Appendix for "What and How does In-Context Learning Learn? Bayesian Model Averaging, Parameterization, and Generalization"

A Conclusion

In this paper, we investigated the theoretical foundations of ICL for the pretrained language models. We proved that the perfectly pretrained LLMs implicitly implements BMA with regret $\mathcal{O}(1/t)$ over a general response generation modeling, which subsumes the models in previous works. Based on this, we showed that the attention mechanism parameterizes the BMA algorithm. Analyzing the pretraining process, we demonstrated that the total variation between the pretrained model and the nominal distribution consists of the approximation error and the generalization error. The combination of the ICL regret and the pretraining performance gives the full description of ICL ability of pretrained LLMs. We mainly focus on the prompts that comprise several examples in this work and leave the analysis of instruction-based prompts for future works.

B More Related Works

Transformers. Our work is also related to the works that theoretically analyze the performance of transformers. For the analytic properties of transformers, Vuckovic et al. (2020) proved that attention is Lipschitz-continuous via the view of interacting particles. Noci et al. (2022) provided the theoretical justification of the rank collapse phenomenon in transformers. Yun et al. (2019) demonstrated that transformers are universal approximators. For the statistical properties of transformers, Malladi et al. (2022), Hron et al. (2020), and Yang (2020) analyzed the training of transformers within the neural tangent kernel framework. Wei et al. (2022a) presented the approximation and generalization bounds for learning boolean circuits and Turing machines with transformers. Edelman et al. (2021) and Li et al. (2023) derived the generalization error bound of transformers. In our work, we analyze transformers from both the analytic and statistical sides. We show that attention essentially implements the BMA algorithm in the ICL setting. Furthermore, we derive the approximation and generalization bounds for transformers in the pretraining phase.

Generalization. Our analysis of the pretraining is also related to the generalization analysis of the neural networks. This topic has attracted a lot of interests for a long time. Anthony et al. (1999) derived the uniform generalization bound for fully-connected neural networks with the help pf VC dimension. Bartlett et al. (2017) sharpened this generalization bound for classification problem by adopting the Dudley's integral and calculating of the covering number of neural network class. At the same time, Neyshabur et al. (2017) derived a similar as Bartlett et al. (2017) from PAC-Bayes framework. Following this line, Liao et al. (2020) , Ledent et al. (2021) and Lin and Zhang (2019) built the generalization bound for graph neural networks and convolutional neural network. These results respected the underlying graph structure and the translation-invariance in the networks. Edelman et al. (2021) established the generalization bound for transformer, but this result did not reflect the permutation-invariance, still depending on the channel number. Our work focuses on the analysis of Maximum Likelihood Estimate (MLE) with transformer function class, which is not covered by previous works. Our bounds are sharper than that of Edelman et al. (2021) on the channel number dependency.

C Experimental Results

We conduct five experiments to verify our theoretical findings, including the Bayesian view (Propositions 4.1 and (4.7)), the regret upper-bounded in Corollary 4.2 and Theorem 6.2, and the constant ratio between $\texttt{attn}_\dagger$ and $\texttt{attn}$ in Proposition 4.3. The implementation details are provided in Appendix D.

C.1 Verification of the Bayesian View

To verify the Bayesian view that we adopt in the paper, we implemented two experiments. In the first experiment, we explicitly construct the hidden concept vectors that are found by LLMs.

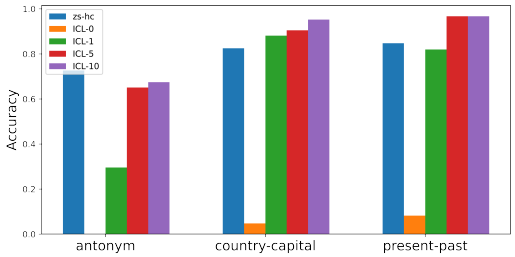

Figure 1: Accuracies of LLMs with and without explicit hidden concepts.

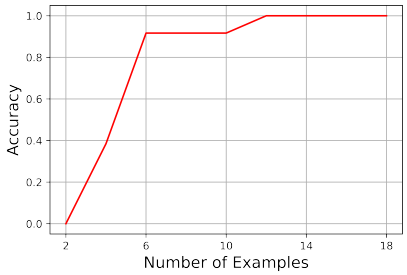

Figure 2: Accuracy of LLMs to find the best arm in the bandit instance with an informative arm.

Motivated by (4.7), we construct the hidden concept vector as the average sum over prompts of the values of twenty selected attention heads, i.e., we compress the hidden concept into a vector with dimension 4096. To demonstrate the effectiveness of the constructed hidden concepts, we add these hidden concept vectors at a layer of LLMs when the model resolves the prompt with zero-shot. In Figure. 1, "zs-hc" refers to the results of LLMs that infers with learned hidden concept vectors and zero-shot prompt, and "ICL-i" refers to the results of LLMs prompted with i examples. We consider the tasks of finding antonyms, finding the capitals of countries, and finding the past tense of words. The results indicate that the LLMs with learned hidden concept vectors have comparable performance with the LLMs prompted with several examples. This indicates that the learned hidden concept vectors are indeed efficient compression of the hidden concepts, which proves that LLMs deduce hidden concepts for ICL. This result strongly corroborates with (4.7).

In the second experiment, we aim to verify that LLMs implement inference with the Bayesian framework, not with gradient descent (Akyürek et al., 2022; von Oswald et al., 2022; Bai et al., 2023) on some tasks. We prompt the LLMs with the history data of a set of similar multi-armed bandit instances with 100 arms, and let LLMs indicate which arm to pull in a similar new bandit instance. In these similar bandit instances, there is an informative arm, whose reward is exactly the index of the arm with the highest rewards. We also provide the side information that "Some arm may directly tell you the arm with the highest reward, even itself does not have the highest reward". In each example provided in the prompt, there are the rewards of six arms, including the informative arm and the best arm, in one bandit instance. As shown in Figure 2, the LLMs can efficiently implement ICL even with only 6 examples. We note that the gradient descent algorithms in the previous works cannot explain this performance, since the gradient descent algorithms need at least 100 data points, where each data point is the reward of one arm, to learn. In contrast, the Bayesian view can clearly explain Figure 2, where LLMs make use of the side information to calculate a better posterior for ICL.

C.2 Verification of the Regret Bound

To verify Corollary 4.2 and Theorem 6.2, we implement experiments to evaluate the regret in two settings. In the first setting, the LLMs is trained for the linear regression task from scratch, which is a representative setting studied in Garg et al. (2022); Akyürek et al. (2022). The examples in the prompt are $\{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$, $d = 20$ and $y_i = w^T x_i$ for some w sampled from Gaussian distribution. Given the Gaussian model, we adopt the squared error to approximate the logarithm of the probability. Then the $t \times$ regret of the LLMs can be well approximated by the sum of the squared error till time t. The results in Figure 3 strongly corroborate our theoretical findings. First, the results verify our claim in Corollary 4.2 and Theorem 6.2 that $t \cdot$ regret can be upper bounded by a constant. Second, the line of squared error indicates that the ICL of LLMs only has a significant error when $T \leq d$, i.e., the regret only increases in this region. Thus, the regret of the ICL by LLMs is at most linear in $O(d/T)$. From the view of our theoretical result, discretizing the set $\{z \in \mathbb{R}^d \,|\, \|z\|_2 \leq d\}$ with approximation error $\delta > 0$ will result in a set with $(C/\delta)^d$ elements, where $C > 0$ is an absolute constant. Corollary 4.2 and Theorem 6.2 imply that the regret is the sum of the $\log |\mathfrak{Z}|/T = d \log(C/\delta)/T$ and the pretraining error, which matches the simulation results.

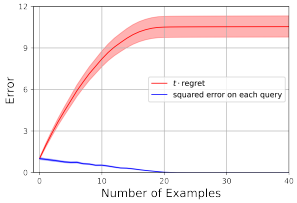

Figure 3: Squared error and regret of LLMs trained for linear regression.

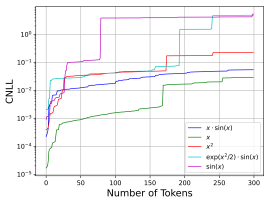

Figure 4: Cumulative negative log-likelihoods of pretrained LLMs for function value prediction.

Figure 5: The ratio between $\texttt{attn}_\dagger$ and $\texttt{attn}$.

In the second experiment, we directly evaluate the regret of pretrained LLMs on the function value prediction task. The prompt consists of the values of a function on the points with fixed intervals. Since the values are real numbers, we adopt the method in Gruver et al. (2023) to transfer a real number to a token sequence. For the pretrained model, we cannot calculate $\mathbb{P}(r_i \,|\, \texttt{prompt}_{i-1}, z)$ due to the unknown nominal distributions. Thus, we calculate the cumulative negative log-likelihood $\text{CNLL}_t = -\sum_{i=1}^{t} \widehat{\mathbb{P}}(r_i \,|\, \texttt{prompt}_{i-1})$, and CNLL_t/t is an upper bound of the regret. In Figure 4, we indicate the cumulative negative log-likelihoods of predicting the values of five functions. The results show that the cumulative negative log-likelihoods are stepped, which means that the cumulative negative log-likelihoods are upper-bounded by constants in a long period. This corroborates with Corollary 4.2 and Theorem 6.2. In addition to the mentioned property, we also observe that there are connections between the cumulative negative log-likelihood and the prediction error. We let the LLMs to predict the value given the prompt that contains the past values. Figures 6 and 7 show that the larger cumulative negative log-likelihood implies a larger prediction error.

C.3 Verification of the Constant Ratio between $\texttt{attn}_\dagger$ and $\texttt{attn}$

To verify Proposition 4.3, we directly calculate the ratio between $\texttt{attn}_\dagger$ and $\texttt{attn}$. We consider the case $d_v = 1$ and $d_k = d$ for some $d > 0$. The entries in K of (4.8) are i.i.d. samples of Gaussian distribution, and the $i-$th entry of V is calculated as the inner product between a Gaussian vector and the $i-$th column. Figure 5 shows the results for $d = 2$ and $d = 3$. It shows that the ratio between $\texttt{attn}_\dagger$ and $\texttt{attn}$ will converge to a constant. This constant depends on the dimension d, which originates from Proposition F.1.

D Implementation Details of Experiments

In this section, we provide the implementation details of the experiments. In the hidden concepts construction experiment, we explicitly calculate the hidden concept vector for Llama2-7b with the method in Todd et al. (2023). Given the prompts generated from the same hidden concept, we calculate the average value of each attention head by prompting the LLM with different prompts. Then we select the attention head according to its average indirect effect, which is defined in Todd et al. (2023). The hidden concept vector is the sum of the average value of the selected attention heads. We test the performance of the learned hidden concept vectors on tasks: (1) Antonym: Given an input word, generate the word with the opposite meaning. (2) Country-Capital. Given a country name, generate the capital city. (3) Present-Past. Given a verb in the present tense, generate the verb's simple past inflection. To test the effectiveness of the learned hidden concept vector, we prompt the LLM only with the query, i.e., the zero-shot case, and set the attention head values at some layer as the learned hidden concept vector.

In the bandit experiment, we ask GPT-4 for the procedures to find the arm with the highest reward. In each bandit instance, there is an informative arm, whose reward is exactly the index of the best arm. When prompting models, we provide the historical data of several bandit instances that share the same informative arm and ask models to specify how we should play in a similar bandit instance. A prompt sample with two examples is provided as follows.

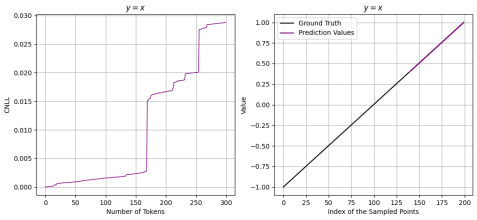

Figure 6: Cumulative negative log-likelihood and the prediction values for $y = x$.

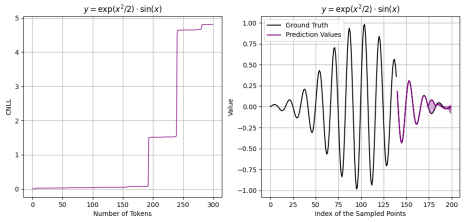

Figure 7: Cumulative negative log-likelihood and the prediction values for $y = \exp(x^2/2) \cdot \sin(x)$.

> Your goal is to find the index of the arm with the highest reward, but the pulled arm may not have the highest reward. I will provide you with the past pull history on other bandits. The format of the history data on each bandit is [arm, reward]. Different pulls are separated by a comma. For example, [5,6] indicates that arm 5 will give us a reward of 6 by pulling it.
> You should learn from history and tell me which arm to pull in the current bandit to find the arm with the highest reward. The history data is as follows.
> Bandit:
> [77, 871], [95, 613], [75, 655], [17, 449], [31, 13], [13, 1028]
> Bandit:
> [40, 698], [44, 88], [80, 147], [94, 265], [24, 1063], [31, 24]
> Different bandits can have different rewards for each arm, but all bandits share a common pattern. Some arm may directly tells you the arm with the highest reward, even itself does not have the highest reward. Now I am playing a new bandit. This bandit will have different rewards than the bandits in history, but they share the same pattern. Tell me which arm to pull to find the arm with the highest reward. Tell me the final answer that only contains the index of the arm in a single line without any additional text.

In the above prompt, the arm 31 always returns the index of the best arm. Thus, we expect LLMs to tell us to pull arm 31 to find the best arm. The number of arms in each instance is 100, and each example only provides information about six arms in each instance. We repeat the prompt with different data ten times to plot Figure 2.

For the linear regression task, the model is trained with the loss

$$L(f) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - f(\texttt{prompt}_t)\right)^2,$$

where $\texttt{prompt}_t = (x_1, y_1, \cdots, x_{t-1}, y_{t-1}, x_t)$, $y_t = w^{\mathrm{T}}x_t$, $\{x_t\}_{t=1}^T$ and w are i.i.d. samples of Gaussian distribution (Garg et al., 2022). The model is designed based on GPT-2, and we add linear layers as the first and last layers to accommodate it for the value prediction task. In the testing phase, we sample w^* and $\{x_t\}_{t=1}^T$ from the Gaussian distribution and let the model predict the response value of a query x_{t+1} given the previous examples $\{x_i, y_i\}_{i=1}^t$. We reuse the code and model in Garg et al. (2022) for the experiments. The error bar in Figure 3 is derived from 90% confidence intervals over 1000 bootstrap trials.

In the function value prediction task, we adopt the method in Gruver et al. (2023) to transfer the real number into tokens. We separate the digits with spaces and add commas ',' between the function values at different times. We calculate the negative likelihood of text-DaVinci-003 by extracting the probability value in the last layer of it. We note that the negative likelihood in Figure 4 takes every token into account, including the separating spaces between the digits.

In the experiment about the ratio between $\texttt{attn}_\dagger$ and $\texttt{attn}$, we set W_Q, W_K and W_V in $\texttt{attn}$ all as the identity matrix. The entries in the K of (4.8) are i.i.d. samples of the normal distribution, and the $i-$th entry of V is calculated as the inner product between a Gaussian vector and the $i-$th column. The Gaussian vector is sampled from $\mathcal{N}(0, I)$. The error bar in Figure 3 is derived from 75th and 25th percentiles over 500 trials.

E FIGURE FOR PRETRAINING AND ICL

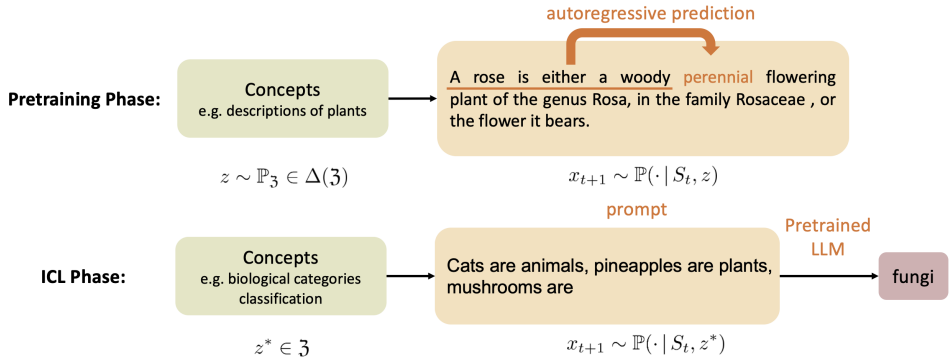

Figure 8: To form the pretraining dataset, a hidden concept z is first sampled according to $\mathbb{P}_{\mathfrak{Z}}$, and a document is generated from the concept. Taking the token sequence S_t up to position $t \in [T]$ as the input, the LLM is pretrained to maximize the next token x_{t+1}. During the ICL phase, the pretrained LLM is prompted with several examples to predict the response of the query.

F PROOFS FOR SECTION 4.1

F.1 INTRODUCTION OF CONDITIONAL MEAN EMBEDDING

Let $\mathcal{H}_k$ and $\mathcal{H}_v$ be the two RKHSs over the spaces $\mathfrak{Q}$ and $\mathfrak{V}$ with the kernels $\mathfrak{K}$ and $\mathfrak{L}$, respectively. We denote by $\phi : \mathfrak{Q} \to \ell_2$ and $\varphi : \mathfrak{V} \to \ell_2$ the feature mappings associated with $\mathcal{H}_k$ and $\mathcal{H}_v$, respectively. Here l_2 is the space of the square-integrable function class. Then it holds for any $k, k' \in \mathfrak{Q}$ and $v, v' \in \mathfrak{V}$ that

$$\phi(k)^\top \phi(k') = \mathfrak{K}(k, k'), \qquad \varphi(v)^\top \varphi(v) = \mathfrak{L}(v, v'). \tag{F.1}$$

Let $\mathbb{P}_{\mathcal{K},\mathcal{V}}$ be the joint distribution of the two random variables $\mathcal{K}$ and $\mathcal{V}$ taking values in $\mathfrak{Q}$ and $\mathfrak{V}$, respectively. Then the conditional mean embedding $\texttt{CME}(q, \mathbb{P}_{\mathcal{K},\mathcal{V}}) \in \mathcal{H}_v$ of the conditional distribution $\mathbb{P}_{\mathcal{V} \mid \mathcal{K}}$ is defined as

$$\texttt{CME}(q, \mathbb{P}_{\mathcal{K},\mathcal{V}}) = \mathbb{E}\big[\mathfrak{L}(\mathcal{V}, \cdot) \,\big|\, \mathcal{K} = q\big].$$

The conditional mean embedding operator $C_{\mathcal{V} \mid \mathcal{K}} : \mathcal{H}_k \to \mathcal{H}_v$ is a linear operator such that

$$C_{\mathcal{V} \mid \mathcal{K}} \mathfrak{K}(q, \cdot) = \texttt{CME}(q, \mathbb{P}_{\mathcal{K},\mathcal{V}}),$$

for any $q \in \mathfrak{Q}$. We define the (uncentered) covariance operator $C_{\mathcal{K}\mathcal{K}} : \mathcal{H}_k \to \mathcal{H}_k$ and the (uncentered) cross-covariance operator $C_{\mathcal{V}\mathcal{K}} : \mathcal{H}_k \to \mathcal{H}_v$ as follows,

$$C_{\mathcal{K}\mathcal{K}} = \mathbb{E}\big[\mathfrak{K}(\mathcal{K}, \cdot) \otimes \mathfrak{K}(\mathcal{K}, \cdot)\big], \qquad C_{\mathcal{V}\mathcal{K}} = \mathbb{E}\big[\mathfrak{L}(\mathcal{V}, \cdot) \otimes \mathfrak{K}(\mathcal{K}, \cdot)\big].$$

Here $\otimes$ is the tensor product. Song et al. (2009) shows that $C_{\mathcal{V} \mid \mathcal{K}} = C_{\mathcal{V}\mathcal{K}} C_{\mathcal{K}\mathcal{K}}^{-1}$. Thus, we have that

$$\texttt{CME}(c, \mathbb{P}_{\mathcal{K},\mathcal{V}}) = C_{\mathcal{V}\mathcal{K}} C_{\mathcal{K}\mathcal{K}}^{-1} \mathfrak{K}(c, \cdot). \tag{F.2}$$

For i.i.d. samples $\{(k^\ell, v^\ell)\}_{\ell \in [L]}$ of $\mathbb{P}_{\mathcal{K},\mathcal{V}}$, $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert-Schmidt norm, we write $\phi(K) = (\phi(k^1), \ldots, \phi(k^L))^\top \in \mathbb{R}^{L \times d_\phi}$ and $\varphi(V) = (\phi(v^1), \ldots, \phi(v^L))^\top \in \mathbb{R}^{L \times d_\varphi}$. Then the empirical covariance operator $\widehat{C}_{\mathcal{K}\mathcal{K}}$ and empirical cross-covariance operator $\widehat{C}_{\mathcal{V}\mathcal{K}}$ are defined as

$$\begin{aligned}
\widehat{C}_{\mathcal{K}\mathcal{K}} &= L^{-1} \sum_{\ell=1}^{L} \phi(k^\ell) \phi(k^\ell)^\top = L^{-1} \phi(K)^\top \phi(K) \in \mathbb{R}^{d_\phi \times d_\phi} \\
\widehat{C}_{\mathcal{V}\mathcal{K}} &= L^{-1} \sum_{\ell=1}^{L} \varphi(v^\ell) \varphi(k^\ell)^\top = L^{-1} \varphi(V) \phi(K)^\top \in \mathbb{R}^{d_\varphi \times d_\phi}.
\end{aligned} \tag{F.3}$$

The empirical version of the conditional operator is

$$\widehat{C}_{\mathcal{V} \mid \mathcal{K}}^{\lambda} = \varphi(Y)^\top \phi(X) \big(\phi(X)^\top \phi(X) + \lambda \mathcal{I}\big)^{-1} = \widehat{C}_{\mathcal{V}\mathcal{K}} (\widehat{C}_{\mathcal{K}\mathcal{K}} + L^{-1} \lambda \mathcal{I})^{-1} \in \mathbb{R}^{d_\varphi \times d_\phi}.$$

F.2 Proof of Proposition 4.1

Proof. By (4.1), we have that

$$\begin{aligned}\mathbb{P}(r_{t+1} \,|\, \texttt{prompt}_t) &= \int \mathbb{P}(r_{t+1} \,|\, h_{t+1}, \texttt{prompt}_t)\mathbb{P}(h_{t+1} \,|\, \texttt{prompt}_t)\mathrm{d}h_{t+1} \qquad \text{(F.4)}\\ &= \int \mathbb{P}(r_{t+1} \,|\, \widetilde{c}_{t+1}, h_{t+1})\mathbb{P}(h_{t+1} \,|\, S_t)\mathrm{d}h_{t+1}\\ &= \int \mathbb{P}(r_{t+1} \,|\, \widetilde{c}_{t+1}, h_{t+1})\mathbb{P}(h_{t+1} \,|\, S_t, z)\mathbb{P}(z \,|\, S_t)\mathrm{d}h_{t+1}\mathrm{d}z\\ &= \int \mathbb{P}(r_{t+1} \,|\, \widetilde{c}_{t+1}, h_{t+1}, S_t, z)\mathbb{P}(h_{t+1} \,|\, S_t, z)\mathrm{d}h_{t+1}\mathbb{P}(z \,|\, S_t)\mathrm{d}z\\ &= \int \mathbb{P}(r_{t+1} \,|\, \widetilde{c}_{t+1}, S_t, z)\mathbb{P}(z \,|\, S_t)\mathrm{d}z, \qquad \text{(F.5)}\end{aligned}$$

where the first inequality results from the Bayes rule, the second equality results from the fact that r_{t+1} is conditionally independent with the previous history given $h_{t+1}, \widetilde{c}_{t+1}$ and the fact that h_{t+1} only parameterizes the transition kernel of r_{t+1} given c_{t+1} in (4.1), the fourth equality results from the fact that r_{t+1} is conditionally independent with the other variables given $h_{t+1}, \widetilde{c}_{t+1}$, and the last equality results from the Bayes' rule.

□

F.3 Proof of Corollary 4.2

Proof. Note that

$$\mathbb{P}(z \,|\, S_t) = \frac{\mathbb{P}(S_t \,|\, z)\mathbb{P}_{\mathcal{Z}}(z)}{\int \mathbb{P}(S_t \,|\, z')\mathbb{P}_{\mathcal{Z}}(z')\mathrm{d}z'} = \frac{\prod_{i=1}^{t}\mathbb{P}(r_i \,|\, z, S_t, c_i)\mathbb{P}_{\mathcal{Z}}(z)}{\int \prod_{i=1}^{t}\mathbb{P}(r_i \,|\, z', S_{i-1}, c_i)\mathbb{P}_{\mathcal{Z}}(z')\mathrm{d}z'},$$

where the second equality results from the fact that the hidden variable z only parameterizes the **conditional probability** of r_t given c_t, c_t and z are independent. Then, by Bayesian model averaging, we have the following density estimation,

$$\begin{aligned}\mathbb{P}(r_{t+1} \,|\, S_t, c_{t+1}) &= \int \mathbb{P}(r_{t+1} \,|\, z, S_t, c_{t+1})\mathbb{P}(z \,|\, S_t)\mathrm{d}z\\ &= \frac{\int \prod_{i=1}^{t+1}\mathbb{P}(r_i \,|\, z, S_{i-1}, c_i)\mathbb{P}_{\mathcal{Z}}(z)\mathrm{d}z}{\int \prod_{i=1}^{t}\mathbb{P}(r_i \,|\, z', S_{i-1}, c_i)\mathbb{P}_{\mathcal{Z}}(z')\mathrm{d}z'}.\end{aligned}$$

Thus, it holds that

$$\begin{aligned}-\sum_{t=0}^{T}\log\mathbb{P}(r_{t+1} \,|\, c_{t+1}, S_t) &= -\sum_{i=1}^{t}\left(\log\int\prod_{i=1}^{t+1}\mathbb{P}(r_i \,|\, z, S_{i-1}, c_i)\mathbb{P}_{\mathcal{Z}}(z)\mathrm{d}z - \log\int\prod_{i=1}^{t}\mathbb{P}(r_i \,|\, z, S_{i-1}, c_i)\mathbb{P}_{\mathcal{Z}}(z)\mathrm{d}z\right)\\ &= -\log\int\prod_{t=0}^{T}\mathbb{P}(r_i \,|\, z, S_{i-1}, c_i)\mathbb{P}_{\mathcal{Z}}(z)\mathrm{d}z\\ &= \inf_q \mathbb{E}_{z\sim q}\left[-\sum_{i=1}^{T+1}\log\mathbb{P}(r_i \,|\, z, S_{i-1}, c_i)\right] + \mathbb{E}_{z\sim q}\left[\log\frac{q(z)}{\mathbb{P}_{\mathcal{Z}}(z)}\right],\end{aligned}$$

where the second equality results from the fact that $\mathbb{P}(r_{t+1} \,|\, c_{t+1}, S_t) = \frac{\int \mathbb{P}(r_1 \,|\, c_1, z)\mathbb{P}(z)\mathrm{d}z}{1}$., and the last equality results from the standard Lagrangian arguments.

We consider q to be in the class of all Dirac measures. Then, we have that

$$-\frac{1}{T}\sum_{t=1}^{T}\log\mathbb{P}(r_t \,|\, c_t, S_{t-1}) \le \frac{1}{T}\inf_z\Big(-\sum_{t=1}^{T}\log\mathbb{P}(r_t \,|\, z, S_{t-1}, c_t) - \log\mathbb{P}_{\mathcal{Z}}(z)\Big).$$

Thus, the statistical convergence rate of the Bayesian posterior averaging is $\mathcal{O}(1/T)$. □

F.4 Proof of Proposition 4.3

Proof. The proof of Proposition 4.3 mainly involves two steps

- Build the relationship between $\texttt{attn}_\dagger$ and conditional mean embedding.
- Build the relationship between the $\texttt{attn}$ and conditional mean embedding.

Step 1: Build the relationship between $\texttt{attn}_\dagger$ and conditional mean embedding.

In the following, we adopt $\mathcal{H}_k$ and $\mathcal{H}_v$ to denote the RKHSs for the key and the value with the kernel functions $\mathfrak{K}$ and $\mathfrak{L}$, respectively. Also, we use $\|\cdot\|$ to denote the norm of RKHS for an element in the corresponding RKHS and the operator norm of the operators that transform elements between RKHSs. For the value space, we adopt the Euclidean kernel $\mathfrak{L}(v, v') = v^\top v'$, and the feature mapping φ is the identity mapping. Recall the definition of the empirical covariance operator and the empirical cross-covariance operator in Appendix F.1. For keys and values, we correspondingly define them as

$$\widehat{C}_{\mathcal{K}\mathcal{K}} = L^{-1}\phi(K)^\top\phi(K),\ \widehat{C}_{\mathcal{V}\mathcal{K}} = L^{-1}\varphi(V)^\top\phi(K),\ \widehat{C}_{\mathcal{V}\mathcal{V}} = L^{-1}\varphi(V)^\top\varphi(V),$$

where $\phi(K) = (\phi(k^1), \ldots, \phi(k^L))^\top \in \mathbb{R}^{L\times d_\phi}$ and $\varphi(V) = (\varphi(v^1), \ldots, \varphi(v^L))^\top \in \mathbb{R}^{L\times d_\varphi}$ By the definition of the newly defined attention in Section 4.1, we have that

$$\texttt{attn}_\dagger(q, K, V) = \widehat{C}_{\mathcal{V}\mathcal{K}}(\widehat{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\phi(q),$$

which implies that $\texttt{attn}_\dagger$ recovers the empirical conditional mean embedding. By (F.2), it holds that

$$\begin{aligned}
&\left\|\texttt{attn}_\dagger(q, K, V) - \texttt{CME}(q, \mathbb{P}_{\mathcal{K},\mathcal{V}})\right\| \\
&\quad\le \underbrace{\left\|\widehat{C}_{\mathcal{V}\mathcal{K}}(\widehat{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\phi(q) - C_{\mathcal{V}\mathcal{K}}(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\phi(q)\right\|}_{\text{(i)}} \\
&\qquad + \underbrace{\left\|C_{\mathcal{V}\mathcal{K}}(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\mathfrak{K}(q,\cdot) - C_{\mathcal{V}\mathcal{K}}C_{\mathcal{K}\mathcal{K}}^{-1}\mathfrak{K}(q,\cdot)\right\|}_{\text{(ii)}}.
\end{aligned} \tag{F.6}$$

Upper bounding term (i) of (F.6). Following the proof from Song et al. (2009), we only need to upper bound $\|\widehat{C}_{\mathcal{V}\mathcal{K}}(\widehat{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} - C_{\mathcal{V}\mathcal{K}}(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\|$. It holds that

$$\begin{aligned}
&\left\|\widehat{C}_{\mathcal{V}\mathcal{K}}(\widehat{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} - C_{\mathcal{V}\mathcal{K}}(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\right\| \\
&\quad\le \left\|\widehat{C}_{\mathcal{V}\mathcal{K}}(\widehat{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}(\widehat{C}_{\mathcal{K}\mathcal{K}} - C_{\mathcal{K}\mathcal{K}})(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\right\| + \left\|(\widehat{C}_{\mathcal{V}\mathcal{K}} - C_{\mathcal{V}\mathcal{K}})(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\right\|.
\end{aligned} \tag{F.7}$$

Considering the first term on the right-hand side of (F.7), we have the operator decomposition $\widehat{C}_{\mathcal{V}\mathcal{K}} = \widehat{C}_{\mathcal{V}\mathcal{V}}^{1/2}\mathcal{W}\widehat{C}_{\mathcal{K}\mathcal{K}}^{1/2}$ for $\mathcal{W}$ such that $\|\mathcal{W}\| \le 1$. This decomposition implies that

$$\begin{aligned}
&\left\|\widehat{C}_{\mathcal{V}\mathcal{K}}(\widehat{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}(\widehat{C}_{\mathcal{K}\mathcal{K}} - C_{\mathcal{K}\mathcal{K}})(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\right\| \\
&\le \|\widehat{C}_{\mathcal{V}\mathcal{V}}\|^{1/2}\cdot\left\|\widehat{C}_{\mathcal{K}\mathcal{K}}^{1/2}(\widehat{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1/2}\right\|\cdot\left\|(\widehat{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1/2}\right\|\cdot\left\|(\widehat{C}_{\mathcal{K}\mathcal{K}} - C_{\mathcal{K}\mathcal{K}})(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\right\| \\
&\le (L^{-1}\lambda)^{-1/2}\cdot\left\|(\widehat{C}_{\mathcal{K}\mathcal{K}} - C_{\mathcal{K}\mathcal{K}})(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\right\|,
\end{aligned} \tag{F.8}$$

where the last inequality follows from the fact that

$$\|\widehat{C}_{\mathcal{V}\mathcal{V}}\|^2 = L^{-1}\sum_{\ell=1}^{L}\|v^\ell\|_2^2 \le 1,\quad \widehat{C}_{\mathcal{K}\mathcal{K}}(\widehat{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} \le \mathcal{I},\quad (\widehat{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} \le (L^{-1}\lambda)^{-1}\mathcal{I}.$$

Combining (F.8) and (F.7), we have

$$\begin{aligned}
&\left\|\widehat{C}_{\mathcal{V}\mathcal{K}}(\widehat{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} - C_{\mathcal{V}\mathcal{K}}(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\right\| \\
&\quad\le (L^{-1}\lambda)^{-1/2}\cdot\left\|(\widehat{C}_{\mathcal{K}\mathcal{K}} - C_{\mathcal{K}\mathcal{K}})(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\right\| + \left\|(\widehat{C}_{\mathcal{V}\mathcal{K}} - C_{\mathcal{V}\mathcal{K}})(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\right\|.
\end{aligned} \tag{F.9}$$

In the following, we will upper bound the second term on the right-hand side of (F.9) with Lemma J.1. For this purpose, we define $\xi : \mathbb{R}^{d_{\mathrm{P}}} \times \mathbb{R}^{d} \to \mathcal{H}_k \otimes \mathcal{H}_v$ as follows,

$$\xi(k, v) = \varphi(v)\phi(k)^\top (C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}.$$

Since the operator norm of $(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}$ is upper bounded by $(L^{-1}\lambda)^{-1}$, we have that

$$\big\|\xi(k, v)\big\| = \big\|(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\| \cdot \big\|\varphi(v)\big\| \cdot \big\|\phi(k)\big\| \le C \cdot (L^{-1}\lambda)^{-1},$$

where $C > 0$ is an absolute constant. Additionally, we can bound the expectation of the squared norm of $\xi(k, v)$ as

$$\begin{aligned}
\mathbb{E}\Big[\big\|\xi(k, v)\big\|^2\Big] &= \mathbb{E}\Big[\big\|\phi(k)^\top (C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\|^2 \cdot \big\|\varphi(v)\big\|^2\Big] \\
&\le \mathbb{E}\Big[\big\|(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\phi(k)\big\|^2\Big] \\
&\le (L^{-1}\lambda)^{-1} \cdot \mathbb{E}\Big[\big\langle (C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\phi(k), \phi(k)\big\rangle\Big].
\end{aligned}$$

Using the definition of the trace operator, we have

$$\begin{aligned}
\mathbb{E}\Big[\big\|\xi(k, v)\big\|^2\Big] &\le \mathbb{E}\Big[\mathrm{tr}\big((C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-2}\phi(k)\phi(k)^\top\big)\Big] \\
&\le (L^{-1}\lambda)^{-1} \cdot \mathrm{tr}\big((C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}C_{\mathcal{K}\mathcal{K}}\big) \\
&= (L^{-1}\lambda)^{-1} \cdot \Gamma(L^{-1}\lambda).
\end{aligned}$$

Here $\Gamma(L^{-1}\lambda)$ is the effective dimension of $C_{\mathcal{K}\mathcal{K}}$ in Caponnetto and De Vito (2007), which is defined as follows,

$$\Gamma(L^{-1}\lambda) = \mathrm{tr}\big((C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}C_{\mathcal{K}\mathcal{K}}\big).$$

We apply Lemma J.1 with $B = C(L^{-1}\lambda)^{-1}$ and $\sigma^2 = (L^{-1}\lambda)^{-1} \cdot \Gamma(L^{-1}\lambda)$, then we have that with probability at least $1 - \delta$, the following holds

$$\big\|\widehat{C}_{\mathcal{V}\mathcal{K}}(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} - C_{\mathcal{V}\mathcal{K}}(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\| \le C \cdot \left(\frac{2}{\lambda} + \sqrt{\frac{\Gamma(L^{-1}\lambda)}{\lambda}}\right)\log\frac{2}{\delta}, \tag{F.10}$$

where $C > 0$ is an absolute constant. Similarly, we can prove that with probability at least $1 - \delta$, the following holds

$$\big\|\widehat{C}_{\mathcal{K}\mathcal{K}}(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} - C_{\mathcal{K}\mathcal{K}}(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\| \le C' \cdot \left(\frac{2}{\lambda} + \sqrt{\frac{\Gamma(L^{-1}\lambda)}{\lambda}}\right)\log\frac{2}{\delta}. \tag{F.11}$$

Here $C' > 0$ is an absolute constant. Combining (F.9), (F.10), and (F.11), we have with probability at least $1 - \delta$ that

$$\begin{aligned}
&\big\|\widehat{C}_{\mathcal{V}\mathcal{K}}(\widehat{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} - C_{\mathcal{V}\mathcal{K}}(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\| \\
&\quad\le C'' \cdot \sqrt{\frac{L}{\lambda}} \cdot \left(\frac{2}{\lambda} + \sqrt{\frac{\Gamma(L^{-1}\lambda)}{\lambda}}\right)\log\frac{2}{\delta}.
\end{aligned} \tag{F.12}$$

Upper bounding term (ii) of (F.6). We follow the procedures in the proof from Fukumizu (2015). For any $g \in \mathcal{H}_k$, we have that

$$\begin{aligned}
\langle C_{\mathcal{V}\mathcal{K}}(g), C_{\mathcal{V}\mathcal{K}}(g)\rangle &= \mathbb{E}\big[\mathfrak{L}(\mathcal{V}, \bar{\mathcal{V}})g(\mathcal{K})g(\bar{\mathcal{K}})\big] \\
&= \Big\langle (C_{\mathcal{K}\mathcal{K}} \otimes C_{\mathcal{K}\mathcal{K}})\mathbb{E}\big[\mathfrak{L}(\mathcal{V}, \bar{\mathcal{V}}) \,\big|\, \mathcal{K} = \cdot, \bar{\mathcal{K}} = \ddagger\big], g \otimes g\Big\rangle.
\end{aligned}$$

Similarly, for any $q \in \mathbb{R}^{d_{\mathrm{P}}}$ and any $g \in \mathcal{H}_k$, we have that

$$\begin{aligned}
\Big\langle C_{\mathcal{V}\mathcal{K}}, \mathbb{E}\big[\mathfrak{L}(\mathcal{V}, \cdot) \,\big|\, \mathcal{K} = q\big]\Big\rangle &= \Big\langle \mathbb{E}\big[\mathfrak{L}(\mathcal{V}, \bar{\mathcal{V}}) \,\big|\, \mathcal{K} = q, \mathcal{K} = \ddagger\big], C_{\mathcal{K}\mathcal{K}}g\Big\rangle \\
&= \Big\langle (\mathcal{I} \otimes C_{\mathcal{K}\mathcal{K}})\mathbb{E}\big[\mathfrak{L}(\mathcal{V}, \bar{\mathcal{V}}) \,\big|\, \mathcal{K} = \cdot, \bar{\mathcal{K}} = \ddagger\big], \mathfrak{L}(\cdot, q) \otimes g\Big\rangle.
\end{aligned}$$

Taking $g = (C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\mathfrak{K}(q,\cdot)$, we have that

$$\begin{aligned}
&\left\|C_{\mathcal{V}\mathcal{K}}(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\mathfrak{K}(q,\cdot) - C_{\mathcal{V}\mathcal{K}}C_{\mathcal{K}\mathcal{K}}^{-1}\mathfrak{K}(q,\cdot)\right\|^2 \\
&\quad= \Big\langle \Big((C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}C_{\mathcal{K}\mathcal{K}} \otimes (C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}C_{\mathcal{K}\mathcal{K}} - \mathcal{I}\otimes(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}C_{\mathcal{K}\mathcal{K}} \\
&\qquad (C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}C_{\mathcal{K}\mathcal{K}} \otimes \mathcal{I} + \mathcal{I}\otimes\mathcal{I}\Big)\mathbb{E}\big[\mathfrak{L}(\mathcal{V},\bar{\mathcal{V}}) \,\big|\, \mathcal{K} = \cdot, \bar{\mathcal{K}} = \ddagger\big], \mathfrak{K}(q,\cdot)\otimes\mathfrak{K}(q,\dagger)\Big\rangle.
\end{aligned}$$

We note that $\mathbb{E}[\mathfrak{L}(v,\bar{v}) \,|\, k = \cdot, \bar{k} = \ddagger] \in \mathcal{H}_k \otimes \mathcal{H}_k$ is in the range spanned by $C_{\mathcal{K}\mathcal{K}} \otimes C_{\mathcal{K}\mathcal{K}}$. Thus, we can define $\widetilde{\mathcal{C}} \in \mathcal{H}_k \times \mathcal{H}_k$ such that $(C_{\mathcal{K}\mathcal{K}} \otimes C_{\mathcal{K}\mathcal{K}})\widetilde{\mathcal{C}} = \mathbb{E}[\mathfrak{L}(v,\bar{v}) \,|\, k = \cdot, \bar{k} = \ddagger]$. Let $\{\lambda_i\}_{i=1}^{\infty}$ and $\{\varphi_i\}_{i=1}^{\infty}$ be the eigenvalues and eigenvectors of $C_{\mathcal{K}\mathcal{K}}$, respectively. We then have that

$$\begin{aligned}
&\left\|C_{\mathcal{V}\mathcal{K}}(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\mathfrak{K}(q,\cdot) - C_{\mathcal{V}\mathcal{K}}C_{\mathcal{K}\mathcal{K}}^{-1}\mathfrak{K}(q,\cdot)\right\|^4 \\
&\quad\le \left\| \Big((C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}C_{\mathcal{K}\mathcal{K}} \otimes (C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}C_{\mathcal{K}\mathcal{K}} - \mathcal{I}\otimes(C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}C_{\mathcal{K}\mathcal{K}} \right.\\
&\qquad \left. (C_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}C_{\mathcal{K}\mathcal{K}} \otimes \mathcal{I} + \mathcal{I}\otimes\mathcal{I}\Big)\mathbb{E}\big[\mathfrak{L}(\mathcal{V},\bar{\mathcal{V}}) \,\big|\, \mathcal{K} = \cdot, \bar{\mathcal{K}} = \ddagger\big] \right\|^2 \\
&\quad= \sum_{i,j}\left(\frac{\lambda_i\lambda_j(L^{-1}\lambda)^2}{(\lambda_i + L^{-1}\lambda)(\lambda_j + L^{-1}\lambda)}\right)^2 \cdot \langle \varphi_i\otimes\varphi_j, \widetilde{\mathcal{C}}\rangle^2 \\
&\quad\le (L^{-1}\lambda)^4 \cdot \|\widetilde{\mathcal{C}}\|^2.
\end{aligned}$$

Thus, we have

$$\left\|C_{\mathcal{V}\mathcal{K}}(C_{\mathcal{K}\mathcal{K}} + \lambda\mathcal{I})^{-1}\mathfrak{K}(q,\cdot) - C_{\mathcal{V}\mathcal{K}}C_{\mathcal{K}\mathcal{K}}^{-1}\mathfrak{K}(q,\cdot)\right\|_2 \le C\cdot\lambda L^{-1}, \tag{F.13}$$

where $C > 0$ is an absolute constant.

Combining (F.6), (F.12), and (F.13), we have with probability at least $1-\delta$, the following holds

$$\left\|\mathtt{attn}_{\dagger}(q,K,V) - \mathtt{CME}(q,\mathbb{P}_{\mathcal{K},\mathcal{V}})\right\| \le \mathcal{O}\left(\sqrt{\frac{L}{\lambda}}\cdot\left(\frac{2}{\lambda} + \sqrt{\frac{\Gamma(L^{-1}\lambda)}{\lambda}}\right)\log\frac{1}{\delta} + \lambda L^{-1}\right). \tag{F.14}$$

Since $\mathfrak{K}$ is Gaussian RBF kernel, we have that $\Gamma(L^{-1}\lambda) = \mathcal{O}(L/\lambda)$.

Step 2: Build the relationship between the `attn` and conditional mean embedding.

We achieve our goal in two sub-steps. In the first step, we prove that there exists a constant $C > 0$ such that

$$\mathtt{attn}_{\mathtt{SM}}(q,K,V) = C\int_{\mathbb{S}^{d-1}} v\widehat{\mathbb{P}}^{\mathfrak{K}}_{\mathcal{V}\,|\,\mathcal{K}}(v\,|\,q)\mathrm{d}v, \tag{F.15}$$

where $\mathbb{S}^{d-1}$ is the $(d-1)$-dimensional unit sphere. Here $\widehat{\mathbb{P}}^{\mathfrak{K}}_{\mathcal{V}\,|\,\mathcal{K}}$ is the kernel conditional density estimation of $\mathbb{P}_{\mathcal{V}\,|\,\mathcal{K}}$ defined as follows,

$$\widehat{\mathbb{P}}^{\mathfrak{K}}_{\mathcal{V}\,|\,\mathcal{K}}(v\,|\,q) = \frac{\sum_{\ell=1}^{L}\mathfrak{K}(k^\ell,q)\cdot\mathfrak{K}(v^\ell,v)}{\sum_{\ell=1}^{L}\mathfrak{K}(k^\ell,q)},$$

where $\iota = 1/\int_{\mathbb{S}^{d-1}}\mathfrak{K}(k,q)\mathrm{d}q$ is a normalization constant. Note that ι does not depend on the value of k by symmetry. We transform the right-hand side of this equality as

$$\begin{aligned}
\int v\widehat{\mathbb{P}}^{\mathfrak{K}}_{\mathcal{V}\,|\,\mathcal{K}}(v\,|\,q)\mathrm{d}v &= \iota\cdot\int_{\mathbb{S}^{d-1}} v\cdot\frac{\sum_{\ell=1}^{L}\mathfrak{K}(k^\ell,q)\cdot\mathfrak{K}(v^\ell,v)}{\sum_{\ell=1}^{L}\mathfrak{K}(k^\ell,q)}\mathrm{d}v \\
&= \frac{\iota\cdot\sum_{\ell=1}^{L}\mathfrak{K}(k^\ell,q)\cdot\int_{\mathbb{S}^{d-1}} v\cdot\mathfrak{K}(v^\ell,v)\mathrm{d}v}{\sum_{\ell=1}^{L}\mathfrak{K}(k^\ell,q)}.
\end{aligned} \tag{F.16}$$

Thus, it suffices to calculate the integration term $\int_{\mathbb{S}^{d-1}} v\cdot\mathfrak{K}(v^\ell,v)\mathrm{d}v$. To this end, we have the following lemma.

Proposition F.1. Let $\mathfrak{K}(a,b) = \exp(a^\top b/\gamma)$ be the exponential kernel with a fixed $\gamma > 0$. It holds for any $b \in \mathbb{S}^{d-1}$ that

$$\int_{\mathbb{S}^{d-1}} a \cdot \mathfrak{K}(a,b)\mathrm{d}a = C_1 \cdot b,$$

where $C_1 > 0$ is an absolute constant.

Proof. See Section I.1 for a detailed proof. □

Thus, it holds for the right-hand side of (F.16) that

$$\iota \cdot C_1 \cdot \frac{\sum_{\ell=1}^L \mathfrak{K}(k^\ell, q) \cdot v^\ell}{\sum_{\ell=1}^L \mathfrak{K}(k^\ell, q)} = \iota \cdot C_1 \cdot V^\top \mathtt{softmax}(Kq/\gamma) = \iota \cdot C_1 \cdot \mathtt{attn}_{\mathtt{SM}}(q, K, V),$$

where the first equality follows from the definition of the softmax function and the second equality follows from the definition of the softmax attention.

The second step is to relate the right-hand side of (F.15) to conditional mean embedding. In fact, under the condition that $\widehat{\mathbb{P}}^{\mathfrak{K}}_{\mathcal{V}\,|\,\mathcal{K}}(v\,|\,q) \to \mathbb{P}(v\,|\,q)$ uniformly for any $q \in \mathbb{S}^{d_{\mathrm{p}}-1}$ as $L \to \infty$, we have

$$\int v\widehat{\mathbb{P}}^{\mathfrak{K}}_{\mathcal{V}\,|\,\mathcal{K}}(v\,|\,q)\mathrm{d}v \to \mathbb{E}[\mathcal{V}\,|\,\mathcal{K} = q] \qquad \text{as} \quad L \to \infty.$$

Thus, we have that

$$\mathtt{attn}_{\mathtt{SM}}(q, K, V) \to C \cdot \mathbb{E}[\mathcal{V}\,|\,\mathcal{K} = q] \qquad \text{as} \quad L \to \infty \tag{F.17}$$

for some constant $C > 0$. Combining (F.17) and (F.14) and choosing $\lambda = L^{3/4}$, we complete the proof of Proposition 4.3.

□

G Appendix for Section 5

G.1 Supplemental Definitions for Markov Chains

We follow the notations in Paulin (2015). Let Ω be a Polish space. The transition kernel for a time-homogeneous Markov chain $\{X_i\}_{i=1}^\infty$ supported on Ω is a probability distribution $\mathbb{P}(x, \mathrm{d}y)$ for every $x \in \Omega$. Given $X_1 = x_1, \cdots, X_{t-1} = x_{t-1}$, the conditional distribution of X_t equals $\mathbb{P}(x_{t-1}, \mathrm{d}y)$. A distribution π is said to be a stationary distribution of this Markov chain if $\int_{x\in\Omega} \mathbb{P}(x, \mathrm{d}y)\pi(\mathrm{d}x) = \pi(\mathrm{d}y)$. We adopt $\mathbb{P}^t(x, \cdot)$ to denote the distribution of X_t conditioned on $X_1 = x$. The *mixing time* of the chain is defined by

$$d(t) = \sup_{x\in\Omega} \mathtt{TV}\big(P^t(x,\cdot), \pi\big), \quad t_{\mathrm{mix}}(\varepsilon) = \min\{t\,|\,d(t) \le \varepsilon\}, \quad t_{\mathrm{mix}} = t_{\mathrm{mix}}(1/4).$$

G.2 Proof of Theorem 5.3

Proof of Theorem 5.3. Our proof mainly involves three steps.

- Error decomposition with the PAC-Bayes framework.
- Control each term in the error decomposition.
- Conclude the proof.

Step 1: Error decomposition with the PAC-Bayes framework.

For ease of notation, we temporarily write T_{p} and N_{p} as T and N, respectively. Recall that the pretraining dataset is $\mathcal{D} = \{(S_t^n, x_{t+1}^n)\}_{n,t=1}^{N,T}$, which consists of N trajectories (essays), and each essay have $T+1$ words. Given S_t^n, the next word is generated as $x_{t+1}^n \sim \mathbb{P}(\cdot\,|\,S_t^n)$, and $S_{t+1}^n =$

(S_t^n, x_{t+1}^n). Here, we construct a ghost sample $\widetilde{\mathcal{D}} = \{(\widetilde{S}_t^n, \widetilde{x}_{t+1}^n)\}_{n,t=1}^{N,T}$ as $\widetilde{S}_t^n = S_t^n$ and $\widetilde{x}_{t+1}^n \sim \mathbb{P}(\cdot \,|\, \widetilde{S}_t^n)$ independently from $\mathcal{D}$. We define function $g(\theta) = L(\theta, D) - \log \mathbb{E}_{\widetilde{\mathcal{D}}}[\exp(L(\theta, \widetilde{\mathcal{D}})) \,|\, \mathcal{D}]$, where

$$L(\theta, \widetilde{D}) = -\frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T} \log \frac{\mathbb{P}(\widetilde{x}_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_\theta(\widetilde{x}_{t+1}^n \,|\, S_t^n)}.$$

For distributions $Q, P \in \Delta(\Theta)$, where P can potentially depends on $\mathcal{D}$, Lemma J.3 shows that

$$\mathbb{E}_P\big[g(\theta)\big] \le \mathtt{KL}(P \| Q) + \log \mathbb{E}_Q\big[\exp\big(g(\theta)\big)\big].$$

Substituting the definition of $g(\theta)$ and taking expectation with respect to the distribution of $\mathcal{D}$ on the both sides of the inequality, we can derive that

$$\mathbb{E}_{\mathcal{D}}\bigg[\exp\Big\{\mathbb{E}_P\Big[L(\theta, \mathcal{D}) - \log \mathbb{E}_{\widetilde{\mathcal{D}}}\big[\exp\big(L(\theta, \widetilde{\mathcal{D}})\big) \,|\, \mathcal{D}\big]\Big] - \mathtt{KL}(P \,\|\, Q)\Big\}\bigg] \le 1.$$

With Chernoff inequality, we can show that with probability at least $1 - \delta$, the following holds

$$-\mathbb{E}_{\theta\sim P}\Big[\log \mathbb{E}_{\widetilde{\mathcal{D}}}\big[\exp\big(L(\theta, \widetilde{\mathcal{D}})\big) \,|\, \mathcal{D}\big]\Big] \le -\mathbb{E}_P\big[L(\theta, \mathcal{D})\big] + \mathtt{KL}(P \,\|\, Q) + \log \frac{1}{\delta}. \tag{G.1}$$

We first cope with the left-hand side of (G.1).

$$\begin{aligned}
&-\mathbb{E}_P\Big[\log \mathbb{E}_{\widetilde{\mathcal{D}}}\big[\exp\big(L(\theta, \widetilde{\mathcal{D}})\big) \,|\, \mathcal{D}\big]\Big] \\
&\ge -\frac{1}{2}\log \mathbb{E}_{\widetilde{\mathcal{D}}}\bigg[\exp\bigg(-\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T}\log\frac{\mathbb{P}(\widetilde{x}_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\widehat{\theta}}(\widetilde{x}_{t+1}^n \,|\, S_t^n)}\bigg)\,\bigg|\,\mathcal{D}\bigg] \\
&\quad -\frac{1}{2}\mathbb{E}_{\theta\sim P}\bigg[\log \mathbb{E}_{\widetilde{\mathcal{D}}}\bigg[\exp\bigg(-\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T}\log\frac{\mathbb{P}_{\widehat{\theta}}(\widetilde{x}_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_\theta(\widetilde{x}_{t+1}^n \,|\, S_t^n)}\bigg)\,\bigg|\,\mathcal{D}\bigg]\bigg] \\
&= -\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T}\log \mathbb{E}_{\widetilde{x}_{t+1}^n\sim\mathbb{P}(\cdot \,|\, S_t^n)}\bigg[\exp\bigg(-\frac{1}{2}\log\frac{\mathbb{P}(\widetilde{x}_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\widehat{\theta}}(\widetilde{x}_{t+1}^n \,|\, S_t^n)}\bigg)\,\bigg|\,\mathcal{D}\bigg] \\
&\quad -\frac{1}{2}\mathbb{E}_{\theta\sim P}\bigg[\log \mathbb{E}_{\widetilde{\mathcal{D}}}\bigg[\exp\bigg(-\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T}\log\frac{\mathbb{P}_{\widehat{\theta}}(\widetilde{x}_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_\theta(\widetilde{x}_{t+1}^n \,|\, S_t^n)}\bigg)\,\bigg|\,\mathcal{D}\bigg]\bigg] \\
&\ge \frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathtt{TV}\big(\mathbb{P}(\cdot \,|\, S_t^n), \mathbb{P}_{\widehat{\theta}}(\cdot \,|\, S_t^n)\big)^2 - \frac{1}{2}\mathbb{E}_{\theta\sim P}\bigg[\log \mathbb{E}_{\widetilde{\mathcal{D}}}\bigg[\exp\bigg(-\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T}\log\frac{\mathbb{P}_{\widehat{\theta}}(\widetilde{x}_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_\theta(\widetilde{x}_{t+1}^n \,|\, S_t^n)}\bigg)\,\bigg|\,\mathcal{D}\bigg]\bigg],
\end{aligned} \tag{G.2}$$

where the first inequality results from the definition of $L(\theta, \mathcal{D})$ and Cauchy-Schwarz inequality, the equality results from that the transitions of $\widetilde{x}_{t+1}^n$ are independent given $\mathcal{D}$, and the last inequality results from Lemma J.5. The second term in the right-hand side of (G.2) can be controlled if the distribution P is chosen to concentrate around $\widehat{\theta}$. This will be done in Step 2. Now we consider the right-hand side of (G.1). For any $\theta^* \in \Theta$, we can decompose it as

$$\begin{aligned}
&-\mathbb{E}_P\big[L(\theta, \mathcal{D})\big] \\
&= \mathbb{E}_P\bigg[\frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T}\log\frac{\mathbb{P}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta^*}(x_{t+1}^n \,|\, S_t^n)} + \log\frac{\mathbb{P}_{\theta^*}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n \,|\, S_t^n)} + \log\frac{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_\theta(x_{t+1}^n \,|\, S_t^n)}\bigg] \\
&\le \frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T}\log\frac{\mathbb{P}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta^*}(x_{t+1}^n \,|\, S_t^n)} + \frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathbb{E}_P\bigg[\log\frac{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_\theta(x_{t+1}^n \,|\, S_t^n)}\bigg],
\end{aligned} \tag{G.3}$$

where the inequality results from the fact that $\widehat{\theta}$ maximizes the likelihood. We will choose θ^* as the projection of $\mathbb{P}$ onto $\{\mathbb{P}_\theta \,|\, \theta \in \Theta\}$, i.e., $\mathbb{P}_\theta^*$ is the best approximation of $\mathbb{P}$ with respect to the KL divergence. Thus, the first term in the right-hand side of (G.3) is the approximation error. The

second term in the right-hand side of (G.3) can be controlled in the same way as the second term in the right-hand side of (G.2). Combining inequalities (G.1), (G.2), and (G.3), we have that

$$\begin{aligned}
&\frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathtt{TV}\big(\mathbb{P}(\cdot\,|\,S_t^n),\mathbb{P}_{\widehat{\theta}}(\cdot\,|\,S_t^n)\big)^2 \\
&\quad\le \underbrace{\frac{1}{2}\mathbb{E}_{\theta\sim P}\left[\log\mathbb{E}_{\widetilde{\mathcal{D}}}\left[\exp\left(-\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T}\log\frac{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n\,|\,S_t^n)}{\mathbb{P}_{\theta}(x_{t+1}^n\,|\,S_t^n)}\right)\,\middle|\,\mathcal{D}\right]\right]+\frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathbb{E}_P\left[\log\frac{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n\,|\,S_t^n)}{\mathbb{P}_{\theta}(x_{t+1}^n\,|\,S_t^n)}\right]}_{\text{(I)}} \\
&\qquad+\underbrace{\frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T}\log\frac{\mathbb{P}(x_{t+1}^n\,|\,S_t^n)}{\mathbb{P}_{\theta^*}(x_{t+1}^n\,|\,S_t^n)}}_{\text{(II)}}+\underbrace{\mathtt{KL}(P\,\|\,Q)}_{\text{(III)}}+\log\frac{1}{\delta},
\end{aligned}\tag{G.4}$$

where term (I) is the fluctuation error induced by $\theta\sim P$, term (II) is the approximation error, and term (III) is the KL divergence between P and Q.

Step 2: Control each term in the error decomposition.

We first consider term (I). Since $\widehat{\theta}$ is a deterministic function of $\mathcal{D}$ and that $\log(\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n\,|\,S_t^n)/\mathbb{P}_{\theta}(x_{t+1}^n\,|\,S_t^n))$ is close to 0 if θ is close to $\widehat{\theta}$, we need to design P for any $\widehat{\theta}\in\Theta$ such that $\theta\sim P$ is close to $\widehat{\theta}$ almost surely.

We need to quantify the fluctuation of $\mathbb{P}_\theta$ when θ is changing, i.e., how $\mathbb{P}_\theta$ is close to $\mathbb{P}_{\widehat{\theta}}$ when θ is close to $\widehat{\theta}$.

Proposition G.1. For any input $X\in\mathbb{R}^{L\times d}$ and $\theta,\widetilde{\theta}\in\Theta$, we have that

$$\begin{aligned}
&\mathtt{TV}\big(\mathbb{P}_\theta(\cdot\,|\,X),\mathbb{P}_{\widetilde{\theta}}(\cdot\,|\,X)\big)\\
&\quad\le\frac{2}{\tau}\big\|A^{(D+1),\top}-\widetilde{A}^{(D+1),\top}\big\|_{1,2}+\sum_{t=1}^{D}\alpha_t(\beta_t+\iota_t+\kappa_t+\rho_t),
\end{aligned}$$

where

$$\begin{aligned}
\alpha_t&=\frac{2}{\tau}B_A(1+B_{A,1}\cdot B_{A,2})\big(1+hB_V(1+4B_QB_K)\big)^{D-t}\\
\beta_t&=|\gamma_2^{(t)}-\widetilde{\gamma}_2^{(t)}|+(1+B_{A,1}\cdot B_{A,2})\cdot\big(1+(\|X^\top\|_{2,\infty}-1)\mathbb{I}_{t=1}\big)\cdot|\gamma_1^{(t)}-\widetilde{\gamma}_1^{(t)}|\\
\iota_t&=B_{A,2}\cdot\|A_1^{(t)}-\widetilde{A}_1^{(t)}\|_{\mathrm{F}}+B_{A,1}\cdot\|A_2^{(t)}-\widetilde{A}_2^{(t)}\|_{\mathrm{F}}\\
\kappa_t&=(1+B_{A,1}\cdot B_{A,2})\cdot\big(1+(\|X^\top\|_{2,\infty}-1)\mathbb{I}_{t=1}\big)\cdot\sum_{i=1}^{h}\big\|W_i^{V,(t)}-\widetilde{W}_i^{V,(t)}\|_{\mathrm{F}}\\
\rho_t&=2(1+B_{A,1}\cdot B_{A,2})\cdot\big(1+(\|X^\top\|_{2,\infty}-1)\mathbb{I}_{t=1}\big)\cdot B_V\\
&\qquad\cdot\sum_{i=1}^{h}B_K\cdot\|W_i^{Q,(t+1)}-\widetilde{W}_i^{Q,(t+1)}\|_{\mathrm{F}}+B_Q\cdot\|W_i^{K,(t+1)}-\widetilde{W}_i^{K,(t+1)}\|_{\mathrm{F}}
\end{aligned}$$

for all $t\in[D]$.

Proof of Proposition G.1 . See Appendix I.3. □

Proposition G.1 implies that the difference between $\mathbb{P}_\theta$ and $\mathbb{P}_{\widetilde{\theta}}$ can be upper-bounded by the difference between the parameters of each layer. Thus, for any $\widehat{\theta}\in\mathcal{D}$, we set the distribution P as uniform distribution on the neighborhood of parameters, and the radius of the neighborhood is set proportional to $1/NT$ shown in Figure 9.

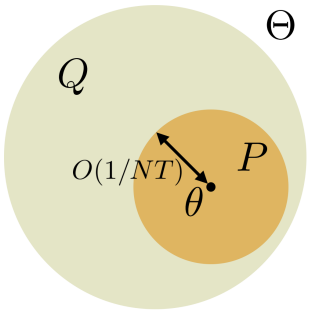

Figure 9: The distribution P in (G.5) is the uniform distribution on the neighborhood of θ with radius proportional to $1/NT$, and Q in (G.8) is the uniform distribution on Θ.

$$P = \prod_{t=1}^{D+1} \mathcal{L}_P\big(\theta^{(t)}\big) \tag{G.5}$$

$$\begin{aligned}
\mathcal{L}_P\big(\theta^{(D+1)}\big) &= \texttt{Unif}\Big(\mathbb{B}\big(\widehat{A}^{(D+1)}, r^{(D+1)}, \|\cdot\|_{1,2}\big)\Big) \\
\mathcal{L}_P\big(\theta^{(t)}\big) &= \texttt{Unif}\Big(\mathbb{B}\big(\widehat{\gamma}_1^{(t)}, r_{\gamma,1}^{(t)}, |\cdot|\big)\Big) \cdot \texttt{Unif}\Big(\mathbb{B}\big(\widehat{\gamma}_2^{(t)}, r_{\gamma,2}^{(t)}, |\cdot|\big)\Big) \cdot \mathcal{L}_P(A^{(t)}) \cdot \mathcal{L}_P(W^{(t)}) \\
\mathcal{L}_P(A^{(t)}) &= \texttt{Unif}\Big(\mathbb{B}\big(\widehat{A}_1^{(t)}, r_{A,1}^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \texttt{Unif}\Big(\mathbb{B}\big(\widehat{A}_2^{(t)}, r_{A,2}^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big) \\
\mathcal{L}_P(W^{(t)}) &= \prod_{i=1}^{h} \texttt{Unif}\Big(\mathbb{B}\big(\widehat{W}_i^{Q,(t)}, r_Q^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \texttt{Unif}\Big(\mathbb{B}\big(\widehat{W}_i^{K,(t)}, r_K^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \texttt{Unif}\Big(\mathbb{B}\big(\widehat{W}_i^{V,(t)}, r_V^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big)
\end{aligned}$$

for $t \in [D]$, where $\texttt{Unif}$ denotes the uniform distribution on the set, $\mathbb{B}(a, r, \|\cdot\|) = \{x \,|\, \|x-a\| \le r\}$ denotes the ball centered in a with radius r, the radius is set as

$$\begin{aligned}
r_{\gamma,1}^{(t)} &= R^{-1}(1 + B_{A,1} \cdot B_{A,2})^{-1}\alpha_t^{-1}/NT, & r_{\gamma,2}^{(t)} &= R^{-1}\alpha_t^{-1}/NT \\
r_{A,1}^{(t)} &= R^{-1}B_{A,2}^{-1}\alpha_t^{-1}/NT, & r_{A,2}^{(t)} &= R^{-1}B_{A,1}^{-1}\alpha_t^{-1}/NT, \\
r_V^{(t)} &= R^{-1}h^{-1}(1 + B_{A,1} \cdot B_{A,2})^{-1}\alpha_t^{-1}/NT, & r_Q^{(t)} &= R^{-1}h^{-1}(1 + B_{A,1} \cdot B_{A,2})^{-1}B_V^{-1}B_K^{-1}\alpha_t^{-1}/NT \\
r_K^{(t)} &= R^{-1}h^{-1}(1 + B_{A,1} \cdot B_{A,2})^{-1}B_V^{-1}B_Q^{-1}\alpha_t^{-1}/NT, & r^{(D+1)} &= \tau B_A^{-1}/NT.
\end{aligned}$$

Under this assignment, we now bound $|\log \mathbb{P}_{\widehat{\theta}}(x \,|\, S)/\mathbb{P}_{\theta}(x \,|\, S)|$ for any $S \in \mathbb{R}^{L \times d}$ and $x \in \mathbb{R}^{d_y}$. We first note that

$$\mathbb{P}_{\widehat{\theta}}(x \,|\, S) \ge b_y = (1 + d_y \exp(B_A/\tau))^{-1} \tag{G.6}$$

for any S and x, which results from the softmax layer defined below (5.1). This results from the fact that the last layer of the transformer is softmax with inverse temperature parameter τ and that

$$\left\| \frac{1}{L\tau} \mathbb{I}_L^\top X^{(D)} A^{(D+1)} \right\|_1 \le \left\| A^{(D+1),\top} \right\|_{1,2} \le B_A.$$

If $\texttt{TV}(\mathbb{P}_{\theta}(\cdot \,|\, S), \mathbb{P}_{\widehat{\theta}}(\cdot \,|\, S)) = \varepsilon \le b_y/2$, some basic calculations show that

$$\frac{b_y}{b_y + \varepsilon} \le \frac{\mathbb{P}_{\widehat{\theta}}(x \,|\, S)}{\mathbb{P}_{\theta}(x \,|\, S)} \le 1 + \frac{2\varepsilon}{b_y}.$$

Thus, if we set the distribution P as the uniform distribution on the neighborhood around $\widehat{\theta}$ with radius proportional to $1/NT$, i.e., (G.5), then for $\theta \sim P$ we have that

$$\left| \log \frac{\mathbb{P}_{\widehat{\theta}}(x \,|\, S)}{\mathbb{P}_{\theta}(x \,|\, S)} \right| \le \frac{2\varepsilon}{b_y} = \mathcal{O}\left(\frac{1}{NT}\right) \quad \text{for } P \text{ a.s.}$$

Based on this, we conclude that

$$\text{(I)} = \mathcal{O}(1). \tag{G.7}$$

Next, we control term (III) in (G.4). In order to upper bound $\mathtt{KL}(P \,\|\, Q)$, we need to make sure that the support of P is a subset of that of Q. Thus, we take Q as the uniform distribution on the parameter space.

$$Q = \prod_{t=1}^{D+1} \mathcal{L}_Q\big(\theta^{(t)}\big) \tag{G.8}$$

$$\begin{aligned}
\mathcal{L}_Q\big(\theta^{(D+1)}\big) &= \mathtt{Unif}\Big(\mathbb{B}\big(0, B_A, \|\cdot\|_{1,2}\big)\Big) \\
\mathcal{L}_Q\big(\theta^{(t)}\big) &= \mathtt{Unif}\Big(\mathbb{B}\big(1/2, 1/2, |\cdot|\big)\Big) \cdot \mathtt{Unif}\Big(\mathbb{B}\big(1/2, 1/2, |\cdot|\big)\Big) \cdot \mathcal{L}_Q(A^{(t)}) \cdot \mathcal{L}_Q(W^{(t)}) \\
\mathcal{L}_Q(A^{(t)}) &= \mathtt{Unif}\Big(\mathbb{B}\big(0, B_{A,1}, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \mathtt{Unif}\Big(\mathbb{B}\big(0, B_{A,2}, \|\cdot\|_{\mathrm{F}}\big)\Big) \\
\mathcal{L}_Q(W^{(t)}) &= \prod_{i=1}^{h} \mathtt{Unif}\Big(\mathbb{B}\big(0, B_Q, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \mathtt{Unif}\Big(\mathbb{B}\big(0, B_K, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \mathtt{Unif}\Big(\mathbb{B}\big(0, B_V, \|\cdot\|_{\mathrm{F}}\big)\Big).
\end{aligned}$$

Then the KL divergence between P and Q is

$$\mathtt{KL}(P \,\|\, Q) = \mathcal{O}\Big((D^2 \cdot d \cdot (d_F + d_h + d) + d \cdot d_y) \cdot \log\big(1 + NT\tau^{-1} R h B_A B_{A,1} B_{A,2} B_Q B_K B_V\big)\Big). \tag{G.9}$$

Finally, we control term (II) in (G.4). This term can be controlled as

$$\begin{aligned}
&\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} \log \frac{\mathbb{P}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta^*}(x_{t+1}^n \,|\, S_t^n)} \\
&= \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} \log \frac{\mathbb{P}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta^*}(x_{t+1}^n \,|\, S_t^n)} - \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} \mathbb{E}_{S_t^n} \mathtt{KL}\big(\mathbb{P}(\cdot \,|\, S_t^n) \,\|\, \mathbb{P}_{\theta^*}(\cdot \,|\, S_t^n)\big) \\
&\quad + \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} \mathbb{E}_{S_t^n} \mathtt{KL}\big(\mathbb{P}(\cdot \,|\, S_t^n) \,\|\, \mathbb{P}_{\theta^*}(\cdot \,|\, S_t^n)\big).
\end{aligned}$$

The first two terms in the right-hand side of the equality is the generalization error, which can be bounded with Lemma J.4. With Assumption 5.2, we note that

$$\left|\log \frac{\mathbb{P}(x \,|\, S)}{\mathbb{P}_{\theta^*}(x \,|\, S)}\right| \le b^* = \log \max\{c_0^{-1}, b_y^{-1}\}, \tag{G.10}$$

so the function satisfies the condition in Lemma J.4 with $c_i = 2b^*$. Using the moment generating function bound in Lemma J.4 and Chernoff bound, we have that

$$\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} \log \frac{\mathbb{P}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta^*}(x_{t+1}^n \,|\, S_t^n)} - \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} \mathbb{E}_{S_t^n} \mathtt{KL}\big(\mathbb{P}(\cdot \,|\, S_t^n) \,\|\, \mathbb{P}_{\theta^*}(\cdot \,|\, S_t^n)\big) \le \sqrt{\frac{t_{\min} b^{*,2}}{2NT}} \log \frac{1}{\delta} \tag{G.11}$$

with probability at least $1 - \delta$.

Step 3: Conclude the proof.

Combining inequalities (G.4), (G.7), (G.9), and (G.11), we have that

$$\begin{aligned}
&\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} \mathtt{TV}\big(\mathbb{P}(\cdot \,|\, S_t^n), \mathbb{P}_{\widehat{\theta}}(\cdot \,|\, S_t^n)\big) \\
&\le \sqrt{\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} \mathtt{TV}\big(\mathbb{P}(\cdot \,|\, S_t^n), \mathbb{P}_{\widehat{\theta}}(\cdot \,|\, S_t^n)\big)^2} \\
&= \mathcal{O}\Bigg(\frac{t_{\min}^{1/4}}{(NT)^{1/4}} \log \frac{1}{\delta} + \frac{\sqrt{D^2 d(d_F + d_h + d) + d \cdot d_y}}{\sqrt{NT}} \cdot \log\big(1 + NT\bar{B}\big) \\
&\quad + \inf_{\theta^* \in \Theta} \sqrt{\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} \mathbb{E}_{S_t^n} \mathtt{KL}\big(\mathbb{P}(\cdot \,|\, S_t^n) \,\|\, \mathbb{P}_{\theta^*}(\cdot \,|\, S_t^n)\big)}\Bigg),
\end{aligned}$$

where we take θ^* as the best approximation parameters. Finally, we will change the left-hand side of this inequality to the expectation of it. In fact, we have that

Proposition G.2. Let $\mathcal{F}$ be the collection of functions of $f : \mathbb{R}^n \to \mathbb{R}$, and we assume that $|f| \leq b$ for any function $f \in \mathcal{F}$. For a Markov chain $X = (X_1, \cdot, X_N)$, we define $f(X) = \sum_{i=1}^N f(X_i)/N$. The mixing time of this Markov chain is denoted as $t_{\text{mix}}(\varepsilon)$. Given a distribution Q on $\mathcal{F}$, with probability at least $1 - \delta$, we have

$$\left|\mathbb{E}_P\Big[\mathbb{E}_X\big[f(X)\big] - f(X)\Big]\right| \leq \sqrt{\frac{b^2 \cdot t_{\min}}{2 \log 2N}}\left[\mathtt{KL}(P \,\|\, Q) + \log\frac{4}{\delta}\right],$$

for any distribution P on $\mathcal{F}$ simultaneously with probability at least $1 - \delta$, where

$$t_{\min} = \inf_{0 \leq \varepsilon < 1} t_{\text{mix}}(\varepsilon) \cdot \left(\frac{2-\varepsilon}{1-\varepsilon}\right)^2.$$

Proof of Proposition G.2. See Appendix I.2. □

We note that Proposition G.2 is indeed an uniform convergence bound, since it holds simultaneously for all P. Thus, we can set P and Q as those in equalities (G.5) and (G.8), then we have that

$$\begin{aligned}
&\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathbb{E}_{S_t^n}\Big[\mathtt{TV}\big(\mathbb{P}(\cdot \mid S_t^n), \mathbb{P}_{\widehat{\theta}}(\cdot \mid S_t^n)\big)\Big] - \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathtt{TV}\big(\mathbb{P}(\cdot \mid S_t^n), \mathbb{P}_{\widehat{\theta}}(\cdot \mid S_t^n)\big) \\
&\quad = \mathcal{O}\left(\frac{\sqrt{t_{\min}}}{\sqrt{NT}}\left(\bar{D}\log(1 + NT\bar{B}) + \log\frac{1}{\delta}\right)\right).
\end{aligned}$$

Thus, we have that

$$\begin{aligned}
&\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathbb{E}_{S_t^n}\Big[\mathtt{TV}\big(\mathbb{P}(\cdot \mid S_t^n), \mathbb{P}_{\widehat{\theta}}(\cdot \mid S_t^n)\big)\Big] \\
&\quad = \mathcal{O}\left(\frac{t_{\min}^{1/4}}{(NT)^{1/4}}\log\frac{1}{\delta} + \frac{\sqrt{t_{\min}}}{\sqrt{NT}}\left(\bar{D}\log(1 + NT\bar{B}) + \log\frac{1}{\delta}\right)\right. \\
&\qquad \left. + \inf_{\theta^* \in \Theta}\sqrt{\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathbb{E}_{S_t^n}\mathtt{KL}\big(\mathbb{P}(\cdot \mid S_t^n) \,\|\, \mathbb{P}_{\theta^*}(\cdot \mid S_t^n)\big)}\right).
\end{aligned}$$

We conclude the proof of Theorem 5.3.

□

G.3 Formal Statement and Proof of Proposition 5.4

Denote the alphabet of the language as $\mathfrak{X} \subseteq \mathbb{R}$ ($d = 1$), then the conditional distribution $\mathbb{P}^*$ can be viewed as a function $g^* : \mathfrak{X}^L \to \mathbb{R}^{d_y}$, where L is the maximal length of a sentence, and the output is the distribution of the next word. Since $\mathcal{A}$ is finite, Theorem 2 in Zaheer et al. (2017) shows that there exist $\rho^* : \mathbb{R} \to \mathbb{R}^{d_y}$ and $\phi^* : \mathfrak{X} \to \mathbb{R}$ such that

$$g^*(X) = \rho^*\left(\frac{1}{L}\sum_{i=1}^{L}\phi^*(x_i)\right),$$

where $X = [x_1, \cdots, x_L]$. The i^{th} component of ρ^* is denoted as ρ_i^* for $i \in [d_y]$. For a function f defined on Ω, the L^∞ norm of it is defined as $\|f\|_\infty = \sup_{x \in \Omega}|f(x)|$. The set of the real-valued smooth functions on it is denoted as $\mathcal{S}^\infty(\Omega, \mathbb{R})$, Then we denote the set of the smooth functions with bounded derivatives as

$$\mathcal{S}_B = \Big\{f \in \mathcal{S}^\infty([-B, B], \mathbb{R}) \,\Big|\, \big\|f^{(n)}(x)\big\| \leq n! \text{ for all } n \in \mathbb{N}\Big\},$$

where $f^{(n)}$ is the n^{th}-order derivative of f.

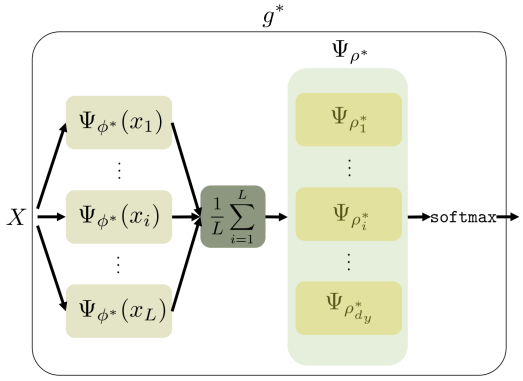

Figure 10: The construction in Proposition G.4 mainly consists of three parts: the approximation of ϕ^*, the approximation of ρ^*, and the realization of $\frac{1}{L}\sum_{i=1}^{L}$.

Assumption G.3. There exists $B > 0$ such that $\phi^*, \tau \log \rho_i^* \in \mathcal{S}_B$ for $i \in [d_y]$.

This assumption states that the function g^* is smooth enough for transformers to approximate.

Proposition G.4. Under Assumptions 5.2 and G.3, if $d_F \geq 16 d_y$, $B_{A,1} \geq 16 R d_y$, $B_{A,2} \geq d_F$ $B_A \geq \sqrt{d_y}$, and $B_V \geq \sqrt{d}$, then

$$\max_{\|S^\top\|_{2,\infty} \leq R} \mathtt{KL}\big(\mathbb{P}^*(\cdot \mid S) \,\|\, \mathbb{P}_{\theta^*}(\cdot \mid S)\big) = \mathcal{O}\bigg(d_y \exp\bigg(-\frac{D^{1/4}}{\sqrt{C^2 B^2 \log B_{A,1}}}\bigg)\bigg),$$

for some constant $C > 0$.

Proof of Proposition G.4. Our proof mainly involves three steps.

- The high-level introduction of transformer approximator for g^*.
- Build the approximators in the transformer for ϕ^* and ρ_i^* separately.
- Conclude the proof.

Step 1: The high-level introduction of transformer approximator for g^*.

Without loss of generality, we assume that $B > 1$ in Assumption G.7. We would like to first introduce our construction in a high-level way. As shown in Figure 10, we will construct Ψ_{ϕ^*} and Ψ_{ρ^*} to respectively approximate ϕ^* and $\tau \log \rho^*$.

To approximate ϕ^* with Ψ_{ϕ^*}, we will make use of the universal approximation property of the fully-connected networks and ignore the attention module in the transformer by setting $W_i^{V,(t)} = 0$, $\gamma_1^{(t)} = 1$, $\gamma_2^{(t)} = 0$ for all $i \in [h]$. We further set $A_2^{(t)} = I_{d_F} \in \mathbb{R}^{d_F \times d_F}$, which is the identity matrix. The network structure for Ψ_{ϕ^*} is

$$X^{(t+1)} = \Pi_{\mathrm{norm}}\big[\mathtt{ReLU}(X^{(t)} A_1^{(t+1)} + b^{(t+1)} \cdot \mathbb{I}_L)\big],$$

where $b^{(t+1)} \in \mathbb{R}$ is the bias term. In Step 2, we will use this fully-connected network to approximate ϕ^*.

To approximate the average $\frac{1}{L}\sum_{i=1}^{L} \phi^*(x_i)$, we take $W_i^{Q,(t)} = 0$, $W_i^{K,(t)} = 0$, and $W_i^{V,(t)} = \mathbb{I}_d$, $\gamma_1^{(t)} = 0$, $\gamma_2^{(t)} = 1$, $A_2^{(t)} = 0$.

After this average aggregation, we still take $W_i^{V,(t)} = 0$, $\gamma_1^{(t)} = 1$, $\gamma_2^{(t)} = 0$ for all $i \in [h]$ and $A_2^{(t)} = I_{d_F} \in \mathbb{R}^{d_F \times d_F}$ to approximate ρ_i^* for $i \in [d_y]$. We stack the approximators for $\tau \log \rho_i^*$ to approximate $\tau \log \rho^*$, multiplying the width of the networks by d_F.

Step 2: Build the approximators in the transformer for ϕ^* and ρ_i^* separately.

In the first and the D^{th} layer, we take $A_1^{(1),\prime} = A_1^{(1)}/R$ and $A_1^{(D),\prime} = A_1^{(D)} \cdot R$ to normalize and retrieve the magnitudes of inputs, where R is the range of the inputs. This will keep the magnitudes of the intermediate outputs small. Next, we will use Lemma J.9 to construct the networks. In the proof of Lemma J.9, the norm of the outputs of the intermediate layers do not excess the range of the inputs, so the layer normalization in our networks will not influence the constructed approximators. In this case, we can respectively approximate ϕ^* and $\tau \log \rho_i^*$ with fully-connected networks Ψ_{ϕ^*} and $\Psi_{\rho_i^*}$ for $i \in [d_y]$ as

$$\|\phi^* - \Psi_{\phi^*}\|_\infty \le \varepsilon_\phi, \quad \|\tau \log \rho_i^* - \Psi_{\rho_i^*}\|_\infty \le \varepsilon_\rho \text{ for } i \in [d_y],$$

where the depth $D(\cdot)$, the width $W(\cdot)$, and the maximal weight $B(\cdot)$ of the networks satisfy that

$$D' = D(\Psi_{\phi^*}) \le C \cdot B \cdot (\log \varepsilon_\phi^{-1})^2 + \log B, \quad D'' = \max_{i\in[d_y]} D(\Psi_{\rho_i^*}) \le C \cdot B \cdot (\log \varepsilon_\rho^{-1})^2 + \log B,$$

$$W(\Psi_{\phi^*}) \le 16, \quad W(\Psi_{\rho_i^*}) \le 16, \quad B(\Psi_{\phi^*}) \le 1, \quad B(\Psi_{\rho_i^*}) \le 1$$

for some constant $C > 0$. The bounds for width and maximal weight require that $d_F \ge 16 d_y$ and $B_{A,1} \ge \sqrt{d_F \cdot d_F} \ge 16 d_y$. Then we have that for any $X = (x_1, \cdots, x_L)$

$$\begin{aligned}
&\left\| \rho^* \left(\frac{1}{L} \sum_{i=1}^L \phi^*(x_i) \right) - \texttt{softmax}\left(\frac{1}{\tau} \Psi_{\rho^*} \left(\frac{1}{L} \sum_{i=1}^L \Psi_{\phi^*}(x_i) \right) \right) \right\|_1 \\
&\le \left\| \rho^* \left(\frac{1}{L} \sum_{i=1}^L \phi^*(x_i) \right) - \texttt{softmax}\left(\frac{1}{\tau} \Psi_{\rho^*} \left(\frac{1}{L} \sum_{i=1}^L \phi^*(x_i) \right) \right) \right\|_1 \\
&\quad + \left\| \texttt{softmax}\left(\frac{1}{\tau} \Psi_{\rho^*} \left(\frac{1}{L} \sum_{i=1}^L \phi^*(x_i) \right) \right) - \texttt{softmax}\left(\frac{1}{\tau} \Psi_{\rho^*} \left(\frac{1}{L} \sum_{i=1}^L \Psi_{\phi^*}(x_i) \right) \right) \right\|_1 \\
&\le d_y \varepsilon_\rho + C' \cdot d_y \cdot (B_{A,1})^{D''} \cdot \varepsilon_\phi, \qquad \text{(G.12)}
\end{aligned}$$

where $C' > 0$ is a constant, the first inequality results from the triangle inequality, $(B_{A,1})^{D''}$ in the second inequality results from the error propagation through a depth-D'' network and the Lipschitzeness of $\texttt{softmax}$ in Lemma J.6. This bound reflects that the later modules will amplify the approximation error in the previous modules. In the following, we will balance the depths of different modules to handle the amplification. Lemma J.9 indicates the approximation error ε of a fully-connected network will depth D can be upper bounded as

$$\varepsilon \le \exp(-\sqrt{\frac{D - \log B}{B}}).$$

Thus, defining the left-hand side of (G.12) as approx err, we have that

$$\text{approx err} \le d_y \exp\left(-\sqrt{\frac{D'' - \log B}{B}} \right) + d_y B_{A,1}^{D''} \exp\left(-\sqrt{\frac{D' - \log B}{B}} \right).$$

We note the fact that: for any $l > 0, c > 0$, we have $\exp(-l\sqrt{x - c}) = O(\exp(-l\sqrt{x}))$, which follows from the direct calculation. Then we can further upper bound the approximation error as

$$\text{approx err} = O\left(d_y \exp\left(-\sqrt{\frac{D''}{B}} \right) + d_y \exp\left(\frac{1}{\sqrt{B}} \left[D'' \sqrt{B} \log B_{A,1} - \sqrt{D'} \right] \right) \right).$$

To handle the second term in the right-hand side of this inequality, we require that

$$k \cdot D'' - \sqrt{D'} \le -\sqrt{D''},$$

where $k = \sqrt{B} \log B_{A,1}$. This is equivalent to

$$D' \ge (k \cdot D'' + \sqrt{D''})^2.$$

Since $D' + D'' \le D$, where D is the depth of the whole network, we can set

$$D'' = \sqrt{D}/(2\sqrt{B} \log B_{A,1}), \quad D' = D - 1 - D'' \ge D/2 + D^{3/4}$$

when D is large. This assignments ensure that $D' \geq (k \cdot D'' + \sqrt{D''})^2$. Thus, we have that

$$\text{approx err} = O\bigg(d_y \exp\bigg(-\sqrt{\frac{D''}{B}}\bigg)\bigg) = O\bigg(d_y \exp\bigg(-\frac{D^{1/4}}{\sqrt{C^2 B^2 \log B_{A,1}}}\bigg)\bigg)$$

for some constant $C > 0$. Here we relax the dependency on B a little for the notational clearness, and the relaxation results from the fact that $B \geq 1$ usually.

Step 3: Conclude the proof.

We denote $\Psi_{\rho^*}(\sum_{i=1}^L \Psi_{\phi^*}(x_i)/L)$ as $\mathbb{P}_{\theta^*}$. Then if $\mathtt{TV}(\mathbb{P}(\cdot \,|\, X), \mathbb{P}_{\theta^*}(\cdot \,|\, X)) = \varepsilon \leq c_0/2$, some basic calculations show that

$$\frac{c_0}{c_0 + \varepsilon} \leq \frac{\mathbb{P}(x \,|\, S)}{\mathbb{P}_{\theta^*}(x \,|\, S)} \leq 1 + \frac{2\varepsilon}{c_0}.$$

Thus, we have

$$\max_{\|S^\top\|_{2,\infty} \leq R} \mathtt{KL}\big(\mathbb{P}(\cdot \,|\, S) \,\|\, \mathbb{P}_{\theta^*}(\cdot \,|\, S)\big) \leq \frac{2\varepsilon}{c_0} = \mathcal{O}\bigg(d_y \exp\bigg(-\frac{D^{1/4}}{\sqrt{C^2 B^2 \log B_{A,1}}}\bigg)\bigg).$$

□

G.4 Pretraining Results for ℓ_2 Loss

G.4.1 Pretraining Algorithm with ℓ_2 Loss

Training with ℓ_2 loss is common in the CV community, e.g. Radford et al. (2021). The network structure is largely similar to those in Brown et al. (2020) and Devlin et al. (2018). Here, we modify the network structure of the last layer. The network derives the final output as $Y^{(D+1)} = \frac{1}{L}\mathbb{I}_L^\top X^{(D)} A^{(D+1)}$, where $\mathbb{I}_L \in \mathbb{R}^L$ is the vector with all ones, $A^{(D+1)} \in \mathbb{R}^{d \times d_y}$. The parameters in each layer are $\theta^{(t)} = (\gamma_1^{(t)}, \gamma_2^{(t)}, W^{(t)}, A^{(t)})$ for $t \in [D]$, and $\theta^{(D+1)} = A^{(D+1)}$, and the parameters of the whole network is $\theta = (\theta^{(1)}, \cdots, \theta^{(D+1)})$. Similar to Section 5.1, we consider the transformer with bounded weights. The set of parameters is

$$\Theta = \Big\{\theta \,|\, \big\|A^{(D+1)}\big\|_{\mathrm{F}} \leq B_A, \max\big\{|\gamma_1^{(t)}|, |\gamma_2^{(t)}|\big\} \leq 1, \big\|A_1^{(t)}\big\|_{\mathrm{F}} \leq B_{A,1}, \big\|A_2^{(t)}\big\|_{\mathrm{F}} \leq B_{A,2},$$
$$\big\|W_i^{Q,(t)}\big\|_{\mathrm{F}} \leq B_Q, \big\|W_i^{K,(t)}\big\|_{\mathrm{F}} \leq B_K, \big\|W_i^{V,(t)}\big\|_{\mathrm{F}} \leq B_V \text{ for all } t \in [D], i \in [h]\Big\},$$

where B_A, $B_{A,1}$, $B_{A,2}$, B_Q, B_K, and B_V are the bounds of parameter. We only consider the non-trivial case where these bounds are larger than 1, otherwise the magnitude of the output in D^{th} layer decades exponentially with growing depth. We denote the transformer with parameter θ as f_θ.

In such case, we focus on the pretraining setting in CV tasks, i.e., the pretraining set $\mathcal{D} = \{(S^i, x^i)\}_{i=1}^N$ consists of i.i.d. pairs. The underlying distribution is denoted as $(S, x) \sim \mu \in \Delta(\mathfrak{X}^* \times \mathfrak{X})$. In such case, $d = d_y$, i.e., the transformer directly predicts the musked token. The training algorithm is

$$\widehat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^N \big\|x^i - f_\theta(S^i)\big\|_2^2 \tag{G.13}$$

From the population version of (G.13), it is easy to see that the function $f^*(S) = \mathbb{E}[x \,|\, S]$ achieves the minimal population error, where the conditional expectation is defined from μ. In the following, we will quantify the error between $f_{\widehat{\theta}}$ and f^*.

G.4.2 Performance Guarantee for Pretraining with ℓ_2 Loss

We first state the assumptions for the pretraining setting.

Assumption G.5. There exists a constant $R > 0$ such that for $(S, x) \sim \mu$, we have $\|S^\top\|_{2,\infty} \leq R$ and $\|x\|_2 \leq B_x$ almost surely.

Then the performance guarantee for the pretraining result $\widehat{\theta}$ can be derived as following.

Theorem G.6. Let $\bar{B} = B_x R h B_A B_{A,1} B_{A,2} B_Q B_K B_V$ and $\bar{D} = D^2 d(d_F + d_h + d) + d \cdot d_y$. If Assumption G.5 holds, the pretrained model $f_{\widehat{\theta}}$ by the algorithm in (G.13) satisfies

$$\mathbb{E}_{S,x}\Big[\big\|f^*(S) - f_{\widehat{\theta}}(S)\big\|_2^2\Big] \le \underbrace{\frac{3}{2}\min_{\theta\in\Theta}\mathbb{E}\Big[\big\|f^*(S) - f_\theta(S)\big\|_2^2\Big]}_{\text{approximation error}} + \underbrace{\mathcal{O}\bigg(\frac{B_x^2}{N}\Big[\bar{D}\log(1+N\bar{B}) + \log\frac{2}{\delta}\Big]\bigg)}_{\text{generalization error}}$$

with probability at least $1-\delta$.

The first term is the approximation error. It measures the proximity between the nominal function f^* and the functions induced by the parameter set Θ. The second term is the generalization error. Similar as Theorem 5.3, the generalization error is independent of the token sequence length.

Since the neural networks are universal approximators, we will explicitly approximate f^* from the transformer function class. Theorem 2 in Zaheer et al. (2017) shows that there exist $\rho^* : \mathbb{R} \to \mathbb{R}^{d_y}$ and $\phi^* : \mathbb{R} \to \mathbb{R}$ such that

$$f^*(X) = \rho^*\bigg(\frac{1}{L}\sum_{i=1}^{L}\phi^*(x_i)\bigg),$$

where $X = [x_1, \cdots, x_L]$. The i^{th} component of ρ^* is denoted as ρ_i^* for $i \in [d_y]$. For a function f defined on Ω, the L^∞ norm of it is defined as $\|f\|_\infty = \sup_{x\in\Omega}|f(x)|$. The set of the real-valued smooth functions on it is denoted as $\mathcal{S}^\infty(\Omega, \mathbb{R})$, Then we denote the set of the smooth functions with bounded derivatives as

$$\mathcal{S}_B = \Big\{f \in \mathcal{S}^\infty([-B, B], \mathbb{R}) \,|\, \big\|f^{(n)}(x)\big\| \le n! \text{ for all } n \in \mathbb{N}\Big\},$$

where $f^{(n)}$ is the n^{th}-order derivative of f.

Assumption G.7. There exists $B > 0$ such that $\phi^*, \rho_i^* \in \mathcal{S}_B$ for $i \in [d_y]$.

This assumption states that the function f^* is smooth enough. Then we have that

Proposition G.8. Under G.7, if $d_F \ge 16 d_y$, $B_{A,1} \ge 16 R d_y$, $B_{A,2} \ge d_F$ $B_A \ge \sqrt{d_y}$, and $B_V \ge \sqrt{d}$, then

$$\max_{\|S^\top\|_{2,\infty}\le R}\big\|f^*(S) - f_{\theta^*}(S)\big\|_2 = \mathcal{O}\bigg(d_y \exp\Big(-\frac{D^{1/4}}{\sqrt{C^2 B^2 \log B_{A,1}}}\Big)\bigg)$$

for some constant $C > 0$.

G.4.3 Proof of Theorem G.6

Proof of Theorem G.6. For ease of notation, we respectively define the empirical risk and the population risk as

$$\widehat{\mathcal{L}}(f, \mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\big\|x^i - f_\theta(S^i)\big\|_2^2, \quad \mathcal{L}(f) = \mathbb{E}_{S,x}\Big[\big\|x - f_\theta(S)\big\|_2^2\Big].$$

The our proof mainly involves three steps.

- Error decomposition for the excess population risk.
- Control each term in the error decomposition.
- Conclude the proof.

Step 1: Error decomposition for the excess population risk. The excess population risk for the estimate $\widehat{\theta}$ can be decomposed to the sum of the generalization error and the approximation error as

$$\begin{aligned}
&\mathcal{L}(f_{\widehat{\theta}}) - \mathcal{L}(f^*) \\
&= \mathcal{L}(f_{\widehat{\theta}}) - \mathcal{L}(f^*) - 2\big(\widehat{\mathcal{L}}(f_{\widehat{\theta}}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D})\big) + 2\big(\widehat{\mathcal{L}}(f_{\widehat{\theta}}, \mathcal{D}) - \widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D})\big) + 2\big(\widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D})\big) \\
&\le \underbrace{\mathcal{L}(f_{\widehat{\theta}}) - \mathcal{L}(f^*) - 2\big(\widehat{\mathcal{L}}(f_{\widehat{\theta}}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D})\big)}_{\text{generalization error}} + \underbrace{2\big(\widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D})\big)}_{\text{approximation error}},
\end{aligned} \tag{G.14}$$

where $\theta^* = \operatorname{argmin}_{\theta\in\Theta} \mathcal{L}(f_\theta)$, and the inequality results from that $\widehat{\theta}$ achieves the minimal empirical risk.

Step 2: Control each term in the error decomposition.

We first consider the generalization error and will adapt Lemma J.2 to bound it. Define the function

$$g(S, x, \theta) = \left\|x - f_\theta(S)\right\|_2^2 - \left\|x - f^*(S)\right\|_2^2.$$

To verify the conditions in Lemma J.2, we notice that $|g(S, x, \theta)| \le (B_x + B_f)^2$ and that

$$\begin{aligned}
\mathbb{E}\big[g(S, x, \theta)\big] &= \mathbb{E}\Big[\left\|x - f_\theta(S)\right\|_2^2 - \left\|x - f^*(S)\right\|_2^2\Big] \\
&= \mathbb{E}\Big[\left\|f^*(S) - f_\theta(S)\right\|_2^2\Big] \\
\mathbb{E}\Big[\big(g(S, x, \theta) - \mathbb{E}\big[g(S, x, \theta)\big]\big)^2\Big] &\le \mathbb{E}\Big[\big(g(S, x, \theta)\big)^2\Big] \\
&\le \mathbb{E}\Big[\left\|2x - f^*(S) - f_\theta(S)\right\|_2^2 \cdot \left\|f^*(S) - f_\theta(S)\right\|_2^2\Big] \\
&\le (3B_x + B_f)^2 \cdot \mathbb{E}\Big[\left\|f^*(S) - f_\theta(S)\right\|_2^2\Big],
\end{aligned}$$

where the second equality results from the definition of f^*, the second inequality results from Cauchy–Schwarz inequality, and the last inequality result from the boundedness of x, f^*, and f_θ. Then Lemma J.2 shows that for a distribution $Q \in \Delta(\Theta)$ and $0 < \lambda \le 1/(2(B_x + B_f)^2)$, the following holds with probability at least $1 - \delta$ simultaneously for all $P \in \Delta(\Theta)$

$$\begin{aligned}
&\left|\mathbb{E}_{\theta\sim P}\left[\mathbb{E}\big[g(S, x, \theta)\big] - \frac{1}{N}\sum_{i=1}^N g(S^i, x^i, \theta)\right]\right| \\
&\quad\le \lambda(3B_x + B_f)^2 \mathbb{E}_{\theta\sim P}\Big[\mathbb{E}\big[g(S, x, \theta)\big]\Big] + \frac{1}{N\lambda}\left[\mathtt{KL}(P \,\|\, Q) + \log\frac{2}{\delta}\right].
\end{aligned}$$

Taking $\lambda = 1/(2(3B_x + B_f)^2)$, we have

$$\begin{aligned}
&\left|\mathbb{E}_{\theta\sim P}\Big[\mathcal{L}(f_\theta) - \mathcal{L}(f^*) - \big(\widehat{\mathcal{L}}(f_\theta, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D})\big)\Big]\right| \\
&\quad\le \frac{1}{2}\mathbb{E}_{\theta\sim P}\big[\mathcal{L}(f_\theta) - \mathcal{L}(f^*)\big] + \frac{2(3B_x + B_f)^2}{N}\left[\mathtt{KL}(P \,\|\, Q) + \log\frac{2}{\delta}\right].
\end{aligned}$$

Next, we will take proper P and Q to relate this equation and the generalization error. For this purpose, we quantify how the perturbation of network parameters influence the output of the network.

Proposition G.9. For any input $X \in \mathbb{R}^{L\times d}$ and $\theta, \widetilde{\theta} \in \Theta$, we have that

$$\|f_\theta(X) - f_{\widetilde{\theta}}(X)\|_2 \le \left\|A^{(D+1)} - \widetilde{A}^{(D+1)}\right\|_{\mathrm{F}} + \sum_{t=1}^D \alpha_t(\beta_t + \iota_t + \kappa_t + \rho_t),$$

where

$$\begin{aligned}
\alpha_t &= B_A(1 + B_{A,1}\cdot B_{A,2})\big(1 + hB_V(1 + 4B_Q B_K)\big)^{D-t} \\
\beta_t &= |\gamma_2^{(t)} - \widetilde{\gamma}_2^{(t)}| + (1 + B_{A,1}\cdot B_{A,2})\cdot\big(1 + (\|X^\top\|_{2,\infty} - 1)\mathbb{I}_{t=1}\big)\cdot|\gamma_1^{(t)} - \widetilde{\gamma}_1^{(t)}| \\
\iota_t &= B_{A,2}\cdot\|A_1^{(t)} - \widetilde{A}_1^{(t)}\|_{\mathrm{F}} + B_{A,1}\cdot\|A_2^{(t)} - \widetilde{A}_2^{(t)}\|_{\mathrm{F}} \\
\kappa_t &= (1 + B_{A,1}\cdot B_{A,2})\cdot\big(1 + (\|X^\top\|_{2,\infty} - 1)\mathbb{I}_{t=1}\big)\cdot\sum_{i=1}^h \left\|W_i^{V,(t)} - \widetilde{W}_i^{V,(t)}\right\|_{\mathrm{F}} \\
\rho_t &= 2(1 + B_{A,1}\cdot B_{A,2})\cdot\big(1 + (\|X^\top\|_{2,\infty} - 1)\mathbb{I}_{t=1}\big)\cdot B_V \\
&\qquad\cdot\sum_{i=1}^h B_K\cdot\|W_i^{Q,(t+1)} - \widetilde{W}_i^{Q,(t+1)}\|_{\mathrm{F}} + B_Q\cdot\|W_i^{K,(t+1)} - \widetilde{W}_i^{K,(t+1)}\|_{\mathrm{F}}
\end{aligned}$$

for all $t \in [D]$.

Proof of Proposition G.9. See Appendix I.4. □

With the help of Proposition G.9, we set the distribution P as

$$P = \prod_{t=1}^{D+1} \mathcal{L}_P\big(\theta^{(t)}\big) \tag{G.15}$$

$$\begin{aligned}
\mathcal{L}_P\big(\theta^{(D+1)}\big) &= \texttt{Unif}\Big(\mathbb{B}\big(\widehat{A}^{(D+1)}, r^{(D+1)}, \|\cdot\|_{\mathrm{F}}\big)\Big) \\
\mathcal{L}_P\big(\theta^{(t)}\big) &= \texttt{Unif}\Big(\mathbb{B}\big(\widehat{\gamma}_1^{(t)}, r_{\gamma,1}^{(t)}, |\cdot|\big)\Big) \cdot \texttt{Unif}\Big(\mathbb{B}\big(\widehat{\gamma}_2^{(t)}, r_{\gamma,2}^{(t)}, |\cdot|\big)\Big) \cdot \mathcal{L}_P(A^{(t)}) \cdot \mathcal{L}_P(W^{(t)}) \\
\mathcal{L}_P(A^{(t)}) &= \texttt{Unif}\Big(\mathbb{B}\big(\widehat{A}_1^{(t)}, r_{A,1}^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \texttt{Unif}\Big(\mathbb{B}\big(\widehat{A}_2^{(t)}, r_{A,2}^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big) \\
\mathcal{L}_P(W^{(t)}) &= \prod_{i=1}^{h} \texttt{Unif}\Big(\mathbb{B}\big(\widehat{W}_i^{Q,(t)}, r_Q^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \texttt{Unif}\Big(\mathbb{B}\big(\widehat{W}_i^{K,(t)}, r_K^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \texttt{Unif}\Big(\mathbb{B}\big(\widehat{W}_i^{V,(t)}, r_V^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big)
\end{aligned}$$

for $t \in [D]$, where $\texttt{Unif}$ denotes the uniform distribution on the set, $\mathbb{B}(a, r, \|\cdot\|) = \{x \,|\, \|x-a\| \le r\}$ denotes the ball centered in a with radius r, the radius is set as

$$\begin{aligned}
r_{\gamma,1}^{(t)} &= (B_x + B_f)^{-1} R^{-1} (1 + B_{A,1} \cdot B_{A,2})^{-1} \alpha_t^{-1} / N, & r_{\gamma,2}^{(t)} &= (B_x + B_f)^{-1} R^{-1} \alpha_t^{-1} / N \\
r_{A,1}^{(t)} &= (B_x + B_f)^{-1} R^{-1} B_{A,2}^{-1} \alpha_t^{-1} / N, & r_{A,2}^{(t)} &= (B_x + B_f)^{-1} R^{-1} B_{A,1}^{-1} \alpha_t^{-1} / N, \\
r_V^{(t)} &= (B_x + B_f)^{-1} R^{-1} h^{-1} (1 + B_{A,1} \cdot B_{A,2})^{-1} \alpha_t^{-1} / N, & r^{(D+1)} &= (B_x + B_f)^{-1} B_A^{-1} / N, \\
r_K^{(t)} &= (B_x + B_f)^{-1} R^{-1} h^{-1} (1 + B_{A,1} \cdot B_{A,2})^{-1} B_V^{-1} B_Q^{-1} \alpha_t^{-1} / N, \\
r_Q^{(t)} &= (B_x + B_f)^{-1} R^{-1} h^{-1} (1 + B_{A,1} \cdot B_{A,2})^{-1} B_V^{-1} B_K^{-1} \alpha_t^{-1} / N.
\end{aligned}$$

Under this assignment, we now bound $\mathbb{E}_{\theta\sim P}[\|x - f_\theta(S)\|_2^2 - \|x - f_{\widehat{\theta}}(S)\|_2^2]$ as

$$\Big|\mathbb{E}_{\theta\sim P}\Big[\big\|x - f_\theta(S)\big\|_2^2 - \big\|x - f_{\widehat{\theta}}(S)\big\|_2^2\Big]\Big| \le 2(B_x + B_f)\Big|\mathbb{E}_{\theta\sim P}\Big[\big\|f_\theta(S) - f_{\widehat{\theta}}(S)\big\|_2\Big]\Big| = \mathcal{O}\Big(\frac{B_x + B_f}{N}\Big),$$

where the inequality results from Cauchy-Schwarz inequality, and the equality results from Proposition G.9. Thus, we have that

$$\begin{aligned}
&\mathcal{L}(f_{\widehat{\theta}}) - \mathcal{L}(f^*) - \big(\widehat{\mathcal{L}}(f_{\widehat{\theta}}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D})\big) \\
&\quad \le \frac{1}{2}\big(\mathcal{L}(f_{\widehat{\theta}}) - \mathcal{L}(f^*)\big) + \mathcal{O}\Big(\frac{B_x + B_f}{N}\Big) + \frac{2(3B_x + B_f)^2}{N}\Big[\mathtt{KL}(P \,\|\, Q) + \log\frac{2}{\delta}\Big].
\end{aligned} \tag{G.16}$$

To access to the value of $\mathtt{KL}(P \,\|\, Q)$, we take Q as the distribution in (G.8) except that

$$\mathcal{L}_Q\big(\theta^{(D+1)}\big) = \texttt{Unif}\Big(\mathbb{B}\big(0, B_A, \|\cdot\|_{\mathrm{F}}\big)\Big). \tag{G.17}$$

Then the KL divergence between P and Q is

$$\mathtt{KL}(P \,\|\, Q) = \mathcal{O}\Big((D^2 \cdot d \cdot (d_F + d_h + d) + d \cdot d_y) \cdot \log\big(1 + NB_x RhB_A B_{A,1} B_{A,2} B_Q B_K B_V\big)\Big).$$

Combining this equality with (G.16), we have that with probability at least $1 - \delta$, the generalization error can be bounded as

$$\mathcal{L}(f_{\widehat{\theta}}) - \mathcal{L}(f^*) - 2\big(\widehat{\mathcal{L}}(f_{\widehat{\theta}}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D})\big) = \mathcal{O}\Big(\frac{B_x^2}{N}\Big[\bar{D}\log(1 + N\bar{B}) + \log\frac{2}{\delta}\Big]\Big). \tag{G.18}$$

Next we control the approximation error in (G.14).

$$\begin{aligned}
&\widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D}) \\
&\quad = \widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D}) - \frac{3}{2}\big(\mathcal{L}(f_{\theta^*}) - \mathcal{L}(f^*)\big) + \frac{3}{2}\big(\mathcal{L}(f_{\theta^*}) - \mathcal{L}(f^*)\big) \\
&\quad = \widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D}) - \frac{3}{2}\big(\mathcal{L}(f_{\theta^*}) - \mathcal{L}(f^*)\big) + \frac{3}{2}\mathbb{E}\Big[\big\|f^*(S) - f_{\theta^*}(S)\big\|_2^2\Big],
\end{aligned} \tag{G.19}$$

where the second equality results from the definition of f^*. To bound the first two terms in the right-hand side of (G.19), we use Lemma J.2 and take P and Q as (G.15) and (G.17), replacing $\widehat{\theta}$ by θ^*. Then we have that

$$\widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D}) - \frac{3}{2}\big(\mathcal{L}(f_{\theta^*}) - \mathcal{L}(f^*)\big) = \mathcal{O}\bigg(\frac{B_x^2}{N}\bigg[\bar{D}\log(1 + N\bar{B}) + \log\frac{2}{\delta}\bigg]\bigg). \tag{G.20}$$

Step 3: Conclude the proof.

Combining inequalities (G.14), (G.18), (G.19), and (G.20), we have that

$$\mathcal{L}(f_{\widehat{\theta}}) - \mathcal{L}(f^*) = \frac{3}{2}\mathbb{E}\Big[\big\|f^*(S) - f_{\theta^*}(S)\big\|_2^2\Big] + \mathcal{O}\bigg(\frac{B_x^2}{N}\bigg[\bar{D}\log(1 + N\bar{B}) + \log\frac{2}{\delta}\bigg]\bigg).$$

Thus, we conclude the proof of Theorem G.6.

□

G.4.4 Proof of Proposition G.8

Proof of Proposition G.8. Our proof mainly involves three steps.

- Build the high-level transformer approximator for f^*.
- Build the approximators in the transformer for ϕ^* and ρ_i^* separately.
- Conclude the proof.

The first two steps follow the procedures of the proof of Proposition G.4 exactly. Now we present the final step.

Step 3: Conclude the proof.

In the final layer, we just take $A^{(D+1)} = I_{d_y}$ as the identity matrix. Denoting the derived parameters as θ^* we have that

$$\max_{\|X^\top\|_{2,\infty}\le R}\bigg\|\rho^*\bigg(\frac{1}{L}\sum_{i=1}^{L}\phi^*(x_i)\bigg) - f_{\theta^*}(X)\bigg\|_2 = \mathcal{O}\bigg(d_y\exp\bigg(-\frac{D^{1/4}}{\sqrt{C^2B^2\log B_{A,1}}}\bigg)\bigg).$$

Thus, we conclude the proof of Proposition G.8. □

H Proofs and Formal Statements for §6

H.1 Proof of Theorem 6.2

Proof. By Corollary 4.2 and the fact that $\log(1/p_0(z_*)) \le \beta$, we have that

$$T^{-1}\cdot\mathbb{E}_{\mathcal{D}_{\texttt{ICL}}}\bigg[\sum_{t=1}^{T}\log\mathbb{P}(r_t\,|\,z^*, \texttt{prompt}_{t-1}) - \sum_{t=1}^{T}\log\mathbb{P}(r_t\,|\,\texttt{prompt}_{t-1})\bigg] \le \beta/T. \tag{H.1}$$

In addition, we have that

$$T^{-1}\cdot\mathbb{E}_{\mathcal{D}_{\texttt{ICL}}}\bigg[\sum_{t=1}^{T}\log\mathbb{P}(r_t\,|\,\texttt{prompt}_{t-1}) - \sum_{t=1}^{T}\log\mathbb{P}_{\widehat{\theta}}(r_t\,|\,\texttt{prompt}_{t-1})\bigg] = \mathbb{E}_{\mathcal{D}_{\texttt{ICL}}}\Big[\texttt{KL}\big(\mathbb{P}(\cdot\,|\,\texttt{prompt})\,\big\|\,\mathbb{P}_{\widehat{\theta}}(\cdot\,|\,\texttt{prompt})\big)\Big]. \tag{H.2}$$

Similar to (G.10), we have that

$$\Big|\log\big(\mathbb{P}(r\,|\,\texttt{prompt})/\mathbb{P}_{\widehat{\theta}}(r\,|\,\texttt{prompt})\big)\Big| \le b^* = \log\max\{c_0^{-1}, b_y^{-1}\}.$$

By Lemma J.10, we have that

$$\texttt{KL}\big(\mathbb{P}(\cdot\,|\,\texttt{prompt})\,\|\,\mathbb{P}_{\widehat{\theta}}(\cdot\,|\,\texttt{prompt})\big) \le (3 + b^*)/2\cdot\texttt{TV}\big(\mathbb{P}(\cdot\,|\,\texttt{prompt}), \mathbb{P}_{\widehat{\theta}.}(\cdot\,|\,\texttt{prompt})\big). \tag{H.3}$$

By Assumption 6.1, we have that $\mathbb{P}_{\mathcal{D}_{\text{ICL}}}(\texttt{prompt}) \le \kappa \mathbb{P}_{\mathcal{D}}(\texttt{prompt})$. Thus, by Theorem 5.3, we have with probability at least $1-\delta$ that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{\text{ICL}}}\Big[\texttt{KL}\big(\mathbb{P}(\cdot \mid \texttt{prompt}) \,\|\, \mathbb{P}_{\widehat{\theta}}(\cdot \mid \texttt{prompt})\big)\Big] \\
&\quad \le C \cdot b^* \cdot \kappa \cdot \mathbb{E}_{S\sim\mathcal{D}}\Big[\texttt{TV}\big(\mathbb{P}(\cdot \mid S), \mathbb{P}_{\widehat{\theta}}(\cdot \mid S)\big)\Big] \le C \cdot b^* \cdot \kappa \cdot \Delta_{\text{pre}}(N, T, \delta). \qquad \text{(H.4)}
\end{aligned}$$

Combining (H.4), (H.1), and (H.2), we have with probability at least $1-\delta$ that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{\text{ICL}}}\Big[T^{-1} \cdot \sum_{t=1}^{T} \log \mathbb{P}(r_t \mid z^*, \texttt{prompt}_{t-1}) - T^{-1} \cdot \sum_{t=1}^{T} \log \mathbb{P}_{\widehat{\theta}}(r_t \mid \texttt{prompt}_{t-1})\Big] \\
&\quad \le \beta/T + \mathbb{E}_{S\sim\mathcal{D}}\Big[\texttt{KL}\big(\mathbb{P}(\cdot \mid S) \,\|\, \mathbb{P}_{\widehat{\theta}}(\cdot \mid S)\big)\Big] \\
&\quad \le \mathcal{O}\big(\beta/T + b^* \cdot \kappa \cdot \Delta_{\text{pre}}(N, T, \delta)\big), \qquad \text{(H.5)}
\end{aligned}$$

which completes the proof of Theorem 6.2. □

H.2 Assumptions and Formal Statement for Prompting With Wrong Input-Output Mappings

We first state assumptions for this setting.

Assumption H.1. Conditioned on any $z \in \mathfrak{Z}$, the input-output pairs are independent, i.e., for any two input-output pair sequences $S_t, S'_{t'} \in \mathfrak{X}^*$, we have $\mathbb{P}((S_t, S'_{t'}) \mid z) = \mathbb{P}(S_t \mid z) \cdot \mathbb{P}(S'_{t'} \mid z)$.
This assumption states that for any task $z \in \mathfrak{Z}$, the input-output pairs are independently generated. This largely holds in realistic applications since the examples usually are independently produced. It can be relaxed when there are more structures in the token generation process, e.g. the hidden Markov model in Xie et al. (2021).

Assumption H.2. There exists a constant $c_1 > 0$ such that $\mathbb{P}_{\mathcal{Z}}(z_*) \ge c_1$.
This assumption states that the prior distribution of the hidden concept z_* is strictly larger than 0, otherwise this concept can never be deduced. For two concepts $z, z' \in \mathfrak{Z}$, we define the KL divergence between the conditional distributions of input-output pair on them as $\texttt{KL}_{\text{pair}}(\mathbb{P}(\cdot \mid z) \| \mathbb{P}(\cdot \mid z')) = \mathbb{E}_{X,y\sim\mathbb{P}(\cdot \mid z)}[\log(\mathbb{P}(X, y \mid z)/\mathbb{P}(X, y \mid z'))]$. This divergence measures the distance between distributions of input-output pairs conditioned on different tasks z and z'.

Assumption H.3. The concept z_* satisfies that $\min_{z\neq z_*} \texttt{KL}_{\text{pair}}(\mathbb{P}(\cdot \mid z_*) \,\|\, \mathbb{P}(\cdot \mid z)) > 2 \log 1/c_0$, where c_0 is the constant in Assumption 5.2.
This distinguishability assumption requires that the divergence between z_* and other concepts z is large enough to infer the concept z_* from the prompt. We denote the pretraining error in Theorem 5.3 as $\Delta_{\text{pre}}(N_{\text{p}}, T_{\text{p}}, \delta)$, then we have the following result.

Proposition H.4. Under Assumptions 5.2, 6.1 H.1, H.2, and H.3, the pretrained model $\mathbb{P}_{\widehat{\theta}}$ in (5.2) predicts the outputs with the prompt containing wrong mappings as

$$\begin{aligned}
&\mathbb{E}_{\texttt{prompt}'}\Big[\texttt{KL}\big(\mathbb{P}(\cdot \mid \widetilde{c}_{t+1}, z_*) \| \mathbb{P}_{\widehat{\theta}}(\cdot \mid S'_t, \widetilde{c}_{t+1})\big)\Big] \\
&= \mathcal{O}\Bigg(\kappa \Delta_{\text{pre}}(N_{\text{p}}, T_{\text{p}}, \delta) + \exp\Bigg(-\frac{\sqrt{t}}{2(1+l)\log 1/c_0}\Big(\min_{z\neq z^*} \texttt{KL}_{\text{pair}}\big(\mathbb{P}(\cdot \mid z^*) \,\|\, \mathbb{P}(\cdot \mid z)\big) + 2\log c_0\Big)\Bigg)\Bigg)
\end{aligned}$$

with probability at least $1-\delta$.

H.3 Proof of Proposition H.4

Proof of Proposition H.4. From Bayesian model averaging, the output distribution is

$$\begin{aligned}
&\mathbb{P}(r_{t+1} \mid S'_t, \widetilde{c}_{t+1}) \\
&= \sum_{z\in\mathfrak{Z}} \mathbb{P}(r_{t+1} \mid \widetilde{c}_{t+1}, z) \cdot \mathbb{P}_{\mathcal{Z}}(z \mid S'_t) \\
&= \mathbb{P}(r_{t+1} \mid \widetilde{c}_{t+1}, z^*) + \sum_{z\neq z^*} \big(\mathbb{P}(r_{t+1} \mid \widetilde{c}_{t+1}, z) - \mathbb{P}(r_{t+1} \mid \widetilde{c}_{t+1}, z^*)\big) \cdot \mathbb{P}_{\mathcal{Z}}(z \mid S'_t) \\
&= \mathbb{P}(r_{t+1} \mid \widetilde{c}_{t+1}, z^*) + \sum_{z\neq z^*} \big(\mathbb{P}(r_{t+1} \mid \widetilde{c}_{t+1}, z) - \mathbb{P}(r_{t+1} \mid \widetilde{c}_{t+1}, z^*)\big) \cdot \mathbb{P}_{\mathcal{Z}}(z^* \mid S'_t) \cdot \frac{\mathbb{P}_{\mathcal{Z}}(z) \cdot \mathbb{P}(S'_t \mid z)}{\mathbb{P}_{\mathcal{Z}}(z^*) \cdot \mathbb{P}(S'_t \mid z^*)}, \qquad \text{(H.6)}
\end{aligned}$$

where the first equality results from Bayesian model averaging, the last equality results from Bayes' theorem. Next, we upperbound the ratio $\mathbb{P}(S'_t \mid z)/\mathbb{P}(S'_t \mid z^*)$ in the right-hand side of Eqn. (H.6). We have that

$$\frac{1}{t}\log\frac{\mathbb{P}(S'_t \mid z)}{\mathbb{P}(S'_t \mid z^*)} = \frac{1}{t}\sum_{i=1}^{t}\log\frac{\mathbb{P}\big((\widetilde{c}_i, r'_i) \mid z\big)}{\mathbb{P}\big((\widetilde{c}_i, r'_i) \mid z^*\big)} \leq -2\log c_0 + \frac{1}{t}\sum_{i=1}^{t}\log\frac{\mathbb{P}\big((\widetilde{c}_i, r_i) \mid z\big)}{\mathbb{P}\big((\widetilde{c}_i, r_i) \mid z^*\big)},$$

where the equality results from Assumption H.1, and the inequality results from Assumption 5.2. Assumption 5.2 also implies that $|\log \mathbb{P}((\widetilde{c}_i, r_i) \mid z)/\mathbb{P}((\widetilde{c}_i, r_i) \mid z^*)| \leq (1+l)\log 1/c_0$. Hoeffding inequality shows that with probability at least $1-\delta$, we have

$$\frac{1}{t}\sum_{i=1}^{t}\log\frac{\mathbb{P}\big((\widetilde{c}_i, r_i) \mid z\big)}{\mathbb{P}\big((\widetilde{c}_i, r_i) \mid z^*\big)} + \mathtt{KL}_{\mathrm{pair}}\big(\mathbb{P}(\cdot \mid z^*) \,\|\, \mathbb{P}(\cdot \mid z)\big) \leq \frac{(1+l)}{\sqrt{t}}\log\frac{1}{c_0}\cdot\log\frac{1}{\delta}.$$

Thus, we have that with probability at least $1-\delta$, the following holds for all $z \neq z^*$

$$\frac{\mathbb{P}(S'_t \mid z)}{\mathbb{P}(S'_t \mid z^*)} \leq \exp\left(-t\left(\mathtt{KL}_{\mathrm{pair}}\big(\mathbb{P}(\cdot \mid z^*) \,\|\, \mathbb{P}(\cdot \mid z)\big) + 2\log c_0 - \frac{(1+l)}{\sqrt{t}}\log\frac{1}{c_0}\cdot\log\frac{|\mathfrak{Z}|}{\delta}\right)\right).$$

Combining this inequality with Eqn. (H.6), we have that

$$\begin{aligned}
&\mathtt{TV}\big(\mathbb{P}(\cdot \mid S'_t, \widetilde{c}_{t+1}), \mathbb{P}(\cdot \mid \widetilde{c}_{t+1}, z^*)\big) \\
&\quad = \mathcal{O}\left(\frac{1}{c_1}\exp\left(-t\left(\min_{z\neq z^*}\mathtt{KL}_{\mathrm{pair}}\big(\mathbb{P}(\cdot \mid z^*) \,\|\, \mathbb{P}(\cdot \mid z)\big) + 2\log c_0 - \frac{(1+l)}{\sqrt{t}}\log\frac{1}{c_0}\cdot\log\frac{|\mathfrak{Z}|}{\delta}\right)\right)\right).
\end{aligned} \tag{H.7}$$

Taking expectations with respect to the distribution of $S'_t, \widetilde{c}_{t+1}$ on the both sides in (H.7), we have that

$$\begin{aligned}
&\mathbb{E}_{\mathtt{prompt}'}\Big[\mathtt{TV}\big(\mathbb{P}(\cdot \mid S'_t, \widetilde{c}_{t+1}), \mathbb{P}(\cdot \mid \widetilde{c}_{t+1}, z^*)\big)\Big] \\
&\quad = \mathcal{O}\left(\frac{1}{c_1}\exp\left(-t\left(\min_{z\neq z^*}\mathtt{KL}_{\mathrm{pair}}\big(\mathbb{P}(\cdot \mid z^*) \,\|\, \mathbb{P}(\cdot \mid z)\big) + 2\log c_0 - \frac{(1+l)}{\sqrt{t}}\log\frac{1}{c_0}\cdot\log\frac{|\mathfrak{Z}|}{\delta}\right)\right)\right) + \delta.
\end{aligned} \tag{H.8}$$

We set $\delta = |\mathfrak{Z}\exp(-a\sqrt{t}/2b)|$, where $a = \min_{z\neq z^*}\mathtt{KL}_{\mathrm{pair}}\big(\mathbb{P}(\cdot \mid z^*) \,\|\, \mathbb{P}(\cdot \mid z)\big) + 2\log c_0$, $b = -(1+l)\log c_0$. Then the right-hand side of (H.8) can be upper bounded as

$$\begin{aligned}
&\mathbb{E}_{\mathtt{prompt}'}\Big[\mathtt{TV}\big(\mathbb{P}(\cdot \mid S'_t, \widetilde{c}_{t+1}), \mathbb{P}(\cdot \mid \widetilde{c}_{t+1}, z^*)\big)\Big] \\
&\quad = \mathcal{O}\left(\exp\left(-\frac{\sqrt{t}}{2(1+l)\log 1/c_0}\left(\min_{z\neq z^*}\mathtt{KL}_{\mathrm{pair}}\big(\mathbb{P}(\cdot \mid z^*) \,\|\, \mathbb{P}(\cdot \mid z)\big) + 2\log c_0\right)\right)\right).
\end{aligned}$$

Let $\mathbb{E}_{\mathtt{prompt}'}[\mathtt{TV}(\mathbb{P}(\cdot \mid S'_t, \widetilde{c}_{t+1}), \mathbb{P}_{\widehat{\theta}}(\cdot \mid S'_t, \widetilde{c}_{t+1}))] \leq \kappa\Delta_{\mathrm{pre}}(N_{\mathrm{p}}, T_{\mathrm{p}}, \delta)$, where $\Delta_{\mathrm{pre}}(N_{\mathrm{p}}, T_{\mathrm{p}}, \delta)$ is the bound in Theorem 5.3. Then we have that

$$\begin{aligned}
&\mathbb{E}_{\mathtt{prompt}'}\Big[\mathtt{KL}\big(\mathbb{P}(\cdot \mid \widetilde{c}_{t+1}, z^*) \| \mathbb{P}_{\widehat{\theta}}(\cdot \mid S'_t, \widetilde{c}_{t+1})\big)\Big] \\
&\quad \leq \mathcal{O}\Big(\mathbb{E}_{\mathtt{prompt}'}\Big[\mathtt{TV}\big(\mathbb{P}_{\widehat{\theta}}(\cdot \mid S'_t, \widetilde{c}_{t+1}), \mathbb{P}(\cdot \mid \widetilde{c}_{t+1}, z^*)\big)\Big]\Big) \\
&\quad = \mathcal{O}\left(\kappa\Delta_{\mathrm{pre}}(N_{\mathrm{p}}, T_{\mathrm{p}}, \delta) + \exp\left(-\frac{\sqrt{t}}{2(1+l)\log 1/c_0}\left(\min_{z\neq z^*}\mathtt{KL}_{\mathrm{pair}}\big(\mathbb{P}(\cdot \mid z^*) \,\|\, \mathbb{P}(\cdot \mid z)\big) + 2\log c_0\right)\right)\right),
\end{aligned}$$

where the first equality results from Assumption 5.2. Thus, we conclude the proof of Proposition H.4. □

I Proof of Supporting Propositions

I.1 Proof of Proposition F.1

Proof. Let a, b be two vectors in the $(d-1)$-dimensional unit sphere $\mathbb{S}^{d-1}$. We first define the following vector,

$$c = (a^\top b)\cdot b - \big(a - (a^\top b)\cdot b\big) \in \mathbb{S}^{d-1}. \tag{I.1}$$

By direct calculation, we have the following property of c defined in (I.1),

$$c^\top b = (a^\top b) \cdot \|b\|_2^2 - a^\top b + (a^\top b) \cdot \|b\|_2^2 = a^\top b. \tag{I.2}$$

By (I.1) and (I.2), we have that

$$a + c = 2(a^\top b) \cdot b = 2(c^\top b) \cdot b = (a^\top b) \cdot b + (c^\top b) \cdot b. \tag{I.3}$$

We now calculate the desired integration. Note that

$$\int_{\mathbb{S}^{d-1}} a \cdot \exp(a^\top b) \mathrm{d}a = b \cdot \int_{\mathbb{S}^{d-1}} (a^\top b) \exp(a^\top b) \mathrm{d}a + \int_{\mathbb{S}^{d-1}} \big(a - (a^\top b) \cdot b\big) \cdot \exp(a^\top b) \mathrm{d}a. \tag{I.4}$$

For the second term on the right-hand side of (I.4), it follows from (I.1) and (I.2) and (I.3) that

$$\int_{\mathbb{S}^{d-1}} \big(a - (a^\top b) \cdot b\big) \cdot \exp(a^\top b) \mathrm{d}a = -\int_{\mathbb{S}^{d-1}} \big(c - (c^\top b) \cdot b\big) \cdot \exp(c^\top b) \mathrm{d}c, \tag{I.5}$$

where the equality follows from the fact that $\mathrm{d}c = 2\|b\|_2^2 \mathrm{d}a - \mathrm{d}a = \mathrm{d}a$. By replacing c by a on the right-hand side of (I.5), we have

$$\int_{\mathbb{S}^{d-1}} \big(a - (a^\top b) \cdot b\big) \cdot \exp(a^\top b) \mathrm{d}a = -\int_{\mathbb{S}^{d-1}} \big(a - (a^\top b) \cdot b\big) \cdot \exp(a^\top b) \mathrm{d}a = 0 \tag{I.6}$$

Finally, by plugging (I.6) into (I.4), we obtain that

$$\int_{\mathbb{S}^{d-1}} a \cdot \exp(a^\top b) \mathrm{d}a = b \cdot \int_{\mathbb{S}^{d-1}} (a^\top b) \exp(a^\top b) \mathrm{d}a.$$

Thus, by setting

$$C_1 = \int_{\mathbb{S}^{d-1}} (a^\top b) \exp(a^\top b) \mathrm{d}a, \quad \forall b \in \mathbb{S}^{d-1},$$

we complete the proof of Proposition F.1. Note that here C_1 is an absolute constant that does not depend on b due to the symmetry on the unit sphere. □

I.2 Proof of Proposition G.2

Proof of Proposition G.2. We note that $f(X)$ satisfies the condition in Lemma J.4 with $c_i = 2b/N$ for $i \in [N]$. Then Lemma J.4 shows that

$$\mathbb{E}_{f \sim P_0}\Big[\mathbb{E}_X\Big(\exp\big[\lambda(f(X) - \mathbb{E}f(X))\big]\Big)\Big] \le \exp\left(\frac{\lambda^2 \cdot b^2 \cdot t_{\min}}{2N}\right).$$

Take $\lambda = \sqrt{2N \log 2/(b^2 t_{\min})}$. The Markov inequality shows that

$$P\left(\mathbb{E}_{f \sim P_0}\Big(\exp\big[\lambda(f(X) - \mathbb{E}f(X))\big]\Big) \ge \frac{2}{\delta}\right) \le \delta$$

for any $0 < \delta < 1$. We note that this probability inequality does not involve P. Take the function g in Lemma J.3 as $g(f) = \lambda(f(X) - \mathbb{E}f(X))$, then it shows that

$$\log \mathbb{E}_{P_0}\Big[\exp\big(g(X)\big)\Big] + \mathtt{KL}(P \,\|\, P_0) \ge \mathbb{E}_P\big[g(X)\big]$$

for any P simultaneously. Combining these inequalities, we have

$$\Big|\mathbb{E}_P\Big[\mathbb{E}_X\big[f(X)\big] - f(X)\Big]\Big| \le \sqrt{\frac{b^2 \cdot t_{\min}}{2 \log 2N}}\left[\mathtt{KL}(P \,\|\, P_0) + \log\frac{4}{\delta}\right],$$

for any distribution P on $\mathcal{F}$ simultaneously with probability at least $1 - \delta$. Thus, we conclude the proof of Proposition G.2. □

I.3 Proof of Proposition G.1

Proof of Proposition G.1 . We analyze the error layer by layer in the neural network. Denote the outputs of each layer in the networks parameterized by θ and $\widetilde{\theta}$ as $X^{(t)}$ and $\widetilde{X}^{(t)}$, respectively. In the final layer, we have that

$$\begin{aligned}
&\mathtt{TV}\big(P_\theta(\cdot \,|\, X), P_{\widetilde{\theta}}(\cdot \,|\, X)\big) \\
&\quad \le 2\bigg\| \frac{1}{L\tau} \mathbb{I}_L^\top X^{(D)} A^{(D+1)} - \frac{1}{L\tau} \mathbb{I}_L^\top \widetilde{X}^{(D)} \widetilde{A}^{(D+1)} \bigg\|_\infty \\
&\quad \le \frac{2}{\tau} \Big[\big\| A^{(D+1),\top} \big\|_{1,2} \cdot \big\| X^{(D),\top} - \widetilde{X}^{(D),\top} \big\|_{2,\infty} + \big\| A^{(D+1),\top} - \widetilde{A}^{(D+1),\top} \big\|_{1,2} \Big],
\end{aligned}$$

where the first inequality results from Lemma J.6, and the second inequality results from Lemma J.7 and that $\|X^{(D),\top}\|_{2,\infty} \le 1$ due to the layer normalization. In the following, we build the recursion relationship between $\|X^{(t),\top} - \widetilde{X}^{(t),\top}\|_{2,\infty}$ for $t \in [D]$.

$$\begin{aligned}
&\|X^{(t+1),\top} - \widetilde{X}^{(t+1),\top}\|_{2,\infty} \\
&\quad \le \big\| \mathtt{ffn}(Y^{(t+1)}, A^{(t+1)})^\top - \mathtt{ffn}(\widetilde{Y}^{(t+1)}, \widetilde{A}^{(t+1)})^\top \big\|_{2,\infty} + |\gamma_2^{(t+1)} - \widetilde{\gamma}_2^{(t+1)}| + \big\| Y^{(t+1),\top} - \widetilde{Y}^{(t+1),\top} \big\|_{2,\infty} \\
&\quad \le |\gamma_2^{(t+1)} - \widetilde{\gamma}_2^{(t+1)}| + \big\| Y^{(t+1),\top} - \widetilde{Y}^{(t+1),\top} \big\|_{2,\infty} + B_{A,1} \cdot B_{A,2} \cdot \|Y^{(t+1),\top} - \widetilde{Y}^{(t+1),\top}\|_{2,\infty} \\
&\qquad + B_{A,2} \cdot \|A_1^{(t+1)} - \widetilde{A}_1^{(t+1)}\|_{\mathrm{F}} + B_{A,1} \cdot \|A_2^{(t+1)} - \widetilde{A}_2^{(t+1)}\|_{\mathrm{F}}, \qquad \text{(I.7)}
\end{aligned}$$

where the first inequality results from the triangle inequality and that Π_{norm} is not expansive, the second inequality results from the following proposition

Proposition I.1. For any $X, \widetilde{X} \in \mathbb{R}^{L\times d}$, $A_1, \widetilde{A}_1 \in \mathbb{R}^{d\times d_F}$, and $A_2, \widetilde{A}_2 \in \mathbb{R}^{d_F \times d}$, we have that

$$\begin{aligned}
&\big\| \mathtt{ffn}(X, A)^\top - \mathtt{ffn}(\widetilde{X}, \widetilde{A})^\top \big\|_{2,\infty} \\
&\quad \le \|A_1\|_{\mathrm{F}} \cdot \|A_2\|_{\mathrm{F}} \cdot \|X^\top - \widetilde{X}^\top\|_{2,\infty} + \|A_1 - \widetilde{A}_1\|_{\mathrm{F}} \cdot \|A_2\|_{\mathrm{F}} \cdot \|\widetilde{X}^\top\|_{2,\infty} \\
&\qquad + \|\widetilde{A}_1\|_{\mathrm{F}} \cdot \|A_2 - \widetilde{A}_2\|_{\mathrm{F}} \cdot \|\widetilde{X}^\top\|_{2,\infty}.
\end{aligned}$$

Proof of Proposition I.1. See Appendix I.5. □

Next, we build the relationship between $\|Y^{(t+1),\top} - \widetilde{Y}^{(t+1),\top}\|_{2,\infty}$ in the right-hand side of inequality (I.7) and $\|X^{(t),\top} - \widetilde{X}^{(t),\top}\|_{2,\infty}$.

$$\begin{aligned}
&\|Y^{(t+1),\top} - \widetilde{Y}^{(t+1),\top}\|_{2,\infty} \\
&\quad \le \big\| \mathtt{mha}(X^{(t)}, W^{(t+1)})^\top - \mathtt{mha}(\widetilde{X}^{(t)}, \widetilde{W}^{(t+1)})^\top \big\|_{2,\infty} + |\gamma_1^{(t+1)} - \widetilde{\gamma}_1^{(t+1)}| + \big\| X^{(t),\top} - \widetilde{X}^{(t),\top} \big\|_{2,\infty} \\
&\quad \le |\gamma_1^{(t+1)} - \widetilde{\gamma}_1^{(t+1)}| + \big\| X^{(t),\top} - \widetilde{X}^{(t),\top} \big\|_{2,\infty} \\
&\qquad + h \cdot B_V \big(1 + 4B_Q B_K\big) \|X^{(t),\top} - \widetilde{X}^{(t),\top}\|_{2,\infty} + \sum_{i=1}^h \big\| W_i^{V,(t+1)} - \widetilde{W}_i^{V,(t+1)} \|_{\mathrm{F}} \\
&\qquad + 2B_V \cdot B_K \sum_{i=1}^h \|W_i^{Q,(t+1)} - \widetilde{W}_i^{Q,(t+1)}\|_{\mathrm{F}} + 2B_V \cdot B_Q \sum_{i=1}^h \|W_i^{K,(t+1)} - \widetilde{W}_i^{K,(t+1)}\|_{\mathrm{F}}, \qquad \text{(I.8)}
\end{aligned}$$

where the first inequality results from the triangle inequality, and the second inequality results from Lemma J.8. Combining inequalities (I.7) and (I.8), we derive that

$$\begin{aligned}
&\|X^{(t+1),\top} - \widetilde{X}^{(t+1),\top}\|_{2,\infty} \\
&\quad \le \big(1 + B_{A,1} \cdot B_{A,2}\big)\big(1 + hB_V(1 + 4B_Q B_K)\big) \|X^{(t),\top} - \widetilde{X}^{(t),\top}\|_{2,\infty} + \beta_{t+1} + \iota_{t+1} + \kappa_{t+1} + \rho_{t+1}.
\end{aligned}$$

This concludes the proof of Proposition G.1. □

I.4 Proof of Proposition G.9

Proof of Proposition G.9. We analyze the error layer by layer in the neural network. Denote the outputs of each layer in the networks parameterized by θ and $\widetilde{\theta}$ as $X^{(t)}$ and $\widetilde{X}^{(t)}$, respectively. In the final layer, we have that

$$\begin{aligned}
&\|f_\theta(X) - f_{\widetilde{\theta}}(X)\|_2 \\
&\quad \le \big\|\widetilde{A}^{(D+1)}\big\|_{\mathrm{F}} \cdot \big\|X^{(D),\top} - \widetilde{X}^{(D),\top}\big\|_{2,\infty} + \big\|A^{(D+1)} - \widetilde{A}^{(D+1)}\big\|_{\mathrm{F}},
\end{aligned}$$

where the inequality results from Lemma J.7 and that $\|X^{(D),\top}\|_{2,\infty} \le 1$ due to the layer normalization. The remaining proof just follows the procedures in the proof of Proposition G.1, and we have that

$$\begin{aligned}
&\|f_\theta(X) - f_{\widetilde{\theta}}(X)\|_2 \\
&\quad \le \big\|A^{(D+1)} - \widetilde{A}^{(D+1)}\big\|_{\mathrm{F}} + \sum_{t=1}^{D} \alpha_t(\beta_t + \iota_t + \kappa_t + \rho_t).
\end{aligned}$$

Thus, we conclude the proof of Proposition G.9. □

I.5 Proof of Proposition I.1

Proof of Proposition I.1. We have that

$$\begin{aligned}
&\big\|\texttt{ffn}(X,A)^\top - \texttt{ffn}(\widetilde{X},\widetilde{A})^\top\big\|_{2,\infty} \\
&\le \max_{i\in[L]} \Big[\big\|\texttt{ReLU}(X_{i,:}A_1)A_2 - \texttt{ReLU}(\widetilde{X}_{i,:}A_1)A_2\big\|_2 + \big\|\texttt{ReLU}(\widetilde{X}_{i,:}A_1)A_2 - \texttt{ReLU}(\widetilde{X}_{i,:}\widetilde{A}_1)\widetilde{A}_2\big\|_2\Big] \\
&\le \max_{i\in[L]} \Big[\|A_1\|_{\mathrm{F}} \cdot \|A_2\|_{\mathrm{F}} \cdot \|X_{i,:} - \widetilde{X}_{i,:}\|_2 + \big\|\texttt{ReLU}(\widetilde{X}_{i,:}A_1)A_2 - \texttt{ReLU}(\widetilde{X}_{i,:}\widetilde{A}_1)A_2\big\|_2 \\
&\qquad + \big\|\texttt{ReLU}(\widetilde{X}_{i,:}\widetilde{A}_1)A_2 - \texttt{ReLU}(\widetilde{X}_{i,:}\widetilde{A}_1)\widetilde{A}_2\big\|_2\Big] \\
&\le \max_{i\in[L]} \Big[\|A_1\|_{\mathrm{F}} \cdot \|A_2\|_{\mathrm{F}} \cdot \|X_{i,:} - \widetilde{X}_{i,:}\|_2 + \|A_1 - \widetilde{A}_1\|_{\mathrm{F}} \cdot \|A_2\|_{\mathrm{F}} \cdot \|\widetilde{X}_{i,:}\|_2 \\
&\qquad + \|\widetilde{A}_1\|_{\mathrm{F}} \cdot \|A_2 - \widetilde{A}_2\|_{\mathrm{F}} \cdot \|\widetilde{X}_{i,:}\|_2\Big],
\end{aligned}$$

where the first inequality results from the triangle inequality, the second and the last inequalities result from Lemma J.7 and that ReLU is not expansive. Thus, we conclude the proof of Proposition I.1. □

J Technical Lemmas

Lemma J.1 (Caponnetto and De Vito (2007))**.** Let (Ω, ν) be a probability space and ξ be a random variable on Ω taking value in a real separable Hilbert space $\mathcal{H}$. We assume that there exists constants $B, \sigma > 0$ such that

$$\big\|\xi(w)\big\|_{\mathcal{H}} \le B/2, \text{ a.s.}, \quad \mathbb{E}\big[\|\xi\|_{\mathcal{H}}^2\big] \le \sigma^2.$$

Then, it holds with probability at least $1 - \delta$ that

$$\left\|L^{-1}\sum_{i=1}^{L}\xi(\omega_i) - \mathbb{E}[\xi]\right\| \le 2\left(\frac{B}{L} + \frac{\sigma}{\sqrt{L}}\right)\log\frac{2}{\delta}.$$

Lemma J.2 (Proposition 4.5 in Duchi (2019))**.** Let $\mathcal{F}$ be the collection of functions of $f : \mathbb{R}^n \to \mathbb{R}$. For any $f \in \mathcal{F}$, we define

$$\mu(f) = \mathbb{E}_X\big[f(X)\big], \quad \sigma^2(f) = \mathbb{E}_X\big[(f(X) - \mathbb{E}_X[f(X)])^2\big],$$

where the expectation is taken with respect to a random variable $X \sim \nu$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Assume that $|f(X) - \mu(f)| \leq b$ a.s. for some constant $b \in \mathbb{R}$ for all $f \in \mathcal{F}$. Then for any $0 < \lambda \leq 1/(2b)$, given a distribution P_0 on $\mathcal{F}$, with probability at least $1 - \delta$, we have

$$\left| \mathbb{E}_Q \left[\mathbb{E}_X[f(X)] - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right] \right| \leq \lambda \mathbb{E}_Q\big[\sigma^2(f)\big] + \frac{1}{n\lambda} \left[\mathtt{KL}(Q \,\|\, P_0) + \log \frac{2}{\delta} \right],$$

for any distribution Q on $\mathcal{F}$, where X_i are i.i.d. samples of ν. If the function class $\mathcal{F}$ further satisfies $\sigma^2(f) \leq c\mu(f)$ for some constant $c \in \mathbb{R}$ for all $f \in \mathcal{F}$, we have

$$\left| \mathbb{E}_Q \left[\mathbb{E}_X\big[f(X)\big] - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right] \right| \leq \lambda c \mathbb{E}_Q\big[\mu(f)\big] + \frac{1}{n\lambda} \left[\mathtt{KL}(Q \,\|\, P_0) + \log \frac{2}{\delta} \right],$$

with probability at least $1 - \delta$.

Lemma J.3 (Donsker–Varadhan representation in Belghazi et al. (2018))**.** Let P and Q be distributions on a common space $\mathcal{X}$. Then

$$\mathtt{KL}(P \,\|\, Q) = \sup_{g \in \mathcal{G}} \left\{ \mathbb{E}_P\big[g(X)\big] - \log \mathbb{E}_Q\Big[\exp\big(g(X)\big) \Big] \right\},$$

where $\mathcal{G} = \{g : \mathcal{X} \to \mathbb{R} \mid \mathbb{E}_Q[\exp(g(X))] < \infty\}$.

Lemma J.4 (Corollary 2.11 in Paulin (2015))**.** Let $X = (X_1, \cdots, X_N)$ be a Markov chain, taking values in $\Lambda = \prod_{i=1}^{N} \Lambda_i$ with mixing time $t_{\mathrm{mix}}(\varepsilon)$ for $\varepsilon \in [0, 1]$. Let

$$t_{\min} = \inf_{0 \leq \varepsilon < 1} t_{\mathrm{mix}}(\varepsilon) \cdot \left(\frac{2 - \varepsilon}{1 - \varepsilon} \right)^2.$$

If function $f : \Lambda \to \mathbb{R}$ is such that $f(x) - f(y) \leq \sum_{i=1}^{N} c_i \mathbb{I}_{x_i \neq y_i}$ for every $x, y \in \Lambda$, then for any $\lambda \in \mathbb{R}$,

$$\log \mathbb{E}\Big(\exp\big[\lambda(f(X) - \mathbb{E}f(X))\big] \Big) \leq \frac{\lambda^2 \cdot \|c\|_2^2 \cdot t_{\min}}{8}.$$

For any $t \geq 0$, we have

$$P\Big(\big| f(X) - \mathbb{E}f(X) \big| \geq t \Big) \leq 2 \exp\left(\frac{-2t^2}{\|c\|_2^2 \cdot t_{\min}} \right).$$

Lemma J.5 (Lemma 25 in Agarwal et al. (2020))**.** For any two conditional probability densities $P(\cdot \,|\, X), P'(\cdot \,|\, X)$ and any distribution $\nu \in \Delta(\mathcal{X})$,we have

$$\mathbb{E}_\nu \Big[\mathtt{TV}\big(P(\cdot \,|\, X), P'(\cdot \,|\, X)\big)^2 \Big] \leq -2 \log \left(\mathbb{E}_{X \sim \nu, Y \sim P(\cdot \,|\, X)} \left[\exp\left(-\frac{1}{2} \log \frac{P(Y \,|\, X)}{P'(Y \,|\, X)} \right) \right] \right).$$

Lemma J.6 (Corollary A.7 in Edelman et al. (2021))**.** For any $x, y \in \mathbb{R}^d$, we have

$$\|\mathtt{softmax}(x) - \mathtt{softmax}(y)\|_1 \leq 2\|x - y\|_\infty.$$

Lemma J.7 (Lemma 17 in Zhang et al. (2022a))**.** Given any two conjugate numbers $u, v \in [1, \infty]$, i.e., $\frac{1}{u} + \frac{1}{v} = 1$, and $1 \leq p \leq \infty$, for any $A \in \mathbb{R}^{r \times c}$ and $x \in \mathbb{R}^c$, we have

$$\|Ax\|_p \leq \|A\|_{p,u} \|x\|_v \quad \text{and} \quad \|Ax\|_p \leq \|A^\top\|_{u,p} \|x\|_v.$$

Lemma J.8 (Propositions 20 and 21 in Zhang et al. (2022a))**.** For any $X, \widetilde{X} \in \mathbb{R}^{L \times d}$, and any $W_i^Q, \widetilde{W}_i^Q, W_i^K, \widetilde{W}_i^K \in \mathbb{R}^{d \times d_h}, W_i^V, \widetilde{W}_i^V \in \mathbb{R}^{d \times d}$ for $i \in [h]$, if $\|X^\top\|_{p,\infty}, \|\widetilde{X}^\top\|_{2,\infty} \leq B_X$, $\|W_i^Q\|_{\mathrm{F}}, \|\widetilde{W}_i^Q\|_{\mathrm{F}} \leq B_Q$, $\|W_i^K\|_{\mathrm{F}}, \|\widetilde{W}_i^K\|_{\mathrm{F}} \leq B_K$, $\|W_i^V\|_{\mathrm{F}}, \|\widetilde{W}_i^V\|_{\mathrm{F}} \leq B_V$ for $i \in [h]$, then we have

$$\begin{aligned}
&\Big\| \big(\mathtt{mha}(X, W) - \mathtt{mha}(\widetilde{X}, \widetilde{W}) \big)^\top \Big\|_{2,\infty} \\
&\leq h \cdot B_V \big(1 + 4B_X^2 \cdot B_Q B_K\big) \|X^\top - \widetilde{X}^\top\|_{2,\infty} + B_X \sum_{i=1}^{h} \big\| W_i^V - \widetilde{W}_i^V \big\|_{\mathrm{F}} \\
&\quad + 2B_X^3 \cdot B_V \cdot B_K \sum_{i=1}^{h} \|W_i^Q - \widetilde{W}_i^Q\|_{\mathrm{F}} + 2B_X^3 \cdot B_V \cdot B_Q \sum_{i=1}^{h} \|W_i^K - \widetilde{W}_i^K\|_{\mathrm{F}}.
\end{aligned}$$

Lemma J.9 (Lemma A.6 in Elbrächter et al. (2021))**.** For $a, b \in \mathbb{R}$ with $a < b$, let

$$\mathcal{S}_{[a,b]} = \left\{ f \in \mathcal{S}^{\infty}([a,b], \mathbb{R}) \,|\, \left\| f^{(n)}(x) \right\| \leq n! \text{ for all } n \in \mathbb{N} \right\}.$$

There exists a constant $C > 0$ such that for all $a, b \in \mathbb{R}$ with $a < b$, $f \in \mathcal{S}_{[a,b]}$, and $\varepsilon \in (0, 1/2)$, there is a fully connect network Ψ_f such that

$$\|f - \Psi_f\|_{\infty} \leq \varepsilon,$$

with the depth of the network as $D(\Psi_f) \leq C \max\{2, b-a\}(\log \varepsilon^{-1})^2 + \log(\lceil \max\{|a|, |b|\} \rceil) + \log(\lceil 1/(b-a) \rceil)$, the width of the network as $W(\Psi_f) \leq 16$, and the maximal weight in the network as $B(\Psi_f) \leq 1$.

Lemma J.10. Let $b = \sup_x \log(p(x)/q(x))$. We have that

$$\mathtt{KL}(p \,\|\, q) \leq 2(3 + b) \cdot \mathtt{TV}(p, q). \tag{J.1}$$

Proof. We let $f(t) = \log t$ and $g(t) = |1/t - 1|$. Then, for $0 \leq t \leq \exp(b)$, we have that

$$\sup_{0 \leq t \leq \exp(b)} \frac{f(t)}{g(t)} = \sup_{0 \leq t \leq \exp(b)} \frac{\log t}{|1/t - 1|} = \sup_{1 \leq t \leq \exp(b)} \frac{t \log t}{t - 1} \leq 2(b + 3).$$

Note that $\mathtt{KL}(p \,\|\, q) = \mathbb{E}_p[f(p(x)/q(x))]$ and $\mathtt{TV}(p, q) = \mathbb{E}_p[g(p(x)/q(x))]$, which concludes the proof. □