

Supplementary Materials

| | | |
|-----|--|-----------|
| 803 | Contents | |
| 804 | 1 Introduction | 1 |
| 805 | 2 Related Works | 2 |
| 806 | 3 Preliminaries | 3 |
| 807 | 3.1 Transformer Architecture | 3 |
| 808 | 3.2 Augmented In-context Learning | 4 |
| 809 | 3.3 Chain-of-Thought Prompting for Augmented ICL | 4 |
| 810 | 4 Expressiveness with CoT Prompting for Augmented ICL | 5 |
| 811 | 5 Training Dynamics with Teacher Forcing | 7 |
| 812 | 6 Experimental Results | 9 |
| 813 | 7 Conclusion | 9 |
| 814 | A Broader Impacts | 21 |
| 815 | B Proof of Expressiveness | 21 |
| 816 | B.1 Proof of Theorem 4.1 | 21 |
| 817 | B.2 Proof of Theorem 4.2 | 24 |
| 818 | B.3 Proof of Corollary 4.1 | 29 |
| 819 | C Proof of Training Dynamics | 30 |
| 820 | D Auxiliary Lemmas | 33 |
| 821 | E Limitations | 33 |

822 A Broader Impacts

823 This work provides theoretical insights into how transformers can leverage unlabeled data to improve
 824 in-context learning, a core capability underlying many recent advances in language models. By
 825 improving data efficiency and adaptability, our findings could enable more accessible and capable
 826 AI systems, particularly in low-resource settings where labeled data is limited. These advances may
 827 benefit a range of applications, including next-generation wireless communications and networking,
 828 healthcare, and financial services. Given the theoretical nature of this work, we anticipate minimal
 829 direct negative societal impact. Nonetheless, we recognize that future practical implementations
 830 inspired by this research should adhere to responsible AI principles.

831 B Proof of Expressiveness

832 First, we restate the theorem:

833 **Theorem B.1.** *There exists a 4-layer transformer, such that its output sequence at the $(t + 1)$ -th*
 834 *CoT step satisfies*

$$\hat{\boldsymbol{\mu}}_i^{(t+1)} = \hat{\boldsymbol{\mu}}_i^{(t)} - \frac{\eta^{(t)}}{M} \sum_{j=N+1}^{N+M} p_{ij}^{(t)} (\hat{\boldsymbol{\mu}}_i^{(t)} - \mathbf{x}_j) + \mathbf{1}_{\{t=0\}} \cdot \frac{C}{N} \sum_{j=1}^N (\mathbf{e}_i^\top \mathbf{y}_j) \mathbf{x}_j, \quad (\text{B.1})$$

835 for any $i \in [C]$, where $\eta^{(t)} = \alpha/(T' + t)$ for some positive constants α and T' , $p_{ij}^{(t)}$ is the normalized
 836 weight

$$p_{ij}^{(t)} = \frac{\sum_{\tau=0}^t \exp\left(-\frac{1}{2} \|\hat{\boldsymbol{\mu}}_i^{(\tau)} - \mathbf{x}_j\|_{\Sigma^{-1}}^2 + \beta\tau\right)}{\sum_{\tau=0}^t \sum_{c=1}^C \exp\left(-\frac{1}{2} \|\hat{\boldsymbol{\mu}}_c^{(\tau)} - \mathbf{x}_j\|_{\Sigma^{-1}}^2 + \beta\tau\right)}, \quad (\text{B.2})$$

837 and β is a positive constant.

838 We start from the proof of Theorem 4.1, which shows the transformer’s capability of implementing
 839 an EM-style algorithm.

840 B.1 Proof of Theorem 4.1

841 Recall that the input sequence at the t -th CoT step is formulated as

$$\hat{\mathbf{H}}^{(t-1)} = \begin{bmatrix} \mathbf{X}_\ell & \mathbf{X}_u & \mathbf{0} & \star & \cdots & \star \\ \mathbf{Y}_\ell & \mathbf{0} & \mathbf{0} & \star & \cdots & \star \\ \mathbf{P}_\ell & \mathbf{P}_u & \mathbf{Q}^{(0)} & \mathbf{Q}^{(1)} & \cdots & \mathbf{Q}^{(t-1)} \end{bmatrix},$$

842 where

$$\mathbf{P}_\ell = [\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_N], \quad (\text{B.3})$$

$$\mathbf{P}_u = [\mathbf{p}_{N+1}, \mathbf{p}_{N+2}, \cdots, \mathbf{p}_{N+M}], \quad (\text{B.4})$$

$$\mathbf{Q}^{(\tau)} = [\mathbf{q}_1^{(\tau)}, \mathbf{q}_2^{(\tau)}, \cdots, \mathbf{q}_C^{(\tau)}], \quad \tau \in [0 : t-1]. \quad (\text{B.5})$$

843 We specify \mathbf{p}_j and $\mathbf{q}_i^{(\tau)}$ as follows. For each data sample $j \in [N + M]$, we denote

$$\mathbf{p}_j = \begin{bmatrix} \mathbf{0}_C \\ \mathbf{0}_d \\ \mathbf{0}_C \\ \mathbf{0} \\ \mathbf{1}_{j \in [N]} \\ \mathbf{1}_{j \in [N+1:N+M]} \\ \mathbf{0} \end{bmatrix}.$$

844 For each class $i \in [C]$, the corresponding $q_i^{(\tau)}$ has the following form

$$\mathbf{q}_i^{(\tau)} = \begin{bmatrix} \mathbf{e}_i \\ \hat{\boldsymbol{\mu}}_i^{(\tau)} \\ \mathbf{0}_C \\ \mathbf{0}_C \\ u_i^{(\tau)} \\ 0 \\ 0 \\ \tau \end{bmatrix},$$

845 where $\hat{\boldsymbol{\mu}}_i^{(\tau)}$ stores the estimate of the mean vector of class i from the τ -th CoT step, and u_i^τ stores a
846 rescaled norm of $\hat{\boldsymbol{\mu}}_i^{(\tau)}$, i.e., $u_i^{(\tau)} = -\frac{\sigma}{2} \|\hat{\boldsymbol{\mu}}_i^{(\tau)}\|^2$.

847 Next, we specify the parameters of each layer of the transformer as follows.

848 **Layer 1:** The first layer of the transformer consists of an attention layer with a softmax activation
849 function, and an MLP layer. Let the parameters of the attention layer satisfy

$$\mathbf{Q}_1 \mathbf{K}_1 = \begin{bmatrix} \mathbf{0}_{d \times (d+2C)} & \boldsymbol{\Sigma}^{-1} & & & \\ & \mathbf{0}_{(4C+d+2) \times 2C} & & & \\ & & 1 & \mathbf{0}_{1 \times 2} & \beta \\ & & & & 0 \end{bmatrix},$$

$$\mathbf{V}_1 = \begin{bmatrix} \mathbf{0}_{(d+2C) \times (d+C)} & & \\ & \mathbf{I}_C & \\ & & \mathbf{0}_{(d+C+4) \times (d+2C+4)} \end{bmatrix}.$$

850 Denote $\text{attn}_1(\mathbf{p}_j)$ as the output token after passing \mathbf{p}_j through the first attention layer, and let
851 $\gamma_i := \text{attn}_1(\mathbf{p}_j)[d+C+1 : d+2C]$. Then, we have

$$\gamma_j = \frac{\sum_{\tau \in [0:t-1]} \sum_{i \in [C]} \exp \left(-\frac{\sigma}{2} \|\hat{\boldsymbol{\mu}}_i^{(\tau)}\|_{\boldsymbol{\Sigma}^{-1}}^2 + (\hat{\boldsymbol{\mu}}_i^{(\tau)})^\top \mathbf{x}_j + \beta \tau \right) \mathbf{e}_i}{\sum_{\tau \in [0:t-1]} \sum_{i \in [C]} \exp \left(-\frac{\sigma}{2} \|\hat{\boldsymbol{\mu}}_i^{(\tau)}\|_{\boldsymbol{\Sigma}^{-1}}^2 + (\hat{\boldsymbol{\mu}}_i^{(\tau)})^\top \mathbf{x}_j + \beta \tau \right)}.$$

852 Other entries in $\hat{\mathbf{H}}^{(t-1)}$ remain unchanged after this attention layer.

853 Subsequent to the first attention layer, a token-wise MLP is applied. Similar to Kim and Suzuki
854 (2024b), in this work, we assume the MLP layer can realize any deterministic token-wise link func-
855 tion with negligible error. The first MLP layer transforms input representations \mathbf{p} such that

$$\text{mlp}_1(\text{attn}_1(\mathbf{p}_j)) = \gamma_j \cdot \text{attn}_1(\mathbf{p}_j)[3C+d+3]$$

$$\text{mlp}_1(u_i^{(\tau)}) = -\frac{\sigma}{2} \|\hat{\boldsymbol{\mu}}_i^{(\tau)}\|^2.$$

856 Since $\mathbf{p}_j[3C+d+3] = 0$ for $j \in [N]$ and $\mathbf{p}_j[3C+d+3] = 1$ for $j \in [N+1 : N+M]$, and
857 the corresponding entries remain unchanged after passing through the first attention layer, this MLP
858 layer only keeps γ_j for tokens corresponding to the unlabeled dataset (i.e., $j \in [N]$), and set γ_j to
859 zero for all other tokens (i.e., $j \in [N+1 : M]$).

860 **Layer 2:** The second layer of the transformer consists of an attention layer with a linear activation
861 function, and an MLP layer. The parameters of the attention layer are set to satisfy

$$\mathbf{Q}_2 \mathbf{K}_2 = \begin{bmatrix} \mathbf{0}_{(2d+4C+2) \times (2d+4C+2)} & & \\ & 0 & 0 \\ & \alpha_1 & 0 \end{bmatrix}$$

$$\mathbf{V}_2 = \begin{bmatrix} \mathbf{0}_{(2d+3C) \times (d+2C)} & & \\ & \mathbf{I}_C & \\ & & \mathbf{0}_{4 \times (d+C+4)} \end{bmatrix}.$$

862 We denote $\mathbf{s}_i^{(\tau)} := \text{attn}_2(\mathbf{q}_i^{(\tau)})[d+2C+1 : d+3C]$ as the vector extracted from the output token
863 after passing $\mathbf{q}_i^{(\tau)}$ through the second attention layer. Then, $\mathbf{s}_i^{(\tau)} = \tau \alpha_1 \sum_{j=N+1}^{N+M} \gamma_j$, where α_1 is a
864 fixed scalar embedded in $\mathbf{Q}_2 \mathbf{K}_2$.

865 We let the subsequent MLP layer realize the following token-wise Lipschitz function:

$$\begin{aligned}\text{mlp}_2(\hat{\boldsymbol{\mu}}_i^{(\tau)}) &= \hat{\boldsymbol{\mu}}_i^{(\tau)} - \frac{1}{\tau(\tau + \alpha_2)} \hat{\boldsymbol{\mu}}_i^{(\tau)} \mathbf{e}_i^\top \mathbf{s}_i^{(\tau)} = \hat{\boldsymbol{\mu}}_i^{(\tau)} - \frac{\alpha_1}{\tau + \alpha_2} \hat{\boldsymbol{\mu}}_i^{(\tau)} \mathbf{e}_i^\top \sum_{j \in [N+1]}^{N+M} \gamma_j \\ \text{mlp}_2(\mathbf{e}_i) &= \frac{\alpha_1}{\tau + \alpha_2} \mathbf{e}_i.\end{aligned}$$

866 **Layer 3:** Similar to the second transformer layer, the third layer also consists of a linear attention
867 layer and an MLP layer. Consider the following parameterization for the attention layer:

$$\begin{aligned}\mathbf{Q}_3 \mathbf{K}_3 &= \begin{bmatrix} \mathbf{0}_{(d+C) \times (d+2C)} & \mathbf{I}_C & \\ & & \mathbf{0}_{(d+2C+4) \times (d+C+4)} \end{bmatrix} \\ \mathbf{V}_3 &= \begin{bmatrix} \mathbf{0}_{(d+3C) \times d} & \\ \mathbf{I}_d & \\ & & \mathbf{0}_{C+4; d+4C+4} \end{bmatrix}.\end{aligned}$$

868 Therefore, this attention layer realizes the following updating process:

$$\text{attn}_3(\hat{\boldsymbol{\mu}}_i^{(\tau)}) = \text{mlp}_2(\hat{\boldsymbol{\mu}}_i^{(\tau)}) + \frac{\alpha_1}{\tau + \alpha_2} \sum_{j \in [M]} \mathbf{x}_j \mathbf{e}_i^\top \gamma_j^{(\tau)}.$$

869 After this linear attention layer, we let the MLP layer realize the following function

$$\text{mlp}_3(\mathbf{e}_i) = \frac{\tau + \alpha_2}{\alpha_1} \mathbf{e}_i$$

870 **Layer 4:** For the last layer, we introduce a transformer layer with a ReLU-activated attention layer
871 followed by an MLP layer. We parameterize the attention layer as:

$$\begin{aligned}\mathbf{Q}_4 \mathbf{K}_4 &= \begin{bmatrix} \mathbf{0}_{(d+C) \times d} & \mathbf{I}_C & & \\ & & \mathbf{0}_{(d+2C+1) \times (d+3C+3)} & \\ & & & 1 \\ & & & \mathbf{0}_2 \end{bmatrix} \\ \mathbf{V}_4 &= \begin{bmatrix} \mathbf{0}_{(d+2C) \times d} & \\ \mathbf{I}_d & \\ & & \mathbf{0}_{(2C+4) \times (4C+d+4)} \end{bmatrix}.\end{aligned}$$

872 The corresponding updating rule of this layer gives

$$\text{attn}_4(\boldsymbol{\mu}_i^{(\tau)}) = \text{attn}_3(\boldsymbol{\mu}_i^{(\tau)}) + \frac{C}{N} \sum_{j \in [N]} \mathbf{x}_j \text{ReLU}(-\tau + \mathbf{e}_i^\top \mathbf{y}_j).$$

873 Therefore, we can further reformulate it as

$$\text{attn}_4(\boldsymbol{\mu}_i^{(\tau)}) = \begin{cases} \frac{C}{N} \sum_{j \in [N]} \mathbf{x}_j \cdot (\mathbf{e}_i^\top \mathbf{y}_j), & \text{if } \tau = 0, \\ \text{attn}_3(\boldsymbol{\mu}_i^{(\tau)}), & \text{if } \tau > 0. \end{cases}$$

874 Given the above 4-layer transformer structure, by setting $\alpha_1 = \alpha/M$ and $\alpha_2 = T'$ for fixed $\alpha > 0$,
875 $T' > 0$, the output sequence corresponding to the $\mathbf{Q}^{(t-1)}$ block in the input sequence that satisfies:

$$\hat{\boldsymbol{\mu}}_i^{(t+1)} = \hat{\boldsymbol{\mu}}_i^{(t)} - \frac{\eta^{(t)}}{M} \sum_{j=N+1}^{N+M} p_{ij}^{(t)} (\hat{\boldsymbol{\mu}}_i^{(t)} - \mathbf{x}_j) + \mathbf{1}_{\{t=0\}} \cdot \frac{C}{N} \sum_{j=1}^N (\mathbf{e}_i^\top \mathbf{y}_j) \mathbf{x}_j, \quad (\text{B.6})$$

876 for any $i \in [C]$, where $\eta^{(t)} = \alpha/(T' + t)$ for some positive constants α and T' , $p_{ij}^{(t)}$ is the normalized
877 weight

$$p_{ij}^{(t)} = \sum_{\tau=0}^t \exp\left(-\frac{1}{2} \|\hat{\boldsymbol{\mu}}_i^{(\tau)} - \mathbf{x}_j\|_{\Sigma^{-1}}^2 + \beta\tau\right) \Bigg/ \sum_{\tau=0}^t \sum_{c=1}^C \exp\left(-\frac{1}{2} \|\hat{\boldsymbol{\mu}}_c^{(\tau)} - \mathbf{x}_j\|_{\Sigma^{-1}}^2 + \beta\tau\right).$$

878 The proof is thus complete.

879 B.2 Proof of Theorem 4.2

880 In this section, we show the detailed proof of Theorem 4.2. We start by restating the theorem:

881 **Theorem B.2** (Class Mean Estimation Error.). *Given the transformer described in Theorem 4.1,*
 882 *when $N \geq 36\alpha^2 L^2 \log 1/\epsilon$, $M \geq \max\{(T')^4, \log^2 1/\epsilon\}$, and $t \geq \max\{\sqrt[4]{M}, T'\}$, with probability*
 883 *at least $1 - \epsilon$, the output of the transformer after t CoT steps satisfies*

$$\|\widehat{\mathbf{M}}^{(t)} - \mathbf{M}\|_F^2 \leq c \frac{\log(1/\epsilon)}{N \sqrt[4]{M}},$$

884 where c, α, L, T' are positive constants.

885 **Step 1: First, we ensure that the initial estimation of the class mean vectors obtained from the**
 886 **labeled data gives a small estimation error.**

887 **Lemma 1** (Initial estimation error from labeled data.). *Consider the initial class mean estimates*

$$\boldsymbol{\mu}_i^{(1)} = \frac{C}{N} \sum_{j \in [N]} \mathbf{x}_j \cdot (\mathbf{e}_i^\top \mathbf{y}_j), \quad \forall i \in [C].$$

888 Then, for fixed $K \geq 1$ and any positive constant $T' \geq 4K$, we have

$$\mathbb{P} \left[\|\boldsymbol{\mu}_i^{(1)} - \boldsymbol{\mu}_i\|^2 > \frac{K}{T'} \right] \leq \exp(-cNK/T'),$$

889 where c is a positive constant.

890 *Proof.* We denote n_i as the number of samples drawn from class i in the N labeled data. Under the
 891 assumption that $\mathbf{y}_j \sim \text{Uniform}(\mathcal{Y})$, $\forall j \in [N]$, we have $n_i \sim \text{Binomial}(N, 1/C)$. Then, according
 892 to Chernoff's inequality, for any $\epsilon \in (0, 1)$, we have

$$\mathbb{P} \left(\left| n_i - \frac{N}{C} \right| > \epsilon \frac{N}{C} \right) \leq 2 \exp \left(-t \frac{\epsilon^2 N}{3C} \right).$$

893 For any $u \geq 0$, Let $\epsilon = u \sqrt{K/T'}$, we obtain

$$\mathbb{P} \left(\left| n_i - \frac{N}{C} \right| > u \sqrt{\frac{K}{T'} \frac{N}{C}} \right) \leq 2 \exp \left(-\frac{u^2 NK}{3CT'} \right).$$

894 Therefore,

$$\mathbb{P} \left(\left| \frac{C}{N} n_i - 1 \right| > u \sqrt{\frac{K}{T'}} \right) \leq 2 \exp \left(-\frac{u^2 NK}{3CT'} \right). \quad (\text{B.7})$$

895 Conditional on n_i , we have $\frac{1}{n_i} \sum_{j: \mathbf{y}_j = \mathbf{e}_i} \mathbf{x}_j - \boldsymbol{\mu}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}/n_i)$. We assume $\boldsymbol{\Sigma}$ is an isotropic
 896 matrix in the form of $\sigma^2 \mathbb{I}$. Then, $\|\boldsymbol{\Sigma}\|_2 = \sigma^2$, and we obtain the following inequality based on
 897 Hoeffding's inequality.

$$\mathbb{P} \left(\left\| \frac{1}{n_i} \sum_{j: \mathbf{y}_j = \mathbf{e}_i} \mathbf{x}_j - \boldsymbol{\mu}_i \right\| > \sigma \sqrt{\frac{2t}{n_i}} \middle| n_i \right) \leq 2e^{-t}.$$

898 For any $v \geq 0$, by setting $t = v^2 n_i K / (2\sigma^2 T')$, we have

$$\mathbb{P} \left(\left\| \frac{1}{n_i} \sum_{j: \mathbf{y}_j = \mathbf{e}_i} \mathbf{x}_j - \boldsymbol{\mu}_i \right\| > v \sqrt{\frac{K}{T'}} \right) \leq 2 \exp(-v^2 n_i K / (8\sigma^2 T')) \quad (\text{B.8})$$

$$\leq 2 \exp(-v^2 (1 - \frac{K}{T'}) \frac{NK}{C\sigma^2 T'}) \quad (\text{B.9})$$

$$\leq 2 \exp(-v^2 \frac{NK}{2C\sigma^2 T'}). \quad (\text{B.10})$$

899 Then,

$$\begin{aligned}
\mathbb{P}\left(\|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\|^2 > \frac{K}{T'}\right) &= \mathbb{P}\left(\left\|\frac{C}{N} \sum_{j:\mathbf{y}_j=\mathbf{e}_i} \mathbf{x}_j - \frac{Cn_i}{N} \boldsymbol{\mu}_i - \left(1 - \frac{Cn_i}{N}\right) \boldsymbol{\mu}_i\right\| > \sqrt{\frac{K}{T'}}\right) \\
&\leq \mathbb{P}\left(\frac{Cn_i}{N} \left\|\frac{1}{n_i} \sum_{j:\mathbf{y}_j=\mathbf{e}_i} \mathbf{x}_j - \boldsymbol{\mu}_i\right\| + \left|1 - \frac{Cn_i}{N}\right| \|\boldsymbol{\mu}_i\| \geq \sqrt{\frac{K}{T'}}\right) \\
&\leq \mathbb{P}\left(\frac{Cn_i}{N} \left\|\frac{1}{n_i} \sum_{j:\mathbf{y}_j=\mathbf{e}_i} \mathbf{x}_j - \boldsymbol{\mu}_i\right\| \geq \sqrt{\frac{K}{T'}}, \text{ or } \left|1 - \frac{Cn_i}{N}\right| \|\boldsymbol{\mu}_i\| \geq \sqrt{\frac{K}{T'}}\right) \\
&\stackrel{(a)}{\leq} 4 \exp(-c \frac{NK}{T'})
\end{aligned}$$

900 for positive constant c . The inequality (a) holds by setting $u = 1/\|\boldsymbol{\mu}_i\|$ in Equation (B.10) and
 901 setting $v = N/Cn_i$ in Equation (B.14). The proof is thus complete. \square

902 **Step 2: Next, we bound the discrepancy between the gradient obtained from each CoT step for**
 903 **a given input sequence, and the gradient of the population loss.**

904 We define the population loss for any given set of class mean vectors $\{\boldsymbol{\mu}_i\}_{i \in [C]}$ (i.e., any given \mathbf{M})
 905 as:

$$\mathcal{L}(\{\boldsymbol{\mu}_i\}) = \mathbb{E}_{\mathbf{x}} \left[\log \left(\frac{1}{C} \sum_{i=1}^C \exp \left(-\frac{1}{2} \|\mathbf{x} - \boldsymbol{\mu}_i\|_{\Sigma^{-1}}^2 \right) \right) \right], \quad (\text{B.11})$$

906 where the expectation is taken over the randomly generated data \mathbf{x} for given \mathbf{M} , as specified in
 907 Equation (3.1).

908 We first characterize an important property of $\mathcal{L}(\{\boldsymbol{\mu}_i\})$ as follows.

909 **Lemma 2.** *The Jacobian of $\nabla_{\boldsymbol{\mu}_i} \mathcal{L}$ at $\boldsymbol{\mu}_i$ for all $i \in [C]$ is negative definite, i.e., $\nabla_{\boldsymbol{\mu}_i}^2 \mathcal{L} \prec \mathbf{0}$.*

910 *Proof.* Define

$$p_{\mathbf{x}}(\boldsymbol{\mu}_i) = \frac{\exp \left(-\frac{1}{2} \|\mathbf{x} - \boldsymbol{\mu}_i\|_{\Sigma^{-1}}^2 \right)}{\sum_{c=1}^C \exp \left(-\frac{1}{2} \|\mathbf{x} - \boldsymbol{\mu}_c\|_{\Sigma^{-1}}^2 \right)},$$

911 so that $p_{\mathbf{x}}(\boldsymbol{\mu}_i)$ is a softmax weight depending on \mathbf{x} and the centers $\{\boldsymbol{\mu}_c\}_{c=1}^C$. Note that $\nabla_{\boldsymbol{\mu}_i}^2 \mathcal{L}$ is the
 912 Hessian of $\nabla \mathcal{L}$ at $\boldsymbol{\mu}_i$, given by

$$\nabla_{\boldsymbol{\mu}_i}^2 \mathcal{L} = \mathbb{E}_{\mathbf{x}} \left[p_{\mathbf{x}}(\boldsymbol{\mu}_i) (1 - p_{\mathbf{x}}(\boldsymbol{\mu}_i)) \Sigma^{-1} (\boldsymbol{\mu}_i - \mathbf{x}) (\boldsymbol{\mu}_i - \mathbf{x})^\top \Sigma^{-1} - p_{\mathbf{x}}(\boldsymbol{\mu}_i) \Sigma^{-1} \right],$$

913 where and the expectation is taken with respect to the distribution of \mathbf{x} . Therefore, there exists a
 914 constant $0 \leq \alpha < 1$ such that

$$\nabla_{\boldsymbol{\mu}_i}^2 \mathcal{L} \preceq \mathbb{E}_{\mathbf{x}} \left[\alpha p_{\mathbf{x}}(\boldsymbol{\mu}_i) \Sigma^{-1} (\boldsymbol{\mu}_i - \mathbf{x}) (\boldsymbol{\mu}_i - \mathbf{x})^\top \Sigma^{-1} - p_{\mathbf{x}}(\boldsymbol{\mu}_i) \Sigma^{-1} \right].$$

915 Now, for any nonzero vector $\mathbf{u} \in \mathbb{R}^d$, consider the quadratic form $\mathbf{u}^\top \nabla_{\boldsymbol{\mu}_i}^2 \mathcal{L} \mathbf{u}$, using the above
 916 matrix inequality, we have

$$\mathbf{u}^\top \nabla_{\boldsymbol{\mu}_i}^2 \mathcal{L} \mathbf{u} \leq \mathbf{u}^\top \mathbb{E}_{\mathbf{x}} \left[\alpha p_{\mathbf{x}}(\boldsymbol{\mu}_i) \Sigma^{-1} (\boldsymbol{\mu}_i - \mathbf{x}) (\boldsymbol{\mu}_i - \mathbf{x})^\top \Sigma^{-1} - p_{\mathbf{x}}(\boldsymbol{\mu}_i) \Sigma^{-1} \right] \mathbf{u}.$$

917 Therefore, rewriting the expectation as an integral yields

$$\begin{aligned}
\mathbf{u}^\top \nabla_{\boldsymbol{\mu}_i}^2 \mathcal{L} \mathbf{u} &\leq \frac{1}{C} \int_{\mathbb{R}^d} \alpha \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_i, \mathbf{I}) \mathbf{u}^\top \Sigma^{-1} (\boldsymbol{\mu}_i - \mathbf{x}) (\boldsymbol{\mu}_i - \mathbf{x})^\top \Sigma^{-1} \mathbf{u} \, d\mathbf{x} - \frac{1}{C} \mathbf{u}^\top \Sigma^{-1} \mathbf{u} \\
&\leq \frac{\alpha}{C} \mathbf{u}^\top \Sigma^{-1} \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma^{-1})} \left[(\boldsymbol{\mu}_i - \mathbf{x}) (\boldsymbol{\mu}_i - \mathbf{x})^\top \right] \Sigma^{-1} \mathbf{u} - \frac{1}{C} \mathbf{u}^\top \Sigma^{-1} \mathbf{u} < 0.
\end{aligned}$$

918 Thus, the quadratic form is negative for every nonzero \mathbf{u} , and the matrix $\nabla_{\boldsymbol{\mu}_i}^2 \mathcal{L}$ is negative definite.
 919 This completes the proof. \square

920 We note that for each CoT step $t > 0$, the updating induced by the constructed transformer is

$$\hat{\boldsymbol{\mu}}_i^{(t+1)} = \hat{\boldsymbol{\mu}}_i^{(t)} - \frac{\eta^{(t)}}{M} \sum_{j=N+1}^{N+M} p_{ij}^{(t)} (\hat{\boldsymbol{\mu}}_i^{(t)} - \mathbf{x}_j) \quad (\text{B.12})$$

921 where $p_{ij}^{(t)}$ is defined in Equation (B.2).

922 To simplify notation, denote

$$\frac{1}{M} \sum_{j=N+1}^{N+M} p_{ij}^{(t)} (\hat{\boldsymbol{\mu}}_i^{(t)} - \mathbf{x}_j) := \nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \hat{\mathcal{L}}. \quad (\text{B.13})$$

923 We note that $\hat{\mathcal{L}}$ itself is not an explicit loss function. We use the notation $\nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \hat{\mathcal{L}}$ to represent the
924 equivalent gradient for the updating determined by the t -th CoT step.

925 In the following, we characterize $\nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \hat{\mathcal{L}}$ and compare it with $\nabla \mathcal{L}$, i.e., the gradient if GD is per-
926 formed on the population loss. We have the following lemma.

927 **Lemma 3** (Properties of the CoT gradient descent). *Fix an epoch t and a component index $i \in [C]$,
928 there exist constants $c_1, c_2 > 0$ such that, for every $M \geq 1$,*

$$\Pr\left(\left\|\nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \hat{\mathcal{L}} - \nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \mathcal{L}\right\| \leq c_1 M^{-1/4}\right) \geq 1 - \exp(-\sqrt{M}),$$

929 and

$$\Pr\left(\left\|\nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \hat{\mathcal{L}}\right\|^2 \leq c_2 + c_3 M^{-1/2}\right) \geq 1 - \exp(-\sqrt{M}).$$

930 *Proof.* Recall that

$$\nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \hat{\mathcal{L}} = \frac{1}{M} \sum_{j=N+1}^{N+M} p_{ij}^{(t)} (\hat{\boldsymbol{\mu}}_i^{(t)} - \mathbf{x}_j)$$

931 where $\hat{p}_{ij}^{(k,t)}$ is given by

$$p_{ij}^{(t)} = \sum_{\tau=0}^t \exp\left(-\frac{1}{2} \|\hat{\boldsymbol{\mu}}_i^{(\tau)} - \mathbf{x}_j\|_{\Sigma^{-1}}^2 + \beta\tau\right) \Bigg/ \sum_{\tau=0}^t \sum_{c=1}^C \exp\left(-\frac{1}{2} \|\hat{\boldsymbol{\mu}}_c^{(\tau)} - \mathbf{x}_j\|_{\Sigma^{-1}}^2 + \beta\tau\right).$$

932 By choosing $\beta \rightarrow \infty$, we further have

$$p_{ij}^{(t)} = \exp\left(-\frac{1}{2} \|\hat{\boldsymbol{\mu}}_i^{(t)} - \mathbf{x}_j\|_{\Sigma^{-1}}^2\right) \Bigg/ \sum_{c=1}^C \exp\left(-\frac{1}{2} \|\hat{\boldsymbol{\mu}}_c^{(t)} - \mathbf{x}_j\|_{\Sigma^{-1}}^2\right).$$

933 where the samples $\{\mathbf{x}_j\}_{j \geq N+1}$ are drawn from a Gaussian mixture distribution.

934 Therefore, given $\hat{\boldsymbol{\mu}}_{ij}^{(t)}$, the random variable $p_{ij}^{(t)} (\hat{\boldsymbol{\mu}}_i^{(t)} - \mathbf{x}_j)$ admits a sub-Gaussian tail bound since
935 \mathbf{x}_j are Gaussian random vectors and $p_{ij}^{(t)} (\hat{\boldsymbol{\mu}}_i^{(t)} - \mathbf{x}_j)$ is Lipschitz continuous over \mathbf{x}_j .

936 Then, by the Bernstein's inequality, for any fixed $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have

$$\begin{aligned} \left\|\nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \hat{\mathcal{L}} - \nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \mathcal{L}\right\| &= \left\|\frac{1}{M} \sum_{j=N+1}^{N+M} p_{ij}^{(t)} (\hat{\boldsymbol{\mu}}_i^{(t)} - \mathbf{x}_j) - \mathbb{E}_{\mathbf{x}_j} \left[p_{ij}^{(t)} (\hat{\boldsymbol{\mu}}_i^{(t)} - \mathbf{x}_j)\right]\right\| \\ &\leq \frac{c_4}{\sqrt{M}} \sqrt{\log\left(\frac{2}{\delta}\right)}, \end{aligned}$$

937 where $c_4 > 0$ is some absolute constant.

938 By choosing $\delta = \exp(-\sqrt{M})$, we obtain that with probability at least $1 - \exp(-\sqrt{M})$,

$$\left\|\nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \hat{\mathcal{L}} - \nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \mathcal{L}\right\| \leq c' M^{-\frac{1}{4}}.$$

939 for another constant $c' > 0$. This completes the proof of the first inequality.

940 Next, we show that $\|\nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \hat{\mathcal{L}}\|$ itself is bounded with high probability.

941 Consequently,

$$\begin{aligned}
\|\nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \hat{\mathcal{L}}\| &= \left\| \frac{1}{M} \sum_{j=N+1}^{N+M} p_{ij}^{(t)} (\hat{\boldsymbol{\mu}}_i^{(t)} - \mathbf{x}_j) \right\| \\
&\leq \frac{1}{M} \sum_{j=N+1}^{N+M} \left\| p_{ij}^{(t)} (\hat{\boldsymbol{\mu}}_i^{(t)} - \mathbf{x}_j) \right\| \\
&\stackrel{(a)}{\leq} \frac{1}{M} \sum_{j \geq N+1} \left\| \hat{\boldsymbol{\mu}}_i^{(t)} - \mathbf{x}_j \right\| \\
&\leq \frac{1}{M} \sum_{j \geq N+1} (\|\hat{\boldsymbol{\mu}}_i^{(t)}\| + \|\mathbf{x}_j\|) \\
&= \|\hat{\boldsymbol{\mu}}_i^{(t)}\| + \frac{1}{M} \sum_{j \geq N+1} \|\mathbf{x}_j\|, \tag{B.14}
\end{aligned}$$

942 where inequality (a) holds since $p_{ij}^{(t)} \leq 1$. Note that

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_i^{(t)} &= \hat{\boldsymbol{\mu}}_i^{(t-1)} - \frac{\eta^{(t-1)}}{M} \sum_{j=N+1}^{N+M} p_{ij}^{(t-1)} (\hat{\boldsymbol{\mu}}_i^{(t-1)} - \mathbf{x}_j) \\
&= \left(1 - \frac{\eta^{(t-1)}}{M}\right) \hat{\boldsymbol{\mu}}_i^{(t-1)} + \frac{\eta^{(t-1)}}{M} \sum_{j=N+1}^{N+M} p_{i,j}^{(t-1)} \mathbf{x}_j.
\end{aligned}$$

943 Therefore, we have

$$\begin{aligned}
\|\hat{\boldsymbol{\mu}}_i^{(t)}\| &\leq \|\hat{\boldsymbol{\mu}}_i^{(t-1)}\| + \frac{1}{M} \sum_{j \geq N+1} \|\mathbf{x}_j\| \\
&\leq \|\hat{\boldsymbol{\mu}}_i^{(1)}\| + \frac{t-1}{M} \sum_{j \geq N+1} \|\mathbf{x}_j\|
\end{aligned}$$

944 Combining with Equation (B.14), we have

$$\|\nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \hat{\mathcal{L}}\| \leq \|\hat{\boldsymbol{\mu}}_i^{(1)}\| + \frac{t}{M} \sum_{j \geq N+1} \|\mathbf{x}_j\|$$

945 Applying the Bernstein's inequality, with probability at least $1 - \exp(-\sqrt{M})$, we have

$$\frac{1}{M} \sum_{j \geq N+1} \|\mathbf{x}_j\| \leq \frac{1}{C} \sum_{i=1}^C \boldsymbol{\mu}_i + c_5 M^{-\frac{1}{4}},$$

946 where c_5 is a positive constant.

947 Therefore, for any $t \leq T$ where T is total number of CoT steps, we have

$$\|\nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \hat{\mathcal{L}}\| \leq \|\hat{\boldsymbol{\mu}}_i^{(1)}\| + \frac{T}{C} \sum_{i=1}^C \boldsymbol{\mu}_i + c_5 t M^{-\frac{1}{4}},$$

948 which implies

$$\|\nabla_{\hat{\boldsymbol{\mu}}_i^{(t)}} \hat{\mathcal{L}}\|^2 \leq c_2 + c_3 M^{-\frac{1}{2}},$$

949 where c_2 and c_3 are positive constants depends on T , $\|\hat{\boldsymbol{\mu}}_i^{(1)}\|$ and $\frac{T}{C} \sum_{i=1}^C \boldsymbol{\mu}_i$. The proof is thus
950 complete. \square

951 **Step 3: Finally, we show the convergence of the class mean estimation error.**

952 Expanding the squared error $\|\hat{\mu}_i^{(t+1)} - \mu_i\|^2$ gives

$$\begin{aligned} \|\hat{\mu}_i^{(t+1)} - \mu_i\|^2 &= \|\hat{\mu}_i^{(t)} - \mu_i\|^2 + 2\eta^{(t)} \langle \hat{\mu}_i^{(t)} - \mu_i, \nabla_{\hat{\mu}_i^{(t)}} \hat{\mathcal{L}} \rangle + (\eta^{(t)})^2 \|\nabla_{\hat{\mu}_i^{(t)}} \hat{\mathcal{L}}\|^2 \\ &\leq \|\hat{\mu}_i^{(t)} - \mu_i\|^2 + 2\eta^{(t)} \langle \hat{\mu}_i^{(t)} - \mu_i, \nabla_{\hat{\mu}_i^{(t)}} \mathcal{L} \rangle + 2\eta^{(t)} \|\nabla \mathcal{L} - \nabla \hat{\mathcal{L}}\| \\ &\quad + (\eta^{(t)})^2 \|\nabla_{\hat{\mu}_i^{(t)}} \hat{\mathcal{L}}\|^2. \end{aligned} \quad (\text{B.15})$$

953 Denote $\hat{\mu}^{(t)}$ and μ as the vectors obtained by stacking $\{\hat{\mu}_i^{(t)}\}_{i=1}^C$ and $\{\mu_i\}_{i=1}^C$, respectively. There-
954 fore, we have

$$\|\hat{\mu}^{(t+1)} - \mu\|^2 \leq \|\hat{\mu}^{(t)} - \mu\|^2 + 2\eta^{(t)} \langle \hat{\mu}^{(t)} - \mu, \nabla_{\hat{\mu}^{(t)}} \mathcal{L} \rangle + 2\eta^{(t)} \|\nabla \mathcal{L} - \nabla \hat{\mathcal{L}}\| + (\eta^{(t)})^2 \|\nabla_{\hat{\mu}^{(t)}} \hat{\mathcal{L}}\|^2.$$

955 To control the inner product term $\langle \hat{\mu}^{(t)} - \mu, \nabla_{\hat{\mu}^{(t)}} \mathcal{L} \rangle$, we perform a first-order Taylor expansion of
956 $\nabla_{\hat{\mu}^{(t)}} \mathcal{L}$ around μ as

$$\begin{aligned} \nabla_{\hat{\mu}^{(t)}} \mathcal{L} &= \nabla_{\mu} \mathcal{L} + (\nabla_{\mu}^2 \mathcal{L}) (\hat{\mu}^{(t)} - \mu) + \mathbf{R}(\hat{\mu}^{(t)}, \mu) \\ &\stackrel{(a)}{=} (\nabla_{\mu}^2 \mathcal{L}) (\hat{\mu}^{(t)} - \mu) + \mathbf{R}(\hat{\mu}^{(t)}, \mu), \end{aligned}$$

957 where equality (a) holds since μ is the global minimizer of \mathcal{L} and \mathcal{L} is differentiable on \mathbb{R}^d , thus
958 $\nabla_{\mu} \mathcal{L} = 0$, and $\mathbf{R}(\hat{\mu}^{(t)}, \mu)$ is the remainder term.

959 For the remainder term, we have

$$\begin{aligned} &\langle \mathbf{R}(\hat{\mu}^{(t)}, \mu), \hat{\mu}^{(t)} - \mu \rangle \\ &= \int_0^1 (\hat{\mu}^{(t)} - \mu)^\top \left(\nabla_{\mu + \xi(\hat{\mu}^{(t)} - \mu)}^2 \mathcal{L} - \nabla_{\mu}^2 \mathcal{L} \right) (\hat{\mu}^{(t)} - \mu) d\xi \\ &\leq \int_0^1 \left\| \nabla_{\mu + \xi(\hat{\mu}^{(t)} - \mu)}^2 \mathcal{L} - \nabla_{\mu}^2 \mathcal{L} \right\| \|\hat{\mu}^{(t)} - \mu\|^2 d\xi \\ &\stackrel{(b)}{\leq} \int_0^1 L \xi \|\hat{\mu}^{(t)} - \mu\|^3 d\xi = L \|\hat{\mu}^{(t)} - \mu\|^3, \end{aligned}$$

960 where Inequality (b) follows from the fact that $\nabla^2 \mathcal{L}$ is twice continuously differentiable, its Jacobian
961 is Lipchitz continuous in a neighborhood of μ , and L is the Lipchitz constant.

962 Therefore, there exists a constant $\lambda > 0$ such that

$$\begin{aligned} &\|\hat{\mu}^{(t)} - \mu\|^2 + \langle \hat{\mu}^{(t)} - \mu, 2\eta^{(t)} \nabla_{\mu} \mathcal{L}^{(t)} \rangle \\ &\leq \|\hat{\mu}^{(t)} - \mu\|^2 + 2\eta^{(t)} (\hat{\mu}^{(t)} - \mu)^\top \nabla_{\mu}^2 \mathcal{L} (\hat{\mu}^{(t)} - \mu) + 2\eta^{(t)} L \|\hat{\mu}^{(t)} - \mu\|^3 \\ &\stackrel{(c)}{\leq} \left(1 - 2\eta^{(t)} \lambda \right) \|\hat{\mu}^{(t)} - \mu\|^2 + 2\eta^{(t)} L \|\hat{\mu}^{(t)} - \mu\|^3, \end{aligned} \quad (\text{B.16})$$

963 where Inequality (c) follows from Lemma 2 which proves $\nabla_{\mu}^2 \mathcal{L}$ is negative definite.

964 Meanwhile, Lemma 3 ensures with probability at least $1 - \exp(-\sqrt{M})$,

$$\eta^{(t)} \|\nabla \mathcal{L} - \nabla \hat{\mathcal{L}}\| \leq c_1 \eta^{(t)} M^{-\frac{1}{4}}, \quad (\text{B.17})$$

$$(\eta^{(t)})^2 \|\nabla_{\hat{\mu}^{(t)}} \hat{\mathcal{L}}\|^2 \leq c_2 (\eta^{(t)})^2 M^{-\frac{1}{2}} + c_3 (\eta^{(t)})^2. \quad (\text{B.18})$$

965 Substituting (B.16), (B.17), and (B.18) into (B.15) then yields the one-step error recursion

$$\begin{aligned} \|\hat{\mu}^{(t+1)} - \mu\|^2 &\leq \left(1 - 2\eta^{(t)} \lambda \right) \|\hat{\mu}^{(t)} - \mu\|^2 + 2\eta^{(t)} L \|\hat{\mu}^{(t)} - \mu\|^3 \\ &\quad + c_1 \eta^{(t)} M^{-\frac{1}{4}} + c_2 (\eta^{(t)})^2 M^{-\frac{1}{2}} + c_3 (\eta^{(t)})^2. \end{aligned} \quad (\text{B.19})$$

966 Next, we aim prove $\|\hat{\boldsymbol{\mu}}^{(t)} - \boldsymbol{\mu}\|^2 \leq K/t$ for a positive constant K by induction.

967 Let $\eta^{(t)} = \frac{\alpha}{t}$ and $M^{(t)} = t^p$ for some $p > 4$. First, assume $\|\hat{\boldsymbol{\mu}}^{(t)} - \boldsymbol{\mu}\|^2 \leq K/t$ for a fixed $t \geq 1$.
 968 From Equation (B.19), we note that there exists a constant $c_4 > 0$ such that

$$\begin{aligned} \|\hat{\boldsymbol{\mu}}^{(t+1)} - \boldsymbol{\mu}\|^2 &\leq \left(1 - 2\frac{\alpha\lambda}{t}\right) \|\hat{\boldsymbol{\mu}}^{(t)} - \boldsymbol{\mu}\|^2 + c_3 \frac{\alpha^2}{t^2} + 2\frac{\alpha L}{t} \|\hat{\boldsymbol{\mu}}^{(t)} - \boldsymbol{\mu}\|^3 + c_4 \alpha t^{-(1+\frac{p}{4})} \\ &\leq \left(1 - 2\frac{\alpha\lambda}{t}\right) \frac{K}{t} + 2\frac{\alpha L}{t} \left(\frac{K}{t}\right)^{\frac{3}{2}} + c_4 \alpha t^{-(1+\frac{p}{4})} + c_3 \frac{\alpha^2}{t^2}. \end{aligned}$$

969 Therefore, we have

$$\begin{aligned} \|\hat{\boldsymbol{\mu}}^{(t+1)} - \boldsymbol{\mu}\|^2 - \frac{K}{t+1} &\leq \left(1 - 2\frac{\alpha\lambda}{t}\right) \frac{K}{t} + 2\frac{\alpha L_s}{t} \left(\frac{K}{t}\right)^{\frac{3}{2}} + c_4 \alpha t^{-(1+\frac{p}{4})} + c_3 \frac{\alpha^2}{t^2} - \frac{K}{t} + \frac{K}{t^2} \\ &= (-2\alpha\lambda + 1) \frac{K}{t^2} + 2\frac{\alpha L}{t} \left(\frac{K}{t}\right)^{\frac{3}{2}} + c_4 \alpha t^{-(1+\frac{p}{4})} + c_3 \frac{\alpha^2}{t^2}. \end{aligned} \quad (\text{B.20})$$

970 By choosing $\alpha \geq 1/\lambda$, $K \geq \max\{3c_3\alpha^2, 3c_4\alpha\}$ and $t \geq 36\alpha^2 L^2 K$, we have

$$(-2\alpha\lambda + 1) \frac{K}{t^2} \leq -\frac{K}{t^2}, \quad 2\frac{\alpha L}{t} \left(\frac{K}{t}\right)^{\frac{3}{2}} \leq \frac{K}{3t^2}, \quad c_4 \alpha t^{-(1+\frac{p}{4})} \leq \frac{K}{3t^2}, \quad c_3 \frac{\alpha^2}{t^2} \leq \frac{K}{3t^2}. \quad (\text{B.21})$$

971 Note that we assume $\|\hat{\boldsymbol{\mu}}^{(t)} - \boldsymbol{\mu}\|^2 \leq K/t$. Therefore, by substituting Equation (B.21) into Equa-
 972 tion (B.20), we have

$$\|\hat{\boldsymbol{\mu}}^{(t+1)} - \boldsymbol{\mu}\|^2 - \frac{K}{t+1} \leq 0, \quad \forall t \geq 36\alpha^2 L^2 K.$$

973 Select $T' = 36\alpha^2 L^2 K$ and let $\eta^{(t)} = \alpha/(t + T')$. If $\|\hat{\boldsymbol{\mu}}^{(t)} - \boldsymbol{\mu}\|^2 \leq K/(T' + t)$, it must have

$$\|\hat{\boldsymbol{\mu}}^{(t+1)} - \boldsymbol{\mu}\|^2 \leq \frac{K}{T' + t + 1}, \quad \forall t \geq 1.$$

974 Recall Lemma 1 indicates that, with probability at least $1 - \exp(-cNK/T')$, for some constant c ,
 975 it holds that $\|\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}\| \leq K/T'$. Therefore, for any fixed $\epsilon \in [0, 1]$, if

$$\begin{aligned} N &\geq 36\alpha^2 L^2 \log 1/\epsilon, \\ M &\geq \max\{(T')^4, \log^2 1/\epsilon\}, \\ t &\geq \sqrt[4]{M}, \end{aligned}$$

976 with probability at least $1 - \epsilon$, the estimation error is upper bounded by

$$\|\hat{\boldsymbol{\mu}}^{(t)} - \boldsymbol{\mu}\|^2 \leq c \frac{\log(1/\epsilon)}{N \sqrt[4]{M}},$$

977 where c is a positive constant. This completes the proof of Theorem 4.2.

978 B.3 Proof of Corollary 4.1

979 First, we restate the corollary below.

980 **Corollary B.1** (Restatement of Corollary 4.1). *Let $\hat{\mathbf{y}}_j$ be the predicted label for \mathbf{x}_j according to*
 981 *Equation (3.5). Let \mathcal{R}^* be the prediction error under the Bayes-optimal classifier with known class*
 982 *mean vectors $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_C$. Then, under the same conditions as described in Theorem 4.2, we have*

$$\mathbb{P}[\hat{\mathbf{y}}_j \neq \mathbf{y} | \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_C] - \mathcal{R}^* \leq \mathcal{O}(1/\sqrt{N \text{poly}(M)}).$$

983 *Proof.* First, we define $\Delta = \|\widehat{\mathbf{M}} - \mathbf{M}\|_F$, define \hat{g} as the Bayes-optimal classifier given estimated
 984 class means $\widehat{\mathbf{M}}$ and define g as the Bayes-optimal classifier given ground truth class means \mathbf{M}

Suppose $\hat{g}(\mathbf{x}) \neq g(\mathbf{x})$. Then, there exist indices $i \neq k$ such that $g(\mathbf{x}) = i$ and $\hat{g}(\mathbf{x}) = k$. Because $g(\mathbf{x}) = i$ is Bayes-optimal, we have

$$\|\mathbf{x} - \boldsymbol{\mu}_i\| \leq \|\mathbf{x} - \boldsymbol{\mu}_k\| \text{ and } \|\mathbf{x} - \hat{\boldsymbol{\mu}}_k\| \leq \|\mathbf{x} - \hat{\boldsymbol{\mu}}_i\|.$$

Denote $\zeta = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\|$. Therefore, from the geometric observation, the misclassification only happens when \mathbf{x} is in the dihedral cone with angle θ , where $\tan(\theta) = \Delta/\zeta$ (Diakonikolas et al., 2018). Thus, the probability for misclassification is upper bounded

$$\mathbb{P}[\hat{g}(\mathbf{x}) \neq g(\mathbf{x})] \leq c'\theta,$$

for a positive constant c' . Since $\mathbb{P}[\hat{\mathbf{y}}_j \neq \mathbf{y} | \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_C] - \mathcal{R}^* = \mathbb{P}[\hat{g}(\mathbf{x}) \neq g(\mathbf{x})]$ and from Theorem 4.2 we have $\Delta \leq c' \sqrt{1/N} \sqrt[4]{M}$ for positive constant c' , the proof is thus complete. \square

C Proof of Training Dynamics

First, we restate Theorem 5.1 below.

Theorem C.1 (Restatement of Theorem 5.1). *Let $\{\mathbf{Q}^{(k)}, \mathbf{K}^{(k)}, \mathbf{V}^{(k)}\}_{k \geq 0}$ be the parameters of the first attention layer of the transformer after applying k iterations of gradient descent on the population loss defined in Equation (5.2) with step size $\eta^{(k)} = 1/k$. Then, with the initialization specified in Assumption 1, we have*

$$\|\mathbf{W}^{(k)} - \boldsymbol{\Sigma}^{-1}\|_F^2 \leq \frac{c}{\sqrt{k}},$$

for some positive constant c , while the other parameters in $\mathbf{Q}^{(0)}$, $\mathbf{K}^{(0)}$ and $\mathbf{V}^{(0)}$ remain unchanged.

We assume ground truth means are IID sampled from standard Gaussian distribution: $\boldsymbol{\mu}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for all i . Then, we introduce the following quantities: 1) the formulation of class mean estimations given by the transformer during teacher-forcing training; 2) the reference class mean estimations given by the reference policy; and 3) the formulation of the gradient of the teacher-forcing training loss.

At the k -th GD iteration during training, we denote the set of reference class mean estimations as $\boldsymbol{\mu}_{\text{ref},1}^{(k,t)}, \dots, \boldsymbol{\mu}_{\text{ref},C}^{(k,t)}$ for the CoT steps $t \in [T]$. Given the reference class mean estimations, the estimation given by the transformer throughout teacher-forcing satisfies

$$\hat{\boldsymbol{\mu}}_i^{(k,t+1)} = \boldsymbol{\mu}_{\text{ref},i}^{(k,t)} - \frac{\eta^{(t)}}{M} \sum_{j=N+1}^{N+M} \hat{p}_{ij}^{(k,t)} (\boldsymbol{\mu}_{\text{ref},i}^{(k,t)} - \mathbf{x}_j)$$

where $\hat{p}_{ij}^{(k,t)}$ is given by

$$\hat{p}_{ij}^{(k,t)} = \frac{\sum_{\tau=0}^t \exp\left(-\frac{w}{2} \|\hat{\boldsymbol{\mu}}_i^{(\tau)}\|^2 + \mathbf{x}_j^\top \mathbf{W}^{(k)} \hat{\boldsymbol{\mu}}_i^{(\tau)} + \beta\tau\right)}{\sum_{\tau=0}^t \sum_{c=1}^C \exp\left(-\frac{w}{2} \|\hat{\boldsymbol{\mu}}_c^{(\tau)}\|^2 + \mathbf{x}_j^\top \mathbf{W}^{(k)} \hat{\boldsymbol{\mu}}_c^{(\tau)} + \beta\tau\right)}.$$

By choosing $\beta \rightarrow \infty$, we further have

$$\hat{p}_{ij}^{(k,t)} = \frac{\exp\left(-\frac{w}{2} \|\hat{\boldsymbol{\mu}}_i^{(\tau)}\|^2 + \mathbf{x}_j^\top \mathbf{W}^{(k)} \hat{\boldsymbol{\mu}}_i^{(\tau)}\right)}{\sum_{c=1}^C \exp\left(-\frac{w}{2} \|\hat{\boldsymbol{\mu}}_c^{(\tau)}\|^2 + \mathbf{x}_j^\top \mathbf{W}^{(k)} \hat{\boldsymbol{\mu}}_c^{(\tau)}\right)}.$$

We choose the reference policy under which

$$\boldsymbol{\mu}_{\text{ref},i}^{(k,t+1)} = \boldsymbol{\mu}_{\text{ref},i}^{(k,t)} - \frac{\eta^{(t)}}{M} \sum_{j=N+1}^{N+M} p_{ij}^{(k,t)} (\boldsymbol{\mu}_{\text{ref},i}^{(k,t)} - \mathbf{x}_j),$$

with

$$p_{ij}^{(k,t)} = \frac{\exp\left(-\frac{1}{2} \|\boldsymbol{\mu}_{\text{ref},i}^{(k,t)} - \mathbf{x}_j\|_{\boldsymbol{\Sigma}^{-1}}^2\right)}{\sum_{c=1}^C \exp\left(-\frac{1}{2} \|\boldsymbol{\mu}_{\text{ref},c}^{(k,t)} - \mathbf{x}_j\|_{\boldsymbol{\Sigma}^{-1}}^2\right)}.$$

1011 To simplify the notation, when there is no ambiguity, we drop the superscript (k) for the training
 1012 iteration. Denote $\hat{\mathbf{q}}_j^{(t)} = [\hat{p}_{1j}^{(t)} \cdots \hat{p}_{Cj}^{(t)}]$ and $\mathbf{q}_j^{(t)} = [p_{1j}^{(t)} \cdots p_{Cj}^{(t)}]$. At the k -th training iteration, the
 1013 CoT training loss with teacher-forcing is

$$\begin{aligned}\hat{\mathcal{L}}_{\text{CoT-train}}(\boldsymbol{\Theta}; \mathcal{I}_{\mathbf{M}}) &= \frac{1}{T} \sum_{t=1}^T \sum_{j=N+1}^{N+M} \text{CE} \left(\mathbf{q}_j^{(t)}, [\text{TF}_{\boldsymbol{\Theta}}(\mathbf{H}_{\text{ref}}^{(t-1)})]_{2d+2c+1:2d+3c, N+j} \right) \\ &= \frac{1}{T} \sum_{t=1}^T \sum_{j=N+1}^{N+M} \text{CE} \left(\mathbf{q}_j^{(t)}, \hat{\mathbf{q}}_j^{(t)} \right),\end{aligned}$$

1014 where CE is the cross entropy loss function.

1015 Define $s_{ij}^{(t)} = -\frac{1}{2} \|\hat{\boldsymbol{\mu}}_i^{(\tau)}\|^2 + \mathbf{x}_j^\top \mathbf{W}^{(k)} \hat{\boldsymbol{\mu}}_i^{(\tau)}$ and $\mathbf{s}_j^{(t)} = [s_{1j}^{(t)} \cdots s_{Cj}^{(t)}]$. Note that the derivative can be
 1016 written as

$$\frac{\partial \text{CE} \left(\mathbf{q}_j^{(t)}, \hat{\mathbf{q}}_j^{(t)} \right)}{\partial s_{ij}^{(t)}} = \frac{\partial \frac{\exp(s_{ij}^{(t)})}{\sum_{k=1}^C \exp(s_{kj}^{(t)})}}{\partial s_{ij}^{(t)}} = \hat{p}_{ij}^{(t)} - p_{ij}^{(t)}.$$

1017 Furthermore, since $\partial s_{ij}^{(t)} / \partial \mathbf{W}_{ab} = \mathbf{M}_{a,i} \mathbf{x}_{jb}$, where $a, b \in [d]$, by the chain rule, we have

$$\frac{\partial \text{CE} \left(\mathbf{q}_j^{(t)}, \hat{\mathbf{q}}_j^{(t)} \right)}{\partial \mathbf{W}_{ab}} = \frac{\partial \text{CE} \left(\mathbf{q}_j^{(t)}, \hat{\mathbf{q}}_j^{(t)} \right)}{\partial \mathbf{s}_j^{(t)}} \frac{\partial \mathbf{s}_j^{(t)}}{\partial \mathbf{W}_{ab}} = \sum_i (\hat{p}_{ij}^{(t)} - p_{ij}^{(t)}) \mathbf{M}_{a,i} \mathbf{x}_{jb}. \quad (\text{C.1})$$

1018 Based on the notations, we will prove Theorem 5.1 as follows.

1019 **Step 1: Given the gradient of the cross entropy loss with respect to the learnable parameter**
 1020 **matrix \mathbf{W} , our first step is to provide a decomposition of the gradient so that it becomes**
 1021 **analytically tractable.**

In the matrix form, Equation (C.1) can be written as

$$\nabla_{\mathbf{W}} \text{CE} \left(\mathbf{q}_j^{(t)}, \hat{\mathbf{q}}_j^{(t)} \right) = \mathbf{M}(\hat{\mathbf{q}}_j^{(t)} - \mathbf{q}_j^{(t)}) \mathbf{x}_j^T.$$

1022 By the Stein's lemma, we have

$$\begin{aligned}\mathbb{E}[\mathbf{M}(\hat{\mathbf{q}}_j^{(t)} - \mathbf{q}_j^{(t)}) \mathbf{x}_j^T] &= \mathbb{E}_{\mathbf{M}} \left[\mathbf{M} \left[\mathbb{E}_{\mathbf{x}_j} [\hat{\mathbf{q}}_j^{(t)} - \mathbf{q}_j^{(t)}] \mathbb{E}[\mathbf{x}_j^T] + \mathbb{E}_{\mathbf{x}_j} [\nabla \hat{\mathbf{q}}_j^{(t)} - \nabla \mathbf{q}_j^{(t)}] \boldsymbol{\Sigma} \right] \right] \\ &= \underbrace{\mathbb{E}_{\mathbf{M}} \left[\mathbf{M} \mathbb{E}_{\mathbf{x}_j} [\hat{\mathbf{q}}_j^{(t)} - \mathbf{q}_j^{(t)}] \mathbb{E}[\mathbf{x}_j^T] \right]}_{\mathcal{A}_1} + \underbrace{\mathbb{E}_{\mathbf{M}} \left[\mathbf{M} \mathbb{E}_{\mathbf{x}_j} [\nabla \hat{\mathbf{q}}_j^{(t)} - \nabla \mathbf{q}_j^{(t)}] \boldsymbol{\Sigma} \right]}_{\mathcal{A}_2}.\end{aligned}$$

1023 **Step 2: Based on the decomposition, we aim to show that $\mathcal{A}_1 = 0$.**

1024 We note that when taking the expectation over the labeled dataset, we have

$$\mathbb{E} \left[\frac{C}{N} \sum_{j \in [N]} \mathbf{x}_j \cdot (\mathbf{e}_i^\top \mathbf{y}_j) \right] = \boldsymbol{\mu}_i.$$

1025 Therefore, $\boldsymbol{\mu}_{\text{ref},i}^0 = \boldsymbol{\mu}_i$. When the reference class mean estimations are generated by gradient
 1026 descent over the population loss, we have $\boldsymbol{\mu}_{\text{ref},i}^{(t)} = \boldsymbol{\mu}_i$ for any $i \in [C]$ and $t \in [T]$ the gradient

1027 over the population loss is zero:

$$\begin{aligned}
& \mathbb{E}_{\mathbf{x}} \left[\frac{\exp\left(-\frac{1}{2}\|\boldsymbol{\mu}_{\text{ref},i}^{(t)} - \mathbf{x}_j\|_{\Sigma^{-1}}^2\right)}{\sum_{c=1}^C \exp\left(-\frac{1}{2}\|\boldsymbol{\mu}_{\text{ref},c}^{(t)} - \mathbf{x}_j\|_{\Sigma^{-1}}^2\right)} (\boldsymbol{\mu}_{\text{ref},i}^{(t)} - \mathbf{x}_j) \right] \\
&= \int_{\mathbb{R}^d} \frac{\exp\left(-\frac{1}{2}\|\boldsymbol{\mu}_{\text{ref},i}^{(t)} - \mathbf{x}\|_{\Sigma^{-1}}^2\right)}{\sum_{c=1}^C \exp\left(-\frac{1}{2}\|\boldsymbol{\mu}_{\text{ref},c}^{(t)} - \mathbf{x}\|_{\Sigma^{-1}}^2\right)} \left[\sum_{k=1}^C \frac{1}{C} \varphi_k(\mathbf{x}) \right] (\boldsymbol{\mu}_{\text{ref},i}^{(t)} - \mathbf{x}) d\mathbf{x}, \\
&\stackrel{(a)}{=} \int_{\mathbb{R}^d} \frac{1}{C} \varphi_i(\mathbf{x}) (\boldsymbol{\mu}_i - \mathbf{x}) d\mathbf{x} = 0,
\end{aligned}$$

1028 where $\varphi_i(\mathbf{x})$ is the pdf of Gaussian distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix Σ , and equal-
1029 ity (a) holds since $\boldsymbol{\mu}_{\text{ref},i}^{(k,t)} = \boldsymbol{\mu}_i$. Given the above-discussed property of the reference class mean
1030 estimations, for $\mathbb{E}_{\mathbf{x}_j}[\hat{\mathbf{q}}_j^{(t)} - \mathbf{q}_j^{(t)}]$ in \mathcal{A}_1 , it is obvious that $\mathbb{E}_{\mathbf{x}_j}[\mathbf{q}_j^{(t)}] = 1/C$. For $\mathbb{E}_{\mathbf{x}_j}[\hat{\mathbf{q}}_j^{(t)}]$, we let
1031 $\mathbf{W}^{(0)}$ initialize form a isotropic matrix $w\mathbf{I}$, and we assume at training iteration step t , it preserve the
1032 isotropic as $w^{(t)}$. Therefore, since the ground truth Σ is an isotropic matrix, the temperature acts
1033 identically on all classes:

$$\mathbb{E} \left[\frac{\exp\left(-\frac{\alpha}{2}\|\hat{\boldsymbol{\mu}}_i^{(\tau)} - \mathbf{x}_j\|^2\right)}{\sum_{c=1}^C \exp\left(-\frac{\alpha}{2}\|\boldsymbol{\mu}_c - \mathbf{x}_j\|^2\right)} \right] = \mathbb{E} \left[\frac{\exp\left(-\frac{1}{2}\|\hat{\boldsymbol{\mu}}_i^{(\tau)} - \mathbf{x}_j\|_{\Sigma^{-1}}^2\right)}{\sum_{c=1}^C \exp\left(-\frac{1}{2}\|\boldsymbol{\mu}_c - \mathbf{x}_j\|_{\Sigma^{-1}}^2\right)} \right].$$

1034 Therefore, we have $\mathbb{E}_{\mathbf{x}_j}[\hat{\mathbf{p}}_j^{(t)} - \mathbf{p}_j^{(t)}] = 0$, which gives $\mathcal{A}_1 = 0$.

1035 **Step 3: Finally, we analyze the properties of \mathcal{A}_2 , and obtain the final results afterwards.** We
1036 will prove that $\mathbf{W}^{(t)}$ preserves isotropic by induction. Note that we assume training iteration step t ,
1037 $\mathbf{W}^{(t)}$ is isotropic. Besides, we initialize $\mathbf{W}^{(0)}$ as an isotropic matrix.

1038 \mathcal{A}_2 can be rewritten as

$$\begin{aligned}
\mathcal{A}_2 &= \mathbb{E}_{\mathbf{M}} \left[\mathbf{M} \mathbb{E}_{\mathbf{x}_j} [\nabla \hat{\mathbf{q}}_j^{(t)} - \nabla \mathbf{q}_j^{(t)}] \Sigma \right] \\
&= \mathbb{E}_{\mathbf{M}} \left[\mathbf{M} \left(\left(\text{diag}(\mathbb{E}[\hat{\mathbf{q}}_j^{(t)}]) - \mathbb{E}_{\mathbf{x}}[\hat{\mathbf{q}}_j^{(t)} (\hat{\mathbf{q}}_j^{(t)})^\top] \right) \mathbf{M}^\top \mathbf{W}^{(k)} \Sigma^{-1} - \left(\text{diag}(\mathbf{q}_j^{(t)}) - \mathbb{E}_{\mathbf{x}}[\mathbf{q}_j^{(t)} (\mathbf{q}_j^{(t)})^\top] \right) \mathbf{M}^\top \right) \right].
\end{aligned}$$

1039 Because the class prior is uniform and the isotropic initialisation, we have

$$\mathbb{E}_{\mathbf{x}_j}[\hat{\mathbf{q}}_j^{(t)}] = \mathbb{E}_{\mathbf{x}_j}[\mathbf{q}_j^{(t)}] = \frac{1}{C} \mathbf{1}.$$

1040 Since each coordinate of $\hat{\mathbf{q}}_j^{(t)}$ (or $\mathbf{q}_j^{(t)}$) has the same marginal distribution and any two distinct
1041 coordinates have the same joint distribution, we have

$$\text{diag}(\mathbb{E}[\hat{\mathbf{p}}_j^{(t)}]) = \text{diag}(\mathbf{q}_j^{(t)}) = \frac{1}{C} \mathbf{I}, \quad \mathbb{E}_{\mathbf{x}}[\hat{\mathbf{q}}_j^{(t)} (\hat{\mathbf{q}}_j^{(t)})^\top] = \mathbb{E}_{\mathbf{x}}[\mathbf{q}_j^{(t)} (\mathbf{q}_j^{(t)})^\top] = \frac{1}{C^2} \mathbf{1} \mathbf{1}^\top.$$

1042 Therefore, we have

$$\mathcal{A}_2 = \mathbb{E}_{\mathbf{M}} \left[\mathbf{M} \left(\text{diag}(1/C) - \frac{1}{C^2} \mathbf{1} \mathbf{1}^\top \right) \mathbf{M}^\top \left(\mathbf{W}^{(k)} \Sigma^{-1} - \mathbf{I} \right) \right].$$

1043 Since all columns in \mathbf{M} are sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{W}^{(k)}$ is assumed to be an isotropic matrix,
1044 it's obvious that \mathcal{A}_2 is also an isotropic matrix. It follows that

$$\begin{aligned}
& \langle \mathbf{W}^{(k)} - \Sigma, \nabla_{\mathbf{W}} L_{\text{CoT}} \rangle \\
&= \mathbb{E}_{\mathbf{M}} \left[\text{trace} \left(\mathbf{M} \left(\text{diag}(1/C) - \frac{1}{C^2} \mathbf{1} \mathbf{1}^\top \right) \mathbf{M}^\top \left(\mathbf{W}^{(k)} - \Sigma \right) \Sigma^{-1} \left(\mathbf{W}^{(k)} - \Sigma \right)^\top \right) \right] \\
&\stackrel{(a)}{=} \frac{1}{\sigma^{-1}} \text{trace} \left(\mathbb{E}_{\mathbf{M}} \left[\frac{1}{C} \mathbf{M} \mathbf{M}^\top \right] \left(\mathbf{W}^{(k)} - \Sigma \right) \left(\mathbf{W}^{(k)} - \Sigma \right)^\top - \mathbb{E}_{\mathbf{M}} \left[\frac{1}{C^2} \mathbf{M} \mathbf{1} \mathbf{1}^\top \mathbf{M}^\top \right] \left(\mathbf{W}^{(k)} - \Sigma \right) \left(\mathbf{W}^{(k)} - \Sigma \right)^\top \right) \\
&= \frac{1}{\sigma^{-1}} \text{trace} \left(\left(\mathbf{W}^{(k)} - \Sigma \right) \left(\mathbf{W}^{(k)} - \Sigma \right)^\top - \frac{1}{C} \left(\mathbf{W}^{(k)} - \Sigma \right) \left(\mathbf{W}^{(k)} - \Sigma \right)^\top \right) \\
&= \frac{1}{\sigma^{-1}} \left(1 - \frac{1}{C} \right) \|\mathbf{W}^{(k)} - \Sigma\|_F^2.
\end{aligned}$$

1045 where equation (a) follows from the assumption that $\Sigma = \sigma^2 \mathbf{I}$.

1046 Set $G := \sup_k \|\nabla_{\mathbf{W}} L_{\text{CoT}}(\mathbf{W}^{(k)})\|_F$, and let $\gamma = (1 - 1/C)/\sigma^{-1}$. Then,

$$\|\mathbf{W}^{(k+1)} - \Sigma\|_F^2 \leq \|\mathbf{W}^{(k+1)} - \Sigma\|_F^2 - 2\gamma\eta^{(k)}\|\mathbf{W}^{(k)} - \Sigma\|_F^2 + (\eta^{(k)})^2 G^2.$$

1047 With the step size $\eta^{(k)} = 1/k$, this becomes

$$\|\mathbf{W}^{(k+1)} - \Sigma\|_F^2 \leq \left(1 - \frac{2\gamma}{k}\right) \|\mathbf{W}^{(k)} - \Sigma\|_F^2 + \frac{G^2}{k^2},$$

1048 which yields

$$\|\mathbf{W}^{(k)} - \Sigma\|_F^2 \leq \frac{\max\{\|\mathbf{W}^{(1)} - \Sigma\|_F^2, G^2/\gamma\}}{\sqrt{k}}.$$

1049 Let $c = \max\{\|\mathbf{W}^{(1)} - \Sigma\|_F^2, G^2/\gamma\}$, we have

$$\|\mathbf{W}^{(k)} - \Sigma\|_F^2 \leq \frac{c}{\sqrt{k}}.$$

1050 Thus, the proof is complete.

1051 D Auxiliary Lemmas

1052 **Lemma 4** (Stein’s Lemma). *Let $X \in \mathbb{R}^d$ be a random vector with*

$$X \sim \mathcal{N}(\mu, \Sigma),$$

1053 *where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ is a positive definite matrix. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ be a continuously*
 1054 *differentiable function such that*

$$\mathbb{E}[\|f(X)\|] < \infty \quad \text{and} \quad \mathbb{E}[\|\nabla f(X)\|_F] < \infty,$$

1055 *where $\|\cdot\|$ denotes the Euclidean norm in \mathbb{R}^k and $\|\cdot\|_F$ is the Frobenius norm. Then, the following*
 1056 *identity holds:*

$$\mathbb{E}[(X - \mu) f(X)^T] = \Sigma \mathbb{E}[\nabla f(X)],$$

1057 *where $\nabla f(X)$ is the $k \times d$ Jacobian matrix of f evaluated at X .*

1058 E Limitations

1059 Our analysis and experiments possess certain limitations. Below, we outline these limitations and
 1060 propose directions for future research.

1061 First, our analysis tracks parameter updates only in the *first* transformer layer, leaving all other
 1062 layers frozen. As a result, we cannot say how weights in non-linear hidden layers evolve under
 1063 teacher–forcing. To the best of our knowledge, the training dynamics of multi-layer transformer
 1064 with *non-linear* activation is still lacking investigation. A full, multi-layer treatment for end-to-end
 1065 training remains an open problem.

1066 Second, this paper is the first theoretical investigation of the influence of unlabeled data in in-context
 1067 learning, therefore, we restricted the experiments to a synthetic data set. However, whether the same
 1068 behavior emerges in real-world tasks, and how unlabeled examples influence in-context learning
 1069 for large, fully-trained transformers, is still unknown. Empirically understanding such impact is a
 1070 promising future direction.