## ON THE TRAINING CONVERGENCE OF TRANSFORM-ERS FOR IN-CONTEXT CLASSIFICATION

Anonymous authors

Paper under double-blind review

### ABSTRACT

While transformers have demonstrated impressive capacities for in-context learning (ICL) in practice, theoretical understanding of the underlying mechanism enabling transformers to perform ICL is still in its infant stage. This work aims to theoretically study the training dynamics of transformers for in-context classification tasks. We demonstrate that, for in-context classification of Gaussian mixtures under certain assumptions, a single-layer transformer trained via gradient descent converges to a globally optimal model at a linear rate. We further quantify the impact of the training and testing prompt lengths on the ICL inference error of the trained transformer. We show that when the lengths of training and testing prompts are sufficiently large, the prediction of the trained transformer approaches the Bayes-optimal classifier. Experimental results corroborate the theoretical findings.

### 1 INTRODUCTION

026 027

004

010 011

012

013

014

015

016

017

018

019

021

023

025

Large language models (LLMs) based on the transformer architecture (Vaswani et al., 2017) have demonstrated remarkable in-context learning (ICL) abilities (Brown et al., 2020). When given a prompt consisting of examples of a learning task, these models can learn to solve this task for new test examples without any parameter updating. This behavior has been empirically demonstrated in state-of-the-art models on real-world tasks (OpenAI, 2023; Touvron et al., 2023).

033 This impressive capacity of transformer-based models has inspired many recent works aiming to 034 understand the ICL abilities of transformers. A more comprehensive literature review can be found in Appendix B. Garg et al. (2022) was the first to study the ICL abilities of transformers for various function classes. They empirically showed that transformers can learn linear regression models in 036 context. Later on, a line of research was developed to theoretically explain how transformers perform 037 in-context linear regression. For example, Akyürek et al. (2022); Von Oswald et al. (2023); Bai et al. (2024); Fu et al. (2023); Giannou et al. (2024) showed by construction that, some specially-designed transformers can perform linear regression in context. Moreover, some recent works like Zhang 040 et al. (2023a); Huang et al. (2023); Chen et al. (2024) studied the training dynamics of a single-041 layer transformer for in-context linear regression. They proved the convergence of their single-layer 042 transformers and showed their trained transformer are able to perform linear regression in context. 043

Building on the earlier works that largely focus on linear regression problems, several recent pa-044 pers have started to investigate the ICL capabilities of transformers for non-linear problems such as classification. For instance, Bai et al. (2024) showed that, by construction, multi-layer transformers 046 can be approximately viewed as multiple steps of gradient descents for logistic regression. Gian-047 nou et al. (2024) further showcased that the constructed transformers can approximately perform 048 Newton's method for logistic regression. Recently, Li et al. (2024) studied the training dynamics of transformers for in-context binary classification. However, their analysis requires the data to be pairwise orthogonal and the possible distribution of their data is highly limited. The learning dy-051 namics of transformers for more general in-context classification problems is not well understood. Moreover, to the best of our knowledge, existing literature (Bai et al., 2024; Giannou et al., 2024; Li 052 et al., 2024) studying the in-context classification of transformers focus only on binary classification. How transformers perform in-context multi-class classification remains unexplored.

In this work, we study the learning dynamics of a singly-layer transformer for both in-context binary and multi-class classification of Gaussian mixtures, a fundamental problem in machine learning. Our main contributions can be summarized as follows:

To the best of our knowledge, we are the first to study the learning dynamics of transformers for in-context classification of Gaussian mixtures, and we are the first to prove the training convergence of transformers for in-context multi-class classification. We prove that with appropriately distributed training data (Assumptions 3.1, 4.1), a single-layer transformer trained via gradient descent will converge to its global minimizer at a linear rate (Theorems 3.1, 4.1) for both in-context binary or multi-class classification problems.

- Due to the high non-linearity of our loss function, we cannot directly find the closed-form expression of the global minimizer. Instead, we prove an important property that the global minimizer consists of a constant plus an error term that is induced by the finite training prompt length (N). We further show that the max norm of this error term is bounded, and converges to zero at a rate of O(1/N).
- With properly distributed test prompts (Assumptions 3.2, 4.2), we establish an upper bound of the inference error (defined in Equation (3)) of the trained transformer and quantify the impact of the training and testing prompt lengths on this error. We further prove that when the lengths of training prompts (N) and testing prompts (M) approach infinity, this error converges to zero at a rate of  $O(1/N + 1/\sqrt{M})$  (Theorems 3.2, 4.2), and the prediction of the trained transformer is Bayes-optimal, i.e., the optimal classifier given the data distribution.

### 2 PRELIMINARIES

074

075 076

084

085

087

088

090

**Notations.** We denote  $[n] = \{1, 2, ..., n\}$ . For a matrix  $A \in \mathbb{R}^{m \times n}$ , we denote its Frobenius norm as  $||A||_F$ , and its max norm as  $||A||_{\max} = \max_{i \in [m], j \in [n]} |A_{ij}|$ . We use  $A_{a,b}$  (or  $A_{ab}$ ) to represent the element of matrix A at the *a*-th row and *b*-th column, and use  $A_{a:c,b}$  to represent a vector of dimension c - a + 1 whose *i*-th element is  $A_{(a+i-1),b}$ . We denote the  $l_2$  norm of a vector as  $|| \cdot ||_2$ . We denote the all-zero vector of size *n* as  $0_n$  and the all-zero matrix of size  $m \times n$  as  $0_{m \times n}$ . We use  $\sigma(x) := 1/(1 + \exp(-x))$  to denote the sigmoid function. We define  $\operatorname{softmax}(\cdot) : \mathbb{R}^k \to (0, 1)^k$ , and its *i*-th element as  $\operatorname{softmax}(\cdot)_i$ , where  $\operatorname{softmax}(x)_i = \exp(x_i)/(\sum_{j=1}^k \exp(x_j))$ .

### 2.1 SINGLE-LAYER TRANSFORMER

Given an input embedding matrix  $E \in \mathbb{R}^{d_e \times d_n}$ , a single head self-attention module  $F_{SA}$  (Vaswani et al., 2017) with width  $d_e$  will output

$$F_{SA}(E; W^V, W^K, W^Q) = E + W^V E \cdot f_{\text{attn}} \left(\frac{(W^K E)^\top W^Q E}{\rho}\right),\tag{1}$$

where  $W^V, W^K, W^Q \in \mathbb{R}^{d_e \times d_e}$  are the value, key, and query weight matrices, respectively,  $\rho > 0$ is a normalization factor, and  $f_{\text{attn}}$  is an activation function for attention. There are different choices of  $f_{\text{attn}}$ ; for example Vaswani et al. (2017) adopts softmax.

In this work, similar to Zhang et al. (2023a); Wu et al. (2023), we set  $f_{\text{attn}}(x) = x$  and define  $W^{KQ} = (W^K)^\top W^Q \in \mathbb{R}^{d_e \times d_e}$ . We use F to denote this simplified model. Then, the output of F with an input embedding matrix  $E \in \mathbb{R}^{d_e \times d_n}$  can be expressed as

$$F(E; W^V, W^{KQ}) = E + W^V E \cdot \frac{E^\top W^{KQ} E}{\rho}.$$
(2)

In the following theoretical study and the subsequent experiments (Section 5.2), we show that this
 simplified transformer model has sufficient capability to approach the Bayes-optimal classifier for
 in-context classification of Gaussian mixtures.

104 105

106

098

099 100

### 2.2 IN-CONTEXT LEARNING FRAMEWORK

107 We adopt a framework for in-context learning similar to that used in Bai et al. (2024). Under this framework, the model receives a prompt  $P = (\mathcal{D}, x_{query})$  comprising a set of demonstrations  $\mathcal{D} =$ 

 $\{(x_i, y_i)\}_{i \in [N]} \overset{\text{i.i.d.}}{\sim} \mathcal{P} \text{ and a query } \sim \mathcal{P}_x, \text{ where } \mathcal{P} \text{ is the joint distribution of } (x, y) \text{ and } \mathcal{P}_x \text{ is } \mathcal{P}_y \text{ of } \mathcal{$ the marginal distribution of x. Here,  $x_i \in \mathbb{R}^d$  is an in-context example, and  $y_i$  is the corresponding label for  $x_i$ . For instance, in regression tasks,  $y_i \in \mathbb{R}$  is a scalar. In this paper, we focus on classification tasks. Thus, the range of  $y_i$  can be any set containing c different elements, such as  $\{1, \ldots, c\}$ , for classification problems involving c classes. The objective is to generate an output  $\widehat{y}_{query}$  that approximates the target  $y_{query} \sim \mathcal{P}_{y|x_{query}}$ . 

Since  $y_{query}$  is a discrete random variable, we use the total variation distance to measure the differ-ence between  $\hat{y}_{query}$  and  $y_{query}$ : 

$$\Delta(y_{query}, \hat{y}_{query}) = \sup_{z \in R(y_{query})} |\mathbb{P}(y_{query} = z) - \mathbb{P}(\hat{y}_{query} = z)|,$$
(3)

where  $R(y_{query})$  is the range of  $y_{query}$ . When  $\Delta(y_{query}, \hat{y}_{query}) = 0$ ,  $\hat{y}_{query}$  has the same distribution as  $y_{query}$ , which means the output of the model perfectly approximates  $y_{query}$ . 

Unlike standard supervised learning, in ICL, each prompt  $P_{\tau}$  can be sampled from a different distribution  $\mathcal{P}_{\tau}$ . We say that a model has the *ICL capability* if it can approximate  $y_{\tau,query}$  for a broad range of  $\mathcal{P}_{\tau}$ 's with fixed parameters. 

#### IN-CONTEXT BINARY CLASSIFICATION

In this section, we study the learning dynamics of a single-layer transformer for in-context binary classification. It is a special case of the general multi-class classification. As a result, the analysis is more concise. The general in-context multi-class classification problem is studied in Section 4.

We first introduce the prompt and the transformer structure we will use for in-context The prompt for in-context binary classification is denoted as binary classification.  $P = (x_1, y_1, \dots, x_N, y_N, x_{query})$ , where  $x_i \in \mathbb{R}^d$  and  $y_i \in \{-1, 1\}$ . We can convert this prompt P into its corresponding embedding matrix E(P) in the following form: 

$$E = E(P) = \begin{pmatrix} x_1 & x_2 & \cdots & x_N & x_{query} \\ y_1 & y_2 & \cdots & y_N & 0 \end{pmatrix} \in \mathbb{R}^{(d+1) \times (N+1)}.$$
 (4)

Similar to Huang et al. (2023); Wu et al. (2023); Ahn et al. (2024), we set some of the parameters in our model to 0 or 1 to simplify the optimization problem, and consider the parameters of our model  $(W^V, W^{KQ})$  in the following sparse form: 

$$W^{V} = \begin{pmatrix} 0_{d \times d} & 0_{d} \\ 0_{d}^{\top} & 1 \end{pmatrix}, \qquad W^{KQ} = \begin{pmatrix} W & 0_{d} \\ 0_{d}^{\top} & 0 \end{pmatrix},$$
(5)

where  $W \in \mathbb{R}^{d \times d}$ . We set the normalization factor  $\rho$  equal to the length of the prompt N. Let F(E(P); W) be the output matrix of the transformer. We then read out the bottom-right entry of the output matrix through a sigmoid function, and denote this output as  $\hat{y}_{out}$ . The output  $\hat{y}_{out}$  of the transformer with prompt P and parameters W can be expressed as

$$\begin{split} \widehat{y}_{\mathsf{out}} &= \sigma \left( [F(E(P);W)]_{(d+1),(N+1)} \right) \\ &= \sigma \Bigg( \begin{pmatrix} 0_d^\top & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{N} \sum_{i=1}^N x_i x_i^\top + \frac{1}{N} x_{\mathsf{query}} x_{\mathsf{query}}^\top & \frac{1}{N} \sum_{i=1}^N x_i y_i \\ \frac{1}{N} \sum_{i=1}^N x_i^\top y_i & \frac{1}{N} \sum_{i=1}^N y_i^2 \end{pmatrix} \begin{pmatrix} W & 0_d \\ 0_d^\top & 0 \end{pmatrix} \begin{pmatrix} x_{\mathsf{query}} \\ 0 \end{pmatrix} \Bigg) \\ &= \sigma \left( \left( \frac{1}{N} \sum_{i=1}^N y_i x_i^\top \right) W x_{\mathsf{query}} \right). \end{split}$$

We denote the prediction of our model for  $x_{query}$  as  $\hat{y}_{query}$ , which is a random variable depending on  $\hat{y}_{out}$ . Consider generating a random variable u uniformly on [0, 1]. If  $u \leq \hat{y}_{out}$ , we output  $\hat{y}_{query} = 1$ ; if  $u > \hat{y}_{out}$ , we output  $\hat{y}_{query} = -1$ . Then, we have  $\mathbb{P}(\hat{y}_{query} = 1) = \hat{y}_{out}$ ,  $\mathbb{P}(\hat{y}_{query} = -1) = \hat{y}_{out}$ .  $1 - \widehat{y}_{out}$ .

#### 162 3.1 TRAINING PROCEDURE 163

166

167

168

173 174

175 176

177

178 179

181

182

183

193

194

195 196

197

200 201 202

203

204 205

206

207

208

164 We study the binary classification of two Gaussian mixtures and use the following definition.

**Definition 3.1** We say a data pair  $(x, y) \sim \mathcal{P}^b(\mu_0, \mu_1, \Lambda)$  if y follows a Bernoulli distribution with  $\mathbb{P}(y = -1) = \mathbb{P}(y = 1) = 1/2$  and  $f(x|y = -1) = \mathsf{N}(\mu_0, \Lambda)$ ,  $f(x|y = 1) = \mathsf{N}(\mu_1, \Lambda)$ , where  $\mu_0, \mu_1 \in \mathbb{R}^d$  and  $\Lambda \in \mathbb{R}^{d \times d}$  is a positive definite matrix.

169 We consider the case of B training tasks indexed by  $\tau \in [B]$ . Each training task  $\tau$  is associated with 170 a prompt  $P_{\tau} = (x_{\tau,1}, y_{\tau,1}, \dots, x_{\tau,N}, y_{\tau,N}, x_{\tau,query})$  and a corresponding label  $y_{\tau,query}$ . We make 171 the following assumption in this section. 172

**Assumption 3.1** For each learning task  $\tau \in [B]$ , we assume

(1)  $\{x_{\tau,i}, y_{\tau,i}\}_{i=1}^{N}, \{x_{\tau,query}, y_{\tau,query}\} \overset{\text{i.i.d.}}{\sim} \mathcal{P}^{b}(\mu_{\tau,0}, \mu_{\tau,1}, \Lambda).$ (2)  $\mu_{\tau,0}$  is randomly sampled from N(0, I<sub>d</sub>), and  $\mu_{\tau,1} = U_{\tau,\Lambda}\mu_{\tau,0}$  where  $U_{\tau,\Lambda} = U_{\tau,\Lambda}\mu_{\tau,0}$  $\Lambda^{1/2}U_{\tau}\Lambda^{-1/2}$ , and  $U_{\tau}$  is uniformly distributed over the closed set of real unitary matrices such that  $U_{\tau}U_{\tau}^{\top} = I_d$ .

We denote the distribution of  $(\mu_{\tau,0},\mu_{\tau,1})$  as  $\mathcal{P}^b_{\Omega}(\Lambda)$ . Note that  $U_{\tau,\Lambda} = \Lambda^{1/2} U_{\tau} \Lambda^{-1/2}$  can be viewed as a linear transformation that preserves the inner product of vectors in  $\Lambda^{-1}$ -weighted norm, and we have  $\mu_{\tau,0}^{\dagger} \Lambda^{-1} \mu_{\tau,0} - \mu_{\tau,1}^{\dagger} \Lambda^{-1} \mu_{\tau,1} = 0.$ 

Let  $\hat{y}_{\tau,\text{out}} = \sigma([F(E(P_{\tau}); W)]_{(d+1), (N+1)})$  be the output of our transformer for task  $\tau$ . We define the empirical risk over B independent tasks as

$$\widehat{L}(W) = \frac{1}{2B} \sum_{\tau=1}^{B} -(1 + y_{\tau, query}) \log(\widehat{y}_{\tau, out}) - (1 - y_{\tau, query}) \log(1 - \widehat{y}_{\tau, out}).$$
(6)

Taking the limit of infinite training tasks  $B \to \infty$ , the expected training loss can be defined as

$$L(W) = \lim_{B \to \infty} \widehat{L}(W) = -\frac{1}{2} \mathbb{E}\left[ (1 + y_{\tau, \mathsf{query}}) \log(\widehat{y}_{\tau, \mathsf{out}}) + (1 - y_{\tau, \mathsf{query}}) \log(1 - \widehat{y}_{\tau, \mathsf{out}}) \right], \quad (7)$$

where the expectation is taken over  $(\mu_{\tau,0},\mu_{\tau,1}) \sim \mathcal{P}^b_{\Omega}(\Lambda), \{x_{\tau,i},y_{\tau,i}\}_{i=1}^N, \{x_{\tau,\mathsf{query}},y_{\tau,\mathsf{query}}\} \overset{\text{i.i.d.}}{\sim}$  $\mathcal{P}^{b}(\mu_{\tau,0},\mu_{\tau,1},\Lambda).$ 

Applying gradient descent over the expected training loss (7), we have the following theorem.

**Theorem 3.1** Under Assumption 3.1, the following statements hold.

(1) Optimizing the training loss L(W) in (7) with training prompt length N via gradient descent  $W^{t+1} = W^t - \eta \nabla L(W^t)$ , we have that for any  $t \ge 1$ 

$$\|W^t - W^*\|_F^2 \le \exp(-t/\kappa) \|W^0 - W^*\|_F^2, \tag{8}$$

where  $W^0$  is the initial parameter and  $W^*$  is the global minimizer of L(W), and  $\kappa = l/\alpha$ . Here  $\alpha$ , l are constants satisfying

$$0 < \alpha \le \lambda_{\min}(\nabla^2 L(W)) \le \lambda_{\max}(\nabla^2 L(W)) \le l, \text{ for all } W \in R_W, \tag{9}$$

where  $R_W = \{ W \in \mathbb{R}^{d \times d} \mid ||W - W^*||_F \le ||W^0 - W^*||_F \}.$ 

(2) Denote  $W^* = 2(\Lambda^{-1} + G), q = x_{\tau, query}, \mu = \mu_{\tau,1} - \mu_{\tau,0}, u = 2(\mu_{\tau,1} + \mu_{\tau,0}), a = \mu^{\top} \Lambda^{-1} q$ for simplicity. Then we have

$$||G||_{\max} \leq \frac{1}{N} ||S^{-1}(\mathbb{E}[\sigma'(a)(4qq^{\top} + \frac{1}{4}uu^{\top}\Lambda^{-1}qq^{\top}) + \sigma''(a)(\frac{1}{8}(u^{\top}\Lambda^{-1}q)^{2}\mu q^{\top} + 2q^{\top}\Lambda^{-1}q\mu q^{\top})])||_{\max} + o(1/N), \quad (10)$$

where  $S = 4\nabla^2 \widetilde{L}(2\Lambda^{-1})$ ,  $\widetilde{L}(2\Lambda^{-1}) = \lim_{N \to \infty} L(2\Lambda^{-1})$ ,  $\sigma'(\cdot)$  and  $\sigma''(\cdot)$  are the first-214 and second-order derivatives of  $\sigma(\cdot)$ , respectively, and the expectation is taken over  $(\mu_{\tau,0},$ 215  $\mu_{\tau,1}$   $\sim \mathcal{P}^b_{\Omega}(\Lambda), x_{\tau, \mathsf{query}} \sim \mathcal{P}^b_x(\mu_{\tau,0}, \mu_{\tau,1}, \Lambda).$ 

The detailed proof of Theorem 3.1 can be found in Appendix E. In the following, we provide a brief proof sketch to highlight the key ideas.

**Proof sketch for Theorem 3.1.** As a first step, we prove in Lemma E.2 that the expected loss 219 function L(W) in (7) is strictly convex with respect to (w.r.t.) W and is strongly convex in any 220 compact set of  $\mathbb{R}^{d \times d}$ . Moreover, we prove L(W) has one unique global minimizer  $W^*$ . Since 221 the loss function L(W) we consider is highly non-linear, we cannot directly find the closed-form 222 expression of  $W^*$ , as is often done in the prior literature. We address this technical challenge via 223 the following method. First, in Lemma E.3, by analyzing the Taylor expansion of L(W), we prove 224 that as  $N \to \infty$ , our loss function L(W) converges to  $\widetilde{L}(W)$  pointwisely (defined in (25)), and the global minimizer  $W^*$  converges to  $2\Lambda^{-1}$ . Thus, we denote  $W^* = 2(\Lambda^{-1} + G)$ , and prove  $||G||_{\text{max}}$ 225 is bounded and scales as  $\|G\|_{\max} = O(N^{-1/2})$ . Next, in Lemma E.4, by further analyzing the 226 227 Taylor expansion of the equation  $\nabla L(W^*) = 0$  at the point  $2\Lambda^{-1}$ , we establish a tighter bound 228  $\|G\|_{\max} = O(N^{-1})$ . In Lemma E.5, we prove that our loss function is *l*-smooth and provide an upper bound for *l*. Thus, in a compact set  $R_W$ , our loss function is  $\alpha$ -strongly convex and *l*-smooth. 229 Finally, leveraging the standard results from the convex optimization, we prove Theorem 3.1. 230

According to Theorem 3.1, we have  $W^t = W^* + H^t$  where  $||H^t||_{\max} \le \exp(-t/(2\kappa))||W^0 - W^*||_F$ . If we set  $T \ge 2\kappa \log(N \cdot ||W^0 - W^*||_F)$ , we have  $||H^T||_{\max} \le 1/N$ . Denoting  $\widehat{W} = W^T$ , we have  $\widehat{W} = 2(\Lambda^{-1} + G + H^T/2) = 2(\Lambda^{-1} + \widehat{G})$ , where  $\widehat{G} = G + H^T/2$ ,  $||\widehat{G}||_{\max} \le ||G||_{\max} + ||H^T||_{\max} = O(1/N)$ . Thus, we have the following corollary.

**Corollary 3.1** If we optimize the expected loss L(W) in (7) via gradient descent with training prompt length N, initial parameters  $W^0$ , and learning rate  $\eta = 1/l$ , then, under Assumption 3.1, after  $T \ge 2\kappa \log(N || W^0 - W^* ||_F)$  steps, the updated model  $\widehat{W}$  satisfies

$$\widehat{W} = 2(\Lambda^{-1} + \widehat{G}),\tag{11}$$

where  $\|\widehat{G}\|_{\max} = O(1/N)$ ,  $\kappa = l/\alpha$ , and  $\alpha$ , l are constants defined in (9).

Theorem 3.1 and Corollary 3.1 show that training a single-layer transformer with properly distributed data (Assumption 3.1) for binary classification via gradient descent can *linearly* converge to its global minimum  $W^* = 2(\Lambda^{-1} + G)$ . Furthermore, when the prompt length N grows, this global minimum  $W^*$  will converge to  $2\Lambda^{-1}$  at a rate of O(1/N).

### 3.2 IN-CONTEXT INFERENCE

236

237

238

239 240 241

242 243

248

249 250

251

253

254

255

256 257

263

Next, we analyze the performance of the trained transformer (11) for in-context binary classification tasks. We make the following assumption.

**Assumption 3.2** For an in-context test prompt  $P_{\text{test}} = (x_1, y_1, \ldots, x_M, y_M, x_{\text{query}})$ , we assume

(1) 
$$\{x_i, y_i\}_{i=1}^{M} \overset{\text{i.i.d.}}{\sim} \mathcal{P}^b(\mu_0, \mu_1, \Lambda), x_{query} \in \mathbb{R}^d$$
  
(2)  $\mu_0^{\top} \Lambda^{-1} \mu_0 = \mu_1^{\top} \Lambda^{-1} \mu_1.$ 

With this assumption, for  $y_{query} \sim \mathcal{P}^b_{y|x_{query}}(\mu_0, \mu_1, \Lambda)$ , according to the Bayes' theorem, we have

$$\mathbb{P}\left(y_{\mathsf{query}} = 1 | x_{\mathsf{query}}\right) = \frac{f(x_{\mathsf{query}} | y_{\mathsf{query}} = 1) \mathbb{P}\left(y_{\mathsf{query}} = 1\right)}{\sum_{z \in \{\pm 1\}} f(x_{\mathsf{query}} | y_{\mathsf{query}} = z) \mathbb{P}\left(y_{\mathsf{query}} = z\right)} = \sigma((\mu_1 - \mu_0)^\top \Lambda^{-1} x_{\mathsf{query}}).$$

If we test the trained transformer with parameters  $\widehat{W}$  in (11) and  $P_{\text{test}}$ , by a simple calculation, we have

$$\widehat{y}_{\mathsf{out}} = \sigma \left( \left( \frac{2}{M} \sum_{i=1}^{M} y_i x_i^{\mathsf{T}} \right) (\Lambda^{-1} + \widehat{G}) x_{\mathsf{query}} \right).$$
(12)

Intuitively, when the training prompt length  $N \to \infty$ , we have  $\widehat{G} \to 0$ , and when the test prompt length  $M \to \infty$ , we have  $\frac{2}{M} \sum_{i=1}^{M} y_i x_i^\top \to (\mu_1 - \mu_0)^\top$ . Thus, when  $N, M \to \infty$ ,

270  $\mathbb{P}(\widehat{y}_{query} = 1) = \widehat{y}_{out} \rightarrow \sigma((\mu_1 - \mu_0)^{\top} \Lambda^{-1} x_{query}) = \mathbb{P}(y_{query} = 1 | x_{query})$ , and the prediction of the trained transformer  $\widehat{y}_{query}$  perfectly matches with the distribution of the ground truth label  $y_{query}$ . 271 272

By analyzing the Taylor expansion of  $\hat{y}_{out}$  at point  $\sigma((\mu_1 - \mu_0)^{\top} \Lambda^{-1} x_{query})$ , we formally present 273 the aforementioned intuition in the following theorem, which establishes an upper bound of the total 274 variation distance between  $y_{query}$  and  $\hat{y}_{query}$ . 275

**Theorem 3.2** Consider a test prompt  $P_{\text{test}}$  satisfying Assumption 3.2, and let  $y_{\text{query}} \sim$  $\mathcal{P}^b_{u|x_{numer}}(\mu_0,\mu_1,\Lambda)$ . Let  $\widehat{y}_{query}$  be the prediction of the trained transformer with parameters W in 278 (11). Then, for the inference error defined in (3), we have

281

279

276

277

$$\begin{split} & \mathbb{E}[\Delta(y_{\mathsf{query}}, \widehat{y}_{\mathsf{query}})] \\ & \leq \sigma'(\mu^{\top} \Lambda^{-1} q) \left[ \|\widehat{G}\|_{\max} \sum_{i,j \in [d]} |\mu_i q_j| + \frac{1}{\sqrt{M}} \left( \frac{1}{2} |u^{\top} \Lambda^{-1} q| + \frac{2\sqrt{2}}{\sqrt{\pi}} \sum_{i,j \in [d]} |\Lambda_{ij}^{-1/2} q_j| \right) \right] \\ & + o\left( \frac{1}{N} + \frac{1}{\sqrt{M}} \right), \end{split}$$

287

288

289 290

284

where  $\mu = \mu_1 - \mu_0$ ,  $u = 2(\mu_1 + \mu_0)$ ,  $q = x_{query}$ , and the expectation is taken over  $\{x_i, y_i\}_{i=1}^{M} \stackrel{\text{i.i.d.}}{\sim}$  $\mathcal{P}^b(\mu_0, \mu_1, \Lambda).$ 

The proof of Theorem 3.2 can be found in Appendix F. Since  $\|\widehat{G}\|_{\max} = O(1/N)$ , Theorem 3.2 291 suggests that if we ignore the constants regarding  $\mu_0, \mu_1, \Lambda, x_{query}$ , the expected total variation dis-292 tance between  $y_{query}$  and  $\hat{y}_{query}$  is at most  $O(1/N + 1/\sqrt{M})$ . On the other hand, for data pair 293  $(x,y) \sim \mathcal{P}^b(\mu_0,\mu_1,\Lambda)$ , the Bayes-optimal classifier is  $\mathbb{P}(y=1|x) = f(x|y)\mathbb{P}(y=1)/f(x) = f(x|y)\mathbb{P}(y=1)/f(x)$  $\sigma((\mu_1 - \mu_0)^{\dagger} \Lambda^{-1} x)$ , which corresponds to the logistic regression model  $\sigma(w^{\dagger} x + b)$  with param-295 eters  $w = \Lambda^{-1}(\mu_1 - \mu_0)$  and b = 0. Therefore, when  $N, M \to \infty$ , the prediction of the trained 296 transformer is Bayes-optimal, and is equivalent to the optimal logistic regressor for binary classi-297 fication problems with distribution  $\mathcal{P}^b(\mu_0, \mu_1, \Lambda)$ . Note that different from Assumption 3.1 which 298 states that  $\mu_{\tau,0}, \mu_{\tau,1}, x_{\tau,query}$  are sampled according to some specific distributions during training, 299 Assumption 3.2 does not impose strong distributional constraints on  $\mu_0, \mu_1$  and  $x_{query}$ , which shows 300 the strong generalization ability of the trained transformer. We also discuss the consequences when 301 Assumption 3.2 does not hold in Remark F.1, which highlights the necessity of Assumption 3.2. 302 Moreover, even if  $M \to \infty$ , the distribution variation between  $y_{query}$  and  $\hat{y}_{query}$  does not disappear 303 unless  $N \to \infty$ . Thus, the ICL ability of trained transformers for binary classification is limited 304 by the finite length of training prompts. Similar behaviors have also been observed in Zhang et al. (2023a) for in-context linear regression. 305

306 307

308

#### 4 IN-CONTEXT MULTI-CLASS CLASSIFICATION

309 We now extend the study of the learning dynamics of a single-layer transformer to in-context multi-310 class classification, generalizing the results of the previous section. We will present the detailed 311 formulation and then focus on the main differences to binary classification. 312

We first introduce the prompt and the transformer structure that will be used for in-context multi-313 class classification. The prompt for in-context multi-class classification involving  $c \ge 2$  classes can 314 be expressed as  $P = (x_1, y_1, \ldots, x_N, y_N, x_{query})$ , where  $x_i \in \mathbb{R}^d$ ,  $y_i \in \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_c\}$ , and  $\mathbf{e}_i$  is 315 the *i*-th standard unit vector of  $\mathbb{R}^{c}$ . Its embedding matrix can be formulated as 316

$$E = E(P) = \begin{pmatrix} x_1 & x_2 & \cdots & x_N & x_{query} \\ y_1 & y_2 & \cdots & y_N & 0_c \end{pmatrix} \in \mathbb{R}^{(d+c) \times (N+1)}.$$
 (13)

318 319

323

317

Similar to the binary case, we set some of the parameters in our model as 0 and 1 to simplify 320 the optimization problem and consider the parameters of our model  $(W^V, W^{KQ})$  in the following 321 sparse form: 322

$$W^{V} = \begin{pmatrix} 0_{d \times d} & 0_{d \times c} \\ 0_{c \times d} & I_{c} \end{pmatrix}, \qquad W^{KQ} = \begin{pmatrix} W & 0_{d \times c} \\ 0_{c \times d} & 0_{c \times c} \end{pmatrix}, \tag{14}$$

where  $W \in \mathbb{R}^{d \times d}$ . We set the normalization factor  $\rho$  equal to the length of the prompt N. We read out the bottom-right c-dimensional column vector from the output matrix with a softmax function as the output, denoted as  $\hat{y}_{out}$ . With parameters W and a prompt  $P = (x_1, y_1, \dots, x_N, y_N, x_{query})$ , the output can be expressed as

$$\widehat{y}_{\mathsf{out}} = \operatorname{softmax}\left([F(E(P); W)]_{(d+1):(d+c),(N+1)}\right) = \operatorname{softmax}\left(\left(\frac{1}{N}\sum_{i=1}^{N} y_i x_i^{\mathsf{T}}\right) W x_{\mathsf{query}}\right).$$

We denote the prediction of the model for  $x_{query}$  as  $\hat{y}_{query}$ , which is a random variable depending on  $\hat{y}_{out}$ . Randomly sample a random variable u that is uniformly distributed on [0,1]. If  $u \in$  $\left[\sum_{j=1}^{i-1} (\widehat{y}_{out})_j, \sum_{j=1}^{i} (\widehat{y}_{out})_j\right]$ , where  $(\widehat{y}_{out})_j$  is the *j*-th element of  $\widehat{y}_{out}$ , we let  $\widehat{y}_{query} = \mathbf{e}_i$ . Thus,  $\mathbb{P}\left(\widehat{y}_{\mathsf{query}} = \mathbf{e}_i\right) = (\widehat{y}_{\mathsf{out}})_i.$ 

4.1 TRAINING PROCEDURE

 We focus on the multi-class classification of Gaussian mixtures and use the following definition.

**Definition 4.1** We say a data pair  $(x, y) \sim \mathcal{P}^m(\mu, \Lambda)$  if  $\mathbb{P}(y = e_i) = 1/c$  and  $f(x|y = e_i) = \mathbb{N}(\mu_i, \Lambda)$  for  $i \in [c]$ , where  $\mu = (\mu_1, \dots, \mu_c) \in \mathbb{R}^{d \times c}$  and  $\Lambda \in \mathbb{R}^{d \times d}$  is a positive definite matrix.

We consider the case of B training tasks indexed by  $\tau \in [B]$ . Each training task  $\tau$  is associated with a prompt  $P_{\tau} = (x_{\tau,1}, y_{\tau,1}, \dots, x_{\tau,N}, y_{\tau,N}, x_{\tau,query})$  and a corresponding label  $y_{\tau,query}$ . We make the following assumption in this section. 

**Assumption 4.1** For each learning task  $\tau \in [B]$ , we assume

(1)  $\{x_{\tau,i}, y_{\tau,i}\}_{i=1}^{N}, \{x_{\tau,query}, y_{\tau,query}\} \overset{\text{i.i.d.}}{\sim} \mathcal{P}^{m}(\mu_{\tau} = (\mu_{\tau,1}, \dots, \mu_{\tau,c}), \Lambda).$ (2)  $\mu_{\tau,1}$  is sampled from N(0,  $I_d$ ), and  $\mu_{\tau,k} = U_{\tau,k,\Lambda}\mu_{\tau,1}, k = 2, 3, \dots, c$ , where  $U_{\tau,k,\Lambda} = U_{\tau,k,\Lambda}$  $\Lambda^{1/2}U_{\tau,k}\Lambda^{-1/2}$ , and  $U_{\tau,k}$  are uniformly distributed over the closed set of real unitary matrices such that  $U_{\tau,k}U_{\tau,k}^{\top} = I_d$ .

We denote the distribution of  $\mu_{\tau}$  as  $\mathcal{P}^m_{\Omega}(\Lambda)$ . Note that  $U_{\tau,k,\Lambda} = \Lambda^{1/2} U_{\tau,k} \Lambda^{-1/2}$  can be viewed as linear transformations that preserve the inner product of vectors in the  $\Lambda^{-1}$ weighted norm, and we have  $\mu_{\tau,i}^{\top}\Lambda^{-1}\mu_{\tau,i} = \mu_{\tau,j}^{\top}\Lambda^{-1}\mu_{\tau,j}$ , for  $i,j \in [c]$ . Let  $\widehat{y}_{\tau,\text{out}} =$ softmax  $([F(E(P_{\tau}); W)]_{(d+1):(d+c),(N+1)})$  be the output of the transformer for task  $\tau$ . We define the empirical risk over B independent tasks as

$$\widehat{L}(W) = \frac{1}{B} \sum_{\tau=1}^{B} \sum_{k=1}^{c} -(y_{\tau, \mathsf{query}})_k \log((\widehat{y}_{\tau, \mathsf{out}})_k).$$
(15)

Taking the limit of infinite training tasks  $B \to \infty$ , the expected training loss can be defined as

$$L(W) = \lim_{B \to \infty} \widehat{L}(W) = -\mathbb{E}\left[\sum_{k=1}^{c} (y_{\tau, \mathsf{query}})_k \log((\widehat{y}_{\tau, \mathsf{out}})_k)\right],\tag{16}$$

where the expectation is taken over  $\mu_{\tau} \sim \mathcal{P}_{\Omega}^{m}(\Lambda), \{x_{\tau,i}, y_{\tau,i}\}_{i=1}^{N}, \{x_{\tau,query}, y_{\tau,query}\}$  $\overset{\rm i.i.d.}{\sim}$  $\mathcal{P}^m(\mu_{\tau}, \Lambda).$ 

Applying gradient descent over the expected training loss (16), we have the following theorem.

**Theorem 4.1** (Informal) Under Assumption 4.1, the following statements hold.

(1) Optimizing training loss L(W) in (16) with training prompt length N via gradient descent  $W^{t+1} = W^t - \eta \nabla L(W^t)$ , for any  $t \ge 1$ , we have

$$\|W^{t} - W^{*}\|_{F}^{2} \le \exp(-t/\kappa)\|W^{0} - W^{*}\|_{F}^{2},$$
(17)

where  $W^0$  is the initial parameter and  $W^*$  is the global minimizer of L(W),  $\kappa = l/\alpha$ . *Here,*  $\alpha$ *, l are constants such that* 

$$0 < \alpha \le \lambda_{\min}(\nabla^2 L(W)) \le \lambda_{\max}(\nabla^2 L(W)) \le l, \text{ for all } W \in R_W,$$

$$(18)$$

$$here P = (W \in \mathbb{R}^{d \times d + ||W|}, W^*|| \le ||W|^0, W^*||)$$

where  $R_W = \{ W \in \mathbb{R}^{d \times d} \mid ||W - W^*||_F \le ||W^0 - W^*||_F \}.$ 

(2) Denoting  $W^* = c(\Lambda^{-1} + G)$ , we have  $||G||_{\max} = O(c/N)$ . (3) After  $T \ge 2\kappa \log(N \cdot ||W^0 - W^*||_F)$  steps, denoting the updated model  $\widehat{W}$  satisfies

> $\widehat{W} = c(\Lambda^{-1} + \widehat{G}).$ (19)

where 
$$\|\widehat{G}\|_{\max} = O(c/N)$$

The formal statement and proof of Theorem 4.1 can be found in Appendix G. Technically, the proof of Theorem 4.1 builds on that of Theorem 3.1, but the more complicated cross terms in the Taylor expansions of the softmax functions, which are due to the nature of *multi-class* classification, bring new challenges to the analysis. To address these issues, we derived new bounds on the expected errors of the cross terms in Lemma G.1, G.2, which may be of independent interest to other similar problems.

Theorem 4.1 shows that training a single-layer transformer with properly distributed data (Assumption 4.1) for in-context multi-class classification via gradient descent can linearly converge to its global minimum  $W^* = c(\Lambda^{-1} + G)$ . When the prompt length N grows, this global minimum  $W^*$ 393 will converge to  $c\Lambda^{-1}$  at a rate of O(c/N). Compared to the binary case, the new results establish 394 the scaling behavior w.r.t. the number of classes c. 395

#### 4.2 IN-CONTEXT INFERENCE

**Assumption 4.2** For an in-context test prompt  $P_{\text{test}} = (x_1, y_1, \dots, x_M, y_M, x_{\text{query}})$ , we assume

(1) 
$$\{x_i, y_i\}_{i=1}^{M} \stackrel{\text{i.i.d.}}{\sim} \mathcal{P}^m(\mu, \Lambda), \ \mu = (\mu_1, \dots, \mu_c) \in \mathbb{R}^{d \times c}, \ x_{query} \in \mathbb{R}^d.$$
  
(2)  $\mu_i^\top \Lambda^{-1} \mu_i = \mu_j^\top \Lambda^{-1} \mu_j, \ \text{for } i, j \in [c].$ 

With this assumption, for  $y_{query} \sim \mathcal{P}_{y|x_{query}}^m(\mu, \Lambda)$ , according to the Bayes' theorem, we have

$$\mathbb{P}\left(y_{\mathsf{query}} = \mathbf{e}_{k} | x_{\mathsf{query}}\right) = \frac{f(x_{\mathsf{query}} | y_{\mathsf{query}} = \mathbf{e}_{k}) \mathbb{P}\left(y_{\mathsf{query}} = \mathbf{e}_{k}\right)}{\sum_{j=1}^{c} f(x_{\mathsf{query}} | y_{\mathsf{query}} = \mathbf{e}_{j}) \mathbb{P}\left(y_{\mathsf{query}} = \mathbf{e}_{j}\right)} = \operatorname{softmax}(\mu^{\top} \Lambda^{-1} x_{\mathsf{query}})_{k}.$$

If we test the trained transformer with parameters  $\widehat{W}$  in (19) and prompt  $P_{\text{test}}$ , by a simple calculation, we have

412

425 426

428

378

379

380

381

382

384

386

387

388

389

390

391

392

397 398

$$\widehat{y}_{\mathsf{out}} = \operatorname{softmax}\left(\left(\frac{c}{M}\sum_{i=1}^{M} y_i x_i^{\mathsf{T}}\right) (\Lambda^{-1} + \widehat{G}) x_{\mathsf{query}}\right).$$
(20)

413 Note that, when the training prompt length  $N \to \infty$ , we have  $\widehat{G} \to 0$ , and when the test prompt 414 length  $M \to \infty$ , we have  $\frac{c}{M} \sum_{i=1}^{M} y_i x_i^\top \to \mu^\top$ . Thus, when  $N, M \to \infty$ ,  $\mathbb{P}(\widehat{y}_{query} = \mathbf{e}_k) = \mathbf{e}_k$ 415  $(\widehat{y}_{out})_k \rightarrow \operatorname{softmax}(\mu^{\top} \Lambda^{-1} \widehat{x}_{query})_k = \mathbb{P}(y_{query} = \mathbf{e}_k | x_{query}), \text{ i.e., the prediction of the trained}$ 416 transformer  $\hat{y}_{query}$  matches the ground truth label  $y_{query}$ . 417

By analyzing the Taylor expansion of  $\hat{y}_{out}$  at point  $\operatorname{softmax}(\mu^{\top}\Lambda^{-1}x_{query})$ , we crystallize the afore-418 mentioned intuition in the following theorem, which establishes an upper bound of the total variation 419 distance between  $y_{query}$  and  $\hat{y}_{query}$ . 420

421 **Theorem 4.2** (Informal) Let  $P_{\text{test}}$  satisfy Assumption 4.2 and  $y_{\text{query}} \sim \mathcal{P}_{y|x_{\text{query}}}^m(\mu, \Lambda)$ . Denote  $\widehat{y}_{\text{query}}$ 422 as the prediction of the trained transformer with parameter  $\widehat{W}$  in (19). Then, for the inference error 423 defined in (3), we have 424

 $\mathbb{E}[\Delta(y_{\mathsf{query}},\widehat{y}_{\mathsf{query}})] = O(c^2N^{-1} + c^{3/2}M^{-1/2}),$ 

where the expectation is taken over  $\{x_i, y_i\}_{i=1}^{M} \stackrel{\text{i.i.d.}}{\sim} \mathcal{P}^m(\mu, \Lambda)$ . 427

The formal statement and proof of Theorem 4.2 can be found in Appendix H. We can see that 429 the convergence rate of the inference error in multi-class classification w.r.t. N and M is sim-430 ilar to that in the binary classification, except for the constant coefficient c. This suggests that 431 classification tasks with more classes may have higher errors than those with fewer classes. On

the other hand, for data pair  $(x, y) \sim \mathcal{P}^m(\mu, \Lambda)$ , the Bayes-optimal classifier is  $\mathbb{P}(y = \mathbf{e}_k | x) = f(x|y)\mathbb{P}(y = \mathbf{e}_k)/f(x) = \operatorname{softmax}(\mu^{\top}\Lambda^{-1}x)_k$ , which corresponds to a softmax regression model  $\operatorname{softmax}(Wx + b)$  with parameters  $W = \mu^{\top}\Lambda^{-1}$  and b = 0. When  $N, M \to \infty$ , the prediction of the trained transformer is Bayes-optimal, and is equivalent to the optimal softmax regressor for multi-class classification problems with distribution  $\mathcal{P}^m(\mu, \Lambda)$ . Note that different from Assump-tion 4.1 which states that  $\mu_{\tau}, x_{\tau, query}$  are sampled according to some specific distributions during training, Assumption 4.2 impose strong distributional constraints on  $\mu$  or  $x_{guerv}$ , which shows the strong generalization ability of the trained transformer. We also discuss the consequences when Assumption 4.2 does not hold in Remark H.1, which highlights the necessity of Assumption 4.2. Moreover, even if  $M \to \infty$ , the distribution variation between  $y_{query}$  and  $\hat{y}_{query}$  does not disappear unless  $N \to \infty$ . Thus, the ICL ability of the trained transformers for multi-class classification is limited by the finite length of training prompts. Similar behaviors have also been observed in Zhang et al. (2023a) for in-context linear regression and in Section 3.2 for in-context binary classification. 

# 

### EXPERIMENTS

In this section, we report the experiment results on multi-layer, nonlinear transformers to investigate their similarities and differences to the single-layer, linear transformer we theoretically analyzed in the pervious sections. Detailed experimental settings and additional results can be found in Appendix I.



Figure 1: '1-layer': single-layer transformer defined in Section 4, '3-layer': 3-layer transformers with softmax
 attention. N: training prompt length. c: number of Gaussian mixtures.

We train single-layer and multi-layer transformers for in-context classification of Gaussian mixtures with different numbers of Gaussian mixtures c, different lengths of training prompts N, and test them with different test prompt lengths M. The results are reported in Figure 1. We can see that for both single-layer and multi-layer transformers, the inference errors decrease as N and M increase, and they increase as c increases, which not only verify our theoretical claims but also show that, the simplified model we have studied indeed exhibits behavioral similarities to the more complex multi-layer, nonlinear transformers, and some of our observations for this simplified model also hold for more complex transformers. 

### 5.1 VARYING COVARIANCES AND NORMS

Note that in Assumption 3.1, 4.1, 3.2, 4.2, we assume that the covariance  $\Lambda$  during pre-training and during inference are the same, and the means of all Gaussian components  $\{\mu_{\tau,i}, i \in [c]\}$  have the same  $\Lambda^{-1}$  weighted norm. In Remark F.1, H.1, we also discuss the situation when Assumption 3.2, 4.2 does not hold and show the necessities of Assumption 3.2, 4.2. In this subsection, we consider training transformers with data of varying covariances  $\Lambda$  and with Gaussian component means of unequal  $\Lambda^{-1}$  weighted norms, and examine how these factors affect the ICL abilities of transformers. Results are shown in Figure 2. From Figure 2 (a), we can see that both models perform better when their  $\mu_{\tau,i}$  have the same  $\Lambda^{-1}$  weighted norm ('same norm'), however, in the 'different norms' setting, the performance of '1-layer' deteriorates more significantly, while transformers with a more complex structure ('3-layer') show better robustness under this distribution shift. Similar situations also happen in Figure 2 (b), where '3-layer' also shows better tolerance to the covariance shifts than '1-layer'. Experimental results in Figure 2 show the necessities of Assumption 3.1, 4.1, 3.2, 4.2 for
 the single-layer transformers we considered in this paper, also demonstrates the better robustness of
 multi-layer, nonlinear transformers. Developing a better understanding of the robustness of more
 complex transformers is an intriguing direction for future research.



Figure 2: All models are trained with prompt length N = 100, tested with prompts satisfying Assumption 4.2 with  $\Lambda$ . c = 3. (a): (same norm): pre-training data are sampled according to Assumption 4.1 with  $\Lambda$ . 'different norms': For each  $\tau$ , with probability  $\mathbb{P}(k = j) = 1/10, \mu_{\tau,i} \sim N(k, I_d), j = 0, 1, ..., 9$ . (b): (same covariance): pre-training data are sampled according to Assumption 4.1 for the fixed  $\Lambda$ . (different covariances): Sample additional  $\Lambda_1, \Lambda_2, \Lambda_3$ . Then, generate pre-training data according to Assumption 4.1 with  $\Lambda$ . (different with  $\Lambda, \Lambda_1, \Lambda_2, \Lambda_3$ .

### 5.2 COMPARISON OF TRANSFORMERS WITH OTHER MACHINE LEARNING ALGORITHMS



Figure 3: '1-layer, sparse': single-layer transformer defined in Section 4, '1-layer, full': single-layer transformer with full parameters (59), '3-layer': a 3-layer transformer with softmax attention, 'softmax': softmax regression, 'SVM, linear': SVM with linear kernel, 'SVM, gaussian': SVM with Gaussian kernel, '1-NN': 1nearest neighbor, '3-NN': 3-nearest neighbor. All three transformers are trained with prompt length N = 100.

Additionally, we conduct experiments comparing the ICL performances of the transformers with other machine learning algorithms for the classification of three Gaussian mixtures. Form Figure 3, we can see that all three transformer models significantly outperform the classical methods (softmax regression, SVM, *K*-nearest neighbor), demonstrating the strong ICL capacities of transformers.

### 6 CONCLUSION

We studied the learning dynamics of transformers for in-context classification of Gaussian mixtures, and showed that with properly distributed data, a single-layer transformer trained via gradient de-scent converges to its global minimum. Moreover, we established the upper bounds of the inference errors of the trained transformers and discussed how the training and test prompt lengths influence the performance of the model. Experimental results also corroborated the theoretical claims. There are some directions worth further exploring. One potential avenue is to investigate whether the assumptions regrading the training and test prompts can be relaxed. Additionally, we have only examined single-layer transformers with linear attention and sparse parameters. The learning dynam-ics of multi-layer transformers with nonlinear attention (e.g., softmax) for in-context classification problems remain an interesting area for future investigation.

## 540 REFERENCES

548

570

571

572

573

577

578

579

- 542 Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to imple 543 ment preconditioned gradient descent for in-context learning. Advances in Neural Information
   544 Processing Systems, 36, 2024.
- Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.
- Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians:
   Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36, 2024.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
   Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are
   few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.
- 555 556 Sébastien Bubeck. Convex optimization: Algorithms and complexity, 2015.
- Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.
- Xiang Cheng, Yuxin Chen, and Suvrit Sra. Transformers implement functional gradient descent to
   learn non-linear functions in context. *arXiv preprint arXiv:2312.06528*, 2023.
- Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.
- Deqing Fu, Tian-Qi Chen, Robin Jia, and Vatsal Sharan. Transformers learn higher-order op timization methods for in-context learning: A study with linear models. arXiv preprint
   arXiv:2310.17086, 2023.
  - 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.
- Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris
   Papailiopoulos. Looped transformers as programmable computers. In *International Conference* on *Machine Learning*, pp. 11398–11442. PMLR, 2023.
  - Angeliki Giannou, Liu Yang, Tianhao Wang, Dimitris Papailiopoulos, and Jason D Lee. How well can transformers emulate in-context newton's method? *arXiv preprint arXiv:2403.03183*, 2024.
- Tianyu Guo, Wei Hu, Song Mei, Huan Wang, Caiming Xiong, Silvio Savarese, and Yu Bai. How do transformers learn in-context beyond simple functions? a case study on learning with representations. *arXiv preprint arXiv:2310.10616*, 2023.
- 583
   584
   584
   585
   Michael Hahn and Navin Goyal. A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*, 2023.
- Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. *arXiv preprint* arXiv:2310.05249, 2023.
- Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximalgradient methods under the polyak-lojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pp. 795–811. Springer, 2016.
- <sup>593</sup> Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. *arXiv preprint arXiv:2402.01258*, 2024.

594

Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. Training nonlinear trans-595 formers for efficient in-context learning: A theoretical learning and generalization analysis. arXiv 596 preprint arXiv:2402.15607, 2024. 597 Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers 598 as algorithms: Generalization and stability in in-context learning. In International Conference on Machine Learning, pp. 19565–19594. PMLR, 2023a. 600 601 Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a 602 mechanistic understanding. In International Conference on Machine Learning, pp. 19689–19729. 603 PMLR, 2023b. 604 Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforce-605 ment learning via supervised pretraining. arXiv preprint arXiv:2310.08566, 2023. 606 607 Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is 608 provably the optimal in-context learner with one layer of linear self-attention. arXiv preprint arXiv:2307.03576, 2023. 609 610 Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with 611 gradient descent. arXiv preprint arXiv:2402.14735, 2024. 612 613 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 614 your work: Scratchpads for intermediate computation with language models. arXiv preprint 615 arXiv:2112.00114, 2021. 616 617 OpenAI. GPT-4 technical report, 2023. 618 Reese Pathak, Rajat Sen, Weihao Kong, and Abhimanyu Das. Transformers can optimally learn 619 regression mixture models. arXiv preprint arXiv:2311.08362, 2023. 620 621 Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers 622 as support vector machines. arXiv preprint arXiv:2308.16898, 2023. 623 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko-624 lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open founda-625 tion and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023. 626 627 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, 628 Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural informa-629 tion processing systems, 30, 2017. 630 Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordv-631 intsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient 632 descent. In International Conference on Machine Learning, pp. 35151–35174. PMLR, 2023. 633 634 Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large lan-635 guage models are implicitly topic models: Explaining and finding good demonstrations for incontext learning. In Workshop on Efficient Systems for Foundation Models@ ICML2023, 2023. 636 637 Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yo-638 gatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language 639 models. arXiv preprint arXiv:2206.07682, 2022. 640 Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter L Bartlett. 641 How many pretraining tasks are needed for in-context learning of linear regression? arXiv 642 preprint arXiv:2310.08391, 2023. 643 644 Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context 645 learning as implicit bayesian inference. arXiv preprint arXiv:2111.02080, 2021. 646 Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. In-context learning with representations: 647

Contextual generalization of trained transformers. arXiv preprint arXiv:2408.10147, 2024.

- 648 Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. 649 arXiv preprint arXiv:2306.09927, 2023a. 650
- Ruiqi Zhang, Jingfeng Wu, and Peter L Bartlett. In-context learning of a linear transformer block: benefits of the mlp component and one-step gd initialization. arXiv preprint arXiv:2402.14951, 652 2024. 653
- 654 Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christo-655 pher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer 656 language models. arXiv preprint arXiv:2205.01068, 2022.
  - Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. arXiv preprint arXiv:2305.19420, 2023b.

660 661 662

663 664

665

666

667

657

658

659

651

#### APPENDIX А

The Appendix is organized as follows. In Section B, we provide a literature review of the related works that studied the ICL abilities of transformers. In Section C, we introduce the additional notations for the proofs in the Appendix. In Section D, we introduce some useful Lemmas we adopt from previous literature. In Sections E, F, G, H, we present the proofs of Theorem 3.1, 3.2, 4.1, 4.2 respectively. In Section I, we provide additional results and details of our experiments.

668 669 670

671

#### В **RELATED WORK**

672 It has been observed that transformer-based models have impressive ICL abilities in natural language 673 processing (Brown et al., 2020; Nye et al., 2021; Wei et al., 2022; Dasgupta et al., 2022; Zhang et al., 674 2022). Garg et al. (2022) first initiated the study of the ICL abilities of transformers in a mathemati-675 cal framework and they empirically showed that transformers can in-context learn linear regression, 676 two-layer ReLU networks, and decision trees. Subsequently, numerous works have been developed 677 to explain the ICL capacities of transformers in solving in-context mathematical problems. These works mainly use two approaches: constructing specific transformers capable of performing certain 678 in-context learning tasks, and studying the training dynamics of transformers for such tasks. 679

680 Constructions of transformers. Akyürek et al. (2022); Von Oswald et al. (2023) showed by con-681 struction that multi-layer transformers can be viewed as multiple steps of gradient descent for lin-682 ear regression. Akyürek et al. (2022) also showed that constructed transformers can implement closed-form ridge regression. Guo et al. (2023) showed that constructed transformers can perform 683 in-context learning with representations. Bai et al. (2024) proved that constructed transformers 684 can perform various statistical machine learning algorithms through in-context gradient descent and 685 showed that constructed transformers can perform in-context model selection. Lin et al. (2023) 686 demonstrated that constructed transformers can approximate several in-context reinforcement learn-687 ing algorithms. Fu et al. (2023); Giannou et al. (2024) further proved that constructed transform-688 ers can perform higher-order optimization algorithms like Newton's method. Pathak et al. (2023) 689 showed that transformers can learn mixtures of linear regressions. Giannou et al. (2023) proved that 690 looped transformers that can emulate various in-context learning algorithms. Cheng et al. (2023) 691 showed that transformers can perform functional gradient descent for learning non-linear functions 692 in context. Zhang et al. (2024) showed that a linear attention layer followed by a linear layer can learn and encode a mean signal vector for in-context linear regression. 693

694 Training dynamics of transformers. Mahankali et al. (2023); Ahn et al. (2024) proved that the 695 global minimizer of the in-context learning loss of linear transformer can be equivalently viewed 696 as one-step preconditioned gradient descent for linear regression. Zhang et al. (2023a) proved the 697 convergence of gradient flow on a single-layer linear transformer and discussed how training and 698 test prompt length will influence the prediction error of transformers for linear regression. Huang 699 et al. (2023) proved the convergence of gradient descent on a single-layer transformer with softmax attention with certain orthogonality assumptions on the data features. Li et al. (2023b) showed that 700 trained transformers can learn topic structure. Wu et al. (2023) analyzed the task complexity bound 701 for pretraining single-layer linear transformers on in-context linear regression tasks. Tarzanagh

et al. (2023) built the connections between single-layer transformers and support vector machines
(SVMs). Nichani et al. (2024) showed that transformers trained via gradient descent can learn causal
structure. Chen et al. (2024) proved the convergence of gradient flow on a multi-head softmax
attention model for in-context multi-task linear regression. Kim & Suzuki (2024); Yang et al. (2024)
proved that trained transformers can learn nonlinear features in context.

Recently, Li et al. (2024) studied the training dynamics of a single layer transformer for in-context 708 classification problems. However, they only studied the binary classification tasks with *finite* patterns. They generated their data as  $x = \mu_j + \kappa v_k$ , where  $\{\mu_j\}_{j=1}^{M_1}$  are in-domain-relevant patterns and  $\{\nu_k\}_{k=1}^{M_2}$  are in-domain-irrelevant patterns,  $M_1 \ge M_2$  and these patterns are all pairwise orthogonal. Thus, the possible distribution of their data is finite and highly limited. In contrast, our 709 710 711 712 work explores the ICL capabilities of transformers for both binary and multi-class classification of 713 Gaussian mixtures. Specifically, our data is drawn according to  $\mathcal{P}^b(\mu_0, \mu_1, \Lambda)$  or  $\mathcal{P}^m(\mu, \Lambda)$ , and 714 the range and possible distributions of our data are *infinite*. Furthermore, the transformer architec-715 tures analyzed in their work also differ from those in our study, thereby highlighting the distinct 716 contributions and independent interests of our work.

717 Some works also studied the ICL from other perspectives. To name a few, Xie et al. (2021) explained 718 the ICL as implicit Bayesian inference; Wang et al. (2023) explained the LLMs as latent variable 719 models; Zhang et al. (2023b) explained the ICL abilities of transformers as implicitly implementing 720 a Bayesian model averaging algorithm; and Li et al. (2023a) studied the generalization and stabil-721 ity of the ICL abilities of transformers. Hahn & Goyal (2023) showed that ICL can arise through 722 recombination of compositional structure found in linguistic data. They derived an information-723 theoretic bound showing how ICL abilities arise from generic next-token prediction and provided a theoretical justification for the benefits of chain-of-thought. They also observed that as prompt 724 length increases, their information-theoretic bound converge to zero. However, they considered the 725 ICL with data generated by Compositional Attribute Grammar (CAG) and for an idealized predic-726 tor, which is not a predictor of actual transformers or LLMs, while we prove the convergence of 727 transformers for ICL of classification of Gaussian mixtures and derived the ICL error respect to the 728 trained transformer. Thus, our paper has its own independent contributions and intellectual merits. 729

730 731

732 733

734 735

736

737 738 739

### C ADDITIONAL NOTATIONS

We denote  $X \sim Bin(n,p)$  if a random variable X follows the binomial distribution with parameters  $n \in \mathbb{N}$  and  $p \in [0,1]$ , which means  $\mathbb{P}(X=k) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$ . We denote  $X \sim Multin(n,p)$  if random variables  $X = (X_1, X_2, \ldots, X_k)$  follow the Multinomial distribution with parameters  $n \in \mathbb{N}$  and  $p_1 = p_2 = \cdots = p_k = 1/k$ , which means  $\mathbb{P}(X = (x_1, x_2, \ldots, x_k)) = \frac{n!}{\prod_{i=1}^k x_i!}k^{-n}$ . We denote  $\zeta_i(x) = \operatorname{softmax}(x)_i = \exp(x_i)/(\sum_{j=1}^k \exp(x_j))$  for simplicity. We define  $\delta_{ii} = 1$ ,  $\delta_{ij} = 0$ ,  $i \neq j$ . For  $x \in \mathbb{N}$ , we define  $t_1(x) = \lfloor (x-1)/d \rfloor + 1$ ,  $t_2(x) = ((x-1) \mod d) + 1$ .

740 741 742

743 744

749

754 755

### D USEFUL LEMMAS

**Lemma D.1** ((Karimi et al., 2016)) If  $f : \mathbb{R}^d \to \mathbb{R}$  is  $\mu$ -strongly convex, then

$$f(x) - \min_{x} f(x) \ge \frac{\mu}{2} \|x^* - x\|_{2}^{2}$$

where  $x^* = \arg \min_x f(x)$ .

**Lemma D.2 ((Bubeck, 2015))** Suppose  $f : \mathbb{R}^d \to \mathbb{R}$  is  $\alpha$ -strongly convex and  $\beta$ -smooth for some  $0 < \alpha \leq \beta$ . Then, the gradient descent iterating  $w^{t+1} = w^t - \eta \nabla f(w^t)$  with learning rate  $\eta = 1/\beta$  and initialization  $w^0 \in \mathbb{R}^d$  satisfies that for any  $t \geq 1$ ,

 $||w^t - w^*||_2^2 \le \exp(-t/\kappa)||w^0 - w^*||_2^2$ 

where  $\kappa = \beta / \alpha$  is the condition number of f, and  $w^* = \arg \min_{w \in \mathbb{R}^d} f(w)$  is the minimizer of f.

#### E TRAINING PROCEDURE FOR IN-CONTEXT BINARY CLASSIFICATION

In this section, we present the proof of Theorem 3.1.

#### E.1 PROOF SKETCH

First, we prove in Lemma E.2 that the expected loss function L(W) in (7) is strictly convex w.r.t. W and is strongly convex in a compact set of  $\mathbb{R}^{d \times d}$ . Moreover, we prove L(W) has one unique global minimizer  $W^*$ . Then, in Lemma E.3, by analyzing the Taylor expansion of L(W), we prove that as  $N \to \infty$ , our loss function L(W) point wisely converges to L(W) (defined in (25)), and the global minimizer  $W^*$  converge to  $2\Lambda^{-1}$ . We denote  $W^* = 2(\Lambda^{-1} + G)$ , and prove  $||G||_{\max} = O(N^{-1/2})$ . Next, in Lemma E.4, by further analyzing the Taylor expansion of the equation  $\nabla L(W^*) = 0$  at the point  $2\Lambda^{-1}$ , we establish a tighter bound  $||G||_{\max} = O(N^{-1})$ . In Lemma E.5, we prove that our loss function is *l*-smooth and provide an upper bound for *l*. Thus, in a compact set  $R_W$ , our loss function is  $\alpha$ -strongly convex and *l*-smooth. Finally, leveraging the standard results from convex optimization, we prove Theorem 3.1 in subsection E.4.

In this section, we use the following notations.

#### E.2 NOTATIONS

Recall the expected loss function (7) is 

$$L(W) = -\frac{1}{2} \mathbb{E}\left[ (1 + y_{\tau,\mathsf{query}}) \log(\widehat{y}_{\tau,\mathsf{out}}) + (1 - y_{\tau,\mathsf{query}}) \log(1 - \widehat{y}_{\tau,\mathsf{out}}) \right], \tag{21}$$

where

 is the output of the transformer, and the label of the data follows the distribution

$$\mathbb{P}\left(y_{\tau,\mathsf{query}} = 1 | x_{\tau,\mathsf{query}}\right) = \sigma((\mu_{\tau,1} - \mu_{\tau,0})^\top \Lambda^{-1} x_{\tau,\mathsf{query}})).$$

 $\widehat{y}_{\tau,\mathsf{out}} = \sigma \left( \left( \frac{2}{N} \sum_{i=1}^{N} y_{\tau,i} x_{\tau,i}^{\top} \right) \frac{W}{2} x_{\tau,\mathsf{query}} \right)$ 

In this section, we introduce the following notations to analyze (7). We denote  $\mu = \mu_{\tau}, \mu_1 = \mu_{\tau,1}$ ,  $\mu_0 = \mu_{\tau,0}$  and  $q = x_{\tau,query}$ . Then with probability  $\mathbb{P}(y_{\tau,query} = 1) = 1/2$  we have  $q = \mu_1 + v$ , and with probability  $\mathbb{P}(y_{\tau,query}=0) = 1/2$  we have  $q = \mu_0 + v$ , where  $v \sim N(0, \Lambda)$ . We define  $p = \frac{2}{N} \sum_{i=1}^{N} y_{\tau,i} x_{\tau,i}.$  Since with probability  $\mathbb{P}(y_{\tau,i} = 1) = 1/2$  we have  $x_{\tau,i} = \mu_1 + v_i$ , and with probability  $\mathbb{P}(y_{\tau,i} = 0) = 1/2$  we have  $x_{\tau,i} = \mu_0 + v_i$ , where  $v_i \sim \mathsf{N}(0,\Lambda)$ , we known  $p = 2N_1\mu_1/N - 2N_0\mu_0/N + g$ , where  $g = \frac{2}{N} \sum_{i=1}^{N} v_i, g \sim \mathsf{N}(0, 4\Lambda/N), N_1 \sim \mathsf{Bin}(N, 1/2).$  Defining  $h = N_1/N - 1/2, u = 2(\mu_1 + \mu_0)$ , we have  $N_0/N = 1/2 - h$  and 

$$p = \mu + hu + g. \tag{22}$$

Then, the expected loss function (7) can be expressed as

$$L(W) = \mathbb{E}[-\sigma(\mu^{\top}\Lambda^{-1}q)\log(\sigma(p^{\top}Wq/2)) - (1 - \sigma(\mu^{\top}\Lambda^{-1}q))\log(1 - \sigma(p^{\top}Wq/2))].$$
(23)

The gradient of the loss function (7) can be expressed as

$$\nabla L(W) = \frac{1}{2} \mathbb{E}[(\sigma(p^{\top}Wq/2) - \sigma(\mu^{\top}\Lambda^{-1}q))pq^{\top}].$$
(24)

Moreover, we define a function L(W) as

$$\widetilde{L}(W) = \mathbb{E}\left[-\sigma(\mu^{\top}\Lambda^{-1}q)\log(\sigma(\mu^{\top}Wq/2)) - (1 - \sigma(\mu^{\top}\Lambda^{-1}q))\log(1 - \sigma(\mu^{\top}Wq/2))\right].$$
 (25)

In Lemma E.3, we show that as  $N \to \infty$ , L(W) will point wisely converge to  $\widetilde{L}(W)$ .

## 810 E.3 LEMMAS

**Lemma E.1** Suppose  $N_1 \sim Bin(N, 1/2)$ . Defining  $h = N_1/N - 1/2$ , we have

814  $\mathbb{E}[h] = 0$ 815  $\mathbb{E}[h^2] = \frac{1}{4N}$ 

 $\mathbb{E}[h^3] = 0$  $\mathbb{E}[h^n] = O(N^{-2}), \text{ for } n > 4$ 

$$[h^n] = O(N^{-2}), \text{ for } n \ge 1$$

$$\mathbb{E}[|h|] \le \frac{1}{2N^{1/2}}$$

$$\mathbb{E}[|h^3|] = O(N^{-3/2}).$$

**Proof** Since  $N_1 \sim Bin(N, 1/2)$ , the moment-generating function of  $N_1$  is

$$M_{N_1}(t) = \left(\frac{1}{2} + \frac{1}{2}\exp(t)\right)^N$$

We can compute the moment-generating function of h as follows:

$$M_h(t) = \exp\left(-\frac{t}{2}\right) M_{N_1}\left(\frac{t}{N}\right) = \left(\frac{\exp\frac{-t}{2N} + \exp\frac{t}{2N}}{2}\right)^N = \left(\cosh\left(\frac{t}{2N}\right)\right)^N$$
$$= \left(1 + \frac{t^2}{8N^2} + \sum_{i=2}^{\infty} \frac{t^{2i}}{(2i)!(2N)^{2i}}\right)^N.$$

Thus, we know the coefficients of t,  $t^2$ ,  $t^3$  are 0, 1/(8N), 0 respectively, and the coefficients of  $t^n, n \ge 4$  are  $O(1/N^2)$ . We have

$$\begin{split} \mathbb{E}[h] &= 0\\ \mathbb{E}[h^2] &= \frac{1}{4N}\\ \mathbb{E}[h^3] &= 0\\ \mathbb{E}[h^n] &= O(1/N^2), \text{ for } n \geq 4. \end{split}$$

Moreover, according to the Jensen's inequality, we have

$$\mathbb{E}[|h|] \le \left(\mathbb{E}[h^2]\right)^{1/2} = \frac{1}{2N^{1/2}}$$
$$\mathbb{E}[|h^3|] \le \left(\mathbb{E}[h^4]\right)^{3/4} = O(N^{-3/2}).$$

**Lemma E.2** For the loss function L(W) (7), we have  $\nabla^2 L(W) \succ 0$ . For any compact set  $R_W$  of  $\mathbb{R}^{d \times d}$ , when  $W \in R_W$ , we have  $\nabla^2 L(W) \succ \gamma I_d$  for some  $\gamma > 0$ . Additionally, L(W) has one unique global minimizer on  $\mathbb{R}^{d \times d}$ .

For  $\widetilde{L}(W)$  defined in (25), we also have  $\nabla^2 \widetilde{L}(W) \succ 0$ . For any compact set  $R_W$  of  $\mathbb{R}^{d \times d}$ , when W  $\in R_W$ , we have  $\nabla^2 \widehat{L}(W) \succ \gamma I_d$  for some  $\gamma > 0$ . Additionally,  $\widetilde{L}(W)$  has one unique global minimizer on  $\mathbb{R}^{d \times d}$ .

Proof We vectorize W as  $\operatorname{Vec}(W) \in \mathbb{R}^{d^2}$ , where  $\operatorname{Vec}(W)_i = W_{t_1(i), t_2(i)}, t_1(x) = \lfloor (x-1)/d \rfloor + 1, t_2(x) = ((x-1) \mod d) + 1$ . Then, we have

$$(\nabla L(W))_i = \mathbb{E}_{p,q} \left[ \frac{1}{2} (\sigma(p^\top W q/2) - \sigma(\mu^\top \Lambda^{-1} q)) p_{t_1(i)} q_{t_2(i)} \right].$$
(26)

The Hessian matrix of the loss function (7) is

$$(\nabla^2 L(W))_{ij} = \mathbb{E}_{p,q} \left[ \frac{1}{4} \sigma(p^\top W q/2) (1 - \sigma(p^\top W q/2)) p_{t_1(i)} q_{t_2(i)} p_{t_1(j)} q_{t_2(j)} \right].$$

Considering  $z \in \mathbb{R}^{d^2}$  such that  $z \neq 0$ , we have

Т

$$z^{\top} \nabla^{2} L(W) z = \mathbb{E}_{q,p} \left[ \frac{1}{4} \sigma(p^{\top} Wq/2) (1 - \sigma(p^{\top} Wq/2)) \sum_{ab} z_{a} z_{b} p_{t_{1}(a)} q_{t_{2}(a)} p_{t_{1}(b)} q_{t_{2}(b)} \right]$$
$$= \int \frac{1}{4} \sigma(p^{\top} Wq/2) (1 - \sigma(p^{\top} Wq/2)) \left( \sum_{a \in [d^{2}]} z_{a} p_{t_{1}(a)} q_{t_{2}(a)} \right)^{2} f_{pq}(p,q) dp dq,$$

where  $f_{pq}(p,q)$  are the probability density function (PDF) function of p,q. Since for any p,q,  $\sigma(p^{\top}Wq/2)(1 - \sigma(p^{\top}Wq/2)) > 0$ , we have  $z^{\top}\nabla^2 L(W)z \ge 0$ . Thus,  $\nabla^2 L(W) \succeq 0$  and L(W) is convex.

881 Moreover, for any  $z \neq 0$ , we denote  $z_{ij} = z_{((i-1)d+j)}$ ,  $i, j \in [d]$ . Suppose  $a, b \in \arg \max_{i,j} |z_{ij}|$ , 882 we consider a set of constants  $\{c_{1pi}, c_{2pi}\}, \{c_{1qi}, c_{2qi}\}, i, j \in [d]$ , where  $c_{1pa} = d, c_{2pa} = d + 1$ , 883  $c_{1qb} = d, c_{2qb} = d + 1$ , and  $c_{1pi} = 1/16, c_{2pi} = 1/8, i \neq a, c_{1qj} = 1/16, c_{2qj} = 1/8, j \neq b$ . Then, 884 for any  $c_{pi} \in [c_{1pi}, c_{2pi}], c_{qj} \in [c_{1qj}, c_{2qj}]$ . We have

$$\left| \sum_{i,j \in [d]} z_{ij} c_{pi} c_{qj} \right| \ge \left[ d^2 - 2(d+1)(d-1)/8 - (d-1)^2/64 \right] \max_{ij} |z_{ij}| \ge d^2 \max_{ij} |z_{ij}|/2.$$

Then, we define region  $\Omega(a,b) \triangleq \{p = \sum_i c_{pi} \mathbf{e}_i, q = \sum_j c_{qj} \mathbf{e}_j, c_{pi} \in [c_{1pi}, c_{2pi}], c_{qj} \in [c_{1qj}, c_{2qj}]\}$ . We have

$$\min_{\Omega(a,b)} \left( \sum_{c \in [d^2]} z_c p_{t_1(c)} q_{t_2(c)} \right)^2 \ge d^4 \max_{ij} |z_{ij}|^2 / 4 \ge ||z||_2^2 / 4$$

Defining

I.

$$C(\Omega) = \min_{a \in [d], b \in [d]} \int_{\Omega(a,b)} f_{pq}(p,q) dp dq,$$
  
$$S(\Omega, W) = \min_{a \in [d], b \in [d]} \min_{\Omega(a,b)} \left\{ \frac{1}{4} \sigma(p^{\top} Wq/2) (1 - \sigma(p^{\top} Wq/2)) \right\},$$

ſ

902 we have  $S(\Omega, W) > 0$ . Since with probability  $\mathbb{P}(y_{\tau,query} = 1) = 1/2, q = \mu_1 + v$ , with probability 903  $\mathbb{P}(y_{\tau,query} = 0) = 1/2, q = \mu_0 + v$ , where  $v \sim N(0, \Lambda)$  and  $p = \mu + hu + g$ , where  $g \sim N(0, 4\Lambda/N), v \sim N(0, \Lambda), \mu_0 \sim N(0, I_d)$ , the covariance matrices of p, q are positive definite and 904 we have  $f_{pq}(p,q) > 0$  for all  $p, q \in \mathbb{R}^d$ . Moreover,  $\Omega(a, b)$  are non-zero measures on  $\mathbb{R}^{d \times d}$ . Thus, 906 we have  $C(\Omega) > 0$ . Then, for any  $z \neq 0$ , we have

$$z^{\top} \nabla^{2} L(W) z \geq \int_{\Omega(a,b)} \frac{1}{4} \sigma(p^{\top} W q/2) (1 - \sigma(p^{\top} W q/2)) \left(\sum_{l} z_{l} p_{t_{1}(l)} q_{t_{2}(l)}\right)^{2} f_{pq}(p,q) dp dq$$
$$\geq C(\Omega) S(\Omega, W) \|z\|_{2}^{2}/4$$
$$> 0.$$

913 Thus, we have  $\nabla^2 L(W) \succ 0$ . L(W) is strictly convex. 

915 Moreover, for any compact set  $R_W$  of  $\mathbb{R}^{d \times d}$ , for any  $W \in R_W$ , we have

$$S(\Omega) = \min_{W \in R_W} \min_{a \in [d], b \in [d]} \min_{\Omega(a,b)} \left\{ \frac{1}{4} \sigma(p^\top W q/2) (1 - \sigma(p^\top W q/2)) \right\} > 0$$

Then, for any  $W \in R_W$ , for any  $z \neq 0$ , we have

$$z^{\top} \nabla^2 L(W) z \ge \int_{\Omega(a,b)} \frac{1}{4} \sigma(p^{\top} Wq/2) (1 - \sigma(p^{\top} Wq/2)) \left(\sum_l z_l p_{t_1(l)} q_{t_2(l)}\right)^2 f_{pq}(p,q) dp dq$$

$$\geq \frac{1}{4}C(\Omega)S(\Omega)\|z\|_2^2.$$

Thus, when  $W \in R_W$ , where  $R_W$  is a compact set, we have  $\nabla^2 L(W) \succ C(\Omega)S(\Omega)I_d/4$  and the loss function L(W) is  $\gamma$ -strongly convex, where  $\gamma = C(\Omega)S(\Omega)/4$ .

Because our loss function is strictly convex in  $\mathbb{R}^{d \times d}$ , it has at most one global minimizer in  $\mathbb{R}^{d \times d}$ . Next, we prove all level sets of our loss function are compact, i.e.  $V_{\alpha} = \{W \in \mathbb{R}^{d \times d} | L(W) \leq \alpha\}$ is compact for all  $\alpha$ . We prove it by contradiction. Suppose  $V_{\alpha}$  is not compact for some  $\alpha$ . Since our loss function is continuous and convex,  $V_{\alpha}$  is an unbounded convex set. Since the dimension of  $V_{\alpha}$  is  $d^2$ , consider a point  $W^{\alpha} \in V_{\alpha}$ , there must exists a  $W^k \neq 0_{d \times d}$  such that  $\{W^{\alpha} + tW^k | t = [0, \infty)\} \in V_{\alpha}$ . For this  $W^k \neq 0_{d \times d}$ , there must exist a set of constants  $0 < c_{3pi} < c_{4pi}, 0 < c_{3qj} < c_{4qj}$  such that for any  $c_{pi} \in [c_{3pi}, c_{4pi}], c_{qj} \in [c_{3qj}, c_{4qj}]$ , we have

$$|\sum_{ij} c_{pi} c_{qj} W_{ij}^k| \neq 0$$

Thus, we have

$$\lim_{t \to \infty} |\sum_{ij} c_{pi} c_{qj} (W_{ij}^{\alpha} + t W_{ij}^{k})| = \infty.$$

We define  $\Omega_0 = \{p = \sum_i c_{pi} \mathbf{e}_i, q = \sum_j c_{qj} \mathbf{e}_j, c_{pi} \in [c_{3pi}, c_{4pi}], c_{qj} \in [c_{3qj}, c_{4qj}], \|\mu\|_2^2 \leq \sum_i c_{4pi}^2 + c_{4qj}^2 \}$ . Then, defining

$$\begin{split} C(\Omega_0) &= \int_{\Omega_0} f_{pq}(p,q) dp dq, \\ S(\Omega_0) &= \min_{\Omega_0} \left\{ \min\{\sigma(\mu^\top \Lambda^{-1}q), (1 - \sigma(\mu^\top \Lambda^{-1}q))\} \right\} \end{split}$$

we have  $S(\Omega_0) > 0$ . Since  $\Omega_0$  are non-zero measures for p, q, we have  $C(\Omega_0) > 0$ . Then, we have

$$\begin{split} &\lim_{t\to\infty} L(W^{\alpha} + tW^{k}) \\ &= \lim_{t\to\infty} \mathbb{E}[-\sigma(\mu^{\top}\Lambda^{-1}q)\log(\sigma(p^{\top}(W^{\alpha} + tW^{k})q/2)) - (1 - \sigma(\mu^{\top}\Lambda^{-1}q))\log(1 - \sigma(p^{\top}(W^{\alpha} + tW^{k})q/2))] \\ &\geq \lim_{t\to\infty} \int_{\Omega_{0}} [-\sigma(\mu^{\top}\Lambda^{-1}q)\log(\sigma(\sum_{ij}c_{pi}c_{qj}(W^{\alpha}_{ij} + tW^{k}_{ij})/2))]f_{pq}(p,q)dpdq \\ &\quad + \lim_{t\to\infty} \int_{\Omega_{0}} [-(1 - \sigma(\mu^{\top}\Lambda^{-1}q))\log(1 - \sigma(\sum_{ij}c_{pi}c_{qj}(W^{\alpha}_{ij} + tW^{k}_{ij})/2))]f_{pq}(p,q)dpdq \\ &\geq C(\Omega_{0})S(\Omega_{0}) \cdot \min_{\Omega_{0}} \left\{ \lim_{t\to\infty} [-\log(\sigma(\sum_{ij}c_{pi}c_{qj}(W^{\alpha}_{ij} + tW^{k}_{ij})/2))] \right\} \\ &\quad + C(\Omega_{0})S(\Omega_{0}) \cdot \min_{\Omega_{0}} \left\{ \lim_{t\to\infty} [-\log(1 - \sigma(\sum_{ij}c_{pi}c_{qj}(W^{\alpha}_{ij} + tW^{k}_{ij})/2))] \right\} \\ &= \infty. \end{split}$$

This contradicts the assumption  $L(W^{\alpha} + tW^k) \leq \alpha$ . Thus, all level sets of the loss function L(W) are compact, which means there exists a global minimizer for L(W). Together with the fact that L(W) is strictly convex, L(W) has one unique global minimizer on  $\mathbb{R}^{d \times d}$ .

Similarly, we can prove the same conclusions for L(W).

972

**Lemma E.3** Denoting the global minimizer of the loss function (7) as  $W^*$ , we have  $W^* = 2(\Lambda^{-1} +$ 973 G), where  $||G||_{\max} = O(N^{-1/2})$ . 974 **Proof** Let  $a = \mu^{\top} \Lambda^{-1} q$ ,  $s = \mu^{\top} W q/2$ ,  $r = (hu + q)^{\top} W q/2$ . Performing the Taylor expansion 975 on (7), we have 976  $L(W) = \mathbb{E}\left[-\sigma(a)\log(\sigma(s+r)) - (1 - \sigma(a))\log(1 - \sigma(s+r))\right]$ 977 978  $= \mathbb{E}\left[-\sigma(a)\log(\sigma(s)) - (1 - \sigma(a))\log(1 - \sigma(s))\right]$ 979  $-\mathbb{E}\left[\left(\sigma(a)(1-\sigma(s))-(1-\sigma(a))\sigma(s)\right)\right)r\right]$ 980 +  $\mathbb{E}\left[\sigma(\xi(s,r))(1-\sigma(\xi(s,r)))r^2/2\right]$ 981  $=\widetilde{L}(W) - \mathbb{E}\left[\left(\sigma(a)(1 - \sigma(s)) - (1 - \sigma(a))\sigma(s)\right)\right)r\right]$ 982 983 +  $\mathbb{E}\left[\sigma(\xi(s,r))(1-\sigma(\xi(s,r)))r^2/2\right]$ , 984 where  $\xi(s, r)$  are real numbers between s and s + r. According to Lemma E.1, we have  $\mathbb{E}[r] =$ 985  $\mathbb{E}\left[(hu+g)^{\top}Wq/2\right] = 0$ . Thus, we have 986  $\mathbb{E}\left[\left(\sigma(a)(1-\sigma(s)) - (1-\sigma(a))\sigma(s)\right)\right)r\right] = \mathbb{E}_{\mu,u,q}\left[\left(\sigma(a)(1-\sigma(s)) - (1-\sigma(a))\sigma(s)\right)\right)\mathbb{E}_{q,h}[r]\right] = 0.$ 987 Moreover, we have 988 989  $\mathbb{E}\left[\sigma(\xi(s,r))(1-\sigma(\xi(s,r))r^2/2\right]$ 990  $< \mathbb{E}\left[r^2\right]$ 991  $= \mathbb{E}[h^2 u^\top W q u^\top W q + q^\top W q q^\top W q]$ 992 993  $\stackrel{(a)}{=} \mathbb{E}[u^{\top}Wqu^{\top}Wq/(4N) + 4(\Lambda Wq)^{\top}Wq/N]$ 994  $\leq C_l \|W\|_{\max}^2 / N,$ 995 where (a) is due to Lemma E.1,  $g^{\top}Wqg^{\top}Wq = \sum_{i,j,k,l \in [d]} g_i W_{ij} q_j g_k W_{kl} q_l$ 996  $\sum_{i,j,k,l \in [d]} g_i g_k W_{kl} q_l W_{ij} q_j = (gg^\top Wq)^\top Wq \text{ and } \mathbb{E}[gg^\top] = 4\Lambda/N. C_l \text{ is a constant independent of } C_l = 0$ 997 998 dent of N and W. Thus, we have 999  $\left|\widetilde{L}(W) - L(W)\right| \le C_l \|W\|_{\max}^2 / N.$ 1000 1001 This shows that L(W) point wisely converges to  $\widetilde{L}(W)$ . 1002 According to Lemma E.2,  $\tilde{L}(W)$  has one unique global minimizer. Consider the equation: 1003 1004  $\nabla \widetilde{L}(W) = \mathbb{E}[\sigma(\mu^{\top}Wq/2) - \sigma(\mu^{\top}\Lambda^{-1}q)] = 0.$ 1005 We can easily find that  $\nabla \widetilde{L}(2\Lambda^{-1}) = 0$  and  $W = 2\Lambda^{-1}$  is the global minimizer of  $\widetilde{L}(W)$ . 1006 Considering a compact set  $R_W = \{W \mid ||W - 2\Lambda^{-1}||_F \leq \rho_W\}$ , we have  $||W||_{\max} \leq C_W$  for  $W \in R_W$ . Here  $\rho_W, C_W$  are some positive finite constants. Then, we have 1008 1009  $\left|\widetilde{L}(W) - L(W)\right| \le C_l'/N, \ W \in R_W,$ 1010 where  $C'_l = C_l C_W^2$  is a constant independent of N and W. This shows that, for  $W \in R_W$ , our loss 1011 1012 function L(W) uniformly converge to L(W). 1013 Denote  $W^*$  as the global minimizer of the loss function L(W) with prompt length N. Then, we 1014 show that, when N is sufficiently large,  $W^* \in R_W$ . We first denote  $\partial R_W = \{W \mid ||W - 2\Lambda^{-1}||_F = 0\}$ 1015  $\rho_W$  and  $\Delta = \min_{W \in \partial R_W} \widetilde{L}(W) - \widetilde{L}(2\Lambda^{-1}) > 0$ . Then, for  $N \ge 4C_l'/\Delta$ , and for any  $W \in R_W$ , 1016 we have 1017  $\left|\widetilde{L}(W) - L(W)\right| \le \Delta/4,$ 1018 This means  $\min_{W \in \partial R_W} L(W) - \min_{W \in R_W} L(W)$ 1020 1021  $\geq \min_{W \in \partial B_W} L(W) - L(2\Lambda^{-1})$  $\geq \min_{W \in \partial R_W} \widetilde{L}(W) - \widetilde{L}(2\Lambda^{-1}) - \Delta/2$ 1023 1024 1025  $>\Delta/2 > 0.$ 

1026 Since L(W) is strictly convex, we have  $W^* = \arg \min_W L(W) \in R_W$ . 1027 Then, we have 1028 1029  $|\widetilde{L}(W^*) - L(W^*)| < C_1'/N$ 1030  $|\widetilde{L}(2\Lambda^{-1}) - L(2\Lambda^{-1})| < C_1'/N$ 1031 1032  $\widetilde{L}(W^*) \le L(W^*) + C_l'/N \le L(2\Lambda^{-1}) + C_l'/N \le \widetilde{L}(2\Lambda^{-1}) + 2C_l'/N.$ 1033 1034 According to Lemma E.2, for  $W \in R_W$ , we have  $\nabla^2 \widetilde{L}(W) \succ \gamma I_d$ , where  $\gamma$  is a positive constant 1035 independent of N. Thus,  $\hat{L}(W)$  is  $\gamma$ -strongly convex in  $R_W$ . According to Lemma D.1, we have 1036  $||W^* - 2\Lambda^{-1}||_F^2 \le \frac{2}{2} (\widetilde{L}(W^*) - \widetilde{L}(2\Lambda^{-1})) \le \frac{4C_l'}{2N}.$ 1037 1038 1039 Thus, when  $N \to \infty$ , we have  $W^* \to 2\Lambda^{-1}$ . Denoting  $W^* = 2(\Lambda^{-1} + G)$ , we have  $\|G\|_{\max} =$ 1040  $O(1/\sqrt{N}).$ 1041 1042 **Lemma E.4** The global minimizer of the loss function (7) is  $W^* = 2(\Lambda^{-1} + G)$ , where 1043 1044  $\|G\|_{\max} \leq \frac{1}{N} \|S^{-1}(\mathbb{E}[\sigma'(a)(4qq^{\top} + uu^{\top}\Lambda^{-1}qq^{\top}/4)$ 1045 1046  $+ \sigma''(a)((u^{\top}\Lambda^{-1}q)^{2}\mu q^{\top}/8 + 2q^{\top}\Lambda^{-1}q\mu q^{\top})])\|_{\max} + o(1/N).$ 1047  $a = u^{\top} \Lambda^{-1} q, S = 4 \nabla^2 \widetilde{L}(2 \Lambda^{-1}).$ 1048 1049 **Proof** According to Lemma E.2, the loss function L(W) has a unique global minimizer  $W^*$ . We 1050 have 1051 1052  $\nabla L(W^*) = \mathbb{E}\left[ (\sigma(p^\top W^* q/2) - \sigma(\mu^\top \Lambda^{-1} q)) p q^\top \right] = 0.$ (27)1053 Let  $W^* = 2(\Lambda^{-1} + G), a = \mu^{\top} \Lambda^{-1} q, b = (\mu + hu + g)^{\top} G g + (hu + g)^{\top} \Lambda^{-1} g$ . We have 1054 1055  $n^{\top}W^*a/2$ 1056  $=(\mu + hu + q)^{\top}(\Lambda^{-1} + G)q$ 1057  $=(\mu + hu + q)^{\top}Gq + (hu + q)^{\top}\Lambda^{-1}q + \mu^{\top}\Lambda^{-1}q = a + b.$ 1058 The Taylor expansion of  $\sigma(a+b)$  at point a with an Lagrange form of remainder is  $\sigma(a+b)pq^{\top} = \sigma(a)pq^{\top} + \sigma'(a)bpq^{\top} + \frac{\sigma''(a)}{2}b^2pq^{\top} + \frac{\sigma'''(\xi(a,b))}{2!}b^3pq^{\top},$ 1061 1062 where  $\xi(a, b)$  are real numbers between a and a + b. Thus, our equation (27) become 1063 1064  $\mathbb{E}_{\mu,u,g,h,q}\left[\sigma'(a)bpq^{\top} + \frac{\sigma''(a)}{2}b^2pq^{\top} + \frac{\sigma'''(\xi(a,b))}{3!}b^3pq^{\top}\right] = 0.$ (28)

Note that  $\mathbb{E}[\sigma'(a)bpq^{\top}] = \mathbb{E}_{\mu,u,q} \left[\sigma'(a)\mathbb{E}_{g,h}\left[bpq^{\top}\right]\right]$ . For  $\mathbb{E}_{g,h}[bpq^{\top}]$ , according to Lemma E.1 and  $g \sim N(0, 4\Lambda/N)$ , we have

$$\mathbb{E}_{g,h}[bpq^{\top}] = \mathbb{E}[\mu^{\top}Gq\mu q^{\top} + g^{\top}\Lambda^{-1}qgq^{\top} + g^{\top}Gqgq^{\top} + h^{2}u^{\top}Gquq^{\top} + h^{2}u^{\top}\Lambda^{-1}quq^{\top}] = \mu\mu^{\top}Gqq^{\top} + 4qq^{\top}/N + 4\Lambda Gqq^{\top}/N + uu^{\top}Gqq^{\top}/(4N) + uu^{\top}\Lambda^{-1}qq^{\top}/(4N).$$
(29)

1074 Then, we have

1070 1071 1072

1075 1076

1079

$$\|\mathbb{E}_{\mu,u,q}[\sigma'(a)(4\Lambda Gqq^{+}/N + uu^{+}Gqq^{+}/(4N))]\|_{\max} \le c_{1}\|G\|_{\max}/N,$$

where  $c_1 = \max_{ij} |\mathbb{E} \left[ \sum_{kl} 4\sigma'(a) \left( \Lambda_{ik} q_l q_j \right) + \sum_{kl} \sigma'(a) \left( u_i u_k q_l q_j / 4 \right) \right] |$  is a constant independent of N. According to Lemma E.3,  $||G||_{\max} = O(1/\sqrt{N}) = o(1)$ , we have

$$\|\mathbb{E}_{\mu,u,q}[\sigma'(a)(4\Lambda Gqq^{\top}/N + uu^{\top}Gqq^{\top}/(4N))]\|_{\max} = o(1/N),$$
(30)

1080 Similarly for  $\mathbb{E}[\sigma''(a)b^2pq^{\top}/2]$ , we have 1081  $\mathbb{E}_{a,h}[b^2 p q^{\top}]$ 1082  $=\underbrace{\mathbb{E}[\mu^{\top}Gq\mu^{\top}Gq\mu q^{\top} + h^{2}u^{\top}Gqu^{\top}Gq\mu q^{\top} + g^{\top}Gqg^{\top}Gq\mu q^{\top} + 2h^{2}u^{\top}Gq\mu^{\top}Gquq^{\top} + 2g^{\top}Gq\mu^{\top}Gqgq^{\top}]}_{(2)}$ 1084  $+\underbrace{\mathbb{E}[2h^{2}u^{\top}Gqu^{\top}\Lambda^{-1}q\mu q^{\top}+2g^{\top}Gqg^{\top}\Lambda^{-1}q\mu q^{\top}+2h^{2}\mu^{\top}Gqu^{\top}\Lambda^{-1}quq^{\top}+2\mu^{\top}Gqg^{\top}\Lambda^{-1}qgq^{\top}]}_{(ii)}$ 1087 1088  $+\underbrace{\mathbb{E}[h^2 u^{\top} \Lambda^{-1} q u^{\top} \Lambda^{-1} q \mu q^{\top} + g^{\top} \Lambda^{-1} q g^{\top} \Lambda^{-1} q \mu q^{\top}]}_{(iii)}.$ 1089 1090 1091 For each term in (i), it contains two G. Thus, their max norms are at most smaller than  $O(||G||^2_{\text{max}})$ . 1092 For each term in (*ii*), it contains one G and  $h^2$  or contains one G and two g. According to  $\mathbb{E}[h^2] =$ 1093 1/(4N) in Lemma E.1, the max norm of terms with one G and  $h^2$  are smaller than  $O(\|G\|_{\max}/N)$ . 1094 Defining  $\bar{q} = N^{1/2} \Lambda^{-1/2} q/2$ , we have  $\bar{q} \sim N(0, I_d)$  and  $q = 2N^{-1/2} \Lambda^{1/2} \bar{q}$ . Thus, converting two 1095 g to  $\bar{g}$ , we have a coefficient of  $N^{-1}$ . Therefore, the max norms of terms with one G and two g are 1096 also smaller than  $O(||G||_{\max}/N)$ . Therefore, for terms (i), (ii), we have 1097  $\|\mathbb{E}[\sigma''(a)(i)/2]\|_{\max} \le O(\|G\|_{\max}^2) = o(\|G\|_{\max}),$ (31)1099  $\|\mathbb{E}[\sigma''(a)(ii)/2]\|_{\max} \le O(\|G\|_{\max}/N) = o(1/N).$ (32)1100 For term (*iii*), according to Lemma E.1 and  $g \sim N(0, 4\Lambda/N)$ , we have 1101 1102  $\|\mathbb{E}[\sigma''(a)(iii)/2]\|_{\max}$ 1103  $= \|\mathbb{E}\left[\sigma''(a)(h^2 u^{\top} \Lambda^{-1} q u^{\top} \Lambda^{-1} q \mu q^{\top} + q^{\top} \Lambda^{-1} q q^{\top} \Lambda^{-1} q \mu q^{\top})/2\right]\|_{\max}$ (33)1104  $= \frac{1}{N} \|\mathbb{E}\left[\sigma^{\prime\prime}(a)((u^{\top}\Lambda^{-1}q)^{2}\mu q^{\top}/8 + 2q^{\top}\Lambda^{-1}q\mu q^{\top})\right]\|_{\max}.$ 1105 (34)1106 1107 For  $\mathbb{E}[\sigma'''(\xi(a,b))b^3pq^{\top}/3!]$ , we have 1108 1109  $\|\mathbb{E}[\sigma'''(\xi(a,b))b^3pq^{\top}/3!]\|_{\max}$ 1110  $\leq \max_{z \in \mathbb{D}} |\sigma^{\prime\prime\prime}(z)|/3! \cdot \max_{ij} \mathbb{E}\left[ |b^3 p_i q_j| \right]$ 1111 1112 1113 1114 1115 1116 1117 1118 1119 1120 1121 1122 1123 1124 1125 1126 1127 1128 For terms in (\*) containing two or three G, these terms' expected absolute values are at most smaller 1129

than  $O(\|G\|_{\max}^2)$ . For terms in (\*\*) containing two of the only index on the contains expected absolute values are at most smaller than  $O(\|G\|_{\max}^2)$ . For terms in (\*\*) containing one G, these terms must contain  $n_1$  number of hand  $n_2$  number of elements of g, where  $n_1 + n_2 = 2, 3, 4, n_1, n_2 \in \mathbb{N}$ . According to Lemma E.1, we know that for  $n_1 = 1, 2, 3, 4, \mathbb{E}[h^{n_1}] \leq O(N^{-n_1/2})$ . Defining  $\bar{g} = N^{1/2} \Lambda^{-1/2} g/2$ , we have  $\bar{g} \sim N(0, I_d)$  and  $g = 2N^{-1/2} \Lambda^{1/2} \bar{g}$ . Converting g to  $\bar{g}$ , we have a coefficient of  $N^{-n_2/2}$ . Thus, for terms in (\*\*), these terms' expected absolute values are at most smaller than  $O(\|G\|_{\max}N^{-(n_1+n_2)/2}) \le O(\|G\|_{\max}N^{-1})$ . For terms in (\*\*\*) without G, these terms must 1135 contain  $n_1$  number of h and  $n_2$  number of elements of g, we have  $n_1+n_2=3, 4, n_1, n_2 \in \mathbb{N}$ . Simi-1136 larly, these term's expected absolute values are at most smaller than  $O(N^{-(n_1+n_2)/2}) \le O(N^{-3/2})$ . 1137 Therefore, we have

$$\begin{split} &\|\mathbb{E}[\sigma^{\prime\prime\prime}(\xi(a,b))b^{3}pq^{\top}/3!]\|_{\max} \\ &\leq \max_{ij} \mathbb{E}\left[|b^{3}p_{i}q_{j}|\right] \cdot \max_{z} |\sigma^{\prime\prime\prime\prime}(z)|/3! \\ &= O(\|G\|_{\max}^{2}) + O(\|G\|_{\max}/N) + O(1/N^{-3/2}) \\ &= o(\|G\|_{\max}) + o(1/N). \end{split}$$
(35)

1145 Moreover, we have

$$\left\{\mathbb{E}_{\mu,u,q}[\sigma'(a)\mu\mu^{\top}Gqq^{\top}]\right\}_{ij} = \sum_{kl} s_{ijkl}G_{kl},$$
(36)

where  $s_{ijkl} = \mathbb{E}\sigma'(a)\mu_i\mu_kq_lq_j$ . We vectorize G as  $\operatorname{Vec}(G)_i = G_{t_1(i),t_2(i)}$ . Define  $S \in \mathbb{R}^{d^2 \times d^2}$ , where  $S_{ij} = s_{t_1(i),t_2(i),t_1(j),t_2(j)} = \mathbb{E}\sigma'(a)\mu_{t_1(i)}q_{t_2(i)}\mu_{t_1(j)}q_{t_2(j)}$ . Then (36) can be expressed as

$$\left\{\mathbb{E}_{\mu,\nu}[\sigma'(a)\mu\mu^{\top}Gqq^{\top}]\right\} = SG.$$
(37)

Note that  $S = 4\nabla^2 \tilde{L}(2\Lambda^{-1})$ . According to Lemma E.2, S is positive definite. Thus, combining (28), (29), (30), (31), (32), (34), (35), (37), we have

1165 Lemma E.5 The loss function (7) is *l*-smooth, where  $l \leq \frac{1}{4} \sum_{i \in [d^2]} \mathbb{E}[(p_{t_1(i)}q_{t_2(i)})^2]$ .

**Proof** The Hessian matrix of the loss function is

$$(\nabla^2 L(W))_{ij} = \frac{1}{4} \mathbb{E}[\sigma(p^\top Wq/2)(1 - \sigma(p^\top Wq/2))p_{t_1(i)}q_{t_2(i)}p_{t_1(j)}q_{t_2(j)}].$$

1170 Considering  $z \in \mathbb{R}^{d^2}$  such that  $z \neq 0$ , we have

$$z^{\top} \nabla^2 L(W) z = \mathbb{E} \left[ \frac{1}{4} \sigma(p^{\top} W q/2) (1 - \sigma(p^{\top} W q/2)) \sum_{ab} z_a z_b p_{t_1(a)} q_{t_2(a)} p_{t_1(b)} q_{t_2(b)} \right]$$

$$= \mathbb{E} \left[ \frac{1}{4} \sigma(p^\top Wq/2) (1 - \sigma(p^\top Wq/2)) \left( \sum_{a \in [d^2]} z_a p_{t_1(a)} q_{t_2(a)} \right)^2 \right]$$

1179  
1180 
$$\leq \mathbb{E} \left| \frac{1}{2} \left( \sum_{x \in \mathcal{D}_{k}} z_{x} p_{k}(x) q_{k}(x) \right) \right|$$

$$\begin{bmatrix} \overline{4} \\ a \in [d^2] \end{bmatrix} \begin{bmatrix} \overline{4} \\ a \in [d^2] \end{bmatrix} \begin{bmatrix} \overline{4} \\ a \in [d^2] \end{bmatrix} \begin{bmatrix} 2aPt_1(a)Qt_2(a) \\ a \in [d^2] \end{bmatrix}$$

 $\overset{(a)}{\leq} \frac{1}{4} \|z\|_{2}^{2} \sum_{i \in [d^{2}]} \mathbb{E}[(p_{t_{1}(i)}q_{t_{2}(i)})^{2}]$ 

where (a) is due to the Cauchy–Schwarz inequality. Thus,  $\nabla^2 L(W) \leq lI_d$  and L(W) is *l*-smooth, where *l* is a constant smaller than  $\frac{1}{4} \sum_{i \in [d^2]} \mathbb{E}[(p_{t_1(i)}q_{t_2(i)})^2]$ .

#### E.4 PROOF OF THEOREM 3.1

**Proof** According to Lemma E.4, the global minimizer of L(W) is  $W^* = 2(\Lambda^{-1} + G)$ , where  $\|G\|_{\max} \leq \frac{1}{N} \|S^{-1}(\mathbb{E}[\sigma'(a)(4qq^{\top} + \frac{1}{4}uu^{\top}\Lambda^{-1}qq^{\top})$  $+ \sigma''(a) (\frac{1}{8} (u^{\top} \Lambda^{-1} q)^2 \mu q^{\top} + 2q^{\top} \Lambda^{-1} q \mu q^{\top})]) \|_{\max} + o(1/N).$ (38)Define  $R_W = \{W \in \mathbb{R}^{d \times d} \mid ||W - W^*||_F \le ||W^0 - W^*||_F\}$ .  $R_W$  is a compact set. Then, according to Lemma E.2, for  $W \in R_W$ , we have  $\nabla^2 \overline{L}(W) \succeq \alpha I_d$ . Here  $\alpha > 0$  is a positive constant number. Thus, L(W) is  $\alpha$ -strongly convex in  $R_W$ . Moreover, according to Lemma E.5, L(W) is *l*-smooth. Then according to Lemma D.2, applying gradient descent with  $\eta = 1/l$ , for any  $t \ge 1$ , we have  $||W^t - W^*||_F^2 < \exp(-t/\kappa) \cdot ||W^0 - W^*||_F^2$ where  $\kappa = l/\alpha$ . F **IN-CONTEXT INFERENCE OF BINARY CLASSIFICATION** F.1 NOTATIONS In this section, we use the following notations. We denote  $\mu = \mu_1 - \mu_0$ ,  $u = 2(\mu_1 + \mu_0)$ ,  $q = x_{query}$ . Define  $p = \frac{2}{M} \sum_{i=1}^{M} y_i x_i$ . Since with probability  $\mathbb{P}(y_i = 1) = 1/2, x_i = \mu_1 + v_i$ , with probability  $\mathbb{P}(y_i = 0) = \frac{1}{2} \sqrt{2}, x_i = \mu_0 + v_i$ , where  $v_i \sim N(0, \Lambda)$ , we have  $p = 2M_1 \mu_1 / M - 2M_0 \mu_0 / M + g$ , where  $g = \frac{2}{M} \sum_{i=1}^{M} v_i$ ,  $g \sim N(0, 4\Lambda/M)$ ,  $M_1 \sim Bin(M, 1/2)$ . Defining  $h = M_1/N - 1/2$ ,  $u = 2(\mu_1 + \mu_0)$ , we have  $M_0/N = 1/2 - h$  and  $p = \mu + hu + g.$ (39)F.2 PROOF OF THEOREM 3.2 **Proof** The output of the trained transformer is  $\widehat{y}_{\mathsf{out}} = \sigma\left(\left(\frac{2}{M}\sum_{i=1}^{M} y_i x_i^{\top}\right)(\Lambda^{-1} + \widehat{G}) x_{\mathsf{query}}\right) = \sigma(p^{\top}(\Lambda^{-1} + \widehat{G})q).$ (40)The probability of  $y_{query} = 1$  given  $x_{query}$  is 

$$\mathbb{P}\left(y_{\mathsf{query}} = 1 | x_{\mathsf{query}}\right) = \sigma((\mu_1 - \mu_0)^\top \Lambda^{-1} x_{\mathsf{query}}) = \sigma(\mu^\top \Lambda^{-1} q)$$

Defining  $a = \mu^{\top} \Lambda^{-1} q$ ,  $b = (\mu + hu + q)^{\top} \widehat{G} q + (hu + q)^{\top} \Lambda^{-1} q$ , we have 

1229  
1230  
1230  
1231  
1232  

$$p^{\top}(\Lambda^{-1} + \widehat{G})q$$
  
 $=(\mu + hu + g)^{\top}(\Lambda^{-1} + \widehat{G})q$   
 $=(\mu + hu + g)^{\top}\widehat{G}q + (hu + g)^{\top}\Lambda^{-1}q + \mu^{\top}\Lambda^{-1}q = a + b,$ 

and 

$$\mathbb{E}\left[\sigma(p^{\top}(\Lambda^{-1}+\widehat{G})q)\right] = \mathbb{E}\left[\sigma(a+b)\right] = \mathbb{E}[\sigma(a) + \sigma'(a)b + \sigma''(\xi(a,b))b^2/2],$$

where  $\xi$  are real numbers between a and a + b. Thus, we have

 $\mathbb{E}[|\sigma(a+b) - \sigma(a)|]$ 

$$\leq \mathbb{E}[\left|\sigma'(a)b + \sigma''(\xi(a,b))b^2/2\right|]$$

1241 
$$\leq \sigma'(a)\mathbb{E}[|b|] + \mathbb{E}[b^2]$$

$$\begin{array}{ll} & \text{We first consider the term } \sigma'(a) \mathbb{E}[|b|]. \text{ Defining } \bar{g} = \Lambda^{-1/2} M^{1/2} g/2, \text{ we have} \\ & \sigma'(a) \mathbb{E}[|b|] \\ & \leq \sigma'(a) \left[ |\mu^{\top} \widehat{G}q| + \mathbb{E}[|hu^{\top} \widehat{G}q]] + \mathbb{E}[|g^{\top} \widehat{G}q]] + \mathbb{E}[|hu^{\top} \Lambda^{-1}q]] + \mathbb{E}[|g^{\top} \Lambda^{-1}q|] \right] \\ & \leq \sigma'(a) \left[ |\mu^{\top} \widehat{G}q| + \frac{1}{2M^{1/2}} |u^{\top} \widehat{G}q| + \frac{2}{M^{1/2}} \mathbb{E}[]\overline{g}^{\top} \Lambda^{1/2} \widehat{G}q]] + \frac{1}{2M^{1/2}} |u^{\top} \Lambda^{-1}q| + \frac{2}{M^{1/2}} \mathbb{E}[]\overline{g}^{\top} \Lambda^{-1/2}q] \right] \\ & \leq \sigma'(a) \left[ |\|\widehat{G}\|_{\max} \sum_{i,j \in [d]} ||\mu_i q_j| + \frac{1}{M^{1/2}} \left( \frac{1}{2} |u^{\top} \Lambda^{-1}q| + \frac{2\sqrt{2}}{\sqrt{\pi}} \sum_{i,j \in [d]} ||\Lambda_{ij}^{-1/2}q_j| \right) \right] + o\left( \frac{1}{N} + \frac{1}{\sqrt{M}} \right), \\ & \text{where } (a) \text{ is due to } \mathbb{E}[|h|] \leq 1/(2M^{1/2}) \text{ in Lemma E.1. } (b) \text{ is because that } \overline{g}_i \sim N(0, 1) \text{ and} \\ & \mathbb{E}[|\overline{g}_i|] = \sqrt{2}/\sqrt{\pi}, \text{ for } i \in [d]. \\ & \text{For } \mathbb{E}[b^2], \text{ we have} \\ & \mathbb{E}[b^2] \leq \mathbb{E}\left[ [(\mu + hu + g)^{\top} \widehat{G}q]^2 \right] + \mathbb{E}\left[ [(hu + g)^{\top} \Lambda^{-1}q]^2 \right] + 2\mathbb{E}\left[ (\mu + hu + g)^{\top} \widehat{G}q(hu + g)^{\top} \Lambda^{-1}q \right]. \\ & \text{Notice that terms in } \mathbb{E}\left[ [(\mu + hu + g)^{\top} \widehat{G}q]^2 \right] \text{ contain two } \widehat{G}. \text{ Thus, they are at most smaller than} \\ & O(||\widehat{G}||_{\max}^2) = O(1/N^2). \text{ Terms in } \mathbb{E}\left[ [(hu + g)^{\top} \Lambda^{-1}q]^2 \right]/2 \text{ contain two } h, \text{ or two } g, \text{ or one } h \\ & \text{ and one } g. \text{ According to Lemma D.1, we have } \mathbb{E}[|h|] = O(1/\sqrt{M}), \mathbb{E}[h^2] = 1/(4M). \text{ Moreover,} \\ & g = 2M^{-1/2}\Lambda^{1/2}\overline{g}. \text{ Converting one } g \text{ to } \overline{g}, \text{ we have a coefficient of } M^{-1/2}. \text{ Thus, terms in} \\ & \mathbb{E}\left[ (hu + g)^{\top} \Lambda^{-1}q]^2 \right]/2 \text{ contain two } h, \text{ or two } g, \text{ or one } h \text{ and one } g \text{ are O}(1/M). \text{ Terms in} \\ & \mathbb{E}\left[ (\mu + hu + g)^{\top} \widehat{G}q(hu + g)^{\top} \Lambda^{-1}q \right] \text{ contain at least one } \widehat{G} \text{ and one } h \text{ or one } \widehat{G} \text{ and one } g. \\ & \text{Thus, they are at most smaller than } O(||\widehat{G}||_{\max}/\sqrt{M}) = O(1/(N\sqrt{M})). \text{ Therefore, we have} \\ & \mathbb{E}[\Delta[2]/2 = O(1/N^2 + 1/M + 1/(N\sqrt{M})) = o(1/N + 1/\sqrt{M}). \\ & \text{Finally, we have} \\ & \mathbb{E}[\Delta(y_{query}, \widehat{y}_{query})] = \mathbb{E}[|\widehat{y}_{out} - \mathbb{P}(y_{query} = 1|x_{query})|] = \mathbb{E}$$

1276 1277

**Remark F.1** We note that Theorem 3.2 requires Assumption 3.2 to hold. For example, we need the covariance  $\Lambda$  in training and testing to be the same. A similar consistency requirement of the covariance  $\Lambda$  in training and testing had also been observed for in-context linear regression in Zhang et al. (2023a).

*Here, we discuss the consequences when Assumption 3.2 does not hold. For example, suppose the labels of our data in test prompts are not balanced where*  $\mathbb{P}(y=1) = p_1, \mathbb{P}(y=-1) = p_0$ . *Besides,*  $\mu_0, \mu_1$  *do not have the same*  $\Lambda^{-1}$  *weighted norm, and the covariance matrix of test data is*  $\Gamma \neq \Lambda$ . Then, as  $N, M \to \infty$ , we have

1286 1287

1288

1290

1292 1293

$$\frac{2}{M} \sum_{i=1}^{M} y_i x_i^{\top} \to 2(p_1 \mu_1 - p_0 \mu_0)^{\top},$$

1289 and

$$\mathbb{P}(\widehat{y}_{\mathsf{query}} = 1) \to \sigma(2(p_1\mu_1 - p_0\mu_0)^{\top}\Lambda^{-1}x_{\mathsf{query}}).$$

1291 On the other hand, the distribution of the ground truth label is

$$\mathbb{P}(y_{query} = 1) = \sigma((\mu_1 - \mu_0)^\top \Gamma^{-1} x_{query} + (\mu_1^\top \Lambda^{-1} \mu_1 - \mu_0^\top \Lambda^{-1} \mu_0)/2 + \log(p_1/p_0)).$$

1294 Define  $z \triangleq (\mu_1 - \mu_0)^\top \Gamma^{-1} x_{query} + (\mu_1^\top \Lambda^{-1} \mu_1 - \mu_0^\top \Lambda^{-1} \mu_0)/2 + \log(p_1/p_0)$  and  $\hat{z} \triangleq 2(p_1\mu_1 - p_0\mu_0)^\top \Lambda^{-1} x_{query}$ . Then, we can notice that unless  $\hat{z} = z$  or  $|\sigma(\hat{z}) - \sigma(z)|$  is sufficiently small, the transformer cannot correctly perform the in-context binary classification.

## <sup>1296</sup> G TRAINING PROCEDURE FOR IN-CONTEXT MULTI-CLASS CLASSIFICATION

<sup>1298</sup> In this section, we present the proof of Theorem 4.1.

### <sup>1300</sup> G.1 PROOF SKETCH 1301

1302 First, we prove in Lemma G.3 that the expected loss function L(W) (16) is strictly convex w.r.t. W and is strongly convex in a compact set of  $\mathbb{R}^{d \times d}$ . Moreover, we prove L(W) has one unique 1303 1304 global minimizer  $W^*$ . Then, in Lemma G.4, by analyzing the Taylor expansion of L(W), we prove that as  $N \to \infty$ , our loss function L(W) point wisely converges to  $\tilde{L}(W)$  (defined in (44)), and 1305 the global minimizer  $W^*$  converge to  $2\Lambda^{-1}$ . Thus, we denote  $W^* = 2(\Lambda^{-1} + G)$ , and prove 1306  $||G||_{\max} = O(N^{-1/4})$ . Next, in Lemma G.5, by further analyzing the Taylor expansion of the 1307 equation  $\nabla L(W^*) = 0$  at the point  $2\Lambda^{-1}$ , we establish a tighter bound  $||G||_{\max} = O(cN^{-1})$ . In 1308 Lemma G.6, we prove that our loss function is l-smooth and provide an upper bound for l. Thus, 1309 in a compact set  $R_W$ , our loss function is  $\alpha$ -strongly convex and *l*-smooth. Finally, leveraging the 1310 standard results from the convex optimization, we prove Theorem 4.1 in subsection G.3. 1311

1312 In this section, we use the following notations.

### 1314 G.2 NOTATIONS

1316 Recall the expected loss function (16) is

1317

1318 1319

1321 1322 1323

1325

1313

1315

$$L(W) = -\mathbb{E}\left[\sum_{k=1}^{c} (y_{\tau,\mathsf{query}})_k \log((\widehat{y}_{\tau,\mathsf{out}})_k)\right],\tag{41}$$

where

$$(\widehat{y}_{\tau,\mathsf{out}})_k = \operatorname{softmax}\left(\frac{1}{c} \left(\frac{c}{N} \sum_{i=1}^N y_{\tau,i} x_{\tau,i}^{\mathsf{T}}\right) W x_{\tau,\mathsf{query}}\right)_k$$

is the output of the transformer, and the label of the data follows the distribution

 $\mathbb{P}\left(y_{\tau,\mathsf{query}} = \mathbf{e}_k | x_{\tau,\mathsf{query}}\right) = \operatorname{softmax}(\mu_{\tau}^{\top} \Lambda^{-1} x_{\tau,\mathsf{query}}))_k.$ 

1326 In this section, we introduce the following notations to analyze (16). We denote  $\mu_k = \mu_{\tau,k}, \mu =$ 1327  $(\mu_1, \mu_2, \dots, \mu_k) \in \mathbb{R}^{d \times c}$  and  $q = x_{\tau, query}$ . Then with probability  $\mathbb{P}(y_{\tau, query} = \mathbf{e}_k) = 1/c, q = 1/c$ 1328  $\mu_k + v$ , where  $v \sim \mathsf{N}(0,\Lambda)$ . We define  $p_k = \frac{c}{N} \sum_{i=1}^N (y_{\tau,i})_k x_{\tau,i} \in \mathbb{R}^d$  and  $P = (p_1, p_2, \dots, p_c) \in \mathbb{R}^d$ 1329  $\mathbb{R}^{d \times c}$ . We have  $P^{\top} = \frac{c}{N} \sum_{i=1}^{N} y_i x_{\tau,i}^{\top} \in \mathbb{R}^{c \times d}$ . Since with probability  $\mathbb{P}(y_{\tau,i} = \mathbf{e}_k) = 1/c$  we 1330 have  $x_{\tau,i} = \mu_k + v_i$ , where  $v_i \sim \mathsf{N}(0,\Lambda)$ , we known  $p_k = \frac{c}{N} \sum_{i=1}^{N} (y_{\tau,i})_k x_{\tau,i} = cN_k \mu_k / N + g_k$ , where  $g_k = \frac{c}{N} \sum_{i \in \{i | y_{\tau,i} = \mathbf{e}_k\}} v_i, g_k \sim \mathsf{N}(0, c^2 N_k \Lambda / N^2)$  and  $(N_1, N_2, \dots, N_c) \sim \mathsf{Multin}(n, 1/c)$ . Defining  $h_k = N_k / N - 1/c$ , we have  $N_k / N = 1/c + h_k$  and  $p_k = \mu_k + ch_k \mu_k + g_k$ . Defining  $\bar{g}_k = \Lambda^{-1/2} g_k$ , we have  $\bar{g}_k \sim \mathsf{N}(0, c^2 N_k I_d / N^2)$ . Defining  $\mu_h = (h_1 \mu_1, h_2 \mu_2, \dots, h_k \mu_k) \in \mathbb{R}^{d \times c}$ and  $g = (g_1, g_2, \dots, g_k) \in \mathbb{R}^{d \times c}$ , we have  $P = \mu + c\mu_h + g$ . 1331 1332 1333 1334 1335 1336

Then, the expected loss function (16) can be expressed as

$$L(W) = \mathbb{E}\left[\sum_{k=1}^{c} -\operatorname{softmax}(\mu^{\top}\Lambda^{-1}q)_{k} \log(\operatorname{softmax}(P^{\top}Wq/c)_{k})\right].$$
(42)

The gradient of the loss function (16) can be expressed as

$$\nabla L(W) = \mathbb{E}\left[\sum_{k=1}^{c} \left[ (\operatorname{softmax}(P^{\top}Wq/c)_{k} - \operatorname{softmax}(\mu^{\top}\Lambda^{-1}q)_{k})p_{k}q^{\top}/c \right] \right].$$
(43)

1345 Moreover, we define a function  $\widetilde{L}(W)$  as

$$\widetilde{L}(W) = \mathbb{E}\left[\sum_{k=1}^{c} -\operatorname{softmax}(\mu^{\top}\Lambda^{-1}q)_{k} \log(\operatorname{softmax}(\mu^{\top}Wq/c)_{k})\right].$$
(44)

In Lemma G.4, we show that as  $N \to \infty$ , L(W) will point wisely converge to  $\widetilde{L}(W)$ .

1342 1343 1344

1347 1348 1349

1338 1339 1340 **Lemma G.1** Suppose  $(N_1, N_2, \ldots, N_c) \sim \text{Multin}(N, 1/c)$ . Defining  $h_k = N_k/N - 1/c$ , we have 

- $\mathbb{E}[h_k] = 0$  $\mathbb{E}[h_k^2] = \frac{1}{N} \left( \frac{1}{c} - \frac{1}{c^2} \right)$
- $\mathbb{E}[h_i h_j] = -\frac{1}{Nc^2}, i \neq j$
- $\mathbb{E}\left[\prod_{k=1}^{c} h_{k}^{n_{k}}\right] = O\left(N^{-2}\right), \sum_{k} n_{k} \ge 3$  $\mathbb{E}\left[|h_{j}|\right] \le N^{-1/2} c^{-1/2} (1 1/c)^{1/2}$
- $\mathbb{E}[|h_i h_j|] = O(N^{-1})$
- $\mathbb{E}[|h_i h_j h_k|] = O\left(N^{-3/2}\right)$
- $\mathbb{E}[|h_i h_j h_k h_l|] = O\left(N^{-2}\right),$

where  $i, j, k, l \in [c]$ . 

Proof Since  $(N_1, N_2, \ldots, N_c) \sim \text{Multin}(N, 1/c)$ , the moment-generating function of  $(N_1, N_2, \ldots, N_c)$  is 

$$M_N(t) = \left(\frac{1}{c}\sum_{i=1}^{c}\exp(t_i)\right)^N$$

We can compute the moment-generating function of  $h = (h_1, h_2, \dots, h_c)$  as follows:

Observing the coefficients of h, we have

$$\begin{split} & \mathbb{E}[h_k] = 0 \\ & \mathbb{E}[h_k] = 0 \\ & \mathbb{E}[h_k^2] = \frac{1}{N} \left(\frac{1}{c} - \frac{1}{c^2}\right) \\ & \mathbb{E}[h_i h_j] = -\frac{1}{Nc^2}, i \neq j \\ & \mathbb{E}[h_i h_j] = -\frac{1}{Nc^2}, i \neq j \\ & \mathbb{E}[\prod_{k=1}^c h_k^{n_k}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(N^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(n_k^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(n_k^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(n_k^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(n_k^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(n_k^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(n_k^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(n_k^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}] = O\left(n_k^{-2}\right), \sum_k n_k \geq 3, \\ & \mathbb{E}[n_k^{-1}]$$

where  $i, j, k \in [c]$ .

1404 Iteratively applying the Hölder's inequality, we have 1405  $\mathbb{E}[|h_j|] \le \left(\mathbb{E}[h_j^2]\right)^{1/2} = N^{-1/2}c^{-1/2}(1-1/c)^{1/2}$ 1406 1407  $\mathbb{E}[|h_i h_j|] \le \left(\mathbb{E}[h_i^2 h_j^2]\right)^{1/2} = O(N^{-1})$ 1408  $\mathbb{E}[|h_i|^3] < \mathbb{E}[|h_i|^4]^{3/4} = (N^{-3/2})$ 1409 1410  $\mathbb{E}[|h_i h_j h_k|] \le \mathbb{E}[|h_i|^3]^{1/3} \mathbb{E}[|h_j|^3]^{1/3} \mathbb{E}[|h_k|^3]^{1/3} = O\left(N^{-3/2}\right)$ 1411 1412  $\mathbb{E}[|h_i h_j h_k h_l|] \le \mathbb{E}[|h_i|^4]^{1/4} \mathbb{E}[|h_j|^4]^{1/4} \mathbb{E}[|h_k|^4]^{1/4} \mathbb{E}[|h_l|^4]^{1/4} = O(N^{-2})$ 1413 where  $i, j, k, l \in [c]$ . 1414 1415 1416 **Lemma G.2** Suppose  $g_k \sim N(0, c^2 N_k \Lambda/N^2)$  and  $(N_1, N_2, \ldots, N_c) \sim \text{Multin}(N, 1/c)$ , define 1417  $\bar{g}_k = \Lambda^{-1/2} g_k$  and  $N_k/N = 1/c + h_k$ , we have 1418  $\mathbb{E}[(\bar{g}_k)_i] = 0$ 1419  $\mathbb{E}[(\bar{g}_k)_i(\bar{g}_l)_j] = \delta_{kl}\delta_{ij}c/N$ 1420 1421  $\mathbb{E}[(\bar{g}_{k_1})_{i_1}(\bar{g}_{k_2})_{i_2}(\bar{g}_{k_3})_{i_3}] = 0$ 1422  $\mathbb{E}[(\bar{g}_k)_i^4] = \mathbb{E}[3c^2/N^2(1+ch_k)^2] = O(N^{-2})$ 1423  $\mathbb{E}[h_m(\bar{q}_k)_i(\bar{q}_l)_i] = \mathbb{E}[c^2 \delta_{kl} \delta_{ij} h_m h_k / N] = O(N^{-2})$ 1424  $\mathbb{E}[h_m h_l(\bar{g}_k)_i] = 0$ 1425 1426 where  $i, j, i_1, i_2, i_3 \in [d], k, l, m, k_1, k_2, k_3 \in [c]$ . 1427 For any  $n_{1k}$ ,  $n_{2ki}$  satisfying  $\sum_{k \in [c]} n_{1k} + \sum_{k \in [c], i \in [d]} n_{2ki} = 1, 2, 3$ , we have 1428 1429  $\mathbb{E}[ \prod \quad h_k^{n_{1k}}(\bar{g}_k)_i^{n_{2ki}}] = O(N^{-1})$ 1430  $k \in [c], i \in [d]$ 1431 1432 Moreover, we have 1433  $\mathbb{E}[|(\bar{a}_k)_i|] < \mathbb{E}[(\bar{a}_k)_i^2]^{1/2} = N^{-1/2}c^{1/2}$ 1434 1435  $\mathbb{E}[|(\bar{q}_k)_i|^3] < \mathbb{E}[(\bar{q}_k)_i^4]^{3/4} = O(N^{-3/2})$ 1436 where  $i \in [d], k \in [c]$ . 1437 1438 For any  $n_{1k}$ ,  $n_{2ki}$  satisfying  $\sum_{k \in [c]} n_{1k} + \sum_{k \in [c], i \in [d]} n_{2ki} = n$ , n = 1, 2, 3, 4, we have 1439 1440  $\mathbb{E}\left[\prod_{k \in [c], i \in [d]} |h_k^{n_{1k}}(\bar{g}_k)_i^{n_{2ki}}|\right] = O(N^{-n/2})$ 1441 1442 1443 **Proof** Since  $g_k \sim N(0, c^2 N_k \Lambda/N^2)$  and  $\bar{g}_k \sim N(0, c^2 N_k I_d/N^2) = N(0, (c/N + c^2 h_k/N)I_d)$ , we 1444 have 1445 1446  $\mathbb{E}[(\bar{g}_k)_i] = 0$ 1447  $\mathbb{E}[(\bar{g}_k)_i(\bar{g}_l)_j] = \delta_{kl}\delta_{ij}c/N$ 1448  $\mathbb{E}[(\bar{g}_{k_1})_{i_i}(\bar{g}_{k_2})_{i_2}(\bar{g}_{k_3})_{i_3}] = 0$ 1449  $\mathbb{E}[(\bar{g}_k)_i^4] = \mathbb{E}[3c^2/N^2(1+ch_k)^2] = O(N^{-2})$ 1450 1451  $\mathbb{E}[h_m(\bar{g}_k)_i(\bar{g}_l)_j] = \mathbb{E}[c^2 \delta_{kl} \delta_{ij} h_m h_k / N] = O(N^{-2})$ 1452  $\mathbb{E}[h_m h_l(\bar{q}_k)_i] = 0$ 1453 where  $i, j, i_1, i_2, i_3 \in [d], k, l, m, k_1, k_2, k_3 \in [c]$ . Thus, with the results from Lemma G.1, for any 1454  $n_{1k}, n_{2ki}$  satisfying  $\sum_{k \in [c]} n_{1k} + \sum_{k \in [c], i \in [d]} n_{2ki} = 1, 2, 3$ , we have 1455 1456  $\mathbb{E}[\prod h_k^{n_{1k}}(\bar{g}_k)_i^{n_{2ki}}] = O(N^{-1})$ 1457  $k{\in}[c], i{\in}[d]$ 

Moreover, according to the Jensen's inequality, we have 

$$\begin{aligned} & \mathbb{E}[|(\bar{g}_k)_i|] \le \mathbb{E}[(\bar{g}_k)_i^2]^{1/2} = N^{-1/2} c^{1/2} \\ & \mathbb{E}[|(\bar{g}_k)_i|^3] \le \mathbb{E}[(\bar{g}_k)_i^4]^{3/4} = O(N^{-3/2}) \end{aligned}$$

where  $i \in [d], k \in [c]$ . Thus, with the results from Lemma G.1, for any  $n_{1k}$ ,  $n_{2ki}$  satisfying  $\sum_{k \in [c]} n_{1k} + \sum_{k \in [c], i \in [d]} n_{2ki} = n, n = 1, 2, 3, 4$ , we have 

$$\mathbb{E}[\prod_{k \in [c], i \in [d]} |h_k^{n_{1k}}(\bar{g}_k)_i^{n_{2ki}}|] \le \prod_{k \in [c], i \in [d]} \mathbb{E}[|h_k^n|]^{n_{1k}/n} \mathbb{E}[|(\bar{g}_k)_i^n|]^{n_{2ki}/n} = O(N^{-n/2}).$$

**Lemma G.3** For the loss function L(W) (16), we have  $\nabla^2 L(W) \succ 0$ . For any compact set  $R_W$ , when  $W \in R_W$ , we have  $\nabla^2 L(W) \succ \gamma I_d$  for some  $\gamma > 0$ . Additionally, L(W) has one unique global minimizer on  $\mathbb{R}^{d \times d}$ . 

For  $\widetilde{L}(W)$  defined in (44), we also have  $\nabla^2 \widetilde{L}(W) \succ 0$ . For any compact set  $R_W$ , when  $W \in R_W$ , we have  $\nabla^2 \widehat{L}(W) \succ \gamma I_d$  for some  $\gamma > 0$ . Additionally,  $\widetilde{L}(W)$  has one unique global minimizer on  $\mathbb{R}^{d \times d}$ 

**Proof** We vectorize W as  $\operatorname{Vec}(W) \in \mathbb{R}^{d^2}$ , where  $\operatorname{Vec}(W)_i = W_{t_1(i), t_2(i)}, t_1(x) = |(x-1)/d| +$  $1, t_2(x) = ((x - 1) \mod d) + 1$ . Then, we have 

$$(\nabla L(W))_i = \mathbb{E}\left[\sum_{k=1}^c \left[ (\operatorname{softmax}(P^\top W q/c)_k - \operatorname{softmax}(\mu^\top \Lambda^{-1} q)_k)(p_k)_{t_1(i)} q_{t_2(i)}/c \right] \right]$$
(45)

Note that

$$\nabla \operatorname{softmax}(P^{\top}Wq/c)_{k} = \sigma(a_{k})(1 - \sigma(a_{k}))\nabla a_{k},$$

 $\nabla a_k = \frac{\sum_{l=1,\dots,c,l \neq k} \exp\left((p_l - p_k)^\top W q/c\right) (p_k - p_l) q^\top/c}{\sum_{l=1,\dots,c,l \neq k} \exp\left((p_l - p_k)^\top W q/c\right)}$ 

$$=\frac{\sum_{l=1,\dots,c,l\neq k}\exp\left(p_l^{\top}Wq/c\right)(p_k-p_l)q^{\top}/c}{\sum_{l=1,\dots,c,l\neq k}\exp\left(p_l^{\top}Wq/c\right)}.$$

softmax $(P^{\top}Wa/c)_{i} = \sigma(a_{i})$ 

where  $a_k = -\log(\sum_{l=1,\dots,c,l\neq k} \exp((p_l - p_k)Wq/c)))$ . For  $\nabla a_k$ , we have

Then we have

1496  
1497 
$$\nabla \operatorname{softmax}(P^{\top}Wq/c)_{k} = \operatorname{softmax}(P^{\top}Wq/c)_{k} \frac{\sum_{l=1,\dots,c,l\neq k} \exp\left(p_{l}^{\top}Wq\right)(p_{k}-p_{l})q^{\top}/c}{\sum_{n=1,\dots,c} \exp\left(p_{n}^{\top}Wq/c\right)}$$
1498  
1499 
$$= \sum_{l=1,\dots,c,l\neq k} \operatorname{softmax}(P^{\top}Wq/c)_{k} \operatorname{softmax}(P^{\top}Wq/c)_{l}(p_{k}-p_{l})q^{\top}/c$$
1501

and 

$$(\nabla \operatorname{softmax}(P^{\top}Wq/c)_{k})_{j} = \sum_{l=1,\dots,c,l \neq k} \operatorname{softmax}(P^{\top}Wq/c)_{k} \operatorname{softmax}(P^{\top}Wq/c)_{l}(p_{k} - p_{l})_{t_{1}(j)}q_{t_{2}(j)}/c.$$
1504
1505
We can express the Hessian matrix of the loss function with the following form:

We can express the Hessian matrix of the loss function with the following form:

1510  
1511 
$$= \mathbb{E}\left[\sum_{k=2}^{c}\sum_{l=1}^{k-1}\operatorname{softmax}(P^{\top}Wq/c)_{k}\operatorname{softmax}(P^{\top}Wq/c)_{l}(p_{k}-p_{l})_{t_{1}(i)}q_{t_{2}(i)}(p_{k}-p_{l})_{t_{1}(j)}q_{t_{2}(j)}/c^{2}\right]$$

Considering  $z \in \mathbb{R}^{d^2}$  such that  $z \neq 0$ , we have

 $= \mathbb{E} \left[ \frac{1}{c^2} \sum_{k=2}^{c} \sum_{l=1}^{k-1} \operatorname{softmax}(P^\top Wq/c)_k \operatorname{softmax}(P^\top Wq/c)_l \sum_{ab} z_a z_b (p_k - p_l)_{t_1(a)} q_{t_2(a)} (p_k - p_l)_{t_1(b)} q_{t_2(b)} \right]$   $= \mathbb{E} \left[ \frac{1}{c^2} \sum_{k=2}^{c} \sum_{l=1}^{k-1} \operatorname{softmax}(P^\top Wq/c)_k \operatorname{softmax}(P^\top Wq/c)_l \left( \sum_{a \in [d^2]} z_a (p_k - p_l)_{t_1(a)} q_{t_2(a)} \right)^2 \right]$   $= \mathbb{E} \left[ \frac{1}{c^2} \sum_{k=2}^{c} \sum_{l=1}^{k-1} \operatorname{softmax}(P^\top Wq/c)_k \operatorname{softmax}(P^\top Wq/c)_l \left( \sum_{a \in [d^2]} z_a (p_k - p_l)_{t_1(a)} q_{t_2(a)} \right)^2 \right]$ 

1522 Since for any P, q, k, l, softmax $(P^{\top}Wq/c)_k$  softmax $(P^{\top}Wq/c)_l > 0$ , we have  $z^{\top}\nabla^2 L(W)z \ge 0$ . 1523 Thus,  $\nabla^2 L(W) \succeq 0$  and L(W) is convex.

1525 Defining  $\tilde{p} = p_1 - p_2$ , we have

$$z^{\top} \nabla^{2} L(W) z$$
  

$$\geq \mathbb{E} \left[ \frac{1}{c^{2}} \operatorname{softmax}(P^{\top} Wq/c)_{1} \operatorname{softmax}(P^{\top} Wq/c)_{2} \left( \sum_{a \in [d^{2}]} z_{a}(p_{1} - p_{2})_{t_{1}(a)}q_{t_{2}(a)} \right)^{2} \right]$$
  

$$= \int \frac{1}{c^{2}} \operatorname{softmax}(P^{\top} Wq/c)_{1} \operatorname{softmax}(P^{\top} Wq/c)_{2} \left( \sum_{a \in [d^{2}]} z_{a} \tilde{p}_{t_{1}(a)}q_{t_{2}(a)} \right)^{2} f_{Pq}(P,q) dP dq$$

where  $f_{Pq}(P,q)$  are the PDF function of P,q. For any  $z \neq 0$ , we denote  $z_{ij} = z_{((i-1)d+j)}$ , suppose  $a, b \in \arg \max_{i,j} |z_{ij}|$ , we consider a set of constants  $\{c_{1pi}, c_{2pi}\}, \{c_{1qi}, c_{2qi}\}, i, j \in [d]$ , where  $c_{1pa} = d, c_{2pa} = d + 1$ ,  $c_{1qb} = d, c_{2qb} = d + 1$ , and  $c_{1pi} = 1/16, c_{2pi} = 1/8, i \neq a$ ,  $c_{1qj} = 1/16, c_{2qj} = 1/8, j \neq b$ . Then, for any  $c_{pi} \in [c_{1pi}, c_{2pi}], c_{qj} \in [c_{1qj}, c_{2qj}]$ , we have

$$\left| \sum_{i,j \in [d]} z_{ij} c_{pi} c_{qj} \right| \ge \left[ d^2 - 2(d+1)(d-1)/8 - (d-1)^2/64 \right] \max_{ij} |z_{ij}| \ge d^2 \max_{ij} |z_{ij}|/2.$$

Then, we define region  $\Omega(a,b) = \{\tilde{p} = \sum_i c_{pi} \mathbf{e}_i, q = \sum_j c_{qj} \mathbf{e}_j, c_{pi} \in [c_{1pi}, c_{2pi}], c_{qj} \in [c_{1pj}, c_{2qj}], \|P\|_F^2 \leq c^2 (\sum_i c_{2pi}^2 + c_{2qj}^2) \}.$  We have

$$\min_{\Omega(a,b)} \left( \sum_{l \in [d^2]} z_l \tilde{p}_{t_1(l)} q_{t_2(l)} \right)^2 \ge d^4 \max_{ij} |z_{ij}|^2 / 4 \ge ||z||_2^2 / 4$$

Defining

$$C(\Omega) = \min_{a \in [d], b \in [d]} \int_{\Omega(a,b)} f_{Pq}(P,q) dP dq,$$
  
$$S(\Omega, W) = \min_{a \in [d], b \in [d]} \min_{\Omega(a,b)} \left\{ \frac{1}{c^2} \operatorname{softmax}(P^\top Wq/c)_1 \operatorname{softmax}(P^\top Wq/c)_2 \right\}$$

we have  $S(\Omega, W) > 0$ . Since we have  $f_{Pq}(P,q) > 0$  for all P,q and  $\Omega(a,b)$  are non-zero measures for P,q. Thus, we have  $C(\Omega) > 0$ . Then, for any  $z \neq 0$ , we have

$$z^{\top} \nabla^{2} L(W) z$$

$$\geq \int_{\Omega(a,b)} \frac{1}{c^{2}} \operatorname{softmax}(P^{\top} Wq/c)_{1} \operatorname{softmax}(P^{\top} Wq/c)_{2} \left(\sum_{l \in [d^{2}]} z_{l} \tilde{p}_{t_{1}(l)} q_{t_{2}(l)}\right)^{2} f_{Pq}(P,q) dP dq$$

$$\geq C(\Omega) S(\Omega, W) \|z\|_{2}^{2}/4 > 0$$

Thus, we have  $\nabla^2 L(W) \succ 0$ . L(W) is strictly convex.

1566 Moreover, for any compact set  $R_W$  of  $\mathbb{R}^{d \times d}$ , for any  $W \in R_W$ , we have

$$S(\Omega) = \min_{W \in R_W} \min_{a \in [d], b \in [d]} \min_{\Omega(a,b)} \left\{ \frac{1}{c^2} \operatorname{softmax}(P^\top W q/c)_1 \operatorname{softmax}(P^\top W q/c)_2 \right\} > 0.$$

 $W \in R_W$   $u \in [a], b \in [a]$  S(a,b) (C 1570 Then, for any  $W \in R_W$ , for any  $z \neq 0$ , we have

$$\geq \int_{\Omega(a,b)} \frac{1}{c^2} \operatorname{softmax}(P^\top Wq/c)_1 \operatorname{softmax}(P^\top Wq/c)_2 \left(\sum_{l \in [d^2]} z_l \tilde{p}_{t_1(l)} q_{t_2(l)}\right)^2 f_{Pq}(P,q) dP dq$$

 $\geq C(\Omega)S(\Omega)\|z\|_2^2/4.$ 

 $z^{\top} \nabla^2 L(W) z$ 

1577 Thus, when  $W \in R_W$ ,  $R_W$  is a compact set, we have  $\nabla^2 L(W) \succ C(\Omega)S(\Omega)I_d/4$ , our loss function 1578 is  $\gamma$ -strongly convex, where  $\gamma = C(\Omega)S(\Omega)/4$ .

1579 Because our loss function is strictly convex in  $\mathbb{R}^{d \times d}$ , it has at most one global minimizer in  $\mathbb{R}^{d \times d}$ . 1580 Next, we prove all level sets of our loss function are compact, i.e.  $V_{\alpha} = \{W \in \mathbb{R}^{d \times d} | L(W) \leq \alpha\}$ 1581 is compact for all  $\alpha$ . We prove it by contradiction. Suppose  $V_{\alpha}$  is not compact for some  $\alpha$ . Since our 1582 loss function is continuous and convex,  $V_{\alpha}$  is an unbounded convex set. Since the dimension of  $V_{\alpha}$  is 1583  $d^2$ , consider a point  $W^{\alpha} \in V_{\alpha}$ , there must exists a  $W^k \neq 0_{d \times d}$  that  $\{W^{\alpha} + tW^k | t = [0, \infty)\} \in V_{\alpha}$ . 1584 For this  $W^k \neq 0_{d \times d}$ , there must exist a set of constants  $0 < c_{3pi} < c_{4pi}, 0 < c_{3qj} < c_{4qj}$  such that 1585 for any  $c_{pi} \in [c_{3pi}, c_{4pi}], c_{qj} \in [c_{3qj}, c_{4qj}]$ , we have

$$|\sum_{ij} c_{pi} c_{qj} W_{ij}^k| \neq 0.$$

<sup>1588</sup> Thus, we have

$$\lim_{t \to \infty} |\sum_{ij} c_{pi} c_{qj} (W_{ij}^{\alpha} + t W_{ij}^{k})| = \infty$$

<sup>1591</sup> <sup>1592</sup> We define  $\Omega_0 = \{\tilde{p} = \sum_i c_{pi} \mathbf{e}_i, q = \sum_j c_{qj} \mathbf{e}_j, c_{pi} \in [c_{3pi}, c_{4pi}], c_{qj} \in [c_{3qj}, c_{4qj}], \|P\|_F^2 \le c^2(\sum_i c_{4pi}^2 + c_{4qj}^2), \|\mu\|_F^2 \le c^2(\sum_i c_{4pi}^2 + c_{4qj}^2)\}.$  Then, defining

$$C(\Omega_0) = \int_{\Omega_0} f_{Pq}(P,q) dP dq,$$
  

$$S(\Omega_0) = \min_{\Omega_0} \left\{ \min\{\operatorname{softmax}(\mu^\top Wq/c)_1, \operatorname{softmax}(\mu^\top Wq/c)_2\} \right\}$$

we have  $S(\Omega_0) > 0$ . Since  $\Omega_0$  are non-zero measures for P, q, we have  $C(\Omega_0) > 0$ . Then, we have  $\lim_{t \to \infty} L(W^{\alpha} + tW^k)$ 

$$= \lim_{t \to \infty} \mathbb{E}[\sum_{l=1}^{c} -\operatorname{softmax}(\mu^{\top} \Lambda^{-1}q)_{l} \log(\operatorname{softmax}(P^{\top}(W^{\alpha} + tW^{k})q/c)_{l})]]$$

$$\geq \lim_{t \to \infty} \int_{\Omega_{0}} [-\operatorname{softmax}(\mu^{\top} \Lambda^{-1}q)_{1} \log(\operatorname{softmax}(P^{\top}(W^{\alpha} + tW^{k})q/c)_{1})]f_{Pq}(P,q)dPdq$$

$$+ \lim_{t \to \infty} \int_{\Omega_{0}} [-\operatorname{softmax}(\mu^{\top} \Lambda^{-1}q)_{2} \log(\operatorname{softmax}(P^{\top}(W^{\alpha} + tW^{k})q/c)_{2})]f_{Pq}(P,q)dPdq$$

$$\geq \lim_{t \to \infty} \int_{\Omega_{0}} [-\operatorname{softmax}(\mu^{\top} \Lambda^{-1}q)_{1} \log(\sigma(\tilde{p}^{\top}(W^{\alpha} + tW^{k})q/c))]f_{Pq}(P,q)dPdq$$

$$+ \lim_{t \to \infty} \int_{\Omega_{0}} [-\operatorname{softmax}(\mu^{\top} \Lambda^{-1}q)_{2} \log(\sigma(-\tilde{p}^{\top}(W^{\alpha} + tW^{k})q/c))]f_{Pq}(P,q)dPdq$$

$$\geq C(\Omega_{0})S(\Omega_{0}) \cdot \min_{\Omega_{0}} \left\{ \lim_{t \to \infty} [-\log(\sigma(\sum_{ij} c_{pi}c_{qj}(W^{\alpha}_{ij} + tW^{k}_{ij})/c))] \right\}$$

$$+ C(\Omega_{0})S(\Omega_{0}) \cdot \min_{\Omega_{0}} \left\{ \lim_{t \to \infty} [-\log(\sigma(-\sum_{ij} c_{pi}c_{qj}(W^{\alpha}_{ij} + tW^{k}_{ij})/c))] \right\}$$

$$= \infty$$

1620 This contradicts the assumption  $L(W^{\alpha} + tW^{k}) \leq \alpha$ . Thus, all level sets of the loss function L(W)1621 are compact, which means there exists a global minimizer for L(W). Together with the fact that 1622 L(W) is strictly convex, L(W) has one unique a global minimizer on  $\mathbb{R}^{d \times d}$ .

Similarly, we can prove the same conclusions for  $\widetilde{L}(W)$ . 1624

1626 **Lemma G.4** Denoting the global minimizer of our loss function (16) as  $W^*$ , we have  $W^* =$ 1627  $c(\Lambda^{-1}+G)$ , where  $||G||_{\max} = O(N^{-1/4})$ . 1628

**Proof** Let  $a = \mu^{\top} \Lambda^{-1} q$ ,  $s = \mu^{\top} W q/c$ ,  $r = (\mu_h + g)^{\top} W q/c$ ,  $a_k = \mu_k^{\top} \Lambda^{-1} q$ ,  $s_k = \mu_k^{\top} W q/c$ , 1629  $r_k = (ch_k \mu_k + g_k)^\top Wq/c$ . Performing the Taylor expansion on (16), we have 1630 1631

$$L(W) = \mathbb{E}\left[\sum_{k=1}^{c} -\zeta_{k}(a) \log(\zeta_{k}(s+r))\right]$$
$$= \mathbb{E}\left[\sum_{k=1}^{c} -\zeta_{k}(a) \log(\zeta_{k}(s)) - \sum_{k,l=1}^{c} \zeta_{k}(a) R_{kl}(s,r) r_{l}\right]$$
$$\widetilde{\chi}(W) = \mathbb{E}\left[\sum_{k=1}^{c} -\zeta_{k}(a) \log(\zeta_{k}(s)) - \sum_{k,l=1}^{c} \zeta_{k}(a) R_{kl}(s,r) r_{l}\right]$$

$$=\widetilde{L}(W) - \mathbb{E}\left[\sum_{k,l=1}^{S} \zeta_k(a) R_{kl}(s,r) r_l\right]$$
1640

where  $|R_{kl}(s,r)| \leq \sup_{y} |\frac{\partial \log(\zeta_k(y))}{\partial y_l}| \sup_{y} |\frac{1}{\zeta_k(y)} \frac{\partial \zeta_k(y)}{\partial y_l}| = \sup_{y} |\delta_{kl} - \zeta_l(y)| \leq 1$ . Thus, we have 1642  $\left|\widetilde{L}(W) - L(W)\right|$ 

1643 
$$|\tilde{L}|^{(1)}$$

$$\leq c \sum_{i=1}^{c} \mathbb{E}\left[|r_{l}|\right]$$

1648  
1649
$$\leq \sum_{l=1}^{c} c \mathbb{E}\left[|h_{l}\mu_{l}^{\top}Wq|\right] + \mathbb{E}\left[|g_{l}^{\top}Wq|\right]$$

l=1

$$\leq O(1) \|W\|_{\max} \mathbb{E}[|h_l|] + O(1) \|W\|_{\max} \mathbb{E}[|(\bar{g}_l)_i|]$$

$$\leq C_l \|W\|_{\max} N^{-1}$$

where the last inequality is due to Lemma G.1, G.2.  $C_l$  is a constant independent of N and W. This 1654 shows that L(W) point wisely converge to L(W). 1655

According to Lemma E.2,  $\tilde{L}(W)$  has one unique global minimizer. Considering the equation: 1656

$$\nabla \widetilde{L}(W) = \mathbb{E}\left[\sum_{k=1}^{c} -\operatorname{softmax}(\mu^{\top} \Lambda^{-1} q)_{k} \log(\operatorname{softmax}(\mu^{\top} W q/c)_{k})\right] = 0$$

m [ | ( \_ ) | ]

1658 1659

1657

1652

1623

1625

1660 We can easily find that  $\nabla \widetilde{L}(c\Lambda^{-1}) = 0$  and  $W = c\Lambda^{-1}$  is the global minimizer of  $\widetilde{L}(W)$ . 1661

Considering a compact set  $R_W = \{W \mid ||W - 2\Lambda^{-1}||_F \le \rho_W\}$ , we have  $||W||_{\max} \le C_W$  for any 1662  $W \in R_W$ . Here  $\rho_W, C_W$  are some positive finite constants. Then, we have 1663

1664  $\left|\widetilde{L}(W) - L(W)\right| \le C_l' N^{-1/2}, W \in R_W$ 1665

1666 where  $C'_l = C_l C_W$  is a constant independent of N and W. This shows that, for any  $W \in R_W$ , L(W) uniformly converge to L(W). 1668

Denote  $W^*$  as the global minimizer of L(W) with prompt length N. Then, we show that, when 1669 N is sufficiently large,  $W^* \in R_W$ . We first denote  $\partial R_W = \{W \mid ||W - c\Lambda^{-1}||_F = \rho_W\}$ , 1670  $\Delta = \min_{W \in \partial R_W} \widetilde{L}(W) - \widetilde{L}(c\Lambda^{-1}) > 0$ . Then, for  $N \ge (4C_l^{\prime}/\Delta)^2$ , and for any  $W \in R_W$ , we 1671 have 1672

1673 
$$\left| \widetilde{L}(W) - L(W) \right| \le \Delta/4$$

$$\min_{W \in \partial R_W} L(W) - \min_{W \in R_W} L(W) \ge \min_{W \in \partial R_W} L(W) - L(c\Lambda^{-1}) \ge \Delta/2 > 0$$

Since L(W) is strictly convex, we have  $W^* = \arg \min_W L(W) \in R_W$ . 

Then, we have

$$\begin{split} |\widetilde{L}(W^*) - L(W^*)| &\leq C'_l/N \\ |\widetilde{L}(c\Lambda^{-1}) - L(c\Lambda^{-1})| &\leq C'_l/N \end{split}$$

$$\widetilde{L}(W^*) \le L(W^*) + C'_l/N \le L(c\Lambda^{-1}) + C'_l/N \le \widetilde{L}(c\Lambda^{-1}) + 2C'_lN^{-1/2}$$

According to Lemma E.2, for  $W \in R_W$ , we have  $\nabla^2 \widetilde{L}(W) \succ \gamma I_d$ , where  $\gamma$  is a positive constant independent of N. Thus,  $\hat{L}(W)$  is  $\gamma$ -strongly convex in  $R_W$ . According to Lemma D.1, we have 

$$\|W^* - c\Lambda^{-1}\|_F^2 \le \frac{2}{\gamma} (\widetilde{L}(W^*) - \widetilde{L}(c\Lambda^{-1})) \le \frac{4C_l'}{\gamma N^{1/2}}$$

Thus, when  $N \to \infty$ , we have  $W^* \to c\Lambda^{-1}$ . Denoting  $W^* = c(\Lambda^{-1} + G)$ , we have  $\|G\|_{\max} = c(\Lambda^{-1} + G)$ .  $O(N^{-1/4}).$ 

**Lemma G.5** The global minimizer of the loss function (16) is  $W^* = c(\Lambda^{-1} + G)$ . We have 

$$\begin{aligned} \|698 \\ \|699 \\ \|G\|_{\max} \leq \frac{1}{N} \left\| S^{-1} \mathbb{E} \left[ \sum_{k,l=1}^{c} \frac{\partial \zeta_{k}(a)}{\partial a_{l}} (c\delta_{kl} - 1) \mu_{k} \mu_{l}^{\top} \Lambda^{-1} q q^{\top} + \sum_{k=1}^{c} \frac{\partial \zeta_{k}(a)}{\partial a_{k}} c q q^{\top} \right. \\ \left. + \sum_{k,l,n=1}^{c} \frac{\partial^{2} \zeta_{k}(a)}{\partial a_{l} \partial a_{n}} (c\delta_{ln} - 1) \mu_{l}^{\top} \Lambda^{-1} q \mu_{n}^{\top} \Lambda^{-1} q \mu_{k} q^{\top} / 2 + \sum_{k,l=1}^{c} \frac{\partial^{2} \zeta_{k}(a)}{\partial a_{l}^{2}} c q^{\top} \Lambda^{-1} q \mu_{k} q^{\top} / 2 \right] \right\|_{\max} \\ \left. + o(1/N), \end{aligned}$$

where  $a = \mu^{\top} \Lambda^{-1} q$ ,  $a_k = \mu_k^{\top} \Lambda^{-1} q$ ,  $S = c^2 \nabla^2 \widetilde{L}(c \Lambda^{-1})$ . Ignoring constants other than c, N, we have  $||G||_{\max} \leq O(c/N)$ . 

**Proof** According to Lemma G.3, the loss function L(W) has a unique global minimizer  $W^*$ . We have

$$\nabla L(W^*) = \mathbb{E}\left[\sum_{k=1}^{c} \left[ (\zeta_k (P^\top W^* q/c) - \zeta_k (\mu^\top \Lambda^{-1} q)) p_k q^\top / c \right] \right] = 0.$$
(46)

Let  $W^* = c(\Lambda^{-1} + G)$ ,  $a = \mu^\top \Lambda^{-1}q$ ,  $a_k = \mu_k^\top \Lambda^{-1}q$ ,  $b = (\mu + c\mu_h + g)^\top Gq + (c\mu_h + g)^\top \Lambda^{-1}q$ ,  $b_k = (\mu_k + ch_k\mu_k + g_k)^\top Gq + (ch_k\mu_k + g_k)^\top \Lambda^{-1}q$ . The Taylor expansion of  $\zeta_k(a + b)$  at point a is 

$$\zeta_k(a+b) = \zeta_k(a) + \sum_{l=1}^c \frac{\partial \zeta_k(a)}{\partial a_l} b_l + \sum_{l,n=1}^c \frac{\partial^2 \zeta_k(a)}{\partial a_l \partial a_n} b_l b_n / 2! + \sum_{l,n,m=1}^c R_{klnm}(a,b) b_l b_n b_m / 3!,$$

where  $|R_{klnm}(a,b)| \leq \sup_{x} |\frac{\partial^{3}\zeta_{k}(x)}{\partial x_{l}\partial x_{n}\partial x_{m}}|$ . Thus, our equation (46) become 

$$\mathbb{E} \begin{bmatrix} \sum_{k,l=1}^{c} \frac{\partial \zeta_k(a)}{\partial a_l} b_l p_k q^\top + \sum_{k,l,n=1}^{c} \frac{\partial^2 \zeta(a)}{\partial a_l \partial a_n} b_l b_n p_k q^\top / 2! + \sum_{k,l,n,m=1}^{c} R_{klnm}(a,b) b_l b_n b_m p_k q^\top / 3! \end{bmatrix} = 0$$

$$(47)$$

For the first term  $\sum_{k,l=1}^{c} \frac{\partial \zeta_k(a)}{\partial a_l} b_l p_k q^{\top}$ , according to Lemma G.1, we have  $\mathbb{E}\left[\sum_{k,l=1}^{c} \frac{\partial \zeta_k(a)}{\partial a_l} b_l p_k q^{\top}\right]$  $= \mathbb{E} \left[ \sum_{l=1}^{c} \frac{\partial \zeta_k(a)}{\partial a_l} \left[ \mu_l^\top G q \mu_k q^\top + c^2 h_l h_k \mu_l^\top G q \mu_k q^\top + c^2 h_l h_k \mu_l^\top \Lambda^{-1} q \mu_k q^\top + g_l^\top \Lambda^{-1} q g_k q^\top + g_l^\top G q g_k q^\top \right] \right]$  $= \mathbb{E} \left[ \sum_{l=1}^{c} \frac{\partial \zeta_k(a)}{\partial a_l} \left( \mu_k \mu_l^\top G q q^\top + (c \delta_{kl} - 1) \mu_k \mu_l^\top G q q^\top / N + (c \delta_{kl} - 1) \mu_k \mu_l^\top \Lambda^{-1} q q^\top / N \right) \right]$  $+\sum_{k=1}^{c} \frac{\partial \zeta_k(a)}{\partial a_k} \left( cqq^\top / N + c\Lambda Gqq^\top / N \right) \bigg|.$ (48)According to Lemma G.4,  $O(||G||_{\max}) = O(N^{-1/4}) = o(1)$ , we have  $\left\|\mathbb{E}\left[\sum_{l=l=1}^{c} \frac{\partial \zeta_k(a)}{\partial a_l} (c\delta_{kl} - 1) \mu_k \mu_l^\top Gqq^\top / N + \sum_{l=1}^{c} \frac{\partial \zeta_k(a)}{\partial a_k} c\Lambda Gqq^\top / N\right]\right\|_{\mathcal{H}}$  $\leq O(\|G\|_{\max}/N) = o(1/N)$ (49)For the second term  $\sum_{k,l,n=1}^{c} \frac{\partial^2 \zeta_k(a)}{\partial a_l \partial a_n} b_l b_n p_k q^{\top}/2!$ , we have  $\mathbb{E}\left[\sum_{l=1,n=1}^{c} \frac{\partial^2 \zeta_k(a)}{\partial a_l \partial a_n} b_l b_n p_k q^\top / 2!\right]$  $= \frac{1}{2} \mathbb{E} \left[ \sum_{k,l,n=1}^{c} \frac{\partial^2 \zeta_k(a)}{\partial a_l \partial a_n} \left( \underbrace{\sum_{\substack{\phi_1 \in \{\mu_l, ch_l \mu_l, g_l\}, \phi_2 \in \{\mu_n, ch_n \mu_n, g_n\}}}_{(c)} \phi_1^\top Gq \phi_2^\top Gq p_k q^\top \right) \right]$  $\underbrace{\sum_{\substack{\phi_1 \in \{\mu_l, ch_l \mu_l, g_l\}, \phi_2 \in \{ch_n \mu_n, g_n\}}} 2\phi_1^\top G q \phi_2^\top \Lambda^{-1} q p_k q^\top}_{(ii)}}_{(ii)}$  $\underbrace{\sum_{\substack{\phi_1 \in \{ch_l \mu_l, g_l\}, \phi_2 \in \{ch_n \mu_n, g_n\}}}}_{(\cdots)} \phi_1^\top \Lambda^{-1} q \phi_2^\top \Lambda^{-1} q p_k q^\top} \right) \Bigg].$ For terms (i) having two G, their max norms are at most smaller than  $O(||G||_{\max}^2)$ . For terms (ii) 

having one G, define  $\bar{g}_l = \Lambda^{-1/2} g_l$ , these terms must contain  $n_{1j}$  number of  $h_j$  and  $n_{2ji}$  number of  $(\bar{g}_j)_i$ , we have  $\sum_{j \in [c], i \in [d]} n_{1j} + n_{2ji} = n_t, n_t = 1, 2, 3$ . According to Lemma G.2, we know that for  $n_t = 1, 2, 3$ ,

$$\mathbb{E}[\prod_{j\in[c],i\in[d]} h_j^{n_{1j}}(\bar{g}_j)_i^{n_{2ji}}] = O(N^{-1})$$

Thus, the max norm of expectations of terms in (ii) are at most smaller than  $O(||G||_{\max}N^{-1})$ . Therefore, for terms (i), (ii), we have

1780 
$$\|\mathbb{E}[(i)]\|_{\max} \le O(\|G\|_{\max}^2) = o(\|G\|_{\max})$$
(50)

$$\|\mathbb{E}[(ii)]\|_{\max} \le O(\|G\|_{\max}/N) = o(1/N)$$
(51)

For terms (iii) without G, we have  $\|\mathbb{E}[(iii)]\|_{\max}$  $= \left\| \mathbb{E} \left[ \sum_{k,l,n=1}^{c} \frac{\partial^2 \zeta_k(a)}{\partial a_l \partial a_n} c^2 h_l h_n \mu_l^\top \Lambda^{-1} q \mu_n^\top \Lambda^{-1} q \mu_k q^\top / 2 + \sum_{k,l=1}^{c} \frac{\partial^2 \zeta_k(a)}{\partial a_l^2} g_l^\top \Lambda^{-1} q g_l^\top \Lambda^{-1} q \mu_k q^\top / 2 \right] \right\}$  $+\sum_{k,l=1}^{c} \frac{\partial^2 \zeta_k(a)}{\partial a_l \partial a_k} ch_l \mu_l^{\top} \Lambda^{-1} qg_k^{\top} \Lambda^{-1} qg_k q^{\top} + \sum_{k,l=n=1}^{c} \frac{\partial^2 \zeta_k(a)}{\partial a_l \partial a_n} c^3 h_l h_n h_k \mu_l^{\top} \Lambda^{-1} q\mu_n^{\top} \Lambda^{-1} q\mu_k q^{\top} / 2 \left[ \left[ \left[ \frac{\partial^2 \zeta_k(a)}{\partial a_l \partial a_k} - \frac{\partial^2 \zeta_k(a)}{\partial a_k} - \frac{\partial^2 \zeta_k$  $\leq \frac{1}{2N} \left\| \mathbb{E} \left[ \sum_{k,l,n=1}^{c} \frac{\partial^2 \zeta_k(a)}{\partial a_l \partial a_n} (c \delta_{ln} - 1) \mu_l^\top \Lambda^{-1} q \mu_n^\top \Lambda^{-1} q \mu_k q^\top + \sum_{k,l=1}^{c} \frac{\partial^2 \zeta_k(a)}{\partial a_l^2} c q^\top \Lambda^{-1} q \mu_k q^\top \right] \right\|_{\mathbf{m}_{\mathbf{c}}}$  $+ O(1/N^2)$ (52)where the last inequity is due to Lemma G.1, G.2. For the third term  $\sum_{k,l,n,m=1}^{c} R_{klnm}(a,b) b_l b_n b_m p_k q^{\top}/3!$ , we have  $\left\| \mathbb{E} \left[ \sum_{k,l,n,m=1}^{c} R_{klnm}(a,b) b_l b_n b_m p_k q^\top / 3! \right] \right\|$  $\leq O(1) \max_{l,m \in [d]} \mathbb{E} \left[ \sum_{k_1,k_2,k_3,k_4 \in [c]} |b_{k_1}b_{k_2}b_{k_3}(p_{k_4})_l q_m| \right]$  $\leq O(1)\mathbb{E}\sum_{k_1,k_2,k_3,k_4\in[c]} \left| \underbrace{\sum_{\phi_1\in\{\mu_{k_1},ch_{k_1}\mu_{k_1},g_{k_1}\},\phi_2\in\{\mu_{k_2},ch_{k_2}\mu_{k_2},g_{k_2}\},\phi_3\in\{\mu_{k_3},ch_{k_3}\mu_{k_3},g_{k_3}\}}_{(r)} \phi_1^\top Gq\phi_2^\top Gq\phi_3^\top Gq(p_{k_4})_l q_m \right|$  $\underbrace{\underbrace{\phi_{1} \in \{\mu_{k_{1}}, ch_{k_{1}}\mu_{k_{1}}, g_{k_{1}}\}, \phi_{2} \in \{\mu_{k_{2}}, ch_{k_{2}}\mu_{k_{2}}, g_{k_{2}}\}, \phi_{3} \in \{ch_{k_{3}}\mu_{k_{3}}, g_{k_{3}}\}}_{(*)}}_{(*)}}_{(*)}$  $\phi_1^{\top} G q \phi_2^{\top} G q \phi_3^{\top} \Lambda^{-1} q (p_{k_A})_l q_m$  $\underbrace{\sum_{\substack{\phi_1 \in \{\mu_{k_1}, ch_{k_1}\mu_{k_1}, g_{k_1}\}, \phi_2 \in \{ch_{k_2}\mu_{k_2}, g_{k_2}\}, \phi_3 \in \{ch_{k_3}\mu_{k_3}, g_{k_3}\}}_{(**)}} \phi_1^\top Gq\phi_2^\top \Lambda^{-1}q\phi_3^\top \Lambda^{-1}q(p_{k_4})_l q_m}$ + $\underbrace{\sum_{\phi_1 \in \{ch_{k_1}\mu_{k_1}, g_{k_1}\}, \phi_2 \in \{ch_{k_2}\mu_{k_2}, g_{k_2}\}, \phi_3 \in \{ch_{k_3}\mu_{k_3}, g_{k_3}\}}_{\phi_1^\top \Lambda^{-1} q \phi_2^\top \Lambda^{-1} q \phi_3^\top \Lambda^{-1} q (p_{k_4})_l q_m} \bigg].$ +

For terms in (\*) having two or three G, these terms' expected absolute values are at most smaller than  $O(||G||_{\max}^2)$ . For terms in (\*\*) having one G, these terms must contain  $n_{1j}$  number of  $h_j$  and  $n_{2ji}$  number of  $(\bar{g}_j)_i$ , we have  $\sum_{j \in [c], i \in [d]} n_{1j} + n_{2ji} = n_t, n_t = 2, 3, 4$ . According to Lemma G.2, for  $n_t = 2, 3, 4$ , we have

$$\mathbb{E}\left[\prod_{j\in[c],i\in[d]}|h_k^{n_{1k}}(\bar{g}_j)_i^{n_{2ji}}|\right] = O(N^{-n_t/2}) = O(N^{-1})$$

Thus, these term's expected absolute values are at most smaller than  $O(||G||_{\max}N^{-1})$ . For terms in (\* \* \*) without G, these terms must contain  $n_{1j}$  number of  $h_j$  and  $n_{2ji}$  number of  $(\bar{g}_j)_i$ , we have  $\sum_{j \in [c], i \in [d]} n_{1j} + n_{2ji} = n_t, n_t = 3, 4$ . According to Lemma G.2, for  $n_t = 3, 4$ , we have  $\mathbb{E}[\prod_{j \in [c], i \in [d]} |h_k^{n_{1k}}(\bar{g}_j)_i^{n_{2ji}}|] = O(N^{-n_t/2}) = O(N^{-3/2})$  Thus, these term's expected absolute values are at most smaller than  $O(N^{-3/2})$ . Therefore, we have 

 $\leq O(\|G\|_{max}^2) + O(\|G\|_{\max}N^{-1}) + O(N^{-3/2})$ 

 $\leq o(\|G\|_{\max}) + o(1/N).$ 

Moreover, we have

$$\begin{cases}
\mathbb{E} \left[ \sum_{k,l=1}^{c} \frac{\partial \zeta_{k}(a)}{\partial a_{l}} \mu_{k} \mu_{l}^{\top} Gqq^{\top} \right] \right\}_{ij} \\
\mathbb{E} \left[ \sum_{k,l=1}^{c} \frac{\partial \zeta_{k}(a)}{\partial a_{l}} \mu_{k} \mu_{l}^{\top} Gqq^{\top} \right] \right\}_{ij} \\
\mathbb{E} \left[ \sum_{k=1}^{c} \zeta_{k}(a)(1 - \zeta_{k}(a)) \mu_{k} \mu_{k}^{\top} Gqq^{\top} - \sum_{k,l=1,k\neq l}^{c} \zeta_{k}(a) \zeta_{l}(a) \mu_{k} \mu_{l}^{\top} Gqq^{\top} \right] \right\}_{ij} \\
\mathbb{E} \left[ \sum_{k,l=1,k\neq l}^{c} \zeta_{k}(a) \zeta_{l}(a) \mu_{k} (\mu_{k} - \mu_{l})^{\top} Gqq^{\top} \right] \right\}_{ij} \\
\mathbb{E} \left[ \sum_{k=2}^{c} \sum_{l=1}^{k-1} \zeta_{k}(a) \zeta_{l}(a) (\mu_{k} - \mu_{l}) (\mu_{k} - \mu_{l})^{\top} Gqq^{\top} \right] \right\}_{ij} \\
\mathbb{E} \left[ \sum_{k=2}^{c} \sum_{l=1}^{k-1} \zeta_{k}(a) \zeta_{l}(a) (\mu_{k} - \mu_{l}) (\mu_{k} - \mu_{l})^{\top} Gqq^{\top} \right] \right\}_{ij} \\
\mathbb{E} \left[ \sum_{k=2}^{d} \sum_{l=1}^{d} s_{ijnm} G_{nm}, \\
\mathbb{E} \left[ \sum_{k=2}^{d} \sum_{l=1}^{d} s_{ijnm} G_{lm}, \\
\mathbb{E} \left[ \sum_{k=2}^{d} \sum_{l=1}^{d} \sum_{l=1}^{d} s_{ijnm} G_{lm}, \\
\mathbb{E} \left[ \sum_{k=2}^{d} \sum_{l=1}^{d} \sum_{l=1}^{d} s_{ijnm} G_{lm}, \\
\mathbb{E} \left[ \sum_{k=2}^{d} \sum_{l=1}^{d} \sum_{$$

where  $s_{ijnm} = \mathbb{E}\left[\sum_{k=2}^{c} \sum_{l=1}^{k-1} \zeta_k(a)\zeta_l(a)(\mu_k - \mu_l)_i(\mu_k - \mu_l)_n q_m q_j\right]$ . We vectorize G as  $\operatorname{Vec}(G)_i = G_{t_1(i), t_2(i)}$ . Define  $S \in \mathbb{R}^{d^2 \times d^2}$ , where  $S_{ij} = s_{t_1(i), t_2(i), t_1(j), t_2(j)} = \mathbb{E}\left[\sum_{k=2}^{c} \sum_{l=1}^{k-1} \zeta_k(a)\zeta_l(a)(\mu_k - \mu_l)_{t_1(i)}q_{t_2(i)}(\mu_k - \mu_l)_{t_1(j)}q_{t_2(j)}\right]$ , (54) can be expressed as 

$$\mathbb{E}\left[\sum_{k,l=1}^{c} \frac{\partial \zeta_k(a)}{\partial a_l} \mu_k \mu_l^\top G q q^\top\right] = SG.$$
(55)

(53)

Note that  $S = c^2 \nabla^2 \widetilde{L}(c \Lambda^{-1})$ . According to Lemma G.3, S is positive definite. Thus, combining (47), (48), (49), (50), (51), (52), (53), (55), we have 

$$\begin{aligned} \|G\|_{\max} & \|G\|_{\max} \\ \|S^{77} & \|G\|_{\max} \\ & \leq \frac{1}{N} \left\| S^{-1} \mathbb{E} \left[ \sum_{k,l=1}^{c} \frac{\partial \zeta_{k}(a)}{\partial a_{l}} (c\delta_{kl} - 1) \mu_{k} \mu_{l}^{\top} \Lambda^{-1} q q^{\top} + \sum_{k=1}^{c} \frac{\partial \zeta_{k}(a)}{\partial a_{k}} c q q^{\top} \\ & + \sum_{k,l,n=1}^{c} \frac{\partial^{2} \zeta_{k}(a)}{\partial a_{l} \partial a_{n}} (c\delta_{ln} - 1) \mu_{l}^{\top} \Lambda^{-1} q \mu_{n}^{\top} \Lambda^{-1} q \mu_{k} q^{\top} / 2 + \sum_{k,l=1}^{c} \frac{\partial^{2} \zeta_{k}(a)}{\partial a_{l}^{2}} c q^{\top} \Lambda^{-1} q \mu_{k} q^{\top} / 2 \right] \right\|_{\max} \\ & + o(1/N). \end{aligned}$$

Ignoring constants other than c, N, we have  $||G||_{\max} \leq O(c/N)$ .

**Lemma G.6** The loss function (7) is *l*-smooth, where  $l \leq \frac{1}{c^2} \sum_{k=2}^{c} \sum_{l=1}^{k-1} \sum_{i \in [d^2]} \mathbb{E}[((p_k - 1)^{k-1}) \sum_{i \in [d^2]} \mathbb{E}[(p_k - 1)^{k-1}) \sum_{i \in [d^2]} \mathbb{E}$  $p_l)_{t_1(i)} q_{t_2(i)})^2].$ 

**Proof** The Hessian matrix of the loss function is 

$$(\nabla^2 L(W))_{ij} = \mathbb{E} \left[ \sum_{k=2}^{c} \sum_{l=1}^{k-1} \operatorname{softmax}(P^\top Wq/c)_k \operatorname{softmax}(P^\top Wq/c)_l (p_k - p_l)_{t_1(i)} q_{t_2(i)} (p_k - p_l)_{t_1(j)} q_{t_2(j)} / c^2 \right]$$

$$1894$$

$$1895$$

Considering  $z \in \mathbb{R}^{d^2}$  such that  $z \neq 0$ , we have

$$z^{\top} \nabla^{2} L(W) z$$

$$= \mathbb{E} \left[ \frac{1}{c^{2}} \sum_{k=2}^{c} \sum_{l=1}^{k-1} \operatorname{softmax}(P^{\top} Wq/c)_{k} \operatorname{softmax}(P^{\top} Wq/c)_{l} \left( \sum_{a \in [d^{2}]} z_{a}(p_{k} - p_{l})_{t_{1}(a)}q_{t_{2}(a)} \right)^{2} \right]$$

$$\stackrel{(a)}{\leq} \frac{1}{c^{2}} \|z\|_{2}^{2} \sum_{k=2}^{c} \sum_{l=1}^{k-1} \sum_{i \in [d^{2}]} \mathbb{E}[((p_{k} - p_{l})_{t_{1}(i)}q_{t_{2}(i)})^{2}]$$

where (a) is due to the Cauchy–Schwarz inequality. Thus,  $\nabla^2 L(W) \leq lI_d$  and L(W) is *l*-smooth, where *l* is a constant smaller than  $\frac{1}{c^2} \sum_{k=2}^{c} \sum_{l=1}^{k-1} \sum_{i \in [d^2]} \mathbb{E}[((p_k - p_l)_{t_1(i)} q_{t_2(i)})^2].$ 

**Theorem G.1 (Formal statement of Theorem 4.1)** The following statements hold. 

(1) Optimizing training loss L(W) (16) with training prompt length N via gradient descent  $W^{t+1} =$  $W^t - \eta \nabla L(W^t)$ , we have for any t 

$$||W^{t} - W^{*}||_{F}^{2} \le \exp(-t/\kappa) ||W^{0} - W^{*}||_{F}^{2},$$

where  $W^0$  is the initial parameter and  $W^*$  is the global minimizer of L(W),  $\kappa = l/\alpha$ .  $\alpha$ , l are constants such that

$$0 < \alpha \le \lambda_{\min}(\nabla^2 L(W)) \le \lambda_{\max}(\nabla^2 L(W)) \le l, \text{ for all } W \in R_W,$$
(56)

where  $R_W = \{ W \in \mathbb{R}^{d \times d} \mid ||W - W^*||_F \le ||W^0 - W^*||_F \}.$ 

(2) Denoting  $W^* = c(\Lambda^{-1} + G)$ , we have 

$$\begin{aligned} \|G\|_{\max} &\leq \frac{1}{N} \left\| S^{-1} \mathbb{E} \left[ \sum_{k,l=1}^{c} \frac{\partial \zeta_k(a)}{\partial a_l} (c\delta_{kl} - 1) \mu_k \mu_l^\top \Lambda^{-1} q q^\top + \sum_{k=1}^{c} \frac{\partial \zeta_k(a)}{\partial a_k} c q q^\top \right. \\ &+ \frac{1}{2} \sum_{k,l,n=1}^{c} \frac{\partial^2 \zeta_k(a)}{\partial a_l \partial a_n} (c\delta_{ln} - 1) \mu_l^\top \Lambda^{-1} q \mu_n^\top \Lambda^{-1} q \mu_k q^\top + \frac{1}{2} \sum_{k,l=1}^{c} \frac{\partial^2 \zeta_k(a)}{\partial a_l^2} c q^\top \Lambda^{-1} q \mu_k q^\top \right] \right\|_{\max} \\ &+ o(1/N) \end{aligned}$$

$$= O(c/N)$$

> where  $S = c^2 \nabla^2 \widetilde{L}(2\Lambda^{-1})$ ,  $\widetilde{L}(2\Lambda^{-1}) = \lim_{N \to \infty} L(2\Lambda^{-1})$ . The expectation is taken over  $\mu_{\tau} \sim L(2\Lambda^{-1})$ .  $\mathcal{P}^m_{\Omega}(\Lambda)$ ,  $x_{\tau, \mathsf{query}} \sim \mathcal{P}^m_x(\mu_{\tau}, \Lambda)$ .

(3) After  $T \ge 2\kappa \log(N \cdot ||W^0 - W^*||_F)$  gradient steps, denoting  $\widehat{W}$  as the final model, we have 

$$\widehat{W} = c(\Lambda^{-1} + \widehat{G}),\tag{57}$$

where  $\|\widehat{G}\|_{\max} = O(c/N)$ .

#### G.3 PROOF OF THEOREM 4.1

**Proof** According to Lemma G.5, the global minimizer of L(W) is  $W^* = c(\Lambda^{-1} + G)$ , where

 $||G||_{\max}$ 

$$\leq \frac{1}{N} \left\| S^{-1} \mathbb{E} \left[ \sum_{k,l=1}^{c} \frac{\partial \zeta_k(a)}{\partial a_l} (c \delta_{kl} - 1) \mu_k \mu_l^\top \Lambda^{-1} q q^\top + \sum_{k=1}^{c} \frac{\partial \zeta_k(a)}{\partial a_k} c q q^\top \right. \\ \left. + \sum_{k,l,n=1}^{c} \frac{\partial^2 \zeta_k(a)}{\partial a_l \partial a_n} (c \delta_{ln} - 1) \mu_l^\top \Lambda^{-1} q \mu_n^\top \Lambda^{-1} q \mu_k q^\top / 2 + \sum_{k,l=1}^{c} \frac{\partial^2 \zeta_k(a)}{\partial a_l^2} c q^\top \Lambda^{-1} q \mu_k q^\top / 2 \right] \right\|_{\max} \\ \left. + o(1/N). \right\}$$

Ignoring constants other than c, N, we have  $||G||_{\max} \leq O(c/N)$ .

Define  $R_W = \{W \in \mathbb{R}^{d \times d} | \|W - W^*\|_F \leq \|W^0 - W^*\|_F\}$ , and  $R_W$  is a compact set. Then, according to Lemma G.3, for  $W \in R_W$ , we have  $\nabla^2 L(W) \succeq \alpha I_d$ . Here  $\alpha > 0$  is a positive constant number. Thus, L(W) is  $\alpha$ -strongly convex in  $R_W$ . Moreover, according to Lemma G.6, L(W) is l-smooth. Then according to Lemma D.2, applying gradient descent with  $\eta = 1/l$ , for any  $t \geq 1$ , we have 

$$||W^t - W^*||_F^2 \le \exp(-t/\kappa) \cdot ||W^0 - W^*||_F^2,$$

where  $\kappa = l/\alpha$ .

After  $T \ge 2\kappa \log(N \cdot ||W^0 - W^*||_F)$  gradient steps, we have  $\widehat{W} = W^T = c(\Lambda^{-1} + G + H^T/c) = C(\Lambda^{-1} + G + H^T/c)$  $2(\Lambda^{-1} + \widehat{G})$ , where  $\widehat{G} = G + H^T/c$ ,  $\|H^T\|_{\max} \leq \exp(-T/\kappa) \cdot \|W^0 - W^*\|_F^2 \leq 1/N$ . Thus,  $\|\widehat{G}\|_{\max} \le \|G\|_{\max} + \|H^T\|_{\max} = O(c/N).$ 

#### Η **IN-CONTEXT INFERENCE OF MULTI-CLASS CLASSIFICATION**

#### H.1 NOTATIONS

In this section, we use the following notations. We denote  $\mu = (\mu_1, \mu_2, \dots, \mu_c), q = x_{query}$ . Define  $p_k = \frac{c}{M} \sum_{i=1}^M (y_i)_k x_i$ , and define  $P = (p_1, p_2, \dots, p_c) \in \mathbb{R}^{d \times c}$ . We have  $P^\top = \frac{c}{M} \sum_{i=1}^M y_i x_{\tau, i}^\top \in \mathbb{R}^{d \times c}$ .  $\mathbb{R}^{c \times d}$ . Since with probability  $\mathbb{P}(y_{\tau,i} = \mathbf{e}_k) = 1/c, x_{\tau,i} = \mu_k + v_i$ , where  $v_i \sim \mathsf{N}(0,\Lambda)$ , we have  $p_k = \frac{c}{M} \sum_{i=1}^{M} (y_{\tau,i})_k x_{\tau,i} = c M_k \mu_k / M + g_k$ , where  $g_k = \frac{c}{M} \sum_{i \in \{i | y_{\tau,i} = \mathbf{e}_k\}} v_i$ ,  $g_k \sim \mathsf{N}(0, c^2 M_k \Lambda/M^2)$  and  $(M_1, M_2, \dots, M_c) \sim \mathsf{Multin}(M, 1/c)$ . Defining  $h_k = M_k/M - 1/c$ , we have  $M_k/M = 1/c + h_k$  and  $p_k = \mu_k + ch_k\mu_k + g_k$ . 

**Theorem H.1 (Formal statement of Theorem 4.2)** Let  $\hat{y}_{query}$  be the prediction of the trained transformer with parameters  $\widehat{W}$  in (19) and  $P_{\text{test}}$  satisfying Assumption 4.2, and let  $y_{\text{query}} \sim$  $\mathcal{P}^m_{y|x_{query}}(\mu,\Lambda)$ . Then, for the inference error defined in (3), we have

$$\begin{aligned} & \mathbb{E}[\Delta(y_{\mathsf{query}}, \widehat{y}_{\mathsf{query}})] \\ & \text{1990} \\ & \text{1991} \\ & \text{1992} \\ & \text{1992} \\ & \text{1993} \\ & \text{1993} \\ & \text{1994} \\ & \text{1995} \\ & \text{1996} \end{aligned} \\ & + o\left(\frac{1}{N} + \frac{1}{\sqrt{M}}\right), \end{aligned}$$

where  $a = \mu^{\top} \Lambda^{-1} q$ ,  $a_k = \mu_k^{\top} \Lambda^{-1} q$ . The expectation is taken over  $\{x_i, y_i\}_{i=1}^M \stackrel{\text{i.i.d.}}{\sim} \mathcal{P}^m(\mu, \Lambda)$ .

## <sup>1998</sup> H.2 PROOF OF THEOREM 4.2

**Proof** The output of the trained transformer is

$$\widehat{y}_{\mathsf{out}} = \operatorname{softmax}\left(\left(\frac{c}{M}\sum_{i=1}^{M}y_{i}x_{i}^{\top}\right)(\Lambda^{-1}+\widehat{G})x_{\mathsf{query}}\right) = \operatorname{softmax}(P^{\top}(\Lambda^{-1}+\widehat{G})q)$$
(58)

The probability of  $y_{query} = \mathbf{e}_k$  given  $x_{query}$  is

 $\mathbb{P}\left(y_{\mathsf{query}} = \mathbf{e}_k | x_{\mathsf{query}}\right) = \operatorname{softmax}(\mu^\top \Lambda^{-1} x_{\mathsf{query}})_k = \operatorname{softmax}(\mu^\top \Lambda^{-1} q)_k$ 

Defining  $a = \mu^{\top} \Lambda^{-1} q$ ,  $b = (\mu + \mu_h + g)^{\top} \widehat{G} q + (\mu_h + g)^{\top} \Lambda^{-1} q$ ,  $a_k = \mu_k^{\top} \Lambda^{-1} q$ ,  $b_k = (\mu_k + ch_k \mu_k + g_k)^{\top} \widehat{G} q + (ch_k \mu_k + g_k)^{\top} \Lambda^{-1} q$ , we have

$$\mathbb{E}\left[\operatorname{softmax}(P^{\top}(\Lambda^{-1}+\widehat{G})q)_{k}\right] = \mathbb{E}\left[\zeta_{k}(a+b)\right] = \mathbb{E}[\zeta_{k}(a) + \sum_{l=1}^{c} \frac{\partial\zeta_{k}(a)}{\partial a_{l}}b_{l} + \sum_{l,n=1}^{c} R_{kln}(a,b)b_{l}b_{n}/2]$$

where  $|R_{kln}(a,b)| \leq \sup_{x} |\frac{\partial^2 \zeta_k(x)}{\partial x_l \partial x_n}|$ . Thus, we have

$$\mathbb{E}[|\zeta_k(a+b) - \zeta_k(a)|] \le \mathbb{E}\left[\sum_{l=1}^c \left|\frac{\partial \zeta_k(a)}{\partial a_l}b_l\right|\right] + \mathbb{E}\left[\left|\sum_{l,n=1}^c R_{kln}(a,b)b_lb_n/2\right|\right].$$

We first consider the term  $\mathbb{E}\left[\sum_{l=1}^{c} \left| \frac{\partial \zeta_k(a)}{\partial a_l} b_l \right| \right]$ . Defining  $\bar{g}_l = \Lambda^{-1/2} g_l$ , we have  $\mathbb{E}\left[\sum_{l=1}^{c} \left| \frac{\partial \zeta_k(a)}{\partial a_l} b_l \right| \right]$ 

$$\leq \sum_{l=1}^{c} \frac{\partial \zeta_{k}(a)}{\partial a_{l}} \left( |\mu_{l}^{\top} \widehat{G}q| + \mathbb{E}[|ch_{l}\mu_{l}^{\top} \widehat{G}q|] + \mathbb{E}[|g_{l}^{\top} \widehat{G}q|] + \mathbb{E}[|ch_{l}\mu_{l}^{\top} \Lambda^{-1}q|] + \mathbb{E}[|g_{l}^{\top} \Lambda^{-1}q|] \right)$$

$$\leq \sum_{l=1}^{c} \frac{\partial \zeta_{k}(a)}{\partial a_{l}} \left( |\mu_{l}^{\top} \widehat{G}q| + \frac{\sqrt{c(1-1/c)}}{M^{1/2}} |\mu_{l}^{\top} \widehat{G}q| + \mathbb{E}[|g_{l}^{\top} \Lambda^{1/2} \widehat{G}q|] + \frac{\sqrt{c(1-1/c)}}{M^{1/2}} |\mu_{l}^{\top} \Lambda^{-1}q| + \mathbb{E}[|g_{l}^{\top} \Lambda^{-1/2}q|] \right)$$

$$\leq \sum_{l=1}^{c} \frac{\partial \zeta_{k}(a)}{\partial a_{l}} \left[ \|\widehat{G}\|_{\max} \sum_{i,j \in [d]} |(\mu_{l})_{i}q_{j}| + \frac{1}{M^{1/2}} \left( \sqrt{c(1-1/c)} |\mu_{l}^{\top} \Lambda^{-1}q| + \sqrt{c} \sum_{i,j \in [d]} |\Lambda_{ij}^{-1/2}q_{j}| \right) \right]$$

$$+ o\left(\frac{1}{N} + \frac{1}{\sqrt{M}}\right),$$

where (a) is due to Lemma G.1 that  $\mathbb{E}[|h|] \leq M^{-1/2}c^{-1/2}(1-1/c)^{1/2}$ . (b) is because that  $\bar{g}_l \sim \mathbb{N}(0, c^2 M_l I_d/M^2)$ ,  $\mathbb{E}[|(\bar{g}_l)_i|] \leq \mathbb{E}[(\bar{g}_l)_i^{2}]^{1/2} = (c/M)^{1/2}$ , for  $l \in [c], i \in [d]$ .

For 
$$\mathbb{E}\left[\left|\sum_{l,n=1}^{c} R_{kln}(a,b)b_{l}b_{n}/2\right|\right]$$
, we have  

$$\mathbb{E}\left[\left|\sum_{l,n=1}^{c} R_{kln}(a,b)b_{l}b_{n}/2\right|\right] = O(1)\mathbb{E}\left[\sum_{l,n=1}^{c} \left(\sum_{\substack{\phi_{1} \in \{\mu_{l},ch_{l}\mu_{l},g_{l}\},\phi_{2} \in \{\mu_{n},ch_{n}\mu_{n},g_{n}\}} \left|\phi_{1}^{\top}\widehat{G}q\phi_{2}^{\top}\widehat{G}q\right|\right] + \underbrace{\sum_{\substack{\phi_{1} \in \{\mu_{l},ch_{l}\mu_{l},g_{l}\},\phi_{2} \in \{ch_{n}\mu_{n},g_{n}\}}_{(i)}}_{(ii)}\left|\phi_{1}^{\top}\Lambda^{-1}q\phi_{2}^{\top}\Lambda^{-1}q\right| + \underbrace{\sum_{\substack{\phi_{1} \in \{\mu_{l},ch_{l}\mu_{l},g_{l}\},\phi_{2} \in \{ch_{n}\mu_{n},g_{n}\}}_{(iii)}}_{(iii)}\left|\phi_{1}^{\top}\Lambda^{-1}q\phi_{2}^{\top}\Lambda^{-1}q\right|\right)$$

For terms (i) having two  $\hat{G}$ , they are at most smaller than  $O(\|\hat{G}\|_{\max}^2) = O(1/N^2)$ . For terms (ii) having one G, these terms must contain  $n_{1j}$  number of  $h_j$  and  $n_{2ji}$  number of  $(\bar{g}_j)_i$ , we have  $\sum_{j \in [c], i \in [d]} n_{1j} + n_{2ji} = n_t, n_t = 1, 2$ . According to Lemma G.2, we know that for  $n_t = 1, 2$ ,  $\mathbb{E}[\prod_{j \in [c], i \in [d]} |h_j^{n_{1j}}(\bar{g}_j)_i^{n_{2ji}}|] = O(M^{-1/2}).$  Thus, terms in (ii) are at most smaller than  $O(||G||_{\max}M^{-1/2}) = O(1/(N\sqrt{M}))$ . For terms (*iii*) without G, these terms must contain  $n_{1j}$  number of  $h_j$  and  $n_{2ji}$  number of  $(\bar{g}_j)_i$ , we have  $\sum_{j \in [c], i \in [d]} n_{1j} + n_{2ji} = n_t, n_t = 2$ . According to Lemma G.2, for  $n_t = 2$ , we have

$$\mathbb{E}\left[\prod_{j\in[c],i\in[d]} |h_k^{n_{1k}}(\bar{g}_j)_i^{n_{2ji}}|\right] = O(M^{-n_t/2}) = O(M^{-1})$$

Thus, these term are  $O(M^{-1})$ . Therefore, we have  $\mathbb{E}\left[\left|\sum_{l,n=1}^{c} R_{kln}(a,b)b_lb_n/2\right|\right] = O(1/N^2 + 1/M + 1/(N\sqrt{M})) = o(1/N + 1/\sqrt{M}).$ 

Finally, we have

$$\begin{split} \mathbb{E}[\Delta(y_{\mathsf{query}}, \widehat{y}_{\mathsf{query}})] &= \max_{k} \{ \mathbb{E}[|\operatorname{softmax}(a+b)_{k} - \operatorname{softmax}(a)_{k}|] \} \\ &\leq \max_{k \in [c]} \left\{ \sum_{l=1}^{c} \frac{\partial \zeta_{k}(a)}{\partial a_{l}} \left[ \|\widehat{G}\|_{\max} \sum_{i,j \in [d]} |(\mu_{l})_{i}q_{j}| + \frac{1}{M^{1/2}} \left( \sqrt{c(1-1/c)} |\mu_{l}^{\top} \Lambda^{-1}q| + \sqrt{c} \sum_{i,j \in [d]} |\Lambda_{ij}^{-1/2}q_{j}| \right) \right] \right\} \\ &+ o\left( \frac{1}{N} + \frac{1}{\sqrt{M}} \right). \end{split}$$

2071 2072 2073

2069 2070

2056 2057 2058

2059 2060

2061

**Remark H.1** We note that Theorem 4.2 requires Assumption 4.2 to hold. For example, we need the covariance  $\Lambda$  in training and testing to be the same. A similar consistency requirement of the covariance  $\Lambda$  in training and testing had also been observed for in-context linear regression in Zhang et al. (2023a) and for in-context binary classification in the previous section 3.2.

2078 Here, we discuss the consequences when Assumption 4.2 does not hold. For example, suppose the 2079 labels of our data in test prompts are not balanced  $\mathbb{P}(y = e_k) = p_k$ ,  $\mu$  do not have the same 2080  $\Lambda^{-1}$  weighted norm  $\mu_k^{\top} \Lambda^{-1} \mu_k \triangleq \Psi_k$ , and the covariance matrix of test data is  $\Gamma \neq \Lambda$ , then as 2081  $N, M \to \infty$ , we have

$$\frac{c}{M}\sum_{i=1}^{M}y_ix_i^{\top} \to c(p_1\mu_1, p_2\mu_2, \dots, p_c\mu_c)^{\top},$$

2085 and

2082 2083 2084

2086

2089 2090

2091 2092

2093 2094

$$\mathbb{P}(\widehat{y}_{\mathsf{query}} = 1) \to \operatorname{softmax}(c(p_1\mu_1, p_2\mu_2, \dots, p_c\mu_c)^{\top}\Lambda^{-1}x_{\mathsf{query}}).$$

2087 Denote  $\Psi = (\Psi_1, \dots, \Psi_c)^\top$ ,  $\Phi = (\log(p_1), \dots, \log(p_c))^\top$  and  $z = \mu^\top \Gamma^{-1} x_{query} - \Psi/2 + \Phi$ . Then 2088 distribution of the ground truth label is

$$\mathbb{P}\left(y_{\mathsf{query}} = \boldsymbol{e}_k\right) = \operatorname{softmax}(z)_k$$

Define  $\hat{z} = c(p_1\mu_1, p_2\mu_2, \dots, p_c\mu_c)^{\top}\Lambda^{-1}x_{query}$ . Then, unless  $\hat{z} = z$  or  $\|\operatorname{softmax}(\hat{z}) - \operatorname{softmax}(z)\|_2$  is sufficiently small, the transformer cannot correctly perform the in-context multiclass classification.

### I ADDITIONAL EXPERIMENTS

In this section, we provide additional experimental results and the detailed experimental settings.

## I.1 SINGLE-LAYER TRANSFORMERS

2101 We train single-layer transformers for in-context classification of Gaussian mixtures with different 2102 numbers of classes c, different lengths of training prompts N, and test them with different test 2103 prompt lengths M. The results are reported in Figure 4. We can see from Figure 4 (a,b) that the 2104 inference errors decrease as N and M increase, and they increase as c increases. In Figure 4 (c,d), 2105 we first fix the training prompt length (test prompt length) to a large number 2000, and then vary 2106 the test prompt length (training prompt length) from 20 to 2000. The results show that, as M and



Figure 4: Inference errors of single-layer transformers. (a): Models trained on different training prompt lengths N on classification tasks involving c = 10 classes. (b): Models trained on different classification tasks involving c classes with a fixed training prompt length N = 80. (c): Relationship between the inference error and the test prompt length M in log-log axes. Training prompt length N = 2000 and number of classes c = 6. (d): Relationship between the inference error and the training prompt length N in log-log axes. Test prompt length M = 2000 and number of classes c = 6.

2135 N become sufficiently large, the inference error, which is an approximation of  $\mathbb{E}[\Delta(y_{query}, \hat{y}_{query})]$ 2136 (see Appendix I.2 for detailed definitions), decreases to near-zero. This indicates that the prediction 2137 of the trained transformer approaches the Bayes-optimal classifier. All these experimental results 2138 corroborate our theoretical claims.

2139 2140 2141

2134

I.2 EXPERIMENT DETAILS

2142 For all tasks, we set d = 20 and we randomly generate a covariance matrix  $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ , 2143 where  $\lambda_i = |\hat{\lambda}_i|$  and  $\hat{\lambda}_i \stackrel{\text{i.i.d.}}{\sim} N(3,1)$ . For each training dataset with different training prompt 2144 lengths N, and different class numbers c, we randomly generate B training samples. Training 2145 prompts  $P_{\tau}, \tau \in [B]$  and their corresponding labels  $y_{\tau,query}$  are generated according to Assumption 4.1. Moreover, we also generate testing datasets. For example, for each testing dataset, 2146 2147 we first randomly generate 20 pairs of  $(\mu_j, x_{j,query}, y_{j,prob}), j \in [20]$ , where  $(\mu_j) \stackrel{\text{i.i.d.}}{\sim} \mathcal{P}^m_{\Omega}(\Lambda)$ , 2148  $x_{j,query} \sim \mathcal{P}_x^m(\mu_j, \Lambda)$ .  $y_{j,prob} = \operatorname{softmax}(\mu_j^\top \Lambda^{-1} x_{j,query})$  are the corresponding probability distributions of the ground truth label  $y_{j,query}$ . For each j, we generate 100 testing prompts 2149 2150  $P_{jk} = (x_{jk,1}, y_{jk,1}, \dots, x_{jk,M}, y_{jk,M}, x_{j,query})$ , where  $(x_{jk,i}, y_{jk,i}) \stackrel{\text{i.i.d.}}{\sim} \mathcal{P}^m(\mu_j, \Lambda), j \in [20], k \in [100], i \in [M]$ . We denote a model's output for testing prompts  $P_{jk}$  as  $\hat{y}_{jk}$ . We calculate its infer-2151 2152 ence error with  $\frac{1}{20 \times 100} \sum_{j \in [20], k \in [100]} \max_{l \in [c]} \left| (y_{jk})_l - (y_{j, \text{prob}})_l \right|$ , which serves an approximation 2153 of the expected total variation distance we defined in (3). 2154 2155

For the '3-layer' model, we used the x-transformers library and defined it as an encoder-only transformer with 64 embedding sizes, 3 layers, 2 heads and without positional encoding.

For experiments in Figure 1, we set the size of the training dataset to B = 100,000 and set the batch size to 50. We train the '1-layer' using Adam with learning rate 0.0005 for 10 epochs, and train the '3-layer' using Adam with learning rate 0.0001 for 5 epochs. Each experiment is repeated 3 times with different random seeds. For experiments in Figure 2, we also set the size of the train-ing dataset to B = 100,000 and set the batch size to 50. We train the '1-layer' using Adam with learning rate 0.001 for 5 epochs, and train the '3-layer' using Adam with learning rate 0.0001 for 5 epochs. In 'same norm' and 'same covariance' settings, pre-training data are sampled according to Assumption 4.1 with a fixed  $\Lambda$  that  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$ , where  $\lambda_i = |\hat{\lambda}_i|$  and  $\hat{\lambda}_i \stackrel{\text{i.i.d.}}{\sim} N(3, 1)$ . In 'different norms' setting, for each  $\tau \in [B]$ , with probability  $\mathbb{P}(k = j) = 1/10, \mu_{\tau,i} \sim N(k, I_d), j = 1/10$ 0, 1, ..., 9, then each Gaussian component is sampled according to N( $\mu_{\tau,i}, \Lambda$ ). In (different covari-ances) setting, we randomly generate  $v_1, v_2, v_3 \in \mathbb{R}^d$  that half of their elements are 0.1 and the other half elements are 100. Then, we define  $\Lambda_i = \text{diag}(v_i), i = 1, 2, 3$  and generate pre-training data according to Assumption 4.1 with  $\Lambda$ ,  $\Lambda_1$ ,  $\Lambda_2$ ,  $\Lambda_3$ . Each experiment is repeated 3 times with different random seeds. For experiments in Figure 3, the structure of the transformer with full parameters '1-layer, full' is defined as 

 $F(E; W^V, W^{KQ}) = E + W^V E \cdot \frac{E^\top W^{KQ} E}{\rho},$ (59)

where  $W^V, W^{KQ} \in \mathbb{R}^{(d+c) \times (d+c)}$  are the parameters for optimization. For all three transformer models, we set the size of the training dataset to B = 400,000 and set the batch size to 50. We train the '1-layer, sparse' and '1-layer, full' using Adam with learning rate 0.001 for 5 epochs, and train the GPT2 model using Adam with learning rate 0.0001 for 5 epochs. Each experiment is repeated 3 times with different random seeds. For experiments in Figure 4, we train the single-layer transformers with the sparse-form parameters and structures defined in Section 4. We set the size of the training dataset to B = 10,000 and set the batch size to 50. We train the transformers using SGD with learning rate  $\{0.1, 0.5, 1\}$  for 10 epochs, and get the best model on each training dataset. Then, we test these trained models on different testing datasets. Each experiment is repeated 10 times with different random seeds.