# In-Context Learning with Representations: Contextual Generalization of Trained Transformers

Tong Yang[1] Yu Huang[2] Yingbin Liang[3] Yuejie Chi[1]

## Abstract

This paper investigates the training dynamics of transformers by gradient descent through the lens of non-linear regression tasks. The contextual generalization here can be attained via the in-context learning of the template function for each task, where all template functions lie in a linear space with $m$ basis functions. We analyze the training dynamics of multi-head transformers to in-contextly predict unlabeled inputs given partially labeled prompts where the labels contain Gaussian noise and there may be only a few examples in each prompt which are not sufficient to determine the template. We show that the training loss for a shallow multi-head transformer converge linearly to a global minimum. Moreover, the transformer effectively learns to perform ridge regression. To our knowledge, this study is the first of showing that transformers can learn contextual (i.e., template) information to generalize to unseen examples when prompts contain only a small number of query-answer pairs.

## 1 Introduction

Transformers (Vaswani et al., 2017) have achieved tremendous successes in their applications in machine learning, particularly in natural language processing, by introducing self-attention mechanisms that enable models to capture long-range dependencies and contextualized representations. In particular, these self-attention mechanisms endow the transformer with remarkable in-context learning (ICL) capabilities, allowing it to adapt to new tasks or domains by simply being provided prompts with a few examples that demonstrate the desired behavior, without any explicit fine-tuning or updating of the model's parameters (Brown et al., 2020).

A series of papers have empirically studied the underlying mechanisms behind in-context learning in transformer models (Garg et al., 2022; Von Oswald et al., 2023; Wei et al., 2023; Olsson et al., 2022; Xie et al., 2021; Chen & Zou, 2024; Agarwal et al., 2024), which have shown that transformers can predict unseen examples after being prompted on a few examples. The pioneering work of Garg et al. (2022) showed empirically that transformers can be trained from scratch to perform in-context learning of simple function classes, providing an theoretically tractable in-context learning framework. Following this well-established framework, several works have investigated various aspects of in-context learning properties in transformers. Most pertinent to this work, Huang et al. (2023); Chen et al. (2024); Li et al. (2024); Nichani et al. (2024) studied the training dynamics of shallow transformers with softmax attention in order to in-context learn simple tasks such as linear regression (Huang et al., 2023; Chen et al., 2024), binary classification tasks (Li et al., 2024) or causal graph (Nichani et al., 2024). Their theoretical analysis showed that given an arbitrary query token, how transformers learn to *directly* apply the answer corresponding to it from the query-answer pairs that appear in each prompt. Therefore, they all require the sequence length of each prompt to be large enough so that every query-answer pair has been seen in each prompt with sufficiently high probability, whereas practical prompts are often too short to contain many query examples. This suggests that in-context learning can exploit *inherent contextual* information of the prompt to generalize to *unseen* examples, which further raise the following intriguing theoretical question:

> *How do transformers learn contextual information from more general function classes to predict unseen examples given prompts that contain only partial examples?*

In this paper, we answer the above question by analyzing the training dynamics of a one-layer transformer with multi-head softmax attention through the lens of non-linear regression tasks. In our setting, the template function for each

[*] Equal contribution [1] Department of Electrical and Computer Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States. [2] Wharton School, University of Pennsylvania, Philadelphia, PA 19104, United States. [3] Department of Electrical and Computer Engineering, Ohio State University, 281 W Lane Ave, Columbus, OH 43210, United States.. Correspondence to: Yuejie Chi <yuejiec@andrew.cmu.edu>, Tong Yang <tongyang@andrew.cmu.edu>.

*Work presented at TF2M workshop at ICML 2024*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

task lies in the linear space formed by $m$ arbitrary basis functions that capture representation (i.e., features) of data. Our goal is to provide insights on how transformers trained by gradient descent (GD) acquire template information from more general function classes to generalize to unseen examples when each prompt contains only a small number of query-answer pairs. We summarize our contributions are as follows.

**Our Contributions:**

- We first establish the convergence guarantee of a shallow transformer with multi-head softmax attention trained with gradient descent on general non-linear regression in-context learning tasks. We assume each prompt contains only a few (i.e., partial) examples with their Gaussian noisy labels, which are not sufficient to determine the template.
- We then analyze the transformer's behavior at inference time after training, and shows that the transformer decides its generating template by performing ridge regression. We also provide the iteration complexity for the transformer to reach $\varepsilon$-precision with respect to its own choice preference of template given an arbitrary prompt at inference time.
- Our analysis framework allows us to overcome a handful of assumptions made in previous works such as the large prompt lengths (Huang et al., 2023; Chen et al., 2024; Li et al., 2024; Nichani et al., 2024), orthogonality of data (Huang et al., 2023; Chen et al., 2024; Li et al., 2024; Nichani et al., 2024), restrictive initialization conditions (Chen et al., 2024), special structure of the transformer (Nichani et al., 2024), and super wide models (Li et al., 2024). Further, the function classes we consider are a generalization of those considered in most theoretical works (Huang et al., 2023; Chen et al., 2024; Li et al., 2024; Wu et al., 2023; Zhang et al., 2023a).

To our best knowledge, this is the *first* work to analyze how transformers learn contextual (i.e., template) information to generalize to unseen examples when prompts contain only a small number of query-answer pairs.

## 2 Problem Setup

**In-context learning with representation.** We consider ICL of regression with unknown representation, similar to the setup introduced in Guo et al. (2023). To begin, let $f : \mathbb{R}^d \to \mathbb{R}^m$ be a fixed representation map that $f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \cdots, f_m(\boldsymbol{x}))^\top$ for any $\boldsymbol{x} \in \mathbb{R}^d$. The map $f$ can be quite general, which can be regarded as a feature extractor that will be learned by the transformer. We assume that each ICL task corresponds to a map $\boldsymbol{\lambda}^\top f(\cdot)$ that lies in the linear span of those $m$ basis functions in $f(\cdot)$, where $\boldsymbol{\lambda}$ is generated by the distribution $\mathcal{D}_{\boldsymbol{\lambda}}$. Thus, for each ICL instance, the (noisy) label of an input $\boldsymbol{v}_k$ ($\forall k \in [K]$) is given as

$$y_k = \boldsymbol{\lambda}^\top (f(\boldsymbol{v}_k) + \boldsymbol{\epsilon}_k),\ \boldsymbol{\lambda} \sim \mathcal{D}_{\boldsymbol{\lambda}},\ \boldsymbol{\epsilon}_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \tau \boldsymbol{I}_m) \tag{1}$$

where $\tau > 0$ is the noise level.

The goal of ICL is to form predictions on query $\boldsymbol{x}_{\mathsf{query}}$ given in-context labels in Equation (1) on a few inputs, known as *prompts*. In this paper, we use $\mathcal{V}$ to denote the *dictionary* set that contains all $K$ unit-norm *distinct* tokens, i.e., $\mathcal{V} := \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K\} \subset \mathbb{R}^d$ with each token $\|\boldsymbol{v}_k\|_2 = 1$. We assume that each prompt $P = P_{\boldsymbol{\lambda}}$ provides the first $N$ tokens (with $N \ll K$) and their labels, and is embedded in the following matrix

$$\boldsymbol{E}^P := \begin{pmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_N \\ y_1 & y_2 & \cdots & y_N \end{pmatrix} := \begin{pmatrix} \boldsymbol{V} \\ \boldsymbol{y}^\top \end{pmatrix} \in \mathbb{R}^{(d+1)\times N}, \tag{2}$$

where $\boldsymbol{V} := (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N) \in \mathbb{R}^{d\times N}$ is the collection of prompt tokens, and $\boldsymbol{y} := (y_1, \cdots, y_N)^\top$ is the prompt label. Given the prompt as the input, the transformer predicts the labels for all the $K$ tokens $y_1, \cdots, y_K$ in the dictionary set.

**Transformer architecture(see Figure 1).** We adopt a one-layer transformer with multi-head softmax attention (Chen et al., 2024) to predict the labels of all the tokens in the dictionary, where $H$ is the number of heads. Denote the query embedding as

$$\boldsymbol{E}^Q := \begin{pmatrix} \boldsymbol{v}_{N+1} & \boldsymbol{v}_{N+2} & \cdots & \boldsymbol{v}_K \\ 0 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(K-N)}, \tag{3}$$

and denote the embedding of both the prompt and the query as $\boldsymbol{E} := (\boldsymbol{E}^P, \boldsymbol{E}^Q) \in \mathbb{R}^{(d+1)\times K}$. We define the output of each transformer head as

$$\mathsf{head}_h(\boldsymbol{E}) := \boldsymbol{W}_h^{\mathsf{V}} \boldsymbol{E}^P \cdot \mathsf{softmax}\left( (\boldsymbol{E}^P)^\top (\boldsymbol{W}_h^{\mathsf{K}})^\top \boldsymbol{W}_h^{\mathsf{Q}} \boldsymbol{E} \right),$$

where $h \in [H]$, $\boldsymbol{W}_h^{\mathsf{Q}} \in \mathbb{R}^{d_e \times (d+1)}$, $\boldsymbol{W}_h^{\mathsf{K}} \in \mathbb{R}^{d_e \times (d+1)}$, and $\boldsymbol{W}_h^{\mathsf{V}} \in \mathbb{R}^{K\times(d+1)}$ are the query, key, and value matrices, respectively, and the softmax is applied column-wisely, i.e., given a vector input $\boldsymbol{x}$, the $i$-th entry of $\mathsf{softmax}(\boldsymbol{x})$ is given by $e^{x_i} / \sum_j e^{x_j}$. The attention map of the transformer $\mathcal{T}(\boldsymbol{E})$ is defined as

$$\mathcal{T}(\boldsymbol{E}) := \boldsymbol{W}^{\mathsf{O}} \begin{pmatrix} \mathsf{head}_1(\boldsymbol{E}) \\ \vdots \\ \mathsf{head}_H(\boldsymbol{E}) \end{pmatrix} \in \mathbb{R}^{K\times K}, \tag{4}$$

where $\boldsymbol{W}^{\mathsf{O}}$ is the output matrix. Following recent theoretical literature to streamline analysis (Huang et al., 2023; Nichani et al., 2024; Deora et al., 2023; Chen et al., 2024), we assume that the embedding matrices take the following

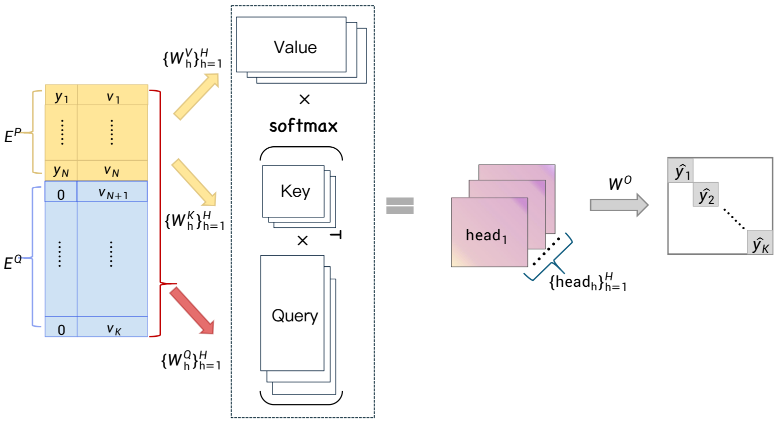

*Figure 1.* The structure of a one-layer transformer with multi-head softmax attention.

forms:

$$\boldsymbol{W}^{\mathsf{O}} := (\boldsymbol{I}_K, \cdots, \boldsymbol{I}_K) \in \mathbb{R}^{K \times HK},$$
$$\boldsymbol{W}_h^{\mathsf{V}} := (\boldsymbol{0}, \boldsymbol{w}_h) \in \mathbb{R}^{K \times (d+1)},$$
$$(\boldsymbol{W}_h^{\mathsf{K}})^\top \boldsymbol{W}_h^{\mathsf{Q}} = \begin{pmatrix} \boldsymbol{Q}_h & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}, \quad \forall h \in [H],$$

where $\boldsymbol{w}_h = (w_{h,1}, \cdots, w_{h,K})^\top \in \mathbb{R}^K$ and $\boldsymbol{Q}_h \in R^{d \times d}$ are trainable parameters for all $h \in [H]$.

The prediction of the labels is provided by the diagonal entries of $\mathcal{T}(\boldsymbol{E})$, which we denote by $\widehat{\boldsymbol{y}} = (\widehat{y}_1, \cdots, \widehat{y}_K) \in \mathbb{R}^K$. Note that $\widehat{y}_k$ takes the following form under our parameter specification:

$$\forall k \in [K]: \quad \widehat{y}_k = \left\langle \boldsymbol{y}, \sum_{h=1}^H w_{h,k} \,\mathsf{softmax}(\boldsymbol{V}^\top \boldsymbol{Q}_h \boldsymbol{v}_k) \right\rangle. \tag{6}$$

**Training via GD.** Let $\boldsymbol{\theta} = \{\boldsymbol{Q}_h, \boldsymbol{w}_h\}_{h=1}^H$ denote all trainable parameters of $\mathcal{T}$. Let $\boldsymbol{\epsilon} := (\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_K) \in \mathbb{R}^{m \times K}$ denote the noise matrix. Given training data over ICL instances, the goal of training is to predict labels $y_k$ for all $\boldsymbol{v}_k \in \mathcal{V}$. Specifically, we train the transformer using gradient descent (GD) by optimizing the following mean-squared population loss:

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\epsilon}}\left[\frac{1}{K}\sum_{k=1}^K (\widehat{y}_k - y_k)^2\right]. \tag{7}$$

We apply different learning rates $\eta_Q, \eta_w > 0$ for updating $\{\boldsymbol{Q}_h\}_{h=1}^H$ and $\{\boldsymbol{w}_h\}_{h=1}^H$, respectively, i.e., at the $t$-th ($t \geq 1$) step, we have

$$\forall h \in [H]: \quad \boldsymbol{Q}_h^{(t)} = \boldsymbol{Q}_h^{(t-1)} - \eta_Q \nabla_{\boldsymbol{Q}_h} \mathcal{L}(\boldsymbol{\theta}^{(t-1)}),$$
$$\boldsymbol{w}_h^{(t)} = \boldsymbol{w}_h^{(t-1)} - \eta_w \nabla_{\boldsymbol{w}_h} \mathcal{L}(\boldsymbol{\theta}^{(t-1)}).$$

**Inference time.** At inference time, given a prompt $P = P_{\boldsymbol{\lambda}}$ with $N$ examples, where $\boldsymbol{\lambda}$ *may not be in the support of the generation distribution* $\mathcal{D}_{\boldsymbol{\lambda}}$, the transformer applies the pretrained parameters and predicts the labels of all $K$ tokens without further parameter updating.

## 3 Theoretical Analysis

### 3.1 Training Time Convergence

In this section, we show that the loss $\mathcal{L}$ converges to its minimum value at a linear rate during training, i.e., the function gap

$$\Delta^{(t)} := \mathcal{L}(\boldsymbol{\theta}^{(t)}) - \inf_{\boldsymbol{\theta}} \mathcal{L} \to 0, \quad t \to \infty. \tag{8}$$

We first state our technical assumptions. The following assumption is on the distribution of generating the coefficient vector $\boldsymbol{\lambda}$ for function maps.

**Assumption 1** (Distribution $\mathcal{D}_{\boldsymbol{\lambda}}$ of generating $\boldsymbol{\lambda}$)**.** *Each entry $\lambda_i$ in (1) is drawn independently and satisfies $\mathbb{E}[\lambda_i] = 0$ and $\mathbb{E}[\lambda_i^2] = 1$ for all $i \in [m]$.*

The following assumption on $\boldsymbol{V}$ is mild to ensure our parameter initialization has good properties. See Appendix B for a discussion on this assumption.

**Assumption 2** (assumption on $\boldsymbol{V}$)**.** *There exists one row vector $\boldsymbol{x} = (x_1, \cdots, x_N)^\top$ of $\boldsymbol{V}$ such that $x_i \neq x_j$, $\forall i \neq j$.*

To guarantee the convergence, we require the initialization of the parameters satisfies the following condition:

**Assumption 3** (parameter initialization)**.** *For all $k \in [K]$, $\boldsymbol{B}_k^{(0)}$ defined in (23) has full row rank.*

**Choice of learning rates.** Define

$$\zeta_0 \coloneqq \min_{k\in[K]} \left\{ \lambda_{\min}\left( \boldsymbol{B}_k^{(0)} \boldsymbol{B}_k^{(0)\top} \right) \right\}, \tag{9}$$

where $\Delta^{(0)}$ is the initial function gap (c.f. (8)), and $\boldsymbol{B}_k^{(0)}$ is defined in (23). Assumption 3 indicates that $\zeta_0 > 0$. Let $\gamma$ be any positive constant that satisfies

$$\gamma \geq \zeta_0^{-5/4} \left( \frac{128\sqrt{2}}{\sqrt{2}-1} \left\| \bar{\boldsymbol{Z}} \right\|_2^2 \sqrt{H} \bar{f}_{\max} K^{3/2} \Delta^{(0)} \right)^{1/2}. \tag{10}$$

We set the learning rates as

$$\eta_Q \leq 1/L \quad \text{and} \quad \eta_w = \gamma^2 \eta_Q, \tag{11}$$

where $L$ is the smoothness constant of the loss specified in (59).

We state our first main result as follows:

**Theorem 1** (Convergence at training time)**.** *Suppose Assumptions 1, 2, 3 hold. We let $\boldsymbol{w}_k^{(0)} = \boldsymbol{0}$ and set the learning rates as in Equation (11). Then for all $t \in \mathbb{N}$, we have*

$$\Delta^{(t)} \leq \left( 1 - \frac{\eta_w \zeta_0}{2K} \right)^t \Delta^{(0)}. \tag{12}$$

Theorem 1 shows that the training loss converges to its minimum value at a linear rate. This gives the *first* convergence result for transformers with multi-head softmax attention trained using GD to perform ICL tasks (see Table 1). Eq. (12) also indicates that the convergence speed decreases as the size $K$ of the dictionary or the number $H$ of attention heads increases, which is intuitive because training with a larger vocabulary size or parameter number is more challenging. However, small $H$ will limit the expressive capacity of the model (see Section B.4 for detailed discussion), and we require $H \geq N$ to guarantee Assumption 3 holds, as stated in Proposition 1 below.

The proof of Theorem 1 is postponed to Appendix E. We provide our proof idea in Appendix B.2.

The following proposition states that when $H \geq N$, we can guarantee that Assumption 3 holds with probability 1 by simply initializing $\{\boldsymbol{Q}_h\}_{h=1}^H$ using Gaussian distribution:

**Proposition 1** (initialization of $\{\boldsymbol{Q}_h\}_{h=1}^H$)**.** *Suppose Assumptions 1, 2 hold and $H \geq N$. For any fixed $\beta > 0$, let $\boldsymbol{Q}_h^{(0)}(i,j) \overset{i.i.d.}{\sim} \mathcal{N}(0, \beta^2)$, then Assumption 3 holds almost surely.*

The proof of Proposition 1 can be found in Appendix G.

### 3.2 Inference Time Performance

We let $\boldsymbol{Z}^Q \in \mathbb{R}^{m\times(K-N)}$ denote

$$\boldsymbol{Z}^Q \coloneqq (f(\boldsymbol{v}_{N+1}), \cdots, f(\boldsymbol{v}_K)) \in \mathbb{R}^{m\times(K-N)}. \tag{13}$$

We assume that the coefficient vectors $\boldsymbol{\lambda}$ at inference time are bounded.

**Assumption 4** (boundedness of $\boldsymbol{\lambda}$ at inference time)**.** *We assume that at inference time $\|\boldsymbol{\lambda}\|_2 \leq B$ for some $B > 0$.*

The following theorem characterizes the performance guarantee of the transformer's output $\widehat{\boldsymbol{y}}$ (after sufficient training) at the inference time, whose proof is deferred to Appendix F.

**Theorem 2** (inference time performance)**.** *Under the assumptions in Theorem 1, for any $\varepsilon > 0$ and $\delta \in (0,1)$, if the number of training iterates $T$ satisfies*

$$T \geq \frac{\log\left( B^2 \Delta^{(0)} \left( \|\boldsymbol{Z}\|_2 + \sqrt{\tau}(\varphi(\delta,N))^{1/2} \right)^2 \Big/ (m\tau\varepsilon) \right)}{\log\left(1/\left(1 - \frac{\eta_w \zeta_0}{2K}\right)\right)}, \tag{14}$$

*where $\varphi(\delta, N) \coloneqq 2\sqrt{N\log(1/\delta)} + 2\log(1/\delta) + N$, then given any prompt $P$ that satisfies Assumption 4 at the inference time, with probability at least $1-\delta$, the output of the transformer $\widehat{\boldsymbol{y}}$ satisfies*

$$\frac{1}{2K} \left\| \widehat{\boldsymbol{y}} - \widehat{\boldsymbol{y}}^\star \right\|_2^2 \leq \varepsilon, \quad \textit{with} \quad \widehat{\boldsymbol{y}}^\star \coloneqq \begin{pmatrix} \boldsymbol{y} \\ \left(\boldsymbol{Z}^Q\right)^\top \widehat{\boldsymbol{\lambda}} \end{pmatrix}, \tag{15}$$

*where $\widehat{\boldsymbol{\lambda}}$ is the solution of the following ridge regression problem:*

$$\widehat{\boldsymbol{\lambda}} \coloneqq \arg\min_{\boldsymbol{\lambda}} \left\{ \tfrac{1}{2N} \textstyle\sum_{i=1}^N (y_i - \boldsymbol{\lambda}^\top f(\boldsymbol{v}_i))^2 + \tfrac{m\tau}{2N} \|\boldsymbol{\lambda}\|_2^2 \right\}. \tag{16}$$

We argue in Appendix B.3 that Theorem 2 actually reflects how the sequence length $N$ affects the transformer's preference for choosing templates and its performance at inference time. To be specific, the closer $m$ is to $N$, the closer the transformer's choice of templates is to the best possible choice, and the better the transformer's prediction will be; when $N < m$, the transformer tends to underfit by choosing a $\boldsymbol{\lambda}$ with small $\ell_2$-norm; when $N > m$, the transformer tends to overfit since it underestimates the noise level and tries to capture noise when making prediction.

## 4 Conclusion

We analyze the training dynamics of a one-layer transformer with multi-head softmax attention trained by gradient descent to solve complex non-linear regression tasks using partially labeled prompts. In this setting, the labels contain Gaussian noise, and each prompt may include only a few examples, which are insufficient to determine the underlying template. Our work overcomes several restrictive assumptions made in previous studies and proves that the training loss converges linearly to its minimum value. Our study provides the first analysis of how transformers can acquire contextual (template) information to generalize to unseen examples when prompts contain a limited number of query-answer pairs.

## Acknowledgements

The work of T. Yang and Y. Chi is supported in part by NSF DMS-2134080. The work of Y. Liang was supported in part by the U.S. National Science Foundation under the grants DMS-2134145 and ECCS-2113860.

## References

Agarwal, R., Singh, A., Zhang, L. M., Bohnet, B., Chan, S., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., et al. Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*, 2024.

Ahuja, K., Panwar, M., and Goyal, N. In-context learning through the bayesian prism. *arXiv preprint arXiv:2306.04891*, 2023.

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Bai, Y., Chen, F., Wang, H., Xiong, C., and Mei, S. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36, 2024.

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

Chen, S., Sheen, H., Wang, T., and Yang, Z. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.

Chen, X. and Zou, D. What can transformer learn with varying depth? case studies on sequence learning tasks. *arXiv preprint arXiv:2404.01601*, 2024.

Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., and Wei, F. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*, 2022.

Deora, P., Ghaderi, R., Taheri, H., and Thrampoulidis, C. On the optimization and generalization of multi-head attention. *arXiv preprint arXiv:2310.12680*, 2023.

Edelman, B. L., Goel, S., Kakade, S., and Zhang, C. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pp. 5793–5831. PMLR, 2022.

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

Guo, T., Hu, W., Mei, S., Wang, H., Xiong, C., Savarese, S., and Bai, Y. How do transformers learn in-context beyond simple functions? a case study on learning with representations. *arXiv preprint arXiv:2310.10616*, 2023.

Hahn, M. and Goyal, N. A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*, 2023.

Han, C., Wang, Z., Zhao, H., and Ji, H. In-context learning of large language models explained as kernel regression. *arXiv preprint arXiv:2305.12766*, 2023.

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

Jeon, H. J., Lee, J. D., Lei, Q., and Van Roy, B. An information-theoretic analysis of in-context learning. *arXiv preprint arXiv:2401.15530*, 2024.

Jiang, H. A latent space theory for emergent abilities in large language models. *arXiv preprint arXiv:2304.09960*, 2023.

Kim, J. and Suzuki, T. Transformers learn nonlinear features in context. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pp. 1302–1338, 2000.

Li, H., Wang, M., Lu, S., Cui, X., and Chen, P.-Y. Training nonlinear transformers for efficient in-context learning: A theoretical learning and generalization analysis. *arXiv preprint arXiv:2402.15607*, 2024.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pp. 19565–19594. PMLR, 2023.

Nguyen, Q. N. and Mondelli, M. Global convergence of deep networks with one wide layer followed by pyramidal topology. *Advances in Neural Information Processing Systems*, 33:11961–11972, 2020.

Nichani, E., Damian, A., and Lee, J. D. How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*, 2024.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

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

Wang, X., Zhu, W., Saxon, M., Steyvers, M., and Wang, W. Y. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*, 2023.

Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., Chen, X., Liu, H., Huang, D., Zhou, D., et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Wies, N., Levine, Y., and Shashua, A. The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Wu, J., Zou, D., Chen, Z., Braverman, V., Gu, Q., and Bartlett, P. L. How many pretraining tasks are needed for in-context learning of linear regression? *arXiv preprint arXiv:2310.08391*, 2023.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Zhang, R., Frei, S., and Bartlett, P. L. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023a.

Zhang, Y., Zhang, F., Yang, Z., and Wang, Z. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023b.

# A Additional Related work

**In-context learning.** Recent research has investigated the theoretical underpinnings of transformers' ICL capabilities from diverse angles. For example, several works focus on explaining the in-context learning of the transformer from the Bayesian perspective (Xie et al., 2021; Ahuja et al., 2023; Han et al., 2023; Jiang, 2023; Wang et al., 2023; Wies et al., 2024; Zhang et al., 2023b; Jeon et al., 2024; Hahn & Goyal, 2023). Li et al. (2023) analyses the generalization and stability in in-context learning of the transformer. Focusing on the expressive capacity of the transformer, Akyürek et al. (2022); Bai et al. (2024) study the expressive power of the transformer on the linear regression task. Akyürek et al. (2022) shows by construction that transformers can represent GD of ridge regression or the closed-form ridge regression solution. Bai et al. (2024) extends Akyürek et al. (2022) and shows that transformers can implement a broad class of standard machine learning algorithms in context. Dai et al. (2022); Von Oswald et al. (2023) show transformers could in-context learn GD.

More pertinent to our work, Guo et al. (2023) considers the ICL problems similar to ours, where the label depends on the input through a basis of possibly complex but fixed template functions, composed with a linear function that differs in each prompt. By construction, the optimal ICL algorithm first transforms the inputs by the representation function, and then performs linear ICL on top of the transformed dataset. Guo et al. (2023) shows the existence of transformers that approximately implement such algorithms, whereas our focus is on analyzing the training dynamics of transformers.

**Training dynamics of transformers.** A line of work initiated by Garg et al. (2022) aims to understand the ICL ability of transformers from an optimization perspective. (Zhang et al., 2023a; Kim & Suzuki, 2024) analyse the training dynamics of transformers with *linear* attention. Huang et al. (2023); Chen et al. (2024); Li et al. (2024) study the optimization dynamics of one-layer softmax attention transformers performing simple in-context learning tasks (linear regression (Huang et al., 2023; Chen et al., 2024) and binary classification (Li et al., 2024)). Table 1 provides a comparison with the existing theoretical works that study the learning dynamics of transformers with softmax attention about the settings, training analysis and generalization of in-context learning.

Among them Huang et al. (2023) was the first to study the training dynamics of softmax attention, where they give the convergence results of a one-layer transformer with a single-head attention on linear regression tasks, assuming context features come from an orthogonal dictionary and each token in the prompts is drawn from a multinomial distribution. In order to the concentration property inherent to multinomial distributions, they require the sequence length $N \gg K^3$. Their analysis indicates that the prompt tokens that are the same as the query will have dominating attention weights, which allows the transformer to copy-paste the correct answer from those prompt tokens.

Li et al. (2024) studies the training of a one-layer single-head transformer in ICL on binary classification tasks. Same as in Huang et al. (2023), they also require the data to be pairwise orthogonal. In addition, they require the width of their model to be larger than the square of the number of data, and they need the number of examples in each prompt $N = \mathcal{O}(\varepsilon^{-2}T)$, where $\varepsilon$ is the precision and $T$ is the number of iterations (Li et al., 2024, Theorem 3.3) to guarantee that the attention weight summation on contexts with the same pattern as the query dominates that on other contexts. This indicates that Li et al. (2024) shares the same copy-paste analysis as in Huang et al. (2023).

Chen et al. (2024) studies the dynamics of gradient flow for training a one-layer multi-head softmax attention model for in-context learning of multi-task linear regression. They consider the linear tasks where they require the coefficient matrix to have certain spectral properties. They also require the sequence length to be sufficiently large (Chen et al., 2024, Assumption 2.1). Moreover, they impose very restrictive initialization conditions (Chen et al., 2024, Definition 3.1). They also use the copy-paste analysis framework as in Huang et al. (2023); Li et al. (2024), but the attention probability vector in their paper is delocalized so that the attention is spread out to capture the information from similar tokens in regression tasks. Nichani et al. (2024) analyzes the optimization dynamics of a simplified two-layer transformer with gradient descent on in-context learning the latent causal graph.

| Reference | nonlinear attention | multi head | task shift | GD convergence | noisy data | representation learning |
|---|---|---|---|---|---|---|
| Wu et al. (2023) | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ |
| Zhang et al. (2023a) | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ |
| Huang et al. (2023) | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| Li et al. (2024) | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ |
| Chen et al. (2024) | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ |
| Ours | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Table 1.* Comparisons with existing theoretical works that study the learning dynamics of transformers in ICL. Note that the last column refers to the fact that the response in our regression task is generated by a linearly weighted unknown representation (feature) model, whereas the other works all consider linear regression tasks without an unknown representation model embedded in their response.

## B Additional Discussion

### B.1 Discussion on Assumption 2

Assumption 2 indicates that $\mathcal{V}$ has distinct tokens, i.e., $\boldsymbol{v}_j \neq \boldsymbol{v}_k$ when $j \neq k$. It is worth noting that Assumption 2 is the only assumption we have on the dictionary $\mathcal{V}$. All other theoretical works in Table 1 impose somewhat unrealistic assumptions on $\mathcal{V}$. For example, Huang et al. (2023); Li et al. (2023); Nichani et al. (2024) assume that the tokens are pairwise orthogonal, which is restrictive since it implies that the dictionary size $K$ should be no larger than the token dimension $d$, whereas in practice it is often the case that $K \gg d$ (Reid et al., 2024; Touvron et al., 2023). Further, Chen et al. (2024); Zhang et al. (2023a); Wu et al. (2023) assume that each token is independently sampled from some Gaussian distribution, which also does not align with practical scenarios where tokens are from a fixed dictionary and there often exist correlations between different tokens.

### B.2 Proof idea of Theorem 1

To prove Theorem 1, we first remove the expectation in the expression of the loss function $\mathcal{L}$ in (7) by reformulating it to a deterministic form (see Lemma 4). With this new form, we show by induction that the loss function $\mathcal{L}$ is smooth (Lemma 10) and satisfies the Polyak-Łojasiewicz condition (PL condition) under Assumption 2 (c.f. (47)). Provided with both smoothness and PL condition, we are able to give our convergence result.

### B.3 Discussion on Theorem 2

In Theorem 2, Equation (15) shows that after training, the transformer learns to output the given labels of the first $N$ tokens in each prompt, and more importantly, predicts the labels of the rest $K - N$ tokens by implementing the ridge regression given in Equation (16). Note that Akyürek et al. (2022) studied the expressive power of the transformer on the linear regression task and showed by construction that transformers can represent the closed-form ridge regression solution. Interestingly, here we show from the optimization perspective that transformers can in fact be trained to do so.

**Two generalization capabilities.** Theorem 2 captures two generalization capabilities that the pretrained transformer can have. (i) *Contextual generalization to unseen examples:* Theorem 2 suggests that the transformer exploits the *inherent contextual* information (to be further discussed in Appendix B.4) of the function template in the given prompt, and can further use such information to predict the unseen tokens. (ii) *Generalization to unseen tasks:* Theorem 2 also suggests that the pretrained transformer can generalize to a function map corresponding to any $\boldsymbol{\lambda} \in \mathbb{R}^m$ at the inference time (albeit satisfying Assumption 4), which is not necessarily sampled from the support of its training distribution $\mathcal{D}_{\boldsymbol{\lambda}}$.

We note that the contextual generalization that the transformer has here is different in nature from the prediction ability shown in the previous work on in-context learning Huang et al. (2023); Chen et al. (2024); Li et al. (2024); Nichani et al. (2024). Those work focuses on a setting where each prompt contains a good portion of tokens similar to the query token, allowing the transformer to *directly* use the label of the corresponding answers from the prompt as the prediction. However, in practical scenarios, prompts often contain only partial information, and our analysis sheds lights on explaining how transformers generalize to unseen examples by leveraging ridge regression to infer the underlying template.

Beyond the above discovery, several questions are yet to be explored. For instance, how good is transformer's choice of the underdetermined template? What's the best choice we could expect? How close is transformer's choice to the best possible choice? We address these questions as follows.

Given any prompt $P$ at inference time, since there is no label information about the rest $K-N$ tokens, the best prediction we could hope to have from the transformer shall be

$$\widehat{\boldsymbol{y}}^{\text{best}} := \begin{pmatrix} \boldsymbol{y} \\ \left(\boldsymbol{Z}^{Q}\right)^{\top} \widehat{\boldsymbol{\lambda}}_{\tau} \end{pmatrix}, \tag{17}$$

where $\boldsymbol{Z}^Q$ is defined in (13), and $\widehat{\boldsymbol{\lambda}}_\tau$ satisfies:

$$\widehat{\boldsymbol{\lambda}}_{\tau} := \arg\min_{\boldsymbol{\lambda}} \mathbb{E}_{\tilde{\boldsymbol{\epsilon}}} \left[\frac{1}{2N} \sum_{i=1}^{N} (y_i - \boldsymbol{\lambda}^{\top} \left(f(\boldsymbol{v}_i) + \boldsymbol{\epsilon}_i\right))^2\right], \tag{18}$$

i.e., we hope the transformer outputs the given $N$ labels as they are. For the rest $K-N$ labels, the best we could hope for is that the transformer estimates the coefficient vector $\boldsymbol{\lambda}$ by solving the above regression problem to obtain $\widehat{\boldsymbol{\lambda}}_\tau$, and predict the $k$-th label by $\widehat{\boldsymbol{\lambda}}_\tau^\top f(\boldsymbol{v}_k)$ for $k = N+1, \cdots, K$. Note that (18) is equivalent to the following ridge regression problem (see Lemma 7 for its derivation):

$$\widehat{\boldsymbol{\lambda}}_{\tau} = \arg\min_{\boldsymbol{\lambda}} \left\{\frac{1}{2N} \sum_{i=1}^{N} (y_i - \boldsymbol{\lambda}^{\top} f(\boldsymbol{v}_i))^2 + \frac{\tau}{2} \|\boldsymbol{\lambda}\|_2^2\right\}. \tag{19}$$

The only difference between the two ridge regression problems (16) and (19) is the coefficient of the regularization term. This indicates that at the training time, the transformer learns to implement ridge regression to predict the labels of the rest $K-N$ tokens, assuming the noise level is given by $\frac{m}{N}\tau$. This observation also reflects how the sequence length $N$ affects the transformer's preference for choosing templates and its performance at inference time:

- the closer $m$ is to $N$, the closer the transformer's choice of templates is to the best possible choice, and the better the transformer's prediction will be;
- when $N < m$, the transformer tends to underfit by choosing a $\boldsymbol{\lambda}$ with small $\ell_2$-norm;
- when $N > m$, the transformer tends to overfit since it underestimates the noise level and tries to capture noise when making prediction.

### B.4 Further Interpretation on Our Main Results

In this section, we provide more interpretation on our results, which may lead to useful insights into the in-context learning ability of the transformer.

**How does the transformer gain in-context ability?** Guo et al. (2023) considers an in-context learning problem very similar to ours, but from the perspective of expressive powers. They showed by construction that a good ICL algorithm should compute the underlying representations $f_i$ ($i \in [m]$) and perform linear regression on the transformed dataset $\{f(\boldsymbol{v}_i), y_i\}_{i=1}^N$ to learn the coefficients $\boldsymbol{\lambda}$ in-context. They further show the existence of transformers that approximately implement such algorithms. Our work is from a different perspective, showing that (pre)training the transformer loss by GD will naturally yield a solution with the aforementioned desirable property characterized in Guo et al. (2023).

Intuitively speaking, our pretrained transformer gains in-context ability by extracting and memorizing some "inherent information" of all basic function maps $f_i$ ($i \in [m]$) during the training. Such information allows it to infer the coefficient vector $\boldsymbol{\lambda}$ from the provided labels in each prompt and calculate inner product $\langle \boldsymbol{\lambda}, f(\boldsymbol{v}_k)\rangle$ to compute $y_k$ given any token $\boldsymbol{v}_k \in \mathcal{V}$ at inference time. To be more specific, the "inherent information" of all basic tasks could be described by the $N$-by-$K$ matrix $\boldsymbol{A}$ (c.f. (31)). During training, the transformer learns to approximate $\boldsymbol{A}_{:,k}$ by $\sum_{h=1}^H w_{h,k} \mathsf{softmax}(\boldsymbol{V}^\top \boldsymbol{Q}_h \boldsymbol{v}_k)$ for each $k \in [K]$.

To further elaborate the underlying idea, we take a look at a special case where the labels do not contain any noise, i.e., $\tau = 0$, and $N \geq m$. In this case, $\boldsymbol{A}$ becomes $\boldsymbol{Z}^\dagger \widehat{\boldsymbol{Z}}$, and given any prompt $P = P_{\boldsymbol{\lambda}}$, the coefficient vector $\boldsymbol{\lambda}$ could be uniquely determined from the provided token-label pairs in the prompt. It is straightforward to verify that the label of each token $\boldsymbol{v}_k$ could be represented by the inner product of the given label vector $\boldsymbol{y}$ and the $k$-th column of $\boldsymbol{Z}^\dagger \widehat{\boldsymbol{Z}}$, i.e.,

$$y_k = \left\langle \boldsymbol{y}, \boldsymbol{Z}^\dagger \widehat{\boldsymbol{Z}}_{:,k} \right\rangle. \tag{20}$$

Comparing the above equation with (6), we can see that in order to gain the in-context ability, the transformer needs to learn an approximation of $\boldsymbol{Z}^{\dagger}\widehat{\boldsymbol{Z}}_{:,k}$ by $\sum_{h=1}^{H} w_{h,k}\mathsf{softmax}(\boldsymbol{V}^{\top}\boldsymbol{Q}_h\boldsymbol{v}_k)$ for each $k \in [K]$.

More generally, in the proof of Theorem 2, we show that

$$\widehat{\boldsymbol{y}}_k^{\star} = \langle \boldsymbol{y}, \boldsymbol{A}_{:,k} \rangle\,, \tag{21}$$

comparing which with (6) suggests that a small training error implies that $\sum_{h=1}^{H} w_{h,k}\mathsf{softmax}(\boldsymbol{V}^{\top}\boldsymbol{Q}_h\boldsymbol{v}_k)$ is close to $\boldsymbol{A}_{:,k}$. In fact, this is the necessary and sufficient condition for the training loss to be small. A rigorous argument is provided in Lemma 4.

**The importance of multi-head mechanism.** Multi-head mechanism is essential in our problem setting. In fact, it is generally impossible to train a shallow transformer with only one attention head to succeed in the in-context learning task considered in our paper. This is because, as we have discussed above, the key for the transformer to gain the in-context learning ability is to approximate $\boldsymbol{A}_{:,k}$ by $\sum_{h=1}^{H} w_{h,k}\mathsf{softmax}(\boldsymbol{V}^{\top}\boldsymbol{Q}_h\boldsymbol{v}_k)$ for each $k \in [K]$. If $H = 1$, the transformer could not approximate each $\boldsymbol{A}_{:,k}$ by $w_{1,k}\mathsf{softmax}(\boldsymbol{V}^{\top}\boldsymbol{Q}_1\boldsymbol{v}_k)$ in general since the entries of the latter vector are either all positive or all negative. Proposition 1 indicates that when $H \geq N$, the weights of the transformer with simple initialization method satisfy our desired property that is crucial to guarantee the fast linear convergence. However, (12) implies that we should not set $H$ to be too large, since larger $H$ yields slower convergence rate.

# C Summary of Notations

**Notation.** Boldface small and capital letters denote vectors and matrices, respectively. Sets are denoted with curly capital letters, e.g., $\mathcal{W}$. We let $(\mathbb{R}^d, \|\cdot\|)$ denote the $d$-dimensional real coordinate space equipped with norm $\|\cdot\|$. $\boldsymbol{I}_d$ is the identity matrix of dimension $d$. The $\ell^p$-norm of $\boldsymbol{v}$ is denoted by $\|\boldsymbol{v}\|_p$, where $1 \leq p \leq \infty$, and the spectral norm and the Frobenius norm of a matrix $\boldsymbol{M}$ are denoted by $\|\boldsymbol{M}\|_2$ and $\|\boldsymbol{M}\|_F$, respectively. $\boldsymbol{M}^{\dagger}$ stands for the Moore-Penrose pseudoinverse of matrix $\boldsymbol{M}$, and $\boldsymbol{M}_{:,i}$ stands for its $i$-th column vector. We let $[N]$ denote $\{1, \ldots, N\}$. We denote by $\boldsymbol{0}$ a vector or a matrix consisting of all 0's. We allow the application of functions such as $\exp(\cdot)$ to vectors or matrices, with the understanding that they are applied in an element-wise manner. We let $\lfloor\cdot\rfloor$ denote the floor function. We use $\boldsymbol{e}_i$ to denote the one-hot vector whose $i$-th entry is 1 and the other entries are all 0. We lists some frequently used notation in our paper in Table 2.

We summarize the frequently used notations in Table 2.

| notation | meaning |
|---|---|
| $K \in \mathbb{N}_+$ | total number of tokens |
| $d \in \mathbb{N}_+$ | token dimension |
| $m \in \mathbb{N}_+$ | number of basic tasks |
| $H \in \mathbb{N}_+$ | number of attention heads |
| $N \in \mathbb{N}_+$ | number of examples in each prompt |
| $\boldsymbol{v}_k \in \mathbb{R}^d$, $k \in [K]$ | the $k$-th token |
| $f_i : \mathbb{R}^d \to \mathbb{R}$, $i \in [m]$ | the $i$-th basic task |
| $\boldsymbol{\lambda} \in \mathbb{R}^m$ | coefficient vector |
| $y_k = \boldsymbol{\lambda}^{\top}(f(\boldsymbol{v}_k) + \boldsymbol{\epsilon}_k), k \in [K]$ | the $k$-th label |

*Table 2.* Notation

We introduce the following notations:

$$\boldsymbol{Z} := (f(\boldsymbol{v}_1) \cdots f(\boldsymbol{v}_N)) \in \mathbb{R}^{m \times N}, \quad \bar{\boldsymbol{Z}} := \left(\boldsymbol{Z}^{\top}\boldsymbol{Z} + m\tau\boldsymbol{I}_N\right)^{1/2} \in \mathbb{R}^{N \times N}, \;\; \bar{f}_{\max} := \max_{i \in [N]} \|\bar{\boldsymbol{z}}_i\|_2\,, \tag{22}$$

where $\bar{\boldsymbol{z}}_i$ is the $i$-th column vector of $\bar{\boldsymbol{Z}}$ for $i \in [N]$. We further define $K$-by-$H$ matrix $\boldsymbol{C}_k^{(t)}$ ($k \in [K]$, $t \in \mathbb{N}_+$) and $N$-by-$H$ matrix $\boldsymbol{B}_k^{(t)}$ as follows:

$$\boldsymbol{C}_k^{(t)} := \mathsf{softmax}(\boldsymbol{V}^{\top}\boldsymbol{Q}_1^{(t)}\boldsymbol{v}_k, \cdots, \boldsymbol{V}^{\top}\boldsymbol{Q}_H^{(t)}\boldsymbol{v}_k) \in \mathbb{R}^{N \times H}, \qquad \boldsymbol{B}_k^{(t)} = \bar{\boldsymbol{Z}}\boldsymbol{C}_k^{(t)} \in \mathbb{R}^{N \times H}. \tag{23}$$

We further define $\boldsymbol{s}_k^h \in \mathbb{R}^N$ as follows:

$$\boldsymbol{s}_k^h := \mathsf{softmax}(\boldsymbol{V}^\top \boldsymbol{Q}_h \boldsymbol{v}_k) = (s_{1k}^h, \cdots, s_{Nk}^h)^\top, \quad \forall k \in [K], h \in [H]. \tag{24}$$

## D Auxiliary Lemmas

In this section, we provide some useful facts that will be repeatedly used later on.

**Lemma 1** (softmax gradient)**.** *For all $j \in [N], k \in [K]$ and $h \in [H]$, we have*

$$\frac{\partial s_{jk}^h}{\partial \boldsymbol{Q}_h} = s_{jk}^h \sum_{i=1}^N s_{ik}^h (\boldsymbol{v}_j - \boldsymbol{v}_i) \boldsymbol{v}_k^\top, \tag{25}$$

*where $s_{jk}^h$ is defined in* (24)*.*

*Proof.* See the proof of Lemma A.1 in (Huang et al., 2023). □

**Lemma 2** (smoothness of softmax)**.** *For vectors $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \mathbb{R}^l$, we have*

$$\|\mathsf{softmax}(\boldsymbol{\xi}_1) - \mathsf{softmax}(\boldsymbol{\xi}_2)\|_1 \le 2 \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|_\infty . \tag{26}$$

*Proof.* See Corollary A.7 in (Edelman et al., 2022). □

We also need to make use of the following form of Young's inequality.

**Lemma 3.** *For any $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_l \in \mathbb{R}^p$, we have*

$$\left\| \sum_{i=1}^l \boldsymbol{x}_i \right\|_2^2 \le l \sum_{i=1}^l \|\boldsymbol{x}_i\|_2^2 . \tag{27}$$

## E Proof of Theorem 1

We define

$$\boldsymbol{\delta}_k^{\boldsymbol{\theta}} := \begin{cases} \sum_{h=1}^H w_{h,k} \boldsymbol{s}_k^h - \left(\boldsymbol{Z}^\top \boldsymbol{Z} + m\tau \boldsymbol{I}\right)^{-1} (\boldsymbol{z}_k + m\tau \boldsymbol{e}_k), & \text{if } k \in [N], \\ \sum_{h=1}^H w_{h,k} \boldsymbol{s}_k^h - \left(\boldsymbol{Z}^\top \boldsymbol{Z} + m\tau \boldsymbol{I}\right)^{-1} \boldsymbol{z}_k, & \text{if } k \in [K] \setminus [N]. \end{cases} \tag{28}$$

We first give a reformat of the loss function to remove the expectation in the population loss.

**Lemma 4** (reformat of the loss function)**.** *Under Assumption 1, the loss function $\mathcal{L}(\boldsymbol{\theta})$ could be rewritten into the following equivalent form:*

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2K} \sum_{k=1}^K \left\| \left(\boldsymbol{Z}^\top \boldsymbol{Z} + m\tau \boldsymbol{I}\right)^{1/2} \boldsymbol{\delta}_k^{\theta} \right\|_2^2 + \mathcal{L}^\star = \frac{1}{2K} \sum_{k=1}^K \|\bar{\boldsymbol{Z}} \boldsymbol{\delta}_k^{\theta}\|_2^2 + \mathcal{L}^\star, \tag{29}$$

*where*

$$\begin{aligned} \mathcal{L}^\star = & \frac{1}{2K} \sum_{k=1}^N \left( - \left(\boldsymbol{Z}^\top \boldsymbol{z}_k + m\tau \boldsymbol{e}_k\right)^\top \left(\boldsymbol{Z}^\top \boldsymbol{Z} + m\tau \boldsymbol{I}\right)^{-1} \left(\boldsymbol{Z}^\top \boldsymbol{z}_k + m\tau \boldsymbol{e}_k\right) + \|\boldsymbol{z}_k\|_2^2 + m\tau \right) \\ & + \frac{1}{2K} \sum_{k=N+1}^K \left( - \left(\boldsymbol{Z}^\top \boldsymbol{z}_k\right)^\top \left(\boldsymbol{Z}^\top \boldsymbol{Z} + m\tau \boldsymbol{I}\right)^{-1} \left(\boldsymbol{Z}^\top \boldsymbol{z}_k\right) + \|\boldsymbol{z}_k\|_2^2 \right) \end{aligned}$$

*is a constant that does not depend on $\boldsymbol{\theta}$, and $\bar{\boldsymbol{Z}}$ is defined in* (22)*.*

*Proof of Lemma 4.* See Appendix H.1. □

Lemma 4 indicates that $\mathcal{L}^\star$ is a lower bound of $\mathcal{L}$. In the proof of Theorem 1 we'll show that $\mathcal{L}^\star$ is actually the infimum of $\mathcal{L}$, i.e., $\mathcal{L}^\star = \inf_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$.

Lemma 4 also indicates that in the noisy case, the necessary and sufficient condition for $\mathcal{L}(\boldsymbol{\theta}^{(t)})$ to converge to $\mathcal{L}^\star$ during training is

$$\forall k \in [K]: \quad \boldsymbol{\delta}_k^{\boldsymbol{\theta}^t} \to \mathbf{0}, \quad t \to \infty. \tag{30}$$

We define the following to matrices:

$$\boldsymbol{A} \coloneqq \left(\boldsymbol{Z}^\top \boldsymbol{Z} + m\tau \boldsymbol{I}_N\right)^{-1} \left(\boldsymbol{Z}^\top \widehat{\boldsymbol{Z}} + (m\tau \boldsymbol{I}_N, \mathbf{0})\right) \in \mathbb{R}^{N\times K}, \tag{31}$$

$$\widehat{\boldsymbol{A}}(\boldsymbol{\theta}) \coloneqq \left(\sum_{h=1}^{H} w_{h,1}\boldsymbol{s}_1^h, \cdots, \sum_{h=1}^{H} w_{h,K}\boldsymbol{s}_K^h\right) \in \mathbb{R}^{N\times K}, \tag{32}$$

where $\widehat{\boldsymbol{Z}} \coloneqq (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_K) \in \mathbb{R}^{m\times K}$. Then by (28) we immediately know that (30) is equivalent to

$$\widehat{\boldsymbol{A}}(\boldsymbol{\theta}^{(t)}) - \boldsymbol{A} \to \mathbf{0}, \quad t \to \infty. \tag{33}$$

To simplify the analysis, we introduce the following reparameterization to unify the learning rates of all parameters, and we'll consider the losses after reparameterization in the subsequent proofs.

**Lemma 5** (reparameterization)**.** *Define*

$$\gamma \coloneqq \sqrt{\eta_w/\eta_Q}, \quad \boldsymbol{\alpha}_h \coloneqq \boldsymbol{w}_h/\gamma, \quad \forall h \in [H], \tag{34}$$

*and let*

$$\boldsymbol{\xi} \coloneqq \{\boldsymbol{Q}_h, \boldsymbol{\alpha}_h\}_{h=1}^H, \quad \ell(\boldsymbol{\xi}) \coloneqq \mathcal{L}(\boldsymbol{\theta}). \tag{35}$$

*Then we have*

$$\boldsymbol{\xi}^{(t)} = \boldsymbol{\xi}^{(t-1)} - \eta_Q \nabla_{\boldsymbol{\xi}} \ell(\boldsymbol{\xi}^{(t-1)}), \quad \forall t \in [T]. \tag{36}$$

*Proof of Lemma 5.* See Appendix H.2. □

We denote $\boldsymbol{\alpha}$ as $\boldsymbol{\alpha} \coloneqq (\alpha_{h,k})_{h\in[H],k\in[K]} \in \mathbb{R}^{H\times K}$.

The following lemma bound the gradient norms by the loss function of the basic, which is crucial to the proof of Theorem 1.

**Lemma 6** (upper bound of the the gradient norms)**.** *Suppose Assumption 1 holds and $|\alpha_{h,k}^{(t)}| \le \alpha$. Then for all $h \in [H]$, we have*

$$\left\| \frac{\partial \ell(\boldsymbol{\xi}^{(t)})}{\partial \boldsymbol{Q}_h^{(t)}} \right\|_F \le 2\sqrt{2}\gamma\alpha \bar{f}_{\max} \sqrt{\ell(\boldsymbol{\xi}^{(t)}) - \mathcal{L}^\star}. \tag{37}$$

*Proof of Lemma 6.* See Appendix H.3. □

The following lemma shows the equivalence between (18) and (19).

**Lemma 7** (equivalence of the regression problems)**.** *Given any prompt $P_{\boldsymbol{\lambda}} \coloneqq (\boldsymbol{v}_1, y_1, \cdots, \boldsymbol{v}_N, y_N)$, we have the following equivalence:*

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[\frac{1}{2N}\sum_{i=1}^{N}(y_i - \boldsymbol{\lambda}^\top (f(\boldsymbol{v}_i) + \boldsymbol{\epsilon}_i))^2\right] = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \boldsymbol{\lambda}^\top f(\boldsymbol{v}_i))^2 + \frac{\tau}{2}\|\boldsymbol{\lambda}\|_2^2. \tag{38}$$

*Proof of Lemma 7.* See Appendix H.4. □

Now we are ready to give the main proof.

*proof of Theorem 1.* To prove Theorem 1, it suffices to prove that under the assumptions made in the theorem, we have:

$$\text{Upper bound of the parameters:} \quad \left\|\boldsymbol{\alpha}_h^{(t)}\right\|_2 \leq \alpha, \tag{39}$$

$$\text{Lower bound of eigenvalues:} \quad \lambda_{\min}\left(\boldsymbol{B}_k^{(t)}\boldsymbol{B}_k^{(t)\top}\right) \geq \frac{\zeta_0}{2}, \tag{40}$$

$$\text{Linear decay of the loss:} \quad \mathcal{L}(\boldsymbol{\theta}^{(t)}) - \mathcal{L}^\star \leq \left(1-\frac{\eta_Q\sigma}{2}\right)^t\left(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^\star\right), \tag{41}$$

where

$$\sigma := \frac{\zeta_0\gamma^2}{K}, \quad \alpha := \sqrt{2K}\frac{4\left\|\bar{\boldsymbol{Z}}\right\|_2}{\gamma\zeta_0}\sqrt{\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^\star}, \tag{42}$$

and $\gamma, \boldsymbol{\alpha}_h$ is defined in (34), $\zeta_0$ is defined in (9).

We prove (39),(40) and (41) by induction. It's apparent that they all hold when $t=0$.

We make the following inductive hypothesis:

**Inductive hypothesis:** when $s \in [t-1]$, (39),(40) and (41) hold.

Below we prove that (39),(40) and (41) hold when $s=t$ by the following steps.

**Step 1: verify the Polyak-Łojasiewicz condition.**

We first compute the gradient of the loss w.r.t. $\boldsymbol{\alpha}$:

$$\begin{aligned}\forall k \in [K]: \quad \frac{\partial \ell(\boldsymbol{\xi})}{\partial \boldsymbol{\alpha}_k} &= \frac{1}{2K}\frac{\partial}{\partial \boldsymbol{\alpha}_k}\left\|\bar{\boldsymbol{Z}}\boldsymbol{\delta}_k^\theta\right\|_2^2 = \frac{1}{2K}\frac{\partial}{\partial \boldsymbol{\alpha}_k}\left\|\bar{\boldsymbol{Z}}\left(\gamma\boldsymbol{C}_k\boldsymbol{\alpha}_k - \boldsymbol{A}_{:k}\right)\right\|_2^2 \\ &= \frac{\gamma}{K}\left(\bar{\boldsymbol{Z}}\boldsymbol{C}_k\right)^\top\bar{\boldsymbol{Z}}\boldsymbol{\delta}_k^\theta = \frac{\gamma}{K}\boldsymbol{B}_k^\top\bar{\boldsymbol{Z}}\boldsymbol{\delta}_k^\theta,\end{aligned} \tag{43}$$

where the first equality follows from Lemma 4, $\boldsymbol{C}_k, \boldsymbol{B}_k$ is defined in (23).

Let $\boldsymbol{b}_k^h$ denote the $h$-th column vector of $\boldsymbol{B}_k$, $h \in [H]$, i.e., $\boldsymbol{B}_k := (\boldsymbol{b}_k^1, \cdots, \boldsymbol{b}_k^H)$,

then for any $k \in [K]$ and $t \in \mathbb{N}_+$, we have

$$\begin{aligned}\left\|(\boldsymbol{b}_k^h)^{(t)} - (\boldsymbol{b}_k^h)^{(0)}\right\|_2 &\leq \left\|\bar{\boldsymbol{Z}}\right\|_2\left\|(\boldsymbol{s}_k^h)^{(t)} - (\boldsymbol{s}_k^h)^{(0)}\right\|_2 \\ &\leq \left\|\bar{\boldsymbol{Z}}\right\|_2\left\|(\boldsymbol{s}_k^h)^{(t)} - (\boldsymbol{s}_k^h)^{(0)}\right\|_1 \\ &\leq 2\left\|\bar{\boldsymbol{Z}}\right\|_2\left\|\boldsymbol{V}^\top(\boldsymbol{Q}_h^{(t)} - \boldsymbol{Q}_h^{(0)})\boldsymbol{v}_k\right\|_\infty \\ &\leq 2\left\|\bar{\boldsymbol{Z}}\right\|_2\max_{j\in[N]}\left|\boldsymbol{v}_j^\top(\boldsymbol{Q}_h^{(t)} - \boldsymbol{Q}_h^{(0)})\boldsymbol{v}_k\right| \\ &\leq 2\left\|\bar{\boldsymbol{Z}}\right\|_2\left\|\boldsymbol{Q}_h^{(t)} - \boldsymbol{Q}_h^{(0)}\right\|_F,\end{aligned} \tag{44}$$

where the third line uses Lemma 2, and that

$$\begin{aligned}\forall h \in [H]: \quad \left\|\boldsymbol{Q}_h^{(t)} - \boldsymbol{Q}_h^{(0)}\right\|_F &\leq \sum_{s=0}^{t-1}\eta\left\|\frac{\partial \ell(\boldsymbol{\xi}^{(s)})}{\partial \boldsymbol{Q}_h^{(s)}}\right\|_F \\ &\leq \sum_{s=0}^{t-1}2\sqrt{2}\eta\gamma\alpha\bar{f}_{\max}\sqrt{\ell(\boldsymbol{\xi}^{(s)}) - \mathcal{L}^\star} \\ &\leq 2\sqrt{2}\eta\gamma\alpha\bar{f}_{\max}\sqrt{\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^\star}\sum_{s=0}^{t-1}\left(\sqrt{1-\frac{\eta\sigma}{2}}\right)^s \\ &\leq \frac{8\sqrt{2}\gamma\alpha\bar{f}_{\max}}{\sigma}\sqrt{\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^\star},\end{aligned} \tag{45}$$

where the second inequality follows from (37) and the third inequality follows from the inductive hypothesis and the fact that $\ell(\boldsymbol{\xi}^{(s)}) = \mathcal{L}(\boldsymbol{\theta}^{(s)}), \forall s$. Combining (45) with (44), we have

$$\begin{aligned}\left\|\boldsymbol{B}_k^{(t)} - \boldsymbol{B}_k^{(0)}\right\|_F &\leq 2\left\|\bar{\boldsymbol{Z}}\right\|_2 \sqrt{\sum_{h=1}^{H}\left\|\boldsymbol{Q}_h^{(t)} - \boldsymbol{Q}_h^{(0)}\right\|_F^2} \\ &\leq \left\|\bar{\boldsymbol{Z}}\right\|_2 \sqrt{H}\frac{16\sqrt{2}\gamma\alpha\bar{f}_{\max}}{\sigma}\sqrt{\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^\star} \\ &\leq \left(1 - 1/\sqrt{2}\right)\sqrt{\zeta_0}, \end{aligned} \tag{46}$$

where the last inequality follows from (10).

(46) indicates that

$$\forall \boldsymbol{x} \in \mathbb{R}^K: \quad \left\|\boldsymbol{x}^\top \boldsymbol{B}_k^{(t)}\right\|_2 \geq \left\|\boldsymbol{x}^\top \boldsymbol{B}_k^{(0)}\right\|_2 - \left\|\boldsymbol{x}^\top(\boldsymbol{B}_k^{(t)} - \boldsymbol{B}_k^{(0)})\right\|_2 \geq \sqrt{\zeta_0/2},$$

which gives (40).

Therefore, we obtain the PL-condition as follows:

$$\begin{aligned}\left\|\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\xi}^{(t)})\right\|_F^2 &\geq \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\frac{\partial \ell(\boldsymbol{\xi})}{\partial \alpha_{h,k}}\right)^2 = \frac{\gamma^2}{K^2}\sum_{k=1}^{K}\left(\bar{\boldsymbol{Z}}\boldsymbol{\delta}_k^{(t)}\right)^\top \boldsymbol{B}_k^{(t)}\boldsymbol{B}_k^{(t)\top}\bar{\boldsymbol{Z}}\boldsymbol{\delta}_k^{(t)} \\ &\geq \frac{\zeta_0\gamma^2}{2K^2}\sum_{k=1}^{K}\left\|\bar{\boldsymbol{Z}}\boldsymbol{\delta}_k^{(t)}\right\|_2^2 = \underbrace{\frac{\zeta_0\gamma^2}{K}}_{:=\sigma}\left(\ell(\boldsymbol{\xi}^{(t)}) - \mathcal{L}^\star\right), \end{aligned} \tag{47}$$

where the equality comes from (43), the last equality follows from (29).

**Step 2: prove the smoothness of the loss function.**

We first give the following lemma that bounds the Lipschitzness of $\boldsymbol{b}_k^h$ and $\boldsymbol{\delta}_k^\theta$, which will be used later on. For notation simplicity, we let $\boldsymbol{B}, \boldsymbol{Q}, \boldsymbol{\alpha}$ denote $\boldsymbol{B}(\boldsymbol{\theta}), \boldsymbol{Q}(\boldsymbol{\theta}), \boldsymbol{\alpha}(\boldsymbol{\theta})$, respectively, and let $\boldsymbol{B}', \boldsymbol{Q}', \boldsymbol{\alpha}'$ denote $\boldsymbol{B}(\boldsymbol{\theta}'), \boldsymbol{Q}(\boldsymbol{\theta}'), \boldsymbol{\alpha}(\boldsymbol{\theta}')$, respectively.

**Lemma 8** (Lipschitzness of $\boldsymbol{b}_k^h$ and $\boldsymbol{\delta}_k^\theta$)**.** *For all $k \in [K]$ and $h \in [H]$, and all transformer parameters $\boldsymbol{\theta}, \boldsymbol{\theta}'$, if $\max\{|\alpha_{h,k}|, |\alpha'_{h,k}|\} \leq \alpha$, then we have*

$$\left\|\boldsymbol{b}_k^h(\boldsymbol{\theta}) - \boldsymbol{b}_k^h(\boldsymbol{\theta}')\right\|_2 \leq 2\left\|\bar{\boldsymbol{Z}}\right\|_2 \left\|\boldsymbol{Q}_h - \boldsymbol{Q}'_h\right\|_F, \tag{48}$$

$$\left\|\boldsymbol{\delta}_k^\theta - \boldsymbol{\delta}_k^{\theta'}\right\|_2 \leq 2\gamma\sqrt{H}\alpha\sqrt{\sum_{h=1}^{H}\left\|\boldsymbol{Q}_h - \boldsymbol{Q}'_h\right\|_F^2} + \gamma\sqrt{H}\left\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}'_k\right\|_2. \tag{49}$$

*Proof.* (48) follows from a similar argument in (44). Regarding the Lipschitzness of $\boldsymbol{\delta}_k^\theta$, we have

$$\begin{aligned}\left\|\boldsymbol{\delta}_k^\theta - \boldsymbol{\delta}_k^{\theta'}\right\|_2 &= \gamma\left\|\sum_{h=1}^{H}\alpha_{h,k}(\boldsymbol{s}_k^h(\boldsymbol{\theta}) - \boldsymbol{s}_k^h(\boldsymbol{\theta}')) + \sum_{h=1}^{H}(\alpha_{h,k} - \alpha'_{h,k})\boldsymbol{s}_k^h(\boldsymbol{\theta}')\right\|_2 \\ &\leq \gamma\sum_{h=1}^{H}|\alpha_{h,k}|\left\|\boldsymbol{s}_k^h(\boldsymbol{\theta}) - \boldsymbol{s}_k^h(\boldsymbol{\theta}')\right\|_2 + \gamma\sum_{h=1}^{H}|\alpha_{h,k} - \alpha'_{h,k}|\left\|\boldsymbol{s}_k^h(\boldsymbol{\theta}')\right\|_2 \\ &\leq 2\gamma\sqrt{H}\alpha\sqrt{\sum_{h=1}^{H}\left\|\boldsymbol{Q}_h - \boldsymbol{Q}'_h\right\|_F^2} + \gamma\sqrt{H}\left\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}'_k\right\|_2, \end{aligned}$$

where we use (44) again to bound the first term in the second line, and use the fact that $\left\|\boldsymbol{s}_k^h(\boldsymbol{\theta}')\right\|_2 \leq 1$ and Cauchy-Schwarz inequality to bound the second term in the second line. □

We also need the following lemma which bounds the norm of $\boldsymbol{B}_k$ and $\boldsymbol{\delta}_k^\theta$.

**Lemma 9** (upper bounds of $\boldsymbol{b}_k^h$ and $\boldsymbol{\delta}_k^\theta$)**.** *For all $k \in [K]$ and $h \in [H]$, if $\max\{|\alpha_{h,k}|, |\alpha'_{h,k}|\} \le \alpha$, then we have*

$$\left\|\boldsymbol{b}_k^h\right\|_2 \le \left\|\bar{\boldsymbol{Z}}\right\|_2, \tag{50}$$

$$\left\|\boldsymbol{\delta}_k^\theta\right\|_2 \le \gamma H \alpha + \|\boldsymbol{A}\|_2, \tag{51}$$

*where $\boldsymbol{A}$ is defined in* (31).

*Proof.* (50) follows from

$$\left\|\boldsymbol{b}_k^h\right\|_2 \le \left\|\bar{\boldsymbol{Z}}\right\|_2 \left\|\boldsymbol{s}_k^h\right\|_2 \le \left\|\bar{\boldsymbol{Z}}\right\|_2.$$

(51) follows from

$$\left\|\boldsymbol{\delta}_k^\theta\right\|_2 \le \gamma \sum_{h=1}^H |\alpha_{h,k}| \left\|\boldsymbol{s}_k^h\right\|_2 + \|\boldsymbol{A}\boldsymbol{e}_k\|_2 \le \gamma H \alpha + \|\boldsymbol{A}\|_2.$$

□

As a consequence of Lemma 8 and Lemma 9, For all $k \in [K]$, and all transformer parameters $\boldsymbol{\theta}, \boldsymbol{\theta}'$, if $\max\{|\alpha_{h,k}|, |\alpha'_{h,k}|\} \le \alpha$, we have

$$\begin{aligned}
&\|\nabla_{\boldsymbol{\alpha}_k}\ell(\boldsymbol{\xi}) - \nabla_{\boldsymbol{\alpha}_k}\ell(\boldsymbol{\xi}')\|_2 \\
&\overset{(43)}{=} \frac{\gamma}{K}\left\|(\boldsymbol{B}_k - \boldsymbol{B}'_k)^\top \bar{\boldsymbol{Z}}\boldsymbol{\delta}_k^\theta + {\boldsymbol{B}'_k}^\top \bar{\boldsymbol{Z}}(\boldsymbol{\delta}_k^\theta - \boldsymbol{\delta}_k^{\theta'})\right\|_2 \\
&\le \frac{\gamma}{K}\left\|\bar{\boldsymbol{Z}}\right\|_2 \|\boldsymbol{B}_k - \boldsymbol{B}'_k\|_F \left\|\boldsymbol{\delta}_k^\theta\right\|_2 + \frac{\gamma}{K}\left\|\bar{\boldsymbol{Z}}\right\|_2 \|\boldsymbol{B}'_k\|_F \left\|\boldsymbol{\delta}_k^\theta - \boldsymbol{\delta}_k^{\theta'}\right\|_2 \\
&\le \frac{\gamma}{K}\cdot 2\left\|\bar{\boldsymbol{Z}}\right\|_2^2 (2\gamma H\alpha + \|\boldsymbol{A}\|_2)\sqrt{\sum_{h=1}^H \|\boldsymbol{Q}_h - \boldsymbol{Q}'_h\|_F^2} + \frac{\gamma^2}{K}H\left\|\bar{\boldsymbol{Z}}\right\|_2^2 \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}'_k\|_2,
\end{aligned} \tag{52}$$

from which we obtain the smoothness of the $\ell$ w.r.t. $\boldsymbol{\alpha}$ as follows:

$$\begin{aligned}
&\|\nabla_{\boldsymbol{\alpha}}\ell(\boldsymbol{\xi}) - \nabla_{\boldsymbol{\alpha}}\ell(\boldsymbol{\xi}')\|_F^2 \\
&= \sum_{k=1}^K \|\nabla_{\boldsymbol{\alpha}_k}\ell(\boldsymbol{\xi}) - \nabla_{\boldsymbol{\alpha}_k}\ell(\boldsymbol{\xi}')\|_2^2 \\
&\le 2K\left(\frac{\gamma}{K}\cdot 2\left\|\bar{\boldsymbol{Z}}\right\|_2^2(2\gamma H\alpha + \|\boldsymbol{A}\|_2)\right)^2 \sum_{h=1}^H \|\boldsymbol{Q}_h - \boldsymbol{Q}'_h\|_F^2 + 2\frac{\gamma^4}{K^2}H^2\left\|\bar{\boldsymbol{Z}}\right\|_2^4 \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_F^2 \\
&\le 2\left(\frac{1}{K}\left(2\gamma\left\|\bar{\boldsymbol{Z}}\right\|_2^2(2\gamma H\alpha + \|\boldsymbol{A}\|_2)\right)^2 + \frac{\gamma^4}{K^2}H^2\left\|\bar{\boldsymbol{Z}}\right\|_2^4\right)\|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2^2,
\end{aligned} \tag{53}$$

where the first inequality uses Young's inequality (c.f. Lemma 3).

To obtain the smoothness of the loss function w.r.t. $\boldsymbol{Q}_h$, we first note that by (80) we have

$$\frac{\partial\ell(\boldsymbol{\xi})}{\partial\boldsymbol{Q}_h} = \frac{\gamma}{K}\sum_{k=1}^K\sum_{j=1}^N \left(\bar{\boldsymbol{Z}}\boldsymbol{\delta}_k^\theta\right)^\top \boldsymbol{z}_j \cdot \alpha_{h,k} s_{jk}^h \sum_{i=1}^N s_{ik}^h(\boldsymbol{v}_j - \boldsymbol{v}_i)\boldsymbol{v}_k^\top. \tag{54}$$

Therefore, if $\max\{|\alpha_{h,k}|, |\alpha'_{h,k}|\} \le \alpha$, we have

$$
\begin{aligned}
\left\| \frac{\partial \ell(\boldsymbol{\xi})}{\partial \boldsymbol{Q}_h} - \frac{\partial \ell(\boldsymbol{\xi}')}{\partial \boldsymbol{Q}_h} \right\|_F \leq \frac{2\gamma \bar{f}_{\max}}{K} \sum_{k=1}^{K} \Bigg\{ & \sum_{j=1}^{N} \|\bar{\boldsymbol{Z}}\|_2 \left\| \boldsymbol{\delta}_k^{\theta} - \boldsymbol{\delta}_k^{\theta'} \right\|_2 \cdot \alpha s_{jk}^h(\boldsymbol{\theta}) \sum_{i=1}^{N} s_{ik}^h(\boldsymbol{\theta}) \\
& + \sum_{j=1}^{N} \|\bar{\boldsymbol{Z}}\|_2 \left\| \boldsymbol{\delta}_k^{\theta'} \right\|_2 |\alpha_{h,k} - \alpha'_{h,k}| s_{jk}^h(\boldsymbol{\theta}) \sum_{i=1}^{N} s_{ik}^h(\boldsymbol{\theta}) \\
& + \sum_{j=1}^{N} \|\bar{\boldsymbol{Z}}\|_2 \left\| \boldsymbol{\delta}_k^{\theta'} \right\|_2 \alpha |s_{jk}^h(\boldsymbol{\theta}) - s_{jk}^h(\boldsymbol{\theta}')| \sum_{i=1}^{N} s_{ik}^h(\boldsymbol{\theta}) \\
& + \sum_{j=1}^{N} \|\bar{\boldsymbol{Z}}\|_2 \left\| \boldsymbol{\delta}_k^{\theta'} \right\|_2 \alpha s_{jk}^h(\boldsymbol{\theta}') \sum_{i=1}^{N} |s_{ik}^h(\boldsymbol{\theta}) - s_{ik}^h(\boldsymbol{\theta}')| \Bigg\} \\
\leq \frac{2\gamma \bar{f}_{\max} \|\bar{\boldsymbol{Z}}\|_2}{K} \sum_{k=1}^{K} \Bigg\{ & \left\| \boldsymbol{\delta}_k^{\theta} - \boldsymbol{\delta}_k^{\theta'} \right\|_2 \alpha + \left\| \boldsymbol{\delta}_k^{\theta'} \right\|_2 |\alpha_{h,k} - \alpha'_{h,k}| \\
& + \left\| \boldsymbol{\delta}_k^{\theta'} \right\|_2 \alpha \sum_{j=1}^{N} |s_{jk}^h(\boldsymbol{\theta}) - s_{jk}^h(\boldsymbol{\theta}')| + \left\| \boldsymbol{\delta}_k^{\theta'} \right\|_2 \alpha \sum_{i=1}^{N} |s_{ik}^h(\boldsymbol{\theta}) - s_{ik}^h(\boldsymbol{\theta}')| \Bigg\} \\
\leq \frac{2\gamma \bar{f}_{\max} \|\bar{\boldsymbol{Z}}\|_2}{K} \sum_{k=1}^{K} \Bigg\{ & \left\| \boldsymbol{\delta}_k^{\theta} - \boldsymbol{\delta}_k^{\theta'} \right\|_2 \alpha + \left\| \boldsymbol{\delta}_k^{\theta'} \right\|_2 |\alpha_{h,k} - \alpha'_{h,k}| \\
& + 2 \left\| \boldsymbol{\delta}_k^{\theta'} \right\|_2 \alpha \sqrt{N} \left\| \boldsymbol{s}_k^h(\boldsymbol{\theta}) - \boldsymbol{s}_k^h(\boldsymbol{\theta}') \right\|_2 \Bigg\},
\end{aligned}
\tag{55}
$$

where the third inequality uses Cauchy-Schwarz inequality. Combining the above inequality (55) with Lemma 8 and Lemma 9, we have

$$
\begin{aligned}
& \left\| \frac{\partial \ell(\boldsymbol{\xi})}{\partial \boldsymbol{Q}_h} - \frac{\partial \ell(\boldsymbol{\xi}')}{\partial \boldsymbol{Q}_h} \right\|_F \\
& \leq \frac{2\gamma \bar{f}_{\max} \|\bar{\boldsymbol{Z}}\|_2}{K} \Bigg\{ \alpha\gamma\sqrt{H} \left( 2K\alpha \sqrt{\sum_{h=1}^{H} \|\boldsymbol{Q}_h - \boldsymbol{Q}'_h\|_F^2} + \sqrt{K} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_F \right) \\
& \quad + (\gamma H \alpha + \|\boldsymbol{A}\|_2) \sqrt{K} \left\| \boldsymbol{\alpha}_{h,:} - \boldsymbol{\alpha}'_{h,:} \right\|_2 \\
& \quad + (\gamma H \alpha + \|\boldsymbol{A}\|_2) \cdot 2\alpha\sqrt{N} \cdot 2K \|\boldsymbol{Q}'_h - \boldsymbol{Q}_h\|_F \Bigg\},
\end{aligned}
\tag{56}
$$

where the last line uses (44) to bound $\left\| \boldsymbol{s}_k^h(\boldsymbol{\theta}) - \boldsymbol{s}_k^h(\boldsymbol{\theta}') \right\|_2$. The above inequality (56) further gives

$$
\begin{aligned}
& \sum_{h=1}^{H} \|\nabla_{\boldsymbol{Q}_h} \ell(\boldsymbol{\xi}) - \nabla_{\boldsymbol{Q}_h} \ell(\boldsymbol{\xi}')\|_F^2 \\
& \leq 8 \cdot \frac{\gamma \bar{f}_{\max} \|\bar{\boldsymbol{Z}}\|_2}{K} \Bigg\{ (2K\alpha)^2 \left[ (\alpha\gamma H)^2 + 4N (\alpha\gamma H + \|\boldsymbol{A}\|_2)^2 \right] \sum_{h=1}^{H} \|\boldsymbol{Q}_h - \boldsymbol{Q}'_h\|_F^2 \\
& \quad + K \left[ (\alpha\gamma H)^2 + (\alpha\gamma H + \|\boldsymbol{A}\|_2)^2 \right] \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_F^2 \Bigg\} \\
& \leq 8\gamma \bar{f}_{\max} \|\bar{\boldsymbol{Z}}\|_2 \cdot \max\left\{ 1, (2\sqrt{K}\alpha)^2 \right\} \left[ (\alpha\gamma H)^2 + 4N (\alpha\gamma H + \|\boldsymbol{A}\|_2)^2 \right] \|\boldsymbol{\xi}' - \boldsymbol{\xi}\|_2^2 ,
\end{aligned}
\tag{57}
$$

where the first inequality makes use of Young's inequality (c.f. Lemma 3).

Combining the above two relations (53) and (57), we obtain the smoothness of $\ell$ w.r.t. $\boldsymbol{\xi}$ as follows:

**Lemma 10** (smoothness of the loss, noisy case)**.** *Let* $\gamma := \sqrt{\eta_w/\eta_Q}$*. For all transformer parameters* $\boldsymbol{\xi}, \boldsymbol{\xi}'$*, if* $\max\{|\alpha_{h,k}|, |\alpha'_{h,k}|\} \leq \alpha$*, then we have*

$$\left\|\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\xi}) - \nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\xi}')\right\|_2 \leq L \left\|\boldsymbol{\xi} - \boldsymbol{\xi}'\right\|_2, \tag{58}$$

*where*

$$\begin{aligned} L^2 = 2 &\left(\frac{1}{K}\left(2\gamma \left\|\bar{\boldsymbol{Z}}\right\|_2^2 (2\gamma H\alpha + \|\boldsymbol{A}\|_2)\right)^2 + \frac{\gamma^4}{K^2} H^2 \left\|\bar{\boldsymbol{Z}}\right\|_2^4\right) \\ &+ 8\gamma \bar{f}_{\max} \left\|\bar{\boldsymbol{Z}}\right\|_2 \cdot \max\left\{1, (2\sqrt{K}\alpha)^2\right\} \left[(\alpha\gamma H)^2 + 4N\left(\alpha\gamma H + \|\boldsymbol{A}\|_2\right)^2\right]. \end{aligned} \tag{59}$$

**Step 3: verify (39).** (43) implies

$$\frac{\partial \ell(\boldsymbol{\xi})}{\partial \alpha_{h,k}} = \frac{\gamma}{K}(\boldsymbol{b}_k^h)^\top \bar{\boldsymbol{Z}} \boldsymbol{\delta}_k^\theta,$$

which, combining with (50), gives

$$\forall k \in [K], h \in [H]: \quad \left(\frac{\partial \ell(\boldsymbol{\xi})}{\partial \alpha_{h,k}}\right)^2 \leq \frac{\gamma^2}{K^2} \left\|\bar{\boldsymbol{Z}}\right\|_2^2 \left\|\bar{\boldsymbol{Z}}\boldsymbol{\delta}_k^\theta\right\|_2^2.$$

Combining this with (29) we obtain

$$\left\|\frac{\ell(\boldsymbol{\xi})}{\partial \boldsymbol{\alpha}_h}\right\|_2^2 \leq \left\|\bar{\boldsymbol{Z}}\right\|_2^2 \frac{2\gamma^2}{K}\left(\ell(\boldsymbol{\xi}) - \mathcal{L}^\star\right),$$

which indicates

$$\left\|\frac{\partial \ell(\boldsymbol{\xi})}{\partial \boldsymbol{\alpha}_h}\right\|_2 \leq \left\|\bar{\boldsymbol{Z}}\right\|_2 \gamma \sqrt{\frac{2}{K}\left(\ell(\boldsymbol{\xi}) - \mathcal{L}^\star\right)}. \tag{60}$$

Therefore, we have

$$\begin{aligned} \left\|\boldsymbol{\alpha}_h^{(t)}\right\|_2 &= \left\|\boldsymbol{\alpha}_h^{(0)} - \eta_Q \sum_{i=0}^{t-1} \frac{\partial \ell(\boldsymbol{\xi}^{(i)})}{\partial \boldsymbol{\alpha}_h}\right\|_2 \\ &\leq \left\|\boldsymbol{\alpha}_h^{(0)}\right\|_2 + \eta_Q \sum_{i=0}^{t-1} \left\|\frac{\partial \ell(\boldsymbol{\xi}^{(i)})}{\partial \boldsymbol{\alpha}_h}\right\|_2 \\ &\leq \left\|\boldsymbol{\alpha}_h^{(0)}\right\|_2 + \eta_Q \left\|\bar{\boldsymbol{Z}}\right\|_2 \sqrt{\frac{2\gamma^2}{K}} \sum_{i=0}^{t-1} \sqrt{\ell(\boldsymbol{\xi}^{(i)}) - \mathcal{L}^\star} \\ &\leq \left\|\boldsymbol{\alpha}_h^{(0)}\right\|_2 + \eta_Q \left\|\bar{\boldsymbol{Z}}\right\|_2 \sqrt{\frac{2\gamma^2\left(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^\star\right)}{K}} \sum_{i=0}^{t-1} \left(\sqrt{1 - \frac{\eta_Q \sigma}{2}}\right)^i \\ &\leq \left\|\boldsymbol{\alpha}_h^{(0)}\right\|_2 + \eta_Q \left\|\bar{\boldsymbol{Z}}\right\|_2 \sqrt{\frac{2\gamma^2\left(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^\star\right)}{K}} \cdot \frac{4}{\eta_Q \sigma}, \end{aligned}$$

where the second inequality follows from (60) and the third inequality follows from the induction hypothesis (41). (39) follows from plugging $\sigma$ defined in (42) into the above inequality and using the initializtion condition that $\boldsymbol{\alpha}^{(0)} = \frac{1}{\gamma}\boldsymbol{w}^{(0)} = \boldsymbol{0}$.

**Step 4: give the linear convergence rate.** Combining (39), (58) and Lemma 4.3 in (Nguyen & Mondelli, 2020), we have

$$\ell(\boldsymbol{\xi}^{(t)}) - \mathcal{L}^\star \leq \ell(\boldsymbol{\xi}^{(t-1)}) - \mathcal{L}^\star + \eta_Q \langle \nabla_{\boldsymbol{\xi}} \ell(\boldsymbol{\xi}^{(t-1)}), \boldsymbol{\xi}^{(t)} - \boldsymbol{\xi}^{(t-1)} \rangle + \frac{L}{2}\left\|\boldsymbol{\xi}^{(t)} - \boldsymbol{\xi}^{(t-1)}\right\|_2^2, \tag{61}$$

which indicates when $\eta_Q \leq 1/L$, we have

$$\ell(\boldsymbol{\xi}^{(t)}) - \mathcal{L}^\star \leq \ell(\boldsymbol{\xi}^{(t-1)}) - \mathcal{L}^\star - \frac{\eta_Q}{2}\left\|\nabla_{\boldsymbol{\xi}} \ell(\boldsymbol{\xi}^{(t-1)})\right\|_F^2 \overset{(47)}{\leq} \left(1 - \frac{\eta_Q \sigma}{2}\right)\left(\ell(\boldsymbol{\xi}^{(t-1)}) - \mathcal{L}^\star\right), \tag{62}$$

which, combined with the fact that $\mathcal{L}(\boldsymbol{\theta}^{(s)}) = \ell(\boldsymbol{\xi}^{(s)})$ for all $s$ (see Lemma 5), verifies (41).

Note that (29) implies that $\quad \mathcal{L}^\star \leq \mathcal{L}(\boldsymbol{\theta})$ holds for all $\boldsymbol{\theta}$. And from (41) we know that $\mathcal{L}(\boldsymbol{\theta}^{(t)}) \to \mathcal{L}^\star$ as $t \to \infty$. Therefore, there must be

$$\mathcal{L}^\star = \inf_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}).$$

Consequently, (41) is equivalent to (12). □

# F Proof of Theorem 2

By (41) we know that $\mathcal{L}(\boldsymbol{\theta}^{(t)}) \to \mathcal{L}^\star$ as $t \to \infty$. Thus from (29) we know that (30) and (33) hold.

By Sherman-Morrison-Woodbury formula, we have

$$\left(m\tau \boldsymbol{I}_N + \boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1} = \frac{1}{m\tau} \boldsymbol{I}_N - \frac{1}{m\tau} \boldsymbol{Z}^\top \left(m\tau \boldsymbol{I}_m + \boldsymbol{Z}\boldsymbol{Z}^\top\right)^{-1} \boldsymbol{Z}. \tag{63}$$

Thus we have

$$\begin{aligned}
\boldsymbol{A} &\overset{(31)}{=} \left(\boldsymbol{Z}^\top \boldsymbol{Z} + m\tau \boldsymbol{I}_N\right)^{-1} \left(\boldsymbol{Z}^\top \widehat{\boldsymbol{Z}} + (m\tau \boldsymbol{I}_N, \boldsymbol{0})\right) \\
&\overset{(63)}{=} \frac{1}{m\tau} \left(\boldsymbol{I}_N - \boldsymbol{Z}^\top \left(m\tau \boldsymbol{I}_m + \boldsymbol{Z}\boldsymbol{Z}^\top\right)^{-1} \boldsymbol{Z}\right) \left(\boldsymbol{Z}^\top \widehat{\boldsymbol{Z}} + (m\tau \boldsymbol{I}_N, \boldsymbol{0})\right) \\
&= \frac{1}{m\tau} \Big[\boldsymbol{Z}^\top \widetilde{\boldsymbol{Z}} + (m\tau \boldsymbol{I}_N, \boldsymbol{0}) - \boldsymbol{Z}^\top \left(m\tau \boldsymbol{I}_m + \boldsymbol{Z}\boldsymbol{Z}^\top\right)^{-1} \left(m\tau \boldsymbol{I}_m + \boldsymbol{Z}\boldsymbol{Z}^\top\right) \widetilde{\boldsymbol{Z}} \\
&\quad + m\tau \boldsymbol{Z}^\top \left(m\tau \boldsymbol{I}_m + \boldsymbol{Z}\boldsymbol{Z}^\top\right)^{-1} \widetilde{\boldsymbol{Z}} - m\tau \boldsymbol{Z}^\top \left(m\tau \boldsymbol{I}_m + \boldsymbol{Z}\boldsymbol{Z}^\top\right)^{-1} (\boldsymbol{Z}, \boldsymbol{0})\Big] \\
&= (\boldsymbol{I}_N, \boldsymbol{0}) + \boldsymbol{Z}^\top \left(m\tau \boldsymbol{I}_m + \boldsymbol{Z}\boldsymbol{Z}^\top\right)^{-1} (\boldsymbol{0}, \boldsymbol{Z}^Q) \\
&= \left(\boldsymbol{I}_N, \boldsymbol{Z}^\top \left(m\tau \boldsymbol{I}_m + \boldsymbol{Z}\boldsymbol{Z}^\top\right)^{-1} \boldsymbol{Z}^Q\right),
\end{aligned} \tag{64}$$

where $\boldsymbol{Z}^Q$ is defined in (13).

On the other hand, it's straightforward to verify that $\widehat{\boldsymbol{\lambda}}$ defined in (16) yields the following closed form:

$$\widehat{\boldsymbol{\lambda}} = \left(m\tau \boldsymbol{I}_m + \boldsymbol{Z}\boldsymbol{Z}^\top\right)^{-1} \boldsymbol{Z}\boldsymbol{y}. \tag{65}$$

Combining the above two equations, we obtain

$$\boldsymbol{A}^\top \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y} \\ \left(\boldsymbol{Z}^Q\right)^\top \left(m\tau \boldsymbol{I}_m + \boldsymbol{Z}\boldsymbol{Z}^\top\right)^{-1} \boldsymbol{Z}\boldsymbol{y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{y} \\ \left(\boldsymbol{Z}^Q\right)^\top \widehat{\boldsymbol{\lambda}} \end{pmatrix} = \widehat{\boldsymbol{y}}^\star,$$

where the last equality follows from (15).

Now we give the iteration complexity for the mean-squared error between the prediction $\widehat{\boldsymbol{y}}$ and the limit point $\widehat{\boldsymbol{y}}^\star$ to be less than $\varepsilon$.

Given any prompt $P = P_{\boldsymbol{\lambda}}$, where $\boldsymbol{\lambda}$ satisfies Assumption 4, we have

$$y_i = \boldsymbol{\lambda}^\top (\boldsymbol{z}_i + \boldsymbol{\epsilon}_i) \sim \mathcal{N}(\boldsymbol{\lambda}^\top \boldsymbol{z}_i, \|\boldsymbol{\lambda}\|_2^2 \tau).$$

Letting $x_i = \frac{y_i - \boldsymbol{\lambda}^\top \boldsymbol{z}_i}{\|\boldsymbol{\lambda}\|_2 \sqrt{\tau}}$, we have $x_i \sim \mathcal{N}(0, 1)$. Define

$$Z = \sum_{i=1}^{N} \|\boldsymbol{\lambda}\|_2^2 \tau (x_i^2 - 1) = \left\|\boldsymbol{y} - \boldsymbol{Z}^\top \boldsymbol{\lambda}\right\|_2^2 - N\tau \|\boldsymbol{\lambda}\|_2^2.$$

By Laurent & Massart (2000, Lemma 1), we have

$$\forall s > 0: \quad \mathbb{P}\left(Z \geq 2\sqrt{N} \|\boldsymbol{\lambda}\|_2^2 \tau \sqrt{s} + 2 \|\boldsymbol{\lambda}\|_2^2 \tau s\right) \leq \exp(-s).$$

By letting $s = \log(1/\delta)$ and using the definition of $Z$, we have

$$\mathbb{P}\left(\left\|\boldsymbol{y} - \boldsymbol{Z}^\top \boldsymbol{\lambda}\right\|_2^2 \geq N\tau \left\|\boldsymbol{\lambda}\right\|_2^2 + 2\sqrt{N\log(1/\delta)}\left\|\boldsymbol{\lambda}\right\|_2^2 \tau + 2\left\|\boldsymbol{\lambda}\right\|_2^2 \tau \log(1/\delta)\right) \leq \delta. \tag{66}$$

Thus with probability at least $1-\delta$, we have

$$\begin{aligned}
\|\boldsymbol{y}\|_2 &\leq \left\|\boldsymbol{Z}^\top \boldsymbol{\lambda}\right\|_2 + \left\|\boldsymbol{y} - \boldsymbol{Z}^\top \boldsymbol{\lambda}\right\|_2 \\
&\leq \left\|\boldsymbol{Z}^\top \boldsymbol{\lambda}\right\|_2 + \|\boldsymbol{\lambda}\|_2 \sqrt{\tau}\left(N + 2\sqrt{N\log(1/\delta)} + 2\log(1/\delta)\right)^{1/2} \\
&\leq B\left(\|\boldsymbol{Z}\|_2 + \sqrt{\tau}\left(N + 2\sqrt{N\log(1/\delta)} + 2\log(1/\delta)\right)^{1/2}\right).
\end{aligned} \tag{67}$$

where we use (66) in the second relation, and the third relation follows from Assumption 4.

On the other hand, by (29) we have

$$\mathcal{L}(\boldsymbol{\theta}^{(t)}) = \frac{1}{2K}\left\|\bar{\boldsymbol{Z}}(\widehat{\boldsymbol{A}} - \boldsymbol{A})\right\|_2^2 + \mathcal{L}^\star \geq \frac{m\tau}{2K}\left\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\right\|_2^2 + \mathcal{L}^\star,$$

which gives

$$\left\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\right\|_2 \leq \sqrt{\frac{2K}{m\tau}\left(\mathcal{L}(\boldsymbol{\theta}^{(T)}) - \mathcal{L}^\star\right)} \leq \sqrt{\frac{2K}{m\tau}\left(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^\star\right)}\left(1 - \frac{\gamma^2 \eta_Q \zeta_0}{2K}\right)^{T/2}. \tag{68}$$

Thus we know that w.p. at least $1-\delta$, we have

$$\frac{1}{2K}\left\|\widehat{\boldsymbol{y}} - \widehat{\boldsymbol{y}}^\star\right\|_2^2 = \frac{1}{2K}\left\|\left(\widehat{\boldsymbol{A}} - \boldsymbol{A}\right)^\top \boldsymbol{y}\right\|_2^2 \leq \frac{1}{2K}\left\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\right\|_2^2 \|\boldsymbol{y}\|_2^2 \leq \varepsilon,$$

where the last relation follows from (67), (68) and (14).

## G Proof of Proposition 1

For notation simplicity we drop the superscript $(0)$ in the subsequent proof.

Let $\boldsymbol{D}_k := \left(\boldsymbol{V}^\top \boldsymbol{Q}_1 \boldsymbol{v}_k, \cdots, \boldsymbol{V}^\top \boldsymbol{Q}_H \boldsymbol{v}_k\right) \in \mathbb{R}^{N\times H}$. Note that

$$\boldsymbol{D}_k = \boldsymbol{V}^\top \boldsymbol{Q} = \boldsymbol{V}^\top(\boldsymbol{q}_1, \cdots, \boldsymbol{q}_H), \quad \text{where } \boldsymbol{Q}(i,j) \overset{i.i.d.}{\sim} \mathcal{N}(0, \beta^2 \|\boldsymbol{v}_k\|_2^2), \quad \forall i \in [d], j \in [H]. \tag{69}$$

This suggests the column vectors of $\boldsymbol{D}_k$ are i.i.d. and the density of each column vector is positive at any point $\boldsymbol{x} \in \mathcal{R}(\boldsymbol{V})$, where $\mathcal{R}(\boldsymbol{V}) \subset \mathbb{R}^N$ is the row space of $\boldsymbol{V}$.

Since $\bar{\boldsymbol{Z}}$ has full rank, to prove $\boldsymbol{B}_k$ has full rank a.s., we only need to argue that $\boldsymbol{C}_k(:, 1:N)$ has full rank w.p. 1. Below we prove this by contradiction (recall that by definition $\boldsymbol{C}_k = \mathsf{softmax}(\boldsymbol{D}_k)$, and we assume $H \geq N$).

Suppose w.p. larger than $0$, there exists one of $\boldsymbol{C}_k(:, 1:N)$'s colomn vector that could be linearly represented by its other $N-1$ colomn vectors. Without loss of generality, we assume this colomn vector is $\boldsymbol{C}_k(:, 1) = \mathsf{softmax}(\boldsymbol{D}_k(:, 1))$. Let $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{q}_1) := \exp(\boldsymbol{D}_k(:, 1)) = \exp(\boldsymbol{V}^\top \boldsymbol{q}_1)$. Then $\boldsymbol{x}$ could be linearly represented by $\exp(\boldsymbol{D}_k(:, i))$, $i = 2, \cdots, N$.

Let $\tilde{\boldsymbol{A}} := \exp(\boldsymbol{D}_k(:, 2:N))$, then w.p. larger than 0, $\boldsymbol{x} \in \mathcal{C}(\tilde{\boldsymbol{A}})$, where $\mathcal{C}(\tilde{\boldsymbol{A}})$ is the column vector space of $\tilde{\boldsymbol{A}}$. i.e., we have

$$\int_{\mathbb{R}^{N\times(m-1)}} \mathbb{P}(\boldsymbol{x} \in \mathcal{C}(\tilde{\boldsymbol{A}}) | \tilde{\boldsymbol{A}}) d\mu(\tilde{\boldsymbol{A}}) > 0,$$

which further indicates that there exists $\tilde{\boldsymbol{A}} \in \mathbb{R}^{N\times(N-1)}$ such that $\mathbb{P}(\boldsymbol{x} \in \mathcal{C}(\tilde{\boldsymbol{A}})) > 0$. Since the dimension of $\mathcal{C}(\tilde{\boldsymbol{A}})$ is at most $N-1$, there exists $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{y} \neq \boldsymbol{0}$ such that $\boldsymbol{y} \perp \mathcal{C}(\tilde{\boldsymbol{A}})$. Therefore, we have

$$\mathbb{P}(\boldsymbol{y}^\top \boldsymbol{x} = 0) > 0. \tag{70}$$

By Assumption 2, without loss of generality, we assume that $\boldsymbol{u}_1 = (v_{11}, v_{12}, \cdots, v_{1N})^\top$ has different entries.

For any vector $\boldsymbol{w} = (w_1, \cdots, w_d)^\top \in \mathbb{R}^d$, we let $\tilde{\boldsymbol{w}} = (w_2, \cdots, w_d)^\top \in \mathbb{R}^{d-1}$ denote the vector formed by deleting the first entry of $\boldsymbol{w}$. Let $\boldsymbol{q}_1 = (q, \tilde{\boldsymbol{q}}_1^\top)^\top$. For any fixed $\tilde{\boldsymbol{q}}_1 \in \mathbb{R}^{d-1}$, the function $g(\cdot|\tilde{\boldsymbol{q}}_1) : \mathbb{R} \to \mathbb{R}$ defined by

$$g(q|\tilde{\boldsymbol{q}}_1) := \sum_{i=1}^N y_i e^{qv_{1i} + \tilde{\boldsymbol{q}}_1^\top \tilde{\boldsymbol{v}}_i} = \sum_{i=1}^N y_i e^{\tilde{\boldsymbol{q}}_1^\top \tilde{\boldsymbol{v}}_i} e^{qv_{1i}} = \left\langle \boldsymbol{y}, \exp(\boldsymbol{V}^\top \boldsymbol{q}_1) \right\rangle = \langle \boldsymbol{y}, \boldsymbol{x}(\boldsymbol{q}_1) \rangle$$

has finite zero points and thus $\{q \in \mathbb{R} | g(q|\tilde{\boldsymbol{q}}_1) = 0\}$ is a zero-measure set. Therefore, we have

$$\mathbb{P}(\langle \boldsymbol{y}, \boldsymbol{x} \rangle = 0) = \int_{\mathbb{R}^{d-1}} \mathbb{P}(g(q|\tilde{\boldsymbol{q}}_1) = 0 | \tilde{\boldsymbol{q}}_1) d\mu(\tilde{\boldsymbol{q}}_1) = 0,$$

which contradicts (70).

Therefore, $\boldsymbol{C}_k(:, 1:N)$ has full rank with probability 1.

# H Proof of Key Lemmas

## H.1 Proof of Lemma 4

We let $\boldsymbol{\epsilon}^P := (\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_N) \in \mathbb{R}^{m \times N}$, $\boldsymbol{\epsilon} := (\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_K) \in \mathbb{R}^{m \times K}$. Recall that $\boldsymbol{y} = (y_1, \cdots, y_N)^\top \in \mathbb{R}^N$. Then we have

$$\boldsymbol{y} = (\boldsymbol{Z} + \boldsymbol{\epsilon}^P)^\top \boldsymbol{\lambda}, \tag{71}$$

and

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{K} \sum_{k=1}^K \mathcal{L}_k(\boldsymbol{\theta}) = \frac{1}{2} \mathbb{E}_{\boldsymbol{\lambda}, \boldsymbol{\epsilon}} \left[ \frac{1}{K} \sum_{k=1}^K (\widehat{y}_k - y_k)^2 \right] && (72) \\
&= \frac{1}{2K} \sum_{k=1}^K \mathbb{E}_{\boldsymbol{\lambda}, \boldsymbol{\epsilon}} \left\| \boldsymbol{y}^\top \widehat{\boldsymbol{a}}_k - \boldsymbol{\lambda}^\top (\boldsymbol{z}_k + \boldsymbol{\epsilon}_k) \right\|_2^2 \\
&= \frac{1}{2K} \sum_{k=1}^K \mathbb{E}_{\boldsymbol{\lambda}, \boldsymbol{\epsilon}} \left\| \boldsymbol{\lambda}^\top (\boldsymbol{Z} + \boldsymbol{\epsilon}^P) \widehat{\boldsymbol{a}}_k - \boldsymbol{\lambda}^\top (\boldsymbol{z}_k + \boldsymbol{\epsilon}_k) \right\|_2^2 \\
&= \frac{1}{2K} \sum_{k=1}^K \mathbb{E}_{\boldsymbol{\lambda}, \boldsymbol{\epsilon}} \left[ (\boldsymbol{Z} + \boldsymbol{\epsilon}^P) \widehat{\boldsymbol{a}}_k - (\boldsymbol{z}_k + \boldsymbol{\epsilon}_k) \right]^\top \boldsymbol{\lambda} \boldsymbol{\lambda}^\top \left[ (\boldsymbol{Z} + \boldsymbol{\epsilon}^P) \widehat{\boldsymbol{a}}_k - (\boldsymbol{z}_k + \boldsymbol{\epsilon}_k) \right] \\
&= \frac{1}{2K} \sum_{k=1}^K \mathbb{E}_{\boldsymbol{\epsilon}} \left[ (\boldsymbol{Z} + \boldsymbol{\epsilon}^P) \widehat{\boldsymbol{a}}_k - (\boldsymbol{z}_k + \boldsymbol{\epsilon}_k) \right]^\top \left[ (\boldsymbol{Z} + \boldsymbol{\epsilon}^P) \widehat{\boldsymbol{a}}_k - (\boldsymbol{z}_k + \boldsymbol{\epsilon}_k) \right] \\
&= \frac{1}{2K} \sum_{k=1}^K \mathbb{E}_{\boldsymbol{\epsilon}} \left[ \| \boldsymbol{Z} \widehat{\boldsymbol{a}}_k - \boldsymbol{z}_k \|_2^2 + 2(\boldsymbol{Z} \widehat{\boldsymbol{a}}_k - \boldsymbol{z}_k)^\top (\boldsymbol{\epsilon}^P \widehat{\boldsymbol{a}}_k - \boldsymbol{\epsilon}_k) + \left\| \boldsymbol{\epsilon}^P \widehat{\boldsymbol{a}}_k - \boldsymbol{\epsilon}_k \right\|_2^2 \right], && (73)
\end{aligned}$$

where the fifth line uses Assumption 1.

Note that for all $k \in [K]$, we have

$$\mathbb{E}_{\boldsymbol{\epsilon}} (\boldsymbol{Z} \widehat{\boldsymbol{a}}_k - \boldsymbol{z}_k)^\top (\boldsymbol{\epsilon}^P \widehat{\boldsymbol{a}}_k - \boldsymbol{\epsilon}_k) = 0, \tag{74}$$

and that

$$\mathbb{E}_{\boldsymbol{\epsilon}} \left\| \boldsymbol{\epsilon}^P \widehat{\boldsymbol{a}}_k - \boldsymbol{\epsilon}_k \right\|_2^2 = m\tau \left( \| \widehat{\boldsymbol{a}}_k \|_2^2 + 1 \right) - 2m\tau \widehat{a}_{kk} \mathbb{1} \{ k \in [N] \}, \tag{75}$$

where $\mathbb{1}\{k \in [N]\}$ is the indicator function that equals 1 if $k \in [N]$ and 0 otherwise, and we have made use of the assumption that $\boldsymbol{\epsilon}_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \tau^2 \boldsymbol{I}_m)$.

Combining the above two equations with (73), we know that for $k \in [N]$, it holds that

$$\mathcal{L}_k(\boldsymbol{\theta}) = \frac{1}{2}\left(\|\boldsymbol{Z}\widehat{\boldsymbol{a}}_k - \boldsymbol{z}_k\|_2^2 + m\tau\|\widehat{\boldsymbol{a}}_k - \boldsymbol{e}_k\|_2^2\right).$$

Reorganizing the terms in the RHS of the above equation, we obtain that

$$\mathcal{L}_k(\boldsymbol{\theta}) = \frac{1}{2}\left\|\left(\boldsymbol{Z}^\top\boldsymbol{Z} + m\tau\boldsymbol{I}\right)^{1/2}\left(\widehat{\boldsymbol{a}}_k - \left(\boldsymbol{Z}^\top\boldsymbol{Z} + m\tau\boldsymbol{I}\right)^{-1}\left(\boldsymbol{Z}^\top\boldsymbol{z}_k + m\tau\boldsymbol{e}_k\right)\right)\right\|_2^2 + \frac{1}{2}c_k, \tag{76}$$

where $c_k = -\left(\boldsymbol{Z}^\top\boldsymbol{z}_k + m\tau\boldsymbol{e}_k\right)^\top\left(\boldsymbol{Z}^\top\boldsymbol{Z} + m\tau\boldsymbol{I}\right)^{-1}\left(\boldsymbol{Z}^\top\boldsymbol{z}_k + m\tau\boldsymbol{e}_k\right) + \|\boldsymbol{z}_k\|_2^2 + m\tau$.

By a similar argument, we can show that for $k \in [K]\backslash[N]$, it holds thet

$$\mathcal{L}_k(\boldsymbol{\theta}) = \frac{1}{2}\left\|\left(\boldsymbol{Z}^\top\boldsymbol{Z} + m\tau\boldsymbol{I}\right)^{1/2}\left(\widehat{\boldsymbol{a}}_k - \left(\boldsymbol{Z}^\top\boldsymbol{Z} + m\tau\boldsymbol{I}\right)^{-1}\boldsymbol{Z}^\top\boldsymbol{z}_k\right)\right\|_2^2 + \frac{1}{2}c_k', \tag{77}$$

where $c_k' = -\left(\boldsymbol{Z}^\top\boldsymbol{z}_k\right)^\top\left(\boldsymbol{Z}^\top\boldsymbol{Z} + m\tau\boldsymbol{I}\right)^{-1}\left(\boldsymbol{Z}^\top\boldsymbol{z}_k\right) + \|\boldsymbol{z}_k\|_2^2$.

(76), (77) together with (28) and the definition of $\mathcal{L}^\star$ give (29).

### H.2 Proof of Lemma 5

First, it holds that

$$\boldsymbol{Q}_h^{(t)} = \boldsymbol{Q}_h^{(t-1)} - \eta_Q\nabla_{\boldsymbol{Q}_h}\ell(\boldsymbol{\xi}^{(t-1)}) = \boldsymbol{Q}_h^{(t-1)} - \eta_Q\nabla_{\boldsymbol{Q}_h}\ell(\boldsymbol{\xi}^{(t-1)}). \tag{78}$$

Second, note that

$$\begin{aligned}\boldsymbol{w}_h^{(t)} &= \boldsymbol{w}_h^{(t-1)} - \eta_w\nabla_{\boldsymbol{w}_h}\mathcal{L}(\boldsymbol{\theta}^{(t-1)}) \\ &= \gamma\boldsymbol{\alpha}_h^{(t-1)} - \gamma^2\cdot\frac{1}{\gamma}\eta_Q\nabla_{\boldsymbol{\alpha}_h}\ell(\boldsymbol{\xi}^{(t-1)}) \\ &= \gamma\left(\boldsymbol{\alpha}_h^{(t-1)} - \eta_Q\nabla_{\boldsymbol{\alpha}_h}\ell(\boldsymbol{\xi}^{(t-1)})\right).\end{aligned}$$

Dividing both sides of the above equality by $\gamma$, we have

$$\boldsymbol{\alpha}_h^{(t)} = \boldsymbol{\alpha}_h^{(t-1)} - \eta_Q\nabla_{\boldsymbol{\alpha}_h}\ell(\boldsymbol{\xi}^{(t-1)}). \tag{79}$$

(36) follows from combining (78) and (79).

### H.3 Proof of Lemma 6

In this proof we omit the superscript $(t)$ for simplicity. We first compute the gradient of $\mathcal{L}$ w.r.t. $\boldsymbol{Q}_h$.

By (29) we know that

$$\ell(\boldsymbol{\xi}) = \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2K}\sum_{k=1}^K\left\|\bar{\boldsymbol{Z}}\boldsymbol{\delta}_k\right\|_2^2,$$

and thus we have

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\xi})}{\partial\boldsymbol{Q}_h} &= \frac{1}{K}\sum_{k=1}^K\sum_{j=1}^N\frac{\partial}{\partial\delta_{jk}}\left[\frac{1}{2}\left\|\sum_{i=1}^N\delta_{ik}\bar{\boldsymbol{z}}_i\right\|_2^2\right]\frac{\partial\delta_{jk}}{\partial\boldsymbol{Q}_h} \\ &= \frac{\gamma}{K}\sum_{k=1}^K\sum_{j=1}^N\left(\bar{\boldsymbol{Z}}\boldsymbol{\delta}_k\right)^\top\bar{\boldsymbol{z}}_j\cdot\underbrace{\alpha_{h,k}s_{jk}^h\sum_{i=1}^N s_{ik}^h(\boldsymbol{v}_j - \boldsymbol{v}_i)\boldsymbol{v}_k^\top}_{:=\boldsymbol{G}^{h,jk}}.\end{aligned} \tag{80}$$

Note that

$$\left\|\boldsymbol{G}^{h,jk}\right\|_F \leq 2\alpha s_{jk}^h, \tag{81}$$

where we use the fact that $\left\|(\boldsymbol{v}_j - \boldsymbol{v}_i)\boldsymbol{v}_k^\top\right\|_2 \leq 2$ (recall that we suppose each $\boldsymbol{v}_k$ has unit norm, $k \in [K]$.) Combining (80) and (81), we have

$$\begin{aligned}
\left\|\frac{\partial \ell(\boldsymbol{\xi})}{\partial \boldsymbol{Q}_h}\right\|_F &\leq \frac{\gamma}{K}\sum_{k=1}^{K}\sum_{j=1}^{N}\left\|\bar{\boldsymbol{Z}}\boldsymbol{\delta}_k\right\|_2 \left\|\bar{\boldsymbol{z}}_j\right\|_2 \left\|\boldsymbol{G}^{h,jk}\right\|_F \\
&\leq \frac{2\gamma}{K}\sum_{k=1}^{K}\sum_{j=1}^{N}\left\|\bar{\boldsymbol{Z}}\boldsymbol{\delta}_k\right\|_2 \bar{f}_{\max}\alpha s_{jk}^h \\
&\leq \frac{2\gamma \bar{f}_{\max}\alpha}{K}\sqrt{K}\sqrt{\sum_{k=1}^{K}\left\|\bar{\boldsymbol{Z}}\boldsymbol{\delta}_k\right\|_2^2} \\
&\leq 2\sqrt{2}\gamma \bar{f}_{\max}\alpha\sqrt{\ell(\boldsymbol{\xi}) - \mathcal{L}^\star},
\end{aligned} \tag{82}$$

where $\bar{f}_{\max}$ is defined in (22) and the third line follows from Cauchy-Schwarz inequality. (82) gives the desired result.

### H.4 Proof of Lemma 7

Lemma 7 can be verified by the following direct computation (recall that the noise in each label satisfies $\boldsymbol{\epsilon}_i \overset{i.i.d}{\sim} \mathcal{N}(0, \tau \boldsymbol{I}_m)$, $\forall i \in [N]$):

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}}\left[\frac{1}{2N}\sum_{i=1}^{N}(y_i - \boldsymbol{\lambda}^\top(f(\boldsymbol{v}_i) + \boldsymbol{\epsilon}_i))^2\right] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}\left[\frac{1}{2N}\sum_{i=1}^{N}\left((y_i - \boldsymbol{\lambda}^\top f(\boldsymbol{v}_i))^2 - 2\boldsymbol{\lambda}^\top\boldsymbol{\epsilon}_i(y_i - \boldsymbol{\lambda}^\top f(\boldsymbol{v}_i)) + \boldsymbol{\lambda}^\top\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top\boldsymbol{\lambda}\right)\right] \\
&= \frac{1}{2N}\sum_{i=1}^{N}\left((y_i - \boldsymbol{\lambda}^\top f(\boldsymbol{v}_i))^2 + \tau\left\|\boldsymbol{\lambda}\right\|_2^2\right) \\
&= \frac{1}{2N}\sum_{i=1}^{N}(y_i - \boldsymbol{\lambda}^\top f(\boldsymbol{v}_i))^2 + \frac{\tau}{2}\left\|\boldsymbol{\lambda}\right\|_2^2.
\end{aligned}$$