Superiority of Multi-Head Attention: A Theoretical Study in Shallow Transformers in In-Context Linear Regression

Yingqian Cui
Michigan State University

Jie Ren
Michigan State University

Pengfei He
Michigan State University

Hui Liu
Michigan State University

Jiliang Tang
Michigan State University

Yue Xing
Michigan State University

Abstract

We present a theoretical analysis of the performance of transformer with softmax attention in in-context learning with linear regression tasks. While the existing theoretical literature predominantly focuses on providing convergence upper bounds to show that trained transformers with single-/multi-head attention can obtain a good in-context learning performance, our research centers on comparing the exact convergence of single- and multi-head attention more rigorously. We conduct an exact theoretical analysis to demonstrate that multi-head attention with a substantial embedding dimension performs better than single-head attention. When the number of in-context examples D increases, the prediction loss using single-/multi-head attention is in $O(1/D)$, and the one for multi-head attention has a smaller multiplicative constant. In addition to the simplest data distribution setting, our technical framework in calculating the exact convergence further facilitates studying more scenarios, e.g., noisy labels, local examples, correlated features, and prior knowledge. We observe that, in general, multi-head attention is preferred over single-head attention. Our results verify the effectiveness of the design of multi-head attention in the transformer architecture.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

In-context learning (ICL) emerged as a new concept in natural language processing (NLP). With the rise of transformer architecture, NLP models become increasingly powerful and show their ability to learn new knowledge even without tuning the model parameters. Given prompts with several examples, these models can generate improved responses, showcasing their ability to adapt and 'learn' from the provided context (Dong et al., 2024).

The mechanism of transformers has been widely studied in the theoretical literature, with a main focus on linear attention (Katharopoulos et al., 2020; Choromanski et al., 2021; Schlag et al., 2021; Liu et al., 2023; Ahn et al., 2024), and emerging interest in the effectiveness and superiority of softmax attention function (Deng et al., 2023b,a; Trauger and Tewari, 2024; Hahn, 2020; Chiang and Cholak, 2022). In recent literature, people have started to work on the theoretical understanding of ICL. (Zhang et al., 2024; Oymak et al., 2023; Li et al., 2023a; Huang et al., 2024; Mahankali et al., 2020; Wu et al., 2024). Besides, Von Oswald et al. (2023); Ahn et al. (2023); Akyürek et al. (2022) and Zhang et al. (2024) explain how ICL learns gradient descent and linear regression models. Bai et al. (2023) studies ICL in generalized linear models, ridge regression, and LASSO. Cheng et al. (2024) investigate the ability of transformers to conduct ICL on non-linear functions. According to Von Oswald et al. (2023) and Dai et al. (2023), ICL can be connected with the gradient descent method.

Besides, some other studies work on multi-head attention. For example, Mahdavi et al. (2024) explored the memorization capacities of multi-head attention, and An et al. (2020) indicates a trade-off between the approximation accuracy and number of heads. Another work (Li et al., 2023b) studies the effectiveness of

ReLU-activated transformers and shows the existence of multi-layer large transformers that can conduct various regression tasks. In addition, Deora et al. (2023) investigates the convergence and generalization performance of multi-head attention in classification tasks.

However, we notice that existing theoretical literature focuses on either single-head or multi-head attention, and there is limited theoretical understanding of their difference. Furthermore, although the superiority of multi-head attentions has been widely observed by empirical studies, there is a lack of theoretical explanation about why with the same number of parameters, the structure of multi-head attentions can provide better performance than single-head attentions. However, upper bound results from previous literature may not suffice for the comparison, and to compare two methods, it is essential to derive the exact convergence rate.

In contrast to the existing literature, this work bridges the above gap by answering the following two questions. First, what are the key factors behind the superiority of multi-head attention in ICL? Second, how much is the exact benefit brought by multi-head attention over single-head attention? Our work answers the two questions by considering a transformer with **single/multi-head softmax attention** to study their ICL performance in linear regression tasks. We provide a clear comparison to quantify the superiority of multi-head attention over single-head attention. Different from Zhang et al. (2024), we do not consider linear multi-attention because linear-activated single-layer single-head attention is sufficient to learn linear regression tasks.

Our contributions are summarized as follows:

First, in Section 4.3, we show that multi-head attention is better than single-head attention by figuring out the exact prediction risk of multi-head attention. With a high input embedding dimension, multi-head attention improves the flexibility of the transformer by allowing positive and negative aggregated attention scores from different heads, thereby potentially enhancing the model's predictive performance. To facilitate the comparison between single- and multi-head attention, we present the optimal solution of single-head attention in Section 4.2. The difference between our work and the work of Ahn et al. (2023) is that their work derives the optimal solution of single-head attention and extends the analysis to multi-layer attention. In our paper, we follow our exact formulation to derive the optimal solution of single-head attention to facilitate the comparison between single-/multi-head attention. Our derived single-head result aligns with the results of Ahn et al. (2023).

Second, in Section 5, we also investigate the scenarios where the training data include prior knowledge, noisy responses, correlated features, or local examples. While our analysis shows that in most scenarios, multi-head attention is preferred over single-head attention, we also reveal some interesting behaviors of ICL when the data consists of local examples or have prior knowledge. Specifically, we observe that (1) when there is a "strong" prior knowledge, predicting using this prior knowledge leads to good performance; (2) whether local examples help or not depends on their distance to the query.

2 OTHER RELATED WORKS

In addition to the aforementioned theoretical studies, we review some empirical studies below:

The initial work that Zhang et al. (2024) builds upon was conducted by Garg et al. (2022). They empirically show the effectiveness of the transformer in performing ICL, with performance matching the optimal least squares estimator. Furthermore, Akyürek et al. (2022) demonstrate that the ICL done by transformers implicitly applies standard learning algorithms to conduct the in-context tasks.

Following these works, Panwar et al. (2024) extend the setting of Garg et al. (2022) by considering a mixture of in-context tasks in the pre-training and demonstrating the ability of the transformer to resemble the effect of Bayesian predictor under the multi-task setting. Raventós et al. (2023) empirically investigates how the diversity of the tasks in the pre-training dataset influences the performance of the transformer to do in-context tasks that are unseen in the pre-training stage. Some other related works can also be found in Fu et al. (2023); von Oswald et al. (2023); Shi et al. (2023); Saparov and He (2023); Lu et al. (2022); Liu et al. (2022); Min et al. (2022c,a); Zhang et al. (2022); Chen et al. (2022); Min et al. (2022b).

3 NOTATIONS

To mathematically define ICL, instead of merely passing a query $x_q \in \mathbb{R}^d$ (or a test sample) to the transformer to make a prediction, ICL passes a prompt, i.e., a few examples with their labels $\{(x_i, y_i)\}_{i=1,\ldots,D}$ together with the query x_q, to the transformer. Using the prompt in the format of

$$E = \begin{pmatrix} x_1 & x_2 & \ldots & x_D & x_q \\ y_1 & y_2 & \ldots & y_D & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(D+1)}, \quad (1)$$

the transformer can learn from the examples to infer the prediction for x_q. Following Zhang et al. (2024), we consider the following simplified neural network architecture

$$f(E) = E + W_{out}H, \quad (2)$$

where H denotes the attention node and W_{out} represents a fully-connected layer. Here, $H = \text{concat}(H_1, \ldots, H_h)$, with h being the number of heads in the multi-head attention. Each attention head H_j is given by

$$H_j = W_j^V E \cdot \phi\left(\frac{(W_j^K E)^\top W_j^Q E}{\rho_j}\right) \tag{3}$$

where ρ_j is a normalization factor, and the activation function ϕ is the column-wise softmax function. For each j, $W_j^V, W_j^K, W_j^Q \in \mathbb{R}^{m\times(d+1)}$, and $m = d/h$. When $h = 1$, the attention is **single-head**. When $h > 1$, the structure is called **multi-head** attention.

To train the model, we fetch the last element of the last row in $f(E)$ as the predicted value of y_q (denote as $\widehat{y}_q$), then minimize

$$L(\Omega) = \mathbb{E}_{\{x_i\},x_q,\theta}(\widehat{y}_q - y_q)^2, \tag{4}$$

where Ω is the set of parameters, and θ denotes the parameter that captures the relationship between the in-context examples and their corresponding responses, e.g., $y_q = \theta^\top x_q$ in Assumption 4.1.

4 SUPERIORITY OF MULTI-HEAD ATTENTION

In this section, we introduce the assumptions, present the optimal solution of single-head attention in ICL, and demonstrate the benefit of multi-head attention.

4.1 Assumptions

Before showing the main results, we first introduce the data generation model:

Assumption 4.1 (Data Generation Model)**.** In each prompt, the examples (x_i, y_i) and (x_q, y_q) are i.i.d. samples from the following noiseless regression model:

- The "input" $x \sim N(0, I_d)$.
- The "output" $y = \theta^\top x$.
- The coefficients θ are the same for the samples in the same prompt and are different across different prompts. In addition, $\theta \sim N(0, I_d/d)$.

Assumption 4.1 follows the work of Zhang et al. (2024) on the data generation model. For simplicity, we use Gaussian distribution to avoid tedious discussions on potential heavy tail issues, and our proofs, in general, can be extended to other data generation models.

4.2 Optimal Solution for Single-Head Attention

In the following, we figure out the optimal solution of single-head attention and summarize it in Theorem 4.1.

Theorem 4.1 (Optimal Solution of Single-Head Attention)**.** Under Assumption 4.1, assume (1) there is infinite training prompts, (2) $(W_{out}W^V)_{d+1,:} = (0, \ldots, 0, v)$, and (3) $(W^K)^\top W_Q$ is in a format of

$$(W^K)^\top W^Q = \begin{bmatrix} A & 0 \\ b & 0 \end{bmatrix},^1$$

then when $D \to \infty$, the loss value is

$$L(A, b, v) = \frac{1}{d} tr((vA - I_d)^2) + v^2\|b\|^2\mathbb{E}\|\theta\|^4 + O(1/D),$$

and the optimal solution satisfies that $\|vA - I_d\|_F^2 = O(1/D)$, and $\|vb\|^2 = O(1/D)$. In addition, when taking $A = I_d/v$ and $b = 0$,

$$L(I_d/v, 0, v) = \frac{v^2}{D}\left(\frac{v^2}{v^2-2}\right)^{\frac{d}{2}} + \frac{1}{D}\left(\frac{v^2}{v^2-2}\right)^{\frac{d}{2}+1} + o(1/D). \tag{5}$$

Denoting the optimal solution as A^*, b^*, for any $v^2 > 2$,

$$L(A^*, b^*, v) - L(I_d/v, 0, v) = o(1/D).$$

Theorem 4.1 shows the optimal solution of the single-head attention when fixing $(W_{out}W^V)_{d+1,:}$. To prove Theorem 4.1, we use Taylor expansion to separate the denominator and numerator of the attention scores. Since there are infinitely many training samples, we directly calculate the expectation of the output. In addition, it is also observed that the loss function is a quadratic function of A and b. The formal proof can be found in Appendix C.1. We also provide justifications for the assumptions on the optimal configurations considered in Theorem 4.1 in the Appendix.

In Theorem 4.1, we reduce the optimal loss of ICL to depend on only one parameter, i.e., v. Generally, v affects the prediction loss in two ways. First, as stated in Theorem 4.1, it is essential that $v^2 > 2$. When taking $v^2 \le 2$ and $A = I_d/v$, $\exp(x_q^\top A x_q) = \exp(\|x_q\|^2/v^2) \ge \exp(\|x_q\|^2/2)$, thus $\exp(x_q^\top A x_q)$ has no finite expectation, and the attention score of $(x_q, 0)$ towards itself becomes predominantly high. Second, when taking Taylor expansion on attention scores, we need the remainder terms to be negligible.

[1]We consider this parameters formats because, when we use the final element of the last row in $f(E)$ as the prediction of y_q, the values in the last column of the matrix $(W^K)^\top W^Q$ have no impact on the prediction of y_q. To simplify, we designate these values as 0.

Note that the mechanism of ICL is different from the common supervised learning methods such as Ordinary Least Square (OLS). Details can be found in Appendix B.1.

Remark 4.1. In addition to the optimal solution in Theorem 4.1, since the prediction loss is approximately a convex function of (A, b), numerical methods such as gradient descent can successfully approximate the optimal solution.

Simulation. While Theorem 4.1 presents the ICL performance of single-head attention given a fixed v, we also conduct some simulation study to investigate the role of v. In the simulation, we take different choices of (d, D, v) and set $(A, b) = (I_d/v, 0)$ to calculate the corresponding prediction loss (MSE). We run 200k repetitions for each setting to get an average and an error bar. The results are in Figure 1 and 2.

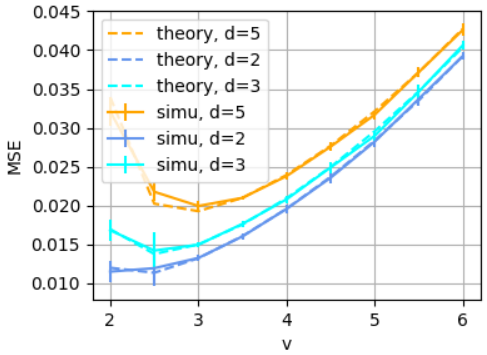


Figure 1: ICL performance of single-head attention with $(A, b) = (I_d/v, 0)$ and $D = 1000$.

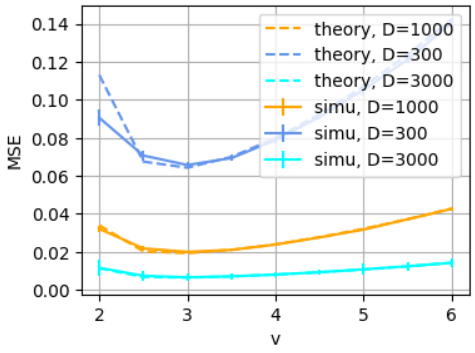


Figure 2: ICL performance of single-head attention with $(A, b) = (I_d/v, 0)$ and $d = 5$.

The figures show that the simulation of prediction loss aligns well with theoretical values. Besides, there are two main observations. First, with fixed (d, D), the MSE exhibits a U-shaped behavior as a function of v. In Figure 1, when d increases, the optimal v increases as well. Second, when fixing v, the MSE increases with d (Figure 1) and decreases with D (Figure 2).

4.3 Multi-Head Attention is Better

While Section 4.2 shows the effectiveness of single-head attention, in this subsection, through deriving the exact performance, we show that multi-head attention is better than single-head attention.

In the implementation of Garg et al. (2022), a linear layer is used to transform E into a space with a higher dimension before feeding the input into the transformer. In the last layer of the transformer, another linear layer is added so that the network outputs a single number. This increases the flexibility of the transformer.

We denote the transformation matrix applied before the transformer as $W_{in} \in \mathbb{R}^{p\times(d+1)}$ with $p \geq d+1$. In single-head attention, introducing the linear layer does not change the results. This is because the rank of $W_{in}^\top (W^K)^\top W^Q W_{in}$ is still $d+1$, meaning that the additional layer does not enlarge the representational capacity of single-head attention. In contrast, multi-head attention benefits from the dimension increase provided by W_{in}. Specifically, when increasing the dimension from d to p, each individual head has keys of dimensionality of p/h. When $p \gg d$, each head of multi-head attention take $\geq d$ dimensions to learn a function of all the dimensions of x_i, which potentially improve predictions. To explain this, in single-head attention, there is only one attention score matrix, and all the attention scores are non-negative. In contrast, we can combine the attention scores from different heads in multi-head attention so that some weights can negatively contribute to the final prediction. This flexibility is beneficial in linear regression, as negative weights and positive weights together can provide a better fit for the data. More discussions on the influence of the dimensionality can be found in Appendix B.2.

We consider a two-head attention in the following theorem to illustrate the superiority:

Theorem 4.2 (Multi-head Attention is Better)**.** Consider a two-head attention with

$$(W_1^K)^\top W_1^Q = \begin{bmatrix} A_1 & 0 \\ b_1 & 0 \end{bmatrix}, \ (W_2^K)^\top W_2^Q = \begin{bmatrix} A_2 & 0 \\ b_2 & 0 \end{bmatrix}.$$

The parameters W_1^V, W_2^V, W_{in} and W_{out} satisfy

$$\begin{aligned} f(E)_{d+1,D+1} \ = \ & vmE_{d+1,:}\phi((W_1^K E)^\top W_1^Q E_{:,D+1}) \\ & -vnE_{d+1,:}\phi((W_2^K E)^\top W_2^Q E_{:,D+1}). \end{aligned}$$

Then the optimal solution satisfies that $\|vmA_1 - vnA_2\|_F^2 = O(1/D)$ and $\|mb_1 - nb_2\|^2 = O(1/D)$.

Considering a specific case when $m = 2$, $n = 1$, $b_1 = b_2 = 0$, and setting $A_1 = (c/v)I_d$ for some $0 < c < 1$, we find that $A_2 = ((2c-1)/v)I_d$. Consequently, for any $v^2 > \max\{2c^2, 2(2c-1)^2\}$,

$$L(A_1, A_2, b_1, b_2, v)$$

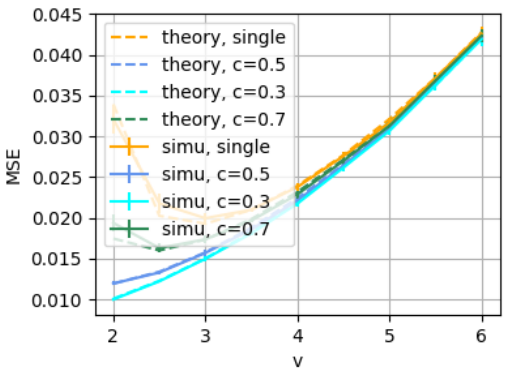


Figure 3: ICL performance of multi-head attention with $(m, n) = (2, 1)$, $(A_1, A_2, b_1, b_2) = ((c/v)I_d, ((2c-1)/v)I_d, 0, 0)$, and $(d, D) = (5, 1000)$.

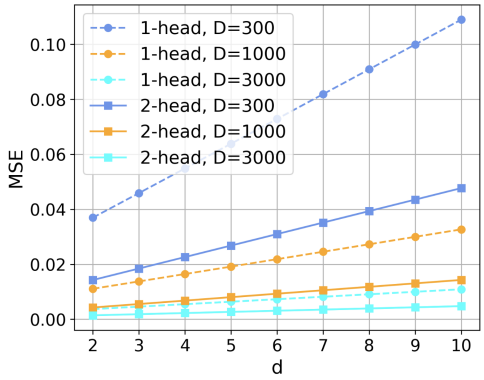


Figure 4: Optimal MSE (theory) achieved by single-/multi-head across different values of d and D.

$$
\begin{aligned}
= \; & \frac{4v^2}{D}\left(\left(\frac{v^2}{v^2-2c^2}\right)^{\frac{d}{2}} - \left(\frac{v^2}{v^2-2c(2c-1)}\right)^{\frac{d}{2}}\right) \\
& + \frac{v^2}{D}\left(\frac{v^2}{v^2-2(2c-1)^2}\right)^{\frac{d}{2}} \\
& + \frac{(2c-1)^2}{D}\left(\frac{v^2}{v^2-2(2c-1)^2}\right)\left(\frac{v^2}{v^2-2(2c-1)^2}\right)^{\frac{d}{2}} \\
& + \frac{4c^2}{D}\left(\frac{v^2}{v^2-2c^2}\right)\left(\frac{v^2}{v^2-2c^2}\right)^{\frac{d}{2}} \\
& - \frac{(8c-4)c}{D}\left(\frac{v^2}{v^2-2c(2c-1)}\right)\left(\frac{v^2}{v^2-2c(2c-1)}\right)^{\frac{d}{2}} \\
& + o\left(1/D\right).
\end{aligned}
$$

In Theorem 4.2, the condition $v^2 > \max\{2c^2, 2(2c-1)^2\}$ guarantees that $\mathbb{E}(x_q^\top A x_q)$ is finite. The proof of Theorem 4.2 is similar Theorem 4.1, and the main difficulty lies in the calculations regarding the cross terms of the two heads. Details of the proof can be found in Appendix C.2.

In Theorem 4.2, we reduce the optimal loss of ICL to a function with two parameters c and v. Due to the complex representation, deriving an exact analytical result for the minimal loss and the corresponding c and v is challenging. Therefore, we retain these parameters in the loss formulas to reflect how they influence the loss. Meanwhile, we use the following proposition to show that the loss achieved by multi-head attention is smaller than that of the optimal single-head attention:

Proposition 4.1. Following the setting of Theorem 4.1 and 4.2, multi-head attention can be reduced to single-head attention when taking $c = 1$, and $c = 1$ is not the optimal choice for multi-head attention to achieve the minimal loss.

The proof of Proposition 4.1 and some simulations can the found in Appendix C.3.

Simulation. We also conduct some simulations to compare the prediction loss of single- and multi-head attention. We use the setting in Theorem 4.2, i.e., $m = 2, n = 1$ with $A_1 = (c/v)I_d$ and $A_2 = (2c-1)I_d/v$.

From Figure 3, we can see that the simulation result is close to the theoretical value for every choice of (c, v). In addition, the MSE of multi-head attention is smaller than that of single-head attention.

While Proposition 4.1 shows that the gap between the loss of single-/multi-head attention is strictly greater than 0, Figure 4 provides quantitative results showing the exact difference in optimal loss when d and D taking different values. Specifically, we search among different parameter values to obtain the minimal loss of single-/multi-head attention. From the figure, in general, the optimal loss increase as d increase. In addition, the advantage of multi-head attention (the relative difference) is more pronounced when d is smaller.

An implication of the theories is that, with a fixed number of heads, once p/h is higher than d, further increasing p cannot enhance the model's prediction performance. This indicates that the optimal number of heads in a multi-head attention mechanism is not determined by the ambient dimension of the input data (p) but rather by the intrinsic dimension from the data (d).

5 OTHER SCENARIOS

In addition to the simplest scenario in Section 4.2 and 4.3, in this section, we relax the data generation model in Assumption 4.1 and discuss some other scenarios to understand the corresponding optimal solution for single-head attention, and verify that multi-head attention again gives better ICL performance.

Specifically, we consider scenarios involving a non-zero mean of θ (prior knowledge, Section 5.1), and local

examples x_i conditioned on x_q (Section 5.2). The prior knowledge scenario reflects a practical case where an LLM may have learned certain knowledge from its pre-training data, yet the user provides contradictory information at inference. This often occurs in retrieval-augmented generation (RAG) settings, where the LLM is connected to an external database that supplies additional information (Lewis et al., 2020). Similarly, the local examples scenario is also motivated by RAG, where retrieved information is selected based on its similarity to the user input. Thus, we incorporate local examples in ICL to better mimic this setting. Additionally, we discuss cases involving noisy responses and correlated features, though due to space constraints, the analysis of these two scenarios is deferred to Appendix A.1 and A.2.

5.1 Prior Knowledge

From the results in Section 4, the trained transformer only learns to compare the similarity of different examples, rather than learning any particular knowledge from the dataset. In this subsection, we explore whether or not the transformer can learn prior knowledge from the training data where θ is not fully random.

Assumption 5.1. For each prompt, assume that θ follows $\theta_0 + N(0, \sigma^2 I_d/d)$ for some $\|\theta_0\| = \Theta(1)$. The value of θ_0 is the same in all prompts.

The following theorem presents how the trained transformer learns θ_0:

Theorem 5.1. Denote $(W_{out}W^V)_{d+1,:} = [u, v]$ for some vector u and value v. Assume there are infinite training prompts. Under Assumption 5.1, for single-head attention, when $\sigma^2 = \Theta(1)$, the population loss is minimized when $\|u\|^2 = O(1/D)$, $\|b\|^2 = O(1/D)$, and $\|vA - I_d\|_F = O(1/D)$. For the optimal solution (u^*, b^*, A^*) at a fixed v such that $v^2 > 2$, the population loss is given by

$$\begin{aligned} & L(u^*, b^*, A^*, v) \\ = \; & \mathbb{E}\left(y_q - (W_{out}W^V_{d+1,:})^\top E\phi\left(E^\top (W^K)^\top W^Q \begin{bmatrix} x_q \\ 0 \end{bmatrix}\right)\right)^2 \\ = \; & (\|\theta_0\|^2 + \sigma^2) L_{noprior}(A^*, b^*, v) + o(1/D), \end{aligned}$$

where $L_{noprior}(A^*, b^*, v)$ denotes the optimal population loss in Theorem 4.1. When $\sigma^2 = O(1/D)$, there exists infinitely many choices of u such that $\|u\| = \Theta(1)$ and $L(u, b, A, v) = O(1/D)$. The specific conditions are in equation (11) in Appendix C.4. For multi-head attention, under the same setting as Theorem 4.2, we denote the population loss in Theorem 4.2 as $L_{noprior}$. Then, when taking $u = 0$,

$$\begin{aligned} & L(A_1, A_2, b_1, b_2, v) \\ = \; & (\|\theta_0\|^2 + \sigma^2) L_{noprior}(A_1, A_2, b_1, b_2, v) + o(1/D). \end{aligned}$$

The proof of the theorem is done by computing the partial derivatives of the loss with respect to the parameters and identifying the points where the derivatives equal zero. More details are shown in Appendix C.4 together with some simulation results.

There are several implications from Theorem 5.1. First, when the prior knowledge is weak, i.e., $\sigma^2 = \Theta(1)$, the best single-head attention does not learn θ_0. Rather, it still makes predictions by comparing the similarity between x_q and x_is. Second, when the prior knowledge is strong, i.e., $\sigma^2 = O(1/D)$, we can obtain good prediction performance when u learns from θ_0. These observations indicate that, whether the transformer uses prior knowledge to make predictions or not depends on whether there is a big distribution gap between the in-context examples and the prior knowledge. Finally, multi-head attention can still be better than single-head attention.

5.2 Local Examples

While ICL can learn from the examples chosen from the whole population, we are also interested in its efficiency when the in-context samples are selected from the neighbors of x_q.

The following two theorems indicate the prediction performance when the prompt is constructed with local examples. In Theorem 5.2, we consider the scenario where x_is are neighbors of x_q in both training stage and inference stage. In Theorem 5.3, we consider another scenario with distribution shift: x_is are totally random in the training stage, and are neighbors of x_q in the inference stage. We provide the proof of the two theorems in Appendix 5.2 and 5.3

Theorem 5.2. Assuming that for both training and test prompts, the in-context examples in the prompt are generated from $x_i \sim N(x_q, \sigma_x^2 I_d)$, and the response $y_i = x_i^\top \theta$ with $\theta \sim N(0, I_d/d)$. Then when $\mathbb{E}(x_q^\top A x_q) < \infty$, the optimal solution of the single-head transformer satisfies

$$v[\sigma_x^2(A + \theta b^\top) + I_d] \to I_d,$$

and the minimal population risk is

$$L(A^*, b^*, v) = O(\sigma_x^2/D) + o(1/D).$$

Theorem 5.2 indicates that the optimal solution for local examples is different from the one when x_is are fully random. We do not consider multi-head attention because: (1) if $\sigma_x^2 \to 0$, the single-head attention is effective enough with the overall prediction risk in $o(1/D)$; (2) if σ_x^2 is large enough, the signal x_q is much smaller than the noise size σ_x, and the problem is similar to the scenario of Theorem 4.1 and 4.2. Another observation is that, when taking different σ_x^2s in the training and inference stages, as long as $\sigma_x^2 = o(1)$ in the two stages, the ICL in the inference stage can still

achieve good performance. These observations indicate that whether local examples help ICL or not depends on their distance to the query example.

While the above result shows that a small distribution shift in σ_x^2 does not hurt the ICL performance, the following theorem considers a large distribution shift:

Theorem 5.3. Assume the training prompts are sampled in the same way as Theorem 4.1, i.e., x_is are randomly selected from the whole population. Besides, in the inference stage, in each prompt, $x_q \sim N(0, I_d)$, and the other examples $x_i \sim N(x_q, \sigma_x^2 I_d)$ for some $\sigma_x^2 > 0$. Then for single-head attention, the prediction loss goes to zero only when $\sigma_x^2 + v - 1 = 0$.

While Theorem 5.2 demonstrates the benefit of local examples, Theorem 5.3 reveals that ICL may not be consistent when facing distribution shifts. From simulations in Section 6, the actual v obtained in training does not satisfy $\sigma_x^2 + v - 1 = 0$. As a result, it is expected that generally ICL cannot perform well in the scenario that only the inference stage involves local examples but training stage does not. Without local examples in the training stage, the transformer does not learn how to handle neighbor examples.

6 EXPERIMENTS

While the simulations in previous sections directly calculate the prediction loss of ICL given specific parameter weights, in this section, we conduct experiments starting from training the transformer. The experiments for noisy response and correlated features is shown in Appendix E.

6.1 Experimental Settings

We modify the implementation of Garg et al. (2022)[2] to conduct the experiments. In particular, we adjust the input format to E as defined in (1). We consider single-layer attention and remove the positional encoding and attention mask in the transformer structure. In each training iteration, we generate a new batch of 64 prompts to train the transformer. In terms of the loss to be minimized during the training, we use the one defined in (4), i.e., we optimize the loss between y_q and $\widehat{y}_q$. We train the transformers with 500k iterations and use Adam optimizer with 0.0001 learning rate.

In the inference stage, we randomly sample 1280 prompts to obtain the average and error bar of the loss. Instead of only using x_q to calculate the loss, for each in-context example $i \in [D]$, we also make the ICL prediction and calculate the corresponding loss.

[2]Their code is available at https://github.com/dtsip/in-context-learning/tree/main under the MIT license.

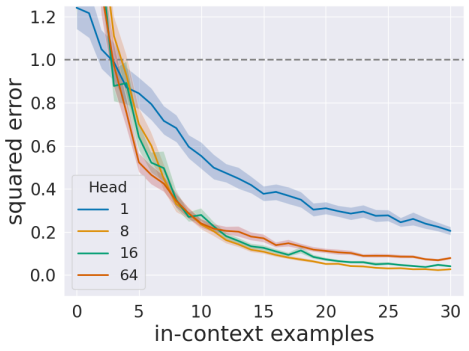


Figure 5: A comparison between single-head and multi-head with the input embedding dimension $p = 64$.

In the experiment, we compare the performance of single-head and multi-head attention. We set the input embedding dimension to $p = 64$, the dimension of in-context examples to $d = 5$, and vary the number of heads for analysis. The results are summarized in Figure 5. Furthermore, to demonstrate that the insights of the superiority of the multi-head mechanism is not limited to single-layer transformers or the Gaussian assumption on x, we provide additional simulation results in Appendix E.2 considering two different settings: (1) using multi-layer transformers, and (2) considering non-Gaussian distributions for x. The results demonstrate the consistency of our insights.

6.2 Single-head vs Multi-head

Figure 5 shows that single-head attention has a worse ICL performance than multi-head attention. In addition, although our theorems do not consider such a scenario, for multi-head attention, when h is too large so that $p/h < d$, the ICL performance can be affected. When taking $h = 64$, the ICL performance gets worse.

In addition to the ICL performance, we also conduct another experiment to examine $(W^K)^\top W^Q$. We remove the read-in layer, train the transformer, and print out $(W^K)^\top W^Q$. We repeat the experiment 10 times to see the value of $(W^K)^\top W^Q$. As in Theorem 4.1, for single-head attention without the read-in layer, $(W^K)^\top W^Q$ is expected to be in the form of I/v when $v^2 > 2$. In the 10 trials, 9 of them observe such a result, where 5 trials have $v > 0$ as in Figure 6 and 4 trials have $v < 0$ as in Figure 7. We also visualize the attention score corresponding to these two cases in Figure 16 (See Appendix E). These results indicates that the theoretical global minimum is highly likely to be attained in the real practice of transformer training.

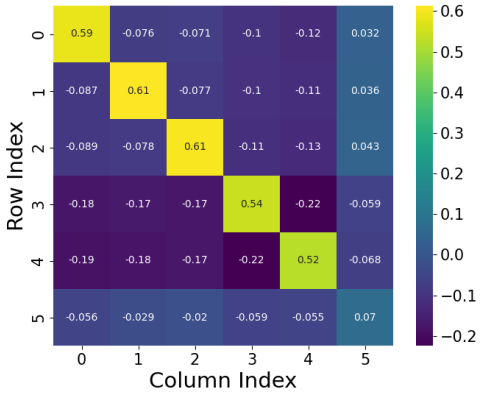


Figure 6: An illustration of the matrix $(W^K)^\top W^Q$ for the no read-in case. It is expected to be some kinds of αI_d. 4 of 10 trials are like this.

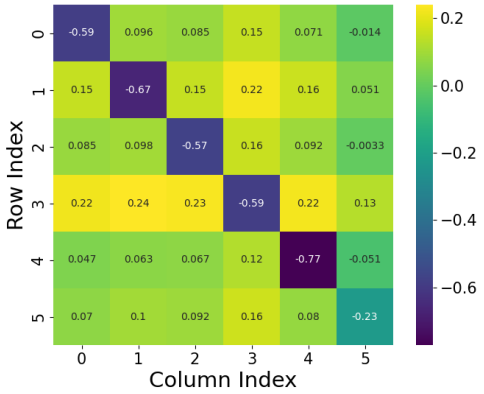


Figure 7: An illustration of the matrix $(W^K)^\top W^Q$ for the no read-in case. It is expected to be some kinds of αI_d. 5 of 10 trials are like this.

6.3 Input Embedding Dimension

As mentioned in Section 4.3, increasing the input embedding dimension p provides the flexibility of multi-head attention to achieve better ICL performance. In this experiment, we change p to examine the performance.

In Figure 8, we fix the dimension in each head (p/h), and increase h. We can observe that when the dimension is sufficient, the increasing h leads to a smaller prediction loss.

In addition, we also run different p/h for different p. As shown in Table 1, we can also see that for all $p = 64, 128, 256$, the following setting gives good ICL performance: (1) $p/h \geq d$ and (2) h is as large as possible.

6.4 Prior Knowledge

In the experiment about prior knowledge, we study the inference-stage performance under different choices of θ. Before training, we randomly generate a $\theta_0 \sim N(0, I_d)$.

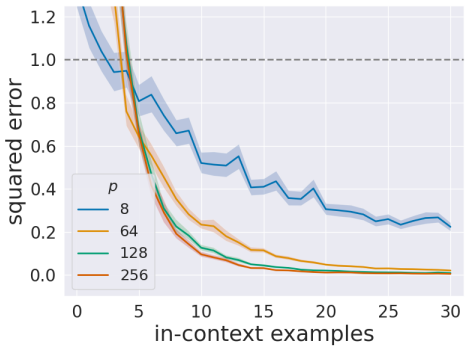


Figure 8: Results of fixing the dimension allocated in each head ($p/h = 8$), and increasing the number of heads.

Table 1: Different choices of head.

p	h	p/h	ICL	p	h	p/h	ICL
6	1	6	0.41878	64	1	64	0.18983
	2	3	0.29825		8	8	0.00769
	3	2	0.58036		16	4	0.01724
	6	1	0.56292		64	1	0.04899
128	1	128	0.16619	256	1	256	0.16141
	8	16	0.00577		8	32	0.00587
	16	8	0.00244		16	16	0.00144
	64	2	0.00611		64	4	0.00134
	128	1	0.01254		128	2	0.00159
					256	1	0.00549

Table 2: ICL performance for local examples in inference stage with/without distribution shift in the training data. More results can be found in Figure 22 in Appendix E.4.

Training	Testing	ICL	
		1 head	16 heads
Same as testing	$\sigma_x^2 = 1^2$	0.01464	0.00285
	$\sigma_x^2 = 0.1^2$	0.00049	0.00096
	$\sigma_x^2 = 0.01^2$	2.50e-05	9.79e-06
Fully random (not local examples)	$\sigma_x^2 = 1^2$	0.29317	0.60400
	$\sigma_x^2 = 0.1^2$	0.39023	1.23142
	$\sigma_x^2 = 0.01^2$	0.41253	1.12165

During the training, to generate one training prompt, we generate $\theta = \theta_0 + N(0, \sigma^2 I_d/d)$, and then generate the examples (x_i, y_i) based on θ. In the test stage, we generate different prompts following different θ. The prediction results can be found in Figure 9 for single-head attention and Figure 10 for multi-head attention with $\sigma^2 = 1$ and $\alpha = 0.1$.

We make the following observations. First, comparing Figure 9 with Figure 10, we note that multi-head attention gives better ICL performance than single-head attention. Second, as shown in Figure 9 and Figure 10, when $\theta \parallel \theta_0$, a smaller $\|\theta\|$ implies better

ICL performance. To explain this, since the ICL loss is in $O(1/D)$, a smaller $\|\theta\|$ indicates less variation among the response of different examples; thus, the multiplicative constant of the $O(1/D)$ is smaller. Finally, comparing $\eta \perp \theta_0$ with $\eta = \theta_0$, although $\|\theta\|$ for $\eta \perp \theta_0$ is smaller, the ICL performance is worse. This observation implies that the transformer learns the prior knowledge θ_0.

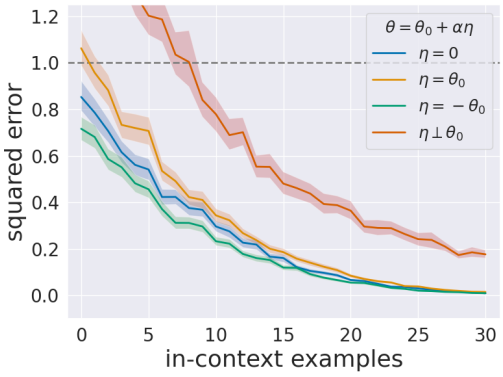


Figure 9: Head 1 prior knowledge. More results can be found from Figure 20 in Appendix E.4.

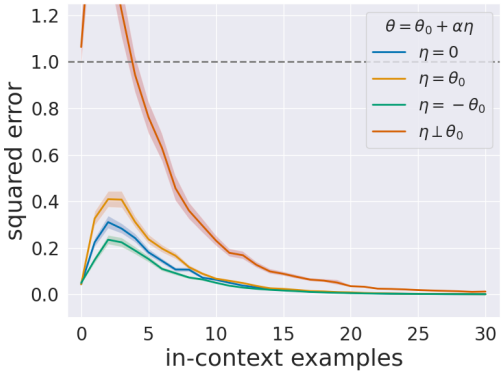


Figure 10: Head 16 prior knowledge. More results can be found in Appendix E Figure 21.

6.5 Local Examples

As discussed in Theorem 5.2 and 5.3, when both the training and inference stage use local examples with the same distribution (i.e., same σ_x^2), ICL leads to consistent predictions. When there is a large distribution shift, the prediction is not consistent. In Table 2, we demonstrate the ICL performance in the inference stage with local examples. As expected, the prediction is more accurate when the training and testing data have the same distribution, with a diminishing σ_x^2. On the other hand, when training with fully random prompts (i.e., not local examples), the prediction is inconsistent.

7 CONCLUSION

This study explicitly calculates the ICL performance in linear regression tasks to show that multi-head attention is preferred over single-head attention. In addition to the simplest case of noiseless regression, we extend the analysis to other scenarios. When the data contain prior knowledge, a transformer that learns the prior knowledge can perform well in ICL prediction. When the in-context examples are neighbors of x_q, the ICL prediction can be very efficient if there is no distribution shift.

There are several future directions. First, according to our theoretical results, the optimal number of heads in a multi-head attention mechanism is not determined by the ambient dimension but the intrinsic dimension of the input data. Future work could continue to explore how the inclusion of multiple layers affects the optimal number of heads, or how to efficiently estimate the intrinsic dimension of language data in the training of LLMs. Second, although we consider different scenarios of the data, we always consider linear regression tasks. We may extend the analysis to other problems such as non-parametric models or embedding matrix with discrete data. Finally, the statistical analysis tool applied in this paper can also be externally applied to the analysis of a wider range of topics about language models and prompting, such as the vulnerability of language models to adversarial attacks and the benefits of Chain-of-thought Prompting.

Acknowledgements

Yingqian Cui, Jie Ren, Pengfei He, Hui Liu and Jiliang Tang are supported by the National Science Foundation (NSF) under grant numbers CNS2321416, IIS2212032, IIS2212144, IOS2107215, DUE2234015, CNS2246050, DRL2405483 and IOS2035472, the Army Research Office (ARO) under grant number W911NF-21-1-0198, Amazon Faculty Award, JP Morgan Faculty Award, Meta, Microsoft and SNAP.

References

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

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

Ekin Akyürek, Dale Schuurmans, Jacob Andreas,

Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2022.

Bang An, Jie Lyu, Zhenyi Wang, Chunyuan Li, Changwei Hu, Fei Tan, Ruiyi Zhang, Yifan Hu, and Changyou Chen. Repulsive attention: Rethinking multi-head attention as bayesian inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 236–255, 2020.

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

Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573, 2022.

Xiang Cheng, Yuxin Chen, and Suvrit Sra. Transformers implement functional gradient descent to learn non-linear functions in context. In *International Conference on Machine Learning*, pages 8002–8037. PMLR, 2024.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7654–7664, 2022.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Yichuan Deng, Zhihang Li, and Zhao Song. Attention scheme inspired softmax regression. *arXiv preprint arXiv:2304.10411*, 2023a.

Yichuan Deng, Zhao Song, and Tianyi Zhou. Superiority of softmax: Unveiling the performance edge over linear attention. *arXiv preprint arXiv:2310.11685*, 2023b.

Puneesh Deora, Rouzbeh Ghaderi, Hossein Taheri, and Christos Thrampoulidis. On the optimization and generalization of multi-head attention. *CoRR*, 2023.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, 2024.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

Deqing Fu, Tian-Qi Chen, Robin Jia, and Vatsal Sharan. Transformers learn higher-order optimization methods for in-context learning: A study with linear models. *arXiv preprint arXiv:2310.17086*, 2023.

Keinosuke Fukunaga and David R Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on computers*, 100(2):176–183, 1971.

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

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pages 19660–19722, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In *International Conference on Learning Representations (ICLR 2023)*, 2023a.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context

learning. In *International conference on machine learning*, pages 19565–19594. PMLR, 2023b.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, 2022.

Langming Liu, Liu Cai, Chi Zhang, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Yifu Lv, Wenqi Fan, Yiqi Wang, Ming He, et al. Linrec: Linear attention mechanism for long-term sequential recommender systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 289–299, 2023.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022.

Arvind V Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The Twelfth International Conference on Learning Representations*, 2020.

Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. Memorization capacity of multi-head attention in transformers. In *The Twelfth International Conference on Learning Representations*, 2024.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, 2022a.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, 2022b.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022c.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In *International Conference on Machine Learning*, pages 26724–26768. PMLR, 2023.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the bayesian prism. In *The Twelfth International Conference on Learning Representations*, 2024.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 14228–14246, 2023.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, pages 9355–9366. PMLR, 2021.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR, 2024.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, 36:39257–39276, 2023.

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

Johannes von Oswald, Eyvind Niklasson, Maximilian Schlegel, Seijin Kobayashi, Nicolas Zucchet, Nino

Scherrer, Nolan Miller, Mark Sandler, Max Vladymyrov, Razvan Pascanu, et al. Uncovering mesa-optimization algorithms in transformers. *arXiv preprint arXiv:2309.05858*, 2023.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *The Twelfth International Conference on Learning Representations*, 2024.

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

Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, 2022.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] A clear description of the mathematical setting and assumptions are presented in Section 3 and 4.1.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable] The paper not propose new algorithm.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No] The experiments involved in this paper are mainly based on the code implemented by Garg et al. (2022). In Section 6.1, we have provided a detailed explanation of how we modified their code to suit our research.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] The statements of the full set of assumptions are presented in Section 4.1.
   (b) Complete proofs of all theoretical results. [Yes] The complete proofs of all theoretical results are presented in Appendix C.
   (c) Clear explanations of any assumptions. [Yes] The explanations of the assumptions are included in Section 4.1.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] We have provided sufficient details about the settings of our experiments in Section 6.1 to ensure the reproducibility.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] We have provided sufficient details about the setting of our experiments in Section 6.1.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] We have clearly defined the measure and provided prediction intervals when reporting the results in figures.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No] Our experimental setup is straightforward, involving only simple data and a single-layer model structure. It doesn't require significant computational resources and can be implemented on any GPU or CPU. As a result, specific resource requirements were not included.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] We implement existing code packages in our experiments. We have properly cited the code and the original paper.
   (b) The license information of the assets, if applicable. [Yes] We implement existing code packages in our experiments. We have properly included their license.
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable] The paper does not introduce new assets.
   (d) Information about consent from data providers/curators. [Yes] The license information we included indicates the consent from the data provider.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable] The paper does not involve sensible content.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable] The paper does not involve crowdsourcing nor research with human subjects.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Additional Scenarios

In this section, we provide analysis on the performance of single-/multi-head attention in additional scenarios including noisy examples and correlated features. The experiments for these two scenarios is shown in Appendix E.

A.1 Noisy Response

We consider linear regression tasks with noisy responses, i.e., $y_i = x_i^\top \theta + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. The following theorem demonstrates the effect of the response noise.

Theorem A.1. Assume infinite training prompts and $y_i = x_i^\top \theta + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. The optimal solution of single-head attention satisfies $tr((I_d - A/v)^2) = O(1/D)$ and $\|b\|^2 = O(1/D)$. When taking $A = I_d/v$, where $v^2 > 2$, and $b = 0$,

$$\begin{aligned} & L(I_d/v, 0, v) \\ = \; & \sigma_\epsilon^2 + \frac{v^2\sigma_\epsilon^2}{D}\left(\frac{v^2}{v^2-2}\right)^{\frac{d}{2}} + \frac{1}{D}\frac{v^4 - v^2}{v^2-2}\left(\frac{v^2}{v^2-2}\right)^{\frac{d}{2}} + o(1/D). \end{aligned}$$

For multi-head attention, taking the same parameter values as Theorem 4.2,

$$\begin{aligned} & L(A_1, A_2, b_1, b_2, v) \\ = \; & \frac{4v^2(1+\sigma_\epsilon^2)}{D}\left(\left(\frac{v^2}{v^2-2c^2}\right)^{\frac{d}{2}} - \left(\frac{v^2}{v^2-2c(2c-1)}\right)^{\frac{d}{2}}\right) \\ & + \frac{(2c-1)^2}{D}\left(\frac{v^2}{v^2-2(2c-1)^2}\right)\left(\frac{v^2}{v^2-2(2c-1)^2}\right)^{\frac{d}{2}} \\ & - \frac{(8c-4)c}{D}\left(\frac{v^2}{v^2-2c(2c-1)}\right)\left(\frac{v^2}{v^2-2c(2c-1)}\right)^{\frac{d}{2}} \\ & + \frac{4c^2}{D}\left(\frac{v^2}{v^2-2c^2}\right)\left(\frac{v^2}{v^2-2c^2}\right)^{\frac{d}{2}} + \frac{v^2(1+\sigma_\epsilon^2)}{D}\left(\frac{v^2}{v^2-2(2c-1)^2}\right)^{\frac{d}{2}} + \sigma_\epsilon^2 + o(1/D). \end{aligned}$$

The proof of Theorem A.1 is similar to that of Theorem 4.1 and 4.2, which can be found in Appendix C.6. Theorem A.1 indicates that the existence of the noise ϵ_i does not significantly change the optimal solution. For both single- and multi-head attention, there are some additional terms in the prediction loss associated with σ_ϵ^2.

Another difference from the noiseless case is the optimal v. Specifically, with a larger σ_ϵ^2, the optimal v should ensure v^2 is smaller. To explain this, denoting w_i as the attention score for each example i, and w_q as the attention score for itself, then the predicted value is $\widehat{y}_q = \sum_i v w_i (x_i^\top \theta + \epsilon_i) = \sum_i v w_i x_i^\top \theta + \sum_i v w_i \epsilon_i$, and $Var(\sum_i v w_i \epsilon_i) = v^2\sigma_\epsilon^2 \sum w_i^2$. Therefore, a smaller v^2 is required to achieve a smaller variance of prediction.

In terms of the difference between single- and multi-head attention, from the theorem it is evident that multi-head attention is still superior to single-head attention.

A.2 Correlated Features

In this subsection, we consider a scenario where x has some correlated features, i.e. $x \sim N(0, \Sigma)$ for some general $\Sigma \in \mathbb{R}^{d\times d}$. The following theorem presents the ICL performance of the transformer in this situation.

Theorem A.2. Assume $x \sim N(0, \Sigma)$ and the read-in layer is $W_{in} = \Sigma^{-1/2}$. For single-head attention, when $\mathbb{E}(x_q^\top A x_q) < \infty$, the optimal solution satisfies $\mathbb{E}\theta^\top (I_d - vA)^2\theta = O(1/D)$ and $\|b\|^2\mathbb{E}\|\theta\|^4 = O(1/D)$ where $\theta \sim N(0, \Sigma^{-1/2}/d)$. For multi-head attention, the best ICL performance is not worse than single-head attention.

To show Theorem A.2, instead of directly deriving the loss starting from correlated features, we show the equivalence of (1) the problem with correlated features and (2) the problem with isotropic features and a new θ distribution. Detailed discussions can be found in Appendix C.7.

Theorem A.2 implies some changes in the prediction loss when considering correlated features. In detail, following the setting of Theorem 4.1, i.e., $\theta \sim N(0, I_d/d)$, $\mathbb{E}\theta^\top (I_d - vA)^2\theta = tr((I_d - vA)^2)$. But in Theorem A.2, the value of $\mathbb{E}\theta^\top (I_d - vA)^2\theta$ depends on the exact distribution of θ. However, similar to Theorem 4.1, we still have $A = I_d/v$ and $b = 0$ close to the same optimal solution.

B Additional Discussions

B.1 Comparison with ordinary least squares (OLS)

As we consider noiseless linear regression between the input x and response y, when applying OLS, the MSE will always be 0. However, we would like to highlight that the setting of applying in-context learning in conducting linear regression is different from simply solving linear regression. With a pretrained transformer. Linear regression explicitly optimizes parameters for each specific regression task. For every new regression relationship, the model must undergo optimization to adjust its parameters accordingly. In contrast, ICL operates differently: once the model is trained, it can adapt to different regression relationships without requiring any parameter updates. Instead, the transformer learns to infer the underlying regression relationship directly from the examples provided in the context and uses this implicit understanding to make predictions. Therefore, it is not a fair comparison to directly evaluate ICL against OLS, as the two approaches address different paradigms of learning and adaptation. OLS solves a task-specific optimization problem, while ICL leverages its pretrained capability to generalize across tasks based solely on contextual examples.

B.2 The Dimensional Consistency Between Single-Head and Multi-Head Attention

In this subsection, we provide justification for why, in our theoretical analysis in Section 4, we assume that each head in multi-head attention uses the same dimension as single-head attention.

Firstly, for the number of parameters in the model, in real practice, an embedding matrix is often used to project the original data into a high-dimensional space. If we follow such a design and consider an input embedding dimension of $h * (d + 1)$, then the **total number of parameters** in single-head attention and multi-head attention are the same. On the other hand, when looking at the effective dimension used in each head, for single-head attention, since the original (x, y) is of $d + 1$ dimension, there is no additional information in the other $(h - 1) * (d + 1)$ dimension. For multi-head attention, each head deals with a $d + 1$-dimensional subspace of the whole $h * (d + 1)$-dimensional space.

Secondly, for the dimension of the data, in real practice, the intrinsic dimension of real text data is often much smaller than the embedding dimension. To provide an evidence, we conduct preliminary experiments to estimate the token-wise intrinsic dimension. There are various existing works dedicated to estimating intrinsic dimension of data (Fukunaga and Olsen, 1971; Facco et al., 2017; Tulchinskii et al., 2023). In the following Table 3, we present the ambient dimension of the token embedding of LLaMA (Touvron et al., 2023) and the intrinsic dimension of the embedding derived by Local Principal Components Analysis (LPCA) (Fukunaga and Olsen, 1971). Based on the calculation of LPCA, the intrinsic dimension of LLaMA's token embedding is around 77. Given that the number of heads in LLaMA is 32, the optimal dimension for the token embedding should be 77*32=2464, which is much lower than the actual token embedding (4096) dimension used in LLaMA. This suggests that in real practice, the dimension learned by each head is higher than the intrinsic dimension. This suggests that, in practice, the dimensions utilized by each head exceed the intrinsic dimension of the data, enabling multi-head attention to fully exploit the real data's information and achieve greater expressive power. This suggests that our theoretical setting aligns with this practical scenario.

Table 3: Ambient dimension and intrinsic dimension derived by LPCA

Ambient	**LPCA**
4096	77

C Proofs

C.1 Theorem 4.1

Proof of Theorem 4.1. The proof starts by breaking down the mean square loss into distinct terms involving attention scores, which are then simplified using Taylor expansion to separate their numerators and denominators. Expectations of these terms are computed separately, assuming an infinite number of training samples, and aggregated to represent the total expected loss, incorporating a $O(\frac{1}{D})$ term. Optimal values for parameters A

and b are derived to minimize this loss. These parameters are then substituted back to refine the loss expression further, explicitly deriving the rate of the $O(\frac{1}{D})$ term and adjusting the final expression to include a smaller term $o(\frac{1}{D})$.

When taking infinite many training samples (prompts), the loss function becomes

$$
\begin{aligned}
& \mathbb{E}\left(y_q-(W_{out}W_{d+1,:}^V)^\top E\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2\\
= \;& \mathbb{E}\left(y_q-v\left[y_1,y_2,\ldots,y_D,0\right]\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2\\
= \;& \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(y_q-v\left[y_1,y_2,\ldots,y_D,0\right]\phi\left(\begin{bmatrix}x_1^\top Ax_q+y_1b^\top x_q\\ \cdots\\ x_q^\top Ax_q+0\end{bmatrix}\right)\right)^2\\
= \;& \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(y_q-\frac{v\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)+\exp(x_q^\top Ax_q)}\right)^2\\
= \;& \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(y_q^2\underbrace{-2y_q\left(\frac{v\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)+\exp(x_q^\top Ax_q)}\right)}_{=A_1}\right)\\
& +\mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\underbrace{\left(\left(\frac{v\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)+\exp(x_q^\top Ax_q)}\right)^2\right)}_{=A_2}.
\end{aligned}
$$

When $D\to\infty$, we have

$$
\begin{aligned}
& \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}A_1\\
= \;& \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(-2v\theta^\top x_q)\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)+\exp(x_q^\top Ax_q)-\underbrace{D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q)}_{C}}+C\\
= \;& \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(-2v\theta^\top x_q)\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q)}\\
& +\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(2v\theta^\top x_q)\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^2}\\
& \qquad\times\left(\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)-D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q)\right)\\
& -\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(2v\theta^\top x_q)\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^3}\\
& \qquad\times\left(\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)-D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q)\right)^2+o(\frac{1}{D})\\
= \;& A_{11}+A_{12}+A_{13}+o(\frac{1}{D}).
\end{aligned}
$$

Based on Assumption 4.1, $x_i\sim N(0,I_d)$, $y_i=\theta^\top x_i$ and $\theta\sim N(0,I_d/d)$. Therefore, when taking expectation w.r.t. $\{x_i,y_i\}_{i\in[D]}$, we have

$$
\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q)}=\mathbb{E}_{\{x_1,y_1\}}\frac{D\mathbb{E}\theta^\top x_1\exp(x_1^\top Ax_q+y_1b^\top x_q)}{\exp(x_q^\top Ax_q)+D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q)},
$$

where

$$
\mathbb{E}_{\{x_1,y_1\}}\exp(x_1^\top Ax_q+y_1b^\top x_q) \;=\; \exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2),
$$

$$\begin{aligned}\mathbb{E}_{\{x_1,y_1\}}x_1\exp(x^\top Ax_q+y_1b^\top x_q) &= \mathbb{E}\frac{\partial}{\partial((A+\theta b^\top)x_q)}\exp(x_1^\top Ax_q+y_1b^\top x_q)\\ &= (A+\theta b^\top)x_q\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2).\end{aligned}$$

Therefore, we have

$$\begin{aligned}A_{11} &= \mathbb{E}_{\{x_1,y_1\}}\left(-\frac{D(2v\theta^\top x_q)\mathbb{E}\theta^\top x_1\exp(x_1^\top Ax_q+y_1b^\top x_q)}{\exp(x_q^\top Ax_q)+D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q)}\right)\\ &= \mathbb{E}_{\{x_1,y_1\}}\Bigg(-\frac{D(2v\theta^\top x_q)\mathbb{E}\theta^\top x_1\exp(x_1^\top Ax_q+y_1b^\top x_q)}{D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q)}\\ &\quad+\frac{D(2v\theta^\top x_q)\mathbb{E}\theta^\top x_1\exp(x_1^\top Ax_q+y_1b^\top x_q)\exp(x_q^\top Ax_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))^2}\Bigg)\\ &\quad+\underbrace{\mathbb{E}_{\{x_1,y_1\}}\left(-\frac{D(2v\theta^\top x_q)\mathbb{E}\theta^\top x_1\exp(x_1^\top Ax_q+y_1b^\top x_q)\exp(2x_q^\top Ax_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))^3}\right)}_{=o(1/D)}\\ &= -\frac{D(2v\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)}\\ &\quad+\frac{D(2v\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)\exp(x_q^\top Ax_q)}{D^2\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)}+o(\frac{1}{D})\\ &= -(2v\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q+\frac{(2v\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)}+o(\frac{1}{D}).\end{aligned}$$

Similarly,

$$\begin{aligned}A_{12} &= \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(2v\theta^\top x_q\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q))(\sum\exp(x_i^\top Ax_q+y_ib^\top x_q))}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp(x_1{}^\top Ax_q+y_1b^\top x_q))^2}\\ &\quad-\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(2v\theta^\top x_q\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q))(D\mathbb{E}\exp(x_1{}^\top Ax_q+y_1b^\top x_q))}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp(x_1{}^\top Ax_q+y_1b^\top x_q))^2}\\ &= \frac{(2Dv\theta^\top x_q)\mathbb{E}\theta^\top x_1\exp(2(x_1^\top Ax_q+y_1b^\top x_q)))}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp(x_1{}^\top Ax_q+y_1b^\top x_q))^2}\\ &\quad+\frac{(2D(D-1)v\theta^\top x_q\mathbb{E}_{x_1,x_2}\theta^\top x_1\exp(x_1^\top(A+\theta b^\top)x_q+x_2^\top(A+\theta b^\top)x_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp(x_1{}^\top Ax_q+y_1b^\top x_q))^2}\\ &\quad-\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(2v\theta^\top x_q\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q))(D\mathbb{E}\exp(x_1{}^\top Ax_q+y_1b^\top x_q))}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp(x_1{}^\top Ax_q+y_1b^\top x_q))^2}\\ &= A_{121}+A_{122}+A_{123},\end{aligned}$$

where

$$\begin{aligned}A_{121} &= \frac{4Dv(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(2x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)}{(\exp(x_q^\top Ax_q)+D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)^2}\\ &= \frac{4Dv(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(2x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)}{(D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)^2}+o(\frac{1}{D})\\ &= \frac{4v}{D}(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)+o(\frac{1}{D}),\end{aligned}$$

$$
\begin{aligned}
A_{122} &= \frac{2(D(D-1)v(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^2}\\
&= 2v(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q-\frac{4Dv(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top Ax_q)\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^2}\\
&\quad-\frac{2Dv(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^2}\\
&\quad-\frac{2v(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(2x_q^\top Ax_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^2}\\
&= 2v(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q-\frac{4v(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top Ax_q)\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)}{(D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^2}\\
&\quad-(\frac{2Dv(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)}{(D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^2}+o(\frac{1}{D})\\
&= 2v(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q-\frac{4v(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top Ax_q)}{(D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2))}-\frac{2v}{D}(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q+o(\frac{1}{D}),
\end{aligned}
$$

$$
\begin{aligned}
A_{123} &= -\frac{(2Dv(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2))(D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))}{\exp(x_q^\top Ax_q)+D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)^2}\\
&= -\frac{2D^2v(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)}{\exp(x_q^\top Ax_q)+D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)^2}\\
&= -2v(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q+\frac{(4Dv(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top Ax_q)\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)/2)}{(\exp(x_q^\top Ax_q)+D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)^2}+o(\frac{1}{D})\\
&= -2v(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q+\frac{(4Dv(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top Ax_q)\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2))}{(D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)^2}+o(\frac{1}{D})\\
&= -2v(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q+\frac{4v(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top x_q/2))}+o(\frac{1}{D}),
\end{aligned}
$$

and

$$
\begin{aligned}
A_{13} &= -\frac{2D^2v(\theta^\top x_q)\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top Ax_q-y_ib^\top x_q)(\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^2}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^3}\\
&\quad-\frac{2Dv(\theta^\top x_q)\mathbb{E}_{x_1}\theta^\top x_1\exp(3{x_1}^\top Ax_q+3y_1b^\top x_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^3}\\
&\quad-\frac{2D(D-1)(\theta^\top x_q)\mathbb{E}_{x_1,x_2}\theta^\top x_1\exp({x_1}^\top Ax_q+y_1b^\top x_q)\exp(2{x_2}^\top Ax_q+2y_2b^\top x_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^3})\\
&\quad+\frac{4D^2v(\theta^\top x_q)\mathbb{E}_{x_1}\theta^\top x_1\exp(2{x_1}^\top Ax_q+2y_1b^\top x_q)\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^3}\\
&\quad+\frac{4D^2(D-1)v(\theta^\top x_q)\mathbb{E}_{x_1,x_2}\theta^\top x_1\exp({x_1}^\top Ax_q+y_1b^\top x_q)\exp({x_2}^\top Ax_q+y_2b^\top x_q)\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^3}\\
&\quad-\frac{4D(D-1)v(\theta^\top x_q)\mathbb{E}_{x_1}\theta^\top x_1\exp(2{x_1}^\top Ax_q+2y_1b^\top x_q)\mathbb{E}_{x_2}\exp({x_2}^\top Ax_q+y_2b^\top x_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^3}\\
&\quad-\frac{2D(D-1)(D-2)v(\theta^\top x_q)\mathbb{E}_{x_1}\theta^\top x_1\exp({x_1}^\top Ax_q+y_1b^\top x_q)\mathbb{E}_{x_2}\exp({x_2}^\top Ax_q+y_2b^\top x_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^3}\\
&\qquad\qquad\times\mathbb{E}_{x_3}\exp({x_3}^\top Ax_q+y_3b^\top x_q)\\
&= \frac{2v}{D}(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q-\frac{2v}{D}(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q).
\end{aligned}
$$

To sum up, we have

$$\begin{aligned} A_1 &= A_{11} + A_{121} + A_{122} + A_{123} + A_{13} \\ &= -(2v\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q + \frac{(2v\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q \exp(x_q^\top A x_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)} \\ &\quad + \frac{2v}{D}(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q \exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q). \end{aligned}$$

In terms of the second-order term, since x_is are independent of each other, we have

$$\begin{aligned} &\mathbb{E}_{\{x_i\}_{i\in[D]}} A_2 \\ =\ & \mathbb{E}_{\{x_i\}_{i\in[D]}} \left(\frac{v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top A x_q + y_i b^\top x_q)}{D\mathbb{E}_{x_1}\exp(x_1^\top A x_q + y_1 b^\top x_q) + \exp(x_q^\top A x_q)} \right)^2 \\ & -2\mathbb{E}_{\{x_i\}_{i\in[D]}} \frac{\left(v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top A x_q + y_i b^\top x_q)\right)^2}{\left(D\mathbb{E}_{x_1}\exp(x_1^\top A x_q + y_1 b^\top x_q) + \exp(x_q^\top A x_q)\right)^3} \\ & \qquad \times \left(\sum \exp(x_i^\top A x_q + y_i b^\top x_q) - (D\mathbb{E}_{x_1}\exp(x_1^\top A x_q + y_1 b^\top x_q))\right) \\ & +3\mathbb{E}_{\{x_i\}_{i\in[D]}} \frac{\left(v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top A x_q + y_i b^\top x_q)\right)^2}{\left(D\mathbb{E}_{x_1}\exp(x_1^\top A x_q + y_1 b^\top x_q) + \exp(x_q^\top A x_q)\right)^4} \\ & \qquad \times \left(\sum \exp(x_i^\top A x_q + y_i b^\top x_q) - (D\mathbb{E}_{x_1}\exp(x_1^\top A x_q + y_1 b^\top x_q))\right)^2 \\ =\ & \frac{Dv^2\mathbb{E}_{x_1}\theta^\top x_1 x_1^\top \theta \exp(2x_1^\top(A+\theta b^\top)x_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top A x_q + y_1 b^\top x_q) + \exp(x_q^\top A x_q))^2} \\ & + \frac{D(D-1)v^2\mathbb{E}_{x_1,x_2}\theta^\top x_1 x_2^\top \theta \exp(x_1^\top(A+\theta b^\top)x_q + x_2^\top(A+\theta b^\top)x_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top A x_q + y_1 b^\top x_q) + \exp(x_q^\top A x_q))^2} \\ & -\mathbb{E}_{\{x_i\}_{i\in[D]}} \frac{2D(D-1)v^2\theta^\top x_1 x_2^\top \theta \exp(x_1^\top(A+\theta b^\top)x_q + x_2^\top(A+\theta b^\top)x_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top A x_q + y_1 b^\top x_q) + \exp(x_q^\top A x_q))^3} \\ & \qquad \times \left(\sum \exp(x_i^\top A x_q + y_i b^\top x_q) - (D\mathbb{E}_{x_1}\exp(x_1^\top A x_q + y_1 b^\top x_q))\right) \\ & +\mathbb{E}_{\{x_i\}_{i\in[D]}} \frac{3D(D-1)v^2\theta^\top x_1 x_2^\top \theta \exp(x_1^\top(A+\theta b^\top)x_q + x_2^\top(A+\theta b^\top)x_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top A x_q + y_1 b^\top x_q) + \exp(x_q^\top A x_q))^4} \\ & \qquad \times \left(\sum \exp(x_i^\top A x_q + y_i b^\top x_q) - (D\mathbb{E}_{x_1}\exp(x_1^\top A x_q + y_1 b^\top x_q))\right)^2 + o(\frac{1}{D}) \\ =\ & A_{21} + A_{22} + A_{23} + A_{24}. \end{aligned}$$

For the terms A_{21} to A_{24}, we have

$$\begin{aligned} A_{21} &= \frac{Dv^2\theta^\top(I_d + 4(A+\theta b^\top)x_q x_q^\top(A+\theta b^\top)^\top)\theta \exp(2x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)}{(D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2) + \exp(x_q^\top A x_q))^2} \\ &= \frac{Dv^2\theta^\top(I_d + 4(A+\theta b^\top)x_q x_q^\top(A+\theta b^\top)^\top)\theta \exp(2x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)}{(D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2))^2} + o(\frac{1}{D}) \\ &= \frac{v^2}{D}\theta^\top(I_d + 4(A+\theta b^\top)x_q x_q^\top(A+\theta b^\top)^\top)\theta \exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q) + o(\frac{1}{D}), \end{aligned}$$

$$A_{22} = \frac{D(D-1)v^2\mathbb{E}_{x_1,x_2}\theta^\top x_1 x_2^\top \theta \exp(x_1^\top(A+\theta b^\top)x_q + x_2^\top(A+\theta b^\top)x_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top A x_q + y_1 b^\top x_q))^2}$$

$$
\begin{aligned}
&-\frac{2D(D-1)v^2\mathbb{E}_{x_1,x_2}\theta^\top x_1x_2^\top\theta\exp(x_1^\top(A+\theta b^\top)x_q+x_2^\top(A+\theta b^\top)x_q)\exp(x_q^\top Ax_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))^3}+o(\frac{1}{D})\\
=\;& v^2(1-\frac{1}{D})(\theta^\top(A+\theta b^\top)x_q)^2-\frac{2v^2(\theta^\top(A+\theta b^\top)x_q)^2\exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)}+o(\frac{1}{D}),
\end{aligned}
$$

$$
A_{23} = \frac{4v^2}{D}(\theta^\top(A+\theta b^\top)x_q)^2-\frac{8v^2}{D}(\theta^\top(A+\theta b^\top)x_q)^2\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q),
$$

and

$$
\begin{aligned}
A_{24} =\;& \frac{3D^3(D-1)v^2\mathbb{E}_{x_1,x_2}\theta^\top x_1x_2^\top\theta\exp(x_1^\top(A+\theta b^\top)x_q+x_2^\top(A+\theta b^\top)x_q)(\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))^2}{(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q)+\exp(x_q^\top Ax_q))^4}\\
&+\frac{3D(D-1)(D-2)v^2\mathbb{E}_{x_1,x_2}\theta^\top x_1x_2^\top\theta\exp(x_1^\top(A+\theta b^\top)x_q+x_2^\top(A+\theta b^\top)x_q)\mathbb{E}_{x_3}\exp(2x_3^\top Ax_q+2y_3b^\top x_q)}{(D\mathbb{E}_{x_3}\exp(x_1^\top Ax_q+y_1b^\top x_q)+\exp(x_q^\top Ax_q))^4}\\
&+\frac{12D(D-1)(D-2)v^2\mathbb{E}_{x_1,x_2}\theta^\top x_1x_2^\top\theta\exp(2x_1^\top(A+\theta b^\top)x_q+x_2^\top(A+\theta b^\top)x_q)(\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q)+\exp(x_q^\top Ax_q))^4}\\
&+\frac{3D(D-1)(D-2)(D-3)v^2\mathbb{E}_{x_1,x_2}\theta^\top x_1x_2^\top\theta\exp(x_1^\top(A+\theta b^\top)x_q+x_2^\top(A+\theta b^\top)x_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q)+\exp(x_q^\top Ax_q))^4}\\
&(\mathbb{E}_{x_3,x_4}\exp(x_3^\top(A+\theta b^\top)x_q+x_4^\top(A+\theta b^\top)x_q))\\
&-\frac{6D^2(D-1)(D-2)v^2\mathbb{E}_{x_1,x_2}\theta^\top x_1x_2^\top\theta\exp(x_1^\top(A+\theta b^\top)x_q+x_2^\top(A+\theta b^\top)x_q)(\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))^2}{(D\mathbb{E}_{x_3}\exp(x_1^\top Ax_q+y_1b^\top x_q)+\exp(x_q^\top Ax_q))^4}\\
&-\frac{12D^2(D-1)v^2\mathbb{E}_{x_1,x_2}\theta^\top x_1x_2^\top\theta\exp(2x_1^\top(A+\theta b^\top)x_q+x_2^\top(A+\theta b^\top)x_q)(\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))}{(D\mathbb{E}_{x_3}\exp(x_1^\top Ax_q+y_1b^\top x_q)+\exp(x_q^\top Ax_q))^4}\\
=\;& -\frac{3v^2}{D}(\theta^\top(A+\theta b^\top)x_q)^2+\frac{3v^2}{D}(\theta^\top(A+\theta b^\top)x_q)^2\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)+o\left(\frac{1}{D}\right).
\end{aligned}
$$

To sum up,

$$
\begin{aligned}
\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}A_2 =\;& \frac{v^2}{D}\theta^\top(I_d-(A+\theta b^\top)x_qx_q^\top(A+\theta b^\top)^\top)\theta\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)\\
&+\;v^2(\theta^\top(A+\theta b^\top)x_q)^2-\frac{2v^2(\theta^\top(A+\theta b^\top)x_q)^2\exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)}+o(\frac{1}{D}).
\end{aligned}
$$

Based on the results of A_1 and A_2, we have

$$
\begin{aligned}
&\mathbb{E}\left(y_q-(W_{d+1,:}^V)^\top E\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2\\
=\;&\mathbb{E}_{(x_q,\theta)}(x_q^\top\theta)^2+v^2(\theta^\top(A+\theta b^\top)x_q)^2-2v(x_q^\top\theta)(\theta^\top(A+\theta b^\top)x_q)+O\left(\frac{1}{D}\right)\\
=\;&\mathbb{E}_{(x_q,\theta)}\left(\theta^\top(v(A+\theta b^\top)-I_d)x_q\right)^2+O\left(\frac{1}{D}\right)\\
=\;&\mathbb{E}_{(x_q,\theta)}\left[\left(\theta^\top(vA-I_d)x_q\right)^2+v^2(\|\theta\|^2b^\top x_q)^2+2v\theta^\top(vA-I_d)x_qx_q^\top b\|\theta\|^2\right]+O\left(\frac{1}{D}\right)\\
=\;&\frac{1}{d}tr\left((vA-I_d)^2\right)+v^2\|b\|^2\mathbb{E}\|\theta\|^4+O\left(\frac{1}{D}\right). \qquad (6)
\end{aligned}
$$

Therefore, to minimize the loss, the optimal A satisfies $tr\left((vA-I_d)^2\right)=O(d/D)$, and $\|b\|^2=O(1/D)$.

Furthermore, we have

$$
\mathbb{E}\left(y_q-(W_{d+1,:}^V)^\top E\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2
$$

$$
\begin{aligned}
&= \mathbb{E}_{(x_q,\theta)}\Big[(x_q^\top\theta)^2 + A_1 + A_2\Big] \\
&= \mathbb{E}_{(x_q,\theta)}\Big[(x_q^\top\theta)^2 - (2v\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q + \frac{2v}{D}(\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q \exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q) \\
&\quad + \frac{v^2}{D}\theta^\top(I_d - (A+\theta b^\top)x_q x_q^\top(A+\theta b^\top)^\top)\theta \exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q) + v^2(\theta^\top(A+\theta b^\top)x_q)^2 \\
&\quad + \frac{(2v\theta^\top x_q)\theta^\top(A+\theta b^\top)x_q \exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)} - \frac{2v^2(\theta^\top(A+\theta b^\top)x_q)^2\exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)} + o(\frac{1}{D})\Big] \\
&= \frac{1}{d}tr\left((Av-I_d)^2\right) + v^2\|b\|^2\mathbb{E}\|\theta\|^4 + \mathbb{E}_\theta\Big(\frac{v^2}{D}det(\Sigma_1)^{\frac{1}{2}}\|\theta\|^2 + \frac{2v}{D}det(\Sigma_1)^{\frac{1}{2}}\theta^\top(A+\theta b^\top)\Sigma_1\theta \qquad (7) \\
&\quad + \frac{2v}{D}det(\Sigma_2)^{\frac{1}{2}}\theta^\top(A+\theta b^\top)\Sigma_2\theta - \frac{v^2}{D}det(\Sigma_1)^{\frac{1}{2}}\theta^\top(A+\theta b^\top)\Sigma_1(A+\theta b^\top)^\top\theta \\
&\quad - \frac{2v^2}{D}det(\Sigma_2)^{\frac{1}{2}}\theta^\top(A+\theta b^\top)\Sigma_2(A+\theta b^\top)^\top\theta\Big) + o(\frac{1}{D}), \qquad (8)
\end{aligned}
$$

where $\Sigma_1 = (I - 2(A+\theta b^\top)^\top(A+\theta b^\top))^{-1}$ and $\Sigma_2 = (I + (A+\theta b^\top)^\top(A+\theta b^\top) - 2A)^{-1}$. □

Assuming that $A^* = \frac{I_d}{v} + \Delta_A$, $b^* = \Delta_b$, where $\Delta_A = O(\frac{1}{\sqrt{D}})$ and $\Delta_b = O(\frac{1}{\sqrt{D}})$, we will have

$$
\begin{aligned}
&\frac{v^2}{D}det(\Sigma_1)^{\frac{1}{2}}\|\theta\|^2\big|_{A=\frac{I_d}{v}+\Delta_A, b=\Delta_b} \\
&= \frac{v^2}{D}det\left(\left(I - 2(I_d/v + \Delta_A + \theta\Delta_b^\top)^\top(I_d/v + \Delta_A + \theta\Delta_b^\top)\right)^{-1}\right)^{\frac{1}{2}}\|\theta\|^2 \\
&= \frac{v^2}{D}det\left(\left((1-\frac{2}{v^2})I - \frac{2}{v}(\Delta_A + \theta\Delta_b^\top)^\top - \frac{2}{v}(\Delta_A + \theta\Delta_b^\top) - 2(\Delta_A + \theta\Delta_b^\top)^\top(\Delta_A + \theta\Delta_b^\top)\right)^{-1}\right)^{\frac{1}{2}}\|\theta\|^2 \\
&= \frac{v^2}{D}(\frac{v^2-2}{v^2})^{\frac{d}{2}}det\left(\left((1-\frac{2}{v^2})I - \frac{2}{v}(\Delta_A + \theta\Delta_b^\top)^\top - \frac{2}{v}(\Delta_A + \theta\Delta_b^\top) - 2(\Delta_A + \theta\Delta_b^\top)^\top(\Delta_A + \theta\Delta_b^\top)\right)^{-1}\right)\|\theta\|^2.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}_\theta\left(\frac{v^2}{D}det(\Sigma_1)^{\frac{1}{2}}\|\theta\|^2\big|_{A=\frac{I_d}{v}+\Delta_A, b=\Delta_b} - \frac{v^2}{D}det(\Sigma_1)^{\frac{1}{2}}\|\theta\|^2\big|_{A=\frac{I_d}{v}, b=0}\right) \\
&= \mathbb{E}_\theta\frac{v^2}{D}(\frac{v^2-2}{v^2})^{\frac{d}{2}}\left(-(\frac{v^2}{v^2-2})^{d+1}tr(-\frac{2}{v}(\Delta_A + \theta\Delta_b^\top)^\top - \frac{2}{v}(\Delta_A + \theta\Delta_b^\top) - 2(\Delta_A + \theta\Delta_b^\top)^\top(\Delta_A + \theta\Delta_b^\top))\right)\|\theta\|^2 \\
&= \mathbb{E}_\theta\frac{v^2}{D}(\frac{v^2}{v^2-2})^{\frac{d}{2}+1}(\frac{4}{v}tr(\Delta_A) + 2\|\ \Delta_A\|_F^2 + 2\|\Delta_b\|^2\|\theta\|^2)\|\theta\|^2 = o(\frac{1}{D}).
\end{aligned}
$$

Furthermore, we have:

$$
\begin{aligned}
&\Sigma_2\big|_{A=\frac{I_d}{v}+\Delta_A, b=\Delta_b} = \left(I + (I_d/v + \Delta_A + \theta\Delta_b^\top)^\top(I_d/v + \Delta_A + \theta\Delta_b^\top) - 2(I_d/v + \Delta_A)\right)^{-1} \\
&= \Big(\frac{(v-1)^2}{v^2}(I + \frac{v}{(v-1)^2}(\Delta_A + \theta\Delta_b^\top) + \frac{v}{(v-1)^2}(\Delta_A + \theta\Delta_b^\top)^\top - \frac{v^2}{(v-1)^2}2\Delta_A \\
&\quad + \frac{v^2}{(v-1)^2}(\Delta_A + \theta\Delta_b^\top)^\top(\Delta_A + \theta\Delta_b^\top))\Big)^{-1} \\
&= \frac{v^2}{(v-1)^2}\Big(I - \frac{v}{(v-1)^2}(\Delta_A + \theta\Delta_b^\top) - \frac{v}{(v-1)^2}(\Delta_A + \theta\Delta_b^\top)^\top + \frac{v^2}{(v-1)^2}2\Delta_A \\
&\quad - \frac{v^2}{(v-1)^2}(\Delta_A + \theta\Delta_b^\top)^\top(\Delta_A + \theta\Delta_b^\top))\Big)
\end{aligned}
$$

$$= \Sigma_2^*.$$

Therefore,

$$\begin{aligned}
&\frac{-2v^2}{D}\left(\det(\Sigma_2)^{\frac{1}{2}}\theta^\top(A+\theta b^\top)\Sigma_2(A+\theta b^\top)^\top\theta\big|_{A=\frac{I_d}{v}+\Delta_A,b=\Delta_b} - \det(\Sigma_2)^{\frac{1}{2}}\theta^\top(A+\theta b^\top)\Sigma_2(A+\theta b^\top)^\top\theta\big|_{A=\frac{I_d}{v},b=0}\right)\\
= &-\frac{2v^2}{D}\left(\det(\Sigma_2^*)^{\frac{1}{2}}\theta^\top(I/v+\Delta_A+\theta\Delta_b^\top)\Sigma_2^*(I/v+\Delta_A+\theta\Delta_b^\top)^\top\theta - \det(\Sigma_2^*)^{\frac{1}{2}}\frac{1}{(v-1)^2}\theta^\top\theta\right)\\
&-\frac{2v^2}{D}\left(\det(\Sigma_2^*)^{\frac{1}{2}}\frac{1}{(v-1)^2}\theta^\top\theta - \left[\det(\frac{v^2}{(v-1)^2}I)^{\frac{1}{2}}\right]\frac{1}{(v-1)^2}\theta^\top\theta\right)\\
= &\ o(\frac{1}{D}).
\end{aligned}$$

We can obtain similar results for other terms. Therefore, we have $L(A^*,b^*)-L(I_d/v,0)=o(\frac{1}{D})$.

When $A=\frac{I_d}{v}$ and $b=0$,

$$\begin{aligned}
&\mathbb{E}\left(y_q-(W_{d+1,:}^V)^\top E\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2\\
= &\ \mathbb{E}_{(x_q,\theta)}\left[(\frac{1}{D}(\theta^\top x_q)\theta^\top x_q\exp(x_q^\top x_q/v^2)+\frac{v^2}{D}\theta^\top\theta\exp(x_q^\top x_q/v^2)\right]\\
= &\ \frac{v^2}{D}(\frac{v^2}{v^2-2})^{\frac{d}{2}}+\frac{v^2}{D(v^2-2)}(\frac{v^2}{v^2-2})^{\frac{d}{2}}+o(\frac{1}{D}),
\end{aligned}\tag{9}$$

and v should satisfies $v^2>2$.

C.2 Theorem 4.2

Proof of Theorem 4.2. The steps in proving Theorem 4.2 closely follow those of Theorem 4.1, with the primary distinction being the treatment of parameter values in the multi-head attention model. After deriving approximations for the optimal parameters, instead of specific parameter assignments, a relational condition among the parameters is established. Specific values are then selected to satisfy this relational condition. These selected values are substituted back into the expected loss equation to compute the rate of the term $O(\frac{1}{D})$.

$$\begin{aligned}
&\mathbb{E}\left(y_q-f(E)_{d+1,D+1}\right)^2\\
= &\ \mathbb{E}\left(y_q-vmE_{d+1,:}\phi((W_1^KE)^\top W_1^QE_{:,D+1})+vnE_{d+1,:}\phi((W_2^KE)^\top W_2^QE_{:,D+1})\right)^2\\
= &\ \mathbb{E}\left(y_q-vm\left[y_1,y_2,\ldots,y_D,0\right]\phi\left(E^\top(W_1^K)^\top W_1^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)+vn\left[y_1,y_2,\ldots,y_D,0\right]\phi\left(E^\top(W_2^K)^\top W_2^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2\\
= &\ \mathbb{E}\left(y_q-\frac{vm\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top(A_1+\theta b_1^\top)x_q)}{\sum\exp(x_i^\top A_1x_q+y_ib_1^\top x_q)+\exp(x_q^\top A_1x_q)}+\frac{vn\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top(A_2+\theta b_2^\top)x_q)}{\sum\exp(x_i^\top A_2x_q+y_ib_2^\top x_q)+\exp(x_q^\top A_2x_q)}\right)^2\\
= &\ \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(y_q^2+\underbrace{\left(\frac{vm\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top A_1x_q+y_ib_1^\top x_q)}{\sum\exp(x_i^\top A_1x_q+y_ib_1^\top x_q)+\exp(x_q^\top A_1x_q)}\right)^2}_{B_1}\right)\\
&+\mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\underbrace{\left(\frac{vn\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top A_2x_q+y_ib_2^\top x_q)}{\sum\exp(x_i^\top A_2x_q+y_ib_2^\top x_q)+\exp(x_q^\top A_2x_q)}\right)^2}_{B_2}
\end{aligned}$$

$$
\begin{aligned}
&+ \quad \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\underbrace{-2y_q\left(\frac{vm\sum_{i=1}^{D}\theta^\top x_i\exp(x_i^\top A_1x_q+y_ib_1^\top x_q)}{\sum\exp(x_i^\top A_1x_q+y_ib_1^\top x_q)+\exp(x_q^\top A_1x_q)}\right)}_{B_2}\right)\\
&\quad + \left(\underbrace{2y_q\left(\frac{vn\sum_{i=1}^{D}\theta^\top x_i\exp(x_i^\top A_2x_q+y_ib_2^\top x_q)}{\sum\exp(x_i^\top A_2x_q+y_ib_2^\top x_q)+\exp(x_q^\top A_2x_q)}\right)}_{B_4}\right).\\
&+ \quad \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\underbrace{-\frac{2vm\sum_{i=1}^{D}\theta^\top x_i\exp(x_i^\top A_1x_q+y_ib_1^\top x_q)}{\sum\exp(x_i^\top A_1x_q+y_ib_1^\top x_q)+\exp(x_q^\top A_1x_q)}\frac{vn\sum_{i=1}^{D}\theta^\top x_i\exp(x_i^\top A_2x_q+y_ib_2^\top x_q)}{\sum\exp(x_i^\top A_2x_q+y_ib_2^\top x_q)+\exp(x_q^\top A_2x_q)}}_{B_5}\right).
\end{aligned}
$$

Similar to the $\mathbb{E}_{\{x_i\}_{i\in[D]}}A_2$ of C.1, we have

$$
\begin{aligned}
\mathbb{E}_{\{x_i\}_{i\in[D]}}B_1 &= \frac{v^2m^2}{D}\theta^\top(I_d-(A_1+\theta b_1^\top)x_qx_q^\top(A_1+\theta b_1^\top)^\top)\theta\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q)\\
&\quad +v^2m^2(\theta^\top(A_1+\theta b_1^\top)x_q)^2-\frac{2v^2m^2(\theta^\top(A_1+\theta b_1^\top)x_q)^2\exp(x_q^\top A_1x_q)}{D\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q/2)}+o(\frac{1}{D}),\\
\mathbb{E}_{\{x_i\}_{i\in[D]}}B_2 &= \frac{v^2n^2}{D}\theta^\top(I_d-(A_2+\theta b_2^\top)x_qx_q^\top(A_2+\theta b_2^\top)^\top)\theta\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q)\\
&\quad +v^2n^2(\theta^\top(A_2+\theta b_2^\top)x_q)^2-\frac{2v^2n^2(\theta^\top(A_2+\theta b_2^\top)x_q)^2\exp(x_q^\top A_2x_q)}{D\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q/2)}+o(\frac{1}{D}),\\
\mathbb{E}_{\{x_i\}_{i\in[D]}}B_3 &= -(2mv\theta^\top x_q)\theta^\top(A_1+\theta b_1^\top)x_q+\frac{(2vm\theta^\top x_q)\theta^\top(A_1+\theta b_1^\top)x_q\exp(x_q^\top A_1x_q)}{D\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q/2)}\\
&\quad +\frac{2mv}{D}(\theta^\top x_q)\theta^\top(A_1+\theta b_1^\top)x_q\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q)+o(\frac{1}{D}),
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{\{x_i\}_{i\in[D]}}B_4 &= +(2nv\theta^\top x_q)\theta^\top(A_2+\theta b_2^\top)x_q-\frac{(2vn\theta^\top x_q)\theta^\top(A_2+\theta b_2^\top)x_q\exp(x_q^\top A_2x_q)}{D\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q/2)}\\
&\quad -\frac{2nv}{D}(\theta^\top x_q)\theta^\top(A_2+\theta b+2^\top)x_q\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q)+o(\frac{1}{D}),
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i\in[D]}}B_5\\
&= -\frac{2v^2mnD(D-1)(\mathbb{E}_{x_1}\theta^\top x_1\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)(\mathbb{E}_{x_2}\theta^\top x_2\exp(x_2^\top A_2x_q+y_2b_2^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)+\exp(x_q^\top A_1x_q))(D\mathbb{E}_{x_2}\exp(x_2^\top A_2x_q+y_1b_2^\top x_q)+\exp(x_q^\top A_2x_q))}\\
&- \frac{2v^2mnD(\mathbb{E}_{x_1}(\theta^\top x_1)^2\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)\exp(x_1^\top A_2x_q+y_1b_2^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)+\exp(x_q^\top A_1x_q))(D\mathbb{E}_{x_2}\exp(x_2^\top A_2x_q+y_2b_2^\top x_q)+\exp(x_q^\top A_2x_q))}\\
&+ \mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\frac{2v^2mnD(D-1)(\theta^\top x_1\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)(\theta^\top x_2\exp(x_2^\top A_2x_q+y_2b_2^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)+\exp(x_q^\top A_1x_q))(D\mathbb{E}_{x_2}\exp(x_2^\top A_2x_q+y_2b_2^\top x_q)+\exp(x_q^\top A_2x_q))^2}\right)\\
&\quad \times\left(\sum_{i=1}^{D}\exp(x_i^\top A_2x_q+y_ib_2^\top x_q)-D\mathbb{E}_{x_2}\exp(x_2^\top A_2x_q+y_2b_2^\top x_q)\right)\\
&+ \mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\frac{2v^2mnD(D-1)(\theta^\top x_1\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)(\theta^\top x_2\exp(x_2^\top A_2x_q+y_2b_2^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)+\exp(x_q^\top A_1x_q))^2(D\mathbb{E}_{x_2}\exp(x_2^\top A_2x_q+y_2b_2^\top x_q)+\exp(x_q^\top A_2x_q))}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\times\left(\sum_{i=1}^{D}\exp(x_i^\top A_1 x_q + y_i b_1^\top x_q) - D\mathbb{E}_{x_1}\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q)\right)\\
-\quad&\mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{2v^2mnD(D-1)(\theta^\top x_1\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q)(\theta^\top x_2\exp(x_2^\top A_2 x_q + y_2 b_2^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q) + \exp(x_q^\top A_1 x_q))^2(D\mathbb{E}_{x_2}\exp(x_2^\top A_2 x_q + y_1 b_2^\top x_q) + \exp(x_q^\top A_2 x_q))^2}\\
&\times\left(\sum_{i=1}^{D}\exp(x_i^\top A_1 x_q + y_i b_1^\top x_q) - D\mathbb{E}_{x_1}\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q)\right)\\
&\times\left(\sum_{i=1}^{D}\exp(x_i^\top A_2 x_q + y_i b_2^\top x_q) - D\mathbb{E}_{x_2}\exp(x_2^\top A_2 x_q + y_2 b_2^\top x_q)\right)\\
-\quad&\mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{2v^2mnD(D-1)(\theta^\top x_1\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q)(\theta^\top x_2\exp(x_2^\top A_2 x_q + y_2 b_2^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q) + \exp(x_q^\top A_1 x_q))(D\mathbb{E}_{x_2}\exp(x_2^\top A_2 x_q + y_1 b_2^\top x_q) + \exp(x_q^\top A_2 x_q))^3}\\
&\times\left(\sum_{i=1}^{D}\exp(x_i^\top A_2 x_q + y_i b_2^\top x_q) - D\mathbb{E}_{x_2}\exp(x_2^\top A_2 x_q + y_2 b_2^\top x_q)\right)^2\\
-\quad&\mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{2v^2mnD(D-1)(\theta^\top x_1\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q)(\theta^\top x_2\exp(x_2^\top A_2 x_q + y_2 b_2^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q) + \exp(x_q^\top A_1 x_q))^3(D\mathbb{E}_{x_2}\exp(x_2^\top A_2 x_q + y_1 b_2^\top x_q) + \exp(x_q^\top A_2 x_q))}\\
&\times\left(\sum_{i=1}^{D}\exp(x_i^\top A_1 x_q + y_i b_1^\top x_q) - D\mathbb{E}_{x_1}\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q)\right)^2 + o(\frac{1}{D})\\
=\quad& B_{51}+B_{52}+B_{53}+B_{54}+B_{55}+B_{56}+B_{57}.
\end{aligned}
$$

For the terms B_{51} to B_{57}, we have

$$
\begin{aligned}
B_{51} \quad=\quad& -2v^2mn(1-\frac{1}{D})\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q + 2v^2mn\frac{1}{D}\frac{\exp(x_q^\top A_1 x_q)\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q}{\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q/2)}\\
&+2v^2mn\frac{1}{D}\frac{\exp(x_q^\top A_2 x_q)\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q}{\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q/2)},
\end{aligned}
$$

$$
\begin{aligned}
B_{52} \quad=\quad& -\frac{2v^2mn}{D}\theta^\top(I+(A_1+\theta b_1^\top+A_2+\theta b_2^\top)x_q^\top x_q(A_1+\theta b_1^\top+A_2+\theta b_2^\top)^\top)\theta\\
&\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_2+\theta b_2^\top)x_q/2 + x_q^\top(A_2+\theta b_2^\top)^\top(A_1+\theta b_1^\top)x_q/2),
\end{aligned}
$$

$$
\begin{aligned}
B_{53} \quad=\quad& \frac{2v^2mnD(D-1)(\mathbb{E}_{x_1}\theta^\top x_1\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q)(\mathbb{E}_{x_2}\theta^\top x_2\exp(2x_2^\top A_2 x_q + 2y_2 b_2^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q) + \exp(x_q^\top A_1 x_q))(D\mathbb{E}_{x_2}\exp(x_2^\top A_2 x_q + y_2 b_2^\top x_q) + \exp(x_q^\top A_2 x_q))^2}\\
&+\frac{2v^2mnD(D-1)(\mathbb{E}_{x_1}\theta^\top x_1\exp(x_1^\top(A_1+A_2)x_q + y_1(b_1+b_2)^\top x_q)(\mathbb{E}_{x_2}\theta^\top x_2\exp(x_2^\top A_2 x_q + y_2 b_2^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q) + \exp(x_q^\top A_1 x_q))(D\mathbb{E}_{x_2}\exp(x_2^\top A_2 x_q + y_2 b_2^\top x_q) + \exp(x_q^\top A_2 x_q))^2}\\
&-\frac{4v^2mnD(D-1)(\mathbb{E}_{x_1}\theta^\top x_1\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q)(\mathbb{E}_{x_2}\theta^\top x_2\exp(x_2^\top A_2 x_q + y_2 b_2^\top x_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1 x_q + y_1 b_1^\top x_q) + \exp(x_q^\top A_1 x_q))(D\mathbb{E}_{x_2}\exp(x_2^\top A_2 x_q + y_2 b_2^\top x_q) + \exp(x_q^\top A_2 x_q))^2}\\
&\times\mathbb{E}_{x_2}\exp(x_2^\top A_2 x_q + y_2 b_2^\top x_q))\\
=\quad& \frac{2v^2mn}{D}\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q\left(2\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q) - 2\right)\\
&+\frac{2v^2mn}{D}\theta^\top(A_1+\theta b_1+A_2+\theta b_2)x_q\theta^\top(A_2+\theta b_2)x_q\\
&\times\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_2+\theta b_2^\top)x_q/2 + x_q^\top(A_2+\theta b_2^\top)^\top(A_1+\theta b_1^\top)x_q/2),
\end{aligned}
$$

$$
\begin{aligned}
B_{54} &= \frac{2v^2mn}{D}\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q\left(2\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q)-2\right)\\
&\quad+\frac{2v^2mn}{D}\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_1+\theta b_1+A_2+\theta b_2)x_q\\
&\qquad\times\ \exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_2+\theta b_2^\top)x_q/2+x_q^\top(A_2+\theta b_2^\top)^\top(A_1+\theta b_1^\top)x_q/2),
\end{aligned}
$$

$$
\begin{aligned}
B_{55} &= \frac{-2v^2mnD^3(D-1)(\mathbb{E}_{x_1}\theta^\top x_1\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)(\mathbb{E}_{x_2}\theta^\top x_2\exp(x_2^\top A_2x_q+y_2b_2^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)+\exp(x_q^\top A_1x_q))(D\mathbb{E}_{x_2}\exp(x_2^\top A_2x_q+y_1b_2^\top x_q)+\exp(x_q^\top A_2x_q))^3}\\
&\quad\times(\mathbb{E}_{x_2}\exp(x_2^\top A_2x_q+y_2b_2^\top x_q))^2\\
&\quad+\frac{4v^2mnD^2(D-1)(D-2)(\mathbb{E}_{x_1}\theta^\top x_1\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)(\mathbb{E}_{x_2}\theta^\top x_2\exp(x_2^\top A_2x_q+y_2b_2^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)+\exp(x_q^\top A_1x_q))(D\mathbb{E}_{x_2}\exp(x_2^\top A_2x_q+y_1b_2^\top x_q)+\exp(x_q^\top A_2x_q))^3}\\
&\quad\times\mathbb{E}_{x_2,x_3}\exp(x_2^\top A_2x_q+y_2b_2^\top x_q+x_3^\top A_2x_q+y_3b_2^\top x_q)\\
&\quad-\frac{2v^2mnD(D-1)(D-2)(\mathbb{E}_{x_1}\theta^\top x_1\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)(\mathbb{E}_{x_2}\theta^\top x_2\exp(x_2^\top A_2x_q+y_2b_2^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)+\exp(x_q^\top A_1x_q))(D\mathbb{E}_{x_2}\exp(x_2^\top A_2x_q+y_1b_2^\top x_q)+\exp(x_q^\top A_2x_q))^3}\\
&\quad\mathbb{E}_{x_3}\exp(2x_3^\top A_2x_q+2y_3b_2^\top x_q)\\
&\quad-\frac{2v^2mnD(D-1)(D-2)(D-3)(\mathbb{E}_{x_1}\theta^\top x_1\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)(\mathbb{E}_{x_2}\theta^\top x_2\exp(x_2^\top A_2x_q+y_2b_2^\top x_q))}{(D\mathbb{E}_{x_1}\exp(x_1^\top A_1x_q+y_1b_1^\top x_q)+\exp(x_q^\top A_1x_q))(D\mathbb{E}_{x_2}\exp(x_2^\top A_2x_q+y_1b_2^\top x_q)+\exp(x_q^\top A_2x_q))^3}\\
&\quad\mathbb{E}_{x_3}\exp(x_3^\top A_2x_q+y_3b_2^\top x_q+x_4^\top A_2x_q+y_4b_2^\top x_q)\\
&= 2v^2mn\frac{1}{D}\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q\\
&\quad-2v^2mn\frac{1}{D}\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q),
\end{aligned}
$$

$$
\begin{aligned}
B_{56} &= 2v^2mn\frac{1}{D}\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q\\
&\quad-2v^2mn\frac{1}{D}\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q),\\
B_{57} &= 2v^2mn\frac{1}{D}\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q-\big[2v^2mn\frac{1}{D}\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q\\
&\quad\times\exp\left(x_q^\top(A_1+\theta b_1^\top)^\top(A_2+\theta b_2^\top)x_q/2+x_q^\top(A_2+\theta b_2^\top)^\top(A_1+\theta b_1^\top)x_q/2\right)\big].
\end{aligned}
$$

Based on the results of B_1 to B_5, we have

$$
\begin{aligned}
&\mathbb{E}\left(y_q-f(E)_{d+1,D+1}\right)^2\\
=\ &\mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}(x_q^\top\theta)^2+v^2m^2(\theta^\top(A_1+\theta b_1^\top)x_q)^2+v^2n^2(\theta^\top(A_2+\theta b_2^\top)x_q)^2-2vm(x_q^\top\theta)(\theta^\top(A_1+\theta b_1^\top)x_q)\\
&+2vn(x_q^\top\theta)(\theta^\top(A_2+\theta b_2^\top)x_q)-2v^2mn\theta^\top(A_1+\theta b_1^\top)x_q\theta^\top(A_2+\theta b_2^\top)x_q+O(\frac{1}{D})\\
=\ &1+\frac{v^2m^2}{d}tr(A_1A_1^\top)+v^2m^2\|b_1\|^2\mathbb{E}\|\theta\|^4+\frac{v^2n^2}{d}tr(A_2A_2^\top)+v^2n^2\|b_2\|^2\mathbb{E}\|\theta\|^4-\frac{2vm}{d}\cdot tr(A_1)\\
&+\frac{2vn}{d}\cdot tr(A_2)-\frac{2v^2mn}{d}\cdot tr(A_1A_2^\top)-2v^2mnb_1^\top b_2\mathbb{E}\|\theta\|^4+O(\frac{1}{D})\\
=\ &\frac{1}{d}tr(vmA_1-vnA_2-I)^2+v^2(mb_1-nb_2)^2\mathbb{E}\|\theta\|^4+O(\frac{1}{D}).
\end{aligned}
$$

Therefore, the optimal solutions satisfies $\|vmA_1-vnA_2\|_F^2=O(\frac{d}{D})$ and $\|mb_1-nb_2\|^2=O(\frac{1}{D})$.

Furthermore, we have

$$
\begin{aligned}
&\mathbb{E}\left(y_q-f(E)_{d+1,D+1}\right)^2\\
=\ &\mathbb{E}_{(x_q,\theta)}\left[x_q^\top\theta)^2+B_1+B_2+B_3+B_4+B_{51}+B_{52}+B_{53}+B_{54}+B_{55}+B_{56}+B_{57}\right]
\end{aligned}
$$

$$
\begin{aligned}
= \ & \mathbb{E}_{(x_q,\theta)}\Bigg[(x_q^\top\theta)^2+\frac{v^2m^2}{D}\theta^\top(I_d-(A_1+\theta b_1^\top)x_qx_q^\top(A_1+\theta b_1^\top)^\top)\theta\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q)\\
&+\frac{v^2n^2}{D}\theta^\top(I_d-(A_2+\theta b_2^\top)x_qx_q^\top(A_2+\theta b_2^\top)^\top)\theta\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q)\\
&+v^2m^2(\theta^\top(A_1+\theta b_1^\top)x_q)^2+v^2n^2(\theta^\top(A_2+\theta b_2^\top)x_q)^2-(2mv\theta^\top x_q)\theta^\top(A_1+\theta b_1^\top)x_q\\
&+(2nv\theta^\top x_q)\theta^\top(A_2+\theta b_2^\top)x_q\\
&-\frac{2v^2m^2(\theta^\top(A_1+\theta b_1^\top)x_q)^2\exp(x_q^\top A_1x_q)}{D\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q/2)}-\frac{2v^2n^2(\theta^\top(A_2+\theta b_2^\top)x_q)^2\exp(x_q^\top A_2x_q)}{D\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q/2)}\\
&+\frac{(2vm\theta^\top x_q)\theta^\top(A_1+\theta b_1^\top)x_q\exp(x_q^\top A_1x_q)}{D\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q/2)}-\frac{(2vn\theta^\top x_q)\theta^\top(A_2+\theta b_2^\top)x_q\exp(x_q^\top A_2x_q)}{D\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q/2)}\\
&+\frac{2mv}{D}(\theta^\top x_q)\theta^\top(A_1+\theta b_1^\top)x_q\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q)\\
&-\frac{2nv}{D}(\theta^\top x_q)\theta^\top(A_2+\theta b_2^\top)x_q\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q)-2v^2mn\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q\\
&+2v^2mn\frac{1}{D}\frac{\exp(x_q^\top A_1x_q)\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q}{\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q/2)}\\
&+2v^2mn\frac{1}{D}\frac{\exp(x_q^\top A_2x_q)\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q}{\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q/2)}\\
&-\frac{2v^2mn}{D}\theta^\top\theta\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_2+\theta b_2^\top)x_q/2+x_q^\top(A_2+\theta b_2^\top)^\top(A_1+\theta b_1^\top)x_q/2)\\
&-\frac{2v^2mn}{D}\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_2+\theta b_2^\top)x_q/2+x_q^\top(A_2+\theta b_2^\top)^\top(A_1+\theta b_1^\top)x_q/2)\\
&\qquad\times\left((\theta^\top(A_1+\theta b_1)^\top x_q\theta^\top(A_2+\theta b_2)^\top x_q)\right)\\
&+\frac{2v^2mn}{D}\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q)\\
&+\frac{2v^2mn}{D}\theta^\top(A_1+\theta b_1)x_q\theta^\top(A_2+\theta b_2)x_q\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q)\Bigg]+o(\frac{1}{D}).
\end{aligned}
$$

When m,n,A_1,A_2,b_1,b_2,v satisfies $vmA_1=vnA_2$ and $mb_1=mb_2$, we have

$$
\begin{aligned}
& \mathbb{E}\left(y_q-f(E)_{d+1,D+1}\right)^2\\
= \ & \mathbb{E}_{(x_q,\theta)}\Bigg[\frac{v^2m^2}{D}\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q)\|\theta\|^2+\frac{v^2n^2}{D}\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q)\|\theta\|^2\\
&-\frac{2v^2mn}{D}\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_2+\theta b_2^\top)x_q/2+x_q^\top(A_2+\theta b_2^\top)^\top(A_1+\theta b_1^\top)x_q/2)\|\theta\|^2\\
&-\frac{2v^2mn}{D}\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_2+\theta b_2^\top)x_q/2+x_q^\top(A_2+\theta b_2^\top)^\top(A_1+\theta b_1^\top)x_q/2\\
&\qquad\times\left((\theta^\top(A_1+\theta b_1)^\top x_q\theta^\top(A_2+\theta b_2)^\top x_q)\right)\\
&+\frac{v^2m^2}{D}(\theta^\top(A_1+\theta b_1^\top)x_q)^2\exp(x_q^\top(A_1+\theta b_1^\top)^\top(A_1+\theta b_1^\top)x_q)\\
&+\frac{v^2n^2}{D}(\theta^\top(A_2+\theta b_2^\top)x_q)^2\exp(x_q^\top(A_2+\theta b_2^\top)^\top(A_2+\theta b_2^\top)x_q)\Bigg]+o(\frac{1}{D}).
\end{aligned}
$$

Taking $m=2$, $n=1$, $A_1=\frac{c}{v}I$, $A_2=\frac{2c-1}{v}I$ and $b_1=b_2=0$,

$$
\begin{aligned}
& \mathbb{E}\left(y_q-f(E)_{d+1,D+1}\right)^2\\
= \ & \frac{4v^2}{D}\left((\frac{v^2}{v^2-2c^2})^{\frac{d}{2}}-(\frac{v^2}{v^2-2c(2c-1)})^{\frac{d}{2}}\right)+\frac{v^2}{D}(\frac{v^2}{v^2-2(2c-1)^2})^{\frac{d}{2}}
\end{aligned}
$$

$$+ \quad \frac{(2c-1)^2}{D}\left(\frac{v^2}{v^2-2(2c-1)^2}\right)\left(\frac{v^2}{v^2-2(2c-1)^2}\right)^{\frac{d}{2}}$$
$$+ \quad \frac{4c^2}{D}\left(\frac{v^2}{v^2-2c^2}\right)\left(\frac{v^2}{v^2-2c^2}\right)^{\frac{d}{2}} - \frac{4(2c-1)c}{D}\left(\frac{v^2}{v^2-2c(2c-1)}\right)\left(\frac{v^2}{v^2-2c(2c-1)}\right)^{\frac{d}{2}} + o(\frac{1}{D}). \tag{10}$$

Assuming that $v^2 > \max\{2c^2, 2(2c-1)^2\}$, when $0 < c < 1$, we have

$\mathbb{E}\left(y_q - f_{multi}(E)_{d+1,D+1}\right)^2 < \mathbb{E}\left(y_q - f_{sing}(E)_{d+1,D+1}\right)^2$. □

C.3 Proposition 4.1

Proof of Proposition 4.1. To differentiate the loss for single-head and multi-head attention, we use L_{sing} and L_{multi} to denote them respectively.

When $c = 1$, the loss of multi-head attention indicated by Theorem 4.2 can be reduced to the optimal loss of single-head attention:

$$L_{\text{multi}}(A_1, A_2, b_1, b_2, v)\big|_{c=1} = \frac{v^2}{D}\left(\frac{v^2}{v^2-2}\right)^{\frac{d}{2}} + \frac{v^2}{D(v^2-2)}\left(\frac{v^2}{v^2-2}\right)^{\frac{d}{2}} + o(\frac{1}{D}) \approx L_{\text{sing}}(A^*, b^*, v).$$

Upon differentiation, we have $\frac{\partial}{\partial c} L_{\text{multi}}\left(A_1, A_2, b_1, b_2, v\right)\big|_{c=1} = 0$, and

$$\frac{\partial^2}{\partial c^2} L_{\text{multi}}\left(A_1, A_2, b_1, b_2, v\right)\big|_{c=1}$$
$$= \quad -\frac{4v^2\left(d+2\right)^2}{D}\left(\frac{(\frac{v^2}{v^2-2})^{d/2}}{(v^2-2)^2}\right) - \frac{16v^2\left(d+2\right)}{D}\left(\frac{(\frac{v^2}{v^2-2})^{d/2}}{(v^2-2)^3}\right) - \frac{8v^4\left(d+2\right)d}{D}\left(\frac{(\frac{v^2}{v^2-2})^{d/2}}{(v^2-2)^4}\right) < 0.$$

Therefore, when fixing other parameters, $c = 1$ is a local maximum of the loss function, indicating that there must exist some $0 < c^* < 1$ such that $L_{\text{multi}}(A_1, A_2, b_1, b_2, v)\big|_{c=c^*} < L_{\text{multi}}(A_1, A_2, b_1, b_2, v)\big|_{c=1} \approx L_{\text{sing}}(A^*, b^*, v)$.

In Figure 11, we also plot the value of L_{multi} when changing c. One can see that when $c = 1$, for all choices of v, $L_{\text{multi}}(\cdot, \cdot, \cdot, \cdot, v)\big|_{c=1}$ achieves its local maximum. □

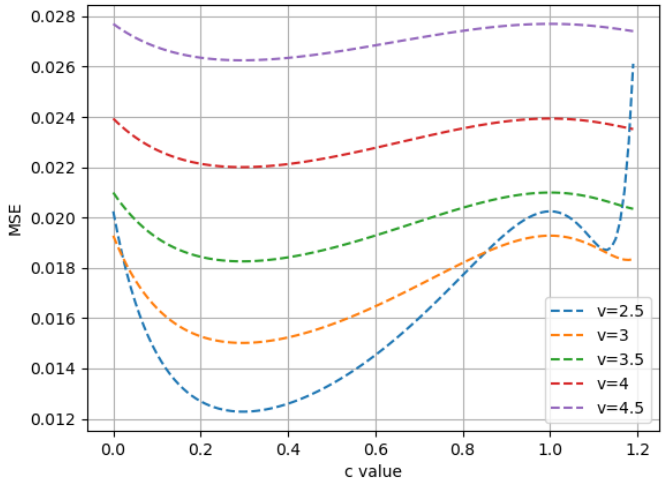


Figure 11: Theoretical loss of multi-head attention when taking different values of c. (D=1000, d=5)

C.4 Prior Knowledge

Proof of Theorem 5.1. The proof of Theorem 5.1 follow the same logic as Theorem 4.1 and Theorem 4.2. A slight difference is that when we are finding the optimal solution for the single-head attention, we first compute the

partial derivatives of the loss with respect to the parameters and then identify the points where the derivatives equal zero. Besides, the determination of the optimal parameters for single-head attention is based on the scale of σ, categorized into whether σ is much larger or smaller than $O(\frac{1}{D})$.

When taking infinite many training samples (prompts), the loss function becomes

$$
\begin{aligned}
& \mathbb{E}\left(y_q-(W_{d+1,:}^V)^\top E\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2\\
= \;& \mathbb{E}\left(y_q-\left[u^\top x_1+vy_1,u^\top x_2+vy_2,\ldots,u^\top x_D+vy_D,u^\top x_q\right]\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2\\
= \;& \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(y_q-\left[u^\top x_1+vy_1,u^\top x_2+vy_2,\ldots,u^\top x_D+vy_D,u^\top x_q\right]\phi\left(\begin{bmatrix}x_1^\top Ax_q+y_1b^\top x_q\\\cdots\\x_q^\top Ax_q+0\end{bmatrix}\right)\right)^2\\
= \;& \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(y_q-\frac{\sum_{i=1}^D(u^\top+v\theta^\top)x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)+u^\top x_q\exp(x_q^\top Ax_q)}{\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)+\exp(x_q^\top Ax_q)}\right)^2\\
= \;& \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\Bigg(y_q^2\underbrace{-2y_q\left(\frac{\sum_{i=1}^D(u^\top+v\theta^\top)x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)+u^\top x_q\exp(x_q^\top Ax_q)}{\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)+\exp(x_q^\top Ax_q)}\right)}_{A_1}\Bigg)\\
+ \;& \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\underbrace{\left(\frac{\sum_{i=1}^D(u^\top+v\theta^\top)x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)+u^\top x_q\exp(x_q^\top Ax_q)}{\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)+\exp(x_q^\top Ax_q)}\right)^2}_{A_2}.
\end{aligned}
$$

When fixing x_q and θ, the terms A_1 becomes

$$
\begin{aligned}
& \mathbb{E}_{y_q}\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}A_1\\
= \;& \mathbb{E}_{y_q}\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}-2y_q\left(\frac{\sum_{i=1}^D(u^\top+v\theta^\top)x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)+u^\top x_q\exp(x_q^\top Ax_q)}{\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)+\exp(x_q^\top Ax_q)}\right)\\
= \;& \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}-2\theta^\top x_q\left(\frac{u^\top x_q\exp(x_q^\top Ax_q)}{\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)+\exp(x_q^\top Ax_q)}\right)\\
& -2\theta^\top x_q\left(\frac{\sum_{i=1}^D(u^\top+v\theta^\top)x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q)}\right)\\
+ \;& 2\theta^\top x_q\left(\frac{\sum_{i=1}^D(u^\top+v\theta^\top)x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^2}\left(\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)-D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q)\right)\right)\\
- \;& 2\theta^\top x_q\left(\frac{\sum_{i=1}^D(u^\top+v\theta^\top)x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{(\exp(x_q^\top Ax_q)+D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q))^3}\left(\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)-D\mathbb{E}\exp({x_1}^\top Ax_q+y_1b^\top x_q)\right)^2\right)\\
= \;& A_{11}+A_{12}+A_{13}+A_{14}.
\end{aligned}
$$

For the terms A_{11} to A_{14}, we have

$$
A_{11} = -2\theta^\top x_q\left(\frac{u^\top x_q\exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)}\right)+o(\frac{1}{D}),
$$

$$
A_{12} = -2\theta^\top x_q(u^\top+v\theta^\top)(A+\theta b^\top)x_q+\frac{2\theta^\top x_q(u^\top+v\theta^\top)(A+\theta b^\top)x_q\exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)}+o(\frac{1}{D}),
$$

$$
A_{13} = -\frac{2}{D}\theta^\top x_q(u^\top+v\theta^\top)(A+\theta b^\top)x_q+\frac{4}{D}\theta^\top x_q((u^\top+v\theta^\top)(A+\theta b^\top)x_q)\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)),
$$

and

$$
A_{14} = \frac{2}{D}\theta^\top x_q(u^\top+v\theta^\top)(A+\theta b^\top)x_q-\frac{2}{D}\theta^\top x_q((u^\top+v\theta^\top)(A+\theta b^\top)x_q)\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)).
$$

In terms of A_2, when fixing x_q and θ, we have

$$\begin{aligned}
& \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}} A_2 \\
= \; & \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}} \left(\frac{\sum_{i=1}^{D}(u^\top + v\theta^\top)x_i \exp(x_i^\top A x_q + y_i b^\top x_q) + u^\top x_q \exp(x_q^\top A x_q)}{\sum \exp(x_i^\top A x_q + y_i b^\top x_q) + \exp(x_q^\top A x_q)} \right)^2 \\
= \; & \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}} \left(\frac{\sum_{i=1}^{D}(u^\top + v\theta^\top)x_i \exp(x_i^\top A x_q + y_i b^\top x_q)}{\sum \exp(x_i^\top A x_q + y_i b^\top x_q) + \exp(x_q^\top A x_q)} \right)^2 \\
+ \; & \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}} \left(\frac{u^\top x_q \exp(x_q^\top A x_q)}{\sum \exp(x_i^\top A x_q + y_i b^\top x_q) + \exp(x_q^\top A x_q)} \right)^2 \\
+ \; & \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}} \left(2u^\top x_q \exp(x_q^\top A x_q) \right) \frac{\sum_{i=1}^{D}(u^\top + v\theta^\top)x_i \exp(x_i^\top A x_q + y_i b^\top x_q)}{\left(\sum \exp(x_i^\top A x_q + y_i b^\top x_q) + \exp(x_q^\top A x_q)\right)^2} \\
= \; & A_{21} + A_{22} + A_{23},
\end{aligned}$$

where

$$\begin{aligned}
A_{21} = \; & \frac{1}{D}(u^\top + v\theta^\top)(u + v\theta)\exp(x_q^\top (A+\theta b^\top)^\top (A+\theta b^\top)x_q) + \left((u^\top + v\theta^\top)(A+\theta b^\top)x_q\right)^2 \\
& - \frac{2\left((u^\top + v\theta^\top)(A+\theta b^\top)x_q\right)^2 \exp\left(x_q^\top A x_q\right)}{D\exp(x_q^\top (A+\theta b^\top)^\top (A+\theta b^\top)x_q/2)} \\
& - \frac{1}{D}\left((u^\top + v\theta^\top)(A+\theta b^\top)x_q\right)^2 \exp(x_q^\top (A+\theta b^\top)^\top (A+\theta b^\top)x_q)
\end{aligned}$$

$$A_{22} = o(\frac{1}{D}),$$

$$A_{23} = \frac{2u^\top x_q \exp(x_q^\top A x_q)(u^\top + v\theta^\top)(A+\theta b^\top)x_q}{D\exp(x_q^\top (A+\theta b^\top)^\top (A+\theta b^\top)x_q/2)} + o(\frac{1}{D}).$$

As a result,

$$\begin{aligned}
& \mathbb{E}\left(y_q - (W_{d+1,:}^V)^\top E\phi\left(E^\top (W^K)^\top W^Q \begin{bmatrix} x_q \\ 0 \end{bmatrix}\right)\right)^2 \\
= \; & \mathbb{E}_{(x_q,\theta)}(x_q^\top\theta)^2 + ((u^\top + v\theta^\top)(A+\theta b^\top)x_q)^2 - 2(x_q^\top\theta)((u^\top + v\theta^\top)(A+\theta b^\top)x_q) + O\left(\frac{1}{D}\right) \\
= \; & \mathbb{E}_{(x_q,\theta)}\left([(u^\top + v\theta^\top)(A+\theta b^\top) - \theta^\top)]x_q\right)^2 + O\left(\frac{1}{D}\right) \\
= \; & u^\top AA^\top u + \mathbb{E}_\theta\left(v^2\theta^\top AA^\top\theta - 2v\theta^\top A^\top\theta + 2\theta^\top(vA-I)A^\top u + \|\theta\|^2 + 2u^\top Ab\theta^\top(u+v\theta)\right) \\
& + \mathbb{E}_\theta\left(2\theta^\top(vA-I)b\theta^\top(u+v\theta)\right) + \|b\|^2\mathbb{E}_\theta\left((u^\top + v\theta^\top)\theta\theta^\top(u+v\theta)\right) \\
= \; & \|A^\top u\|^2 + \frac{\sigma^2}{d}tr((vA-I)^2) + \theta_0^\top((vA-I)^2)\theta_0 + 2\theta_0^\top(vA-I)A^\top u + \underbrace{2(\theta_0^\top u + v\|\theta_0\|^2 + v\sigma^2)u^\top}_{b_1^\top} Ab \\
& + \underbrace{(\frac{4v\sigma^2}{d}\theta_0^\top + 2v\sigma^2\theta_0^\top + 2v\|\theta_0\|^2\theta_0^\top + \frac{2\sigma^2}{d}u^\top + 2u^\top\theta_0\theta_0^\top)}_{b_2^\top}(vA-I)b \\
& + \|b\|^2 \underbrace{(\frac{\sigma^2}{d}\|u\|^2 + \|u^\top\theta_0\|^2 + \frac{2v^2\sigma^4}{d} + \frac{4v^2\sigma^2}{d}\|\theta_0\|^2 + v^2(\sigma^2 + \|\theta_0\|^2)^2 + 2vu^\top(\frac{2\sigma^2}{d}\theta_0 + \sigma^2\theta_0 + \|\theta_0\|^2\theta_0))}_{a_1}
\end{aligned}$$

$$+O(\frac{1}{D}).$$

Therefore, to minimize the loss, assuming that A, u, v are fixed, the optimal b satisfies

$b^* = -\frac{1}{2a_1}\left(A^\top b_1 + (vA-I)^\top b_2\right)$.

Then we have

$$\begin{aligned}
&\mathbb{E}\left(y_q - (W^V_{d+1,:})^\top E\phi\left(E^\top (W^K)^\top W^Q \begin{bmatrix} x_q \\ 0 \end{bmatrix}\right)\right)^2 \\
= \;& \|A^\top u\| + \frac{\sigma^2}{d} tr((vA-I)^2) + \theta_0^\top((vA-I)^2)\theta_0 + 2\theta_0^\top(vA-I)A^\top u \\
& -\frac{1}{4a_1}\left[b_1^\top AA^\top b_1 + b_2^\top (vA-I)^2 b_2 + 2b_1^\top A(vA-I)^\top b_2\right] + O(\frac{1}{D}).
\end{aligned}$$

Assuming that v and θ_0 are fixed, the optimal A^* should satisfies

$$\begin{aligned}
&\left[\left(u^\top u + v\theta_0 u^\top - \frac{1}{4a_1}b_1b_1^\top - \frac{v}{4a_1}b_1b_2^\top\right) + v\left(v\frac{\sigma^2}{d}I + v\theta_0\theta_0^\top + u\theta_0^\top - \frac{v}{4a_1}b_2b_2^\top - \frac{1}{4a_1}b_2b_1^\top\right)\right]A^* \\
& - \left(v\frac{\sigma^2}{d}I + v\theta_0\theta_0^\top + u\theta_0^\top - \frac{v}{4a_1}b_2b_2^\top - \frac{1}{4a_1}b_2b_1^\top\right) = O(\frac{1}{D}).
\end{aligned}$$

- When $\sigma^2 \gg O(\frac{1}{D})$:
  If there exist an optimal A^* which can minimize $\mathbb{E}\left(y_q - f(E)_{d+1,D+1}\right)^2$, it is required that $vA^* - I = O(\frac{1}{D})$ and $\left(u^\top u + v\theta_0 u^\top - \frac{1}{4a_1}b_1b_1^\top - \frac{v}{4a_1}b_1b_2^\top\right) = O(\frac{1}{D})$. From $\left(u^\top u + v\theta_0 u^\top - \frac{1}{4a_1}b_1b_1^\top - \frac{v}{4a_1}b_1b_2^\top\right) = O(\frac{1}{D})$, we have $\left(\|u\|^2 + 2v^2\theta_0^2 + 2vu^\top\theta_0\right)\left(uu^\top + v\theta_0 u^\top\right) - v\left(\theta_0^\top u + v\theta_0^2 + v\sigma^2\right)uu^\top = O(\frac{1}{D})$, which indicates that $u \parallel \theta_0$.

  If we let $u = c_u\theta_0$, we have $c_u(c_u+2v)^2(c_u+v)\|\theta_0\|^2 + c_u^2v^2\sigma^2 = O(\frac{1}{D})$.
  - If $c_u^2 = O(\frac{1}{D})$, substituting $u = c_u\theta_0$ and $vA - I = O(\frac{1}{D})$ to $\mathbb{E}\left(y_q - f(E)_{d+1,D+1}\right)^2$, $\mathbb{E}\left(y_q - f(E)_{d+1,D+1}\right)^2$
    $= \underbrace{\frac{c^2\frac{\sigma^2}{d}\|\theta_0\|^2\left(\|\theta_0\|^2(c+2v)^2 + 2v^2\sigma^2\right)}{v^2\left((c+v)^2\|\theta_0\|^4 + \frac{\sigma^2}{d}\|\theta_0\|^2(c+2v)^2 + v^2\sigma^4(1+\frac{2}{d}) + 2v(c+v)\sigma^2\|\theta_0\|^2\right)}}_{=O(\frac{1}{D})} + O(\frac{1}{D}) = O(\frac{1}{D}).$
  - If $c_u^2 > O(\frac{1}{D})$, we have $(c_u+v)\|\theta_0\|^2 + \frac{c_uv^2\sigma^2}{(c+2v)^2} = O(\frac{1}{D})$.
    $\mathbb{E}\left(y_q - f(E)_{d+1,D+1}\right)^2 = \underbrace{\frac{c^2\frac{\sigma^2}{d}\|\theta_0\|^2\left(\|\theta_0\|^2(c+2v)^2 + 2v^2\sigma^2\right)}{v^2\left((c+v)^2\|\theta_0\|^4 + \frac{\sigma^2}{d}\|\theta_0\|^2(c+2v)^2 + v^2\sigma^4\frac{2}{d} + \frac{v^2\sigma^4(c+v)^2}{(c+2v)^2}\right)}}_{>0 \text{ and } >O(\frac{1}{D})} + O(\frac{1}{D}).$

    Therefore, in order to minimize $\mathbb{E}\left(y_q - f(E)_{d+1,D+1}\right)^2$, it is required that $c_u = O(\frac{1}{\sqrt{D}})$. Then we have $b = O(\frac{1}{\sqrt{D}})$.

- When $\sigma^2 \ll O(\frac{1}{D})$: As long as A, b, u and v satisfies

$$2a_1b + \left(A^\top b_1 + (vA-I)^\top b_2\right) = 0, \tag{11}$$

  we have $\mathbb{E}\left(y_q - f(E)_{d+1,D+1}\right)^2 = O(\frac{1}{D})$.

When taking $A = I_d/v$, $b = 0$ and $u = 0$ we have

$$A_{11} \;=\; o(\frac{1}{D}),$$

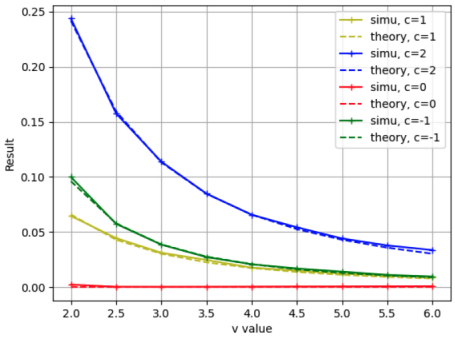


Figure 12: Simultion: when A=I/v, the loss is minimized at c=0.

$$
\begin{aligned}
A_{12} &= -2\mathbb{E}(\theta^\top x_q x_q^\top \theta) + \frac{2}{D}\mathbb{E}\frac{\theta^\top x_q x_q^\top \theta \exp(x_q^\top x_q/v)}{\exp(x_q^\top x_q/2v^2)} + o(\frac{1}{D}),\\
A_{13}+A_{14} &= \frac{2}{D}\mathbb{E}\theta^\top x_q x_q^\top \theta \exp(x_q^\top x_q/v^2) + o(\frac{1}{D}),
\end{aligned}
$$

$$
\begin{aligned}
A_{21} &= \frac{v^2}{D}\mathbb{E}\theta^\top\theta\exp(x_q^\top x_q/v^2) + \mathbb{E}\theta^\top x_q x_q^\top\theta - \mathbb{E}\frac{2}{D}\frac{\left(\theta^\top x_q x_q^\top\theta\right)\exp\left(x_q^\top x_q/v\right)}{\exp(x_q^\top x_q/2v^2)}\\
&\quad -\frac{1}{D}\mathbb{E}\left(\theta^\top x_q x_q^\top\theta\right)\exp(x_q^\top x_q/v^2) + o(\frac{1}{D}),\\
A_{22} &= o(\frac{1}{D}),\\
A_{23} &= o(\frac{1}{D}).
\end{aligned}
$$

As a result,

$$
\begin{aligned}
&\mathbb{E}\left(y_q - (W^V_{d+1,:})^\top E\phi\left(E^\top (W^K)^\top W^Q \begin{bmatrix} x_q \\ 0 \end{bmatrix}\right)\right)^2\\
=\ &\mathbb{E}(\theta^\top x_q x_q^\top\theta) + A_{12} + A_{13} + A_{14} + A_{21}\\
=\ &\frac{1}{D}\mathbb{E}\theta^\top x_q x_q^\top\theta\exp(x_q^\top x_q/v^2) + \frac{v^2}{D}\mathbb{E}\theta^\top\theta\exp(x_q^\top x_q/v^2) + o(\frac{1}{D}).\\
=\ &\frac{v^2(\sigma^2+\|\theta_0\|^2)}{D}(\frac{v^2}{v^2-2})^{\frac{d}{2}} + \frac{v^2(\sigma^2+\|\theta_0\|^2)}{D(v^2-2)}(\frac{v^2}{v^2-2})^{\frac{d}{2}}.
\end{aligned}
$$

Figure 13, 14 below demonstrate the theoretical values and the corresponding simulation results, which indicates that the simulation of prediction loss aligns well with theoretical values.

ICL performance of multi-head attention

$$
\begin{aligned}
&\mathbb{E}\left(y_q - f(E)_{d+1,D+1}\right)^2\\
=\ &\mathbb{E}\Bigg(y_q - m\left[u^\top x_1 + vy_1, u^\top x_2 + vy_2, \ldots, u^\top x_D + vy_D, u^\top x_q\right]\phi\left(E^\top (W_1^K)^\top W_1^Q \begin{bmatrix} x_q \\ 0 \end{bmatrix}\right)\\
&\quad + n\left[u^\top x_1 + vy_1, u^\top x_2 + vy_2, \ldots, u^\top x_D + vy_D, u^\top x_q\right]\phi\left(E^\top (W_2^K)^\top W_2^Q \begin{bmatrix} x_q \\ 0 \end{bmatrix}\right)\Bigg)^2\\
=\ &\mathbb{E}\Bigg(y_q - \frac{m\sum_{i=1}^D (u^\top x_i + vy_i)\exp(x_i^\top(A_1+\theta b_1^\top)x_q) + u^\top x_q\exp(x_q^\top A_1 x_q)}{\sum\exp(x_i^\top A_1 x_q + y_i b_1^\top x_q) + \exp(x_q^\top A_1 x_q)}\\
&\quad + \frac{n\sum_{i=1}^D (u^\top x_i + vy_i)\exp(x_i^\top(A_2+\theta b_2^\top)x_q) + u^\top x_q\exp(x_q^\top A_2 x_q)}{\sum\exp(x_i^\top A_2 x_q + y_i b_2^\top x_q) + \exp(x_q^\top A_2 x_q)}\Bigg)^2
\end{aligned}
$$

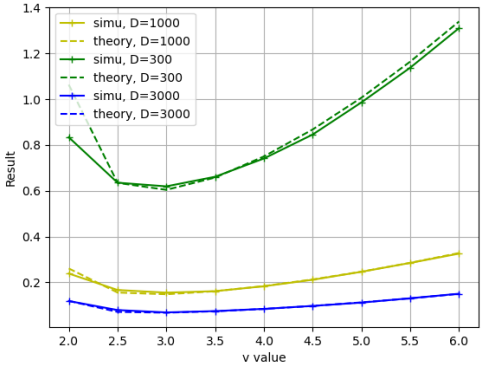


Figure 13: ICL performance of single-head attention with prior knowledge, $(A, b, u) = (I_d/v, 0, 0)$ and $d = 5$.

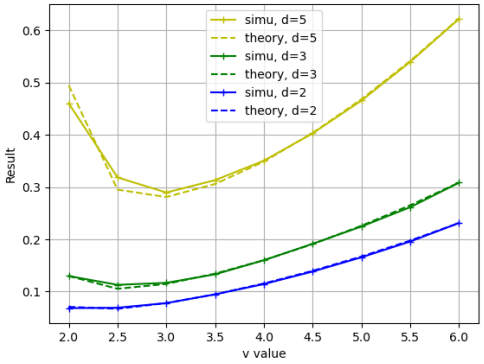


Figure 14: ICL performance of single-head attention with prior knowledge, $(A, b, u) = (I_d/v, 0, 0)$ and $D = 1000$.

$$
\begin{aligned}
= \quad & \mathbb{E}\left(y_q^2 + \left(\frac{m\sum_{i=1}^D (u^\top x_i + v y_i)\exp(x_i^\top (A_1 + \theta b_1^\top) x_q) + u^\top x_q \exp(x_q^\top A_1 x_q)}{\sum \exp(x_i^\top A_1 x_q + y_i b_1^\top x_q) + \exp(x_q^\top A_1 x_q)}\right)^2\right) \\
& + \mathbb{E}\left(\frac{n\sum_{i=1}^D (u^\top x_i + v y_i)\exp(x_i^\top (A_2 + \theta b_2^\top) x_q) + u^\top x_q \exp(x_q^\top A_2 x_q)}{\sum \exp(x_i^\top A_2 x_q + y_i b_2^\top x_q) + \exp(x_q^\top A_2 x_q)}\right)^2 \\
& - \mathbb{E}\left(2y_q\left(\frac{m\sum_{i=1}^D (u^\top x_i + v y_i)\exp(x_i^\top (A_1 + \theta b_1^\top) x_q) + u^\top x_q \exp(x_q^\top A_1 x_q)}{\sum \exp(x_i^\top A_1 x_q + y_i b_1^\top x_q) + \exp(x_q^\top A_1 x_q)}\right)\right) \\
& + \mathbb{E}\left(2y_q\left(\frac{n\sum_{i=1}^D (u^\top x_i + v y_i)\exp(x_i^\top (A_2 + \theta b_2^\top) x_q) + u^\top x_q \exp(x_q^\top A_2 x_q)}{\sum \exp(x_i^\top A_2 x_q + y_i b_2^\top x_q) + \exp(x_q^\top A_2 x_q)}\right)\right) \\
& - \mathbb{E}\left(\frac{m\sum_{i=1}^D (u^\top x_i + v y_i)\exp(x_i^\top (A_1 + \theta b_1^\top) x_q) + u^\top x_q \exp(x_q^\top A_1 x_q)}{\sum \exp(x_i^\top A_1 x_q + y_i b_1^\top x_q) + \exp(x_q^\top A_1 x_q)}\right) \\
& \times \mathbb{E}\left(\frac{n\sum_{i=1}^D (u^\top x_i + v y_i)\exp(x_i^\top (A_2 + \theta b_2^\top) x_q) + u^\top x_q \exp(x_q^\top A_2 x_q)}{\sum \exp(x_i^\top A_2 x_q + y_i b_2^\top x_q) + \exp(x_q^\top A_2 x_q)}\right).
\end{aligned}
$$

When taking $m = 2$, $n = 1$, $A_1 = \frac{c}{v} I$, $A_2 = \frac{2c-1}{v} I$ and $u = b_1 = b_2 = 0$, it becomes

$$
\begin{aligned}
& \mathbb{E}\left(y_q - f(E)_{d+1,D+1}\right)^2 \\
= \quad & \sigma^2 + \|\theta_0\|^2 + \mathbb{E}\left(\left(\frac{2v\sum_{i=1}^D y_i \exp(x_i^\top x_q (c/v))}{\sum \exp(x_i^\top x_q (c/v)) + \exp(\|x_q\|^2 (c/v))}\right)^2 + \left(\frac{v\sum_{i=1}^D y_i \exp(x_i^\top x_q (2c-1)/v)}{\sum \exp(x_i^\top x_q (2c-1)/v) + \exp(\|x_q\|^2 (2c-1)/v)}\right)^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\mathbb{E}\left(2y_q\left(\frac{v\sum_{i=1}^{D} y_i \exp(x_i^\top x_q(2c-1)/v)}{\sum \exp(x_i^\top x_q(2c-1)/v)+\exp(\|x_q\|^2(2c-1)/v)}\right)-2y_q\left(\frac{2v\sum_{i=1}^{D} \theta^\top x_i \exp(x_i^\top x_q(c/v))}{\sum \exp(x_i^\top x_q(c/v))+\exp(\|x_q\|^2(c/v))}\right)\right)\\
&-\mathbb{E}\left(\frac{2v\sum_{i=1}^{D} y_i \exp(x_i^\top x_q(c/v))}{\sum \exp(x_i^\top x_q(c/v))+\exp(\|x_q\|^2(c/v))}\frac{2v\sum_{i=1}^{D} y_i \exp(x_i^\top x_q(2c-1)/v)}{\sum \exp(x_i^\top x_q(2c-1)/v)+\exp(\|x_q\|^2(2c-1)/v)}\right)\\
=\;& \sigma^2+\|\theta_0\|^2+B_1+B_2+B_3.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
B_1 =\;& \mathbb{E}\left(\left(\frac{2v\sum_{i=1}^{D} \theta^\top x_i \exp(x_i^\top x_q(c/v))}{\sum \exp(x_i^\top x_q(c/v))+\exp(\|x_q\|^2(c/v))}\right)^2+\left(\frac{v\sum_{i=1}^{D} \theta^\top x_i \exp(x_i^\top x_q(2c-1)/v)}{\sum \exp(x_i^\top x_q(2c-1)/v)+\exp(\|x_q\|^2(2c-1)/v)}\right)^2\right)\\
=\;& \frac{4}{D}v^2\|\theta\|^2\exp(c^2\|x_q\|^2/v^2)+4c^2(\sigma^2+\|\theta_0\|^2)-\frac{8c^2(\theta^\top x_q x_q^\top\theta)\exp\left(c^2\|x_q\|^2/v^2\right)}{D\exp(c^2\|x_q\|^2/2v^2)}\\
&-\frac{4c}{D}\left(\theta^\top x_q x_q^\top\theta\right)\exp(c^2\|x_q\|^2/v^2)+\frac{1}{D}v^2\|\theta\|^2\exp\left((2c-1)^2\|x_q\|^2/v^2\right)+(2c-1)^2(\sigma^2+\|\theta_0\|^2)\\
&-\frac{2(2c-1)^2(\theta^\top x_q x_q^\top\theta)\exp\left(((2c-1)^2\|x_q\|^2/v^2)\right)}{D\exp((2c-1)^2\|x_q\|^2/2v^2)}-\frac{2c-1}{D}\left(\theta^\top x_q x_q^\top\theta\right)\exp((2c-1)^2\|x_q\|^2/v^2),
\end{aligned}
$$

$$
\begin{aligned}
B_2 =\;& \mathbb{E}\left(2y_q\left(\frac{v\sum_{i=1}^{D} y_i \exp(x_i^\top x_q(2c-1)/v)}{\sum \exp(x_i^\top x_q(2c-1)/v)+\exp(\|x_q\|^2(2c-1)/v)}\right)-2y_q\left(\frac{2v\sum_{i=1}^{D} \theta^\top x_i \exp(x_i^\top x_q(c/v))}{\sum \exp(x_i^\top x_q(c/v))+\exp(\|x_q\|^2(c/v))}\right)\right)\\
=\;& 2(2c-1)\theta^\top x_q x_q^\top\theta-\frac{2(2c-1)\theta^\top x_q x_q^\top\theta\exp((2c-1)\|x_q\|^2/v)}{D\exp((2c-1)^2\|x_q\|^2/2v^2)}-\frac{2(2c-1)}{D}\theta^\top x_q x_q^\top\theta\exp((2c-1)^2\|x_q\|^2/v^2))\\
-\;& 4c\theta^\top x_q x_q^\top\theta+\frac{4c\theta^\top x_q x_q^\top\theta\exp(c\|x_q\|^2/v)}{D\exp(c^2\|x_q\|^2/2v^2)}+\frac{4c}{D}\theta^\top x_q x_q^\top\theta\exp(c^2\|x_q\|^2/v^2)),
\end{aligned}
$$

and

$$
\begin{aligned}
B_3 =\;& -4c(2c-1)(1-\frac{1}{D})\theta^\top x_q x_q^\top\theta+\frac{4}{D}\frac{c(2c-1)\theta^\top x_q x_q^\top\theta\exp(c\|x_q\|^2/v)}{\exp(c^2\|x_q\|^2/2v^2)}+\frac{4}{D}\frac{c(2c-1)\theta^\top x_q x_q^\top\theta\exp((2c-1)\|x_q\|^2/v)}{\exp((2c-1)^2\|x_q\|^2/2v^2}\\
+\;& \frac{4}{D}c(2c-1)\theta^\top x_q x_q^\top\theta\left(2\exp(c^2\|x_q\|^2/v^2)-2\right)+\frac{4}{D}c(2c-1)\theta^\top x_q x_q^\top\theta\left(2\exp((2c-1)^2\|x_q\|^2/v^2)-2\right)\\
-\;& \frac{4v^2}{D}\theta^\top\theta\exp((2c-1)c\|x_q\|^2/v^2)-\frac{4}{D}c(2c-1)(\theta^\top x_q x_q^\top\theta)\exp((2c-1)c\|x_q\|^2/v^2)\\
+\;& \frac{4c^2}{D}(\theta^\top x_q x_q^\top\theta)\exp(c^2\|x_q\|^2/v^2)+\frac{1}{D}(2c-1)^2(\theta^\top x_q x_q^\top\theta)\exp((2c-1)^2\|x_q\|^2/v^2),
\end{aligned}
$$

To sum up, we have

$$
\begin{aligned}
&\mathbb{E}\left(y_q-f(E)_{d+1,D+1}\right)^2\\
=\;& \frac{4}{D}v^2\|\theta\|^2\exp(c^2\|x_q\|^2/v^2)+\frac{1}{D}v^2\|\theta\|^2\exp\left((2c-1)^2\|x_q\|^2/v^2\right)\\
+\;& \frac{4}{D}c^2\theta^\top x_q x_q^\top\theta\exp(c^2\|x_q\|^2/v^2)+\frac{1}{D}(2c-1)^2\theta^\top x_q x_q^\top\theta\exp\left((2c-1)^2\|x_q\|^2/v^2\right)\\
+\;& 4v^2\|\theta\|^2\exp((2c-1)c\|x_q\|^2/v^2)-\frac{4}{D}c(2c-1)(\theta^\top x_q x_q^\top\theta)\exp((2c-1)c\|x_q\|^2/v^2)\\
=\;& \frac{4v^2(\sigma^2+\|\theta_0\|^2)}{D}\left((\frac{v^2}{v^2-2c^2})^{\frac{d}{2}}-(\frac{v^2}{v^2-2c(2c-1)})^{\frac{d}{2}}\right)+\frac{v^2}{D}(\sigma^2+\|\theta_0\|^2)(\frac{v^2}{v^2-2(2c-1)^2})^{\frac{d}{2}}\\
+\;& \frac{(2c-1)^2}{D}(\sigma^2+\|\theta_0\|^2)(\frac{v^2}{v^2-2(2c-1)^2})(\frac{v^2}{v^2-2(2c-1)^2})^{\frac{d}{2}}\\
+\;& \frac{4c^2}{D}(\sigma^2+\|\theta_0\|^2)(\frac{v^2}{v^2-2c^2})(\frac{v^2}{v^2-2c^2})^{\frac{d}{2}}-\frac{4(2c-1)c}{D}(\sigma^2+\|\theta_0\|^2)(\frac{v^2}{v^2-2c(2c-1)})(\frac{v^2}{v^2-2c(2c-1)})^{\frac{d}{2}}.
\end{aligned}
$$

Figure 15 below demonstrates the alignment between the theoretical values and the corresponding simulation results. □

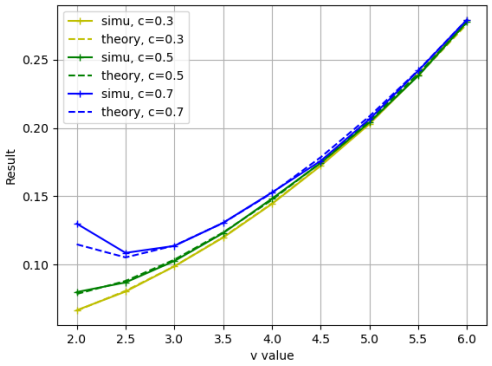


Figure 15: ICL performance of multi-head attention with prior knowledge $(A_1, A_2, b_1, b_2) = ((c/v)I_d, ((2c-1)/v)I_d, 0, 0)$, $(m, n) = (2, 1)$, and $(d, D) = (5, 1000)$.

C.5 Local Examples: Theorem 5.2 and 5.3

C.5.1 Theorem 5.2

Proof of Theorem 5.2. The proof of Theorem 5.2 is almost the same as Theorem 4.1. The only difference is the change on the distribution of the examples (x_i, y_i)s.

When taking infinite many training prompts, the loss function becomes

$$\begin{aligned}
&\mathbb{E}\left(y_q - (W^V_{d+1,:})^\top E\phi\left(E^\top (W^K)^\top W^Q \begin{bmatrix} x_q \\ 0 \end{bmatrix}\right)\right)^2 \\
= \;& \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(y_q^2 \underbrace{-2y_q\left(\frac{v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top A x_q + y_i b^\top x_q)}{\sum \exp(x_i^\top A x_q + y_i b^\top x_q) + \exp(x_q^\top A x_q)}\right)}_{=A_1}\right) \\
& + \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}} \underbrace{\left(\frac{v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top A x_q + y_i b^\top x_q)}{\sum \exp(x_i^\top A x_q + y_i b^\top x_q) + \exp(x_q^\top A x_q)}\right)^2}_{=A_2}.
\end{aligned}$$

For A_1, we have

$$\begin{aligned}
& \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}} A_1 \\
= \;& \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}} \frac{(-2v\theta^\top x_q)\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top A x_q + y_i b^\top x_q)}{D\mathbb{E}\exp({x_1}^\top A x_q + y_1 b^\top x_q)} + o(\frac{1}{D}) \\
& + \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}} \frac{(2v\theta^\top x_q)\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top A x_q + y_i b^\top x_q)}{(D\mathbb{E}\exp({x_1}^\top A x_q + y_1 b^\top x_q))^2} \\
& \qquad \times \left(\exp(x_q^\top A x_q) + \sum \exp(x_i^\top A x_q + y_i b^\top x_q) - D\mathbb{E}\exp({x_1}^\top A x_q + y_1 b^\top x_q)\right) \\
& - \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}} \frac{(2v\theta^\top x_q)\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top A x_q + y_i b^\top x_q)}{D\mathbb{E}\exp({x_1}^\top A x_q + y_1 b^\top x_q))^3} \\
& \qquad \times \left(\sum \exp(x_i^\top A x_q + y_i b^\top x_q) - D\mathbb{E}\exp({x_1}^\top A x_q + y_1 b^\top x_q)\right)^2 \\
& + o\left(\frac{1}{D}\right) \\
= \;& A_{11} + A_{12} + A_{13} + o(\frac{1}{D}).
\end{aligned}$$

Since $x_i \sim N(x_q, \sigma_x^2)$, we have

$$\begin{aligned}
\mathbb{E}_{\{x_1,y_1\}} \exp(x_1^\top A x_q + y_1 b^\top x_q) &= \mathbb{E}_{\{x_1,y_1\}} \exp\left(\left(\sigma_x \frac{x_1 - x_q}{\sigma_x}\right)^\top (A + \theta b^\top) x_q\right) \\
&= \exp\left(\frac{1}{2}\sigma_x^2 x_q^\top (A+\theta b^\top)^\top (A+\theta b^\top) x_q + x_q^\top (A+\theta b^\top) x_q\right), \\
\mathbb{E}_{\{x_1,y_1\}} x_1 \exp(x^\top A x_q + y_1 b^\top x_q) &= \left[\sigma_x^2 (A+\theta b^\top) x_q + x_q\right] \exp(x_q^\top (A+\theta b^\top)^\top (A+\theta b^\top) x_q/2 + x_q^\top (A+\theta b^\top) x_q).
\end{aligned}$$

Therefore,

$$\begin{aligned}
A_{11} &= \mathbb{E}_{\{x_1,y_1\}} \left(-\frac{D(2v\theta^\top x_q)\mathbb{E}\theta^\top x_1 \exp(x_1^\top A x_q + y_1 b^\top x_q)}{D\mathbb{E}_{x_1} \exp(x_1^\top A x_q + y_1 b^\top x_q)}\right) \\
&= -(2v\theta^\top x_q)\theta^\top \left[\sigma_x^2 (A+\theta b^\top) x_q + x_q\right] + o(\frac{1}{D}).
\end{aligned}$$

$$\begin{aligned}
A_{12} &= \mathbb{E}_{\{x_i,y_i\}_{i\in[D]}} \frac{(2v\theta^\top x_q \sum_{i=1}^D \theta^\top x_i \exp(x_i^\top A x_q + y_i b^\top x_q))(\exp(x_q^\top A x_q) + \sum \exp(x_i^\top A x_q + y_i b^\top x_q))}{(D\mathbb{E}\exp({x_1}^\top A x_q + y_1 b^\top x_q))^2} \\
&\quad -\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}} \frac{(2v\theta^\top x_q \sum_{i=1}^D \theta^\top x_i \exp(x_i^\top A x_q + y_i b^\top x_q))(D\mathbb{E}\exp({x_1}^\top A x_q + y_1 b^\top x_q))}{(D\mathbb{E}\exp({x_1}^\top A x_q + y_1 b^\top x_q))^2} \\
&= \frac{2v(\theta^\top y_q)}{D}\theta^\top \left[2\sigma_x^2 (A+\theta b^\top) x_q + x_q\right] \exp\left(x_q^\top (A+\theta b^\top)^\top (A+\theta b^\top) x_q\right) \\
&\quad -\frac{2v}{D}(\theta^\top x_q)\theta^\top \left[\sigma_x^2 (A+\theta b^\top) x_q + x_q\right] + \frac{2v(\theta^\top \left[\sigma_x^2 (A+\theta b^\top) x_q + x_q\right])^2 \exp\left(x_q^\top A x_q\right)}{D\exp(x_q^\top (A+\theta b^\top)^\top (A+\theta b^\top) x_q)}.
\end{aligned}$$

$$A_{13} = \frac{2v}{D}(\theta^\top x_q)\theta^\top \left[\sigma_x^2 (A+\theta b^\top) x_q + x_q\right] - \frac{2v}{D}(\theta^\top x_q)\theta^\top \left[\sigma_x^2 (A+\theta b^\top) x_q + x_q\right] \exp(x_q^\top (A+\theta b^\top)^\top (A+\theta b^\top) x_q).$$

To sum up, we have

$$\begin{aligned}
A_1 &= A_{11} + A_{12} + A_{13} \\
&= -2(v\theta^\top x_q)\theta^\top \left[\sigma_x^2 (A+\theta b^\top) x_q + x_q\right] + \frac{(2v\theta^\top x_q)\theta^\top \left[\sigma_x^2 (A+\theta b^\top) x_q + x_q\right] \exp(x_q^\top A x_q)}{D\exp(x_q^\top (A+\theta b^\top)^\top (A+\theta b^\top) x_q/2)} \\
&\quad +\frac{2v}{D}(\theta^\top x_q)\theta^\top \left[\sigma_x^2 (A+\theta b^\top) x_q\right] \exp(x_q^\top (A+\theta b^\top)^\top (A+\theta b^\top) x_q).
\end{aligned}$$

In terms of A_2, we have

$$\begin{aligned}
&\mathbb{E} x_i x_i^\top \exp(2x_i^\top (A+\theta b^\top) x_q) \\
=\ &\mathbb{E}\left(\frac{x_i - x_q}{\sigma_x}\right)\left(\frac{x_i - x_q}{\sigma_x}\right)^\top \sigma_x^2 \exp\left(2\sigma_x \frac{(x_i - x_q)^\top (A+\theta b^\top) x_q}{\sigma_x} + 2x_q^\top (A+\theta b^\top) x_q\right) \\
&+x_q x_q^\top \exp(\frac{\sigma_x^2}{2v^2}\|x_q\|^2 + 2\|x_q\|^2/v) \\
=\ &\left(\sigma_x^2 \left(I_d + 4\sigma_x^2 (A+\theta b^\top) x_q x_q^\top (A+\theta b^\top)^\top\right) + x_q x_q^\top\right) \exp\left(2\sigma_x \frac{(x_i - x_q)^\top (A+\theta b^\top) x_q}{\sigma_x} + 2x_q^\top (A+\theta b^\top) x_q\right).
\end{aligned}$$

since x_is are independent with each other, we have

$$\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i\in[D]}} A_2 \\
=\ &\mathbb{E}_{\{x_i\}_{i\in[D]}} \left(\frac{v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top A x_q + y_i b^\top x_q)}{D\mathbb{E}_{x_1} \exp(x_1^\top A x_q + y_1 b^\top x_q)}\right)^2
\end{aligned}$$

$$
\begin{aligned}
&-2\mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{\left(v\sum_{i=1}^{D}\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)\right)^2}{\left(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q)\right)^3}\left(\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)-(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))\right)\\
&-2\mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{\left(v\sum_{i=1}^{D}\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)\right)^2}{\left(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q)\right)^3}\exp(x_q^\top Ax_q)\\
&+3\mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{\left(v\sum_{i=1}^{D}\theta^\top x_i\exp(x_i^\top Ax_q+y_ib^\top x_q)\right)^2}{\left(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q)\right)^4}\left(\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)-(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))\right)^2\\
=\;&\frac{v^2D\mathbb{E}(\theta^\top x_i)^2\exp(2x_i^\top Ax_q+2y_ib^\top x_q)+D(D-1)v^2\mathbb{E}^2(\theta^\top x_i)\exp(x_i^\top Ax_q+y_ib^\top x_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))^2}\\
&-\frac{4D(D-1)v^2\mathbb{E}(\theta^\top x_i)\exp(x_i^\top Ax_q+y_ib^\top x_q)\mathbb{E}(\theta^\top x_i)\exp(2x_i^\top Ax_q+2y_ib^\top x_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))^3}\\
&+\frac{4D(D-1)v^2\mathbb{E}(\theta^\top x_i)\exp(x_i^\top Ax_q+y_ib^\top x_q)^2\mathbb{E}\exp(x_i^\top Ax_q+y_ib^\top x_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))^3}\\
&-\frac{D(D-1)v^2\mathbb{E}(\theta^\top x_i)\exp(x_i^\top Ax_q+y_ib^\top x_q)^2\exp(x_q^\top Ax_q)}{(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))^3}\\
&+\frac{3D^2(D-1)v^2\mathbb{E}^2(\theta^\top x_i)\exp(x_i^\top Ax_q+y_ib^\top x_q)\left[\mathbb{E}\exp(2x_i^\top Ax_q+2y_ib^\top x_q)-\mathbb{E}^2\exp(x_i^\top Ax_q+y_ib^\top x_q)\right]}{(D\mathbb{E}_{x_1}\exp(x_1^\top Ax_q+y_1b^\top x_q))^4}\\
&+o\left(\frac{1}{D}\right)\\
=\;&\frac{v^2}{D}\theta^\top(\sigma_x^2I_d+4\sigma_x^4(A+\theta b^\top)x_qx_q^\top(A+\theta b^\top)^\top+x_qx_q^\top)\theta\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)\\
&+v^2(\theta^\top\left[\sigma_x^2(A+\theta b^\top)x_q+x_q\right])^2\\
&+\frac{v^2}{D}\theta^\top(-4(2\sigma_x^2(A+\theta b^\top)x_q+x_q)(\sigma_x^2(A+\theta b^\top)x_q+x_q)^\top+3(\sigma_x^2(A+\theta b^\top)x_q+x_q)(\sigma_x^2(A+\theta b^\top)x_q+x_q)^\top)\theta\\
&\qquad\times\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)\\
&-\frac{2v^2(\theta^\top\left[\sigma_x^2(A+\theta b^\top)x_q+x_q\right])^2\exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)}+o(\frac{1}{D})\\
=\;&\frac{v^2}{D}\theta^\top(\sigma_x^2I_d-\sigma_x^4(A+\theta b^\top)x_qx_q^\top(A+\theta b^\top)^\top)\theta\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)\\
&+v^2(\theta^\top\left[\sigma_x^2(A+\theta b^\top)x_q+x_q\right])^2-\frac{2v^2(\theta^\top\left[\sigma_x^2(A+\theta b^\top)x_q+x_q\right])^2\exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)}+o(\frac{1}{D})
\end{aligned}
$$

Based on the results of A_1 and A_2, we have

$$
\begin{aligned}
&\mathbb{E}\left(y_q-(W_{d+1,:}^V)^\top E\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2\\
=\;&\mathbb{E}_{(x_q,\theta)}\left[(x_q^\top\theta)^2+A_1+A_2\right]\\
=\;&\mathbb{E}_{(x_q,\theta)}\left[(x_q^\top\theta)^2-2(v\theta^\top x_q)\theta^\top\left[\sigma_x^2(A+\theta b^\top)x_q+x_q\right]+\frac{(2v\theta^\top x_q)\theta^\top\left[\sigma_x^2(A+\theta b^\top)x_q+x_q\right]\exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)}\right.\\
&+\frac{2v}{D}(\theta^\top x_q)\theta^\top\left[\sigma_x^2(A+\theta b^\top)x_q\right]\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)\\
&+\frac{v^2}{D}\theta^\top(\sigma_x^2I_d-\sigma_x^4(A+\theta b^\top)x_qx_q^\top(A+\theta b^\top)^\top)\theta\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)\\
&\left.+v^2(\theta^\top\left[\sigma_x^2(A+\theta b^\top)x_q+x_q\right])^2-\frac{2v^2(\theta^\top\left[\sigma_x^2(A+\theta b^\top)x_q+x_q\right])^2\exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q)}\right]+o\left(\frac{1}{D}\right).
\end{aligned}
$$

From the above formulation, one can see that the optimal solution satisfies

$$v[\sigma_x^2(A+\theta b^\top)+I_d]\approx I_d.$$

Taking $v = 1$, $A = 0_{d\times d}$, $b = \mathbf{0}$, we have

$$\mathbb{E}\left(y_q - (W_{d+1,:}^V)^\top E\phi\left(E^\top (W^K)^\top W^Q \begin{bmatrix} x_q \\ 0 \end{bmatrix}\right)\right)^2 = O\left(\frac{\sigma_x^2}{D}\right) + o\left(\frac{1}{D}\right).$$

□

C.5.2 Theorem 5.3

Proof of Theorem 5.3. The proof of Theorem 5.3 is done by substituting the optimal solution of single-head attention derived in Theorem 4.1 to the prediction risk and follow the proof of Theorem 4.1 to calculate the expectation.

Recall that for single-head attention, we take $A = I_d/v$ and $b = 0$. Following the proof of Theorem 4.1, the prediction risk becomes

$$\begin{aligned}
&\mathbb{E}\left(y_q - (W_{d+1,:}^V)^\top E\phi\left(E^\top (W^K)^\top W^Q \begin{bmatrix} x_q \\ 0 \end{bmatrix}\right)\right)^2 \\
= \;& \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(y_q^2 - 2y_q\left(\frac{v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top A x_q + y_i b^\top x_q)}{\sum \exp(x_i^\top A x_q + y_i b^\top x_q) + \exp(x_q^\top A x_q)}\right)\right. \\
&\left. + \left(\frac{v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top A x_q + y_i b^\top x_q)}{\sum \exp(x_i^\top A x_q + y_i b^\top x_q) + \exp(x_q^\top A x_q)}\right)^2\right) \\
= \;& \mathbb{E}_{(x_q,\theta)}\mathbb{E}_{\{x_i\}_{i\in[D]}}\Bigg(\underbrace{y_q^2 - 2y_q\left(\frac{v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top x_q/v)}{\sum \exp(x_i^\top x_q/v) + \exp(x_q^\top x_q/v)}\right)}_{=A_1} + \underbrace{\left(\frac{v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top x_q/v)}{\sum \exp(x_i^\top x_q/v) + \exp(x_q^\top x_q/v)}\right)^2}_{=A_2}\Bigg).
\end{aligned}$$

Recall that in the testing stage, $x_i \sim N(x_q, \sigma_x^2)$. In this case,

$$\mathbb{E}\exp(x_i^\top x_q/v) = \mathbb{E}\exp\left(\frac{(x_i - x_q)^\top x_q}{\sigma_x v}\sigma_x + \|x_q\|^2/v\right) = \exp\left(\frac{\sigma_x^2}{2v^2}\|x_q\|^2 + \|x_q\|^2/v\right),$$

and

$$\mathbb{E}x_i \exp(x_i^\top x_q/v) = \frac{\sigma_x^2}{v}x_q \exp\left(\frac{\sigma_x^2}{2v^2}\|x_q\|^2 + \|x_q\|^2/v\right) + x_q \exp\left(\frac{\sigma_x^2}{2v^2}\|x_q\|^2 + \|x_q\|^2/v\right).$$

Consequently, fixing x_q and θ,

$$\begin{aligned}
\mathbb{E}A_1 &= -2y_q\mathbb{E}\left(\frac{v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top x_q/v)}{\sum \exp(x_i^\top x_q/v) + \exp(x_q^\top x_q/v)}\right) \\
&= -2y_q\mathbb{E}\left(\frac{v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top x_q/v)}{\mathbb{E}\sum \exp(x_i^\top x_q/v) + \exp(x_q^\top x_q/v)}\right) \\
&\quad + 2y_q\mathbb{E}\left(\frac{v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top x_q/v)}{(\mathbb{E}\sum \exp(x_i^\top x_q/v) + \exp(x_q^\top x_q/v))^2}\right)\left(\sum \exp(x_i^\top x_q/v) - \mathbb{E}\sum \exp(x_i^\top x_q/v)\right) \\
&\quad - 2y_q\mathbb{E}\left(\frac{v\sum_{i=1}^D \theta^\top x_i \exp(x_i^\top x_q/v)}{(\mathbb{E}\sum \exp(x_i^\top x_q/v) + \exp(x_q^\top x_q/v))^3}\right)\left(\sum \exp(x_i^\top x_q/v) - \mathbb{E}\sum \exp(x_i^\top x_q/v)\right)^2 + o\left(\frac{1}{D}\right) \\
&:= A_{11} + A_{12} + A_{13},
\end{aligned}$$

where

$$A_{11} = -2y_q^2\left(\sigma_x^2 + v\right)\frac{D\exp(\sigma_x^2\|x_q\|^2/(2v^2))}{D\exp(\sigma_x^2\|x_q\|^2/(2v^2)) + 1} = -2y_q^2(\sigma_x^2 + v) + O\left(\frac{1}{D}\right),$$

$$A_{12} = 2y_q^2\left(2\sigma_x^2\exp\left(\frac{2\sigma_x^2}{v^2}\|x_q\|^2\right) - \sigma_x^2\exp\left(\frac{\sigma_x^2}{v^2}\|x_q\|^2\right)\right)\frac{D}{(D\exp(\sigma_x^2\|x_q\|^2/(2v^2))+1)^2} = O\left(\frac{1}{D}\right),$$

and

$$A_{13} = -2y_q^2(\sigma_x^2+v)\frac{D\exp(\sigma_x^2\|x_q\|^2/(2v^2))}{(D\exp(\sigma_x^2\|x_q\|^2/(2v^2))+1)^3}D\left(\exp\left(\frac{2\sigma_x^2}{v^2}\|x_q\|^2\right) - \exp\left(\frac{\sigma_x^2}{v^2}\|x_q\|^2\right)\right) + o\left(\frac{1}{D}\right) = O\left(\frac{1}{D}\right).$$

For A_2, when fixing x_q and θ, we have

$$\begin{aligned}
A_2 &= \mathbb{E}\left(\frac{v\sum_{i=1}^D \theta^\top x_i\exp(x_i^\top x_q/v)}{\sum\exp(x_i^\top x_q/v)+\exp(x_q^\top x_q/v)}\right)^2 = \mathbb{E}\left(\frac{v\sum_{i=1}^D \theta^\top x_i\exp(x_i^\top x_q/v)}{\mathbb{E}\sum\exp(x_i^\top x_q/v)+\exp(x_q^\top x_q/v)}\right)^2 + O\left(\frac{1}{D}\right)\\
&= \frac{D(D-1)}{(D\exp(\sigma_x^2\|x_q\|^2/(2v^2))+1)^2}\left[(\sigma_x^2+v)x_q^\top\theta\right]^2\exp\left(\frac{\sigma_x^2}{v^2}\|x_q\|^2\right) + O\left(\frac{1}{D}\right)\\
&= (\sigma_x^2+v)^2y_q^2 + +O\left(\frac{1}{D}\right).
\end{aligned}$$

To conclude, when fixing x_q and θ, we obtain

$$\mathbb{E}\left(y_q - (W_{d+1,:}^V)^\top E\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2 = (\sigma_x^2+v-1)^2y_q^2 + O\left(\frac{1}{D}\right).$$

□

C.6 Noisy Response: Theorem A.1

Proof of Theorem A.1. The main logic of the proof is the same as Theorem 4.1 for single-head attention and Theorem 4.2 for multi-head attention. The main difference is the change on the distribution of the labels y_is.

Optimal solution for single-head attention

$$\begin{aligned}
&\mathbb{E}\left(y_q - (W_{d+1,:}^V)^\top E\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2\\
=\ &\mathbb{E}_{(x_q,y_q)}\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\left(y_q - v\left[y_1,y_2,\ldots,y_D,0\right]\phi\left(\begin{bmatrix}x_1^\top Ax_q + y_1b^\top x_q\\ \cdots\\ x_q^\top Ax_q+0\end{bmatrix}\right)\right)^2\\
=\ &\mathbb{E}_{(x_q,y_q)}\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\Bigg(y_q^2 + \underbrace{\left(\frac{v\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)+\exp(x_q^\top Ax_q)}\right)^2}_{:=A_1}\Bigg)\\
&-\mathbb{E}_{(x_q,y_q)}\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\Bigg(\underbrace{2y_q\left(\frac{v\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)+\exp(x_q^\top Ax_q)}\right)}_{:=A_2}\Bigg),
\end{aligned}$$

where $\mathbb{E}y_q^2 = 1+\sigma_\epsilon^2$.

When fixing x_q and θ, the terms A_1 becomes

$$\begin{aligned}
&\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}A_1\\
=\ &v^2\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q))^2}{D^2\mathbb{E}^2\exp(x_i^\top Ax_q+y_ib^\top x_q)}\\
&-2v^2\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q))^2}{D^3\mathbb{E}^3\exp(x_i^\top Ax_q+y_ib^\top x_q)}\left[\sum_{i=1}^D\exp(x_i^\top Ax_q+y_ib^\top x_q) - D\mathbb{E}\exp(x_i^\top Ax_q+y_ib^\top x_q)\right]
\end{aligned}$$

$$
\begin{aligned}
&+3v^2\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q))^2}{D^4\mathbb{E}^4\exp(x_i^\top Ax_q+y_ib^\top x_q)}\left[\sum_{i=1}^D\exp(x_i^\top Ax_q+y_ib^\top x_q)-D\mathbb{E}\exp(x_i^\top Ax_q+y_ib^\top x_q)\right]^2\\
&-2v^2\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q))^2}{D^3\mathbb{E}^3\exp(x_i^\top Ax_q+y_ib^\top x_q)}\exp(x_q^\top Ax_q)+o\left(\frac{1}{D}\right)\\
:=\;& A_{11}+A_{12}+A_{13}+A_{14}+o\left(\frac{1}{D}\right).
\end{aligned}
$$

To figure out A_{11} to A_{13}, we know that

$$
\begin{aligned}
&\mathbb{E}y_i^2\exp(x_i^\top Ax_q+y_ib^\top x_q)^2\\
=\;&\mathbb{E}(x_i^\top\theta+\epsilon_i)^2\exp(2x_i^\top Ax_q+2x_i^\top\theta b^\top x_q+2\epsilon_ib^\top x_q)\\
=\;&\theta^\top\left(I+4(Ax_q+\theta b^\top x_q)(Ax_q+\theta b^\top x_q)^\top\right)\theta\exp(2\|Ax_q+\theta b^\top x_q\|^2+2\sigma_\epsilon^2(b^\top x_q)^2)\\
&+8\sigma_\epsilon(\theta^\top Ax_q)b^\top x_q\exp(2\|Ax_q+\theta b^\top x_q\|^2+2\sigma_\epsilon^2(b^\top x_q)^2)\\
&+\sigma_\epsilon^2(1+4\sigma_\epsilon^2(b^\top x_q)^2)\exp(2\|Ax_q+\theta b^\top x_q\|^2+2\sigma_\epsilon^2(b^\top x_q)^2),
\end{aligned}
$$

and

$$
\mathbb{E}y_i\exp(x_i^\top Ax_q+y_ib^\top x_q)=(\theta^\top Ax_q+\|\theta\|^2b^\top x_q+\sigma_\epsilon(b^\top x_q))\exp(\|Ax_q+\theta b^\top x_q\|^2/2+\sigma_\epsilon^2(b^\top x_q)^2/2).
$$

As a result,

$$
\begin{aligned}
A_{11}\;=\;& v^2\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q))^2}{D^2\mathbb{E}^2\exp(x_i^\top Ax_q+y_ib^\top x_q)}\\
=\;& v^2\frac{D\mathbb{E}y_i^2\exp(x_i^\top Ax_q+y_ib^\top x_q)^2+D(D-1)\mathbb{E}^2y_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{D^2\exp(\|Ax_q+\theta b^\top x_q\|^2+\sigma_\epsilon^2(b^\top x_q)^2)}\\
=\;& v^2\frac{1}{D}\left[\theta^\top\left(I+4(Ax_q+\theta b^\top x_q)(Ax_q+\theta b^\top x_q)^\top\right)\theta+8\sigma_\epsilon(\theta^\top Ax_q)b^\top x_q+\sigma_\epsilon^2(1+4\sigma_\epsilon^2(b^\top x_q)^2)\right]\\
&\qquad\times\exp(\|Ax_q+\theta b^\top x_q\|^2+\sigma_\epsilon^2(b^\top x_q)^2)\\
&+v^2\frac{D-1}{D}(\theta^\top Ax_q+\|\theta\|^2b^\top x_q+\sigma_\epsilon(b^\top x_q))^2,
\end{aligned}
$$

$$
\begin{aligned}
&A_{12}\\
=\;& -2v^2\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q))^2}{D^3\mathbb{E}^3\exp(x_i^\top Ax_q+y_ib^\top x_q)}\left[\sum_{i=1}^D\exp(x_i^\top Ax_q+y_ib^\top x_q)-D\mathbb{E}\exp(x_i^\top Ax_q+y_ib^\top x_q)\right]\\
=\;& -2v^2\mathbb{E}\frac{2D(D-1)}{D^3\mathbb{E}^3\exp(x_i^\top Ax_q+y_ib^\top x_q)}\left[\mathbb{E}y_i\exp(x_i^\top Ax_q+y_ib^\top x_q)\mathbb{E}y_i\exp(x_i^\top Ax_q+y_ib^\top x_q)^2\right]\\
&+2v^2\mathbb{E}\frac{2D(D-1)}{D^3\mathbb{E}^3\exp(x_i^\top Ax_q+y_ib^\top x_q)}\left[\mathbb{E}^2y_i\exp(x_i^\top Ax_q+y_ib^\top x_q)\mathbb{E}\exp(x_i^\top Ax_q+y_ib^\top x_q))\right]+o\left(\frac{1}{D}\right)\\
=\;& -4v^2\mathbb{E}\frac{2D(D-1)}{D^3\mathbb{E}^3\exp(x_i^\top Ax_q+y_ib^\top x_q)}(\theta^\top Ax_q+\|\theta\|^2b^\top x_q+\sigma_\epsilon(b^\top x_q))^2\exp(5\|Ax_q+\theta b^\top x_q\|^2/2+5\sigma_\epsilon^2(b^\top x_q)^2/2)\\
&+2v^2\mathbb{E}\frac{2D(D-1)}{D^3\mathbb{E}^3\exp(x_i^\top Ax_q+y_ib^\top x_q)}(\theta^\top Ax_q+\|\theta\|^2b^\top x_q+\sigma_\epsilon(b^\top x_q))^2\exp(3\|Ax_q+\theta b^\top x_q\|^2/2+3\sigma_\epsilon^2(b^\top x_q)^2/2)\\
=\;& -8v^2\frac{1}{D}(\theta^\top Ax_q+\|\theta\|^2b^\top x_q+\sigma_\epsilon(b^\top x_q))^2\exp(\|Ax_q+\theta b^\top x_q\|^2+\sigma_\epsilon^2(b^\top x_q)^2)\\
&+4v^2\frac{1}{D}(\theta^\top Ax_q+\|\theta\|^2b^\top x_q+\sigma_\epsilon(b^\top x_q))^2+o\left(\frac{1}{D}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
&A_{13}\\
=\;& 3v^2\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}\frac{(\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q))^2}{D^4\mathbb{E}^4\exp(x_i^\top Ax_q+y_ib^\top x_q)}\left[\sum_{i=1}^D\exp(x_i^\top Ax_q+y_ib^\top x_q)-D\mathbb{E}\exp(x_i^\top Ax_q+y_ib^\top x_q)\right]^2
\end{aligned}
$$

$$
\begin{aligned}
&= 3v^2\frac{\mathbb{E}^2 y_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{D\mathbb{E}^4\exp(x_i^\top Ax_q+y_ib^\top x_q)}\left[\mathbb{E}\exp(x_i^\top Ax_q+y_ib^\top x_q)^2-\mathbb{E}^2\exp(x_i^\top Ax_q+y_ib^\top x_q)\right]+o\left(\frac{1}{D}\right)\\
&= \frac{3v^2}{D}(\theta^\top Ax_q+\|\theta\|^2b^\top x_q+\sigma_\epsilon(b^\top x_q))^2\left(\exp(\|Ax_q+\theta b^\top x_q\|^2+\sigma_\epsilon^2(b^\top x_q)^2)-1\right)+o\left(\frac{1}{D}\right),
\end{aligned}
$$

with

$$
A_{14} = -2v^2\frac{1}{D}(\theta^\top Ax_q+\|\theta\|^2b^\top x_q+\sigma_\epsilon(b^\top x_q))^2\exp(x_q^\top Ax_q-\|Ax_q+\theta b^\top x_q\|^2/2-\sigma_\epsilon^2(b^\top x_q)^2/2).
$$

In terms of A_2, when fixing x_q and θ, we have

$$
\begin{aligned}
&\mathbb{E}_{y_q}\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}A_2\\
=\ &\mathbb{E}_{y_q}\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}2y_q\left(\frac{v\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{\sum\exp(x_i^\top Ax_q+y_ib^\top x_q)+\exp(x_q^\top Ax_q)}\right)\\
=\ &\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}2\theta^\top x_q\left(\frac{v\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{D\mathbb{E}\exp(x_i^\top Ax_q+y_ib^\top x_q)}\right)\\
&-\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}2\theta^\top x_q\left(\frac{v\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{D^2\mathbb{E}^2\exp(x_i^\top Ax_q+y_ib^\top x_q)}\right)\\
&\qquad\times\left(\sum_{i=1}^D\exp(x_i^\top Ax_q+y_ib^\top x_q)-D\mathbb{E}\exp(x_i^\top Ax_q+y_ib^\top x_q)\right)\\
&+\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}2\theta^\top x_q\left(\frac{v\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{D^3\mathbb{E}^3\exp(x_i^\top Ax_q+y_ib^\top x_q)}\right)\\
&\qquad\times\left(\sum_{i=1}^D\exp(x_i^\top Ax_q+y_ib^\top x_q)-D\mathbb{E}\exp(x_i^\top Ax_q+y_ib^\top x_q)\right)^2\\
&-\mathbb{E}_{\{x_i,y_i\}_{i\in[D]}}2\theta^\top x_q\left(\frac{v\sum_{i=1}^D y_i\exp(x_i^\top Ax_q+y_ib^\top x_q)}{D^2\mathbb{E}^2\exp(x_i^\top Ax_q+y_ib^\top x_q)}\right)\exp(x_q^\top Ax_q)+o\left(\frac{1}{D}\right)\\
:=\ &A_{21}+A_{22}+A_{23}+A_{24}.
\end{aligned}
$$

For A_{21} to A_{24}, we have

$$
A_{21} = 2v\theta^\top x_q(\theta^\top Ax_q+\|\theta\|^2b^\top x_q+\sigma_\epsilon(b^\top x_q)),
$$

$$
A_{22} = -2v\theta^\top x_q\frac{1}{D}(\theta^\top Ax_q+\|\theta\|^2b^\top x_q+\sigma_\epsilon(b^\top x_q))\left(2\exp(\|Ax_q+\theta b^\top x_q\|^2+\sigma_\epsilon^2(b^\top x_q)^2)-1\right)+o\left(\frac{1}{D}\right),
$$

$$
A_{23} = 2v\theta^\top x_q\frac{1}{D}(\theta^\top Ax_q+\|\theta\|^2b^\top x_q+\sigma_\epsilon(b^\top x_q))\left(\exp(\|Ax_q+\theta b^\top x_q\|^2+\sigma_\epsilon^2(b^\top x_q)^2)-1\right)+o\left(\frac{1}{D}\right),
$$

$$
A_{24} = -2v\theta^\top x_q\frac{1}{D}(\theta^\top Ax_q+\|\theta\|^2b^\top x_q+\sigma_\epsilon(b^\top x_q))\exp(x_q^\top Ax_q-\|Ax_q+\theta b^\top x_q\|^2/2-\sigma_\epsilon^2(b^\top x_q)^2/2).
$$

Inserting A_{11} to A_{24} into A_1 and A_2, we obtain

$$
\begin{aligned}
&\mathbb{E}\left(y_q-(W_{d+1,:}^V)^\top E\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2\\
=\ &\mathbb{E}\left(x_q^\top\theta-v(\theta^\top Ax_q+\|\theta\|^2b^\top x_q+\sigma_\epsilon(b^\top x_q))\right)^2+O\left(\frac{1}{D}\right)\\
=\ &\frac{1}{d}tr\left((I-vA)^2\right)+\mathbb{E}\|\theta\|^4\|b\|^2+\sigma_\epsilon^2\|b\|^2+O\left(\frac{1}{D}\right).
\end{aligned}
\tag{12}
$$

As a result, the optimal solution of A and b satisfies that $tr\left((I-vA)^2\right)/d = O(1/D)$ and $\|b\|^2 = O(1/D)$. In addition, similar to Theorem 4.1, when taking $A = I_d/v$ and $b = 0$, we have

$$\begin{aligned}
A_{11} &= \frac{1}{D}\mathbb{E}\left[v^2\|\theta\|^2 + 4(x_q^\top\theta)^2 + v^2\sigma_\epsilon^2\right]\exp(\|x_q\|^2/v^2) + \frac{D-1}{D},\\
A_{12} &= -\frac{8}{D}\mathbb{E}(x_q^\top\theta)^2\exp(\|x_q\|^2/v^2) + \frac{4}{D},\\
A_{13} &= -\frac{3}{D} + \frac{3}{D}\mathbb{E}(x_q^\top\theta)^2\exp(\|x_q\|^2/v^2),\\
A_{14} &= -\frac{2}{D}\mathbb{E}(x_q^\top\theta)^2\exp(\|x_q\|^2/v - \|x_q\|^2/v^2/2),\\
A_{21} &= 2,\\
A_{22} &= -\frac{2}{D}\mathbb{E}(x_q^\top\theta)^2(2\exp(\|x_q\|^2/v^2) - 1),\\
A_{23} &= \frac{2}{D}\mathbb{E}(x_q^\top\theta)^2(\exp(\|x_q\|^2/v^2) - 1),\\
A_{24} &= -\frac{2}{D}\mathbb{E}(x_q^\top\theta)^2\exp(\|x_q\|^2/v - \|x_q\|^2/v^2/2).
\end{aligned}$$

As a result,

$$\begin{aligned}
&\mathbb{E}\left(y_q - (W_{d+1,:}^V)^\top E\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2\\
=\ & 1 + \sigma_\epsilon^2 + A_{11} + A_{12} + A_{13} + A_{14} - A_{21} - A_{22} - A_{23} - A_{24}\\
=\ & \sigma_\epsilon^2 + \frac{v^2(1+\sigma_\epsilon^2)}{D}\mathbb{E}\exp(\|x_q\|^2/v^2) + \frac{1}{D}\mathbb{E}(x_q^\top\theta)^2\exp(\|x_q\|^2/v^2) + o(\frac{1}{D}).
\end{aligned}$$

ICL performance of multi-head attention

$$\begin{aligned}
&\mathbb{E}\left(y_q - f(E)_{d+1,D+1}\right)^2\\
=\ & \mathbb{E}\left(y_q - vm\left[y_1,y_2,\ldots,y_D,0\right]\phi\left(E^\top(W_1^K)^\top W_1^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right) + vn\left[y_1,y_2,\ldots,y_D,0\right]\phi\left(E^\top(W_2^K)^\top W_2^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2\\
=\ & \mathbb{E}\left(y_q - \frac{vm\sum_{i=1}^D y_i\exp(x_i^\top(A_1+\theta b_1^\top)x_q)}{\sum\exp(x_i^\top A_1x_q + y_ib_1^\top x_q) + \exp(x_q^\top A_1x_q)} + \frac{vn\sum_{i=1}^D y_i\exp(x_i^\top(A_2+\theta b_2^\top)x_q)}{\sum\exp(x_i^\top A_2x_q + y_ib_2^\top x_q) + \exp(x_q^\top A_2x_q)}\right)^2\\
=\ & \mathbb{E}\left(y_q^2 + \left(\frac{vm\sum_{i=1}^D y_i\exp(x_i^\top A_1x_q + y_ib_1^\top x_q)}{\sum\exp(x_i^\top A_1x_q + y_ib_1^\top x_q) + \exp(x_q^\top A_1x_q)}\right)^2 - 2y_q\left(\frac{vm\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top A_1x_q + y_ib_1^\top x_q)}{\sum\exp(x_i^\top A_1x_q + y_ib_1^\top x_q) + \exp(x_q^\top A_1x_q)}\right)\right)\\
+\ & \mathbb{E}\left(\left(\frac{vn\sum_{i=1}^D y_i\exp(x_i^\top A_2x_q + y_ib_2^\top x_q)}{\sum\exp(x_i^\top A_2x_q + y_ib_2^\top x_q) + \exp(x_q^\top A_2x_q)}\right)^2 + 2y_q\left(\frac{vn\sum_{i=1}^D y_i\exp(x_i^\top A_2x_q + y_ib_2^\top x_q)}{\sum\exp(x_i^\top A_2x_q + y_ib_2^\top x_q) + \exp(x_q^\top A_2x_q)}\right)\right)\\
-\ & \mathbb{E}\left(\frac{2vn\sum_{i=1}^D y_i\exp(x_i^\top A_1x_q + y_ib_1^\top x_q)}{\sum\exp(x_i^\top A_1x_q + y_ib_1^\top x_q) + \exp(x_q^\top A_1x_q)}\frac{vm\sum_{i=1}^D y_i\exp(x_i^\top A_2x_q + y_ib_2^\top x_q)}{\sum\exp(x_i^\top A_2x_q + y_ib_2^\top x_q) + \exp(x_q^\top A_2x_q)}\right).
\end{aligned}$$

When taking $m = 2$, $n = 1$, $A_1 = \frac{c}{v}I$, $A_2 = \frac{2c-1}{v}I$ and $b_1 = b_2 = 0$, it becomes

$$\begin{aligned}
&\mathbb{E}\left(y_q - f(E)_{d+1,D+1}\right)^2\\
=\ & 1 + \sigma_\epsilon^2 + \mathbb{E}\left(\left(\frac{2v\sum_{i=1}^D y_i\exp(x_i^\top x_q(c/v))}{\sum\exp(x_i^\top x_q(c/v)) + \exp(\|x_q\|^2(c/v))}\right)^2 - 2y_q\left(\frac{2v\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top x_q(c/v))}{\sum\exp(x_i^\top x_q(c/v)) + \exp(\|x_q\|^2(c/v))}\right)\right)\\
&+\mathbb{E}\Bigg(\left(\frac{v\sum_{i=1}^D y_i\exp(x_i^\top x_q(2c-1)/v)}{\sum\exp(x_i^\top x_q(2c-1)/v) + \exp(\|x_q\|^2(2c-1)/v)}\right)^2\\
&\qquad +2y_q\left(\frac{v\sum_{i=1}^D y_i\exp(x_i^\top x_q(2c-1)/v)}{\sum\exp(x_i^\top x_q(2c-1)/v) + \exp(\|x_q\|^2(2c-1)/v)}\right)\Bigg)\\
&-\mathbb{E}\left(\frac{2v\sum_{i=1}^D y_i\exp(x_i^\top x_q(c/v))}{\sum\exp(x_i^\top x_q(c/v)) + \exp(\|x_q\|^2(c/v))}\frac{2v\sum_{i=1}^D y_i\exp(x_i^\top x_q(2c-1)/v)}{\sum\exp(x_i^\top x_q(2c-1)/v) + \exp(\|x_q\|^2(2c-1)/v)}\right)
\end{aligned}$$

$$:= \quad 1+\sigma_\epsilon^2+B_1+B_2+B_3.$$

Similar to how we calculate A_1 and A_2, for B_1, the terms are similar. We follow the above proof and obtain

$$\begin{aligned}
A_{11} &= \frac{4}{D}\mathbb{E}\left[v^2\|\theta\|^2+4c^2(x_q^\top\theta)^2+v^2\sigma_\epsilon^2\right]\exp(c^2\|x_q\|^2/v^2)+\frac{D-1}{D}4c^2,\\
A_{12} &= -\frac{32c^2}{D}\mathbb{E}(x_q^\top\theta)^2\exp(c^2\|x_q\|^2/v^2)+\frac{16c^2}{D},\\
A_{13} &= -\frac{12c^2}{D}+\frac{12c^2}{D}\mathbb{E}(x_q^\top\theta)^2\exp(c^2\|x_q\|^2/v^2),\\
A_{14} &= -\frac{8c^2}{D}\mathbb{E}(x_q^\top\theta)^2\exp(c\|x_q\|^2/v-c^2\|x_q\|^2/v^2/2),\\
A_{21} &= 4c,\\
A_{22} &= -\frac{4c}{D}\mathbb{E}(x_q^\top\theta)^2(2\exp(c^2\|x_q\|^2/v^2)-1),\\
A_{23} &= \frac{4c}{D}\mathbb{E}(x_q^\top\theta)^2(\exp(c^2\|x_q\|^2/v^2)-1),\\
A_{24} &= -\frac{4c}{D}\mathbb{E}(x_q^\top\theta)^2\exp(c\|x_q\|^2/v-c^2\|x_q\|^2/v^2/2),
\end{aligned}$$

thus

$$\begin{aligned}
B_1 &= \mathbb{E}\left(\left(\frac{2v\sum_{i=1}^D y_i\exp(x_i^\top x_q(c/v))}{\sum\exp(x_i^\top x_q(c/v))+\exp(\|x_q\|^2(c/v))}\right)^2-2y_q\left(\frac{2v\sum_{i=1}^D\theta^\top x_i\exp(x_i^\top x_q(c/v))}{\sum\exp(x_i^\top x_q(c/v))+\exp(\|x_q\|^2(c/v))}\right)\right)\\
&= 4c^2-4c+\frac{4v^2(1+\sigma_\epsilon^2)}{D}\mathbb{E}\exp(c^2\|x_q\|^2/v^2)-\frac{4c^2-4c}{D}\mathbb{E}(x_q^\top\theta)^2\exp(c^2\|x_q\|^2/v^2)\\
&\quad -\frac{2(4c^2-2c)}{D}\mathbb{E}(x_q^\top\theta)^2\exp(c\|x_q\|^2/v-c^2\|x_q\|^2/v^2/2).
\end{aligned}$$

For B_2, similarly, we obtain

$$\begin{aligned}
B_2 &= (2c-1)^2+2(2c-1)+\frac{v^2(1+\sigma_\epsilon^2)}{D}\mathbb{E}\exp((2c-1)^2\|x_q\|^2/v^2)\\
&\quad -\frac{(2c-1)^2+2(2c-1)}{D}\mathbb{E}(x_q^\top\theta)^2\exp((2c-1)^2\|x_q\|^2/v^2)\\
&\quad -\frac{2((2c-1)^2+(2c-1))}{D}\mathbb{E}(x_q^\top\theta)^2\exp((2c-1)\|x_q\|^2/v-(2c-1)(2c-1)^2\|x_q\|^2/v^2/2).
\end{aligned}$$

In terms of B_3,

$$\begin{aligned}
&B_3\\
=\ & -\mathbb{E}\left(\frac{2v\sum_{i=1}^D y_i\exp(x_i^\top x_q(c/v))}{\sum\exp(x_i^\top x_q(c/v))+\exp(\|x_q\|^2(c/v))}\frac{2v\sum_{i=1}^D y_i\exp(x_i^\top x_q(2c-1)/v)}{\sum\exp(x_i^\top x_q(2c-1)/v)+\exp(\|x_q\|^2(2c-1)/v)}\right)\\
=\ & -4v^2\mathbb{E}\frac{\left(\sum_{i=1}^D y_i\exp(x_i^\top x_q(c/v))\right)\left(\sum_{i=1}^D y_i\exp(x_i^\top x_q(2c-1)/v)\right)}{D^2\mathbb{E}\exp(x_i^\top x_q(c/v))\mathbb{E}\exp(x_i^\top x_q(2c-1)/v)}\\
&+4v^2\mathbb{E}\frac{\left(\sum_{i=1}^D y_i\exp(x_i^\top x_q(c/v))\right)\left(\sum_{i=1}^D y_i\exp(x_i^\top x_q(2c-1)/v)\right)}{D^3\mathbb{E}^2\exp(x_i^\top x_q(c/v))\mathbb{E}\exp(x_i^\top x_q(2c-1)/v)}\left(\sum_{i=1}^D\exp(x_i^\top x_q(c/v))-D\mathbb{E}\exp(x_i^\top x_q(c/v))\right)\\
&+4v^2\mathbb{E}\frac{\left(\sum_{i=1}^D y_i\exp(x_i^\top x_q(c/v))\right)\left(\sum_{i=1}^D y_i\exp(x_i^\top x_q(2c-1)/v)\right)}{D^3\mathbb{E}\exp(x_i^\top x_q(c/v))\mathbb{E}^2\exp(x_i^\top x_q(2c-1)/v)}\\
&\qquad\times\left(\sum_{i=1}^D\exp(x_i^\top x_q(2c-1)/v)-D\mathbb{E}\exp(x_i^\top x_q(2c-1)/v)\right)\\
&-4v^2\mathbb{E}\frac{\left(\sum_{i=1}^D y_i\exp(x_i^\top x_q(c/v))\right)\left(\sum_{i=1}^D y_i\exp(x_i^\top x_q(2c-1)/v)\right)}{D^4\mathbb{E}^3\exp(x_i^\top x_q(c/v))\mathbb{E}\exp(x_i^\top x_q(2c-1)/v)}\left(\sum_{i=1}^D\exp(x_i^\top x_q(c/v))-D\mathbb{E}\exp(x_i^\top x_q(c/v))\right)^2\\
&-4v^2\mathbb{E}\frac{\left(\sum_{i=1}^D y_i\exp(x_i^\top x_q(c/v))\right)\left(\sum_{i=1}^D y_i\exp(x_i^\top x_q(2c-1)/v)\right)}{D^4\mathbb{E}\exp(x_i^\top x_q(c/v))\mathbb{E}^3\exp(x_i^\top x_q(2c-1)/v)}
\end{aligned}$$

$$
\begin{aligned}
&\times \left(\sum_{i=1}^{D} \exp(x_i^\top x_q(2c-1)/v) - D\mathbb{E}\exp(x_i^\top x_q(2c-1)/v)\right)^2 \\
&-4v^2\mathbb{E}\frac{\left(\sum_{i=1}^{D} y_i \exp(x_i^\top x_q(c/v))\right)\left(\sum_{i=1}^{D} y_i \exp(x_i^\top x_q(2c-1)/v)\right)}{D^4\mathbb{E}^2\exp(x_i^\top x_q(c/v))\mathbb{E}^2\exp(x_i^\top x_q(2c-1)/v)} \\
&\times \left(\sum_{i=1}^{D} \exp(x_i^\top x_q(c/v)) - D\mathbb{E}\exp(x_i^\top x_q(c/v))\right)\left(\sum_{i=1}^{D} \exp(x_i^\top x_q(2c-1)/v) - D\mathbb{E}\exp(x_i^\top x_q(2c-1)/v)\right) \\
&+4v^2\mathbb{E}\frac{\left(\sum_{i=1}^{D} y_i \exp(x_i^\top x_q(c/v))\right)\left(\sum_{i=1}^{D} y_i \exp(x_i^\top x_q(2c-1)/v)\right)}{D^3\mathbb{E}^2\exp(x_i^\top x_q(c/v))\mathbb{E}\exp(x_i^\top x_q(2c-1)/v)}\exp(\|x_q\|^2 c/v) \\
&+4v^2\mathbb{E}\frac{\left(\sum_{i=1}^{D} y_i \exp(x_i^\top x_q(c/v))\right)\left(\sum_{i=1}^{D} y_i \exp(x_i^\top x_q(2c-1)/v)\right)}{D^3\mathbb{E}\exp(x_i^\top x_q(c/v))\mathbb{E}^2\exp(x_i^\top x_q(2c-1)/v)}\exp(\|x_q\|^2(2c-1)/v) + o\left(\frac{1}{D}\right) \\
:= \quad & \mathbb{E}(B_{31}+B_{32}+B_{33}+B_{34}+B_{35}+B_{36}+B_{37}+B_{38}) + o\left(\frac{1}{D}\right).
\end{aligned}
$$

For B_{31} to B_{38}, we have

$$
\begin{aligned}
B_{31} &= -4c(2c-1)\frac{D-1}{D}(x_q^\top\theta)^2 - 4v^2\frac{1}{D}\frac{\left[\theta^\top(I_d+(3c-1)^2x_qx_q^\top/v^2)\theta+\sigma_\epsilon^2\right]\exp((3c-1)^2\|x_q\|^2/(2v^2))}{\exp((c^2+(2c-1)^2)\|x_q\|^2/(2v^2))} \\
&= -4c(2c-1)\frac{D-1}{D}(x_q^\top\theta)^2 - 4v^2\frac{1}{D}\left[\theta^\top(I_d+(3c-1)^2x_qx_q^\top/v^2)\theta+\sigma_\epsilon^2\right]\exp((2c^2-c)\|x_q\|^2/v^2),
\end{aligned}
$$

$$
\begin{aligned}
B_{32} &= 4\frac{1}{D}(x_q^\top\theta)^2\frac{2c(2c-1)\exp((4c^2+(2c-1)^2)\|x_q\|^2/(2v^2)) - c(2c-1)\exp((2c^2+(2c-1)^2)\|x_q\|^2/(2v^2))}{\exp((2c^2+(2c-1)^2)\|x_q\|^2/(2v^2))} \\
&\quad +4\frac{1}{D}(x_q^\top\theta)^2\frac{c(3c-1)\exp((c^2+(3c-1)^2)\|x_q\|^2/(2v^2)) - c(2c-1)\exp((c^2+2(2c-1)^2)\|x_q\|^2/(2v^2))}{\exp((2c^2+(2c-1)^2)\|x_q\|^2/(2v^2))} \\
&\quad +o\left(\frac{1}{D}\right) \\
&= \frac{4}{D}(x_q^\top\theta)^2[2c(2c-1)]\exp(\|x_q\|^2c^2/v^2) - \frac{8}{D}(x_q^\top\theta)^2[c(2c-1)] \\
&\quad +\frac{4}{D}(x_q^\top\theta)^2[c(3c-1)]\exp((2c^2-c)\|x_q\|^2/v^2) + o\left(\frac{1}{D}\right),
\end{aligned}
$$

$$
\begin{aligned}
B_{33} &= 4\frac{1}{D}(x_q^\top\theta)^2\frac{2c(2c-1)\exp((c^2+4(2c-1)^2)\|x_q\|^2/(2v^2)) - c(2c-1)\exp((c^2+2(2c-1)^2)\|x_q\|^2/(2v^2))}{\exp((c^2+2(2c-1)^2)\|x_q\|^2/(2v^2))} \\
&\quad +4\frac{1}{D}(x_q^\top\theta)^2\frac{(3c-1)(2c-1)\exp(((3c-1)^2+(2c-1)^2)\|x_q\|^2/(2v^2))}{\exp((c^2+2(2c-1)^2)\|x_q\|^2/(2v^2))} \\
&\quad +4\frac{1}{D}(x_q^\top\theta)^2\frac{-c(2c-1)\exp((c^2+2(2c-1)^2)\|x_q\|^2/(2v^2))}{\exp((c^2+2(2c-1)^2)\|x_q\|^2/(2v^2))} + o\left(\frac{1}{D}\right) \\
&= \frac{4}{D}(x_q^\top\theta)^2[2c(2c-1)]\exp(\|x_q\|^2(2c-1)^2/v^2) - \frac{8v^2}{D}(x_q^\top\theta)^2[c(2c-1)] \\
&\quad +\frac{4}{D}(x_q^\top\theta)^2[(3c-1)(2c-1)]\exp((2c^2-c)\|x_q\|^2/v^2) + o\left(\frac{1}{D}\right),
\end{aligned}
$$

$$
B_{34} = -4\frac{c(2c-1)}{D}(x_q^\top\theta)^2\left(\exp(c^2\|x_q\|^2/v^2)-1\right) + o\left(\frac{1}{D}\right),
$$

$$
B_{35} = -4\frac{c(2c-1)}{D}(x_q^\top\theta)^2\left(\exp((2c-1)^2\|x_q\|^2/v^2)-1\right) + o\left(\frac{1}{D}\right),
$$

$$
B_{36} = -4\frac{c(2c-1)}{D}(x_q^\top\theta)^2\left(\frac{\exp((3c-1)^2\|x_q\|^2/(2v^2))}{\exp(c^2\|x_q\|^2/(2v^2))\exp((2c-1)^2\|x_q\|^2/(2v^2))}-1\right) + o\left(\frac{1}{D}\right)
$$

$$= \quad -4\frac{c(2c-1)}{D}(x_q^\top\theta)^2\left(\exp((2c^2-c)\|x_q\|^2/v^2)-1\right)+o\left(\frac{1}{D}\right),$$

and

$$B_{37} = \frac{4}{D}c(2c-1)\exp(\|x_q\|^2 c/v - \|x_q\|^2 c^2/(2v^2)) + o\left(\frac{1}{D}\right),$$

$$B_{38} = \frac{4}{D}c(2c-1)\exp(\|x_q\|^2 (2c-1)/v - \|x_q\|^2 (2c-1)^2/(2v^2)) + o\left(\frac{1}{D}\right).$$

Putting everything together, we have

$$\begin{aligned}
B_3 &= B_{31}+B_{32}+B_{33}+B_{34}+B_{35}+B_{36}+B_{37}+B_{38}\\
&= -4c(2c-1)\frac{D-1}{D}(x_q^\top\theta)^2 - 4v^2\frac{1}{D}\left[\theta^\top(I_d+(3c-1)^2x_qx_q^\top/v^2)\theta+\sigma_\epsilon^2\right]\exp((2c^2-c)\|x_q\|^2/v^2)\\
&\quad +\frac{4}{D}(x_q^\top\theta)^2[2c(2c-1)]\exp(\|x_q\|^2c^2/v^2) - \frac{8}{D}(x_q^\top\theta)^2[c(2c-1)]\\
&\quad +\frac{4}{D}(x_q^\top\theta)^2[c(3c-1)]\exp((2c^2-c)\|x_q\|^2/v^2)\\
&\quad +\frac{4}{D}(x_q^\top\theta)^2[2c(2c-1)]\exp(\|x_q\|^2(2c-1)^2/v^2) - \frac{8}{D}(x_q^\top\theta)^2[c(2c-1)]\\
&\quad +\frac{4}{D}(x_q^\top\theta)^2[(3c-1)(2c-1)]\exp((2c^2-c)\|x_q\|^2/v^2)\\
&\quad -4\frac{c(2c-1)}{D}(x_q^\top\theta)^2\left(\exp(c^2\|x_q\|^2/v^2)-1\right)\\
&\quad -4\frac{c(2c-1)}{D}(x_q^\top\theta)^2\left(\exp((2c-1)^2\|x_q\|^2/v^2)-1\right)\\
&\quad -4\frac{c(2c-1)}{D}(x_q^\top\theta)^2\left(\exp((2c^2-c)\|x_q\|^2/v^2)-1\right)\\
&\quad +\frac{4}{D}c(2c-1)\exp(\|x_q\|^2c/v-\|x_q\|^2c^2/(2v^2))\\
&\quad +\frac{4}{D}c(2c-1)\exp(\|x_q\|^2(2c-1)/v-\|x_q\|^2(2c-1)^2/(2v^2)) + o\left(\frac{1}{D}\right)\\
&= -(8c^2-4c)(x_q^\top\theta)^2 - \frac{6}{D}(x_q^\top\theta)^2c(2c-1) - \frac{4v^2}{D}\left[\|\theta\|^2+\sigma_\epsilon^2\right]\exp((2c^2-c)\|x_q\|^2/v^2)\\
&\quad +\frac{4}{D}(x_q^\top\theta)^2\exp(\|x_q\|^2c^2/v^2)\left[2c(2c-1)-c(2c-1)\right]\\
&\quad +\frac{4}{D}(x_q^\top\theta)^2\exp(\|x_q\|^2(2c-1)^2/v^2)\left[2c(2c-1)-c(2c-1)\right]\\
&\quad +\frac{4}{D}(x_q^\top\theta)^2\exp(\|x_q\|^2(2c^2-c)/v^2)\left[-2c(2c-1)\right]\\
&\quad +\frac{4}{D}c(2c-1)\exp(\|x_q\|^2c/v-\|x_q\|^2c^2/(2v^2))\\
&\quad +\frac{4}{D}c(2c-1)\exp(\|x_q\|^2(2c-1)/v-\|x_q\|^2(2c-1)^2/(2v^2)) + o\left(\frac{1}{D}\right)\\
&= -4(2c^2-c)(x_q^\top\theta)^2 - \frac{4v^2}{D}\left[\|\theta\|^2+\sigma_\epsilon^2\right]\exp((2c^2-c)\|x_q\|^2/v^2)\\
&\quad +\frac{4}{D}(x_q^\top\theta)^2\exp(\|x_q\|^2c^2/v^2)\left[c(2c-1)\right]\\
&\quad +\frac{4}{D}(x_q^\top\theta)^2\exp(\|x_q\|^2(2c-1)^2/v^2)\left[c(2c-1)\right]\\
&\quad +\frac{4}{D}(x_q^\top\theta)^2\exp(\|x_q\|^2(2c^2-c)/v^2)\left[-2c(2c-1)\right]\\
&\quad +\frac{4}{D}c(2c-1)\exp(\|x_q\|^2c/v-\|x_q\|^2c^2/(2v^2))\\
&\quad +\frac{4}{D}c(2c-1)\exp(\|x_q\|^2(2c-1)/v-\|x_q\|^2(2c-1)^2/(2v^2)) + o\left(\frac{1}{D}\right).
\end{aligned}$$

Finally,

$$\mathbb{E}\left(y_q - f(E)_{d+1,D+1}\right)^2$$

$$
\begin{aligned}
&= 1+\sigma_\epsilon^2+B_1+B_2+\mathbb{E}(B_{31}+B_{32}+B_{33}+B_{34}+B_{35}+B_{36}+B_{37}+B_{38})+o\left(\frac{1}{D}\right)\\
&= 1+\sigma_\epsilon^2+4c^2-4c+\frac{4v^2(1+\sigma_\epsilon^2)}{D}\mathbb{E}\exp(c^2\|x_q\|^2/v^2)-\frac{4c^2-4c}{D}\mathbb{E}(x_q^\top\theta)^2\exp(c^2\|x_q\|^2/v^2)\\
&\quad-\frac{2(4c^2-2c)}{D}\mathbb{E}(x_q^\top\theta)^2\exp(c\|x_q\|^2/v-c^2\|x_q\|^2/v^2/2)\\
&\quad+(2c-1)^2+2(2c-1)+\frac{v^2(1+\sigma_\epsilon^2)}{D}\mathbb{E}\exp((2c-1)^2\|x_q\|^2/v^2)\\
&\quad-\frac{(2c-1)^2+2(2c-1)}{D}\mathbb{E}(x_q^\top\theta)^2\exp((2c-1)^2\|x_q\|^2/v^2)\\
&\quad-\frac{2((2c-1)^2+(2c-1))}{D}\mathbb{E}(x_q^\top\theta)^2\exp((2c-1)\|x_q\|^2/v-(2c-1)(2c-1)^2\|x_q\|^2/v^2/2)\\
&\quad-\mathbb{E}4(2c^2-c)(x_q^\top\theta)^2-\mathbb{E}\frac{4v^2}{D}\left[\|\theta\|^2+\sigma_\epsilon^2\right]\exp((2c^2-c)\|x_q\|^2/v^2)\\
&\quad+\mathbb{E}\frac{4}{D}(x_q^\top\theta)^2\exp(\|x_q\|^2c^2/v^2)\left[c(2c-1)\right]\\
&\quad+\mathbb{E}\frac{4}{D}(x_q^\top\theta)^2\exp(\|x_q\|^2(2c-1)^2/v^2)\left[c(2c-1)\right]\\
&\quad+\mathbb{E}\frac{4}{D}(x_q^\top\theta)^2\exp(\|x_q\|^2(2c^2-c)/v^2)\left[-2c(2c-1)\right]\\
&\quad+\mathbb{E}\frac{4}{D}c(2c-1)\exp(\|x_q\|^2c/v-\|x_q\|^2c^2/(2v^2))\\
&\quad+\mathbb{E}\frac{4}{D}c(2c-1)\exp(\|x_q\|^2(2c-1)/v-\|x_q\|^2(2c-1)^2/(2v^2))+o\left(\frac{1}{D}\right)\\
&= \sigma_\epsilon^2+\frac{v^2(1+\sigma_\epsilon^2)}{D}\mathbb{E}\left[4\exp(c^2\|x_q\|^2/v^2)+\exp((2c-1)^2\|x_q\|^2/v^2)-4\exp((2c^2-c)\|x_q\|^2/v^2)\right]\\
&\quad+\frac{4c^2}{D}(x_q^\top\theta)^2\exp(\|x_q\|^2c^2/v^2)+\frac{(2c-1)^2}{D}(x_q^\top\theta)^2\exp(\|x_q\|^2(2c-1)^2/v^2)\\
&\quad+\frac{4}{D}(x_q^\top\theta)^2\exp(\|x_q\|^2(2c^2-c)/v^2)\left[-2c(2c-1)\right]+o\left(\frac{1}{D}\right),
\end{aligned}
$$

thus

$$
\begin{aligned}
&\mathbb{E}\left(y_q-f(E)_{d+1,D+1}\right)^2\\
=\;&\sigma_\epsilon^2+\frac{4v^2(1+\sigma_\epsilon^2)}{D}\left((\frac{v^2}{v^2-2c^2})^{\frac{d}{2}}-(\frac{v^2}{v^2-2c(2c-1)})^{\frac{d}{2}}\right)+\frac{v^2(1+\sigma_\epsilon^2)}{D}(\frac{v^2}{v^2-2(2c-1)^2})^{\frac{d}{2}}\\
&+\frac{(2c-1)^2}{D}(\frac{v^2}{v^2-2(2c-1)^2})(\frac{v^2}{v^2-2(2c-1)^2})^{\frac{d}{2}}\\
&+\frac{4c^2}{D}(\frac{v^2}{v^2-2c^2})(\frac{v^2}{v^2-2c^2})^{\frac{d}{2}}-\frac{4(2c-1)c}{D}(\frac{v^2}{v^2-2c(2c-1)})(\frac{v^2}{v^2-2c(2c-1)})^{\frac{d}{2}}+o(\frac{1}{D}).
\end{aligned}
$$

□

C.7 Correlated Features: Theorem A.2

Proof of Theorem A.2. To figure out the optimal solution of single-head attention, we firstly transform the problem from correlated features to the problem with isotropic features with a new θ distribution. After transforming the problem, since Theorem 4.1 only utilize the distribution of θ in its last derivation step, we can directly utilize the results in Theorem 4.1.

To transform correlated features, denote $z\sim N(0,I_d)$ and $x=\Sigma^{1/2}z$. Recall that the attention score is calculated as

$$
\phi\left((W^KW_{in}E)^\top(W^QW_{in}E)\right)=\phi\left((W^KW_{in}E)^\top(W^QW_{in}E)\right)
$$

Based on Theorem 4.1, we have

$$
\mathbb{E}\left(y_q-(W_{d+1,:}^V)^\top E\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\0\end{bmatrix}\right)\right)^2
$$

$$
\begin{aligned}
&= \mathbb{E}_{(x_q,\theta)}(x_q^\top\theta)^2 + \frac{v^2}{D}\theta^\top(I_d - 4(A+\theta b^\top)x_q x_q^\top(A+\theta b^\top)^\top)\theta \exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q) \\
&+ v^2(1+\frac{3}{D})(\theta^\top(A+\theta b^\top)x_q)^2 - \frac{2v^2(\theta^\top(A+\theta b^\top)x_q)^2\exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)} \\
&- \frac{3v^2}{D}(\theta^\top(A+\theta b^\top)x_q)^2 + \frac{3v^2}{D}(\theta^\top(A+\theta b^\top)x_q)^2\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q) \\
&- 2(x_q^\top\theta)\Bigg(v\theta^\top(A+\theta b^\top)x_q - \frac{v\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top Ax_q)}{D\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q/2)} + \frac{v}{D}\theta^\top(A+\theta b^\top)x_q \\
&\quad -\frac{2v}{D}\theta^\top(A+\theta b^\top)x_q\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q - \frac{v}{D}(\theta^\top(A+\theta b^\top)x_q) \\
&\quad +\frac{v}{D}(\theta^\top(A+\theta b^\top)x_q)\exp(x_q^\top(A+\theta b^\top)^\top(A+\theta b^\top)x_q))\Bigg) + o(\frac{1}{D}),
\end{aligned}
$$

from which the optimal solution satisfies $\mathbb{E}\theta^\top(I_d - vA)^2\theta = O(1/D)$ and $\|b\|^2\mathbb{E}\|\theta\|^4 = O(1/D)$ where $\theta \sim N(0, \Sigma^{-1/2}/d)$.

For multi-head attention, the same argument applies, and we can also transform the correlated features problem to isotropic features with a new θ distribution. Further, due to the flexibility of multi-head attention, when each head is of full rank, i.e., $p/h > d$, the performance of multi-head attention is not worse than single-head attention. There always exists some W_{out} such that the multi-head attention can be reduced to a single-head attention.

□

D Justifications for Assumptions on Optimal Configurations

The assumed configurations format in Theorem 4.1 and 4.2 are not selected specific configurations but indeed the optimal format which minimized the MSE of prediction. To verify this, we provide the justification for these assumptions as follows.

1. $(W_{out}W^V)_{d+1,;} = (0, 0, ..., 0, v)$ in Theorem 4.1:

   We justify this assumption by showing that when having format other than $(W_{out}W^V)_{d+1,;} = (0, 0, ..., 0, v)$, the loss of the prediction will becomes larger.

   Intuitively, since $\theta \sim N(0, I_d/d)$, if $W_{out}W^V = (\beta, v)$ for some nonzero $\beta \in \mathbb{R}^d$, then in addition to the weighted average of y_is, the output also contains another weighted average of $\beta^\top x_i$s, and β and θ are independent with each other, i.e., this part leads to an extra loss.

   To write the formula, denote $W_{out}W_v = (\beta, v)$ and $attn$ as the attention score output. When $\|\beta\| > 0$, we have

$$
\begin{aligned}
& \mathbb{E}_{\theta,\{x_i\}}(\hat{y}_q - y_q)^2 \\
= \;& \mathbb{E}_{\{x_i\}}(\mathbb{E}_\theta(\hat{y}_q - y_q)^2) \\
= \;& \mathbb{E}_{\{x_i\}}(\mathbb{E}_\theta(\beta^\top(\sum attn_i x_i) + v\sum attn_i x_i^\top\theta - x_q^\top\theta)^2) \\
= \;& \mathbb{E}_{\{x_i\}}(\mathbb{E}_\theta(v\sum attn_i x_i^\top\theta - x_q^\top\theta)^2 + (\beta^\top(\sum attn_i x_i))^2) \\
> \;& \mathbb{E}_{\{x_i\}}\mathbb{E}_\theta(v\sum attn_i x_i^\top\theta - x_q^\top\theta)^2.
\end{aligned}
$$

2. Reason for assuming $(W^K)^\top W^Q = \begin{bmatrix} A & 0 \\ b & 0 \end{bmatrix}$ in Theorem 4.1:

   As mentioned in Section 3, we fetch the last element of the last row in $f(E)$ as the predicted value of y_q. This prediction is expressed as: $\hat{y_q} = (W_{out}W^V_{d+1,:})^\top E\phi\left(E^\top(W^K)^\top W^Q\begin{bmatrix} x_q \\ 0 \end{bmatrix}\right)$. From this equation, we can see that the last column of the matrix $(W^K)^\top W^Q$ is multiplied by 0 due to the appended zero in the

vector $\begin{bmatrix} x_q \\ 0 \end{bmatrix}$. Consequently, any value in the last column of $(W^K)^\top W^Q$ has no effect on the outcome, as it contributes nothing to the computation of $\hat{y_q}$. To simplify, we designate these values as 0.

3. $f(E)_{d+1,D+1} = vmE_{d+1,:}\phi((W_1^K E)^\top W_1^Q E_{:,D+1}) - vnE_{d+1,:}\phi((W_2^K E)^\top W_2^Q E_{:,D+1})$ in Theorem 4.2:
Based on the standard definition of the multi-head attention mechanism, we have:

$$\text{Multi-head}(Q,K,V) = [\text{head}_1, \text{head}_2]W_0,$$
$$\text{where} \quad \text{head}_i = W^V E\phi((W_i^K E)^\top W_i^Q E) \quad \text{and} \quad W_0 \in R^{(2D+2)\times(d+1)}$$

Since we take the we use the final element of the last row in f (E) as the prediction of y_q, the above equations can be simplified to

$$\text{head}_i = v_i E_{d+1,:}\phi((W_i^K E)^\top W_i^Q E) \quad \text{and} \quad W_0 \in R^{(2D+2)}$$

Here we assume that $W_0 = [0, \cdots, 0, p, 0, \cdots, 0, q]^\top$. Then by letting $v = v_1$, $m = p$ and $n = -v_2 q/v_1$, we can get $f(E)_{d+1,D+1} = vmE_{d+1,:}\phi((W_1^K E)^\top W_1^Q E_{:,D+1}) - vnE_{d+1,:}\phi((W_2^K E)^\top W_2^Q E_{:,D+1})$. The reason why we assume $W_0 = [0, \cdots, 0, p, 0, \cdots, 0, q]^\top$ is that if $W_0 = [\beta_1, p, \beta_2, q]^\top$ where $\beta_1 \neq 0$ and $\beta_2 \neq 0$, the loss will become larger because the prediction of the response of the query example x_q will include some terms which is independent of x_q and thus introduce unnecessary noise into the model. Let $attn_{a,b}$ be the final element of the output at token x_a's position from head b, then we have:

$$\begin{aligned}
& \mathbb{E}_{\theta,\{x_i\}} \left(y_q - \hat{y_q}\right)^2 \\
= \; & \mathbb{E}_{\{x_i\}}\mathbb{E}_\theta \left(y_q - v_1 p \cdot attn_{q,1} - v_2 q \cdot attn_{q,2} - v_1 \sum_{i=0}^{D} \beta_{1,i} attn_{i,1} - v_2 \sum_{i=0}^{D} \beta_{2,i} attn_{i,2} \right)^2 \\
= \; & \mathbb{E}_{\{x_i\}}\mathbb{E}_\theta \left(\left(y_q - v_1 p \cdot attn_{q,1} - v_2 q \cdot attn_{q,2}\right)^2 \right) + \mathbb{E}_{\{x_i\}}\mathbb{E}_\theta \left(v_1 \sum_{i=0}^{D} \beta_{1,i} attn_{i,1} + v_2 \sum_{i=0}^{D} \beta_{2,i} attn_{i,2} \right)^2 \\
& \text{(The expectation of the cross term is O(1/D))} \\
> \; & \mathbb{E}_{\{x_i\}}\mathbb{E}_\theta \left(\left(y_q - v_1 p \cdot attn_{q,1} - v_2 q \cdot attn_{q,2}\right)^2 \right)
\end{aligned}$$

E Simulation and Experiment Details

E.1 Visualization of Single-Head Attention Score

Based on Theorem 4.1, the optimal A is in the format of $I_d/v + o$. As a result, there are two possible cases. (i) When $v > 0$, the attention score of x_q against itself is usually the largest one as $x_q^\top A x_q = \|x_q\|^2/v$ is always positive. (ii) When $v < 0$, the attention score of x_q against itself is always small. Figure 16 shows these two cases correspondingly.

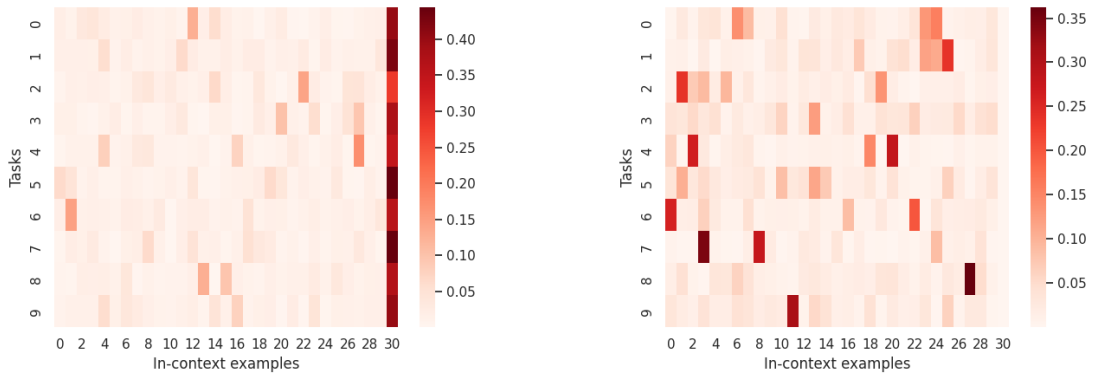
E.2 Simulations Results under Additional Settings.

In Table 4, we demonstrate simulation results when the input data follows Student's t-distribution ($\nu = 5$) or a Gaussian Mixture Model (consisting of 4 components, each with means of $[2,0,0,0,0]$, $[-2,0,0,0,0]$, $[0,2,0,0,0]$ and $[0,-2,0,0,0]$ and each having a spherical covariance matrix). The results in the table demonstrate that the insights on the superiority of multi-head attention are consistent across different distributions of x.

Additionally, we extend our analysis to multi-layer transformers, with the results summarized in Table 5. These findings demonstrate that the superiority of multi-head attention over single-head attention holds across different numbers of layer, which confirms the insights derived from our single-layer analysis generalize well to multi-layer settings.

E.3 Noisy Response and Correlated Features

For noisy response and correlated features, we conduct experiments to verify the effectiveness of multi-head attention. The results for noisy label can be found in Figure 17. While the best prediction loss is away from zero, one can still see that with sufficient input embedding dimension, multi-head attention improves the performance.



(a) $v > 0$ (b) $v < 0$

Figure 16: Single-Head Attention Score for 10 tasks.

Table 4: MSE for different values of h when x follows non-Gaussian distributions.

Distribution of x	h	**MSE**
Student's t-distribution	1	0.6186
	2	0.4760
Gaussian Mixture Model	1	0.7912
	2	0.1745

Table 5: MSE of ICL with Different Number of Heads.

Layers	h	**MSE**
2	1	0.2235
	2	0.0651
4	1	0.2176
	2	0.0496

For correlated features, to generate Σ, we follow the procedure in Zhang et al. (2024) and take the diagonal elements following exp(1) distribution. For the off diagonal elements, we take all of them as 0.1. From Figure 18 we can see that multi-head attention with $p/h > d$ is better than single-head attention.
More simulation results can be found in `https://drive.google.com/file/d/1TjGNtcxxs88ngr6usRhFNYCeapZD4s7R/view?usp=sharing`.