A COMPARATIVE STUDY OF IN-CONTEXT LINEAR REGRESSION BY TRANSFORMERS WITH CHAIN-OF-THOUGHT

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers have demonstrated strong capabilities in multi-step reasoning, with chain-of-thought (CoT) emerging as a powerful prompting and modeling technique. To theoretically understand the capabilities of transformers with CoT, a line of recent work has studied settings in which transformers generate the iterates of an algorithm either through layer-wise propagation, which can be viewed as a form of "implicit CoT" (Saunshi et al., 2025), or through autoregressive "explicit CoT" steps, which directly produce the intermediate iterates (Huang et al., 2025a). However, it remains unclear how the transformer and CoT setup affect the model's capabilities. To better understand these mechanisms in a controlled and mathematically concrete setting, we study a single-layer linear self-attention model on two in-context linear regression tasks: weight prediction and query prediction. Using multi-step gradient descent (GD) as the teacher algorithm, we compare when implicit and explicit CoT can or cannot simulate the target iterative computation. Our empirical and theoretical results reveal a sharp contrast in this stylized setting. On weight prediction, implicit CoT fails whereas explicit CoT succeeds; on query prediction, explicit CoT fails, whereas implicit CoT can simulate the GD trajectory up to a fixed output sign transformation and succeeds directly with an added scratchpad. We identify two corresponding bottlenecks: context overwriting during recurrent updates and insufficient writable state for autoregressive token generation. We further show that adding a scratchpad resolves the failure of implicit CoT on weight prediction and enables explicit CoT on query prediction. In addition, we provide a construction showing that the computation realized by explicit CoT can be exactly implemented by implicit CoT with a scratchpad.

1 INTRODUCTION

Recent advances in large language models have highlighted the remarkable capability of transformers to solve complex reasoning tasks (Achiam et al., 2023; Radford et al., 2019; Vaswani et al., 2017). Specifically, chain-of-thought (CoT) reasoning has emerged as a powerful technique that enables transformers to solve complex problems by generating intermediate reasoning steps (Wang et al., 2023a; Saparov & He, 2022; Shi et al., 2022; Paul et al., 2024).

Research on CoT reasoning have largely focused on two paradigms named "explicit CoT" and "implicit CoT". In explicit CoT, the model generates external intermediate reasoning states before producing the final answer, which form an external reasoning trajectory (Wei et al., 2022). In implicit CoT, intermediate computation is carried out inside the model, without exposing the external reasoning trace (Deng et al., 2023). Specifically, the terminology "implicit CoT" includes several related mechanisms, such as implicit chain-of-thought (Deng et al., 2023; 2024), continuous latent reasoning (Hao et al., 2024), latent tokens (Sun et al., 2025), pause tokens (Goyal et al., 2023), and recurrent-depth reasoning (Saunshi et al., 2025; Geiping et al., 2026; Chen et al., 2025). These mechanisms differ, but they share the same idea that intermediate reasoning need not be externalized as visible tokens. However, the theoretical understanding of the effect of the two CoT paradigms remains unclear. This motivates a controlled comparison of how the two CoT paradigms store, update, and preserve information during multi-step reasoning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of implicit and explicit CoT with and without scratchpad on two tasks in our single-layer linear self-attention setting. "Success" means that the model can simulate the target GD trajectory and training converges to the desired parameters. "Failure" means that the GD algorithm cannot be exactly implemented.

Task	CoT Paradigm	Without Scratchpad	With Scratchpad
Weight Prediction	Implicit	failure	success
	Explicit	success	success
Query Prediction	Implicit	success	success
	Explicit	failure	success

For explicit CoT, we refer to autoregressively generating explicit intermediate reasoning steps (Huang et al., 2025a; Takanami & Pehlevan, 2026). At each step, the model reads out the last column $f_{\mathrm{LSA}}(\hat{\mathbf{Z}}_i)_{[:,-1]}$ and appends it to the sequence as

$$\hat{\mathbf{Z}}_{i+1} = \left[\hat{\mathbf{Z}}_i, f_{\mathrm{LSA}}(\hat{\mathbf{Z}}_i)_{[:,-1]}\right], \quad i = 0, \ldots, L-1,$$

so the generated columns form an external reasoning trajectory. For implicit CoT, we mean using the internal computation in the forward pass of an "unrolled" model to perform reasoning, following Deng et al. (2023); Saunshi et al. (2025); Geiping et al. (2026). Starting from $\hat{\mathbf{Z}}_0$, the model repeatedly applies the same self-attention layer

$$\hat{\mathbf{Z}}_{i+1} = f_{\mathrm{LSA}}(\hat{\mathbf{Z}}_i), \quad i = 0, \ldots, L-1,$$

and reads the prediction from the final internal state $\hat{\mathbf{Z}}_L$. Thus, our setting can be regarded as a stylized theoretical formalization of concrete implicit and explicit CoT paradigms, reflecting the broader distinction between external and internal intermediate reasoning.

In this paper, we provide a comprehensive case study comparing these two CoT formulations in a controlled setting. We focus on a single-layer linear self-attention model and consider two well-defined numerical tasks based on in-context linear regression: weight prediction and query prediction. In this numerical setting, the intermediate reasoning states are inevitably continuous representations. We use multi-step gradient descent (GD) on the regression objective as a concrete teacher trajectory and target iterative computation. Further, we investigate the role of "scratchpad" (additional writable state) and show how it influences each CoT paradigm. We view these settings as a controlled theoretical testbed rather than a direct model of all real-world CoT reasoning.

The main contributions of this paper are as follows:

- We formulate a unified controlled framework for comparing implicit and explicit CoT in a single-layer linear self-attention model. The framework covers eight cases across CoT paradigm (implicit/explicit), task type (weight/query prediction), and encoding design (with/without scratchpad), as summarized in Table 1.
- Within the controlled framework, we characterize when the model can or cannot implement the teacher GD trajectory. For the successful cases, we establish positive expressivity results and training guarantees, including convergence and out-of-distribution (OOD) test error bounds. For the failure cases, we prove expressivity lower bounds showing that exact implementation of the target GD trajectory is impossible under the specified architecture. We also provide experimental support for the theoretical findings.
- We identify the mechanisms behind the negative results. The failure of implicit CoT on weight prediction is due to context overwriting of the original data during recurrent layer updates, and the failure of explicit CoT on query prediction is caused by insufficient space for state updates. These mechanisms point to architectural and encoding bottlenecks behind the expressivity failures.
- We analyze scratchpad as a controlled change to the input encoding and representation space. We demonstrate that an additional scratchpad can overcome context overwriting for implicit CoT and insufficient space for explicit CoT. We also design a construction showing that the reasoning trajectory generated from explicit CoT can be exactly simulated by implicit CoT with a scratchpad.

2 PRELIMINARIES

In this section, we introduce the tasks, model, and training methods used in our comparison. We consider two in-context learning tasks derived from linear regression. For linear regression, the

ground-truth parameter is $\boldsymbol{\beta} \sim N(0, \mathbf{I}_d)$, each covariate is $\mathbf{x}_i \sim N(0, \mathbf{I}_d)$, and the response is $y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \xi_i$, where $\xi_i \sim N(0, \sigma^2)$. Denote $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top$ and $\mathbf{y} = [y_1, \ldots, y_N]^\top$. Given the examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, we consider GD steps as the teacher trajectory, where the updates follow

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \eta_{gd} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta}_t - \mathbf{y}).$$

In-context weight prediction. Given example pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, the goal is to recover the ground-truth weight vector $\boldsymbol{\beta}$. In our teacher-guided formulation, the intermediate reasoning states are the GD iterates $\boldsymbol{\beta}_t$, and the CoT reasoning is to learn the update of $\boldsymbol{\beta}_t$. This task is also considered in recent works Huang et al. (2025a); Takanami & Pehlevan (2026).

In-context query prediction. Given example pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ and a query $\mathbf{x}_q$, the goal is to predict the query label $y_q = \boldsymbol{\beta}^\top \mathbf{x}_q$. Under the same teacher guidance, the model is expected to learn the update of intermediate predictions $\boldsymbol{\beta}_t^\top \mathbf{x}_q$. This task is classical for studying in-context learning and is commonly used in broad literature Garg et al. (2022); Von Oswald et al. (2023); Zhang et al. (2024); Ahn et al. (2023); Gatmiry et al. (2024).

Regarding the model, we consider a single-layer linear self-attention (LSA) model

$$f_{\text{LSA}}(\mathbf{Z}) = \mathbf{Z} + \mathbf{W}^V \mathbf{Z} \cdot \mathbf{Mask}(\mathbf{Z})^\top \mathbf{W}^{KQ} \mathbf{Z},$$

where $\mathbf{Z}$ is the input matrix, $\mathbf{W}^V$ is the value matrix, $\mathbf{W}^{KQ}$ is a single matrix merging the key and query matrices, and the mask function $\mathbf{Mask}(\cdot)$ sets all columns after column N to zero.

This single-layer LSA model provides a tractable setting in which the computation between attention, recurrent updates, and appended intermediate states can be analyzed rigorously. This architecture is commonly considered in theoretical studies of ICL theory (Garg et al., 2022; Von Oswald et al., 2023; Zhang et al., 2024; Huang et al., 2025a).

The mask function serves two purposes depending on the setting. For implicit CoT, it prevents the example tokens from attending to the query token x_q. For explicit CoT, the current generated token plays the role of the query/residual state, while the original N examples remain fixed source tokens. The mask removes older generated tokens from the source set, so the next generated state depends only on the fixed context and the immediately preceding generated state. This gives a Markovian formalization of explicit CoT, aligned with the MCoT framework (Yang et al., 2025). Unlike the literal causal mask used in standard language models, the original N examples remain mutually accessible, as is standard in in-context learning. Since GD updates form a Markov chain, MCoT provides a natural explicit CoT framework for our study. This masking choice does not remove the external reasoning trace in explicit CoT; instead, the generated intermediate states are still read out, appended to the sequence, and preserved as intermediate outputs.

For training, we employ a teacher-guided method in which GD provides the target intermediate trajectory. The choice of GD provides a concrete and mathematically tractable iterative reasoning trajectory, allowing us to study whether each CoT paradigm can implement the same multi-step computation. This teacher-guided setup is aligned with prior work on CoT intermediate step supervision training (Deng et al., 2023; Magister et al., 2023; Ho et al., 2023; Huang et al., 2025a), where models are trained to reproduce intermediate reasoning states generated by a teacher process.

Experiment setup. For experiments in this paper, we set the length of sequences $N = 20$, the dimension $d = 5$, and the learning rate of the teacher GD algorithm $\eta_{gd} = 10^{-2}$. We train the single-layer LSA model for 100000 iterations. All weight matrices are initialized with small random values and are optimized by the Adam optimizer with a learning rate $\eta = 10^{-3}$. We repeat each experiment with five independent random seeds and report the mean and standard deviation across seeds, where the shaded regions in the figures denote the standard deviation.

3 In-Context Weight Prediction

In this section, we present the performance of implicit and explicit CoT on in-context weight prediction. Given the data sequence $\{\mathbf{x}_i, y_i\}_{i=1}^N$, the input matrix is designed as

$$\mathbf{Z}_0 = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_N & \boldsymbol{\beta}_0 \\ y_1 & y_2 & \ldots & y_N & 0 \\ 0 & 0 & \ldots & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top & \boldsymbol{\beta}_0 \\ \mathbf{y}^\top & 0 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{(d+2)\times(N+1)},$$

where $\boldsymbol{\beta}_0$ is the initialization of the weight vector, and the extra 1 is added to separate the original input data and the query token, which is also considered in Bai et al. (2023); Huang et al. (2025a).

3.1 Implicit CoT

For implicit CoT, we loop a single-layer linear self-attention model and read the first d entries of the last column as the prediction, i.e. $f_{\text{LSA}}^{(L)}(\mathbf{Z}_0)_{[:d,-1]}$. Then, we define the population training loss

$$\mathcal{L}(\mathbf{W}^V, \mathbf{W}^{KQ}) = \mathbb{E}_{\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\beta}_0}\left[\|f_{\text{LSA}}(\mathbf{Z}_0)_{[:d,-1]} - \boldsymbol{\beta}_1\|^2\right],$$

where the initialized weight $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I}_d)$, and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 - \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})$ is the ground-truth weight vector after one gradient step on the linear regression objective with a fixed learning rate η_{gd}.

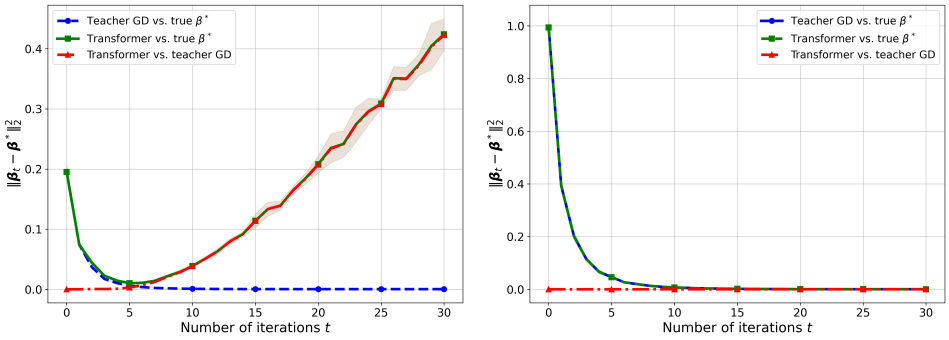
We train a single-layer LSA model to minimize this loss and unroll the trained model iteratively. We get the experimental results as shown in Figure 1a. The red curve illustrates that the prediction obtained from the unrolled model diverges from the ground-truth GD steps, indicating that implicit CoT fails to simulate the teacher algorithm. The following theorem formalizes the lack of expressivity.

Theorem 3.1. *Assume that $\mathbf{x}_i \sim N(0, \mathbf{I}_d)$, $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I}_d)$, $\mathbf{y} \sim N(0, \mathbf{I}_N)$, and $\boldsymbol{\beta}_i$ is the ground-truth weight vector after i GD steps from $\boldsymbol{\beta}_0$ with learning rate η_{gd}. For any $f_{LSA}(\cdot)$ satisfying $f_{LSA}(\mathbf{Z}_0)_{[:d,-1]} \equiv \boldsymbol{\beta}_1$, it holds that*

$$\mathbb{E}\left[\|f_{LSA}^{(2)}(\mathbf{Z}_0)_{[:d,-1]} - \boldsymbol{\beta}_2\|_2^2\right] \geq \eta_{gd}^4 Nd.$$

This theorem demonstrates that even if the first step is perfectly matched, the second step inevitably leads to an error. It indicates that a transformer model with implicit CoT cannot implement a multi-step GD algorithm for weight prediction, revealing a limitation in the expressivity of implicit CoT. Note that the theorem assumes $\mathbf{y} \sim N(0, \mathbf{I}_N)$, which differs from the experimental setting. This assumption is mainly used to simplify the theoretical analysis. In this expressivity theorem, both $\mathbf{X}$ and $\mathbf{y}$ are observed inputs, and the goal is to reproduce the GD update determined by the observed data and the current iterate $\boldsymbol{\beta}_i$. Thus, once $(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}_i)$ are given, the next GD state is fully determined, and the randomness of $\mathbf{y}$ does not prevent the GD update.

We remark that the failure is not merely an artifact of the one-step training objective. We additionally train implicit-CoT models end-to-end at recurrent depths $L \in \{2, \ldots, 8\}$. We illustrate the test losses in the appendix (see Figure 9). The test loss remains far from zero across these recurrent depths, indicating that the failure persists under end-to-end recurrent-depth training. This empirical result is also consistent with the expressivity limitation stated in Theorem 3.1.



(a) Implicit CoT (b) Explicit CoT

Figure 1: Results on weight prediction without scratchpad, evaluated for $L = 30$ steps. Lines denote differences between: model prediction and teacher GD (red), model prediction and ground-truth label (green), and teacher GD and ground-truth label (blue). (a) corresponds to implicit CoT, while (b) corresponds to explicit CoT.

3.2 EXPLICIT COT

For explicit CoT, we generate intermediate sequences $\{\hat{\mathbf{Z}}_i\}_{i=1}^{L}$ as

$$\hat{\mathbf{Z}}_i = \begin{bmatrix} \mathbf{X}^\top & \boldsymbol{\beta}_0 & \hat{\boldsymbol{\beta}}_1 & ... & \hat{\boldsymbol{\beta}}_i \\ \mathbf{y}^\top & 0 & 0 & ... & 0 \\ 0 & 1 & 1 & ... & 1 \end{bmatrix} \in \mathbb{R}^{(d+2)\times(N+i+1)}, i \in [L],$$

where we read the first d entries of the last column as the i-th step prediction for the weight vector. Then, based on the teacher GD algorithm, the CoT training objective is defined as

$$\ell^{\text{CoT}}(\mathbf{X}, \boldsymbol{\beta}; \mathbf{W}^V, \mathbf{W}^{KQ}) = \sum_{i=0}^{L-1} \left\| f_{\text{LSA}}(\mathbf{Z}_i)_{[:,-1]} - (\boldsymbol{\beta}_{i+1}, 0, 1) \right\|^2,$$

where $\{\mathbf{Z}_i\}_{i=1}^{L}$ are the ground-truth sequences, $\boldsymbol{\beta}_i$ is the ground-truth weight vector after i GD steps with learning rate η_{gd}. Further, the corresponding population loss is $\mathcal{L}^{\text{CoT}}(\mathbf{W}^V, \mathbf{W}^{KQ}) = \mathbb{E}_{\mathbf{x}_i, \boldsymbol{\beta}}\left[\ell^{\text{CoT}}(\mathbf{X}, \boldsymbol{\beta}; \mathbf{W}^V, \mathbf{W}^{KQ})\right]$.

Figure 1b compares the prediction of the trained model with explicit CoT and the teacher GD algorithm. We observe that the blue line and the green line are exactly coincident, and the red line remains near 0, which means that the trained model can implement the teacher multi-step GD.

According to the visualization of the trained matrices $\mathbf{W}^{KQ}$ and $\mathbf{W}^V$ (see Figure 5 in appendix), we observe that they take the form $\mathbf{W}^{KQ} = \begin{bmatrix} -\mu\mathbf{I}_d & 0 & 0 \\ 0 & 0 & \mu \\ 0 & 0 & 0 \end{bmatrix}$ and $\mathbf{W}^V = \begin{bmatrix} \nu\mathbf{I}_d & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$. With this structure, we can derive the update of the predicted weight vector as

$$\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_{i-1} - \mu\nu\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}_{i-1} + \mu\nu\mathbf{X}^\top\mathbf{y}.$$

We observe that when $\mu\nu = \eta_{gd}$, the update rule is consistent with the teacher GD algorithm. As a result, we have the following theoretical result about the expressivity of explicit CoT.

Theorem 3.2. *There exist parameter matrices $\mathbf{W}^{KQ}$ and $\mathbf{W}^V$ such that f_{LSA} with explicit CoT and without scratchpad can implement a multi-step GD algorithm for in-context weight prediction task.*

Compared with Theorem 3.1, we observe that explicit CoT shows stronger expressivity than implicit CoT on weight prediction.

Next, we provide training guarantees under gradient flow. Here, we briefly introduce our results, with details stated in Appendix C.2. We demonstrate that with the initialization $w_{23}(0) = 1$ and all other parameters zero, the parameters always remain in the form

$$\mathbf{W}^{KQ}(t) = \begin{bmatrix} w_{11}(t)\cdot\mathbf{I}_d & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & w_{23}(t) \\ \mathbf{0} & 0 & 0 \end{bmatrix}, \mathbf{W}^V(t) = \begin{bmatrix} v_{11}(t)\cdot\mathbf{I}_d & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & 0 \\ \mathbf{0} & 0 & 0 \end{bmatrix},$$

where only three scalars w_{11}, w_{23}, and v_{11} evolve in training. We obtain the convergence result.

Theorem 3.3. *For any $\epsilon > 0$, there exists $T = \Theta(\log(\epsilon^{-1}LN^2d^2))$ such that for all $t > T$, it holds that $|v_{11}(t)w_{11}(t) + \eta_{gd}| < \epsilon$, $|v_{11}(t)w_{23}(t) - \eta_{gd}| < \epsilon$, and $\mathcal{L}^{CoT}(\mathbf{W}^{KQ}(t), \mathbf{W}^V(t)) < \epsilon$.*

This theorem demonstrates that $v_{11}w_{11}$ and $v_{11}w_{23}$ will converge to $-\eta_{gd}$ and η_{gd} respectively. Further, based on the convergence of $v_{11}w_{11}$ and $v_{11}w_{23}$, we demonstrate that gradient flow will nearly converge to the global minimizer.

We further obtain an OOD generalization bound under covariance shift. As the learned GD dynamics approach the ordinary least-squares estimator (OLS), the OOD test error is controlled, where its leading term is of the same order as the asymptotic estimation error of OLS. The precise statement and proof are provided in Appendix C.2 (See Theorem C.12).

3.3 CONTEXT OVERWRITING OF IMPLICIT COT

For weight prediction, we find that explicit CoT successfully simulates the teacher GD algorithm, while implicit CoT fails. The fundamental cause is how the two paradigms interact with the original input data. For GD, computing the i-th step strictly requires both the intermediate weight vector $\boldsymbol{\beta}_i$ and the original data $\mathbf{X}$ and $\mathbf{y}$. Explicit CoT naturally preserves $\mathbf{X}$ since intermediate states are appended as new tokens at the end of the sequence, leaving the historical tokens unchanged. In

contrast, implicit CoT updates the whole representation space directly. Since the $\boldsymbol{\beta}_i$ is in the same row of $\mathbf{X}$, the representation of $\mathbf{X}$ will change simultaneously with the update of $\boldsymbol{\beta}_i$, leading to the context overwriting. Without the essential information contained in $\mathbf{X}$, the model can hardly compute the correct GD steps. We remark that recent works (Liu et al., 2024; Shi et al., 2023; Wang et al., 2023b) also observe the context overwriting of original data and prompting in LLMs.

To empirically validate this hypothesis, we try to recover $\mathbf{X}$ from the intermediate state $\hat{\mathbf{Z}}_L$. In particular, we train a 3-layer transformer model to recover $\mathbf{X}$. Figure 2 illustrates the recovery loss. For implicit CoT (blue line), the loss increases with the number of iterations, suggesting that it becomes increasingly difficult to recover $\mathbf{X}$. For explicit CoT (red line), we can always perfectly recover the original data $\mathbf{X}$ from the intermediate sequences. This supports our hypothesis that implicit CoT gradually loses information about the original data and fails to simulate the GD algorithm, while explicit CoT preserves this information and succeeds.

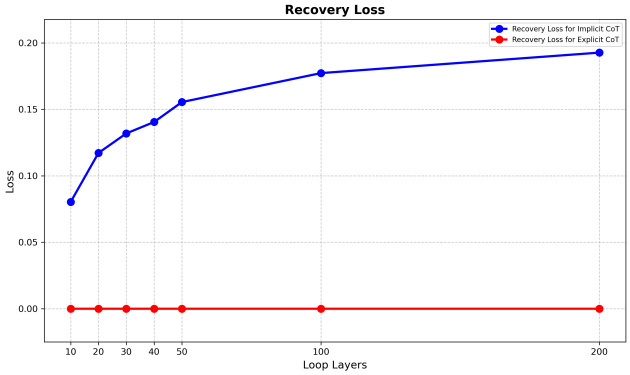


Figure 2: The loss for recovering $\mathbf{X}$ given output matrix after L steps $\hat{\mathbf{Z}}_L$ with $10 \leq L \leq 200$ for the weight prediction task.

4 In-Context Query Prediction

In this section, we consider the in-context query prediction task. The data sequence $\{\mathbf{x}_i, y_i\}_{i=1}^N$ is sampled from $y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \xi_i$ with $\boldsymbol{\beta} \sim N(0, \mathbf{I}_d)$, $\mathbf{x}_i \sim N(0, \mathbf{I}_d)$, and the noise $\xi_i \sim N(0, \sigma^2)$. Then, the input matrix is designed as follows

$$\mathbf{Z}_0 = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_N & \mathbf{x}_q \\ y_1 & y_2 & \dots & y_N & y_{q0} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{y}^\top & y_{q0} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(N+1)},$$

where $y_{q0} = \boldsymbol{\beta}_0^\top \mathbf{x}_q$ is the initialization of the query label. The experiment setup in this section is the same as that in Section 3.

4.1 Implicit CoT

We read the bottom-right entry of the matrix output as the prediction, and define the population training loss as

$$\mathcal{L}(\mathbf{W}^V, \mathbf{W}^{KQ}) = \mathbb{E}_{\mathbf{x}_i, \mathbf{x}_q, \boldsymbol{\beta}_0} \left[(f_{\text{LSA}}(\mathbf{Z}_0)_{[-1,-1]} - \boldsymbol{\beta}_1^\top \mathbf{x}_q)^2 \right],$$

where the initialization $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I}_d)$, and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 - \eta_{gd} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})$ is the ground-truth weight vector after one gradient step on the linear regression objective with a fixed learning rate η_{gd}.

Figure 3a illustrates the experimental results obtained by unrolling the trained model. The red curve shows that the difference between the model's prediction and the teacher GD output remains large, indicating that the learned model with implicit CoT fails to simulate the teacher algorithm.

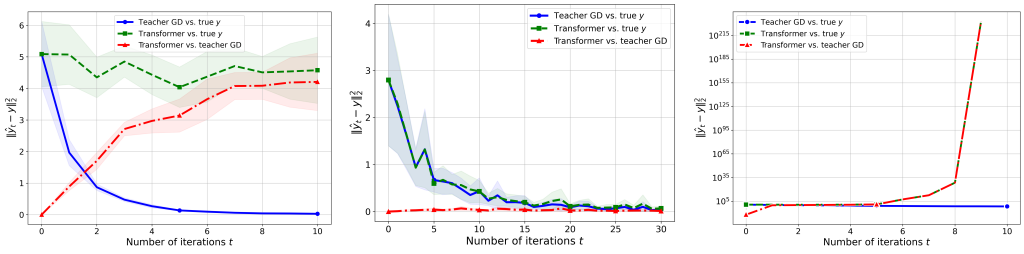
However, some recent works Von Oswald et al. (2023); Ahn et al. (2023); Gatmiry et al. (2024) observed that by artificially adding a minus sign to the output, a one-layer LSA model with implicit CoT can successfully implement a multi-step GD algorithm. In other words, we expect that the intermediate states can simulate the negative value of the ground-truth GD steps. The corresponding training loss is defined as

$$\mathcal{L}_{-1}(\mathbf{W}^{KQ}, \mathbf{W}^V) = \mathbb{E}_{\mathbf{x}_i, \mathbf{x}_q, \boldsymbol{\beta}_0} \left[(f_{\text{LSA}}(\mathbf{Z}_0)_{[-1,-1]} + \boldsymbol{\beta}_1^\top \mathbf{x}_q)^2 \right].$$

Figure 3b empirically shows that the modified model successfully matches the GD steps. The expressivity results and convergence guarantees are similar to Gatmiry et al. (2024), and we show the expressivity result here.

Theorem 4.1. *There exist parameter matrices $\mathbf{W}^{KQ}$ and $\mathbf{W}^{V}$ such that f_{LSA} with implicit CoT and without scratchpad can implement a multi-step GD algorithm for in-context query prediction task.*

In addition, we also provide the positive convergence results (see Appendix C.4).



(a) Implicit CoT (b) Implicit CoT with a minus sign (c) Explicit CoT

Figure 3: Results on query prediction without scratchpad. Lines denote differences between: model prediction and teacher GD (red), model prediction and ground-truth label (green), and teacher GD and ground-truth label (blue). (a) corresponds to implicit CoT; (b) corresponds to implicit CoT with a minus sign; (c) corresponds to explicit CoT. (b) is evaluated for $L = 30$ steps, while (a) and (c) are evaluated for $L = 10$ steps.

4.2 Explicit CoT

For explicit CoT, we can generate L intermediate sequences $\{\hat{\mathbf{Z}}_i\}_{i=1}^{L}$ as follows:

$$\hat{\mathbf{Z}}_i = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q & \mathbf{x}_q & \dots & \mathbf{x}_q \\ \mathbf{y}^\top & y_{q0} & \hat{y}_{q1} & \dots & \hat{y}_{qi} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(N+i+1)}, i \in [L].$$

We read the bottom-right entry as the prediction and design the CoT training objective as

$$\ell^{\text{CoT}}(\mathbf{X}, \boldsymbol{\beta}; \mathbf{W}^V, \mathbf{W}^{KQ}) = \sum_{i=0}^{L-1} (f_{\text{LSA}}(\mathbf{Z}_i)_{[-1,-1]} - \boldsymbol{\beta}_{i+1}^\top \mathbf{x}_q)^2,$$

where $\{\mathbf{Z}_i\}_{i=1}^{L}$ are the ground-truth sequences, $\boldsymbol{\beta}_i$ is the weight at the i-th GD step. Further, the corresponding population loss is

$$\mathcal{L}^{\text{CoT}}(\mathbf{W}^V, \mathbf{W}^{KQ}) = \mathbb{E}_{\mathbf{x}_i, \boldsymbol{\beta}} \left[\ell^{\text{CoT}}(\mathbf{X}, \boldsymbol{\beta}; \mathbf{W}^V, \mathbf{W}^{KQ}) \right].$$

Similar to Figure 3a, Figure 3c also shows the failure of explicit CoT. In Figure 3c, the difference between the prediction and the teacher GD update rapidly blows up within 10 steps. In addition, we can get the following theoretical results about the expressivity of explicit CoT.

Theorem 4.2. *Assume that $\mathbf{x}_i \sim N(0, \mathbf{I}_d)$, $\mathbf{x}_q \sim N(0, \mathbf{I}_d)$, $\mathbf{y} \sim N(0, \mathbf{I}_N)$, $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I}_d)$ and $\boldsymbol{\beta}_1$ is the weight vector after one GD step from $\boldsymbol{\beta}_0$ with learning rate η_{gd}. For any $f_{LSA}(\cdot)$, it holds that*

$$\min_{c \in \mathbb{R}} \mathbb{E} \left[(c \cdot f_{LSA}(\mathbf{Z}_0)_{[-1,-1]} - \boldsymbol{\beta}_1^\top \mathbf{x}_q)^2 \right] \geq \eta_{gd}^2 N d.$$

Theorem 4.2 shows that under certain conditions, explicit CoT cannot implement a single GD step for query prediction. Moreover, the failure persists even if we allow an arbitrary scalar multiplier c on the output. In particular, the sign-changing operation considered in the implicit CoT paradigm does not make explicit CoT succeed this time. Similar to Theorem 3.1, the assumption $\mathbf{y} \sim N(0, \mathbf{I}_N)$ is used for analytical simplification. This already provides enough theoretical evidence that the failure is not accidental but stems from an inherent expressive limitation.

4.3 Limited Expressivity for State Updates

We analyze why explicit CoT fails on in-context query prediction. Simulating multi-step GD for query prediction requires computing $y_{q,i+1} = y_{q,i} - \eta_{gd} \boldsymbol{\beta}_i^\top \mathbf{X}^\top \mathbf{X} \mathbf{x}_q + \eta_{gd} \mathbf{y}^\top \mathbf{X} \mathbf{x}_q$. Thus, each step requires both the current projection $\boldsymbol{\beta}_i^\top \mathbf{X}^\top \mathbf{X} \mathbf{x}_q$ and the fixed term $\mathbf{y}^\top \mathbf{X} \mathbf{x}_q$. Consequently, the model must simultaneously track the intermediate weight vector $\boldsymbol{\beta}_i$ (or relevant terms) and compute the scalar prediction $y_{q,i} = \boldsymbol{\beta}_i^\top \mathbf{x}_q$. Explicit CoT fails here because autoregressive generation updates only a single token at a time, lacking the spatial capacity for parallel state tracking.

In contrast, implicit CoT overcomes this expressivity bottleneck by updating the entire representation matrix at once as

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{y}^\top - \boldsymbol{\beta}_i^\top \mathbf{X}^\top & -\boldsymbol{\beta}_i^\top \mathbf{x}_q \end{bmatrix}.$$

This allows different blocks to track evolving quantities in parallel. Notably, the bottom-left block preserves the original $\mathbf{y}$ while storing the projection $\boldsymbol{\beta}_i^\top \mathbf{X}^\top$, effectively avoiding the context overwriting issue (discussed in Section 3.3) and demonstrating a structural advantage over explicit CoT.

5 Adding a Scratchpad

In this section, we find that the situation changes when we introduce an additional scratchpad, i.e., extra rows appended to each token in the input matrix. In our theoretical formulation, adding a scratchpad is essentially a change of encoding, which can be interpreted as extra dimensions. The amount of scratchpad required is task-specific and depends on the representation and the auxiliary quantities needed for the underlying update. We show how structured auxiliary space can remove the expressivity obstructions in this controlled setting and discuss the influence of this scratchpad on both CoT paradigms. The experiment settings are the same as those in Section 3.

5.1 In-Context Weight Prediction

For in-context weight prediction, we modify only the format of the input matrix, keeping all other settings identical to Section 3. The new embedding includes extra d rows that serve as a scratchpad:

$$\mathbf{Z}_0 = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_N & 0 \\ y_1 & y_2 & \dots & y_N & 0 \\ 0 & 0 & \dots & 0 & \boldsymbol{\beta}_0 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top & 0 \\ \mathbf{y}^\top & 0 \\ 0 & \boldsymbol{\beta}_0 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{(2d+2)\times(N+1)}.$$

Figure 4a and 4b illustrate that both implicit and explicit CoT paradigms succeed after adding the scratchpad, with the trained model strictly aligned with the teacher GD algorithm (red lines near zero). The success of explicit CoT is expected, as it already worked without the scratchpad in Section 3. This setting was also considered by Huang et al. (2025a). However, implicit CoT now succeeds as well, in contrast to its previous failure. As discussed in Section 3.3, the original failure of implicit CoT is mainly due to context overwriting, while the added scratchpad overcomes this issue since the prediction is now in a separate row from $\mathbf{X}$ and $\mathbf{y}$, preserving the original data throughout the updates and thus enhancing the expressivity of implicit CoT. This leads to the following theoretical result regarding the expressivity.

Theorem 5.1. *There exist parameter matrices $\mathbf{W}^{KQ}$ and $\mathbf{W}^{V}$ such that f_{LSA} with implicit CoT and with scratchpad can implement a multi-step GD algorithm for the weight prediction task.*

The analysis of training guarantees is similar to that in Section 3.2. We also analyze the convergence results and OOD generalization bound (See Appendix C.1). Interestingly, we find a construction to transform explicit CoT into implicit CoT with a scratchpad.

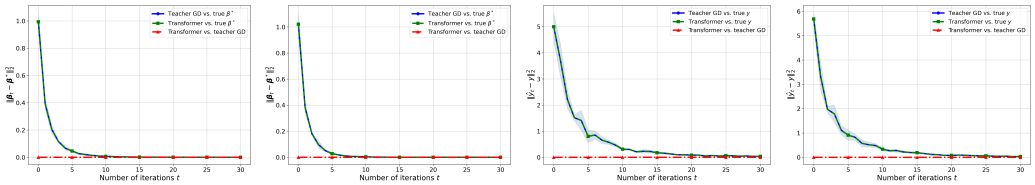
Theorem 5.2. *Any explicit CoT reasoning algorithm generated by a single-layer self-attention model can be simulated by a layer-recurrent implicit CoT model with an additional scratchpad.*

This theorem indicates that under a specific construction, implicit CoT with a scratchpad can exactly implement any algorithm generated by the explicit CoT model. Note that although we focus on linear attention, this result can be applied to softmax attention by replacing the mask from zero with negative infinity; see Appendix B.3 for details.

5.2 In-Context Query Prediction

In Section 4, we have demonstrated that by manually applying a sign-flipping operation to the final output, implicit CoT succeeds in simulating multi-step GD. However, we expect to make the model more natural by removing the artificial minus sign. We now augment the input matrix with a scratchpad as

$$\mathbf{Z}_0 = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{y}^\top & 0 \\ \boldsymbol{\beta}_0^\top \mathbf{X}^\top & \boldsymbol{\beta}_0^\top \mathbf{x}_q \end{bmatrix} \in \mathbb{R}^{(d+2)\times(N+1)}.$$



(a) Implicit CoT (Weight) (b) Explicit CoT (Weight) (c) Implicit CoT (Query) (d) Explicit CoT (Query)

Figure 4: Results with scratchpad on weight prediction and query prediction, evaluated for $L = 30$ steps. Lines denote differences between: model prediction and teacher GD (red), model prediction and ground-truth label (green), and teacher GD and ground-truth label (blue). (a) corresponds implicit CoT on weight prediction; (b) corresponds explicit CoT on weight prediction; (c) corresponds implicit CoT on query prediction; (d) corresponds explicit CoT on query prediction.

Here, we expect the bottom-right element to predict GD updates directly without additional operation. Figure 4c illustrates the success of implicit CoT with a scratchpad. Unlike storing $\boldsymbol{\beta}_i^\top \mathbf{X}^\top$ and $\mathbf{y}$ together as in Section 4.3, the extra scratchpad provides additional space to track $\boldsymbol{\beta}_i^\top \mathbf{X}^\top$ and $\boldsymbol{\beta}_i^\top \mathbf{x}_q$, thereby eliminating the need for the sign-flipping operation.

For explicit CoT, we expect the generated intermediate sequences to simulate

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q & \mathbf{x}_q & \cdots & \mathbf{x}_q \\ \mathbf{y}^\top & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{u}_0 & \mathbf{u}_1 & \cdots & \mathbf{u}_i \\ 0 & 0 & \boldsymbol{\beta}_1^\top \mathbf{x}_q & \cdots & \boldsymbol{\beta}_i^\top \mathbf{x}_q \end{bmatrix} \in \mathbb{R}^{(2d+2)\times(N+i+1)}, i \in [L],$$

where $\mathbf{u}_i = (\mathbf{I}_d - \eta_{gd}\mathbf{X}^\top\mathbf{X})^i \mathbf{x}_q$ is the auxiliary information we track during the generation. Figure 4d shows the success of explicit CoT with the scratchpad. Here, the extra scratchpad provides space for storing the auxiliary information $\mathbf{u}_i$ and the prediction $\boldsymbol{\beta}_i \mathbf{x}_q$, which overcomes the limited expressivity issue discussed in Section 4.3.

Similar to Section 3.2, we also provide the theoretical results of expressivity and training guarantees for implicit CoT and explicit CoT (See Appendix C.3 and Appendix C.5, respectively).

6 Conclusion and Limitation

In this paper, we compared implicit and explicit chain-of-thought reasoning in a single-layer linear self-attention model on two tasks, in-context weight prediction and in-context query prediction. In this stylized setting, we found that the two paradigms exhibit different failure modes: implicit CoT fails on weight prediction because recurrent updates overwrite information needed from the original input, while explicit CoT fails on query prediction because autoregressive generation provides insufficient writable state for tracking all required intermediate quantities. We also showed that adding a scratchpad can enlarge the available state space and alleviate these bottlenecks, with theorem-level support for some regimes and empirical support for others. In addition, we provided a construction showing that explicit CoT can be simulated by implicit CoT with a scratchpad. We view these results as a step toward a more fine-grained theory of how internal state allocation shapes multi-step reasoning in transformers.

Our analysis is conducted in a stylized single-layer linear attention setting, which should be understood as a controlled theoretical testbed rather than a direct characterization of all practical transformer architectures or real-world language models. The training guarantees rely on the simplifying assumption that the query and key matrices can be reparameterized as a single trainable matrix. Extending the full training analysis to more practical transformer architectures remains an important direction for future work. However, we remark that some of our results can be extended to more complex transformer architectures. For positive expressivity results, richer linear-attention architectures with separate $\mathbf{Q}/\mathbf{K}/\mathbf{V}$ projections, multiple heads, or MLP blocks can simulate our constructions by setting the additional components to identity or zero maps. In addition, we believe linear attention is informative as a theoretical testbed, since recent work Ahn et al. (2024) has shown that linear attention transformers can already capture important optimization phenomena observed in practical transformer training.

AI USE STATEMENT

We used generative AI tools to edit the manuscript for readability and language polish. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proofs, write proofs, propose or refine hypotheses, design or provide feedback on the methodology or experiments, implement methods, or interpret results. Dataset cleaning or reformatting and qualitative or thematic data analysis were not applicable to this work. The authors reviewed all AI-assisted edits and take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is a theoretical study and evaluates results on synthetically generated data. It does not involve human participants, personal or sensitive data, or interventions affecting individuals. We do not anticipate ethical risks concerning privacy, consent, or demographic bias.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 2 specifies the data distributions, teacher gradient-descent updates, model, and training setup for the linear experiments. Results are reported as means and standard deviations over five independent random seeds. The appendix provides the assumptions and proofs for the theoretical results, along with the experiment setup of the LEGO task. All experiments use synthetic data and do not depend on private or external datasets.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *International Conference on Learning Representations*, volume 2024, pp. 16193–16205, 2024.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36:57125–57211, 2023.

Yilong Chen, Junyuan Shang, Zhenyu Zhang, Yanxi Xie, Jiawei Sheng, Tingwen Liu, Shuohuan Wang, Yu Sun, Hua Wu, and Haifeng Wang. Inner thinking transformer: Leveraging dynamic depth scaling to foster adaptive internal thinking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28241–28259, 2025.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598, 2022.

Khashayar Gatmiry, Nikunj Saunshi, Sashank J Reddi, Stefanie Jegelka, and Sanjiv Kumar. Can looped transformers learn to implement multi-step gradient descent for in-context learning? In *International Conference on Machine Learning*, pp. 15130–15152. PMLR, 2024.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *Advances in Neural Information Processing Systems*, 38: 41340–41391, 2026.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. *arXiv preprint arXiv:2310.02226*, 2023.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 14852–14882, 2023.

Jianhao Huang, Zixuan Wang, and Jason D Lee. Transformers learn to implement multi-step gradient descent with chain of thought. *arXiv preprint arXiv:2502.21212*, 2025a.

Yu Huang, Zixin Wen, Aarti Singh, Yuejie Chi, and Yuxin Chen. Transformers provably learn chain-of-thought reasoning with length generalization. *arXiv preprint arXiv:2511.07378*, 2025b.

M Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to markov models: Unveiling the dynamics of generative transformers. *arXiv preprint arXiv:2402.13512*, 2024.

Adel Javanmard, Baharan Mirzasoleiman, and Vahab Mirrokni. Understanding the role of training data in test-time scaling. In *International Conference on Learning Representations*, volume 2026, pp. 8566–8588, 2026.

Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1773–1781, 2023.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. Refiner: Reasoning feedback on intermediate representations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1100–1126, 2024.

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

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*, 2022.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J Reddi. Reasoning with latent thoughts: On the power of looped transformers. In *International Conference on Learning Representations*, volume 2025, pp. 14855–14881, 2025.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227. PMLR, 2023.

Yuchang Sun, Yanxi Chen, Yaliang Li, and Bolin Ding. Enhancing latent computation in transformers with latent tokens. *arXiv preprint arXiv:2505.12629*, 2025.

Jihoon Tack, Jack Lanchantin, Jane Yu, Andrew Cohen, Ilia Kulikov, Janice Lan, Shibo Hao, Yuandong Tian, Jason Weston, and Xian Li. Llm pretraining with continuous concepts. *arXiv preprint arXiv:2502.08524*, 2025.

Kaito Takanami and Cengiz Pehlevan. An asymptotic theory of chain-of-thought in in-context learning. *arXiv preprint arXiv:2606.03217*, 2026.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in neural information processing systems*, 36:71911–71947, 2023.

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

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

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2717–2739, 2023a.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36:74764–74786, 2023b.

Zixuan Wang, Stanley Wei, Daniel Hsu, and Jason D Lee. Transformers provably learn sparse token selection while fully-connected nets cannot. *arXiv preprint arXiv:2406.06893*, 2024.

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

Wen Yang, Minpeng Liao, and Kai Fan. Markov chain of thought for efficient mathematical reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7132–7157, 2025.

Jiaran Ye, Zijun Yao, Zhidian Huang, Liangming Pan, Jinxin Liu, Yushi Bai, Amy Xin, Liu Weichuan, Xiaoyin Che, Lei Hou, et al. How does transformer learn implicit reasoning? *arXiv e-prints*, pp. arXiv–2505, 2025.

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

Yi Zhang, Arturs Backurs, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, and Tal Wagner. Unveiling transformers with lego: a synthetic reasoning task. *arXiv preprint arXiv:2206.04301*, 2022.

A Related Works

Chain-of-Thought. Chain-of-Thought (CoT) prompting (Wei et al., 2022) has shown great multi-step reasoning ability. Explicit CoT requires models to generate observable intermediate reasoning tokens before arriving at a final answer, which has been shown to significantly improve performance on complex tasks (Wang et al., 2023a; Saparov & He, 2022; Shi et al., 2022; Paul et al., 2024). Recently, to mitigate the high inference latency and computational costs associated with generating long explicit reasoning traces, researchers have explored implicit CoT (Ye et al., 2025). In the implicit CoT paradigm, reasoning steps are internalized within the hidden continuous representations of the model (Tack et al., 2025; Sun et al., 2025; Hao et al., 2024; Deng et al., 2023; Goyal et al., 2023). Layer-recurrent architectures, which repeatedly apply the same transformer block, have been actively investigated as a mechanism to achieve this implicit internal computation (Chen et al., 2025; Saunshi et al., 2025; Geiping et al., 2026).

In-Context Learning and Algorithm Simulation. To theoretically understand transformers, a rich line of work studies their ability to perform In-Context Learning (Garg et al., 2022). Transformers solve these tasks by internally simulating optimization algorithms, most notably Gradient Descent (Von Oswald et al., 2023; Ahn et al., 2023). In our study, we utilize a single-layer Linear Self-Attention (LSA) model, a simplified architecture widely adopted for theoretical need (Jelassi et al., 2022; Tian et al., 2023; Zhang et al., 2024; Wang et al., 2024; Ildiz et al., 2024). Prior works have shown that LSA models can implement GD for linear regression, sometimes requiring artificial modifications such as appending a minus sign to the output (Gatmiry et al., 2024). We extend this literature by analyzing how different CoT paradigms simulate multi-step GD on two distinct in-context tasks: weight prediction (Huang et al., 2025a) and query prediction (Garg et al., 2022).

Closely related to our work, Huang et al. (2025a) shows that transformers can implement multi-step GD with autoregressive CoT and scratchpads for in-context weight prediction. This corresponds to one case in our broader comparison, and we view this setting as a supporting component rather than as our novelty. Our contribution is to formulate a unified controlled framework comparing eight cases across implicit/explicit CoT, weight/query prediction, and with/without scratchpads. Takanami & Pehlevan (2026) develop an asymptotic theory of CoT in ICL, characterizing how explicit intermediate reasoning changes the limiting behavior of in-context prediction. Javanmard et al. (2026) study test-time scaling from the data perspective, analyzing how properties of the training data influence the benefit of additional test-time computation. In contrast, our focus is on the comparative expressivity and training mechanisms of two concrete CoT formalizations under a fixed GD teacher algorithm. In addition, Nye et al. (2021) provides important empirical evidence that scratchpads support intermediate computation in language models. Our work is complementary. To our knowledge, our paper is the first to provide a systematic theoretical analysis of how scratchpads affect both implicit and explicit CoT paradigms in a unified controlled framework.

B Expressivity

B.1 Weight Prediction

For implicit CoT without scratchpad, we can obtain the following result.

Theorem B.1. *Assume that $\mathbf{x}_i \sim N(0, \mathbf{I}_d)$, $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I}_d)$, $\mathbf{y} \sim N(0, \mathbf{I}_N)$, and $\boldsymbol{\beta}_i$ is the ground-truth weight vector after i GD steps from $\boldsymbol{\beta}_0$ with learning rate η_{gd}. For any $f_{LSA}(\cdot)$ such that $f_{LSA}(\mathbf{Z}_0)_{[:d,-1]} \equiv \boldsymbol{\beta}_1$, it holds that*

$$\mathbb{E}\left[\|f_{LSA}^{(2)}(\mathbf{Z}_0)_{[:d,-1]} - \boldsymbol{\beta}_2\|_2^2\right] \geq \eta_{gd}^4 N d.$$

Proof. Denote that

$$\mathbf{W}^{KQ} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} & \mathbf{W}_{13} \\ \mathbf{W}_{21} & w_{22} & w_{23} \\ \mathbf{W}_{31} & w_{32} & w_{33} \end{bmatrix}, \mathbf{W}^{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} \\ \mathbf{V}_{21} & v_{22} & v_{23} \\ \mathbf{V}_{31} & v_{32} & v_{33} \end{bmatrix}, f_{\text{LSA}}^{(L)}(\mathbf{Z}_0) = \begin{bmatrix} \mathbf{Z}_1^{(L)} & \mathbf{Z}_2^{(L)} \\ \mathbf{Z}_3^{(L)} & \mathbf{Z}_4^{(L)} \\ \mathbf{Z}_5^{(L)} & \mathbf{Z}_6^{(L)} \end{bmatrix}.$$

We can obtain that

$$\begin{aligned} f_{\text{LSA}}(\mathbf{Z}_0)_{[:d,-1]} &= \boldsymbol{\beta}_0 + (\mathbf{V}_{11}\mathbf{X}^\top + \mathbf{V}_{12}\mathbf{y}^\top) \cdot (\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_0 + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_0 + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}w_{23}) \\ &= \boldsymbol{\beta}_0 + (\mathbf{V}_{11}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{11} + \mathbf{V}_{11}\mathbf{X}^\top\mathbf{y}\mathbf{W}_{21} + \mathbf{V}_{12}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{11} + \mathbf{V}_{12}\mathbf{y}^\top\mathbf{y}\mathbf{W}_{21})\boldsymbol{\beta}_0 \end{aligned}$$

$$+ \mathbf{V}_{11}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13} + \mathbf{V}_{11}\mathbf{X}^\top\mathbf{y}w_{23} + \mathbf{V}_{12}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{13} + \mathbf{V}_{12}\mathbf{y}^\top\mathbf{y}w_{23}.$$

The ground truth weight vector $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 - \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y}) = \boldsymbol{\beta}_0 - \eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0 + \eta_{gd}\mathbf{X}^\top\mathbf{y}$. Under the condition that $f_{\text{LSA}}(\mathbf{Z}_0)_{[:d,-1]} \equiv \boldsymbol{\beta}_1$, we can compare the coefficient of $\boldsymbol{\beta}_0$ to get

$$\mathbf{V}_{11}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{11} + \mathbf{V}_{11}\mathbf{X}^\top\mathbf{y}\mathbf{W}_{21} + \mathbf{V}_{12}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{11} + \mathbf{V}_{12}\mathbf{y}^\top\mathbf{y}\mathbf{W}_{21} \equiv -\eta_{gd}\mathbf{X}^\top\mathbf{X},$$
$$\mathbf{V}_{11}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13} + \mathbf{V}_{11}\mathbf{X}^\top\mathbf{y}w_{23} + \mathbf{V}_{12}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{13} + \mathbf{V}_{12}\mathbf{y}^\top\mathbf{y}w_{23} \equiv \eta_{gd}\mathbf{X}^\top\mathbf{y}.$$

Then, based on the number of appearances of $\mathbf{X}$, we can obtain that

$$\mathbf{V}_{11}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{11} \equiv -\eta_{gd}\mathbf{X}^\top\mathbf{X}, \tag{1}$$
$$\mathbf{V}_{11}\mathbf{X}^\top\mathbf{y}\mathbf{W}_{21} + \mathbf{V}_{12}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{11} \equiv 0, \tag{2}$$
$$\mathbf{V}_{12}\mathbf{y}^\top\mathbf{y}\mathbf{W}_{21} \equiv 0, \tag{3}$$
$$\mathbf{V}_{11}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13} \equiv 0, \tag{4}$$
$$\mathbf{V}_{11}\mathbf{X}^\top\mathbf{y}w_{23} + \mathbf{V}_{12}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{13} \equiv \eta_{gd}\mathbf{X}^\top\mathbf{y}, \tag{5}$$
$$\mathbf{V}_{12}\mathbf{y}^\top\mathbf{y}w_{23} \equiv 0. \tag{6}$$

We first prove that $w_{23} \neq 0$. If $w_{23} = 0$, from equation 5, we can get $\mathbf{V}_{12}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{13} \equiv \eta_{gd}\mathbf{X}^\top\mathbf{y}$. Then, $\mathbf{V}_{12}\mathbf{W}_{13}^\top\mathbf{X}^\top\mathbf{y} = \mathbf{V}_{12}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{13} \equiv \eta_{gd}\mathbf{X}^\top\mathbf{y}$, which indicates that $\mathbf{V}_{12}\mathbf{W}_{13}^\top = \eta_{gd}\mathbf{I}_d$. However, $\mathbf{V}_{12}\mathbf{W}_{13}^\top$ cannot be $\eta_{gd}\mathbf{I}_d$ since the vectors $\mathbf{V}_{12}, \mathbf{W}_{13}$ ensure that the rank of $\mathbf{V}_{12}\mathbf{W}_{13}^\top$ is 1. This is a contradiction. Therefore, we have proven that $w_{23} \neq 0$. Then, by equation 6, we get $\mathbf{V}_{12}w_{23} = 0$. With $w_{23} \neq 0$, we can get $\mathbf{V}_{12} = 0$. Plugging $\mathbf{V}_{12} = 0$ into equation 5, we get $w_{23}\mathbf{V}_{11} = \eta_{gd}\mathbf{I}_d$. So, we can denote $\mathbf{V}_{11} = v_{11}\mathbf{I}_d$ with $v_{11} \neq 0$, and $v_{11}w_{23} = \eta_{gd}$. Plugging $\mathbf{V}_{11} = v_{11}\mathbf{I}_d$ into equation 4, we get $\mathbf{W}_{13} = 0$. Plugging $\mathbf{V}_{12} = 0, \mathbf{V}_{11} = v_{11}\mathbf{I}_d$ into equation 2, we get $\mathbf{W}_{21} = 0$. Plugging $\mathbf{V}_{11} = v_{11}\mathbf{I}_d$ into equation 1, we can get $v_{11}\mathbf{W}_{11} = -\eta_{gd}\mathbf{I}_d$. So, we can denote $\mathbf{W}_{11} = w_{11}\mathbf{I}_d$ with $w_{11} \neq 0$, and $v_{11}w_{11} = -\eta_{gd}$.

As a result, under the condition that $f_{\text{LSA}}(\mathbf{Z}_0)_{[:d,-1]} \equiv \boldsymbol{\beta}_1$, we can obtain the following results:

$$\mathbf{V}_{11} = v_{11}\mathbf{I}_d, \mathbf{W}_{11} = w_{11}\mathbf{I}_d, v_{11}w_{23} = \eta_{gd}, v_{11}w_{11} = -\eta_{gd}, \mathbf{V}_{12} = 0, \mathbf{W}_{13} = 0, \mathbf{W}_{21} = 0.$$

Based on these results, we can get

$$\mathbf{Z}_1^{(1)} = \mathbf{X}^\top - \eta_{gd}\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top + v_{11}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{12}\mathbf{y}^\top + w_{22}\mathbf{y}\mathbf{y}^\top) := \mathbf{X}^\top + \mathbf{S}^\top,$$
$$\mathbf{Z}_2^{(1)} = \boldsymbol{\beta}_0 - \eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0 + \eta_{gd}\mathbf{X}^\top\mathbf{y} = \boldsymbol{\beta}_1,$$
$$\mathbf{Z}_3^{(1)} = \mathbf{y}^\top, \mathbf{Z}_4^{(1)} = 0, \mathbf{Z}_5^{(1)} = 0, \mathbf{Z}_6^{(1)} = 1.$$

Here, we denote that $\mathbf{S} = -\eta_{gd}\mathbf{X}\mathbf{X}^\top\mathbf{X} + v_{11}\mathbf{y}\mathbf{W}_{12}^\top\mathbf{X}^\top\mathbf{X} + v_{11}w_{22}\mathbf{y}\mathbf{y}^\top\mathbf{X}$. Then, we can obtain

$$\begin{aligned}\mathbf{Z}_2^{(2)} &= \mathbf{Z}_2^{(1)} - \eta_{gd}\mathbf{Z}_1^{(1)}\mathbf{Z}_1^{(1)\top}\mathbf{Z}_2^{(1)} + \eta_{gd}\mathbf{Z}_1^{(1)}\mathbf{y} \\ &= \boldsymbol{\beta}_1 - \eta_{gd}(\mathbf{X}^\top + \mathbf{S}^\top)(\mathbf{X} + \mathbf{S})\boldsymbol{\beta}_1 + \eta_{gd}(\mathbf{X}^\top + \mathbf{S}^\top)\mathbf{y} \\ &= \boldsymbol{\beta}_1 - \eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_1 + \eta_{gd}\mathbf{X}^\top\mathbf{y} - \eta_{gd}(\mathbf{X}^\top\mathbf{S} + \mathbf{S}^\top\mathbf{X} + \mathbf{S}^\top\mathbf{S})\boldsymbol{\beta}_1 + \eta_{gd}\mathbf{S}^\top\mathbf{y}.\end{aligned}$$

where $\mathbf{S} = -\eta_{gd}\mathbf{X}\mathbf{X}^\top\mathbf{X} + v_{11}\mathbf{y}\mathbf{W}_{12}^\top\mathbf{X}^\top\mathbf{X} + v_{11}w_{22}\mathbf{y}\mathbf{y}^\top\mathbf{X}$. Thus, we can get the lower bound of $\mathbb{E}\left[\|\mathbf{Z}_2^{(2)} - \boldsymbol{\beta}_2\|_2^2\right]$ as follows:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{Z}_2^{(2)} - \boldsymbol{\beta}_2\|_2^2\right] &= \mathbb{E}\left[\left\|-\eta_{gd}(\mathbf{X}^\top\mathbf{S} + \mathbf{S}^\top\mathbf{X} + \mathbf{S}^\top\mathbf{S})\boldsymbol{\beta}_1 + \eta_{gd}\mathbf{S}^\top\mathbf{y}\right\|_2^2\right] \\ &= \mathbb{E}\left[\left\|-\eta_{gd}(\mathbf{X}^\top\mathbf{S} + \mathbf{S}^\top\mathbf{X} + \mathbf{S}^\top\mathbf{S})(\mathbf{I}_d - \eta_{gd}\mathbf{X}^\top\mathbf{X})\boldsymbol{\beta}_0 + \eta_{gd}\mathbf{S}^\top\mathbf{y}\right\|_2^2\right] \\ &= \mathbb{E}\left[\left\|-\eta_{gd}(\mathbf{X}^\top\mathbf{S} + \mathbf{S}^\top\mathbf{X} + \mathbf{S}^\top\mathbf{S})(\mathbf{I}_d - \eta_{gd}\mathbf{X}^\top\mathbf{X})\boldsymbol{\beta}_0\right\|_2^2\right] + \mathbb{E}\left[\|\eta_{gd}\mathbf{S}^\top\mathbf{y}\|_2^2\right] \\ &\geq \mathbb{E}\left[\|\eta_{gd}\mathbf{S}^\top\mathbf{y}\|_2^2\right].\end{aligned}$$

Next, we are going to provide a lower bound of $\mathbb{E}[\|\eta_{gd}\mathbf{S}^\top\mathbf{y}\|_2^2]$. We can get

$$\begin{aligned}&\mathbb{E}[\|\eta_{gd}\mathbf{S}^\top\mathbf{y}\|_2^2] \\ &= \eta_{gd}^2\mathbb{E}[\mathbf{y}^\top\mathbf{S}\mathbf{S}^\top\mathbf{y}] \\ &= \eta_{gd}^2\mathbb{E}\Big[\eta_{gd}^2\mathbf{y}^\top(\mathbf{X}\mathbf{X}^\top)^3\mathbf{y} - \eta_{gd}v_{11}\mathbf{y}^\top(\mathbf{X}\mathbf{X}^\top)^2\mathbf{X}\mathbf{W}_{12}\mathbf{y}^\top\mathbf{y} - \eta_{gd}v_{11}w_{22}\mathbf{y}^\top(\mathbf{X}\mathbf{X}^\top)^2\mathbf{y}\mathbf{y}^\top\mathbf{y}\end{aligned}$$

$$
\begin{aligned}
&- \eta_{gd} v_{11} \mathbf{y}^\top \mathbf{y} \mathbf{W}_{12}^\top \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^2 \mathbf{y} + v_{11}^2 \mathbf{y}^\top \mathbf{y} \mathbf{W}_{12}^\top (\mathbf{X}^\top \mathbf{X})^2 \mathbf{W}_{12} \mathbf{y}^\top \mathbf{y} \\
&+ v_{11}^2 w_{22} \mathbf{y}^\top \mathbf{y} \mathbf{W}_{12}^\top \mathbf{X}^\top \mathbf{X} \mathbf{X}^\top \mathbf{y} \mathbf{y}^\top \mathbf{y} - \eta_{gd} v_{11} w_{22} \mathbf{y}^\top \mathbf{y} \mathbf{y}^\top (\mathbf{X}\mathbf{X}^\top)^2 \mathbf{y} \\
&+ v_{11}^2 w_{22} \mathbf{y}^\top \mathbf{y} \mathbf{y}^\top \mathbf{X} \mathbf{X}^\top \mathbf{X} \mathbf{W}_{12} \mathbf{y}^\top \mathbf{y} + v_{11}^2 w_{22}^2 \mathbf{y}^\top \mathbf{y} \mathbf{y}^\top \mathbf{X} \mathbf{X}^\top \mathbf{y} \mathbf{y}^\top \mathbf{y} \Big] \\
&\overset{(i)}{=} \eta_{gd}^2 \mathbb{E} \Big[\eta_{gd}^2 \mathbf{y}^\top (\mathbf{X}\mathbf{X}^\top)^3 \mathbf{y} - \eta_{gd} v_{11} w_{22} \mathbf{y}^\top (\mathbf{X}\mathbf{X}^\top)^2 \mathbf{y} \mathbf{y}^\top \mathbf{y} + v_{11}^2 \mathbf{y}^\top \mathbf{y} \mathbf{W}_{12}^\top (\mathbf{X}^\top \mathbf{X})^2 \mathbf{W}_{12} \mathbf{y}^\top \mathbf{y} \\
&- \eta_{gd} v_{11} w_{22} \mathbf{y}^\top \mathbf{y} \mathbf{y}^\top (\mathbf{X}\mathbf{X}^\top)^2 \mathbf{y} + v_{11}^2 w_{22}^2 \mathbf{y}^\top \mathbf{y} \mathbf{y}^\top \mathbf{X} \mathbf{X}^\top \mathbf{y} \mathbf{y}^\top \mathbf{y} \Big] \\
&\overset{(ii)}{=} \eta_{gd}^2 \Big[\eta_{gd}^2 \left(Nd^3 + 3N(N+1)d^2 + N(N^2+3N+4)d \right) - 2\eta_{gd} v_{11} w_{22} (N+2) N d (d+N+1) \\
&+ v_{11}^2 \|\mathbf{W}_{12}\|_2^2 N(N+2) N (d+N+1) + v_{11}^2 w_{22}^2 N(N+2)(N+4) d \Big] \\
&\geq \eta_{gd}^2 \left[\eta_{gd}^2 N d + N d \left(\eta_{gd}(d+N+1) - v_{11} w_{22} (N+2) \right)^2 \right] \\
&\geq \eta_{gd}^4 N d,
\end{aligned}
$$

where (i) is due to the fact that some terms are the odd functions of $\mathbf{y}$, and (ii) is by Lemma D.1. □

For explicit CoT without scratchpad, we have the following result.

Theorem B.2. *There exist parameter matrices $\mathbf{W}^{KQ}$ and $\mathbf{W}^{V}$ such that f_{LSA} with explicit CoT and without scratchpad can implement a multi-step GD algorithm for weight prediction task.*

Proof. We set

$$
\mathbf{W}^V = \begin{bmatrix} \nu \mathbf{I}_d & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{W}^{KQ} = \begin{bmatrix} -\mu \mathbf{I}_d & 0 & 0 \\ 0 & 0 & \mu \\ 0 & 0 & 0 \end{bmatrix}.
$$

We know the intermediate sequence is

$$
\hat{\mathbf{Z}}_i = \begin{bmatrix} \mathbf{X}^\top & \boldsymbol{\beta}_0 & \hat{\boldsymbol{\beta}}_1 & ... & \hat{\boldsymbol{\beta}}_i \\ \mathbf{y}^\top & 0 & 0 & ... & 0 \\ 0 & 1 & 1 & ... & 1 \end{bmatrix}.
$$

Given this intermediate input $\hat{\mathbf{Z}}_i$, we can obtain that the output of the single-layer model is

$$
\begin{aligned}
f_{\text{LSA}}(\hat{\mathbf{Z}}_i) &= \hat{\mathbf{Z}}_{i[:,-1]} + \mathbf{W}^V \hat{\mathbf{Z}}_i \cdot \mathbf{Mask}(\hat{\mathbf{Z}}_i^\top) \mathbf{W}^{KQ} \hat{\mathbf{Z}}_{i[:,-1]} \\
&= \begin{bmatrix} \hat{\boldsymbol{\beta}}_i \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} \nu \mathbf{X}^\top & \nu \boldsymbol{\beta}_0 & \nu \hat{\boldsymbol{\beta}}_1 & \cdots & \nu \hat{\boldsymbol{\beta}}_i \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \cdot \begin{bmatrix} -\mu \mathbf{X} \hat{\boldsymbol{\beta}}_i + \mu \mathbf{y} \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \hat{\boldsymbol{\beta}}_i - \mu\nu \mathbf{X}^\top (\mathbf{X} \hat{\boldsymbol{\beta}}_i - \mathbf{y}) \\ 0 \\ 1 \end{bmatrix}.
\end{aligned}
$$

We can easily observe that $\hat{\boldsymbol{\beta}}_i - \mu\nu \mathbf{X}^\top (\mathbf{X} \hat{\boldsymbol{\beta}}_i - \mathbf{y})$ is the ground-truth weight vector after one GD step from $\hat{\boldsymbol{\beta}}_i$ with learning rate $\eta_{gd} = \mu\nu$. It means that the single-layer model can implement a one-step GD algorithm for all i. Therefore, this LSA model with explicit CoT can implement the multi-step GD algorithm in the weight prediction task. □

For implicit CoT with scratchpad, we have the following result.

Theorem B.3. *There exist parameter matrices $\mathbf{W}^{KQ}$ and $\mathbf{W}^{V}$ such that f_{LSA} with implicit CoT and with scratchpad can implement a multi-step GD algorithm for weight prediction task.*

Proof. We set

$$\mathbf{W}^{KQ} = \begin{bmatrix} 0 & 0 & -\mu\mathbf{I}_d & 0 \\ 0 & 0 & 0 & \mu \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \mathbf{W}^{V} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \nu\mathbf{I}_d & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

With this structure, the update rule becomes

$$\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_{i-1} - \mu\nu\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}_{i-1} + \mu\nu\mathbf{X}^\top\mathbf{y}.$$

We can easily observe that when $\mu\nu = \eta_{gd}$, the update rule exactly recovers the teacher GD algorithm. □

For explicit CoT with scratchpad, we have the following result.

Theorem B.4. *There exist parameter matrices $\mathbf{W}^{KQ}$ and $\mathbf{W}^{V}$ such that f_{LSA} with explicit CoT and with scratchpad can implement a multi-step GD algorithm for weight prediction task.*

Proof. We set

$$\mathbf{W}^{KQ} = \begin{bmatrix} 0 & 0 & -\mu\mathbf{I}_d & 0 \\ 0 & 0 & 0 & \mu \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \mathbf{W}^{V} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \nu\mathbf{I}_d & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

With this structure, the update rule becomes

$$\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_{i-1} - \mu\nu\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}_{i-1} + \mu\nu\mathbf{X}^\top\mathbf{y}.$$

We can easily observe that when $\mu\nu = \eta_{gd}$, the update rule exactly recovers the teacher GD algorithm. □

B.2 Query Prediction

For explicit CoT without scratchpad, we design the input matrix as

$$\mathbf{Z}_0 = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N & \mathbf{x}_q \\ y_1 & y_2 & \cdots & y_N & \boldsymbol{\beta}_0^\top\mathbf{x}_q \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{y}^\top & \boldsymbol{\beta}_0^\top\mathbf{x}_q \end{bmatrix} \in \mathbb{R}^{(d+1)\times(N+1)},$$

and we regard the bottom-right block $f_{\text{LSA}}^{(L)}(\mathbf{Z}_0)_{[-1,-1]}$ as the prediction. We can obtain the following result.

Theorem B.5. *Assume that $\mathbf{x}_i \sim N(0, \mathbf{I}_d)$, $\mathbf{x}_q \sim N(0, \mathbf{I}_d)$, $\mathbf{y} \sim N(0, \mathbf{I}_N)$, $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I}_d)$ and $\boldsymbol{\beta}_1$ is the weight vector after one GD step from $\boldsymbol{\beta}_0$ with learning rate η_{gd}. For any $f_{LSA}(\cdot)$, it holds that*

$$\min_{c\in\mathbb{R}} \mathbb{E}\left[(c \cdot f_{LSA}(\mathbf{Z}_0)_{[-1,-1]} - \boldsymbol{\beta}_1^\top\mathbf{x}_q)^2\right] \geq \eta_{gd}^2 Nd,$$

Proof. Denote that

$$\mathbf{W}^{KQ} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & w_{22} \end{bmatrix}, \mathbf{W}^{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & v_{22} \end{bmatrix}.$$

Then, we have

$f_{\text{LSA}}(\mathbf{Z}_0)_{[-1,-1]} = \boldsymbol{\beta}_0^\top\mathbf{x}_q + (\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top)(\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q + \mathbf{y}\mathbf{W}_{21}\mathbf{x}_q + \mathbf{X}\mathbf{W}_{12}\boldsymbol{\beta}_0^\top\mathbf{x}_q + \mathbf{y}w_{22}\boldsymbol{\beta}_0^\top\mathbf{x}_q)$.

The ground truth weight vector $\boldsymbol{\beta}_1$ is $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 - \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y}) = \boldsymbol{\beta}_0 - \eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0 + \eta_{gd}\mathbf{X}^\top\mathbf{y}$.

Then, we can obtain that

$$\begin{aligned}
&\mathbb{E}\left[(f_{\text{LSA}}(\mathbf{Z}_0)_{[-1,-1]} - \boldsymbol{\beta}_1^\top\mathbf{x}_q)^2\right] \\
&= \mathbb{E}\left[\left((c-1)\boldsymbol{\beta}_0^\top\mathbf{x}_q + c(\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top)(\mathbf{X}\mathbf{W}_{11} + \mathbf{y}\mathbf{W}_{21} + \mathbf{X}\mathbf{W}_{12}\boldsymbol{\beta}_0^\top + \mathbf{y}w_{22}\boldsymbol{\beta}_0^\top)\mathbf{x}_q + \eta_{gd}(\boldsymbol{\beta}_0^\top\mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q\right)^2\right] \\
&\overset{(i)}{=} \mathbb{E}\left[\left((c-1)\boldsymbol{\beta}_0^\top\mathbf{x}_q + c(\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top)(\mathbf{X}\mathbf{W}_{12}\boldsymbol{\beta}_0^\top\mathbf{x}_q + \mathbf{y}w_{22}\boldsymbol{\beta}_0^\top\mathbf{x}_q) + \eta_{gd}\boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}\mathbf{x}_q\right)^2\right] \\
&\quad + \mathbb{E}\left[\left(c(\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top)(\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q + \mathbf{y}\mathbf{W}_{21}\mathbf{x}_q) - \eta_{gd}\mathbf{y}^\top\mathbf{X}\mathbf{x}_q\right)^2\right] \\
&\geq \mathbb{E}\left[\left((c-1)\boldsymbol{\beta}_0^\top\mathbf{x}_q + c(\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top)(\mathbf{X}\mathbf{W}_{12}\boldsymbol{\beta}_0^\top\mathbf{x}_q + \mathbf{y}w_{22}\boldsymbol{\beta}_0^\top\mathbf{x}_q) + \eta_{gd}\boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}\mathbf{x}_q\right)^2\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(ii)}{=} \mathbb{E}\left[\left((c-1)\boldsymbol{\beta}_0^\top \mathbf{x}_q + c\mathbf{V}_{21}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{12}\boldsymbol{\beta}_0^\top\mathbf{x}_q + cv_{22}\mathbf{y}^\top\mathbf{y}w_{22}\boldsymbol{\beta}_0^\top\mathbf{x}_q + \eta_{gd}\boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}\mathbf{x}_q\right)^2\right] \\
&\quad + \mathbb{E}\left[\left(c\mathbf{V}_{21}\mathbf{X}^\top\mathbf{y}w_{22}\boldsymbol{\beta}_0^\top\mathbf{x}_q + cv_{22}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{12}\boldsymbol{\beta}_0^\top\mathbf{x}_q\right)^2\right] \\
&\geq \mathbb{E}\left[\left((c-1)\boldsymbol{\beta}_0^\top \mathbf{x}_q + c\mathbf{V}_{21}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{12}\boldsymbol{\beta}_0^\top\mathbf{x}_q + cv_{22}\mathbf{y}^\top\mathbf{y}w_{22}\boldsymbol{\beta}_0^\top\mathbf{x}_q + \eta_{gd}\boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}\mathbf{x}_q\right)^2\right] \\
&= \mathbb{E}\Bigg[(c-1)^2(\boldsymbol{\beta}_0^\top\mathbf{x}_q)^2 + 2c(c-1)(\mathbf{V}_{21}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{12} + v_{22}\mathbf{y}^\top\mathbf{y}w_{22})(\boldsymbol{\beta}_0^\top\mathbf{x}_q)^2 + 2(c-1)\eta_{gd}\boldsymbol{\beta}_0^\top\mathbf{x}_q\boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}\mathbf{x}_q \\
&\quad + c^2(\boldsymbol{\beta}_0^\top\mathbf{x}_q)^2(\mathbf{V}_{21}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{12})^2 + c^2(\boldsymbol{\beta}_0^\top\mathbf{x}_q)^2(v_{22}w_{22}\mathbf{y}^\top\mathbf{y})^2 + 2c^2(\boldsymbol{\beta}_0^\top\mathbf{x}_q)^2 v_{22}w_{22}\mathbf{V}_{21}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{12}\mathbf{y}^\top\mathbf{y} \\
&\quad + \eta_{gd}^2(\boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}\mathbf{x}_q)^2 + 2c\mathbf{V}_{21}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{12}\boldsymbol{\beta}_0^\top\mathbf{x}_q\eta_{gd}\boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}\mathbf{x}_q + 2cv_{22}w_{22}\mathbf{y}^\top\mathbf{y}\boldsymbol{\beta}_0^\top\mathbf{x}_q\eta_{gd}\boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}\mathbf{x}_q\Bigg] \\
&\overset{(iii)}{=} (c-1)^2 d + 2c(c-1)Nd\mathbf{V}_{21}\mathbf{W}_{12} + 2c(c-1)Ndv_{22}w_{22} + 2(c-1)\eta_{gd}Nd \\
&\quad + c^2 d\left(N\|\mathbf{V}_{21}\|_2^2\|\mathbf{W}_{12}\|_2^2 + N(N+1)(\mathbf{V}_{21}\mathbf{W}_{12})^2\right) + c^2 dN(N+2)v_{22}^2w_{22}^2 + 2c^2dN^2v_{22}w_{22}\mathbf{V}_{21}\mathbf{W}_{12} \\
&\quad + \eta_{gd}^2 dN(d+N+1) + 2c\eta_{gd}N(dN+2)\mathbf{V}_{21}\mathbf{W}_{12} + 2c\eta_{gd}dN^2v_{22}w_{22} \\
&= c^2dN\|\mathbf{V}_{21}\|_2^2\|\mathbf{W}_{12}\|_2^2 + d(c-1+cN\mathbf{V}_{21}\mathbf{W}_{12} + cNv_{22}w_{22} + N\eta_{gd})^2 + 2c^2dNv_{22}^2w_{22}^2 \\
&\quad + c^2dN(\mathbf{V}_{21}\mathbf{W}_{12})^2 + 4c\eta_{gd}N\mathbf{V}_{21}\mathbf{W}_{12} + \eta_{gd}^2dN(d+N+1) \\
&= c^2dN\|\mathbf{V}_{21}\|_2^2\|\mathbf{W}_{12}\|_2^2 + d(c-1+cN\mathbf{V}_{21}\mathbf{W}_{12} + cNv_{22}w_{22} + N\eta_{gd})^2 + 2c^2dNv_{22}^2w_{22}^2 \\
&\quad + dN(c\mathbf{V}_{21}\mathbf{W}_{12} + \frac{2\eta_{gd}}{d})^2 + \eta_{gd}^2N\left(d(d+N+1) - \frac{4}{d}\right) \\
&\geq \eta_{gd}^2 Nd,
\end{aligned}
$$

where (i) is by $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I}_d)$, (ii) is by $\mathbf{y} \sim N(0, \mathbf{I}_N)$, and (iii) is by Lemma D.1. □

For implicit CoT without scratchpad, we have the following result.

Theorem B.6. *There exist parameter matrices $\mathbf{W}^{KQ}$ and $\mathbf{W}^V$ such that f_{LSA} with implicit CoT and without scratchpad can implement a multi-step GD algorithm for query prediction task.*

Proof. We set

$$
\mathbf{W}^{KQ} = \begin{bmatrix} \mu\mathbf{I}_d & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{W}^V = \begin{bmatrix} 0 & 0 \\ 0 & -\nu \end{bmatrix}.
$$

With this structure, we can get the prediction after L steps as

$$
f_{\text{LSA}}^{(L)}(\mathbf{Z}_0)_{[-1,-1]} = -\boldsymbol{\beta}_0^\top(\mathbf{I}_d - \mu\nu\mathbf{X}^\top\mathbf{X})^L\mathbf{x}_q - \mu\nu\mathbf{y}^\top\mathbf{X}\sum_{k=0}^{L-1}(\mathbf{I} - \mu\nu\mathbf{X}^\top\mathbf{X})^k\mathbf{x}_q,
$$

while the ground-truth intermediate weight vector in the GD algorithm is

$$
\boldsymbol{\beta}_L = (\mathbf{I}_d - \eta_{gd}\mathbf{X}^\top\mathbf{X})^L\boldsymbol{\beta}_0 + \eta_{gd}\sum_{i=1}^{L-1}(\mathbf{I}_d - \eta_{gd}\mathbf{X}^\top\mathbf{X})^i\mathbf{X}^\top\mathbf{y}.
$$

We can easily observe that when $\mu\nu = \eta_{gd}$, we obtain $f_{\text{LSA}}^{(L)}(\mathbf{Z}_0)_{[-1,-1]} = -\boldsymbol{\beta}_L^\top\mathbf{x}_q$, which demonstrates that the computation process aligns with the teacher GD algorithm. □

For explicit CoT with scratchpad, the generated intermediate sequence is expected to simulate

$$
\mathbf{Z}_i = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q & \mathbf{x}_q & \cdots & \mathbf{x}_q \\ \mathbf{y}^\top & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{u}_0 & \mathbf{u}_1 & \cdots & \mathbf{u}_i \\ 0 & 0 & \boldsymbol{\beta}_1^\top\mathbf{x}_q & \cdots & \boldsymbol{\beta}_i^\top\mathbf{x}_q \end{bmatrix} \in \mathbb{R}^{(2d+2)\times(N+i+1)}, i \in [L],
$$

where $\mathbf{u}_i = (\mathbf{I}_d - \eta_{gd}\mathbf{X}^\top\mathbf{X})^i\mathbf{x}_q$. Here, $\boldsymbol{\beta}_i$ is GD steps from 0 initialization with learning rate η_{gd}. We have the following result of expressivity.

Theorem B.7. *There exist parameter matrices $\mathbf{W}^{KQ}$ and $\mathbf{W}^V$ such that f_{LSA} with explicit CoT and with scratchpad can implement a multi-step GD algorithm for query prediction task.*

Proof. We set

$$\mathbf{W}^{KQ}(t) = \begin{bmatrix} 0 & 0 & \mu\mathbf{I}_d & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \mathbf{W}^{V}(t) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -\nu\mathbf{I}_d & 0 & 0 & 0 \\ 0 & \nu & 0 & 0 \end{bmatrix}.$$

With this structure, we can get the prediction after L steps as

$$f_{\text{LSA}}^{(L)}(\mathbf{Z}_0)_{[-1,-1]} = \boldsymbol{\beta}_0^\top(\mathbf{I}_d - \mu\nu\mathbf{X}^\top\mathbf{X})^L\mathbf{x}_q + \mu\nu\mathbf{y}^\top\mathbf{X}\sum_{k=0}^{L-1}(\mathbf{I} - \mu\nu\mathbf{X}^\top\mathbf{X})^k\mathbf{x}_q.$$

We can observe that when $\mu\nu = \eta_{gd}$, it holds that $f_{\text{LSA}}^{(L)}(\mathbf{Z}_0)_{[-1,-1]} = \boldsymbol{\beta}_L^\top\mathbf{x}_q$ for all $L \geq 0$, which demonstrates that the computation process aligns with the teacher GD algorithm. □

For implicit CoT with scratchpad, we have the following result.

Theorem B.8. *There exist parameter matrices $\mathbf{W}^{KQ}$ and $\mathbf{W}^{V}$ such that f_{LSA} with implicit CoT and with scratchpad can implement a multi-step GD algorithm for query prediction task.*

Proof. We set

$$\mathbf{W}^{KQ} = \begin{bmatrix} \mu\mathbf{I}_d & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{W}^{V} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \nu & -\nu \end{bmatrix}.$$

With this structure, we can get the prediction after L steps as

$$f_{\text{LSA}}^{(L)}(\mathbf{Z}_0)_{[-1,-1]} = \boldsymbol{\beta}_0^\top(\mathbf{I}_d - \mu\nu\mathbf{X}^\top\mathbf{X})^L\mathbf{x}_q + \mu\nu\mathbf{y}^\top\mathbf{X}\sum_{k=0}^{L-1}(\mathbf{I} - \mu\nu\mathbf{X}^\top\mathbf{X})^k\mathbf{x}_q.$$

We can observe that when $\mu\nu = \eta_{gd}$, it holds that $f_{\text{LSA}}^{(L)}(\mathbf{Z}_0)_{[-1,-1]} = \boldsymbol{\beta}_L^\top\mathbf{x}_q$ for all $L \geq 0$, which demonstrates that the computation process aligns with the teacher GD algorithm. □

B.3 Transform from Explicit to Implicit

We build a rigorous construction to transform explicit CoT into implicit CoT with a scratchpad.

Theorem B.9. *Any explicit CoT reasoning algorithm generated by a single-layer linear-attention model can be exactly simulated by a layer-recurrent implicit CoT model with an additional scratchpad.*

Proof. Assume that for explicit CoT, it generates the next token as

$$\begin{aligned} f_{\text{LSA}}([\mathbf{Z}, \mathbf{z}_{q,1}, ..., \mathbf{z}_{q,i}])_{[:,-1]} &= \mathbf{z}_{q,i} + \mathbf{W}^V[\mathbf{Z}, \mathbf{z}_{q,1}, ..., \mathbf{z}_{q,i}] \cdot [\mathbf{Z}, 0, ..., 0]^\top\mathbf{W}^{KQ}\mathbf{z}_{q,i} \\ &= \mathbf{z}_{q,i} + [\mathbf{W}^V\mathbf{Z}, \mathbf{W}^V\mathbf{z}_{q,1}, ..., \mathbf{W}^V\mathbf{z}_{q,i}] \cdot \begin{bmatrix} \mathbf{Z}^\top\mathbf{W}^{KQ}\mathbf{z}_{q,i} \\ 0 \end{bmatrix} \\ &= \mathbf{z}_{q,i} + \mathbf{W}^V\mathbf{Z} \cdot \mathbf{Z}^\top\mathbf{W}^{KQ}\mathbf{z}_{q,i} := \mathbf{z}_{q,i+1}, \end{aligned}$$

where $\mathbf{Z}$ is the original input matrix, and $\mathbf{z}_{q,i}$ is the current intermediate token.

Now, we design the model with implicit CoT, adding a scratchpad. The new format of the input matrix is $\begin{bmatrix} \mathbf{Z} & 0 \\ 0 & \mathbf{z}_{q,i} \end{bmatrix}$, and the modified parameter matrices are $\begin{bmatrix} 0 & 0 \\ \mathbf{W}^V & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & \mathbf{W}^{KQ} \\ 0 & 0 \end{bmatrix}$, respectively. Then, the update is

$$\begin{aligned} f_{\text{LSA}}\left(\begin{bmatrix} \mathbf{Z} & 0 \\ 0 & \mathbf{z}_{q,i} \end{bmatrix}\right) &= \begin{bmatrix} \mathbf{Z} & 0 \\ 0 & \mathbf{z}_{q,i} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \mathbf{W}^V & 0 \end{bmatrix}\begin{bmatrix} \mathbf{Z} & 0 \\ 0 & \mathbf{z}_{q,i} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{Z} & 0 \\ 0 & \mathbf{z}_{q,i} \end{bmatrix}^\top \begin{bmatrix} 0 & \mathbf{W}^{KQ} \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \mathbf{Z} & 0 \\ 0 & \mathbf{z}_{q,i} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Z} & 0 \\ 0 & \mathbf{z}_{q,i} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \mathbf{W}^V\mathbf{Z} & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & \mathbf{Z}^\top\mathbf{W}^{KQ}\mathbf{z}_{q,i} \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Z} & 0 \\ 0 & \mathbf{z}_{q,i} + \mathbf{W}^V\mathbf{Z} \cdot \mathbf{Z}^\top\mathbf{W}^{KQ}\mathbf{z}_{q,i} \end{bmatrix} = \begin{bmatrix} \mathbf{Z} & 0 \\ 0 & \mathbf{z}_{q,i+1} \end{bmatrix}. \end{aligned}$$

This straightforward calculation shows that the implicit model produces exactly the same update, mapping $\mathbf{z}_{q,i}$ to $\mathbf{z}_{q,i+1}$ while preserving $\mathbf{Z}$ unchanged. Hence, we demonstrate that any algorithm implemented by explicit CoT can also be simulated by implicit CoT with the designed scratchpad. Note that although here we consider linear attention, a similar transformation can be applied to softmax attention by replacing the mask from zero with negative infinity. □

This result can also be extended to softmax attention in the following theorem.

Theorem B.10. *Any explicit CoT reasoning algorithm generated by a single-layer softmax-attention model can be exactly simulated by a layer-recurrent implicit CoT model with an additional scratchpad.*

Proof. Assume that for explicit CoT, it generates the next token as

$$\begin{aligned} f_{\text{LSA}}([\mathbf{Z}, \mathbf{z}_{q,1}, ..., \mathbf{z}_{q,i}])_{[:,-1]} &= \mathbf{z}_{q,i} + \mathbf{W}^V[\mathbf{Z}, \mathbf{z}_{q,1}, ..., \mathbf{z}_{q,i}] \cdot \mathcal{S}\left([\mathbf{Z}, -\infty, ..., -\infty]^\top \mathbf{W}^{KQ}\mathbf{z}_{q,i}\right) \\ &= \mathbf{z}_{q,i} + [\mathbf{W}^V\mathbf{Z}, \mathbf{W}^V\mathbf{z}_{q,1}, ..., \mathbf{W}^V\mathbf{z}_{q,i}] \cdot \mathcal{S}\left(\begin{bmatrix} \mathbf{Z}^\top\mathbf{W}^{KQ}\mathbf{z}_{q,i} \\ -\infty \end{bmatrix}\right) \\ &= \mathbf{z}_{q,i} + \mathbf{W}^V\mathbf{Z} \cdot \mathcal{S}\left(\mathbf{Z}^\top\mathbf{W}^{KQ}\mathbf{z}_{q,i}\right) := \mathbf{z}_{q,i+1}, \end{aligned}$$

where $\mathbf{Z}$ is the original input matrix, and $\mathbf{z}_{q,i}$ is the current intermediate token.

Now, we design the model with implicit CoT, adding a scratchpad. The new format of the input matrix is $\begin{bmatrix} \mathbf{Z} & 0 \\ * & \mathbf{z}_{q,i} \end{bmatrix}$, and the modified parameter matrices are $\begin{bmatrix} 0 & 0 \\ \mathbf{W}^V & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & \mathbf{W}^{KQ} \\ 0 & 0 \end{bmatrix}$, respectively. Here, $*$ can be any real vector. Then, the update is

$$\begin{aligned} f_{\text{LSA}}\left(\begin{bmatrix} \mathbf{Z} & 0 \\ * & \mathbf{z}_{q,i} \end{bmatrix}\right) &= \begin{bmatrix} \mathbf{Z} & 0 \\ * & \mathbf{z}_{q,i} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \mathbf{W}^V & 0 \end{bmatrix} \begin{bmatrix} \mathbf{Z} & 0 \\ * & \mathbf{z}_{q,i} \end{bmatrix} \cdot \mathcal{S}\left(\begin{bmatrix} \mathbf{Z} & -\infty \\ * & -\infty \end{bmatrix}^\top \begin{bmatrix} 0 & \mathbf{W}^{KQ} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{Z} & 0 \\ * & \mathbf{z}_{q,i} \end{bmatrix}\right) \\ &= \begin{bmatrix} \mathbf{Z} & 0 \\ * & \mathbf{z}_{q,i} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \mathbf{W}^V\mathbf{Z} & 0 \end{bmatrix} \cdot \mathcal{S}\left(\begin{bmatrix} 0 & \mathbf{Z}^\top\mathbf{W}^{KQ}\mathbf{z}_{q,i} \\ 0 & -\infty \end{bmatrix}\right) \\ &= \begin{bmatrix} \mathbf{Z} & 0 \\ * & \mathbf{z}_{q,i} + \mathbf{W}^V\mathbf{Z} \cdot \mathcal{S}\left(\mathbf{Z}^\top\mathbf{W}^{KQ}\mathbf{z}_{q,i}\right) \end{bmatrix} = \begin{bmatrix} \mathbf{Z} & 0 \\ * & \mathbf{z}_{q,i+1} \end{bmatrix}. \end{aligned}$$

This straightforward calculation shows that the implicit model produces exactly the same update, mapping $\mathbf{z}_{q,i}$ to $\mathbf{z}_{q,i+1}$ while preserving $\mathbf{Z}$ unchanged. Therefore, we demonstrate that any algorithm implemented by explicit CoT can be simulated by implicit CoT with the designed scratchpad. □

C Training Guarantee

In this section, we provide some convergence results and the OOD generalization results for the successful cases.

C.1 Implicit CoT with Scratchpad for Weight Prediction

The input with scratchpad is designed as

$$\tilde{\mathbf{Z}} = \begin{bmatrix} \mathbf{X}^\top & 0 \\ \mathbf{y}^\top & 0 \\ 0 & \boldsymbol{\beta}_0 \\ 0 & 1 \end{bmatrix},$$

where $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I})$ is the initial weight vector. A single linear attention layer is

$$f_{\text{LSA}}(\mathbf{Z}) = \mathbf{Z} + \mathbf{W}^V\mathbf{Z} \cdot \mathbf{Mask}(\mathbf{Z})^\top \mathbf{W}^{KQ}\mathbf{Z},$$

where

$$\mathbf{W}^{KQ} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} & \mathbf{W}_{13} & \mathbf{W}_{14} \\ \mathbf{W}_{21} & w_{22} & \mathbf{W}_{23} & w_{24} \\ \mathbf{W}_{31} & \mathbf{W}_{32} & \mathbf{W}_{33} & \mathbf{W}_{34} \\ \mathbf{W}_{41} & w_{42} & \mathbf{W}_{43} & w_{44} \end{bmatrix}, \mathbf{W}^V = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} & \mathbf{V}_{14} \\ \mathbf{V}_{21} & v_{22} & \mathbf{V}_{23} & v_{24} \\ \mathbf{V}_{31} & \mathbf{V}_{32} & \mathbf{V}_{33} & \mathbf{V}_{34} \\ \mathbf{V}_{41} & v_{42} & \mathbf{V}_{43} & v_{44} \end{bmatrix}.$$

The training loss is

$$\mathcal{L} = \mathbb{E}\left[\frac{1}{2}\|f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[(d+2):(2d+1),-1]} - \boldsymbol{\beta}_1\|_2^2\right],$$

where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 - \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})$ is the ground truth GD update. Then, we can get

$f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[(d+2):(2d+1),-1]} = \boldsymbol{\beta}_0 + (\mathbf{V}_{31}\mathbf{X}^\top + \mathbf{V}_{32}\mathbf{y}^\top)\cdot(\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + \mathbf{X}\mathbf{W}_{14} + \mathbf{y}\mathbf{W}_{23}\boldsymbol{\beta}_0 + \mathbf{y}w_{24})$.
We consider the gradient flow method

$$\frac{d\theta}{dt} = -\nabla\mathcal{L}(\theta)$$

for $\theta := (\mathbf{W}^{KQ}, \mathbf{W}^V)$. We assume that the initialization of w_{24} is $w_{24}(0) = 1$, and all the elements in the other blocks are initialized as 0. Then, we can find the form of $\mathbf{W}^{KQ}(t)$ and $\mathbf{W}^V(t)$ during the training in the following theorem.

Theorem C.1. *For all $t > 0$, $\mathbf{W}^{KQ}(t)$ and $\mathbf{W}^V(t)$ have the following form:*

$$\mathbf{W}^{KQ}(t) = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} & w_{13}(t)\cdot\mathbf{I} & \boldsymbol{0} \\ \boldsymbol{0} & 0 & \boldsymbol{0} & w_{24}(t) \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 0 & \boldsymbol{0} & 0 \end{bmatrix}, \mathbf{W}^V(t) = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 0 & \boldsymbol{0} & 0 \\ v_{31}(t)\cdot\mathbf{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 0 & \boldsymbol{0} & 0 \end{bmatrix}.$$

Proof. Since the prediction is

$f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[(d+2):(2d+1),-1]} = \boldsymbol{\beta}_0 + (\mathbf{V}_{31}\mathbf{X}^\top + \mathbf{V}_{32}\mathbf{y}^\top)\cdot(\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + \mathbf{X}\mathbf{W}_{14} + \mathbf{y}\mathbf{W}_{23}\boldsymbol{\beta}_0 + \mathbf{y}w_{24})$,
we can observe that the blocks except $\mathbf{V}_{31}$, $\mathbf{V}_{32}$, $\mathbf{W}_{13}$, $\mathbf{W}_{14}$, $\mathbf{W}_{23}$, and w_{24} are irrelevant to the training loss. Thus, these blocks stay at 0 for $t > 0$. Next, we prove that $\mathbf{V}_{32}(t) = 0$, $\mathbf{W}_{14}(t) = 0$, $\mathbf{W}_{23}(t) = 0$ for $t > 0$. When $\mathbf{V}_{32}$, $\mathbf{W}_{14}$, $\mathbf{W}_{23}$ are 0, we can calculate their gradients as follows:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{V}_{32}} &= \mathbb{E}\Big[\big((\mathbf{V}_{31}\mathbf{X}^\top + \mathbf{V}_{32}\mathbf{y}^\top)\cdot(\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + \mathbf{X}\mathbf{W}_{14} + \mathbf{y}\mathbf{W}_{23}\boldsymbol{\beta}_0 + \mathbf{y}w_{24}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})\big) \\
&\qquad \cdot(\boldsymbol{\beta}_0^\top\mathbf{W}_{13}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{W}_{14}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}_0^\top\mathbf{W}_{23}^\top\mathbf{y}^\top\mathbf{y} + w_{24}\mathbf{y}^\top\mathbf{y})\Big] \\
&= \mathbb{E}\left[\big(\mathbf{V}_{31}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + w_{24}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})\big)\cdot(\boldsymbol{\beta}_0^\top\mathbf{W}_{13}^\top\mathbf{X}^\top\mathbf{y} + w_{24}\mathbf{y}^\top\mathbf{y})\right] \\
&= \mathbb{E}\Big[(\mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + \mathbf{V}_{31}\mathbf{X}^\top w_{24}\mathbf{X}\boldsymbol{\beta} + \eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0 - \eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta})\cdot\boldsymbol{\beta}_0^\top\mathbf{W}_{13}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\
&\qquad + (\mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + \mathbf{V}_{31}\mathbf{X}^\top w_{24}\mathbf{X}\boldsymbol{\beta} + \eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0 - \eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta})\cdot w_{24}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\Big], \\
\frac{\partial\mathcal{L}}{\partial\mathbf{W}_{14}} &= \mathbb{E}\Big[(\mathbf{X}^\top\mathbf{X}\mathbf{V}_{31}^\top + \mathbf{X}^\top\mathbf{y}\mathbf{V}_{32}^\top) \\
&\qquad \cdot\big((\mathbf{V}_{31}\mathbf{X}^\top + \mathbf{V}_{32}\mathbf{y}^\top)\cdot(\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + \mathbf{X}\mathbf{W}_{14} + \mathbf{y}\mathbf{W}_{23}\boldsymbol{\beta}_0 + \mathbf{y}w_{24}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})\big)\Big] \\
&= \mathbb{E}\left[\mathbf{X}^\top\mathbf{X}\mathbf{V}_{31}^\top\cdot\big(\mathbf{V}_{31}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + w_{24}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})\big)\right] \\
&= \mathbb{E}\left[\mathbf{X}^\top\mathbf{X}\mathbf{V}_{31}^\top\mathbf{V}_{31}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + w_{24}\mathbf{Z}\boldsymbol{\beta}) + \mathbf{X}^\top\mathbf{X}\mathbf{V}_{31}^\top\eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{X}\boldsymbol{\beta})\right], \\
\frac{\partial\mathcal{L}}{\partial\mathbf{W}_{23}} &= \mathbb{E}\Big[(\mathbf{y}^\top\mathbf{X}\mathbf{V}_{31}^\top + \mathbf{y}^\top\mathbf{y}\mathbf{V}_{32}^\top) \\
&\qquad \cdot\big((\mathbf{V}_{31}\mathbf{X}^\top + \mathbf{V}_{32}\mathbf{y}^\top)\cdot(\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + \mathbf{X}\mathbf{W}_{14} + \mathbf{y}\mathbf{W}_{23}\boldsymbol{\beta}_0 + \mathbf{y}w_{24}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})\big)\boldsymbol{\beta}_0^\top\Big] \\
&= \mathbb{E}\left[\mathbf{y}^\top\mathbf{X}\mathbf{V}_{31}^\top\cdot\big(\mathbf{V}_{31}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + w_{24}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})\big)\boldsymbol{\beta}_0^\top\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\mathbf{V}_{31}^\top\mathbf{V}_{31}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^\top + w_{24}\mathbf{Z}\boldsymbol{\beta}\boldsymbol{\beta}_0^\top) + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\mathbf{V}_{31}^\top\eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^\top - \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}_0^\top)\right].
\end{aligned}$$

We can find that each term in these three gradients is an odd function of $\boldsymbol{\beta}$ or $\boldsymbol{\beta}_0$. Since $\boldsymbol{\beta} \sim N(0, \mathbf{I})$ and $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I})$, the expectation of these three gradients should be 0. Therefore, we get that $\mathbf{V}_{32}$, $\mathbf{W}_{14}$, and $\mathbf{W}_{23}$ always stay at 0. Then, we can get

$$f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[(d+2):(2d+1),-1]} = \boldsymbol{\beta}_0 + \mathbf{V}_{31}\mathbf{X}^\top\cdot(\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + w_{24}\mathbf{y}).$$

Next, we prove that $\mathbf{V}_{31}(t) \propto \mathbf{I}$ and $\mathbf{W}_{13}(t) \propto \mathbf{I}$. Assuming that $\mathbf{V}_{31} \propto \mathbf{I}$ and $\mathbf{W}_{13} \propto \mathbf{I}$, their gradients are

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{V}_{31}} &= \mathbb{E}\left[\left(\mathbf{V}_{31}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + \mathbf{y}w_{24}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})\right) \cdot (\boldsymbol{\beta}_0^\top \mathbf{W}_{13}^\top \mathbf{X}^\top \mathbf{X} + w_{24}\mathbf{y}^\top \mathbf{X})\right] \\
&= \mathbb{E}\Big[\mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^\top\mathbf{W}_{13}^\top\mathbf{X}^\top\mathbf{X} + \eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^\top\mathbf{W}_{13}^\top\mathbf{X}^\top\mathbf{X} \\
&\qquad + \mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}w_{24}^2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X} - \eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}w_{24}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\Big] \\
&= \mathbb{E}\left[(\mathbf{V}_{31}\mathbf{W}_{13}\mathbf{W}_{13}^\top + \eta_{gd}\mathbf{W}_{13}^\top + w_{24}^2\mathbf{V}_{31} - \eta_{gd}w_{24}\mathbf{I}) \cdot (\mathbf{X}^\top\mathbf{X})^2\right] \\
&= N(N+d+1)\cdot(\mathbf{V}_{31}\mathbf{W}_{13}\mathbf{W}_{13}^\top + \eta_{gd}\mathbf{W}_{13}^\top + w_{24}^2\mathbf{V}_{31} - \eta_{gd}w_{24}\mathbf{I}), \\
\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{13}} &= \mathbb{E}\left[\mathbf{X}^\top\mathbf{X}\mathbf{V}_{31}^\top \cdot \left(\mathbf{V}_{31}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + \mathbf{y}w_{24}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})\right)\boldsymbol{\beta}_0^\top\right] \\
&= \mathbb{E}\left[\mathbf{X}^\top\mathbf{X}\mathbf{V}_{31}^\top\mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^\top + \mathbf{X}^\top\mathbf{X}\mathbf{V}_{31}^\top\eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^\top\right] \\
&= \mathbb{E}\left[(\mathbf{V}_{31}^\top\mathbf{V}_{31}\mathbf{W}_{13} + \eta_{gd}\mathbf{V}_{31}^\top)\cdot(\mathbf{X}^\top\mathbf{X})^2\right] \\
&= N(N+d+1)\cdot(\mathbf{V}_{31}^\top\mathbf{V}_{31}\mathbf{W}_{13} + \eta_{gd}\mathbf{V}_{31}^\top),
\end{aligned}$$

where we use Lemma D.1. We can find that when $\mathbf{V}_{31}$ and $\mathbf{W}_{13}$ are proportional to $\mathbf{I}$, the expectation of their gradient is also proportional to $\mathbf{I}$. Therefore, with zero initialization, $\mathbf{V}_{31}(t)$ and $\mathbf{W}_{13}(t)$ can be written as $\mathbf{V}_{31}(t) = v_{31}(t)\cdot\mathbf{I}$ and $\mathbf{W}_{13}(t) = w_{13}(t)\cdot\mathbf{I}$. □

Now, we can only consider three parameters v_{31}, w_{13}, and w_{24}. We can calculate the corresponding gradients.

Lemma C.2. *The gradients for v_{31}, w_{13}, w_{24} are*

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial v_{31}} &= c_2 \cdot w_{13}(v_{31}w_{13} + \eta_{gd}) + c_2 \cdot w_{24}(v_{31}w_{24} - \eta_{gd}), \\
\frac{\partial \mathcal{L}}{\partial w_{13}} &= c_2 \cdot v_{31}(v_{31}w_{13} + \eta_{gd}), \\
\frac{\partial \mathcal{L}}{\partial w_{24}} &= c_2 \cdot v_{31}(v_{31}w_{24} - \eta_{gd}),
\end{aligned}$$

where $c_2 = Nd(N+d+1)$.

Proof. Recall that

$$f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[(d+2):(2d+1),-1]} = \boldsymbol{\beta}_0 + \mathbf{V}_{31}\mathbf{X}^\top \cdot (\mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0 + w_{24}\mathbf{y}).$$

According to Lemma D.1, we can obtain

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial v_{31}} &= \mathbb{E}\left[(w_{13}\boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X} + w_{24}\mathbf{y}^\top\mathbf{X}) \cdot \left(v_{31}\mathbf{X}^\top(w_{13}\mathbf{X}\boldsymbol{\beta}_0 + w_{24}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})\right)\right] \\
&= \mathbb{E}\left[w_{13}(v_{31}w_{13} + \eta_{gd})\boldsymbol{\beta}_0^\top(\mathbf{X}^\top\mathbf{X})^2\boldsymbol{\beta}_0 + w_{24}(v_{31}w_{24} - \eta_{gd})\boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})^2\boldsymbol{\beta}\right] \\
&= Nd(N+d+1)\cdot w_{13}(v_{31}w_{13} + \eta_{gd}) + Nd(N+d+1)\cdot w_{24}(v_{31}w_{24} - \eta_{gd}), \\
\frac{\partial \mathcal{L}}{\partial w_{13}} &= \mathbb{E}\left[v_{31}\boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X} \cdot \left(v_{31}\mathbf{X}^\top(w_{13}\mathbf{X}\boldsymbol{\beta}_0 + w_{24}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})\right)\right] \\
&= \mathbb{E}\left[v_{31}(v_{31}w_{13} + \eta_{gd})\boldsymbol{\beta}_0^\top(\mathbf{X}^\top\mathbf{X})^2\boldsymbol{\beta}_0\right] \\
&= Nd(N+d+1)\cdot v_{31}(v_{31}w_{13} + \eta_{gd}), \\
\frac{\partial \mathcal{L}}{\partial w_{24}} &= \mathbb{E}\left[v_{31}\mathbf{y}^\top\mathbf{X} \cdot \left(v_{31}\mathbf{X}^\top(w_{13}\mathbf{X}\boldsymbol{\beta}_0 + w_{24}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})\right)\right] \\
&= \mathbb{E}\left[v_{31}(v_{31}w_{24} - \eta_{gd})\boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})^2\boldsymbol{\beta}\right] \\
&= Nd(N+d+1)\cdot v_{31}(v_{31}w_{24} - \eta_{gd}).
\end{aligned}$$

□

In order to analyze the convergence of three parameters, we first provide their bounded property.

Lemma C.3. *For all $t > 0$, $v_{31}(t)$, $w_{13}(t)$ and $w_{24}(t)$ are bounded, i.e. there exists a positive constant M such that $|v_{31}(t)| \leq M$, $|w_{13}(t)| \leq M$, $|w_{24}(t)| \leq M$.*

Proof. Denote $A = v_{31}w_{13} + \eta_{gd}$, $B = v_{31}w_{24} - \eta_{gd}$. Then, we have the following equations:

$$\frac{dv_{31}}{dt} = -c_2(w_{13}A + w_{24}B), \quad \frac{dw_{13}}{dt} = -c_2 v_{31}A, \quad \frac{dw_{24}}{dt} = -c_2 v_{31}B.$$

By observation, we can get

$$\begin{aligned}\frac{d(w_{13}^2 + w_{24}^2 - v_{31}^2)}{dt} &= 2w_{13}\frac{dw_{13}}{dt} + 2w_{24}\frac{dw_{24}}{dt} - 2v_{31}\frac{dv_{31}}{dt} \\ &= -2c_2 v_{31}w_{13}A - 2c_2 v_{31}w_{24}B + 2c_2 v_{31}(w_{13}A + w_{24}B) \\ &= 0.\end{aligned}$$

Thus, for all $t > 0$, the value of $w_{13}(t)^2 + w_{24}(t)^2 - v_{31}(t)^2$ will not change, which means that $w_{13}(t)^2 + w_{24}(t)^2 - v_{31}(t)^2 = w_{24}(0)^2 = 1$. Next, in order to give the bound of w_{13}, w_{24}, v_{31}, we check the dynamic of $w_{13}^2 + w_{24}^2 + v_{31}^2$:

$$\begin{aligned}\frac{d(w_{13}^2 + w_{24}^2 + v_{31}^2)}{dt} &= 2w_{13}\frac{dw_{13}}{dt} + 2w_{24}\frac{dw_{24}}{dt} + 2v_{31}\frac{dv_{31}}{dt} \\ &= -2c_2 v_{31}w_{13}A - 2c_2 v_{31}w_{24}B - 2c_2 v_{31}(w_{13}A + w_{24}B) \\ &= -4c_2 v_{31}w_{13}A - 4c_2 v_{31}w_{24}B \\ &= -4c_2 v_{31}w_{13}(v_{31}w_{13} + \eta_{gd}) - 4c_2 v_{31}w_{24}(v_{31}w_{24} - \eta_{gd}) \\ &= -4c_2 v_{31}^2(w_{13}^2 + w_{24}^2) + 4c_2\eta_{gd}v_{31}(w_{24} - w_{13}) \\ &\overset{(i)}{\leq} -4c_2 v_{31}^2(w_{13}^2 + w_{24}^2) + 4\sqrt{2}c_2\eta_{gd}|v_{31}|\sqrt{w_{13}^2 + w_{24}^2} \\ &\overset{(ii)}{=} -4c_2 v_{31}^2(v_{31}^2 + 1) + 4\sqrt{2}c_2\eta_{gd}|v_{31}|\sqrt{v_{31}^2 + 1} \\ &= -4c_2|v_{31}|\sqrt{v_{31}^2 + 1}\left(|v_{31}|\sqrt{v_{31}^2 + 1} - \sqrt{2}\eta_{gd}\right),\end{aligned}$$

where (i) is by $w_{24} - w_{13} \leq \sqrt{2(w_{13}^2 + w_{24}^2)}$, and (ii) is by $w_{13}(t)^2 + w_{24}(t)^2 - v_{31}(t)^2 = 1$. Once $v_{31}^2 + w_{13}^2 + w_{24}^2 > 4\eta_{gd} + 1$, we have $v_{31}^2 > 2\eta_{gd}$, then $\frac{d(v_{31}^2 + w_{13}^2 + w_{24}^2)}{dt} < 0$, which indicates that $v_{31}^2 + w_{13}^2 + w_{24}^2$ will never be larger than $4\eta_{gd} + 1$. Thus, we have $v_{31}^2 + w_{13}^2 + w_{24}^2 \leq 4\eta_{gd} + 1$. So, there exists a positive constant M such that $|v_{31}| < M, |w_{13}| < M, |w_{24}| < M$. □

Next, we can derive the following convergence results.

Lemma C.4. *$v_{31}w_{13}$ and $v_{31}w_{24}$ will converge to $-\eta_{gd}$ and η_{gd}, respectively. Furthermore, there exists $T_1 = \Theta(c_2^{-1})$ such that for all $t > T_1$, it holds that $|v_{31}(t)w_{13}(t) + \eta_{gd}| = O(e^{-c_2 t})$ and $|v_{31}(t)w_{24}(t) - \eta_{gd}| = O(e^{-c_2 t})$, where $c_2 = Nd(d + N + 1)$.*

Proof. Denote $A = v_{31}w_{13} + \eta_{gd}$, $B = v_{31}w_{24} - \eta_{gd}$. Then, we have

$$\begin{aligned}\frac{d(A^2 + B^2)}{dt} &= 2A\frac{dA}{dt} + 2B\frac{dB}{dt} \\ &= -2c_2(v_{31}^2 + w_{13}^2)A^2 - 2c_2(v_{31}^2 + w_{24}^2)B^2 - 4c_2 w_{13}w_{24}AB \\ &= -2c_2 v_{31}^2(A^2 + B^2) - 2c_2(w_{13}A + w_{24}B)^2 \\ &\leq 0.\end{aligned}$$

Thus, $A^2 + B^2$ is non-increasing, and $0 \leq A(t)^2 + B(t)^2 \leq A(0)^2 + B(0)^2 = 2\eta_{gd}^2$, so $A^2 + B^2$ will finally converge to a constant C such that $0 \leq C \leq 2\eta_{gd}^2$. When $t \to \infty$, we have $A^2 + B^2 \to C$ and $\frac{d(A^2 + B^2)}{dt} \to 0$. Then, we get

$$v_{31}^2 C + (w_{13}A + w_{24}B)^2 \to 0.$$

Assume, for contradiction, that $C > 0$. Then, we have $v_{31} \to 0$ and $w_{13}A + w_{24}B \to 0$. Since w_{13} and w_{24} are bounded, we can get

$$A = v_{31}w_{13} + \eta_{gd} \to \eta_{gd},$$
$$B = v_{31}w_{24} - \eta_{gd} \to -\eta_{gd}.$$

So $A^2 + B^2 \to 2\eta_{gd}^2$, indicating that $C = 2\eta_{gd}^2$. However,

$$\left.\frac{d(A^2+B^2)}{dt}\right|_{t=0} = -2c_2\eta_{gd}^2 < 0.$$

It means that $A^2 + B^2$ is rigorously decreasing at the initial point, so the final limit C cannot be the initial value $A(0)^2 + B(0)^2 = 2\eta_{gd}^2$. It contradicts the fact that $C = 2\eta_{gd}^2$. Thus, we have proven that $C = 0$, so $A = v_{31}w_{13} + \eta_{gd} \to 0$ and $B = v_{31}w_{24} - \eta_{gd} \to 0$. Next, we calculate the bounds of $A(t)$ and $B(t)$. We first focus on the dynamic of v_{31}. From the previous theorem, we know $|w_{24}| < M$ and $|w_{13}| < M$, and $M \geq 1$. Then, when $v_{31} \leq \frac{\eta_{gd}}{4M}$, we have

$$\begin{aligned}\frac{d(w_{24}-w_{13})}{dt} &= -c_2v_{31}^2(w_{24}-w_{13}) + 2c_2\eta_{gd}v_{31} \\ &= c_2v_{31}(2\eta_{gd} - v_{31}(w_{24}-w_{13})) \\ &\geq 0,\end{aligned}$$

so we can get that $w_{24} - w_{13} \geq w_{24}(0) - w_{13}(0) = 1$ in the initial period that $v_{31} \leq \frac{\eta_{gd}}{4M}$. Therefore, for $v_{31} \leq \min\{\frac{\eta_{gd}}{4M}, 1\}$, we have

$$\begin{aligned}\frac{dv_{31}}{dt} &= -c_2(w_{13}A + w_{24}B) \\ &= -c_2(w_{13}(v_{31}w_{13} + \eta_{gd}) + w_{24}(v_{31}w_{24} - \eta_{gd})) \\ &= -c_2v_{31}(w_{13}^2 + w_{24}^2) + c_2\eta_{gd}(w_{24} - w_{13}) \\ &\overset{(i)}{=} -c_2v_{31}(v_{31}^2 + 1) + c_2\eta_{gd}(w_{24} - w_{13}) \\ &\overset{(ii)}{\geq} \frac{c_2\eta_{gd}}{2},\end{aligned}$$

where (i) is by $w_{13}^2 + w_{24}^2 - v_{31}^2 = 1$, and (ii) is by $v_{31} \leq \min\{\frac{\eta_{gd}}{4M}, 1\}$. Thus, there exists $T_1 = \Theta(c_2^{-1} + c_2^{-1}\eta_{gd}^{-1})$ such that for $t \geq T_1$, it holds that $v_{31}(t) \geq \min\{\frac{\eta_{gd}}{4M}, 1\} := C_v$. Then, since

$$\frac{d(A^2+B^2)}{dt} = -2c_2v_{31}^2(A^2+B^2) - 2c_2(w_{13}A + w_{24}B)^2 \leq -2c_2v_{31}^2(A^2+B^2),$$

we have for $t \geq T_1$,

$$\begin{aligned}A(t)^2 + B(t)^2 &\leq (A(T_1)^2 + B(T_1)^2) \cdot e^{-2c_2C_v^2(t-T_1)} \\ &\leq (A(0)^2 + B(0)^2)e^{2c_2C_v^2T_1} \cdot e^{-2c_2C_v^2t} \\ &= 2\eta_{gd}^2e^{2c_2C_v^2T_1} \cdot e^{-2c_2C_v^2t} \\ &= O(e^{-2c_2t}).\end{aligned}$$

Therefore, $A(t) = O(e^{-c_2t})$ and $B(t) = O(e^{-c_2t})$. □

Theorem C.5. *For any $\epsilon > 0$, there exists $T = \Theta(\log(\epsilon^{-1}N^2d^2))$ such that for all $t > T$, it holds that $|v_{31}(t)w_{13}(t) + \eta_{gd}| < \epsilon$, $|v_{31}(t)w_{24}(t) - \eta_{gd}| < \epsilon$, and $\mathcal{L}(\theta(t)) < \epsilon$.*

Proof. We have already obtained $|v_{31}(t)w_{13}(t) + \eta_{gd}| = O(e^{-c_2t}) < \epsilon$ and $|v_{31}(t)w_{24}(t) - \eta_{gd}| = O(e^{-c_2t}) < \epsilon$. Then, the loss is

$$\begin{aligned}\mathcal{L}(\theta(t)) &= \mathbb{E}\left[\frac{1}{2}\|f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[(d+2):(2d+1),-1]} - \boldsymbol{\beta}_1\|_2^2\right] \\ &= \mathbb{E}\left[\frac{1}{2}\left\|v_{31}\mathbf{X}^\top(w_{13}\mathbf{X}\boldsymbol{\beta}_0 + w_{24}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})\right\|_2^2\right] \\ &= \mathbb{E}\left[\frac{1}{2}\left\|(v_{31}w_{13} + \eta_{gd})\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0 + (v_{31}w_{24} - \eta_{gd})\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\right\|_2^2\right]\end{aligned}$$

$$\leq O(c_2 e^{-2c_2 t})$$
$$< \epsilon.$$

□

Given the trained one-layer model, we unroll it L times to get L-layer model. We consider the out-of-distribution generalization. Assume that $\mathbf{x}_i \sim N(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is not necessarily the identity matrix. The test error is defined as

$$\mathcal{L}_{\mathbf{\Sigma}}^{\text{test}} = \frac{1}{2}\mathbb{E}\left[\left\| f_{\text{LSA}}^{(L)}(\mathbf{Z})_{[(d+2):(2d+1),-1]} - \boldsymbol{\beta}\right\|_2^2\right].$$

Then, we can get the following result regarding this OOD loss.

Theorem C.6. *Assume that $\eta_{gd} = O(N^{-1}d^{-1})$, and the out-of-distribution input data $\mathbf{x}_i \sim N(0, \mathbf{\Sigma})$ where $\lambda_{\min}(\mathbf{\Sigma}), \lambda_{\max}(\mathbf{\Sigma}) = \Theta(1)$. It holds that $\mathcal{L}_{\mathbf{\Sigma}}^{test} \leq O(\sigma^2 dN^{-1} + d(r^L + \epsilon^2))$ where $0 < r < 1$ is a constant depending on $\lambda_{\min}(\mathbf{\Sigma}), \eta_{gd}, N$, and ϵ is the bound considered for the training loss.*

Proof. Denote that $\lambda_0 = \lambda_{\min}(\mathbf{\Sigma})$ and $\lambda_1 = \lambda_{\max}(\mathbf{\Sigma})$. We can obtain that

$$\mathbb{P}\left(\lambda_{\min}(\mathbf{X}^\top\mathbf{X}) \leq \frac{1}{2}\lambda_0 N\right) \leq e^{-c_0\lambda_0^2 N}.$$

Then, we have

$$\mathbb{E}\left[\|\mathbf{I} + v_{31}w_{13}\mathbf{X}^\top\mathbf{X}\|_{op}^L\right] \leq \mathbb{E}\left[(1 + v_{31}w_{13}\lambda_{\min}(\mathbf{X}^\top\mathbf{X}))^L\right] \leq \left(1 + \frac{1}{2}v_{31}w_{13}\lambda_0 N\right)^L + O(e^{-N}),$$

and

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=0}^{L-1}(\mathbf{I} + v_{31}w_{13}\mathbf{X}^\top\mathbf{X})^k\right\|_{op}\right] &= \mathbb{E}\left[\sum_{k=0}^{L-1}(1 + v_{31}w_{13}\lambda_{\min}(\mathbf{X}^\top\mathbf{X}))^k\right] \\
&\leq \sum_{k=0}^{L-1}\left(1 + \frac{1}{2}v_{31}w_{13}\lambda_0 N\right)^k + O(e^{-N}) \\
&= \frac{1 - (1 + v_{31}w_{13}\lambda_0 N/2)^L}{-v_{31}w_{13}\lambda_0 N/2} + O(e^{-N}).
\end{aligned}$$

Therefore, given $|v_{31}w_{24} - \eta_{gd}| < \epsilon$ and $|v_{31}w_{13} + \eta_{gd}| < \epsilon$, we can get

$$\begin{aligned}
\mathcal{L}_{\mathbf{\Sigma}}^{\text{test}} &= \frac{1}{2}\mathbb{E}\left[\left\| f_{\text{LSA}}^{(L)}(\mathbf{Z})_{[(d+2):(2d+1),-1]} - \boldsymbol{\beta}\right\|_2^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left\| v_{31}w_{24}\sum_{k=0}^{L-1}(\mathbf{I} + v_{31}w_{13}\mathbf{X}^\top\mathbf{X})^k\mathbf{X}^\top\mathbf{y} - \boldsymbol{\beta}\right\|_2^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left\| w_{24}w_{13}^{-1}[(\mathbf{I} + v_{31}w_{13}\mathbf{X}^\top\mathbf{X})^L - \mathbf{I}]\boldsymbol{\beta} + v_{31}w_{24}\sum_{k=0}^{L-1}(\mathbf{I} + v_{31}w_{13}\mathbf{X}^\top\mathbf{X})^k\mathbf{X}^\top\boldsymbol{\xi} - \boldsymbol{\beta}\right\|_2^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left\| w_{24}w_{13}^{-1}(\mathbf{I} + v_{31}w_{13}\mathbf{X}^\top\mathbf{X})^L\boldsymbol{\beta} - (w_{24}w_{13}^{-1} + 1)\boldsymbol{\beta}\right\|_2^2\right] \\
&\quad + \mathbb{E}\left[\left\| v_{31}w_{24}\sum_{k=0}^{L-1}(\mathbf{I} + v_{31}w_{13}\mathbf{X}^\top\mathbf{X})^k\mathbf{X}^\top\boldsymbol{\xi}\right\|_2^2\right] \\
&\leq (w_{24}w_{13}^{-1})^2\mathbb{E}\left[\left\|(\mathbf{I} + v_{31}w_{13}\mathbf{X}^\top\mathbf{X})^L\boldsymbol{\beta}\right\|_2^2\right] + (w_{24}w_{13}^{-1} + 1)^2\mathbb{E}\left[\|\boldsymbol{\beta}\|_2^2\right] \\
&\quad + (v_{31}w_{24})^2\mathbb{E}\left[\left\|\sum_{k=0}^{L-1}(\mathbf{I} + v_{31}w_{13}\mathbf{X}^\top\mathbf{X})^k\mathbf{X}^\top\boldsymbol{\xi}\right\|_2^2\right] \\
&\leq (w_{24}w_{13}^{-1})^2\mathbb{E}\left[\|\boldsymbol{\beta}\|_2^2\|\mathbf{I} + v_{31}w_{13}\mathbf{X}^\top\mathbf{X}\|_{op}^{2L}\right] + (w_{24}w_{13}^{-1} + 1)^2\mathbb{E}\left[\|\boldsymbol{\beta}\|_2^2\right]
\end{aligned}$$

$$
\begin{aligned}
&+ (v_{31}w_{24})^2 \mathbb{E}\left[\|\boldsymbol{\xi}\|_2^2 \|\mathbf{X}\|_{op}^2 \left\|\sum_{k=0}^{L-1}(\mathbf{I} + v_{31}w_{13}\mathbf{X}^\top\mathbf{X})^k\right\|_{op}^2\right] \\
&\overset{(i)}{\leq} (w_{24}w_{13}^{-1})^2 d\left(1 + \frac{1}{2}v_{31}w_{13}\lambda_0 N\right)^{2L} + (w_{24}w_{13}^{-1}+1)^2 d + \sigma^2 d \cdot \frac{4\lambda_1}{\lambda_0^2 N} + O(e^{-N}) \\
&\overset{(ii)}{\leq} \left(\frac{\eta_{gd}+\epsilon}{\eta_{gd}-\epsilon}\right)^2 d r^{2L} + \left(\frac{2\epsilon}{\eta_{gd}-\epsilon}\right)^2 d + \sigma^2 \cdot \frac{4\lambda_1 d}{\lambda_0^2 N} + O(e^{-N}) \\
&\leq O\left(\frac{\sigma^2 d}{N} + d(r^L + \epsilon^2)\right),
\end{aligned}
$$

where (i) is by the previous results, (ii) is by $|v_{31}w_{13} + \eta_{gd}| < \epsilon$ and $|v_{31}w_{24} - \eta_{gd}| < \epsilon$. Here, $r = 1 + v_{31}w_{13}\lambda_0 N/2$ such that $0 < r < 1$. □

C.2 Explicit CoT without Scratchpad for Weight Prediction

The input with scratchpad is designed as

$$
\hat{\mathbf{Z}}_0 = \begin{bmatrix} \mathbf{X}^\top & \boldsymbol{\beta}_0 \\ \mathbf{y}^\top & 0 \\ 0 & 1 \end{bmatrix},
$$

where $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I})$ is the initial weight vector. A single linear attention layer is

$$
f_{\text{LSA}}(\mathbf{Z}) = \mathbf{Z} + \mathbf{W}^V\mathbf{Z} \cdot \mathbf{Mask}(\mathbf{Z})^\top \mathbf{W}^{KQ}\mathbf{Z},
$$

where

$$
\mathbf{W}^{KQ} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} & \mathbf{W}_{13} \\ \mathbf{W}_{21} & w_{22} & w_{23} \\ \mathbf{W}_{31} & w_{32} & w_{33} \end{bmatrix}, \mathbf{W}^V = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} \\ \mathbf{V}_{21} & v_{22} & v_{23} \\ \mathbf{V}_{31} & v_{32} & v_{33} \end{bmatrix}.
$$

At the i-th generation step, the model predicts the next state $f_{\text{LSA}}(\hat{\mathbf{Z}}_i)_{[:,-1]}$, which is then appended to the sequence $\hat{\mathbf{Z}}_i$ to form $\hat{\mathbf{Z}}_{i+1}$, i.e. $\hat{\mathbf{Z}}_{i+1} = \left[\hat{\mathbf{Z}}_i, f_{\text{LSA}}(\hat{\mathbf{Z}}_i)_{[:,-1]}\right]$. The training loss is

$$
\mathcal{L} = \mathbb{E}\left[\frac{1}{2}\sum_{i=0}^{L-1} \left\|f_{\text{LSA}}(\mathbf{Z}_i)_{[:,-1]} - (\boldsymbol{\beta}_{i+1}, 0, 1)\right\|_2^2\right],
$$

where $\boldsymbol{\beta}_{i+1} = \boldsymbol{\beta}_i - \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})$ is the ground truth GD update, and

$$
\mathbf{Z}_i = \begin{bmatrix} \mathbf{X}^\top & \boldsymbol{\beta}_0 & \boldsymbol{\beta}_1 & \cdots & \boldsymbol{\beta}_i \\ \mathbf{y}^\top & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 1 \end{bmatrix}
$$

is the ground-truth CoT sequence. We also denote that

$$
\mathcal{L}_i = \mathbb{E}\left[\frac{1}{2}\left\|f_{\text{LSA}}(\mathbf{Z}_i)_{[:,-1]} - (\boldsymbol{\beta}_{i+1}, 0, 1)\right\|_2^2\right]
$$

For the training, we consider the gradient flow method

$$
\frac{d\theta}{dt} = -\nabla\mathcal{L}(\theta)
$$

for $\theta := (\mathbf{W}^{KQ}, \mathbf{W}^V)$. We assume that the initialization of w_{23} is $w_{23}(0) = 1$, and all the elements in the other blocks are initialized as 0. Then, we can find the form of $\mathbf{W}^{KQ}(t)$ and $\mathbf{W}^V(t)$ during the training in the following theorem.

Theorem C.7. *For all $t > 0$, $\mathbf{W}^{KQ}(t)$ and $\mathbf{W}^V(t)$ have the following form:*

$$
\mathbf{W}^{KQ}(t) = \begin{bmatrix} w_{11}(t)\cdot\mathbf{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 0 & w_{23}(t) \\ \boldsymbol{0} & 0 & 0 \end{bmatrix}, \mathbf{W}^V(t) = \begin{bmatrix} v_{11}(t)\cdot\mathbf{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 0 & 0 \\ \boldsymbol{0} & 0 & 0 \end{bmatrix}.
$$

Proof. We first prove that the blocks except $\mathbf{W}_{11}$, w_{23}, and $\mathbf{V}_{11}$ will remain 0. The prediction of each step is

$$f_{\text{LSA}}(\mathbf{Z}_i)_{[:,-1]} = \begin{bmatrix} \boldsymbol{\beta}_i + (\mathbf{V}_{11}\mathbf{X}^\top + \mathbf{V}_{12}\mathbf{y}^\top) \cdot (\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_i + \mathbf{y}w_{23}) \\ (\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top) \cdot (\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_i + \mathbf{y}w_{23}) \\ 1 + (\mathbf{V}_{31}\mathbf{X}^\top + v_{32}\mathbf{y}^\top) \cdot (\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_i + \mathbf{y}w_{23}) \end{bmatrix}.$$

Thus, the loss of the i-th step is

$$\begin{aligned} \mathcal{L}_i = &\frac{1}{2}\left\|(\mathbf{V}_{11}\mathbf{X}^\top + \mathbf{V}_{12}\mathbf{y}^\top) \cdot (\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_i + \mathbf{y}w_{23}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})\right\|_2^2 \\ &+ \frac{1}{2}\left((\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top) \cdot (\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_i + \mathbf{y}w_{23})\right)^2 \\ &+ \frac{1}{2}\left((\mathbf{V}_{31}\mathbf{X}^\top + v_{32}\mathbf{y}^\top) \cdot (\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_i + \mathbf{y}w_{23})\right)^2. \end{aligned}$$

We can observe that the blocks $\mathbf{V}_{13}$, v_{23}, w_{33}, $\mathbf{W}_{13}$, w_{22}, $\mathbf{W}_{31}$, w_{32}, and w_{33} are irrelevant to the loss function, so these blocks will stay at 0. Next, we prove that the blocks $\mathbf{V}_{12}$, $\mathbf{V}_{21}$, v_{22}, $\mathbf{V}_{31}$, v_{32}, $\mathbf{W}_{13}$, $\mathbf{W}_{21}$ will also remain 0 for $t > 0$. When $\mathbf{V}_{12}$, $\mathbf{V}_{21}$, v_{22}, $\mathbf{V}_{31}$, v_{32}, $\mathbf{W}_{13}$, $\mathbf{W}_{21}$ are 0, we can calculate their gradients as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}_i}{\partial \mathbf{V}_{12}} &= \mathbb{E}\Big[\left((\mathbf{V}_{11}\mathbf{X}^\top + \mathbf{V}_{12}\mathbf{y}^\top) \cdot (\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_i + \mathbf{y}w_{23}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})\right) \\ &\quad \cdot (\boldsymbol{\beta}_i^\top \mathbf{W}_{11}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{W}_{13}^\top \mathbf{X}^\top \mathbf{y} + \boldsymbol{\beta}_i^\top \mathbf{W}_{21}^\top \mathbf{y}^\top \mathbf{y} + w_{23}\mathbf{y}^\top \mathbf{y})\Big] \\ &= \mathbb{E}\left[\left(\mathbf{V}_{11}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + w_{23}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})\right) \cdot (\boldsymbol{\beta}_i^\top \mathbf{W}_{11}^\top \mathbf{X}^\top \mathbf{y} + w_{23}\mathbf{y}^\top \mathbf{y})\right] \\ &= \mathbb{E}\Big[(\mathbf{V}_{11}\mathbf{X}^\top \mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{V}_{11}\mathbf{X}^\top w_{23}\mathbf{X}\boldsymbol{\beta} + \eta_{gd}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}_i - \eta_{gd}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}) \cdot \boldsymbol{\beta}_i^\top \mathbf{W}_{11}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} \\ &\quad + (\mathbf{V}_{11}\mathbf{X}^\top \mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{V}_{11}\mathbf{X}^\top w_{23}\mathbf{X}\boldsymbol{\beta} + \eta_{gd}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}_i - \eta_{gd}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}) \cdot w_{23}\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\Big], \\ \frac{\partial \mathcal{L}_i}{\partial \mathbf{V}_{21}} &= \mathbb{E}\Big[(\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top)(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_i + \mathbf{y}w_{23}) \\ &\quad \cdot (\boldsymbol{\beta}_i^\top \mathbf{W}_{11}^\top \mathbf{X}^\top \mathbf{X} + \mathbf{W}_{13}^\top \mathbf{X}^\top \mathbf{X} + \boldsymbol{\beta}_i^\top \mathbf{W}_{21}^\top \mathbf{y}^\top \mathbf{X} + w_{23}\mathbf{y}^\top \mathbf{X})\Big] \\ &= 0, \\ \frac{\partial \mathcal{L}_i}{\partial v_{22}} &= \mathbb{E}\Big[(\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top)(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_i + \mathbf{y}w_{23}) \\ &\quad \cdot (\boldsymbol{\beta}_i^\top \mathbf{W}_{11}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{W}_{13}^\top \mathbf{X}^\top \mathbf{y} + \boldsymbol{\beta}_i^\top \mathbf{W}_{21}^\top \mathbf{y}^\top \mathbf{y} + w_{23}\mathbf{y}^\top \mathbf{y})\Big] \\ &= 0, \\ \frac{\partial \mathcal{L}_i}{\partial \mathbf{V}_{31}} &= \mathbb{E}\Big[(\mathbf{V}_{31}\mathbf{X}^\top + v_{32}\mathbf{y}^\top)(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_i + \mathbf{y}w_{23}) \\ &\quad \cdot (\boldsymbol{\beta}_i^\top \mathbf{W}_{11}^\top \mathbf{X}^\top \mathbf{X} + \mathbf{W}_{13}^\top \mathbf{X}^\top \mathbf{X} + \boldsymbol{\beta}_i^\top \mathbf{W}_{21}^\top \mathbf{y}^\top \mathbf{X} + w_{23}\mathbf{y}^\top \mathbf{X})\Big] \\ &= 0, \\ \frac{\partial \mathcal{L}_i}{\partial v_{32}} &= \mathbb{E}\Big[(\mathbf{V}_{31}\mathbf{X}^\top + v_{32}\mathbf{y}^\top)(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_i + \mathbf{y}w_{23}) \\ &\quad \cdot (\boldsymbol{\beta}_i^\top \mathbf{W}_{11}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{W}_{13}^\top \mathbf{X}^\top \mathbf{y} + \boldsymbol{\beta}_i^\top \mathbf{W}_{21}^\top \mathbf{y}^\top \mathbf{y} + w_{23}\mathbf{y}^\top \mathbf{y})\Big] \\ &= 0, \\ \frac{\partial \mathcal{L}_i}{\partial \mathbf{W}_{13}} &= \mathbb{E}\Big[(\mathbf{X}^\top \mathbf{X}\mathbf{V}_{11}^\top + \mathbf{X}^\top \mathbf{y}\mathbf{V}_{12}^\top) \end{aligned}$$

$$\cdot \left((\mathbf{V}_{11}\mathbf{X}^\top + \mathbf{V}_{12}\mathbf{y}^\top)\cdot(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_i + \mathbf{y}w_{23}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})\right)\Bigg]$$

$$= \mathbb{E}\left[\mathbf{X}^\top\mathbf{X}\mathbf{V}_{11}^\top \cdot \left(\mathbf{V}_{11}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + w_{23}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})\right)\right]$$

$$= \mathbb{E}\left[\mathbf{X}^\top\mathbf{X}\mathbf{V}_{11}^\top\mathbf{V}_{11}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + w_{23}\mathbf{Z}\boldsymbol{\beta}) + \mathbf{X}^\top\mathbf{X}\mathbf{V}_{11}^\top\eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{X}\boldsymbol{\beta})\right],$$

$$\frac{\partial \mathcal{L}_i}{\partial \mathbf{W}_{21}} = \mathbb{E}\Bigg[(\mathbf{y}^\top\mathbf{X}\mathbf{V}_{11}^\top + \mathbf{y}^\top\mathbf{y}\mathbf{V}_{12}^\top)$$

$$\cdot \left((\mathbf{V}_{11}\mathbf{X}^\top + \mathbf{V}_{12}\mathbf{y}^\top)\cdot(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{X}\mathbf{W}_{13} + \mathbf{y}\mathbf{W}_{21}\boldsymbol{\beta}_i + \mathbf{y}w_{23}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})\right)\boldsymbol{\beta}_i^\top\Bigg]$$

$$= \mathbb{E}\left[\mathbf{y}^\top\mathbf{X}\mathbf{V}_{11}^\top \cdot \left(\mathbf{V}_{11}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + w_{23}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})\right)\boldsymbol{\beta}_i^\top\right]$$

$$= \mathbb{E}\left[\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\mathbf{V}_{11}^\top\mathbf{V}_{11}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top + w_{23}\mathbf{Z}\boldsymbol{\beta}\boldsymbol{\beta}_i^\top) + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\mathbf{V}_{11}^\top\eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top - \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}_i^\top)\right].$$

We can find that each term in these three gradients is an odd function of $\boldsymbol{\beta}$ or $\boldsymbol{\beta}_i$. Thus, the expectation of these gradients should be 0. Therefore, we obtain that $\mathbf{V}_{12}$, $\mathbf{V}_{21}$, v_{22}, $\mathbf{V}_{31}$, v_{32}, $\mathbf{W}_{13}$, and $\mathbf{W}_{21}$ always stay at 0. Then, we can get

$$f_{\text{LSA}}(\mathbf{Z}_i)_{[:,-1]} = \begin{bmatrix} \boldsymbol{\beta}_i + \mathbf{V}_{11}\mathbf{X}^\top\cdot(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + w_{23}\mathbf{y}) \\ 0 \\ 1 \end{bmatrix}.$$

Next, we prove that $\mathbf{V}_{11}(t) \propto \mathbf{I}$ and $\mathbf{W}_{11}(t) \propto \mathbf{I}$. When $\mathbf{V}_{11} \propto \mathbf{I}$ and $\mathbf{W}_{11} \propto \mathbf{I}$, their gradients are

$$\frac{\partial \mathcal{L}_i}{\partial \mathbf{V}_{11}} = \mathbb{E}\left[\left(\mathbf{V}_{11}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{y}w_{23}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})\right)\cdot(\boldsymbol{\beta}_i^\top\mathbf{W}_{11}^\top\mathbf{X}^\top\mathbf{X} + w_{23}\mathbf{y}^\top\mathbf{X})\right]$$

$$= \mathbb{E}\Bigg[\mathbf{V}_{11}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top\mathbf{W}_{11}^\top\mathbf{X}^\top\mathbf{X} + \eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top\mathbf{W}_{11}^\top\mathbf{X}^\top\mathbf{X}$$

$$+ \mathbf{V}_{11}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}w_{23}^2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X} - \eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}w_{23}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\Bigg]$$

$$= \mathbb{E}\left[(\mathbf{V}_{11}\mathbf{W}_{11}\mathbf{W}_{13}^\top\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top + \eta_{gd}\mathbf{W}_{11}^\top\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top + w_{23}^2\mathbf{V}_{11} - \eta_{gd}w_{23}\mathbf{I})\cdot(\mathbf{X}^\top\mathbf{X})^2\right]$$

$$= N(N+d+1)\cdot(\mathbf{V}_{11}\mathbf{W}_{11}\mathbf{W}_{11}^\top\mathbb{E}(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top) + \eta_{gd}\mathbf{W}_{11}^\top\mathbb{E}(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top) + w_{23}^2\mathbf{V}_{11} - \eta_{gd}w_{23}\mathbf{I}),$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{11}} = \mathbb{E}\left[\mathbf{X}^\top\mathbf{X}\mathbf{V}_{11}^\top\cdot\left(\mathbf{V}_{11}\mathbf{X}^\top(\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i + \mathbf{y}w_{23}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})\right)\boldsymbol{\beta}_i^\top\right]$$

$$= \mathbb{E}\left[\mathbf{X}^\top\mathbf{X}\mathbf{V}_{11}^\top\mathbf{V}_{11}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{11}\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top + \mathbf{X}^\top\mathbf{X}\mathbf{V}_{11}^\top\eta_{gd}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top\right]$$

$$= \mathbb{E}\left[(\mathbf{V}_{11}^\top\mathbf{V}_{11}\mathbf{W}_{11} + \eta_{gd}\mathbf{V}_{11}^\top)\cdot(\mathbf{X}^\top\mathbf{X})^2\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top\right]$$

$$= N(N+d+1)\cdot(\mathbf{V}_{11}^\top\mathbf{V}_{11}\mathbf{W}_{11} + \eta_{gd}\mathbf{V}_{11}^\top)\mathbb{E}(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top).$$

Since $\mathbb{E}(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top) \propto \mathbf{I}$ for all $i > 0$, we can find that when $\mathbf{V}_{11}$ and $\mathbf{W}_{11}$ are proportional to $\mathbf{I}$, the expectation of their gradient is also proportional to $\mathbf{I}$. Therefore, with zero initialization, $\mathbf{V}_{11}(t)$ and $\mathbf{W}_{11}(t)$ can be written as $\mathbf{V}_{11}(t) = v_{11}(t)\cdot\mathbf{I}$ and $\mathbf{W}_{11}(t) = w_{11}(t)\cdot\mathbf{I}$. □

Now, we only have three parameters v_{11}, w_{11}, and w_{23}. We can calculate the corresponding gradients.

Lemma C.8. *The gradients for v_{31}, w_{13}, w_{24} are*

$$\frac{\partial \mathcal{L}_i}{\partial v_{11}} = c_{1i}\cdot w_{11}(v_{11}w_{11} + \eta_{gd}) + c_{2i}\cdot w_{23}(v_{11}w_{23} - \eta_{gd}),$$

$$\frac{\partial \mathcal{L}_i}{\partial w_{11}} = c_{1i}\cdot v_{11}(v_{11}w_{11} + \eta_{gd}),$$

$$\frac{\partial \mathcal{L}_i}{\partial w_{24}} = c_{2i}\cdot v_{11}(v_{11}w_{23} - \eta_{gd}),$$

where $c_{1i} = N(N+d+1)\mathbb{E}(\mathrm{tr}(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top))$, and $c_{2i} = Nd(N+d+1)$.

Proof. Recall that

$$f_{\text{LSA}}(\mathbf{Z}_i)_{[:d,-1]} = \boldsymbol{\beta}_i + v_{11}w_{11}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_i + v_{11}w_{23}\mathbf{X}^\top\mathbf{y}.$$

Then, based on Lemma D.1, we can calculate the gradients.

$$\begin{aligned}
\frac{\partial \mathcal{L}_i}{\partial v_{11}} &= \mathbb{E}\left[(w_{11}\boldsymbol{\beta}_i^\top \mathbf{X}^\top \mathbf{X} + w_{23}\mathbf{y}^\top \mathbf{X}) \cdot \big(v_{11}\mathbf{X}^\top(w_{11}\mathbf{X}\boldsymbol{\beta}_i + w_{23}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})\big)\right] \\
&= \mathbb{E}\left[w_{11}(v_{11}w_{11} + \eta_{gd})\boldsymbol{\beta}_i^\top(\mathbf{X}^\top\mathbf{X})^2\boldsymbol{\beta}_i + w_{23}(v_{11}w_{23} - \eta_{gd})\boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})^2\boldsymbol{\beta}\right] \\
&= N(N+d+1)\mathbb{E}(\mathrm{tr}(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top)) \cdot w_{11}(v_{11}w_{11} + \eta_{gd}) + Nd(N+d+1) \cdot w_{23}(v_{11}w_{23} - \eta_{gd}),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}_i}{\partial w_{11}} &= \mathbb{E}\left[v_{11}\boldsymbol{\beta}_i^\top \mathbf{X}^\top \mathbf{X} \cdot \big(v_{11}\mathbf{X}^\top(w_{11}\mathbf{X}\boldsymbol{\beta}_i + w_{23}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})\big)\right] \\
&= \mathbb{E}\left[v_{11}(v_{11}w_{11} + \eta_{gd})\boldsymbol{\beta}_i^\top(\mathbf{X}^\top\mathbf{X})^2\boldsymbol{\beta}_i\right] \\
&= N(N+d+1)\mathbb{E}(\mathrm{tr}(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top)) \cdot v_{11}(v_{11}w_{11} + \eta_{gd}),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}_i}{\partial w_{23}} &= \mathbb{E}\left[v_{11}\mathbf{y}^\top \mathbf{X} \cdot \big(v_{11}\mathbf{X}^\top(w_{11}\mathbf{X}\boldsymbol{\beta}_i + w_{23}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})\big)\right] \\
&= \mathbb{E}\left[v_{11}(v_{11}w_{23} - \eta_{gd})\boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})^2\boldsymbol{\beta}\right] \\
&= Nd(N+d+1) \cdot v_{11}(v_{11}w_{23} - \eta_{gd}).
\end{aligned}$$

□

We further denote that $c_1 = \sum_{i=0}^{L-1} c_{i1}$ and $c_2 = \sum_{i=0}^{L-1} c_{i2}$. In order to analyze the convergence of three parameters, we first provide their bounded property.

Lemma C.9. *For all $t > 0$, $v_{11}(t)$, $w_{11}(t)$ and $w_{23}(t)$ are bounded, i.e. there exists a positive constant M such that $|v_{11}(t)| \leq M$, $|w_{11}(t)| \leq M$, $|w_{23}(t)| \leq M$.*

Proof. Denote $A = v_{11}w_{11} + \eta_{gd}$, $B = v_{11}w_{23} - \eta_{gd}$. Then, we have the following equations:

$$\frac{dv_{11}}{dt} = -c_1 w_{11} A - c_2 w_{23} B, \quad \frac{dw_{11}}{dt} = -c_1 v_{11} A, \quad \frac{dw_{23}}{dt} = -c_2 v_{11} B.$$

By observation, we can get

$$\begin{aligned}
\frac{d(w_{11}^2 + w_{23}^2 - v_{11}^2)}{dt} &= 2w_{11}\frac{dw_{11}}{dt} + 2w_{23}\frac{dw_{23}}{dt} - 2v_{11}\frac{dv_{11}}{dt} \\
&= -2c_1 v_{11} w_{11} A - 2c_2 v_{11} w_{23} B + 2c_1 v_{11} w_{11} A + 2c_2 v_{11} w_{23} B \\
&= 0.
\end{aligned}$$

Thus, for all $t > 0$, the value of $w_{11}(t)^2 + w_{23}(t)^2 - v_{11}(t)^2$ will not change, which means that $w_{11}(t)^2 + w_{23}(t)^2 - v_{11}(t)^2 = w_{23}(0)^2 = 1$. Next, we check the dynamic of v_{11}^2:

$$\begin{aligned}
\frac{1}{2} \cdot \frac{dv_{11}^2}{dt} &= v_{11}\frac{dv_{11}}{dt} \\
&= -v_{11}(c_1 w_{11} A + c_2 w_{23} B) \\
&= -c_1 v_{11} w_{11}(v_{11}w_{11} + \eta_{gd}) - c_2 v_{11} w_{23}(v_{11}w_{23} - \eta_{gd}) \\
&= -v_{11}^2(c_1 w_{11}^2 + c_2 w_{23}^2) - \eta_{gd} v_{11}(c_1 w_{11} - c_2 w_{23}) \\
&\leq -c_{\min} v_{11}^2(v_{11}^2 + 1) + \eta_{gd} c_{\max} |v_{11}| \sqrt{v_{11}^2 + 1} \\
&= |v_{11}|\sqrt{v_{11}^2 + 1}\left(-c_{\min}|v_{11}|\sqrt{v_{11}^2 + 1} + \eta_{gd} c_{\max}\right),
\end{aligned}$$

where $c_{\min} = \min\{c_1, c_2\}$ and $c_{\max} = \max\{c_1, c_2\}$. We know when $|v_{11}|\sqrt{v_{11}^2 + 1} > \frac{c_{\max}\eta_{gd}}{c_{\min}}$, $\frac{dv_{11}^2}{dt}$ will be smaller than 0, which indicates that $|v_{11}|$ cannot reach $\sqrt{\frac{c_{\max}\eta_{gd}}{c_{\min}}}$. Since $w_{11}^2 + w_{23}^2 = v_{11}^2 + 1$, we can conclude that there exists a positive constant M such that $|v_{11}| < M, |w_{11}| < M, |w_{23}| < M$. □

Next, we derive the following convergence results.

Lemma C.10. *$v_{11}w_{11}$ and $v_{11}w_{23}$ will converge to $-\eta_{gd}$ and η_{gd}, respectively. Furthermore, there exists $T_1 = \Theta(c_2^{-1})$ such that for all $t > T_1$, it holds that $|v_{11}(t)w_{11}(t) + \eta_{gd}| = O(e^{-t})$ and $|v_{11}(t)w_{23}(t) - \eta_{gd}| = O(e^{-t})$.*

Proof. Denote $A = v_{11}w_{11} + \eta_{gd}$, $B = v_{11}w_{23} - \eta_{gd}$. Then, we have

$$\begin{aligned}\frac{1}{2}\cdot\frac{d(c_1A^2+c_2B^2)}{dt} &= c_1A\frac{dA}{dt} + c_2B\frac{dB}{dt}\\ &= c_1A\left(v_{11}\frac{dw_{11}}{dt} + w_{11}\frac{dv_{11}}{dt}\right) + c_2B\left(v_{11}\frac{dw_{23}}{dt} + w_{23}\frac{dv_{11}}{dt}\right)\\ &= -c_1^2(v_{11}^2+w_{11}^2)A^2 - c_1c_2w_{11}w_{23}AB - c_2^2(v_{11}^2+w_{23}^2)B^2 - c_1c_2w_{11}w_{23}AB\\ &= -v_{11}^2(c_1^2A^2+c_2^2B^2) - (c_1w_{11}A + c_2w_{23}B)^2\\ &\le 0.\end{aligned}$$

Thus, $c_1A^2 + c_2B^2$ is non-increasing, and $0 \le c_1A(t)^2 + c_2B(t)^2 \le c_1A(0)^2 + c_2B(0)^2 = (c_1+c_2)\eta_{gd}^2$, so $c_1A^2+c_2B^2$ will finally converge to a constant C such that $0 \le C \le (c_1+c_2)\eta_{gd}^2$. When $t\to\infty$, we have $c_1A^2+c_2B^2 \to C$ and $\frac{d(c_1A^2+c_2B^2)}{dt} \to 0$. Without loss of generality, we assume that $c_1 \le c_2$. Then, we can get

$$\frac{1}{2}\cdot\frac{d(c_1A^2+c_2B^2)}{dt} = -v_{11}^2c_1C - v_{11}^2(c_2^2-c_1c_2)B^2 - (c_1w_{11}A + c_2w_{23}B)^2 \to 0.$$

Assume, for contradiction, that $C > 0$. Then, we have $v_{11}\to 0$ and $c_1w_{11}A + c_2w_{23}B \to 0$. Since w_{11} and w_{23} are bounded, we can get

$$\begin{aligned}A &= v_{11}w_{11} + \eta_{gd} \to \eta_{gd},\\ B &= v_{11}w_{23} - \eta_{gd} \to -\eta_{gd}.\end{aligned}$$

So $c_1A^2 + c_2B^2 \to (c_1+c_2)\eta_{gd}^2$, indicating that $C = (c_1+c_2)\eta_{gd}^2$. However,

$$\frac{1}{2}\cdot\frac{d(c_1A^2+c_2B^2)}{dt}\bigg|_{t=0} = -c_2^2\eta_{gd}^2 < 0.$$

It means that $c_1A^2+c_2B^2$ is rigorously decreasing at the initial point, so the final limit C cannot be the initial value $c_1A(0)^2 + c_2B(0)^2 = (c_1+c_2)\eta_{gd}^2$. It contradicts the fact that the limit $C = (c_1 + c_2)\eta_{gd}^2$. Thus, we have proven that $C = 0$, so $A = v_{11}w_{11} + \eta_{gd} \to 0$ and $B = v_{11}w_{23} - \eta_{gd} \to 0$. Next, we calculate the bounds of $A(t)$ and $B(t)$. We first focus on the dynamic of v_{11}. From the previous theorem, we know $|w_{11}| < M$ and $|w_{23}| < M$, and $M \ge 1$. Then, when $v_{11} \le \frac{\eta_{gd}}{4M}$, we have

$$\begin{aligned}\frac{d(c_2w_{23}-c_1w_{11})}{dt} &= -c_2^2v_{11}B + c_1^2v_{11}A\\ &= -c_2^2v_{11}(v_{11}w_{23}-\eta_{gd}) + c_1^2v_{11}(v_{11}w_{11}+\eta_{gd})\\ &= v_{11}(\eta_{gd}(c_1^2+c_2^2) - v_{11}(c_2^2w_{23} - c_1^2w_{11}))\\ &\ge v_{11}(\eta_{gd}(c_1^2+c_2^2) - Mv_{11}(c_2^2+c_1^2))\\ &\ge 0,\end{aligned}$$

so we can get that $c_2w_{23} - c_1w_{11} \ge c_2w_{23}(0) - c_1w_{11}(0) = c_2$ in the initial period that $v_{11} \le \frac{\eta_{gd}}{4M}$. Therefore, for $v_{11} \le \min\{\frac{\eta_{gd}}{4M}, 1\}$, we have

$$\begin{aligned}\frac{dv_{11}}{dt} &= -c_1w_{11}A - c_2w_{23}B\\ &= -c_1w_{11}(v_{11}w_{11}+\eta_{gd}) - c_2w_{23}(v_{11}w_{23}-\eta_{gd})\\ &= -v_{11}(c_1w_{11}^2 + c_2w_{23}^2) + \eta_{gd}(c_2w_{23} - c_1w_{11})\\ &\overset{(i)}{\ge} -c_2v_{11}(v_{11}^2+1) + \eta_{gd}(c_2w_{23}-c_1w_{11})\\ &\overset{(ii)}{\ge} \frac{c_2\eta_{gd}}{2},\end{aligned}$$

where (i) is by $w_{11}^2 + w_{23}^2 = v_{11}^2 + 1$, and (ii) is by $v_{11} \le \min\{\frac{\eta_{gd}}{4M}, 1\}$. Thus, there exists $T_1 = \Theta(c_2^{-1} + c_2^{-1}\eta_{gd}^{-1})$ such that for $t \ge T_1$, it holds that $v_{11}(t) \ge \min\{\frac{\eta_{gd}}{4M}, 1\} := C_v$. Then, since

$$\frac{d(c_1A^2+c_2B^2)}{dt} = -2v_{11}^2(c_1A^2+c_2B^2) - v_{11}^2(c_2^2-c_1c_2)B^2 - (c_1w_{11}A + c_2w_{23}B)^2$$

$$\leq -2v_{11}^2(c_1A^2 + c_2B^2),$$

we have for $t \geq T_1$,

$$\begin{aligned} c_1A(t)^2 + c_2B(t)^2 &\leq (c_1A(T_1)^2 + c_2B(T_1)^2) \cdot e^{-2C_v^2(t-T_1)} \\ &\leq (c_1A(0)^2 + c_2B(0)^2)e^{2C_v^2T_1} \cdot e^{-2C_v^2t} \\ &= (c_1 + c_2)\eta_{gd}^2 e^{2C_v^2T_1} \cdot e^{-2C_v^2t}. \end{aligned}$$

Therefore, $A(t) = O(e^{-t})$ and $B(t) = O(e^{-t})$. □

Theorem C.11. *For any $\epsilon > 0$, there exists $T = \Theta(\log(\epsilon^{-1}LN^2d^2))$ such that for all $t > T$, it holds that $|v_{11}(t)w_{11}(t) + \eta_{gd}| < \epsilon$, $|v_{11}(t)w_{23}(t) - \eta_{gd}| < \epsilon$, and $\mathcal{L}(\theta(t)) < \epsilon$.*

Proof. We have already obtained $|v_{11}(t)w_{11}(t) + \eta_{gd}| = O(e^{-t}) < \epsilon$ and $|v_{11}(t)w_{23}(t) - \eta_{gd}| = O(e^{-t}) < \epsilon$. Then, the loss is

$$\begin{aligned} \mathcal{L}(\theta(t)) &= \mathbb{E}\left[\frac{1}{2}\sum_{i=0}^{L-1}\|f_{\text{LSA}}(\mathbf{Z}_i)_{[:d,-1]} - \boldsymbol{\beta}_{i+1}\|_2^2\right] \\ &= \mathbb{E}\left[\frac{1}{2}\sum_{i=0}^{L-1}\left\|v_{11}\mathbf{X}^\top(w_{11}\mathbf{X}\boldsymbol{\beta}_i + w_{23}\mathbf{y}) + \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})\right\|_2^2\right] \\ &= \mathbb{E}\left[\frac{1}{2}\sum_{i=0}^{L-1}\left\|(v_{11}w_{11} + \eta_{gd})\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_i + (v_{11}w_{23} - \eta_{gd})\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\right\|_2^2\right] \\ &\leq O((c_1 + c_2)e^{-2t}) \\ &< \epsilon. \end{aligned}$$

□

Next, we consider the out-of-distribution generalization. Assume that $\mathbf{x}_i \sim N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is not necessarily the identity matrix. The test error is defined as

$$\mathcal{L}_{\boldsymbol{\Sigma}}^{\text{test}} = \frac{1}{2}\mathbb{E}\left[\left\|f_{\text{LSA}}^{(L)}(\mathbf{Z})_{[:,-1]} - (\boldsymbol{\beta}, 0, 1)\right\|_2^2\right].$$

Then, we can get the following result regarding this OOD loss.

Theorem C.12. *Assume that $\eta_{gd} = O(N^{-1}d^{-1})$, and the out-of-distribution input data $\mathbf{x}_i \sim N(0, \boldsymbol{\Sigma})$ where $\lambda_{\min}(\boldsymbol{\Sigma}), \lambda_{\max}(\boldsymbol{\Sigma}) = \Theta(1)$. It holds that $\mathcal{L}_{\boldsymbol{\Sigma}}^{test} \leq O(\sigma^2 dN^{-1} + d(r^L + \epsilon^2))$ where $0 < r < 1$ is a constant depending on $\eta_{gd}, \lambda_{\min}(\boldsymbol{\Sigma}), N$, and ϵ is the bound considered for the training loss.*

Proof. Denote that $\lambda_0 = \lambda_{\min}(\boldsymbol{\Sigma})$ and $\lambda_1 = \lambda_{\max}(\boldsymbol{\Sigma})$. We can obtain that

$$\mathbb{P}\left(\lambda_{\min}(\mathbf{X}^\top\mathbf{X}) \leq \frac{1}{2}\lambda_0 N\right) \leq e^{-c_0\lambda_0^2 N}.$$

Then, we have

$$\mathbb{E}\left[\|\mathbf{I} + v_{11}w_{11}\mathbf{X}^\top\mathbf{X}\|_{op}^L\right] \leq \mathbb{E}\left[(1 + v_{11}w_{11}\lambda_{\min}(\mathbf{X}^\top\mathbf{X}))^L\right] \leq \left(1 + \frac{1}{2}v_{11}w_{11}\lambda_0 N\right)^L + O(e^{-N}),$$

and

$$\begin{aligned} \mathbb{E}\left[\left\|\sum_{k=0}^{L-1}(\mathbf{I} + v_{11}w_{11}\mathbf{X}^\top\mathbf{X})^k\right\|_{op}\right] &= \mathbb{E}\left[\sum_{k=0}^{L-1}(1 + v_{11}w_{11}\lambda_{\min}(\mathbf{X}^\top\mathbf{X}))^k\right] \\ &\leq \sum_{k=0}^{L-1}\left(1 + \frac{1}{2}v_{11}w_{11}\lambda_0 N\right)^k + O(e^{-N}) \\ &= \frac{1 - (1 + v_{11}w_{11}\lambda_0 N/2)^L}{-v_{11}w_{11}\lambda_0 N/2} + O(e^{-N}). \end{aligned}$$

Therefore, given $|v_{11}w_{23} - \eta_{gd}| < \epsilon$ and $|v_{11}w_{11} + \eta_{gd}| < \epsilon$, we can get

$$
\begin{aligned}
\mathcal{L}_{\mathbf{\Sigma}}^{\text{test}} &= \frac{1}{2}\mathbb{E}\left[\left\|f_{\text{LSA}}^{(L)}(\mathbf{Z})_{[:,-1]} - (\boldsymbol{\beta}, 0, 1)\right\|_2^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left\|v_{11}w_{23}\sum_{k=0}^{L-1}(\mathbf{I} + v_{11}w_{11}\mathbf{X}^\top\mathbf{X})^k\mathbf{X}^\top\mathbf{y} - \boldsymbol{\beta}\right\|_2^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left\|w_{23}w_{11}^{-1}[(\mathbf{I} + v_{11}w_{11}\mathbf{X}^\top\mathbf{X})^L - \mathbf{I}]\boldsymbol{\beta} + v_{11}w_{23}\sum_{k=0}^{L-1}(\mathbf{I} + v_{11}w_{11}\mathbf{X}^\top\mathbf{X})^k\mathbf{X}^\top\boldsymbol{\xi} - \boldsymbol{\beta}\right\|_2^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left\|w_{23}w_{11}^{-1}(\mathbf{I} + v_{11}w_{11}\mathbf{X}^\top\mathbf{X})^L\boldsymbol{\beta} - (w_{23}w_{11}^{-1} + 1)\boldsymbol{\beta}\right\|_2^2\right] \\
&\quad + \mathbb{E}\left[\left\|v_{11}w_{23}\sum_{k=0}^{L-1}(\mathbf{I} + v_{11}w_{11}\mathbf{X}^\top\mathbf{X})^k\mathbf{X}^\top\boldsymbol{\xi}\right\|_2^2\right] \\
&\le (w_{23}w_{11}^{-1})^2\mathbb{E}\left[\left\|(\mathbf{I} + v_{11}w_{11}\mathbf{X}^\top\mathbf{X})^L\boldsymbol{\beta}\right\|_2^2\right] + (w_{23}w_{11}^{-1} + 1)^2\mathbb{E}\left[\|\boldsymbol{\beta}\|_2^2\right] \\
&\quad + (v_{11}w_{23})^2\mathbb{E}\left[\left\|\sum_{k=0}^{L-1}(\mathbf{I} + v_{11}w_{11}\mathbf{X}^\top\mathbf{X})^k\mathbf{X}^\top\boldsymbol{\xi}\right\|_2^2\right] \\
&\le (w_{23}w_{11}^{-1})^2\mathbb{E}\left[\|\boldsymbol{\beta}\|_2^2\|\mathbf{I} + v_{11}w_{11}\mathbf{X}^\top\mathbf{X}\|_{op}^{2L}\right] + (w_{23}w_{11}^{-1} + 1)^2\mathbb{E}\left[\|\boldsymbol{\beta}\|_2^2\right] \\
&\quad + (v_{11}w_{23})^2\mathbb{E}\left[\|\boldsymbol{\xi}\|_2^2\|\mathbf{X}\|_{op}^2\left\|\sum_{k=0}^{L-1}(\mathbf{I} + v_{11}w_{11}\mathbf{X}^\top\mathbf{X})^k\right\|_{op}^2\right] \\
&\overset{(i)}{\le} (w_{23}w_{11}^{-1})^2 d\left(1 + \frac{1}{2}v_{11}w_{11}\lambda_0 N\right)^{2L} + (w_{23}w_{11}^{-1} + 1)^2 d + \sigma^2 d\cdot\frac{4\lambda_1}{\lambda_0^2 N} + O(e^{-N}) \\
&\overset{(ii)}{\le} \left(\frac{\eta_{gd} + \epsilon}{\eta_{gd} - \epsilon}\right)^2 d r^{2L} + \left(\frac{2\epsilon}{\eta_{gd} - \epsilon}\right)^2 d + \sigma^2\cdot\frac{4\lambda_1 d}{\lambda_0^2 N} + O(e^{-N}) \\
&\le O\left(\frac{\sigma^2 d}{N} + d(r^L + \epsilon^2)\right),
\end{aligned}
$$

where (i) is by the previous results, (ii) is by $|v_{11}w_{11} + \eta_{gd}| < \epsilon$ and $|v_{11}w_{23} - \eta_{gd}| < \epsilon$. Here, $r = 1 + v_{11}w_{11}\lambda_0 N/2$ such that $0 < r < 1$. □

This theorem bounds the OOD test error. The first term is an irreducible error mainly caused by the noise, which matches the asymptotic error of the ordinary least-squares estimator. The second term is an approximation error caused by the model.

C.3 Implicit CoT with Scratchpad for Query Prediction

In the training process, the input with scratchpad is designed as

$$
\tilde{\mathbf{Z}} = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{y}^\top & 0 \\ \boldsymbol{\beta}_0^\top\mathbf{X}^\top & \boldsymbol{\beta}_0^\top\mathbf{x}_q \end{bmatrix},
$$

where $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I})$ is the initial weight vector. A single linear attention layer is

$$
f_{\text{LSA}}(\mathbf{Z}) = \mathbf{Z} + \mathbf{W}^V\mathbf{Z}\cdot\mathbf{Mask}(\mathbf{Z})^\top\mathbf{W}^{KQ}\mathbf{Z},
$$

where

$$
\mathbf{W}^{KQ} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} & \mathbf{W}_{13} \\ \mathbf{W}_{21} & w_{22} & w_{23} \\ \mathbf{W}_{31} & w_{32} & w_{33} \end{bmatrix}, \mathbf{W}^V = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} \\ \mathbf{V}_{21} & v_{22} & v_{23} \\ \mathbf{V}_{31} & v_{32} & v_{33} \end{bmatrix}.
$$

Based on the teacher GD algorithm, the training loss is designed as

$$\mathcal{L} = \mathbb{E}\left[\frac{1}{2}(f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} - \boldsymbol{\beta}_1^\top \mathbf{x}_q)^2\right],$$

where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 - \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})$ is the ground truth GD update. Then, we can get

$$f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} = \boldsymbol{\beta}_0^\top \mathbf{x}_q + (\mathbf{V}_{31}\mathbf{X}^\top + v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top \mathbf{X}^\top) \cdot [\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q + \mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0^\top \mathbf{x}_q + \mathbf{y}\mathbf{W}_{21}\mathbf{x}_q + \mathbf{y}w_{23}\boldsymbol{\beta}_0^\top \mathbf{x}_q + \mathbf{X}\boldsymbol{\beta}_0\mathbf{W}_{31}\mathbf{x}_q + \mathbf{X}\boldsymbol{\beta}_0 w_{33}\boldsymbol{\beta}_0^\top \mathbf{x}_q].$$

We assume that w_{23} and w_{33} are fixed at 0. The initialization of v_{32} and v_{33} is $v_{32}(0), v_{33}(0) \sim N(0,1)$, and all the elements in the other blocks are initialized as 0. Without loss of generality, we assume $v_{32}(0) > v_{33}(0)$. We use the gradient flow method:

$$\frac{d\theta}{dt} = -\nabla\mathcal{L}(\theta)$$

for $\theta := (\mathbf{W}^{KQ}, \mathbf{W}^V \backslash \{w_{23}, w_{33}\})$. Then, we can find the form of $\mathbf{W}^{KQ}(t)$ and $\mathbf{W}^V(t)$ during the training in the following theorem.

Theorem C.13. *For all $t > 0$, $\mathbf{W}^{KQ}(t)$, $\mathbf{W}^V(t)$ have the following form:*

$$\mathbf{W}^{KQ}(t) = \begin{bmatrix} \mu(t)\cdot\mathbf{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 0 & 0 \\ \boldsymbol{0} & 0 & 0 \end{bmatrix}, \mathbf{W}^V(t) = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 0 & 0 \\ \boldsymbol{0} & v_{32}(t) & v_{33}(t) \end{bmatrix}.$$

Proof. Since the output is

$$f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} = \boldsymbol{\beta}_0^\top \mathbf{x}_q + (\mathbf{V}_{31}\mathbf{X}^\top + v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top \mathbf{X}^\top) \cdot [\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q + \mathbf{X}\mathbf{W}_{13}\boldsymbol{\beta}_0^\top \mathbf{x}_q + \mathbf{y}\mathbf{W}_{21}\mathbf{x}_q + \mathbf{X}\boldsymbol{\beta}_0\mathbf{W}_{31}\mathbf{x}_q],$$

we can observe that $\mathbf{W}_{12}$, w_{22}, w_{32}, $\mathbf{V}_{11}$, $\mathbf{V}_{12}$, $\mathbf{V}_{13}$, $\mathbf{V}_{21}$, v_{22}, v_{23} are irrelevant to the prediction, so they stay at 0 for all $t > 0$. Next, we prove that when $\mathbf{V}_{31}$, $\mathbf{W}_{13}$, $\mathbf{W}_{21}$, $\mathbf{W}_{31}$ are 0, the gradients for these blocks are always 0. We first calculate those gradients as follows:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{V}_{31}} &= \mathbb{E}\left[\left(f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} - \boldsymbol{\beta}_1^\top\mathbf{x}_q\right)\cdot\frac{\partial f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]}}{\partial\mathbf{V}_{31}}\right] \\
&= \mathbb{E}\left[(\boldsymbol{\beta}_0^\top\mathbf{x}_q + (v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top\mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q - \boldsymbol{\beta}_1^\top\mathbf{x}_q)\cdot\mathbf{x}_q^\top\mathbf{W}_{11}^\top\mathbf{X}^\top\mathbf{X}\right] \\
&= \mathbb{E}\left[((v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top\mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}(\boldsymbol{\beta}_0^\top\mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q)\cdot\mathbf{x}_q^\top\mathbf{W}_{11}^\top\mathbf{X}^\top\mathbf{X}\right] \\
&= \mathbb{E}\left[(\mathbf{y}^\top\mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q) + \boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q))\cdot\mathbf{x}_q^\top\mathbf{W}_{11}^\top\mathbf{X}^\top\mathbf{X}\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q)\mathbf{x}_q^\top\mathbf{W}_{11}^\top\mathbf{X}^\top\mathbf{X} + \boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)\mathbf{x}_q^\top\mathbf{W}_{11}^\top\mathbf{X}^\top\mathbf{X}\right],
\end{aligned}$$

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{W}_{13}} &= \mathbb{E}\left[\left(f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} - \boldsymbol{\beta}_1^\top\mathbf{x}_q\right)\cdot\frac{\partial f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]}}{\partial\mathbf{W}_{13}}\right] \\
&= \mathbb{E}\left[(\boldsymbol{\beta}_0^\top\mathbf{x}_q + (v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top\mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q - \boldsymbol{\beta}_1^\top\mathbf{x}_q)\cdot\mathbf{X}^\top(v_{32}\mathbf{y} + v_{33}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\boldsymbol{\beta}_0\right] \\
&= \mathbb{E}\left[((v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top\mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}(\boldsymbol{\beta}_0^\top\mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q)\cdot\mathbf{X}^\top(v_{32}\mathbf{y} + v_{33}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\boldsymbol{\beta}_0\right] \\
&= \mathbb{E}\left[(\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q) + \boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q))\cdot\mathbf{X}^\top\mathbf{X}(v_{32}\boldsymbol{\beta} + v_{33}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\boldsymbol{\beta}_0\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q)v_{32}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\mathbf{x}_q^\top\boldsymbol{\beta}_0 + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q)v_{33}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0\mathbf{x}_q^\top\boldsymbol{\beta}_0\right] \\
&\quad + \mathbb{E}\left[\boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{32}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\mathbf{x}_q^\top\boldsymbol{\beta}_0 + \boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{33}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0\mathbf{x}_q^\top\boldsymbol{\beta}_0\right],
\end{aligned}$$

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{W}_{21}} &= \mathbb{E}\left[\left(f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} - \boldsymbol{\beta}_1^\top\mathbf{x}_q\right)\cdot\frac{\partial f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]}}{\partial\mathbf{W}_{21}}\right] \\
&= \mathbb{E}\left[(\boldsymbol{\beta}_0^\top\mathbf{x}_q + (v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top\mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q - \boldsymbol{\beta}_1^\top\mathbf{x}_q)\cdot\mathbf{y}^\top(v_{32}\mathbf{y} + v_{33}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[((v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top\mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}(\boldsymbol{\beta}_0^\top\mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q)\cdot\mathbf{y}^\top(v_{32}\mathbf{y} + v_{33}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[(\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q) + \boldsymbol{\beta}_0^\top\mathbf{X}^\top\mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q))\cdot\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}(v_{32}\boldsymbol{\beta} + v_{33}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q)v_{32}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\mathbf{x}_q^\top + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q)v_{33}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0\mathbf{x}_q^\top\right]
\end{aligned}$$

$$+ \mathbb{E}\left[\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{32}\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\mathbf{x}_q^\top + \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{33}\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}_0\mathbf{x}_q^\top\right],$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{31}} &= \mathbb{E}\left[\left(f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} - \boldsymbol{\beta}_1^\top \mathbf{x}_q\right) \cdot \frac{\partial f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]}}{\partial \mathbf{W}_{31}}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{\beta}_0^\top \mathbf{x}_q + (v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top \mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q - \boldsymbol{\beta}_1^\top \mathbf{x}_q\right) \cdot \boldsymbol{\beta}_0^\top \mathbf{X}^\top (v_{32}\mathbf{y} + v_{33}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[\left((v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top \mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}(\boldsymbol{\beta}_0^\top \mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q\right) \cdot \boldsymbol{\beta}_0^\top \mathbf{X}^\top (v_{32}\mathbf{y} + v_{33}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q) + \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)\right) \cdot \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{32}\boldsymbol{\beta} + v_{33}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q)v_{32}\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\mathbf{x}_q^\top + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q)v_{33}\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}_0\mathbf{x}_q^\top\right] \\
&\quad + \mathbb{E}\left[\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{32}\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\mathbf{x}_q^\top + \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{33}\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}_0\mathbf{x}_q^\top\right].
\end{aligned}$$

We can find that each term in the gradients for $\mathbf{V}_{31}$, $\mathbf{W}_{13}$, $\mathbf{W}_{21}$, $\mathbf{W}_{31}$ is an odd function of $\boldsymbol{\beta}$ or $\boldsymbol{\beta}_0$. Since $\boldsymbol{\beta} \sim N(0, \mathbf{I})$ and $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I})$, the expectation of these four gradients should be 0. Therefore, we know that $\mathbf{V}_{31}$, $\mathbf{W}_{13}$, $\mathbf{W}_{21}$ and $\mathbf{W}_{31}$ always stay at 0. Then, we can get

$$f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} = \boldsymbol{\beta}_0^\top \mathbf{x}_q + (v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top \mathbf{X}^\top) \cdot \mathbf{X}\mathbf{W}_{11}\mathbf{x}_q.$$

Next, we prove that $\mathbf{W}_{11}(t) \propto \mathbf{I}$ for all $t > 0$. We can calculate the gradient as follows:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{11}} &= \mathbb{E}\left[\left(f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} - \boldsymbol{\beta}_1^\top \mathbf{x}_q\right) \cdot \frac{\partial f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]}}{\partial \mathbf{W}_{11}}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{\beta}_0^\top \mathbf{x}_q + (v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top \mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q - \boldsymbol{\beta}_1^\top \mathbf{x}_q\right) \cdot \mathbf{X}^\top (v_{32}\mathbf{y} + v_{33}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[\left((v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top \mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}(\boldsymbol{\beta}_0^\top \mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q\right) \cdot \mathbf{X}^\top (v_{32}\mathbf{y} + v_{33}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q) + \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)\right) \cdot \mathbf{X}^\top \mathbf{X}(v_{32}\boldsymbol{\beta} + v_{33}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q)v_{32}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\mathbf{x}_q^\top + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q)v_{33}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}_0\mathbf{x}_q^\top\right] \\
&\quad + \mathbb{E}\left[\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{32}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\mathbf{x}_q^\top + \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{33}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}_0\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{32}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}\mathbf{x}_q)v_{32}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\mathbf{x}_q^\top + \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{33}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{33}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}_0\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[v_{32}(\mathbf{X}^\top \mathbf{X})^2(v_{32}\mathbf{W}_{11} - \eta_{gd}\mathbf{I}) + v_{33}(\mathbf{X}^\top \mathbf{X})^2(v_{33}\mathbf{W}_{11} + \eta_{gd}\mathbf{I})\right] \\
&= \mathbb{E}\left[v_{32}N(N + d + 1)(v_{32}\mathbf{W}_{11} - \eta_{gd}\mathbf{I})\right] + \mathbb{E}\left[v_{33}N(N + d + 1)(v_{33}\mathbf{W}_{11} + \eta_{gd}\mathbf{I})\right],
\end{aligned}$$

where the last equation is by Lemma D.1. We can find that when $\mathbf{W}_{11}$ is proportional to $\mathbf{I}$, the expectation of the gradient $\mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{11}}\right]$ is also proportional to $\mathbf{I}$. Therefore, $\mathbf{W}_{11}(t)$ can be written as $\mathbf{W}_{11}(t) = \mu(t) \cdot \mathbf{I}$.

□

Now, we can only consider three parameters μ, v_{32}, and v_{33}. The following lemma shows their gradients.

Lemma C.14. *The gradients for μ, v_{32}, v_{33} are*

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu} &= c_1[v_{32}(v_{32}\mu - \eta_{gd}) + v_{33}(v_{33}\mu + \eta_{gd})], \\
\frac{\partial \mathcal{L}}{\partial v_{32}} &= c_1[\mu(v_{32}\mu - \eta_{gd})], \\
\frac{\partial \mathcal{L}}{\partial v_{33}} &= c_1[\mu(v_{33}\mu + \eta_{gd})],
\end{aligned}$$

where $c_1 = dN(d + N + 1)$.

Proof. Recall that

$$\begin{aligned}
f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} - \boldsymbol{\beta}_1^\top \mathbf{x}_q &= \boldsymbol{\beta}_0^\top \mathbf{x}_q + (v_{32}\mathbf{y}^\top + v_{33}\boldsymbol{\beta}_0^\top \mathbf{X}^\top) \cdot \mu\mathbf{X}\mathbf{x}_q - \boldsymbol{\beta}_0^\top \mathbf{x}_q + \eta_{gd}(\boldsymbol{\beta}_0^\top \mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q \\
&= (v_{32}\mu - \eta_{gd})\mathbf{y}^\top \mathbf{X}\mathbf{x}_q + (v_{33}\mu + \eta_{gd})\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q.
\end{aligned}$$

By Lemma D.1, we can calculate the corresponding gradients.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu} &= \mathbb{E}\left[\left((v_{32}\mu - \eta_{gd})\mathbf{y}^\top \mathbf{X}\mathbf{x}_q + (v_{33}\mu + \eta_{gd})\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right) \cdot \left(v_{32}\mathbf{y}^\top \mathbf{X}\mathbf{x}_q + v_{33}\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right)\right] \\
&= \mathbb{E}\left[v_{32}(v_{32}\mu - \eta_{gd})\mathbf{y}^\top \mathbf{X}\mathbf{x}_q \mathbf{y}^\top \mathbf{X}\mathbf{x}_q + v_{33}(v_{33}\mu + \eta_{gd})\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right] \\
&= \mathbb{E}\left[v_{32}(v_{32}\mu - \eta_{gd})\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q \mathbf{x}_q^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + v_{33}(v_{33}\mu + \eta_{gd})\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q \mathbf{x}_q^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}_0\right] \\
&= \left[v_{32}(v_{32}\mu - \eta_{gd}) + v_{33}(v_{33}\mu + \eta_{gd})\right] \cdot \mathbb{E}\left[\mathrm{tr}((\mathbf{X}^\top \mathbf{X})^2)\right] \\
&= dN(d+N+1)[v_{32}(v_{32}\mu - \eta_{gd}) + v_{33}(v_{33}\mu + \eta_{gd})],
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{v_{32}} &= \mathbb{E}\left[\left((v_{32}\mu - \eta_{gd})\mathbf{y}^\top \mathbf{X}\mathbf{x}_q + (v_{33}\mu + \eta_{gd})\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right) \cdot \mu \mathbf{y}^\top \mathbf{X}\mathbf{x}_q\right] \\
&= \mathbb{E}\left[\mu(v_{32}\mu - \eta_{gd})\mathbf{y}^\top \mathbf{X}\mathbf{x}_q \mathbf{y}^\top \mathbf{X}\mathbf{x}_q\right] \\
&= \mu(v_{32}\mu - \eta_{gd}) \cdot \mathbb{E}\left[\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q \mathbf{x}_q^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\right] \\
&= \mu(v_{32}\mu - \eta_{gd}) \cdot \mathbb{E}\left[\mathrm{tr}((\mathbf{X}^\top \mathbf{X})^2)\right] \\
&= dN(d+N+1)\left[\mu(v_{32}\mu - \eta_{gd})\right], \\
\frac{\partial \mathcal{L}}{v_{33}} &= \mathbb{E}\left[\left((v_{32}\mu - \eta_{gd})\mathbf{y}^\top \mathbf{X}\mathbf{x}_q + (v_{33}\mu + \eta_{gd})\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right) \cdot \mu \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right] \\
&= \mathbb{E}\left[\mu(v_{33}\mu + \eta_{gd})\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right] \\
&= \mu(v_{33}\mu + \eta_{gd}) \cdot \mathbb{E}\left[\mathrm{tr}((\mathbf{X}^\top \mathbf{X})^2)\right] \\
&= dN(d+N+1)\left[\mu(v_{33}\mu + \eta_{gd})\right].
\end{aligned}$$

□

In order to analyze the convergence of three parameters, we first provide their bounded property.

Lemma C.15. *For all $t > 0$, $\mu(t)$, $v_{32}(t)$ and $v_{33}(t)$ are bounded, i.e. there exists a positive constant M such that $|\mu(t)| \le M$, $|v_{32}(t)| \le M$, $|v_{33}(t)| \le M$.*

Proof. Denote $A = v_{32}\mu - \eta_{gd}$, $B = v_{33}\mu + \eta_{gd}$. According to the previous lemma, we have the following equations:

$$\frac{d\mu}{dt} = -c_1(v_{32}A + v_{33}B), \quad \frac{dv_{32}}{dt} = -c_1\mu A, \quad \frac{dv_{33}}{dt} = -c_1\mu B.$$

By observation, we can get

$$\begin{aligned}
\frac{d(v_{32}^2 + v_{33}^2 - \mu^2)}{dt} &= 2v_{32}\frac{dv_{32}}{dt} + 2v_{33}\frac{dv_{33}}{dt} - 2\mu\frac{d\mu}{dt} \\
&= -2c_1 v_{32}\mu A - 2c_1 v_{33}\mu B + 2c_1\mu(v_{32}A + v_{33}B) \\
&= 0.
\end{aligned}$$

Thus, for all $t > 0$, the value of $v_{32}(t)^2 + v_{33}(t)^2 - \mu(t)^2$ will not change, which means that $v_{32}(t)^2 + v_{33}(t)^2 - \mu(t)^2 = v_{32}(0)^2 + v_{33}(0)^2$. Next, in order to give the bound of μ, v_{32}, v_{33}, we check the dynamic of $\mu^2 + v_{32}^2 + v_{33}^2$:

$$\begin{aligned}
\frac{d(\mu^2 + v_{32}^2 + v_{33}^2)}{dt} &= 2\mu\frac{d\mu}{dt} + 2v_{32}\frac{dv_{32}}{dt} + 2v_{33}\frac{dv_{33}}{dt} \\
&= -2c_1\mu(v_{32}A + v_{33}B) - 2c_1 v_{32}\mu A - 2c_1 v_{33}\mu B \\
&= -4c_1 v_{32}\mu A - 4c_1 v_{33}\mu B \\
&= -4c_1 v_{32}\mu(v_{32}\mu - \eta_{gd}) - 4c_1 v_{33}\mu(v_{33}\mu + \eta_{gd}) \\
&= -4c_1\mu^2(v_{32}^2 + v_{33}^2) + 4c_1\eta_{gd}\mu(v_{32} - v_{33}) \\
&\le -4c_1\mu^2(v_{32}^2 + v_{33}^2) + 4\sqrt{2}c_1\eta_{gd}|\mu|\sqrt{v_{32}^2 + v_{33}^2} \\
&= -4c_1\mu^2(\mu^2 + v_{32}(0)^2 + v_{33}(0)^2) + 4\sqrt{2}c_1\eta_{gd}|\mu|\sqrt{\mu^2 + v_{32}(0)^2 + v_{33}(0)^2} \\
&= -4c_1|\mu|\sqrt{\mu^2 + v_{32}(0)^2 + v_{33}(0)^2}(|\mu|\sqrt{\mu^2 + v_{32}(0)^2 + v_{33}(0)^2} - \sqrt{2}\eta_{gd}),
\end{aligned}$$

where the inequality is by $v_{32} - v_{33} \leq \sqrt{2(v_{32}^2 + v_{33}^2)}$. Once $\mu^2 + v_{32}^2 + v_{33}^2 > v_{32}(0)^2 + v_{33}(0)^2 + 4\eta_{gd}$, we have $\mu^2 > 2\eta_{gd}$, then $\frac{d(\mu^2+v_{32}^2+v_{33}^2)}{dt} < 0$. Therefore, we have $\mu^2 + v_{32}^2 + v_{33}^2 \leq v_{32}(0)^2 + v_{33}(0)^2 + 4\eta_{gd}$. So, there exists a positive constant M such that $|\mu| < M, |v_{32}| < M, |v_{33}| < M$. □

Next, we can derive the following convergence results.

Lemma C.16. *$v_{32}\mu$ and $v_{33}\mu$ will converge to η_{gd} and $-\eta_{gd}$, respectively. Furthermore, there exists $T_1 = \Theta\left(c_1^{-1}(v_{32}(0)^2 + v_{33}(0)^2)^{-1} + c_1^{-1}\eta_{gd}^{-1}\right)$ such that for all $t > T_1$, it holds that $|v_{32}(t)\mu(t) - \eta_{gd}| = O(e^{-c_1 t})$ and $|v_{33}(t)\mu(t) + \eta_{gd}| = O(e^{-c_1 t})$, where $c_1 = dN(d+N+1)$.*

Proof. Denote $A = v_{32}\mu - \eta_{gd}$, $B = v_{33}\mu + \eta_{gd}$. Then, we have

$$\begin{aligned}\frac{d(A^2+B^2)}{dt} &= 2A\frac{dA}{dt} + 2B\frac{dB}{dt} \\ &= -2c_1(\mu^2 + v_{32}^2)A^2 - 2c_1(\mu^2 + v_{33}^2)B^2 - 4c_1 v_{32} v_{33} AB \\ &= -2c_1\mu^2(A^2+B^2) - 2c_1(v_{32}A + v_{33}B)^2 \\ &\leq 0.\end{aligned}$$

Thus, A^2+B^2 is non-increasing, and $0 \leq A(t)^2 + B(t)^2 \leq A(0)^2 + B(0)^2 = 2\eta_{gd}^2$, so A^2+B^2 will finally converge to a constant C such that $0 \leq C \leq 2\eta_{gd}^2$. When $t \to \infty$, we have $A^2 + B^2 \to C$ and $\frac{d(A^2+B^2)}{dt} \to 0$. Then, we get

$$\mu^2 C + (v_{32}A + v_{33}B)^2 \to 0.$$

Assume, for contradiction, that $C > 0$. Then, we have $\mu \to 0$ and $v_{32}A + v_{33}B \to 0$. Since v_{32} and v_{33} are bounded, we can get

$$\begin{aligned}A &= v_{32}\mu - \eta_{gd} \to -\eta_{gd}, \\ B &= v_{33}\mu + \eta_{gd} \to \eta_{gd}.\end{aligned}$$

So $A^2 + B^2 \to 2\eta_{gd}^2$, indicating that $C = 2\eta_{gd}^2$. However,

$$\left.\frac{d(A^2+B^2)}{dt}\right|_{t=0} = -2c_1\eta_{gd}^2(v_{32}(0) - v_{33}(0))^2 < 0.$$

It means that $A^2 + B^2$ is rigorously decreasing at the initial point, so the final limit C cannot be the initial value $A(0)^2 + B(0)^2 = 2\eta_{gd}^2$. It contradicts the fact that $C = 2\eta_{gd}^2$. Thus, we have proven that $C = 0$, so $A = v_{32}\mu - \eta_{gd} \to 0$ and $B = v_{33}\mu + \eta_{gd} \to 0$. Next, we calculate the bounds of $A(t)$ and $B(t)$. We first focus on the dynamic of μ. From the previous theorem, we know $|w_{24}| < M$ and $|w_{13}| < M$, and $M \geq 1$. When $\mu \leq \frac{\eta_{gd}}{4M}$, we have

$$\begin{aligned}\frac{d(v_{32} - v_{33})}{dt} &= -c_1\mu^2(v_{32} - v_{33}) + 2c_1\eta_{gd}\mu \\ &= c_1\mu(2\eta_{gd} - \mu(v_{32} - v_{33})) \\ &\geq 0,\end{aligned}$$

so we can get that $v_{32} - v_{33} \geq v_{32}(0) - v_{33}(0)$ when $\mu \leq \frac{\eta_{gd}}{4M}$. Denote that

$$C_\mu = \min\left\{\frac{\eta_{gd}}{4M}, \frac{\eta_{gd}(v_{32}(0) - v_{33}(0))}{4(v_{32}(0)^2 + v_{33}(0)^2)}, \sqrt{v_{32}(0)^2 + v_{33}(0)^2}\right\}$$

Then, for $\mu \leq C_\mu$, we have

$$\begin{aligned}\frac{d\mu}{dt} &= -c_1(v_{32}A + v_{33}B) \\ &= -c_1(v_{32}(v_{32}\mu - \eta_{gd}) + v_{33}(v_{33}\mu + \eta_{gd})) \\ &= -c_1\mu(v_{32}^2 + v_{33}^2) + c_1\eta_{gd}(v_{32} - v_{33}) \\ &\overset{(i)}{=} -c_1\mu(\mu^2 + v_{32}(0)^2 + v_{33}(0)^2) + c_1\eta_{gd}(v_{32} - v_{33}) \\ &\overset{(ii)}{\geq} \frac{c_1\eta_{gd}(v_{32}(0) - v_{33}(0))}{2},\end{aligned}$$

where (i) is by $v_{32}^2 + v_{33}^2 - \mu^2 = v_{32}(0)^2 + v_{33}(0)^2 - \mu(0)^2$, and (ii) is by $\mu \leq C_\mu$. Thus, there exists $T_1 = \Theta(c_1^{-1} + c_1^{-1}\eta_{gd}^{-1})$ such that for $t \geq T_1$, it holds that $\mu(t) \geq C_\mu$. Then, since

$$\frac{d(A^2+B^2)}{dt} = -2c_1\mu^2(A^2+B^2) - 2c_1(v_{32}A + v_{33}B)^2 \leq -2c_1\mu^2(A^2+B^2),$$

we have for $t \geq T_1$,

$$\begin{aligned} A(t)^2 + B(t)^2 &\leq (A(T_1)^2 + B(T_1)^2) \cdot e^{-2c_1C_\mu^2(t-T_1)} \\ &\leq (A(0)^2 + B(0)^2)e^{2c_1C_\mu^2 T_1} \cdot e^{-2c_1C_\mu^2 t} \\ &= 2\eta_{gd}^2 e^{2c_1C_\mu^2 T_1} \cdot e^{-2c_1C_\mu^2 t} \\ &= O(e^{-2c_1 t}). \end{aligned}$$

Therefore, $A(t) = O(e^{-c_1 t})$ and $B(t) = O(e^{-c_1 t})$. □

Theorem C.17. *For any $\epsilon > 0$, there exists $T = \Theta(\log(\epsilon^{-1}N^2d^2))$ such that for all $t > T$, it holds that $|v_{32}(t)\mu(t) - \eta_{gd}| < \epsilon$, $|v_{33}(t)\mu(t) + \eta_{gd}| < \epsilon$, and $\mathcal{L}(\theta(t)) < \epsilon$.*

Proof. We have already obtained $|v_{32}(t)\mu(t) - \eta_{gd}| = O(e^{-c_1 t}) < \epsilon$ and $|v_{33}(t)\mu(t) + \eta_{gd}| = O(e^{-c_1 t}) < \epsilon$. Then, the loss is

$$\begin{aligned} \mathcal{L}(\theta(t)) &= \mathbb{E}\left[(f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} - \boldsymbol{\beta}_1^\top \mathbf{x}_q)^2\right] \\ &= \mathbb{E}\left[\left[(v_{32}(t)\mu(t) - \eta_{gd})\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q + (v_{33}(t)\mu(t) + \eta_{gd})\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right]^2\right] \\ &\leq O(c_1 e^{-2c_1 t}) \\ &< \epsilon. \end{aligned}$$

□

Given the trained one-layer model, we unroll it L times to get L-layer model. We consider the out-of-distribution generalization. Assume that $\mathbf{x}_i \sim N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is not necessarily the identity matrix. The test error is defined as

$$\mathcal{L}_{\boldsymbol{\Sigma}}^{\text{test}} = \frac{1}{2}\mathbb{E}\left[(f_{\text{LSA}}^{(L)}(\mathbf{Z})_{[-1,-1]} - y_q)^2\right].$$

Then, we can get the following result regarding this OOD loss.

Theorem C.18. *Assume that $\eta_{gd} = O(N^{-1}d^{-1})$, and the out-of-distribution input data $\mathbf{x}_i \sim N(0, \boldsymbol{\Sigma})$ where $\lambda_{\min}(\boldsymbol{\Sigma}), \lambda_{\max}(\boldsymbol{\Sigma}) = \Theta(1)$. It holds that $\mathcal{L}_{\boldsymbol{\Sigma}}^{test} \leq O(\sigma^2 dN^{-1} + d(r^L + \epsilon^2))$ where $0 < r < 1$ is a constant depending on $\eta_{gd}, \lambda_{\min}, N$, and ϵ is the bound considered for the training loss.*

Proof. Denote that $\lambda_0 = \lambda_{\min}(\boldsymbol{\Sigma})$ and $\lambda_1 = \lambda_{\max}(\boldsymbol{\Sigma})$. We can obtain that

$$\mathbb{P}\left(\lambda_{\min}(\mathbf{X}^\top\mathbf{X}) \leq \frac{1}{2}\lambda_0 N\right) \leq e^{-c_0\lambda_0^2 N}.$$

Then, we have

$$\mathbb{E}\left[\|\mathbf{I} + v_{33}\mu\mathbf{X}^\top\mathbf{X}\|_{op}^L\right] \leq \mathbb{E}\left[(1 + v_{33}\mu\lambda_{\min}(\mathbf{X}^\top\mathbf{X}))^L\right] \leq \left(1 + \frac{1}{2}v_{33}\mu\lambda_0 N\right)^L + O(e^{-N}),$$

and

$$\begin{aligned} \mathbb{E}\left[\left\|\sum_{k=0}^{L-1}(\mathbf{I} + v_{33}\mu\mathbf{X}^\top\mathbf{X})^k\right\|_{op}\right] &= \mathbb{E}\left[\sum_{k=0}^{L-1}(1 + v_{33}\mu\lambda_{\min}(\mathbf{X}^\top\mathbf{X}))^k\right] \\ &\leq \sum_{k=0}^{L-1}\left(1 + \frac{1}{2}v_{33}\mu\lambda_0 N\right)^k + O(e^{-N}) \\ &= \frac{1 - (1 + v_{33}\mu\lambda_0 N/2)^L}{-v_{33}\mu\lambda_0 N/2} + O(e^{-N}). \end{aligned}$$

Therefore, given $|v_{32}\mu - \eta_{gd}| < \epsilon$ and $|v_{33}\mu + \eta_{gd}| < \epsilon$, we can get

$$
\begin{aligned}
\mathcal{L}_{\mathbf{\Sigma}}^{\text{test}} &= \frac{1}{2}\mathbb{E}\left[(f_{\text{LSA}}^{(L)}(\mathbf{Z})_{[-1,-1]} - y_q)^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[(v_{32}\mu\mathbf{y}^\top\mathbf{X}\sum_{k=0}^{L-1}(\mathbf{I} + v_{33}\mu\mathbf{X}^\top\mathbf{X})^k\mathbf{x}_q - y_q)^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left(v_{32}v_{33}^{-1}\boldsymbol{\beta}^\top[(\mathbf{I} + v_{33}\mu\mathbf{X}^\top\mathbf{X})^L - \mathbf{I}]\mathbf{x}_q + v_{32}\mu\boldsymbol{\xi}^\top\mathbf{X}\sum_{k=0}^{L-1}(\mathbf{I} + v_{33}\mu\mathbf{X}^\top\mathbf{X})^k\mathbf{x}_q - \boldsymbol{\beta}^\top\mathbf{x}_q\right)^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left(v_{32}v_{33}^{-1}\boldsymbol{\beta}^\top(\mathbf{I} + v_{33}\mu\mathbf{X}^\top\mathbf{X})^L\mathbf{x}_q - (v_{32}v_{33}^{-1} + 1)\boldsymbol{\beta}^\top\mathbf{x}_q\right)^2\right] \\
&\quad + \mathbb{E}\left[\left(v_{32}\mu\boldsymbol{\xi}^\top\mathbf{X}\sum_{k=0}^{L-1}(\mathbf{I} + v_{33}\mu\mathbf{X}^\top\mathbf{X})^k\mathbf{x}_q\right)^2\right] \\
&\le (v_{32}v_{33}^{-1})^2\mathbb{E}\left[(\boldsymbol{\beta}^\top(\mathbf{I} + v_{33}\mu\mathbf{X}^\top\mathbf{X})^L\mathbf{x}_q)^2\right] + (v_{32}v_{33}^{-1} + 1)^2\mathbb{E}\left[(\boldsymbol{\beta}^\top\mathbf{x}_q)^2\right] \\
&\quad + (v_{32}\mu)^2\mathbb{E}\left[\left(\boldsymbol{\xi}^\top\mathbf{X}\sum_{k=0}^{L-1}(\mathbf{I} + v_{33}\mu\mathbf{X}^\top\mathbf{X})^k\mathbf{x}_q\right)^2\right] \\
&\le (v_{32}v_{33}^{-1})^2\mathbb{E}\left[(\boldsymbol{\beta}^\top\mathbf{x}_q)^2\|\mathbf{I} + v_{33}\mu\mathbf{X}^\top\mathbf{X}\|_{op}^{2L}\right] + (v_{32}v_{33}^{-1} + 1)^2\mathbb{E}\left[(\boldsymbol{\beta}^\top\mathbf{x}_q)_2^2\right] \\
&\quad + (v_{32}\mu)^2\mathbb{E}\left[(\boldsymbol{\xi}^\top\mathbf{x}_q)^2\|\mathbf{X}\|_{op}^2\left\|\sum_{k=0}^{L-1}(\mathbf{I} + v_{33}\mu\mathbf{X}^\top\mathbf{X})^k\right\|_{op}^2\right] \\
&\overset{(i)}{\le} (v_{32}v_{33}^{-1})^2 d\left(1 + \frac{1}{2}v_{33}\mu\lambda_0 N\right)^{2L} + (v_{32}v_{33}^{-1} + 1)^2 d + \sigma^2 d\cdot\frac{4\lambda_1}{\lambda_0^2 N} + O(e^{-N}) \\
&\overset{(ii)}{\le} \left(\frac{\eta_{gd} + \epsilon}{\eta_{gd} - \epsilon}\right)^2 dr^{2L} + \left(\frac{2\epsilon}{\eta_{gd} - \epsilon}\right)^2 d + \sigma^2\cdot\frac{4\lambda_1 d}{\lambda_0^2 N} + O(e^{-N}) \\
&\le O\left(\frac{\sigma^2 d}{N} + d(r^L + \epsilon^2)\right),
\end{aligned}
$$

where (i) is by previous results, and (ii) is by $|v_{32}\mu - \eta_{gd}| < \epsilon$ and $|v_{33}\mu + \eta_{gd}| < \epsilon$. Here, $r = |1 + v_{33}\mu\lambda_0 N/2|$ such that $0 < r < 1$. □

C.4 Implicit CoT without Scratchpad for Query Prediction

In the training process, the input representation is designed as

$$\tilde{\mathbf{Z}} = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{y}^\top - \boldsymbol{\beta}_0^\top\mathbf{X}^\top & -\boldsymbol{\beta}_0^\top\mathbf{x}_q \end{bmatrix},$$

where $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I})$ is the initial weight vector. A single linear attention layer is

$$f_{\text{LSA}}(\mathbf{Z}) = \mathbf{Z} + \mathbf{W}^V\mathbf{Z}\cdot\mathbf{Mask}(\mathbf{Z})^\top\mathbf{W}^{KQ}\mathbf{Z},$$

where

$$\mathbf{W}^{KQ} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & w_{22} \end{bmatrix}, \mathbf{W}^V = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & v_{22} \end{bmatrix}.$$

Based on the teacher GD algorithm, the training loss is designed as

$$\mathcal{L} = \mathbb{E}\left[\frac{1}{2}(f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} + \boldsymbol{\beta}_1^\top\mathbf{x}_q)^2\right],$$

where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 - \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{y})$ is the ground truth GD update. Then, we can get

$$
\begin{aligned}
f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} &= -\boldsymbol{\beta}_0^\top\mathbf{x}_q + (\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top - v_{22}\boldsymbol{\beta}_0^\top\mathbf{X}^\top) \\
&\quad\cdot[\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q - \mathbf{X}\mathbf{W}_{12}\boldsymbol{\beta}_0^\top\mathbf{x}_q + \mathbf{y}\mathbf{W}_{21}\mathbf{x}_q - \mathbf{y}w_{22}\boldsymbol{\beta}_0^\top\mathbf{x}_q - \mathbf{X}\boldsymbol{\beta}_0\mathbf{W}_{21}\mathbf{x}_q + \mathbf{X}\boldsymbol{\beta}_0 w_{22}\boldsymbol{\beta}_0^\top\mathbf{x}_q].
\end{aligned}
$$

We assume that w_{22} is fixed at 0. The initialization of v_{22} is $v_{22}(0) \sim N(0,1)$, and all the elements in the other blocks are initialized as 0. We use the gradient flow method:

$$\frac{d\theta}{dt} = -\nabla \mathcal{L}(\theta)$$

for $\theta := (\mathbf{W}^{KQ}, \mathbf{W}^{V} \backslash \{w_{22}\})$. Then, we can find the form of $\mathbf{W}^{KQ}(t)$ and $\mathbf{W}^{V}(t)$ during the training in the following theorem.

Theorem C.19. *For all $t > 0$, $\mathbf{W}^{KQ}(t)$, $\mathbf{W}^{V}(t)$ have the following form:*

$$\mathbf{W}^{KQ}(t) = \begin{bmatrix} \mu(t) \cdot \mathbf{I} & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{bmatrix}, \mathbf{W}^{V}(t) = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & v_{22}(t) \end{bmatrix}.$$

Proof. Since the output is

$$\begin{aligned} f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} = &-\boldsymbol{\beta}_0^\top \mathbf{x}_q + (\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top - v_{22}\boldsymbol{\beta}_0^\top \mathbf{X}^\top) \\ &\cdot [\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q - \mathbf{X}\mathbf{W}_{12}\boldsymbol{\beta}_0^\top \mathbf{x}_q + \mathbf{y}\mathbf{W}_{21}\mathbf{x}_q - \mathbf{X}\boldsymbol{\beta}_0\mathbf{W}_{21}\mathbf{x}_q], \end{aligned}$$

we can observe that $\mathbf{V}_{11}$, $\mathbf{V}_{12}$ are irrelevant to the prediction, so they stay at 0 for all $t > 0$. Next, we prove that when $\mathbf{V}_{21}$, $\mathbf{W}_{12}$, $\mathbf{W}_{21}$ are 0, the gradients for these blocks are always 0. We first calculate those gradients as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{V}_{21}} &= \mathbb{E}\left[\left(f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} + \boldsymbol{\beta}_1^\top \mathbf{x}_q\right) \cdot \frac{\partial f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]}}{\partial \mathbf{V}_{21}}\right] \\ &= \mathbb{E}\left[(-\boldsymbol{\beta}_0^\top \mathbf{x}_q + (v_{22}\mathbf{y}^\top - v_{22}\boldsymbol{\beta}_0^\top \mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q + \boldsymbol{\beta}_1^\top \mathbf{x}_q) \cdot \mathbf{x}_q^\top \mathbf{W}_{11}^\top \mathbf{X}^\top \mathbf{X}\right] \\ &= \mathbb{E}\left[((v_{22}\mathbf{y}^\top - v_{22}\boldsymbol{\beta}_0^\top \mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}(\boldsymbol{\beta}_0^\top \mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q) \cdot \mathbf{x}_q^\top \mathbf{W}_{11}^\top \mathbf{X}^\top \mathbf{X}\right] \\ &= \mathbb{E}\left[(\mathbf{y}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q) - \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)) \cdot \mathbf{x}_q^\top \mathbf{W}_{11}^\top \mathbf{X}^\top \mathbf{X}\right] \\ &= \mathbb{E}\left[\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)\mathbf{x}_q^\top \mathbf{W}_{11}^\top \mathbf{X}^\top \mathbf{X} + \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)\mathbf{x}_q^\top \mathbf{W}_{11}^\top \mathbf{X}^\top \mathbf{X}\right], \end{aligned}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{W}_{12}} &= \mathbb{E}\left[\left(f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} + \boldsymbol{\beta}_1^\top \mathbf{x}_q\right) \cdot \frac{\partial f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]}}{\partial \mathbf{W}_{12}}\right] \\ &= \mathbb{E}\left[-\left(-\boldsymbol{\beta}_0^\top \mathbf{x}_q + (v_{22}\mathbf{y}^\top - v_{22}\boldsymbol{\beta}_0^\top \mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q + \boldsymbol{\beta}_1^\top \mathbf{x}_q\right) \cdot \mathbf{X}^\top (v_{22}\mathbf{y} - v_{22}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top \boldsymbol{\beta}_0\right] \\ &= \mathbb{E}\left[-\left((v_{22}\mathbf{y}^\top - v_{22}\boldsymbol{\beta}_0^\top \mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}(\boldsymbol{\beta}_0^\top \mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q\right) \cdot \mathbf{X}^\top (v_{22}\mathbf{y} - v_{22}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top \boldsymbol{\beta}_0\right] \\ &= \mathbb{E}\left[-\left(\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q) - \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)\right) \cdot \mathbf{X}^\top \mathbf{X}(v_{22}\boldsymbol{\beta} - v_{22}\boldsymbol{\beta}_0)\mathbf{x}_q^\top \boldsymbol{\beta}_0\right] \\ &= \mathbb{E}\left[-(\boldsymbol{\beta}^\top - \boldsymbol{\beta}_0^\top)\mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{22}\mathbf{X}^\top \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\mathbf{x}_q^\top \boldsymbol{\beta}_0\right], \end{aligned}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{W}_{21}} &= \mathbb{E}\left[\left(f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} + \boldsymbol{\beta}_1^\top \mathbf{x}_q\right) \cdot \frac{\partial f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]}}{\partial \mathbf{W}_{21}}\right] \\ &= \mathbb{E}\left[(-\boldsymbol{\beta}_0^\top \mathbf{x}_q + (v_{22}\mathbf{y}^\top - v_{22}\boldsymbol{\beta}_0^\top \mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q + \boldsymbol{\beta}_1^\top \mathbf{x}_q) \cdot (\mathbf{y}^\top - \boldsymbol{\beta}_0^\top \mathbf{X}^\top)(v_{22}\mathbf{y} - v_{22}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\ &= \mathbb{E}\left[((v_{22}\mathbf{y}^\top - v_{22}\boldsymbol{\beta}_0^\top \mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}(\boldsymbol{\beta}_0^\top \mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q) \cdot (\mathbf{y}^\top - \boldsymbol{\beta}_0^\top \mathbf{X}^\top)(v_{22}\mathbf{y} - v_{22}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\ &= \mathbb{E}\left[\left(\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q) - \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)\right)(\boldsymbol{\beta}^\top - \boldsymbol{\beta}_0^\top)\mathbf{X}^\top \mathbf{X}(v_{22}\boldsymbol{\beta} - v_{22}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\ &= \mathbb{E}\left[(\boldsymbol{\beta}^\top - \boldsymbol{\beta}_0^\top)\mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{22}(\boldsymbol{\beta}^\top - \boldsymbol{\beta}_0^\top)\mathbf{X}^\top \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\mathbf{x}_q^\top\right]. \end{aligned}$$

We can find that each term in the gradients for $\mathbf{V}_{21}$, $\mathbf{W}_{12}$, $\mathbf{W}_{21}$ is an odd function of $\boldsymbol{\beta}$ or $\boldsymbol{\beta}_0$. Since $\boldsymbol{\beta} \sim N(0, \mathbf{I})$ and $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I})$, the expectation of these four gradients should be 0. Therefore, we know that $\mathbf{V}_{21}$, $\mathbf{W}_{12}$, $\mathbf{W}_{21}$ always stay at 0. Then, we can get

$$f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} = -\boldsymbol{\beta}_0^\top \mathbf{x}_q + v_{22}(\mathbf{y}^\top - \boldsymbol{\beta}_0^\top \mathbf{X}^\top) \cdot \mathbf{X}\mathbf{W}_{11}\mathbf{x}_q.$$

Next, we prove that $\mathbf{W}_{11}(t) \propto \mathbf{I}$ for all $t > 0$. We can calculate the gradient as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{W}_{11}} &= \mathbb{E}\left[\left(f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} + \boldsymbol{\beta}_1^\top \mathbf{x}_q\right) \cdot \frac{\partial f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]}}{\partial \mathbf{W}_{11}}\right] \\ &= \mathbb{E}\left[(-\boldsymbol{\beta}_0^\top \mathbf{x}_q + (v_{22}\mathbf{y}^\top - v_{22}\boldsymbol{\beta}_0^\top \mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q + \boldsymbol{\beta}_1^\top \mathbf{x}_q) \cdot \mathbf{X}^\top (v_{22}\mathbf{y} - v_{22}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\ &= \mathbb{E}\left[((v_{22}\mathbf{y}^\top - v_{22}\boldsymbol{\beta}_0^\top \mathbf{X}^\top)\mathbf{X}\mathbf{W}_{11}\mathbf{x}_q - \eta_{gd}(\boldsymbol{\beta}_0^\top \mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q) \cdot \mathbf{X}^\top (v_{22}\mathbf{y} - v_{22}\mathbf{X}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q) - \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)\right) \cdot \mathbf{X}^\top \mathbf{X}(v_{22}\boldsymbol{\beta} - v_{22}\boldsymbol{\beta}_0)\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{22}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\mathbf{x}_q^\top - \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{22}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}_0\mathbf{x}_q^\top\right] \\
&\quad - \mathbb{E}\left[\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{22}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\mathbf{x}_q^\top - \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{22}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}_0\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{22}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\mathbf{x}_q^\top + \boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}(v_{22}\mathbf{W}_{11}\mathbf{x}_q + \eta_{gd}\mathbf{x}_q)v_{22}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}_0\mathbf{x}_q^\top\right] \\
&= \mathbb{E}\left[2v_{22}(\mathbf{X}^\top \mathbf{X})^2(v_{22}\mathbf{W}_{11} + \eta_{gd}\mathbf{I})\right] \\
&= \mathbb{E}\left[2v_{22}N(N+d+1)(v_{22}\mathbf{W}_{11} + \eta_{gd}\mathbf{I})\right],
\end{aligned}
$$

where the last equation is by Lemma D.1. We can find that when $\mathbf{W}_{11}$ is proportional to $\mathbf{I}$, the expectation of the gradient $\mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{11}}\right]$ is also proportional to $\mathbf{I}$. Therefore, $\mathbf{W}_{11}(t)$ can be written as $\mathbf{W}_{11}(t) = \mu(t) \cdot \mathbf{I}$.

□

Now, we can only consider two parameters μ and v_{22}. The following lemma shows their gradients.

Lemma C.20. *The gradients for μ, v_{22} are*

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu} &= 2c_1 v_{22}(v_{22}\mu + \eta_{gd}), \\
\frac{\partial \mathcal{L}}{\partial v_{22}} &= 2c_1 \mu(v_{22}\mu + \eta_{gd}),
\end{aligned}
$$

where $c_1 = dN(d+N+1)$.

Proof. Recall that

$$
\begin{aligned}
f_{\text{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} + \boldsymbol{\beta}_1^\top \mathbf{x}_q &= -\boldsymbol{\beta}_0^\top \mathbf{x}_q + (v_{22}\mathbf{y}^\top - v_{22}\boldsymbol{\beta}_0^\top \mathbf{X}^\top) \cdot \mu \mathbf{X}\mathbf{x}_q + \boldsymbol{\beta}_0^\top \mathbf{x}_q - \eta_{gd}(\boldsymbol{\beta}_0^\top \mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q \\
&= (v_{22}\mu + \eta_{gd})(\boldsymbol{\beta}^\top - \boldsymbol{\beta}_0^\top)\mathbf{X}^\top \mathbf{X}\mathbf{x}_q.
\end{aligned}
$$

By Lemma D.1, we can calculate the corresponding gradients.

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu} &= \mathbb{E}\left[\left((v_{22}\mu + \eta_{gd})(\boldsymbol{\beta}^\top - \boldsymbol{\beta}_0^\top)\mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right) \cdot \left(v_{22}(\boldsymbol{\beta}^\top - \boldsymbol{\beta}_0^\top)\mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right)\right] \\
&= \mathbb{E}\left[v_{22}(v_{22}\mu + \eta_{gd}) \cdot \left((\boldsymbol{\beta}^\top - \boldsymbol{\beta}_0^\top)\mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right)^2\right] \\
&= 2v_{22}(v_{22}\mu + \eta_{gd}) \cdot \mathbb{E}\left[\operatorname{tr}((\mathbf{X}^\top \mathbf{X})^2)\right] \\
&= 2dN(d+N+1)v_{22}(v_{22}\mu + \eta_{gd}), \\
\frac{\partial \mathcal{L}}{\partial v_{22}} &= \mathbb{E}\left[\left((v_{22}\mu + \eta_{gd})(\boldsymbol{\beta}^\top - \boldsymbol{\beta}_0^\top)\mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right) \cdot \left(\mu(\boldsymbol{\beta}^\top - \boldsymbol{\beta}_0^\top)\mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right)\right] \\
&= \mathbb{E}\left[\mu(v_{22}\mu + \eta_{gd}) \cdot \left((\boldsymbol{\beta}^\top - \boldsymbol{\beta}_0^\top)\mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right)^2\right] \\
&= 2\mu(v_{22}\mu + \eta_{gd}) \cdot \mathbb{E}\left[\operatorname{tr}((\mathbf{X}^\top \mathbf{X})^2)\right] \\
&= 2dN(d+N+1)\mu(v_{22}\mu + \eta_{gd}).
\end{aligned}
$$

□

Next, we can derive the following convergence results.

Lemma C.21. *$v_{22}\mu$ will converge to $-\eta_{gd}$. Furthermore, there exists $T_1 = \Theta(c_1^{-1}\eta_{gd}^{-1})$ such that for all $t > T_1$, it holds that $|v_{22}(t)\mu(t) + \eta_{gd}| = O(e^{-c_1 t})$, where $c_1 = dN(d+N+1)$.*

Proof. Denote $A = v_{22}\mu + \eta_{gd}$. Then, we have

$$
\begin{aligned}
\frac{dA^2}{dt} &= 2A\frac{dA}{dt} \\
&= -2c_1(\mu^2 + v_{22}^2)A^2 \\
&\leq 0.
\end{aligned}
$$

Thus, A^2 is non-increasing, and $0 \le A(t)^2 \le A(0)^2 = \eta_{gd}^2$, so A^2 will finally converge to a constant C such that $0 \le C \le \eta_{gd}^2$. When $t \to \infty$, we have $A^2 \to C$ and $\frac{dA^2}{dt} \to 0$. Then, we get

$$C(\mu^2 + v_{22}^2) \to 0.$$

Assume, for contradiction, that $C > 0$. Then, we have $\mu \to 0$ and $v_{22} \to 0$, so $A^2 \to \eta_{gd}^2$. However,

$$\left.\frac{dA^2}{dt}\right|_{t=0} = -2c_1\eta_{gd}^2(\mu(0)^2 + v_{22}(0)^2) < 0.$$

It means that A^2 is rigorously decreasing at the initial point, so the final limit of A^2 cannot be the initial value $A(0)^2 = \eta_{gd}^2$. It contradicts the fact that $C = \eta_{gd}^2$. Thus, we have proven that $C = 0$, so $A = v_{22}\mu + \eta_{gd} \to 0$.

Next, we show the convergence rate. Under gradient flow, we have

$$\frac{d\mu}{dt} = -2c_1 v_{22} A, \qquad \frac{dv_{22}}{dt} = -2c_1\mu A.$$

Therefore,

$$\frac{d}{dt}\left(\mu^2 - v_{22}^2\right) = 2\mu\frac{d\mu}{dt} - 2v_{22}\frac{dv_{22}}{dt} = 0.$$

Hence, $\mu^2(t) - v_{22}^2(t) = \mu^2(0) - v_{22}^2(0)$. Then, we have $\mu^2(t) + v_{22}^2(t) = 2\mu^2(t) + v_{22}^2(0) \ge v_{22}^2(0)$. Thus,

$$\frac{dA^2}{dt} = -2c_1(\mu^2 + v_{22}^2)A^2 \le -2c_1 v_{22}^2(0)A^2.$$

Therefore,

$$A(t)^2 \le A(0)^2 \exp\{-2c_1 v_{22}^2(0)t\}.$$

Since $v_{22}(0) = \Theta(1)$ with high probability under the random initialization, we have

$$|v_{22}(t)\mu(t) + \eta_{\mathrm{gd}}| = O(e^{-c_1 t}).$$

□

Theorem C.22. *For any $\epsilon > 0$, there exists $T = \Theta(\log(\epsilon^{-1}N^2d^2))$ such that for all $t > T$, it holds that $|v_{22}(t)\mu(t) + \eta_{gd}| < \epsilon$ and $\mathcal{L}(\theta(t)) < \epsilon$.*

Proof. We have already obtained $|v_{22}(t)\mu(t) + \eta_{gd}| = O(e^{-c_1 t}) < \epsilon$. Then, the loss is

$$\begin{aligned}
\mathcal{L}(\theta(t)) &= \mathbb{E}\left[(f_{\mathrm{LSA}}(\tilde{\mathbf{Z}})_{[-1,-1]} + \boldsymbol{\beta}_1^\top \mathbf{x}_q)^2\right] \\
&= \mathbb{E}\left[\left[(v_{22}(t)\mu(t) + \eta_{gd})\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q - (v_{22}(t)\mu(t) + \eta_{gd})\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{X}\mathbf{x}_q\right]^2\right] \\
&\le O(c_1 e^{-2c_1 t}) \\
&< \epsilon.
\end{aligned}$$

□

Given the trained one-layer model, we unroll it L times to get L-layer model. We consider the out-of-distribution generalization. Assume that $\mathbf{x}_i \sim N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is not necessarily the identity matrix. The test error is defined as

$$\mathcal{L}_{\boldsymbol{\Sigma}}^{\text{test}} = \frac{1}{2}\mathbb{E}\left[(f_{\mathrm{LSA}}^{(L)}(\mathbf{Z})_{[-1,-1]} + y_q)^2\right].$$

Then, we can get the following result regarding this OOD loss.

Theorem C.23. *Assume that $\eta_{gd} = O(N^{-1}d^{-1})$, and the out-of-distribution input data $\mathbf{x}_i \sim N(0, \boldsymbol{\Sigma})$ where $\lambda_{\min}(\boldsymbol{\Sigma}), \lambda_{\max}(\boldsymbol{\Sigma}) = \Theta(1)$. It holds that $\mathcal{L}_{\boldsymbol{\Sigma}}^{test} \le O(\sigma^2 dN^{-1} + dr^L)$ where $0 < r < 1$ is a constant depending on $\eta_{gd}, \lambda_{\min}, N$, and ϵ is the bound considered for the training loss.*

Proof. Denote that $\lambda_0 = \lambda_{\min}(\boldsymbol{\Sigma})$ and $\lambda_1 = \lambda_{\max}(\boldsymbol{\Sigma})$. We can obtain that

$$\mathbb{P}\left(\lambda_{\min}(\mathbf{X}^\top\mathbf{X}) \le \frac{1}{2}\lambda_0 N\right) \le e^{-c_0\lambda_0^2 N}.$$

Then, we have

$$\mathbb{E}\left[\|\mathbf{I}+v_{22}\mu\mathbf{X}^\top\mathbf{X}\|_{op}^L\right]\le\mathbb{E}\left[(1+v_{22}\mu\lambda_{\min}(\mathbf{X}^\top\mathbf{X}))^L\right]\le\left(1+\frac{1}{2}v_{22}\mu\lambda_0 N\right)^L+O(e^{-N}),$$

and

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{k=0}^{L-1}(\mathbf{I}+v_{22}\mu\mathbf{X}^\top\mathbf{X})^k\right\|_{op}\right]&=\mathbb{E}\left[\sum_{k=0}^{L-1}(1+v_{22}\mu\lambda_{\min}(\mathbf{X}^\top\mathbf{X}))^k\right]\\&\le\sum_{k=0}^{L-1}\left(1+\frac{1}{2}v_{22}\mu\lambda_0 N\right)^k+O(e^{-N})\\&=\frac{1-(1+v_{22}\mu\lambda_0 N/2)^L}{-v_{22}\mu\lambda_0 N/2}+O(e^{-N}).\end{aligned}$$

Therefore, given $|v_{22}\mu+\eta_{gd}|<\epsilon$, we can get

$$\begin{aligned}\mathcal{L}_{\mathbf{\Sigma}}^{\text{test}}&=\frac{1}{2}\mathbb{E}\left[(f_{\text{LSA}}^{(L)}(\mathbf{Z})_{[-1,-1]}+y_q)^2\right]\\&=\frac{1}{2}\mathbb{E}\left[(v_{22}\mu\mathbf{y}^\top\mathbf{X}\sum_{k=0}^{L-1}(\mathbf{I}+v_{22}\mu\mathbf{X}^\top\mathbf{X})^k\mathbf{x}_q+y_q)^2\right]\\&=\frac{1}{2}\mathbb{E}\left[\left(\boldsymbol{\beta}^\top[(\mathbf{I}+v_{22}\mu\mathbf{X}^\top\mathbf{X})^L-\mathbf{I}]\mathbf{x}_q+v_{22}\mu\boldsymbol{\xi}^\top\mathbf{X}\sum_{k=0}^{L-1}(\mathbf{I}+v_{22}\mu\mathbf{X}^\top\mathbf{X})^k\mathbf{x}_q+\boldsymbol{\beta}^\top\mathbf{x}_q\right)^2\right]\\&=\mathbb{E}\left[\left(\boldsymbol{\beta}^\top(\mathbf{I}+v_{22}\mu\mathbf{X}^\top\mathbf{X})^L\mathbf{x}_q\right)^2\right]+(v_{22}\mu)^2\mathbb{E}\left[\left(\boldsymbol{\xi}^\top\mathbf{X}\sum_{k=0}^{L-1}(\mathbf{I}+v_{22}\mu\mathbf{X}^\top\mathbf{X})^k\mathbf{x}_q\right)^2\right]\\&\le\mathbb{E}\left[(\boldsymbol{\beta}^\top\mathbf{x}_q)^2\|\mathbf{I}+v_{22}\mu\mathbf{X}^\top\mathbf{X}\|_{op}^{2L}\right]+(v_{22}\mu)^2\mathbb{E}\left[(\boldsymbol{\xi}^\top\mathbf{x}_q)^2\|\mathbf{X}\|_{op}^2\left\|\sum_{k=0}^{L-1}(\mathbf{I}+v_{22}\mu\mathbf{X}^\top\mathbf{X})^k\right\|_{op}^2\right]\\&\overset{(i)}{\le}d\left(1+\frac{1}{2}v_{22}\mu\lambda_0 N\right)^{2L}+\sigma^2 d\cdot\frac{4\lambda_1}{\lambda_0^2 N}+O(e^{-N})\\&\overset{(ii)}{\le}\left(\frac{\eta_{gd}+\epsilon}{\eta_{gd}-\epsilon}\right)^2 dr^{2L}+\sigma^2\cdot\frac{4\lambda_1 d}{\lambda_0^2 N}+O(e^{-N})\\&\le O\left(\frac{\sigma^2 d}{N}+dr^L\right),\end{aligned}$$

where (i) is by previous results, and (ii) is by $|v_{22}\mu+\eta_{gd}|<\epsilon$. Here, $r=|1+v_{22}\mu\lambda_0 N/2|$ such that $0<r<1$. □

C.5 Explicit CoT with Scratchpad for Query Prediction

For explicit CoT, the generated intermediate sequence is expected to simulate

$$\mathbf{Z}_i=\begin{bmatrix}\mathbf{X}^\top & \mathbf{x}_q & \mathbf{x}_q & \cdots & \mathbf{x}_q\\ \mathbf{y}^\top & 0 & 0 & \cdots & 0\\ 0 & \mathbf{u}_0 & \mathbf{u}_1 & \cdots & \mathbf{u}_i\\ 0 & 0 & \boldsymbol{\beta}_1^\top\mathbf{x}_q & \cdots & \boldsymbol{\beta}_i^\top\mathbf{x}_q\end{bmatrix}\in\mathbb{R}^{(2d+2)\times(N+i+1)},i\in[L],$$

where $\mathbf{u}_i=(\mathbf{I}_d-\eta_{gd}\mathbf{X}^\top\mathbf{X})^i\mathbf{x}_q$. Here, $\boldsymbol{\beta}_i$ is GD steps from 0 initialization with learning rate η_{gd}. We have

$$\boldsymbol{\beta}_L=\eta_{gd}\sum_{i=1}^{L-1}(\mathbf{I}_d-\eta_{gd}\mathbf{X}^\top\mathbf{X})^i\mathbf{X}^\top\mathbf{y}.$$

A single linear attention layer is

$$f_{\text{LSA}}(\mathbf{Z})=\mathbf{Z}+\mathbf{W}^V\mathbf{Z}\cdot\mathbf{Mask}(\mathbf{Z})^\top\mathbf{W}^{KQ}\mathbf{Z},$$

where

$$\mathbf{W}^{KQ} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} & \mathbf{W}_{13} & \mathbf{W}_{14} \\ \mathbf{W}_{21} & w_{22} & \mathbf{W}_{23} & w_{24} \\ \mathbf{W}_{31} & \mathbf{W}_{32} & \mathbf{W}_{33} & \mathbf{W}_{34} \\ \mathbf{W}_{41} & w_{42} & \mathbf{W}_{43} & w_{44} \end{bmatrix}, \mathbf{W}^{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} & \mathbf{V}_{14} \\ \mathbf{V}_{21} & v_{22} & \mathbf{V}_{23} & v_{24} \\ \mathbf{V}_{31} & \mathbf{V}_{32} & \mathbf{V}_{33} & \mathbf{V}_{34} \\ \mathbf{V}_{41} & v_{42} & \mathbf{V}_{43} & v_{44} \end{bmatrix}.$$

Based on the teacher GD algorithm, the training loss is designed as

$$\mathcal{L} = \mathbb{E}\left[\frac{1}{2}\sum_{i=0}^{L-1}\left\|f_{\text{LSA}}(\mathbf{Z}_i)_{[:,-1]} - (\mathbf{x}_q, 0, \mathbf{u}_{i+1}, \boldsymbol{\beta}_{i+1}^\top \mathbf{x}_q)\right\|_2^2\right],$$

where $\boldsymbol{\beta}_{i+1} = \boldsymbol{\beta}_i - \eta_{gd}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}_i - \mathbf{y})$ is the ground truth GD update. We also denote that

$$\mathcal{L}_i = \frac{1}{2}\mathbb{E}\left[\left\|f_{\text{LSA}}(\mathbf{Z}_i)_{[:,-1]} - (\mathbf{x}_q, 0, \mathbf{u}_{i+1}, \boldsymbol{\beta}_{i+1}^\top \mathbf{x}_q)\right\|_2^2\right]$$

For the training, we consider the gradient flow method

$$\frac{d\theta}{dt} = -\nabla\mathcal{L}(\theta)$$

for $\theta := (\mathbf{W}^{KQ}, \mathbf{W}^{V})$. We assume that the initialization of v_{42} is $v_{42}(0) = 1$, and all the elements in the other blocks are initialized as 0. Then, we can find the form of $\mathbf{W}^{KQ}(t)$ and $\mathbf{W}^{V}(t)$ during the training in the following theorem.

Theorem C.24. *For all $t > 0$, $\mathbf{W}^{KQ}(t)$, $\mathbf{W}^{V}(t)$ have the following form:*

$$\mathbf{W}^{KQ}(t) = \begin{bmatrix} 0 & 0 & w_{13}(t)\cdot\mathbf{I}_d & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \mathbf{W}^{V}(t) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ v_{31}(t)\cdot\mathbf{I}_d & 0 & 0 & 0 \\ 0 & v_{42}(t) & 0 & 0 \end{bmatrix}.$$

Proof. We first prove that the blocks except $\mathbf{W}_{31}$, $\mathbf{V}_{31}$, and v_{42} will remain 0. The prediction of each step is

$$f_{\text{LSA}}(\mathbf{Z}_i)_{[:,-1]} = \begin{bmatrix} \mathbf{x}_q + (\mathbf{V}_{11}\mathbf{X}^\top + \mathbf{V}_{12}\mathbf{y}^\top)\cdot\mathbf{S} \\ (\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top)\cdot\mathbf{S} \\ \mathbf{u}_i + (\mathbf{V}_{31}\mathbf{X}^\top + \mathbf{V}_{32}\mathbf{y}^\top)\cdot\mathbf{S} \\ \boldsymbol{\beta}_i^\top\mathbf{x}_q + (\mathbf{V}_{41}\mathbf{X}^\top + v_{42}\mathbf{y}^\top)\cdot\mathbf{S} \end{bmatrix},$$

where $\mathbf{S} = \mathbf{X}(\mathbf{W}_{11}\mathbf{x}_q + \mathbf{W}_{13}\mathbf{u}_i + \mathbf{W}_{14}\boldsymbol{\beta}_i^\top\mathbf{x}_q) + \mathbf{y}^\top(\mathbf{W}_{21}\mathbf{x}_q + \mathbf{W}_{23}\mathbf{u}_i + w_{24}\boldsymbol{\beta}_i^\top\mathbf{x}_q)$. Thus, the loss of the i-th step is

$$\begin{aligned}\mathcal{L}_i = &\frac{1}{2}\left\|(\mathbf{V}_{11}\mathbf{X}^\top + \mathbf{V}_{12}\mathbf{y}^\top)\cdot\mathbf{S}\right\|_2^2 + \frac{1}{2}\left((\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top)\cdot\mathbf{S}\right)^2 \\ &+ \frac{1}{2}\left\|(\mathbf{V}_{31}\mathbf{X}^\top + \mathbf{V}_{32}\mathbf{y}^\top)\cdot\mathbf{S} + \eta_{gd}\mathbf{X}^\top\mathbf{X}\mathbf{u}_i\right\|_2^2 + \frac{1}{2}\left((\mathbf{V}_{41}\mathbf{X}^\top + v_{42}\mathbf{y}^\top)\cdot\mathbf{S} - \eta_{gd}\mathbf{y}^\top\mathbf{X}\mathbf{u}_i\right)^2\end{aligned}$$

We can observe that the blocks $\mathbf{V}_{13}$, $\mathbf{V}_{14}$, $\mathbf{V}_{23}$, v_{24}, $\mathbf{V}_{33}$, $\mathbf{V}_{34}$, $\mathbf{V}_{43}$, v_{44}, $\mathbf{W}_{12}$, w_{22}, $\mathbf{W}_{31}$, $\mathbf{W}_{32}$, $\mathbf{W}_{33}$, $\mathbf{W}_{34}$, $\mathbf{W}_{41}$, w_{42}, $\mathbf{W}_{43}$, w_{44} are irrelevant to the loss function, so these blocks will stay at 0. Next, we prove that the blocks $\mathbf{V}_{11}$, $\mathbf{V}_{12}$, $\mathbf{V}_{21}$, v_{22}, $\mathbf{V}_{32}$, $\mathbf{V}_{41}$, $\mathbf{W}_{11}$, $\mathbf{W}_{14}$, $\mathbf{W}_{21}$, $\mathbf{W}_{23}$, w_{24} will also remain 0 for $t > 0$. When $\mathbf{V}_{11}$, $\mathbf{V}_{12}$, $\mathbf{V}_{21}$, v_{22}, $\mathbf{V}_{32}$, $\mathbf{V}_{41}$, $\mathbf{W}_{11}$, $\mathbf{W}_{14}$, $\mathbf{W}_{21}$, $\mathbf{W}_{23}$, w_{24} are 0, we can calculate their gradients as follows (we use Δ_{ij} to represent some irrelevant terms):

$$\frac{\partial\mathcal{L}_i}{\partial\mathbf{V}_{11}} = \mathbb{E}\left[(\mathbf{V}_{11}\mathbf{X}^\top + \mathbf{V}_{12}\mathbf{y}^\top)\cdot\mathbf{S}\cdot\Delta_{11}\right] = 0,$$

$$\frac{\partial\mathcal{L}_i}{\partial\mathbf{V}_{12}} = \mathbb{E}\left[(\mathbf{V}_{11}\mathbf{X}^\top + \mathbf{V}_{12}\mathbf{y}^\top)\cdot\mathbf{S}\cdot\Delta_{12}\right] = 0,$$

$$\frac{\partial\mathcal{L}_i}{\partial\mathbf{V}_{21}} = \mathbb{E}\left[(\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top)\cdot\mathbf{S}\cdot\Delta_{21}\right] = 0,$$

$$\frac{\partial\mathcal{L}_i}{\partial v_{22}} = \mathbb{E}\left[(\mathbf{V}_{21}\mathbf{X}^\top + v_{22}\mathbf{y}^\top)\cdot\mathbf{S}\cdot\Delta_{22}\right] = 0,$$

$$\frac{\partial \mathcal{L}_i}{\partial \mathbf{V}_{32}} = \mathbb{E}\left[((\mathbf{V}_{31}\mathbf{X}^\top + \mathbf{V}_{32}\mathbf{y}^\top)\cdot \mathbf{S} + \eta_{gd}\mathbf{X}^\top\mathbf{X}\mathbf{u}_i)\cdot \mathbf{y}^\top\mathbf{S}\right]$$
$$= \mathbb{E}\left[(\mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i + \eta_{gd}\mathbf{X}^\top\mathbf{X}\mathbf{u}_i)\cdot \mathbf{y}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i\right],$$

$$\frac{\partial \mathcal{L}_i}{\partial \mathbf{V}_{41}} = \mathbb{E}\left[((\mathbf{V}_{41}\mathbf{X}^\top + v_{42}\mathbf{y}^\top)\cdot \mathbf{S} - \eta_{gd}\mathbf{y}^\top\mathbf{X}\mathbf{u}_i)\cdot \mathbf{X}^\top\mathbf{S}\right]$$
$$= \mathbb{E}\left[(v_{42}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i - \eta_{gd}\mathbf{y}^\top\mathbf{X}\mathbf{u}_i)\cdot \mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i\right],$$

$$\frac{\partial \mathcal{L}_i}{\partial \mathbf{W}_{11}} = \mathbb{E}\left[(\mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i + \eta_{gd}\mathbf{X}^\top\mathbf{X}\mathbf{u}_i)\mathbf{y}\mathbf{x}_q^\top + (v_{42}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i - \eta_{gd}\mathbf{y}^\top\mathbf{X}\mathbf{u}_i)\boldsymbol{\beta}_i\mathbf{X}\mathbf{x}_q^\top\right],$$

$$\frac{\partial \mathcal{L}_i}{\partial \mathbf{W}_{14}} = \mathbb{E}\left[(\mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i + \eta_{gd}\mathbf{X}^\top\mathbf{X}\mathbf{u}_i)\mathbf{X}^\top\mathbf{X}\mathbf{x}_q^\top\boldsymbol{\beta}_i + (v_{42}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i - \eta_{gd}\mathbf{y}^\top\mathbf{X}\mathbf{u}_i)\mathbf{X}^\top\mathbf{y}v_{42}\mathbf{x}_q^\top\boldsymbol{\beta}_i\right],$$

$$\frac{\partial \mathcal{L}_i}{\partial \mathbf{W}_{21}} = \mathbb{E}\left[(\mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i + \eta_{gd}\mathbf{X}^\top\mathbf{X}\mathbf{u}_i)\mathbf{y}\mathbf{x}_q^\top + (v_{42}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i - \eta_{gd}\mathbf{y}^\top\mathbf{X}\mathbf{u}_i)\mathbf{y}^\top\mathbf{y}\mathbf{x}_q^\top\right],$$

$$\frac{\partial \mathcal{L}_i}{\partial \mathbf{W}_{23}} = \mathbb{E}\left[(\mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i + \eta_{gd}\mathbf{X}^\top\mathbf{X}\mathbf{u}_i)\mathbf{y}\mathbf{u}_i^\top + (v_{42}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i - \eta_{gd}\mathbf{y}^\top\mathbf{X}\mathbf{u}_i)\mathbf{y}^\top\mathbf{y}\mathbf{u}_i^\top\right],$$

$$\frac{\partial \mathcal{L}_i}{\partial w_{24}} = \mathbb{E}\left[(\mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i + \eta_{gd}\mathbf{X}^\top\mathbf{X}\mathbf{u}_i)\mathbf{y}\boldsymbol{\beta}_i^\top\mathbf{x}_q + (v_{42}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i - \eta_{gd}\mathbf{y}^\top\mathbf{X}\mathbf{u}_i)\mathbf{y}^\top\mathbf{y}\boldsymbol{\beta}_i^\top\mathbf{x}_q\right],$$

We can find that each term in these three gradients is an odd function of $\mathbf{y}$ or $\boldsymbol{\beta}_i$. Thus, the expectation of these gradients should be 0. Then, we can get

$$f_{\text{LSA}}(\mathbf{Z}_i)_{[:,-1]} = \begin{bmatrix} \mathbf{x}_q \\ 0 \\ \mathbf{u}_i + \mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i \\ \boldsymbol{\beta}_i^\top\mathbf{x}_q + v_{42}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i \end{bmatrix}.$$

Next, we prove that $\mathbf{V}_{31}(t) \propto \mathbf{I}$ and $\mathbf{W}_{13}(t) \propto \mathbf{I}$. When $\mathbf{V}_{31} \propto \mathbf{I}$ and $\mathbf{W}_{13} \propto \mathbf{I}$, their gradients are

$$\frac{\partial \mathcal{L}_i}{\partial \mathbf{V}_{31}} = \mathbb{E}\left[(\mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i + \eta_{gd}\mathbf{X}^\top\mathbf{X}\mathbf{u}_i)\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i\right],$$

$$\frac{\partial \mathcal{L}_i}{\partial \mathbf{W}_{13}} = \mathbb{E}\left[(\mathbf{V}_{31}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i + \eta_{gd}\mathbf{X}^\top\mathbf{X}\mathbf{u}_i)\mathbf{u}_i^\top\mathbf{X}^\top\mathbf{X}\mathbf{V}_{31} + (v_{42}\mathbf{y}^\top\mathbf{X}\mathbf{W}_{13}\mathbf{u}_i - \eta_{gd}\mathbf{y}^\top\mathbf{X}\mathbf{u}_i)\mathbf{u}_i^\top\mathbf{X}^\top\mathbf{y}v_{42}\right].$$

We can find that when $\mathbf{V}_{31}$ and $\mathbf{W}_{13}$ are proportional to $\mathbf{I}$, the expectation of their gradient is also proportional to $\mathbf{I}$. Therefore, with zero initialization, $\mathbf{V}_{31}(t)$ and $\mathbf{W}_{13}(t)$ can be written as $\mathbf{V}_{31}(t) = v_{31}(t)\cdot\mathbf{I}$ and $\mathbf{W}_{13}(t) = w_{13}(t)\cdot\mathbf{I}$. □

Now, we can only consider three parameters w_{13}, v_{31}, and v_{42}. The following lemma shows their gradients.

Lemma C.25. *The gradients for w_{13}, v_{31}, v_{42} are*

$$\frac{\partial \mathcal{L}_i}{\partial w_{13}} = c_{1i}\cdot v_{31}(v_{31}w_{13} + \eta_{gd}) + c_{2i}\cdot v_{42}(v_{42}w_{13} - \eta_{gd}),$$
$$\frac{\partial \mathcal{L}_i}{\partial v_{31}} = c_{1i}\cdot w_{13}(v_{31}w_{13} + \eta_{gd}),$$
$$\frac{\partial \mathcal{L}_i}{\partial v_{42}} = c_{2i}\cdot w_{13}(v_{42}w_{13} - \eta_{gd}),$$

where $c_{1i} = N(N+d+1)\mathbb{E}(\text{tr}(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top))$, and $c_{2i} = Nd(N+d+1)$.

The proof is entirely the same as Lemma C.8, so we omit it here. Here, we find that the form of the gradients of w_{13}, v_{31}, v_{42} is the same as that of v_{11}, w_{11}, w_{24} in Lemma C.8. Thus, the remaining analysis is highly similar to Section C.2, so we just state the following results and ignore the detailed proof.

We denote that $c_1 = \sum_{i=0}^{L-1} c_{i1}$ and $c_2 = \sum_{i=0}^{L-1} c_{i2}$. In order to analyze the convergence of three parameters, we first provide their boundedness property.

Lemma C.26. *For all $t > 0$, $w_{13}(t)$, $v_{31}(t)$ and $v_{42}(t)$ are bounded, i.e. there exists a positive constant M such that $|w_{13}(t)| \leq M$, $|v_{31}(t)| \leq M$, $|v_{42}(t)| \leq M$.*

Next, we derive the following convergence results.

Lemma C.27. *$w_{13}v_{31}$ and $w_{13}v_{42}$ will converge to $-\eta_{gd}$ and η_{gd}, respectively. Furthermore, there exists $T_1 = \Theta(c_2^{-1})$ such that for all $t > T_1$, it holds that $|w_{13}(t)v_{31}(t) + \eta_{gd}| = O(e^{-t})$ and $|w_{13}(t)v_{42}(t) - \eta_{gd}| = O(e^{-t})$.*

Theorem C.28. *For any $\epsilon > 0$, there exists $T = \Theta(\log(\epsilon^{-1}LN^2d^2))$ such that for all $t > T$, it holds that $|w_{13}(t)v_{31}(t) + \eta_{gd}| < \epsilon$, $|w_{13}(t)v_{42}(t) - \eta_{gd}| < \epsilon$, and $\mathcal{L}(\theta(t)) < \epsilon$.*

In addition, we can get the following result regarding this OOD loss.

Theorem C.29. *Assume that $\eta_{gd} = O(N^{-1}d^{-1})$, and the out-of-distribution input data $\mathbf{x}_i \sim N(0, \mathbf{\Sigma})$ where $\lambda_{\min}(\mathbf{\Sigma}), \lambda_{\max}(\mathbf{\Sigma}) = \Theta(1)$. It holds that $\mathcal{L}_{\mathbf{\Sigma}}^{test} \leq O(\sigma^2 dN^{-1} + d(r^L + \epsilon^2))$ where $0 < r < 1$ is a constant depending on $\eta_{gd}, \lambda_{\min}(\mathbf{\Sigma}), N$, and ϵ is the bound considered for the training loss.*

D Supplementary Lemmas

Lemma D.1. *Assume $\mathbf{x}_i \overset{i.i.d.}{\sim} N(0, \mathbf{I}_d)$ for $1 \leq i \leq N$. Then,*

$$\mathbb{E}\left[\left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^\top\right)^2\right] = N(N+d+1)\mathbf{I}_d,$$

$$\mathbb{E}\left[\left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^\top\right)^3\right] = N(N^2+3N(d+1)+d^2+3d+4)\mathbf{I}_d.$$

Proof. Denote that $S = \sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^\top$. We first compute $\mathbb{E}[S^2]$. For any $p, q \in \{1, ..., d\}$, we have

$$(S^2)_{pq} = \sum_{i,j=1}^{N} (\mathbf{x}_i^\top \mathbf{x}_j) \cdot (\mathbf{x}_i)_p (\mathbf{x}_j)_q.$$

Taking expectation, we have

$$\mathbb{E}[(S^2)_{pq}] = \sum_{i,j} \mathbb{E}\left[(\mathbf{x}_i^\top \mathbf{x}_j)(\mathbf{x}_i)_p(\mathbf{x}_j)_q\right].$$

For $i = j$, by Isserlis' theorem, we have

$$\begin{aligned}\mathbb{E}\left[(\mathbf{x}_i^\top \mathbf{x}_i)(\mathbf{x}_i)_p(\mathbf{x}_i)_q\right] &= \sum_{r=1}^{d} \mathbb{E}\left[(x_{i,r})^2 x_{i,p}x_{i,q}\right] \\ &= \sum_{r=1}^{d} (\delta_{pq} + 2\delta_{rp}\delta_{rq}) \\ &= (d+2)\delta_{pq}.\end{aligned}$$

For $i \neq j$, we have

$$\mathbb{E}\left[(\mathbf{x}_i^\top \mathbf{x}_i)(\mathbf{x}_i)_p(\mathbf{x}_i)_q\right] = \sum_{r=1}^{d} \mathbb{E}\left[x_{i,r}x_{i,p}\right]\mathbb{E}\left[x_{j,r}x_{j,q}\right] = \sum_{r=1}^{d} \delta_{rp}\delta_{rq} = \delta_{pq}.$$

There are N terms with $i = j$ and $N(N-1)$ terms with $i \neq j$. Therefore

$$\mathbb{E}[(S^2)_{pq}] = [N(d+2) + N(N-1)]\delta_{pq} = N(N+d+1)\delta_{pq}.$$

Thus, $\mathbb{E}[S^2] = N(N+d+1)\mathbf{I}_d$.

Next, we compute $\mathbb{E}[S^3]$. We have

$$(S^3)_{pq} = \sum_{i,j,k=1}^{N} (\mathbf{x}_i^\top \mathbf{x}_j)(\mathbf{x}_j^\top \mathbf{x}_k)\,(\mathbf{x}_i)_p(\mathbf{x}_k)_q.$$

There are several conditions for the triples (i, j, k).

Condition 1: For $i = j = k$, we have

$$\mathbb{E}[(\mathbf{x}_i^\top \mathbf{x}_i)^2(\mathbf{x}_i)_p(\mathbf{x}_i)_q] = \mathbb{E}[\|\mathbf{x}_i\|_2^4(\mathbf{x}_i)_p(\mathbf{x}_i)_q] = c\delta_{pq},$$

where $c = \mathbb{E}[\|\mathbf{x}\|_2^4 x_1^2]$ by taking $p = q = 1$. Then, we are going to compute the value of c. Since $\mathbb{E}[x_1^6] = 15$, $\mathbb{E}[x_1^4] = 3$, and $\mathbb{E}[x_1^2] = 1$, we can get

$$\begin{aligned} c &= \sum_r \mathbb{E}[x_r^4 x_1^2] + \sum_{r\neq s} \mathbb{E}[x_r^2 x_s^2 x_1^2] \\ &= 15 + (d-1)\cdot 3 + 2(d-1)\cdot 3 + (d-1)(d-2)\cdot 1 \\ &= d^2 + 6d + 8. \end{aligned}$$

Thus, $\mathbb{E}[(\mathbf{x}_i^\top \mathbf{x}_i)^2(\mathbf{x}_i)_p(\mathbf{x}_i)_q] = (d^2+6d+8)\delta_{pq}$.

Condition 2: Exactly two equals in (i, j, k). Without loss of generality, we assume that $i = j \neq k$. Then, we have

$$\begin{aligned} \mathbb{E}[(\mathbf{x}_i^\top \mathbf{x}_i)(\mathbf{x}_i^\top \mathbf{x}_k)(\mathbf{x}_i)_p(\mathbf{x}_k)_q] &= \sum_{r,s} \mathbb{E}[x_{i,r}^2 x_{i,s} x_{i,p}]\mathbb{E}[x_{k,s}x_{k,q}] \\ &= \sum_r \mathbb{E}[x_{i,r}^2 x_{i,s} x_{i,p}] \\ &= (d+2)\delta_{pq}, \end{aligned}$$

where we use $\mathbb{E}[x_s x_q] = \delta_{sq}$ and the previous result $\sum_r \mathbb{E}[x_r^2 x_p x_q] = (d+2)\delta_{pq}$.

Condition 3: i, j, k all distinct. We have

$$\begin{aligned} \mathbb{E}[(\mathbf{x}_i^\top \mathbf{x}_j)(\mathbf{x}_j^\top \mathbf{x}_k)(\mathbf{x}_i)_p(\mathbf{x}_k)_q] &= \sum_{r,s} \mathbb{E}[x_{i,r}x_{i,p}]\mathbb{E}[x_{j,r}x_{j,s}]\mathbb{E}[x_{k,s}x_{k,q}] \\ &= \sum_{r,s} \delta_{rp}\delta_{rs}\delta_{sq} \\ &= \delta_{pq}. \end{aligned}$$

Now, combining all conditions, we can get

$$\begin{aligned} \mathbb{E}[(S^3)_{pq}] &= N\cdot(d^2+6d+8)\delta_{pq} + 3N(N-1)\cdot(d+2)\delta_{pq} + N(N-1)(N-2)\cdot\delta_{pq} \\ &= N\left[N^2 + 3N(d+1) + d^2 + 3d + 4\right]\delta_{pq}. \end{aligned}$$

Thus,

$$\mathbb{E}[S^3] = N\left[N^2 + 3N(d+1) + d^2 + 3d + 4\right]\mathbf{I}_d.$$

□

E Additional Experiments

E.1 Linear Tasks

We provide some visualizations of trained parameters for different cases, as shown in Figure 5, Figure 6, Figure 7, and Figure 8. Based on these forms, we can easily verify that the computation process is exactly aligned with a multi-step GD algorithm, which supports our theoretical findings. We also illustrate the test loss of implicit CoT models trained end-to-end at recurrent depths $L =$

2, ..., 8 on weight prediction tasks. Each model is trained and evaluated at the matched recurrent depth. Figure 9 shows that the test loss remains far away from 0.

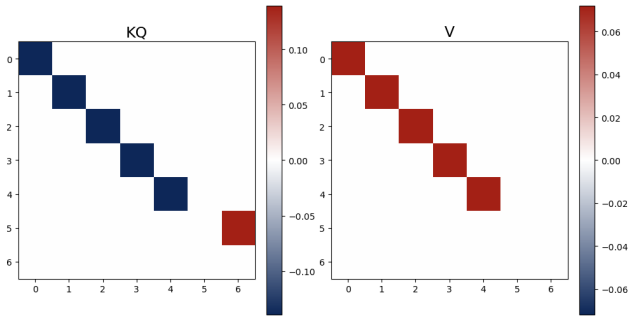


Figure 5: Visualization of trained $\mathbf{W}^{KQ}$ and $\mathbf{W}^{V}$ for the weight prediction task with explicit CoT.

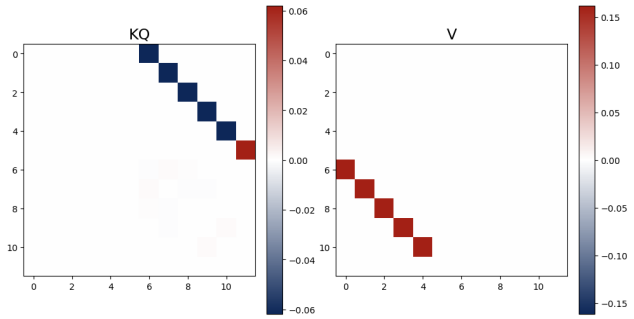


Figure 6: Visualization of trained $\mathbf{W}^{KQ}$ and $\mathbf{W}^{V}$ for the weight prediction task with implicit CoT adding a scratchpad.

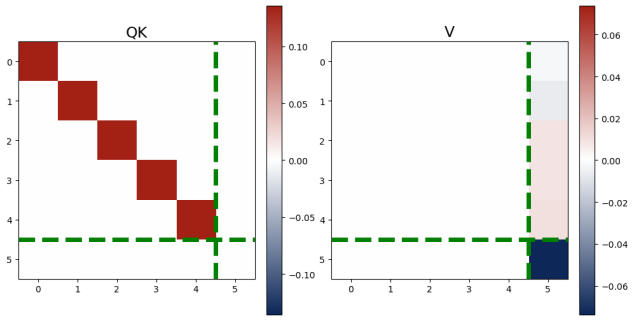


Figure 7: Visualization of trained $\mathbf{W}^{KQ}$ and $\mathbf{W}^{V}$ for the query prediction task with implicit CoT.

E.2 Extension to nonlinear LEGO Experiment

Beyond linear attention, we evaluate the role of scratchpads on the nonlinear LEGO state-tracking task. The LEGO task was introduced by Zhang et al. (2022) as a synthetic reasoning benchmark for studying how transformers track states under a sequence of symbolic transformations. We follow the state-tracking formulation and transformer setup of Huang et al. (2025b), adapting it to compare explicit and implicit CoT with and without scratchpad.

LEGO state-tracking task. Let $\mathcal{X}$, $\mathcal{G}$, and $\mathcal{Y}$ denote finite sets of variables, actions, and values, respectively. Each action $g \in \mathcal{G}$ is a map $g : \mathcal{Y} \to \mathcal{Y}$. A LEGO sentence consists of two types of

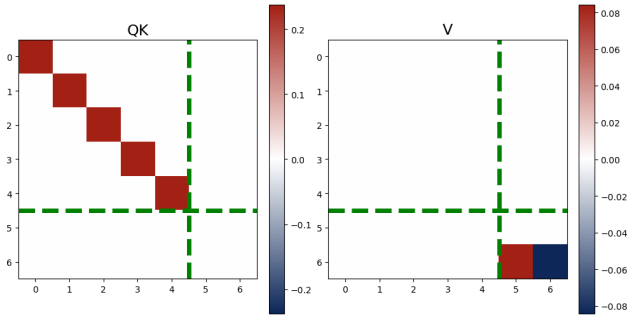


Figure 8: Visualization of trained $\mathbf{W}^{KQ}$ and $\mathbf{W}^{V}$ for the query prediction task with implicit CoT adding a scratchpad.

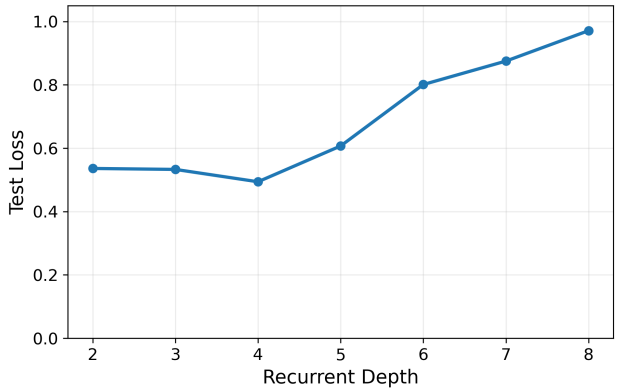


Figure 9: Test loss of implicit CoT models trained end-to-end at recurrent depths $L = 2, \ldots, 8$. The loss remains far from zero across depths, showing that the failure is not caused solely by training at a single layer.

clauses:

$$\text{predicate clause: } x = g(x'), \qquad \text{answer clause: } x = y.$$

The predicate clause specifies that the value of variable x is obtained by applying action g to the value of variable x', while the answer clause assigns value y to variable x. A LEGO instance defines a chain of variables and actions

$$x_0 \xrightarrow{g_1} x_1 \xrightarrow{g_2} x_2 \xrightarrow{g_3} \cdots \xrightarrow{g_L} x_L,$$

together with an initial answer clause $x_0 = y_0$. The corresponding values are generated recursively as

$$y_i = g_i(y_{i-1}), \qquad i = 1, \ldots, L.$$

Given all predicate clauses and the answer clauses up to step L', the model is to predict the next answer clause $x_{L'+1} = y_{L'+1}$. Thus, solving the task requires repeatedly retrieving the next applicable predicate and applying the corresponding group action to the current value.

In our experiments, we use the simply transitive LEGO setting with a cyclic group of order 6, the same as the setting in Huang et al. (2025b). Concretely, $|\mathcal{G}| = |\mathcal{Y}| = 6$, and each action corresponds to a cyclic shift on the value set.

Encoding with scratchpad. Following Huang et al. (2025b), each clause is represented as a fixed-length tuple. In the scratchpad version, predicate and answer clauses are encoded using five blocks:

$$Z_{\text{pred},i} = (x_i, g_i, x_{i-1}, \langle\text{blank}\rangle, \langle\text{blank}\rangle),$$

$$Z_{\text{ans},i} = (\langle\text{blank}\rangle, \langle\text{blank}\rangle, \langle\text{blank}\rangle, x_i, y_i).$$

The first three blocks store predicate information, while the final two blocks store and update answer information. Thus, the final two blocks act as an auxiliary scratchpad that separates the current state (x_i, y_i) from the predicate transition (x_i, g_i, x_{i-1}).

Let

$$Z^{L,L'} = (Z_{\mathrm{pred},1}, \ldots, Z_{\mathrm{pred},L}, Z_{\mathrm{ans},0}, \ldots, Z_{\mathrm{ans},L'})$$

denote the partial LEGO transcript containing all predicate clauses and answer clauses up to step L'. The target next clause is $Z_{\mathrm{ans},L'+1}$. Each discrete symbol is mapped to a vocabulary index and then represented by a one-hot embedding. The blank token is represented by the zero vector. Then, we can get the encoding sequence $\mathbf{Z}^{L,L'} = [\mathbf{Z}_{\mathrm{pred},1}, \ldots, \mathbf{Z}_{\mathrm{pred},L}, \mathbf{Z}_{\mathrm{ans},0}, \ldots, \mathbf{Z}_{\mathrm{ans},L'}]$.

Encoding without scratchpad. To analyze the role of the scratchpad, we also compare against a compressed encoding with only three blocks:

$$Z^{\mathrm{non}}_{\mathrm{pred},i} = (x_i, g_i, x_{i-1}), \qquad Z^{\mathrm{non}}_{\mathrm{ans},i} = (x_i, y_i, \langle \mathrm{blank} \rangle).$$

In this representation, predicate information and answer information must share the same coordinate blocks. In particular, the value y_i and the action g_{i+1} are no longer separated into different blocks after attention aggregation. This compression makes the next-value prediction harder, because the model must preserve the current state and compute the next transition using the same representation space. Similar to the encoding in the case with scratchpad, we can get the encoding sequence $\mathbf{Z}^{\mathrm{non},L,L'} = [\mathbf{Z}^{\mathrm{non}}_{\mathrm{pred},1}, \ldots, \mathbf{Z}^{\mathrm{non}}_{\mathrm{pred},L}, \mathbf{Z}^{\mathrm{non}}_{\mathrm{ans},0}, \ldots, \mathbf{Z}^{\mathrm{non}}_{\mathrm{ans},L'}]$.

Transformer model. We use a one-layer softmax transformer, following the LEGO setup of Huang et al. (2025b). The input is the sequence of clause embeddings $\mathbf{Z}^{L,L'}$, and the model predicts the next answer clause from the latest answer token. The model consists of a softmax attention layer followed by a feed-forward network. The attention score between a query clause $\mathbf{Z}_j$ and a source clause $\mathbf{Z}_k$ is written as

$$\mathrm{score}(j,k) = \mathbf{Z}_j^\top \mathbf{W} \mathbf{Z}_k,$$

and the attention weight is

$$\mathrm{Attn}_{j\to k}(\mathbf{W}, \mathbf{Z}^{L,L'}) = \frac{\exp(\mathbf{Z}_j^\top \mathbf{W} \mathbf{Z}_k)}{\sum_r \exp(\mathbf{Z}_j^\top \mathbf{W} \mathbf{Z}_r)}.$$

The attention output for predicting the next answer clause is computed from the latest answer clause:

$$\mathrm{Attention}(\mathbf{W}, \mathbf{Z}^{L,L'}) = \sum_k \mathrm{Attn}_{\mathrm{ans},L'\to k}(\mathbf{W}, \mathbf{Z}^{L,L'}) \mathbf{Z}_k.$$

The feed-forward network then maps this attention output to logits for the next clause. Equivalently, the model outputs one logit vector for each block of the next clause:

$$F(\mathbf{Z}^{L,L'}) = \big(F_1(\mathbf{Z}^{L,L'}), \ldots, F_B(\mathbf{Z}^{L,L'})\big),$$

where $B = 5$ for the scratchpad encoding and $B = 3$ for the compressed no-scratchpad encoding. Each $F_b(Z^{L,L'})$ is a logit vector over the vocabulary.

In our implementation, we use a single softmax-attention transformer layer with two attention heads and a standard feed-forward block. The standard implementation uses learned $Q/K/V$ projections, an output projection, residual connections, layer normalization, and a dense feed-forward network. This is slightly richer than the minimal bilinear-attention notation above, but follows the same next-clause prediction interface and preserves the intended one-layer softmax transformer structure.

Training objective. Training uses next-clause supervision. Given a ground-truth partial transcript $Z^{L,L'-1}$, the model is trained to predict the next answer clause $Z_{\mathrm{ans},L'}$. The loss is the average cross-entropy over the blocks of the answer clause:

$$\mathcal{L}^{L,L'}(F) = \mathbb{E}_{Z^{L,L'}\sim\mathcal{D}^{L,L'}} \left[-\frac{1}{B} \sum_{b=1}^{B} \log p_F\big(Z_{\mathrm{ans},L',b} \mid Z^{L,L'-1}\big) \right],$$

where $B = 5$ for the scratchpad encoding and $B = 3$ for the no-scratchpad encoding. The full training objective sums this next-clause loss over the answer steps used for training:

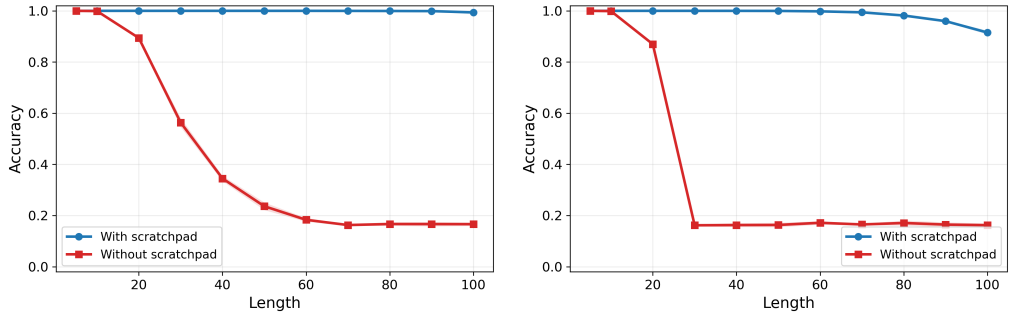
$$\mathcal{L}^{L}(F) = \frac{1}{L} \sum_{L'=1}^{L} \mathcal{L}^{L,L'}(F).$$

Explicit and implicit CoT. For explicit CoT, the model autoregressively generates answer clauses. Starting from

$$(Z_{\mathrm{pred},1}, \ldots, Z_{\mathrm{pred},L}, Z_{\mathrm{ans},0}),$$

the model predicts $\widehat{Z}_{\mathrm{ans},1}$, appends it to the sequence, then predicts $\widehat{Z}_{\mathrm{ans},2}$, and so on. The intermediate reasoning states are therefore externalized as generated answer clauses in the visible sequence.

For implicit CoT, the model does not append generated answer clauses to the sequence. Instead, it repeatedly applies the same transformer block to update the internal answer-state representation. The intermediate computation is therefore kept inside the recurrent internal updates. In the scratchpad setting, the answer blocks provide additional space for storing and updating the current variable-value pair.



(a) Explicit CoT (b) Implicit CoT

Figure 10: Length generalization on the LEGO state-tracking task with and without scratchpad. We report final next-answer value accuracy at lengths $L \in \{5, 10, 20, 30, \ldots, 100\}$.

Training and evaluation. We generate LEGO instances online by sampling $L+1$ distinct variables uniformly from a set of 121 variables, an initial value uniformly from a set of 6 values, and each action uniformly from a set of 6 cyclic shifts. The values are obtained by applying the sampled shifts successively modulo 6. Each symbol is represented by a one-hot vector over a vocabulary of 134 symbols, and blank blocks are encoded as zero vectors. We use a one-layer, two-head softmax transformer, with learned query, key, value, and output projections, residual connections, and a GELU feed-forward network. Training uses batches of 128 newly generated instances and the Adam optimizer for 30,000 updates, with a learning rate 10^{-4}. The explicit CoT model is trained with teacher-guided next-answer supervision, whereas the implicit model is trained to predict each next answer from the corresponding teacher state. For evaluation, we use autoregressive generation for explicit CoT and recurrent internal updates for implicit CoT, and report final-step value accuracy at lengths $L \in \{5, 10, 20, 30, \ldots, 100\}$. Each evaluation uses 20 batches of 128 instances. Curves show the mean and standard deviation across five seeds.

Analysis of results. Figure 10 illustrates the length-generalization results. We train the model with a short sequence length $L = 5$ and evaluate it at longer lengths up to $L = 100$. Figure 10 shows that in both explicit and implicit CoT, the scratchpad substantially improves long-length value accuracy compared with the case without the scratchpad. With the scratchpad, predicate information and answer information are stored in different blocks. Thus, when the current answer token attends to the relevant predicate token, the current value $\mathbf{y}_i$ and the next action $\mathbf{g}_{i+1}$ remain separated in the attention-weighted representation, allowing the model to compose them and recover the next value $\mathbf{y}_{i+1} = \mathbf{g}_{i+1}(\mathbf{y}_i)$. In contrast, the encoding without scratchpad compresses the value and the action into the same block. Thus, even if the answer token $(\mathbf{x}_i, \mathbf{y}_i, 0)$ attends to the correct predicate token $(\mathbf{x}_{i+1}, \mathbf{g}_{i+1}, \mathbf{x}_i)$, the attention-weighted representation mixes $\mathbf{y}_i$ and $\mathbf{g}_{i+1}$ in the same coordinates, making it difficult to predict the next value from this entangled representation. This is similar to the

bottlenecks identified in the linear tasks: without enough auxiliary space, the model may use the same coordinates both to preserve the existing state and to compute the next update. The scratchpad removes this conflict by separating existing information from update information. Although the LEGO task uses softmax attention and nonlinear group actions rather than linear attention and GD update, the failure mechanism is consistent with the insufficient computational space discussed in the previous sections.