify our theoretical results. More experimental results can be found in Appendix E.



(a) Trained parameters (b) Similarity (c) Loss curve

Figure 1: (a) Post-training visualization of matrix product $\boldsymbol{C}^\top\boldsymbol{W}_B$; (b) Cosine similarity evolution between $\boldsymbol{w}$ and $\tilde{\boldsymbol{h}}_l = (\boldsymbol{W}_C^\top)_{[1:d,:]}\boldsymbol{h}_l^{(d+1)}$ across recurrent steps l (after processing prompts $\boldsymbol{e}_{1:l}$); (c) Test loss versus token sequence length N. Blue curve: experimental results; orange curve: theoretical upper bound.

Experiments Setting We follow Section 3 to generate the dateset and initialize the model. Specifically, we set dimension $d = 4$, $d_h = 80$, prompt token length $N = 50$, and train the Mamba model on 3000 sequences by gradient descent. After training, we save the model and test it on 1000 new generated sequences, tracking the cosine similarity between $\tilde{\boldsymbol{h}}_l(:= (\boldsymbol{W}_C^\top)_{[1:d,:]}\boldsymbol{h}_l^{(d+1)})$ and $\boldsymbol{w}$. Moreover, we vary the length of the prompt token N from 4 to 80 and compare the test loss with the theoretical upper bound. For each N, we conduct 10 independent experiments and report the averaged results. All experiments are performed on an NVIDIA A800 GPU.

Experiment Result Recalling that we denote $\boldsymbol{W}_B = [\boldsymbol{B}\, \boldsymbol{b}]$, Figure 1a reveals the convergence of $\boldsymbol{C}^\top \boldsymbol{B}$ to a diagonal matrix and $\boldsymbol{C}^\top \boldsymbol{b}$ to $\mathbf{0}$, confirming the theoretical induction presented in Section 5, also consistent with (Thm 4.1 (b)). Figure 1b shows that the projected hidden state $\tilde{\boldsymbol{h}}_l$ gradually aligns with $\boldsymbol{w}$ as more prompt tokens are processed, consistent with (Thm 4.1 (a)). Figure 1c demonstrates that the experimental loss has an upper bound $\frac{3d(d+1)}{2N}$ that decays linearly with N, aligning with (Thm 4.1 (c)).

7 Conclusion

This paper study Mamba's in-context learning mechanism, and rigorously establish its convergence and loss bound. By analysing the *Vector-coupled Dynamics*, we provide an exponential convergence rate with *Negative Feedback Convergence* technique in a *Fine-grained Induction*, and finally establish a $O(1/N)$ loss bound. The loss bound is comparable to that of Transformer. Our theoretical results reveal the different mechanism between Transformer and Mamba on ICL, where Mamba emulates a variant of *online gradient descent* to perform in-context, while Transformers approximate a single step of gradient descent. Furthermore, our comparison with the S4 model demonstrates that the selection components are essential for Mamba to perform ICL.

Limitations and Social Impact Our analysis focuses on one-layer Mamba model, thus the behavior of Mamba with multi-layer or augmented with other components such as MLP is still unclear. We believe that our work will provide insight for those cases and can be used to study more data models such as nonlinear features. This paper is mainly a theoretical investigation, and we do not see an immediate social impact.

Acknowledgements

We thank the anonymous reviewers for their insightful comments to improve the paper. Miao Zhang was partially sponsored by the National Natural Science Foundation of China under Grant 62306084 and U23B2051, Shenzhen College Stability Support Plan under Grant GXWD20231128102243003, and Shenzhen Science and Technology Program under Grant ZDSYS20230626091203008 and KJZD20230923115113026.

References

Ahamed, M. A. and Cheng, Q. (2024). Timemachine: A time series is worth 4 mambas for long-term forecasting. In *ECAI 2024: 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain-Including 13th Conference on Prestigious Applications of Intelligent Systems. European Conference on Artificial Intelli*, volume 392, page 1688.

Ahn, K., Cheng, X., Daneshmand, H., and Sra, S. (2023). Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650.

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. (2023). What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*.

Arora, S., Cohen, N., Golowich, N., and Hu, W. (2019). A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*.

Bai, Y., Chen, F., Wang, H., Xiong, C., and Mei, S. (2023). Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36:57125–57211.

Behrouz, A., Li, Z., Kacham, P., Daliri, M., Deng, Y., Zhong, P., Razaviyayn, M., and Mirrokni, V. (2025a). Atlas: Learning to optimally memorize the context at test time. *arXiv preprint arXiv:2505.23735*.

Behrouz, A., Razaviyayn, M., Zhong, P., and Mirrokni, V. (2025b). It's all connected: A journey through test-time memorization, attentional bias, retention, and online optimization. *arXiv preprint arXiv:2504.13173*.

Bondaschi, M., Rajaraman, N., Wei, X., Ramchandran, K., Pascanu, R., Gulcehre, C., Gastpar, M., and Makkuva, A. V. (2025). From markov to laplace: How mamba in-context learns markov chains. *arXiv preprint arXiv:2502.10178*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bu, D., Huang, W., Han, A., Nitanda, A., Suzuki, T., Zhang, Q., and Wong, H.-S. (2024). Provably transformers harness multi-concept word semantics for efficient in-context learning. *Advances in Neural Information Processing Systems*, 37:63342–63405.

Bu, D., Huang, W., Han, A., Nitanda, A., Zhang, Q., Wong, H.-S., and Suzuki, T. (2025). Provable in-context vector arithmetic via retrieving task concepts. In *Forty-second International Conference on Machine Learning*.

Chen, Y., Li, X., Liang, Y., Shi, Z., and Song, Z. (2025). The computational limits of state-space models and mamba via the lens of circuit complexity. In *The Second Conference on Parsimony and Learning (Proceedings Track)*.

Cirone, N. M., Orvieto, A., Walker, B., Salvi, C., and Lyons, T. (2024). Theoretical foundations of deep selective state-space models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Dao, T. and Gu, A. (2024). Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*.

Du, S. and Hu, W. (2019). Width provably matters in optimization for deep linear neural networks. In *International Conference on Machine Learning*, pages 1655–1664. PMLR.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. (2022). What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598.

Gatmiry, K., Saunshi, N., Reddi, S., Jegelka, S., and Kumar, S. (2024). Can looped transformers learn to implement multi-step gradient descent for in-context learning? In *Proceedings of the 41st International Conference on Machine Learning*, pages 15130–15152.

Giannou, A., Yang, L., Wang, T., Papailiopoulos, D., and Lee, J. D. (2025). How well can transformers emulate in-context newton's method? In *The 28th International Conference on Artificial Intelligence and Statistics*.

Grazzi, R., Siems, J. N., Schrodi, S., Brox, T., and Hutter, F. (2024). Is mamba capable of in-context learning? In *Proceedings of the Third International Conference on Automated Machine Learning*, volume 256 of *Proceedings of Machine Learning Research*, pages 1/1–26. PMLR.

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

Gu, A., Goel, K., and Re, C. (2022). Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*.

Huang, Y., Cheng, Y., and Liang, Y. (2023). In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*.

Lee, I., Jiang, N., and Berg-Kirkpatrick, T. (2024). Is attention required for icl? exploring the relationship between model architecture and in-context learning ability. In *The Twelfth International Conference on Learning Representations*.

Li, H., Lu, S., Cui, X., Chen, P.-Y., and Wang, M. (2025a). Understanding mamba in in-context learning with outliers: A theoretical generalization analysis. In *High-dimensional Learning Dynamics*.

Li, K., Chen, G., Yang, R., and Hu, X. (2024a). Spmamba: State-space model is all you need in speech separation. *arXiv preprint arXiv:2404.02063*.

Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., and Qiao, Y. (2024b). Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer.

Li, Y., Tarzanagh, D. A., Rawat, A. S., Fazel, M., and Oymak, S. (2025b). Gating is weighting: Understanding gated linear attention through in-context learning. *arXiv preprint arXiv:2504.04308*.

Lin, L., Bai, Y., and Mei, S. (2024). Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *The Twelfth International Conference on Learning Representations*.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., and Liu, Y. (2024). Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063.

Mahankali, A. V., Hashimoto, T., and Ma, T. (2024). 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*.

Nishikawa, N. and Suzuki, T. (2025). State space models are provably comparable to transformers in dynamic token selection. In *The Thirteenth International Conference on Learning Representations*.

Park, J., Park, J., Xiong, Z., Lee, N., Cho, J., Oymak, S., Lee, K., and Papailiopoulos, D. (2024). Can mamba learn how to learn? a comparative study on in-context learning tasks. In *Proceedings of the 41st International Conference on Machine Learning*, pages 39793–39812.

Patro, B. N. and Agneeswaran, V. S. (2024). Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges. *arXiv preprint arXiv:2404.16112*.

Sander, M. E., Giryes, R., Suzuki, T., Blondel, M., and Peyré, G. (2024). How do transformers perform in-context autoregressive learning? In *Proceedings of the 41st International Conference on Machine Learning*, pages 43235–43254.

Shen, W., Zhou, R., Yang, J., and Shen, C. (2024). On the training convergence of transformers for in-context classification of gaussian mixtures. *arXiv preprint arXiv:2410.11778*.

Sushma, N. M., Tian, Y., Mestha, H., Colombo, N., Kappel, D., and Subramoney, A. (2024). State-space models can learn in-context by gradient descent. *arXiv preprint arXiv:2410.11687*.

Tong, W. L. and Pehlevan, C. (2024). Mlps learn in-context on regression and classification tasks. *arXiv preprint arXiv:2405.15618*.

Vankadara, L. C., Xu, J., Haas, M., and Cevher, V. (2024). On feature learning in structured state space models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. (2023). Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.

Wu, J., Zou, D., Chen, Z., Braverman, V., Gu, Q., and Bartlett, P. (2024). How many pretraining tasks are needed for in-context learning of linear regression? In *The Twelfth International Conference on Learning Representations*.

Yang, S., Kautz, J., and Hatamizadeh, A. (2025). Gated delta networks: Improving mamba2 with delta rule. In *The Thirteenth International Conference on Learning Representations*.

Yang, S., Wang, B., Shen, Y., Panda, R., and Kim, Y. (2024). Gated linear attention transformers with hardware-efficient training. In *International Conference on Machine Learning*, pages 56501–56523. PMLR.

Zhang, R., Frei, S., and Bartlett, P. L. (2024). Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55.

Zhang, Y., Singh, A. K., Latham, P. E., and Saxe, A. (2025). Training dynamics of in-context learning in linear attention. *arXiv preprint arXiv:2501.16265*.

Zheng, C., Huang, W., Wang, R., Wu, G., Zhu, J., and Li, C. (2024). On mesa-optimization in autoregressively trained transformers: Emergence and capability. *Advances in Neural Information Processing Systems*, 37:49081–49129.

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 reflects the paper's scope, and the introduction reflects the paper's contributions.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss our limitation in the conclusion Section.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: We provide the assumptions in Assumption 4.1. In the main paper, we provide a proof sketch (Section 5). The detailed proofs are in the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The experiments follows our problem setup and assumption. We also provide the experiments setting in this paper, and the codes are in the supplementary material.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The codes are in the supplementary material. We also provide a readme file.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The experiments follows our problem setup and assumption. We also provide hyperparameters in the experiments setting.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We plot the 1-sigma error bar. Figure 1b, the error bar for experimental loss can be found in Figure 2c.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We report the computer resources in experiments setting. All experiments are performed on an NVIDIA A800 GPU within hours.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This work focuses on theoretical study of Mamba's in-context learning. All the data is synthesized. We see no ethical or potential harms of our work. We will not violate the Code Of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss the societal impacts in Conclusion Section. We do not see an immediate social impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: This paper is mainly a theoretical work. It poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [NA]

    Justification: We perform experiments on synthetic data. It does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [NA]

    Justification: We perform experiments on synthetic data. No new assets are introduced in the paper.

    Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [NA]

    Justification: This paper is mainly a theoretical work. The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

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

Appendix

Contents

Table 1: Key notations

Symbols	Definitions
$\boldsymbol{x}_i, \boldsymbol{x}_q, \boldsymbol{w}, y_i, y_q$	$\boldsymbol{x}_i, \boldsymbol{x}_q, \boldsymbol{w}$ are i.i.d. sampled from Gaussian distribution $\mathcal{N}(0, \boldsymbol{I}_d)$. $y_i = \boldsymbol{w}^\top \boldsymbol{x}_i, \quad y_q = \boldsymbol{w}^\top \boldsymbol{x}_q$.
$\bar{\boldsymbol{b}}_i(t), \bar{\boldsymbol{c}}_i(t), \bar{\boldsymbol{b}}(t)$	$\bar{\boldsymbol{b}}_i(t) = \frac{1}{\eta}\big(\boldsymbol{b}_i(t+1) - \boldsymbol{b}_i(t)\big)$, $\bar{\boldsymbol{c}}_i(t) = \frac{1}{\eta}\big(\boldsymbol{c}_i(t+1) - \boldsymbol{c}_i(t)\big)$, $\bar{\boldsymbol{b}}(t) = \frac{1}{\eta}\big(\boldsymbol{b}(t+1) - \boldsymbol{b}(t)\big)$.
$\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{b}_i, \boldsymbol{c}_i$	Decompose the matrices $\boldsymbol{W}_B, \boldsymbol{W}_C$ into colums of vectors: $\boldsymbol{W}_B = [\boldsymbol{B}\, \boldsymbol{b}] = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d\, \boldsymbol{b}]$, $\boldsymbol{W}_C = [\boldsymbol{C}\, \boldsymbol{c}] = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_d\, \boldsymbol{c}]$ where $\boldsymbol{W}_B, \boldsymbol{W}_C \in \mathbb{R}^{d_h \times (d+1)}$, $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{d_h \times d}$, $\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{b}_i, \boldsymbol{c}_i \in \mathbb{R}^{d_h \times 1}$
$\boldsymbol{b}_i(t)^\top \boldsymbol{b}_j(t), \boldsymbol{c}_i(t)^\top \boldsymbol{c}_j(t), \boldsymbol{b}(t)^\top \boldsymbol{b}(t)$ $\boldsymbol{c}_i(t)^\top \boldsymbol{b}_j(t), \boldsymbol{b}_i(t)^\top \boldsymbol{b}(t), \boldsymbol{c}_i(t)^\top \boldsymbol{b}(t)$	inner product of the vectors $\boldsymbol{b}, \boldsymbol{b}_i, \boldsymbol{c}_i$ with $i, j \in [1, d]$. e.g. $\boldsymbol{b}_i(t)^\top \boldsymbol{b}_j(t)$ is the inner product of $\boldsymbol{b}_i(t)$ and $\boldsymbol{b}_i(t)$.
α	A factor, $\alpha := \exp(-\Delta_l) = \exp((-\ln 2)/N)$.
$\beta_1, \beta_2, \beta_3$	The factors appearing in the gradient equation. Specifically, $\beta_1 = \Big(\alpha^2\big(1-\alpha^N\big)^2 + \frac{(d+1)\alpha^2(1-\alpha)\left(1-\alpha^{2N}\right)}{(1+\alpha)}\Big)$, $\beta_2 = \Big(d^2\alpha^2\big(1-\alpha^N\big)^2 + \frac{(2d^2+6d)\alpha^2(1-\alpha)\left(1-\alpha^{2N}\right)}{(1+\alpha)}\Big)$, $\beta_3 = \alpha\big(1-\alpha^N\big)$
γ	The lower bound of squared norms $\boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t), \boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t)$, and $\boldsymbol{b}^\top(t)\boldsymbol{b}(t)$. Specifically, $\gamma = \frac{1}{2}d_h$.
$\delta(T)$	The upper bound of cross-interactions: $\boldsymbol{b}_i^\top(t)\boldsymbol{b}_j(t), \boldsymbol{c}_i^\top(t)\boldsymbol{c}_j(t)$, and $\boldsymbol{b}_i^\top(t)\boldsymbol{b}(t)$. Specifically, $\delta(t) = \max_{s\in[0,t]}\{2\sqrt{d_h \log(4d(2d+1)/\delta)}$, $|\boldsymbol{b}_i^\top(s)\boldsymbol{b}_j(s)|, |\boldsymbol{c}_i^\top(s)\boldsymbol{c}_j(s)|, |\boldsymbol{b}_i^\top(s)\boldsymbol{b}(s)|\}$.

A Basic Calculations

This Section provide the data statistics related to gaussian distribution, and compute the expressions for the output, loss, gradient, training dynamics (particularly *Vector-coupled Dynamics*) of the Mamba model. Section B presents the *Fine-grained Induction* with *Negative Feedback Convergence* technique, and finally establish the results for Theorem 4.1. Section C details the complete proofs for Section A and Section B. In Section D, we discuss about orthogonal initialization and compare our framework with other techniques. In Section E, we give more experimental results.

A.1 Data Statistics

Lemma A.1 (Concentration Inequalities) *Let $\boldsymbol{b}_i(0)$ be the i-th colum of $\boldsymbol{B}(0)$, $\boldsymbol{c}_i(0)$ be the i-th colum of $\boldsymbol{C}(0)$, and suppose that $\delta > 0$ and $d_h = \Omega(\log(4(2d+1)/\delta))$, with probability at least $1-\delta$, we have:*

$$\frac{3d_h}{4} \le \boldsymbol{b}_i(0)^\top \boldsymbol{b}_i(0), \boldsymbol{c}_i(0)^\top \boldsymbol{c}_i(0), \boldsymbol{b}(0)^\top \boldsymbol{b}(0) \le \frac{5d_h}{4},$$

$$\Big|\boldsymbol{c}_i(0)^\top \boldsymbol{b}_i(0)\Big|, \Big|\boldsymbol{c}_i(0)^\top \boldsymbol{b}_j(0)\Big|, \Big|\boldsymbol{c}_i(0)^\top \boldsymbol{b}(0)\Big| \le 2\sqrt{d_h \log(4d(2d+1)/\delta)},$$

$$\Big|\boldsymbol{b}_i(0)^\top \boldsymbol{b}_j(0)\Big|, \Big|\boldsymbol{c}_i(0)^\top \boldsymbol{c}_j(0)\Big|, \Big|\boldsymbol{b}_i(0)^\top \boldsymbol{b}(0)\Big| \le 2\sqrt{d_h \log(4d(2d+1)/\delta)}$$

for $i, j \in [d], i \neq j$.

Proof of Lemma A.1. By Bernstein's inequality, with probability at least $1 - \delta/2(2d+1)$ we have

$$\Big|\boldsymbol{b}_i(0)^\top \boldsymbol{b}_i(0) - d_h\Big| = O\Big(\sqrt{d_h \log(4(2d+1)/\delta)}\Big).$$

Therefore, as long as $d_h = \Omega(\log(4(2d+1)/\delta))$, we have $3d_h/4 \leq \boldsymbol{b}_i(0)^\top \boldsymbol{b}_i(0) \leq 5d_h/4$. Similarly, we have

$$\frac{3d_h}{4} \leq \boldsymbol{c}_i(0)^\top \boldsymbol{c}_i(0), \boldsymbol{b}(0)^\top \boldsymbol{b}(0) \leq \frac{5d_h}{4}.$$

For $i, j \in [d], i \neq j$, By Bernstein's inequality, with probability at least $1 - \delta/2d(2d+1)$, we have

$$\Big|\boldsymbol{c}_i(0)^\top \boldsymbol{b}_i(0)\Big|, \Big|\boldsymbol{c}_i(0)^\top \boldsymbol{b}_j(0)\Big|, \Big|\boldsymbol{c}_i(0)^\top \boldsymbol{b}(0)\Big| \leq 2\sqrt{d_h \log(4d(2d+1)/\delta)},$$

$$\Big|\boldsymbol{b}_i(0)^\top \boldsymbol{b}_j(0)\Big|, \Big|\boldsymbol{c}_i(0)^\top \boldsymbol{c}_j(0)\Big|, \Big|\boldsymbol{b}_i(0)^\top \boldsymbol{b}(0)\Big| \leq 2\sqrt{d_h \log(4d(2d+1)/\delta)}.$$

We can apply a union bound to complete the proof.

Lemma A.2 *If vectors $\boldsymbol{x}$ and $\boldsymbol{w}$ are iid generated from $\mathcal{N}(0, \boldsymbol{I}_d)$, $y = \boldsymbol{x}^\top \boldsymbol{w}$ we have the following expectations:*

$$\mathbb{E}\Big[\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{w}\boldsymbol{w}^\top \boldsymbol{x}\boldsymbol{x}^\top\Big] = (d+2)\boldsymbol{I},$$

$$\mathbb{E}\Big[y^2\Big] = d,$$

$$\mathbb{E}\Big[y^4\Big] = 3d(d+2).$$

The proof of lemma A.2 is in Section C.1.

Lemma A.3 *If vectors $\boldsymbol{x}_i$ and $\boldsymbol{w}$ are iid generated from $\mathcal{N}(0, \boldsymbol{I}_d)$, $y = \boldsymbol{x}_i^\top \boldsymbol{w}$ we have the following expectations:*

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2} y_{N-i} y_{N-j} \boldsymbol{x}_{N-i} \boldsymbol{x}_{N-j}^\top\Big] = \Big(\frac{\alpha^2\big(1-\alpha^N\big)^2}{(1-\alpha)^2} + \frac{(d+1)\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)}\Big) \cdot \boldsymbol{I},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2} y_{N-i} y_{N-j}^2 \boldsymbol{x}_{N-i}\Big] = \boldsymbol{0},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2} y_{N-i}^2 y_{N-j}^2\Big] = \frac{d^2\alpha^2\Big(1-\alpha^N\Big)^2}{(1-\alpha)^2} + \frac{(2d^2+6d)\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i} \boldsymbol{x}_{N-i} \boldsymbol{w}^\top\Big] = \alpha\Big(\frac{1-\alpha^N}{1-\alpha}\Big) \cdot \boldsymbol{I},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}^2 \boldsymbol{w}\Big] = \boldsymbol{0},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2} \boldsymbol{x}_{N-i} \underbrace{\boldsymbol{x}_{N-i}^\top \boldsymbol{w}}_{y_{N-i}} \underbrace{\boldsymbol{x}_{N-j}^\top \boldsymbol{w}}_{y_{N-j}}\Big] = \boldsymbol{0},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2} y_{N-i}^2 y_{N-j}\Big] = \boldsymbol{0},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2} y_{N-i} y_{N-j}\Big] = \frac{d\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i} \boldsymbol{w}\Big] = \boldsymbol{0}.$$

The proof of lemma A.3 is in Section C.2.

A.2 Output, Loss, Gradient

This section we derive the output of Mamba given sequence $\{\boldsymbol{e}_{1:N}, \boldsymbol{e}_q\}$, and establishe the loss function formulation with its gradient expression.

Linear Recurrence. To start with, we show how the hidden states update when receiving token $\boldsymbol{e}_l = (\boldsymbol{x}_l^\top, y_l)^\top$. By (Eq. (5)) and Assumption 4.1(2), we have $\Delta_l = \mathrm{softplus}(\ln(\exp((\ln 2)/N) - 1)) = (\ln 2)/N$. Combining it with (Eq. (1)(2)) and get:

$$\begin{aligned}
\boldsymbol{h}_l^{(d+1)} &= \overline{\boldsymbol{A}}_l \boldsymbol{h}_{l-1}^{(d+1)} + \overline{\boldsymbol{B}}_l y_l \\
&= \exp(\Delta_l \boldsymbol{A}) \boldsymbol{h}_{l-1}^{(d+1)} + y_l (\Delta_l \boldsymbol{A})^{-1} (\exp(\Delta_l \boldsymbol{A}) - \boldsymbol{I}) \Delta_l \boldsymbol{B}_l \\
&= \exp(-\Delta_l) \boldsymbol{I} \boldsymbol{h}_{l-1}^{(d+1)} - y_l \Delta_l^{-1} (\exp(-\Delta_l) \boldsymbol{I} - \boldsymbol{I}) \Delta_l \boldsymbol{B}_l \\
&= \exp(-\Delta_l) \boldsymbol{h}_{l-1}^{(d+1)} + (1 - \exp(-\Delta_l)) y_l \boldsymbol{B}_l \\
&= \alpha \boldsymbol{h}_{l-1}^{(d+1)} + (1 - \alpha) y_l \boldsymbol{B}_l
\end{aligned} \tag{18}$$

where $\alpha := \exp(-\Delta_l) = \exp((-\ln 2)/N)$, the second equality is by discretization rule (2), the third equality is by Assumption 4.1(2) and $exp(-\Delta_l \boldsymbol{I}) = exp(-\Delta_l)\boldsymbol{I}$. (Eq. (18)) is similar to theorem 1 in Gu and Dao (2024)

Prediction Output. We next derive the expression of $\hat{y}_q$. Based on (Eq.(11)), the hidden state after receiving the first l context prompts $\boldsymbol{e}_{1:l}$ is given by:

$$\begin{aligned}
\boldsymbol{h}_l^{(d+1)} &= \alpha \boldsymbol{h}_{l-1}^{(d+1)} + (1 - \alpha) y_l \boldsymbol{B}_l \\
&= \alpha^2 \boldsymbol{h}_{l-2}^{(d+1)} + (1 - \alpha) y_l \boldsymbol{B}_l + (1 - \alpha) \alpha y_{l-1} \boldsymbol{B}_{l-1} \\
&= \dots \\
&= (1 - \alpha) \sum_{i=0}^{l-1} \alpha^i y_{l-i} \boldsymbol{B}_{l-i}
\end{aligned} \tag{19}$$

Receiving the query token $\boldsymbol{e}_q = (\boldsymbol{x}_q^\top, 0)^\top$, we have:

$$\boldsymbol{h}_{N+1}^{(d+1)} = \alpha \boldsymbol{h}_N^{(d+1)} + (1 - \alpha) \cdot 0 \cdot \boldsymbol{B}_N = (1 - \alpha) \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i} \boldsymbol{B}_{N-i} \tag{20}$$

Finally, the prediction output is as follows

$$\begin{aligned}
\hat{y}_q = \boldsymbol{C}_{N+1}^\top \boldsymbol{h}_{N+1}^{(d+1)} &= (1 - \alpha) \boldsymbol{C}_{N+1}^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i} \boldsymbol{B}_{N-i} \\
&= (1 - \alpha) (\boldsymbol{W}_C \boldsymbol{e}_q + \boldsymbol{b}_C)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i} (\boldsymbol{W}_B \boldsymbol{e}_{N-i} + \boldsymbol{b}_B)
\end{aligned} \tag{21}$$

To handle $\boldsymbol{W}_C \boldsymbol{e}_q$ and $\boldsymbol{W}_B \boldsymbol{e}_{N-i}$, we further denote $\boldsymbol{W}_B = [\boldsymbol{B}\, \boldsymbol{b}]$ and $\boldsymbol{W}_C = [\boldsymbol{C}\, \boldsymbol{c}]$, where $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{d_h \times d}$, $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^{d_h \times 1}$. Then we write another form of (Eq. (21)):

$$\hat{y}_q = (1 - \alpha) (\boldsymbol{C} \boldsymbol{x}_q + \boldsymbol{b}_C)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i} (\boldsymbol{B} \boldsymbol{x}_{N-i} + y_{N-i} \boldsymbol{b} + \boldsymbol{b}_B) \tag{22}$$

The loss becomes:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2} \mathbb{E} \Big[\Big((1 - \alpha) (\boldsymbol{C} \boldsymbol{x}_q + \boldsymbol{b}_C)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i} (\boldsymbol{B} \boldsymbol{x}_{N-i} + y_{N-i} \boldsymbol{b} + \boldsymbol{b}_B) - \boldsymbol{w}^\top \boldsymbol{x}_q \Big)^2 \Big] \tag{23}$$

The following lemma provides the gradient of $\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{b}, \boldsymbol{b}_B, \boldsymbol{b}_C$ with respect to loss (Eq. (23)).

Lemma A.4 (Gradient) *The gradient of trainable parameters $\boldsymbol{\theta}' = \{\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{b}, \boldsymbol{b}_B, \boldsymbol{b}_C\}$ with respect to loss (Eq.* (23)*) are as follows:*

$$\nabla_{\boldsymbol{b}_B} \mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{0},$$

$$\nabla_{\boldsymbol{b}_C} \mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{0},$$

$$\nabla_{\boldsymbol{B}} \mathcal{L}(\boldsymbol{\theta}) = \underbrace{\Big(\alpha^2\big(1-\alpha^N\big)^2 + \frac{(d+1)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)}_{:=\beta_1} \boldsymbol{C}\boldsymbol{C}^\top \boldsymbol{B} - \underbrace{\alpha\big(1-\alpha^N\big)}_{:=\beta_3} \boldsymbol{C},$$

$$\nabla_{\boldsymbol{b}} \mathcal{L}(\boldsymbol{\theta}) = \underbrace{\Big(d^2\alpha^2\big(1-\alpha^N\big)^2 + \frac{(2d^2+6d)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)}_{:=\beta_2} \boldsymbol{C}\boldsymbol{C}^\top \boldsymbol{b},$$

$$\begin{aligned}\nabla_{\boldsymbol{C}} \mathcal{L}(\boldsymbol{\theta}) = &\underbrace{\Big(\alpha^2\big(1-\alpha^N\big)^2 + \frac{(d+1)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)}_{:=\beta_1} \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{C} \\ &+ \underbrace{\Big(d^2\alpha^2\big(1-\alpha^N\big)^2 + \frac{(2d^2+6d)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)}_{:=\beta_2} \boldsymbol{b}\boldsymbol{b}^\top \boldsymbol{C} \\ &- \underbrace{\alpha\big(1-\alpha^N\big)}_{:=\beta_3} \boldsymbol{B}.\end{aligned}$$

Here, we denote $\beta_1 = \Big(\alpha^2\big(1-\alpha^N\big)^2 + \frac{(d+1)\alpha^2(1-\alpha)\left(1-\alpha^{2N}\right)}{(1+\alpha)}\Big)$, $\beta_2 = \Big(d^2\alpha^2\big(1-\alpha^N\big)^2 + \frac{(2d^2+6d)\alpha^2(1-\alpha)\left(1-\alpha^{2N}\right)}{(1+\alpha)}\Big)$, $\beta_3 = \alpha\big(1-\alpha^N\big)$ for simplicity. The proof of lemma A.4 is in Section C.3.

A.3 Training Dynamics

With the gradient in lemma A.4, we further provide the update rule for Mamba's parameters

and the *Vector-coupled Dynamics*.

Using gradient descent algorithm $\boldsymbol{\theta}'(t+1) = \boldsymbol{\theta}'(t) - \eta \nabla_{\boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}(t))$ with training rate η, we have the following update rule base on lemma A.4.

Lemma A.5 (Update Rule, restatement of lemma 5.1) *Let η be the learning rate and we use gradient descent to update the weights $\boldsymbol{W}_B, \boldsymbol{W}_C, \boldsymbol{b}_B, \boldsymbol{b}_C$, for $t \geq 0$ we have*

$$\begin{aligned}\boldsymbol{B}(t+1) &= \boldsymbol{B}(t) + \eta\beta_3 \boldsymbol{C}(t) - \eta\beta_1 \boldsymbol{C}(t)\boldsymbol{C}(t)^\top \boldsymbol{B}(t), \\ \boldsymbol{C}(t+1) &= \boldsymbol{C}(t) + \eta\beta_3 \boldsymbol{B}(t) - \eta\beta_1 \boldsymbol{B}(t)\boldsymbol{B}(t)^\top \boldsymbol{C}(t) - \eta\beta_2 \boldsymbol{b}(t)\boldsymbol{b}(t)^\top \boldsymbol{C}(t), \\ \boldsymbol{b}(t+1) &= \boldsymbol{b}(t) - \eta\beta_2 \boldsymbol{C}(t)\boldsymbol{C}(t)^\top \boldsymbol{b}(t), \\ \boldsymbol{b}_B(t) &= \boldsymbol{b}_C(t) = \boldsymbol{0}.\end{aligned}$$

We decompose $\boldsymbol{B}$ and $\boldsymbol{C}$ as $\boldsymbol{B} = [\boldsymbol{b}_1 \dots \boldsymbol{b}_d]$, $\boldsymbol{C} = [\boldsymbol{c}_1 \dots \boldsymbol{c}_d]$, and provide the update rule for $\boldsymbol{b}_i, \boldsymbol{c}_i$ and $\boldsymbol{b}$ with $i \in [1:d]$ as the following lemma.

Lemma A.6 (Vectors Update Rule, restatement of lemma 5.2) *Let η be the learning rate and we use gradient descent to update the weights $\boldsymbol{W}_B, \boldsymbol{W}_C, \boldsymbol{b}_B, \boldsymbol{b}_C$, for $i \in [d]$, $t \geq 0$ we have*

$$\boldsymbol{b}_i(t+1) = \boldsymbol{b}_i(t) + \eta\Big(\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_i(t) - \beta_1 \sum_{k \neq i}^{d} \boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t) \cdot \boldsymbol{c}_k(t)\Big)$$

$$=: \boldsymbol{b}_i(t) + \eta\bar{\boldsymbol{b}}_i(t)$$

$$\boldsymbol{c}_i(t+1) = \boldsymbol{c}_i(t) + \eta\Big(\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{b}_i(t) - \beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\cdot\boldsymbol{b}_k(t) - \beta_2\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{b}(t)\Big)$$
$$=: \boldsymbol{c}_i(t) + \eta\bar{\boldsymbol{c}}_i(t)$$

$$\boldsymbol{b}(t+1) = \boldsymbol{b}(t) - \eta\Big(\beta_2\sum_{k=1}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{c}_k(t)\Big) =: \boldsymbol{b}(t) + \eta\bar{\boldsymbol{b}}(t)$$

Here, we denote $\eta\bar{\boldsymbol{b}}_i(t) = \boldsymbol{b}_i(t+1) - \boldsymbol{b}_i(t)$, $\eta\bar{\boldsymbol{c}}_i(t) = \boldsymbol{c}_i(t+1) - \boldsymbol{c}_i(t)$, and $\eta\bar{\boldsymbol{b}}(t) = \boldsymbol{b}(t+1) - \boldsymbol{b}(t)$ for simplicity.

Next, we provide the dynamics for the inner products of these vectors.

Lemma A.7 (Vector-coupled Dynamics) *Let η be the learning rate and we use gradient descent to update the weights $\boldsymbol{W}_B, \boldsymbol{W}_C, \boldsymbol{b}_B, \boldsymbol{b}_C$, we have*

$$\begin{aligned}
&\boldsymbol{b}_i^\top(t+1)\boldsymbol{b}_i(t+1)\\
&= \boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t) + 2\eta\Big(\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t) - \beta_1\sum_{k\neq i}^{d}\big(\boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\big)^2\Big)\\
&+ \eta^2\Big\|\bar{\boldsymbol{b}}_i(t)\Big\|_2^2
\end{aligned}$$

$$\begin{aligned}
&\boldsymbol{b}_i^\top(t+1)\boldsymbol{b}_j(t+1)\\
&= \boldsymbol{b}_i^\top(t)\boldsymbol{b}_j(t) + \eta\Big(2\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t) + 2\big(\beta_3 - \beta_1\boldsymbol{c}_j^\top(t)\boldsymbol{b}_j(t)\big)\boldsymbol{c}_j^\top(t)\boldsymbol{b}_i(t)\\
&- \beta_3\big(\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t) + \boldsymbol{c}_j^\top(t)\boldsymbol{b}_i(t)\big) - 2\beta_1\sum_{k\neq i,k\neq j}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\cdot\boldsymbol{c}_k^\top(t)\boldsymbol{b}_j(t)\Big) + \eta^2\bar{\boldsymbol{b}}_i^\top(t)\bar{\boldsymbol{b}}_j(t)
\end{aligned}$$

$$\begin{aligned}
&\boldsymbol{c}_i^\top(t+1)\boldsymbol{c}_i(t+1)\\
&= \boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t) + 2\eta\Big(\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t) - \beta_1\sum_{k\neq i}^{d}\big(\boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\big)^2\\
&- \beta_2\big(\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\big)^2\Big) + \eta^2\Big\|\bar{\boldsymbol{c}}_i(t)\Big\|_2^2
\end{aligned}$$

$$\begin{aligned}
&\boldsymbol{c}_i^\top(t+1)\boldsymbol{c}_j(t+1)\\
&= \boldsymbol{c}_i^\top(t)\boldsymbol{c}_j(t) + \eta\Big(2\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_j^\top(t)\boldsymbol{b}_i(t) + 2\big(\beta_3 - \beta_1\boldsymbol{c}_j^\top(t)\boldsymbol{b}_j(t)\big)\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)\\
&- \beta_3\big(\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t) + \boldsymbol{c}_j^\top(t)\boldsymbol{b}_i(t)\big) - 2\beta_1\sum_{k\neq i,k\neq j}^{d}\boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\cdot\boldsymbol{c}_j^\top(t)\boldsymbol{b}_k(t)\\
&- 2\beta_2\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{c}_j^\top(t)\boldsymbol{b}(t)\Big) + \eta^2\bar{\boldsymbol{c}}_i^\top(t)\bar{\boldsymbol{c}}_j(t)
\end{aligned}$$

$$\begin{aligned}
&\boldsymbol{c}_i^\top(t+1)\boldsymbol{b}_i(t+1)\\
&= \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t) + \eta\Big(\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t) - \beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\cdot\boldsymbol{b}_k^\top(t)\boldsymbol{b}_i(t)\\
&- \beta_2\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{b}_i^\top(t)\boldsymbol{b}(t) + \big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t)
\end{aligned}$$

$$- \beta_1 \sum_{k\neq i}^{d} \boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\cdot \boldsymbol{c}_i^\top(t)\boldsymbol{c}_k(t)\Big) + \eta^2 \bar{\boldsymbol{c}}_i^\top(t)\bar{\boldsymbol{b}}_i(t)$$

$$\begin{aligned}
&\boldsymbol{c}_i^\top(t+1)\boldsymbol{b}_j(t+1)\\
&= \Big(1-\eta\beta_1\big(\boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t)+\boldsymbol{b}_j^\top(t)\boldsymbol{b}_j(t)\big)\Big)\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)\\
&+\eta\Big(\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{b}_i^\top(t)\boldsymbol{b}_j(t)-\beta_1\sum_{k\neq i,k\neq j}^{d}\boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\cdot\boldsymbol{b}_k^\top(t)\boldsymbol{b}_j(t)\\
&-\beta_2\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{b}_j^\top(t)\boldsymbol{b}(t)+\big(\beta_3-\beta_1\boldsymbol{c}_j^\top(t)\boldsymbol{b}_j(t)\big)\boldsymbol{c}_i^\top(t)\boldsymbol{c}_j(t)\\
&-\beta_1\sum_{k\neq i,k\neq j}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}_j(t)\cdot\boldsymbol{c}_i^\top(t)\boldsymbol{c}_k(t)\Big)+\eta^2\bar{\boldsymbol{c}}_i^\top(t)\bar{\boldsymbol{b}}_j(t)
\end{aligned}$$

$$\boldsymbol{b}^\top(t+1)\boldsymbol{b}(t+1)=\boldsymbol{b}^\top(t)\boldsymbol{b}(t)-2\eta\Big(\beta_2\sum_{k=1}^{d}\big(\boldsymbol{c}_k^\top(t)\boldsymbol{b}(t)\big)^2\Big)+\eta^2\Big\|\bar{\boldsymbol{b}}(t)\Big\|_2^2$$

$$\begin{aligned}
&\boldsymbol{b}_i^\top(t+1)\boldsymbol{b}(t+1)\\
&=\boldsymbol{b}_i^\top(t)\boldsymbol{b}(t)+\eta\Big(\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)-\beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\cdot\boldsymbol{c}_k^\top(t)\boldsymbol{b}(t)\\
&-\beta_2\sum_{k=1}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\Big)+\eta^2\bar{\boldsymbol{b}}_i^\top(t)\bar{\boldsymbol{b}}(t)
\end{aligned}$$

$$\begin{aligned}
&\boldsymbol{c}_i^\top(t+1)\boldsymbol{b}(t+1)\\
&=\Big(1-\eta\beta_2\big(\boldsymbol{b}^\top(t)\boldsymbol{b}(t)+\boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t)\big)\Big)\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\\
&+\eta\Big(\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{b}_i^\top(t)\boldsymbol{b}(t)-\beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\cdot\boldsymbol{b}_k^\top(t)\boldsymbol{b}(t)\\
&-\beta_2\sum_{k\neq i}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{c}_k^\top(t)\boldsymbol{c}_i(t)\Big)+\eta^2\bar{\boldsymbol{c}}_i^\top(t)\bar{\boldsymbol{b}}(t)
\end{aligned}$$

Lemma A.7 is derive by calculating the inner products of the vectors update rule in lemma A.6. For example, $\boldsymbol{b}^\top(t+1)\boldsymbol{b}(t+1)$ is derived as follow:

$$\begin{aligned}
\boldsymbol{b}^\top(t+1)\boldsymbol{b}(t+1)&=\Big(\boldsymbol{b}(t)+\eta\bar{\boldsymbol{b}}(t)\Big)^\top\Big(\boldsymbol{b}(t)+\eta\bar{\boldsymbol{b}}(t)\Big)\\
&=\boldsymbol{b}^\top(t)\boldsymbol{b}(t)-2\eta\bar{\boldsymbol{b}}(t)^\top\boldsymbol{b}(t)+\eta^2\Big\|\bar{\boldsymbol{b}}(t)\Big\|_2^2\\
&=\boldsymbol{b}^\top(t)\boldsymbol{b}(t)-2\eta\Big(\beta_2\sum_{k=1}^{d}\big(\boldsymbol{c}_k^\top(t)\boldsymbol{b}(t)\big)^2\Big)+\eta^2\Big\|\bar{\boldsymbol{b}}(t)\Big\|_2^2
\end{aligned}$$

The other equations are similar to it.

B Proof of Theorem 4.1

In this section, we present the framework of *Fine-grained Induction*, and establish the results of Theorem 4.1 after convergence.

Fine-gained Induction Specifically, denoting $\delta(t) = \max_{s\in[0,t]}\{|\boldsymbol{b}_i^\top(s)\boldsymbol{b}_j(s)|, |\boldsymbol{c}_i^\top(s)\boldsymbol{c}_j(s)|, |\boldsymbol{b}_i^\top(s)\boldsymbol{b}(s)|\}$ and $\gamma = \min_{t\geq 0}\{\boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t), \boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t), \boldsymbol{b}^\top(t)\boldsymbol{b}(t)\}$, we establish the following three properties $\mathcal{A}(t)$, $\mathcal{B}(t)$, and $\mathcal{C}(t)$ simultaneously for $t \geq 0$:

- $\mathcal{A}(t)$:

$$d_h/2 \leq \boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t), \boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t), \boldsymbol{b}^\top(t)\boldsymbol{b}(t) \leq 2d_h$$

- $\mathcal{B}(t)$:

$$|\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)| \leq \delta(t)\exp(-\eta\beta_1\gamma t)$$

$$|\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)| \leq 2\delta(t)\exp(-\eta\beta_1\gamma t)$$

$$|\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)| \leq 2\delta(t)\exp(-\eta\beta_2\gamma t) + \frac{\delta(t)}{\beta_2}\exp(-\eta\beta_1\gamma t)$$

- $\mathcal{C}(t)$:

$$|\boldsymbol{b}_i^\top(t)\boldsymbol{b}_j(t)|, |\boldsymbol{c}_i^\top(t)\boldsymbol{c}_j(t)|, |\boldsymbol{b}_i^\top(t)\boldsymbol{b}(t)| \leq \delta(t) \leq 3\sqrt{d_h\log(4d(2d+1)/\delta)} =: \delta_{\max}$$

Here, $i, j \in [1, d], i \neq j$. The initial conditions $\mathcal{A}(0)$, $\mathcal{B}(0)$, and $\mathcal{C}(0)$ are established with high probability by concentration inequalities (lemma A.1). We also provide the following claims to establish $\mathcal{A}(t)$, $\mathcal{B}(t)$, and $\mathcal{C}(t)$ for $t \geq 0$:

Claim B.1 $\mathcal{A}(0), \ldots, \mathcal{A}(T), \mathcal{B}(0), \ldots, \mathcal{B}(T), \mathcal{C}(0), \ldots, \mathcal{C}(T) \Longrightarrow \mathcal{A}(T+1)$

Claim B.2 $\mathcal{A}(0), \ldots, \mathcal{A}(T), \mathcal{B}(0), \ldots, \mathcal{B}(T), \mathcal{C}(0), \ldots, \mathcal{C}(T) \Longrightarrow \mathcal{B}(T+1)$

Claim B.3 $\mathcal{A}(0), \ldots, \mathcal{A}(T), \mathcal{B}(0), \ldots, \mathcal{B}(T), \mathcal{C}(0), \ldots, \mathcal{C}(T) \Longrightarrow \mathcal{C}(T+1)$

Remark. Property $\mathcal{A}(t)$ establishes the stability of quadratic norms:

$$\min\left\{\boldsymbol{b}_i(t)^\top\boldsymbol{b}_i(t),\ \boldsymbol{c}_i(t)^\top\boldsymbol{c}_i(t),\ \boldsymbol{b}(t)^\top\boldsymbol{b}(t)\right\} \geq d_h/2.$$

This norm lower bound induces two critical effects:

1. Convergence Rate: As we can see in property $\mathcal{B}(t)$, The upper bound of $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)$, $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)$ and $\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)$ is related to γ (lower bound of the squared norms), thus the stability of quadratic norms ensure a stable rapid convergence rate for property $\mathcal{B}(t)$.
2. Saddle Point Avoidance: The strict positivity (> 0) of $|\boldsymbol{b}_i|^2$ and $|\boldsymbol{c}_i|^2$ prevents the dynamics collapse to undesirable solutions $\boldsymbol{b}_i = \boldsymbol{c}_i = \boldsymbol{0}$, which would permanently make $\boldsymbol{c}_i^\top\boldsymbol{b}_i = 0$ (saddle points).

Property $\mathcal{B}(t)$ establishes a rapid exponential convergence rate:

$$\boldsymbol{C}^\top\boldsymbol{B} \to \frac{\beta_3}{\beta_1}\boldsymbol{I}, \quad \boldsymbol{C}^\top\boldsymbol{b} \to \boldsymbol{0}$$

The rapid convergence rate ensures that the variations of *Squared norms* (in property $\mathcal{A}(t)$) and *Cross-interactions* (in property $\mathcal{C}(t)$) remain bounded, thereby establishing their constraints. For example, at initialization, $\boldsymbol{b}_i^\top(0)\boldsymbol{b}_i(0)$ is bounded by $3d_h/4 \leq \boldsymbol{b}_i^\top(0)\boldsymbol{b}_i(0) \leq 5d_h/4$. Further, thanks to the exponential convergence rate in property $\mathcal{B}(t)$, we can prove that $|\boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t) - \boldsymbol{b}_i^\top(0)\boldsymbol{b}_i(0)| \leq d_h/4$, and therefore $d_h/2 \leq \boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t), \boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t), \boldsymbol{b}^\top(t)\boldsymbol{b}(t) \leq 3d_h/2 \leq 2d_h$.

Property $\mathcal{C}(t)$ establishes the upper bound for the *Cross-interactions*. As we discuss in section 5.2, if the *Squared norms* ($\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)$, $\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)$ and $\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)$) are larger enough than the *Cross-interactions* ($\boldsymbol{b}_i^\top(t)\boldsymbol{b}_j(t), \boldsymbol{c}_i^\top(t)\boldsymbol{c}_j(t)$, and $\boldsymbol{b}_i^\top(t)\boldsymbol{b}(t)$), we can make use of the *negative feedback term* to establish an exponential convergence rate. Thus property $\mathcal{C}(t)$ is also important.

The proof of claim B.1, claim B.2, and claim B.3 are in section C.4, section C.5, and section C.6 respectively.

Proof of Theorem 4.1 After convergence ($t \to 0$), we will have $\boldsymbol{C}^\top \boldsymbol{B} = \frac{\beta_3}{\beta_1}\boldsymbol{I}, \boldsymbol{C}^\top \boldsymbol{b} = \boldsymbol{0}$ (by property $\mathcal{B}(t)$), and $\boldsymbol{b}_B(t) = \boldsymbol{b}_C(t) = \boldsymbol{0}$ (by lemma A.5).

We will restate some equality for ease of reference.

Linear Recurrence (restatement of (Eq. (18)))

$$\boldsymbol{h}_l^{(d+1)} = \alpha \boldsymbol{h}_{l-1}^{(d+1)} + (1-\alpha) y_l \boldsymbol{B}_l \tag{24}$$

Prediction Output (restatement of (Eq. (22)))

$$\hat{y}_q = (1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q + \boldsymbol{b}_C)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b} + \boldsymbol{b}_B) \tag{25}$$

Loss (restatement of (Eq. (23)))

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}\Big[\Big((1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q + \boldsymbol{b}_C)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b} + \boldsymbol{b}_B) - \boldsymbol{w}^\top \boldsymbol{x}_q\Big)^2\Big] \tag{26}$$

Based on (Eq. (24)), we have

$$\begin{aligned}
&(\boldsymbol{W}_C^\top)_{[1:d,:]}(t)\boldsymbol{h}_l^{(d+1)} = \alpha(\boldsymbol{W}_C^\top)_{[1:d,:]}(t)\boldsymbol{h}_{l-1}^{(d+1)} + (1-\alpha) y_l (\boldsymbol{W}_C^\top)_{[1:d,:]}(t)\boldsymbol{B}_l \\
&= \alpha(\boldsymbol{W}_C^\top)_{[1:d,:]}(t)\boldsymbol{h}_{l-1}^{(d+1)} + (1-\alpha) y_l \boldsymbol{C}^\top(t)(\boldsymbol{B}(t)\boldsymbol{x}_l + y_l \boldsymbol{b}(t) + \boldsymbol{b}_B(t)) \\
&= \alpha(\boldsymbol{W}_C^\top)_{[1:d,:]}(t)\boldsymbol{h}_{l-1}^{(d+1)} + (1-\alpha) y_l \boldsymbol{C}^\top(t)\boldsymbol{B}(t)\boldsymbol{x}_l + (1-\alpha) y_l^2 \boldsymbol{C}^\top(t)\boldsymbol{b}(t) \\
&= \alpha(\boldsymbol{W}_C^\top)_{[1:d,:]}(t)\boldsymbol{h}_{l-1}^{(d+1)} + (1-\alpha)\frac{\beta_3}{\beta_1} y_l \boldsymbol{x}_l \\
&= \alpha(\boldsymbol{W}_C^\top)_{[1:d,:]}(t)\boldsymbol{h}_{l-1}^{(d+1)} + \frac{2(1+\alpha)(1-\alpha)}{\alpha\big(3(1-\alpha)d + 4 - 2\alpha\big)} y_l \boldsymbol{x}_l
\end{aligned} \tag{27}$$

where the second equality is by selection rule $\boldsymbol{B}_l = \boldsymbol{W}_B \boldsymbol{e}_l + \boldsymbol{b}_B$ (Eq. (3)), and $\boldsymbol{W}_B = [\boldsymbol{B}\, \boldsymbol{b}]$, $\boldsymbol{e}_l = (\boldsymbol{x}_l^\top, y_l)^\top$. The third equality is by $\boldsymbol{b}_B(t) = \boldsymbol{0}$. The fourth equality is by $\boldsymbol{C}^\top \boldsymbol{B} = \frac{\beta_3}{\beta_1}\boldsymbol{I}$ and $\boldsymbol{C}^\top \boldsymbol{b} = \boldsymbol{0}$. (Eq. (27)) establish the first equation (Thm 4.1 (a)) of the Theorem.

Based on (Eq. (25)), we have

$$\begin{aligned}
\hat{y}_q &= (1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q + \boldsymbol{b}_C)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b} + \boldsymbol{b}_B) \\
&= \boldsymbol{x}_q^\top \boldsymbol{C}^\top \sum_{i=0}^{N-1} (1-\alpha)\alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b}) \\
&= \boldsymbol{x}_q^\top \sum_{i=0}^{N-1} (1-\alpha)\alpha^{i+1} y_{N-i} \boldsymbol{C}^\top \boldsymbol{B}\boldsymbol{x}_{N-i} + \boldsymbol{x}_q^\top \sum_{i=0}^{N-1} (1-\alpha)\alpha^{i+1} y_{N-i}^2 \boldsymbol{C}^\top \boldsymbol{b} \\
&= \boldsymbol{x}_q^\top \sum_{i=0}^{N-1} (1-\alpha)\alpha^{i+1} \frac{\beta_3}{\beta_1} y_{N-i} \boldsymbol{x}_{N-i} \\
&= \boldsymbol{x}_q^\top \sum_{i=0}^{N-1} \frac{2\alpha^i(1+\alpha)(1-\alpha)}{\big(3(1-\alpha)d + 4 - 2\alpha\big)} y_{N-i} \boldsymbol{x}_{N-i}
\end{aligned} \tag{28}$$

where the second equality is by $\boldsymbol{b}_B(t) = \boldsymbol{b}_C(t) = \boldsymbol{0}$. The fourth equality is by $\boldsymbol{C}^\top \boldsymbol{B} = \frac{\beta_3}{\beta_1}\boldsymbol{I}$ and $\boldsymbol{C}^\top \boldsymbol{b} = \boldsymbol{0}$. (Eq. (28)) establish the second equation (Thm 4.1 (b)) of the Theorem.

Based on (Eq. (25)), we have

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{2}\mathbb{E}\Big[\Big((1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q+\boldsymbol{b}_C)^\top \sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i}+y_{N-i}\boldsymbol{b}+\boldsymbol{b}_B)-\boldsymbol{w}^\top\boldsymbol{x}_q\Big)^2\Big] \\
&= \frac{1}{2}\mathbb{E}\Big[\Big(\frac{\beta_3}{\beta_1}\boldsymbol{x}_q^\top\sum_{i=0}^{N-1}(1-\alpha)\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}-\boldsymbol{w}^\top\boldsymbol{x}_q\Big)^2\Big] \\
&= \underbrace{\frac{1}{2}\mathbb{E}\Big[\Big(\frac{\beta_3}{\beta_1}\boldsymbol{x}_q^\top\sum_{i=0}^{N-1}(1-\alpha)\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}\Big)^2\Big]}_{\spadesuit} \\
&\quad \underbrace{-\mathbb{E}\Big[\frac{\beta_3}{\beta_1}\boldsymbol{x}_q^\top\sum_{i=0}^{N-1}(1-\alpha)\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}\cdot\boldsymbol{w}^\top\boldsymbol{x}_q\Big]}_{\clubsuit} \\
&\quad + \frac{1}{2}\underbrace{\mathbb{E}\Big[\Big(\boldsymbol{w}^\top\boldsymbol{x}_q\Big)^2\Big]}_{=d\text{ (by lemma A.2)}}
\end{aligned}
\tag{29}
$$

We compute terms ♠ and ♣ as follows:

$$
\begin{aligned}
\spadesuit &= \mathbb{E}\Big[\Big(\frac{\beta_3}{\beta_1}\boldsymbol{x}_q^\top\sum_{i=0}^{N-1}(1-\alpha)\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}\Big)^2\Big] \\
&= \frac{\beta_3^2}{\beta_1^2}\mathbb{E}\Big[\boldsymbol{x}_q^\top\big(\sum_{i=0}^{N-1}(1-\alpha)\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}\big)\big(\sum_{i=0}^{N-1}(1-\alpha)\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}\big)^\top\boldsymbol{x}_q\Big] \\
&= \frac{(1-\alpha)^2\beta_3^2}{\beta_1^2}\mathbb{E}_{\boldsymbol{x}_q}\Big[\boldsymbol{x}_q^\top\mathbb{E}_{\boldsymbol{x}_{N-i},\boldsymbol{x}_{N-j},\boldsymbol{w}}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-j}^\top\Big]\boldsymbol{x}_q\Big] \\
&= \frac{(1-\alpha)^2\beta_3^2}{\beta_1^2}\cdot\Big(\frac{\alpha^2\big(1-\alpha^N\big)^2}{(1-\alpha)^2}+\frac{(d+1)\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)}\Big)\mathbb{E}\Big[\boldsymbol{x}_q^\top\boldsymbol{I}\boldsymbol{x}_q\Big] \\
&= \frac{d\beta_3^2}{\beta_1}
\end{aligned}
$$

For the fourth equality, $\mathbb{E}_{\boldsymbol{x}_{N-i},\boldsymbol{x}_{N-j},\boldsymbol{w}}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-j}^\top\Big] = \Big(\frac{\alpha^2\big(1-\alpha^N\big)^2}{(1-\alpha)^2}+\frac{(d+1)\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)}\Big)\cdot\boldsymbol{I}$ by lemma A.3. The last equality is by $\beta_1 = \Big(\alpha^2\big(1-\alpha^N\big)^2+\frac{(d+1)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)$

$$
\begin{aligned}
\clubsuit &= \mathbb{E}\Big[\frac{\beta_3}{\beta_1}\boldsymbol{x}_q^\top\sum_{i=0}^{N-1}(1-\alpha)\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}\cdot\boldsymbol{w}^\top\boldsymbol{x}_q\Big] \\
&= \frac{(1-\alpha)\beta_3}{\beta_1}\mathbb{E}_{\boldsymbol{x}_q}\Big[\boldsymbol{x}_q^\top\mathbb{E}_{\boldsymbol{x}_{N-i},\boldsymbol{w}}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}\boldsymbol{w}^\top\Big]\boldsymbol{x}_q\Big] \\
&= \frac{(1-\alpha)\beta_3}{\beta_1}\cdot\alpha\Big(\frac{1-\alpha^N}{1-\alpha}\Big)\mathbb{E}\Big[\boldsymbol{x}_q^\top\boldsymbol{I}\boldsymbol{x}_q\Big] \\
&= \frac{d\beta_3^2}{\beta_1}
\end{aligned}
$$

For the third equality, $\mathbb{E}_{\boldsymbol{x}_{N-i},\boldsymbol{w}}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}\boldsymbol{w}^\top\Big] = \alpha\Big(\frac{1-\alpha^N}{1-\alpha}\Big)\boldsymbol{I}$ by lemma A.3. The last equality is by $\beta_3 = \alpha\big(1-\alpha^N\big)$.

Substituting ♠ and ♣ into (Eq. (29)) and get:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \frac{d\beta_3^2}{2\beta_1} - \frac{d\beta_3^2}{\beta_1} + \frac{1}{2}d \\ &= \frac{d}{2}\Big(1 - \frac{\beta_3^2}{\beta_1}\Big) \\ &= \frac{d(d+1)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{2(1+\alpha)\beta_1} \\ &= d(d+1)(1-\alpha)\cdot\frac{\alpha^2}{\beta_1}\cdot\frac{\big(1-\alpha^{2N}\big)}{2(1+\alpha)} \\ &\le \frac{d(d+1)}{N}\cdot 4\cdot\frac{3}{8} \\ &= \frac{3d(d+1)}{2N}\end{aligned} \tag{30}$$

Recall $\beta_1 = \Big(\alpha^2\big(1-\alpha^N\big)^2 + \frac{(d+1)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)$, $\beta_3 = \alpha\big(1-\alpha^N\big)$ and $\alpha = \exp((-\ln 2)/N)$. For the inequality, $1-\alpha = 1-\exp((-\ln 2)/N) \le \frac{\ln 2}{N} \le \frac{1}{N}$, $\quad \beta_1 \ge \alpha^2\big(1-\alpha^N\big)^2 = \frac{1}{4}\alpha^2$, $1-\alpha^{2N} = 1-\frac{1}{4} = \frac{3}{4}$, thus $d(d+1)(1-\alpha) \le \frac{d(d+1)}{N}$, $\quad \frac{\alpha^2}{\beta_1} \le 4$, $\quad \frac{\big(1-\alpha^{2N}\big)}{2(1+\alpha)} \le \frac{3}{8(1+\alpha)} \le \frac{3}{8}$. (Eq. (30)) establish the third equation (Thm 4.1 (c)) of the Theorem.

C Complete Proof

This section presents the complete proof for the above results. To begin with, we provide the exact assumptions for N, η and d_h as part of Assumption 4.1.

Assumption

$$N = \Omega(d) \ge \max\{\frac{2\ln 2}{\ln 6 - \ln 5}, \frac{3(d+1)\ln 2}{2}\}$$

$$\eta = O(d^{-2}d_h^{-1}) \le \frac{1}{2d^2 d_h} \le \frac{\ln 2}{\beta_2 d_h}$$

$$d_h = \widetilde{\Omega}(d^2) \ge \max\{\lambda_1, \ldots, \lambda_{11}\}$$

where

$$\lambda_1 = \big(1728\log(4d(2d+1)/\delta) + 576(d-1)\beta_1\log(4d(2d+1)/\delta)\big)/\beta_1$$

$$\lambda_2 = \big(576\log(4d(2d+1)/\delta) + 192\log(4d(2d+1)/\delta)\big)/\beta_1$$

$$\lambda_3 = \big(1728\log(4d(2d+1)/\delta) + (576d+1872)\beta_1\log(4d(2d+1)/\delta)\big)/\beta_1$$

$$\begin{aligned}\lambda_4 = {} & 576\log(4d(2d+1)/\delta)/\beta_1 + 192(d-1)\log(4d(2d+1)/\delta) \\ & + 384\log(4d(2d+1)/\delta)/\beta_1 + 3840\ln 2\log(4d(2d+1)/\delta)\end{aligned}$$

$$\lambda_5 = 2448d\log(4d(2d+1)/\delta)$$

$$\lambda_6 = 816d\log(4d(2d+1)/\delta) + 768\ln 2d^2\log(4d(2d+1)/\delta) + 48\log(4d(2d+1)/\delta)/\beta_1$$

$$\lambda_7 = \Big(\frac{1}{\sqrt{\log(4d(2d+1)/\delta)}} + 24\sqrt{\log(4d(2d+1)/\delta)}\big(8\beta_1(d-1) + 10 + 6\beta_1 + 12\eta\beta_1/d\big)\Big)^2$$

$$\lambda_8 = 36\log(4d(2d+1)/\delta)\Big(\frac{8}{\beta_1} + 8(d-2) + 6 + \frac{12}{d}\Big)^2$$

$$\lambda_9 = 36\log(4d(2d+1)/\delta)\Big(4(d-1) + 56d\ln 2\Big)^2$$

$$\lambda_{10} = 36\log(4d(2d+1)/\delta)\Big(6 + 4\beta_1(d-1) + 2(d-1)\Big)^2$$

$$\lambda_{11} = \frac{36}{(\ln(3/2))^2}\log(4d(2d+1)/\delta)\Big(\frac{32d}{\beta_1} + \frac{8\beta_3}{\beta_1} + 32d + \frac{8}{\beta_1} + \frac{4}{\beta_1\beta_2} + 80\ln 2\Big)^2$$

Note that under the assumption of $N \geq \max\{\frac{2\ln 2}{\ln 6 - \ln 5}, \frac{3(d+1)\ln 2}{2}\}$ and combining $\alpha = \exp((-\ln 2)/N)$, we have the following:

$$\frac{5}{6} \leq \alpha^2, \quad 4\beta_1 \leq \beta_2, \quad \frac{5}{24} \leq \beta_1 \leq \frac{3}{4}, \quad 1 \leq \frac{1}{2}d \leq \beta_2, \quad \beta_3 = \Theta(1)$$

These condition will be use to prove some bounds.

C.1 Proof of Lemma A.2

Lemma C.1 (restatement of lemma A.2) *If vectors $\boldsymbol{x}$ and $\boldsymbol{w}$ are iid generated from $\mathcal{N}(0, \boldsymbol{I}_d)$, $y = \boldsymbol{x}^\top \boldsymbol{w}$ we have the following expectations:*

$$\mathbb{E}\Big[\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{w}\boldsymbol{w}^\top \boldsymbol{x}\boldsymbol{x}^\top\Big] = (d+2)\boldsymbol{I},$$

$$\mathbb{E}\Big[y^2\Big] = d,$$

$$\mathbb{E}\Big[y^4\Big] = 3d(d+2).$$

Proof. For (i,j)-th element of $\mathbb{E}\Big[\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{w}\boldsymbol{w}^\top \boldsymbol{x}\boldsymbol{x}^\top\Big]$, we have:

$$\mathbb{E}\Big[\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{w}\boldsymbol{w}^\top \boldsymbol{x}\boldsymbol{x}^\top\Big]_{[i,j]} = \mathbb{E}\Big[\boldsymbol{x}_{[i]}\sum_{k=1}^{d}\big(\boldsymbol{x}_{[k]}\boldsymbol{w}_{[k]}\big)\sum_{l=1}^{d}\big(\boldsymbol{w}_{[l]}\boldsymbol{x}_{[l]}\big)\boldsymbol{x}_{[j]}\Big]$$
$$= \sum_{k=1}^{d}\sum_{l=1}^{d}\mathbb{E}\Big[\boldsymbol{x}_{[i]}\boldsymbol{x}_{[k]}\boldsymbol{x}_{[l]}\boldsymbol{x}_{[j]}\Big]\mathbb{E}\Big[\boldsymbol{w}_{[k]}\boldsymbol{w}_{[l]}\Big]$$

According to the distribution of $\boldsymbol{w}$, we have $\mathbb{E}\Big[\boldsymbol{w}_{[k]}\boldsymbol{w}_{[l]}\Big] = \delta_{kl}$, where δ_{kl} is the Kronecker delta defined as:

$$\delta_{kl} = \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{if } k \neq l, \end{cases}$$

By Isserlis Theorem, we have:

$$\mathbb{E}\Big[\boldsymbol{x}_{[i]}\boldsymbol{x}_{[k]}\boldsymbol{x}_{[l]}\boldsymbol{x}_{[j]}\Big]$$
$$= \mathbb{E}\Big[\boldsymbol{x}_{[i]}\boldsymbol{x}_{[k]}\Big]\mathbb{E}\Big[\boldsymbol{x}_{[l]}\boldsymbol{x}_{[j]}\Big] + \mathbb{E}\Big[\boldsymbol{x}_{[i]}\boldsymbol{x}_{[l]}\Big]\mathbb{E}\Big[\boldsymbol{x}_{[k]}\boldsymbol{x}_{[j]}\Big] + \mathbb{E}\Big[\boldsymbol{x}_{[i]}\boldsymbol{x}_{[j]}\Big]\mathbb{E}\Big[\boldsymbol{x}_{[k]}\boldsymbol{x}_{[l]}\Big]$$
$$= \delta_{ik}\delta_{lj} + \delta_{il}\delta_{kj} + \delta_{ij}\delta_{kl}$$

Then we have:

$$\mathbb{E}\Big[\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{w}\boldsymbol{w}^\top \boldsymbol{x}\boldsymbol{x}^\top\Big]_{[i,j]}$$
$$= \sum_{k=1}^{d}\sum_{l=1}^{d}\mathbb{E}\Big[\boldsymbol{x}_{[i]}\boldsymbol{x}_{[k]}\boldsymbol{x}_{[l]}\boldsymbol{x}_{[j]}\Big]\mathbb{E}\Big[\boldsymbol{w}_{[k]}\boldsymbol{w}_{[l]}\Big]$$
$$= \sum_{k=1}^{d}\sum_{l=1}^{d}\big(\delta_{ik}\delta_{lj} + \delta_{il}\delta_{kj} + \delta_{ij}\delta_{kl}\big)\delta_{kl}$$
$$= \sum_{k=1}^{d}(2\delta_{ik}\delta_{kj} + \delta_{ij}\delta_{kk})$$
$$= (d+2)\delta_{ij}$$

Then we have:

$$\mathbb{E}\Big[\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{w}\boldsymbol{w}^\top \boldsymbol{x}\boldsymbol{x}^\top\Big] = (d+2)\boldsymbol{I}$$

$$\begin{aligned}
\mathbb{E}\Big[y^2\Big] &= \mathbb{E}\Big[\boldsymbol{x}^\top\boldsymbol{w}\cdot\boldsymbol{x}^\top\boldsymbol{w}\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{d}\big(\boldsymbol{x}_{[i]}\boldsymbol{w}_{[i]}\big)\sum_{j=1}^{d}\big(\boldsymbol{x}_{[j]}\boldsymbol{w}_{[j]}\big)\Big] \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\mathbb{E}\Big[\boldsymbol{x}_{[i]}\boldsymbol{x}_{[j]}\Big]\mathbb{E}\Big[\boldsymbol{w}_{[i]}\boldsymbol{w}_{[j]}\Big] \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\delta_{ij}^2 \\
&= d
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\Big[y^4\Big] &= \mathbb{E}\Big[\boldsymbol{x}^\top\boldsymbol{w}\cdot\boldsymbol{x}^\top\boldsymbol{w}\cdot\boldsymbol{x}^\top\boldsymbol{w}\cdot\boldsymbol{x}^\top\boldsymbol{w}\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{d}\big(\boldsymbol{x}_{[i]}\boldsymbol{w}_{[i]}\big)\sum_{j=1}^{d}\big(\boldsymbol{x}_{[j]}\boldsymbol{w}_{[j]}\big)\sum_{k=1}^{d}\big(\boldsymbol{x}_{[k]}\boldsymbol{w}_{[k]}\big)\sum_{l=1}^{d}\big(\boldsymbol{x}_{[l]}\boldsymbol{w}_{[l]}\big)\Big] \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{l=1}^{d}\mathbb{E}\Big[\boldsymbol{x}_{[i]}\boldsymbol{x}_{[j]}\boldsymbol{x}_{[k]}\boldsymbol{x}_{[l]}\Big]\mathbb{E}\Big[\boldsymbol{w}_{[i]}\boldsymbol{w}_{[j]}\boldsymbol{w}_{[k]}\boldsymbol{w}_{[l]}\Big] \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{l=1}^{d}\big(\delta_{ik}\delta_{lj}+\delta_{il}\delta_{kj}+\delta_{ij}\delta_{kl}\big)^2 \\
&= \Big(\underbrace{\sum_{i=j=k=l}}_{=d} + \underbrace{\sum_{i=j\neq k=l}+\sum_{i=k\neq j=l}+\sum_{i=k\neq j=l}}_{=3\cdot d(d-1)}\Big)\cdot\Big(\delta_{ik}\delta_{lj}+\delta_{il}\delta_{kj}+\delta_{ij}\delta_{kl}\Big)^2 \\
&= d\cdot 3^2 + 3\cdot d(d-1)\cdot 1^2 \\
&= 3d(d+2)
\end{aligned}$$

C.2 Proof of Lemma A.3

Lemma C.2 (restatement of lemma A.3) *If vectors $\boldsymbol{x}_i$ and $\boldsymbol{w}$ are iid generated from $\mathcal{N}(0,\boldsymbol{I}_d)$, $y=\boldsymbol{x}_i^\top\boldsymbol{w}$ we have the following expectations:*

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-j}^\top\Big] = \Big(\frac{\alpha^2\big(1-\alpha^N\big)^2}{(1-\alpha)^2}+\frac{(d+1)\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)}\Big)\cdot\boldsymbol{I},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}^2\boldsymbol{x}_{N-i}\Big] = \boldsymbol{0},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}^2y_{N-j}^2\Big] = \frac{d^2\alpha^2\Big(1-\alpha^N\Big)^2}{(1-\alpha)^2}+\frac{(2d^2+6d)\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}\boldsymbol{w}^\top\Big] = \alpha\Big(\frac{1-\alpha^N}{1-\alpha}\Big)\cdot\boldsymbol{I},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}^2\boldsymbol{w}\Big]=\mathbf{0},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-i}^\top\boldsymbol{w}\boldsymbol{x}_{N-j}^\top\boldsymbol{w}\Big]=\mathbf{0},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}^2y_{N-j}\Big]=\mathbf{0},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}\Big]=\frac{d\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)},$$

$$\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{w}\Big]=\mathbf{0}.$$

Proof.

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-j}^\top\Big]\\
&=\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}\boldsymbol{x}_{N-i}\underbrace{\boldsymbol{x}_{N-i}^\top\boldsymbol{w}}_{y_{N-i}}\underbrace{\boldsymbol{w}^\top\boldsymbol{x}_{N-j}}_{y_{N-j}}\boldsymbol{x}_{N-j}^\top\Big]\\
&=\sum_{i\neq j}\alpha^{i+j+2}\underbrace{\mathbb{E}\Big[\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-i}^\top\Big]}_{=\boldsymbol{I}}\underbrace{\mathbb{E}\Big[\boldsymbol{w}\boldsymbol{w}^\top\Big]}_{=\boldsymbol{I}}\underbrace{\mathbb{E}\Big[\boldsymbol{x}_{N-j}\boldsymbol{x}_{N-j}^\top\Big]}_{=\boldsymbol{I}}\\
&+\sum_{i=j}\alpha^{i+j+2}\underbrace{\mathbb{E}\Big[\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-i}^\top\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{x}_{N-j}\boldsymbol{x}_{N-j}^\top\Big]}_{=(d+2)\ \boldsymbol{I},\ \text{by lemma } A.2}\\
&=\Big(\Big(\sum_{i=0}^{N-1}\alpha^{i+1}\Big)^2-\Big(\sum_{i=0}^{N-1}\alpha^{2i+2}\Big)\Big)\boldsymbol{I}\\
&+\Big(\sum_{i=0}^{N-1}\alpha^{2i+2}\Big)(d+2)\boldsymbol{I}\\
&=\Big(\frac{\alpha^2\big(1-\alpha^N\big)^2}{(1-\alpha)^2}+\frac{(d+1)\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)}\Big)\cdot\boldsymbol{I}
\end{aligned}$$

Here, $\sum_{i\neq j}\alpha^{i+j+2}=\Big(\sum_{i=0}^{N-1}\alpha^{i+1}\Big)^2-\Big(\sum_{i=0}^{N-1}\alpha^{2i+2}\Big)$ for the third equality.

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}^2\boldsymbol{x}_{N-i}\Big]\\
&=\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}\boldsymbol{x}_{N-i}\underbrace{\boldsymbol{x}_{N-i}^\top\boldsymbol{w}}_{y_{N-i}}\underbrace{\boldsymbol{x}_{N-j}^\top\boldsymbol{w}}_{y_{N-j}}\underbrace{\boldsymbol{x}_{N-j}^\top\boldsymbol{w}}_{y_{N-j}}\Big]\\
&=\mathbf{0}
\end{aligned}$$

Notice that $\boldsymbol{w}$ appears three (odd) times in the second equality, if we define a function $g(\boldsymbol{w})=\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-i}^\top\boldsymbol{w}\boldsymbol{x}_{N-j}^\top\boldsymbol{w}\boldsymbol{x}_{N-j}^\top\boldsymbol{w}$, we can see that $g(-\boldsymbol{w})=-g(\boldsymbol{w})$, and further $\mathbb{E}_{\boldsymbol{w}}\Big[g(\boldsymbol{w})\Big]=\mathbf{0}$. Therefore, the above expectation equals to $\mathbf{0}$. We will use the similar property in some of the following equations.

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}^2y_{N-j}^2\Big]\\
&=\sum_{i\neq j}\alpha^{i+j+2}\mathbb{E}\Big[y_{N-i}^2\Big]\mathbb{E}\Big[y_{N-j}^2\Big]+\sum_{i=j}\alpha^{i+j+2}\mathbb{E}\Big[y_{N-i}^4\Big]\\
&=\Big(\Big(\sum_{i=0}^{N-1}\alpha^{i+1}\Big)^2-\Big(\sum_{i=0}^{N-1}\alpha^{2i+2}\Big)\Big)d^2+\Big(\sum_{i=0}^{N-1}\alpha^{2i+2}\Big)\cdot 3d(d+2)\\
&=\frac{d^2\alpha^2\Big(1-\alpha^N\Big)^2}{(1-\alpha)^2}+\frac{(2d^2+6d)\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)}
\end{aligned}
$$

where the second equality is by $\mathbb{E}\Big[y_{N-i}^2\Big]=\mathbb{E}\Big[y_{N-j}^2\Big]=d$ and $\mathbb{E}\Big[y_{N-i}^4\Big]=3d(d+2)$ (lemma A.2).

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}\boldsymbol{w}^\top\Big]\\
&=\sum_{i=0}^{N-1}\alpha^{i+1}\mathbb{E}\Big[\boldsymbol{x}_{N-i}\underbrace{\boldsymbol{x}_{N-i}^\top\boldsymbol{w}}_{y_{N-i}}\boldsymbol{w}^\top\Big]\\
&=\sum_{i=0}^{N-1}\alpha^{i+1}\underbrace{\mathbb{E}\Big[\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-i}^\top\Big]}_{=\boldsymbol{I}}\underbrace{\mathbb{E}\Big[\boldsymbol{w}\boldsymbol{w}^\top\Big]}_{=\boldsymbol{I}}\\
&=\sum_{i=0}^{N-1}\alpha^{i+1}\cdot\boldsymbol{I}\\
&=\alpha\Big(\frac{1-\alpha^N}{1-\alpha}\Big)\cdot\boldsymbol{I}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}^2\boldsymbol{w}\Big]\\
&=\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}\boldsymbol{w}\underbrace{\boldsymbol{x}_{N-i}^\top\boldsymbol{w}}_{y_{N-i}}\underbrace{\boldsymbol{x}_{N-i}^\top\boldsymbol{w}}_{y_{N-i}}\Big]\\
&=\mathbf{0}
\end{aligned}
$$

Notice that $\boldsymbol{w}$ appears three (odd) times in the second equality, thus this expectation equals to $\mathbf{0}$.

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-i}^\top\boldsymbol{w}\boldsymbol{x}_{N-j}^\top\boldsymbol{w}\Big]\\
&=\sum_{i=j}\alpha^{i+j+2}\mathbb{E}\Big[\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-i}^\top\boldsymbol{w}\boldsymbol{x}_{N-i}^\top\boldsymbol{w}\Big]\\
&+\sum_{i\neq j}\alpha^{i+j+2}\mathbb{E}\Big[\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-i}^\top\Big]\mathbb{E}\Big[\boldsymbol{w}\boldsymbol{x}_{N-j}^\top\boldsymbol{w}\Big]\\
&=\mathbf{0}+\mathbf{0}=\mathbf{0}
\end{aligned}
$$

Notice that $\boldsymbol{x}_{N-i}$ appears three (odd) times in $\mathbb{E}\Big[\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-i}^\top\boldsymbol{w}\boldsymbol{x}_{N-i}^\top\boldsymbol{w}\Big]$, and $\boldsymbol{x}_{N-j}^\top$ appears once (odd) in $\mathbb{E}\Big[\boldsymbol{w}\boldsymbol{x}_{N-j}^\top\boldsymbol{w}\Big]$, thus this expectation equals to $\mathbf{0}$.

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}^2 y_{N-j}\Big] \\
&=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}\mathbb{E}\Big[\underbrace{\boldsymbol{x}_{N-i}^\top \boldsymbol{w}}_{y_{N-i}}\underbrace{\boldsymbol{x}_{N-i}^\top \boldsymbol{w}}_{y_{N-i}}\underbrace{\boldsymbol{x}_{N-j}^\top \boldsymbol{w}}_{y_{N-j}}\Big] \\
&=\mathbf{0}
\end{aligned}
$$

Notice that $\boldsymbol{w}$ appears three (odd) times in the second equality, thus this expectation equals to $\mathbf{0}$.

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}\Big] \\
&=\sum_{i\neq j}\alpha^{i+j+2}\mathbb{E}\Big[y_{N-i}y_{N-j}\Big]+\sum_{i=j}\alpha^{i+j+2}\mathbb{E}\Big[y_{N-i}^2\Big] \\
&=\sum_{i\neq j}\alpha^{i+j+2}\mathbb{E}\Big[\underbrace{\boldsymbol{x}_{N-i}^\top \boldsymbol{w}}_{y_{N-i}}\underbrace{\boldsymbol{x}_{N-j}^\top \boldsymbol{w}}_{y_{N-j}}\Big]+\sum_{i=0}^{N-1}\alpha^{2i+2}\mathbb{E}\Big[y_{N-i}^2\Big] \\
&=\frac{d\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)}
\end{aligned}
$$

Notice that $\boldsymbol{x}_{N-i}^\top$ appears once (odd) in $\mathbb{E}\Big[\boldsymbol{x}_{N-i}^\top \boldsymbol{w}\boldsymbol{x}_{N-j}^\top \boldsymbol{w}\Big]$ where $i \neq j$, thus $\sum_{i\neq j}\alpha^{i+j+2}\mathbb{E}\Big[y_{N-i}y_{N-j}\Big]=0$. Moreover, $\mathbb{E}\Big[y_{N-i}^2\Big]=d$ by lemma A.2.

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{w}\Big] \\
&=\sum_{i=0}^{N-1}\alpha^{i+1}\mathbb{E}\Big[\boldsymbol{w}\underbrace{\boldsymbol{w}^\top \boldsymbol{x}_{N-i}}_{y_{N-i}}\Big] \\
&=\sum_{i=0}^{N-1}\alpha^{i+1}\mathbb{E}\Big[\boldsymbol{w}\boldsymbol{w}^\top\Big]\mathbb{E}\Big[\boldsymbol{x}_{N-i}\Big] \\
&=\mathbf{0}
\end{aligned}
$$

C.3 Proof of Lemma A.4

Lemma C.3 (restatement of lemma A.4) *The gradient of trainable parameters $\boldsymbol{\theta}' = \{\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{b}, \boldsymbol{b}_B, \boldsymbol{b}_C\}$ with respect to loss (Eq. (23)) are as follows:*

$$
\nabla_{\boldsymbol{b}_B}\mathcal{L}(\boldsymbol{\theta})=\mathbf{0},
$$

$$
\nabla_{\boldsymbol{b}_C}\mathcal{L}(\boldsymbol{\theta})=\mathbf{0},
$$

$$
\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{\theta})=\underbrace{\Big(\alpha^2\big(1-\alpha^N\big)^2+\frac{(d+1)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)}_{:=\beta_1}\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{B}-\underbrace{\alpha\big(1-\alpha^N\big)}_{:=\beta_3}\boldsymbol{C},
$$

$$
\nabla_{\boldsymbol{b}}\mathcal{L}(\boldsymbol{\theta})=\underbrace{\Big(d^2\alpha^2\big(1-\alpha^N\big)^2+\frac{(2d^2+6d)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)}_{:=\beta_2}\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{b},
$$

$$\nabla_{\boldsymbol{C}}\mathcal{L}(\boldsymbol{\theta}) = \underbrace{\Big(\alpha^2\big(1-\alpha^N\big)^2 + \frac{(d+1)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)}_{:=\beta_1}\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{C}$$
$$+ \underbrace{\Big(d^2\alpha^2\big(1-\alpha^N\big)^2 + \frac{(2d^2+6d)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)}_{:=\beta_2}\boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{C}$$
$$- \underbrace{\alpha\big(1-\alpha^N\big)}_{:=\beta_3}\boldsymbol{B}.$$

Proof of lemma C.3. Recalling the loss:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}\Big[\Big((1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q+\boldsymbol{b}_C)^\top\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i}+y_{N-i}\boldsymbol{b}+\boldsymbol{b}_B)-\boldsymbol{w}^\top\boldsymbol{x}_q\Big)^2\Big]$$

We will compute the gradient of $\{\boldsymbol{B},\boldsymbol{C},\boldsymbol{b},\boldsymbol{b}_B,\boldsymbol{b}_C\}$ with respect to $\mathcal{L}(\boldsymbol{\theta})$. Some expectation calculation are detailed in Section C.3.1.

$$\nabla_{\boldsymbol{b}_C}\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}\Big[(1-\alpha)\Big((1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q+\boldsymbol{b}_C)^\top\underbrace{\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i}+y_{N-i}\boldsymbol{b}+\boldsymbol{b}_B)}_{:=\boldsymbol{v}}-\boldsymbol{w}^\top\boldsymbol{x}_q\Big)$$
$$\cdot\underbrace{\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i}+y_{N-i}\boldsymbol{b}+\boldsymbol{b}_B)}_{:=\boldsymbol{v}}\Big]$$
$$= (1-\alpha)^2\mathbb{E}\Big[\boldsymbol{v}\boldsymbol{v}^\top(\boldsymbol{C}\boldsymbol{x}_q+\boldsymbol{b}_C)\Big]-(1-\alpha)\mathbb{E}\Big[\boldsymbol{v}\boldsymbol{w}^\top\boldsymbol{x}_q\Big]$$
$$= (1-\alpha)^2\mathbb{E}\Big[\boldsymbol{v}\boldsymbol{v}^\top\boldsymbol{C}\Big]\mathbb{E}\Big[\boldsymbol{x}_q\Big]+(1-\alpha)^2\mathbb{E}\Big[\boldsymbol{v}\boldsymbol{v}^\top\Big]\boldsymbol{b}_C-(1-\alpha)\mathbb{E}\Big[\boldsymbol{v}\boldsymbol{w}^\top\Big]\mathbb{E}\Big[\boldsymbol{x}_q\Big]$$

It is clear that $\mathbb{E}\Big[\boldsymbol{x}_q\Big]=\boldsymbol{0}$. Thus, if $\boldsymbol{b}_C=\boldsymbol{0}$, then $\nabla_{\boldsymbol{b}_C}\mathcal{L}(\boldsymbol{\theta})=\boldsymbol{0}$. Notice that we assume $\boldsymbol{b}_C(0)=\boldsymbol{0}$ at initialization, so by induction, $\boldsymbol{b}_C(t)=\boldsymbol{0}$ and $\nabla_{\boldsymbol{b}_C}\mathcal{L}(\boldsymbol{\theta}(t))=\boldsymbol{0}$ for $t\geq 0$. We will consider $\boldsymbol{b}_C=\boldsymbol{0}$ when computing other gradients.

$$\nabla_{\boldsymbol{b}_B}\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}\Big[(1-\alpha)\Big((1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q+\boldsymbol{b}_C)^\top\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i}+y_{N-i}\boldsymbol{b}+\boldsymbol{b}_B)-\boldsymbol{w}^\top\boldsymbol{x}_q\Big)$$
$$\cdot(\boldsymbol{C}\boldsymbol{x}_q+\boldsymbol{b}_C)\cdot\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\Big]$$
$$= \mathbb{E}\Big[(1-\alpha)\Big((1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q)^\top\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i}+y_{N-i}\boldsymbol{b}+\boldsymbol{b}_B)-\boldsymbol{w}^\top\boldsymbol{x}_q\Big)$$
$$\cdot\boldsymbol{C}\boldsymbol{x}_q\cdot\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\Big]$$
$$= (1-\alpha)^2\mathbb{E}\Big[\boldsymbol{x}_q^\top\boldsymbol{C}^\top\Big(\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i}+y_{N-i}\boldsymbol{b}+\boldsymbol{b}_B)\Big)\cdot\boldsymbol{C}\boldsymbol{x}_q\cdot\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\Big]$$
$$-(1-\alpha)\mathbb{E}\Big[\boldsymbol{w}^\top\boldsymbol{x}_q\cdot\boldsymbol{C}\boldsymbol{x}_q\cdot\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\Big]$$
$$= \frac{d\alpha^2(1-\alpha)\Big(1-\alpha^{2N}\Big)}{(1+\alpha)}\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{b}_B$$

The last equality follows from lemma C.4 where we have: $\mathbb{E}\Big[\boldsymbol{x}_q^\top \boldsymbol{C}^\top \Big(\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b} + \boldsymbol{b}_B)\Big) \cdot \boldsymbol{C}\boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}\Big] = \frac{d\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)}\boldsymbol{C}\boldsymbol{C}^\top \boldsymbol{b}_B$, and $\mathbb{E}\Big[\boldsymbol{w}^\top \boldsymbol{x}_q \cdot \boldsymbol{C}\boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}\Big] = \boldsymbol{0}$. Similar to $\boldsymbol{b}_C$, notice that $\boldsymbol{b}_B$ is initialized as $\boldsymbol{0}$, thus by induction, $\boldsymbol{b}_B(t) = \boldsymbol{0}$ and $\nabla_{\boldsymbol{b}_B}\mathcal{L}(\boldsymbol{\theta}(t)) = \boldsymbol{0}$ for $t \geq 0$. We will consider $\boldsymbol{b}_B = \boldsymbol{b}_C = \boldsymbol{0}$ when computing other gradients.

$$
\begin{aligned}
\nabla_{\boldsymbol{C}}\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}\Big[&(1-\alpha)\Big((1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q + \boldsymbol{b}_C)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b} + \boldsymbol{b}_B) - \boldsymbol{w}^\top \boldsymbol{x}_q\Big) \\
&\cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b} + \boldsymbol{b}_B)\boldsymbol{x}_q^\top\Big] \\
= (1-\alpha)^2 \mathbb{E}\Big[&\boldsymbol{x}_q^\top \boldsymbol{C}^\top \Big(\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\Big) \\
&\cdot \Big(\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\Big)\boldsymbol{x}_q^\top\Big] \\
&- (1-\alpha)\mathbb{E}\Big[\boldsymbol{w}^\top \boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\boldsymbol{x}_q^\top\Big] \\
= &\underbrace{\Big(\alpha^2\big(1-\alpha^N\big)^2 + \frac{(d+1)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)}_{:=\beta_1}\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{C} \\
&+ \underbrace{\Big(d^2\alpha^2\big(1-\alpha^N\big)^2 + \frac{(2d^2+6d)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)}_{:=\beta_2}\boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{C} \\
&- \underbrace{\alpha\big(1-\alpha^N\big)}_{:=\beta_3}\boldsymbol{B}
\end{aligned}
$$

The last equality follows from lemma C.4, where we have:

$$
\begin{aligned}
&\mathbb{E}\Big[\boldsymbol{x}_q^\top \boldsymbol{C}^\top \Big(\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\Big) \cdot \Big(\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\Big)\boldsymbol{x}_q^\top\Big] \\
&= \Big(\frac{\alpha^2\big(1-\alpha^N\big)^2}{(1-\alpha)^2} + \frac{(d+1)\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{C} \\
&+ \Big(\frac{d^2\alpha^2\Big(1-\alpha^N\Big)^2}{(1-\alpha)^2} + \frac{(2d^2+6d)\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{C}
\end{aligned}
$$

and $\mathbb{E}\Big[\boldsymbol{w}^\top \boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\boldsymbol{x}_q^\top\Big] = \alpha\Big(\frac{1-\alpha^N}{1-\alpha}\Big)\boldsymbol{B}$

$$
\begin{aligned}
\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}\Big[&(1-\alpha)\Big((1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q + \boldsymbol{b}_C)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b} + \boldsymbol{b}_B) - \boldsymbol{w}^\top \boldsymbol{x}_q\Big) \\
&\cdot (\boldsymbol{C}\boldsymbol{x}_q + \boldsymbol{b}_C) \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}\boldsymbol{x}_{N-i}^\top\Big]
\end{aligned}
$$

$$
\begin{aligned}
&= (1-\alpha)^2\mathbb{E}\Big[\boldsymbol{x}_q^\top \boldsymbol{C}^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b}) \cdot \boldsymbol{C}\boldsymbol{x}_q \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}\boldsymbol{x}_{N-i}^\top\Big] \\
&- (1-\alpha)\mathbb{E}\Big[\boldsymbol{w}^\top \boldsymbol{x}_q \cdot \boldsymbol{C}\boldsymbol{x}_q \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}\boldsymbol{x}_{N-i}^\top\Big] \\
&= \underbrace{\Big(\alpha^2\big(1-\alpha^N\big)^2 + \frac{(d+1)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)}_{:=\beta_1} \boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{B} \\
&- \underbrace{\alpha\big(1-\alpha^N\big)}_{:=\beta_3}\boldsymbol{C}
\end{aligned}
$$

The last equality follows from lemma C.4, where we have:

$$
\begin{aligned}
&\mathbb{E}\Big[\boldsymbol{x}_q^\top \boldsymbol{C}^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b}) \cdot \boldsymbol{C}\boldsymbol{x}_q \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}\boldsymbol{x}_{N-i}^\top\Big] \\
&= \Big(\frac{\alpha^2\big(1-\alpha^N\big)^2}{(1-\alpha)^2} + \frac{(d+1)\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{B}
\end{aligned}
$$

and $\mathbb{E}\Big[\boldsymbol{w}^\top \boldsymbol{x}_q \cdot \boldsymbol{C}\boldsymbol{x}_q \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}\boldsymbol{x}_{N-i}^\top\Big] = \alpha\Big(\frac{1-\alpha^N}{1-\alpha}\Big)\boldsymbol{C}$.

$$
\begin{aligned}
\nabla_{\boldsymbol{b}}\mathcal{L}(\boldsymbol{\theta}) &= \mathbb{E}\Big[(1-\alpha)\Big((1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q + \boldsymbol{b}_C)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b} + \boldsymbol{b}_B) - \boldsymbol{w}^\top\boldsymbol{x}_q\Big) \\
&\quad \cdot (\boldsymbol{C}\boldsymbol{x}_q + \boldsymbol{b}_C) \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}^2\Big] \\
&= \mathbb{E}\Big[(1-\alpha)\Big((1-\alpha)(\boldsymbol{C}\boldsymbol{x}_q)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b}) - \boldsymbol{w}^\top\boldsymbol{x}_q\Big) \\
&\quad \cdot \boldsymbol{C}\boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}^2\Big] \\
&= (1-\alpha)^2\mathbb{E}\Big[\Big(\boldsymbol{x}_q^\top \boldsymbol{C}^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\Big) \cdot \boldsymbol{C}\boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}^2\Big] \\
&\quad - (1-\alpha)\mathbb{E}\Big[\boldsymbol{w}^\top \boldsymbol{x}_q \cdot \boldsymbol{C}\boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}^2\Big] \\
&= \underbrace{\Big(d^2\alpha^2\big(1-\alpha^N\big)^2 + \frac{(2d^2+6d)\alpha^2(1-\alpha)\big(1-\alpha^{2N}\big)}{(1+\alpha)}\Big)}_{:=\beta_2} \boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{b}
\end{aligned}
$$

The last equality follows from lemma C.4, where we have:

$$
\begin{aligned}
&\mathbb{E}\Big[\Big(\boldsymbol{x}_q^\top \boldsymbol{C}^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\Big) \cdot \boldsymbol{C}\boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}^2\Big] \\
&= \Big(\frac{d^2\alpha^2\Big(1-\alpha^N\Big)^2}{(1-\alpha)^2} + \frac{(2d^2+6d)\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{b}
\end{aligned}
$$

and $\mathbb{E}\Big[\boldsymbol{w}^\top \boldsymbol{x}_q \cdot \boldsymbol{C}\boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}^2\Big] = \boldsymbol{0}$.

C.3.1 Auxiliary Lemma for Lemma A.4

Lemma C.4 *If vectors $\boldsymbol{x}_i$, $\boldsymbol{x}_q$ and $\boldsymbol{w}$ are iid generated from $\mathcal{N}(0, \boldsymbol{I}_d)$, $y = \boldsymbol{x}_i^\top \boldsymbol{w}$ we have the following expectations:*

$$\mathbb{E}\Big[\boldsymbol{x}_q^\top \boldsymbol{C}^\top \Big(\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\Big) \cdot \Big(\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\Big)\boldsymbol{x}_q^\top\Big]$$
$$= \Big(\frac{\alpha^2\big(1-\alpha^N\big)^2}{(1-\alpha)^2} + \frac{(d+1)\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{C}$$
$$+ \Big(\frac{d^2\alpha^2\Big(1-\alpha^N\Big)^2}{(1-\alpha)^2} + \frac{(2d^2+6d)\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{C}$$

$$\mathbb{E}\Big[\boldsymbol{w}^\top\boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\boldsymbol{x}_q^\top\Big] = \alpha\Big(\frac{1-\alpha^N}{1-\alpha}\Big)\boldsymbol{B}$$

$$\mathbb{E}\Big[\boldsymbol{x}_q^\top \boldsymbol{C}^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b}) \cdot \boldsymbol{C}\boldsymbol{x}_q \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}\boldsymbol{x}_{N-i}^\top\Big]$$
$$= \Big(\frac{\alpha^2\big(1-\alpha^N\big)^2}{(1-\alpha)^2} + \frac{(d+1)\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{B}$$

$$\mathbb{E}\Big[\boldsymbol{w}^\top\boldsymbol{x}_q \cdot \boldsymbol{C}\boldsymbol{x}_q \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}\boldsymbol{x}_{N-i}^\top\Big] = \alpha\Big(\frac{1-\alpha^N}{1-\alpha}\Big)\boldsymbol{C}$$

$$\mathbb{E}\Big[\Big(\boldsymbol{x}_q^\top \boldsymbol{C}^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\Big) \cdot \boldsymbol{C}\boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}^2\Big]$$
$$= \Big(\frac{d^2\alpha^2\Big(1-\alpha^N\Big)^2}{(1-\alpha)^2} + \frac{(2d^2+6d)\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{b}$$

$$\mathbb{E}\Big[\boldsymbol{w}^\top\boldsymbol{x}_q \cdot \boldsymbol{C}\boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}^2\Big] = \boldsymbol{0}$$

$$\mathbb{E}\Big[\boldsymbol{x}_q^\top \boldsymbol{C}^\top \Big(\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b} + \boldsymbol{b}_B)\Big) \cdot \boldsymbol{C}\boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}\Big]$$
$$= \frac{d\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)}\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{b}_B$$

$$\mathbb{E}\Big[\boldsymbol{w}^\top\boldsymbol{x}_q \cdot \boldsymbol{C}\boldsymbol{x}_q \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}\Big] = \boldsymbol{0}$$

Proof of lemma C.4. We will use the results of lemma A.3 to prove the above equation.

$$
\begin{aligned}
&\mathbb{E}\Big[\underbrace{\boldsymbol{x}_q^\top \boldsymbol{C}^\top \Big(\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\Big)}_{\spadesuit} \\
&\cdot \Big(\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\Big)\boldsymbol{x}_q^\top\Big] \\
&= \mathbb{E}\Big[\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}\big(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b}\big) \\
&\underbrace{\Big(\boldsymbol{x}_q^\top \boldsymbol{C}^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\Big)^\top}_{\spadesuit} \boldsymbol{x}_q^\top\Big] \\
&= \mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2}\Big(\big(\boldsymbol{B} y_{N-i}\boldsymbol{x}_{N-i} + y_{N-i}^2\boldsymbol{b}\big) \\
&\big(y_{N-j}\boldsymbol{x}_{N-j}^\top \boldsymbol{B}^\top + y_{N-j}^2 \boldsymbol{b}^\top\big)\Big)\Big]\boldsymbol{C}\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big] \\
&= \boldsymbol{B}\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2} y_{N-i}y_{N-j}\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-j}^\top\Big]\boldsymbol{B}^\top\boldsymbol{C} \\
&+ \boldsymbol{B}\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2} y_{N-i}y_{N-j}^2\boldsymbol{x}_{N-i}\Big]\boldsymbol{b}^\top\boldsymbol{C} \\
&+ \boldsymbol{b}\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2} y_{N-j}y_{N-i}^2\boldsymbol{x}_{N-j}\Big]\boldsymbol{B}^\top\boldsymbol{C} \\
&+ \mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2} y_{N-i}^2y_{N-j}^2\Big]\boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{C} \\
&= \Big(\frac{\alpha^2\big(1-\alpha^N\big)^2}{(1-\alpha)^2} + \frac{(d+1)\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{C} \\
&+ \Big(\frac{d^2\alpha^2\Big(1-\alpha^N\Big)^2}{(1-\alpha)^2} + \frac{(2d^2+6d)\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{C}
\end{aligned}
$$

The last equality follows from lemma A.3, where we have: $\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2} y_{N-i}y_{N-j}\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-j}^\top\Big] = \Big(\frac{\alpha^2\big(1-\alpha^N\big)^2}{(1-\alpha)^2} + \frac{(d+1)\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{I}$, $\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2} y_{N-i}y_{N-j}^2\boldsymbol{x}_{N-i}\Big] = \mathbf{0}$ and $\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \alpha^{i+j+2} y_{N-i}^2y_{N-j}^2\Big] = \Big(\frac{d^2\alpha^2\Big(1-\alpha^N\Big)^2}{(1-\alpha)^2} + \frac{(2d^2+6d)\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)}\Big)$.

$$
\begin{aligned}
&\mathbb{E}\Big[\underbrace{\boldsymbol{w}^\top\boldsymbol{x}_q}_{\spadesuit} \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\boldsymbol{x}_q^\top\Big] \\
&= \mathbb{E}\Big[\sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\underbrace{\boldsymbol{w}^\top\boldsymbol{x}_q}_{\spadesuit}\boldsymbol{x}_q^\top\Big]
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{B}\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}\boldsymbol{w}^\top\Big]\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big] \\
&+ \boldsymbol{b}\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}^2\boldsymbol{w}^\top\Big]\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big] \\
&= \alpha\Big(\frac{1-\alpha^N}{1-\alpha}\Big)\boldsymbol{B}
\end{aligned}
$$

The last equality follows from lemma A.3, where we have: $\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}\boldsymbol{w}^\top\Big] = \alpha\Big(\frac{1-\alpha^N}{1-\alpha}\Big)\boldsymbol{I}$, $\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}^2\boldsymbol{w}\Big] = \boldsymbol{0}$, and $\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big] = \boldsymbol{I}$.

$$
\begin{aligned}
&\mathbb{E}\Big[\underbrace{\boldsymbol{x}_q^\top\boldsymbol{C}^\top\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i}+y_{N-i}\boldsymbol{b})}_{\spadesuit}\cdot\boldsymbol{C}\boldsymbol{x}_q\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}^\top\Big] \\
&= \mathbb{E}\Big[\boldsymbol{C}\boldsymbol{x}_q\underbrace{\Big(\boldsymbol{x}_q^\top\boldsymbol{C}^\top\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i}+y_{N-i}\boldsymbol{b})\Big)}_{\spadesuit}\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}^\top\Big] \\
&= \boldsymbol{C}\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big]\boldsymbol{C}^\top\boldsymbol{B}\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-j}^\top\Big] \\
&= \Big(\frac{\alpha^2\big(1-\alpha^N\big)^2}{(1-\alpha)^2}+\frac{(d+1)\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{B}
\end{aligned}
$$

The last equality follows from lemma A.3, where we have: $\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}\boldsymbol{x}_{N-i}\boldsymbol{x}_{N-j}^\top\Big] = \Big(\frac{\alpha^2\big(1-\alpha^N\big)^2}{(1-\alpha)^2}+\frac{(d+1)\alpha^2\big(1-\alpha^{2N}\big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{I}$, and $\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big] = \boldsymbol{I}$.

$$
\begin{aligned}
&\mathbb{E}\Big[\underbrace{\boldsymbol{w}^\top\boldsymbol{x}_q}_{\spadesuit}\cdot\boldsymbol{C}\boldsymbol{x}_q\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}^\top\Big] \\
&= \boldsymbol{C}\mathbb{E}\Big[\boldsymbol{x}_q\underbrace{\boldsymbol{x}_q^\top\boldsymbol{w}}_{\spadesuit}\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}^\top\Big] \\
&= \boldsymbol{C}\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big]\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{w}\boldsymbol{x}_{N-i}^\top\Big] \\
&= \alpha\Big(\frac{1-\alpha^N}{1-\alpha}\Big)\boldsymbol{C}
\end{aligned}
$$

The last equality follows from lemma A.3, where we have: $\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{x}_{N-i}\boldsymbol{w}^\top\Big] = \alpha\Big(\frac{1-\alpha^N}{1-\alpha}\Big)\boldsymbol{I}$, and $\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big] = \boldsymbol{I}$.

$$
\mathbb{E}\Big[\underbrace{\Big(\boldsymbol{x}_q^\top\boldsymbol{C}^\top\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i}+y_{N-i}\boldsymbol{b})\Big)}_{\spadesuit}\cdot\boldsymbol{C}\boldsymbol{x}_q\cdot\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}^2\Big]
$$

$$
\begin{aligned}
&= \mathbb{E}\Big[\boldsymbol{C}\boldsymbol{x}_q \underbrace{\Big(\boldsymbol{x}_q^\top \boldsymbol{C}^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i} + y_{N-i}\boldsymbol{b})\Big)}_{\spadesuit} \cdot \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}^2\Big] \\
&= \boldsymbol{C}\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big]\boldsymbol{C}^\top\boldsymbol{B}\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2} y_{N-i}y_{N-j}^2\boldsymbol{x}_{N-i}\Big] \\
&+ \boldsymbol{C}\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big]\boldsymbol{C}^\top\boldsymbol{b}\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2} y_{N-i}^2y_{N-j}^2\Big] \\
&= \Big(\frac{d^2\alpha^2\Big(1-\alpha^N\Big)^2}{(1-\alpha)^2} + \frac{(2d^2+6d)\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)}\Big)\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{b}
\end{aligned}
$$

The last equality follows from lemma A.3, where we have: $\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}^2\boldsymbol{x}_{N-i}\Big] = \mathbf{0}$, $\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}^2y_{N-j}^2\Big] = \Big(\frac{d^2\alpha^2\Big(1-\alpha^N\Big)^2}{(1-\alpha)^2} + \frac{(2d^2+6d)\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)}\Big)$, and $\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big] = \boldsymbol{I}$.

$$
\begin{aligned}
&\mathbb{E}\Big[\underbrace{\boldsymbol{w}^\top\boldsymbol{x}_q}_{\spadesuit}\cdot\boldsymbol{C}\boldsymbol{x}_q\cdot\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}^2\Big] \\
&= \mathbb{E}\Big[\boldsymbol{C}\boldsymbol{x}_q\underbrace{\boldsymbol{x}_q^\top\boldsymbol{w}}_{\spadesuit}\cdot\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}^2\Big] \\
&= \boldsymbol{C}\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big]\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}^2\boldsymbol{w}\Big] \\
&= \mathbf{0}
\end{aligned}
$$

The last equality follows from lemma A.3, where we have $\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}^2\boldsymbol{w}\Big] = \mathbf{0}$.

$$
\begin{aligned}
&\mathbb{E}\Big[\underbrace{\boldsymbol{x}_q^\top\boldsymbol{C}^\top\Big(\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i}+y_{N-i}\boldsymbol{b}+\boldsymbol{b}_B)\Big)}_{\spadesuit}\cdot\boldsymbol{C}\boldsymbol{x}_q\cdot\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\Big] \\
&= \mathbb{E}\Big[\boldsymbol{C}\boldsymbol{x}_q\cdot\underbrace{\boldsymbol{x}_q^\top\boldsymbol{C}^\top\Big(\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}(\boldsymbol{B}\boldsymbol{x}_{N-i}+y_{N-i}\boldsymbol{b}+\boldsymbol{b}_B)\Big)}_{\spadesuit}\cdot\sum_{j=0}^{N-1}\alpha^{j+1}y_{N-j}\Big] \\
&= \boldsymbol{C}\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big]\boldsymbol{C}^\top\boldsymbol{B}\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}\boldsymbol{x}_{N-i}\Big] \\
&+ \boldsymbol{C}\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big]\boldsymbol{C}^\top\boldsymbol{b}\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}^2y_{N-j}\Big] \\
&+ \boldsymbol{C}\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big]\boldsymbol{C}^\top\boldsymbol{b}_B\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}\Big] \\
&= \frac{d\alpha^2\Big(1-\alpha^{2N}\Big)}{(1-\alpha)(1+\alpha)}\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{b}_B
\end{aligned}
$$

The last equality follows from lemma A.3, where we have: $\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}\boldsymbol{x}_{N-i}\underbrace{\boldsymbol{x}_{N-i}^\top\boldsymbol{w}}_{y_{N-i}}\underbrace{\boldsymbol{x}_{N-j}^\top\boldsymbol{w}}_{y_{N-j}}\Big]=\mathbf{0}$, $\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}^2y_{N-j}\Big]=0$, $\mathbb{E}\Big[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{i+j+2}y_{N-i}y_{N-j}\Big]=\frac{d\alpha^2\left(1-\alpha^{2N}\right)}{(1-\alpha)(1+\alpha)}$, and $\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big]=\boldsymbol{I}$.

$$
\begin{aligned}
&\mathbb{E}\Big[\underbrace{\boldsymbol{w}^\top\boldsymbol{x}_q}_{\spadesuit}\cdot\boldsymbol{C}\boldsymbol{x}_q\cdot\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\Big]\\
&=\mathbb{E}\Big[\boldsymbol{C}\boldsymbol{x}_q\underbrace{\boldsymbol{x}_q^\top\boldsymbol{w}}_{\spadesuit}\cdot\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\Big]\\
&=\boldsymbol{C}\mathbb{E}\Big[\boldsymbol{x}_q\boldsymbol{x}_q^\top\Big]\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{w}\Big]\\
&=\mathbf{0}
\end{aligned}
$$

The last equality follows from lemma A.3, where we have $\mathbb{E}\Big[\sum_{i=0}^{N-1}\alpha^{i+1}y_{N-i}\boldsymbol{w}\Big]=\mathbf{0}$.

C.4 Proof of claim B.1

This Section presents the bounds for terms $\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}_i(T+1)$, $\boldsymbol{c}_i^\top(Tt+1)\boldsymbol{c}_i(T+1)$ and $\boldsymbol{b}^\top(T+1)\boldsymbol{b}(T+1)$, establishing the property $\mathcal{A}(T+1)$.

Recurring the *Vector-coupled Dynamics* equations of $\boldsymbol{b}_i^\top(t+1)\boldsymbol{b}_i(t+1)$, $\boldsymbol{c}_i^\top(t+1)\boldsymbol{c}_i(t+1)$ and $\boldsymbol{b}^\top(t+1)\boldsymbol{b}(t+1)$ in lemma A.7, we have:

$$
\begin{aligned}
&\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}_i(T+1)=\boldsymbol{b}_i^\top(T)\boldsymbol{b}_i(T)+2\eta\Big(\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\\
&-\beta_1\sum_{k\neq i}^{d}\big(\boldsymbol{c}_k^\top(T)\boldsymbol{b}_i(T)\big)^2\Big)+\eta^2\Big\|\bar{\boldsymbol{b}}_i(T)\Big\|_2^2\\
&=\boldsymbol{b}_i^\top(0)\boldsymbol{b}_i(0)+\sum_{s=0}^{T}\Big(2\eta\Big(\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\big)\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)-\beta_1\sum_{k\neq i}^{d}\big(\boldsymbol{c}_k^\top(s)\boldsymbol{b}_i(s)\big)^2\Big)\\
&+\eta^2\Big\|\bar{\boldsymbol{b}}_i(s)\Big\|_2^2\Big)\\
&=\boldsymbol{b}_i^\top(0)\boldsymbol{b}_i(0)+2\eta\underbrace{\sum_{s=0}^{T}\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\big)\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)}_{\text{term I}}-2\eta\beta_1\underbrace{\sum_{k\neq i}^{d}\sum_{s=0}^{T}\big(\boldsymbol{c}_k^\top(s)\boldsymbol{b}_i(s)\big)^2}_{\text{term II}}\\
&+\eta^2\underbrace{\sum_{s=0}^{T}\Big\|\bar{\boldsymbol{b}}_i(s)\Big\|_2^2}_{\text{term III}}
\end{aligned}
$$

$$
\begin{aligned}
&\boldsymbol{c}_i^\top(T+1)\boldsymbol{c}_i(T+1)=\boldsymbol{c}_i^\top(T)\boldsymbol{c}_i(T)+2\eta\Big(\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\\
&-\beta_1\sum_{k\neq i}^{d}\big(\boldsymbol{c}_i^\top(T)\boldsymbol{b}_k(T)\big)^2-\beta_2\big(\boldsymbol{c}_i^\top(T)\boldsymbol{b}(T)\big)^2\Big)+\eta^2\Big\|\bar{\boldsymbol{c}}_i(T)\Big\|_2^2\\
&=\boldsymbol{c}_i^\top(T)\boldsymbol{c}_i(T)+\sum_{s=0}^{T}\Big(2\eta\Big(\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\big)\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)-\beta_1\sum_{k\neq i}^{d}\big(\boldsymbol{c}_i^\top(s)\boldsymbol{b}_k(s)\big)^2
\end{aligned}
$$

$$
\begin{aligned}
&- \beta_2\big(\boldsymbol{c}_i^\top(s)\boldsymbol{b}(s)\big)^2\Big) + \eta^2\Big\|\bar{\boldsymbol{c}}_i(s)\Big\|_2^2\Big) \\
&= \boldsymbol{c}_i^\top(0)\boldsymbol{c}_i(0) + 2\eta\underbrace{\sum_{s=0}^{T}\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\big)\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)}_{\text{term I}} - 2\eta\beta_1\sum_{k\neq i}^{d}\underbrace{\sum_{s=0}^{T}\big(\boldsymbol{c}_i^\top(s)\boldsymbol{b}_k(s)\big)^2}_{=\text{term II}} \\
&- 2\eta\beta_2\underbrace{\sum_{s=0}^{T}\big(\boldsymbol{c}_i^\top(s)\boldsymbol{b}(s)\big)^2}_{\text{term IV}} + \eta^2\underbrace{\sum_{s=0}^{T}\Big\|\bar{\boldsymbol{c}}_i(s)\Big\|_2^2}_{\text{term V}}
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{b}^\top(T+1)\boldsymbol{b}(T+1) &= \boldsymbol{b}^\top(T)\boldsymbol{b}(T) - 2\eta\Big(\beta_2\sum_{k=1}^{d}\big(\boldsymbol{c}_k^\top(T)\boldsymbol{b}(T)\big)^2\Big) + \eta^2\Big\|\bar{\boldsymbol{b}}(T)\Big\|_2^2 \\
&= \boldsymbol{b}^\top(0)\boldsymbol{b}(0) + 2\eta\beta_2\sum_{k=1}^{d}\underbrace{\sum_{s=0}^{T}\big(\boldsymbol{c}_k^\top(s)\boldsymbol{b}(s)\big)^2}_{=\text{term IV}} + \eta^2\underbrace{\sum_{s=0}^{T}\Big\|\bar{\boldsymbol{b}}(s)\Big\|_2^2}_{\text{term VI}}
\end{aligned}
$$

To bound terms I - VI, we will use some inequalities from property $\mathcal{B}(t)$ and lemma C.7 as following with $i, j, k \in [1, d], i \neq j$:

$$
\begin{aligned}
|\beta_3 - \beta_1\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)| &\leq \delta(t)\exp(-\eta\beta_1\gamma t) \\
|\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)| &\leq 2\delta(t)\exp(-\eta\beta_1\gamma t) \\
|\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)| &\leq 2\delta(t)\exp(-\eta\beta_2\gamma t) + \frac{\delta(t)}{\beta_2}\exp(-\eta\beta_1\gamma t) \\
\Big|\bar{\boldsymbol{b}}_i(t)^\top\bar{\boldsymbol{b}}_k(t)\Big| &\leq 8d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t), \\
\Big|\bar{\boldsymbol{c}}_i(t)^\top\bar{\boldsymbol{c}}_k(t)\Big| &\leq 8d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t) + 40\beta_2^2 d_h\delta(t)^2\exp(-\eta\beta_2\gamma t), \\
\Big\|\bar{\boldsymbol{b}}(t)\Big\|_2^2 &\leq 16d_h\beta_2^2 d^2\delta(t)^2\exp(-\eta\beta_2\gamma t) + 2d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t),
\end{aligned}
$$

Next we begin bounding terms I - VI.

Bound of term I: By $\Big|\beta_3 - \beta_1\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\Big| \leq \delta(s)\exp(-\eta\beta_1\gamma s)$, we have:

$$
\begin{aligned}
\Big|\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\Big| &\leq \frac{\delta(s)\exp(-\eta\beta_1\gamma s) + \beta_3}{\beta_1} \\
&\leq \frac{4\delta(s) + 2\alpha}{\alpha^2} \\
&\leq \frac{5\delta(s)}{\alpha^2} \\
&\leq 6\delta(s)
\end{aligned}
$$

The third inequality is by $\delta(s) \geq 2\sqrt{d_h\log(4d(2d+1)/\delta)} \geq 2\alpha = 2\alpha = 2\exp((-\ln 2)/N)$. For the last inequality, as long as $N \geq \frac{2\ln 2}{\ln 6 - \ln 5}$, we have $\frac{5}{\alpha^2} \leq 6$.

$$
\begin{aligned}
&\Big|\sum_{s=0}^{T}\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\big)\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\Big| \\
&\leq \sum_{s=0}^{T}6\delta(s)^2\exp(-\eta\beta_1\gamma s)
\end{aligned}
$$

$$
\begin{aligned}
&\leq 6\delta_{\max}^2 \int_{-1}^{\infty} \exp(-\eta\beta_1\gamma s) ds \\
&\leq \frac{6\delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\eta\beta_1\gamma}
\end{aligned}
$$

The second inequality is due to $\exp(-\eta\beta_1\gamma s)$ is monotone decreasing.

Bound of term II:

$$
\begin{aligned}
&\sum_{s=0}^{T} \left(\boldsymbol{c}_k^\top(s)\boldsymbol{b}_i(s)\right)^2 \\
&\leq \sum_{s=0}^{T} \left(2\delta(s)\exp(-\eta\beta_1\gamma s)\right)^2 \\
&\leq 4\delta_{\max}^2 \int_{-1}^{\infty} \exp(-2\eta\beta_1\gamma s) ds \\
&\leq \frac{2\delta_{\max}^2 \exp(2\eta\beta_1\gamma)}{\eta\beta_1\gamma}
\end{aligned}
$$

The second inequality is due to $\exp(-2\eta\beta_1\gamma s)$ is monotone decreasing.

Bound of term III:

$$
\begin{aligned}
&\sum_{s=0}^{T} \left\|\bar{\boldsymbol{b}}_i(s)\right\|_2^2 \\
&\leq \sum_{s=0}^{T} 8 d_h d^2 \delta(t)^2 \exp(-\eta\beta_1\gamma t) \\
&\leq 8 d_h d^2 \delta_{\max}^2 \int_{-1}^{\infty} \exp(-\eta\beta_1\gamma s) ds \\
&\leq \frac{8 d_h d^2 \delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\eta\beta_1\gamma}
\end{aligned}
$$

The second inequality is due to $\exp(-\eta\beta_1\gamma s)$ is monotone decreasing.

Bound of term IV:

$$
\begin{aligned}
&\sum_{s=0}^{T} \left(\boldsymbol{c}_i^\top(s)\boldsymbol{b}(s)\right)^2 \\
&= \sum_{s=0}^{T} \left(2\delta(s)\exp(-\eta\beta_2\gamma s) + \frac{\delta(s)}{\beta_2}\exp(-\eta\beta_1\gamma s)\right)^2 \\
&\leq \delta_{\max}^2 \cdot \left(4\sum_{s=0}^{T}\exp(-2\eta\beta_2\gamma s) + \frac{4}{\beta_2}\sum_{s=0}^{T}\exp(-\eta(\beta_1+\beta_2)\gamma s) + \frac{1}{\beta_2^2}\sum_{s=0}^{T}\exp(-2\eta\beta_1\gamma s)\right) \\
&\leq \delta_{\max}^2 \cdot \left(4\int_{-1}^{\infty}\exp(-2\eta\beta_2\gamma s) ds + \frac{4}{\beta_2}\int_{-1}^{\infty}\exp(-\eta(\beta_1+\beta_2)\gamma s) ds\right. \\
&\left. + \frac{1}{\beta_2^2}\int_{-1}^{\infty}\exp(-2\eta\beta_1\gamma s) ds\right) \\
&= \frac{2\delta_{\max}^2\exp(2\eta\beta_2\gamma)}{\eta\beta_2\gamma} + \frac{4\delta_{\max}^2\exp(\eta(\beta_1+\beta_2)\gamma)}{\eta\beta_2(\beta_1+\beta_2)\gamma} + \frac{\delta_{\max}^2\exp(2\eta\beta_1\gamma)}{2\eta\beta_1\beta_2^2\gamma} \\
&\leq \frac{17\delta_{\max}^2}{\eta\beta_2\gamma}
\end{aligned}
$$

The second inequality is due to $\exp(-2\eta\beta_2\gamma s)$, $\exp(-\eta(\beta_1+\beta_2)\gamma s)$ and $\exp(-2\eta\beta_1\gamma s)$ are monotone decreasing. The last inequality is by $\frac{\exp(\eta(\beta_1+\beta_2)\gamma)}{(\beta_1+\beta_2)} \leq 2$ and $\frac{\exp(2\eta\beta_1\gamma)}{2\beta_1\beta_2} \leq 7$ since $\exp(2\eta\beta_1\gamma) \leq \exp(\eta(\beta_1+\beta_2)\gamma) \leq 2$ and $\beta_1+\beta_2 \geq 1$, $\beta_1\beta_2 \geq \frac{1}{7}$.

Bound of term V:

$$
\begin{aligned}
&\sum_{s=0}^{T} \left\| \bar{\boldsymbol{c}}_i(s) \right\|_2^2 \\
&\leq \sum_{s=0}^{T} \Big(8d_h d^2 \delta(t)^2 \exp(-\eta\beta_1\gamma t) + 40\beta_2^2 d_h \delta(t)^2 \exp(-\eta\beta_2\gamma t) \Big) \\
&\leq 8d_h d^2 \delta_{\max}^2 \int_{-1}^{\infty} \exp(-\eta\beta_1\gamma s) ds + 40\beta_2^2 d_h \delta_{\max}^2 \int_{-1}^{\infty} \exp(-\eta\beta_2\gamma s) ds \\
&= \frac{8d_h d^2 \delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\eta\beta_1\gamma} + \frac{40\beta_2 d_h \delta_{\max}^2 \exp(\eta\beta_2\gamma)}{\eta\gamma} \\
&\leq \frac{16d_h d^2 \delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\eta\beta_1\gamma} + \frac{80\beta_2 d_h \delta_{\max}^2 \exp(\eta\beta_2\gamma)}{\eta\gamma}
\end{aligned}
$$

The second inequality is due to $\exp(-\eta\beta_2\gamma s)$ and $\exp(-\eta\beta_1\gamma s)$ are monotone decreasing.

Bound of term VI:

$$
\begin{aligned}
&\sum_{s=0}^{T} \left\| \bar{\boldsymbol{b}}(s) \right\|_2^2 \\
&\leq \sum_{s=0}^{T} \Big(16d_h \beta_2^2 d^2 \delta(t)^2 \exp(-\eta\beta_2\gamma t) + 2d_h d^2 \delta(t)^2 \exp(-\eta\beta_1\gamma t) \Big) \\
&\leq 16d_h \beta_2^2 d^2 \delta_{\max}^2 \int_{-1}^{\infty} \exp(-\eta\beta_2\gamma s) ds + 2d_h d^2 \delta_{\max}^2 \int_{-1}^{\infty} \exp(-\eta\beta_1\gamma s) ds \\
&\leq \frac{16d_h \beta_2 d^2 \delta_{\max}^2 \exp(\eta\beta_2\gamma)}{\eta\gamma} + \frac{2d_h d^2 \delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\eta\beta_1\gamma}
\end{aligned}
$$

The second inequality is due to $\exp(-\eta\beta_2\gamma s)$ and $\exp(-\eta\beta_1\gamma s)$ are monotone decreasing.

We next use the bounds of I - VI to bound $\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}_i(T+1)$, $\boldsymbol{c}_i^\top(T+1)\boldsymbol{c}_i(T+1)$ and $\boldsymbol{b}^\top(T+1)\boldsymbol{b}(T+1)$.

Lower bound of $\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}_i(T+1)$

$$
\begin{aligned}
&\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}_i(T+1) \\
&= \underbrace{\boldsymbol{b}_i^\top(0)\boldsymbol{b}_i(0)}_{\geq \frac{3d_h}{4}} + 2\eta \underbrace{\sum_{s=0}^{T} \big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\big)\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)}_{\text{term I}} - 2\eta\beta_1 \sum_{k\neq i}^{d} \underbrace{\sum_{s=0}^{T} \big(\boldsymbol{c}_k^\top(s)\boldsymbol{b}_i(s)\big)^2}_{\text{term II}} \\
&+ \eta^2 \underbrace{\sum_{s=0}^{T} \left\| \bar{\boldsymbol{b}}_i(s) \right\|_2^2}_{\geq 0} \\
&\geq \frac{3d_h}{4} - 2\eta \cdot \frac{6\delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\eta\beta_1\gamma} - 2\eta\beta_1(d-1) \cdot \frac{2\delta_{\max}^2 \exp(2\eta\beta_1\gamma)}{\eta\beta_1\gamma} \\
&\geq \frac{3d_h}{4} - \frac{12\delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\beta_1\gamma} - \frac{4(d-1)\delta_{\max}^2 \exp(2\eta\beta_1\gamma)}{\gamma}
\end{aligned}
$$

$$
\begin{aligned}
&\geq \frac{3d_h}{4} - \frac{2 * 12 * 9d_h \log(4d(2d+1)/\delta)}{\beta_1 \frac{1}{2} d_h} - \frac{2 * 4(d-1) * 9d_h \log(4d(2d+1)/\delta)}{\frac{1}{2} d_h} \\
&\geq \frac{d_h}{2}
\end{aligned}
$$

The third inequality is by $\delta_{\max} = 3\sqrt{d_h \log(4d(2d+1)/\delta)}$, $\exp(\eta\beta_1\gamma) \leq \exp(2\eta\beta_1\gamma) \leq 2$ and $\gamma = \frac{1}{2} d_h$. The last inequality follows from $d_h = \widetilde{\Omega}(d^2) \geq \big(1728 \log(4d(2d+1)/\delta) + 576(d-1)\beta_1 \log(4d(2d+1)/\delta)\big)/\beta_1$.

Upper bound of $\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}_i(T+1)$

$$
\begin{aligned}
&\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}_i(T+1) \\
&= \underbrace{\boldsymbol{b}_i^\top(0)\boldsymbol{b}_i(0)}_{\leq \frac{5d_h}{4}} - 2\eta \underbrace{\sum_{s=0}^{T} \big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\big)\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)}_{\text{term I}} - 2\eta\beta_1 \sum_{k\neq i}^{d} \underbrace{\sum_{s=0}^{T} \big(\boldsymbol{c}_k^\top(s)\boldsymbol{b}_i(s)\big)^2}_{\text{term II}} \\
&+ \eta^2 \underbrace{\sum_{s=0}^{T} \Big\|\overline{\boldsymbol{b}}_i(s)\Big\|_2^2}_{\text{term III}} \\
&\leq \frac{5d_h}{4} + 2\eta \cdot \frac{6\delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\eta\beta_1\gamma} + \eta^2 \cdot \frac{8d_h d^2 \delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\eta\beta_1\gamma} \\
&\leq \frac{5d_h}{4} + \frac{12\delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\beta_1\gamma} + \frac{8\eta d_h d^2 \delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\beta_1\gamma} \\
&\leq \frac{5d_h}{4} + \frac{2 * 12 * 9d_h \log(4d(2d+1)/\delta)}{\beta_1 \frac{1}{2} d_h} + \frac{2 * 8\eta d_h d^2 * 9d_h \log(4d(2d+1)/\delta)}{\beta_1 \frac{1}{2} d_h} \\
&\leq 2d_h
\end{aligned}
$$

The third inequality is by $\delta_{\max} = 3\sqrt{d_h \log(4d(2d+1)/\delta)}$, $\exp(\eta\beta_1\gamma) \leq 2$ and $\gamma = \frac{1}{2} d_h$. The last inequality follows from

$$
\begin{aligned}
d_h &= \widetilde{\Omega}(d^2) \\
&\geq \big(576 \log(4d(2d+1)/\delta) + 192 \log(4d(2d+1)/\delta)\big)/\beta_1 \\
&\geq \big(576 \log(4d(2d+1)/\delta) + 384\eta d_h d^2 \log(4d(2d+1)/\delta)\big)/\beta_1
\end{aligned}
$$

Lower bound of $\boldsymbol{c}_i^\top(T+1)\boldsymbol{c}_i(T+1)$

$$
\begin{aligned}
&\boldsymbol{c}_i^\top(T+1)\boldsymbol{c}_i(T+1) \\
&= \underbrace{\boldsymbol{c}_i^\top(0)\boldsymbol{c}_i(0)}_{\geq \frac{3d_h}{4}} + 2\eta \underbrace{\sum_{s=0}^{T} \big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\big)\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)}_{\text{term I}} - 2\eta\beta_1 \sum_{k\neq i}^{d} \underbrace{\sum_{s=0}^{T} \big(\boldsymbol{c}_i^\top(s)\boldsymbol{b}_k(s)\big)^2}_{=\text{ term II}} \\
&- 2\eta\beta_2 \underbrace{\sum_{s=0}^{T} \big(\boldsymbol{c}_i^\top(s)\boldsymbol{b}(s)\big)^2}_{\text{term IV}} + \eta^2 \underbrace{\sum_{s=0}^{T} \Big\|\overline{\boldsymbol{c}}_i(s)\Big\|_2^2}_{\geq 0} \\
&\geq \frac{3d_h}{4} - 2\eta \cdot \frac{6\delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\eta\beta_1\gamma} - 2\eta\beta_1(d-1) \cdot \frac{2\delta_{\max}^2 \exp(2\eta\beta_1\gamma)}{\eta\beta_1\gamma} - 2\eta\beta_2 \cdot \frac{17\delta_{\max}^2}{\eta\beta_2\gamma} \\
&\geq \frac{3d_h}{4} - \frac{2 * 12 * 9d_h \log(4d(2d+1)/\delta)}{\beta_1 \frac{1}{2} d_h} - \frac{2 * 4(d-1) * 9d_h \log(4d(2d+1)/\delta)}{\frac{1}{2} d_h} \\
&- \frac{34 * 9d_h \log(4d(2d+1)/\delta)}{\frac{1}{2} d_h}
\end{aligned}
$$

$$\geq \frac{d_h}{2}$$

The second inequality is by $\delta_{\max} = 3\sqrt{d_h \log(4d(2d+1)/\delta)}$, $\exp(\eta\beta_1\gamma) \leq \exp(2\eta\beta_1\gamma) \leq 2$ and $\gamma = \frac{1}{2}d_h$. The last inequality follows from $d_h = \widetilde{\Omega}(d^2) \geq \big(1728\log(4d(2d+1)/\delta) + (576d + 1872)\beta_1 \log(4d(2d+1)/\delta)\big)/\beta_1$.

Upper bound of $\boldsymbol{c}_i^\top(T+1)\boldsymbol{c}_i(T+1)$

$$\begin{aligned}
&\boldsymbol{c}_i^\top(T+1)\boldsymbol{c}_i(T+1) \\
&= \underbrace{\boldsymbol{c}_i^\top(0)\boldsymbol{c}_i(0)}_{\leq \frac{5d_h}{4}} + 2\eta \underbrace{\sum_{s=0}^{T} \big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\big)\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)}_{\text{term I}} - 2\eta\beta_1 \sum_{k\neq i}^{d} \underbrace{\sum_{s=0}^{T} \big(\boldsymbol{c}_i^\top(s)\boldsymbol{b}_k(s)\big)^2}_{=\text{term II}} \\
&- 2\eta\beta_2 \underbrace{\sum_{s=0}^{T} \big(\boldsymbol{c}_i^\top(s)\boldsymbol{b}(s)\big)^2}_{\geq 0} + \eta^2 \underbrace{\sum_{s=0}^{T} \Big\|\bar{\boldsymbol{c}}_i(s)\Big\|_2^2}_{\text{term V}} \\
&\leq \frac{5d_h}{4} + 2\eta \cdot \frac{6\delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\eta\beta_1\gamma} + 2\eta\beta_1(d-1) \cdot \frac{2\delta_{\max}^2 \exp(2\eta\beta_1\gamma)}{\eta\beta_1\gamma} \\
&+ \eta^2 \cdot \big(\frac{16 d_h d^2 \delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\eta\beta_1\gamma} + \frac{80\beta_2 d_h \delta_{\max}^2 \exp(\eta\beta_2\gamma)}{\eta\gamma}\big) \\
&\leq \frac{5d_h}{4} + \frac{2 * 12 * 9 d_h \log(4d(2d+1)/\delta)}{\beta_1 \frac{1}{2} d_h} + \frac{2 * 4(d-1) * 9 d_h \log(4d(2d+1)/\delta)}{\frac{1}{2} d_h} \\
&+ \frac{2 * 16\eta d_h d^2 * 9 d_h \log(4d(2d+1)/\delta)}{\beta_1 \frac{1}{2} d_h} + \frac{2 * 80\eta\beta_2 d_h * 9 d_h \log(4d(2d+1)/\delta)}{\frac{1}{2} d_h} \\
&\leq 2d_h
\end{aligned}$$

The second inequality is by $\delta_{\max} = 3\sqrt{d_h \log(4d(2d+1)/\delta)}$, $\exp(\eta\beta_1\gamma) \leq \exp(2\eta\beta_1\gamma) \leq 2$ and $\gamma = \frac{1}{2}d_h$. The last inequality follows from

$$\begin{aligned}
d_h &= \widetilde{\Omega}(d^2) \\
&\geq 576\log(4d(2d+1)/\delta)/\beta_1 + 192(d-1)\log(4d(2d+1)/\delta) \\
&+ 384\log(4d(2d+1)/\delta)/\beta_1 + 3840\ln 2 \log(4d(2d+1)/\delta) \\
&\geq 576\log(4d(2d+1)/\delta)/\beta_1 + 192(d-1)\log(4d(2d+1)/\delta) \\
&+ 768\eta d_h d^2 \log(4d(2d+1)/\delta)/\beta_1 + 3840\eta\beta_2 d_h \log(4d(2d+1)/\delta)
\end{aligned}$$

Lower bound of $\boldsymbol{b}^\top(T+1)\boldsymbol{b}(T+1)$

$$\begin{aligned}
\boldsymbol{b}^\top(T+1)\boldsymbol{b}(T+1) &= \underbrace{\boldsymbol{b}^\top(0)\boldsymbol{b}(0)}_{\geq \frac{3d_h}{4}} - 2\eta\beta_2 \sum_{k=1}^{d} \underbrace{\sum_{s=0}^{T} \big(\boldsymbol{c}_k^\top(s)\boldsymbol{b}(s)\big)^2}_{=\text{term IV}} + \eta^2 \underbrace{\sum_{s=0}^{T} \Big\|\bar{\boldsymbol{b}}(s)\Big\|_2^2}_{\geq 0} \\
&\geq \frac{3d_h}{4} - 2\eta\beta_2 d \cdot \frac{17\delta_{\max}^2}{\eta\beta_2\gamma} \\
&\geq \frac{3d_h}{4} - \frac{34 d \delta_{\max}^2}{\gamma} \\
&\geq \frac{3d_h}{4} - \frac{34 d * 9 d_h \log(4d(2d+1)/\delta)}{\frac{1}{2} d_h} \\
&\geq \frac{d_h}{2}
\end{aligned}$$

The third inequality is by $\delta_{\max} = 3\sqrt{d_h \log(4d(2d+1)/\delta)}$ and $\gamma = \frac{1}{2}d_h$. The last inequality follows from $d_h = \widetilde{\Omega}(d^2) \geq 2448 d \log(4d(2d+1)/\delta)$.

Upper bound of $\boldsymbol{b}^\top(T+1)\boldsymbol{b}(T+1)$

$$\begin{aligned}
\boldsymbol{b}^\top(T+1)\boldsymbol{b}(T+1) &= \underbrace{\boldsymbol{b}^\top(0)\boldsymbol{b}(0)}_{\leq \frac{5d_h}{4}} - 2\eta\beta_2 \sum_{k=1}^{d} \underbrace{\sum_{s=0}^{T} \big(\boldsymbol{c}_k^\top(s)\boldsymbol{b}(s)\big)^2}_{=\text{term IV}} + \eta^2 \underbrace{\sum_{s=0}^{T} \Big\|\bar{\boldsymbol{b}}(s)\Big\|_2^2}_{\text{term VI}} \\
&\leq \frac{5d_h}{4} + 2\eta\beta_2 d \cdot \frac{17\delta_{\max}^2}{\eta\beta_2\gamma} + \eta^2 \cdot \Big(\frac{16 d_h \beta_2 d^2 \delta_{\max}^2 \exp(\eta\beta_2\gamma)}{\eta\gamma} + \frac{2 d_h d^2 \delta_{\max}^2 \exp(\eta\beta_1\gamma)}{\eta\beta_1\gamma}\Big) \\
&\leq \frac{5d_h}{4} + \frac{34 d \delta_{\max}^2}{\gamma} + \frac{2 * 16 \eta d_h \beta_2 d^2 \delta_{\max}^2}{\gamma} + \frac{2 * 2\eta d_h d^2 \delta_{\max}^2}{\beta_1\gamma} \\
&\leq \frac{5d_h}{4} + \frac{34 d * 9 d_h \log(4d(2d+1)/\delta)}{\frac{1}{2}d_h} + \frac{32\eta d_h \beta_2 d^2 * 9 d_h \log(4d(2d+1)/\delta)}{\frac{1}{2}d_h} \\
&+ \frac{4\eta d_h d^2 * 9 d_h \log(4d(2d+1)/\delta)}{\beta_1 \frac{1}{2} d_h} \\
&\leq 2d_h
\end{aligned}$$

The second inequality is by $\exp(\eta\beta_1\gamma) \leq \exp(\eta\beta_2\gamma) \leq 2$ The third inequality is by $\delta_{\max} = 3\sqrt{d_h \log(4d(2d+1)/\delta)}$ and $\gamma = \frac{1}{2}d_h$. The last inequality follows from

$$\begin{aligned}
&d_h = \widetilde{\Omega}(d^2) \\
&\geq 816 d \log(4d(2d+1)/\delta) + 768 \ln 2 d^2 \log(4d(2d+1)/\delta) + 48 \log(4d(2d+1)/\delta)/\beta_1 \\
&\geq 816 d \log(4d(2d+1)/\delta) + 768\eta\beta_2 d_h d^2 \log(4d(2d+1)/\delta) + 96\eta d_h d^2 \log(4d(2d+1)/\delta)/\beta_1
\end{aligned}$$

C.5 Proof of claim B.2

This Section presents the exponential decay bounds for terms $\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_i(T+1)\big)$, $\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_j(T+1)$ and $\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}(T+1)$, establishing the property $\mathcal{B}(T+1)$.

Bound of $\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_i(T+1)\big)$

Recall the following equation from lemma A.7.

$$\begin{aligned}
\boldsymbol{c}_i^\top(t+1)\boldsymbol{b}_i(t+1) = \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t) + \eta\Big(&\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t) - \beta_1 \sum_{k\neq i}^{d} \boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\cdot \boldsymbol{b}_k^\top(t)\boldsymbol{b}_i(t) \\
&- \beta_2 \boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\cdot \boldsymbol{b}_i^\top(t)\boldsymbol{b}(t) + \big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t) \\
&- \beta_1 \sum_{k\neq i}^{d} \boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\cdot \boldsymbol{c}_i^\top(t)\boldsymbol{c}_k(t)\Big) + \eta^2 \bar{\boldsymbol{c}}_i^\top(t)\bar{\boldsymbol{b}}_i(t)
\end{aligned}$$

Based on the above equation, we have:

$$\begin{aligned}
&\Big|\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_i(T+1)\big)\Big| = \Big|\underline{\beta_3 - \beta_1 \boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)} \\
&\quad - \eta\beta_1\Big(\underline{\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)}\boldsymbol{b}_i^\top(T)\boldsymbol{b}_i(T) - \beta_1 \sum_{k\neq i}^{d} \boldsymbol{c}_i^\top(T)\boldsymbol{b}_k(T)\cdot \boldsymbol{b}_k^\top(T)\boldsymbol{b}_i(T) \\
&\quad - \beta_2 \boldsymbol{c}_i^\top(T)\boldsymbol{b}(T)\cdot \boldsymbol{b}_i^\top(T)\boldsymbol{b}(T) + \underline{\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)}\boldsymbol{c}_i^\top(T)\boldsymbol{c}_i(T) \\
&\quad - \beta_1 \sum_{k\neq i}^{d} \boldsymbol{c}_k^\top(T)\boldsymbol{b}_i(T)\cdot \boldsymbol{c}_i^\top(T)\boldsymbol{c}_k(T)\Big) - \eta^2\beta_1 \bar{\boldsymbol{c}}_i^\top(T)\bar{\boldsymbol{b}}_i(T)\Big| \qquad (31) \\
&= \Big|\Big(1 - \eta\beta_1\big(\boldsymbol{b}_i^\top(T)\boldsymbol{b}_i(T) + \boldsymbol{c}_i^\top(T)\boldsymbol{c}_i(T)\big)\Big)\underline{\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)} \\
&\quad + \eta\beta_1^2 \sum_{k\neq i}^{d} \boldsymbol{c}_i^\top(T)\boldsymbol{b}_k(T)\cdot \boldsymbol{b}_k^\top(T)\boldsymbol{b}_i(T) + \eta\beta_1^2 \sum_{k\neq i}^{d} \boldsymbol{c}_k^\top(T)\boldsymbol{b}_i(T)\cdot \boldsymbol{c}_i^\top(T)\boldsymbol{c}_k(T) \\
&\quad + \eta\beta_1\beta_2 \boldsymbol{c}_i^\top(T)\boldsymbol{b}(T)\cdot \boldsymbol{b}_i^\top(T)\boldsymbol{b}(T) - \eta^2\beta_1 \bar{\boldsymbol{c}}_i^\top(T)\bar{\boldsymbol{b}}_i(T)\Big|
\end{aligned}$$

The term $\beta_3 - \beta_1 \boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)$ is highlighted with underline, and we collect its *negative feedback* terms together. The factor $\Big(1-\eta\beta_1\big(\boldsymbol{b}_i^\top(T)\boldsymbol{b}_i(T)+\boldsymbol{c}_i^\top(T)\boldsymbol{c}_i(T)\big)\Big) \le 1$ will drive $\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_i(T+1)\big)$ to converge to zero.

By Recurring (Eq. (31)) from 0 to T, we have:

$$
\begin{aligned}
&\Big|\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_i(T+1)\big)\Big| \\
&= \Big|\prod_{s=0}^{T}\Big(1-\eta\beta_1\big(\boldsymbol{b}_i^\top(s)\boldsymbol{b}_i(s)+\boldsymbol{c}_i^\top(s)\boldsymbol{c}_i(s)\big)\Big)\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(0)\boldsymbol{b}_i(0)\big) \\
&\quad + \sum_{s=0}^{T}\prod_{s'=s+1}^{T}\Big(1-\eta\beta_1\big(\boldsymbol{b}_i^\top(s')\boldsymbol{b}_i(s')+\boldsymbol{c}_i^\top(s')\boldsymbol{c}_i(s')\big)\Big) \\
&\quad \cdot \Big(\underbrace{\eta\beta_1^2\sum_{k\neq i}^{d}\boldsymbol{c}_i^\top(s)\boldsymbol{b}_k(s)\cdot\boldsymbol{b}_k^\top(s)\boldsymbol{b}_i(s) + \eta\beta_1^2\sum_{k\neq i}^{d}\boldsymbol{c}_k^\top(s)\boldsymbol{b}_i(s)\cdot\boldsymbol{c}_i^\top(s)\boldsymbol{c}_k(s)}_{\spadesuit} \\
&\quad + \underbrace{\eta\beta_1\beta_2\boldsymbol{c}_i^\top(s)\boldsymbol{b}(s)\cdot\boldsymbol{b}_i^\top(s)\boldsymbol{b}(s)}_{\clubsuit} - \underbrace{\eta^2\beta_1\bar{\boldsymbol{c}}_i^\top(s)\bar{\boldsymbol{b}}_i(s)}_{\diamondsuit}\Big)\Big|
\end{aligned}
\tag{32}
$$

Here $\prod_{s=0}^{T}\Big(1-\eta\beta_1\big(\boldsymbol{b}_i^\top(s)\boldsymbol{b}_i(s)+\boldsymbol{c}_i^\top(s)\boldsymbol{c}_i(s)\big)\Big) \le (1-2\eta\beta_1\gamma)^{T+1}$ since $\gamma \le \boldsymbol{b}_i^\top(s)\boldsymbol{b}_i(s), \boldsymbol{c}_i^\top(s)\boldsymbol{c}_i(s)$. Besides, from property $\mathcal{B}(0),\dots,\mathcal{B}(T)$ and lemma C.7 we know that $\boldsymbol{c}_i^\top(s)\boldsymbol{b}_k(s)$, $\boldsymbol{c}_i^\top(s)\boldsymbol{b}(s)$ and $\bar{\boldsymbol{c}}_i^\top(s)\bar{\boldsymbol{b}}_i(s)$ have bounds with exponential decreasing rate. Therefore, it is easy to derive an exponential decreasing upper bound for $\Big|\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_i(T+1)\big)\Big|$.

By substituting the bounds of $\boldsymbol{c}_i^\top(s)\boldsymbol{b}_k(s)$, $\boldsymbol{c}_i^\top(s)\boldsymbol{b}(s)$, $\bar{\boldsymbol{c}}_i^\top(s)\bar{\boldsymbol{b}}_i(s)$, $\boldsymbol{b}_k^\top(s)\boldsymbol{b}_i(s)$, $\boldsymbol{c}_i^\top(s)\boldsymbol{c}_k(s)$ and $\boldsymbol{b}_i^\top(s)\boldsymbol{b}(s)$, we have:

$$
\begin{aligned}
&\Big|\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_i(T+1)\big)\Big| \\
&\le (1-2\eta\beta_1\gamma)^{T+1}\Big|\beta_3 - \beta_1\boldsymbol{c}_i^\top(0)\boldsymbol{b}_i(0)\Big| \\
&\quad + \sum_{s=0}^{T}(1-2\eta\beta_1\gamma)^{T-s}\cdot\Big(\underbrace{2\eta\beta_1^2(d-1)\cdot 2\delta(s)^2\exp(-\eta\beta_1\gamma s)}_{\spadesuit} \\
&\quad \underbrace{+\eta\beta_1\beta_2\cdot\big(2\delta(s)^2\exp(-\eta\beta_2\gamma s) + \frac{\delta(s)^2}{\beta_2}\exp(-\eta\beta_1\gamma s)\big)}_{\clubsuit} \\
&\quad + \underbrace{\eta^2\beta_1\big(8d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t) + 8\beta_2 d_h d\delta(t)^2\exp(-\eta\beta_2\gamma t)\big)}_{\diamondsuit}\Big)
\end{aligned}
\tag{33}
$$

The notations $\spadesuit$, $\clubsuit$ and $\diamondsuit$ highlight the corresponding terms between (Eq. (32)) and (Eq. (33)) for refference.

We further have the following:

$$
\begin{aligned}
&\Big|\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_i(T+1)\big)\Big| \\
&\le (1-2\eta\beta_1\gamma)^{T+1}\Big|\beta_3 - \beta_1 \boldsymbol{c}_i^\top(0)\boldsymbol{b}_i(0)\Big| \\
&\quad + \sum_{s=0}^{T}(1-2\eta\beta_1\gamma)^{T-s}\cdot\Big(\underbrace{2\eta\beta_1^2(d-1)\cdot 2\delta(s)^2\exp(-\eta\beta_1\gamma s)}_{\spadesuit} \\
&\quad + \underbrace{\eta\beta_1\beta_2\cdot\big(2\delta(s)^2\exp(-\eta\beta_2\gamma s) + \frac{\delta(s)^2}{\beta_2}\exp(-\eta\beta_1\gamma s)\big)}_{\clubsuit} \\
&\quad + \underbrace{\eta^2\beta_1\big(8d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t) + 8\beta_2 d_h d\delta(t)^2\exp(-\eta\beta_2\gamma t)\big)}_{\diamondsuit}\Big) \\
&\le \exp(-2\eta\beta_1\gamma(T+1))\Big|\beta_3 - \beta_1\boldsymbol{c}_i^\top(0)\boldsymbol{b}_i(0)\Big| \\
&\quad + \sum_{s=0}^{T}\exp(2\eta\beta_1\gamma(s-T))\cdot\Big(\big(4\eta\beta_1^2(d-1)\delta(s)^2 + \eta\beta_1\delta(s)^2 + 8\eta^2\beta_1 d_h d^2\delta(t)^2\big)\exp(-\eta\beta_1\gamma s) \\
&\quad + \big(2\eta\beta_1\beta_2\delta(s)^2 + 8\eta^2\beta_1\beta_2 d_h d\delta(s)^2\big)\exp(-\eta\beta_2\gamma s)\Big) \\
&\le (\beta_3 + \beta_1\delta(0))\exp(-2\eta\beta_1\gamma(T+1)) \\
&\quad + \Big(4\eta\beta_1^2(d-1)\delta(T)^2 + \eta\beta_1\delta(T)^2 + 8\eta^2\beta_1 d_h d^2\delta(T)^2\Big)\cdot\frac{2}{\eta\beta_1\gamma}\exp(-\eta\beta_1\gamma(T+1)) \\
&\quad + \Big(2\eta\beta_1\beta_2\delta(T)^2 + 8\eta^2\beta_1\beta_2 d_h d\delta(T)^2\Big)\cdot\frac{3}{\eta\beta_2\gamma}\exp(-\eta\beta_1\gamma(T+1)) \\
&\le \Big(\frac{\beta_3}{\delta(0)} + \beta_1 + \frac{8\beta_1(d-1)\delta(T)}{\gamma} + \frac{2\delta(T)}{\gamma} + \frac{16\eta d_h d^2\delta(T)}{\gamma} + \frac{6\beta_1\delta(T)}{\gamma} + \frac{24\eta\beta_1 d_h d\delta(T)}{\gamma}\Big) \\
&\quad \cdot\delta(T)\cdot\exp(-\eta\beta_1\gamma(T+1)) \\
&\le \delta(T)\exp(-\eta\beta_1\gamma(T+1))
\end{aligned}
\tag{34}
$$

The second inequality is derived by factoring out the factors $\exp(-\eta\beta_1\gamma s)$ and $\exp(-\eta\beta_2\gamma s)$. The third inequality is due to $\sum_{s=0}^{T}\exp(2\eta\beta_1\gamma(s-T))\cdot\exp(-\eta\beta_1\gamma s) \le \frac{2}{\eta\beta_1\gamma}\exp(-\eta\beta_1\gamma(T+1))$ and $\sum_{s=0}^{T}\exp(2\eta\beta_1\gamma(s-T))\cdot\exp(-\eta\beta_2\gamma s) \le \frac{3}{\eta\beta_2\gamma}\exp(-\eta\beta_1\gamma(T+1))$ in lemma C.5. The fourth inequality is by $\delta(0)\le\delta(T)$, $\exp(-2\eta\beta_1\gamma(T+1)) \le \exp(-\eta\beta_1\gamma(T+1))$, and we consider $\beta_3 = \frac{\beta_3}{\delta(0)}\cdot\delta(0) \le \frac{\beta_3}{\delta(0)}\cdot\delta(T)$. The fifth inequality is by proving $\Big(\frac{\beta_3}{\delta(0)} + \beta_1 + \frac{8\beta_1(d-1)\delta(T)}{\gamma} + \frac{2\delta(T)}{\gamma} + \frac{16\eta d_h d^2\delta(T)}{\gamma} + \frac{6\beta_1\delta(T)}{\gamma} + \frac{24\eta\beta_1 d_h d\delta(T)}{\gamma}\Big) \le 1$ as follows:

$$
\begin{aligned}
&\frac{\beta_3}{\delta(0)} + \beta_1 + \frac{8\beta_1(d-1)\delta(T)}{\gamma} + \frac{2\delta(T)}{\gamma} + \frac{16\eta d_h d^2\delta(T)}{\gamma} + \frac{6\beta_1\delta(T)}{\gamma} + \frac{24\eta\beta_1 d_h d\delta(T)}{\gamma} \\
&\le \frac{\beta_3}{\delta(0)} + \frac{3}{4} + \frac{\delta(T)}{\gamma}\cdot\Big(8\beta_1(d-1) + 2 + 16\eta d_h d^2 + 6\beta_1 + 24\eta\beta_1 d_h d\Big) \\
&\le \frac{\beta_3}{2\sqrt{d_h\log(4d(2d+1)/\delta)}} + \frac{3}{4} \\
&\quad + \frac{3\sqrt{d_h\log(4d(2d+1)/\delta)}}{\frac{1}{2}d_h}\cdot\Big(8\beta_1(d-1) + 2 + 16\eta d_h d^2 + 6\beta_1 + 24\eta\beta_1 d_h d\Big) \\
&\le 1
\end{aligned}
$$

The first inequality is by $\beta_1 \le \frac{3}{4}$. The second inequality is by $\delta(0) \ge 2\sqrt{d_h\log(4d(2d+1)/\delta)}$, $\delta(T) \le 3\sqrt{d_h\log(4d(2d+1)/\delta)}$ and $\gamma = \frac{1}{2}d_h$. The last inequality hold as long as $d_h =$

$\widetilde{\Omega}(d^2) \geq \Big(\frac{1}{\sqrt{\log(4d(2d+1)/\delta)}} + 24\sqrt{\log(4d(2d+1)/\delta)}\big(8\beta_1(d-1)+2+8+6\beta_1+12\eta\beta_1/d\big)\Big)^2 \geq \Big(\frac{1}{\sqrt{\log(4d(2d+1)/\delta)}} + 24\sqrt{\log(4d(2d+1)/\delta)}\big(8\beta_1(d-1)+2+16\eta d_h d^2+6\beta_1+24\eta\beta_1 d_h d\big)\Big)^2$.
Therefore, we have

$$\Big|\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_i(T+1)\big)\Big| \leq \delta(T)\exp(-\eta\beta_1\gamma(T+1)) \leq \delta(T+1)\exp(-\eta\beta_1\gamma(T+1)) \quad (35)$$

where the last inequality is by $\delta(T) \leq \delta(T+1)$.

The proof for the bounds of $\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_j(T+1)$ and $\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}(T+1)$ are similar to that of $\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_i(T+1)\big)$. We presents the calculation as follows.

Bound of $\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_j(T+1)$

$$\begin{aligned}
&\Big|\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}_j(T+1)\Big| \\
=&\Big|\boldsymbol{c}_i^\top(T)\boldsymbol{b}_j(T)+\eta\Big(\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)\boldsymbol{b}_i^\top(T)\boldsymbol{b}_j(T)-\beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_i^\top(T)\boldsymbol{b}_k(T)\cdot\boldsymbol{b}_k^\top(T)\boldsymbol{b}_j(T) \\
&-\beta_2\boldsymbol{c}_i^\top(T)\boldsymbol{b}(T)\cdot\boldsymbol{b}_j^\top(T)\boldsymbol{b}(T)+\big(\beta_3-\beta_1\boldsymbol{c}_j^\top(T)\boldsymbol{b}_j(T)\big)\boldsymbol{c}_i^\top(T)\boldsymbol{c}_j(T) \\
&-\beta_1\sum_{k\neq j}^{d}\boldsymbol{c}_k^\top(T)\boldsymbol{b}_j(T)\cdot\boldsymbol{c}_i^\top(T)\boldsymbol{c}_k(T)\Big)+\eta^2\bar{\boldsymbol{c}}_i^\top(T)\bar{\boldsymbol{b}}_j(T)\Big| \\
=&\Big|\Big(1-\eta\beta_1\big(\boldsymbol{c}_i^\top(T)\boldsymbol{c}_i(T)+\boldsymbol{b}_j^\top(T)\boldsymbol{b}_j(T)\big)\Big)\boldsymbol{c}_i^\top(T)\boldsymbol{b}_j(T) \\
&+\eta\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)\boldsymbol{b}_i^\top(T)\boldsymbol{b}_j(T)-\eta\beta_1\sum_{k\neq i,k\neq j}^{d}\boldsymbol{c}_i^\top(T)\boldsymbol{b}_k(T)\cdot\boldsymbol{b}_k^\top(T)\boldsymbol{b}_j(T) \\
&-\eta\beta_2\boldsymbol{c}_i^\top(T)\boldsymbol{b}(T)\cdot\boldsymbol{b}_j^\top(T)\boldsymbol{b}(T)+\eta\big(\beta_3-\beta_1\boldsymbol{c}_j^\top(T)\boldsymbol{b}_j(T)\big)\boldsymbol{c}_i^\top(T)\boldsymbol{c}_j(T) \\
&-\eta\beta_1\sum_{k\neq i,k\neq j}^{d}\boldsymbol{c}_k^\top(T)\boldsymbol{b}_j(T)\cdot\boldsymbol{c}_i^\top(T)\boldsymbol{c}_k(T)+\eta^2\bar{\boldsymbol{c}}_i^\top(T)\bar{\boldsymbol{b}}_j(T)\Big| \\
=&\Big|\prod_{s=0}^{T}\Big(1-\eta\beta_1\big(\boldsymbol{c}_i^\top(s)\boldsymbol{c}_i(s)+\boldsymbol{b}_j^\top(s)\boldsymbol{b}_j(s)\big)\Big)\boldsymbol{c}_i^\top(0)\boldsymbol{b}_j(0) \\
&+\sum_{s=0}^{T}\prod_{s'=s+1}^{T}\Big(1-\eta\beta_1\big(\boldsymbol{c}_i^\top(s')\boldsymbol{c}_i(s')+\boldsymbol{b}_j^\top(s')\boldsymbol{b}_j(s')\big)\Big) \\
&\cdot\Big(\underbrace{\eta\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\big)\boldsymbol{b}_i^\top(s)\boldsymbol{b}_j(s)+\eta\big(\beta_3-\beta_1\boldsymbol{c}_j^\top(s)\boldsymbol{b}_j(s)\big)\boldsymbol{c}_i^\top(s)\boldsymbol{c}_j(s)}_{\spadesuit} \\
&\underbrace{-\eta\beta_1\sum_{k\neq i,k\neq j}^{d}\boldsymbol{c}_i^\top(s)\boldsymbol{b}_k(s)\cdot\boldsymbol{b}_k^\top(s)\boldsymbol{b}_j(s)-\eta\beta_1\sum_{k\neq i,k\neq j}^{d}\boldsymbol{c}_k^\top(s)\boldsymbol{b}_j(s)\cdot\boldsymbol{c}_i^\top(s)\boldsymbol{c}_k(s)}_{\clubsuit} \\
&-\underbrace{\eta\beta_2\boldsymbol{c}_i^\top(s)\boldsymbol{b}(s)\cdot\boldsymbol{b}_j^\top(s)\boldsymbol{b}(s)}_{\diamondsuit}+\underbrace{\eta^2\bar{\boldsymbol{c}}_i^\top(s)\bar{\boldsymbol{b}}_j(s)}_{\heartsuit}\Big)\Big| \\
\leq&(1-2\eta\beta_1\gamma)^{T+1}\Big|\boldsymbol{c}_i^\top(0)\boldsymbol{b}_j(0)\Big| \\
&+\sum_{s=0}^{T}(1-2\eta\beta_1\gamma)^{T-s}\big(\underbrace{2\eta\delta(s)^2\exp(-\eta\beta_1\gamma s)}_{\spadesuit} \\
&+\underbrace{2\eta\beta_1(d-2)\cdot 2\delta(s)^2\exp(-\eta\beta_1\gamma s)}_{\clubsuit}
\end{aligned}$$

$$
\begin{aligned}
&+ \underbrace{\eta\beta_2 \cdot \big(2\delta(s)^2 \exp(-\eta\beta_2\gamma s) + \frac{\delta(s)^2}{\beta_2} \exp(-\eta\beta_1\gamma s)\big)}_{\diamondsuit} \\
&+ \underbrace{\eta^2 \big(8d_h d^2 \delta(t)^2 \exp(-\eta\beta_1\gamma t) + 8\beta_2 d_h d\delta(t)^2 \exp(-\eta\beta_2\gamma t)\big)}_{\heartsuit}\big) \\
&\le \exp(-2\eta\beta_1\gamma(T+1))\Big|\boldsymbol{c}_i^\top(0)\boldsymbol{b}_j(0)\Big| \\
&\quad + \sum_{s=0}^{T} \exp(2\eta\beta_1\gamma(s-T))\Big(\big(2\eta\delta(s)^2 + 4\eta\beta_1(d-2)\delta(s)^2 + \eta\delta(s)^2 + 8\eta^2 d_h d^2\delta(s)^2\big)\exp(-\eta\beta_1\gamma s) \\
&\quad + \big(2\eta\beta_2\delta(s)^2 + 8\eta^2\beta_2 d_h d\delta(s)^2\big)\exp(-\eta\beta_2\gamma s)\Big) \\
&\le \delta(T)\exp(-\eta\beta_1\gamma(T+1)) \\
&\quad + \Big(2\eta\delta(T)^2 + 4\eta\beta_1(d-2)\delta(T)^2 + \eta\delta(T)^2 + 8\eta^2 d_h d^2\delta(T)^2\Big)\cdot\frac{2}{\eta\beta_1\gamma}\exp(-\eta\beta_1\gamma(T+1)) \\
&\quad + \Big(2\eta\beta_2\delta(T)^2 + 8\eta^2\beta_2 d_h d\delta(T)^2\Big)\cdot\frac{3}{\eta\beta_2\gamma}\exp(-\eta\beta_1\gamma(T+1)) \\
&= \Big(1 + \frac{4\delta(T)}{\beta_1\gamma} + \frac{8(d-2)\delta(T)}{\gamma} + \frac{2\delta(T)}{\beta_1\gamma} + \frac{16\eta d_h d^2\delta(T)}{\beta_1\gamma} + \frac{6\delta(T)}{\gamma} + \frac{24\eta d_h d\delta(T)}{\gamma}\Big) \\
&\quad \cdot \delta(T)\exp(-\eta\beta_1\gamma(T+1)) \\
&\le 2\delta(T)\exp(-\eta\beta_1\gamma(T+1)) \\
&\le 2\delta(T+1)\exp(-\eta\beta_1\gamma(T+1)) \qquad (36)
\end{aligned}
$$

This bound requires $\frac{\delta(T)}{\gamma}\Big(\frac{4}{\beta_1} + 8(d-2) + \frac{2}{\beta_1} + \frac{16\eta d_h d^2}{\beta_1} + 6 + 24\eta d_h d\Big) \le 1$, which can be verified by $d_h = \widetilde{\Omega}(d^2) \ge 36\log(4d(2d+1)/\delta)\Big(\frac{4}{\beta_1} + 8(d-2) + \frac{2}{\beta_1} + \frac{2}{\beta_1} + 6 + \frac{12}{d}\Big)^2 \ge 36\log(4d(2d+1)/\delta)\Big(\frac{4}{\beta_1} + 8(d-2) + \frac{2}{\beta_1} + \frac{16\eta d_h d^2}{\beta_1} + 6 + 24\eta d_h d\Big)^2$.

Bound of $\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}(T+1)$

$$
\begin{aligned}
&\Big|\boldsymbol{c}_i^\top(T+1)\boldsymbol{b}(T+1)\Big| \\
&= \Big|\boldsymbol{c}_i^\top(T)\boldsymbol{b}(T) + \eta\Big(\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)\boldsymbol{b}_i^\top(T)\boldsymbol{b}(T) - \beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_i^\top(T)\boldsymbol{b}_k(T)\cdot\boldsymbol{b}_k^\top(T)\boldsymbol{b}(T) \\
&\quad - \beta_2\boldsymbol{c}_i^\top(T)\boldsymbol{b}(T)\cdot\boldsymbol{b}^\top(T)\boldsymbol{b}(T) - \beta_2\sum_{k=1}^{d}\boldsymbol{c}_k^\top(T)\boldsymbol{b}(T)\cdot\boldsymbol{c}_k^\top(T)\boldsymbol{c}_i(T)\Big) + \eta^2\bar{\boldsymbol{c}}_i^\top(T)\bar{\boldsymbol{b}}(T)\Big| \\
&= \Big|\Big(1 - \eta\beta_2\big(\boldsymbol{b}^\top(T)\boldsymbol{b}(T) + \boldsymbol{c}_i^\top(T)\boldsymbol{c}_i(T)\big)\Big)\boldsymbol{c}_i^\top(T)\boldsymbol{b}(T) \\
&\quad + \eta\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)\boldsymbol{b}_i^\top(T)\boldsymbol{b}(T) - \eta\beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_i^\top(T)\boldsymbol{b}_k(T)\cdot\boldsymbol{b}_k^\top(T)\boldsymbol{b}(T) \\
&\quad - \eta\beta_2\sum_{k\neq i}^{d}\boldsymbol{c}_k^\top(T)\boldsymbol{b}(T)\cdot\boldsymbol{c}_k^\top(T)\boldsymbol{c}_i(T) + \eta^2\bar{\boldsymbol{c}}_i^\top(T)\bar{\boldsymbol{b}}(T)\Big| \\
&= \Big|\prod_{s=0}^{T}\Big(1 - \eta\beta_2\big(\boldsymbol{b}^\top(s)\boldsymbol{b}(s) + \boldsymbol{c}_i^\top(s)\boldsymbol{c}_i(s)\big)\Big)\boldsymbol{c}_i^\top(0)\boldsymbol{b}(0) \\
&\quad + \sum_{s=0}^{T}\prod_{s'=s+1}^{T}\Big(1 - \eta\beta_2\big(\boldsymbol{b}^\top(s')\boldsymbol{b}(s') + \boldsymbol{c}_i^\top(s')\boldsymbol{c}_i(s')\big)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&\cdot \Big(\underbrace{\eta\big(\beta_3 - \beta_1 \boldsymbol{c}_i^\top(s)\boldsymbol{b}_i(s)\big)\boldsymbol{b}_i^\top(s)\boldsymbol{b}(s)}_{\spadesuit} - \underbrace{\eta\beta_1 \sum_{k\neq i}^{d} \boldsymbol{c}_i^\top(s)\boldsymbol{b}_k(s)\cdot \boldsymbol{b}_k^\top(s)\boldsymbol{b}(s)}_{\clubsuit} \\
&\quad \underbrace{- \eta\beta_2 \sum_{k\neq i}^{d} \boldsymbol{c}_k^\top(s)\boldsymbol{b}(s)\cdot \boldsymbol{c}_k^\top(s)\boldsymbol{c}_i(s)}_{\diamondsuit} + \underbrace{\eta^2 \bar{\boldsymbol{c}}_i^\top(s)\bar{\boldsymbol{b}}(s)}_{\heartsuit} \Big)\Big| \\
&\leq (1-2\eta\beta_2\gamma)^{T+1}\Big|\boldsymbol{c}_i^\top(0)\boldsymbol{b}(0)\Big| \\
&\quad + \sum_{s=0}^{T}(1-2\eta\beta_2\gamma)^{T-s}\big(\underbrace{\eta\delta(s)^2\exp(-\eta\beta_1\gamma s)}_{\spadesuit} \\
&\quad + \underbrace{\eta\beta_1(d-1)\cdot 2\delta(s)^2\exp(-\eta\beta_1\gamma s)}_{\clubsuit} \\
&\quad + \underbrace{\eta\beta_2(d-1)\cdot\big(2\delta(s)^2\exp(-\eta\beta_2\gamma s) + \frac{\delta(s)^2}{\beta_2}\exp(-\eta\beta_1\gamma s)\big)}_{\diamondsuit} \\
&\quad + \underbrace{\eta^2\big(4d_h d^2\delta(s)^2\exp(-\eta\beta_1\gamma s) + 28\beta_2^2 d_h d\delta(s)^2\exp(-\eta\beta_2\gamma s)\big)}_{\heartsuit}\big) \\
&\leq \exp(-2\eta\beta_2\gamma(T+1))\Big|\boldsymbol{c}_i^\top(0)\boldsymbol{b}(0)\Big| \\
&\quad + \sum_{s=0}^{T}\exp(2\eta\beta_2\gamma(s-T))\Big(\big(\eta\delta(T)^2 + 2\eta\beta_1(d-1)\delta(T)^2 \\
&\quad + \eta(d-1)\delta(T)^2 + 4\eta^2 d_h d^2\delta(T)^2\big)\exp(-\eta\beta_1\gamma s) \\
&\quad + \big(2\eta\beta_2(d-1)\delta(T)^2 + 28\eta^2\beta_2^2 d_h d\delta(s)^2\big)\exp(-\eta\beta_2\gamma s)\Big) \\
&\leq \delta(T)\exp(-\eta\beta_2\gamma(T+1)) \\
&\quad + \Big(\eta\delta(T)^2 + 2\eta\beta_1(d-1)\delta(T)^2 + \eta(d-1)\delta(T)^2 + 4\eta^2 d_h d^2\delta(T)^2\Big)\cdot\frac{2}{\eta\beta_2\gamma}\exp(-\eta\beta_1\gamma(T+1)) \\
&\quad + \Big(2\eta\beta_2(d-1)\delta(T)^2 + 28\eta^2\beta_2^2 d_h d\delta(T)^2\Big)\cdot\frac{2}{\eta\beta_2\gamma}\exp(-\eta\beta_2\gamma(T+1)) \\
&= \delta(T)\Big(1 + \frac{4(d-1)\delta(T)}{\gamma} + \frac{56\eta\beta_2 d_h d\delta(T)}{\gamma}\Big)\exp(-\eta\beta_2\gamma(T+1)) \\
&\quad + \frac{\delta(T)}{\beta_2}\Big(\frac{2\delta(T)}{\gamma} + \frac{4\beta_1(d-1)\delta(T)}{\gamma} + \frac{2(d-1)\delta(T)}{\gamma} + \frac{8\eta d_h d^2\delta(T)}{\gamma}\Big)\exp(-\eta\beta_1\gamma(T+1)) \\
&\leq 2\delta(T)\exp(-\eta\beta_2\gamma(T+1)) + \frac{\delta(T)}{\beta_2}\exp(-\eta\beta_1\gamma(T+1)) \\
&\leq 2\delta(T+1)\exp(-\eta\beta_2\gamma(T+1)) + \frac{\delta(T+1)}{\beta_2}\exp(-\eta\beta_1\gamma(T+1)) \qquad (37)
\end{aligned}
$$

This bound requires $\frac{\delta(T)}{\gamma}\Big(4(d-1)+56\eta\beta_2 d_h d\Big) \leq 1$ and $\frac{\delta(T)}{\gamma}\Big(2+4\beta_1(d-1)+2(d-1)+8\eta d_h d^2\Big) \leq 1$, which can be verified by $d_h = \widetilde{\Omega}(d^2) \geq 36\log(4d(2d+1)/\delta)\Big(4(d-1)+56d\ln 2\Big)^2 \geq 36\log(4d(2d+1)/\delta)\Big(4(d-1)+56\eta\beta_2 d_h d\Big)^2$
and
$d_h = \widetilde{\Omega}(d^2) \geq 36\log(4d(2d+1)/\delta)\Big(2+4\beta_1(d-1)+2(d-1)+4\Big)^2 \geq 36\log(4d(2d+1)/\delta)\Big(2+4\beta_1(d-1)+2(d-1)+8\eta d_h d^2\Big)^2$.

Property $\mathcal{B}(T+1)$ is established by (Eq. (35)), (Eq. (36)) and (Eq. (37)).

C.5.1 Auxiliary lemma

Lemma C.5 *As long as* $2\eta\beta_1\gamma \le \ln 2$*, and* $2\eta\beta_2\gamma \le \ln 2$*, we have*

$$\sum_{s=0}^{T} \exp(2\eta\beta_1\gamma(s-T)) \cdot \exp(-\eta\beta_1\gamma s) \le \frac{2}{\eta\beta_1\gamma} \exp(-\eta\beta_1\gamma(T+1))$$

$$\sum_{s=0}^{T} \exp(2\eta\beta_1\gamma(s-T)) \cdot \exp(-\eta\beta_2\gamma s) \le \frac{3}{\eta\beta_2\gamma} \exp(-\eta\beta_1\gamma(T+1))$$

$$\sum_{s=0}^{T} \exp(2\eta\beta_2\gamma(s-T)) \cdot \exp(-\eta\beta_1\gamma s) \le \frac{2}{\eta\beta_2\gamma} \exp(-\eta\beta_1\gamma(T+1))$$

$$\sum_{s=0}^{T} \exp(2\eta\beta_2\gamma(s-T)) \cdot \exp(-\eta\beta_2\gamma s) \le \frac{2}{\eta\beta_2\gamma} \exp(-\eta\beta_2\gamma(T+1))$$

Proof of lemma C.5.

$$\begin{aligned}
&\sum_{s=0}^{T} \exp(2\eta\beta_1\gamma(s-T)) \cdot \exp(-\eta\beta_1\gamma s) \\
&= \sum_{s=0}^{T} \exp(\eta\beta_1\gamma s - 2\eta\beta_1\gamma T) \\
&\le \int_{0}^{T+1} \exp(\eta\beta_1\gamma s - 2\eta\beta_1\gamma T) ds \\
&\le \frac{1}{\eta\beta_1\gamma} \big(\exp(-\eta\beta_1\gamma(T-1)) - \exp(-2\eta\beta_1\gamma T) \big) \\
&\le \frac{1}{\eta\beta_1\gamma} \exp(-\eta\beta_1\gamma(T-1)) \\
&\le \frac{\exp(2\eta\beta_1\gamma)}{\eta\beta_1\gamma} \exp(-\eta\beta_1\gamma(T+1)) \\
&\le \frac{2}{\eta\beta_1\gamma} \exp(-\eta\beta_1\gamma(T+1))
\end{aligned}$$

The first inequality is due to $\exp(\eta\beta_1\gamma s)$ is monotone increasing. The last inequality is due to $2\eta\beta_1\gamma \le \ln 2$.

$$\begin{aligned}
&\sum_{s=0}^{T} \exp(2\eta\beta_1\gamma(s-T)) \cdot \exp(-\eta\beta_2\gamma s) \\
&= \sum_{s=0}^{T} \exp(-\eta(\beta_2 - 2\beta_1)\gamma s - 2\eta\beta_1\gamma T) \\
&\le \int_{-1}^{T} \exp(-\eta(\beta_2 - 2\beta_1)\gamma s - 2\eta\beta_1\gamma T) ds \\
&\le \frac{1}{\eta(\beta_2 - 2\beta_1)\gamma} \big(\exp(\eta(\beta_2 - 2\beta_1)\gamma - 2\eta\beta_1\gamma T) - \exp(-\eta\beta_2\gamma T) \big) \\
&\le \frac{1}{\eta(\beta_2 - 2\beta_1)\gamma} \exp(\eta(\beta_2 - 2\beta_1)\gamma - 2\eta\beta_1\gamma T) \\
&\le \frac{\exp(\eta\beta_2\gamma)}{\eta(\beta_2 - 2\beta_1)\gamma} \exp(-2\eta\beta_1\gamma(T+1))
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{2\exp(\eta\beta_2\gamma)}{\eta\beta_2\gamma}\exp(-2\eta\beta_1\gamma(T+1)) \\
&\leq \frac{3}{\eta\beta_2\gamma}\exp(-\eta\beta_1\gamma(T+1))
\end{aligned}
$$

The first inequality is due to $\exp(-\eta(\beta_2-2\beta_1)\gamma s)$ is monotone decreasing. The fourth inequality is due to $\beta_2 \geq 4\beta_1$, thus $\frac{1}{\beta_2-2\beta_1} \leq \frac{2}{\beta 2}$. The last inequality is due to $\eta\beta_2\gamma \leq (\ln 2)/2 \leq \ln(3/2)$.

$$
\begin{aligned}
&\sum_{s=0}^{T}\exp(2\eta\beta_2\gamma(s-T))\cdot\exp(-\eta\beta_1\gamma s) \\
&=\sum_{s=0}^{T}\exp(\eta(2\beta_2-\beta_1)\gamma s-2\eta\beta_2\gamma T) \\
&\leq \int_0^{T+1}\exp(\eta(2\beta_2-\beta_1)\gamma s-2\eta\beta_2\gamma T)ds \\
&\leq \frac{1}{\eta(2\beta_2-\beta_1)\gamma}\big(\exp(2\eta\beta_2\gamma-\eta\beta_1\gamma(T+1))-\exp(-2\eta\beta_2\gamma T)\big) \\
&\leq \frac{1}{\eta(2\beta_2-\beta_1)\gamma}\exp(2\eta\beta_2\gamma-\eta\beta_1\gamma(T+1)) \\
&\leq \frac{\exp(2\eta\beta_2\gamma)}{\eta(2\beta_2-\beta_1)\gamma}\exp(-\eta\beta_1\gamma(T+1)) \\
&\leq \frac{2}{\eta\beta_2\gamma}\exp(-\eta\beta_1\gamma(T+1))
\end{aligned}
$$

The first inequality is due to $\exp(\eta(2\beta_2-\beta_1)\gamma s)$ is monotone increasing. The last inequality is due to $\beta_2 \geq \beta_1$ and $2\eta\beta_2\gamma \leq \ln 2$.

The proof of $\sum_{s=0}^{T}\exp(2\eta\beta_2\gamma(s-T))\cdot\exp(-\eta\beta_2\gamma s) \leq \frac{2}{\eta\beta_2\gamma}\exp(-\eta\beta_2\gamma(T+1))$ is similar to $\sum_{s=0}^{T}\exp(2\eta\beta_1\gamma(s-T))\cdot\exp(-\eta\beta_1\gamma s) \leq \frac{2}{\eta\beta_1\gamma}\exp(-\eta\beta_1\gamma(T+1))$. Just replace β_1 with β_2, and consider $2\eta\beta_2\gamma \leq \ln 2$.

C.6 Proof of claim B.3

This Section presents the bounds for terms $\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}_j(T+1)$, $\boldsymbol{c}_i^\top(T+1)\boldsymbol{c}_j(T+1)$ and $\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}(T+1)$ with $i,j\in[1,d], i\neq j$, establishing the property $\mathcal{C}(T+1)$.

Recall the *Vector-coupled Dynamics* equations of $\boldsymbol{b}_i^\top(t+1)\boldsymbol{b}_j(t+1)$, $\boldsymbol{c}_i^\top(t+1)\boldsymbol{c}_j(t+1)$ and $\boldsymbol{b}_i^\top(t+1)\boldsymbol{b}(t+1)$ in lemma A.7:

$$
\begin{aligned}
&\boldsymbol{b}_i^\top(t+1)\boldsymbol{b}_j(t+1) \\
&=\boldsymbol{b}_i^\top(t)\boldsymbol{b}_j(t)+\eta\Big(2\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)+2\big(\beta_3-\beta_1\boldsymbol{c}_j^\top(t)\boldsymbol{b}_j(t)\big)\boldsymbol{c}_j^\top(t)\boldsymbol{b}_i(t) \\
&-\beta_3\big(\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)+\boldsymbol{c}_j^\top(t)\boldsymbol{b}_i(t)\big)-2\beta_1\sum_{k\neq i,k\neq j}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\cdot\boldsymbol{c}_k^\top(t)\boldsymbol{b}_j(t)\Big)+\eta^2\bar{\boldsymbol{b}}_i^\top(t)\bar{\boldsymbol{b}}_j(t)
\end{aligned}
\tag{38}
$$

$$
\begin{aligned}
&\boldsymbol{c}_i^\top(t+1)\boldsymbol{c}_j(t+1) \\
&=\boldsymbol{c}_i^\top(t)\boldsymbol{c}_j(t)+\eta\Big(2\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_j^\top(t)\boldsymbol{b}_i(t)+2\big(\beta_3-\beta_1\boldsymbol{c}_j^\top(t)\boldsymbol{b}_j(t)\big)\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t) \\
&-\beta_3\big(\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)+\boldsymbol{c}_j^\top(t)\boldsymbol{b}_i(t)\big)-2\beta_1\sum_{k\neq i,k\neq j}^{d}\boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\cdot\boldsymbol{c}_j^\top(t)\boldsymbol{b}_k(t) \\
&-2\beta_2\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{c}_j^\top(t)\boldsymbol{b}(t)\Big)+\eta^2\bar{\boldsymbol{c}}_i^\top(t)\bar{\boldsymbol{c}}_j(t)
\end{aligned}
\tag{39}
$$

$$
\begin{aligned}
&\boldsymbol{b}_i^\top(t+1)\boldsymbol{b}(t+1)\\
&=\boldsymbol{b}_i^\top(t)\boldsymbol{b}(t)+\eta\Big(\big(\beta_3-\beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)-\beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\cdot\boldsymbol{c}_k^\top(t)\boldsymbol{b}(t)\\
&-\beta_2\sum_{k=1}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\Big)+\eta^2\bar{\boldsymbol{b}}_i^\top(t)\bar{\boldsymbol{b}}(t)
\end{aligned}
\tag{40}
$$

To give bounds for the above three terms, we will use the following bounds from property $\mathcal{B}(t)$ and lemma C.7:

$$
|\beta_3-\beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)|\leq\delta(t)\exp(-\eta\beta_1\gamma t)
$$

$$
|\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)|\leq 2\delta(t)\exp(-\eta\beta_1\gamma t)
$$

$$
|\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)|\leq 2\delta(t)\exp(-\eta\beta_2\gamma t)+\frac{\delta(t)}{\beta_2}\exp(-\eta\beta_1\gamma t)
$$

$$
\left|\bar{\boldsymbol{b}}_i(t)^\top\bar{\boldsymbol{b}}_j(t)\right|\leq 8d_hd^2\delta(t)^2\exp(-\eta\beta_1\gamma t),
$$

$$
\left|\bar{\boldsymbol{c}}_i(t)^\top\bar{\boldsymbol{c}}_j(t)\right|\leq 8d_hd^2\delta(t)^2\exp(-\eta\beta_1\gamma t)+40\beta_2^2d_h\delta(t)^2\exp(-\eta\beta_2\gamma t),
$$

$$
\left|\bar{\boldsymbol{b}}_i^\top(t)\bar{\boldsymbol{b}}(t)\right|\leq 8\beta_2d_hd^2\delta(t)^2\exp(-\eta\beta_2\gamma t)+4d_hd^2\delta(t)^2\exp(-\eta\beta_1\gamma t)
$$

Besides, by $\left|\beta_3-\beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\right|\leq\delta(t)\exp(-\eta\beta_1\gamma t)$, we have:

$$
\begin{aligned}
\left|\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\right|&\leq\frac{\delta(t)\exp(-\eta\beta_1\gamma t)+\beta_3}{\beta_1}\\
&\leq\frac{4\delta(t)+2\alpha}{\alpha^2}\\
&\leq\frac{5\delta(t)}{\alpha^2}\\
&\leq 6\delta(t)
\end{aligned}
$$

The third inequality is by $\delta(s)\geq 2\sqrt{d_h\log(4d(2d+1)/\delta)}\geq 2\alpha=2\alpha=2\exp((-\ln 2)/N)$. For the last inequality, as long as $N\geq\frac{2\ln 2}{\ln 6-\ln 5}$, we have $\frac{5}{\alpha^2}\leq 6$.

We will provide the upper bounds for $\left|\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}_j(T+1)\right|$, $\left|\boldsymbol{c}_i^\top(T+1)\boldsymbol{c}_j(T+1)\right|$ and $\left|\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}(T+1)\right|$ by substituting the above bounds into (Eq. 38), (Eq. 39) and (Eq. 40) respectively.

Bound of $\left|\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}_j(T+1)\right|$

$$
\begin{aligned}
&\left|\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}_j(T+1)\right| \\
&= \Big|\boldsymbol{b}_i^\top(T)\boldsymbol{b}_j(T) + \eta\Big(2\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)\boldsymbol{c}_i^\top(T)\boldsymbol{b}_j(T) + 2\big(\beta_3 - \beta_1\boldsymbol{c}_j^\top(T)\boldsymbol{b}_j(T)\big)\boldsymbol{c}_j^\top(T)\boldsymbol{b}_i(T) \\
&- \beta_3\big(\boldsymbol{c}_i^\top(T)\boldsymbol{b}_j(T) + \boldsymbol{c}_j^\top(T)\boldsymbol{b}_i(T)\big) - 2\beta_1\sum_{k\neq i,k\neq j}^{d}\boldsymbol{c}_k^\top(T)\boldsymbol{b}_i(T)\cdot\boldsymbol{c}_k^\top(T)\boldsymbol{b}_j(T)\Big) + \eta^2\bar{\boldsymbol{b}}_i^\top(T)\bar{\boldsymbol{b}}_j(T)\Big| \\
&\leq \left|\boldsymbol{b}_i^\top(T)\boldsymbol{b}_j(T)\right| + 2\eta\left|\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)\boldsymbol{c}_i^\top(T)\boldsymbol{b}_j(T)\right| + 2\eta\left|\big(\beta_3 - \beta_1\boldsymbol{c}_j^\top(T)\boldsymbol{b}_j(T)\big)\boldsymbol{c}_j^\top(T)\boldsymbol{b}_i(T)\right| \\
&+ \eta\beta_3\left|\big(\boldsymbol{c}_i^\top(T)\boldsymbol{b}_j(T) + \boldsymbol{c}_j^\top(T)\boldsymbol{b}_i(T)\big)\right| + 2\eta\beta_1\sum_{k\neq i,k\neq j}^{d}\left|\boldsymbol{c}_k^\top(T)\boldsymbol{b}_i(T)\cdot\boldsymbol{c}_k^\top(T)\boldsymbol{b}_j(T)\right| + \eta^2\left|\bar{\boldsymbol{b}}_i^\top(T)\bar{\boldsymbol{b}}_j(T)\right| \\
&\leq \delta(T) + 4\eta\cdot\delta(T)\exp(-\eta\beta_1\gamma T)\cdot 2\delta(T)\exp(-\eta\beta_1\gamma T) + 2\eta\beta_3\cdot 2\delta(T)\exp(-\eta\beta_1\gamma T) \\
&+ 2\eta\beta_1(d-2)\cdot\Big(2\delta(T)\exp(-\eta\beta_1\gamma T)\Big)^2 + \eta^2\cdot 8d_h d^2\delta(T)^2\exp(-\eta\beta_1\gamma T) \\
&\leq \delta(T) + \delta(T)\Big(8\eta\delta(T)\exp(-\eta\beta_1\gamma T) + 4\eta\beta_3 + 8\eta\beta_1(d-2)\delta(T)\exp(-\eta\beta_1\gamma T) \\
&+ 8\eta^2 d_h d^2\delta(T)\Big)\cdot\exp(-\eta\beta_1\gamma T) \\
&\leq \delta(T) + \delta(T)\Big(8\eta\delta_{\max} + 4\eta\beta_3 + 8\eta\beta_1(d-2)\delta_{\max} + 8\eta^2 d_h d^2\delta_{\max}\Big)\exp(-\eta\beta_1\gamma T)
\end{aligned}
\tag{41}
$$

The first inequality is derived by triangle inequality. The second inequality is derived by $\left|\boldsymbol{b}_i^\top(T)\boldsymbol{b}_j(T)\right| \leq \delta(T)$ and substituting the bounds of $|\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)|$, $|\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)|$ and $\left|\bar{\boldsymbol{b}}_i(t)^\top\bar{\boldsymbol{b}}_j(t)\right|$. The third inequality is derived by factoring out the common factor $\delta(T)$. The last inequality is derived by $\delta(T) \leq \delta_{\max}$ and $\exp(-\eta\beta_1\gamma T) \leq 1$.

Bound of $\left|\boldsymbol{c}_i^\top(T+1)\boldsymbol{c}_j(T+1)\right|$

$$
\begin{aligned}
&\left|\boldsymbol{c}_i^\top(T+1)\boldsymbol{c}_j(T+1)\right| \\
&= \Big|\boldsymbol{c}_i^\top(T)\boldsymbol{c}_j(T) + \eta\Big(2\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)\boldsymbol{c}_j^\top(T)\boldsymbol{b}_i(T) + 2\big(\beta_3 - \beta_1\boldsymbol{c}_j^\top(T)\boldsymbol{b}_j(T)\big)\boldsymbol{c}_i^\top(T)\boldsymbol{b}_j(T) \\
&- \beta_3\big(\boldsymbol{c}_i^\top(T)\boldsymbol{b}_j(T) + \boldsymbol{c}_j^\top(T)\boldsymbol{b}_i(T)\big) - 2\beta_1\sum_{k\neq i,k\neq j}^{d}\boldsymbol{c}_i^\top(T)\boldsymbol{b}_k(T)\cdot\boldsymbol{c}_j^\top(T)\boldsymbol{b}_k(T) \\
&- 2\beta_2\boldsymbol{c}_i^\top(T)\boldsymbol{b}(T)\cdot\boldsymbol{c}_j^\top(T)\boldsymbol{b}(T)\Big) + \eta^2\bar{\boldsymbol{c}}_i^\top(T)\bar{\boldsymbol{c}}_j(T)\Big| \\
&\leq \left|\boldsymbol{c}_i^\top(T)\boldsymbol{c}_j(T)\right| + 2\eta\Big|\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)\boldsymbol{c}_j^\top(T)\boldsymbol{b}_i(T)\Big| + 2\eta\Big|\big(\beta_3 - \beta_1\boldsymbol{c}_j^\top(T)\boldsymbol{b}_j(T)\big)\boldsymbol{c}_i^\top(T)\boldsymbol{b}_j(T)\Big| \\
&+ \eta\beta_3\Big|\big(\boldsymbol{c}_i^\top(T)\boldsymbol{b}_j(T) + \boldsymbol{c}_j^\top(T)\boldsymbol{b}_i(T)\big)\Big| + 2\eta\beta_1\sum_{k\neq i,k\neq j}^{d}\left|\boldsymbol{c}_i^\top(T)\boldsymbol{b}_k(T)\cdot\boldsymbol{c}_j^\top(T)\boldsymbol{b}_k(T)\right| \\
&+ 2\eta\beta_2\Big|\boldsymbol{c}_i^\top(T)\boldsymbol{b}(T)\cdot\boldsymbol{c}_j^\top(T)\boldsymbol{b}(T)\Big| + \eta^2\Big|\bar{\boldsymbol{c}}_i^\top(T)\bar{\boldsymbol{c}}_j(T)\Big| \\
&\leq \delta(T) + 4\eta\cdot\delta(T)\exp(-\eta\beta_1\gamma T)\cdot 2\delta(T)\exp(-\eta\beta_1\gamma T) + 2\eta\beta_3\cdot 2\delta(T)\exp(-\eta\beta_1\gamma T) \\
&+ 2\eta\beta_1(d-2)\cdot\Big(2\delta(T)\exp(-\eta\beta_1\gamma T)\Big)^2 \\
&+ 2\eta\beta_2\Big(2\delta(T)\exp(-\eta\beta_2\gamma T) + \frac{\delta(T)}{\beta_2}\exp(-\eta\beta_1\gamma T)\Big)^2 \\
&+ \eta^2\cdot\Big(8d_h d^2\delta(T)^2\exp(-\eta\beta_1\gamma T) + 40\beta_2^2 d_h\delta(T)^2\exp(-\eta\beta_2\gamma T)\Big) \\
&\leq \delta(T) + \delta(T)\Big(8\eta\delta(T)\exp(-\eta\beta_1\gamma T) + 4\eta\beta_3 + 8\eta\beta_1(d-2)\delta(T)\exp(-\eta\beta_1\gamma T) + 8\eta^2 d_h d^2\delta(T) \\
&+ \frac{2\eta\delta(T)}{\beta_2}\exp(-\eta\beta_1\gamma T) + 8\eta\delta(T)\exp(-\eta\beta_2\gamma T)\Big)\cdot\exp(-\eta\beta_1\gamma T) \\
&+ \delta(T)\Big(8\eta\beta_2\delta(T)\exp(-\eta\beta_2\gamma T) + 40\eta^2\beta_2^2 d_h\delta(T)\Big)\cdot\exp(-\eta\beta_2\gamma T) \\
&\leq \delta(T) + \delta(T)\Big(16\eta\delta_{\max} + 4\eta\beta_3 + 8\eta\beta_1(d-2)\delta_{\max} + 8\eta^2 d_h d^2\delta_{\max} + \frac{2\eta\delta_{\max}}{\beta_2}\Big)\cdot\exp(-\eta\beta_1\gamma T) \\
&+ \delta(T)\Big(8\eta\beta_2\delta_{\max} + 40\eta^2\beta_2^2 d_h\delta_{\max}\Big)\cdot\exp(-\eta\beta_2\gamma T)
\end{aligned}
\tag{42}
$$

The first inequality is derived by triangle inequality. The second inequality is derived by $\left|\boldsymbol{b}_i^\top(T)\boldsymbol{b}_j(T)\right| \leq \delta(T)$ and substituting the bounds of $|\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)|$, $|\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)|$ and $\left|\bar{\boldsymbol{c}}_i(t)^\top\bar{\boldsymbol{c}}_j(t)\right|$. The third inequality is derived by factoring out the common factor $\delta(T)$. The last inequality is derived by $\delta(T) \leq \delta_{\max}$, $\exp(-\eta\beta_1\gamma T) \leq 1$ and $\exp(-\eta\beta_2\gamma T) \leq 1$.

Bound of $\left|\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}(T+1)\right|$

$$
\begin{aligned}
&\left|\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}(T+1)\right| \\
&= \Big|\boldsymbol{b}_i^\top(T)\boldsymbol{b}(T) + \eta\Big(\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)\boldsymbol{c}_i^\top(T)\boldsymbol{b}(T) - \beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_k^\top(T)\boldsymbol{b}_i(T)\cdot\boldsymbol{c}_k^\top(T)\boldsymbol{b}(T) \\
&- \beta_2\sum_{k=1}^{d}\boldsymbol{c}_k^\top(T)\boldsymbol{b}(T)\cdot\boldsymbol{c}_k^\top(T)\boldsymbol{b}_i(T)\Big) + \eta^2\bar{\boldsymbol{b}}_i^\top(T)\bar{\boldsymbol{b}}(T)\Big| \\
&= \Big|\boldsymbol{b}_i^\top(T)\boldsymbol{b}(T) + \eta\Big(\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)\boldsymbol{c}_i^\top(T)\boldsymbol{b}(T) - \beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_k^\top(T)\boldsymbol{b}_i(T)\cdot\boldsymbol{c}_k^\top(T)\boldsymbol{b}(T) \\
&- \beta_2\sum_{k\neq i}^{d}\boldsymbol{c}_k^\top(T)\boldsymbol{b}(T)\cdot\boldsymbol{c}_k^\top(T)\boldsymbol{b}_i(T) - \beta_2\boldsymbol{c}_i^\top(T)\boldsymbol{b}(T)\cdot\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\Big) + \eta^2\bar{\boldsymbol{b}}_i^\top(T)\bar{\boldsymbol{b}}(T)\Big| \\
&\leq \Big|\boldsymbol{b}_i^\top(T)\boldsymbol{b}(T)\Big| + \eta\Big|\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\big)\boldsymbol{c}_i^\top(T)\boldsymbol{b}(T)\Big| + \eta\beta_1\sum_{k\neq i}^{d}\Big|\boldsymbol{c}_k^\top(T)\boldsymbol{b}_i(T)\cdot\boldsymbol{c}_k^\top(T)\boldsymbol{b}(T)\Big| \\
&+ \eta\beta_2\sum_{k\neq i}^{d}\Big|\boldsymbol{c}_k^\top(T)\boldsymbol{b}(T)\cdot\boldsymbol{c}_k^\top(T)\boldsymbol{b}_i(T)\Big| + \eta\beta_2\Big|\boldsymbol{c}_i^\top(T)\boldsymbol{b}(T)\cdot\boldsymbol{c}_i^\top(T)\boldsymbol{b}_i(T)\Big| + \eta^2\Big|\bar{\boldsymbol{b}}_i^\top(T)\bar{\boldsymbol{b}}(T)\Big| \\
&\leq \delta(T) + \eta\cdot\delta(T)\exp(-\eta\beta_1\gamma T)\cdot\Big(2\delta(T)\exp(-\eta\beta_2\gamma T) + \frac{\delta(T)}{\beta_2}\exp(-\eta\beta_1\gamma T)\Big) \\
&+ \eta(\beta_1+\beta_2)(d-1)\cdot 2\delta(T)\exp(-\eta\beta_1\gamma T)\cdot\Big(2\delta(T)\exp(-\eta\beta_2\gamma T) + \frac{\delta(T)}{\beta_2}\exp(-\eta\beta_1\gamma T)\Big) \\
&+ \eta\beta_2\cdot 6\delta(T)\cdot\Big(2\delta(T)\exp(-\eta\beta_2\gamma T) + \frac{\delta(T)}{\beta_2}\exp(-\eta\beta_1\gamma T)\Big) \\
&+ \eta^2\cdot\Big(8\beta_2 d_h d^2\delta(T)^2\exp(-\eta\beta_2\gamma T) + 4d_h d^2\delta(T)^2\exp(-\eta\beta_1\gamma T)\Big) \\
&\leq \delta(T) + \delta(T)\Big(\eta\big(2(\beta_1+\beta_2)(d-1)+1\big)\cdot\frac{\delta(T)}{\beta_2}\exp(-\eta\beta_1\gamma T) \\
&+ 6\eta\beta_2\cdot\frac{\delta(T)}{\beta_2} + 4\eta^2 d_h d^2\delta(T)\Big)\cdot\exp(-\eta\beta_1\gamma T) \\
&+ \delta(T)\Big(2\eta\big(2(\beta_1+\beta_2)(d-1)+1\big)\delta(T)\exp(-\eta\beta_1\gamma T) \\
&+ 12\eta\beta_2\delta(T) + 8\eta^2\beta_2 d_h d^2\delta(T)\Big)\cdot\exp(-\eta\beta_2\gamma T) \\
&\leq \delta(T) + \delta(T)\Big(4\eta d\delta_{\max} + 4\eta^2 d_h d^2\delta_{\max}\Big)\cdot\exp(-\eta\beta_1\gamma T) \\
&+ \delta(T)\Big(8\eta\beta_2 d\delta_{\max} + 12\eta\beta_2\delta_{\max} + 8\eta^2\beta_2 d_h d^2\delta_{\max}\Big)\cdot\exp(-\eta\beta_2\gamma T)
\end{aligned}
\tag{43}
$$

The first inequality is derived by triangle inequality. The second inequality is derived by $\left|\boldsymbol{b}_i^\top(T)\boldsymbol{b}_j(T)\right| \leq \delta(T)$ and substituting the bounds of $|\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)|$, $|\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)|$, $|\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)|$ and $\left|\bar{\boldsymbol{b}}_i(t)^\top\bar{\boldsymbol{b}}(t)\right|$. The third inequality is derived by factoring out the common factor $\delta(T)$. The last inequality is derived by $\delta(T) \leq \delta_{\max}$, $\exp(-\eta\beta_1\gamma T) \leq 1$ and $2(\beta_1+\beta_2)(d-1)+1 \leq 4\beta_2 d$ since $\beta_2 \geq \beta_1$ and $\beta_2 \geq 1$.

We next provide the upper bound for $\delta(T+1)$.

$$\begin{aligned}
&\delta(T+1) = \max\{|\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}_j(T+1)|, |\boldsymbol{c}_i^\top(T+1)\boldsymbol{c}_j(T+1)|, |\boldsymbol{b}_i^\top(T+1)\boldsymbol{b}(T+1)|\} \\
&\leq \delta(T) + \delta(T)\Big(16\eta d\delta_{\max} + 4\eta\beta_3 + 8\eta\beta_1(d-2)\delta_{\max} + 8\eta^2 d_h d^2\delta_{\max} + \frac{2\eta\delta_{\max}}{\beta_2}\Big)\cdot\exp(-\eta\beta_1\gamma T) \\
&+ \delta(T)\Big(8\eta\beta_2 d\delta_{\max} + 40\eta^2\beta_2^2 d_h\delta_{\max} + 12\eta\beta_2\delta_{\max} + 8\eta^2\beta_2 d_h d^2\delta_{\max}\Big)\cdot\exp(-\eta\beta_2\gamma T)
\end{aligned} \tag{44}$$

This inequality can be verified by comparing with (Eq. (41)), (Eq. (42)), (Eq. (43)). To give more precise bound, we introduce the following lemma:

Lemma C.6 *If $y(t+1) \leq y(t) + cy(t)\exp(-at) + dy(t)\exp(-bt)$, with $a, b, c, d > 0, t \geq 0$ and $a, b \leq \ln 2$, then $y(t)$ satisfies:*

$$y(t) \leq y(0)\exp(\frac{2c}{a} + \frac{2d}{b})$$

Proof of lemma C.6.

$$\begin{aligned}
&y(t+1) \leq y(t) + cy(t)\exp(-at) + dy(t)\exp(-bt) \\
&\Longrightarrow y(t+1) \leq y(t)(1 + c\exp(-at) + d\exp(-bt)) \\
&\Longrightarrow y(t+1) \leq y(0)\prod_{s=0}^{t}(1 + c\exp(-as) + d\exp(-bs)) \\
&\Longrightarrow \ln y(t+1) \leq \ln y(0) + \sum_{s=0}^{t}\ln(1 + c\exp(-as) + d\exp(-bs)) \\
&\Longrightarrow \ln y(t+1) \leq \ln y(0) + \sum_{s=0}^{t}(c\exp(-as) + d\exp(-bs)) \\
&\Longrightarrow \ln y(t+1) \leq \ln y(0) + \int_{-1}^{t}(c\exp(-as) + d\exp(-bs))ds \\
&\Longrightarrow \ln y(t+1) \leq \ln y(0) + (\frac{c}{a}(\exp(a) - \exp(-at)) + \frac{d}{b}(\exp(b) - \exp(-bt))) \\
&\Longrightarrow \ln y(t+1) \leq \ln y(0) + (\frac{2c}{a} + \frac{2d}{b}) \\
&\Longrightarrow y(t+1) \leq y(0)\exp(\frac{2c}{a} + \frac{2d}{b})
\end{aligned}$$

The fourth arrow is due to $\ln(1+x) \leq x$ for $x \geq 0$. The fifth arrow is due to $\exp(-as), \exp(-bs)$ are monotone decreasing. the 7-th arrow is due to $a, b \leq \ln 2$ and $-\exp(-at) \leq 0$, $-\exp(-bt) \leq 0$.

Lemma C.6 presents the core idea of establishing property $\mathcal{C}(T+1)$. If $a \gg c$ and $b \gg d$ in the above lemma, we will have $y(t+1) \leq y(0)\cdot O(1)$. Similarly, as Mamba converges quickly ($\boldsymbol{C}^\top\boldsymbol{B} \to \frac{\beta_3}{\beta_1}\boldsymbol{I}$, $\boldsymbol{C}^\top\boldsymbol{b} \to \boldsymbol{0}$), we can prove that $|\boldsymbol{b}_i^\top(t)\boldsymbol{b}_j(t)|, |\boldsymbol{c}_i^\top(t)\boldsymbol{c}_j(t)|, |\boldsymbol{b}_i^\top(t)\boldsymbol{b}(t)|$ hold their magnitudes around their initial states.

We next combine (Eq. (44)) and lemma C.6 to give bound for $\delta(T+1)$.

$$\begin{aligned}
&\delta(T+1) \\
&\leq \delta(T) + \delta(T)\Big(16\eta d\delta_{\max} + 4\eta\beta_3 + 8\eta\beta_1(d-2)\delta_{\max} + 8\eta^2 d_h d^2\delta_{\max} + \frac{2\eta\delta_{\max}}{\beta_2}\Big)\cdot\exp(-\eta\beta_1\gamma T) \\
&+ \delta(T)\Big(8\eta\beta_2 d\delta_{\max} + 40\eta^2\beta_2^2 d_h\delta_{\max} + 12\eta\beta_2\delta_{\max} + 8\eta^2\beta_2 d_h d^2\delta_{\max}\Big)\cdot\exp(-\eta\beta_2\gamma T) \\
&\leq \delta(0)\cdot\exp\Big(\frac{2\big(16\eta d\delta_{\max} + 4\eta\beta_3 + 8\eta\beta_1(d-2)\delta_{\max} + 8\eta^2 d_h d^2\delta_{\max} + \frac{2\eta\delta_{\max}}{\beta_2}\big)}{\eta\beta_1\gamma}
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{2\big(8\eta\beta_2 d\delta_{\max} + 40\eta^2\beta_2^2 d_h\delta_{\max} + 12\eta\beta_2\delta_{\max} + 8\eta^2\beta_2 d_h d^2\delta_{\max}\big)}{\eta\beta_2\gamma}\Big) \\
&\le \delta(0)\cdot\exp\Big(\frac{3\sqrt{d_h\log(4d(2d+1)/\delta)}}{\frac{1}{2}d_h}\cdot\Big(\frac{32d}{\beta_1} + \frac{8\beta_3}{\beta_1} + 16(d-2) + \frac{16\eta d_h d^2}{\beta_1} + \frac{4}{\beta_1\beta_2} \\
&+ 16d + 80\eta\beta_2 d_h + 24 + 16\eta d_h d^2\Big)\Big) \\
&\le \frac{3}{2}\cdot\delta(0) \\
&\le 3\sqrt{d_h\log(4d(2d+1)/\delta)}
\end{aligned}
$$

The first inequality is derived by (Eq. (44)). The second inequality is derived by lemma C.6. The third inequality is derived by $\gamma = \frac{1}{2}d_h$. The last inequality is derived by

$$
\begin{aligned}
d_h &= \widetilde{\Omega}(d^2) \\
&\ge \frac{36}{(\ln(3/2))^2}\log(4d(2d+1)/\delta)\Big(\frac{32d}{\beta_1} + \frac{8\beta_3}{\beta_1} \\
&+ 16(d-2) + \frac{8}{\beta_1} + \frac{4}{\beta_1\beta_2} + 16d + 80\ln 2 + 24 + 8\Big)^2 \\
&\ge \frac{36}{(\ln(3/2))^2}\log(4d(2d+1)/\delta)\Big(\frac{32d}{\beta_1} + \frac{8\beta_3}{\beta_1} \\
&+ 16(d-2) + \frac{16\eta d_h d^2}{\beta_1} + \frac{4}{\beta_1\beta_2} + 16d + 80\eta\beta_2 d_h + 24 + 16\eta d_h d^2\Big)^2
\end{aligned}
$$

C.7 Bounds of η^2 terms

This Section presents the bounds for $\bar{\boldsymbol{b}}_i(t)^\top\bar{\boldsymbol{b}}_j(t)$, $\bar{\boldsymbol{c}}_i(t)^\top\bar{\boldsymbol{c}}_j(t)$, $\Big\|\bar{\boldsymbol{b}}(t)\Big\|_2^2$, $\bar{\boldsymbol{c}}_i^\top(t)\bar{\boldsymbol{b}}_j(t)$, $\bar{\boldsymbol{c}}_i^\top(t)\bar{\boldsymbol{b}}(t)$, $\bar{\boldsymbol{b}}_i^\top(t)\bar{\boldsymbol{b}}(t)$ (these terms usually appear in the *Vector-coupled Dynamics* equations with a η^2 factor) with $i,j\in[1,d]$ under the assumption that $\mathcal{A}(t)$, $\mathcal{B}(t)$, and $\mathcal{C}(t)$ hold.

Lemma C.7 *Under the assumption that $\mathcal{A}(t)$, $\mathcal{B}(t)$, and $\mathcal{C}(t)$ hold, we have the following bounds:*

$$
\begin{aligned}
\Big|\bar{\boldsymbol{b}}_i(t)^\top\bar{\boldsymbol{b}}_j(t)\Big| &\le 8d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t), \\
\Big|\bar{\boldsymbol{c}}_i(t)^\top\bar{\boldsymbol{c}}_j(t)\Big| &\le 8d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t) + 24\beta_2^2 d_h\delta(t)^2\exp(-\eta\beta_2\gamma t), \\
\Big\|\bar{\boldsymbol{b}}(t)\Big\|_2^2 &\le 16d_h\beta_2^2 d^2\delta(t)^2\exp(-\eta\beta_2\gamma t) + 2d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t), \\
\Big|\bar{\boldsymbol{c}}_i^\top(t)\bar{\boldsymbol{b}}_j(t)\Big| &\le 8d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t) + 8\beta_2 d_h d\delta(t)^2\exp(-\eta\beta_2\gamma t), \\
\Big|\bar{\boldsymbol{c}}_i^\top(t)\bar{\boldsymbol{b}}(t)\Big| &\le 4d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t) + 28\beta_2^2 d_h d\delta(t)^2\exp(-2\eta\beta_2\gamma t), \\
\Big|\bar{\boldsymbol{b}}_i^\top(t)\bar{\boldsymbol{b}}(t)\Big| &\le 8\beta_2 d_h d^2\delta(t)^2\exp(-\eta\beta_2\gamma t) + 4d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t)
\end{aligned}
$$

where $i,j\in[1,d]$. Note that this lemma does not require $i\neq j$.

Firstly, recall the following dynamics equation in lemma A.6:

$$
\begin{aligned}
\boldsymbol{b}_i(t+1) &= \boldsymbol{b}_i(t) + \eta\Big(\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_i(t) - \beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\cdot\boldsymbol{c}_k(t)\Big) \\
&=: \boldsymbol{b}_i(t) + \eta\bar{\boldsymbol{b}}_i(t)
\end{aligned}
$$

$$
\boldsymbol{c}_i(t+1) = \boldsymbol{c}_i(t) + \eta\Big(\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{b}_i(t) - \beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\cdot\boldsymbol{b}_k(t)
$$

$$
\begin{aligned}
&- \beta_2 \boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{b}(t)\Big)\\
&=: \boldsymbol{c}_i(t) + \eta\bar{\boldsymbol{c}}_i(t)
\end{aligned}
$$

$$
\boldsymbol{b}(t+1) = \boldsymbol{b}(t) - \eta\Big(\beta_2\sum_{k=1}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{c}_k(t)\Big) =: \boldsymbol{b}(t) + \eta\bar{\boldsymbol{b}}(t)
$$

Thus we have

$$
\bar{\boldsymbol{b}}_i(t) = \big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_i(t) - \beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\cdot\boldsymbol{c}_k(t)
$$

$$
\bar{\boldsymbol{c}}_i(t) = \big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{b}_i(t) - \beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\cdot\boldsymbol{b}_k(t) - \beta_2\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{b}(t)
$$

$$
\bar{\boldsymbol{b}}(t) = \beta_2\sum_{k=1}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{c}_k(t)
$$

Recalling the properties $\mathcal{A}(t)$ and $\mathcal{B}(t)$:

$\mathcal{A}(t)$:

$$
d_h/2 \le \boldsymbol{b}_i^\top(t)\boldsymbol{b}_i(t), \boldsymbol{c}_i^\top(t)\boldsymbol{c}_i(t), \boldsymbol{b}^\top(t)\boldsymbol{b}(t) \le 2d_h
$$

$\mathcal{B}(t)$:

$$
\begin{aligned}
|\beta_3/\beta_1 - \boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)| &\le \delta(t)\exp(-\eta\beta_1\gamma t)\\
|\boldsymbol{c}_i^\top(t)\boldsymbol{b}_j(t)| &\le 2\delta(t)\exp(-\eta\beta_1\gamma t)\\
|\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)| &\le 2\delta(t)\exp(-\eta\beta_2\gamma t) + \frac{\delta(t)}{\beta_2}\exp(-\eta\beta_1\gamma t)
\end{aligned}
$$

We can derive the follow bounds for the norm of $\bar{\boldsymbol{b}}_i(t)$, $\bar{\boldsymbol{c}}_i(t)$ and $\bar{\boldsymbol{b}}(t)$:

$$
\begin{aligned}
\Big\|\bar{\boldsymbol{b}}_i(t)\Big\| &= \Big\|\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{c}_i(t) - \beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\cdot\boldsymbol{c}_k(t)\Big\|\\
&\le \Big\|\boldsymbol{c}_i(t)\Big\|\cdot\Big|\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\Big| + \beta_1\sum_{k\neq i}^{d}\Big\|\boldsymbol{c}_k(t)\Big\|\cdot\Big|\boldsymbol{c}_k^\top(t)\boldsymbol{b}_i(t)\Big|\\
&\le \sqrt{2d_h}\delta(t)\exp(-\eta\beta_1\gamma t) + \sqrt{2d_h}\beta_1(d-1)\cdot 2\delta(t)\exp(-\eta\beta_1\gamma t)\\
&\le 2\sqrt{2d_h}d\delta(t)\exp(-\eta\beta_1\gamma t)
\end{aligned}
$$

The last inequality is by $\beta_1 \le 1$.

$$
\begin{aligned}
\Big\|\bar{\boldsymbol{c}}_i(t)\Big\| &= \Big\|\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\boldsymbol{b}_i(t) - \beta_1\sum_{k\neq i}^{d}\boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\cdot\boldsymbol{b}_k(t) - \beta_2\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{b}(t)\Big\|\\
&\le \Big\|\boldsymbol{b}_i(t)\Big\|\cdot\Big|\big(\beta_3 - \beta_1\boldsymbol{c}_i^\top(t)\boldsymbol{b}_i(t)\big)\Big| + \beta_1\sum_{k\neq i}^{d}\Big\|\boldsymbol{b}_k(t)\Big\|\cdot\Big|\boldsymbol{c}_i^\top(t)\boldsymbol{b}_k(t)\Big| + \beta_2\Big\|\boldsymbol{b}(t)\Big\|\cdot\Big|\boldsymbol{c}_i^\top(t)\boldsymbol{b}(t)\Big|\\
&\le \sqrt{2d_h}\delta(t)\exp(-\eta\beta_1\gamma t) + \sqrt{2d_h}\beta_1(d-1)\cdot 2\delta(t)\exp(-\eta\beta_1\gamma t)\\
&+ \sqrt{2d_h}\beta_2\cdot\Big(2\delta(t)\exp(-\eta\beta_2\gamma t) + \frac{\delta(t)}{\beta_2}\exp(-\eta\beta_1\gamma t)\Big)\\
&\le 2\sqrt{2d_h}d\delta(t)\exp(-\eta\beta_1\gamma t) + 2\sqrt{2d_h}\beta_2\delta(t)\exp(-\eta\beta_2\gamma t)
\end{aligned}
$$

The last inequality is by $\beta_1 \le 1$.

$$
\Big\|\bar{\boldsymbol{b}}(t)\Big\| = \Big\|\beta_2\sum_{k=1}^{d}\boldsymbol{c}_k^\top(t)\boldsymbol{b}(t)\cdot\boldsymbol{c}_k(t)\Big\|
$$

$$
\begin{aligned}
&\leq \beta_2 \sum_{k=1}^{d} \left\| \boldsymbol{c}_k(t) \right\| \cdot \left| \boldsymbol{c}_k^\top(t) \boldsymbol{b}(t) \right| \\
&\leq \sqrt{2d_h}\beta_2 d \cdot \left(2\delta(t)\exp(-\eta\beta_2\gamma t) + \frac{\delta(t)}{\beta_2}\exp(-\eta\beta_1\gamma t) \right) \\
&\leq 2\sqrt{2d_h}\beta_2 d\delta(t)\exp(-\eta\beta_2\gamma t) + \sqrt{2d_h} d\delta(t)\exp(-\eta\beta_1\gamma t)
\end{aligned}
$$

By multiplying them pairwise, we obtain:

$$
\left| \bar{\boldsymbol{b}}_i(t)^\top \bar{\boldsymbol{b}}_j(t) \right| \leq \left(2\sqrt{2d_h} d\delta(t)\exp(-\eta\beta_1\gamma t) \right)^2 \leq 8d_h d^2 \delta(t)^2 \exp(-\eta\beta_1\gamma t)
$$

The last inequality is by $\exp(-2\eta\beta_1\gamma t) \leq \exp(-\eta\beta_1\gamma t)$.

$$
\begin{aligned}
&\left| \bar{\boldsymbol{c}}_i(t)^\top \bar{\boldsymbol{c}}_j(t) \right| \leq \left(2\sqrt{2d_h} d\delta(t)\exp(-\eta\beta_1\gamma t) + 2\sqrt{2d_h}\beta_2\delta(t)\exp(-\eta\beta_2\gamma t) \right)^2 \\
&= 8d_h d^2\delta(t)^2\exp(-2\eta\beta_1\gamma t) + 8\beta_2^2 d_h \delta(t)^2 \exp(-2\eta\beta_2\gamma t) \\
&+ 16 d_h \beta_2 d\delta(t)^2 \exp(-\eta(\beta_1+\beta_2)\gamma t) \\
&\leq 8d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t) + 40\beta_2^2 d_h \delta(t)^2 \exp(-\eta\beta_2\gamma t)
\end{aligned}
$$

The last inequality is by $\beta_2 d \leq 2\beta_2^2$ because $\beta_2 = \Omega(d^2) \geq d$, and $\exp(-2\eta\beta_1\gamma t) \leq \exp(-\eta\beta_1\gamma t)$, $\exp(-2\eta\beta_2\gamma t) \leq \exp(-\eta\beta_2\gamma t)$, $\exp(-\eta(\beta_1+\beta_2)\gamma t) \leq \exp(-\eta\beta_2\gamma t)$.

$$
\begin{aligned}
&\left\| \bar{\boldsymbol{b}}(t) \right\|_2^2 \leq \left(2\sqrt{2d_h}\beta_2 d\delta(t)\exp(-\eta\beta_2\gamma t) + \sqrt{2d_h} d\delta(t)\exp(-\eta\beta_1\gamma t) \right)^2 \\
&= 8d_h\beta_2^2 d^2\delta(t)^2\exp(-2\eta\beta_2\gamma t) + 2d_h d^2\delta(t)^2\exp(-2\eta\beta_1\gamma t) \\
&+ 8d_h\beta_2 d^2\delta(t)^2\exp(-\eta(\beta_1+\beta_2)\gamma t) \\
&\leq 16 d_h\beta_2^2 d^2\delta(t)^2\exp(-\eta\beta_2\gamma t) + 2d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t)
\end{aligned}
$$

The last inequality is by $\beta_2 \leq \beta_2^2$ because $\beta_2 \geq 1$, and $\exp(-2\eta\beta_1\gamma t) \leq \exp(-\eta\beta_1\gamma t)$, $\exp(-2\eta\beta_2\gamma t) \leq \exp(-\eta\beta_2\gamma t)$, $\exp(-\eta(\beta_1+\beta_2)\gamma t) \leq \exp(-\eta\beta_2\gamma t)$.

$$
\begin{aligned}
&\left| \bar{\boldsymbol{c}}_i^\top(t) \bar{\boldsymbol{b}}_j(t) \right| \leq \left\| \bar{\boldsymbol{c}}_i(t) \right\| \cdot \left\| \bar{\boldsymbol{b}}_j(t) \right\| \\
&\leq \left(2\sqrt{2d_h} d\delta(t)\exp(-\eta\beta_1\gamma t) + 2\sqrt{2d_h}\beta_2\delta(t)\exp(-\eta\beta_2\gamma t) \right) \cdot 2\sqrt{2d_h} d\delta(t)\exp(-\eta\beta_1\gamma t) \\
&\leq 8d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t) + 8\beta_2 d_h d\delta(t)^2\exp(-\eta\beta_2\gamma t)
\end{aligned}
$$

The last inequality is by $\exp(-2\eta\beta_1\gamma t) \leq \exp(-\eta\beta_1\gamma t)$, $\exp(-\eta(\beta_1+\beta_2)\gamma t) \leq \exp(-\eta\beta_2\gamma t)$.

$$
\begin{aligned}
&\left| \bar{\boldsymbol{c}}_i^\top(t) \bar{\boldsymbol{b}}(t) \right| \leq \left\| \bar{\boldsymbol{c}}_i(t) \right\| \cdot \left\| \bar{\boldsymbol{b}}(t) \right\| \\
&\leq \left(2\sqrt{2d_h} d\delta(t)\exp(-\eta\beta_1\gamma t) + 2\sqrt{2d_h}\beta_2\delta(t)\exp(-\eta\beta_2\gamma t) \right) \\
&\cdot \left(2\sqrt{2d_h}\beta_2 d\delta(t)\exp(-\eta\beta_2\gamma t) + \sqrt{2d_h} d\delta(t)\exp(-\eta\beta_1\gamma t) \right) \\
&= 4d_h d^2\delta(t)^2\exp(-2\eta\beta_1\gamma t) + 8\beta_2^2 d_h d\delta(t)^2\exp(-2\eta\beta_2\gamma t) \\
&+ 8\beta_2 d_h d^2\delta(t)^2\exp(-\eta(\beta_1+\beta_2)\gamma t) + 4\beta_2 d_h d\delta(t)^2\exp(-\eta(\beta_1+\beta_2)\gamma t) \\
&\leq 4d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t) + 28\beta_2^2 d_h d\delta(t)^2\exp(-2\eta\beta_2\gamma t)
\end{aligned}
$$

The last inequality is by $\beta_2 d \leq 2\beta_2^2$, $\beta_2 \leq \beta_2^2$, and $\exp(-2\eta\beta_1\gamma t) \leq \exp(-\eta\beta_1\gamma t)$, $\exp(-2\eta\beta_2\gamma t) \leq \exp(-\eta\beta_2\gamma t)$, $\exp(-\eta(\beta_1+\beta_2)\gamma t) \leq \exp(-\eta\beta_2\gamma t)$.

$$
\begin{aligned}
&\left| \bar{\boldsymbol{b}}_i^\top(t) \bar{\boldsymbol{b}}(t) \right| \leq \left\| \bar{\boldsymbol{b}}_i(t) \right\| \cdot \left\| \bar{\boldsymbol{b}}(t) \right\| \\
&\leq 2\sqrt{2d_h} d\delta(t)\exp(-\eta\beta_1\gamma t) \cdot \left(2\sqrt{2d_h}\beta_2 d\delta(t)\exp(-\eta\beta_2\gamma t) + \sqrt{2d_h} d\delta(t)\exp(-\eta\beta_1\gamma t) \right) \\
&\leq 8\beta_2 d_h d^2\delta(t)^2\exp(-\eta\beta_2\gamma t) + 4d_h d^2\delta(t)^2\exp(-\eta\beta_1\gamma t)
\end{aligned}
$$

The last inequality is by $\exp(-2\eta\beta_1\gamma t) \leq \exp(-\eta\beta_1\gamma t)$, $\exp(-\eta(\beta_1+\beta_2)\gamma t) \leq \exp(-\eta\beta_2\gamma t)$.

D Discussion

In this section, we show that orthogonal initialization Mamba can be trained to ICL solution, and compare our method with some previous works.

Orthogonal Initialization Now we assume that each column of $\boldsymbol{W}_B$ and $\boldsymbol{W}_C$ are initialized with orthogonal columns of unit norm. Then we have

$$\boldsymbol{C}^\top(0)\boldsymbol{C}(0) = \boldsymbol{B}^\top(0)\boldsymbol{B}(0) = \boldsymbol{I}, \quad \boldsymbol{B}^\top(0)\boldsymbol{b}(0) = \boldsymbol{C}^\top(0)\boldsymbol{b}(0) = \boldsymbol{0}.$$

Consider the following update rule as part of lemma 5.1.

$$\boldsymbol{B}(t+1) = \boldsymbol{B}(t) + \eta\beta_3\boldsymbol{C}(t) - \eta\beta_1\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{B}(t), \tag{45}$$

$$\boldsymbol{C}(t+1) = \boldsymbol{C}(t) + \eta\beta_3\boldsymbol{B}(t) - \eta\beta_1\boldsymbol{B}(t)\boldsymbol{B}(t)^\top\boldsymbol{C}(t) - \eta\beta_2\boldsymbol{b}(t)\boldsymbol{b}(t)^\top\boldsymbol{C}(t), \tag{46}$$

$$\boldsymbol{b}(t+1) = \boldsymbol{b}(t) - \eta\beta_2\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{b}(t). \tag{47}$$

By (Eq. (45)), (Eq. (46)) and (Eq. (47)), we have:

$$\begin{aligned}
&\boldsymbol{B}^\top(t+1)\boldsymbol{b}(t+1) = \boldsymbol{B}(t)^\top\boldsymbol{b}(t) + \eta\beta_3\boldsymbol{C}(t)^\top\boldsymbol{b}(t) - \eta\beta_1\boldsymbol{B}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{b}(t) \\
&- \eta\beta_2\boldsymbol{B}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{b}(t) - \eta^2\beta_2\beta_3\boldsymbol{C}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{b}(t) + \eta^2\beta_1\beta_2\boldsymbol{B}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{b}(t)
\end{aligned}$$

$$\begin{aligned}
&\boldsymbol{C}(t+1)^\top\boldsymbol{b}(t+1) = \boldsymbol{C}(t)^\top\boldsymbol{b}(t) + \eta\beta_3\boldsymbol{B}(t)^\top\boldsymbol{b}(t) - \eta\beta_1\boldsymbol{C}(t)^\top\boldsymbol{B}(t)\boldsymbol{B}(t)^\top\boldsymbol{b}(t) - \eta\beta_2\boldsymbol{C}(t)^\top\boldsymbol{b}(t)\boldsymbol{b}(t)^\top\boldsymbol{b}(t) \\
&- \eta\beta_2\boldsymbol{C}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{b}(t) - \eta^2\beta_2\beta_3\boldsymbol{B}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{b}(t) \\
&+ \eta^2\beta_1\beta_2\boldsymbol{C}(t)^\top\boldsymbol{B}(t)\boldsymbol{B}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{b}(t) + \eta^2\beta_2^2\boldsymbol{C}(t)^\top\boldsymbol{b}(t)\boldsymbol{b}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{b}(t)
\end{aligned}$$

Combining $\boldsymbol{B}^\top(0)\boldsymbol{b}(0) = \boldsymbol{C}^\top(0)\boldsymbol{b}(0) = \boldsymbol{0}$ with induction, we can derive that $\boldsymbol{B}^\top(t)\boldsymbol{b}(t) = \boldsymbol{C}^\top(t)\boldsymbol{b}(t) = \boldsymbol{0}$ for $t \geq 0$. Thus we only need to consider the following dynamics.

$$\boldsymbol{B}(t+1) = \boldsymbol{B}(t) + \eta\beta_3\boldsymbol{C}(t) - \eta\beta_1\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{B}(t), \tag{48}$$

$$\boldsymbol{C}(t+1) = \boldsymbol{C}(t) + \eta\beta_3\boldsymbol{B}(t) - \eta\beta_1\boldsymbol{B}(t)\boldsymbol{B}(t)^\top\boldsymbol{C}(t) \tag{49}$$

Denote $\boldsymbol{B}(t)^\top\boldsymbol{B}(t) = D(t)$, $\boldsymbol{C}(t)^\top\boldsymbol{C}(t) = E(t)$ and $\boldsymbol{C}(t)^\top\boldsymbol{B}(t) = F(t)$ then by (Eq. (48)) and (Eq. (49)), we have

$$\begin{aligned}
\boldsymbol{F}(t+1) = \boldsymbol{F}(t) &+ \eta\beta_3(\boldsymbol{D}(t) + \boldsymbol{E}(t)) - \eta\beta_1\boldsymbol{F}(t)\boldsymbol{D}(t) \\
&+ \eta^2\beta_3^2\boldsymbol{F}(t)^\top - 2\eta^2\beta_1\beta_3\boldsymbol{F}(t)\boldsymbol{F}(t)^\top - \eta\beta_1\boldsymbol{E}(t)\boldsymbol{F}(t) \\
&+ \eta^2\beta_1^2\boldsymbol{F}(t)\boldsymbol{F}(t)^\top\boldsymbol{F}(t)
\end{aligned} \tag{50}$$

$$\begin{aligned}
\boldsymbol{D}(t+1) = \boldsymbol{D}(t) &+ \eta\beta_3(\boldsymbol{F}(t) + \boldsymbol{F}(t)^\top) - \eta\beta_1(\boldsymbol{F}(t)^\top\boldsymbol{F}(t) + \boldsymbol{F}(t)^\top\boldsymbol{F}(t)) \\
&+ \eta^2\beta_3^2\boldsymbol{E}(t) - \eta^2\beta_1\beta_3\boldsymbol{F}(t)^\top\boldsymbol{E}(t) \\
&- \eta^2\beta_1\beta_3\boldsymbol{E}(t)\boldsymbol{F}(t) + \eta^2\beta_1^2\boldsymbol{F}(t)^\top\boldsymbol{E}(t)\boldsymbol{F}(t)
\end{aligned} \tag{51}$$

$$\begin{aligned}
\boldsymbol{E}(t+1) = \boldsymbol{E}(t) &+ \eta\beta_3(\boldsymbol{F}(t) + \boldsymbol{F}(t)^\top) - \eta\beta_1(\boldsymbol{F}(t)^\top\boldsymbol{F}(t) + \boldsymbol{F}(t)^\top\boldsymbol{F}(t)) \\
&+ \eta^2\beta_3^2\boldsymbol{D}(t) - \eta^2\beta_1\beta_3\boldsymbol{F}(t)^\top\boldsymbol{D}(t) \\
&- \eta^2\beta_1\beta_3\boldsymbol{D}(t)\boldsymbol{F}(t) + \eta^2\beta_1^2\boldsymbol{F}(t)^\top\boldsymbol{D}(t)\boldsymbol{F}(t)
\end{aligned} \tag{52}$$

Note that $D(0) = E(0) = \boldsymbol{I}$ and $F(0) = \boldsymbol{0}$. By induction we can see that $D(t), E(t)$ and $F(t)$ are diagonal matrix for $t > 0$. Because of the symmetry, we have $D(t) = E(t)$. Now we denote $D(t) = E(t) = g(t)\boldsymbol{I}$ and $F(t) = h(t)\boldsymbol{I}$. Then based on (Eq. (50)), (Eq. (51)) and (Eq. (52)), we have:

$$g(t+1) = g(t) + \eta(2h(t) + \eta\beta_3 g(t) - \eta\beta_1 g(t)h(t))(\beta_3 - \beta_1 h(t)) \tag{53}$$

$$h(t+1) = h(t) + \eta g(t)(\beta_3 - \beta_1 h(t)) + \eta^2 h(t)(\beta_3 - \beta_1 h(t))^2 \tag{54}$$

Since $g(0) = 1$ and $h(0) = 0$ at initialization, $h(t)$ will converge to $\frac{\beta_3}{\beta_1}$ (i.e. $\boldsymbol{C}^\top\boldsymbol{B} \to \frac{\beta_3}{\beta_1}\boldsymbol{I}$).

Compare with Other Techniques (Eq. (45)), (Eq. (46)) and (Eq. (47)) can be viewed as the gradient descent that minimize the following target:

$$\frac{1}{2}\|\boldsymbol{C}^\top \boldsymbol{W}_B \boldsymbol{X} - \boldsymbol{Y}\|_F^2 \tag{55}$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ satisfy:

$$\boldsymbol{X}\boldsymbol{X}^\top = \begin{bmatrix} \beta_1 & & & \\ & \ddots & & \\ & & \beta_1 & \\ & & & \beta_2 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}, \boldsymbol{X}\boldsymbol{Y}^\top = \begin{bmatrix} \beta_3 & & \\ & \ddots & \\ & & \beta_3 \\ & & \mathbf{0}_{1\times d} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d)}.$$

To establish convergence for this problem under gaussian initialization, Arora et al. (2019) require the standard deviation to be small enough, while Du and Hu (2019) require larger dimension d_h because their method relies on the condition number of $\boldsymbol{X}$. Our method balances the requirements on initialization and dimension. The fine-grained nature of our analysis (particularly the *Vector-coupled Dynamics*) enables extension to various problem beyong (Eq. (55)).

E Additional Experimental Results

(a) Initialization (b) Trained parameter (c) Loss curve

Figure 2: (a) Visualization of matrix product $\boldsymbol{C}^\top \boldsymbol{W}_B$ before training; (b) Post-training visualization of matrix product $\boldsymbol{C}^\top \boldsymbol{W}_B$; (c) Test loss versus token sequence length N. Blue curve: experimental loss; orange dashed line: theoretical loss $\frac{d}{2}\left(1 - \frac{\beta_3^2}{\beta_1}\right)$.

Experiments Setting We follow Section 3 to generate the dateset and initialize the model. Specifically, we set dimension $d = 4$, $d_h = 80$, prompt token length $N = 50$, and train the Mamba model on 3000 sequences by gradient descent. Moreover, we vary the length of the prompt token N from 4 to 80 and compare the test loss with the theoretical loss. For each N, we conduct 10 independent experiments and report the averaged results. All experiments are performed on an NVIDIA A800 GPU.

Experiment Result Figure 2a and Figure 2b show that $\boldsymbol{C}^\top \boldsymbol{B}$ can be trained to diagonal matrix from random initialization. Figure 2c show that the experimental loss aligns with the theoretical loss $\mathcal{L}(\boldsymbol{\theta}) = \frac{d}{2}\left(1 - \frac{\beta_3^2}{\beta_1}\right)$, noting that the theoretical loss $\left(1 - \frac{\beta_3^2}{\beta_1}\right)$ has an upper bound $\frac{3d(d+1)}{2N}$ that decays linearly with N. These experimental results further verified our theoretical proof.

Mamba vs Linear Attention Optimal linear attention outperforms Mamba under our construction, and they have $O(1/N)$ error upper bound with different constant factors. We provide a comprarison of loss between optimal Mamba (under our Assumption 4.1) with optimal linear attention as in Table 2 with setting $d = 10, N = 10, 20, \ldots, 80$.

When N is smaller than d We also test the case when $N \leq d$ in Table 3 with setting $d = 20, N = 4, 6, \ldots, 20$.

Convergence of $\boldsymbol{w}_\Delta$ We set $\boldsymbol{w}_\Delta = 0$ in the assumption. Now we show that random initializd $\boldsymbol{w}_\Delta = 0$ can converge to 0 experimental. The results is in Table 4.

Different d_h Table 5 shows the mean value and standard deviation of the loss for smaller d_h (in 10 repeated experiments). We set $d = 4, N = 30$, and the theoretical loss is 0.2954.

Table 2: Comparison of Mamba and Linear Attention

N	10	20	30	40	50	60	70	80
Mamba	2.6671	1.8189	1.3800	1.1117	0.9308	0.8005	0.7022	0.6254
Linear Attention	2.6190	1.7742	1.3415	1.0784	0.9016	0.7746	0.6790	0.6044

Table 3: Experiment for $N \leq d$

N	4	6	8	10	12	14	16	18	20
Experimental Loss	8.5911	7.8292	7.7009	6.8235	6.4004	6.0612	5.9689	5.6193	5.1426
Theoretical Loss	8.4484	7.8425	7.3173	6.8579	6.4526	6.0926	5.7706	5.4810	5.2190

Table 4: Convergence of $\boldsymbol{w}_\Delta$

Epoch	0	10	20	30	40	50	60	70	80
$\|\boldsymbol{w}_\Delta\|_2$	0.8883	0.7513	0.4821	0.3331	0.2444	0.2026	0.1868	0.1799	0.1773
$\|\boldsymbol{w}_\Delta\|_2^2$	0.7891	0.5645	0.2324	0.1109	0.0597	0.0410	0.0349	0.0324	0.0314

Table 5: Different d_h

d_h	6	8	10	12	14	16	18	20
mean(loss)	0.2912	0.2933	0.2899	0.2887	0.2929	0.2951	0.2967	0.2959
std(loss)	0.0075	0.0055	0.0116	0.0052	0.0105	0.0097	0.0110	0.0142