NORTHWESTERN UNIVERSITY

Essay on Foundation Models and Reinforcement Learning

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Industrial Engineering and Management Sciences

By

Yufeng Zhang

EVANSTON, ILLINOIS

September 2023

ABSTRACT

Essay on Foundation Models and Reinforcement Learning

Yufeng Zhang

In this dissertation, we aim to develop a theoretical understanding of foundation models and reinforcement learning. We delve into a comprehensive analysis of specific aspects within these domains. The focal points of our study are as follows:

- **Generative Adversarial Imitation Learning (GAIL) with Neural Networks**: GAIL is poised to execute tasks based on expert demonstrations. By parameterizing both the reward function and policy using neural networks, we develop a gradient-based algorithm with alternating updates for GAIL. Through rigorous analysis, we establish that this algorithm converges to the global optimum at a sublinear rate.
- **Temporal-Difference (TD) Learning and Q-learning with Neural Networks**: We dissect the fundamental reason behind the empirical success of deep TD learning and deep Q-learning: the learned feature representation. Utilizing mean-field analysis, we scrutinize the evolution of this representation. We demonstrate that, when implemented through an overparameterized two-layer neural

network, both TD learning and Q-learning algorithms are capable of globally minimizing the mean-squared projected Bellman error at a sublinear rate.

- **Attention Mechanisms and Transformers**: Analyzing attention mechanisms and transformers through the lens of exchangeability, we first establish the existence of a representation for input tokens that is sufficient and minimal. We then ascertain that the attention mechanism with the appropriate parameters is able to infer the latent posterior within a margin of approximation error that diminishes as input sizes increase. Additionally, we prove that employing either supervised or self-supervised objectives enables empirical risk minimization to learn the optimal parameters within a generalization error that remains independent of input sizes.
- **In-Context Learning (ICL)**: We execute an exhaustive investigation into ICL by addressing several pertinent questions. Firstly, from a Bayesian view, we show that the language models learn an ICL estimator by implementing Bayesian model averaging. Subsequently, we evaluate the performance of the ICL algorithm from an online learning standpoint and establish a regret bound decreasing with the length of the ICL input sequence. Then, we demonstrate that during pretraining, the total variation distance between the learned model and the underlying true model is constrained by a generalization error decreasing with the number of token sequences and the length of each sequence during pretraining, respectively. Finally, by combining this two results, we show that the learned model is capable in ICL.

This dissertation aspires to enrich the academic discourse on foundation models and reinforcement learning by offering novel insights and rigorous proofs that may serve as building blocks for future research in these rapidly evolving fields.

Acknowledgements

I would like to express my deepest appreciation to my advisor, Zhaoran Wang, whose expertise, understanding, and patience, added considerably to my graduate experience. I appreciate his vast knowledge and skills in reinforcement learning, language models, and related areas, as well as his continuous encouragement during this challenging journey.

I would also like to express my gratitude to my committee members, Ethan Xingyuan Fang and Chang-Han Rhee, for their insightful comments and suggestions. Their expertise and contributions in a variety of perspectives have led me to develop a much more comprehensive understanding of my research area.

I extend my heartfelt thanks to my colleagues and friends in the Department of Industrial Engineering and Management Sciences for their friendship, intellectual discussions and the vibrant environment they created.

I am also immensely grateful to Northwestern University for providing me with the necessary resources and environment conducive for research. Special mention goes to the staff and administration who facilitated all necessary processes and made sure everything needed was taken care of.

To my mother, Junping Zhao, thank you for supporting me emotionally and financially. Your belief in me and my abilities was a source of strength that I drew from time and again. To my partner, Wei Wang, my deepest thanks for being an integral part of my life. The moments shared with you have been the most enriching aspect of my academic journey at Northwestern University.

Finally, I would like to express my gratitude to my collaborators, Zhuoran Yang, Qi Cai, Yongxin Chen, Zuyue Fu, Lingxiao Wang, Boyi Liu, and Junwei Lu, who made this research possible.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Generative Adversarial Imitation Learning with Neural Networks: Global Optimality and Convergence Rate

Generative adversarial imitation learning (GAIL) demonstrates tremendous success in practice, especially when combined with neural networks. Different from reinforcement learning, GAIL learns both policy and reward function from expert (human) demonstration. Despite its empirical success, it remains unclear whether GAIL with neural networks converges to the globally optimal solution. The major difficulty comes from the nonconvex-nonconcave minimax optimization structure. To bridge the gap between practice and theory, we analyze a gradient-based algorithm with alternating updates and establish its sublinear convergence to the globally optimal solution. To the best of our knowledge, our analysis establishes the global optimality and convergence rate of GAIL with neural networks for the first time.

1.1. Introduction

The goal of imitation learning (IL) is to learn to perform a task based on expert demonstration (Ho and Ermon, 2016). In contrast to reinforcement learning (RL), the agent only has access to the expert trajectories but not the rewards. The most straightforward approach of IL is behavioral cloning (BC) (Pomerleau, 1991). BC treats IL as the supervised learning problem of predicting the actions based on the states. Despite its simplicity, BC suffers from the compounding errors caused by covariate shift (Ross et al., 2011; Ross and Bagnell, 2010). Another approach of IL is inverse reinforcement learning (IRL) (Russell, 1998; Ng and Russell, 2000; Levine and Koltun, 2012; Finn et al., 2016), which jointly learns the reward function and the corresponding optimal policy. IRL formulates IL as a bilevel optimization problem. Specifically, IRL solves an RL subproblem given a reward function at the inner level and searches for the reward function which makes the expert policy optimal at the outer level. However, IRL is computationally inefficient as it requires fully solving an RL subproblem at each iteration of the outer level. Moreover, the desired reward function may be nonunique. To address such issues of IRL, Ho and Ermon (2016) propose generative adversarial imitation learning (GAIL), which searches for the optimal policy without fully solving an RL subproblem given a reward function at each iteration. GAIL solves IL via minimax optimization with alternating updates. In particular, GAIL alternates between (i) minimizing the discrepancy in expected cumulative reward between the expert policy and the learned policy and (ii) maximizing such a discrepancy over the reward function class. Such an alternating update scheme mirrors the training of generative adversarial networks (GANs) (Goodfellow et al., 2014; Arjovsky et al., 2017), where the learned policy acts as the generator while the reward function acts as the discriminator.

Incorporated with neural networks, which parameterize the learned policy and the reward function, GAIL achieves significant empirical success in challenging applications, such as natural language processing (Yu et al., 2016), autonomous driving (Kuefler et al., 2017), human behavior modeling (Merel et al., 2017), and robotics (Tai et al., 2018). Despite its empirical success, GAIL with neural networks remains less understood in theory. The major difficulty arises from the following aspects: (i) GAIL involves minimax optimization, while the existing analysis of policy optimization with neural networks (Anthony and Bartlett, 2009; Liu et al., 2019; Bhandari and Russo, 2019; Wang et al., 2019) only focuses on a minimization or maximization problem. (ii) GAIL with neural networks is nonconvex-nonconcave, and therefore, the existing analysis of convex-concave optimization with alternating updates is not applicable (Nesterov, 2013). There is an emerging body of literature (Rafique et al., 2018; Zhang et al., 2019b) that casts nonconvex-nonconcave optimization as bilevel optimization, where the inner level is solved to a high precision as in IRL. However, such analysis is not applicable to GAIL as it involves alternating updates.

In this paper, we bridge the gap between practice and theory by establishing the global optimality and convergence of GAIL with neural networks. Specifically, we parameterize the learned policy and the reward function with two-layer neural networks and consider solving GAIL by alternatively updating the learned policy via a step of natural policy gradient (Kakade, 2002; Peters and Schaal, 2008) and the reward function via a step of gradient ascent. In particular, we parameterize the state-action value function (also known as the Q-function) with a two-layer neural network and apply a variant of the temporal difference algorithm (Sutton and Barto, 2018) to solve the policy evaluation subproblem in natural policy gradient. We prove that the learned policy $\bar{\pi}$ converges to the expert policy π_{E} at a $1/\sqrt{T}$ rate in the $\mathcal{R}$-distance (Chen et al., 2020a), which is defined as $\mathbb{D}_{\mathcal{R}}(\pi_{\mathrm{E}}, \bar{\pi}) = \max_{r \in \mathcal{R}} J(\pi_{\mathrm{E}}; r) - J(\bar{\pi}; r)$. Here $J(\pi; r)$ is the expected cumulative reward of a policy π given a reward function $r(s, a)$ and $\mathcal{R}$ is the reward function class. The core of our analysis is constructing a potential function that tracks the $\mathcal{R}$-distance. Such a rate of convergence implies that the learned policy $\bar{\pi}$ (approximately) outperforms the expert policy π_{E} given any reward function $r \in \mathcal{R}$ within a finite number of iterations T. In other words, the learned policy $\bar{\pi}$ is globally optimal. To the best of our knowledge, our analysis establishes the global optimality and convergence of GAIL with neural networks for the first time. It is worth mentioning that our analysis is straightforwardly applicable to linear and tabular settings, which, however, are not our focus.

Related works. Our work extends an emerging body of literature on RL with neural networks (Xu et al., 2019a; Zhang et al., 2019a; Bhandari and Russo, 2019; Liu et al., 2019; Wang et al., 2019; Agarwal et al., 2019) to IL. This line of research analyzes the global optimality and convergence of policy gradient for solving RL, which is a minimization or maximization problem. In contrast, we analyze GAIL, which is a minimax optimization problem.

Our work is also related to the analysis of apprenticeship learning (Syed et al., 2008) and GAIL (Cai et al., 2019a; Chen et al., 2020a). Syed et al. (2008) analyze the convergence and generalization of apprenticeship learning. They assume the state space to be finite, and thus, do not require function approximation for the policy and the reward function. In contrast, we assume the state space to be infinite and employ function approximation based on neural networks. Cai et al. (2019a) study the global optimality and convergence of GAIL in the setting of linear-quadratic regulators. In contrast, our analysis handles general MDPs without restrictive assumptions on the transition kernel and the reward function. Chen et al. (2020a) study the convergence and generalization of GAIL in the setting of general MDPs. However, they only establish the convergence to a stationary point. In contrast, we establish the global optimality of GAIL.

Notations. Let $[n] = \{1, \ldots, n\}$ for $n \in \mathbb{N}_+$ and $[m : n] = \{m, m+1, \ldots, n\}$ for $m < n$. Also, let $N(\mu, \Sigma)$ be the Gaussian distribution with mean μ and covariance Σ. We denote by $\mathscr{P}(\mathcal{X})$ the set of all probability measures over the space $\mathcal{X}$. For a function $f : \mathcal{X} \to \mathbb{R}$, a constant $p \geq 1$, and a probability measure $\mu \in \mathscr{P}(\mathcal{X})$, we denote by $\|f\|_{p,\mu} = (\int_{\mathcal{X}} |f(x)|^p \mathrm{d}\mu(x))^{1/p}$ the $L_p(\mu)$ norm of the function f. For any two functions $f, g : \mathcal{X} \to \mathbb{R}$, we denote by $\langle f, g \rangle_{\mathcal{X}} = \int_{\mathcal{X}} f(x) \cdot g(x) \mathrm{d}x$ the inner product on the space $\mathcal{X}$.

1.2. Background

1.2.1. Reinforcement Learning

We consider a Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, r, P, \rho, \gamma)$. Here $\mathcal{S} \subseteq \mathbb{R}^{d_1}$ is the state space, $\mathcal{A} \subseteq \mathbb{R}^{d_2}$ is the action space, which is assumed to be finite throughout this paper, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $P : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathcal{S})$ is the transition kernel, $\rho \in \mathscr{P}(\mathcal{S})$ is the initial state distribution, and $\gamma \in (0, 1)$ is the discount factor. Without loss of generality, we assume that $\mathcal{S} \times \mathcal{A}$ is compact and that $\|(s, a)\|_2 \leq 1$ for any $(s, a) \in \mathcal{S} \times \mathcal{A} \subseteq \mathbb{R}^d$, where $d = d_1 + d_2$. An agent following a policy $\pi : \mathcal{S} \to \mathscr{P}(\mathcal{A})$ interacts with the environment in the following manner. At the state $s_t \in \mathcal{S}$, the agent takes the action $a_t \in \mathcal{A}$ with probability $\pi(a_t \,|\, s_t)$ and receives the reward $r_t = r(s_t, a_t)$. The environment then transits into the next state s_{t+1} with probability $P(s_{t+1} \,|\, s_t, a_t)$. Given a policy π and a reward function $r(s, a)$, we define the state-action value function $Q_r^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as follows,

$$Q_r^\pi(s, a) = \mathbb{E}_\pi\biggl[(1 - \gamma) \cdot \sum_{t=0}^{\infty} \gamma^t \cdot r(s_t, a_t) \,\bigg|\, s_0 = s, a_0 = a\biggr]. \tag{1.2.1}$$

Here the expectation $\mathbb{E}_\pi$ is taken with respect to $a_t \sim \pi(\cdot \,|\, s_t)$ and $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$. Correspondingly, we define the state value function $V_r^\pi : \mathcal{S} \to \mathbb{R}$ and the advantage function $A_r^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as follows,

$$V_r^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot \,|\, s)}\bigl[Q_r^\pi(s,a)\bigr], \quad A_r^\pi(s,a) = Q_r^\pi(s,a) - V_r^\pi(s).$$

The goal of RL is to maximize the following expected cumulative reward,

$$J(\pi; r) = \mathbb{E}_{s \sim \rho}\bigl[V_r^\pi(s)\bigr]. \tag{1.2.2}$$

The policy π induces a state visitation measure $d_\pi \in \mathscr{P}(\mathcal{S})$ and a state-action visitation measure $\nu_\pi \in \mathscr{P}(\mathcal{S} \times \mathcal{A})$, which take the forms of

$$d_\pi(s) = (1-\gamma) \cdot \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{P}\bigl(s_t = s \,\big|\, s_0 \sim \rho, a_t \sim \pi(\cdot \,|\, s_t)\bigr), \quad \nu_\pi(s,a) = d_\pi(s) \cdot \pi(a \,|\, s). \tag{1.2.3}$$

It then holds that $J(\pi; r) = \mathbb{E}_{(s,a) \sim \nu_\pi}[r(s,a)]$. Meanwhile, we assume that the policy π induces a state stationary distribution ϱ_π over $\mathcal{S}$, which satisfies that

$$\varrho_\pi(s) = \mathbb{P}\bigl(s_{t+1} = s \,\big|\, s_t \sim \rho_\pi, a_t \sim \pi(\cdot \,|\, s_t)\bigr).$$

We denote by $\rho_\pi(s,a) = \varrho(s) \cdot \pi(a \,|\, s)$ the state-action stationary distribution over $\mathcal{S} \times \mathcal{A}$.

1.2.2. Generative Adversarial Imitation Learning

The goal of imitation learning (IL) is to learn a policy that outperforms the expert policy π_{E} based on the trajectory $\{(s_i^{\mathrm{E}}, a_i^{\mathrm{E}})\}_{i \in [T_{\mathrm{E}}]}$ of π_{E}. We denote by $\nu_{\mathrm{E}} = \nu_{\pi_{\mathrm{E}}}$ and $d_{\mathrm{E}} = d_{\pi_{\mathrm{E}}}$

the state-action and state visitation measures induced by the expert policy, respectively, and assume that the expert trajectory $\{(s_i, a_i)\}_{i\in[T_{\mathrm{E}}]}$ is drawn from ν_{E}. To this end, we parameterize the policy and the reward function by π_θ for $\theta \in \mathcal{X}_\Pi$ and $r_\beta(s,a)$ for $\beta \in \mathcal{X}_R$, respectively, and solve the following minimax optimization problem known as GAIL (Ho and Ermon, 2016),

$$\min_{\theta\in\mathcal{X}_\Pi}\max_{\beta\in\mathcal{X}_R} L(\theta,\beta), \quad \text{where } L(\theta,\beta) = J(\pi_{\mathrm{E}}; r_\beta) - J(\pi_\theta; r_\beta) - \lambda\cdot\psi(\beta). \tag{1.2.4}$$

Here $J(\pi; r)$ is the expected cumulative reward defined in (1.2.2), $\psi : \mathcal{X}_R \to \mathbb{R}_+$ is the regularizer, and $\lambda \geq 0$ is the regularization parameter. Given a reward function class $\mathcal{R}$, we define the $\mathcal{R}$-distance between two policies π_1 and π_2 as follows,

$$\mathbb{D}_{\mathcal{R}}(\pi_1,\pi_2) = \max_{r\in\mathcal{R}} J(\pi_1; r) - J(\pi_2; r) = \max_{r\in\mathcal{R}} \mathbb{E}_{\nu_{\pi_1}}\big[r(s,a)\big] - \mathbb{E}_{\nu_{\pi_2}}\big[r(s,a)\big]. \tag{1.2.5}$$

When $\mathcal{R}$ is the class of 1-Lipschitz functions, $\mathbb{D}_{\mathcal{R}}(\pi_1,\pi_2)$ is the Wasserstein-1 metric between the state-action visitation measures induced by π_1 and π_2. However, $\mathbb{D}_{\mathcal{R}}(\pi_1,\pi_2)$ is not a metric in general. When $\mathbb{D}_{\mathcal{R}}(\pi_1,\pi_2) \leq 0$, the policy π_2 outperforms the policy π_1 for any reward function $r \in \mathcal{R}$. Such a notion of $\mathcal{R}$-distance is used in Chen et al. (2020a). We denote by $\mathcal{R}_\beta = \{r_\beta(s,a)\,|\,\beta\in\mathcal{X}_R\}$ the reward function class parameterized with β. Hence, the optimization problem in (1.2.4) minimizes the $\mathcal{R}_\beta$-distance $\mathbb{D}_{\mathcal{R}_\beta}(\pi_{\mathrm{E}},\pi_\theta)$ (up to the regularizer $\lambda\cdot\psi(\beta)$), which searches for a policy $\bar{\pi}$ that (approximately) outperforms the expert policy given any reward function $r_\beta \in \mathcal{R}_\beta$.

1.3. Algorithm

In this section, we introduce an algorithm with alternating updates for GAIL with neural networks, which employs natural policy gradient (NPG) to update the policy π_θ and gradient ascent to update the reward function $r_\beta(s, a)$.

1.3.1. Parameterization with Neural Networks

We define a two-layer neural network with rectified linear units (ReLU) as follows,

$$(1.3.1)\quad u_{W,b}(s,a) = \frac{1}{\sqrt{m}} \sum_{l=1}^{m} b_l \cdot \mathbb{1}\left\{(s,a)^\top [W]_l > 0\right\} \cdot (s,a)^\top [W]_l = \sum_{l=1}^{m} \big[\phi_{W,b}(s,a)\big]_l^\top [W]_l.$$

Here $m \in \mathbb{N}_+$ is the width of the neural network, $b = (b_1, \ldots, b_m)^\top \in \mathbb{R}^m$ and $W = ([W]_1^\top, \ldots, [W]_m^\top)^\top \in \mathbb{R}^{md}$ are the parameters, and $\phi_{W,b}(s,a) = ([\phi_{W,b}(s,a)]_1^\top, \ldots, [\phi_{W,b}(s,a)]_m^\top)^\top \in \mathbb{R}^{md}$ is called the feature vector in the sequel, where

$$(1.3.2)\qquad \big[\phi_{W,b}(s,a)\big]_l = m^{-1/2} \cdot b_l \cdot \mathbb{1}\big\{(s,a)^\top [W]_l > 0\big\} \cdot (s,a).$$

It then holds that $u_{W,b}(s,a) = W^\top \phi_{W,b}(s,a)$. Note that the feature vector $\phi_{W,b}(s,a)$ depends on the parameters W and b. We consider the following random initialization,

$$(1.3.3)\qquad b_l \overset{\text{i.i.d.}}{\sim} \text{Unif}\big(\{-1,1\}\big), \quad [W_0]_l \overset{\text{i.i.d.}}{\sim} N(0, I_d/d), \quad \forall l \in [m].$$

Throughout the training process, we keep the parameter b fixed while updating W. For notational simplicity, we write $u_{W,b}(s,a)$ as $u_W(s,a)$ and $\phi_{W,b}(s,a)$ as $\phi_W(s,a)$ in the sequel. We denote by $\mathbb{E}_{\text{init}}$ the expectation taken with respect to the random initialization in (1.3.3). For an absolute constant $B > 0$, we define the parameter domain as

$$S_B = \big\{W \in \mathbb{R}^{md} \,\big|\, \|W - W_0\|_2 \leq B\big\}, \tag{1.3.4}$$

which is the ball centered at W_0 with the domain radius B.

In the sequel, we consider the following energy-based policy,

$$\pi_\theta(a \,|\, s) = \frac{\exp\big(\tau \cdot u_\theta(s, a)\big)}{\sum_{a' \in \mathcal{A}} \exp\big(\tau \cdot u_\theta(s, a')\big)}, \tag{1.3.5}$$

where $\tau \geq 0$ is the inverse temperature parameter and $u_\theta(s, a)$ is the energy function parameterized by the neural network defined in (1.3.1) with $W = \theta$. In the sequel, we call θ the policy parameter. Meanwhile, we parameterize the reward function $r_\beta(s, a)$ as follows,

$$r_\beta(s, a) = (1 - \gamma)^{-1} \cdot u_\beta(s, a), \tag{1.3.6}$$

where $u_\beta(s, a)$ is the neural network defined in (1.3.1) with $W = \beta$ and γ is the discount factor. Here we use the scaling parameter $(1 - \gamma)^{-1}$ to ensure that for any policy π the state-action value function $Q^\pi_{r_\beta}(s, a)$ defined in (1.2.1) is well approximated by the neural network defined in (1.3.1). In the sequel, we call β the reward parameter and define the reward function class as

$$\mathcal{R}_\beta = \{r_\beta(s, a) \,|\, \beta \in S_{B_\beta}\},$$

where S_{B_β} is the parameter domain of β defined in (1.3.4) with domain radius B_β. To facilitate algorithm design, we establish the following proposition, which calculates the explicit expressions of the gradients $\nabla L(\theta, \beta)$ and the Fisher information $\mathcal{I}(\theta)$. Recall that the Fisher information is defined as

$$\mathcal{I}(\theta) = \mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}\big[\nabla_\theta \log \pi_\theta(s,a) \nabla_\theta \log \pi_\theta(s,a)^\top\big]. \tag{1.3.7}$$

Proposition 1.3.1 (Gradients and Fisher Information)**.** We call $\iota_\theta(s,a) = \tau^{-1}\cdot\nabla_\theta \log \pi_\theta(a \,|\, s)$ the temperature-adjusted score function. It holds that

$$\iota_\theta(s,a) = \phi_\theta(s,a) - \mathbb{E}_{a'\sim\pi_\theta(\cdot\,|\,s)}\big[\phi_\theta(s,a')\big]. \tag{1.3.8}$$

It then holds that

$$\mathcal{I}(\theta) = \tau^2 \cdot \mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}\big[\iota_\theta(s,a)\,\iota_\theta(s,a)^\top\big], \tag{1.3.9}$$

$$\nabla_\theta L(\theta, \beta) = -\tau \cdot \mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}\big[Q^{\pi_\theta}_{r_\beta}(s,a) \cdot \iota_\theta(s,a)\big], \tag{1.3.10}$$

and

$$\nabla_\beta L(\theta,\beta) = (1-\gamma)^{-1} \cdot \mathbb{E}_{(s,a)\sim\nu_{\mathrm{E}}}\big[\phi_\beta(s,a)\big] - (1-\gamma)^{-1} \cdot \mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}\big[\phi_\beta(s,a)\big] - \lambda \cdot \nabla_\beta \psi(\beta), \tag{1.3.11}$$

where $Q^{\pi_\theta}_{r_\beta}(s,a)$ is the state-action value function defined in (1.2.1) with $\pi = \pi_\theta$ and $r = r_\beta$, ν_{π_θ} is the state-action visitation measure defined in (1.2.3) with $\pi = \pi_\theta$, and $\mathcal{I}(\theta)$ is the Fisher information defined in (1.3.7).

Proof. See §A.3.1 for a detailed proof. □

Note that the expression of the policy gradient $\nabla_\theta L(\theta, \beta)$ in (1.3.10) of Proposition 1.3.1 involves the state-action value function $Q^{\pi_\theta}_{r_\beta}(s,a)$. To this end, we estimate the state-action value function $Q_r^\pi(s,a)$ by $\widehat{Q}_\omega(s,a)$, which is parameterized as follows,

$$\widehat{Q}_\omega(s,a) = u_\omega(s,a). \tag{1.3.12}$$

Here $u_\omega(s,a)$ is the neural network defined in (1.3.1) with $W=\omega$. In the sequel, we call ω the value parameter.

1.3.2. GAIL with Alternating Updates

We employ an actor-critic scheme with alternating updates of the policy and the reward function, which is presented in Algorithm 1. Specifically, we update the policy parameter θ via natural policy gradient and update the reward parameter β via gradient ascent in the actor step, while we estimate the state-action value function $Q_r^\pi(s,a)$ via neural temporal difference (TD) (Cai et al., 2019c) in the critic step.

Actor Step. In the k-th actor step, we update the policy parameter θ and the reward parameter β as follows,

$$\theta_{k+1} = \tau_{k+1}^{-1} \cdot (\tau_k \cdot \theta_k - \eta \cdot \delta_k), \tag{1.3.13}$$

$$\beta_{k+1} = \mathrm{Proj}_{S_{B_\beta}}\bigl\{\beta_k + \eta \cdot \widehat{\nabla}_\beta L(\theta_k, \beta_k)\bigr\}, \tag{1.3.14}$$

where

$$\tau_{k+1} = \eta + \tau_k, \quad \delta_k \in \mathop{\mathrm{argmin}}_{\delta \in S_{B_\theta}} \bigl\|\widehat{\mathcal{I}}(\theta_k)\delta - \tau_k \cdot \widehat{\nabla}_\theta L(\theta_k, \beta_k)\bigr\|_2. \tag{1.3.15}$$

Here $\eta > 0$ is the stepsize, S_{B_θ} and S_{B_β} are the parameter domains of θ and β defined in (1.3.4) with domain radii B_θ and B_β, respectively, $\mathrm{Proj}_{S_{B_\beta}} : \mathbb{R}^{md} \to S_{B_\beta}$ is the projection operator, τ_k is the inverse temperature parameter of π_{θ_k}, and $\widehat{\mathcal{I}}(\theta_k), \widehat{\nabla}_\theta L(\theta_k, \beta_k), \widehat{\nabla}_\beta L(\theta_k, \beta_k)$ are the estimators of $\mathcal{I}(\theta_k), \nabla_\theta L(\theta_k, \beta_k), \nabla_\beta L(\theta_k, \beta_k)$, respectively, which are defined in the sequel. In (1.3.13), we update the policy parameter θ_k along the direction δ_k, which approximates the natural policy gradient $\mathcal{I}(\theta)^{-1} \cdot \nabla_\theta L(\theta, \beta)$, and in (1.3.15) we update the inverse temperature parameter τ_k to ensure that $\theta_{k+1} \in S_{B_\theta}$. Meanwhile, in (1.3.14), we update the reward parameter β via (projected) gradient ascent. Following from (1.3.9) and (1.3.10) of Proposition 1.3.1, we construct the estimators of $\mathcal{I}(\theta_k)$ and $\nabla_\theta L(\theta_k, \beta_k)$ as follows,

$$\widehat{\mathcal{I}}(\theta_k) = \frac{\tau_k^2}{N} \sum_{i=1}^{N} \iota_{\theta_k}(s_i, a_i)\, \iota_{\theta_k}(s_i, a_i)^\top, \tag{1.3.16}$$

$$\widehat{\nabla}_\theta L(\theta_k, \beta_k) = -\frac{\tau_k}{N} \sum_{i=1}^{N} \widehat{Q}_{\omega_k}(s_i, a_i) \cdot \iota_{\theta_k}(s_i, a_i), \tag{1.3.17}$$

where $\{(s_i, a_i)\}_{i \in [N]}$ is sampled from the state-action visitation measure $\nu_{\pi_{\theta_k}}$ given θ_k with the batch size N, and $\widehat{Q}_{\omega_k}(s, a)$ is the estimator of $Q^{\pi_{\theta_k}}_{r_{\beta_k}}(s, a)$ computed in the critic step. Meanwhile, following from (1.3.11) of Proposition 1.3.1, we construct the estimator of $\nabla_\beta L(\theta_k, \beta_k)$ as follows,

$$\widehat{\nabla}_\beta L(\theta, \beta) = \frac{1}{N \cdot (1-\gamma)} \sum_{i=1}^{N} \bigl[\phi_{\beta_k}(s_i^{\mathrm{E}}, a_i^{\mathrm{E}}) - \phi_{\beta_k}(s_i, a_i)\bigr] - \lambda \cdot \nabla_\beta \psi(\beta_k), \tag{1.3.18}$$

where $\{(s_i^{\mathrm{E}}, a_i^{\mathrm{E}})\}_{i \in [N]}$ is the expert trajectory. For notational simplicity, we write $\pi_k = \pi_{\theta_k}$, $r_k(s, a) = r_{\beta_k}(s, a)$, $d_k = d_{\pi_k}$ and $\nu_k = \nu_{\pi_k}$ for the k-th step hereafter, where π_θ is the policy, $r_\beta(s,a)$ is the reward function, and d_π, ν_π are the visitation measures defined in (1.2.3).

Critic Step. Note that the estimator $\widehat{\nabla}_\theta L(\theta,\beta)$ in (1.3.17) involves the estimator $\widehat{Q}_{\omega_k}(s,a)$ of $Q^{\pi_k}_{r_k}(s,a)$. To this end, we parameterize $\widehat{Q}_\omega(s,a)$ as in (1.3.12) and adapt neural TD (Cai et al., 2019c), which solves the following minimization problem,

$$\omega_k = \operatorname*{argmin}_{\omega \in S_{B_\omega}} \mathbb{E}_{(s,a)\sim\rho_k}\big[\widehat{Q}_\omega(s,a) - \mathcal{T}^{\pi_k}_{r_k}\widehat{Q}_\omega(s,a)\big]^2. \tag{1.3.19}$$

Here S_{B_ω} is the parameter domain with domain radius B_ω, ρ_k is the state-action stationary distribution induced by π_k, and $\mathcal{T}^{\pi_k}_{r_k}$ is the Bellman operator. Note that the Bellman operator $\mathcal{T}^\pi_r$ is defined as follows,

$$\mathcal{T}^\pi_r Q(s,a) = (1-\gamma)\cdot r(s,a) + \gamma \cdot \mathbb{E}_\pi\big[Q(s',a') \,\big|\, s,a\big],$$

where the expectation is taken with respect to $s' \sim P(\cdot \,|\, s,a)$ and $a' \sim \pi(\cdot \,|\, s')$. In neural TD, we iteratively update the value parameter ω via

$$\begin{aligned} \delta(j) &= Q_{\omega(j)}(s,a) - r(s,a) - \gamma \cdot Q_{\omega(j)}(s',a'), \\ \omega(j+1) &= \mathrm{Proj}_{S_{B_\omega}}\big\{\omega(j) - \alpha\cdot\delta(j)\cdot\nabla_\omega Q_{\omega(j)}(s,a)\big\}, \end{aligned} \tag{1.3.20}$$

where $\delta(j)$ is the Bellman residual, $\alpha > 0$ is the stepsize, (s,a) is sampled from the state-action stationary distribution ρ_k, and $s' \sim P(\cdot \,|\, s,a), a' \sim \pi_k(\cdot \,|\, s')$ are the subsequent state and action. We defer the detailed discussion of neural TD to §A.2.

To approximately obtain the compatible function approximation (Sutton et al., 2000; Wang et al., 2019), we share the random initialization among the policy π_θ, the reward function $r_\beta(s,a)$, and the state-action value function $\widehat{Q}_\omega(s,a)$. In other words, we set $\theta_0 = \beta_0 = \omega(0) = W_0$ in our algorithm, where W_0 is the random initialization in (1.3.3). The output of GAIL is the mixed policy $\bar{\pi}$ (Altman, 1999). Specifically, the mixed policy $\bar{\pi}$ of $\pi_0, \ldots, \pi_{T-1}$ is executed by randomly selecting a policy π_k for $k \in [0 : T-1]$ with equal probability before time $t = 0$ and exclusively following π_k thereafter. It then holds for any reward function $r(s,a)$ that

$$J(\bar{\pi}; r) = \frac{1}{T} \sum_{k=0}^{T-1} J(\pi_k; r). \tag{1.3.21}$$

Algorithm 1 GAIL

```
Require: Expert trajectory $\{(s_i^{\mathrm{E}}, a_i^{\mathrm{E}})\}_{i\in[T_{\mathrm{E}}]}$, number of iterations $T$, number of iterations
    $T_{\mathrm{TD}}$ of neural TD, stepsize $\eta$, stepsize $\alpha$ of neural TD, batch size $N$, and domain radii
    $B_\theta, B_\omega, B_\beta$.
Initialization. Initialize $b_l \sim \mathrm{Unif}(\{-1,1\})$ and $[W_0]_l \sim N(0, I_d/d)$ for any $l \in [m]$
    and set $\tau_0 \leftarrow 0$, $\theta_0 \leftarrow W_0$, and $\beta_0 \leftarrow W_0$.
for $k = 0, 1, \ldots, T-1$ do
    Update value parameter $\omega_k$ via Algorithm 4 with $\pi_k$, $r_k$, $W_0$, $b$, $T_{\mathrm{TD}}$, and $\alpha$ as the
        input.
    Sample $\{(s_i, a_i)\}_{i=1}^N$ from the state-action visitation measure $\nu_k$, and estimate $\widehat{\mathcal{I}}(\theta_k)$,
        $\widehat{\nabla}_\theta L(\theta_k, \beta_k)$, and $\widehat{\nabla}_\beta L(\theta_k, \beta_k)$ via (1.3.16), (1.3.17), and (1.3.18), respectively.
    Solve $\delta_k \leftarrow \mathrm{argmin}_{\delta\in S_\theta} \big\|\widehat{\mathcal{I}}(\theta_k)\cdot\delta - \tau_k \cdot \widehat{\nabla}_\theta L(\theta_k, \beta_k)\big\|_2$ and set $\tau_{k+1} \leftarrow \tau_k + \eta$.
    Update policy parameter $\theta$ via $\theta_{k+1} \leftarrow \tau_{k+1}^{-1} \cdot (\tau_k \cdot \theta_k - \eta \cdot \delta_k)$.
    Update reward parameter $\beta$ via $\beta_{k+1} \leftarrow \mathrm{Proj}_{S_{B_\beta}}\{\beta_k + \eta \cdot \widehat{\nabla}_\beta L(\theta_k, \beta_k)\}$.
end for
Ensure: Mixed policy $\bar{\pi}$ of $\pi_0, \ldots, \pi_{T-1}$.
```

1.4. Main Results

In this section, we first present the assumptions for our analysis. Then, we establish the global optimality and convergence of Algorithm 1.

1.4.1. Assumptions

We impose the following assumptions on the stationary distributions $\varrho_k \in \mathscr{P}(\mathcal{S}), \rho_k \in \mathscr{P}(\mathcal{S} \times \mathcal{A})$ and the visitation measures $d_k \in \mathscr{P}(\mathcal{S}), \nu_k \in \mathscr{P}(\mathcal{S} \times \mathcal{A})$.

Assumption 1.4.1. We assume that the following properties hold.

(a) Let μ be either ρ_k or ν_k. We assume for an absolute constant $c > 0$ that

$$\mathbb{E}_{(s,a)\sim\mu}\Big[\mathbb{1}\big\{|W^\top(s,a)| \le y\big\}\Big] \le \frac{c \cdot y}{\|W\|_2}, \quad \forall y > 0, W \neq 0.$$

(b) We assume for an absolute constant $C_h > 0$ that

$$\max_{k\in\mathbb{N}}\left\{\left\|\frac{\mathrm{d}d_{\mathrm{E}}}{\mathrm{d}d_k}\right\|_{2,d_k} + \left\|\frac{\mathrm{d}\nu_{\mathrm{E}}}{\mathrm{d}\nu_k}\right\|_{2,\nu_k}\right\} \le C_h,$$
$$\max_{k\in\mathbb{N}}\left\{\left\|\frac{\mathrm{d}d_{\mathrm{E}}}{\mathrm{d}\varrho_k}\right\|_{2,\varrho_k} + \left\|\frac{\mathrm{d}\nu_{\mathrm{E}}}{\mathrm{d}\rho_k}\right\|_{2,\rho_k}\right\} \le C_h.$$

Here $\mathrm{d}d_{\mathrm{E}}/\mathrm{d}d_k$, $\mathrm{d}\nu_{\mathrm{E}}/\mathrm{d}\nu_k$, $\mathrm{d}d_{\mathrm{E}}/\mathrm{d}\varrho_k$, and $\mathrm{d}\nu_{\mathrm{E}}/\mathrm{d}\rho_k$ are the Radon-Nikodym derivatives.

Assumption 1.4.1 (a) holds when the probability density functions of ρ_k and ν_k are uniformly upper bounded across k. Assumption 1.4.1 (b) states that the concentrability coefficients are uniformly upper bounded across k, which is commonly used in the analysis of RL (Szepesvári and Munos, 2005; Munos and Szepesvári, 2008; Antos et al., 2008; Farahmand et al., 2010; Scherrer et al., 2015; Farahmand et al., 2016; Lazaric et al., 2016).

For notational simplicity, we write $u_0(s,a) = u_{W_0}(s,a)$ and $\phi_0(s,a) = \phi_{W_0}(s,a)$, where $u_{W_0}(s,a)$ is the neural network defined in (1.3.1) with $W = W_0$, $\phi_{W_0}(s,a)$ is the feature vector defined in (1.3.2) with $W = W_0$, and W_0 is the random initialization in (1.3.3). We impose the following assumptions on the neural network $u_0(s,a)$ and the transition kernel P.

Assumption 1.4.2. We assume that the following properties hold.

(a) Let $\bar{U} = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |u_0(s,a)|$. We assume for absolute constants $M_0 > 0$ and $v > 0$ that

$$\mathbb{E}_{\text{init}}[\bar{U}^2] \le M_0^2, \quad \mathbb{P}(\bar{U} > t) \le \exp(-v \cdot t^2), \quad \forall t > 2M_0. \tag{1.4.1}$$

(b) We assume that the transition kernel P belongs to the following class,

$$\widetilde{\mathcal{M}}_{\infty,B_P} = \biggl\{ P(s' \,|\, s,a) = \int \vartheta(s,a;w)^\top \varphi(s';w)\,\mathrm{d}q(w) \biggm| \sup_w \biggl\| \int \varphi(s;w)\mathrm{d}s \biggr\|_2 \le B_P \biggr\}.$$

Here $B_P > 0$ is an absolute constant, q is the probability density function of $N(0, I_d/d)$, and $\vartheta(s,a;w)$ is defined as $\vartheta(s,a;w) = \mathbb{1}\{w^\top(s,a) > 0\} \cdot (s,a)$.

Assumption 1.4.2 (b) states that the MDP belongs to (a variant of) the class of linear MDPs (Yang and Wang, 2019a,b; Jin et al., 2019; Cai et al., 2019b). However, our class of transition kernels is infinite-dimensional, and thus, captures a rich class of MDPs. To understand Assumption 1.4.2 (a), recall that we initialize the neural network with $[W_0]_l \sim N(0, I_d/d)$ and $b_l \sim \text{Unif}(\{-1,1\})$ for any $l \in [m]$. Thus, the neural network $u_0(s,a)$ defined in (1.3.1) with $W = W_0$ converges to a Gaussian process indexed by $(s,a) \in \mathcal{S}\times\mathcal{A}$ as m goes to infinity. Following from the facts that the maximum of a Gaussian process over a compact index set is sub-Gaussian (van de Geer and Muro, 2014) and that $\mathcal{S}\times\mathcal{A}$ is compact, it is reasonable to assume that $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |u_0(s,a)|$ is sub-Gaussian, which further implies the existence of the absolute constants M_0 and v in

(1.4.1) of Assumption 1.4.2 (a). Moreover, if we assume that m is even and initialize the parameters W_0, b as follows,

$$\begin{cases} [W_0]_l \overset{\text{i.i.d.}}{\sim} N(0, I_d/d), \quad b_l \sim \mathrm{Unif}\big(\{-1,1\}\big), & \forall l = 1, \ldots, m/2, \\ [W_0]_l = [W_0]_{l-m/2}, \quad b_l = -b_{l-m/2}, & \forall l = m/2+1, \ldots, m, \end{cases} \tag{1.4.2}$$

we have that $u_0(s,a) = 0$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, which allows us to set $M_0 = 0$ and $v = +\infty$ in Assumption 1.4.2 (a). Also, it holds that $0 = u_0(s,a) \in \mathcal{R}_\beta$, which implies that $\mathbb{D}_{\mathcal{R}_\beta}(\pi_1, \pi_2) \geq 0$ for any π_1 and π_2. The proof of our results with the random initialization in (1.4.2) is identical.

Finally, we impose the following assumption on the regularizer $\psi(\beta)$ and the variances of the estimators $\widehat{\mathcal{I}}(\theta)$, $\widehat{\nabla}_\theta L(\theta, \beta)$, and $\widehat{\nabla}_\beta L(\theta, \beta)$ defined in (1.3.16), (1.3.17), and (1.3.18), respectively.

Assumption 1.4.3. We assume that the following properties hold.

(a) We assume for an absolute constant $\sigma > 0$ that

$$\mathbb{E}_k\Big[\Big\|\widehat{\mathcal{I}}(\theta_k)W - \mathbb{E}_k\big[\widehat{\mathcal{I}}(\theta_k)W\big]\Big\|_2^2\Big] \leq \tau_k^4 \cdot \sigma^2/N, \quad \forall W \in S_{B_\theta}, \tag{1.4.3}$$

$$\mathbb{E}_k\Big[\Big\|\widehat{\nabla}_\theta L(\theta_k, \beta_k) - \mathbb{E}_k\big[\widehat{\nabla}_\theta L(\theta_k, \beta_k)\big]\Big\|_2^2\Big] \leq \tau_k^2 \cdot \sigma^2/N, \tag{1.4.4}$$

$$\mathbb{E}_k\Big[\Big\|\widehat{\nabla}_\beta L(\theta_k, \beta_k) - \mathbb{E}_k\big[\widehat{\nabla}_\beta L(\theta_k, \beta_k)\big]\Big\|_2^2\Big] \leq \sigma^2/N, \tag{1.4.5}$$

where τ_k is the inverse temperature parameter in (1.3.5), $N \in \mathbb{N}_+$ is the batch size, and S_{B_θ} is the parameter domain of θ defined in (1.3.4) with the domain radius B_θ. Here the expectation $\mathbb{E}_k$ is taken with respect to the k-th batch, which is drawn from ν_k given θ_k.

(b) We assume that the regularizer $\psi(\beta)$ in (1.2.4) is convex and L_ψ-Lipschitz continuous over the compact parameter domain S_{B_β}.

Assumption 1.4.3 (a) holds when $\widehat{Q}_{\omega_k}(s_i, a_i) \cdot \iota_{\theta_k}(s_i, a_i)$, $\iota_{\theta_k}(s_i, a_i)\iota_{\theta_k}(s_i, a_i)^\top$, and $\phi_{\beta_k}(s_i, a_i)$ have uniformly upper bounded variances across $i \in [m]$ and k, and the Markov chain that generates $\{(s_i, a_i)\}_{i\in[N]}$ mixes sufficiently fast (Zhang et al., 2019a). Similar assumptions are also used in the analysis of policy optimization (Xu et al., 2019a,b). Also, Assumption 1.4.3 (b) holds for most commonly used regularizers (Ho and Ermon, 2016).

1.4.2. Global Optimality and Convergence

In this section, we establish the global optimality and convergence of Algorithm 1. The following proposition adapted from Cai et al. (2019c) characterizes the global optimality and convergence of neural TD, which is presented in Algorithm 4.

Proposition 1.4.4 (Global Optimality and Convergence of Neural TD)**.** In Algorithm 4, we set $T_{\mathrm{TD}} = \Omega(m)$, $\alpha = \min\{(1-\gamma)/8, m^{-1/2}\}$, and $B_\omega = c \cdot (B_\beta + B_P \cdot (M_0 + B_\beta))$ for an absolute constant $c > 0$. Let π_k, r_k be the input and ω_k be the output of Algorithm 4. Under Assumptions 1.4.1 and 1.4.2, it holds for an absolute constant $C_v > 0$ that

(1.4.6)

$$\mathbb{E}_{\mathrm{init}}\Big[\big\|Q_{\omega_k}(s,a) - Q^{\pi_k}_{r_k}(s,a)\big\|^2_{2,\rho_k}\Big] = \mathcal{O}\big(B^3_\omega \cdot m^{-1/2} + B^{5/2}_\omega \cdot m^{-1/4} + B^2_\omega \cdot \exp(-C_v \cdot B^2_\omega)\big).$$

Here the expectation $\mathbb{E}_{\mathrm{init}}$ is taken with respect to the random initialization in (1.3.3).

Proof. See §A.2.1 for a detailed proof. □

The term $B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2)$ in (1.4.6) of Proposition 1.4.4 characterizes the hardness of estimating the state-action value function $Q_{r_k}^{\pi_k}(s,a)$ by the neural network defined in (1.3.1), which arises because $\|Q_{r_k}^{\pi_k}(s,a)\|_\infty$ is not uniformly upper bounded across k. Note that if we employ the random initialization in (1.4.2), we have that $C_v = +\infty$. And consequently, such a term vanishes. We are now ready to establish the global optimality and convergence of Algorithm 1.

Theorem 1.4.5 (Global Optimality and Convergence of GAIL)**.** We set $\eta = 1/\sqrt{T}$ and $B_\omega = c \cdot (B_\beta + B_P \cdot (M_0 + B_\beta))$ for an absolute constant $c > 0$, and $B_\theta = B_\omega$ in Algorithm 1. Let $\bar{\pi}$ be the output of Algorithm 1. Under Assumptions 1.4.1-1.4.3, it holds that

$$\mathbb{E}\big[\mathbb{D}_{\mathcal{R}_\beta}(\pi_{\mathrm{E}}, \bar{\pi})\big] \le \underbrace{\frac{(1-\gamma)^{-1} \cdot \log|\mathcal{A}| + 13\bar{B}^2 + M_0^2 + 8}{\sqrt{T}}}_{\text{(i)}} + \underbrace{2\lambda \cdot L_\psi \cdot \bar{B}}_{\text{(ii)}} + \underbrace{\frac{1}{T}\sum_{k=0}^{T-1} \varepsilon_k}_{\text{(iii)}}. \tag{1.4.7}$$

Here $\bar{B} = \max\{B_\theta, B_\omega, B_\beta\}$, $\mathbb{D}_{\mathcal{R}_\beta}$ is the $\mathcal{R}_\beta$-distance defined in (1.2.5) with $\mathcal{R}_\beta = \{r_\beta(s,a) \,|\, \beta \in S_{B_\beta}\}$, the expectation is taken with respect to the random initialization in (1.3.3) and the T batches, and the error term ε_k satisfies that

$$\varepsilon_k = \underbrace{2\sqrt{2} \cdot C_h \cdot \bar{B} \cdot \sigma \cdot N^{-1/2}}_{\text{(iii.a)}} + \underbrace{\epsilon_{Q,k}}_{\text{(iii.b)}} + \underbrace{\mathcal{O}(k \cdot \bar{B}^{3/2} \cdot m^{-1/4} + \bar{B}^{5/4} \cdot m^{-1/8})}_{\text{(iii.c)}}, \tag{1.4.8}$$

where C_h is defined in Assumption 1.4.1, L_ψ and σ are defined in Assumption 1.4.3, and $\epsilon_{Q,k} = \mathcal{O}(B_\omega^3 \cdot m^{-1/2} + B_\omega^{5/2} \cdot m^{-1/4} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2))$ is the error induced by neural TD (Algorithm 4).

Proof. See §1.5 for a detailed proof. □

The optimality gap in (1.4.7) of Theorem 1.4.5 is measured by the expected $\mathcal{R}_\beta$-distance $\mathbb{D}_{\mathcal{R}_\beta}(\pi_{\mathrm{E}}, \bar{\pi})$ between the expert policy π_{E} and the learned policy $\bar{\pi}$. Thus, by showing that the optimality gap is upper bounded by $\mathcal{O}(1/\sqrt{T})$, we prove that $\bar{\pi}$ (approximately) outperforms the expert policy π_{E} in expectation when the number of iterations T goes to infinity. As shown in (1.4.7) of Theorem 1.4.5, the optimality gap is upper bounded by the sum of the three terms. Term (i) corresponds to the $1/\sqrt{T}$ rate of convergence of Algorithm 1. Term (ii) corresponds to the bias induced by the regularizer $\lambda \cdot \psi(\beta)$ in the objective function $L(\theta, \beta)$ defined in (1.2.4). Term (iii) is upper bounded by the sum of the three terms in (1.4.8) of Theorem 1.4.5. In detail, term (iii.a) corresponds to the error induced by the variances of $\widehat{\mathcal{I}}(\theta)$, $\widehat{\nabla}_\theta L(\theta, \beta)$, and $\widehat{\nabla}_\beta L(\theta, \beta)$ defined in (1.4.3), (1.4.4), and (1.4.5) of Assumption 1.4.3, which vanishes as the batch size N in Algorithm 1 goes to infinity. Term (iii.b) is the error of estimating $Q_r^\pi(s, a)$ by $\widehat{Q}_\omega(s, a)$ using neural TD (Algorithm 4). As shown in Proposition 1.4.4, the estimation error $\epsilon_{Q,k}$ vanishes as m and B_ω go to infinity. Term (iii.c) corresponds to the linearization error of the neural network defined in (1.3.1), which is characterized in Lemma A.1.2. Following from Theorem 1.4.5, it holds for $B_\omega = \Omega((C_v^{-1} \cdot \log T)^{1/2})$, $m = \Omega(\bar{B}^{10} \cdot T^6)$, and $N = \Omega(\bar{B}^2 \cdot T \cdot \sigma^2)$ that $\mathbb{E}\big[\mathbb{D}_{\mathcal{R}_\beta}(\pi_{\mathrm{E}}, \bar{\pi})\big] = \mathcal{O}(T^{-1/2} + \lambda)$, which implies the $1/\sqrt{T}$ rate of convergence of Algorithm 1 (up to the bias induced by the regularizer).

1.5. Proof of Main Results

In this section, we present the proof of Theorem 1.4.5, which establishes the global optimality and convergence of Algorithm 1. For notational simplicity, we write $\pi^s(a) = \pi(a \,|\, s)$ for any policy π, state $s \in \mathcal{S}$, and action $a \in \mathcal{A}$. For any policies π_1, π_2 and distribution μ over $\mathcal{S}$, we denote the expected Kullback-Leibler (KL) divergence by KL^{μ}, which is defined as $\mathrm{KL}^{\mu}(\pi_1 \,\|\, \pi_2) = \mathbb{E}_{s\sim\mu}[\mathrm{KL}(\pi_1^s \,\|\, \pi_2^s)]$. For any visitation measures $d_\pi \in \mathscr{P}(\mathcal{S})$ and $\nu_\pi \in \mathscr{P}(\mathcal{S}\times\mathcal{A})$, we denote by $\mathbb{E}_{d_\pi}$ and $\mathbb{E}_{\nu_\pi}$ the expectations taken with respect to $s \sim d_\pi$ and $(s,a) \sim \nu_\pi$, respectively.

Following from the property of the mixed policy $\bar{\pi}$ in (1.3.21), we have that

$$\begin{aligned}\mathbb{E}\big[\mathbb{D}_{\mathcal{R}_\beta}(\pi_{\mathrm{E}},\bar{\pi})\big] &= \mathbb{E}\big[\max_{\beta'\in S_{B_\beta}} J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\bar{\pi}; r_{\beta'})\big] \\ &= \mathbb{E}\biggl[\max_{\beta'\in S_{B_\beta}} \frac{1}{T}\sum_{k=0}^{T-1} J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\pi_k; r_{\beta'})\biggr]. \end{aligned} \tag{1.5.1}$$

We now upper bound the optimality gap in (1.5.1) by upper bounding the following difference of expected cumulative rewards,

$$J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\pi_k; r_{\beta'}) = \underbrace{J(\pi_{\mathrm{E}}; r_k) - J(\pi_k; r_k)}_{\text{(i)}} + \underbrace{L(\theta_k, \beta') - L(\theta_k, \beta_k)}_{\text{(ii)}} + \underbrace{\lambda\cdot\big(\psi(\beta') - \psi(\beta_k)\big)}_{\text{(iii)}}, \tag{1.5.2}$$

where $\beta' \in S_{B_\beta}$ is chosen arbitrarily and $L(\theta,\beta)$ is the objective function defined in (1.2.4). Following from Assumption 1.4.3 and the fact that $\beta_k, \beta' \in S_{B_\beta}$, we have that

$$\lambda\cdot\big(\psi(\beta') - \psi(\beta_k)\big) \le \lambda \cdot L_\psi \cdot \|\beta' - \beta_k\|_2 \le \lambda\cdot L_\psi \cdot 2B_\beta, \tag{1.5.3}$$

which upper bounds term (iii) of (1.5.2). It remains to upper bound terms (i) and (ii) of (1.5.2), which hinges on the one-point convexity of $J(\pi; r)$ with respect to π and the (approximate) convexity of $L(\theta,\beta)$ with respect to β.

Upper bound of term (i) in (1.5.2)**.** In what follows, we upper bound term (i) of (1.5.2). We first introduce the following cost difference lemma (Kakade and Langford, 2002), which corresponds to the one-point convexity of $J(\pi; r)$ with respect to π. Recall that $d_{\mathrm{E}} \in \mathscr{P}(\mathcal{S})$ is the state visitation measure induced by the expert policy π_{E}.

Lemma 1.5.1 (Cost Difference Lemma, Lemma 6.1 in Kakade and Langford (2002))**.** For any policy π and reward function $r(s, a)$, it holds that

$$J(\pi_{\mathrm{E}}; r) - J(\pi; r) = (1-\gamma)^{-1} \cdot \mathbb{E}_{d_{\mathrm{E}}}\Bigl[\bigl\langle Q_r^{\pi}(s, \cdot), \pi_{\mathrm{E}}^s - \pi^s \bigr\rangle_{\mathcal{A}}\Bigr], \tag{1.5.4}$$

where γ is the discount factor.

Furthermore, we establish the following lemma, which upper bounds the right-hand side of (1.5.4) in Lemma 1.5.1.

Lemma 1.5.2. Under Assumptions 1.4.1-1.4.3, we have that

$$\mathbb{E}_{d_{\mathrm{E}}}\Bigl[\bigl\langle Q_{r_k}^{\pi_k}(s, \cdot), \pi_{\mathrm{E}}^s - \pi_k^s \bigr\rangle_{\mathcal{A}}\Bigr] = \eta^{-1} \cdot \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_k) - \eta^{-1} \cdot \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_{k+1}) + \Delta_k^{(\mathrm{i})},$$

where

$$\begin{aligned} \mathbb{E}\bigl[|\Delta_k^{(\mathrm{i})}|\bigr] = {} & 2\sqrt{2} \cdot C_h \cdot B_\theta^{1/2} \cdot \sigma^{1/2} \cdot N^{-1/4} + \epsilon_{Q,k} + \eta \cdot (M_0^2 + 9B_\theta^2) \\ & + \mathcal{O}(\eta^{-1} \cdot \tau_{k+1} \cdot B_\theta^{3/2} \cdot m^{-1/4} + B_\theta^{5/4} \cdot m^{-1/8}). \end{aligned} \tag{1.5.5}$$

Here M_0 is defined in Assumption 1.4.2, σ is defined in Assumption 1.4.3, N is the batch size in (1.3.16)-(1.3.18), and $\epsilon_{Q,k} = \mathcal{O}(B_\omega^3 \cdot m^{-1/2} + B_\omega^{5/2} \cdot m^{-1/4} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2))$ for an absolute constant $C_v > 0$, which depends on the absolute constant v in Assumption 1.4.2.

Proof. See §A.3.2 for a detailed proof. □

Combining Lemmas 1.5.1 and 1.5.2, we have that

$$J(\pi_{\rm E}; r_k) - J(\pi_k; r_k) \le \frac{{\rm KL}^{d_{\rm E}}(\pi_{\rm E} \,\|\, \pi_k) - {\rm KL}^{d_{\rm E}}(\pi_{\rm E} \,\|\, \pi_{k+1}) + \eta \cdot \Delta_k^{(\rm i)}}{\eta \cdot (1-\gamma)}, \tag{1.5.6}$$

which upper bounds term (i) of (1.5.2). Here $\Delta_k^{(\rm i)}$ is upper bounded in (1.5.5) of Lemma 1.5.2.

Upper bound of term (ii) in (1.5.2)**.** In what follows, we upper bound term (ii) of (1.5.2). We first establish the following lemma, which characterizes the (approximate) convexity of $L(\theta, \beta)$ with respect to β.

Lemma 1.5.3. Under Assumption 1.4.1, it holds for any $\beta' \in S_{B_\beta}$ that

$$\mathbb{E}_{\rm init}\big[L(\theta_k, \beta') - L(\theta_k, \beta_k)\big] = \mathbb{E}_{\rm init}\big[\nabla_\beta L(\theta_k, \beta_k)^\top (\beta' - \beta_k)\big] + \mathcal{O}(B_\beta^{3/2} \cdot m^{-1/4}).$$

Proof. See §A.3.3 for a detailed proof. □

The term $\mathcal{O}(B_\beta^{3/2} \cdot m^{-1/4})$ in Lemma 1.5.3 arises from the linearization error of the neural network, which is characterized in Lemma A.1.2. Based on Lemma 1.5.3, we establish the following lemma, which upper bounds term (ii) of (1.5.2).

Lemma 1.5.4. Under Assumptions 1.4.1 and 1.4.3, we have that

$$L(\theta_k, \beta') - L(\theta_k, \beta_k) \le \eta^{-1} \cdot \|\beta_k - \beta'\|_2^2 - \eta^{-1} \cdot \|\beta_{k+1} - \beta'\|_2^2 - \eta^{-1} \cdot \|\beta_{k+1} - \beta_k\|_2^2 + \Delta_k^{(\rm ii)},$$

where

$$\mathbb{E}\big[|\Delta_k^{(\rm ii)}|\big] = \eta \cdot \big((2 + \lambda \cdot L_\psi)^2 + \sigma^2 \cdot N^{-1}\big) + 2B_\beta \cdot \sigma \cdot N^{-1/2} + \mathcal{O}(B_\beta^{3/2} \cdot m^{-1/4}). \tag{1.5.7}$$

Proof. See §A.3.4 for a detailed proof. □

By Lemma 1.5.4, we have that

$$L(\theta_k, \beta') - L(\theta_k, \beta_k) \leq \eta^{-1} \cdot \big(\|\beta_k - \beta'\|_2^2 - \|\beta_{k+1} - \beta'\|_2^2 - \|\beta_{k+1} - \beta_k\|_2^2\big) + \Delta_k^{(\mathrm{ii})}, \tag{1.5.8}$$

which upper bounds term (ii) of (1.5.2). Here $\Delta_k^{(\mathrm{ii})}$ is upper bounded in (1.5.7) of Lemma 1.5.4.

Plugging (1.5.3), (1.5.6), and (1.5.8) into (1.5.2), we obtain that

(1.5.9)

$$\begin{aligned} &J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\pi_k; r_{\beta'}) \\ &\quad \leq \frac{\mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_k) - \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_{k+1})}{\eta \cdot (1-\gamma)} + \eta^{-1} \cdot \big(\|\beta_k - \beta'\|_2^2 - \|\beta_{k+1} - \beta'\|_2^2\big) + 2\lambda \cdot L_\psi \cdot B_\beta + \Delta_k. \end{aligned}$$

Here $\Delta_k = \Delta_k^{(\mathrm{i})} + \Delta_k^{(\mathrm{ii})}$, where $\Delta_k^{(\mathrm{i})}$ and $\Delta_k^{(\mathrm{ii})}$ are upper bounded in (1.5.5) and (1.5.7) of Lemmas 1.5.2 and 1.5.4, respectively. Note that the upper bound of Δ_k does not depend on θ and β. Upon telescoping (1.5.9) with respect to k, we obtain that

(1.5.10)

$$\begin{aligned} J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\bar{\pi}; r_{\beta'}) &= \frac{1}{T}\sum_{k=0}^{T-1} \big[J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\pi_k; r_{\beta'})\big] \\ &\leq \frac{(1-\gamma)^{-1} \cdot \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_0) + \|\beta_0 - \beta'\|_2^2}{\eta \cdot T} + 2\lambda \cdot L_\psi \cdot B_\beta + \frac{1}{T}\sum_{k=0}^{T-1} |\Delta_k|. \end{aligned}$$

Following from the fact that $\tau_0 = 0$ and the parameterization of π_θ in (1.3.5), it holds that π_0^s is the uniform distribution over $\mathcal{A}$ for any $s \in \mathcal{S}$. Thus, we have $\mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_0) \leq \log|\mathcal{A}|$. Meanwhile, following from the fact that $\beta' \in S_{B_\beta}$, it holds that $\|\beta' - \beta_0\|_2 \leq B_\beta$. Finally, by setting $\eta = T^{-1/2}$, $\tau_k = k \cdot \eta$, and $\bar{B} = \max\{B_\theta, B_\beta\}$ in (1.5.10), we have that

$$\begin{aligned}
\mathbb{E}\big[\mathbb{D}_{\mathcal{R}_\beta}(\pi_{\mathrm{E}}, \bar{\pi})\big] &= \mathbb{E}\big[\max_{\beta' \in S_{B_\beta}} J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\bar{\pi}; r_{\beta'})\big] \\
&\le \frac{(1-\gamma)^{-1} \cdot \log|\mathcal{A}| + 4B_\beta^2}{\eta \cdot T} + 2\lambda \cdot L_\psi \cdot B_\beta + \frac{\mathbb{E}\big[\max_{\beta'} \sum_{k=0}^{T-1} |\Delta_k|\big]}{T} \\
&= \frac{(1-\gamma)^{-1} \cdot \log|\mathcal{A}| + 13\bar{B}^2 + M_0^2 + 8}{\sqrt{T}} + 2\lambda \cdot L_\psi \cdot \bar{B} + \frac{\sum_{k=0}^{T-1} \varepsilon_k}{T}.
\end{aligned}$$

Here ε_k is upper bounded as follows,

$$\varepsilon_k = 2\sqrt{2} \cdot C_h \cdot \bar{B} \cdot \sigma \cdot N^{-1/2} + \epsilon_{Q,k} + \mathcal{O}(k \cdot \bar{B}^{3/2} \cdot m^{-1/4} + \bar{B}^{5/4} \cdot m^{-1/8}),$$

where $\epsilon_{Q,k} = \mathcal{O}(B_\omega^3 \cdot m^{-1/2} + B_\omega^{5/2} \cdot m^{-1/4} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2))$ for an absolute constant $C_v > 0$. Thus, we complete the proof of Theorem 1.4.5.

CHAPTER 2

Can Temporal-Difference and Q-Learning Learn Representation? A Mean-Field Theory

Temporal-difference and Q-learning play a key role in deep reinforcement learning, where they are empowered by expressive nonlinear function approximators such as neural networks. At the core of their empirical successes is the learned feature representation, which embeds rich observations, e.g., images and texts, into the latent space that encodes semantic structures. Meanwhile, the evolution of such a feature representation is crucial to the convergence of temporal-difference and Q-learning.

In particular, temporal-difference learning converges when the function approximator is linear in a feature representation, which is fixed throughout learning, and possibly diverges otherwise. We aim to answer the following questions: *When the function approximator is a neural network, how does the associated feature representation evolve? If it converges, does it converge to the optimal one?*

We prove that, utilizing an overparameterized two-layer neural network, temporal-difference and Q-learning globally minimize the mean-squared projected Bellman error at a sublinear rate. Moreover, the associated feature representation converges to the optimal one, generalizing the previous analysis of Cai et al. (2019c) in the neural tangent kernel regime, where the associated

feature representation stabilizes at the initial one. The key to our analysis is a mean-field perspective, which connects the evolution of a finite-dimensional parameter to its limiting counterpart over an infinite-dimensional Wasserstein space. Our analysis generalizes to soft Q-learning, which is further connected to policy gradient.

2.1. Introduction

Deep reinforcement learning achieves phenomenal empirical successes, especially in challenging applications where an agent acts upon rich observations, e.g., images and texts. Examples include video gaming (Mnih et al., 2015), visuomotor manipulation (Levine et al., 2016), and language generation (He et al., 2015). Such empirical successes are empowered by expressive nonlinear function approximators such as neural networks, which are used to parameterize both policies (actors) and value functions (critics) (Konda and Tsitsiklis, 2000). In particular, the neural network learned from interacting with the environment induces a data-dependent feature representation, which embeds rich observations into a latent space encoding semantic structures (Hinton, 1986; Bengio, 2012; Yosinski et al., 2014; LeCun et al., 2015). In contrast, classical reinforcement learning mostly relies on a handcrafted feature representation that is fixed throughout learning (Sutton and Barto, 2018).

In this paper, we study temporal-difference (TD) (Sutton, 1988) and Q-learning (Watkins and Dayan, 1992), two of the most prominent algorithms in deep reinforcement learning, which are further connected to policy gradient (Williams, 1992) through its equivalence to soft Q-learning (O'Donoghue et al., 2016; Schulman et al., 2017; Nachum et al., 2017; Haarnoja et al., 2017). In particular, we aim to characterize how an overparameterized two-layer neural network and its induced feature representation evolve in TD and Q-learning, especially their rate of convergence and global optimality. A fundamental obstacle, however, is that such an evolving feature representation possibly leads to the divergence of TD and Q-learning. For example, TD converges when the value function approximator is linear in a feature representation, which is fixed throughout learning, and possibly diverges otherwise (Baird, 1995; Boyan and Moore, 1995; Tsitsiklis and Van Roy, 1997).

To address such an issue of divergence, nonlinear gradient TD (Bhatnagar et al., 2009) explicitly linearizes the value function approximator locally at each iteration, that is, using its gradient with respect to the parameter as an evolving feature representation. Although nonlinear gradient TD converges, it is unclear whether the attained solution is globally optimal. On the other hand, when the value function approximator in TD is an overparameterized multi-layer neural network, which is required to be properly scaled, such a feature representation stabilizes at the initial one (Cai et al., 2019c), making the explicit local linearization in nonlinear gradient TD unnecessary. Moreover, the implicit local linearization enabled by overparameterization allows TD (and Q-learning) to converge to the globally optimal solution. However, such a required scaling, also known as the neural tangent kernel (NTK) regime (Jacot et al., 2018), effectively constrains the evolution of the induced feature presentation to an infinitesimal neighborhood of the initial one, which is not data-dependent.

Contribution. Going beyond the NTK regime, we prove that, when the value function approximator is an overparameterized two-layer neural network, TD and Q-learning globally minimize the mean-squared projected Bellman error (MSPBE) at a sublinear rate. Moreover, in contrast to the NTK regime, the induced feature representation is able to deviate from the initial one and subsequently evolve into the globally optimal one, which corresponds to the global minimizer of the MSPBE. We further extend our analysis to soft Q-learning, which is connected to policy gradient.

The key to our analysis is a mean-field perspective, which allows us to associate the evolution of a finite-dimensional parameter with its limiting counterpart over an infinite-dimensional Wasserstein space (Villani, 2003, 2008; Ambrosio et al., 2008; Ambrosio and Gigli, 2013). Specifically, by exploiting the permutation invariance of the parameter, we associate the neural network and its induced feature representation with an empirical distribution, which, at the infinite-width limit, further corresponds to a population distribution. The evolution of such a population distribution is characterized by a partial differential equation (PDE) known as the continuity equation. In particular, we develop a generalized notion of strongly monotonicity (Harker and Pang, 1990), which is tailored to the Wasserstein space, especially the first variation formula therein (Ambrosio et al., 2008), to characterize the evolution of such a PDE solution, which, by a discretization argument, further quantifies the evolution of the induced feature representation.

Related Work. When the value function approximator is linear, the convergence of TD is extensively studied in both continuous-time (Jaakkola et al., 1994; Tsitsiklis and Van Roy, 1997; Borkar and Meyn, 2000; Kushner and Yin, 2003; Borkar, 2009) and discrete-time (Bhandari et al., 2018; Lakshminarayanan and Szepesvári, 2018; Dalal et al., 2018; Srikant and Ying, 2019) settings. See Dann et al. (2014) for a detailed survey. Also, when the value function approximator is linear, Melo et al. (2008); Zou et al. (2019); Chen et al. (2019b) study the convergence of Q-learning. When the value function approximator is nonlinear, TD possibly diverges (Baird, 1995; Boyan and Moore, 1995; Tsitsiklis and Van Roy, 1997). Bhatnagar et al. (2009) propose nonlinear gradient TD, which converges but only to a locally optimal solution. See Geist and Pietquin (2013); Bertsekas (2019) for a detailed survey. When the value function approximator is an overparameterized multi-layer neural network, Cai et al. (2019c) prove that TD converges to the globally optimal solution in the NTK regime. See also the independent work of Brandfonbrener and Bruna (2019a,b); Agazzi and Lu (2019); Sirignano and Spiliopoulos (2019), where the state space is required to be finite. In contrast to the previous analysis in the NTK regime, our analysis allows TD to attain a data-dependent feature representation that is globally optimal.

Meanwhile, our analysis is related to the recent breakthrough in the mean-field analysis of stochastic gradient descent (SGD) for the supervised learning of an overparameterized two-layer neural network (Chizat and Bach, 2018b; Mei et al., 2018, 2019; Javanmard et al., 2019; Wei et al., 2019; Fang et al., 2019a,b; Chen et al., 2020b). See also the previous analysis in the NTK regime (Daniely, 2017; Chizat and Bach, 2018a; Jacot et al., 2018; Li and Liang, 2018; Allen-Zhu et al., 2018a,b; Du et al., 2018a,b; Zou et al., 2018; Arora et al., 2019a,b; Lee et al., 2019b; Cao and Gu, 2019a; Chen et al., 2019a; Zou and Gu, 2019; Ji and Telgarsky, 2019; Bai and Lee, 2019). Specifically, the previous mean-field analysis casts SGD as the Wasserstein gradient flow of an energy functional, which corresponds to the objective function in supervised learning. In contrast, TD follows the stochastic semigradient of the MSPBE (Sutton and Barto, 2018), which is biased. As a result, there does not exist an energy functional for casting TD as its Wasserstein gradient flow. Instead, our analysis combines a generalized notion of strongly monotonicity (Harker and Pang,

1990) and the first variation formula in the Wasserstein space (Ambrosio et al., 2008), which is of independent interest.

Notations. We denote by $\mathscr{B}(\mathcal{X})$ the Borel σ-algebra over the space $\mathcal{X}$. Let $\mathscr{P}(\mathcal{X})$ be the set of Borel probability measures over the measurable space $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$. We denote by $[N] = \{1, 2, \ldots, N\}$ for any $N \in \mathbb{N}_+$. Also, we denote by $\mathcal{B}^n(x; r) = \{y \in \mathbb{R}^n \,|\, \|y - x\| \leq r\}$ the closed ball in $\mathbb{R}^n$. Given a curve $\rho : \mathbb{R} \to \mathcal{X}$, we denote by $\rho'_s = \partial_t \rho_t \,|_{\,t=s}$ its derivative with respect to the time. For a function $f : \mathcal{X} \to \mathbb{R}$, we denote by $\mathrm{Lip}(f) = \sup_{x,y \in \mathcal{X}, x \neq y} |f(x) - f(y)| / \|x - y\|$ its Lipschitz constant. For an operator $F : \mathcal{X} \to \mathcal{X}$ and a measure $\mu \in \mathscr{P}(\mathcal{X})$, we denote by $F_\sharp \mu = \mu \circ F^{-1}$ the push forward of μ through F. We denote by D_{KL} and D_{χ^2} the Kullback-Leibler (KL) divergence and the χ^2 divergence, respectively.

2.2. Background

2.2.1. Policy Evaluation

We consider a Markov decision process $(\mathcal{S}, \mathcal{A}, P, R, \gamma, \mathcal{D}_0)$, where $\mathcal{S} \subseteq \mathbb{R}^{d_1}$ is the state space, $\mathcal{A} \subseteq \mathbb{R}^{d_2}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathcal{S})$ is the transition kernel, $R : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathbb{R})$ is the reward distribution, $\gamma \in (0, 1)$ is the discount factor, and $\mathcal{D}_0 \in \mathscr{P}(\mathcal{S})$ is the initial state distribution. An agent following a policy $\pi : \mathcal{S} \to \mathscr{P}(\mathcal{A})$ interacts with the environment in the following manner. At a state s_t, the agent takes an action a_t according to $\pi(\cdot \,|\, s_t)$ and receives from the environment a random reward r_t following $R(\cdot \,|\, s_t, a_t)$. Then, the environment transits into the next state s_{t+1} according to $P(\cdot \,|\, s_t, a_t)$. We measure the performance of a policy π via the expected cumulative reward $J(\pi)$, which is defined as follows,

$$J(\pi) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \cdot r_t \,\Big|\, s_0 \sim \mathcal{D}_0, a_t \sim \pi(\cdot \,|\, s_t), r_t \sim R(\cdot \,|\, s_t, a_t), s_{t+1} \sim P(\cdot \,|\, s_t, a_t)\Big]. \tag{2.2.1}$$

In policy evaluation, we are interested in the state-action value function (Q-function) $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which is defined as follows,

$$Q^\pi(s, a) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \cdot r_t \,\Big|\, s_0 = s, a_0 = a, a_t \sim \pi(\cdot \,|\, s_t), r_t \sim R(\cdot \,|\, s_t, a_t), s_{t+1} \sim P(\cdot \,|\, s_t, a_t)\Big].$$

We learn the Q-function by minimizing the mean-squared Bellman error (MSBE), which is defined as follows,

$$\mathrm{MSBE}(Q) = \frac{1}{2} \cdot \mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[\big(Q(s, a) - \mathcal{T}^\pi Q(s, a)\big)^2\Big].$$

Here $\mathcal{D} \in \mathscr{P}(\mathcal{S} \times \mathcal{A})$ is the stationary distribution induced by the policy π of interest and $\mathcal{T}^\pi$ is the corresponding Bellman operator, which is defined as follows,

$$\mathcal{T}^\pi Q(s, a) = \mathbb{E}\big[r + \gamma \cdot Q(s', a') \,\big|\, r \sim R(\cdot \,|\, s, a), s' \sim P(\cdot \,|\, s, a), a' \sim \pi(\cdot \,|\, s')\big].$$

However, $\mathcal{T}^\pi Q$ may be not representable by a given function class $\mathcal{F}$. Hence, we turn to minimizing a surrogate of the MSBE over $Q \in \mathcal{F}$, namely the mean-squared projected Bellman error (MSPBE), which is defined as follows,

$$\mathrm{MSPBE}(Q) = \frac{1}{2} \cdot \mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[\big(Q(s, a) - \Pi_{\mathcal{F}} \mathcal{T}^\pi Q(s, a)\big)^2\Big], \tag{2.2.2}$$

where $\Pi_{\mathcal{F}}$ is the projection onto $\mathcal{F}$ with respect to the $\mathcal{L}_2(\mathcal{D})$-norm. The global minimizer of the MSPBE is the fixed point solution to the projected Bellman equation $Q = \Pi_{\mathcal{F}} \mathcal{T}^\pi Q$.

In temporal-difference (TD) learning, corresponding to the MSPBE defined in (2.2.2), we parameterize the Q-function with $\widehat{Q}(\cdot;\theta)$ and update the parameter θ via stochastic semigradient descent (Sutton and Barto, 2018),

$$\theta' = \theta - \epsilon \cdot \bigl(\widehat{Q}(s,a;\theta) - r - \gamma \cdot \widehat{Q}(s',a';\theta)\bigr) \cdot \nabla_\theta \widehat{Q}(s,a;\theta), \tag{2.2.3}$$

where $\epsilon > 0$ is the stepsize and $(s,a,r,s',a') \sim \widetilde{\mathcal{D}}$. Here we denote by $\widetilde{\mathcal{D}} \in \mathscr{P}(\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S} \times \mathcal{A})$ the distribution of (s,a,r,s',a'), where $(s,a) \sim \mathcal{D}$, $r \sim R(\cdot \,|\, s,a)$, $s' \sim P(\cdot \,|\, s,a)$, and $a' \sim \pi(\cdot \,|\, s')$.

2.2.2. Wasserstein Space

Let $\Theta \subseteq \mathbb{R}^D$ be a Polish space. We denote by $\mathscr{P}_2(\Theta) \subseteq \mathscr{P}(\Theta)$ the set of probability measures with finite second moments. Then, the Wasserstein-2 distance between $\mu, \nu \in \mathscr{P}_2(\Theta)$ is defined as follows,

$$\mathcal{W}_2(\mu,\nu) = \inf\Bigl\{\mathbb{E}\bigl[\|X - Y\|^2\bigr]^{1/2} \Bigm| \mathrm{law}(X) = \mu, \mathrm{law}(Y) = \nu\Bigr\}, \tag{2.2.4}$$

where the infimum is taken over the random variables X and Y on Θ. Here we denote by $\mathrm{law}(X)$ the distribution of a random variable X. We call $\mathcal{M} = (\mathscr{P}_2(\Theta), \mathcal{W}_2)$ the Wasserstein space, which is an infinite-dimensional manifold (Villani, 2008). In particular, such a structure allows us to write any tangent vector at $\mu \in \mathcal{M}$ as ρ_0' for a corresponding curve $\rho : [0,1] \to \mathscr{P}_2(\Theta)$ that satisfies $\rho_0 = \mu$. Here ρ_0' denotes $\partial_t \rho_t \,|_{\, t=0}$. Specifically, under certain regularity conditions, for any curve $\rho : [0,1] \to \mathscr{P}_2(\Theta)$, the continuity equation $\partial_t \rho_t = -\operatorname{div}(\rho_t v_t)$ corresponds to a vector field $v : [0,1] \times \Theta \to \mathbb{R}^D$, which endows the infinite-dimensional manifold $\mathscr{P}_2(\Theta)$ with a weak Riemannian structure in the following sense (Villani, 2008). Given any tangent vectors u and $\widetilde{u}$ at $\mu \in \mathcal{M}$ and the corresponding vector fields $v, \widetilde{v}$, which satisfy $u + \operatorname{div}(\mu v) = 0$ and $\widetilde{u} + \operatorname{div}(\mu \widetilde{v}) = 0$, respectively, we define the inner product of u and $\widetilde{u}$ as follows,

$$\langle u, \widetilde{u} \rangle_\mu = \int \langle v, \widetilde{v} \rangle \, \mathrm{d}\mu, \tag{2.2.5}$$

which yields a Riemannian metric. Here $\langle v, \widetilde{v} \rangle$ is the inner product on $\mathbb{R}^D$. Such a Riemannian metric further induces a norm $\|u\|_\mu = \langle u, u \rangle_\mu^{1/2}$ for any tangent vector $u \in T_\mu \mathcal{M}$ at any $\mu \in \mathcal{M}$, which allows us to write the Wasserstein-2 distance defined in (2.2.4) as follows,

$$\mathcal{W}_2(\mu, \nu) = \inf \biggl\{ \biggl(\int_0^1 \|\rho_t'\|_{\rho_t}^2 \, \mathrm{d}t \biggr)^{1/2} \biggm| \rho : [0, 1] \to \mathcal{M}, \rho_0 = \mu, \rho_1 = \nu \biggr\}. \tag{2.2.6}$$

Here ρ_s' denotes $\partial_t \rho_t \,|_{\, t=s}$ for any $s \in [0, 1]$. In particular, the infimum in (2.2.6) is attained by the geodesic $\widetilde{\rho} : [0, 1] \to \mathscr{P}_2(\Theta)$ connecting $\mu, \nu \in \mathcal{M}$. Moreover, the geodesics on $\mathcal{M}$ are constant-speed, that is,

$$\|\widetilde{\rho}_t'\|_{\widetilde{\rho}_t} = \mathcal{W}_2(\mu, \nu), \quad \forall t \in [0, 1]. \tag{2.2.7}$$

2.3. Temporal-Difference Learning

For notational simplicity, we write $\mathbb{R}^d = \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, $\mathcal{X} = \mathcal{S} \times \mathcal{A} \subseteq \mathbb{R}^d$, and $x = (s, a) \in \mathcal{X}$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Parameterization of Q-Function. We consider the parameter space $\mathbb{R}^D$ and parameterize the Q-function with the following two-layer neural network,

$$\widehat{Q}(x;\theta^{(m)}) = \frac{\alpha}{m}\sum_{i=1}^{m}\sigma(x;\theta_i), \tag{2.3.1}$$

where $\theta^{(m)} = (\theta_1,\ldots,\theta_m) \in \mathbb{R}^{D\times m}$ is the parameter, $m \in \mathbb{N}_+$ is the width, $\alpha > 0$ is the scaling parameter, and $\sigma : \mathbb{R}^d \times \mathbb{R}^D \to \mathbb{R}$ is the activation function. Assuming the activation function in (2.3.1) takes the form of $\sigma(x;\theta) = b \cdot \widetilde{\sigma}(x;w)$ for $\theta = (w,b)$, we recover the standard form of two-layer neural networks, where $\widetilde{\sigma}$ is the rectified linear unit or the sigmoid function. Such a parameterization is also used in Chizat and Bach (2018a); Mei et al. (2019); Chen et al. (2020b). For $\{\theta_i\}_{i=1}^m$ independently sampled from a distribution $\rho \in \mathscr{P}(\mathbb{R}^D)$, we have the following infinite-width limit of (2.3.1),

$$Q(x;\rho) = \alpha \cdot \int \sigma(x;\theta)\,\mathrm{d}\rho(\theta). \tag{2.3.2}$$

For the empirical distribution $\widehat{\rho}^{(m)} = m^{-1}\cdot\sum_{i=1}^m \delta_{\theta_i}$ corresponding to $\{\theta_i\}_{i=1}^m$, we have $Q(x;\widehat{\rho}^{(m)}) = \widehat{Q}(x;\theta^{(m)})$.

TD Dynamics. In what follows, we consider the TD dynamics,

$$\theta_i(k+1) = \theta_i(k) - \eta\epsilon\cdot\alpha\cdot\Big(\widehat{Q}\big(x_k;\theta^{(m)}(k)\big) - r_k - \gamma\cdot\widehat{Q}\big(x_k';\theta^{(m)}(k)\big)\Big)\cdot\nabla_\theta\sigma\big(x_k;\theta_i(k)\big), \tag{2.3.3}$$

where $i \in [m]$, $(x_k, r_k, x_k') \sim \widetilde{\mathcal{D}}$, and $\epsilon > 0$ is the stepsize with the scaling parameter $\eta > 0$. Without loss of generality, we assume that (x_k, r_k, x_k') is independently sampled from $\widetilde{\mathcal{D}}$, while our analysis straightforwardly generalizes to the setting of Markov sampling

(Bhandari et al., 2018; Zou et al., 2019; Xu et al., 2019c). For an initial distribution $\rho_0 \in \mathscr{P}(\mathbb{R}^D)$, we initialize $\{\theta_i\}_{i=1}^m$ as $\theta_i \overset{\text{i.i.d.}}{\sim} \rho_0$ $(i \in [m])$. See Algorithm 2 for a detailed description.

Algorithm 2 Temporal-Difference Learning with Two-Layer Neural Network for Policy Evaluation

```
**Initialization:** $\theta_i(0) \overset{\text{i.i.d.}}{\sim} \rho_0$ $(i \in [m])$, number of iterations $K = \lfloor T/\epsilon \rfloor$, and policy $\pi$ of interest.
**for** $k = 0, \ldots, K-1$ **do**
    Sample the state-action pair $(s, a)$ from the stationary distribution $\mathcal{D}$ of $\pi$, receive the reward $r$, and obtain the subsequent state-action pair $(s', a')$.
    Calculate the Bellman residual $\delta = \widehat{Q}(x; \theta^{(m)}(k)) - r - \gamma \cdot \widehat{Q}(x'; \theta^{(m)}(k))$, where $x = (s, a)$ and $x' = (s', a')$.
    Perform the TD update $\theta_i(k+1) \leftarrow \theta_i(k) - \eta\epsilon \cdot \alpha \cdot \delta \cdot \nabla_\theta \sigma(x; \theta_i(k))$ $(i \in [m])$.
**end for**
**Ensure:** $\{\theta^{(m)}(k)\}_{k=0}^{K-1}$
```

Mean-Field Limit. Corresponding to $\epsilon \to 0^+$ and $m \to \infty$, the continuous-time and infinite-width limit of the TD dynamics in Algorithm 2 is characterized by the following partial differential equation (PDE) with ρ_0 as the initial distribution,

$$\partial_t \rho_t = -\eta \cdot \operatorname{div}\bigl(\rho_t \cdot g(\cdot; \rho_t)\bigr). \tag{2.3.4}$$

Here $g(\cdot; \rho_t) : \mathbb{R}^D \to \mathbb{R}^D$ is a vector field, which is defined as follows,

$$g(\theta; \rho) = -\alpha \cdot \mathbb{E}_{(x,r,x') \sim \widetilde{\mathcal{D}}}\Bigl[\bigl(Q(x; \rho) - r - \gamma \cdot Q(x'; \rho)\bigr) \cdot \nabla_\theta \sigma(x; \theta)\Bigr]. \tag{2.3.5}$$

Note that (2.3.4) holds in the sense of distributions (Ambrosio et al., 2008). See Mei et al. (2018, 2019); Araújo et al. (2019) for the existence, uniqueness, and regularity of the PDE solution ρ_t in (2.3.4). In the sequel, we refer to the continuous-time and infinite-width limit with $\epsilon \to 0^+$ and $m \to \infty$ as the mean-field limit. Let $\widehat{\rho}_k^{(m)} = m^{-1} \cdot \sum_{i=1}^m \delta_{\theta_i(k)}$ be the empirical distribution corresponding to $\{\theta_i(k)\}_{i=1}^m$ in (2.3.3). The following proposition proves that the PDE solution ρ_t in (2.3.4) well approximates the TD dynamics $\theta^{(m)}(k)$ in (2.3.3).

Proposition 2.3.1 (Informal Version of Proposition B.2.1)**.** Let the initial distribution ρ_0 be the standard Gaussian distribution $N(0, I_D)$. Under certain regularity conditions, $\widehat{\rho}^{(m)}_{\lfloor t/\epsilon \rfloor}$ weakly converges to ρ_t as $\epsilon \to 0^+$ and $m \to \infty$.

The proof of Proposition 2.3.1 is based on the propagation of chaos (Sznitman, 1991; Mei et al., 2018, 2019). In contrast to Mei et al. (2018, 2019), the PDE in (2.3.4) can not be cast as a gradient flow, since there does not exist a corresponding energy functional. Thus, their analysis is not directly applicable to our setting. We defer the detailed discussion on the approximation analysis to §B.2. Proposition 2.3.1 allows us to convert the TD dynamics over the finite-dimensional parameter space to its counterpart over the infinite-dimensional Wasserstein space, where the infinitely wide neural network $Q(\cdot; \rho)$ in (2.3.2) is linear in the distribution ρ.

Feature Representation. We are interested in the evolution of the feature representation

$$\Big(\nabla_\theta \sigma\big(x; \theta_1(k)\big)^\top, \ldots, \nabla_\theta \sigma\big(x; \theta_m(k)\big)^\top\Big)^\top \in \mathbb{R}^{Dm} \tag{2.3.6}$$

corresponding to $\theta^{(m)}(k) = (\theta_1(k), \ldots, \theta_m(k)) \in \mathbb{R}^{D \times m}$. Such a feature representation is used to analyze the TD dynamics $\theta^{(m)}(k)$ in (2.3.3) in the NTK regime (Cai et al., 2019c), which corresponds to setting $\alpha = \sqrt{m}$ in (2.3.1). Meanwhile, the nonlinear gradient TD dynamics (Bhatnagar et al., 2009) explicitly uses such a feature representation at each iteration to locally linearize the Q-function. Moreover, up to a rescaling, such a feature representation corresponds to the kernel

$$\mathbb{K}(x, x'; \widehat{\rho}_k^{(m)}) = \int \nabla_\theta \sigma(x; \theta)^\top \nabla_\theta \sigma(x'; \theta) \, \mathrm{d}\widehat{\rho}_k^{(m)}(\theta),$$

which by Proposition 2.3.1 further induces the kernel

$$\mathbb{K}(x, x'; \rho_t) = \int \nabla_\theta \sigma(x; \theta)^\top \nabla_\theta \sigma(x'; \theta) \, \mathrm{d}\rho_t(\theta) \tag{2.3.7}$$

at the mean-field limit with $\epsilon \to 0^+$ and $m \to \infty$. Such a correspondence allows us to use the PDE solution ρ_t in (2.3.4) as a proxy for characterizing the evolution of the feature representation in (2.3.6).

2.4. Main Results

We first introduce the assumptions for our analysis. In §2.4.1, we establish the global optimality and convergence of the PDE solution ρ_t in (2.3.4). In §2.4.2, we further invoke Proposition 2.3.1 to establish the global optimality and convergence of the TD dynamics $\theta^{(m)}(k)$ in (2.3.3).

Assumption 2.4.1. We assume that the state-action pair $x = (s, a)$ satisfies $\|x\| \leq 1$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Assumption 2.4.1 can be ensured by normalizing all state-action pairs. Such an assumption is commonly used in the mean-field analysis of neural networks (Chizat and Bach, 2018b; Mei et al., 2018, 2019; Araújo et al., 2019; Fang et al., 2019a,b; Chen et al., 2020b). We remark that our analysis straightforwardly generalizes to the setting where $\|x\| \leq C$ for an absolute constant $C > 0$.

Assumption 2.4.2. We assume that the activation function σ in (2.3.1) satisfies

$$\big|\sigma(x;\theta)\big| \le B_0, \quad \big\|\nabla_\theta \sigma(x;\theta)\big\| \le B_1 \cdot \|x\|, \quad \big\|\nabla^2_{\theta\theta}\sigma(x;\theta)\big\|_{\mathrm{F}} \le B_2 \cdot \|x\|^2 \tag{2.4.1}$$

for any $x \in \mathcal{X}$. Also, we assume that the reward r satisfies $|r| \le B_r$.

Assumption 2.4.2 holds for a broad range of neural networks. For example, let $\theta = (w,b) \in \mathbb{R}^{D-1} \times \mathbb{R}$. The activation function

$$\sigma^\dagger(x;\theta) = B_0 \cdot \tanh(b) \cdot \mathrm{sigmoid}(w^\top x) \tag{2.4.2}$$

satisfies (2.4.1) in Assumption 2.4.2. Moreover, the infinitely wide neural network in (2.3.2) with the activation function $\sigma^\dagger$ in (2.4.2) induces the following function class,

$$\mathcal{F}^\dagger = \bigg\{ \int \beta \cdot \mathrm{sigmoid}(w^\top x)\,\mathrm{d}\mu(w,\beta) \,\bigg|\, \mu \in \mathscr{P}\big(\mathbb{R}^{D-1} \times [-B_0, B_0]\big) \bigg\},$$

where $\beta = B_0 \cdot \tanh(b) \in [-B_0, B_0]$. By the universal approximation theorem (Barron, 1993; Pinkus, 1999), $\mathcal{F}^\dagger$ captures a rich class of functions.

2.4.1. Global Optimality and Convergence of PDE Solution

Throughout the rest of this paper, we consider the following function class,

$$\mathcal{F} = \bigg\{ \int \sigma_0(b) \cdot \sigma_1(x;w)\,\mathrm{d}\rho(w,b) \,\bigg|\, \rho \in \mathscr{P}_2(\mathbb{R}^{D-1} \times \mathbb{R}) \bigg\}, \tag{2.4.3}$$

which is induced by the infinitely wide neural network in (2.3.2) with $\theta = (w, b) \in \mathbb{R}^{D-1} \times \mathbb{R}$ and the following activation function,

$$\sigma(x; \theta) = \sigma_0(b) \cdot \sigma_1(x; w).$$

We assume that σ_0 is an odd function, that is, $\sigma_0(b) = -\sigma_0(-b)$, which implies $\int \sigma(x; \theta) \, \mathrm{d}\rho_0(\theta) = 0$. Note that the set of infinitely wide neural networks taking the forms of (2.3.2) is $\alpha \cdot \mathcal{F}$, which is larger than $\mathcal{F}$ in (2.4.3) by the scaling parameter $\alpha > 0$. Thus, α can be viewed as the degree of "overrepresentation". Without loss of generality, we assume that $\mathcal{F}$ is complete. The following theorem characterizes the global optimality and convergence of the PDE solution ρ_t in (2.3.4).

Theorem 2.4.3. There exists a unique fixed point solution to the projected Bellman equation $Q = \Pi_{\mathcal{F}} \mathcal{T}^\pi Q$, which takes the form of $Q^*(x) = \int \sigma(x; \theta) \, \mathrm{d}\bar{\rho}(\theta)$. Also, Q^* is the global minimizer of the MSPBE defined in (2.2.2). We assume that $D_{\chi^2}(\bar{\rho} \,\|\, \rho_0) < \infty$ and $\bar{\rho}(\theta) > 0$ for any $\theta \in \mathbb{R}^D$. Under Assumptions 2.4.1 and 2.4.2, it holds for $\eta = \alpha^{-2}$ in (2.3.4) that

$$\inf_{t \in [0,T]} \mathbb{E}_{x \sim \mathcal{D}}\Big[\big(Q(x; \rho_t) - Q^*(x)\big)^2\Big] \le \frac{D_{\chi^2}(\bar{\rho} \,\|\, \rho_0)}{2(1-\gamma) \cdot T} + \frac{C_*}{(1-\gamma) \cdot \alpha}, \tag{2.4.4}$$

where $C_* > 0$ is a constant that depends on $D_{\chi^2}(\bar{\rho} \,\|\, \rho_0)$, B_1, B_2, and B_r.

Theorem 2.4.3 proves that the optimality gap $\mathbb{E}_{x \sim \mathcal{D}}[(Q(x; \rho_t) - Q^*(x))^2]$ decays to zero at a sublinear rate up to the error of $O(\alpha^{-1})$, where $\alpha > 0$ is the scaling parameter in (2.3.1). Varying α leads to a tradeoff between such an error of $O(\alpha^{-1})$ and the deviation of ρ_t from ρ_0. Specifically, in §2.5 we prove that ρ_t deviates from ρ_0 by the divergence $D_{\chi^2}(\rho_t \,\|\, \rho_0) \le O(\alpha^{-2})$. Hence, a smaller α allows ρ_t to move further away from ρ_0, inducing a feature representation that is more different from the initial one (Fang et al., 2019a,b). See (2.3.6)-(2.3.7) for the correspondence of ρ_t with the feature representation and the kernel that it induces. On the other hand, a smaller α yields a larger error of $O(\alpha^{-1})$ in (2.4.4) of Theorem 2.4.3. In contrast, the NTK regime (Cai et al., 2019c), which corresponds to setting $\alpha = \sqrt{m}$ in (2.3.1), only allows ρ_t to deviate from ρ_0 by the divergence $D_{\chi^2}(\rho_t \,\|\, \rho_0) \leq O(m^{-1}) = o(1)$. In other words, the NTK regime fails to induce a feature representation that is significantly different from the initial one. In summary, our analysis goes beyond the NTK regime, which allows us to characterize the evolution of the feature representation towards the (near-)optimal one.

2.4.2. Global Optimality and Convergence of TD Dynamics

As a result of Proposition 2.3.1, we establish the following lemma, which characterizes the error of approximating the optimality gap in Theorem 2.4.3 by that of the TD dynamics $\theta^{(m)}(k)$ in (2.3.3).

Lemma 2.4.4. Let B be a constant that depends on α, η, γ, B_0, B_1, and B_2. Under Assumptions 2.4.1 and 2.4.2, it holds for any $k \leq T/\epsilon$ ($k \in \mathbb{N}$) that

$$\begin{aligned} &\mathbb{E}_{x\sim\mathcal{D}}\Big[\Big(\widehat{Q}\big(x;\theta^{(m)}(k)\big) - Q^*(x)\Big)^2\Big] \\ &\qquad \leq \mathbb{E}_{x\sim\mathcal{D}}\Big[\big(Q(x;\rho_{k\epsilon}) - Q^*(x)\big)^2\Big] + B\cdot e^{BT}\cdot\Big(\sqrt{m^{-1}\cdot\log(m/\delta)} + \sqrt{\epsilon\cdot\big(D+\log(m/\delta)\big)}\Big) \end{aligned}$$

with probability at least $1-\delta$.

Proof. See §B.2.2 for a detailed proof. □

Based on Theorem 2.4.3 and Lemma 2.4.4, we establish the following corollary, which characterizes the global optimality and convergence of the TD dynamics $\theta^{(m)}(k)$ in (2.3.3).

Corollary 2.4.5. Under the same conditions of Theorem 2.4.3, it holds with probability at least $1-\delta$ that

$$\min_{\substack{k\leq T/\epsilon \\ (k\in\mathbb{N})}} \mathbb{E}_{x\sim\mathcal{D}}\Big[\Big(\widehat{Q}\big(x;\theta^{(m)}(k)\big)-Q^*(x)\Big)^2\Big] \leq \frac{D_{\chi^2}(\bar{\rho}\,\|\,\rho_0)}{2(1-\gamma)\cdot T}+\frac{C_*}{(1-\gamma)\cdot\alpha}+\Delta(\epsilon,m,\delta,T), \tag{2.4.5}$$

where $C_*>0$ is the constant in (2.4.4) of Theorem 2.4.3 and $\Delta(\epsilon,m,\delta,T)>0$ is an error term such that

$$\lim_{m\to\infty}\lim_{\epsilon\to 0^+}\Delta(\epsilon,m,\delta,T)=0.$$

Proof. Combining Theorem 2.4.3 and Lemma 2.4.4 implies Corollary 2.4.5. □

In (2.4.5) of Corollary 2.4.5, the error term $\Delta(\epsilon,m,\delta,T)$ characterizes the error of approximating the TD dynamics $\theta^{(m)}(k)$ in (2.3.3) using the PDE solution ρ_t in (2.3.4). In particular, such an error vanishes at the mean-field limit.

2.5. Proof of Main Results

We first introduce two technical lemmas. Recall that $\mathcal{F}$ is defined in (2.4.3), $Q(x;\rho)$ is defined in (2.3.2), and $g(\theta;\rho)$ is defined in (2.3.5).

Lemma 2.5.1. There exists a unique fixed point solution to the projected Bellman equation $Q=\Pi_{\mathcal{F}}\mathcal{T}^\pi Q$, which takes the form of $Q^*(x)=\int\sigma(x;\theta)\,\mathrm{d}\bar{\rho}(\theta)$. Also, there exists $\rho^*\in\mathscr{P}_2(\mathbb{R}^D)$ that satisfies the following properties,

(i) $Q(x;\rho^*)=Q^*(x)$ for any $x\in\mathcal{X}$,

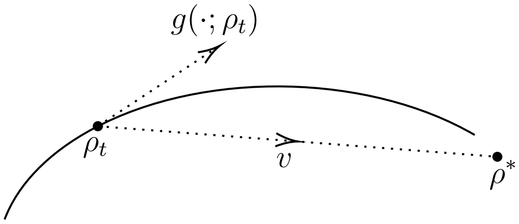


Figure 2.1. We illustrate the first variation formula $\frac{\mathrm{d}\mathcal{W}_2(\rho_t,\rho^*)^2}{2} = -\langle g(\cdot;\rho_t), v\rangle_{\rho_t}$, where v is the vector field corresponding to the geodesic that connects ρ_t and ρ^*. See Lemma B.3.2 for details.

(ii) $g(\cdot;\rho^*) = 0$ for $\bar{\rho}$-a.e., and

(iii) $\mathcal{W}_2(\rho^*, \rho_0) \le \alpha^{-1} \cdot \bar{D}$, where $\bar{D} = D_{\chi^2}(\bar{\rho}\,\|\,\rho_0)^{1/2}$.

Proof. See §B.1.1 for a detailed proof. □

Lemma 2.5.1 establishes the existence of the fixed point solution Q^* to the projected Bellman equation $Q = \Pi_{\mathcal{F}}\mathcal{T}^\pi Q$. Furthermore, such a fixed point solution Q^* can be parameterized with the infinitely wide neural network $Q(\cdot;\rho^*)$ in (2.3.2). Meanwhile, the Wasserstein-2 distance between ρ^* and the initial distribution ρ_0 is upper bounded by $O(\alpha^{-1})$. Based on the existence of Q^* and the property of ρ^* in Lemma 2.5.1, we establish the following lemma that characterizes the evolution of $\mathcal{W}_2(\rho_t, \rho^*)$, where ρ_t is the PDE solution in (2.3.4).

Lemma 2.5.2. We assume that $\mathcal{W}_2(\rho_t, \rho^*) \le 2\mathcal{W}_2(\rho_0, \rho^*)$, $D_{\chi^2}(\bar{\rho}\,\|\,\rho_0) < \infty$, and $\bar{\rho}(\theta) > 0$ for any $\theta \in \mathbb{R}^D$. Under Assumptions 2.4.1 and 2.4.2, it holds that

$$(2.5.1)\qquad \frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\rho_t,\rho^*)^2}{2} \le -(1-\gamma)\cdot\eta\cdot\mathbb{E}_{x\sim\mathcal{D}}\Big[\big(Q(x;\rho_t) - Q^*(x)\big)^2\Big] + C_*\cdot\alpha^{-1}\cdot\eta,$$

where $C_* > 0$ is a constant depending on $D_{\chi^2}(\bar{\rho}\,\|\,\rho_0)$, B_1, B_2, and B_r.

Proof. See §B.1.2 for a detailed proof. □

The proof of Lemma 2.5.2 is based on the first variation formula of the Wasserstein-2 distance (Lemma B.3.2), which is illustrated in Figure 2.1, and the strongly monotonicity of $g(\cdot;\beta_t)$ along a curve β on the Wasserstein space (Lemma B.1.1). When the right-hand side of (2.5.1) is nonpositive, Lemma 2.5.2 characterizes the decay of $\mathcal{W}_2(\rho_t,\rho^*)$. We are now ready to present the proof of Theorem 2.4.3.

Proof. We use a continuous counterpart of the induction argument. We define

$$t^* = \inf\biggl\{\tau \in \mathbb{R}_+ \biggm| \mathbb{E}_{x\sim\mathcal{D}}\Bigl[(1-\gamma)\cdot\bigl(Q(x;\rho_\tau)-Q^*(x)\bigr)^2\Bigr] < C_*\cdot\alpha^{-1}\biggr\}. \tag{2.5.2}$$

In other words, the right-hand side of (2.5.1) in Lemma 2.5.2 is nonpositive for any $t \le t^*$, that is,

$$-(1-\gamma)\cdot\mathbb{E}_{x\sim\mathcal{D}}\Bigl[\bigl(Q(x;\rho_t)-Q^*(x)\bigr)^2\Bigr] + C_*\cdot\alpha^{-1} \le 0. \tag{2.5.3}$$

Also, we define

$$t_* = \inf\bigl\{\tau \in \mathbb{R}_+ \bigm| \mathcal{W}_2(\rho_\tau,\rho^*) > 2\mathcal{W}_2(\rho_0,\rho^*)\bigr\}. \tag{2.5.4}$$

In other words, (2.5.1) of Lemma 2.5.2 holds for any $t \le t_*$. Thus, for any $0 \le t \le \min\{t^*, t_*\}$, it holds that $\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\rho_t,\rho_*)^2}{2} \le 0$. Figure 2.2 illustrates the definition of t^* and t_* in (2.5.2) and (2.5.4), respectively.

We now prove that $t_* \ge t^*$ by contradiction. By the continuity of $\mathcal{W}_2(\rho_t,\rho^*)^2$ with respect to t (Ambrosio et al., 2008), it holds that $t_* > 0$, since $\mathcal{W}_2(\rho_0,\rho^*) < 2\mathcal{W}_2(\rho_0,\rho^*)$. For the sake of contradiction, we assume that $t_* < t^*$, by (2.5.1) of Lemma 2.5.2 and

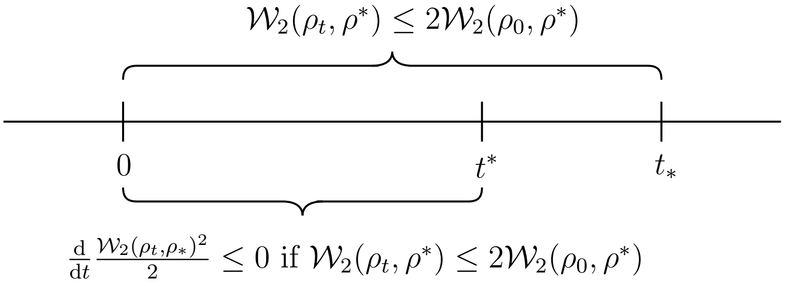

$\mathcal{W}_2(\rho_t, \rho^*) \le 2\mathcal{W}_2(\rho_0, \rho^*)$
0
t^*
t_*
$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\rho_t,\rho_*)^2}{2} \le 0$ if $\mathcal{W}_2(\rho_t, \rho^*) \le 2\mathcal{W}_2(\rho_0, \rho^*)$

Figure 2.2. For any $0 \le t \le \min\{t^*, t_*\}$, (2.5.1) of Lemma 2.5.2 holds and $\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\rho_t,\rho_*)^2}{2} \le 0$.

(2.5.3), it holds for any $0 \le t \le t_*$ that

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\rho_t, \rho^*)^2}{2} \le 0,$$

which implies that $\mathcal{W}_2(\rho_t, \rho^*) \le \mathcal{W}_2(\rho_0, \rho^*)$ for any $0 \le t \le t_*$. This contradicts the definition of t_* in (2.5.4). Thus, it holds that $t_* \ge t^*$, which implies that (2.5.1) of Lemma 2.5.2 holds for any $0 \le t \le t^*$.

If $t^* \le T$, (2.5.3) implies Theorem 2.4.3. If $t^* > T$, by (2.5.1) of Lemma 2.5.2, it holds for any $0 \le t \le T$ that

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\rho_t, \rho^*)^2}{2} \le -(1-\gamma) \cdot \eta \cdot \mathbb{E}_{x\sim\mathcal{D}}\Big[\big(Q(x;\rho_t) - Q^*(x)\big)^2\Big] + C_* \cdot \alpha^{-1} \cdot \eta \le 0,$$

which further implies that

$$\mathbb{E}_{x\sim\mathcal{D}}\Big[\big(Q(x;\rho_t) - Q^*(x)\big)^2\Big] \le -(1-\gamma)^{-1} \cdot \eta^{-1} \cdot \frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\rho_t, \rho^*)^2}{2} + C_* \cdot (1-\gamma)^{-1} \cdot \alpha^{-1}. \tag{2.5.5}$$

Upon telescoping (2.5.5) and setting $\eta = \alpha^{-2}$, we obtain that

$$\begin{aligned}
&\inf_{t\in[0,T]} \mathbb{E}_{\mathcal{D}}\Bigl[\bigl(Q(x;\rho_t) - Q^*(x)\bigr)^2\Bigr] \\
&\quad\le T^{-1} \cdot \int_0^T \mathbb{E}_{x\sim\mathcal{D}}\Bigl[\bigl(Q(x;\rho_t) - Q^*(x)\bigr)^2\Bigr]\,\mathrm{d}t \\
&\quad\le 1/2 \cdot (1-\gamma)^{-1} \cdot \eta^{-1} \cdot T^{-1} \cdot \mathcal{W}_2(\rho_0, \rho^*)^2 + C_* \cdot (1-\gamma)^{-1} \cdot \alpha^{-1} \\
&\quad\le 1/2 \cdot (1-\gamma)^{-1} \cdot \bar{D}^2 \cdot T^{-1} + C_* \cdot (1-\gamma)^{-1} \cdot \alpha^{-1},
\end{aligned}$$

where the last inequality follows from the fact that $\eta = \alpha^{-2}$ and (iii) of Lemma 2.5.1. Thus, we complete the proof of Theorem 2.4.3. □

2.6. Extension to Q-Learning and Policy Gradient

In this section, we extend our analysis of TD to Q-learning and policy gradient. In §2.6.1, we introduce Q-learning and its mean-field limit. In §2.6.2, we establish the global optimality and convergence of Q-learning. In §2.6.3, we further extend our analysis to soft Q-learning, which is equivalent to policy gradient.

2.6.1. Q-Learning

Q-learning aims to solve the following projected Bellman optimality equation,

$$Q = \Pi_{\mathcal{F}}\mathcal{T}^* Q. \tag{2.6.1}$$

Here $\mathcal{T}^*$ is the Bellman optimality operator, which is defined as follows,

$$\mathcal{T}^* Q(s,a) = \mathbb{E}\bigl[r + \gamma \cdot \max_{\underline{a}\in\mathcal{A}} Q(s', \underline{a}) \,\big|\, r \sim R(\cdot\,|\,s,a), s' \sim P(\cdot\,|\,s,a)\bigr].$$

When $\Pi_{\mathcal{F}}$ is the identity mapping, the fixed point solution to (2.6.1) is the Q-function Q^{π^*} of the optimal policy π^*, which maximizes the expected total reward $J(\pi)$ defined in (2.2.1) (Sutton and Barto, 2018). We consider the parameterization of the Q-function in (2.3.1) and update the parameter $\theta^{(m)}$ as follows,

$$\begin{aligned} &(2.6.2)\\ &\theta_i(k+1)\\ &\quad = \theta_i(k) - \eta\epsilon \cdot \alpha \cdot \Big(\widehat{Q}\big(s_k, a_k; \theta^{(m)}(k)\big) - r_k - \gamma \cdot \max_{\underline{a}\in\mathcal{A}} \widehat{Q}\big(s_k', \underline{a}; \theta^{(m)}(k)\big)\Big) \cdot \nabla_\theta \sigma\big(s_k, a_k; \theta_i(k)\big), \end{aligned}$$

where $i \in [m]$, (s_k, a_k) is sampled from the stationary distribution $\mathcal{D}_{\mathrm{E}} \in \mathscr{P}(\mathcal{S} \times \mathcal{A})$ of an exploration policy π_{E}, $r_k \sim R(\cdot \,|\, s_k, a_k)$ is the reward, and $s_k' \sim P(\cdot \,|\, s_k, a_k)$ is the subsequent state. For notational simplicity, we denote by $\widetilde{\mathcal{D}}_{\mathrm{E}} \in \mathscr{P}(\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S})$ the distribution of (s_k, a_k, r_k, s_k'). For an initial distribution $\nu_0 \in \mathscr{P}(\mathbb{R}^D)$, we initialize $\{\theta_i\}_{i=1}^m$ as $\theta_i \overset{\text{i.i.d.}}{\sim} \rho_0$ $(i \in [m])$. See Algorithm 3 for a detailed description.

Algorithm 3 Q-Learning with Two-Layer Neural Network for Policy Improvement

```
**Initialization.** $\theta_i(0) \overset{\text{i.i.d.}}{\sim} \nu_0$ $(i \in [m])$, number of iterations $K = \lfloor T/\epsilon \rfloor$, and exploration policy $\pi_{\mathrm{E}}$.
**for** $k = 0, \ldots, K-1$ **do**
    Sample the state-action pair $(s, a)$ from the stationary distribution $\mathcal{D}_{\mathrm{E}}$ of $\pi_{\mathrm{E}}$, receive the reward $r$, and obtain the subsequent state $s'$.
    Calculate the Bellman residual $\delta = \widehat{Q}(x; \theta^{(m)}(k)) - r - \gamma \cdot \widehat{Q}(x'; \theta^{(m)}(k))$, where $x = (s, a)$ and $x' = (s', \mathrm{argmax}_{\underline{a}\in\mathcal{A}} \widehat{Q}(s', \underline{a}; \theta^{(m)}(k)))$.
    Perform the Q-learning update $\theta_i(k+1) \leftarrow \theta_i(k) - \eta\epsilon \cdot \alpha \cdot \delta \cdot \nabla_\theta \sigma(x; \theta_i(k))$ $(i \in [m])$.
**end for**
**Ensure:** $\{\theta^{(m)}(k)\}_{k=0}^{K-1}$
```

Mean-Field Limit. Corresponding to $\epsilon \to 0^+$ and $m \to \infty$, the mean-field limit of the Q-learning dynamics in Algorithm 3 is characterized by the following PDE with ν_0 as the initial distribution,

$$\partial_t \nu_t = -\eta \cdot \operatorname{div}\big(\nu_t \cdot h(\cdot;\nu_t)\big). \tag{2.6.3}$$

Here $h(\cdot;\nu_t) : \mathbb{R}^D \to \mathbb{R}^D$ is a vector field, which is defined as follows,

$$h(\theta;\nu) = -\alpha \cdot \mathbb{E}_{(s,a,r,s')\sim\widetilde{\mathcal{D}}_{\rm E}}\Big[\big(Q(s,a;\nu) - r - \gamma \cdot \max_{\underline{a}\in\mathcal{A}} Q(s',\underline{a};\nu)\big) \cdot \nabla_\theta \sigma(s,a;\theta)\Big]. \tag{2.6.4}$$

In parallel to Proposition 2.3.1, the empirical distribution $\widehat{\nu}_k^{(m)} = m^{-1} \cdot \sum_{i=1}^m \delta_{\theta_i(k)}$ weakly converges to $\nu_{k\epsilon}$ as $\epsilon \to 0^+$ and $m \to \infty$.

2.6.2. Global Optimality and Convergence of Q-Learning

The max operator in the Bellman optimality operator $\mathcal{T}^*$ makes the analysis of Q-learning more challenging than that of TD. Correspondingly, we lay out an extra regularity condition on the exploration policy $\pi_{\rm E}$. Recall that the function class $\mathcal{F}$ is defined in (2.4.3).

Assumption 2.6.1. We assume for an absolute constant $\kappa > 0$ and any $Q^1, Q^2 \in \mathcal{F}$ that

$$\mathbb{E}_{(s,a)\sim\mathcal{D}_{\rm E}}\Big[\big(Q^1(s,a) - Q^2(s,a)\big)^2\Big] \geq (\gamma+\kappa)^2 \cdot \mathbb{E}_{(s,a)\sim\mathcal{D}_{\rm E}}\Big[\big(\max_{\underline{a}\in\mathcal{A}} Q^1(s,\underline{a}) - \max_{\underline{a}\in\mathcal{A}} Q^2(s,\underline{a})\big)^2\Big].$$

Although Assumption 2.6.1 is strong, we are not aware of any weaker regularity condition in the literature, even in the linear setting (Melo et al., 2008; Zou et al., 2019; Chen et al., 2019b) and the NTK regime (Cai et al., 2019c). Let the initial distribution ν_0 be the standard Gaussian distribution $N(0, I_D)$. In parallel to Theorem 2.4.3, we establish the following theorem, which characterizes the global optimality and convergence of Q-learning. Recall that we write $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ and $x = (s, a) \in \mathcal{X}$. Also, ν_t is the PDE solution in (2.6.3), while $\theta^{(m)}(k)$ is the Q-learning dynamics in (2.6.2).

Theorem 2.6.2. There exists a unique fixed point solution to the projected Bellman optimality equation $Q = \Pi_{\mathcal{F}}\mathcal{T}^*Q$, which takes the form of $Q^\dagger(x) = \int \sigma(x;\theta)\,\mathrm{d}\bar{\nu}(\theta)$. We assume that $D_{\chi^2}(\bar{\nu}\,\|\,\nu_0) < \infty$ and $\bar{\nu}(\theta) > 0$ for any $\theta \in \mathbb{R}^D$. Under Assumptions 2.4.1, 2.4.2, and 2.6.1, it holds for $\eta = \alpha^{-2}$ that

$$\inf_{t\in[0,T]} \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{E}}}\Big[\big(Q(x;\nu_t) - Q^\dagger(x)\big)^2\Big] \le \frac{(\kappa+\gamma)\cdot D_{\chi^2}(\bar{\nu}\,\|\,\nu_0)}{2\kappa\cdot T} + \frac{(\kappa+\gamma)\cdot C_*}{\kappa\cdot\alpha}, \tag{2.6.5}$$

where $C_* > 0$ is a constant depending on $D_{\chi^2}(\bar{\nu}\,\|\,\nu_0)$, B_1, B_2, and B_r. Moreover, it holds with probability at least $1-\delta$ that

$$\begin{aligned}&\min_{\substack{k\le T/\epsilon\\(k\in\mathbb{N})}} \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{E}}}\bigg[\Big(\widehat{Q}\big(x;\theta^{(m)}(k)\big) - Q^\dagger(x)\Big)^2\bigg]\\ &\qquad\le \frac{(\kappa+\gamma)\cdot D_{\chi^2}(\bar{\nu}\,\|\,\nu_0)}{2\kappa\cdot T} + \frac{(\kappa+\gamma)\cdot C_*}{\kappa\cdot\alpha} + \Delta(\epsilon, m, \delta, T),\end{aligned} \tag{2.6.6}$$

where $\Delta(\epsilon, m, \delta, T) > 0$ is an error term such that

$$\lim_{m\to\infty}\lim_{\epsilon\to 0^+} \Delta(\epsilon, m, \delta, T) = 0.$$

Proof. See §B.1.3 for a detailed proof. □

Theorem 2.6.2 proves that the optimality gap $\mathbb{E}_{x\sim\mathcal{D}_{\mathrm{E}}}[(Q(x;\nu_t) - Q^\dagger(x))^2]$ decays to zero at a sublinear rate up to the error of $O(\alpha^{-1})$, where $\alpha > 0$ is the scaling parameter in (2.3.1). In parallel to Theorem 2.4.3, varying α leads to a tradeoff between such an error of $O(\alpha^{-1})$ and the deviation of ν_t from ν_0. Moreover, based on the counterparts of Proposition 2.3.1 and Lemma 2.4.4, Theorem 2.6.2 gives the global optimality and convergence of the Q-learning dynamics $\theta^{(m)}(k)$ in (2.6.2), which is in parallel to Corollary 2.4.5.

2.6.3. Soft Q-Learning and Policy Gradient

Theorem 2.6.2 straightforwardly generalizes to soft Q-learning, where the max operator is replaced by the softmax operator. Specifically, we define the soft Bellman optimality operator as follows,

$$\mathcal{T}_\beta Q(s,a) = \mathbb{E}\big[r + \gamma \cdot \texttt{softmax}_{\underline{a}\in\mathcal{A}}{}^{\beta} Q(s', \underline{a}) \,\big|\, r \sim R(\cdot \,|\, s, a), s' \sim P(\cdot \,|\, s, a)\big],$$

where the softmax operator is defined as follows,

$$\texttt{softmax}_{\underline{a}\in\mathcal{A}}{}^{\beta} Q(s, \underline{a}) = \beta \cdot \log \mathbb{E}_{\underline{a}\sim\bar{\pi}(\cdot \,|\, s)}\Big[\exp\big(\beta^{-1} \cdot Q(s, \underline{a})\big)\Big].$$

Here $\bar{\pi}(\cdot \,|\, s)$ is the uniform policy. Soft Q-learning aims to find the fixed point solution to the projected soft Bellman optimality equation $Q = \Pi_{\mathcal{F}}\mathcal{T}_\beta Q$. In parallel to the Q-learning dynamics in (2.6.2), we consider the following soft Q-learning dynamics,

$$(2.6.7)$$

$$\theta_i(k+1)$$

$$= \theta_i(k) - \eta\epsilon \cdot \alpha \cdot \Big(\widehat{Q}\big(s_k, a_k; \theta^{(m)}(k)\big) - r_k - \gamma \cdot \texttt{softmax}_{\underline{a}\in\mathcal{A}}{}^{\beta} \widehat{Q}\big(s'_k, \underline{a}; \theta^{(m)}(k)\big)\Big) \cdot \nabla_\theta \sigma\big(s_k, a_k; \theta_i(k)\big),$$

whose mean-field limit is characterized by the following PDE,

$$\partial_t \nu_t = -\eta \cdot \mathrm{div}\big(\nu_t \cdot h(\cdot\,;\nu_t)\big). \tag{2.6.8}$$

In parallel to (2.6.4), $h(\cdot\,;\nu_t) : \mathbb{R}^D \to \mathbb{R}^D$ is a vector field, which is defined as follows,

$$h(\theta;\nu) = -\alpha \cdot \mathbb{E}_{(s,a,r,s')\sim\widetilde{\mathcal{D}}_{\mathrm{E}}}\Big[\big(Q(s,a;\nu) - r - \gamma \cdot \mathtt{softmax}_{\underline{a}\in\mathcal{A}}{}^{\beta} Q(s',\underline{a};\nu)\big) \cdot \nabla_\theta \sigma(s,a;\theta)\Big].$$

In parallel to Assumption 2.6.1, we lay out the following regularity condition.

Assumption 2.6.3. We assume for an absolute constant $\kappa > 0$ and any $\nu^1, \nu^2 \in \mathscr{P}(\mathbb{R}^D)$ that

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\mathcal{D}_{\mathrm{E}}}\Big[\big(Q(s,a;\nu^1) - Q(s,a;\nu^2)\big)^2\Big] \\
&\qquad \geq (\gamma+\kappa)^2 \cdot \mathbb{E}_{(s,a)\sim\mathcal{D}_{\mathrm{E}}}\Big[\big(\mathtt{softmax}_{\underline{a}\in\mathcal{A}}{}^{\beta} Q(s,\underline{a};\nu^1) - \mathtt{softmax}_{\underline{a}\in\mathcal{A}}{}^{\beta} Q(s,\underline{a};\nu^2)\big)^2\Big].
\end{aligned}$$

The following proposition parallels Theorem 2.6.2, which characterizes the global optimality and convergence of soft Q-learning. Recall that ν_t is the PDE solution in (2.6.8) and $\theta^{(m)}(k)$ is the soft Q-learning dynamics in (2.6.7).

Proposition 2.6.4. There exists a unique fixed point solution to the projected soft Bellman optimality equation $Q = \Pi_{\mathcal{F}}\mathcal{T}_\beta Q$, which takes the form of $Q^\ddagger(x) = \int \sigma(x;\theta)\,\mathrm{d}\underline{\nu}(\theta)$. We assume that $D_{\chi^2}(\underline{\nu}\,\|\,\nu_0) < \infty$ and $\underline{\nu}(\theta) > 0$ for any $\theta \in \mathbb{R}^D$. Under Assumptions 2.4.1, 2.4.2, and 2.6.3, it holds for $\eta = \alpha^{-2}$ that

$$\inf_{t\in[0,T]} \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{E}}}\Big[\big(Q(x;\nu_t) - Q^\ddagger(x)\big)^2\Big] \leq \frac{(\kappa+\gamma)\cdot D_{\chi^2}(\underline{\nu}\,\|\,\nu_0)}{2\kappa\cdot T} + \frac{(\kappa+\gamma)\cdot C_*}{\kappa\cdot\alpha},$$

where $C_* > 0$ is a constant depending on $D_{\chi^2}(\underline{\nu} \,\|\, \nu_0)$, B_1, B_2, and B_r. Moreover, it holds with probability at least $1 - \delta$ that

$$\min_{\substack{k \le T/\epsilon \\ (k \in \mathbb{N})}} \mathbb{E}_{x \sim \mathcal{D}_{\mathrm{E}}} \Big[\Big(\widehat{Q}\big(x; \theta^{(m)}(k)\big) - Q^{\ddagger}(x) \Big)^2 \Big] \le \frac{(\kappa + \gamma) \cdot D_{\chi^2}(\underline{\nu} \,\|\, \nu_0)}{2\kappa \cdot T} + \frac{(\kappa + \gamma) \cdot C_*}{\kappa \cdot \alpha} + \Delta(\epsilon, m, \delta, T),$$

where $\Delta(\epsilon, m, \delta, T) > 0$ is an error term such that

$$\lim_{m \to \infty} \lim_{\epsilon \to 0^+} \Delta(\epsilon, m, \delta, T) = 0.$$

Proof. Replacing the max operator by the softmax operator in the proof of Theorem 2.6.2 implies Proposition 2.6.4. □

Moreover, soft Q-learning is equivalent to a variant of policy gradient (O'Donoghue et al., 2016; Schulman et al., 2017; Nachum et al., 2017; Haarnoja et al., 2017). Hence, Proposition 2.6.4 also characterizes the global optimality and convergence of such a variant of policy gradient.

CHAPTER 3

An Analysis of Attention via the Lens of Exchangeability and Latent Variable Models

With the attention mechanism, transformers achieve significant empirical successes in natural language processing and computer vision. Despite the intuitive understanding that transformers perform relational inference (or "inductive reasoning") over long sequences to produce desirable representations, we lack a rigorous theory on how the attention mechanism achieves it. In particular, several intriguing questions remain open: (a) What makes a desirable representation? (b) How does the attention mechanism infer the desirable representation within the forward pass? (c) How does a pretraining procedure learn to infer the desirable representation through the backward pass?

We aim to answer the three questions via the lens of exchangeability. Specifically, we observe that, as is the case in BERT and ViT, input tokens are often exchangeable since they already include positional encodings. The notion of exchangeability induces a latent variable model that is invariant to input sizes, which enables our theoretical analysis.

- To answer (a) on representation, we establish the existence of a sufficient and minimal representation of input tokens. In particular, such a representation instantiates the posterior distribution of the latent variable (or "concept") given input tokens, which plays a central role in predicting output labels and solving downstream tasks.

- To answer (b) on inference, we prove that attention with the desired parameter infers the latent posterior up to an approximation error, which is decreasing in input sizes. In detail, we quantify how attention approximates the conditional mean of the value given the key, which characterizes how it performs relational inference over long sequences.

- To answer (c) on learning, we prove that both supervised and self-supervised objectives allow empirical risk minimization to learn the desired parameter up to a generalization error, which is independent of input sizes. Particularly, in the self-supervised setting, we identify a condition number that is pivotal to solving downstream tasks.

Our theoretical analysis gives a complete characterization of the attention mechanism as a "greybox" design, which unifies the handcrafted architecture induced by the latent variable model ("whitebox") and the learnable parameter estimated from data ("blackbox") with provable approximation, generalization, and optimization guarantees.

3.1. Introduction

Transformers are the state-of-the-art architecture for a variety of tasks in natural language processing (Vaswani et al., 2017), computer vision (Dosovitskiy et al., 2020), and multimodal generation (Ramesh et al., 2021). At the core of their significant empirical successes is the attention mechanism, which is defined by a computation graph for the forward pass. In particular, the computation graph performs a specific form of message passing across input tokens (Bronstein et al., 2021). It is commonly believed that the attention mechanism is capable of handling long sequences and performing relational inference (or "inductive reasoning"), which appears to be the key advantage of transformers. However, the intuitive understanding lacks a quantitative justification, which leaves many intriguing questions open:

(a) What makes a desirable representation? Ideally, the desirable representation of input tokens is sufficient and minimal in the sense that it preserves all relevant information for predicting output labels or solving downstream tasks (sufficiency) while it neglects all irreverent information (minimality). However, we lack a quantitative definition of sufficiency and minimality, which requires a probabilistic model.

(b) How does the attention mechanism infer the desirable representation within the forward pass? Intuitively, the attention mechanism is defined by a computation graph that resembles kernel smoothing or kernel regression (Shawe-Taylor et al., 2004) for predicting the value given the key. However, we lack a formal characterization of what function class the attention mechanism parameterizes or approximates. Also, it remains unclear why the specific form of message passing produces the desirable representation of input tokens, that is, one with sufficiency and minimality.

(c) How does a pretraining procedure learn the desirable representation through the backward pass? Empirically, the pretraining procedure that minimizes empirical risks for predicting masked tokens (Devlin et al., 2018; Dosovitskiy et al., 2020; He et al., 2022) appears to succeed in the presence of long sequences. However, we lack a theoretical justification of whether the pretraining procedure with the masked objective attains a desirable estimator that generalizes and why the generalization error does not appear to degrade for long sequences. In particular, it remains unclear to what degree the estimated representation facilitates solving downstream tasks.

In this paper, we answer the three questions via the lens of exchangeability. The key observation is that, as is the case in BERT (Devlin et al., 2018) and ViT (Dosovitskiy et al., 2020), input tokens are exchangeable since they include positional encodings. In other words, the joint distribution of input tokens, e.g., vector embeddings of words in a paragraph or patches in an image with positional encodings, remains the same upon permuting their orders. Meanwhile, the attention mechanism and entrywise feedforward neural networks preserve the notion of exchangeability throughout all transformer layers. By the de Finetti Theorem (de Finetti, 1937), the notion of exchangeability induces a latent variable model that is invariant to input sizes. Unlike classical Bayesian settings, where the latent variable model is defined across many data points, ours is defined over input tokens within one data point (in an "in-context" manner), which captures a fine-grained structure of interactions as relational inductive biases (Battaglia et al., 2018). The latent variable model enables our theoretical analysis, which is summarized in the following:

- To answer (a) on representation, we establish the existence of a sufficient and minimal representation of input tokens based on the latent variable model, which is induced by exchangeability. In particular, we leverage the latent variable model to define sufficiency and minimality following the factorization theorem and the sufficiency principle (Fisher, 1922). Moreover, we prove that the posterior distribution of the latent variable given input tokens is a sufficient and minimal representation, which plays a central role in predicting output labels and solving downstream tasks. Intuitively, the latent variable instantiates the "concept" of a paragraph or an image, which is "summarized" over words or patches.

In detail, the “summarization” process is formalized by the mapping from input tokens to the posterior distribution of the latent variable, that is, inferring the “concept” in a Bayesian manner within the forward pass.

Given the answer to (a), which defines the desirable representation as the latent posterior, it remains unclear how to parameterize or approximate the latent posterior, which is addressed by the answer to (b).

- To answer (b) on inference, we prove that the attention mechanism with the desired parameter infers the latent posterior up to an approximation error, which is decreasing in input sizes. In particular, we prove that a specific parameterization of the latent posterior yields a variant of the attention mechanism based on kernel conditional mean embedding (CME), namely the CME attention, which infers the conditional mean of the value given the key. Here the value and the key (or the query) are obtained from a parameterized transformation of input tokens, where the unknown parameter requires learning.

Although the CME attention recovers the latent posterior for any input sizes, it differs from the commonly used softmax attention by a normalization matrix. To this end, we prove that the CME attention and the softmax attention are equivalent at the infinite limit of input sizes by drawing a connection to nonparametric conditional density estimation. In other words, the softmax attention recovers the latent posterior up to an approximation error that is decreasing in input sizes, which characterizes how it performs relational inference over long sequences. As byproducts, we justify the necessity of multiple attention heads in transformers and provide a causal interpretation of the inferred representation through instrumental variables.

Given the answer to (b), which quantifies the approximation error for the latent posterior, it remains unclear how to learn the desired parameter of the attention mechanism, which is addressed by the answer to (c).

- To answer (c) on learning, we prove that both supervised and self-supervised objectives allow empirical risk minimization to learn the desired parameter up to a generalization error, which is independent of input sizes. In particular, through maximum likelihood estimation, we establish the connection between the latent posterior and the masked objective, which is defined by the empirical risk for predicting masked tokens.

Moreover, we prove that the global minimizer of the masked objective attains a generalization error that is independent of input sizes, which justifies why transformers allow long sequences. Our proof exploits the invariance and equivariance of the attention mechanism and entrywise feedforward neural networks, which deviates from most existing analyses of the generalization error. Particularly, in the self-supervised setting, e.g., as in MAE (He et al., 2022), we identify a condition number that is pivotal to solving downstream tasks. Intuitively, the condition number quantifies the amount of information that is transferred from the pretraining task to a new task.

Meanwhile, in the overparameterized regime, we prove that any stationary point of the masked objective is almost globally optimal when the attention mechanism and entrywise feedforward neural networks have sufficient expressive power. As a result, stochastic gradient descent finds the global minimizer of the masked objective, which generalizes as discussed above.

Combining the above analysis of the approximation, generalization, and optimization errors in the answer to (a)-(c), we provide a complete characterization of the attention mechanism.

Contribution: In summary, our theoretical contribution is threefold:

(i) We identify a general principle for parameterizing function classes and constructing learning objectives based on latent posterior inference, which requires a minimal assumption of data. In contrast to classical learning paradigms, the latent variable model is defined over input tokens within one data point, which captures relational inductive biases.

(ii) We recover the attention mechanism from a specific parameterization of latent posterior inference based on kernel conditional mean embedding and nonparametric conditional density estimation. In particular, we demonstrate how the attention mechanism combines the handcrafted architecture, which is induced by latent posterior inference, and the learnable parameter, which determines the kernel function.

(iii) We characterize the approximation, generalization, and optimization errors for estimating the learnable parameter of the attention mechanism through minimizing the masked objective. In particular, we prove that input sizes do not degrade the approximation and generalization errors, which justifies why transformers allow long sequences.

Discussion: Our theoretical analysis casts the attention mechanism as a “greybox” approach to modeling, that is, it combines the handcrafted architecture, which is coined by a probabilistic model over input tokens within one data point (“whitebox”), and the learnable parameter, which is estimated in an end-to-end manner through empirical risk minimization (“blackbox”). It is worth mentioning that our theoretical analysis studies the class of transformers like BERT (Devlin et al., 2018) and ViT (Dosovitskiy et al., 2020) ("encoder-only"), which does not exploit the autoregressive structure as in GPT (Brown et al., 2020) ("decoder-only"). On the other hand, the general principle identified in (i) is applicable to other probabilistic models like hidden Markov models or general graphical models over trees and grids, which motivates other principled architectures beyond the attention mechanism. We leave it as a future direction.

Related Works

Transformers and Attention. The pioneering work (Vaswani et al., 2017) proposes transformers for the first time and highlights the key role of the attention mechanism. Subsequently, there are a vast body of works that propose various transformer architectures and different pretraining paradigms. See, e.g., Devlin et al. (2018); Radford et al. (2018, 2019); Dai et al. (2019); Brown et al. (2020); Dosovitskiy et al. (2020); He et al. (2022) and the references therein. Transformers demonstrate significant empirical successes in natural language processing (Wolf et al., 2020), computer vision (Dosovitskiy et al., 2020), protein structure prediction (Jumper et al., 2021), and sequential decision making (Chen et al., 2021). Our work provides a theoretical justification of transformers and the attention mechanism, that is, how a latent variable model induced by exchangeability allows us to derive transformer architectures and pretraining paradigms in a principled manner.

Analysis of Transformers and Attention. Our work is related to a recent line of works that analyze transformers and the attention mechanism (Tsai et al., 2019; Vuckovic et al., 2020; Hron et al., 2020; Yang, 2020; Yang and Littwin, 2021; Edelman et al., 2021; Wei et al., 2021a; Xie et al., 2021; Malladi et al., 2022; Garg et al., 2022; Zhang et al., 2022b).

Specifically, Tsai et al. (2019) demonstrate that the attention mechanism can be viewed as a kernel smoother over input tokens. Vuckovic et al. (2020) establish the Lipschitz continuity of transformers via the lens of interacting particle systems. Hron et al. (2020); Yang (2020); Yang and Littwin (2021); Malladi et al. (2022) characterize the infinite-width limit of transformers under the framework of neural tangent kernels (Jacot et al., 2018). Among them, Malladi et al. (2022) demonstrate that neural tangent kernels can capture the parameter update in the fine-tuning phase. Edelman et al. (2021) prove that transformers can represent a sparse function of input tokens and establish a sample complexity that scales logarithmically in input sizes. Wei et al. (2021a) characterize the approximation and generalization errors for learning a Turing machine with transformers. Xie et al. (2021) prove that transformers can infer a latent variable (or "concept") assuming that the data distribution is a mixture of hidden Markov models. Garg et al. (2022) demonstrate that transformers can learn to perform linear predictions within one data point (in an "in-context" manner). Zhang et al. (2022b) evaluate the empirical performance of transformers for learning equality and group operations.

Our work provides a complete characterization of the representation, inference, and learning aspects of the attention mechanism via the lens of exchangeability and latent variable models, which requires a minimal assumption on the data distribution (exchangeability). Specifically, in comparison with Tsai et al. (2019), we demonstrate that the attention mechanism not only parameterizes nonparametric conditional density estimation but also approximates kernel conditional mean embedding, which infers the conditional mean of the value given the key. Moreover, we invoke the latent variable model induced by exchangeability to justify the attention mechanism as a specific parameterization of latent posterior inference. Meanwhile, we leverage the latent variable model to derive the common choice of both supervised and self-supervised objectives, e.g., the masked objective. In comparison with Xie et al. (2021), we do not assume that the data distribution takes a specific form (a mixture of hidden Markov models) or the latent posterior is given a priori by a specific parameter (no learning required). Instead, we prove that the attention mechanism is capable of instantiating latent posterior inference up to an approximation error and the masked objective allows us to learn to infer the latent posterior up to the generalization and optimization errors. Also, it is worth mentioning that Xie et al. (2021) focus on the class of transformers like GPT (Brown et al., 2020) ("decoder-only"), while we focus on the class of transformers like BERT (Devlin et al., 2018) and ViT (Dosovitskiy et al., 2020) ("encoder-only"). In comparison to Edelman et al. (2021), we exploit the invariance and equivariance of transformers and establish a generalization error that is independent of input sizes.

Generalization of Deep Neural Networks. Our work is related to the vast body of works that analyze the generalization error of deep neural networks. See, e.g., Jiang et al. (2019); Valle-Pérez and Louis (2020) for a comprehensive introduction. However, most of them do not exploit invariance and equivariance. As a result, a direct application of such results yields a vacuous bound as input sizes increase. On the other hand, Sokolic et al. (2017); Sannai et al. (2021); Elesedy (2021); Zhu et al. (2021) establish a generalization error that captures the improvement from invariance and equivariance, which, however, is not applicable to the attention mechanism. Our theoretical analysis of the generalization error follows the framework of Bartlett et al. (2017), which stems from Bartlett (1996); Bartlett and Mendelson (2002). In addition, the concurrent work (Zhang et al., 2022a) provides a PAC-Bayes analysis of the generalization error of the attention mechanism in the context of multiagent reinforcement learning.

Optimization of Deep Neural Networks. Our work is built on the vast body of works that analyze the optimization error of deep neural networks (Allen-Zhu et al., 2019a,c,b; Arora et al., 2019b; Du et al., 2018b, 2019; Zhang et al., 2019c; Zou et al., 2018; Zou and Gu, 2019; Allen-Zhu et al., 2019c; Cao and Gu, 2019b; Li and Liang, 2018; Chizat et al., 2019; Mei et al., 2018, 2019; Rotskoff and Vanden-Eijnden, 2018; Nguyen, 2019; Sirignano and Spiliopoulos, 2020). Most of them focus on overparameterized neural networks in the neural tangent kernel (Jacot et al., 2018) or mean-field regime (Mei et al., 2018). Our work analyzes the optimization error in the neural tangent kernel regime, which is similar to Malladi et al. (2022). Meanwhile, it is worth mentioning that our theoretical analysis of the approximation and generalization errors is not restricted to the neural tangent kernel regime.

Invariance and Equivariance in Deep Neural Networks. Our work is related to a recent line of works on deep neural networks with invariance or equivariance with respect to permutations and other group operations. See, e.g., Scarselli et al. (2008); Zaheer et al. (2017); Lee et al. (2019a); Keriven and Peyré (2019); Romero and Cordonnier (2020); Bloem-Reddy and Teh (2020); Hutchinson et al. (2021); Satorras et al. (2021); Kossen et al. (2021) and the references therein. Also, see Valle-Pérez and Louis (2020); Han et al. (2022) for a detailed survey. In comparison, we exploit the latent variable model induced by exchangeability to provide a complete characterization of the representation, inference, and learning aspects of the attention mechanism.

3.2. Preliminary

Notations. We denote by $[L]$ the index set $\{1, 2, \ldots, L\}$ for any $L \in \mathbb{N}_+$. For any vector $v \in \mathbb{R}^L$, we denote by $\texttt{softmax}(v) = (\exp(v^\ell)/(\sum_{\ell'=1}^{L} \exp(v^{\ell'})))_{\ell\in[L]} \in \mathbb{R}^L$ the softmax function. We denote by $\|\cdot\|_2$ the spectral norm, which becomes the ℓ_2-norm when it operates on a vector. We denote by $\|\cdot\|_\mathrm{F}$ the Frobenius norm. For any $d \in \mathbb{N}_+$, we denote by $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \,|\, \|x\|_2 = 1\}$ the $(d-1)$-dimensional unit sphere.

Reproducing Kernel Hilbert Space. Let $\mathcal{H}_x$ be a Hilbert space over a domain $\mathfrak{X}$, which contains functions $f : \mathfrak{X} \to \mathbb{R}$ and is equipped with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_x}$. We say that $\mathcal{H}_x$ is a reproducing kernel Hilbert space (RKHS) with the kernel function $\mathfrak{K} : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$ if we have the reproducing property that $\langle f, \mathfrak{K}(x,\cdot)\rangle_{\mathcal{H}_x} = f(x)$ for any $f \in \mathcal{H}_x$ and $x \in \mathfrak{X}$. An RKHS $\mathcal{H}_x$ is associated with a feature mapping $\phi : \mathfrak{X} \to \ell_2$ such that $\mathfrak{K}(x, x') = \phi(x)^\top \phi(x')$ for any $x, x' \in \mathfrak{X}$ (Muandet et al., 2016). Here we denote by ℓ_2 the space of all square-summable series.

Attention Mechanism. For an input sequence $X = \{x^\ell\}_{\ell\in[L]}$ with the input tokens $x^\ell \in \mathbb{R}^d$, we consider the key matrix $K \in \mathbb{R}^{L\times d_\mathrm{p}}$ and the value matrix $V \in \mathbb{R}^{L\times d}$ defined as

$$K = (k^1, \ldots, k^L)^\top = \big(k_\theta(x^1), \ldots, k_\theta(x^L)\big)^\top \in \mathbb{R}^{L\times d_\mathrm{p}},$$
$$V = (v^1, \ldots, v^L)^\top = \big(v_\theta(x^1), \ldots, v_\theta(x^L)\big)^\top \in \mathbb{R}^{L\times d}.$$

Here $k_\theta : \mathbb{R}^d \to \mathbb{R}^{d_\mathrm{p}}$ and $v_\theta : \mathbb{R}^d \to \mathbb{R}^d$ map the ℓ-th input token x^ℓ to the key k^ℓ and the value v^ℓ, respectively, where $\theta \in \Theta$ is the the learnable parameter. For any query $q \in \mathbb{R}^{d_\mathrm{p}}$, we define the attention mechanism as follows,

$$\mathtt{attn}(q, K, V) = V^\top \mathtt{norm}\big(\mathfrak{K}(K, q)\big) \in \mathbb{R}^d, \tag{3.2.1}$$

where $\mathfrak{K} : \mathbb{R}^{d_\mathrm{p}} \times \mathbb{R}^{d_\mathrm{p}} \to \mathbb{R}$ is a kernel function and we write $\mathfrak{K}(K, q) = (\mathfrak{K}(k^\ell, q))_{\ell \in L} \in \mathbb{R}^L$. Here we denote by $\mathtt{norm} : \mathbb{R}^L \to \mathbb{R}^L$ a normalization mapping.

A common example of the attention mechanism is the softmax attention (Vaswani et al., 2017), where the kernel function is the exponential kernel $\mathfrak{K}_{\mathtt{EXP}}(q, k) = \exp(q^\top k/\gamma)$ with a fixed $\gamma > 0$ and the normalization mapping is the following softmax normalization,

$$\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}(K, q)\big) = \big(\mathbf{1}^\top \mathfrak{K}(K, q)\big)^{-1} \cdot \mathfrak{K}(K, q).$$

The attention mechanism in (3.2.1) with the exponential kernel and the softmax normalization is the softmax attention (Vaswani et al., 2017), which takes the following form,

$$\mathtt{attn}_{\mathtt{SM}}(q, K, V) = V^\top \mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{EXP}}(K, q)\big) = \mathtt{softmax}(Kq/\gamma).$$

3.3. Representation, Inference, and Estimation via Latent Variable Model

From Exchangeability to Latent Variable Model. We consider the input sequence $X = \{x^\ell\}_{\ell \in [L]}$, where $x^\ell \in \mathbb{R}^d$ is an input token and $L \in \mathbb{N}_+$ is the sequence length. In natural language processing (NLP), such a sequence consists of embeddings of words in a paragraph, while in computer vision (CV), such a sequence consists of embeddings of patches in an image.

As is the case in BERT (Devlin et al., 2018) and ViT (Dosovitskiy et al., 2020), the input sequence is exchangeable since it includes positional encodings. Specifically, we say that a random variable sequence $\{x^\ell\}_{\ell\in\mathbb{N}_+}$ is exchangeable if and only if it holds for any sequence length $L \in \mathbb{N}_+$ and any index permutation $\pi : [L] \to [L]$ that

$$\mathbb{P}(x^1, x^2, \ldots, x^L) = \mathbb{P}(x^{\pi(1)}, x^{\pi(2)}, \ldots, x^{\pi(L)}).$$

In other words, permuting the index order within the random variable sequence does not affect its joint distribution. The following proposition states that the exchangeability of a random variable sequence induces a latent variable.

Proposition 3.3.1 (de Finetti Representation Theorem (de Finetti, 1937))**.** Let $\{x^\ell\}_{\ell\in\mathbb{N}_+}$ be an exchangeable sequence. Then, there exists a latent variable z such that for any sequence length $L \in \mathbb{N}_+$,

$$\begin{aligned}\mathbb{P}(x^1, \ldots, x^L) &= \int \prod_{\ell=1}^{L} \mathbb{P}(x^\ell \,|\, z) \cdot \mathbb{P}(z)\mathrm{d}z,\\ \mathbb{P}(x^\ell \,|\, x^1, \ldots, x^{\ell-1}, x^{\ell+1}, \ldots, x^L) &= \int \mathbb{P}(x^\ell \,|\, z) \cdot \mathbb{P}(z \,|\, x^1, \ldots, x^{\ell-1}, x^{\ell+1}, \ldots, x^L)\mathrm{d}z, \quad \forall \ell \in [L].\end{aligned}$$

We remark that Proposition 3.3.1 requires an infinite-length exchangeable sequence. Up to an approximation error, a finite-length exchangeable sequence also induces a latent variable (Diaconis and Freedman, 1980). In what follows, we consider the former case where the input sequence includes positional encodings and is thus exchangeable (Devlin et al., 2018; Dosovitskiy et al., 2020). See Figure 3.1 for an illustration of the exchangeability.

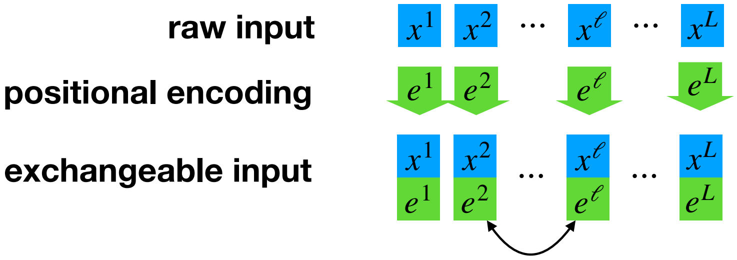


Figure 3.1. The input sequence (the raw version without positional encodings) becomes exchangeable with positional encodings. In practice, the positional encoding is incorporated in an additive manner (instead of concatenation).

Proposition 3.3.1 guarantees the existence of a latent variable, which forms the basis of our theoretical analysis. See Figure 3.2a for an illustration. Intuitively, the latent variable can be viewed as the "concept" of the input sequence, which is "summarized" over words or patches. For instance, in NLP, the latent variable instantiates the "meaning" of a paragraph, while in CV, the latent variable instantiates the "theme" of an image. In particular, the latent posterior $\mathbb{P}(z \mid X)$ plays a key role in solving downstream tasks (Song et al., 2014; Xie et al., 2021), as it is a desired representation of the input sequence X. See Figure 3.2b for an illustration. In the following lemma, we prove that the latent posterior $b_z(X) = \mathbb{P}(z = \cdot \mid X)$ is a minimal sufficient statistic (Fisher, 1922).

Lemma 3.3.2 (Minimal Sufficiency of Latent Posterior)**.** Let z be the latent variable induced by the exchangeability of the input sequence X. The latent posterior $b_z(X) = \mathbb{P}(z = \cdot \mid X)$ is a minimal sufficient statistic of the input sequence X for the latent variable z. Meanwhile, for any target variable y that is independent of the input sequence X conditioning on the latent variable z, we assume the invertibility of the operator $\mathcal{T}$ defined by

$$(\mathcal{T}f)(y) = \int \mathbb{P}(y \,|\, z) f(z) \mathrm{d}z. \tag{3.3.1}$$

Then, the latent posterior $b_z(X) = \mathbb{P}(z = \cdot \,|\, X)$ is a minimal sufficient statistic of the input sequence X for the target variable y.

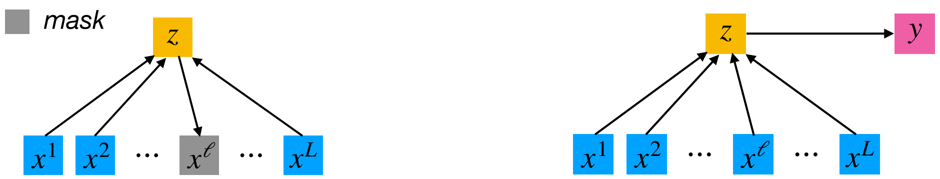
Proof. See §C.2.3 for a detailed proof. □

From Latent Variable Model to Learning Objectives. In what follows, we consider the prediction task in BERT (Devlin et al., 2018) and ViT (Dosovitskiy et al., 2020). Let y be the target variable and $X = \{x^\ell\}_{\ell \in [L]}$ is the input sequence. In particular, in self-supervised learning (BERT), the target variable y is a masked token of the input sequence, while in supervised learning (ViT), y is the unknown label corresponding to the class encoding. We remark that in ViT, the unknown label y corresponds to the masked token in BERT, while the input class encoding corresponds to the mask in BERT. In both cases, the concatenation $\{y, x^1, \ldots, x^L\}$ is treated as an exchangeable sequence since it includes the positional encodings. By Proposition 3.3.1, we have

$$\mathbb{P}(y \,|\, X) = \int \mathbb{P}(y \,|\, z) \cdot \mathbb{P}(z \,|\, X) \mathrm{d}z, \tag{3.3.2}$$

where z is the latent variable induced by the exchangeability of X. See Figure 3.2 for an illustration.

In what follows, we treat y as a target variable that satisfies (3.3.2), which specifies that y is independent of the input sequence X conditioning on the latent variable z. By Lemma 3.3.2, the latent posterior $b_z(X)$ is a minimal sufficient statistic of X for y. In other words, the latent posterior $b_z(X)$ is a desired representation of the input sequence X. According to (3.3.2), the prediction of the target variable y from the input sequence X (forward pass) takes two implicit steps: i) the inference of the latent posterior $\mathbb{P}(z \,|\, X)$, and ii) the prediction of y based on the generative distribution $\mathbb{P}(y \,|\, z)$ integrated with the latent posterior $\mathbb{P}(z \,|\, X)$.



(a) Prediction of a masked token x^ℓ.

(b) Prediction of the target variable y, which can be viewed as a masked token.

Figure 3.2. The forward pass for the prediction of the masked token x^ℓ and the target variable y. The prediction of y takes two steps: i) the inference of the latent posterior $\mathbb{P}(z \,|\, X)$, and ii) the prediction of y based on the generative distribution $\mathbb{P}(y \,|\, z)$ integrated with the latent posterior $\mathbb{P}(z \,|\, X)$.

To construct the learning objective, we consider the distribution of the target variable y conditioning on the input sequence X and parameterize it by $\mathbb{P}_\theta(y \,|\, X)$, where θ is the learnable parameter. Given a dataset $\mathcal{D}_n = \{(X_i, y_i)\}_{i \in [n]}$, where X_i is the i-th input sequence and y_i is the i-th target variable, the maximum likelihood estimation (MLE) objective takes the following form,

$$\begin{aligned} &\max_\theta \widehat{\mathbb{E}}_{(X,y)\sim\mathcal{D}_n}\big[\log \mathbb{P}_\theta(y \,|\, X)\big] \\ &\quad = \widehat{\mathbb{E}}_{(X,y)\sim\mathcal{D}_n}\bigg[\log \int \mathbb{P}_\theta(y \,|\, z)\mathbb{P}_\theta(z \,|\, X)\mathrm{d}z\bigg]. \end{aligned} \tag{3.3.3}$$

We define $\widehat{\mathbb{E}}_{(X,y)\sim\mathcal{D}_n}[\,\cdot\,]$ as the empirical expectation with respect to the dataset $\mathcal{D}_n$.

3.3.1. Preliminary Finite-Dimensional Example

Latent Variable Model. We provide a finite-dimensional Gaussian-distributed example to illustrate the latent variable model and the MLE objective. Specifically, we consider the setting with the input sequence $X = \{x^\ell\}_{\ell \in [L]}$ and the target variable y, where $x^\ell \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$. For the input sequence X, we consider the following example of the latent variable model in (3.3.2),

$$r^\ell = zc^\ell + \epsilon^\ell, \qquad \text{where} \quad c^\ell = c_*(x^\ell), \quad r^\ell = r_*(x^\ell), \quad \forall \ell \in [L]. \tag{3.3.4}$$

Here $z \in \mathbb{R}^{d_r \times d_c}$ is the latent variable induced by the exchangeability of the input sequence, $c^\ell \in \mathbb{R}^{d_c}$ and $r^\ell \in \mathbb{R}^{d_r}$ are the covariate and response, respectively, which are determined by two unknown functions $c_* : \mathbb{R}^d \to \mathbb{R}^{d_c}$ and $r_* : \mathbb{R}^d \to \mathbb{R}^{d_r}$, and $\epsilon^\ell \sim N(0, \sigma^2 I)$ is the noise, which is independent of c^ℓ. In practice, the covariate c^ℓ instantiates the contextual information, while the response r^ℓ instantiates the semantic information. We consider the prediction of the (unknown) target variable y based on its (known) input mask $\texttt{msk}$. In the self-supervised setting (Devlin et al., 2018), y is a masked token of the input sequence, while $\texttt{msk}$ is the positional encoding. In the supervised setting (Dosovitskiy et al., 2020), y is the label of the input sequence, while $\texttt{msk}$ is the class encoding. Specifically, corresponding to (3.3.4), we consider the prediction model with $y = r^{\texttt{msk}}$, where $r^{\texttt{msk}}$ is the response corresponding to the covariate $c^{\texttt{msk}}$ of the input mask such that

$$y = r^{\texttt{msk}} = zc^{\texttt{msk}} + \epsilon, \qquad \text{where} \quad c^{\texttt{msk}} = c_*^{\texttt{m}}(\texttt{msk}). \tag{3.3.5}$$

Here $c_*^{\mathtt{m}} : \mathbb{R}^d \to \mathbb{R}^{d_c}$ is an unknown function and $\epsilon \sim N(0, \sigma^2 I)$ is the noise, which is independent of $c^{\mathtt{msk}}$. For example, (3.3.5) holds when we consider an exchangeable sequence $\{x^1, \ldots, x^L, x^{L+1}\}$ satisfying (3.3.4) for any $\ell \in [L+1]$ with the input sequence $X = \{x^1, \ldots, x^L\}$, $c^{\mathtt{msk}} = c^{L+1}$, and the target variable $y = r^{\mathtt{msk}} = r^{L+1}$. In the next section, we consider an advanced infinite-dimensional example and show that $c^{\mathtt{msk}}$, c^ℓ, and r^ℓ correspond to the query, the key, and the value in the attention mechanism, respectively.

Note that the regression model in (3.3.4) is a conditional model. Instead of modeling the conditional distribution of y given X as in (3.3.2), we model the conditional distribution of y given X and $\mathtt{msk}$. Recall that $y = r^{\mathtt{msk}}$. Corresponding to (3.3.2), the generative distribution takes the following form,

$$\mathbb{P}(y \,|\, \mathtt{msk}, z) \propto \exp\Big(-\big\|y - z c_*^{\mathtt{m}}(\mathtt{msk})\big\|_2^2 / 2\sigma^2\Big). \tag{3.3.6}$$

We take the Gaussian distribution $N(0, \lambda I)$ as the prior of z. By (3.3.4), the latent posterior $\mathbb{P}(z \,|\, X)$ is a Gaussian distribution, which (approximately) takes the following form,

$$\mathbb{P}(z \,|\, X) \propto \exp\Big(-\big\|z - \bar{z}(X)\big\|_2^2 / 2\iota^2\Big). \tag{3.3.7}$$

Here the covariance of the latent posterior is approximated by $\iota^2 I$ and the mean $\bar{z}(X)$ of the latent posterior takes the following form,

$$\bar{z}(X) = \mathbb{E}[z \,|\, X] = R^\top (CC^\top + \lambda I)^{-1} C,$$

where we define $C = (c^1, \ldots, c^L)^\top \in \mathbb{R}^{L\times d_c}$ and $R = (r^1, \ldots, r^L)^\top \in \mathbb{R}^{L\times d_r}$. Combining (3.3.6) and (3.3.7), we obtain

$$\text{(3.3.8)} \quad \mathbb{P}(y \,|\, \texttt{msk}, X) = \int \mathbb{P}(y \,|\, \texttt{msk}, z) \cdot \mathbb{P}(z \,|\, X) \mathrm{d}z \propto \exp\Bigl(-\bigl\|y - \bar{z}(X) c_*^{\texttt{m}}(\texttt{msk})\bigr\|_2^2 / 2\widetilde{\sigma}^2\Bigr),$$

which corresponds to (3.3.2), Here we approximate the covariance of y conditioning on X and msk by $\widetilde{\sigma}^2 I$, where $\widetilde{\sigma}$ does not depend on X. We remark that (3.3.8) is a form of Bayesian model averaging (Wasserman, 2000) within one data point.

Parameterization of Latent Variable Model. Recall that c_*, r_*, and $c_*^{\texttt{m}}$ in (3.3.4) and (3.3.5) are unknown. We parameterize them with c_θ, r_θ, and $c_\theta^{\texttt{m}}$, where $\theta \in \Theta$ is a learnable parameter. With the ideal parameter $\theta^* \in \Theta$, it holds for any $\ell \in [L]$ that

$$\text{(3.3.9)} \qquad c_{\theta^*}(x^\ell) = c_*(x^\ell) = c^\ell, \quad r_{\theta^*}(x^\ell) = r_*(x^\ell) = r^\ell, \quad c_{\theta^*}^{\texttt{m}}(\texttt{msk}) = c_*^{\texttt{m}}(\texttt{msk}) = c^{\texttt{msk}}.$$

By (3.3.7), we parameterize the latent posterior $\mathbb{P}(z \,|\, X)$ as follows,

$$\text{(3.3.10)} \qquad \mathbb{P}_\theta(z \,|\, X) \propto \exp\Bigl(-\bigl\|z - \bar{z}_\theta(X)\bigr\|_2^2 / 2\iota^2\Bigr),$$

where $\bar{z}_\theta(X)$ is calculated as follows,

$$\bar{z}_\theta(X) = r_\theta(X)^\top \bigl(c_\theta(X) c_\theta(X)^\top + \lambda I\bigr)^{-1} c_\theta(X).$$

Here $c_\theta(X) = (c_\theta(x^1), \ldots, c_\theta(x^L))^\top \in \mathbb{R}^{L\times d_c}$ and $r_\theta(X) = (r_\theta(x^1), \ldots, r_\theta(x^L))^\top \in \mathbb{R}^{L\times d_r}$. By (3.3.6), we parameterize the generative distribution $\mathbb{P}(y \,|\, \texttt{msk}, z)$ as follows,

$$\mathbb{P}_\theta(y \,|\, \texttt{msk}, z) \propto \exp\Bigl(-\bigl\|y - z c_\theta^{\texttt{m}}(\texttt{msk})\bigr\|_2^2 / 2\sigma^2\Bigr).$$

By (3.3.8), we define the conditional likelihood $\mathbb{P}(y \,|\, \mathtt{msk}, X)$ as follows,

$$\mathbb{P}_\theta(y \,|\, \mathtt{msk}, X) \propto \exp\Big(-\big\|y - \bar{z}_\theta(X) c_\theta^{\mathtt{m}}(\mathtt{msk})\big\|_2^2 / 2\widetilde{\sigma}^2\Big). \tag{3.3.11}$$

Training and Testing. In the training phase, given the dataset $\mathcal{D}_n = \{(X_i, y_i)\}_{i\in[n]}$, we aim to maximize the MLE objective in (3.3.3). By (3.3.11), maximizing the MLE objective is equivalent to minimizing the mean-squared error as follows,

$$\min_\theta \widehat{\mathbb{E}}_{(X,y)\sim\mathcal{D}_n}\Big[\big\|y - \bar{z}_\theta(X) c_\theta^{\mathtt{m}}(\mathtt{msk})\big\|_2^2\Big]. \tag{3.3.12}$$

Note that the learnable parameter θ is estimated across different data points in the dataset $\mathcal{D}_n$ through the backward pass, while the latent variable z is inferred within one data point (X_i, y_i) through the forward pass. We remark that by learning θ, the model learns to perform Bayesian model averaging. Suppose that we solve (3.3.12) and obtain the estimator $\widehat{\theta}$. In the testing phase, given an input sequence $X_\dagger$ and an input mask $\mathtt{msk}_\dagger$, we predict the target variable $y_\dagger$ by maximizing the posterior of y,

$$\widehat{y} = \operatorname*{argmax}_y \mathbb{P}_{\widehat{\theta}}(y \,|\, \mathtt{msk}_\dagger, X_\dagger) = \mathbb{E}[r^{\mathtt{msk}} \,|\, \mathtt{msk}_\dagger, X_\dagger] = \bar{z}_{\widehat{\theta}}(X_\dagger) c_{\widehat{\theta}}^{\mathtt{m}}(\mathtt{msk}_\dagger). \tag{3.3.13}$$

We remark that the learning process for the attention mechanism involves two aspects. In the forward pass, within one data point, we infer the latent posterior $\mathbb{P}(z \,|\, X)$ to predict the target variable y. In the backward pass, we estimate the learnable parameter θ across different data points. See Figure 3.3 for an illustration.

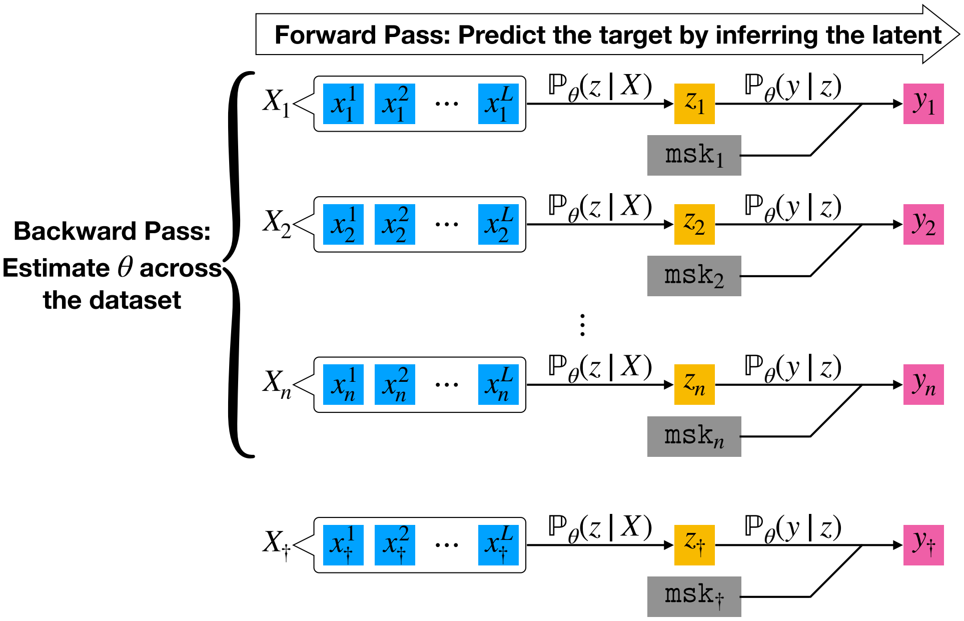

Forward Pass: Predict the target by inferring the latent
Backward Pass: Estimate θ across the dataset
X_1
x_1^1 x_1^2 ⋯ x_1^L
$\mathbb{P}_\theta(z \mid X)$
z_1
$\mathbb{P}_\theta(y \mid z)$
y_1
$\mathtt{msk}_1$
X_2
x_2^1 x_2^2 ⋯ x_2^L
z_2
y_2
$\mathtt{msk}_2$
X_n
x_n^1 x_n^2 ⋯ x_n^L
z_n
y_n
$\mathtt{msk}_n$
$X_\dagger$
$x_\dagger^1$ $x_\dagger^2$ ⋯ $x_\dagger^L$
$z_\dagger$
$y_\dagger$
$\mathtt{msk}_\dagger$

Figure 3.3. Forward pass: within one data point (X, y), we infer the latent posterior $\mathbb{P}_\theta(z \mid X)$ by (3.3.10). We predict $y_\dagger$ by $\widehat{y}$ in (3.3.13). Backward pass: across different data points in the dataset $\mathcal{D}_n$, we estimate the learnable parameter θ by (3.3.12).

The finite-dimensional example illustrates a "greybox" approach to modeling, that is, it combines the handcrafted architecture in (3.3.8) ("whitebox"), and the learnable parameter in (3.3.12), which is estimated in an end-to-end manner through empirical risk minimization ("blackbox"). As shown in Figure 3.4, the forward pass first infers the latent variable z and then utilizes the latent variable z to predict the masked token (in the self-supervised setting) or the unknown label (in the supervised setting). Meanwhile, the backward pass estimatess the learnable parameter. In the following section, we extend the finite-dimensional example to the infinite-dimensional setting, which recovers the attention mechanism. In particular, we demonstrate that the attention mechanism infers the latent posterior within a data point. Also, we show that the covariate corresponds to the query and key and that the response corresponds to the value in the attention mechanism.

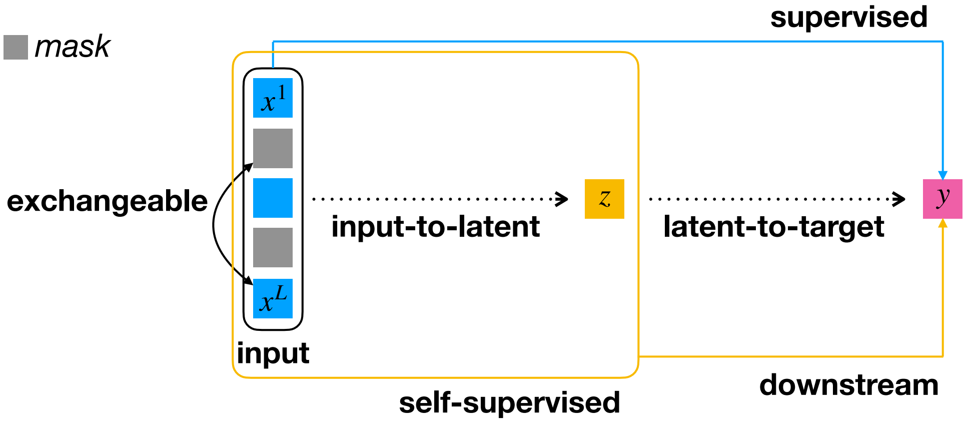


Figure 3.4. The forward and backward passes in transformers. Dotted arrows stand for forward passes (input→latent→target). Solid arrows stand for backward passes (training). Masks (grey tokens) are only used to illustrate the self-supervised setting (yellow box).

3.4. Attention as Latent Posterior Inference

In what follows, we demonstrate how the attention mechanism performs latent posterior inference for the latent variable model, which is induced by the exchangeability of the input sequence. In §3.4.1, we extend the finite-dimensional example in §3.3 to an RKHS to induce a variant of the softmax attention, namely, the conditional mean embedding (CME) attention. In particular, we prove that it infers the latent posterior in the forward pass. In §3.4.2, we prove that the softmax attention has the same limit as the CME attention when the sequence length goes to infinity, which implies that the softmax attention approximately infers the latent posterior.

3.4.1. Attention as Kernel Conditional Mean Embedding

Advanced Infinite-Dimensional Example. We present an infinite-dimensional version of the preliminary example in §3.3.1, which motivates us to study the CME attention. Similarly to (3.3.4), we consider the following model for the input sequence $X = \{x^\ell\}_{\ell \in [L]}$ with input token $x^\ell \in \mathbb{R}^d$,

$$r^\ell = z\phi(c^\ell) + \epsilon^\ell, \qquad \text{where} \quad c^\ell = c_*(x^\ell), \quad r^\ell = r_*(x^\ell), \quad \forall \ell \in [L]. \tag{3.4.1}$$

Here $c^\ell \in \mathbb{R}^{d_{\mathrm{p}}}$ and $r^\ell \in \mathbb{R}^d$ are the covariate and the response, respectively, which are determined by two unknown functions $c_* : \mathbb{R}^d \to \mathbb{R}^{d_{\mathrm{p}}}$ and $r_* : \mathbb{R}^d \to \mathbb{R}^d$, $\phi : \mathbb{R}^{d_{\mathrm{p}}} \to \mathcal{H}_c$ is the feature mapping of the RKHS $\mathcal{H}_c$, $z : \mathcal{H}_c \to \mathbb{R}^d$ is a linear mapping, which is viewed as the latent variable induced by the exchangeability of the input sequence X, and $\epsilon^\ell \sim N(0, \sigma^2 I)$ is the Gaussian noise, which is independent of the covariate c^ℓ. Note that $\phi(c^1)^\top \phi(c^2) = \mathfrak{K}(c^1, c^2)$ for any $c^1, c^2 \in \mathbb{R}^{d_{\mathrm{p}}}$, where $\mathfrak{K} : \mathbb{R}^{d_{\mathrm{p}}} \times \mathbb{R}^{d_{\mathrm{p}}} \to \mathbb{R}$ is the kernel function of the RKHS $\mathcal{H}_c$. A common example is the Gaussian radial basis function (RBF) kernel $\mathfrak{K}_{\mathtt{RBF}}(q, k) = \exp(-\|q - k\|_2^2 / 2\gamma)$ with $\gamma > 0$. Similarly to (3.3.5), the (unknown) target variable y is determined by its (known) input mask $\mathtt{msk}$, which satisfies

$$y = r^{\mathtt{msk}} = z\phi(c^{\mathtt{msk}}) + \epsilon, \qquad \text{where} \quad c^{\mathtt{msk}} = c_*^{\mathtt{m}}(\mathtt{msk}). \tag{3.4.2}$$

Here we denote by $c^{\mathtt{msk}}$ and $r^{\mathtt{msk}}$ the covariate and the response corresponding to the input $\mathtt{msk}$, respectively, $c_*^{\mathtt{m}} : \mathbb{R}^d \to \mathbb{R}^{d_{\mathrm{p}}}$ is an unknown function, and $\epsilon \sim N(0, \sigma^2 I)$ is a Gaussian noise, which is independent of $c^{\mathtt{msk}}$. To simplify the presentation, we view the RKHS $\mathcal{H}_c$ as a vector space $\mathbb{R}^{d_\phi}$ with $d_\phi = \infty$. Correspondingly, we view the latent variable z as a matrix in $\mathbb{R}^{d\times d_\phi}$. We present a rigorous characterization of z in §C.2.1 based on Gaussian process regression. Similarly to (3.3.9), we parameterize c_*, r_*, and $c_*^{\mathtt{m}}$ by c_θ, r_θ, and $c_\theta^{\mathtt{m}}$, where $\theta \in \Theta$ is a learnable parameter. Similarly to (3.3.13), we predict $r^{\mathtt{msk}}$ in the forward pass via

$$\begin{aligned} \widehat{r} &= \mathbb{E}[r^{\mathtt{msk}} \,|\, \mathtt{msk}, X] \\ &= r_\theta(X)^\top \Big(\phi\big(c_\theta(X)\big)\phi\big(c_\theta(X)\big)^\top + \lambda I\Big)^{-1} \phi\big(c_\theta(X)\big)\phi\big(c_\theta^{\mathtt{m}}(\mathtt{msk})\big) \\ &= r_\theta(X)^\top \Big(\mathfrak{K}\big(c_\theta(X), c_\theta(X)\big) + \lambda I\Big)^{-1} \mathfrak{K}\big(c_\theta(X), c_\theta^{\mathtt{m}}(\mathtt{msk})\big), \end{aligned} \tag{3.4.3}$$

where we define $\mathfrak{K}(c_\theta(X), c_\theta(X)) = (\mathfrak{K}(c_\theta(x^i), c_\theta(x^j)))_{i,j\in[L]} \in \mathbb{R}^{L\times L}$, $\mathfrak{K}(c_\theta(X), c_\theta^{\mathtt{m}}(\mathtt{msk})) = (\mathfrak{K}(c_\theta(x^\ell), c_\theta^{\mathtt{m}}(\mathtt{msk})))_{\ell\in[L]} \in \mathbb{R}^L$, $\phi(c_\theta(X)) = (\phi(c_\theta(x^1)), \ldots, \phi(c_\theta(x^L)))^\top \in \mathbb{R}^{L\times d_\phi}$, and $r_\theta(X) = (r_\theta(x^1), \ldots, r_\theta(x^L))^\top \in \mathbb{R}^{L\times d}$. We remark that (3.4.3) recovers the empirical version of the kernel conditional mean embedding of $\mathbb{P}_{\mathcal{R}\,|\,\mathcal{C}}$ (Song et al., 2009), where we denote by $\mathbb{P}_{\mathcal{R}\,|\,\mathcal{C}}$ the conditional distribution of r^ℓ given c^ℓ (as two random variables) within one data point, and $\mathcal{H}_r = (\mathbb{R}^d)^* = \mathbb{R}^d$ is the dual space of $\mathbb{R}^d$ equipped with the Euclidean kernel $\langle\cdot,\cdot\rangle$.

From Latent Variable Model to Attention. Recall that the attention mechanism is defined in (3.2.1) with $q \in \mathbb{R}^{d_{\mathrm{p}}}$, $K = (k^1, \ldots, k^L)^\top \in \mathbb{R}^{L\times d_{\mathrm{p}}}$, and $V = (v^1, \ldots, v^L)^\top \in \mathbb{R}^{L\times d}$. The kernel conditional mean embedding in (3.4.3) motivates us to consider the following CME normalization,

$$\mathtt{norm}_{\mathtt{CME}}\big(\mathfrak{K}(K, q)\big) = \big(\mathfrak{K}(K, K) + \lambda I\big)^{-1}\mathfrak{K}(K, q), \tag{3.4.4}$$

where we write $\mathfrak{K}(K, q) = (\mathfrak{K}(k^\ell, q))_{\ell\in[L]} \in \mathbb{R}^L$ and $\mathfrak{K}(K, K) = (\mathfrak{K}(k^i, k^j))_{(i,j)\in[L]\times[L]} \in \mathbb{R}^{L\times L}$ We call the attention mechanism with the CME normalization in (3.4.4) the CME attention and denote it by $\mathtt{attn}_{\mathtt{CME}}$. In particular, the CME attention takes the following form,

$$\text{(3.4.5)}\quad \mathtt{attn}_{\mathtt{CME}}(q, K, V) = V^\top \mathtt{norm}_{\mathtt{CME}}\big(\mathfrak{K}(K, q)\big) = V^\top \big(\mathfrak{K}(K, K) + \lambda I\big)^{-1} \mathfrak{K}(K, q) \in \mathbb{R}^d.$$

We see that the CME attention recovers (3.4.3) when

$$\text{(3.4.6)}\qquad q = c^{\mathtt{m}}_\theta(\mathtt{msk}), \qquad k^\ell = c_\theta(x^\ell), \qquad v^\ell = r_\theta(x^\ell), \qquad \forall \ell \in [L].$$

We remark that (3.4.6) establishes a connection between the latent variable model and the attention mechanism. In other words, the covariate $c^{\mathtt{msk}}$ of the input mask $\mathtt{msk}$ corresponds to the query q, the covariate c^ℓ of the input token x^ℓ corresponds to the key k^ℓ for $\ell \in [L]$, and the response r^ℓ of the input token x^ℓ corresponds to the value v^ℓ for $\ell \in [L]$. In the attention mechanism, we denote by $q_\theta : \mathbb{R}^d \to \mathbb{R}^{d_{\mathrm{p}}}$, $k_\theta : \mathbb{R}^d \to \mathbb{R}^{d_{\mathrm{p}}}$, and $v_\theta : \mathbb{R}^d \to \mathbb{R}^d$ the mappings from the input token to the query, the key, and the value, respectively, with the learnable parameter θ. In particular, we have the following correspondence,

$$q_\theta = c^{\mathtt{m}}_\theta, \qquad k_\theta = c_\theta, \qquad v_\theta = r_\theta.$$

In an common example, we instantiate q_θ, k_θ, and v_θ for $x \in \mathbb{R}^d$ as follows,

(3.4.7)

$$q_\theta(x) = (W^{\mathrm{q}})^\top \mathtt{nn}(x; A), \qquad k_\theta(x) = (W^{\mathrm{k}})^\top \mathtt{nn}(x; A), \qquad v_\theta(x) = (W^{\mathrm{v}})^\top \mathtt{nn}(x; A),$$

where $W^{\mathrm{q}}, W^{\mathrm{k}} \in \mathbb{R}^{d\times d_{\mathrm{p}}}$ and $W^{\mathrm{v}} \in \mathbb{R}^{d\times d}$ are learnable parameters. Here we denote by $\mathtt{nn}(\cdot; A) : \mathbb{R}^d \to \mathbb{R}^d$ the feedforward neural network with the learnable parameter A and summarize the learnable parameter as $\theta = (A, W^{\mathrm{q}}, W^{\mathrm{k}}, W^{\mathrm{v}})$. Similarly to (3.3.12), the MLE objective takes the following form,

$$\min_{\theta} \widehat{\mathbb{E}}_{(X,y)\sim\mathcal{D}_n}\Bigl[\Bigl\|y - \mathtt{attn}_{\mathtt{CME}}\bigl(q_\theta(\mathtt{msk}), k_\theta(X), v_\theta(X)\bigr)\Bigr\|_2^2\Bigr], \tag{3.4.8}$$

where we write $k_\theta(X) = (k_\theta(x^1), \ldots, k_\theta(x^L))^\top \in \mathbb{R}^{L\times d_{\mathrm{p}}}$ and $v_\theta(X) = (v_\theta(x^1), \ldots, v_\theta(x^L))^\top \in \mathbb{R}^{L\times d}$.

Limit of CME Attention with $L \to \infty$. Given an input sequence $X = \{x^\ell\}_{\ell\in[L]}$, we consider the key-value pairs $\{(k^\ell, v^\ell)\}_{\ell\in[L]}$ obtained from $k^\ell = k_\theta(x^\ell)$ and $v^\ell = v_\theta(x^\ell)$ for a fixed θ. For notational simplicity, we denote by $\mathcal{K}$ and $\mathcal{V}$ the random variables with the same distribution as (k^ℓ, v^ℓ) within one data point. Recall that we define the CME attention in (3.4.5). Also, we define the covariance operator $\mathcal{C}_{\mathcal{K}\mathcal{K}} = \mathbb{E}[\mathfrak{K}(\mathcal{K}, \cdot) \otimes \mathfrak{K}(\mathcal{K}, \cdot)]$. In the following proposition, we prove that the CME attention approximates the kernel conditional mean embedding of $\mathbb{P}_{\mathcal{V}\,|\,\mathcal{K}}$ as $L \to \infty$. Note that the following proposition does not depend on the latent variable model in (3.4.1).

Proposition 3.4.1 (CME Attention Converges to Kernel Conditional Mean Embedding)**.** Let $\mathfrak{K}$ be a positive definite kernel function. We assume that $\{x^\ell\}_{\ell\in[L]}$ in the input sequence X are independent and identically distributed (within one data point) and the value $\|v^\ell\|_2$ is upper bounded by 1 for any $\ell \in [L]$. It holds with probability at least $1 - \delta$ that

$$\bigl\|\mathtt{attn}_{\mathtt{CME}}(q, K, V) - \mathbb{E}[\mathcal{V}\,|\,\mathcal{K} = q]\bigr\|_2 = \mathcal{O}\biggl(\sqrt{\frac{L}{\lambda}} \cdot \biggl(\frac{2}{\lambda} + \sqrt{\frac{\Gamma(L^{-1}\lambda)}{\lambda}}\biggr) \log\frac{1}{\delta} + \lambda L^{-1}\biggr).$$

Here $\Gamma(L^{-1}\lambda)$ is the effective dimension of the covariance operator $\mathcal{C}_{\mathcal{K}\mathcal{K}}$, which is defined as $\Gamma(L^{-1}\lambda) = \mathrm{Tr}((L^{-1}\lambda\mathcal{I} + \mathcal{C}_{\mathcal{K}\mathcal{K}})^{-1}\mathcal{C}_{\mathcal{K}\mathcal{K}})$.

Proof. See §C.2.4 for a detailed proof. □

We remark that when we use the Gaussian RBF kernel $\mathfrak{K}_{\mathtt{RBF}}$ in the CME attention, it holds that $\Gamma(L^{-1}\lambda) \leq \mathcal{O}(L/\lambda)$ (Zhang et al., 2015). We then have $\big\|\mathtt{attn}_{\mathtt{CME}}(q, K, V) - \mathbb{E}[\mathcal{V} \,|\, \mathcal{K} = q]\big\|_2 \leq \mathcal{O}(L \cdot \lambda^{-3/2} \cdot \log(1/\delta) + \lambda L^{-1})$. Note that the CME attention $\mathtt{attn}_{\mathtt{CME}}$ is a variant of the softmax attention (Vaswani et al., 2017) with a different normalization. In the following section, we prove that the softmax attention has the same limit as the CME attention when the sequence length L goes to infinity.

3.4.2. Softmax Attention Infers Latent Posterior

In §3.4.1, we demonstrate how the latent variable model motivates the design of the CME attention. Recall that we consider the attention mechanism in the form of (3.2.1) with $q \in \mathbb{R}^{d_{\mathrm{p}}}$, $K \in \mathbb{R}^{L\times d_{\mathrm{p}}}$, and $V \in \mathbb{R}^{L\times d}$. In practice, a common normalization is defined as follows,

$$\mathtt{norm}_{\mathtt{SM}}(\mathfrak{K}(K, q)) = \big(\mathbf{1}_L^\top \mathfrak{K}(K, q)\big)^{-1} \cdot \mathfrak{K}(K, q), \tag{3.4.9}$$

where $\mathbf{1}_L \in \mathbb{R}^L$ is the L-dimensional all-one vector and we recall that $\mathfrak{K}(K, q) = (\mathfrak{K}(k^\ell, q))_{\ell\in[L]} \in \mathbb{R}^L$. We denote by $\mathtt{attn}_{\mathtt{SM}}$ the attention mechanism with the normalization in (3.4.9). When the kernel function is the exponential kernel $\mathfrak{K}_{\mathtt{EXP}}(q, k) = \exp(k^\top q/\gamma)$ for any $q, k \in \mathbb{R}^{d_{\mathrm{p}}}$

and a fixed $\gamma > 0$, the attention mechanism in (3.2.1) takes the following form,

$$\texttt{attn}_{\texttt{SM}}(q, K, V) = V^\top \texttt{norm}_{\texttt{SM}}\big(\mathfrak{K}_{\texttt{EXP}}(K, q)\big) = V^\top \texttt{softmax}(Kq/\gamma), \tag{3.4.10}$$

which recovers the softmax attention in Vaswani et al. (2017). In what follows, we prove that as the sequence length L goes to infinity, the softmax attention $\texttt{attn}_{\texttt{SM}}$ has the same limit as the CME attention $\texttt{attn}_{\texttt{CME}}$.

Softmax Attention Has the Same Limit as CME Attention with $L \to \infty$. We demonstrate that the softmax attention in (3.4.10) is a conditional kernel density estimator of $\mathbb{P}_{\mathcal{V}\,|\,\mathcal{K}}$. We define the conditional kernel density estimator (KDE) as follows,

$$\widehat{\mathbb{P}}^{\mathfrak{K}}_{\mathcal{V}\,|\,\mathcal{K}}(v\,|\,q) = \iota \cdot \frac{\sum_{\ell=1}^{L} \mathfrak{K}(k^\ell, q) \cdot \mathfrak{K}(v^\ell, v)}{\sum_{\ell=1}^{L} \mathfrak{K}(k^\ell, q)}, \tag{3.4.11}$$

where $\iota > 0$ is the normalization factor such that $\int \widehat{\mathbb{P}}^{\mathfrak{K}}_{\mathcal{V}\,|\,\mathcal{K}}(v\,|\,q)\mathrm{d}v = 1$. We remark that although the definition of the KDE in (3.4.11) involves the kernel function $\mathfrak{K}(\cdot,\cdot)$, it is not associated with any RKHS. A common choice of the kernel function is the Gaussian RBF kernel $\mathfrak{K}_{\texttt{RBF}}(q, k) = \exp(-\|q-k\|_2^2/2\gamma)$. In what follows, we normalize the query q, the key k, and the value v so that $q, k \in \mathbb{S}^{d_{\mathrm{p}}-1}$ and $v \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d_{\mathrm{p}}-1}$ and $\mathbb{S}^{d-1}$ are the $(d_{\mathrm{p}}-1)$-dimensional and $(d-1)$-dimensional unit spheres, respectively. On the unit sphere, the exponential kernel is equivalent to the Gaussian RBF kernel. Specifically, it holds for a given rescaling $\gamma > 0$ that $\mathfrak{K}_{\texttt{EXP}}(q, k) = \exp(q^\top k/\gamma) = C \cdot \exp(-\|q-k\|_2^2/2\gamma) = C \cdot \mathfrak{K}_{\texttt{RBF}}(q, k)$ for any $q, k \in \mathbb{R}^{d_{\mathrm{p}}}$, where $C > 0$ is an absolute constant. Moreover, when we use the exponential kernel in (3.4.11), the value of ι does not depend on q. To see this, note that $\int_{\mathbb{S}^{d-1}} \mathfrak{K}_{\texttt{EXP}}(v^1, v)\mathrm{d}v = \int_{\mathbb{S}^{d-1}} \mathfrak{K}_{\texttt{EXP}}(v^2, v)\mathrm{d}v$ for any $v^1, v^2 \in \mathbb{S}^{d-1}$ due to the symmetry.

The following proposition proves that the attention mechanism in (3.4.10) outputs the conditional kernel density estimator in (3.4.11) and has the same limit as the CME attention as $L \to \infty$.

Proposition 3.4.2 (Softmax Attention Converges to Kernel Conditional Mean Embedding)**.** Recall that the softmax attention is defined in (3.4.10). It holds for any $q \in \mathbb{S}^{d_{\mathrm{p}}-1}$ that

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

where $C > 0$ is an absolute constant. Meanwhile, under the condition that $\widehat{\mathbb{P}}^{\mathfrak{K}}_{\mathcal{V} \,|\, \mathcal{K}}(v \,|\, k) \to \mathbb{P}_{\mathcal{V} \,|\, \mathcal{K}}(v \,|\, k)$ uniformly for any k as $L \to \infty$, it holds for $L \to \infty$ that

$$\mathtt{attn}_{\mathtt{SM}}(q, K, V) \to C \cdot \mathbb{E}[\mathcal{V} \,|\, \mathcal{K} = q].$$

Proof. See §C.2.5 for a detailed proof. □

We remark that the uniform convergence $\widehat{\mathbb{P}}^{\mathfrak{K}}_{\mathcal{V} \,|\, \mathcal{K}}(v \,|\, k) \to \mathbb{P}_{\mathcal{V} \,|\, \mathcal{K}}(v \,|\, k)$ holds when the density of $\mathbb{P}_{\mathcal{K}}$ is bounded from below (De Gooijer and Zerom, 2003). As shown in Propositions 3.4.1 and 3.4.2, the softmax attention $\mathtt{attn}_{\mathtt{SM}}$ and the CME attention $\mathtt{attn}_{\mathtt{CME}}$ have the same limit as $L \to \infty$. Since the CME attention captures the latent posterior, which is proved in §3.4.1, we conclude that the softmax attention also captures the latent posterior approximately. Moreover, in terms of the limiting expectation $\mathbb{E}[\mathcal{V} \,|\, \mathcal{K} = q]$, we highlight that it implies the necessity of using the multiple heads and connects the attention mechanism with causal inference. See §C.2.2 for a detailed discussion.

$$\begin{tikzcd}
& \mathbb{E}[\mathscr{V} \mid \mathscr{K} = q] & \\
\mathtt{attn}_{\mathtt{CME}} \arrow[ur, "\textbf{Proposition 4.1}"] \arrow[rr, leftrightarrow, dashed] & & \mathtt{attn}_{\mathtt{SM}} \arrow[ul, "\textbf{Proposition 4.2}"']
\end{tikzcd}$$

Figure 3.5. As shown in Propositions 3.4.1 and 3.4.2, the softmax attention $\mathtt{attn}_{\mathtt{SM}}$ and the CME attention $\mathtt{attn}_{\mathtt{CME}}$ have the same limit $\mathbb{E}[\mathcal{V} \,|\, \mathcal{K} = q]$ as $L \to \infty$.

3.5. Excess Risk Analysis

To demonstrate the theoretical benefit of incorporating latent posterior inference into the transformer architecture, we present a compact version of excess risk analysis for one-layer single-head softmax attention neural networks without skip connections. See §C.3 for a detailed analysis of the complete setup of the transformer architecture.

Attention Neural Network. We specify the feedforward neural network in (3.4.7) as $\mathtt{nn}(x; A) = \mathtt{ReLU}(Ax)$[1], where $\mathtt{ReLU}(\cdot)$ is the rectified linear unit (ReLU) activation that operates elementwise. In the rest of the paper, we consider the attention neural networks with a final aggregation layer to allow for the proper scaling of the outputs in the supervised setting and the transfer capability to diverse downstream tasks in the self-supervised setting, which is discussed in §3.6. We define the following function class of attention neural networks,

$$(3.5.1)\quad \mathcal{F}_{\mathtt{attn}} = \Big\{ \mathtt{agg}_{\theta_0} \circ \mathtt{attn}_{\mathtt{SM}}\big(q_\theta(\mathtt{msk}), k_\theta(X), v_\theta(X)\big) : \theta = (\theta_0, A, W^{\mathrm{q}}, W^{\mathrm{k}}, W^{\mathrm{v}}) \in \Theta \Big\},$$

where $\mathtt{agg}_{\theta_0} : \mathbb{R}^d \to \mathbb{R}^{d_{\mathrm{y}}}$ is the aggregation layer parameterized by θ_0 and $\mathtt{attn}_{\mathtt{SM}}$ is the softmax attention defined in (3.4.10) with the learnable parameters $(A, W^{\mathrm{q}}, W^{\mathrm{k}}, W^{\mathrm{v}})$

[1] Here, for ease of presentation, we consider feedforward neural networks without bias terms.

defined in (3.4.7). To characterize the excess risk, we specify the parameter space as follows, which grants $\mathcal{F}_{\texttt{attn}}$ a finite capacity.

Assumption 3.5.1 (Parameter Space)**.** We assume for all $\theta = (\theta_0, A, W^{\mathrm{q}}, W^{\mathrm{k}}, W^{\mathrm{v}}) \in \Theta$ that

$$\|W^{\mathrm{q}}\|_2 \leq \omega^{\mathrm{q}}, \qquad \|W^{\mathrm{k}}\|_2 \leq \omega^{\mathrm{k}}, \qquad \|W^{\mathrm{v}}\|_2 \leq \omega^{\mathrm{v}}, \qquad \|A\|_2 \leq \alpha^{\mathtt{nn}},$$
$$\|W^{\mathrm{q}}\|_{\mathrm{F}} \leq R^{\mathrm{q}}, \qquad \|W^{\mathrm{k}}\|_{\mathrm{F}} \leq R^{\mathrm{k}}, \qquad \|W^{\mathrm{v}}\|_{\mathrm{F}} \leq R^{\mathrm{v}}, \qquad \|A\|_{\mathrm{F}} \leq R^{\mathtt{nn}},$$

where $\omega^{\mathrm{q}}, \omega^{\mathrm{k}}, \omega^{\mathrm{v}}, \alpha^{\mathtt{nn}}, R^{\mathrm{q}}, R^{\mathrm{k}}, R^{\mathrm{v}}, R^{\mathtt{nn}} > 0$.

Excess Risk. Following (3.4.8), we consider the learning objective $\mathcal{L}((X, y), f) = \|y - f(X)\|_2^2$ for $f \in \mathcal{F}_{\texttt{attn}}$, where y is the target variable. We make the following assumption on the training dataset.

Assumption 3.5.2 (Data Distribution)**.** We assume that the training dataset $\mathcal{D}_n = \{(X_i, y_i)\}_{i\in[n]}$ is independently and identically drawn from the data distribution $\mathcal{D}$, which is supported on the product space $\mathfrak{X}^L \times \mathfrak{Y}$, where

$$\mathfrak{X}^L = \big\{X \in \mathbb{R}^{L\times d} : \max_{\ell\in[L]}\|x^\ell\|_2 \leq R\big\}, \qquad \mathfrak{Y} = \big\{y \in \mathbb{R}^{d_{\mathrm{y}}} : \|y\|_2 \leq 1/2\big\}. \tag{3.5.2}$$

We consider the excess risk $\mathcal{E} = \mathbb{E}[\mathcal{L}((X, y), \widehat{f})] - \mathbb{E}[\mathcal{L}((X, y), f^*)]$, where $\mathbb{E}[\,\cdot\,]$ is the population expectation over the data distribution $\mathcal{D}$. Here $\widehat{f} \in \mathcal{F}_{\texttt{attn}}$ is the attention neural network obtained from minimizing the empirical risk $\widehat{\mathbb{E}}[\mathcal{L}((X, y), f)]$ in the training process, where $\widehat{\mathbb{E}}[\,\cdot\,]$ is the empirical expectation over the training dataset $\mathcal{D}_n$. Here $f^*(X) = \mathbb{E}[y \,|\, X]$ is the regression function that we aim to approximate. In other words, f^* is the optimal model that minimizes the population risk $\mathbb{E}[\mathcal{L}((X, y), f)]$.

To analyze the excess risk $\mathcal{E}$, we decompose it into three terms,

$$\begin{aligned}
\mathcal{E} &= \underbrace{\mathbb{E}\Big[\mathcal{L}\big((X,y),\widehat{f}\big)\Big] - \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),\widehat{f}\big)\Big] + \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),\widetilde{f}\big)\Big] - \min_{f\in\mathcal{F}_{\texttt{attn}}} \mathbb{E}\Big[\mathcal{L}\big((X,y),f\big)\Big]}_{\mathcal{E}_{\text{gen}}\text{: Generalization Error}} \\
&\quad + \underbrace{\min_{f\in\mathcal{F}_{\texttt{attn}}} \mathbb{E}\Big[\mathcal{L}\big((X,y),f\big)\Big] - \mathbb{E}\Big[\mathcal{L}\big((X,y),f^*\big)\Big]}_{\mathcal{E}_{\text{approx}}\text{: Approximation Error}} + \underbrace{\widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),\widehat{f}\big)\Big] - \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),\widetilde{f}\big)\Big]}_{\mathcal{E}_{\text{opt}}\text{: Optimization Error}}.
\end{aligned} \tag{3.5.3}$$

where $\widetilde{f} = \mathrm{argmin}_{f\in\mathcal{F}_{\texttt{attn}}} \widehat{\mathbb{E}}[\mathcal{L}((X,y),f)]$ is the attention neural network that minimizes the empirical risk over $\mathcal{F}_{\texttt{attn}}$.

In §3.5.1-3.5.3, we analyze the three terms on the right-hand side of (3.5.3) in the supervised setting. In §3.6, we extend the following analysis of the approximation error $\mathcal{E}_{\text{approx}}$ to the self-supervised setting.

3.5.1. Generalization Error Analysis

Recall that the softmax attention in (3.4.10) is instantiated via the exponential kernel, which is equivalent to the Gaussian RBF kernel when q and k are on the unit sphere $\mathbb{S}^{d_{\mathrm{p}}-1}$. Also, note that vector ℓ_2-norm scales with the dimension d_{p} at the rate of $\sqrt{d_{\mathrm{p}}}$. In the rest of the paper, we consider the Gaussian RBF kernel with inputs rescaled by $1/\sqrt{d_{\mathrm{p}}}$, i.e.,

$$\mathfrak{K}_{\texttt{RBF}}(q,k) = \exp\big(-\|q/\sqrt{d_{\mathrm{p}}} - k/\sqrt{d_{\mathrm{p}}}\|_2^2/2\big) = \exp\big(-\|q-k\|_2^2/2d_{\mathrm{p}}\big). \tag{3.5.4}$$

Under Assumption 3.5.1, we define

$$\gamma = \max\{\alpha^{\mathrm{nn}}, \omega^{\mathrm{v}}\}, \qquad \kappa = \max\bigg\{\frac{R^{\mathrm{nn}}}{\alpha^{\mathrm{nn}}}, \frac{R^{\mathrm{v}}}{\omega^{\mathrm{v}}}, \frac{R^{\mathrm{k}}+R^{\mathrm{q}}}{(\omega^{\mathrm{q}}+\omega^{\mathrm{k}})\cdot\omega^{\mathrm{v}}}\bigg\}, \qquad \zeta = \frac{(\omega^{\mathrm{q}}+\omega^{\mathrm{k}})^2\cdot R^{\mathrm{v}}}{\omega^{\mathrm{v}}}.$$

Recall that $\mathcal{F}_{\texttt{attn}}$ is the family of attention neural networks defined in (3.5.1). Let $\texttt{agg}_{\theta_0,j}$ be the j-th entry of the aggregation layer $\texttt{agg}_{\theta_0}$ with $j \in [d_{\mathrm{y}}]$. We provide the following characterization of the generalization error $\mathcal{E}_{\mathrm{gen}}$.

Theorem 3.5.3 (Generalization Error)**.** Let $D = \max\{d, d_{\mathrm{p}}, d_{\mathrm{y}}\}$. Suppose that Assumptions 3.5.1-3.5.2 hold. We assume that $\texttt{agg}_{\theta_0}$ has the output range within $\mathfrak{Y}$ and $\texttt{agg}_{\theta_0,j}$ is 1-Lipschitz with respect to the $\|\cdot\|_{\mathrm{F}}$-norm for all $j \in [d_{\mathrm{y}}]$. Then, for any $\delta > 0$, it holds with probability at least $1 - \delta$ that

$$\mathcal{E}_{\mathrm{gen}} = O\biggl(\frac{D^2}{\sqrt{n}} \cdot \bigl[\sqrt{\log(1+\gamma)} + \sqrt{\log(1+\zeta R)} + \sqrt{\log(1+\kappa/\zeta)}\bigr] + \sqrt{\frac{\log(1/\delta)}{n}}\biggr).$$

Proof. See §C.3 for a detailed proof. □

An important implication of Theorem 3.5.3 is that the generalization error for attention neural networks does not degrade as the sequence length L goes to infinity. It is also worth mentioning that the constants α, ω,$R^{\texttt{nn}}$, $R^{\texttt{attn}}$, and R play crucial roles in the theoretical analysis of the generalization error and justify the architecture design of the original transformer. In specific, we observe that (i) skip connections help reducing α, ω, $R^{\texttt{nn}}$, and $R^{\texttt{attn}}$, and (ii) layer normalizations help reducing R when there is multilayer composition of many attention mechanisms. See §C.3 for a more involved analysis of the generalization error of the complete setup of the transformer architecture and the related discussion.

3.5.2. Approximation Error Analysis

In what follows, we characterize the approximation error in the supervised setting.

Approximation Target. We aim to approximate the regression function $f^*(X) = \mathbb{E}[y \,|\, X]$ with the attention neural network $f_\theta \in \mathcal{F}_{\texttt{attn}}$, which is defined in (3.5.1). The regression function $f^*(X)$ is the optimal model in the sense that it minimizes the population risk $\mathbb{E}[\mathcal{L}((X, y), f)]$. By Lemma 3.3.2, when the input mask $\texttt{msk}$ and the latent variable z are given, the target variable y is independent of the input sequence X. Thus, the regression function $f^*(X)$ can be decomposed as follows,

$$\begin{aligned} f^*(X) &= \mathbb{E}[y \,|\, X] = \mathbb{E}_{z \,|\, X}\big[\mathbb{E}[y \,|\, \texttt{msk}, z]\big] \\ &= \int \underbrace{\mathbb{E}[y \,|\, \texttt{msk}, z]}_{g^*(z;\texttt{msk})\,:\,\text{latent-to-target}} \cdot \mathbb{P}(z \,|\, X)\mathrm{d}z. \end{aligned} \tag{3.5.5}$$

Here the latent-to-target mapping $g^*(z;\texttt{msk})$ can be viewed as a decoding function, which maps the latent variable z to the target variable y given the input mask $\texttt{msk}$. On the other hand, the latent posterior $\mathbb{P}(z \,|\, X)$ encodes the input sequence X into the latent variable z. We note that the input mask $\texttt{msk}$ describes the prediction task and is fixed throughout. For example, the input mask $\texttt{msk}$ corresponds to the class encoding in the supervised setting or the positional encoding in the self-supervised setting.

From (3.5.5), we see that approximating the regression function $f^*(X)$ involves capturing (i) the latent posterior $\mathbb{P}(z \,|\, X)$ and (ii) the latent-to-target mapping $g^*(z;\texttt{msk})$. Corresponding to (i), the latent variable z summarizes the "concept" of the input sequence X, while corresponding to (ii), the target variable y and the input mask $\texttt{msk}$ specify the prediction task. In the following, we demonstrate the central role of the latent-to-target mapping $g^*(z;\texttt{msk})$, which attention neural networks aim to approximate.

Approximation Surrogate. We define the reweighted CME attention

$$f_W^\dagger(X;\texttt{msk}) = W^\top \texttt{attn}_{\texttt{CME}}\big(q_*(\texttt{msk}), k_*(X), v_*(X)\big) \tag{3.5.6}$$

as a surrogate function for approximating the regression function $f^*(X)$ in (3.5.5). Here the reweighting parameter $W \in \mathbb{R}^{d\times d_\mathrm{y}}$ satisfies $\|W\|_\mathrm{F} < \infty$. In the sequel, we demonstrate that the latent-to-target function contained in $f_W^\dagger(X;\texttt{msk})$ approximates the latent-to-target mapping $g^*(z;\texttt{msk})$, which is a key component of the regression function $f^*(X)$. By (3.4.3) and (3.4.5), we have

$$\begin{aligned} W^\top \texttt{attn}_{\texttt{CME}}\big(q_*(\texttt{msk}), k_*(X), v_*(X)\big) &= W^\top \mathbb{E}[v^{\texttt{msk}} \,|\, \texttt{msk}, X] \\ &= W^\top \int \underbrace{\mathbb{E}[v^{\texttt{msk}} \,|\, \texttt{msk}, z]}_{\psi(z;\texttt{msk})\text{: latent-to-value}} \cdot \mathbb{P}(z \,|\, X)\mathrm{d}z, \end{aligned} \tag{3.5.7}$$

where $q_*(\texttt{msk})$ and $v^{\texttt{msk}}$ replace $c_*^{\texttt{m}}(\texttt{msk})$ and $r^{\texttt{msk}}$ in (3.4.2), respectively. Taking (3.5.7) into (3.5.6), we obtain

$$f_W^\dagger(X;\texttt{msk}) = \int \underbrace{W^\top \psi(z;\texttt{msk})}_{g_W^\dagger(X;\texttt{msk})} \cdot \mathbb{P}(z \,|\, X)\mathrm{d}z = \mathbb{E}_{z\,|\,X}\big[g_W^\dagger(z;\texttt{msk})\big], \tag{3.5.8}$$

where $g_W^\dagger(z;\texttt{msk})$ is a latent-to-target function parameterized by $W \in \mathbb{R}^{d\times d_\mathrm{y}}$.

Following the infinite-dimensional counterpart of (3.3.7), the reweighted CME attention captures the latent posterior $\mathbb{P}(z \,|\, X)$ under the latent variable model in (3.4.1), where the latent prior is Gaussian. Comparing (3.5.5) and (3.5.8), we see that the reweighted CME attention $f_W^\dagger(X;\texttt{msk})$ performs the latent-to-target decoding via $g_W^\dagger(z;\texttt{msk})$, which plays the same role as the latent-to-target mapping $g^*(z;\texttt{msk})$. Thus, it remains to characterize the expressity of the function class

$$\mathcal{G}^{\dagger} = \big\{ g_W^{\dagger}(z;\mathtt{msk}) = W^\top \psi(z;\mathtt{msk}) : W \in \mathbb{R}^{d \times d_{\mathrm{y}}}, \|W\|_{\mathrm{F}} < \infty \big\} \tag{3.5.9}$$

in terms of approximating the latent-to-target mapping $g^*(z;\mathtt{msk})$ in (3.5.5).

To characterize the function class $\mathcal{G}^{\dagger}$ defined in (3.5.9), we define the function class

$$\mathcal{G}_i^{\dagger} = \{ g_{W,i}^{\dagger}(z;\mathtt{msk}) = w_i^\top \psi(z;\mathtt{msk}) : w_i \in \mathbb{R}^d, \|w_i\|_2 < \infty \}, \tag{3.5.10}$$

which is formed by the i-th entry of the latent-to-target function $g_W^{\dagger}(z;\mathtt{msk}) \in \mathcal{G}^{\dagger}$. Here $i \in [d_{\mathrm{y}}]$ and $W = [w_1, \ldots, w_{d_{\mathrm{y}}}]^\top$. Note that the function class $\mathcal{G}_i^{\dagger}$ is the RKHS $\mathcal{H}_{\mathtt{LTV}}$ induced by the kernel function $\mathfrak{K}_{\mathtt{LTV}}(z, z';\mathtt{msk}) = \psi(z;\mathtt{msk})^\top \psi(z';\mathtt{msk})$, which is a reproducing kernel. Here the latent-to-value ($\mathtt{LTV}$) mapping $\psi(z;\mathtt{msk})$ is defined in (3.5.7). See §C.5.1 for a detailed discussion. See Figure 3.6 for a visualization of the construction of $\mathcal{H}_{\mathtt{LTV}}$.

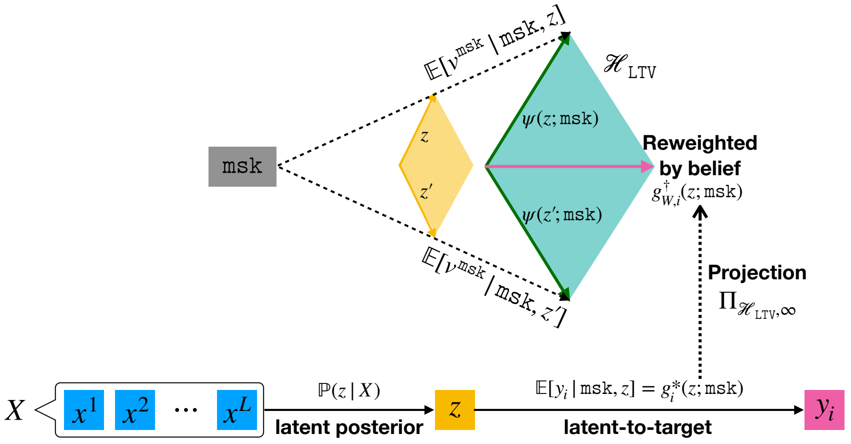

$\mathbb{E}[\nu^{\mathtt{msk}} \,|\, \mathtt{msk}, z]$
$\mathscr{H}_{\mathtt{LTV}}$
$\mathtt{msk}$
z
z'
$\psi(z;\mathtt{msk})$
$\psi(z';\mathtt{msk})$
Reweighted by belief
$g_{W,i}^{\dagger}(z;\mathtt{msk})$
$\mathbb{E}[\nu^{\mathtt{msk}} \,|\, \mathtt{msk}, z']$
Projection
$\Pi_{\mathscr{H}_{\mathtt{LTV}},\infty}$
X
x^1 x^2 ⋯ x^L
$\mathbb{P}(z \,|\, X)$
latent posterior
z
$\mathbb{E}[y_i \,|\, \mathtt{msk}, z] = g_i^*(z;\mathtt{msk})$
latent-to-target
y_i

Figure 3.6. The RKHS $\mathcal{H}_{\mathtt{LTV}}$ induced by the latent-to-value mapping $\psi(z;\mathtt{msk})$. The input mask $\mathtt{msk}$ describes the prediction task and determines the RKHS $\mathcal{H}_{\mathtt{LTV}}$.

Therefore, the reweighted CME attention $f_W^\dagger(X;\texttt{msk})$ in (3.5.5) aims to capture the i-th entry $g_i^*(z;\texttt{msk})$ of the latent-to-target mapping $g^*(z;\texttt{msk})$ within the RKHS $\mathcal{H}_{\texttt{LTV}}$. To this end, we make the following assumption on the fundamental hardness of the recovery task.

Assumption 3.5.4 (Recovery Gap)**.** For any fixed input mask $\texttt{msk}$, let

$$g_{W,i}^\dagger(\cdot;\texttt{msk}) = \Pi_{\mathcal{H}_{\texttt{LTV}},\infty}\big(g_i^*(\cdot;\texttt{msk})\big) = \operatorname*{argmin}_{g_i(\cdot;\texttt{msk})\in\mathcal{H}_{\texttt{LTV}}} \big\|g_i^*(\cdot;\texttt{msk}) - g_i(\cdot;\texttt{msk})\big\|_\infty$$

be the ℓ_∞-norm projection of the i-th entry $g_i^*(z;\texttt{msk})$ of the latent-to-target mapping $g^*(z;\texttt{msk})$ onto the RKHS $\mathcal{H}_{\texttt{LTV}}$. We assume that there exists $\epsilon_g(\texttt{msk}) \in [0,+\infty)$ such that

$$\sum_{i=1}^{d_{\mathrm{y}}} \big\|g_i^*(\cdot;\texttt{msk}) - g_{W,i}^\dagger(\cdot;\texttt{msk})\big\|_\infty^2 \le \epsilon_g^2(\texttt{msk}).$$

Here the ℓ_∞-norm is taken over the latent variable z.

Recall that the function class of attention neural networks $\mathcal{F}_{\texttt{attn}}$ is defined in (3.5.1). We have the following theorem characterizing the approximation error $\mathcal{E}_{\text{approx}}$ defined in (3.5.3).

Theorem 3.5.5 (Approximation Error)**.** Let $\{g_{W,i}^\dagger(z;\texttt{msk}) = w_i^\top\psi(z;\texttt{msk})\}_{i\in[d_{\mathrm{y}}]}$ be a function class satisfying Assumption 3.5.4. We define $W = [w_1^\top,\cdots,w_{d_{\mathrm{y}}}^\top]^\top$. Suppose that there exists $f_\theta \in \mathcal{F}_{\texttt{attn}}$ and $\epsilon_{\texttt{attn}} \in [0,+\infty)$ such that

$$\sup_{X\in\mathfrak{X}^L} \Big\|f_\theta(X;\texttt{msk}) - W^\top \texttt{attn}_{\texttt{CME}}\big(q_*(\texttt{msk}), k_*(X), v_*(X)\big)\Big\|_2 \le \epsilon_{\texttt{attn}}, \tag{3.5.11}$$

where $\mathfrak{X}^L$ is defined in (3.5.2). Then, we have

$$\mathcal{E}_{\text{approx}} \le 2\epsilon_g^2(\texttt{msk}) + 2\epsilon_{\texttt{attn}}^2.$$

Proof. See §C.5.2 for a detailed proof. □

The approximation error bound in Theorem 3.5.5 involves the recovery gap $\epsilon_g(\texttt{msk})$ and the surrogate approximation error $\epsilon_{\texttt{attn}}$. Since the latent posterior $\mathbb{P}(z \mid X)$ is captured by the reweighted CME attention, the recovery gap $\epsilon_g(\texttt{msk})$ between the function class $\mathcal{G}^\dagger$ in (3.5.9) and the latent-to-target mapping $g^*(z; \texttt{msk})$ in (3.5.5) plays the central role in the approximation error bound. On the other hand, the approximation error $\epsilon_{\texttt{attn}}$ between attention neural networks in $\mathcal{F}_{\texttt{attn}}$ and the reweighted CME attention is characterized in Proposition 3.4.2.

3.5.3. Optimization Error Analysis

Since the learning objective of attention neural networks is nonconvex with respect to the parameter θ, we consider the property of the stationary points. Let $\widehat{\theta} = (\widehat{\theta}_0, \widehat{A}, \widehat{W}^{\text{q}}, \widehat{W}^{\text{k}}, \widehat{W}^{\text{v}})$ be the stationary point of the empirical risk $\widehat{\mathbb{E}}[\mathcal{L}((X, y), f)]$, that is,

$$\Big\langle \nabla_\theta \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X, y), f_{\widehat{\theta}}\big)\Big], \theta - \widehat{\theta} \Big\rangle \ge 0, \quad \forall \theta \in \Theta, \tag{3.5.12}$$

which is the learnable parameter obtained in the training process, i.e., $\widehat{f} = f_{\widehat{\theta}}$. Recall that the regression function $f^*(X) = \mathbb{E}[y \mid X]$ is the minimizer of the population risk $\mathbb{E}[\mathcal{L}((X, y), f)]$. We have the following proposition characterizing the optimization error $\mathcal{E}_{\text{opt}}$, which is defined in (3.5.3).

Proposition 3.5.6 (Optimization Error)**.** Suppose that Assumption 3.5.2 holds. Then, it holds that

$$\mathcal{E}_{\mathrm{opt}} \leq 2 \cdot \min_{\theta \in \Theta} \widehat{\mathbb{E}}\Big[\big\| f_{\widehat{\theta}}(X) + \nabla_\theta f_{\widehat{\theta}}(X)^\top (\theta - \widehat{\theta}) - f^*(X) \big\|_2\Big]. \tag{3.5.13}$$

Proof. See §C.4 for a detailed proof. □

The right-hand side of (3.5.13) quantifies the expressity of the function class defined by the local linearization,

$$\big\{ f_{\widehat{\theta}}(X) + \nabla_\theta f_{\widehat{\theta}}(X)^\top (\theta - \widehat{\theta}) : \theta \in \Theta \big\}.$$

In the neural tangent kernel (NTK) regime (Yang, 2020; Yang and Littwin, 2021; Jacot et al., 2018), it is known that,

$$f^*(X) = f_{\widehat{\theta}}(X) + \nabla_\theta f_{\widehat{\theta}}(X)^\top (\theta - \widehat{\theta}) + o(1), \quad \forall X \in \mathbb{R}^{L \times d},$$

where the $o(1)$ error captures the local linearization error in the NTK-based analysis. As a consequence, the optimization error satisfies $\mathcal{E}_{\mathrm{opt}} = o(1)$, that is, the stationary point $\widehat{\theta}$ is (approximately) global optimal. Such a result shows the theoretical benefit of incorporating feedforward neural networks in the architecture design. While NTK-based analysis involves a random initialization in the supervised setting, Malladi et al. (2022) provide an NTK-based analysis for the downstream training of the transformer with a pretrained initialization in the self-supervised setting.

3.6. From Supervised Learning to Self-Supervised Learning

An important aspect of the attention mechanism is that one can obtain a sequence embedding by pretraining in a self-supervised manner, which gives rise to the transfer capability for diverse downstream tasks.

Self-Supervised Learning. The attention mechanism enables embedding learning and downstream prediction via the self-supervised learning (SSL) process as follows.

(PT) Pretraining process: We train an attention neural network $\widehat{f}_{\texttt{PT}}(\overline{X};\texttt{msk}_{\texttt{PT}}) = f_{\widehat{\theta}_{\texttt{PT}}}(\overline{X};\texttt{msk}_{\texttt{PT}}) \in \mathcal{F}_{\texttt{PT}}$ with the learned parameter $\widehat{\theta}_{\texttt{PT}}$ to predict the masked token $x^L \in \mathbb{R}^d$, which is denoted by $y_{\texttt{PT}}$, from the truncated input sequence $\overline{X} = \{x^\ell\}_{\ell\in[L-1]}$ and the input mask $\texttt{msk}_{\texttt{PT}}$. Here the function class of attention neural networks for the pretraining process is defined as follows,

$$\mathcal{F}_{\texttt{PT}} = \Big\{\texttt{agg}_\theta^{\texttt{PT}} \circ \texttt{attn}_{\texttt{SM}}\big(q_\theta(\texttt{msk}_{\texttt{PT}}), k_\theta(\overline{X}), v_\theta(\overline{X})\big) : \theta \in \Theta_{\texttt{PT}}\Big\}, \tag{3.6.1}$$

where $\texttt{agg}_\theta^{\texttt{PT}} : \mathbb{R}^d \to \mathbb{R}^d$ is the aggregation layer. For the pretraining process, the input mask $\texttt{msk}_{\texttt{PT}}$ is the positional encoding of the masked token x^L.

(DS) Downstream task: We freeze the learned parameter $\widehat{\theta}_{\texttt{PT}}$ and train another attention neural network $\widehat{f}_{\texttt{DS}}(\overline{X};\texttt{msk}_{\texttt{DS}}) = f_{\widehat{\theta}_{\texttt{DS}}}(\overline{X};\texttt{msk}_{\texttt{DS}}) \in \mathcal{F}_{\texttt{DS}}$ with the learned parameter $\widehat{\theta}_{\texttt{DS}}$ to predict another target variable $y_{\texttt{DS}} \in \mathbb{R}^{d_{\rm y}}$ from the truncated input sequence $\overline{X} = \{x^\ell\}_{\ell\in[L-1]}$ and another input mask $\texttt{msk}_{\texttt{DS}}$. Here the function class of attention neural networks for the downstream task is defined as follows,

$$\mathcal{F}_{\texttt{DS}} = \Big\{\texttt{agg}_\theta^{\texttt{DS}} \circ \texttt{attn}_{\texttt{SM}}\big(q_{\widehat{\theta}_{\texttt{PT}}}(\texttt{msk}_{\texttt{DS}}), k_{\widehat{\theta}_{\texttt{PT}}}(\overline{X}), v_{\widehat{\theta}_{\texttt{PT}}}(\overline{X})\big) : \theta \in \Theta_{\texttt{DS}}\Big\}, \tag{3.6.2}$$

which means that the aggregation layer $\mathtt{agg}^{\mathtt{DS}}_{\widehat{\theta}_{\mathtt{DS}}} : \mathbb{R}^d \to \mathbb{R}^{d_y}$ replaces the aggregation layer $\mathtt{agg}^{\mathtt{PT}}_{\widehat{\theta}_{\mathtt{PT}}} : \mathbb{R}^d \to \mathbb{R}^d$ obtained in the pretraining process. For the downstream task, the input mask $\mathtt{msk}_{\mathtt{DS}}$ is the class encoding of the target variable $y_{\mathtt{DS}}$.

With the full input sequence X replaced by the truncated input sequence $\overline{X}$, the attention neural network $\widehat{f}_{\mathtt{DS}}(\overline{X}; \mathtt{msk})$ obtained in the SSL process has the same decomposition of the excess risk as that in (3.5.3). In the risk decomposition, for the SSL process, we have the same characterization of the generalization error and the optimization error as those in the supervised setting. When the downstream task is trained using the same set of truncated input sequences as that in the pretraining process, our previous analysis of the generalization error in the supervised setting is applicable to the SSL process. On the other hand, when the downstream task is trained using an independent set of truncated input sequences, we can modify our previous analysis to prove that the generalization error only scales with the complexity measure (e.g., the covering number) of the function class $\{\mathtt{agg}^{\mathtt{DS}}_{\theta} : \theta \in \Theta_{\mathtt{DS}}\}$ of aggregation layers without depending on that of the attention mechanism, as $\widehat{\theta}_{\mathtt{PT}}$ is frozen. Also, the attention neural network $\widehat{f}_{\mathtt{PT}}(\overline{X}; \mathtt{msk})$ obtained in the pretraining process has the same approximation error as that in the supervised setting. To characterize the approximation error for the SSL process, we analyze the approximation error for the downstream task by connecting it to the approximation error for the pretraining process.

Approximation Error. In parallel to the supervised setting, we define the regression function and the latent-to-target mapping for the pretraining process as follows,

$$f^*_{\mathtt{PT}}(\overline{X}) = \mathbb{E}[y_{\mathtt{PT}} \,|\, \overline{X}], \qquad g^*_{\mathtt{PT}}(z; \mathtt{msk}_{\mathtt{PT}}) = \mathbb{E}[y_{\mathtt{PT}} \,|\, \mathtt{msk}_{\mathtt{PT}}, z]. \tag{3.6.3}$$

Correspondingly, we defined the regression function and the latent-to-target mapping for the downstream task as follows,

$$f^*_{\mathtt{DS}}(\overline{X}) = \mathbb{E}[y_{\mathtt{DS}} \,|\, \overline{X}], \qquad g^*_{\mathtt{DS}}(z; \mathtt{msk}_{\mathtt{DS}}) = \mathbb{E}[y_{\mathtt{DS}} \,|\, \mathtt{msk}_{\mathtt{DS}}, z]. \tag{3.6.4}$$

In parallel to the reweighted CME attention defined in (3.5.6), we defined the surrogate functions for the pretraining process and the downstream task as follows,

$$\begin{aligned} f^\dagger_{W_{\mathtt{PT}}}(\overline{X}; \mathtt{msk}_{\mathtt{PT}}) &= W_{\mathtt{PT}}^\top \mathtt{attn}_{\mathtt{CME}}\big(q_*(\mathtt{msk}_{\mathtt{PT}}), k_*(\overline{X}), v_*(\overline{X})\big), \\ f^\dagger_{W_{\mathtt{DS}}}(\overline{X}; \mathtt{msk}_{\mathtt{DS}}) &= W_{\mathtt{DS}}^\top \mathtt{attn}_{\mathtt{CME}}\big(q_*(\mathtt{msk}_{\mathtt{DS}}), k_*(\overline{X}), v_*(\overline{X})\big), \end{aligned} \tag{3.6.5}$$

where $W_{\mathtt{PT}} \in \mathbb{R}^{d\times d}$ and $W_{\mathtt{DS}} \in \mathbb{R}^{d\times d_{\mathrm{y}}}$ are the reweighting parameters. We use the surrogate function to bridge the regression function and the attention neural network, which is illustrated in Figure 3.7. In parallel to the latent-to-value mapping $\psi(z; \mathtt{msk})$ defined in (3.5.7), we define the latent-to-value mappings for the pretraining process and the downstream task as follows,

$$\psi_{\mathtt{PT}}(z; \mathtt{msk}_{\mathtt{PT}}) = \mathbb{E}[v^{\mathtt{msk}_{\mathtt{PT}}} \,|\, \mathtt{msk}_{\mathtt{PT}}, z], \qquad \psi_{\mathtt{DS}}(z; \mathtt{msk}_{\mathtt{DS}}) = \mathbb{E}[v^{\mathtt{msk}_{\mathtt{DS}}} \,|\, \mathtt{msk}_{\mathtt{DS}}, z],$$

where $v^{\mathtt{msk}_{\mathtt{PT}}}$ and $v^{\mathtt{msk}_{\mathtt{DS}}}$ replace $r^{\mathtt{msk}}$ in (3.4.2). The latent-to-value mappings induce the kernel functions as follows,

$$\begin{aligned} \mathfrak{K}_{\mathtt{PT}}(z, z'; \mathtt{msk}_{\mathtt{PT}}) &= \psi_{\mathtt{PT}}(z; \mathtt{msk}_{\mathtt{PT}})^\top \psi_{\mathtt{PT}}(z'; \mathtt{msk}_{\mathtt{PT}}), \\ \mathfrak{K}_{\mathtt{DS}}(z, z'; \mathtt{msk}_{\mathtt{DS}}) &= \psi_{\mathtt{DS}}(z; \mathtt{msk}_{\mathtt{DS}})^\top \psi_{\mathtt{DS}}(z'; \mathtt{msk}_{\mathtt{DS}}), \end{aligned}$$

which induce the RKHSs $\mathcal{H}_{\texttt{PT}}$ and $\mathcal{H}_{\texttt{DS}}$. Corresponding to (3.5.8), we have

$$\begin{aligned} f^{\dagger}_{W_{\texttt{PT}}}(\overline{X};\texttt{msk}_{\texttt{PT}}) &= \int \underbrace{W_{\texttt{PT}}^{\top}\psi_{\texttt{PT}}(z;\texttt{msk}_{\texttt{PT}})}_{g^{\dagger}_{W_{\texttt{PT}}}(\overline{X};\texttt{msk}_{\texttt{PT}})} \cdot \mathbb{P}(z \,|\, \overline{X})\mathrm{d}z = \mathbb{E}_{z\,|\,\overline{X}}\big[g^{\dagger}_{W_{\texttt{PT}}}(z;\texttt{msk}_{\texttt{PT}})\big], \\ f^{\dagger}_{W_{\texttt{DS}}}(\overline{X};\texttt{msk}_{\texttt{DS}}) &= \int \underbrace{W_{\texttt{DS}}^{\top}\psi_{\texttt{DS}}(z;\texttt{msk}_{\texttt{DS}})}_{g^{\dagger}_{W_{\texttt{DS}}}(\overline{X};\texttt{msk}_{\texttt{DS}})} \cdot \mathbb{P}(z \,|\, \overline{X})\mathrm{d}z = \mathbb{E}_{z\,|\,\overline{X}}\big[g^{\dagger}_{W_{\texttt{DS}}}(z;\texttt{msk}_{\texttt{DS}})\big]. \end{aligned} \tag{3.6.6}$$

Note that $f^{\dagger}_{W_{\texttt{PT}}}(\overline{X};\texttt{msk}_{\texttt{PT}})$ and $f^{\dagger}_{W_{\texttt{DS}}}(\overline{X};\texttt{msk}_{\texttt{DS}})$ share the same latent posterior since the attention mechanism is frozen for the downstream task. By our previous arguments following (3.5.8)-(3.5.10), it remains to characterize how the reweighted CME attentions in (3.6.5) recover the latent-to-target mappings in (3.6.3)-(3.6.4) within the RKHSs $\mathcal{H}_{\texttt{PT}}$ and $\mathcal{H}_{\texttt{DS}}$. See Figure 3.7 for an illustration of the construction of the RKHSs $\mathcal{H}_{\texttt{PT}}$ and $\mathcal{H}_{\texttt{DS}}$.

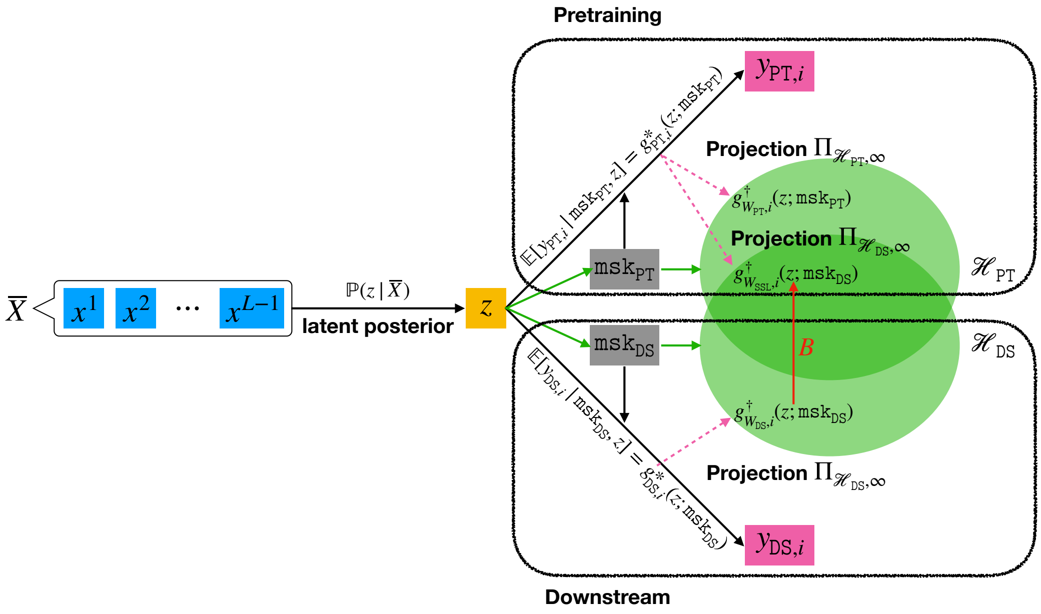

Pretraining
$y_{\mathtt{PT},i}$
$\mathbb{E}[y_{\mathtt{PT},i} \mid \mathtt{msk}_{\mathtt{PT}}, z] = g^*_{\mathtt{PT},i}(z; \mathtt{msk}_{\mathtt{PT}})$
Projection $\Pi_{\mathscr{H}_{\mathtt{PT}},\infty}$
$g^\dagger_{W_{\mathtt{PT}},i}(z; \mathtt{msk}_{\mathtt{PT}})$
Projection $\Pi_{\mathscr{H}_{\mathtt{DS}},\infty}$
$g^\dagger_{W_{\mathtt{SSL}},i}(z; \mathtt{msk}_{\mathtt{DS}})$
$\mathscr{H}_{\mathtt{PT}}$
$\overline{X}$
x^1
x^2
$\cdots$
x^{L-1}
$\mathbb{P}(z \mid \overline{X})$
latent posterior
z
$\mathtt{msk}_{\mathtt{PT}}$
$\mathtt{msk}_{\mathtt{DS}}$
B
$\mathscr{H}_{\mathtt{DS}}$
$\mathbb{E}[y_{\mathtt{DS},i} \mid \mathtt{msk}_{\mathtt{DS}}, z] = g^*_{\mathtt{DS},i}(z; \mathtt{msk}_{\mathtt{DS}})$
$g^\dagger_{W_{\mathtt{DS}},i}(z; \mathtt{msk}_{\mathtt{DS}})$
Projection $\Pi_{\mathscr{H}_{\mathtt{DS}},\infty}$
$y_{\mathtt{DS},i}$
Downstream

Figure 3.7. The RKHSs $\mathcal{H}_{\mathtt{PT}}$ and $\mathcal{H}_{\mathtt{DS}}$ induced by the latent-to-value mappings $\psi_{\mathtt{PT}}(z; \mathtt{msk}_{\mathtt{PT}})$ and $\psi_{\mathtt{DS}}(z; \mathtt{msk}_{\mathtt{DS}})$, respectively. The input masks $\mathtt{msk}_{\mathtt{PT}}$ and $\mathtt{msk}_{\mathtt{DS}}$ describe the pretraining process and the downstream task, respectively, and determine the RKHSs correspondingly. The ℓ_∞-norm projections $\Pi_{\mathcal{H}_{\mathtt{PT}},\infty}$ and $\Pi_{\mathcal{H}_{\mathtt{DS}},\infty}$ are defined in Assumption 3.6.1.

In parallel to Assumption 3.5.4, we introduce the following assumption on the fundamental hardness of approximating the latent-to-target mappings within the RKHSs $\mathcal{H}_{\mathtt{PT}}$ and $\mathcal{H}_{\mathtt{DS}}$.

Assumption 3.6.1 (SSL Recovery Gap)**.** For any fixed input masks $\mathtt{msk}_{\mathtt{PT}}$ and $\mathtt{msk}_{\mathtt{DS}}$, let

$$g^\dagger_{W_{\mathtt{PT}},i}(\cdot; \mathtt{msk}_{\mathtt{PT}}) = \Pi_{\mathcal{H}_{\mathtt{PT}},\infty}\big(g^*_{\mathtt{PT},i}(\cdot; \mathtt{msk}_{\mathtt{PT}})\big) = \underset{g_i(\cdot;\mathtt{msk}_{\mathtt{PT}}) \in \mathcal{H}_{\mathtt{PT}}}{\operatorname{argmin}} \left\| g^*_{\mathtt{PT},i}(\cdot; \mathtt{msk}_{\mathtt{PT}}) - g_i(\cdot; \mathtt{msk}_{\mathtt{PT}}) \right\|_\infty,$$

$$g^\dagger_{W_{\mathtt{DS}},i}(\cdot; \mathtt{msk}_{\mathtt{DS}}) = \Pi_{\mathcal{H}_{\mathtt{DS}},\infty}\big(g^*_{\mathtt{DS},i}(\cdot; \mathtt{msk}_{\mathtt{DS}})\big) = \underset{g_i(\cdot;\mathtt{msk}_{\mathtt{DS}}) \in \mathcal{H}_{\mathtt{DS}}}{\operatorname{argmin}} \left\| g^*_{\mathtt{DS},i}(\cdot; \mathtt{msk}_{\mathtt{DS}}) - g_i(\cdot; \mathtt{msk}_{\mathtt{DS}}) \right\|_\infty,$$

$$g^\dagger_{W_{\mathtt{SSL}},i}(\cdot; \mathtt{msk}_{\mathtt{DS}}) = \Pi_{\mathcal{H}_{\mathtt{DS}},\infty}\big(g^*_{\mathtt{PT},i}(\cdot; \mathtt{msk}_{\mathtt{PT}})\big) = \underset{g_i(\cdot;\mathtt{msk}_{\mathtt{DS}}) \in \mathcal{H}_{\mathtt{DS}}}{\operatorname{argmin}} \left\| g^*_{\mathtt{PT},i}(\cdot; \mathtt{msk}_{\mathtt{PT}}) - g_i(\cdot; \mathtt{msk}_{\mathtt{DS}}) \right\|_\infty$$

be the ℓ_∞-norm projections of the i-th entry $g^*_{\texttt{PT},i}(z;\texttt{msk}_{\texttt{PT}})$ of the latent-to-target mapping $g^*_{\texttt{PT}}(z;\texttt{msk}_{\texttt{PT}})$ onto the RKHS $\mathcal{H}_{\texttt{PT}}$, the i-th entry $g^*_{\texttt{DS},i}(z;\texttt{msk}_{\texttt{DS}})$ of the latent-to-target mapping $g^*_{\texttt{DS}}(z;\texttt{msk}_{\texttt{DS}})$ onto the RKHS $\mathcal{H}_{\texttt{DS}}$, and the i-th entry $g^*_{\texttt{PT},i}(z;\texttt{msk}_{\texttt{PT}})$ of the latent-to-target mapping $g^*_{\texttt{PT}}(z;\texttt{msk}_{\texttt{PT}})$ onto the RKHS $\mathcal{H}_{\texttt{DS}}$, respectively. We assume the following statements hold.

(PT) There exists $\epsilon_g(\texttt{msk}_{\texttt{PT}}) \in [0,+\infty)$ such that

$$\sum_{i=1}^{d} \big\| g^*_{\texttt{PT},i}(\cdot;\texttt{msk}_{\texttt{PT}}) - g^\dagger_{W_{\texttt{PT}},i}(\cdot;\texttt{msk}_{\texttt{PT}}) \big\|_\infty^2 \le \epsilon_g^2(\texttt{msk}_{\texttt{PT}}). \tag{3.6.7}$$

(DS) There exists $\epsilon_g(\texttt{msk}_{\texttt{DS}}) \in [0,+\infty)$ such that

$$\sum_{i=1}^{d_{\mathrm{y}}} \big\| g^*_{\texttt{DS},i}(\cdot;\texttt{msk}_{\texttt{DS}}) - g^\dagger_{W_{\texttt{DS}},i}(\cdot;\texttt{msk}_{\texttt{DS}}) \big\|_\infty^2 \le \epsilon_g^2(\texttt{msk}_{\texttt{DS}}). \tag{3.6.8}$$

(SSL) There exists $\epsilon_{\texttt{SSL}}(\texttt{msk}_{\texttt{PT}},\texttt{msk}_{\texttt{DS}}) \in [0,+\infty)$ such that

$$\sum_{i=1}^{d} \big\| g^*_{\texttt{PT},i}(\cdot;\texttt{msk}_{\texttt{PT}}) - g^\dagger_{W_{\texttt{SSL}},i}(\cdot;\texttt{msk}_{\texttt{DS}}) \big\|_\infty^2 \le \epsilon_{\texttt{SSL}}^2(\texttt{msk}_{\texttt{PT}},\texttt{msk}_{\texttt{DS}}). \tag{3.6.9}$$

Here the ℓ_∞-norms are taken over the latent variable z.

Intuitively, the feature $\psi_{\texttt{PT}}(z;\texttt{msk}_{\texttt{PT}})$ is obtained in the pretraining process, while the feature $\psi_{\texttt{DS}}(z;\texttt{msk}_{\texttt{DS}})$ is desired by the downstream task. Meanwhile, (3.6.9) characterizes the fundamental hardness of recovering the latent-to-target mapping $g^*_{\texttt{PT}}(z;\texttt{msk}_{\texttt{PT}})$ for the pretraining process within the RKHS $\mathcal{H}_{\texttt{DS}}$. Thus, the transfer error $\epsilon_{\texttt{SSL}}(\texttt{msk}_{\texttt{PT}},\texttt{msk}_{\texttt{DS}})$ captures the transfer capability of the sequence embedding obtained in the pretraining process to the downstream task. In other words, when the pretraining process is sufficiently related to the downstream task, the transfer error $\epsilon_{\texttt{SSL}}(\texttt{msk}_{\texttt{PT}},\texttt{msk}_{\texttt{DS}})$ is small, which allows us to approximate the i-th entry of the latent-to-target mapping $g^*_{\mathtt{PT}}(z;\mathtt{msk}_{\mathtt{PT}})$ within the RKHS $\mathcal{H}_{\mathtt{DS}}$ up to the approximation error $\epsilon_{\mathtt{SSL}}(\mathtt{msk}_{\mathtt{PT}},\mathtt{msk}_{\mathtt{DS}})$.

We introduce the following assumption on the condition number that characterizes the alignment between the reweighting parameter desired by the downstream task and the reweighting parameter obtained in the pretraining process.

Assumption 3.6.2 (SSL Condition Number)**.** Let $\{g^\dagger_{W_{\mathtt{DS}},i}(z;\mathtt{msk}_{\mathtt{DS}}) = w_{\mathtt{DS},i}^\top \psi_{\mathtt{DS}}(z;\mathtt{msk}_{\mathtt{DS}})\}_{i\in[d]}$ and $\{g^\dagger_{W_{\mathtt{SSL}},i}(z;\mathtt{msk}_{\mathtt{DS}}) = w_{\mathtt{SSL},i}^\top \psi_{\mathtt{DS}}(z;\mathtt{msk}_{\mathtt{DS}})\}_{i\in[d_{\mathrm{y}}]}$ be the function classes satisfying (3.6.8) and (3.6.9) in Assumption 3.6.1, respectively. Also, let $W_{\mathtt{DS}} = [w_{\mathtt{DS},1},\ldots,w_{\mathtt{DS},d_{\mathrm{y}}}]^\top \in \mathbb{R}^{d\times d_{\mathrm{y}}}$, $W_{\mathtt{SSL}} = [w_{\mathtt{SSL},1},\ldots,w_{\mathtt{SSL},d}]^\top \in \mathbb{R}^{d\times d}$, and[2]

$$B = W_{\mathtt{DS}}^\top (W_{\mathtt{SSL}} W_{\mathtt{SSL}}^\top)^{-1} W_{\mathtt{SSL}} \in \mathbb{R}^{d_{\mathrm{y}}\times d}. \tag{3.6.10}$$

We assume that there exists $\mu \in [0,+\infty)$ such that $\|B\|_2^2 \le \mu$.

The condition number μ plays a critical role in our subsequent analysis. To see the intuition behind μ, let $W_{\mathtt{DS}} = W_{\mathtt{SSL}}$, which implies that B is a projection matrix and $\mu = 1$. Also, let the row vectors of $W_{\mathtt{SSL}}$ be an orthonormal basis of $\mathbb{R}^d$, which implies that $W_{\mathtt{SSL}} W_{\mathtt{SSL}}^\top = I_{d_{\mathrm{p}}}$ and $B = W_{\mathtt{DS}}^\top W_{\mathtt{SSL}}$. In this case, B measures the subspace alignment between the reweighting parameter $W_{\mathtt{DS}}$ desired by the downstream task and the reweighting parameter $W_{\mathtt{SSL}}$ obtained in the pretraining process. In general cases where $W_{\mathtt{SSL}}$ is nonorthonormal, we have a similar interpretation through the eigenvalue decomposition of $W_{\mathtt{SSL}} W_{\mathtt{SSL}}^\top$.

Recall that $\widehat{f}_{\mathtt{PT}}(\overline{X};\mathtt{msk}_{\mathtt{PT}})$ is the attention neural network obtained in the pretraining process. For any $U \in \mathbb{R}^{d_{\mathrm{y}}\times d}$, we define the following quantity that characterizes the

[2] For ease of presentation, we assume that $W_{\mathtt{SSL}} W_{\mathtt{SSL}}^\top \in \mathbb{R}^{d\times d}$ is invertible. When $W_{\mathtt{SSL}} W_{\mathtt{SSL}}^\top$ is not invertible, our subsequent analysis can be generalized using the pseudoinverse of $W_{\mathtt{SSL}} W_{\mathtt{SSL}}^\top$.

expressity of the function class $\{\mathtt{agg}_{\theta}^{\mathtt{DS}} : \theta \in \Theta_{\mathtt{DS}}\}$ of aggregation layers for the downstream task,

$$
\begin{aligned}
\epsilon_{\mathtt{agg}}(U) &= \inf_{f_{\mathtt{DS}} \in \mathcal{F}_{\mathtt{DS}}} \sup_{\overline{X} \in \mathfrak{X}^{L-1}} \big\| f_{\mathtt{DS}}(\overline{X}; \mathtt{msk}_{\mathtt{DS}}) - U \widehat{f}_{\mathtt{PT}}(\overline{X}; \mathtt{msk}_{\mathtt{PT}}) \big\|_2 \\
&= \inf_{\theta \in \Theta_{\mathtt{DS}}} \sup_{\overline{X} \in \mathfrak{X}^{L-1}} \big\| \mathtt{agg}_{\theta}^{\mathtt{DS}} \circ \mathtt{attn}_{\mathtt{SM}}\big(q_{\widehat{\theta}_{\mathtt{PT}}}(\mathtt{msk}_{\mathtt{DS}}), k_{\widehat{\theta}_{\mathtt{PT}}}(\overline{X}), v_{\widehat{\theta}_{\mathtt{PT}}}(\overline{X})\big) \\
&\qquad - U \mathtt{agg}_{\widehat{\theta}_{\mathtt{PT}}}^{\mathtt{PT}} \circ \mathtt{attn}_{\mathtt{SM}}\big(q_{\widehat{\theta}_{\mathtt{PT}}}(\mathtt{msk}_{\mathtt{PT}}), k_{\widehat{\theta}_{\mathtt{PT}}}(\overline{X}), v_{\widehat{\theta}_{\mathtt{PT}}}(\overline{X})\big) \big\|_2, \qquad (3.6.11)
\end{aligned}
$$

where $\mathfrak{X}^{L-1}$ is defined in (3.5.2). Since the attention mechanism is frozen for the downstream task, the trainable part of the attention neural network is the aggregation layer $\mathtt{agg}_{\theta}^{\mathtt{DS}}$. Thus, the aggregation approximation error $\epsilon_{\mathtt{agg}}(U)$ characterizes the expressity of the function class $\{\mathtt{agg}_{\theta}^{\mathtt{DS}} : \theta \in \Theta_{\mathtt{DS}}\}$ of aggregation layers in terms of approximating the composition of (i) the linear transformation $U\mathtt{agg}_{\widehat{\theta}_{\mathtt{PT}}}^{\mathtt{PT}}$ of the aggregation layer obtained in the pretraining process and (ii) the output variation induced by switching the input mask $\mathtt{msk}_{\mathtt{PT}}$ to another input mask $\mathtt{msk}_{\mathtt{DS}}$ in the attention mechanism, which is frozen. To see the intuition behind $\epsilon_{\mathtt{agg}}(U)$, let $\mathtt{msk}_{\mathtt{PT}} = \mathtt{msk}_{\mathtt{DS}}$, which implies that $\epsilon_{\mathtt{agg}}(U) = 0$ as long as the function class of aggregation layers takes the form of $\{\mathtt{agg}_{\theta}^{\mathtt{DS}} = U\mathtt{agg}_{\widehat{\theta}_{\mathtt{PT}}}^{\mathtt{PT}} : \theta = U \in \mathbb{R}^{d_{\mathrm{y}} \times d}\}$. In this case, $\epsilon_{\mathtt{agg}}(U)$ characterizes the compatibility between $\mathtt{agg}_{\theta}^{\mathtt{DS}}$ and $\mathtt{agg}_{\widehat{\theta}_{\mathtt{PT}}}^{\mathtt{PT}}$ under a linear transformation parameterized by θ. In general cases where $\mathtt{msk}_{\mathtt{PT}} \neq \mathtt{msk}_{\mathtt{DS}}$, $\epsilon_{\mathtt{agg}}(U)$ additionally characterizes the capability of $\mathtt{agg}_{\theta}^{\mathtt{DS}}$ to capture the output variation induced by switching the input mask.

Recall that the function class $\mathcal{F}_{\texttt{PT}}$ of attention neural networks for the pretraining process is defined in (3.6.1). Let

$$\mathcal{E}^{\texttt{PT}}_{\text{approx}} = \min_{f \in \mathcal{F}_{\texttt{PT}}} \mathbb{E}\Big[\mathcal{L}\big((\overline{X}, y_{\texttt{PT}}), f\big)\Big] - \mathbb{E}\Big[\mathcal{L}\big((\overline{X}, y_{\texttt{PT}}), f^*_{\texttt{PT}}\big)\Big] \tag{3.6.12}$$

be the approximation error for the pretraining process, which is characterized in Theorem 3.5.5. Recall that the function class $\mathcal{F}_{\texttt{DS}}$ of attention neural networks for the downstream task is defined in (3.6.2). For the downstream task, the approximation error in (3.5.3) takes the following form,

$$\mathcal{E}_{\text{approx}} = \min_{f \in \mathcal{F}_{\texttt{DS}}} \mathbb{E}\Big[\mathcal{L}\big((\overline{X}, y_{\texttt{DS}}), f\big)\Big] - \mathbb{E}\Big[\mathcal{L}\big((\overline{X}, y_{\texttt{DS}}), f^*_{\texttt{DS}}\big)\Big]. \tag{3.6.13}$$

The following theorem characterizes the approximation error $\mathcal{E}_{\text{approx}}$ for the SSL process.

Theorem 3.6.3 (SSL Approximation Error)**.** Under Assumptions 3.6.1 and 3.6.2, it holds that

$$\mathcal{E}_{\text{approx}} = O\Big(\mu \cdot \big(\mathcal{E}^{\texttt{PT}}_{\text{approx}} + \epsilon^2_{\texttt{SSL}}(\texttt{msk}_{\texttt{PT}}, \texttt{msk}_{\texttt{DS}})\big) + \epsilon^2_g(\texttt{msk}_{\texttt{DS}}) + \epsilon^2_{\texttt{agg}}(B)\Big),$$

where $\mathcal{E}^{\texttt{PT}}_{\text{approx}}$, $\epsilon_{\texttt{SSL}}(\texttt{msk}_{\texttt{PT}}, \texttt{msk}_{\texttt{DS}})$, $\epsilon_g(\texttt{msk}_{\texttt{DS}})$, and $\epsilon_{\texttt{agg}}(B)$ are defined in (3.6.12), (3.6.9), (3.6.8), and (3.6.11), respectively, and B is defined in (3.6.10).

Proof. See §C.5.3 for a detailed proof. □

Theorem 3.6.3 demonstrates that the attention neural network enables the transfer capability to diverse downstream tasks, where the approximation error is subsume from that in the supervised setting with a few extra error terms. We interpret the approximation error in Theorem 3.6.3 as follows.

(i) The condition number μ characterizes the the alignment between the reweighting parameter $W_{\mathtt{DS}}$ desired by the downstream task and the reweighting parameter $W_{\mathtt{SSL}}$ obtained in the pretraining process. When $W_{\mathtt{DS}} = W_{\mathtt{SSL}}$, we have $\mu = 1$.

(ii) The approximation error $\mathcal{E}^{\mathtt{PT}}_{\text{approx}}$ for the pretraining process is characterized in Theorem 3.5.5. Specifically, $\mathcal{E}^{\mathtt{PT}}_{\text{approx}}$ involves the pretraining recovery gap $\epsilon_g(\mathtt{msk}_{\mathtt{PT}})$ defined in (3.6.7), which characterizes the fundamental hardness of approximating the i-th entry $g^*_{\mathtt{PT},i}(z;\mathtt{msk}_{\mathtt{PT}})$ of the latent-to-target mapping $g^*_{\mathtt{PT}}(z;\mathtt{msk}_{\mathtt{PT}})$ defined in (3.6.3) within the RKHS $\mathcal{H}_{\mathtt{PT}}$, and the attention approximation error $\epsilon_{\mathtt{attn}}$ defined in (3.5.11), which is characterized in Proposition 3.4.2.

(iii) The transfer error $\epsilon_{\mathtt{SSL}}(\mathtt{msk}_{\mathtt{PT}}, \mathtt{msk}_{\mathtt{DS}})$ captures the transfer capability of the sequence embedding obtained in the pretraining process to the downstream task. By our previous arguments following Assumption 3.6.1, $\epsilon_{\mathtt{SSL}}(\mathtt{msk}_{\mathtt{PT}}, \mathtt{msk}_{\mathtt{DS}})$ is small as long as the pretraining process is sufficiently related to the downstream task.

(iv) The downstream recovery gap $\epsilon_g(\mathtt{msk}_{\mathtt{DS}})$ defined in (3.6.8) characterizes the fundamental hardness of approximating the i-th entry $g^*_{\mathtt{DS},i}(z;\mathtt{msk}_{\mathtt{DS}})$ of the latent-to-target mapping $g^*_{\mathtt{DS}}(z;\mathtt{msk}_{\mathtt{DS}})$ defined in (3.6.4) within the RKHS $\mathcal{H}_{\mathtt{DS}}$.

(v) The aggregation approximation error $\epsilon_{\mathsf{agg}}(B)$ measures the expressity of the function class of aggregation layers for the downstream task. By our previous arguments following (3.6.11), $\epsilon_{\mathsf{agg}}(B)$ is small as long as the aggregation layer $\mathsf{agg}^{\mathtt{DS}}_{\theta}$ for the downstream task can approximate the composition of the linear transformation $B\mathsf{agg}^{\mathtt{PT}}_{\widehat{\theta}_{\mathtt{PT}}}$ of the aggregation layer obtained in the pretraining process and the variation induced by switching the input mask $\mathtt{msk}_{\mathtt{PT}}$ to another input mask $\mathtt{msk}_{\mathtt{DS}}$.

CHAPTER 4

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

In this paper, we conduct a comprehensive study of In-Context Learning (ICL) by addressing several open questions: (a) What type of ICL estimator is learned within language models? (b) What are suitable performance metrics to evaluate ICL accurately and what are the error rates? (c) How does the transformer architecture enable ICL? To answer (a), we take a Bayesian view and demonstrate that ICL implicitly implements the Bayesian model averaging algorithm. This Bayesian model averaging algorithm is proven to be approximately parameterized by the attention mechanism. For (b), we analyze the ICL performance from an online learning perspective and establish a regret bound $\mathcal{O}(1/T)$, where T is the ICL input sequence length. To address (c), in addition to the encoded Bayesian model averaging algorithm in attention, we show that during pertaining, the total variation distance between the learned model and the nominal model is bounded by a sum of an approximation error and a generalization error of $\widetilde{\mathcal{O}}(1/\sqrt{N_{\mathrm{p}}T_{\mathrm{p}}})$, where N_{p} and T_{p} are the number of token sequences and the length of each sequence in pretraining, respectively. Our results provide a unified understanding of the transformer and its ICL ability with bounds on ICL regret, approximation, and generalization, which deepens our knowledge of these essential aspects of modern language models.

4.1. Introduction

With the ever-increasing sizes of model capacity and corpus, LLM have achieved tremendous successes across a wide range of tasks, including natural language understanding (Dong et al., 2019; Jiao et al., 2019), symbolic reasoning (Wei et al., 2022c; Kojima et al., 2022), and conversations (Brown et al., 2020; Ouyang et al., 2022). Recent studies have revealed that these LLMs possess immense potential, as their large capacity allows for a series of *emergent abilities* (Wei et al., 2022b; Liu et al., 2023). One such ability is ICL, which enables an LLM to learn from just a few examples, without tuning parameters. Since the examples are provided in natural language, ICL offers an interpretable way for humans to communicate with and impart knowledge to LLMs (Liu et al., 2021; Dong et al., 2022).

Despite the immense empirical successes of ICL, its theoretical understanding remains limited. Specifically, existing works fail to explain why LLMs the ability for ICL, how the attention mechanism is related to the ICL ability, and how pretraining influences ICL. Although the optimality of ICL is investigated in Xie et al. (2021) and Wies et al. (2023), these works both make unrealistic assumptions on the pretrained models, and their results cannot clarify the importance of the attention mechanism in ICL.

We focus on the scenario where transformers are first pretrained on a large dataset and then prompted to perform ICL. Our goal is to analyze both the pretraining phase and the ICL performance of the pretrained model, aiming to understand why LLMs possess such a strong ability for ICL and how this ability relates to the pretraining process. This boils down to three questions: (a) What type of ICL estimator is learned by LLMs? (b) What are suitable performance metrics to evaluate ICL accurately and what are the error rates? (c) How does the transformer structure enable ICL? The first and the third questions demand distilling the ICL process from the transformer structure itself. It relies on analytically analyzing the inference in transformers. The second question then requires statistically analyzing the extracted ICL process. Beyond the mentioned analytic analysis, the third question also necessitates a statistical analysis of the pretrained transformer.

To address the first question, we show that the perfectly pretrained LLMs perform ICL in the form of Bayesian model averaging and that the attention mechanism in the transformer parameterizes the Bayesian model averaging. For the second question, we adopt an online learning setting and analyze the ICL regret of this extracted Bayesian model averaging algorithm. Finally, for the third question, we apply a PAC-Bayes framework to analyze pretraining error and build a connection between pretraining error and ICL regret.

In this paper, we first show that the perfectly pretrained LLMs performs Bayesian model averaging over a general model in ICL in Theorem 4.4.1, which subsumes the models in previous works. Based on this, we derive the ICL regret of the Bayesian model averaging as $\mathcal{O}(1/T)$. Here T is the number of ICL examples. To build a connection with the transformer, we show that the attention mechanism approximately parameterizes the Bayesian model averaging as T goes to infinity in Proposition 4.4.3. In addition, we analyze the pretraining of transformers and show that the total variation distance of the learned model and the nominal distribution is bounded by the sum of approximation error and generalization error in Theorem 4.5.3. The generalization error is $\mathcal{O}(1/\sqrt{N_{\mathrm{p}}T_{\mathrm{p}}})$, where N_{p} and T_{p} are the number of token sequences and the length of each sequence in the pretraining dataset, respectively. The approximation error decays exponentially with the depth of the transformer. This features the first pretraining analysis of transformers in total variation distance, which also takes the approximation error into account. In Theorem 4.6.2, we show that the ICL regret of the pretrained model is the sum of the pretraining error and the $\mathcal{O}(1/T)$ regret for Bayesian model averaging with the true distribution. Moreover, even when the input-output mappings are wrong in the example, we show that LLMs can identify the nominal concept if the nominal concept is separated from other concepts with respect to KL divergence, which is called the distinguishable case. The performance guarantee for ICL in the distinguishable case is provided in Proposition 4.6.7.

4.2. Related Work

In-Context Learning. After Brown et al. (2020) showcased the in-context learning capacity of GPT-3, there has been a notable surge in interest towards enhancing and comprehending this particular ability (Dong et al., 2022). The in-context learning ability has seen enhancements through the incorporation of extra training stages (Min et al., 2021; Wei et al., 2021b; Iyer et al., 2022), carefully selecting and arranging informative demonstrations (Liu et al., 2021; Kim et al., 2022; Rubin et al., 2021; Lu et al., 2021), giving explicit instructions (Honovich et al., 2022; Zhou et al., 2022b; Wang et al., 2022), and prompting a chain of thoughts (Wei et al., 2022c; Zhang et al., 2022c; Zhou et al., 2022a). In efforts to comprehend the mechanisms of ICL ability, researchers have also conducted extensive work. Empirically, Chan et al. (2022) demonstrated that the distributional properties, including the long-tailedness, are important for ICL. Garg et al. (2022)

investigated the function class that ICL can approximate. Min et al. (2022) showed that providing wrong mappings between the input-output pairs in examples does not degrade the ICL. Theoretically, Akyürek et al. (2022), von Oswald et al. (2022), and Dai et al. (2022) indicated that ICL implicitly implements the gradient descent or least-square algorithms from the function approximation perspective. However, the first two works only showed that transformers are able to approximate these two algorithms, which may not align with the pretrained model. The last work ignored the softmax module, which turns out to be important in practical implementation. Li et al. (2023) viewed ICL from the multi-task learning perspective and derived the generalization bound. Xie et al. (2021) analyzed ICL within the Bayesian framework, assuming the access to the nominal language distribution and that the tokens are generated from Hiddn Markov Model (HMM)s. However, the first assumption hides the relationship between pretraining and ICL, and the second assumption is restrictive. Following this thread, Wies et al. (2023) relaxed the HMM assumption and assumed access to a pretrained model that is close to the nominal distribution conditioned on any token sequence, which is also unrealistic. Two concurrent works Wang et al. (2023), and Jiang (2023) also provide the Bayesian analysis of ICL. Unfortunately, these Bayesian works cannot explain the importance of the attention mechanism for ICL and clarify how pretraining relates to ICL. In contrast, we prove that the structure of attention enables Bayesian model averaging and related the pretraining error of transformers to the ICL regret.

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 Bayesian model averaging algorithm in the ICL setting. Furthermore, we derive the approximation and generalization bounds for transformers in the pretraining phase.

4.3. Preliminary

Notation. We denote $\{1, \cdots, N\}$ as $[N]$. For a Polish space $\mathcal{S}$, we denote the collection of all the probability measures on it as $\Delta(\mathcal{S})$. The total variation distance between two distributions $P, Q \in \Delta(\mathcal{S})$ is $\mathrm{TV}(P, Q) = \sup_{A \subseteq \mathcal{S}} |P(A) - Q(A)|$. The i^{th} entry of a vector x is denoted as x_i or $[x]_i$. For a matrix $X \in \mathbb{R}^{T \times d}$, we index its i^{th} row and column as $X_{i,:}$ and $X_{:,i}$ respectively. The $\ell_{p,q}$ norm of X is defined as $\|X\|_{p,q} = (\sum_{i=1}^{d} \|X_{:,i}\|_p^q)^{1/q}$, and the *Frobenius norm* of it is defined as $\|X\|_{\mathrm{F}} = \|X\|_{2,2}$.

Attention and Transformers. Attention mechanism has been the most powerful and popular neural network module in both Computer Vision (CV) and Natural Language Processing (NLP) communities, and it is the backbone of the LLMs (Devlin et al., 2018; Brown et al., 2020). Assume that we have a query vector $q \in \mathbb{R}^{d_k}$. With T key vectors in

$K \in \mathbb{R}^{T\times d_k}$ and T value vectors in $V \in \mathbb{R}^{T\times d_v}$, the attention mechanism maps the query vector q to $\texttt{attn}(q, K, V) = V^\top \texttt{softmax}(Kq)$, where $\texttt{softmax}$ normalizes a vector via the exponential function, i.e., for $x \in \mathbb{R}^d$, $[\texttt{softmax}(x)]_i = \exp(x_i)/\sum_{j=1}^{d}\exp(x_j)$ for $i \in [d]$. The output is a weighted sum of V, and the weights reflect the closeness between W and q. For t query vectors, we stack them into $Q \in \mathbb{R}^{t\times d_k}$. Attention maps these queries using the function $\texttt{attn}(Q, K, V) = \texttt{softmax}(QK^\top)V \in \mathbb{R}^{t\times d_v}$, where $\texttt{softmax}$ is applied row-wisely. In the practical design of transformers, practitioners usually use Multi-Head Attention (MHA) instead of single attention to express sophisticated functions, which forwards the inputs through h attention modules in parallel and outputs the sum of these sub-modules. Here $h \in \mathbb{N}$ is a hyperparameter. Taking $X \in \mathbb{R}^{T\times d}$ as the input, MHA outputs $\texttt{mha}(X, W) = \sum_{i=1}^{h}\texttt{attn}(XW_i^Q, XW_i^K, XW_i^V)$, where $W = (W_i^Q, W_i^K, W_i^V)_{i=1}^h$ is the parameters set of h attention modules, $W_i^Q \in \mathbb{R}^{d\times d_h}$, $W_i^K \in \mathbb{R}^{d\times d_h}$, and $W_i^V \in \mathbb{R}^{d\times d}$ for $i \in [h]$ are weight matrices for queries, keys, and values, and d_h is usually set to be d/h (Michel et al., 2019). The transformer is the concatenation of the attention modules and the fully-connected layers, which is widely adopted in LLMs (Devlin et al., 2018; Brown et al., 2020).

Large Language Models and In-Context Learning. Many LLMs are *autoregressive*, such as GPT (Brown et al., 2020). It means that the model continuously predicts future tokens based on its own previous values. For example, starting from a token $x_1 \in \mathfrak{X}$, where $\mathfrak{X}$ is the alphabet of tokens, a LLM $\mathbb{P}_\theta$ with parameter $\theta \in \Theta$ continuously predicts the next token according to $x_{t+1} \sim \mathbb{P}_\theta(\cdot \,|\, S_t)$ based on the past $S_t = (x_1, \cdots, x_t)$ for $t \in \mathbb{N}$. Here, each token represents a word and the position of the word (Ke et al., 2020), and the token sequences S_t for $t \in \mathbb{N}$ live in the sequences space $\mathfrak{X}^*$. LLMs are first *pretrained* on a

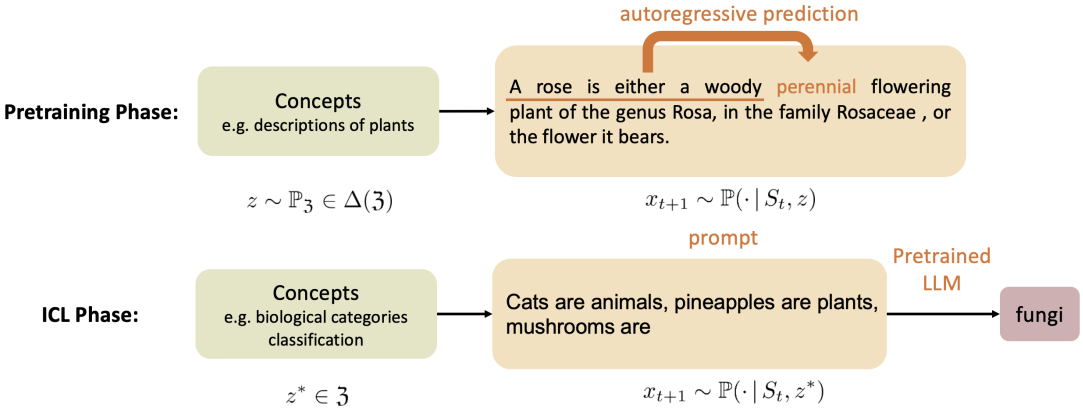

autoregressive prediction
Pretraining Phase:
Concepts
e.g. descriptions of plants
A rose is either a woody perennial flowering plant of the genus Rosa, in the family Rosaceae , or the flower it bears.
$z \sim \mathbb{P}_{\mathfrak{Z}} \in \Delta(\mathfrak{Z})$
$x_{t+1} \sim \mathbb{P}(\cdot \mid S_t, z)$
prompt
Pretrained LLM
ICL Phase:
Concepts
e.g. biological categories classification
Cats are animals, pineapples are plants, mushrooms are
fungi
$z^* \in \mathfrak{Z}$
$x_{t+1} \sim \mathbb{P}(\cdot \mid S_t, z^*)$

Figure 4.1. 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.

huge body of corpus, making the prediction $x_{t+1} \sim \mathbb{P}_\theta(\cdot \mid S_t)$ accurate, and then prompted to perform downstream tasks. During the pretraining phase, we aim to maximize the conditional probability $\mathbb{P}_\theta(x \mid S)$ over the nominal token x, which can be the tokens inside a sentence (Devlin et al., 2018) or the next token (Brown et al., 2020).

After pretraining, LLMs are prompted to perform downstream tasks without tuning parameters. Different from the finetuned models that learn the task explicitly (Liu et al., 2023), LLMs can implicitly learn from the examples in the *prompt*, which is known as ICL (Brown et al., 2020). Concretely, as shown in Figure fig:pipeline, pretrained LLMs are provided with a prompt $\texttt{prompt}_t = (\widetilde{c}_1, r_1, \ldots, \widetilde{c}_t, r_t, \widetilde{c}_{t+1})$ with t examples and a query as inputs, where each pair $(\widetilde{c}_i, r_i) \in \mathfrak{X}^* \times \mathfrak{X}$ is an example of the task, and $\widetilde{c}_{t+1}$ is the query. For example, the $\texttt{prompt}_t$ with $t = 2$ can be "Cats are animals, pineapples are plants, mushrooms are". Here $\widetilde{c}_1 \in \mathfrak{X}^*$ is a token sequence "Cats are", while r_1 is the response "animals". The query $\widetilde{c}_{t+1}$ is "mushrooms are", and the desired response is "fungi". The prompts are generated from a hidden concept $z_* \in \mathfrak{Z}$, e.g., z_* can be the classification of biological categories, where $\mathfrak{Z}$ is the concept space. The generation process is $\widetilde{c}_i \sim \mathbb{P}(\cdot \,|\, \widetilde{c}_1, r_1, \cdots, \widetilde{c}_{i-1}, r_{i-1}, z_*)$ and $r_i \sim \mathbb{P}(\cdot \,|\, \mathtt{prompt}_{i-1}, z_*)$ for the nominal distribution $\mathbb{P}$ and $i \in [t]$. Thus, in ICL, LLMs aim to estimate the conditional distribution $\mathbb{P}(r_{t+1}|\mathtt{prompt}_t, z_*)$. It is widely conjectured that the pretrained LLMs can implicitly identify the hidden concept $z_* \in \mathfrak{Z}$ from the examples, and thus perform ICL. In the following, we will provide theoretical justifications for this claim. We note that delimiters are omitted in our work, and our results can be generalized to handle this case. Since LLMs are autoregressive, the definition of the notation $\mathbb{P}(\cdot \,|\, S)$ with $S \in \mathfrak{X}^*$ may be ambiguous. Unless explicitly specified, we adopt $\mathbb{P}(\cdot \,|\, S)$ to denote the distribution of the next single token conditioned on S.

4.4. In-Context Learning via Bayesian Model Averaging

In this section, we show that LLMs perform ICL implicitly via Bayesian model averaging. Given a sequence $S = \{(\widetilde{c}_t, r_t)\}_{t=1}^T$ with T examples generated from a hidden concept $z_* \in \mathfrak{Z}$, we use $S_t = \{(\widetilde{c}_i, r_i)\}_{i=1}^t$ to represent the first t ICL examples in the sequence. Here $\widetilde{c}_t$ and r_t respectively denote the ICL covariate and response. During the ICL phase, a LLM is sequentially prompted with $\mathtt{prompt}_t = (S_t, \widetilde{c}_{t+1})$ for $t \in [T-1]$, i.e., the first t examples and the $(t+1)$-th covariate. The prompted LLM aims to predict the response r_{t+1} based on $\mathtt{prompt}_t = (S_t, \widetilde{c}_{t+1})$ whose ground-truth distribution is $r_{t+1} \sim \mathbb{P}(\cdot \,|\, \mathtt{prompt}_t, z_*)$. We temporarily consider the setting where we have access to the nominal pretraining distribution $\mathbb{P}(r_{t+1} \,|\, \mathtt{prompt}_t)$, i.e., the pretrained model *perfectly* learns the distribution, and we relax this condition by specifying the pretraining error in Section 4.5. For the analysis of ICL, we take the following model to further specify $r_{t+1} \sim \mathbb{P}(\cdot \,|\, \texttt{prompt}_t, z_*)$ as

$$r_t = f(\widetilde{c}_t, h_t, \xi_t), \quad \forall t \in [T], \tag{4.4.1}$$

where the hidden variable $h_t \in \mathcal{H}$ determines the relation between c_t and r_t, $\xi_t \in \Xi$ for $t \in [T]$ are i.i.d. random noises, and $f : \mathcal{X} \times \mathcal{H} \times \Xi \to \mathfrak{X}$ is a function that relates response r_t to $\widetilde{c}_t, h_t$, and ξ_t. The hidden variables $\{h_t\}_{t=1}^T$ form a stochastic process, whose distribution is determined by the hidden concept $z_* \in \mathfrak{Z}$. The model in (4.4.1) essentially assumes that the hidden concept z_* implicitly determines the transition of the belief $b_t = \mathbb{P}(r_t = \cdot \,|\, \widetilde{c}_t)$, and it does not impose any assumption on the distribution of $\widetilde{c}$. This model is quite general, and it subsumes the models in previous works. When f is the emission function in HMM and $h_t = h$ for $t \in [T]$ is the values of hidden states that depend on z, model in (4.4.1) recovers the HMM assumption in Xie et al. (2021). When $h_t = z$ for $t \in [T]$ degenerate to the hidden concept, this recovers the casual graph model in Wang et al. (2023) and the ICL model in Jiang (2023). Under the model in (4.4.1), we show that perfectly pretrained LLMs perform Bayesian model averaging (Wasserman, 2000).

Theorem 4.4.1 (LLMs Perform Bayesian Model Averaging)**.** Under the model in (4.4.1), it holds that

$$\mathbb{P}(r_{t+1} \,|\, \texttt{prompt}_t) = \int \mathbb{P}(r_{t+1} \,|\, \widetilde{c}_{t+1}, S_t, z)\mathbb{P}(z \,|\, S_t)\mathrm{d}z. \tag{4.4.2}$$

We note that the right-hand side of (4.4.2) is exactly the Bayesian model averaging algorithm that takes S_t and r_{t+1} as the training set and the test sample, respectively. Thus, this theorem implies that the perfectly trained LLMs perform Bayesian model averaging in ICL. The proof is in Appendix D.2.1.

Next, we study the performance of ICL from an online learning perspective. Recall that LLMs are continuously prompted with S_t and aim to predict the $(t+1)$-th covariate r_{t+1} for $t \in [T-1]$. This can be viewed as an online learning problem. For a sequence of density estimators $\{\widehat{\mathbb{P}}(r_t)\}_{t=1}^{T}$, we take the following ICL regret as its performance metric,

$$\texttt{regret}_t = t^{-1} \sup_z \sum_{i=1}^{t} \log \mathbb{P}(r_i \,|\, \texttt{prompt}_{i-1}, z) - t^{-1} \sum_{i=1}^{t} \log \widehat{\mathbb{P}}(r_i). \tag{4.4.3}$$

This ICL regret measures the performance of the estimator $\widehat{\mathbb{P}}$ compared with the best hidden concept in hindsight. For the perfectly trained LLMs, the estimator is exactly $\widehat{\mathbb{P}}(r_t) = \mathbb{P}(r_{t+1} \,|\, \texttt{prompt}_t)$.

Theorem 4.4.2 (ICL Regret of Perfectly Pretrained Model)**.** Under the model in (4.4.1), we have for any $t \in [T]$ that

$$t^{-1} \sum_{i=1}^{t} \log \mathbb{P}(r_i \,|\, \texttt{prompt}_{i-1}) \geq \sup_{z \in \mathcal{Z}} \Big(t^{-1} \sum_{i=1}^{t} \log \mathbb{P}(r_i \,|\, z, \texttt{prompt}_{i-1}) + t^{-1} \log \mathbb{P}_{\mathcal{Z}}(z) \Big).$$

Here $\mathbb{P}_{\mathcal{Z}}$ is the prior of the hidden concept $z \in \mathcal{Z}$. When the hidden concept space $\mathfrak{Z}$ is finite and the prior $\mathbb{P}_{\mathcal{Z}}(z)$ is the uniform distribution on $\mathfrak{Z}$, we have that $\texttt{regret}_t \leq \log |\mathfrak{Z}|/t$. When the nominal concept z_* satisfies that $\sup_z \sum_{i=1}^{t} \mathbb{P}(r_i \,|\, z, \texttt{prompt}_{i-1}) = \sum_{i=1}^{t} \mathbb{P}(r_i \,|\, z_*, \texttt{prompt}_{i-1})$ for any $t \in [T]$, the regret is bounded as $\texttt{regret}_t \leq \log(1/\mathbb{P}_{\mathcal{Z}}(z_*))/t$.

This theorem state that the ICL regret of the perfectly pretrained model is the minus logarithmic prior probability of concept divided by T. This is intuitive, since the regret is relatively large if the concept z_* rarely appears according to the prior distribution. The proof of Theorem 4.4.2 is in Appendix D.2.2. Theorems 4.4.1 and 4.4.2 show that the perfectly pretrained LLMs perform Bayesian model averaging in ICL and have a ICL regret of $\mathcal{O}(1/t)$. In Section 4.5, we characterize the deviation between the learned model and the underlying true model. Next, we show how transformers parameterize Bayesian model averaging.

4.4.1. Attention Parameterizes Bayesian Model Averaging

To simplify the presentation, we consider the case where the covariate $\widetilde{c}_t \in \mathfrak{X}^*$ is a single token $c_t \in \mathfrak{X}$ in this subsection. During the ICL phase, pretrained LLMs are prompted with $\texttt{prompt}_t = (S_t, c_{t+1})$ and tasked with predicting the $(t+1)$-th response r_{t+1}. We assume the existence of two learnable mappings $k : \mathbb{R}^d \to \mathbb{R}^{d_k}$ and $v : \mathbb{R}^d \to \mathbb{R}^{d_v}$, which are parameterized by fully connected layers, and their nominal versions k_* and v_* satisfy the following relation:

$$v_t = z\phi(k_t) + \epsilon_t, \quad \forall t \in [T], \tag{4.4.4}$$

where $v_t = v_*(r_t)$ represents the value, $k_t = k_*(c_t)$ denotes the key, $\phi : \mathbb{R}^{d_k} \to \mathbb{R}^{d_\phi}$ refers to the feature mapping in some Reproducing Kernel Hilbert Space (RKHS), $z \in \mathbb{R}^{d_v \times d_\phi}$ corresponds to the hidden concept, and $\epsilon_t \sim N(0, \sigma^2 I)$ is Gaussian noise with variance σ^2. We assume that ϵ_t is independent across $t \in [T]$. The mappings v and k represent the feature extraction process in the high-dimensional space induced by transformers. In such space, the hidden concept z represents a transformation between the value v and the key k. Here, we simply take this as the transformation by a matrix, which can be easily generalized by building a bijection between concepts z and complex transformations. The pretraining of the transformer essentially learns the nominal mappings v_* and k_*. Note that (4.4.4) can be written as

$$r_t = v_*^{-1}\Big(z\phi\big(k_*(c_t)\big) + \epsilon_t\Big), \tag{4.4.5}$$

which is a realization of (4.4.1) with $h_t = z$, $\xi_t = \epsilon_t$, and $f(c, h, \xi) = v_*^{-1}(h\phi(k_*(c)) + \xi)$. In the following, we adopt the prior of z as $N(0, \lambda I)$ and denote the kernel function of the RKHS induced by ϕ as $\mathfrak{K} : \mathbb{R}^{d_k} \times \mathbb{R}^{d_k} \to \mathbb{R}$. The stacks of the values and keys are denoted as $K_t = (k_1, \ldots, k_t)^\top \in \mathbb{R}^{t\times d_k}$ and $V_t = (v_1, \ldots, v_t)^\top \in \mathbb{R}^{t\times d_v}$, respectively. Additionally, we denote the query for the $(t+1)$-th token as $q_{t+1} = k_{t+1} = k_*(c_{t+1})$. Consequently, the model in (4.4.4) implies that

$$\mathbb{P}(v_{t+1} \,|\, \texttt{prompt}_t) = \int \mathbb{P}(v_{t+1} \,|\, z, q_{t+1})\mathbb{P}(z \,|\, S_t)\mathrm{d}z \propto \exp\Big(-\big\|v_{t+1} - \bar{z}_t\phi(q_{t+1})\big\|_{\Sigma_t^{-1}}^2/2\Big), \tag{4.4.6}$$

where we denote by Σ_t the covariance of $v_{t+1} \sim \mathbb{P}(\cdot \,|\, S_t, q_{t+1})$, and the mean concept $\bar{z}_t$ is

$$\bar{z}_t = V_t\big(\phi(K_t)\phi(K_t)^\top + \lambda I\big)^{-1}\phi(K_t) = V_t\big(\mathfrak{K}(K_t, K_t) + \lambda I\big)^{-1}\phi(K_t). \tag{4.4.7}$$

Combining (4.4.6) and (4.4.7), we can see that $\bar{z}_t\phi(q_{t+1})$ essentially measures the similarity between the query and keys, which is quite similar to the attention mechanism defined in Section 4.3. However, here the similarity is normalization according to (4.4.7), not by softmax. This motivates us to define a new structure of attention and explore the relationship between the newly defined attention and the original one. For any $q \in \mathbb{R}^{d_k}$, $K \in \mathbb{R}^{t\times d_k}$, and $V \in \mathbb{R}^{t\times d_v}$, we define a variant of the attention mechanism as follows,

$$\texttt{attn}_\dagger(q, K, V) = V^\top\big(\mathfrak{K}(K, K) + \lambda I\big)^{-1}\mathfrak{K}(K, q). \tag{4.4.8}$$

From (4.4.6), (4.4.7), and (4.4.8), it holds that the response v_{t+1} for $(t+1)$-th query is distributed as $v_{t+1} \sim N(\mathtt{attn}_\dagger(q_{t+1}, K_t, V_t), \Sigma_t)$. Recall that we define the softmax attention (Vaswani et al., 2017) for any $q \in \mathbb{R}^{d_k}$, $K \in \mathbb{R}^{t\times d_k}$, and $V \in \mathbb{R}^{t\times d_v}$ as

$$\mathtt{attn}(q, K, V) = V^\top \mathtt{softmax}(Kq)$$

In the following proposition, we demonstrate that the attention in (4.4.8) converges to the same limit as the softmax attention as the sequence length goes to infinity.

Proposition 4.4.3. We assume that the key-value pairs $\{(k_t, v_t)\}_{t=1}^T$ are independent and identically distributed, and we adopt Gaussian RBF kernel $\mathfrak{K}_{\mathtt{RBF}}$. In addition, we assume that $\|k_t\|_2 = \|v_t\| = 1$. Then, it holds for an absolute constant $C > 0$ and any $q \in \mathbb{R}^{d_k}$ with $\|q\| = 1$ that

$$\lim_{T\to\infty} \mathtt{attn}_\dagger(q, K_T, V_T) = C \cdot \lim_{T\to\infty} \mathtt{attn}(q, K_T, V_T).$$

The proof is in Appendix D.2.3. Combined with the conditional probability of v_{t+1} in (4.4.6), this proposition shows how the attention mechanism parameterizes the Bayesian model averaging in long token sequences (Wasserman, 2000).

4.5. Theoretical Analysis of Pretraining

4.5.1. Pretraining Algorithm

In this section, we describe the pretraining setting. We largely follow the transformer structures in Devlin et al. (2018) and Brown et al. (2020). The whole network is a composition of D sub-modules, and each sub-module consists of a MHA and a Feed-Forward (FF) fully-connected layer. Here, $D > 0$ is the depth of the network. The whole network takes $X^{(0)} = X \in \mathbb{R}^{L\times d}$ as its input. In the t-th layer for $t \in [D]$, it first takes the output $X^{(t-1)}$ of the $(t-1)$-th layer as the input and forward it through MHA with a residual link and a layer normalization $\Pi_{\mathrm{norm}}(\cdot)$ to output $Y^{(t)}$, which projects each row of the input into the unit ℓ_2-ball. Here we take $d_h = d$ in MHA, and the generalization of our result to general cases is trivial. Then the intermediate output $Y^{(t)}$ is forwarded to the FF module. It maps each row of the input $Y^{(t)} \in \mathbb{R}^{L\times d}$ through a same single-hidden layer neural network with d_F neurons, that is $\mathtt{ffn}(Y^{(t)}, A^{(t)}) = \mathtt{ReLU}(Y^{(t)}A_1^{(t)})A_2^{(t)}$, where $A_1^{(t)} \in \mathbb{R}^{d\times d_F}$, and $A_2^{(t)} \in \mathbb{R}^{d_F\times d}$ are the weight matrices. Combined with a residual link and layer normalization, it outputs the output of layer t as $X^{(t)}$, that is

$$Y^{(t)}=\Pi_{\mathrm{norm}}\big[\mathtt{mha}(X^{(t-1)},W^{(t)}) + \gamma_1^{(t)}X^{(t-1)}\big],\ X^{(t)}=\Pi_{\mathrm{norm}}\big[\mathtt{ffn}(Y^{(t)},A^{(t)}) + \gamma_2^{(t)}Y^{(t)}\big]. \tag{4.5.1}$$

Here we allocate weights $\gamma_1^{(t)}$ and $\gamma_2^{(t)}$ to residual links only for the convenience of theoretical analysis. In the last layer, the network outputs the probability of the next token via a softmax module, that is $Y^{(D+1)} = \mathtt{softmax}(\mathbb{I}_L^\top X^{(D)}A^{(D+1)}/(L\tau)) \in \mathbb{R}^{d_y}$, where $\mathbb{I}_L \in \mathbb{R}^L$ is the vector with all ones, $A^{(D+1)} \in \mathbb{R}^{d\times d_y}$ is the weight matrix, $\tau \in (0,1]$ is the fixed temperature parameter, and d_y is the output dimension. The parameters of each layer are denoted as $\theta^{(t)} = (\gamma_1^{(t)}, \gamma_2^{(t)}, W^{(t)}, A^{(t)})$ for $t \in [D]$ and $\theta^{(D+1)} = A^{(D+1)}$, and the parameter of the whole network is the concatenation of these parameters, i.e., $\theta = (\theta^{(1)}, \cdots, \theta^{(D+1)})$. We consider the transformers with bounded parameters. The set of parameters is

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

where B_A, $B_{A,1}$, $B_{A,2}$, B_Q, B_K, and B_V are the bounds of parameter. Here 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. The probability induced by the transformer with parameter θ is denoted as $\mathbb{P}_\theta$.

The pretraining dataset consists of N_{p} independent trajectories. For the n- th trajectory with $n \in [N_{\text{p}}]$, a hidden concept $z^n \sim \mathbb{P}_{\mathcal{Z}}(z) \in \Delta(\mathfrak{Z})$ is first sampled, which is the hidden variables of the token sequence to generate, e.g., the theme, the sentiment, and the style. Then the tokens are sequentially sampled from the Markov chain induced by z^n as $x_{t+1}^n \sim \mathbb{P}(\cdot \,|\, S_t^n, z^n)$ and $S_{t+1}^n = (S_t^n, x_{t+1}^n)$, where $x_{t+1}^n \in \mathfrak{X}$, and $S_t^n, S_{t+1}^n \in \mathfrak{X}^*$. Here the Markov chain is defined with respect to the state S_t^n, which obviously satisfies the Markov property since S_i^n for $i \in [t-1]$ are contained in S_t^n. The pretraining dataset is $\mathcal{D}_{N_{\text{p}},T_{\text{p}}} = \{(S_t^n, x_{t+1}^n)\}_{n,t=1}^{N_{\text{p}},T_{\text{p}}}$ where the concepts z^n is hidden from the context and thus unobserved. Here each token sequence is divided into T_{p} pieces $\{(S_t^n, x_{t+1}^n)\}_{t=1}^{T_{\text{p}}}$. We highlight that this pretraining dataset collecting process subsumes those for BERT, GPT, and even Masked AutoEncoders (MAE) (Radford et al., 2021). For GPT, each trajectory corresponds to a paragraph or an article in the pretraining dataset, and $z^n \sim \mathbb{P}_{\mathcal{Z}}(z)$ is realized by the selection process of these contexts from the Internet. For BERT, we just take $T_{\text{p}} = 1$. Then S_1^n and x_2^n respectively correspond to the sampled sentence and the masked token. For MAE, we take $T_{\text{p}} = 1$, and S_1^n and x_2^n respectively correspond to the image and the masked token.

To pretrain the transformer, we adopt the cross-entropy as the loss function, which is widely used in the training of BERT and GPT. The corresponding pretraining algorithm is

$$\widehat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} -\frac{1}{N_{\mathrm{p}} T_{\mathrm{p}}} \sum_{n=1}^{N_{\mathrm{p}}} \sum_{t=1}^{T_{\mathrm{p}}} \log \mathbb{P}_\theta(x_{t+1}^n \,|\, S_t^n). \tag{4.5.2}$$

We first analyze the population version of (4.5.2). In the training set, the conditional distribution of x_{t+1}^n conditioned on S_t^n is $\mathbb{P}(x_{t+1}^n \,|\, S_t^n) = \int_{\mathfrak{Z}} \mathbb{P}(x_{t+1}^n \,|\, S_t^n, z) \mathbb{P}_{\mathcal{Z}}(z \,|\, S_t^n) \mathrm{d}z$, where the unobserved hidden concept is weighed via its posterior distribution. Thus, the population risk of (4.5.2) is $\mathbb{E}_t[\mathbb{E}_{S_t}[\mathrm{KL}(\mathbb{P}(\cdot \,|\, S_t) \| \mathbb{P}_\theta(\cdot \,|\, S_t^n)) + H(\mathbb{P}(\cdot \,|\, S_t))]]$, where $t \sim \mathrm{Unif}([T_{\mathrm{p}}])$, and $H(p) = -\langle p, \log p \rangle$ is the entropy. Thus, we expect that $\mathbb{P}_\theta$ will converge to $\mathbb{P}$. For MAE, the network training adopts ℓ_2-loss, and we defer the analysis of this case to Appendix D.3.4.

4.5.2. Performance Guarantee for Pretraining

We first state the assumptions for the pretraining setting.

Assumption 4.5.1. There exists a constant $R > 0$ such that for any $z \in \mathfrak{Z}$ and $S_t \sim \mathbb{P}(\cdot \,|\, z)$, we have $\|S_t^\top\|_{2,\infty} \leq R$ almost surely.

This assumption states that the ℓ_2-norm of the magnitude of each token in the token sequence is upper bounded by $R > 0$. This assumption holds in most machine learning settings. For BERT and GPT, each token consists of word embedding and positional embedding. For MAE, each token consists of a patch of pixels. The ℓ_2-norm of each token is bounded in these cases.

Assumption 4.5.2. There exists a constant $c_0 > 0$ such that for any $z \in \mathfrak{Z}$, $x \in \mathfrak{X}$ and $S \in \mathfrak{X}^*$, we have $\mathbb{P}(x \,|\, S, z) \geq c_0$.

This assumption states that the conditional probability of x conditioned on S and z is lower bounded. This comes from the ambiguity of language, that is, a sentence can take lots of words as its next word. Similar regularity assumptions are also widely adopted in ICL literature (Xie et al., 2021; Wies et al., 2023). To state our result, we respectively use $\mathbb{E}_{S\sim\mathcal{D}}$ and $\mathbb{P}_{\mathcal{D}}$ to denote the expectation and the distribution of the average distribution of S_t^n in $\mathcal{D}_{N_{\mathrm{p}},T_{\mathrm{p}}}$, i.e., $\mathbb{E}_{S\sim\mathcal{D}}[f(S)] = \sum_{t=1}^{T_{\mathrm{p}}} \mathbb{E}_{S_t}[f(S_t)]/T_{\mathrm{p}}$ for any function $f : \mathfrak{X}^* \to \mathbb{R}$.

Theorem 4.5.3. Let $\bar{B} = \tau^{-1} R h B_A B_{A,1} B_{A,2} B_Q B_K B_V$ and $\bar{D} = D^2 d(d_F + d_h + d) + d \cdot d_y$. Under Assumptions 4.5.1 and 4.5.2, the pretrained model $\mathbb{P}_{\widehat{\theta}}$ by the algorithm in (4.5.2) satisfies

$$\mathbb{E}_{S\sim\mathcal{D}}\Big[\mathrm{TV}\left(\mathbb{P}(\cdot|\, S), \mathbb{P}_{\widehat{\theta}}(\cdot|\, S)\right)\Big]$$

$$= O\Bigg(\underbrace{\inf_{\theta^*\in\Theta} \sqrt{\mathbb{E}_{S\sim\mathcal{D}} \mathrm{KL}\big(\mathbb{P}(\cdot|S)\|\mathbb{P}_{\theta^*}(\cdot|S)\big)} + \frac{t_{\mathrm{mix}}^{1/4}\log 1/\delta}{(N_{\mathrm{p}}T_{\mathrm{p}})^{1/4}}}_{\text{approximation error}} + \underbrace{\frac{\sqrt{t_{\mathrm{mix}}}}{\sqrt{N_{\mathrm{p}}T_{\mathrm{p}}}}\Big(\bar{D}\log(1+N_{\mathrm{p}}T_{\mathrm{p}}\bar{B}) + \log\frac{1}{\delta}\Big)}_{\text{generalization error}}\Bigg)$$

with probability at least $1-\delta$, where t_{mix} is the mixing time of the Markov chains induced by $\mathbb{P}$, formally defined in Appendix D.3.1.

We define the right-hand side of the equation as $\Delta_{\mathrm{pre}}(N_{\mathrm{p}}, T_{\mathrm{p}}, \delta)$. The first and the second terms in the bound are the approximation error. It measures the distance between the nominal distribution $\mathbb{P}$ and the distributions induced by transformers with respect to KL divergence. If the nominal model $\mathbb{P}$ can be represented by transformers exactly, i.e., the realizable case, these two terms will vanish. The third term is the generalization error, and it does not increase with the growing sequence length T_{p}. If we use each token sequence once in pretraining, like BERT, this term is independent of T_{p}.

This pretraining analysis is missing in most existing theoretical works about ICL. Xie et al. (2021), Wies et al. (2023), and Jiang (2023) all assume access to an arbitrarily precise pretraining model. Although the generalization bound in Li et al. (2023) can be adapted to the pretraining analysis, the risk definition therein can not capture the approximation error in our result. Furthermore, their analysis cannot fit the maximum likelihood algorithm in (4.5.2). Concretely, their result can only show that the convergence rate of KL divergence is $O((N_{\mathrm{p}}T_{\mathrm{p}})^{-1/2})$ with a realizable function class. Combined with Pinsker's inequality, this gives the convergence rate for total variation as $O((N_{\mathrm{p}}T_{\mathrm{p}})^{-1/4})$ even in the realizable case.

The deep neural networks are shown to be universal approximators for many function classes (Cybenko, 1989; Hornik, 1991; Yarotsky, 2017). Thus, the approximation error in Theorem 4.5.3 should vanish with the increasing size of the transformer. To achieve this, we slightly change the structure of the transformer by admitting a bias term in feed-forward modules, taking $A_2^{(t)} \in \mathbb{R}^{d_F \times d_F}$, and admitting d_F to vary across layers. This mildly affects the generalization error by replacing $D \cdot d_F$ by the sum of d_F of all the layers in Theorem 4.5.3. We derive the approximation error bound when the dimension of each word is 1, i.e., $\mathfrak{X} \subseteq \mathbb{R}$. Our method can carry over the case $d > 1$.

Proposition 4.5.4 (Informal). Under certain smoothness conditions, 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

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

for some constant $C > 0$.

The formal statement and proof are deferred to Appendix D.3.3. This proposition states that the approximation error decays exponentially with the increasing depth. Combined with this result, Theorem 4.5.3 provides the full description of the pretraining performance.

4.6. ICL Regret under Practical Settings

4.6.1. ICL Regret with an Imperfectly Pretrained Model

In Section 4.4, we study the ICL regret with a perfect pretrained model. In what follows, we characterize the ICL regret when the pretrained model has an error. Note that the distribution $\mathcal{D}_{\texttt{ICL}}$ of the prompts of ICL tasks can be different from that of pretraining. We impose the following assumption on their relation.

Assumption 4.6.1. We assume that there exists an absolute constant $\kappa > 0$ such that for any ICL prompt, it holds that $\mathbb{P}_{\mathcal{D}_{\texttt{ICL}}}(\texttt{prompt}) \leq \kappa \cdot \mathbb{P}_{\mathcal{D}}(\texttt{prompt})$.

This assumption states that the prompt distribution is covered by the pretraining distribution. Intuitively, the pretrained model cannot precisely inference on the datapoint that is outside the support of the pretraining distribution. For example, if the pretraining data does not contain any mathematical symbols and numbers, it is difficult for the pretrained model to calculate 2 × 3 in ICL precisely. We then have the following theorem characterizing the ICL regret of the pretrained model.

Theorem 4.6.2 (ICL Regret of Pretrained Model)**.** We assume that the underlying hidden concept z_* maximizes $\sum_{i=1}^{t} \log \mathbb{P}(r_i \,|\, \texttt{prompt}_{i-1}, z)$ for any $t \in [T]$ and there exists an absolute constant $\beta > 0$ such that $\log(1/p_0(z_*)) \leq \beta$. Under Assumptions 4.5.1, 4.5.2, and 4.6.1, we have with probability at least $1-\delta$ that

$$\mathbb{E}_{\texttt{prompt}\sim\mathcal{D}_{\texttt{ICL}}}\Big[T^{-1}\cdot\sum_{t=1}^{T}\log\mathbb{P}(r_t\,|\,z^*,\texttt{prompt}_{t-1})-T^{-1}\cdot\sum_{t=1}^{T}\log\mathbb{P}_{\widehat{\theta}}(r_t\,|\,\texttt{prompt}_{t-1})\Big]$$
$$\leq\mathcal{O}\big(\beta/T+\kappa\cdot b^*\cdot\Delta_{\mathrm{pre}}(N_{\mathrm{p}},T_{\mathrm{p}},\delta)\big).$$

Here we denote by $\Delta_{\mathrm{pre}}(N_{\mathrm{p}},T_{\mathrm{p}},\delta)$ the pretraining error in Theorem 4.5.3.

Theorem 4.6.2 shows that the expected ICL regret for the pretrained model is upper bounded by the sum of two terms: (a) the ICL regret for the underlying true model and the error and (b) the pretraining error.

4.6.2. Prompting with Wrong Input-Output Mappings

In the real-world implementations of ICL, the provided input-output examples may not conform to the nominal distribution induced by z_*, and the outputs in examples can be *perturbed*. We temporarily take concept space $\mathfrak{Z}$ as a finite space, and our results can be generalized with the cover number argument. We denote the prompt considered in Section 4.4 as $\texttt{prompt}_t=(S_t,\widetilde{c}_{t+1})$, $S_t=(\widetilde{c}_1,r_1,\cdots,\widetilde{c}_t,r_t)\in\mathfrak{X}^*$, and $(\widetilde{c}_{i+1},r_{i+1})\sim\mathbb{P}(\cdot\,|\,S_i,z_*)$ for $i\in[t-1]$. Here, each input $\widetilde{c}_i\in\mathfrak{X}^l$ is a l-length token sequence, and each output $r_i\in\mathfrak{X}$ is a single token. The perturbed prompt is then denoted as $\texttt{prompt}'=(S'_t,\widetilde{c}_{t+1})$, where $S'_t=(\widetilde{c}_1,r'_1,\cdots,\widetilde{c}_t,r'_t)\in\mathfrak{X}^*$, and r'_i for $i\in[t]$ is the modified output. We denote the perturbed prompt distribution as $\mathbb{P}'$. Next, we state assumptions for this setting.

Assumption 4.6.3. 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 $\mathbb{P}((S_t,S_{t'})\,|\,z)=\mathbb{P}(S_t\,|\,z)\cdot\mathbb{P}(S_{t'}\,|\,z)$.

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 4.6.4. There exists a constant $c_1 > 0$ such that $\mathbb{P}_{\mathcal{Z}}(z_*) \geq c_1$.

This assumption states that the prior distribution of the hidden concept z_* is strictly larger than 0, otherwise this concept can never be deduced.

Assumption 4.6.5. There exists a constant $c_2 > 0$ such that for any $\texttt{prompt}' \in \mathfrak{X}^*$, it holds that $\mathbb{P}'(\texttt{prompt}')/\mathbb{P}_{\mathcal{D}}(\texttt{prompt}') \leq c_2$.

Similar to Assumption 4.6.1, this assumption states that the distribution of the perturbed prompt is covered by the pretraining distribution. For two concepts $z, z' \in \mathfrak{Z}$, we define the KL divergence between the conditional distributions of input-output pair on them as $\mathrm{KL}_{\mathrm{pair}}(\mathbb{P}(\cdot \mid z) \| \mathbb{P}(\cdot \mid z')) = \mathbb{E}_{X,y \sim \mathbb{P}(\cdot \mid z)}[\log(\mathbb{P}(X, y \mid z)/\mathbb{P}(X, y \mid z'))]$. This divergence measures the distance between distributions of input-output pairs conditioned on different tasks z and z'.

Assumption 4.6.6. The concept z_* satisfies that $\min_{z \neq z_*} \mathrm{KL}_{\mathrm{pair}}(\mathbb{P}(\cdot \mid z_*) \,\|\, \mathbb{P}(\cdot \mid z)) > 2 \log 1/c_0$, where c_0 is the constant in Assumption 4.5.2.

This distinguishability assumption requires that the divergence between z_* and other concepts z is large enough to infer the concept z_* from the prompt.

Proposition 4.6.7. Under Assumptions 4.5.2, 4.6.3, 4.6.4, 4.6.5 and 4.6.6, the pretrained model $\mathbb{P}_{\widehat{\theta}}$ in (4.5.2) predicts the outputs with the prompt containing wrong mappings as

$$\mathbb{E}_{\texttt{prompt}' \sim \mathbb{P}'}\Big[\mathrm{KL}\big(\mathbb{P}(\cdot \mid \widetilde{c}_{t+1}, z_*) \| \mathbb{P}_{\widehat{\theta}}(\cdot \mid S'_t, \widetilde{c}_{t+1})\big)\Big]$$
$$= \mathcal{O}\bigg(c_2 \Delta_{\mathrm{pre}}(N_{\mathrm{p}}, T_{\mathrm{p}}, \delta) + \exp\bigg(- t\bigg(\min_{z \neq z_*} \mathrm{KL}_{\mathrm{pair}}\big(\mathbb{P}(\cdot \mid z_*) \| \mathbb{P}(\cdot \mid z)\big) + 2 \log c_0 - \frac{2l \log 1/c_0}{\sqrt{t}} \log \frac{|\mathfrak{Z}|}{\delta}\bigg)\bigg)\bigg)$$

with probability at least $1-\delta$.

The first term is the pretraining error, which is related to the size of the pretraining set and the capacity of the neural networks. The second term is the ICL error. Intuitively, this term represents the concept identification error. If the considered task z_* is distinguishable, i.e., satisfying Assumption 4.6.6, this term decays to 0 exponentially in t.

4.7. 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 Bayesian model averaging 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 Bayesian model averaging 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.

References

Agarwal, A., Kakade, S., Krishnamurthy, A. and Sun, W. (2020). Flambe: Structural complexity and representation learning of low rank MDPs. *Advances in Neural Information Processing Systems*, **33** 20095–20107.

Agarwal, A., Kakade, S. M., Lee, J. D. and Mahajan, G. (2019). Optimality and approximation with policy gradient methods in Markov decision processes. *arXiv preprint arXiv:1908.00261*.

Agazzi, A. and Lu, J. (2019). Temporal-difference learning for nonlinear value function approximation in the lazy training regime. *arXiv preprint arXiv:1905.10917*.

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

Allen-Zhu, Z., Li, Y. and Liang, Y. (2018a). Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*.

Allen-Zhu, Z., Li, Y. and Liang, Y. (2019a). Learning and generalization in overparameterized neural networks, going beyond two layers. In *Neural Information Processing Systems*.

Allen-Zhu, Z., Li, Y. and Song, Z. (2018b). A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*.

Allen-Zhu, Z., Li, Y. and Song, Z. (2019b). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*.

Allen-Zhu, Z., Li, Y. and Song, Z. (2019c). On the convergence rate of training recurrent neural networks. In *Neural Information Processing Systems*.

Altman, E. (1999). *Constrained Markov decision processes*, vol. 7. CRC Press.

Ambrosio, L. and Gigli, N. (2013). A user's guide to optimal transport. In *Modelling and Optimisation of Flows on Networks*. Springer, 1–155.

Ambrosio, L., Gigli, N. and Savaré, G. (2008). *Gradient flows: In metric spaces and in the space of probability measures*. Springer.

Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. Cambridge University Press.

Anthony, M., Bartlett, P. L., Bartlett, P. L. et al. (1999). *Neural network learning: Theoretical foundations*, vol. 9. cambridge university press Cambridge.

Antos, A., Szepesvári, C. and Munos, R. (2008). Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, **71** 89–129.

Araújo, D., Oliveira, R. I. and Yukimura, D. (2019). A mean-field limit for certain deep neural networks. *arXiv preprint arXiv:1906.00193*.

Arjovsky, M., Chintala, S. and Bottou, L. (2017). Wasserstein GAN. *arXiv preprint arXiv:1701.07875*.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R. and Wang, R. (2019a). On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*.

Arora, S., Du, S. S., Hu, W., Li, Z. and Wang, R. (2019b). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*.

Bai, Y. and Lee, J. D. (2019). Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*.

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*. Elsevier, 30–37.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, **39** 930–945.

Bartlett, P. (1996). For valid generalization the size of the weights is more important than the size of the network. *Neural Information Processing Systems*.

Bartlett, P., Helmbold, D. and Long, P. (2018a). Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In *International Conference on Machine Learning*.

Bartlett, P. L., Evans, S. N. and Long, P. M. (2018b). Representing smooth functions as compositions of near-identity functions with implications for deep network optimization. *arXiv preprint arXiv:1804.05012*.

Bartlett, P. L., Foster, D. J. and Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R. et al. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A. and Hjelm, D. (2018). Mutual information neural estimation. In *International Conference on Machine Learning*. PMLR.

Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. In *ICML Workshop on Unsupervised and Transfer Learning*.

Bertsekas, D. P. (2019). Feature-based aggregation and deep reinforcement learning: A survey and some new implementations. *IEEE/CAA Journal of Automatica Sinica*, **6** 1–31.

Bhandari, J. and Russo, D. (2019). Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*.

Bhandari, J., Russo, D. and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*.

Bhatnagar, S., Precup, D., Silver, D., Sutton, R. S., Maei, H. R. and Szepesvári, C. (2009). Convergent temporal-difference learning with arbitrary smooth function approximation. In *Advances in Neural Information Processing Systems*.

Bloem-Reddy, B. and Teh, Y. W. (2020). Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*.

Borkar, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint*, vol. 48. Springer.

Borkar, V. S. and Meyn, S. P. (2000). The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, **38** 447–469.

Boyan, J. A. and Moore, A. W. (1995). Generalization in reinforcement learning: Safely approximating the value function. In *Advances in Neural Information Processing Systems*.

Brandfonbrener, D. and Bruna, J. (2019a). Geometric insights into the convergence of nonlinear TD learning. *arXiv preprint arXiv:1905.12185*.

Brandfonbrener, D. and Bruna, J. (2019b). On the expected dynamics of nonlinear TD learning. *arXiv preprint arXiv:1905.12185*.

Bronstein, M. M., Bruna, J., Cohen, T. and Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Neural Information Processing Systems*.

Cai, Q., Hong, M., Chen, Y. and Wang, Z. (2019a). On the global convergence of imitation learning: A case for linear quadratic regulator. *arXiv preprint arXiv:1901.03674*.

Cai, Q., Yang, Z., Jin, C. and Wang, Z. (2019b). Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*.

Cai, Q., Yang, Z., Lee, J. D. and Wang, Z. (2019c). Neural temporal-difference learning converges to global optima. *arXiv preprint arXiv:1905.10027*.

Cao, Y. and Gu, Q. (2019a). Generalization bounds of stochastic gradient descent for wide and deep neural networks. *arXiv preprint arXiv:1905.13210*.

Cao, Y. and Gu, Q. (2019b). Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Neural Information Processing Systems*.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*.

Chan, S. C., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A., Richemond, P. H., McClelland, J. and Hill, F. (2022). Data distributional properties drive emergent few-shot learning in transformers. *arXiv preprint arXiv:2205.05055*.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A. and Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Neural Information Processing Systems*.

Chen, M., Wang, Y., Liu, T., Yang, Z., Li, X., Wang, Z. and Zhao, T. (2020a). On computation and generalization of generative adversarial imitation learning. *arXiv preprint arXiv:2001.02792*.

Chen, Z., Cao, Y., Gu, Q. and Zhang, T. (2020b). Mean-field analysis of two-layer neural networks: Non-asymptotic rates and generalization bounds. *arXiv preprint arXiv:2002.04026*.

Chen, Z., Cao, Y., Zou, D. and Gu, Q. (2019a). How much over-parameterization is sufficient to learn deep ReLU networks? *arXiv preprint arXiv:1911.12360*.

Chen, Z., Zhang, S., Doan, T. T., Maguluri, S. T. and Clarke, J.-P. (2019b). Performance of q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv:1905.11425*.

Chizat, L. and Bach, F. (2018a). A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*.

Chizat, L. and Bach, F. (2018b). On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*.

Chizat, L., Oyallon, E. and Bach, F. (2019). On lazy training in differentiable programming. *Neural Information Processing Systems*.

Conway, J. B. (2019). *A course in functional analysis*, vol. 96. Springer.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, **2** 303–314.

Dai, D., Sun, Y., Dong, L., Hao, Y., Sui, Z. and Wei, F. (2022). Why can GPT learn In-Context? Language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V. and Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Dalal, G., Szörényi, B., Thoppe, G. and Mannor, S. (2018). Finite sample analyses for TD(0) with function approximation. In *AAAI Conference on Artificial Intelligence*.

Daniely, A. (2017). SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*.

Dann, C., Neumann, G. and Peters, J. (2014). Policy evaluation with temporal differences: A survey and comparison. *The Journal of Machine Learning Research*, **15** 809–883.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*.

De Gooijer, J. G. and Zerom, D. (2003). On conditional density estimation. *Statistica Neerlandica*.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequences. *Annals of Probability*.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M. and Hon, H.-W. (2019). Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, **32**.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J. and Sui, Z. (2022). A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, S., Lee, J., Li, H., Wang, L. and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*.

Du, S. S., Lee, J. D., Li, H., Wang, L. and Zhai, X. (2018a). Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*.

Du, S. S., Zhai, X., Poczos, B. and Singh, A. (2018b). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*.

Duchi, J. C. (2019). Information theory and statistics. *Lecture Notes for Statistics*, **311** 304.

Edelman, B. L., Goel, S., Kakade, S. and Zhang, C. (2021). Inductive biases and variable creation in self-attention mechanisms. *arXiv preprint arXiv:2110.10090*.

Elbrächter, D., Perekrestenko, D., Grohs, P. and Bölcskei, H. (2021). Deep neural network approximation theory. *IEEE Transactions on Information Theory*, **67** 2581–2623.

Elesedy, B. (2021). Provably strict generalisation benefit for invariance in kernel methods. *Neural Information Processing Systems*.

Fang, C., Dong, H. and Zhang, T. (2019a). Over parameterized two-level neural networks can learn near optimal feature representations. *arXiv preprint arXiv:1910.11508*.

Fang, C., Gu, Y., Zhang, W. and Zhang, T. (2019b). Convex formulation of overparameterized deep neural networks. *arXiv preprint arXiv:1911.07626*.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C. and Mannor, S. (2016). Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, **17** 4809–4874.

Farahmand, A.-m., Szepesvári, C. and Munos, R. (2010). Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*.

Finn, C., Levine, S. and Abbeel, P. (2016). Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, **222** 309–368.

Fukumizu, K. (2015). Nonparametric bayesian inference with kernel mean embedding. In *Modern Methodology and Applications in Spatial-Temporal Modeling*. Springer, 1–24.

Garg, S., Tsipras, D., Liang, P. and Valiant, G. (2022). What can transformers learn in-context? A case study of simple function classes. *arXiv preprint arXiv:2208.01066*.

Geist, M. and Pietquin, O. (2013). Algorithmic survey of parametric value function approximation. *IEEE Transactions on Neural Networks and Learning Systems*, **24** 845–867.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*.

Haarnoja, T., Tang, H., Abbeel, P. and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*.

Han, J., Rong, Y., Xu, T. and Huang, W. (2022). Geometrically equivariant graph neural networks: A survey. *arXiv preprint arXiv:2202.07230*.

Hardt, M. and Ma, T. (2016). Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*.

Harker, P. T. and Pang, J.-S. (1990). Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications. *Mathematical Programming*, **48** 161–220.

He, J., Chen, J., He, X., Gao, J., Li, L., Deng, L. and Ostendorf, M. (2015). Deep reinforcement learning with a natural language action space. *arXiv preprint arXiv:1511.04636*.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Computer Vision and Pattern Recognition*.

Hinton, G. (1986). Learning distributed representations of concepts. In *Annual Conference of Cognitive Science Society*.

Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*.

Hofmann, T., Schölkopf, B. and Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics* 1171–1220.

Holte, J. M. (2009). Discrete Gronwall lemma and applications. In *MAA-NCS meeting at the University of North Dakota*, vol. 24.

Honovich, O., Shaham, U., Bowman, S. R. and Levy, O. (2022). Instruction induction: From few examples to natural language task descriptions. *arXiv preprint arXiv:2205.10782*.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, **4** 251–257.

Hron, J., Bahri, Y., Sohl-Dickstein, J. and Novak, R. (2020). Infinite attention: NNGP and NTK for deep attention networks. In *International Conference on Machine Learning*.

Hutchinson, M. J., Le Lan, C., Zaidi, S., Dupont, E., Teh, Y. W. and Kim, H. (2021). Lietransformer: Equivariant self-attention for Lie groups. In *International Conference on Machine Learning*.

Iyer, S., Lin, X. V., Pasunuru, R., Mihaylov, T., Simig, D., Yu, P., Shuster, K., Wang, T., Liu, Q., Koura, P. S. et al. (2022). OPT-IML: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Jaakkola, T., Jordan, M. I. and Singh, S. P. (1994). Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems*.

Jacot, A., Gabriel, F. and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*.

Javanmard, A., Mondelli, M. and Montanari, A. (2019). Analysis of a two-layer neural network via displacement convexity. *arXiv preprint arXiv:1901.01375*.

Ji, Z. and Telgarsky, M. (2019). Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. *arXiv preprint arXiv:1909.12292*.

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

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D. and Bengio, S. (2019). Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F. and Liu, Q. (2019). Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Jin, C., Yang, Z., Wang, Z. and Jordan, M. I. (2019). Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*.

Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, vol. 2.

Kakade, S. M. (2002). A natural policy gradient. In *Advances in Neural Information Processing Systems*.

Ke, G., He, D. and Liu, T.-Y. (2020). Rethinking positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595*.

Keriven, N. and Peyré, G. (2019). Universal invariant and equivariant graph neural networks. *Neural Information Processing Systems*.

Kim, H. J., Cho, H., Kim, J., Kim, T., Yoo, K. M. and Lee, S.-g. (2022). Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. *arXiv preprint arXiv:2206.08082*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in Neural Information Processing Systems*.

Kossen, J., Band, N., Lyle, C., Gomez, A. N., Rainforth, T. and Gal, Y. (2021). Self-attention between datapoints: Going beyond individual input-output pairs in deep learning. *Neural Information Processing Systems*.

Kuefler, A., Morton, J., Wheeler, T. and Kochenderfer, M. (2017). Imitating driver behavior with generative adversarial networks. In *IEEE Intelligent Vehicles Symposium*. IEEE.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*. Springer.

Lakshminarayanan, C. and Szepesvári, C. (2018). Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*.

Lazaric, A., Ghavamzadeh, M. and Munos, R. (2016). Analysis of classification-based policy iteration algorithms. *The Journal of Machine Learning Research*, **17** 583–612.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, **521** 436–444.

Ledent, A., Mustafa, W., Lei, Y. and Kloft, M. (2021). Norm-based generalisation bounds for deep multi-class convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: isoperimetry and processes*, vol. 23. Springer Science & Business Media.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S. and Teh, Y. W. (2019a). Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J. and Pennington, J. (2019b). Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*.

Levine, S., Finn, C., Darrell, T. and Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, **17** 1334–1373.

Levine, S. and Koltun, V. (2012). Continuous inverse optimal control with locally optimal examples. *arXiv preprint arXiv:1206.4617*.

Li, Y., Ildiz, M. E., Papailiopoulos, D. and Oymak, S. (2023). Transformers as algorithms: Generalization and stability in in-context learning. *arXiv preprint arXiv:2301.07067*.

Li, Y. and Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*.

Liao, R., Urtasun, R. and Zemel, R. (2020). A pac-bayesian approach to generalization bounds for graph neural networks. *arXiv preprint arXiv:2012.07690*.

Lin, S. and Zhang, J. (2019). Generalization bounds for convolutional neural networks. *arXiv preprint arXiv:1910.01487*.

Liu, B., Cai, Q., Yang, Z. and Wang, Z. (2019). Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L. and Chen, W. (2021). What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, **55** 1–35.

Lu, Y., Bartolo, M., Moore, A., Riedel, S. and Stenetorp, P. (2021). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Malladi, S., Wettig, A., Yu, D., Chen, D. and Arora, S. (2022). A kernel-based view of language model fine-tuning. *arXiv preprint arXiv:2210.05643*.

Maurer, A. (2016). A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*. Springer.

Mei, S., Misiakiewicz, T. and Montanari, A. (2019). Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit. *arXiv preprint arXiv:1902.06015*.

Mei, S., Montanari, A. and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, **115** E7665–E7671.

Melo, F. S., Meyn, S. P. and Ribeiro, M. I. (2008). An analysis of reinforcement learning with function approximation. In *International Conference on Machine Learning*.

Merel, J., Tassa, Y., TB, D., Srinivasan, S., Lemmon, J., Wang, Z., Wayne, G. and Heess, N. (2017). Learning human behaviors from motion capture by adversarial imitation. *arXiv preprint arXiv:1707.02201*.

Michel, P., Levy, O. and Neubig, G. (2019). Are sixteen heads really better than one? *Advances in neural information processing systems*, **32**.

Min, S., Lewis, M., Zettlemoyer, L. and Hajishirzi, H. (2021). Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H. and Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G. et al. (2015). Human-level control through deep reinforcement learning. *Nature*, **518** 529–533.

Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Muandet, K., Fukumizu, K., Sriperumbudur, B. and Schölkopf, B. (2016). Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*.

Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *The Journal of Machine Learning Research*, **9** 815–857.

Nachum, O., Norouzi, M., Xu, K. and Schuurmans, D. (2017). Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*.

Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer Science & Business Media.

Neyshabur, B., Bhojanapalli, S., McAllester, D. and Srebro, N. (2017). A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*.

Ng, A. Y. and Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*.

Nguyen, P.-M. (2019). Mean field limit of the learning dynamics of multilayer neural networks. *arXiv preprint arXiv:1902.02880*.

Noci, L., Anagnostidis, S., Biggio, L., Orvieto, A., Singh, S. P. and Lucchi, A. (2022). Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *arXiv preprint arXiv:2206.03126*.

O'Donoghue, B., Munos, R., Kavukcuoglu, K. and Mnih, V. (2016). Combining policy gradient and Q-learning. *arXiv preprint arXiv:1611.01626*.

Otto, F. and Villani, C. (2000). Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, **173** 361–400.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, **35** 27730–27744.

Paulin, D. (2015). Concentration inequalities for markov chains by marton couplings and spectral methods.

Pearl, J. (2009). *Causality*. Cambridge University press.

Peters, J. and Schaal, S. (2008). Natural actor-critic. *Neurocomputing*, **71** 1180–1190.

Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numerica*, **8** 143–195.

Pomerleau, D. A. (1991). Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, **3** 88–97.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018). Improving language understanding by generative pre-training. *Technical Report*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*.

Rafique, H., Liu, M., Lin, Q. and Yang, T. (2018). Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv:1810.02060*.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*.

Rahimi, A. and Recht, B. (2009). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in Neural Information Processing Systems* 1313–1320.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I. (2021). Zero-shot text-to-image generation. In *International Conference on Machine Learning*.

Romero, D. W. and Cordonnier, J.-B. (2020). Group equivariant stand-alone self-attention for vision. *arXiv preprint arXiv:2010.00977*.

Ross, S. and Bagnell, D. (2010). Efficient reductions for imitation learning. In *International Conference on Artificial Intelligence and Statistics*.

Ross, S., Gordon, G. and Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*.

Rotskoff, G. and Vanden-Eijnden, E. (2018). Parameters as interacting particles: Long time convergence and asymptotic error scaling of neural networks. *Neural Information*

Processing Systems.

Rubin, O., Herzig, J. and Berant, J. (2021). Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633.*

Russell, S. (1998). Learning agents for uncertain environments. In *Conference on Learning Theory.*

Sannai, A., Imaizumi, M. and Kawano, M. (2021). Improved generalization bounds of group invariant/equivariant deep networks via quotient feature spaces. In *Uncertainty in Artificial Intelligence.*

Satorras, V. G., Hoogeboom, E., Fuchs, F. B., Posner, I. and Welling, M. (2021). $E(n)$ equivariant normalizing flows. *arXiv preprint arXiv:2105.09016.*

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. and Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks.*

Scherrer, B., Ghavamzadeh, M., Gabillon, V., Lesner, B. and Geist, M. (2015). Approximate modified policy iteration and its application to the game of Tetris. *The Journal of Machine Learning Research*, **16** 1629–1676.

Schulman, J., Chen, X. and Abbeel, P. (2017). Equivalence between policy gradients and soft Q-learning. *arXiv preprint arXiv:1704.06440.*

Schulz, E., Speekenbrink, M. and Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology.*

Shawe-Taylor, J., Cristianini, N. et al. (2004). *Kernel methods for pattern analysis.* Cambridge University Press.

Singh, R., Sahani, M. and Gretton, A. (2019). Kernel instrumental variable regression. *Advances in Neural Information Processing Systems.*

Sirignano, J. and Spiliopoulos, K. (2019). Asymptotics of reinforcement learning with neural networks. *arXiv preprint arXiv:1911.07304.*

Sirignano, J. and Spiliopoulos, K. (2020). Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, **80** 725–752.

Sokolic, J., Giryes, R., Sapiro, G. and Rodrigues, M. (2017). Generalization error of invariant classifiers. In *Artificial Intelligence and Statistics.*

Song, L., Anandkumar, A., Dai, B. and Xie, B. (2014). Nonparametric estimation of multi-view latent variable models. In *International Conference on Machine Learning*.

Song, L., Huang, J., Smola, A. and Fukumizu, K. (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *International Conference on Machine Learning*.

Srikant, R. and Ying, L. (2019). Finite-time error bounds for linear stochastic approximation and TD learning. *arXiv preprint arXiv:1902.00923*.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, **3** 9–44.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Sutton, R. S., McAllester, D. A., Singh, S. P. and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*.

Syed, U., Bowling, M. and Schapire, R. E. (2008). Apprenticeship learning using linear programming. In *International Conference on Machine Learning*.

Szepesvári, C. and Munos, R. (2005). Finite time bounds for sampling based fitted value iteration. In *International Conference on Machine Learning*. ACM.

Sznitman, A.-S. (1991). Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*. Springer, 165–251.

Tai, L., Zhang, J., Liu, M. and Burgard, W. (2018). Socially compliant navigation through raw depth inputs with generative adversarial imitation learning. In *IEEE International Conference on Robotics and Automation*. IEEE.

Tsai, Y.-H. H., Bai, S., Yamada, M., Morency, L.-P. and Salakhutdinov, R. (2019). Transformer dissection: A unified understanding of transformer's attention via the lens of kernel. *arXiv preprint arXiv:1908.11775*.

Tsitsiklis, J. N. and Van Roy, B. (1997). Analysis of temporal-diffference learning with function approximation. In *Advances in Neural Information Processing Systems*.

Valle-Pérez, G. and Louis, A. A. (2020). Generalization bounds for deep learning. *arXiv preprint arXiv:2012.04115*.

van de Geer, S. and Muro, A. (2014). On higher order isotropy conditions and lower bounds for sparse quadratic forms. *Electronic Journal of Statistics*, **8** 3031–3061.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need. In *Neural Information Processing Systems*.

Villani, C. (2003). *Topics in optimal transportation*. American Mathematical Society.

Villani, C. (2008). *Optimal transport: Old and new*. Springer.

von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A. and Vladymyrov, M. (2022). Transformers learn in-context by gradient descent. *arXiv preprint arXiv:2212.07677*.

Vuckovic, J., Baratin, A. and Combes, R. T. d. (2020). A mathematical theory of attention. *arXiv preprint arXiv:2007.02876*.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.

Wang, L., Cai, Q., Yang, Z. and Wang, Z. (2019). Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*.

Wang, X., Zhu, W. and Wang, W. Y. (2023). Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D. and Hajishirzi, H. (2022). Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44** 92–107.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, **8** 279–292.

Wei, C., Chen, Y. and Ma, T. (2021a). Statistically meaningful approximation: A case study on approximating Turing machines with transformers. *arXiv preprint arXiv:2107.13163*.

Wei, C., Chen, Y. and Ma, T. (2022a). Statistically meaningful approximation: a case study on approximating turing machines with transformers. *Advances in Neural Information Processing Systems*, **35** 12071–12083.

Wei, C., Lee, J. D., Liu, Q. and Ma, T. (2019). Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M. and Le, Q. V. (2021b). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. et al. (2022b). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q. and Zhou, D. (2022c). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Wies, N., Levine, Y. and Shashua, A. (2023). The learnability of in-context learning. *arXiv preprint arXiv:2303.07895*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, **8** 229–256.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. et al. (2020). Transformers: State-of-the-art natural language processing. In *Empirical Methods in Natural Language Processing*.

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

Xu, P., Gao, F. and Gu, Q. (2019a). An improved convergence analysis of stochastic variance-reduced policy gradient. *arXiv preprint arXiv:1905.12615*.

Xu, P., Gao, F. and Gu, Q. (2019b). Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*.

Xu, T., Zou, S. and Liang, Y. (2019c). Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. In *Advances in Neural Information Processing Systems*.

Yang, G. (2020). Tensor programs II: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548.*

Yang, G. and Littwin, E. (2021). Tensor programs IIb: Architectural universality of neural tangent kernel training dynamics. In *International Conference on Machine Learning.*

Yang, L. and Wang, M. (2019a). Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning.*

Yang, L. F. and Wang, M. (2019b). Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389.*

Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, **94** 103–114.

Yosinski, J., Clune, J., Bengio, Y. and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems.*

Yu, L., Zhang, W., Wang, J. and Yu, Y. (2016). SeqGAN: Sequence generative adversarial nets with policy gradient. *arXiv preprint arXiv:1609.05473.*

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S. J. and Kumar, S. (2019). Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077.*

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R. and Smola, A. J. (2017). Deep sets. *Neural Information Processing Systems.*

Zhang, F., Liu, B., Wang, K., Tan, V. Y., Yang, Z. and Wang, Z. (2022a). Relational reasoning via set transformers: Provable efficiency and applications to MARL. *arXiv preprint arXiv:2209.09845.*

Zhang, K., Koppel, A., Zhu, H. and Başar, T. (2019a). Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383.*

Zhang, K., Yang, Z. and Başar, T. (2019b). Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. *arXiv preprint arXiv:1906.00729.*

Zhang, X., Yu, Y., Wang, L. and Gu, Q. (2019c). Learning one-hidden-layer ReLU networks via gradient descent. In *International Conference on Machine Learning.*

Zhang, Y., Backurs, A., Bubeck, S., Eldan, R., Gunasekar, S. and Wagner, T. (2022b). Unveiling transformers with LEGO: A synthetic reasoning task. *arXiv preprint* *arXiv:2206.04301.*

Zhang, Y., Duchi, J. and Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, **16** 3299–3340.

Zhang, Z., Zhang, A., Li, M. and Smola, A. (2022c). Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493.*

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Bousquet, O., Le, Q. and Chi, E. (2022a). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625.*

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H. and Ba, J. (2022b). Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910.*

Zhu, S., An, B. and Huang, F. (2021). Understanding the generalization benefit of model invariance from a data perspective. *Neural Information Processing Systems.*

Zou, D., Cao, Y., Zhou, D. and Gu, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *arXiv preprint arXiv:1811.08888.*

Zou, D. and Gu, Q. (2019). An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems.*

Zou, S., Xu, T. and Liang, Y. (2019). Finite-sample analysis for sarsa and q-learning with linear function approximation. *arXiv preprint arXiv:1902.02234.*

APPENDIX A

Generative Adversarial Imitation Learning with Neural Networks: Global Optimality and Convergence Rate

A.1. Neural Networks

In what follows, we present the properties of the neural network defined in (1.3.1). First, we define the following function class.

Definition A.1.1 (Function Class)**.** For $B > 0$ and $m \in \mathbb{N}_+$, we define

$$\mathcal{F}_{B,m} = \big\{ W^\top \phi_0(s,a) \,\big|\, W \in \mathbb{R}^{md},\ \|W - W_0\|_2 \le B \big\},$$

where $\phi_0(s,a)$ is the feature vector defined in (1.3.2) with $W = W_0$.

As shown in Rahimi and Recht (2008), the feature $\phi_0(s,a)$ induces a reproducing kernel Hilbert space (RKHS), namely $\mathcal{H}$. When m goes to infinity, $\mathcal{F}_{B,m}$ approximates a ball in $\mathcal{H}$, which captures a rich class of functions (Hofmann et al., 2008; Rahimi and Recht, 2008). Furthermore, we obtain the following lemma from Cai et al. (2019c), which characterizes the linearization error of the neural network defined in (1.3.1).

Lemma A.1.2 (Linearization Error, Lemma 5.1 in Cai et al. (2019c))**.** Under Assumption 1.4.1, it holds for any $W, W_1, W_2 \in S_B$ that,

$$\mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_1}(s,a) - W^\top \phi_{W_2}(s,a)\big\|_{2,\mu}^2\Big] = \mathcal{O}(B^3 \cdot m^{-1/2}),$$
$$\mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_1}(s,a) - W^\top \phi_{W_2}(s,a)\big\|_{1,\mu}\Big] = \mathcal{O}(B^{3/2} \cdot m^{-1/4}),$$

where $\phi_W(s,a)$ is the feature vector defined in (1.3.2) and $\mu \in \mathscr{P}(\mathcal{S} \times \mathcal{A})$ is a distribution that satisfies Assumption 1.4.1.

Proof. See §A.1.1 for a detailed proof. □

Following from Lemma A.1.2, the function class $\mathcal{F}_{B,m}$ defined in Definition A.1.1 is a first-order approximation of the class of the neural networks defined in (1.3.1). Meanwhile, we establish the following lemma to characterize the sub-Gaussian property of the neural network defined in (1.3.1).

Lemma A.1.3. Under Assumption 1.4.2, for any $W, W' \in S_B$, it holds that $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |W^\top \phi_{W'}(s,a)|$ is sub-Gaussian, where the randomness comes from the random initialization W_0 in the definition of S_B in (1.3.4). Moreover, it holds that

$$\mathbb{E}_{\text{init}}\Big[\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|W^\top \phi_{W'}(s,a)\big|^2\Big] \le 2M_0^2 + 18B^2$$

and that

$$\mathbb{P}\Big(\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|W^\top \phi_{W'}(s,a)\big| > t\Big) \le \exp(-v \cdot t^2/2), \quad \forall t > 2M_0 + 6B.$$

Proof. See §A.1.2 for a detailed proof. □

A.1.1. Proof of Lemma A.1.2

Proof. We consider any $W, W' \in S_B$. By the definition of $\phi_W(s,a)$ in (1.3.2) and the triangle inequality, we have that

$$\begin{aligned} &\big|W^\top \phi_{W'}(s,a) - W^\top \phi_0(s,a)\big| \\ &\quad \le \frac{1}{\sqrt{m}} \sum_{l=1}^{m} \big|[W]_l^\top (s,a)\big| \cdot \Big|\mathbb{1}\big\{(s,a)^\top [W']_l > 0\big\} - \mathbb{1}\big\{(s,a)^\top [W_0]_l > 0\big\}\Big|. \end{aligned} \tag{A.1.1}$$

We now upper bound the right-hand side of (A.1.1). For the term $|[W]_l^\top (s,a)|$ in (A.1.1), we have that

$$\begin{aligned} \big|[W]_l^\top (s,a)\big| &\le \big|[W_0]_l^\top (s,a)\big| + \Big|\big([W]_l - [W_0]_l\big)^\top (s,a)\Big| \\ &\le \big|[W_0]_l^\top (s,a)\big| + \big\|[W]_l - [W_0]_l\big\|_2, \end{aligned} \tag{A.1.2}$$

where the first inequality follows from the triangle inequality and the second inequality follows from the Cauchy-Schwartz inequality and the fact that $\|(s,a)\|_2 \le 1$. To upper bound the term $|\,\mathbb{1}\{(s,a)^\top [W']_l > 0\} - \mathbb{1}\{(s,a)^\top [W_0]_l > 0\}|$ on the right-hand side of (A.1.1), note that $\mathbb{1}\{(s,a)^\top [W']_l > 0\} \ne \mathbb{1}\{(s,a)^\top [W_0]_l > 0\}$ implies that

$$\big|[W_0]_l^\top (s,a)\big| \le \big|[W']_l^\top (s,a) - [W_0]_l^\top (s,a)\big| \le \big\|[W']_l - [W_0]_l\big\|_2.$$

Thus, we have that

$$\Big|\mathbb{1}\big\{(s,a)^\top [W']_l > 0\big\} - \mathbb{1}\big\{(s,a)^\top [W_0]_l > 0\big\}\Big| \le \mathbb{1}\Big\{\big|(s,a)^\top [W_0]_l\big| \le \big\|[W']_l - [W_0]_l\big\|_2\Big\}. \tag{A.1.3}$$

Plugging (A.1.2) and (A.1.3) into (A.1.1), we have that

$$
\begin{aligned}
&\bigl|W^\top \phi_{W'}(s,a) - W^\top \phi_0(s,a)\bigr| \\
&\quad \le \frac{1}{\sqrt{m}} \sum_{l=1}^m \mathbb{1}\Bigl\{\bigl|(s,a)^\top [W_0]_l\bigr| \le \bigl\|[W']_l - [W_0]_l\bigr\|_2\Bigr\} \cdot \Bigl(\bigl|(s,a)^\top [W_0]_l\bigr| + \bigl\|[W]_l - [W_0]_l\bigr\|_2\Bigr) \\
&\quad \le \frac{1}{\sqrt{m}} \sum_{l=1}^m \mathbb{1}\,\Bigl\{\bigl|(s,a)^\top [W_0]_l\bigr| \le \bigl\|[W']_l - [W_0]_l\bigr\|_2\Bigr\} \cdot \Bigl(\bigl\|[W']_l - [W_0]_l\bigr\|_2 + \bigl\|[W]_l - [W_0]_l\bigr\|_2\Bigr).
\end{aligned}
$$

By the fact that $W, W' \in S_B$, we obtain that

$$
\bigl|W^\top \phi_{W'}(s,a) - W^\top \phi_0(s,a)\bigr|^2 \le \frac{4B^2}{m} \sum_{l=1}^m \mathbb{1}\Bigl\{\bigl|(s,a)^\top [W_0]_l\bigr| \le \bigl\|[W']_l - [W_0]_l\bigr\|_2\Bigr\}.
$$

By setting $y = \|[W']_l - [W_0]_l\|_2$ in Assumption 1.4.1, we have that

$$
\bigl\|W^\top \phi_{W'}(s,a) - W^\top \phi_0(s,a)\bigr\|_{2,\mu}^2 \le \frac{8B^2}{m} \sum_{l=1}^m \frac{c \cdot \bigl\|[W']_l - [W_0]_l\bigr\|_2}{\bigl\|[W_0]_l\bigr\|_2}.
$$

Taking the expectation with respect to the random initialization in (1.3.3) and using the Cauchy-Schwartz inequality, we have that

$$
\begin{aligned}
&\mathbb{E}_{\text{init}}\Bigl[\bigl\|W^\top \phi_{W'}(s,a) - W^\top \phi_0(s,a)\bigr\|_{2,\mu}^2\Bigr] \\
&\quad \le \mathbb{E}_{\text{init}}\biggl[\frac{8cB^2}{m}\Bigl(\sum_{l=1}^m \bigl\|[W']_l - [W_0]_l\bigr\|_2^2\Bigr)^{1/2} \cdot \Bigl(\sum_{l=1}^m 1/\bigl\|[W_0]_l\bigr\|_2^2\Bigr)^{1/2}\biggr] \\
&\quad \le \frac{8cB^3}{m}\mathbb{E}_{\text{init}}\biggl[\Bigl(\sum_{l=1}^m 1/\bigl\|[W_0]_l\bigr\|_2^2\Bigr)^{1/2}\biggr] \\
&\quad \le \frac{8cB^3}{\sqrt{m}}\Bigl(\mathbb{E}_{w \sim N(0, I_d/d)}\bigl[1/\|w\|_2^2\bigr]\Bigr)^{1/2} \\
&\quad = \mathcal{O}(B^3 \cdot m^{-1/2}),
\end{aligned}
$$

where the second inequality follows from the fact that $\|W'-W_0\|_2 \leq B$, the third inequality follows from Jensen's inequality, and the last inequality follows from Assumption 1.4.1 and Lemma A.1.2. Thus, for any $W, W_1, W_2 \in S_B$, we have that

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_1}(s,a) - W^\top \phi_{W_2}(s,a)\big\|_{2,\mu}^2\Big] \\
&\quad \leq 2\mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_1}(s,a) - W^\top \phi_0(s,a)\big\|_{2,\mu}^2\Big] + 2\mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_2}(s,a) - W^\top \phi_0(s,a)\big\|_{2,\mu}^2\Big] \\
&\quad = \mathcal{O}(B^3 \cdot m^{-1/2}).
\end{aligned}$$

Moreover, following from the Cauchy-Schwartz inequality, we have that $\|\cdot\|_{1,\mu} \leq \|\cdot\|_{2,\mu}$. Thus, by Jensen's inequality, we have that

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_1}(s,a) - W^\top \phi_{W_2}(s,a)\big\|_{1,\mu}\Big] \\
&\quad \leq \mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_1}(s,a) - W^\top \phi_{W_2}(s,a)\big\|_{2,\mu}\Big] \\
&\quad = \mathcal{O}(B^{3/2} \cdot m^{-1/4}),
\end{aligned}$$

which completes the proof of Lemma A.1.2. □

A.1.2. Proof of Lemma A.1.3

In what follows, we present the proof of Lemma A.1.3.

Proof. Recall that we write $u_W(s,a) = W^\top \phi_W(s,a)$ and $u_0(s,a) = u_{W_0}(s,a)$. Then, we have

$$\begin{aligned}\big|W^\top \phi_{W'}(s,a)\big| &\le \big|u_0(s,a)\big| + \big|(W-W')^\top \phi_{W'}(s,a)\big| + \big|u_{W'}(s,a) - u_0(s,a)\big| \\ &\le \big|u_0(s,a)\big| + \|W-W'\|_2 \cdot \big\|\phi_{W'}(s,a)\big\|_2 + \big|u_{W'}(s,a) - u_0(s,a)\big|, \qquad \text{(A.1.4)}\end{aligned}$$

where the last inequality follows from the Cauchy-Schwartz inequality. It suffices to upper bound the three terms on the right-hand side of (A.1.4). Note that we have $W, W' \in S_B$ and $\|\phi_{W'}(s,a)\|_2 \le 1$. We have that

$$\|W-W'\|_2 \cdot \big\|\phi_{W'}(s,a)\big\|_2 \le 2B. \qquad \text{(A.1.5)}$$

It remains to upper bound the term $|u_{W'}(s,a) - u_0(s,a)|$ in (A.1.4). Note that $u_W(s,a)$ is almost everywhere differentiable with respect to W. Also, it holds that $\nabla_W u_W(s,a) = \phi_W(s,a)$. Thus, following from the mean-value theorem and the Cauchy-Schwartz inequality, we have that

$$\big|u_{W'}(s,a) - u_0(s,a)\big| \le \sup_{W \in S_B} \big\|\phi_W(s,a)\big\|_2 \cdot \|W' - W_0\|_2 \le B, \qquad \text{(A.1.6)}$$

where the second inequality follows from the fact that $\|\phi_W(s,a)\|_2 \le 1$ and $W' \in S_B$. Plugging (A.1.5) and (A.1.6) into (A.1.4), we have that

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|W^\top \phi_{W'}(s,a)\big| \le \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|u_0(s,a)\big| + 3B.$$

Following from Assumption 1.4.2, we have that $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}|W^\top\phi_{W'}(s,a)|$ is sub-Gaussian. Furthermore, it holds that

$$\mathbb{E}_{\text{init}}\Big[\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\big|W^\top\phi_{W'}(s,a)\big|^2\Big] \le 2\mathbb{E}_{\text{init}}\Big[\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\big|u_0(s,a)\big|^2\Big] + 18B^2 \le 2M_0^2 + 18B^2$$

and that

$$\begin{aligned}\mathbb{P}\Big(\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\big|W^\top\phi_{W'}(s,a)\big| > t\Big) &\le \mathbb{P}\Big(\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\big|u_0(s,a)\big| + 3B > t\Big)\\ &\le \exp\big(-v\cdot(t-3B)^2\big) \le \exp(-v\cdot t^2/2)\end{aligned}$$

for $t > 2M_0 + 6B$. Thus, we complete the proof of Lemma A.1.3. □

A.2. Neural Temporal Difference

In this section, we introduce neural TD (Cai et al., 2019c), which computes ω_k in Algorithm 1. Specifically, neural TD solves the optimization problem in (1.3.19) via the update in (1.3.20), which is presented in Algorithm 4.

Algorithm 4 Neural TD

```
Require: Policy $\pi$, reward function $r$, initialization $W_0, b$, number of iterations $T_{\text{TD}}$ of
    neural TD, and stepsize $\alpha$ of neural TD.
Initialization. Set $S_{B_\omega} \leftarrow \{W \in \mathbb{R}^{md} \,|\, \|W - W_0\|_2 \le B_\omega\}$ and $\omega(0) \leftarrow W_0$.
for $j = 0, \ldots, T_{\text{TD}} - 1$ do
    Sample $(s, a, s', a')$, where $(s,a) \sim \rho_\pi$, $s' \sim P(\cdot \,|\, s, a)$, and $a' \sim \pi(\cdot \,|\, s')$.
    Compute the Bellman residual $\delta(j) = Q_{\omega(j)}(s,a) - (1-\gamma)\cdot r(s,a) - \gamma \cdot Q_{\omega(j)}(s', a')$.
    Update $\omega$ via $\omega(j+1) \leftarrow \mathrm{Proj}_{S_{B_\omega}}\big\{\omega(j) - \eta \cdot \delta(j) \cdot \phi_{\omega(j)}(s,a)\big\}$.
end for
Ensure: Output $\bar{\omega} = T^{-1}\sum_{t=0}^{T_{\text{TD}}-1} \omega(j)$.
```

A.2.1. Proof of Proposition 1.4.4

Proof. We obtain the following proposition from Cai et al. (2019c), which characterizes the convergence of Algorithm 4.

Proposition A.2.1 (Proposition 4.7 in Cai et al. (2019c))**.** We set $\alpha = \min\{(1-\gamma)/8, T_{\mathrm{TD}}^{-1/2}\}$ in Algorithm 4. Let $Q_{\bar{\omega}}(s,a)$ be the state-action value function associated with the output $\bar{\omega}$. Under Assumption 1.4.1, it holds for any policy π and reward function $r(s,a)$ that

$$\begin{aligned}\mathbb{E}_{\mathrm{init}}\Big[\big\|Q_{\bar{\omega}}(s,a) - Q_r^{\pi}(s,a)\big\|_{2,\rho_\pi}^2\Big] &= 2\mathbb{E}_{\mathrm{init}}\Big[\big\|\mathrm{Proj}_{\mathcal{F}_{B_\omega,m}} Q_r^{\pi}(s,a) - Q_r^{\pi}(s,a)\big\|_{2,\rho_\pi}^2\Big] \\ &\quad + \mathcal{O}(B_\omega^2 \cdot T_{\mathrm{TD}}^{-1/2} + B_\omega^3 \cdot m^{-1/2} + B_\omega^{5/2} \cdot m^{-1/4}), \end{aligned} \tag{A.2.1}$$

where $\mathcal{F}_{B_\omega,m}$ is defined in Definition A.1.1.

Recall that we denote by $\phi_0(s,a)$ the feature vector corresponding to the random initialization in (1.3.3). We establish the following lemma to upper bound the bias $\mathbb{E}_{\mathrm{init}}[\|\mathrm{Proj}_{\mathcal{F}_{B_\omega,m}} Q_r^{\pi}(s,a) - Q_r^{\pi}(s,a)\|_{2,\rho_\pi}^2]$ in (A.2.1) of Proposition A.2.1 when the reward function $r(s,a)$ belongs to the reward function class $\mathcal{R}_\beta$.

Lemma A.2.2. We consider any reward function $r_\beta(s,a) \in \mathcal{R}_\beta$ and policy π. Under Assumptions 1.4.1 and 1.4.2, it holds for $B_\omega > B_\beta + (1-\gamma)^{-1} \cdot \gamma \cdot B_P \cdot (2M_0 + 3B_\beta)$ and an absolute constant $C_v = (2 \cdot \gamma^2 \cdot B_P^2)^{-1} \cdot (1-\gamma)^2 \cdot v$ that

$$\mathbb{E}_{\mathrm{init}}\Big[\big\|\mathrm{Proj}_{\mathcal{F}_{B_\omega,m}} Q_{r_\beta}^{\pi}(s,a) - Q_{r_\beta}^{\pi}(s,a)\big\|_{2,\rho_\pi}^2\Big] = \mathcal{O}\big(B_\beta^3 \cdot m^{-1/2} + B_\omega^2 \cdot m^{-1} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2)\big).$$

Proof. See §A.2.2 for a detailed proof. □

Combining Proposition A.2.1 and Lemma A.2.2, for $B_\omega > B_\beta + (1-\gamma)^{-1} \cdot \gamma \cdot B_P \cdot (2M_0 + 3B_\beta)$, we have for any π that

$$\begin{aligned}&\mathbb{E}_{\mathrm{init}}\Big[\big\|Q_{\bar{\omega}}(s,a) - Q^\pi_{r_\beta}(s,a)\big\|^2_{2,\rho_\pi}\Big]\\ &\qquad = \mathcal{O}\big(B_\omega^2 \cdot T_{\mathrm{TD}}^{-1/2} + B_\omega^3 \cdot m^{-1/2} + B_\omega^{5/2} \cdot m^{-1/4} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2)\big).\end{aligned}$$

Finally, by setting $T_{\mathrm{TD}} = \Omega(m)$, we have that

$$\mathbb{E}_{\mathrm{init}}\Big[\big\|Q_{\bar{\omega}}(s,a) - Q^\pi_{r_\beta}(s,a)\big\|^2_{2,\rho_\pi}\Big] = \mathcal{O}\big(B_\omega^3 \cdot m^{-1/2} + B_\omega^{5/2} \cdot m^{-1/4} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2)\big),$$

which completes the proof of Proposition 1.4.4. □

A.2.2. Proof of Lemma A.2.2

Proof. For notational simplicity, we write $\vartheta(s,a;w) = \mathbb{1}\left\{|w^\top(s,a)| > 0\right\} \cdot (s,a)$. Under Assumption 1.4.2, we have that

$$P(s' \,|\, s,a) = \int \vartheta(s,a;w)^\top \varphi(s';w)\mathrm{d}q(w), \quad \text{where } \sup_w \Big\|\int \varphi(s;w)\mathrm{d}s\Big\|_2 \leq B_P. \tag{A.2.2}$$

Thus, since $r_\beta = (1-\gamma)^{-1} \cdot u_\beta(s,a)$, we have that

$$\begin{aligned}Q^\pi_{r_\beta}(s,a) &= (1-\gamma) \cdot r_\beta(s,a) + \gamma \cdot \int_{\mathcal{S}} P(s' \,|\, s,a) \cdot V^\pi_{r_\beta}(s')\mathrm{d}s'\\ &= u_\beta(s,a) + \int_{\mathcal{S}} \gamma \cdot V^\pi_{r_\beta}(s') \cdot \int \vartheta(s,a;w)^\top \varphi(s';w)\mathrm{d}q(w)\mathrm{d}s'\\ &= u_\beta(s,a) + \int \vartheta(s,a;w)^\top\Big(\gamma \cdot \int_{\mathcal{S}} \varphi(s';w)V^\pi_{r_\beta}(s')\mathrm{d}s'\Big)\mathrm{d}q(w),\end{aligned}$$

where the second equality follows from (A.2.2) and the last equality follows from Fubini's theorem. In the sequel, we define

$$\alpha(w) = \gamma \cdot \int_{\mathcal{S}} \varphi(s';w) V^{\pi}_{r_\beta}(s') \mathrm{d}s'. \tag{A.2.3}$$

Note that $\alpha(w) \in \mathbb{R}^d$. Then, we have that

$$Q^{\pi}_{r_\beta}(s,a) = u_\beta(s,a) + \int \vartheta(s,a;w)^\top \alpha(w) \mathrm{d}q(w).$$

To prove Lemma A.2.2, we first approximate $Q^{\pi}_{r_\beta}(s,a)$ by

$$\bar{Q}(s,a) = u_\beta(s,a) + \int \vartheta(s,a;w)^\top \bar{\alpha}(w) \mathrm{d}q(w), \tag{A.2.4}$$

where $\bar{\alpha}(w) = \alpha(w) \cdot \mathbb{1}\{\|\alpha(w)\|_2 \le K\}$ for an absolute constant $K > 0$ specified later. Then, it holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$\begin{aligned} \big|\bar{Q}(s,a) - Q^{\pi}_{r_\beta}(s,a)\big| &\le \int \Big|\vartheta(s,a;w)^\top \big(\bar{\alpha}(w) - \alpha(w)\big)\Big| \mathrm{d}q(w) \\ &\le \int \big\|\vartheta(s,a;w)\big\|_2 \cdot \big\|\bar{\alpha}(w) - \alpha(w)\big\|_2 \mathrm{d}q(w) \\ &\le \sup_w \big\|\bar{\alpha}(w) - \alpha(w)\big\|_2, \end{aligned}$$

where the second inequality follows from the Cauchy-Schwartz inequality and the last inequality follows from the fact that $\|\vartheta(s,a;w)\|_2 \le 1$. Thus, we have that

$$\big\|\bar{Q}(s,a) - Q^{\pi}_{r_\beta}(s,a)\big\|_{2,\rho_\pi} \le \big\|\bar{Q}(s,a) - Q^{\pi}_{r_\beta}(s,a)\big\|_\infty \le \sup_w \big\|\bar{\alpha}(w) - \alpha(w)\big\|_2. \tag{A.2.5}$$

We now upper bound the right-hand side of (A.2.5). To this end, we show that $\sup_w \|\alpha(w)\|_2$ is sub-Gaussian in the sequel. By the definition of $\alpha(w)$ in (A.2.3), we have that

$$\begin{aligned}
\sup_w \bigl\|\alpha(w)\bigr\|_2 &= \gamma \cdot \biggl\|\int_{\mathcal{S}} \varphi(s';w) V^{\pi}_{r_\beta}(s') \mathrm{d}s'\biggr\|_2 \\
&\le \gamma \cdot \sup_{s'\in\mathcal{S}} \bigl|V^{\pi}_{r_\beta}(s')\bigr| \cdot \sup_w \biggl\|\int_{\mathcal{S}} \varphi(s';w) \mathrm{d}s'\biggr\|_2 \\
&\le \gamma \cdot B_P \cdot \sup_{s'\in\mathcal{S}} \bigl|V^{\pi}_{r_\beta}(s')\bigr| \\
&\le \gamma \cdot (1-\gamma)^{-1} \cdot B_P \cdot \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bigl|u_\beta(s,a)\bigr|, \qquad \text{(A.2.6)}
\end{aligned}$$

where the second inequality follows from Assumption 1.4.2 and the third inequality follows from the fact that $V^{\pi}_{r_\beta}(s) = \mathbb{E}_{(s',a')\sim\nu_\pi(s)}[r_\beta(s',a')]$. Here we denote by $\nu_\pi(s)$ the state-action visitation measure starting from the state s and following the policy π. Following from Lemma A.1.3, we have that $\sup_w \|\alpha(w)\|_2$ is sub-Gaussian. By Lemma A.1.3 and (A.2.6), it holds for $t > (1-\gamma)^{-1} \cdot \gamma \cdot B_P \cdot (2M_0 + 3B_\beta)$ that

$$\begin{aligned}
\mathbb{P}\Bigl(\sup_w \bigl\|\alpha(w)\bigr\|_2 > t\Bigr) &\le \mathbb{P}\Bigl(\gamma \cdot (1-\gamma)^{-1} \cdot B_P \cdot \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bigl|u_\beta(s,a)\bigr| > t\Bigr) \\
&\le \exp\biggl(-\frac{v \cdot (1-\gamma)^2 \cdot t^2}{2\gamma^2 \cdot B_P^2}\biggr). \qquad \text{(A.2.7)}
\end{aligned}$$

Let the absolute constant K satisfy that $K > (1-\gamma)^{-1} \cdot \gamma \cdot B_P \cdot (2M_0 + 3B_\beta)$ in (A.2.7). For notational simplicity, we write $C_v = (2 \cdot \gamma^2 \cdot B_P^2)^{-1} \cdot v \cdot (1-\gamma)^2$. By the fact that $\|\bar{\alpha}(w) - \alpha(w)\|_2 = \|\alpha(w)\|_2 \cdot \mathbb{1}\{\|\alpha(w)\|_2 > K\}$, we have that

$$\sup_w \bigl\|\bar{\alpha}(w) - \alpha(w)\bigr\|_2 \le \sup_w \bigl\|\alpha(w)\bigr\|_2 \cdot \mathbb{1}\Bigl\{\sup_w \|\alpha(w)\|_2 > K\Bigr\}.$$

Following from (A.2.5) and (A.2.7), we have that

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\Big[\big\|\bar{Q}(s,a) - Q^{\pi}_{r_\beta}(s,a)\big\|_{2,\rho_\pi}\Big] \\
&\quad\le \mathbb{E}\Big[\sup_w \big\|\alpha(w)\big\|_2 \cdot \mathbb{1}\Big\{\sup_w \|\alpha(w)\|_2 > K\Big\}\Big] \\
&\quad\le \int_0^K t\cdot \mathbb{P}\Big(\sup_w \|\alpha(w)\|_2 > K\Big)\mathrm{d}t + \int_K^\infty t\cdot \mathbb{P}\Big(\sup_w \|\alpha(w)\|_2 > t\Big)\mathrm{d}t \\
&\quad= \mathcal{O}\big(K^2 \cdot \exp(-C_v \cdot K^2)\big). \qquad \text{(A.2.8)}
\end{aligned}$$

We now construct $\widehat{Q}(s,a) \in \mathcal{F}_{K,m}$, which approximates $\bar{Q}(s,a)$ defined in (A.2.4). We define

$$f(s,a) = \int \vartheta(s,a;w)^\top \bar{\alpha}(w)\mathrm{d}q(\omega).$$

Then, we have that $\bar{Q}(s,a) = u_\beta(s,a) + f(s,a)$. Note that $f(s,a)$ belongs to the following function class,

$$\widetilde{\mathcal{F}}_{K,\infty} = \bigg\{\int \vartheta(s,a;w)^\top \alpha(w)\mathrm{d}q(\omega) \,\bigg|\, \sup_w \big\|\alpha(w)\big\|_2 \le K\bigg\}.$$

We now show that $f(s,a)$ is well approximated by the following function class,

$$\widetilde{\mathcal{F}}_{K,m} = \bigg\{W^\top \phi_0(s,a) = \frac{1}{\sqrt{m}}\sum_{l=1}^m [W]_l^\top \vartheta\big(s,a;[W]_l\big) \,\bigg|\, \sup_l \big\|[W]_l\big\|_2 \le K/\sqrt{m}\bigg\},$$

where $\phi_0(s,a)$ is the feature vector corresponding to the random initialization. We obtain the following lemma from Rahimi and Recht (2009), which characterizes the approximation error of $\widetilde{\mathcal{F}}_{K,\infty}$ by $\widetilde{\mathcal{F}}_{K,m}$.

Lemma A.2.3 (Lemma 1 in Rahimi and Recht (2009))**.** For any $f(s,a) \in \widetilde{\mathcal{F}}_{K,\infty}$, it holds with probability at least $1-\delta$ that

$$\big\|\mathrm{Proj}_{\widetilde{\mathcal{F}}_{K,m}} f(s,a) - f(s,a)\big\|_{2,\mu} \le K \cdot m^{-1/2} \cdot \big(1 + \sqrt{2\log(1/\delta)}\big),$$

where $\mu \in \mathscr{P}(\mathcal{S} \times \mathcal{A})$.

Lemma A.2.3 implies that there exists $\widehat{f}(s,a) \in \widetilde{\mathcal{F}}_{K,m}$ such that

$$\begin{aligned} \mathbb{E}_{\mathrm{init}}\Big[\big\|\widehat{f}(s,a) - f(s,a)\big\|_{2,\rho_\pi}^2\Big] &= \int_0^\infty \mathbb{P}\Big(\big\|\widehat{f}(s,a) - f(s,a)\big\|_{2,\rho_\pi}^2 > y\Big)\mathrm{d}y \\ &\le \int_0^\infty y \cdot \exp\big(-1/2 \cdot (\sqrt{my}/K - 1)^2\big) = \mathcal{O}(K^2/m). \end{aligned} \tag{A.2.9}$$

By the fact that $\widehat{f}(s,a) \in \widetilde{\mathcal{F}}_{K,m}$ and the definition of $\mathcal{F}_{K,m}$ in Definition A.1.1, we have that $\widehat{f}(s,a) \in \mathcal{F}_{K,m} - u_0(s,a)$. Let

$$\widehat{Q}(s,a) = \beta^\top \phi_0(s,a) + \widehat{f}(s,a) = (\beta + W_f)^\top \phi_0(s,a).$$

We then have that $\widehat{Q}(s,a) \in \mathcal{F}_{B_\beta + K,m}$ and that

$$\begin{aligned} &\mathbb{E}_{\mathrm{init}}\Big[\big\|\bar{Q}(s,a) - \widehat{Q}(s,a)\big\|_{2,\rho_k}^2\Big] \\ &\quad \le 2\mathbb{E}_{\mathrm{init}}\Big[\big\|u_\beta(s,a) - \beta^\top\phi_0(s,a)\big\|_{2,\rho_\pi}^2\Big] + 2\mathbb{E}_{\mathrm{init}}\Big[\big\|\widehat{f}(s,a) - f(s,a)\big\|_{2,\rho_\pi}^2\Big] \\ &\quad = \mathcal{O}(B_\beta^3 \cdot m^{-1/2} + K^2 \cdot m^{-1}), \end{aligned} \tag{A.2.10}$$

where the last inequality follows from Assumption 1.4.1, Lemma A.1.2, and (A.2.9).

Finally, we set $B_\omega = K + B_\beta > B_\beta + (1-\gamma)^{-1} \cdot \gamma \cdot B_P \cdot (2M_0 + 3B_\beta)$. Combining (A.2.8) and (A.2.10), we have that

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\Big[\big\|Q^{\pi}_{r_\beta}(s,a) - \widehat{Q}(s,a)\big\|^2_{2,\rho_k}\Big] \\
&\quad \le 2\mathbb{E}_{\text{init}}\Big[\big\|\bar{Q}(s,a) - \widehat{Q}(s,a)\big\|^2_{2,\rho_k}\Big] + 2\mathbb{E}_{\text{init}}\Big[\big\|\bar{Q}(s,a) - Q^{\pi}_{r_\beta}(s,a)\big\|^2_{2,\rho_k}\Big] \\
&\quad = \mathcal{O}\big(B_\beta^3 \cdot m^{-1/2} + B_\omega^2 \cdot m^{-1} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2)\big),
\end{aligned}$$

where $\widehat{Q}(s,a) \in \mathcal{F}_{B_\omega, m}$. Thus, we complete the proof of Lemma A.2.2. □

A.3. Proofs of Auxiliary Results

In what follows, we present the proofs of the lemmas in §1.3-1.5.

A.3.1. Proof of Proposition 1.3.1

Proof. By the definition of the neural network in (1.3.1), we have for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ that $\nabla_W u_W(s,a) = \phi_W(s,a)$ almost everywhere. We first calculate $\nabla_\theta L(\theta, \beta)$. Following from the policy gradient theorem (Sutton and Barto, 2018) and the definition of $L(\theta,\beta)$ in (1.2.4), we have that

$$\begin{aligned}
\nabla_\theta L(\theta,\beta) &= -\nabla_\theta J(\pi_\theta; r_\beta) \\
&= -\mathbb{E}_{\nu_{\pi_\theta}}\big[Q^{\pi_\theta}_{r_\beta}(s,a) \cdot \nabla_\theta \log \pi_\theta(a \,|\, s)\big]. \qquad \text{(A.3.1)}
\end{aligned}$$

Following from the parameterization of π_θ in (1.3.5) and the definition of $\iota_\theta(s,a)$ in (1.3.8) of Proposition 1.3.1, we have that

$$\begin{aligned}\nabla_\theta \log \pi_\theta(a \,|\, s) &= \tau \cdot \phi_\theta(s,a) - \frac{\sum_{a' \in \mathcal{A}} \tau \cdot \exp\bigl(\tau \cdot \theta^\top \phi_\theta(s,a')\bigr) \cdot \phi_\theta(s,a')}{\sum_{a' \in \mathcal{A}} \exp\bigl(\tau \cdot \theta^\top \phi_\theta(s,a')\bigr)} \\ &= \tau \cdot \Bigl(\phi_\theta(s,a) - \tau \cdot \mathbb{E}_{a' \sim \pi_\theta(\cdot \,|\, s)}\bigl[\phi_\theta(s,a')\bigr]\Bigr) = \tau \cdot \iota_\theta(s,a). \end{aligned} \tag{A.3.2}$$

Plugging (A.3.2) into (A.3.1), we have that

$$\nabla_\theta L(\theta,\beta) = -\tau \cdot \mathbb{E}_{\nu_{\pi_\theta}}\bigl[Q_{r_\beta}^{\pi_\theta}(s,a) \cdot \iota_\theta(s,a)\bigr].$$

It remains to calculate $\mathcal{I}(\theta)$ and $\nabla_\beta L(\theta,\beta)$. By (A.3.2) and the definition of $\mathcal{I}(\theta)$ in (1.3.7), it holds that

$$\begin{aligned}\mathcal{I}(\theta) &= \mathbb{E}_{\nu_{\pi_\theta}}\bigl[\nabla \log \pi_\theta(a \,|\, s) \nabla \log \pi_\theta(a \,|\, s)^\top\bigr] \\ &= \tau^2 \cdot \mathbb{E}_{\nu_{\pi_\theta}}\bigl[\iota_\theta(s,a)\iota_\theta(s,a)^\top\bigr].\end{aligned}$$

By the definition of the objective function $L(\theta,\beta)$ in (1.2.4), it holds that

$$\begin{aligned}\nabla_\beta L(\theta,\beta) &= \nabla_\beta J(\pi_{\mathrm{E}}; r_\beta) - \nabla_\beta J(\pi_\theta; r_\beta) - \lambda \cdot \nabla_\beta \psi(\beta) \\ &= \mathbb{E}_{\nu_{\mathrm{E}}}\bigl[\nabla_\beta r_\beta(s,a)\bigr] - \mathbb{E}_{\nu_{\pi_\theta}}\bigl[\nabla_\beta r_\beta(s,a)\bigr] - \lambda \cdot \nabla_\beta \psi(\beta) \\ &= (1-\gamma)^{-1} \cdot \mathbb{E}_{\nu_{\mathrm{E}}}\bigl[\phi_\beta(s,a)\bigr] - (1-\gamma)^{-1} \cdot \mathbb{E}_{\nu_{\pi_\theta}}\bigl[\phi_\beta(s,a)\bigr] - \lambda \cdot \nabla_\beta \psi(\beta).\end{aligned}$$

Thus, we complete the proof of Proposition 1.3.1. □

A.3.2. Proof of Lemma 1.5.2

Proof. The proof of Lemma 1.5.2 is similar to that of Lemmas 5.4 and 5.5 in Wang et al. (2019). By direct calculation, we have that

$$\eta \cdot \mathbb{E}_{d_{\mathrm{E}}}\Big[\big\langle Q_{r_k}^{\pi_k}(s,\cdot), \pi_{\mathrm{E}}^s - \pi_k^s \big\rangle_{\mathcal{A}}\Big] = \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_k) - \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_{k+1}) + \eta \cdot \Delta_k^{(\mathrm{i})},$$

where $\Delta_k^{(\mathrm{i})}$ takes the form of

$$\begin{aligned}
\Delta_k^{(\mathrm{i})} &= \eta^{-1} \cdot \Big\{ \mathbb{E}_{d_{\mathrm{E}}}\Big[\big\langle \log(\pi_{k+1}^s / \pi_k^s) - \eta \cdot Q_{r_k}^{\pi_k}(s,\cdot), \pi_{\mathrm{E}}^s - \pi_k^s \big\rangle_{\mathcal{A}} \\
&\qquad + \big\langle \log(\pi_{k+1}^s / \pi_k^s), \pi_k^s - \pi_{k+1}^s \big\rangle_{\mathcal{A}}\Big] - \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{k+1}^s \,\|\, \pi_k^s) \Big\} \\
&= \underbrace{\eta^{-1} \cdot \mathbb{E}_{d_{\mathrm{E}}}\Big[\big\langle \log(\pi_{k+1}^s / \pi_k^s) - \eta \cdot \widehat{Q}_{\omega_k}(s,\cdot), \pi_{\mathrm{E}}^s - \pi_k^s \big\rangle_{\mathcal{A}}\Big]}_{\text{(i.a)}} \\
&\qquad + \underbrace{\mathbb{E}_{d_{\mathrm{E}}}\Big[\big\langle \widehat{Q}_{\omega_k}(s,\cdot) - Q_{r_k}^{\pi_k}(s,\cdot), \pi_{\mathrm{E}}^s - \pi_k^s \big\rangle_{\mathcal{A}}\Big]}_{\text{(i.b)}} \\
&\qquad + \underbrace{\eta^{-1} \cdot \mathbb{E}_{d_{\mathrm{E}}}\Big[\big\langle \log(\pi_{k+1}^s / \pi_k^s), \pi_k^s - \pi_{k+1}^s \big\rangle_{\mathcal{A}} - \mathrm{KL}(\pi_{k+1}^s \,\|\, \pi_k^s)\Big]}_{\text{(i.c)}}. \qquad \text{(A.3.3)}
\end{aligned}$$

The following lemmas upper bound $\Delta_k^{(\mathrm{i})}$ by upper bounding terms (i.a), (i.b), and (i.c) on the right-hand side of (A.3.3), respectively. Note that the expectation $\mathbb{E}_{\mathrm{init}, d_{\mathrm{E}}}$ is taken with respect to the random initialization in (1.3.3) and $s \sim d_{\mathrm{E}}$.

Lemma A.3.1 (Upper Bound of Term (i.a) in (A.3.3))**.** Under Assumptions 1.4.1 and 1.4.3, we have that

$$\mathbb{E}_{\mathrm{init},d_{\mathrm{E}}}\Big[\Big|\big\langle \log(\pi^s_{k+1}/\pi^s_k) - \eta\cdot \widehat{Q}_{\omega_k}(s,\cdot), \pi^s_{\mathrm{E}} - \pi^s_k\big\rangle_{\mathcal{A}}\Big|\Big]$$
$$= \eta\cdot 2\sqrt{2}\cdot C_h\cdot B_\theta^{1/2}\cdot \sigma^{1/2}\cdot N^{-1/4} + \mathcal{O}(\tau_{k+1}\cdot B_\theta^{3/2}\cdot m^{-1/4} + \eta\cdot B_\theta^{5/4}\cdot m^{-1/8}),$$

where C_h is defined in Assumption 1.4.1 and σ is defined in Assumption 1.4.3.

Proof. See §A.4.1 for a detailed proof. □

Lemma A.3.2 (Upper Bound of Term (i.b) in (A.3.3))**.** Under Assumption 1.4.1, we have that

$$\mathbb{E}_{\mathrm{init},d_{\mathrm{E}}}\Big[\big\langle \widehat{Q}_{\omega_k}(s,\cdot) - Q^{\pi_k}_{r_k}(s,\cdot), \pi^s_{\mathrm{E}} - \pi^s_k\big\rangle_{\mathcal{A}}\Big] \le C_h\cdot \epsilon_{Q,k},$$

where $\epsilon_{Q,k}$ takes the form of

$$\epsilon_{Q,k} = \mathbb{E}_{\mathrm{init}}\Big[\big\|Q^{\pi_k}_{r_k}(s,a) - \widehat{Q}_{\omega_k}(s,a)\big\|_{2,\rho_k}\Big]. \tag{A.3.4}$$

Proof. See §A.4.2 for a detailed proof. □

Lemma A.3.3 (Upper Bound of Term (i.c) in (A.3.3))**.** Under Assumptions 1.4.1 and 1.4.2, we have that

$$\mathbb{E}_{\mathrm{init},d_{\mathrm{E}}}\Big[\Big|\big\langle \log(\pi^s_{k+1}/\pi^s_k), \pi^s_k - \pi^s_{k+1}\big\rangle_{\mathcal{A}}\Big| - \mathrm{KL}(\pi^s_{k+1}\,\|\,\pi^s_k)\Big]$$
$$= \eta^2\cdot (M_0^2 + 9B_\theta^2) + \mathcal{O}(\tau_{k+1}\cdot B_\theta^{3/2}\cdot m^{-1/4}),$$

where M_0 is defined in Assumption 1.4.2.

Proof. See §A.4.3 for a detailed proof. □

Finally, by Lemmas A.3.1-A.3.3, under Assumptions 1.4.2 and 1.4.3, we obtain from (A.3.3) that

$$\mathbb{E}_{\text{init}}\big[|\Delta_k^{(\text{i})}|\big] = 2\sqrt{2}C_h \cdot B_\theta^{1/2} \cdot \sigma^{1/2} \cdot N^{-1/4} + C_h \cdot \epsilon_{Q,k} + \eta \cdot (M_0^2 + 9B_\theta^2)$$
$$+ \mathcal{O}(\eta^{-1} \cdot \tau_{k+1} \cdot B_\theta^{3/2} \cdot m^{-1/4} + B_\theta^{5/4} \cdot m^{-1/8}).$$

Here M_0 is defined in Assumption 1.4.2, τ_{k+1} is the inverse temperature parameter of π_{k+1} defined in (1.3.5), σ is defined in Assumption 1.4.3, and $\epsilon_{Q,k}$ is defined in (A.3.4) of Lemma A.3.2. Following from Proposition 1.4.4, we have that

$$C_h \cdot \epsilon_{Q,k} = \mathcal{O}\big(B_\omega^3 \cdot m^{-1/2} + B_\omega^{5/2} \cdot m^{-1/4} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2)\big).$$

Thus, we complete the proof of Lemma 1.5.2. □

A.3.3. Proof of Lemma 1.5.3

Proof. We consider a fixed $\beta' \in S_{B_\beta}$. For notational simplicity, we write $r' = r_{\beta'}(s,a)$, $r_k = r_k(s,a)$ and $\phi_\beta = \phi_\beta(s,a)$. By the parameterization of $r_\beta(s,a)$ in (1.3.6), we have that

$$
\begin{aligned}
L(\theta_k, \beta') - L(\theta_k, \beta_k) &= \langle r' - r_k, \nu_{\mathrm{E}} - \nu_k \rangle_{\mathcal{S}\times\mathcal{A}} + \lambda \cdot \psi(\beta_k) - \lambda \cdot \psi(\beta') \\
&= (1-\gamma)^{-1} \cdot \Big(\big\langle \phi_{\beta_k}^\top (\beta' - \beta_k), \nu_{\mathrm{E}} - \nu_k \big\rangle_{\mathcal{S}\times\mathcal{A}} + \langle \phi_{\beta'}^\top \beta' - \phi_{\beta_k}^\top \beta', \nu_{\mathrm{E}} - \nu_k \rangle_{\mathcal{S}\times\mathcal{A}} \Big) \\
&\quad + \lambda \cdot \big(\psi(\beta) - \psi(\beta') \big)
\end{aligned}
$$

$$
\le (\beta' - \beta_k)^\top \nabla_\beta L(\theta_k, \beta_k) + (1-\gamma)^{-1} \cdot \big(\|\phi_{\beta_k}^\top \beta' - \phi_{\beta'}^\top \beta'\|_{1,\nu_k} + \|\phi_{\beta_k}^\top \beta' - \phi_{\beta'}^\top \beta'\|_{1,\nu_{\mathrm{E}}} \big), \tag{A.3.5}
$$

where the last inequality follows from (1.3.10) of Proposition 1.3.1. Then, we have that

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{init}}\big[L(\theta_k, \beta') - L(\theta_k, \beta_k) \big] \\
&\quad \le \mathbb{E}_{\mathrm{init}}\Big[(\beta' - \beta_k)^\top \nabla_\beta L(\theta_k, \beta_k) + (1-\gamma)^{-1} \cdot \big(\|\phi_{\beta_k}^\top \beta' - \phi_{\beta'}^\top \beta'\|_{1,\nu_k} \\
&\qquad + \|\phi_{\beta_k}^\top \beta' - \phi_{\beta'}^\top \beta'\|_{1,\nu_{\mathrm{E}}} \big) \Big] \\
&\quad \le \mathbb{E}_{\mathrm{init}}\big[(\beta' - \beta_k)^\top \nabla_\beta L(\theta_k, \beta_k) \big] + \mathcal{O}(B_\beta^{3/2} \cdot m^{-1/4}),
\end{aligned}
$$

where the last inequality follows from Assumption 1.4.1, Lemma A.1.2, and the fact that $\beta', \beta_k \in S_{B_\beta}$. Thus, we complete the proof of Lemma 1.5.3. □

A.3.4. Proof of Lemma 1.5.4

Proof. By the update of β_k in (1.3.14), it holds for any $\beta' \in S_{B_\beta}$ that

$$
\big(\beta_k + \eta \cdot \widehat{\nabla}_\beta L(\theta_k, \beta_k) - \beta_{k+1} \big)^\top (\beta' - \beta_{k+1}) \le 0,
$$

which further implies that

$$
\begin{aligned}
&\eta \cdot (\beta' - \beta_k)^\top \nabla_\beta L(\theta_k, \beta_k) \\
&\le \|\beta_k - \beta'\|_2^2 - \|\beta_{k+1} - \beta'\|_2^2 - \|\beta_{k+1} - \beta_k\|_2^2 \\
&\qquad + \eta \cdot \Big((\beta_{k+1} - \beta_k)^\top \widehat{\nabla}_\beta L(\theta_k, \beta_k) + (\beta_k - \beta')^\top \big(\widehat{\nabla}_\beta L(\theta_k, \beta_k) - \nabla_\beta L(\theta_k, \beta_k) \big) \Big).
\end{aligned} \tag{A.3.6}
$$

Combining (A.3.5) and (A.3.6), we have that

$$
\eta \cdot \big(L(\theta_k, \beta_k) - L(\theta_k, \beta') \big) \le \|\beta_k - \beta'\|_2^2 - \|\beta_{k+1} - \beta'\|_2^2 - \|\beta_{k+1} - \beta_k\|_2^2 + \eta \cdot \Delta_k^{(\mathrm{ii})},
$$

where $\Delta_k^{(\mathrm{ii})}$ takes the form of

$$
\begin{aligned}
\Delta_k^{(\mathrm{ii})} &= \underbrace{(\beta_{k+1} - \beta_k)^\top \widehat{\nabla}_\beta L(\theta_k, \beta_k)}_{\text{(ii.a)}} + \underbrace{(\beta_k - \beta')^\top \big(\widehat{\nabla}_\beta L(\theta_k, \beta_k) - \nabla_\beta L(\theta_k, \beta_k) \big)}_{\text{(ii.b)}} \\
&\qquad + \underbrace{(1 - \gamma)^{-1} \cdot \big(\|\phi_{\beta_k}^\top \beta' - \phi_{\beta'}^\top \beta'\|_{2,\nu_k} + \|\phi_{\beta_k}^\top \beta' - \phi_{\beta'}^\top \beta'\|_{2,\nu_{\mathrm{E}}} \big)}_{\text{(ii.c)}}
\end{aligned} \tag{A.3.7}
$$

We now upper bound terms (ii.a), (ii.b), and (ii.c) on the right-hand side of (A.3.7). Following from Assumption 1.4.1 and Lemma A.1.2, we have that

$$
\mathbb{E}_{\mathrm{init}} \big[\|\phi_{\beta_k}^\top \beta' - \phi_{\beta'}^\top \beta'\|_{2,\nu_k} + \|\phi_{\beta_k}^\top \beta' - \phi_{\beta'}^\top \beta'\|_{2,\nu_{\mathrm{E}}} \big] = \mathcal{O}(B_\beta^{3/2} \cdot m^{-1/4}), \tag{A.3.8}
$$

which upper bounds term (ii.c) of (A.3.7). For term (ii.b) of (A.3.7), we have that

$$\mathbb{E}\Big[\Big|(\beta_k-\beta')^\top\big(\widehat{\nabla}_\beta L(\theta_k,\beta_k)-\nabla_\beta L(\theta_k,\beta_k)\big)\Big|\Big]$$

$$\text{(A.3.9)}\qquad \leq \mathbb{E}\Big[\big\|\widehat{\nabla}_\beta L(\theta_k,\beta_k)-\nabla_\beta L(\theta_k,\beta_k)\big\|_2\cdot\|\beta'-\beta_k\|_2\Big]\leq 2B_\beta\cdot\mathbb{E}\big[\|\xi'_k\|_2\big]\leq 2B_\beta\cdot(\sigma^2/N)^{1/2},$$

where we write $\xi'_k=\widehat{\nabla}_\beta L(\theta_k,\beta_k)-\nabla_\beta L(\theta_k,\beta_k)$. Here the first inequality follows from the Cauchy-Schwartz inequality, the second inequality follows from the fact that $\beta_k,\beta'\in S_{B_\beta}$, and the last inequality follows from Assumption 1.4.3. To upper bound term (ii.a) in (A.3.7), we have that

$$\begin{aligned}\text{(A.3.10)}\qquad &\mathbb{E}\Big[\big|(\beta_{k+1}-\beta_k)^\top\widehat{\nabla}_\beta L(\theta_k,\beta_k)\big|\Big]\\ &\quad\leq \mathbb{E}\Big[\big\|\widehat{\nabla}_\beta L(\theta_k,\beta_k)\big\|_2\cdot\|\beta_{k+1}-\beta_k\|_2\Big]\leq \eta\cdot\mathbb{E}\Big[\big\|\widehat{\nabla}_\beta L(\theta_k,\beta_k)\big\|_2^2\Big]\\ &\quad= 2\eta\cdot\Big(\big\|\nabla_\beta L(\theta_k,\beta_k)\big\|_2^2+\mathbb{E}\big[\|\xi'_k\|_2^2\big]\Big),\end{aligned}$$

where the first inequality follows from the Cauchy-Schwartz inequality and the second inequality follows from the update of β in (1.3.14). Furthermore, we have

$$\begin{aligned}\big\|\nabla_\beta L(\theta_k,\beta_k)\big\|_2^2 &= \Big\|\mathbb{E}_{\nu_k}\big[\phi_{\beta_k}(s,a)\big]-\mathbb{E}_{\nu_E}\big[\phi_{\beta_k}(s,a)\big]+\lambda\cdot\nabla_\beta\psi(\beta_k)\Big\|_2^2\\ &\leq\Big(\mathbb{E}_{\nu_k}\Big[\big\|\phi_{\beta_k}(s,a)\big\|_2\Big]+\mathbb{E}_{\nu_k}\Big[\big\|\phi_{\beta_k}(s,a)\big\|_2\Big]+\lambda\cdot\big\|\nabla_\beta\psi(\beta_k)\big\|_2\Big)^2\\ \text{(A.3.11)}\qquad &\leq(2+\lambda\cdot L_\psi)^2,\end{aligned}$$

where the first inequality follows from Jensen's inequality and the second inequality follows from the fact that $\|\phi_W(s,a)\|_2\leq 1$ and the Lipschitz continuity of $\psi(\beta)$ in Assumption

1.4.3. By plugging (A.3.11) into (A.3.10), we have that

$$
\begin{aligned}
\mathbb{E}\Big[\big|\widehat{\nabla}_\beta L(\theta_k, \beta_k)^\top(\beta_k - \beta_{k+1})\big|\Big] &\le \eta \cdot \Big((2 + \lambda \cdot L_\psi)^2 + \mathbb{E}\big[\|\xi_k'\|_2^2\big]\Big) \\
&\le \eta \cdot \big((2 + \lambda \cdot L_\psi)^2 + \sigma^2/N\big), \qquad \text{(A.3.12)}
\end{aligned}
$$

where the last inequality follows from Assumption 1.4.3. Finally, by plugging (A.3.8), (A.3.9), and (A.3.12) into (A.3.7), we have that

$$
\mathbb{E}_{\text{init}}\big[|\Delta_k^{(\text{ii})}|\big] = \eta \cdot \big((2 + \lambda \cdot L_\psi)^2 + \sigma^2 \cdot N^{-1}\big) + 2B_\beta \cdot \sigma \cdot N^{-1/2} + \mathcal{O}(B_\beta^{3/2} \cdot m^{-1/4}).
$$

Thus, we complete the proof of Lemma 1.5.4. □

A.4. Proofs of Supporting Lemmas

In what follows, we present the proofs of the lemmas in §A.3.

A.4.1. Proof of Lemma A.3.1

Proof. It holds for any policies π, π' that

$$
\big\langle D(s), \pi^s - (\pi')^s \big\rangle_{\mathcal{A}} = 0, \qquad \text{(A.4.1)}
$$

where $D(s)$ only depends on the state s. Thus, we have that

$$
\begin{aligned}
&\big\langle \log(\pi_{k+1}^s/\pi_k^s) - \eta \cdot \widehat{Q}_{\omega_k}(s, \cdot), \pi_{\mathrm{E}}^s - \pi_k^s \big\rangle_{\mathcal{A}} \\
&\quad = \big\langle \tau_{k+1} \cdot \phi_{\theta_{k+1}}(s, \cdot)^\top \theta_{k+1} - \tau_k \cdot \phi_{\theta_k}(s, \cdot)^\top \theta_k - \eta \cdot \phi_{\omega_k}(s, \cdot)^\top \omega_k, \pi_{\mathrm{E}}^s - \pi_k^s \big\rangle_{\mathcal{A}} \\
&\quad = \big\langle \tau_{k+1} \cdot \iota_{\theta_{k+1}}(s, \cdot)^\top \theta_{k+1} - \tau_k \cdot \iota_{\theta_k}(s, \cdot)^\top \theta_k - \eta \cdot \iota_{\omega_k}(s, \cdot)^\top \omega_k, \pi_{\mathrm{E}}^s - \pi_k^s \big\rangle_{\mathcal{A}},
\end{aligned}
$$

where the first inequality follows from the parameterization of π_θ and $\widehat{Q}_\omega$ in (1.3.5) and (1.3.12), respectively, and the second equality follows from the definition of the temperature-adjusted score function $\iota_\theta(s, a)$ in (1.3.8) of Proposition 1.3.1. Here, with a slight abuse of the notation, we define

$$\iota_{\omega_k}(s, a) = \phi_{\omega_k}(s, a) - \mathbb{E}_{a' \sim \pi_k(\cdot \,|\, s)}\big[\phi_{\omega_k}(s, a')\big]. \tag{A.4.2}$$

Then, following from (A.4.1) and the update $\tau_{k+1} \cdot \theta_{k+1} = \tau_k \cdot \theta_k - \eta \cdot \delta_k$ in (1.3.13), we have that

$$\begin{aligned} &\big\langle \log(\pi^s_{k+1}/\pi^s_k) - \eta \cdot \widehat{Q}_{\omega_k}(s, \cdot), \pi^s_{\mathrm{E}} - \pi^s_k \big\rangle_{\mathcal{A}} \\ &\quad = \big\langle \tau_{k+1} \cdot \iota_{\theta_{k+1}}(s, \cdot)^\top \theta_{k+1} - \tau_k \cdot \iota_{\theta_k}(s, \cdot)^\top \theta_k - \eta \cdot \iota_{\omega_k}(s, \cdot)^\top \omega_k, \pi^s_{\mathrm{E}} - \pi^s_k \big\rangle_{\mathcal{A}} \\ &\quad = \underbrace{\tau_{k+1} \cdot \big\langle \iota_{\theta_{k+1}}(s, \cdot)^\top \theta_{k+1} - \iota_{\theta_k}(s, \cdot)^\top \theta_{k+1}, \pi^s_{\mathrm{E}} - \pi^s_k \big\rangle_{\mathcal{A}}}_{\text{(i)}} \\ &\qquad - \underbrace{\eta \cdot \big\langle \iota_{\theta_k}(s, \cdot)^\top \delta_k + \iota_{\omega_k}(s, \cdot)^\top \omega_k, \pi^s_{\mathrm{E}} - \pi^s_k \big\rangle_{\mathcal{A}}}_{\text{(ii)}}. \end{aligned} \tag{A.4.3}$$

In what follows, we upper bound terms (i) and (ii) on the right-hand side of (A.4.3).

Upper bound of term (i) in (A.4.3)**.** Following from (1.3.8) of Proposition 1.3.1 and (A.4.1) we have that

$$\begin{aligned}&\Big|\big\langle \iota_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1} - \iota_{\theta_k}(s,\cdot)^\top\theta_{k+1}, \pi_{\mathrm{E}}^s - \pi_k^s\big\rangle_{\mathcal{A}}\Big| \\ &\quad = \Big|\big\langle \phi_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1} - \phi_{\theta_k}(s,\cdot)^\top\theta_{k+1}, \pi_{\mathrm{E}}^s - \pi_k^s\big\rangle_{\mathcal{A}}\Big| \\ &\quad \le \big\|\phi_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1} - \phi_{\theta_k}(s,\cdot)^\top\theta_{k+1}\big\|_{1,\pi_{\mathrm{E}}^s} + \big\|\phi_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1} - \phi_{\theta_k}(s,\cdot)^\top\theta_{k+1}\big\|_{1,\pi_k^s},\end{aligned} \tag{A.4.4}$$

where the inequality follows from the triangle inequality. Following from Assumption 1.4.1 and Lemma A.1.2, we have that

$$\mathbb{E}_{\mathrm{init},d_{\mathrm{E}}}\Big[\big\|\phi_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1} - \phi_{\theta_k}(s,\cdot)^\top\theta_{k+1}\big\|_{1,\pi_{\mathrm{E}}^s}\Big] = \mathcal{O}(B_\theta^{3/2}\cdot m^{-1/4}). \tag{A.4.5}$$

Furthermore, following from Assumption 1.4.1, Lemma A.1.2, and the Cauchy-Schwartz inequality, we have that

$$\begin{aligned}&\mathbb{E}_{\mathrm{init},d_{\mathrm{E}}}\Big[\big\|\phi_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1} - \phi_{\theta_k}(s,\cdot)^\top\theta_{k+1}\big\|_{1,\pi_k^s}\Big] \\ &\quad = \mathbb{E}_{\mathrm{init},d_k}\bigg[\big\|\phi_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1} - \phi_{\theta_k}(s,\cdot)^\top\theta_{k+1}\big\|_{1,\pi_k^s}\cdot\frac{\mathrm{d}d_{\mathrm{E}}}{\mathrm{d}d_k}\bigg] \\ &\quad \le \big\|\phi_{\theta_{k+1}}(s,a)^\top\theta_{k+1} - \phi_{\theta_k}(s,a)^\top\theta_{k+1}\big\|_{2,\nu_k}\cdot\bigg\|\frac{\mathrm{d}d_{\mathrm{E}}}{\mathrm{d}d_k}\bigg\|_{2,d_k} \\ &\quad = \mathcal{O}(B_\theta^{3/2}\cdot m^{-1/4}).\end{aligned} \tag{A.4.6}$$

Here the expectation $\mathbb{E}_{\mathrm{init},d_k}$ is taken with respect to the random initialization in (1.3.3) and $s\sim d_k$. Thus, plugging (A.4.5) and (A.4.6) into (A.4.4), we obtain for term (i) of

(A.4.3) that

$$\mathbb{E}_{\mathrm{init},d_{\mathrm{E}}}\Big[\Big|\big\langle \iota_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1} - \iota_{\theta_k}(s,\cdot)^\top\theta_{k+1}, \pi_{\mathrm{E}}^s - \pi_k^s\big\rangle_{\mathcal{A}}\Big|\Big] = \mathcal{O}(B_\theta^{3/2}\cdot m^{-1/4}). \tag{A.4.7}$$

Upper bound of term (ii) in (A.4.3)**.** Following from the Cauchy-Schwartz inequality, we have that

$$\begin{aligned}\mathbb{E}_{d_{\mathrm{E}}}\Big[\Big|\big\langle \iota_{\theta_k}(s,\cdot)^\top\delta_k + \iota_{\omega_k}(s,\cdot)^\top\omega_k, \pi_{\mathrm{E}}^s\big\rangle_{\mathcal{A}}\Big|\Big] &\le \int_{\mathcal{S}\times\mathcal{A}}\big|\iota_{\theta_k}(s,a)^\top\delta_k + \iota_{\omega_k}(s,a)^\top\omega_k\big|\mathrm{d}\nu_{\mathrm{E}}(s,a)\\ &\le \Big\|\frac{\mathrm{d}\nu_{\mathrm{E}}}{\mathrm{d}\nu_k}\Big\|_{2,\nu_k}\cdot\big\|\iota_{\theta_k}(s,a)^\top\delta_k + \iota_{\omega_k}(s,a)^\top\omega_k\big\|_{2,\nu_k}. \end{aligned}\tag{A.4.8}$$

Similarly, we have that

$$\begin{aligned}&\mathbb{E}_{d_{\mathrm{E}}}\Big[\Big|\big\langle \iota_{\theta_k}(s,\cdot)^\top\delta_k + \iota_{\omega_k}(s,\cdot)^\top\omega_k, \pi_k^s\big\rangle_{\mathcal{A}}\Big|\Big]\\ &\quad\le \int_{\mathcal{S}\times\mathcal{A}}\Big|\big\langle \iota_{\theta_k}(s,\cdot)^\top\delta_k + \iota_{\omega_k}(s,\cdot)^\top\omega_k, \pi_k^s\big\rangle_{\mathcal{A}}\Big|\mathrm{d}\pi_k^s(a)\mathrm{d}d_{\mathrm{E}}(s)\\ &\quad= \int_{\mathcal{S}\times\mathcal{A}}\Big|\big\langle \iota_{\theta_k}(s,\cdot)^\top\delta_k + \iota_{\omega_k}(s,\cdot)^\top\omega_k, \pi_k^s\big\rangle_{\mathcal{A}}\Big|\cdot\frac{\mathrm{d}d_{\mathrm{E}}}{\mathrm{d}d_k}(s)\mathrm{d}\nu_k(s,a)\\ &\quad\le \Big\|\frac{\mathrm{d}d_{\mathrm{E}}}{\mathrm{d}d_k}\Big\|_{2,d_k}\cdot\big\|\iota_{\theta_k}(s,a)^\top\delta_k + \iota_{\omega_k}(s,a)^\top\omega_k\big\|_{2,\nu_k},\end{aligned}\tag{A.4.9}$$

where the last inequality follows from the Cauchy-Schwartz inequality. Combining (A.4.8) and (A.4.9), we obtain for term (ii) of (A.4.3) that

$$\begin{aligned}
&\mathbb{E}_{d_{\mathrm{E}}}\Big[\Big|\big\langle \iota_{\theta_k}(s,\cdot)^\top\delta_k + \iota_{\omega_k}(s,\cdot)^\top\omega_k, \pi_{\mathrm{E}}^s - \pi_k^s\big\rangle_{\mathcal{A}}\Big|\Big] \\
&\quad\le \bigg(\Big\|\frac{\mathrm{d}\nu_{\mathrm{E}}}{\mathrm{d}\nu_k}\Big\|_{2,\nu_k} + \Big\|\frac{\mathrm{d}d_{\mathrm{E}}}{\mathrm{d}d_k}\Big\|_{2,d_k}\bigg)\cdot \big\|\iota_{\theta_k}(s,a)^\top\delta_k + \iota_{\omega_k}(s,a)^\top\omega_k\big\|_{2,\nu_k} \\
&\quad\le C_h\cdot \big\|\iota_{\theta_k}(s,a)^\top\delta_k + \iota_{\omega_k}(s,a)^\top\omega_k\big\|_{2,\nu_k}, \qquad \text{(A.4.10)}
\end{aligned}$$

where the last inequality follows from Assumption 1.4.1. To upper bound term (ii) of (A.4.3), it suffices to upper bound the right-hand side of (A.4.10). For notational simplicity, we write $\iota_{\theta_k} = \iota_{\theta_k}(s,a)$, $\iota_{\omega_k} = \iota_{\omega_k}(s,a)$, and $\phi_{\omega_k} = \phi_{\omega_k}(s,a)$. By the triangle inequality, we have that

$$\begin{aligned}
&\|\delta_k^\top\iota_{\theta_k} + \omega_k^\top\iota_{\omega_k}\|_{2,\nu_k} = \Big(\mathbb{E}_{\nu_k}\big[(\delta_k^\top\iota_{\theta_k} + \omega_k^\top\iota_{\omega_k})\cdot(\delta_k^\top\iota_{\theta_k} + \omega_k^\top\iota_{\omega_k})\big]\Big)^{1/2} \\
&\text{(A.4.11)} \\
&\quad\le \underbrace{\Big|(\delta_k-\omega_k)^\top\mathbb{E}_{\nu_k}\big[\iota_{\theta_k}(\delta_k^\top\iota_{\theta_k} + \omega_k^\top\iota_{\omega_k})\big]\Big|^{1/2}}_{\text{(ii.a)}} + \underbrace{\Big|\mathbb{E}_{\nu_k}\big[\omega_k^\top(\iota_{\theta_k}-\iota_{\omega_k})\cdot(\delta_k^\top\iota_{\theta_k} + \omega_k^\top\iota_{\omega_k})\big]\Big|^{1/2}}_{\text{(ii.b)}}.
\end{aligned}$$

We now upper bound the two terms (ii.a) and (ii.b) on the right-hand side of (A.4.11). For term (ii.a) of (A.4.11), following from (1.3.9) of Proposition 1.3.1, we have that

$$\mathcal{I}(\theta_k) = \tau_k^2\cdot\mathbb{E}_{\nu_k}[\iota_{\theta_k}\iota_{\theta_k}^\top]. \qquad \text{(A.4.12)}$$

Recall that the expectation $\mathbb{E}_k$ is taken with respect to the k-th batch. Following from the definition of $\widehat{\nabla}_\theta L(\theta_k, \beta_k)$ in (1.3.17), we have that

$$
\begin{aligned}
\mathbb{E}_k\bigl[\widehat{\nabla}_\theta L(\theta_k, \beta_k)\bigr] &= -\tau_k \cdot \mathbb{E}_{\nu_k}[\omega_k^\top \phi_{\omega_k} \cdot \iota_{\theta_k}] \\
&= -\tau_k \cdot \mathbb{E}_{\nu_k}[\omega_k^\top \iota_{\omega_k} \cdot \iota_{\theta_k}] - \tau_k \cdot w_k^\top \mathbb{E}_{a' \sim \pi_k^s}\bigl[\phi_{\omega_k}(s, a')\bigr] \cdot \mathbb{E}_{\nu_k}[\iota_{\theta_k}] \\
&= -\tau_k \cdot \mathbb{E}_{\nu_k}[\omega_k^\top \iota_{\omega_k} \cdot \iota_{\theta_k}], \qquad \text{(A.4.13)}
\end{aligned}
$$

where the first equality follows from the fact that $\widehat{Q}_{\omega_k}(s, a) = \omega_k^\top \phi_{\omega_k}(s, a)$, while the second and third equalities follow from the definition of $\iota_{\omega_k}(s, a)$ in (A.4.2). Following from (A.4.12) and (A.4.13), we have that

$$
\begin{aligned}
\Bigl|(\delta_k - \omega_k)^\top \mathbb{E}_{\nu_k}\bigl[\iota_{\theta_k}(\delta_k^\top \iota_{\theta_k} + \omega_k^\top \iota_{\omega_k})\bigr]\Bigr| &= \tau_k^{-2} \cdot \Bigl|(\delta_k - \omega_k)^\top \Bigl(\mathcal{I}(\theta_k)\delta_k - \tau_k \cdot \mathbb{E}_k\bigl[\widehat{\nabla}_\theta L(\theta, \beta)\bigr]\Bigr)\Bigr| \\
&\le 2B_\theta \cdot \tau_k^{-2} \cdot \Bigl\|\mathcal{I}(\theta_k)\delta_k - \tau_k \cdot \mathbb{E}_k\bigl[\widehat{\nabla}_\theta L(\theta, \beta)\bigr]\Bigr\|_2. \qquad \text{(A.4.14)}
\end{aligned}
$$

Here the last inequality follows from the Cauchy-Schwartz inequality and the fact that $\|\omega_k - \delta_k\|_2 \le 2B_\theta$ as $\omega_k, \delta_k \in S_{B_\theta}$. For notational simplicity, we define the following error terms,

$$
\xi_k^{(1)} = \widehat{\mathcal{I}}(\theta_k)\delta_k - \mathcal{I}(\theta_k)\delta_k, \qquad \text{(A.4.15)}
$$

$$
\xi_k^{(2)} = \widehat{\nabla}_\theta L(\theta_k, \beta_k) - \mathbb{E}_k\bigl[\widehat{\nabla}_\theta L(\theta_k, \beta_k)\bigr]. \qquad \text{(A.4.16)}
$$

Then, we have for term (ii.a) in (A.4.11) that

(A.4.17)

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\biggl[\Bigl|(\delta_k-\omega_k)^\top\mathbb{E}_{\nu_k}\bigl[\iota_{\theta_k}(\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k})\bigr]\Bigr|^{1/2}\biggr]\\
&\quad\le(2B_\theta)^{1/2}\cdot\tau_k^{-1}\cdot\mathbb{E}_{\text{init}}\biggl[\Bigl\|\mathcal{I}(\theta_k)\delta_k-\tau_k\cdot\mathbb{E}_k\bigl[\widehat{\nabla}_\theta L(\theta,\beta)\bigr]\Bigr\|_2^{1/2}\biggr]\\
&\quad\le(2B_\theta)^{1/2}\cdot\tau_k^{-1}\cdot\mathbb{E}_{\text{init}}\biggl[\Bigl(\bigl\|\widehat{\mathcal{I}}(\theta_k)\delta_k-\tau_k\cdot\widehat{\nabla}_\theta L(\theta,\beta)\bigr\|_2+\|\xi_k^{(1)}\|_2+\tau_k\cdot\|\xi_k^{(2)}\|_2\Bigr)^{1/2}\biggr]\\
&\quad\le(2B_\theta)^{1/2}\cdot\tau_k^{-1}\cdot\biggl(\mathbb{E}_{\text{init}}\Bigl[\bigl\|\widehat{\mathcal{I}}(\theta_k)\delta_k-\tau_k\cdot\widehat{\nabla}_\theta L(\theta,\beta)\bigr\|_2\Bigr]+\mathbb{E}_{\text{init}}\bigl[\|\xi_k^{(1)}\|_2+\tau_k\cdot\|\xi_k^{(2)}\|_2\bigr]\biggr)^{1/2},
\end{aligned}$$

where the first inequality follows from (A.4.14), the second inequality follows from the triangle inequality, and the last inequality follows from Jensen's inequality. Similarly to (A.4.15), we define the following error term,

$$\xi_k^{(3)}=\widehat{\mathcal{I}}(\theta_k)\omega_k-\mathcal{I}(\theta_k)\omega_k. \tag{A.4.18}$$

We now upper bound the right-hand side of (A.4.17). Recall the definition of δ_k in (1.3.15). We have that

(A.4.19)

$$\begin{aligned}
\bigl\|\widehat{\mathcal{I}}(\theta_k)\delta_k-\tau_k\cdot\widehat{\nabla}_\theta L(\theta_k,\beta_k)\bigr\|_2&\le\bigl\|\widehat{\mathcal{I}}(\theta_k)\omega_k-\tau_k\cdot\widehat{\nabla}_\theta L(\theta_k,\beta_k)\bigr\|_2\\
&\le\Bigl\|\mathcal{I}(\theta_k)\omega_k-\tau_k\cdot\mathbb{E}_k\bigl[\widehat{\nabla}_\theta L(\theta_k,\beta_k)\bigr]\Bigr\|_2+\|\xi_k^{(1)}\|_2+\tau_k\cdot\|\xi_k^{(2)}\|_2.
\end{aligned}$$

Following from (A.4.12), (A.4.13), and Jensen's inequality, we have that

$$\begin{aligned}\Big\|\mathcal{I}(\theta_k)\omega_k - \tau_k \cdot \mathbb{E}_k\big[\widehat{\nabla}_\theta L(\theta_k, \beta_k)\big]\Big\|_2 &= \tau_k^2 \cdot \Big\|\mathbb{E}_{\nu_k}\big[\iota_{\theta_k} \cdot \omega_k^\top(\iota_{\theta_k} - \iota_{\omega_k})\big]\Big\|_2 \\ &\le \tau_k^2 \cdot \mathbb{E}_{\nu_k}\Big[\|\iota_{\theta_k}\|_2 \cdot \big|\omega_k^\top(\iota_{\theta_k} - \iota_{\omega_k})\big|\Big] \\ &\le 2\tau_k^2 \cdot \big\|\omega_k^\top(\iota_{\theta_k} - \iota_{\omega_k})\big\|_{1,\nu_k},\end{aligned}$$

where the last inequality follows from the fact that $\|\iota_\theta\|_2 \le 2$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. Following from Assumption 1.4.1 and Lemma A.1.2, we have that

$$\begin{aligned}\mathbb{E}_{\text{init}}\Big[\Big\|\mathcal{I}(\theta_k)\omega_k - \tau_k \cdot \mathbb{E}_k\big[\widehat{\nabla}_\theta L(\theta_k, \beta_k)\big]\Big\|_2\Big] &\le \mathbb{E}_{\text{init}}\Big[2\tau_k^2 \cdot \big\|\omega_k^\top(\iota_{\theta_k} - \iota_{\omega_k})\big\|_{1,\nu_k}\Big] \\ &= \mathcal{O}(\tau_k^2 \cdot B_\theta^{3/2} \cdot m^{-1/4}). \qquad \text{(A.4.20)}\end{aligned}$$

Plugging (A.4.19) and (A.4.20) into (A.4.17), we have that

$$\begin{aligned}&\mathbb{E}_{\text{init}}\Big[\Big|(\delta_k - \omega_k)^\top \mathbb{E}_{\nu_k}\big[\iota_{\theta_k}(\delta_k^\top \iota_{\theta_k} + \omega_k^\top \iota_{\omega_k})\big]\Big|^{1/2}\Big] \\ &\quad = (2B_\theta)^{1/2} \cdot \tau_k^{-1} \cdot \Big(\mathcal{O}(2\tau_k^2 \cdot B_\theta^{3/2} \cdot m^{-1/4}) + \mathbb{E}_{\text{init}}\big[\|\xi_k^{(1)}\|_2 + 2\tau_k \cdot \|\xi_k^{(2)}\|_2 + \|\xi_k^{(3)}\|_2\big]\Big)^{1/2} \\ &\quad = \mathcal{O}(\tau_k \cdot B_\theta^{5/4} \cdot m^{-1/4}) + (2B_\theta)^{1/2} \cdot \tau_k^{-1} \cdot \Big(\mathbb{E}_{\text{init}}\big[\|\xi_k^{(1)}\|_2 + 2\tau_k \cdot \|\xi_k^{(2)}\|_2 + \|\xi_k^{(3)}\|_2\big]\Big)^{1/2} \\ &\text{(A.4.21)} \\ &\quad \le \mathcal{O}(\tau_k \cdot B_\theta^{5/4} \cdot m^{-1/4}) + 2\sqrt{2}B_\theta^{1/2} \cdot (\sigma^2/N)^{1/4},\end{aligned}$$

where the last inequality follows from Assumption 1.4.3. We now upper bound term (ii.a) of (A.4.11). We have that

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\Big[\Big|\mathbb{E}_{\nu_k}\big[\omega_k^\top(\iota_{\theta_k}-\iota_{\omega_k})\cdot(\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k})\big]\Big|^{1/2}\Big] \\
&\quad\le \mathbb{E}_{\text{init},\nu_k}\Big[\big|\omega_k^\top(\iota_{\theta_k}-\iota_{\omega_k})\cdot(\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k})\big|\Big]^{1/2} \\
&\quad\le \mathbb{E}_{\text{init}}\Big[\big\|\omega_k^\top(\iota_{\theta_k}-\iota_{\omega_k})\big\|_{2,\nu_k}\Big]^{1/2}\cdot\mathbb{E}_{\text{init}}\big[\|\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k}\|_{2,\nu_k}\big]^{1/2}, \qquad \text{(A.4.22)}
\end{aligned}$$

where the expectation $\mathbb{E}_{\text{init},\nu_k}$ is taken with respect to the random initialization in (1.3.3) and $(s,a)\sim\nu_k$, the first inequality follows from Jensen's inequality, and the second inequality follows from the Cauchy-Schwartz inequality. Following from Assumption 1.4.1 and Lemma A.1.2, we have that

$$\mathbb{E}_{\text{init}}\Big[\big\|\omega_k^\top(\iota_{\theta_k}-\iota_{\omega_k})\big\|_{2,\nu_k}\Big]=\mathcal{O}(B_\theta^{3/2}\cdot m^{-1/4}). \qquad \text{(A.4.23)}$$

To upper bound the right-hand side of (A.4.22), it remains to upper bound the term $\mathbb{E}_{\text{init}}[\|\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k}\|_{2,\nu_k}]$. We have that

(A.4.24)

$$\mathbb{E}_{\text{init}}\big[\|\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k}\|_{2,\nu_k}\big]\le\mathbb{E}_{\text{init}}\big[\|\delta_k\|_2\cdot\|\iota_{\theta_k}\|_2\big]+\mathbb{E}_{\text{init}}\big[\|\omega_k\|_2\cdot\|\iota_{\omega_k}\|_2\big]=\mathcal{O}(B_\theta),$$

where the inequality follows from the Cauchy-Schwartz inequality and the equality follows from the facts that $\|\iota_{\theta_k}\|_2\le 2$, $\|\iota_{\omega_k}\|_2\le 2$, and $\delta_k,\omega_k\in S_{B_\theta}$. Plugging (A.4.23) and (A.4.24) into (A.4.22), we have that

$$\mathbb{E}_{\text{init}}\Big[\Big|\mathbb{E}_{\nu_k}\big[\omega_k^\top(\iota_{\theta_k}-\iota_{\omega_k})\cdot(\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k})\big]\Big|^{1/2}\Big]=\mathcal{O}(B_\theta^{5/4}\cdot m^{-1/8}), \qquad \text{(A.4.25)}$$

which upper bounds term (ii.b) of (A.4.11). Plugging (A.4.21) and (A.4.25) into (A.4.11), following from (A.4.10), we have that

$$\mathbb{E}_{\mathrm{init},d_{\mathrm{E}}}\Big[\Big|\big\langle \iota_{\theta_k}(s,\cdot)^\top\delta_k - \iota_{\omega_k}(s,\cdot)^\top\omega_k, \pi^s_{\mathrm{E}} - \pi^s_k\big\rangle_{\mathcal{A}}\Big|\Big]$$

$$= \eta\cdot C_h\cdot\big(\mathcal{O}(B_\theta^{5/4}\cdot m^{-1/8}) + 2\sqrt{2}B_\theta^{1/2}\cdot(\sigma^2/N)^{1/4}\big), \tag{A.4.26}$$

which upper bounds term (ii) of (A.4.3).

Finally, plugging (A.4.7) and (A.4.26) into (A.4.3), we have that

$$\mathbb{E}_{\mathrm{init},d_{\mathrm{E}}}\Big[\Big|\big\langle \log(\pi^s_{k+1}/\pi^s_k) - \eta\cdot\widehat{Q}_{\omega_k}(s,\cdot), \pi^s_{\mathrm{E}} - \pi^s_k\big\rangle_{\mathcal{A}}\Big|\Big]$$

$$= \eta\cdot C_h\cdot 2\sqrt{2}B_\theta^{1/2}\cdot(\sigma^2/N)^{1/4} + \mathcal{O}(\tau_{k+1}\cdot B_\theta^{3/2}\cdot m^{-1/4} + \eta\cdot B_\theta^{5/4}\cdot m^{-1/8}),$$

where $\xi_k^{(1)}$, $\xi_k^{(2)}$, and $\xi_k^{(3)}$ are defined in (A.4.15), (A.4.16), and (A.4.18), respectively, and C_h is defined in Assumption 1.4.1. Thus, we complete the proof of Lemma A.3.1. □

A.4.2. Proof of Lemma A.3.2

Proof. For notational simplicity, for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, we denote by $\Delta_{Q,k}(s,a) = \widehat{Q}_{\omega_k}(s,a) - Q^{\pi_k}_{r_k}(s,a)$ the error of estimating $Q^{\pi_k}_{r_k}(s,a)$ by $\widehat{Q}_{\omega_k}(s,a)$. Then, we have that

$$\begin{aligned}
&\mathbb{E}_{d_{\mathrm{E}}}\Big[\Big|\big\langle \Delta_{Q,k}(s,\cdot), \pi^s_{\mathrm{E}} - \pi^s_k\big\rangle_{\mathcal{A}}\Big|\Big]\\
&\quad\le \int_{\mathcal{S}\times\mathcal{A}}\big|\Delta_{Q,k}(s,a)\big|\mathrm{d}\pi^s_{\mathrm{E}}(a)\mathrm{d}d_{\mathrm{E}}(s) + \int_{\mathcal{S}\times\mathcal{A}}\big|\Delta_{Q,k}(s,a)\big|\mathrm{d}\pi^s_k(a)\mathrm{d}d_{\mathrm{E}}(s)\\
&\quad= \int_{\mathcal{S}\times\mathcal{A}}\big|\Delta_{Q,k}(s,a)\big|\cdot\frac{\mathrm{d}\nu_{\mathrm{E}}}{\mathrm{d}\rho_k}(s,a)\mathrm{d}\rho_k(s,a) + \int_{\mathcal{S}\times\mathcal{A}}\big|\Delta_{Q,k}(s,a)\big|\cdot\frac{\mathrm{d}d_{\mathrm{E}}}{\mathrm{d}\varrho_k}(s)\mathrm{d}\rho_k(s,a)\\
&\quad\le C_h\cdot\|\Delta_{Q,k}\|_{2,\rho_k},
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwartz inequality and Assumption 1.4.1. Thus, we complete the proof of Lemma A.3.2. □

A.4.3. Proof of Lemma A.3.3

Proof. Following from (A.4.1) and the parameterization of π_θ in (1.3.5), we have that

$$\begin{aligned}\text{(A.4.27)}\qquad &\bigl\langle \log(\pi_{k+1}^s/\pi_k^s), \pi_k^s - \pi_{k+1}^s\bigr\rangle_{\mathcal{A}} \\ &= \bigl\langle \tau_{k+1}\cdot\theta_{k+1}^\top\phi_{\theta_{k+1}}(s,\cdot) - \tau_k\cdot\theta_k^\top\phi_{\theta_k}(s,\cdot), \pi_k^s - \pi_{k+1}^s\bigr\rangle_{\mathcal{A}} \\ &= \bigl\langle (\tau_{k+1}\cdot\theta_{k+1} - \tau_k\cdot\theta_k)^\top\phi_{\theta_k}(s,\cdot), \pi_k^s - \pi_{k+1}^s\bigr\rangle_{\mathcal{A}} \\ &\quad + \tau_{k+1}\cdot\Bigl\langle \theta_{k+1}^\top\bigl(\phi_{\theta_{k+1}}(s,\cdot) - \phi_{\theta_k}(s,\cdot)\bigr), \pi_k^s - \pi_{k+1}^s\Bigr\rangle_{\mathcal{A}}.\end{aligned}$$

We now upper bound the two terms on the right-hand side of (A.4.27). For the first term on the right-hand side of (A.4.27), recall that we define δ_k in (1.3.15). Thus, we have that

$$\text{(A.4.28)}\qquad \bigl|(\tau_{k+1}\cdot\theta_{k+1} - \tau_k\cdot\theta_k)^\top\phi_{\theta_k}(s,a)\bigr| = \eta\cdot\bigl|\delta_k^\top\phi_{\theta_k}(s,a)\bigr|.$$

Following from (A.4.28) and Hölder's inequality, we have for any $s\in\mathcal{S}$ that

$$\begin{aligned}&\Bigl|\bigl\langle (\tau_{k+1}\cdot\theta_{k+1} - \tau_k\cdot\theta_k)^\top\phi_{\theta_k}(s,\cdot), \pi_k^s - \pi_{k+1}^s\bigr\rangle_{\mathcal{A}}\Bigr| \\ &\quad\le \bigl\|\delta_k^\top\phi_{\theta_k}(s,\cdot)\bigr\|_\infty\cdot\|\pi_k^s - \pi_{k+1}^s\|_1.\end{aligned}$$

Then, following from Pinsker's inequality, we have that

$$
\begin{aligned}
&\Big|\big\langle(\tau_{k+1}\cdot\theta_{k+1}-\tau_k\cdot\theta_k)^\top\phi_{\theta_k}(s,\cdot),\pi_k^s-\pi_{k+1}^s\big\rangle_{\mathcal{A}}\Big| - \mathrm{KL}(\pi_{k+1}^s\,\|\,\pi_k^s)\\
&\quad\le \eta\cdot\big\|\delta_k^\top\phi_{\theta_k}(s,\cdot)\big\|_\infty\cdot\|\pi_k^s-\pi_{k+1}^s\|_1 - 1/2\cdot\|\pi_k^s-\pi_{k+1}^s\|_1^2\\
&\quad\le 1/2\cdot\eta^2\cdot\big\|\delta_k^\top\phi_{\theta_k}(s,\cdot)\big\|_\infty^2. \qquad\text{(A.4.29)}
\end{aligned}
$$

By the update of θ_k in (1.3.13) and the definition of δ_k in (1.3.15), we have that $\theta_k,\delta_k\in S_{B_\theta}$. Thus, by Lemma A.1.3, we have that

$$
\mathbb{E}_{\mathrm{init}}\Big[\big\|\delta_k^\top\phi_{\theta_k}(s,\cdot)\big\|_\infty^2\Big]\le 2M_0+18B_\theta^2. \qquad\text{(A.4.30)}
$$

Plugging (A.4.30) into (A.4.29), we have that

(A.4.31)
$$
\Big|\big\langle(\tau_{k+1}\cdot\theta_{k+1}-\tau_k\cdot\theta_k)^\top\phi_{\theta_k}(s,\cdot),\pi_k^s-\pi_{k+1}^s\big\rangle_{\mathcal{A}}\Big| - \mathrm{KL}(\pi_{k+1}^s\,\|\,\pi_k^s)\le\eta^2\cdot(M_0^2+9B_\theta^2).
$$

For the second term on the right-hand side of (A.4.27), following from Assumption 1.4.1 and Lemma A.1.2, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{init},d_{\mathrm{E}}}\bigg[\Big|\Big\langle\theta_{k+1}^\top\big(\phi_{\theta_{k+1}}(s,\cdot)-\phi_{\theta_k}(s,\cdot)\big),\pi_k^s-\pi_{k+1}^s\Big\rangle_{\mathcal{A}}\Big|\bigg]\\
&\quad\le\mathbb{E}_{\mathrm{init},d_{\mathrm{E}}}\Big[\Big\|\theta_{k+1}^\top\big(\phi_{\theta_{k+1}}(s,\cdot)-\phi_{\theta_k}(s,\cdot)\big)\Big\|_{1,\pi_k^s}\Big]\\
&\qquad+\mathbb{E}_{\mathrm{init},d_{\mathrm{E}}}\Big[\Big\|\theta_{k+1}^\top\big(\phi_{\theta_{k+1}}(s,\cdot)-\phi_{\theta_k}(s,\cdot)\big)\Big\|_{1,\pi_{k+1}^s}\Big]\\
&\quad=\mathcal{O}(B_\theta^{3/2}\cdot m^{-1/4}). \qquad\text{(A.4.32)}
\end{aligned}
$$

Finally, plugging (A.4.31) and (A.4.32) into (A.4.27), we have that

$$\mathbb{E}_{\mathrm{init},d_{\mathrm{E}}}\Big[\Big|\big\langle\log(\pi^s_{k+1}/\pi^s_k),\pi^s_k-\pi^s_{k+1}\big\rangle_{\mathcal{A}}\Big|-\mathrm{KL}(\pi^s_{k+1}\,\|\,\pi^s_k)\Big]$$
$$=\eta^2\cdot(M_0^2+9B_\theta^2)+\mathcal{O}(\tau_{k+1}\cdot B_\theta^{3/2}\cdot m^{-1/4}),$$

which completes the proof of Lemma A.3.3. □

APPENDIX B

Can Temporal-Difference and Q-Learning Learn Representation? A Mean-Field Theory

B.1. Proofs for §2.5-2.6

For notational simplicity, we denote by $\mathbb{E}_{\mathcal{D}}$ the expectation with respect to $x \sim \mathcal{D}$ and $\mathbb{E}_{\widetilde{\mathcal{D}}}$ the expectation with respect to $(x, r, x') \sim \widetilde{\mathcal{D}}$. Also, with a slight abuse of notations, we write $\theta^{(m)} = \{\theta_i\}_{i=1}^m$.

B.1.1. Proof of Lemma 2.5.1

Proof. Existence and uniqueness of Q^*. To establish the existence of the fixed point solution Q^* to the projected Bellman equation $Q = \Pi_{\mathcal{F}} \mathcal{T}^\pi Q$, it suffices to show that $\Pi_{\mathcal{F}} \mathcal{T}^\pi : \mathcal{F} \to \mathcal{F}$ is a contraction mapping. It holds for any $Q^1, Q^2 \in \mathcal{F}$ that

$$
\begin{aligned}
\|\Pi_{\mathcal{F}} \mathcal{T}^\pi Q^1 - \Pi_{\mathcal{F}} \mathcal{T}^\pi Q^2\|_{\mathcal{L}_2(\mathcal{D})}^2 &\leq \gamma^2 \cdot \mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\big(Q^1(x') - Q^2(x')\big)^2\Big] \\
&= \gamma^2 \cdot \big\|Q^1 - Q^2\big\|_{\mathcal{L}_2(\mathcal{D})}^2,
\end{aligned}
$$

where the last equality follows from the fact that $\mathcal{D}$ is the stationary distribution. Thus, $\Pi_{\mathcal{F}} \mathcal{T}^\pi : \mathcal{F} \to \mathcal{F}$ is a contraction mapping. Note that $\mathcal{F}$ is complete. Following from the Banach fixed point theorem (Conway, 2019), there exists a unique $Q^* \in \mathcal{F}$ that solves the projected Bellman equation $Q = \Pi_{\mathcal{F}} \mathcal{T}^\pi Q$. Moreover, by the definition of $\mathcal{F}$ in (2.4.3), there exists $\bar{\rho} \in \mathscr{P}_2(\mathbb{R}^D)$ such that

$$Q^*(x) = \int \sigma(x;\theta)\,\mathrm{d}\bar{\rho}(\theta).$$

Proof of (i) in Lemma 2.5.1. We define

$$\rho^* = \rho_0 + \alpha^{-1} \cdot (\bar{\rho} - \rho_0). \tag{B.1.1}$$

By the definition of $Q(\cdot;\rho)$ in (2.3.2) and the fact that $Q(x;\rho_0) = 0$, we have that $Q(x;\rho^*) = Q^*(x)$, which completes the proof of (i) in Lemma 2.5.1.

Proof of (ii) in Lemma 2.5.1. For (ii) of Lemma 2.5.1, note that $Q(\cdot;\rho^*) = \Pi_{\mathcal{F}}\mathcal{T}^{\pi}Q(\cdot;\rho^*)$. Thus, we have that

$$\big\langle Q(\cdot;\rho^*) - \mathcal{T}^{\pi}Q(\cdot;\rho^*), f(\cdot) - Q(\cdot;\rho^*)\big\rangle_{\mathcal{D}} \geq 0, \quad \forall f \in \mathcal{F},$$

which further implies that

$$\mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\big(Q(x;\rho^*) - r - \gamma \cdot Q(x';\rho^*)\big) \cdot \int \sigma(x;\theta)\,\mathrm{d}(\rho - \bar{\rho})(\theta)\Big] \geq 0, \quad \forall \rho \in \mathscr{P}_2(\mathbb{R}^D). \tag{B.1.2}$$

Let $\rho = (\mathrm{id} + h \cdot v)_\sharp \bar{\rho}$ for a sufficiently small scaling parameter $h \in \mathbb{R}_+$ and any Lipschitz-continuous mapping $v : \mathbb{R}^D \to \mathbb{R}^D$. Then, following from (B.1.2), we have that

$$\int \mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\big(Q(x;\rho^*) - r - \gamma \cdot Q(x';\rho^*)\big) \cdot \Big(\sigma\big(x;\theta + h \cdot v(\theta)\big) - \sigma(x;\theta)\Big)\Big]\,\mathrm{d}\bar{\rho}(\theta) \geq 0 \tag{B.1.3}$$

for any $v : \mathbb{R}^D \to \mathbb{R}^D$. Dividing the both sides of (B.1.3) by h and letting $h \to 0^+$, we have for any $v : \mathbb{R}^D \to \mathbb{R}^D$ that

$$\begin{aligned}0 &\le \int \mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\big(Q(x;\rho^*) - r - \gamma \cdot Q(x';\rho^*)\big) \cdot \big\langle \nabla_\theta \sigma(x;\theta), v(\theta)\big\rangle\Big] \,\mathrm{d}\bar{\rho}(\theta)\\ &= -\alpha^{-1} \cdot \int \big\langle g(\theta;\rho^*), v(\theta)\big\rangle \,\mathrm{d}\bar{\rho}(\theta),\end{aligned}$$

where the equality follows from the definition of g in (2.3.5). Thus, we have that $g(\theta;\rho^*) = 0$ for $\bar{\rho}$-a.e., which completes the proof of (ii) in Lemma 2.5.1.

Proof of (iii) in Lemma 2.5.1. Following from the definition of ρ^* in (B.1.1), we have that

$$\begin{aligned}&D_{\chi^2}(\rho^* \,\|\, \rho_0)\\ &\quad= \int \bigg(\frac{\rho^*(\theta)}{\rho_0(\theta)} - 1\bigg)^2 \mathrm{d}\rho_0(\theta) = \int \bigg(\frac{(1-\alpha^{-1}) \cdot \rho_0(\theta) + \alpha^{-1} \cdot \bar{\rho}(\theta)}{\rho_0(\theta)} - 1\bigg) \mathrm{d}\rho_0(\theta) = \alpha^{-2} \cdot \bar{D}^2,\end{aligned}$$

where $\bar{D} = D_{\chi^2}(\bar{\rho} \,\|\, \rho_0)^{1/2}$. By Lemma B.3.3, we have that

$$\mathcal{W}_2(\rho^*, \rho_0) \le D_{\mathrm{KL}}(\rho^* \,\|\, \rho_0)^{1/2} \le D_{\chi^2}(\rho^* \,\|\, \rho_0)^{1/2} \le \alpha^{-1} \cdot \bar{D},$$

which completes the proof of (iii) in Lemma 2.5.1. □

B.1.2. Proof of Lemma 2.5.2

We first introduce the following lemmas. The first lemma establishes the strongly monotonicity of $g(\cdot;\beta_t)$ along a curve $\beta : [0,1] \to \mathscr{P}_2(\mathbb{R}^D)$ on the Wasserstein space.

Lemma B.1.1. Let $\beta : [0,1] \to \mathscr{P}_2(\mathbb{R}^D)$ be a curve such that $\partial_t \beta_t = -\operatorname{div}(\beta_t \cdot v_t)$ for a vector field v. We have that

$$\big\langle \partial_t g(\cdot;\beta_t), v_t \big\rangle_{\beta_t} \le -(1-\gamma) \cdot \mathbb{E}_{\mathcal{D}}\Big[\big(\partial_t Q(x;\beta_t)\big)^2\Big].$$

Furthermore, we have that

$$\int_0^1 \big\langle \partial_s g(\cdot;\beta_s), v_s \big\rangle_{\beta_s} \,\mathrm{d}s \le -(1-\gamma) \cdot \mathbb{E}_{\mathcal{D}}\Big[\big(Q(x;\beta_0) - Q(x;\beta_1)\big)^2\Big]. \tag{B.1.4}$$

Proof. Following from the definition of g in (2.3.5), we have that

$$\partial_t g(\theta;\beta_t) = -\alpha \cdot \mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\partial_t\big(Q(x;\beta_t) - \gamma \cdot Q(x';\beta_t)\big) \cdot \nabla_\theta \sigma(x;\theta)\Big].$$

Thus, following from integration by parts and the continuity equation $\partial_t \beta_t = -\operatorname{div}(\beta_t \cdot v_t)$, we have that

$$\begin{aligned}
\big\langle \partial_t g(\cdot;\beta_t), v_t \big\rangle_{\beta_t} &= -\int \Big\langle \alpha \cdot \mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\partial_t\big(Q(x;\beta_t) - \gamma \cdot Q(x';\beta_t)\big) \cdot \nabla_\theta \sigma(x;\theta)\Big], v_t(\theta) \cdot \beta_t(\theta) \Big\rangle \,\mathrm{d}\theta \\
&= -\int \alpha \cdot \mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\partial_t\big(Q(x;\beta_t) - \gamma \cdot Q(x';\beta_t)\big) \cdot \sigma(x;\theta)\Big] \cdot \partial_t \beta_t(\theta) \,\mathrm{d}\theta \\
&= -\mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\partial_t\big(Q(x;\beta_t) - \gamma \cdot Q(x';\beta_t)\big) \cdot \partial_t Q(x;\beta_t)\Big], \qquad \text{(B.1.5)}
\end{aligned}$$

where the last equality follows from the definition of Q in (2.3.2). Applying the Cauchy-Schwartz inequality to (B.1.5), we have that

$$\begin{aligned}\big\langle \partial_t g(\cdot;\beta_t), v_t\big\rangle_{\beta_t} &= -\mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\big(\partial_t Q(x;\beta_t)\big)^2\Big] + \gamma\cdot\mathbb{E}_{\widetilde{\mathcal{D}}}\big[\partial_t Q(x';\beta_t)\cdot\partial_t Q(x;\beta_t)\big]\\ &\le -\mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\big(\partial_t Q(x;\beta_t)\big)^2\Big] + \gamma\cdot\mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\big(\partial_t Q(x;\beta_t)\big)^2\Big]^{1/2}\cdot\mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\big(\partial_t Q(x';\beta_t)\big)^2\Big]^{1/2}\\ &= -(1-\gamma)\cdot\mathbb{E}_{\mathcal{D}}\Big[\big(\partial_t Q(x;\beta_t)\big)^2\Big],\end{aligned}\tag{B.1.6}$$

where the last equality follows from the fact that the marginal distributions of $\widetilde{\mathcal{D}}$ with respect to x and x' are $\mathcal{D}$, since $\mathcal{D}$ is the stationary distribution. Furthermore, we have that

$$\begin{aligned}\int_0^1 \big\langle \partial_s g(\cdot;\beta_s), v_s\big\rangle_{\beta_s}\,\mathrm{d}s &\le -(1-\gamma)\cdot\int_0^1 \mathbb{E}_{\mathcal{D}}\Big[\big(\partial_s Q(x;\beta_s)\big)^2\Big]\,\mathrm{d}s\\ &\le -(1-\gamma)\cdot\mathbb{E}_{\mathcal{D}}\Big[\Big(\int_0^1 \partial_s Q(x;\beta_s)\,\mathrm{d}s\Big)^2\Big]\\ &= -(1-\gamma)\cdot\mathbb{E}_{\mathcal{D}}\Big[\big(Q(x;\beta_1)-Q(x;\beta_0)\big)^2\Big],\end{aligned}$$

which completes the proof of Lemma B.1.1. □

The following lemma upper bounds the norms of Q and $\nabla_\theta g$.

Lemma B.1.2. Under Assumptions 2.4.1 and 2.4.2, it holds for any $\rho\in\mathscr{P}_2(\mathbb{R}^D)$ that

$$\sup_{x\in\mathcal{X}}\big|Q(x;\rho)\big| \le \alpha\cdot\min\big\{B_1\cdot\mathcal{W}_2(\rho,\rho_0),\ B_0\big\},\tag{B.1.7}$$

$$\sup_{\theta\in\mathbb{R}^D}\big\|\nabla_\theta g(\theta;\rho)\big\|_{\mathrm{F}} \le \alpha\cdot B_2\cdot\min\big\{2\alpha\cdot B_1\cdot\mathcal{W}_2(\rho,\rho_0)+B_r,\ 2\alpha\cdot B_0+B_r\big\}.\tag{B.1.8}$$

Proof. We introduce the Wasserstein-1 distance, which is defined as

$$\mathcal{W}_1(\mu^1, \mu^2) = \inf\Big\{\mathbb{E}\big[\|X - Y\|\big] \,\Big|\, \text{law}(X) = \mu^1, \text{law}(Y) = \mu^2\Big\}$$

for any $\mu^1, \mu^2 \in \mathscr{P}(\mathbb{R}^D)$ with finite first moments. Thus, we have that $\mathcal{W}_1(\mu^1, \mu^2) \leq \mathcal{W}_2(\mu^1, \mu^2)$. The Wasserstein-1 distance has the following dual representation (Ambrosio et al., 2008),

$$\mathcal{W}_1(\mu^1, \mu^2) = \sup\bigg\{\int f(x)\,\mathrm{d}(\mu^1 - \mu^2)(x) \,\bigg|\, \text{continuous } f : \mathbb{R}^D \to \mathbb{R}, \text{Lip}(f) \leq 1\bigg\}. \tag{B.1.9}$$

Following from Assumptions 2.4.1 and 2.4.2, we have that $\|\nabla_\theta \sigma(x; \theta)\| \leq B_1$ for any $x \in \mathcal{X}$ and $\theta \in \mathbb{R}^D$, which implies that $\text{Lip}(\sigma(x; \cdot)/B_1) \leq 1$ for any $x \in \mathcal{X}$. Note that $Q(x; \rho_0) = 0$ for any $x \in \mathcal{X}$. Thus, by (B.1.9) we have for any $\rho \in \mathscr{P}_2(\mathbb{R}^D)$ and $x \in \mathcal{X}$ that

$$\big|Q(x; \rho)\big| = \alpha \cdot \bigg|\int \sigma(x; \theta) \cdot \,\mathrm{d}(\rho - \rho_0)(\theta)\bigg| \leq \alpha \cdot B_1 \cdot \mathcal{W}_1(\rho, \rho_0) \leq \alpha \cdot B_1 \cdot \mathcal{W}_2(\rho, \rho_0). \tag{B.1.10}$$

Meanwhile, following from Assumptions 2.4.1 and 2.4.2, we have for any $x \in \mathcal{X}$ and $\rho \in \mathscr{P}_2(\mathbb{R}^D)$ that

$$\big|Q(x; \rho)\big| = \alpha \cdot \bigg|\int \sigma(x; \theta)\,\mathrm{d}\rho(\theta)\bigg| \leq \alpha \cdot B_0. \tag{B.1.11}$$

Combining (B.1.10) and (B.1.11), we have for any $\rho \in \mathscr{P}_2(\mathbb{R}^D)$ that

$$\sup_{x \in \mathcal{X}} \big|Q(x; \rho)\big| \leq \alpha \cdot \min\big\{B_1 \cdot \mathcal{W}_2(\rho, \rho_0),\, B_0\big\}, \tag{B.1.12}$$

which completes the proof of (B.1.7) in Lemma B.1.2. Following from the definition of g in (2.3.5), we have for any $x \in \mathcal{X}$ and $\rho \in \mathscr{P}_2(\mathbb{R}^D)$ that

$$
\begin{aligned}
\big\|\nabla_\theta g(\theta;\rho)\big\|_{\mathrm{F}} &\leq \alpha \cdot \mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\big|Q(x;\rho) - r - \gamma \cdot Q(x';\rho)\big| \cdot \big\|\nabla^2_{\theta\theta}\sigma(x;\theta)\big\|_{\mathrm{F}}\Big] \\
&\leq \alpha \cdot \min\big\{2\alpha \cdot B_1 \cdot \mathcal{W}_2(\rho,\rho_0) + B_r,\, 2\alpha \cdot B_0 + B_r\big\} \cdot B_2.
\end{aligned}
$$

Here the last inequality follows from (B.1.12) and the fact that $\|\nabla^2_{\theta\theta}\sigma(x;\theta)\|_{\mathrm{F}} \leq B_2$ for any $x \in \mathcal{X}$ and $\rho \in \mathscr{P}_2(\mathbb{R}^D)$, which follows from Assumptions 2.4.1 and 2.4.2. Thus, we complete the proof of Lemma B.1.2. □

We are now ready to present the proof of Lemma 2.5.2.

Proof. Recall that ρ_t is the PDE solution in (2.3.4), that is,

$$
\partial_t \rho_t = -\eta \cdot \operatorname{div}\big(\rho_t \cdot g(\cdot;\rho_t)\big),
$$

where

$$
g(\theta;\rho) = -\alpha \cdot \mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\big(Q(x;\rho) - r - \gamma \cdot Q(x';\rho)\big) \cdot \nabla_\theta \sigma(x;\theta)\Big].
$$

We fix a $t \in [0,T]$. We denote by $\beta : [0,1] \to \mathscr{P}_2(\mathbb{R}^D)$ the geodesic connecting ρ_t and ρ^*. Specifically, β satisfies that $\beta'_s = -\operatorname{div}(\beta_s \cdot v_s)$ for a vector field v. Following from Lemma

B.3.2, we have that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\rho_t,\rho^*)^2}{2} &= -\eta\cdot\big\langle g(\cdot;\rho_t),v_0\big\rangle_{\rho_t}\\ &= \eta\cdot\int_0^1 \partial_s\big\langle g(\cdot;\beta_s),v_s\big\rangle_{\beta_s}\,\mathrm{d}s - \eta\cdot\big\langle g(\cdot;\rho^*),v_1\big\rangle_{\rho^*}\\ &= \eta\cdot\underbrace{\int_0^1\big\langle\partial_s g(\cdot;\beta_s),v_s\big\rangle_{\beta_s}\,\mathrm{d}s}_{\text{(i)}}+\eta\cdot\underbrace{\int_0^1\int\big\langle g(\theta;\beta_s),\partial_s(v_s\cdot\beta_s)(\theta)\big\rangle\,\mathrm{d}\theta\,\mathrm{d}s}_{\text{(ii)}},\end{aligned}\tag{B.1.13}$$

where the last equality follows from (ii) of Lemma 2.5.1.

For term (i) of (B.1.13), following from (B.1.4) of Lemma B.1.1, we have that

$$\begin{aligned}\int_0^1\big\langle\partial_s g(\cdot;\beta_s),v_s\big\rangle_{\beta_s}\,\mathrm{d}s &\le -(1-\gamma)\cdot\mathbb{E}_{\mathcal{D}}\Big[\big(Q(x;\beta_0)-Q(x;\beta_1)\big)^2\Big]\\ &= -(1-\gamma)\cdot\mathbb{E}_{\mathcal{D}}\Big[\big(Q(x;\rho_t)-Q^*(x)\big)^2\Big].\end{aligned}\tag{B.1.14}$$

For term (ii) of (B.1.14), we have that

$$\begin{aligned}\int\Big|\big\langle g(\theta;\beta_s),\partial_s(v_s\cdot\beta_s)(\theta)\big\rangle\Big|\,\mathrm{d}\theta &= \int\Big|\big\langle\nabla_\theta g(\theta;\beta_s),\beta_s(\theta)\cdot v_s(\theta)\otimes v_s(\theta)\big\rangle\Big|\,\mathrm{d}\theta\\ &\le \sup_{\theta\in\mathbb{R}^D}\big\|\nabla_\theta g(\theta;\beta_s)\big\|_{\mathrm{F}}\cdot\|v_s\|_{\beta_s}^2,\end{aligned}$$

where the equality follows from integration by parts and Lemma B.3.4. Since β is the geodesic connecting ρ_t and ρ^*, (2.2.7) implies that $\|v_s\|_{\beta_s}^2=\mathcal{W}_2(\beta_0,\beta_1)^2=\mathcal{W}_2(\rho_t,\rho^*)^2$ for any $s\in[0,1]$. Applying (B.1.8) of Lemma B.1.2, we have that

$$\begin{aligned}\int\Big|\big\langle g(\theta;\beta_s),\partial_s(v_s\cdot\beta_s)(\theta)\big\rangle\Big|\,\mathrm{d}\theta &\le \alpha\cdot B_2\cdot\big(2\alpha\cdot B_1\cdot\mathcal{W}_2(\rho_t,\rho_0)+B_r\big)\cdot\mathcal{W}_2(\rho_t,\rho^*)^2\\ &\le 4\alpha\cdot B_2\cdot\big(6\alpha\cdot B_1\cdot\mathcal{W}_2(\rho_0,\rho^*)+B_r\big)\cdot\mathcal{W}_2(\rho_0,\rho^*)^2,\end{aligned}\tag{B.1.15}$$

where the last inequality follows from the condition of Lemma 2.5.2 that $\mathcal{W}_2(\rho_t, \rho^*) \leq 2\mathcal{W}_2(\rho_0, \rho^*)$ and the fact that $\mathcal{W}_2(\rho_t, \rho_0) \leq \mathcal{W}_2(\rho_t, \rho^*) + \mathcal{W}_2(\rho_0, \rho^*)$. Then, applying (iii) of Lemma 2.5.1 to (B.1.15), we have that

$$\begin{aligned} \int_0^1 \int \Big| \big\langle g(\theta; \beta_s), \partial_s (v_s \cdot \beta_s)(\theta) \big\rangle \Big| \, \mathrm{d}\theta \, \mathrm{d}s &\leq 4\alpha^{-1} \cdot B_2 \cdot \bar{D}^2 \cdot (6B_1 \cdot \bar{D} + B_r) \\ &= C_* \cdot \alpha^{-1}, \end{aligned} \tag{B.1.16}$$

where $C_* > 0$ is a constant depending on $\bar{D}$, B_1, B_2, and B_r.

Finally, plugging (B.1.14) and (B.1.16) into (B.1.13), we have that

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{\mathcal{W}_2(\rho_t, \rho^*)^2}{2} \leq -(1-\gamma) \cdot \eta \cdot \mathbb{E}_{\mathcal{D}} \Big[\big(Q(x; \rho_t) - Q^*(x) \big)^2 \Big] + C_* \cdot \alpha^{-1} \cdot \eta,$$

which completes the proof of Lemma 2.5.2. □

B.1.3. Proof of Theorem 2.6.2

Proof. In parallel to the proof of Lemma 2.5.1 in §B.1.1, to establish the existence and uniqueness of the fixed point solution to the projected Bellman optimality equation $Q = \Pi_{\mathcal{F}} \mathcal{T}^* Q$, it suffices to show that $\Pi_{\mathcal{F}} \mathcal{T}^* : \mathcal{F} \to \mathcal{F}$ is a contraction mapping. In particular, it holds for any $Q^1, Q^2 \in \mathcal{F}$ that

$$\begin{aligned} \|\Pi_{\mathcal{F}} \mathcal{T}^* Q^1 - \Pi_{\mathcal{F}} \mathcal{T}^* Q^2\|^2_{\mathcal{L}_2(\mathcal{D}_{\mathrm{E}})} &\leq \gamma^2 \cdot \mathbb{E}_{\widetilde{\mathcal{D}}_{\mathrm{E}}} \Big[\big(\max_{a \in \mathcal{A}} Q^1(s', \underline{a}) - \max_{a \in \mathcal{A}} Q^2(s', \underline{a}) \big)^2 \Big] \\ &= \gamma^2 \cdot \mathbb{E}_{\mathcal{D}_{\mathrm{E}}} \Big[\big(\max_{a \in \mathcal{A}} Q^1(s, \underline{a}) - \max_{a \in \mathcal{A}} Q^2(s, \underline{a}) \big)^2 \Big] \\ &\leq \frac{\gamma^2}{(\gamma + \kappa)^2} \cdot \mathbb{E}_{\mathcal{D}_{\mathrm{E}}} \Big[\big(Q^1(s, a) - Q^2(s, a) \big)^2 \Big], \end{aligned}$$

where the equality follows from the fact that $\mathcal{D}_{\mathrm{E}}$ is the stationary distribution and the last inequality follows from Assumption 2.6.1. Thus, $\Pi_{\mathcal{F}}\mathcal{T}^* : \mathcal{F} \to \mathcal{F}$ is a contraction mapping. Following from the Banach fixed point theorem (Conway, 2019), there exists a unique fixed point solution $Q^\dagger \in \mathcal{F}$ to the projected Bellman optimality equation $Q = \Pi_{\mathcal{F}}\mathcal{T}^*Q$. Moreover, in parallel to the proof of Lemma 2.5.1 in §B.1.1, there exists $\nu^\dagger \in \mathscr{P}_2(\mathbb{R}^D)$ such that $Q(x;\nu^\dagger) = Q^\dagger(x)$, $h(x;\nu^\dagger) = 0$, and $\mathcal{W}_2(\nu^\dagger, \nu_0) \le \alpha^{-1} \cdot \bar{D}$, where $\bar{D} = D_{\chi^2}(\bar{\nu} \,\|\, \nu_0)^{1/2}$.

For notational simplicity, we define $Q^{\mathcal{A}}(x) = \max_{\underline{a} \in \mathcal{A}} Q(s, \underline{a})$. In parallel to (B.1.13) in the proof of Lemma 2.5.2 in §B.1.2, we have that

(B.1.17)

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\nu_t, \nu^\dagger)^2}{2} = \eta \cdot \underbrace{\int_0^1 \big\langle \partial_s h(\cdot;\beta_s), v_s \big\rangle_{\beta_s} \,\mathrm{d}s}_{\text{(i)}} + \eta \cdot \underbrace{\int_0^1 \int \big\langle h(\theta;\beta_s), \partial_s(v_s \cdot \beta_s)(\theta) \big\rangle \,\mathrm{d}\theta \,\mathrm{d}s}_{\text{(ii)}},$$

where $\beta : [0,1] \to \mathscr{P}_2(\mathbb{R}^D)$ is the geodesic connecting ν_t and $\nu^\dagger$ with $\partial_s\beta_s = -\operatorname{div}(\beta_s \cdot v_s)$.

Upper bounding term (i) of (B.1.17)**.** In parallel to (B.1.5) and (B.1.6) in the proof of Lemma B.1.1, we have that

(B.1.18)

$$\begin{aligned}\big\langle \partial_s h(\cdot;\beta_s), v_s \big\rangle_{\beta_s} &= -\mathbb{E}_{\tilde{\mathcal{D}}_{\mathrm{E}}}\Big[\partial_s\big(Q(x;\beta_s) - \gamma \cdot Q^{\mathcal{A}}(x';\beta_s)\big) \cdot \partial_s Q(x;\beta_s)\Big] \\ &\le -\mathbb{E}_{\mathcal{D}_{\mathrm{E}}}\Big[\big(\partial_s Q(x;\beta_s)\big)^2\Big] + \gamma \cdot \mathbb{E}_{\mathcal{D}_{\mathrm{E}}}\Big[\big(\partial_s Q(x;\beta_s)\big)^2\Big]^{1/2} \cdot \mathbb{E}_{\mathcal{D}_{\mathrm{E}}}\Big[\big(\partial_s Q^{\mathcal{A}}(x;\beta_s)\big)^2\Big]^{1/2}.\end{aligned}$$

For the second term on the right-hand side of (B.1.18), we have that

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}_{\mathrm{E}}}\Big[\big(\partial_s Q^{\mathcal{A}}(x;\beta_s)\big)^2\Big] &= \lim_{u\to 0}\mathbb{E}_{\mathcal{D}_{\mathrm{E}}}\Big[\Big(u^{-1}\cdot\big(Q^{\mathcal{A}}(x;\beta_{s+u}) - Q^{\mathcal{A}}(x;\beta_s)\big)\Big)^2\Big]\\
&\le (\gamma+\kappa)^{-2}\cdot\lim_{u\to 0}u^{-2}\cdot\mathbb{E}_{\mathcal{D}_{\mathrm{E}}}\Big[\big(Q(x;\beta_{s+u}) - Q(x;\beta_s)\big)^2\Big]\\
&= (\gamma+\kappa)^{-2}\cdot\mathbb{E}_{\mathcal{D}_{\mathrm{E}}}\Big[\big(\partial_s Q(x;\beta_s)\big)^2\Big], \qquad \text{(B.1.19)}
\end{aligned}
$$

where the inequality follows from Assumption 2.6.1 and the fact that $Q(\cdot;\nu)\in\alpha\cdot\mathcal{F}$. Plugging (B.1.19) into (B.1.18), we have that

$$
\big\langle \partial_s h(\cdot;\beta_s), v_s\big\rangle_{\beta_s} \le -\frac{\kappa}{\gamma+\kappa}\cdot\mathbb{E}_{\mathcal{D}_{\mathrm{E}}}\Big[\big(\partial_s Q(x;\beta_s)\big)^2\Big],
$$

which further implies that

$$
\begin{aligned}
\int_0^1\big\langle \partial_s h(\cdot;\beta_s), v_s\big\rangle_{\beta_s}\,\mathrm{d}s &\le -\frac{\kappa}{\gamma+\kappa}\cdot\int_0^1\mathbb{E}_{\mathcal{D}_{\mathrm{E}}}\Big[\big(\partial_s Q(x;\beta_s)\big)^2\Big]\,\mathrm{d}s\\
&\le -\frac{\kappa}{\gamma+\kappa}\cdot\mathbb{E}_{\mathcal{D}_{\mathrm{E}}}\Big[\Big(\int_0^1\partial_s Q(x;\beta_s)\,\mathrm{d}s\Big)^2\Big]\\
&= -\frac{\kappa}{\gamma+\kappa}\cdot\mathbb{E}_{\mathcal{D}_{\mathrm{E}}}\Big[\big(Q(x;\nu_t) - Q(x;\nu^\dagger)\big)^2\Big]. \qquad \text{(B.1.20)}
\end{aligned}
$$

Upper bounding term (ii) of (B.1.17)**.** In parallel to the proof of Lemma B.1.2 in §B.1.2, noting that $|Q^{\mathcal{A}}(x;\nu)|\le\sup_{x\in\mathcal{X}}|Q(x;\nu)|$ for any $\nu\in\mathscr{P}_2(\mathbb{R}^D)$, we have that

$$
\big\|\nabla_\theta h(\theta;\nu_t)\big\|_{\mathrm{F}} \le \alpha\cdot B_2\cdot\big(2\alpha\cdot B_1\cdot\mathcal{W}_2(\nu_t,\nu_0) + B_r\big).
$$

In parallel to (B.1.15) and (B.1.16), we have that

$$\int_0^1 \int \Big| \big\langle h(\theta;\beta_s), \partial_s(v_s \cdot \beta_s)(\theta) \big\rangle \Big| \, \mathrm{d}\theta \, \mathrm{d}s \le C_* \cdot \alpha^{-1}, \tag{B.1.21}$$

where $C_* > 0$ is a constant that depends on $\bar{D}$, B_1, B_2, and B_r.

Plugging (B.1.20) and (B.1.21) into (B.1.17), we have that

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{\mathcal{W}_2(\nu_t, \nu^\dagger)^2}{2} \le -\frac{\eta \cdot \kappa}{\gamma + \kappa} \cdot \mathbb{E}_{\mathcal{D}_\mathrm{E}} \Big[\big(Q(x;\nu_t) - Q(x;\nu^\dagger) \big)^2 \Big] + C_* \cdot \eta \cdot \alpha^{-1}.$$

Thus, in parallel to the proof of Theorem 2.4.3 in §2.5, we have that

$$\inf_{t \in [0,T]} \mathbb{E}_{\mathcal{D}} \Big[\big(Q(x;\nu_t) - Q^\dagger(x) \big)^2 \Big] \le \frac{(\kappa + \gamma) \cdot D_{\chi^2}(\bar{\nu} \,\|\, \nu_0)}{2\kappa \cdot T} + C_* \cdot \alpha^{-1} \cdot \frac{\kappa + \gamma}{\kappa},$$

which completes the proof of (2.6.5) in Theorem 2.6.2. Meanwhile, in parallel to the proof of Lemma 2.4.4 in §B.2.2, we upper bound the error of approximating $\widehat{\nu}_k$ by $\nu_{k\epsilon}$, which further implies (2.6.6) of Theorem 2.6.2. □

B.2. Mean-Field Limit of Neural Networks

In this section, we prove Proposition 2.3.1, whose formal version is presented as follows. Recall that ρ_t is the PDE solution in (2.3.4) and $\widehat{\rho}_k = m^{-1} \cdot \sum_{i=1}^m \theta_i(k)$ is the empirical distribution of $\theta^{(m)}(k) = \{\theta_i(k)\}_{i=1}^m$. Note that we omit the dependence of $\widehat{\rho}_k$ on m and ϵ for notational simplicity.

Proposition B.2.1 (Formal Version of Proposition 2.3.1)**.** Let $f : \mathbb{R}^D \to \mathbb{R}$ be any continuous function such that $\|f\|_\infty \le 1$ and $\mathrm{Lip}(f) \le 1$. Under Assumptions 2.4.1 and

2.4.2, it holds that

$$\sup_{\substack{k \le T/\epsilon \\ (k \in \mathbb{N})}} \biggl| \int f(\theta)\,\mathrm{d}\rho_{k\epsilon}(\theta) - \int f(\theta)\,\mathrm{d}\widehat{\rho}_k(\theta) \biggr| \\ \le B \cdot e^{BT} \cdot \Bigl(\sqrt{\log(m/\delta)/m} + \sqrt{\epsilon \cdot \bigl(D + \log(m/\delta)\bigr)} \Bigr)$$

with probability at least $1-\delta$. Here B is a constant that depends on α, η, γ, B_r, and B_j $(j \in \{0,1,2\})$.

The proof of Proposition B.2.1 is based on Mei et al. (2018, 2019); Araújo et al. (2019), which utilizes the propagation of chaos (Sznitman, 1991). Recall that $g(\cdot;\rho)$ is a vector field defined as follows,

$$g(\theta;\rho) = -\alpha \cdot \mathbb{E}_{\widetilde{\mathcal{D}}}\Bigl[\bigl(Q(x;\rho) - r - \gamma \cdot Q(x';\rho)\bigr) \cdot \nabla_\theta \sigma(x;\theta)\Bigr].$$

Correspondingly, we define the finite-width and stochastic counterparts of $g(\theta;\rho)$ as follows,

$$\widehat{g}(\theta;\theta^{(m)}) = -\alpha \cdot \mathbb{E}_{\widetilde{\mathcal{D}}}\Bigl[\bigl(\widehat{Q}(x;\theta^{(m)}) - r - \gamma \cdot \widehat{Q}(x';\theta^{(m)})\bigr) \cdot \nabla_\theta \sigma(x;\theta)\Bigr], \tag{B.2.1}$$

$$\widehat{G}_k(\theta;\theta^{(m)}) = -\alpha \cdot \bigl(\widehat{Q}(x_k;\theta^{(m)}) - r_k - \gamma \cdot \widehat{Q}(x'_k;\theta^{(m)})\bigr) \cdot \nabla_\theta \sigma(x_k;\theta), \tag{B.2.2}$$

where $(x_k, r_k, x'_k) \sim \widetilde{\mathcal{D}}$. Following from Mei et al. (2019); Araújo et al. (2019), we consider the following four dynamics.

- **Temporal-difference (TD).** We consider the following TD dynamics $\theta^{(m)}(k)$, where $k \in \mathbb{N}$, with $\theta_i(0) \overset{\text{i.i.d.}}{\sim} \rho_0$ ($i \in [m]$) as its initialization,

$$\begin{aligned}\theta_i(k+1) &= \theta_i(k) - \eta\epsilon \cdot \alpha \cdot \Big(\widehat{Q}\big(x_k; \theta^{(m)}(k)\big) - r_k - \gamma \cdot \widehat{Q}\big(x_k'; \theta^{(m)}(k)\big)\Big) \cdot \nabla_\theta \sigma\big(x_k; \theta_i(k)\big) \\ &= \theta_i(k) + \eta\epsilon \cdot \widehat{G}_k\big(\theta_i(k); \theta^{(m)}(k)\big), \qquad \text{(B.2.3)}\end{aligned}$$

where $(x_k, r_k, x_k') \sim \widetilde{\mathcal{D}}$. Note that this definition is equivalent to (2.2.3).

- **Expected temporal-difference (ETD).** We consider the following expected TD dynamics $\breve{\theta}^{(m)}(k)$, where $k \in \mathbb{N}$, with $\breve{\theta}_i(0) = \theta_i(0)$ ($i \in [m]$) as its initialization,

$$\begin{aligned}\breve{\theta}_i(k+1) &= \breve{\theta}_i(k) - \eta\epsilon \cdot \alpha \cdot \mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\Big(\widehat{Q}\big(x; \breve{\theta}^{(m)}(k)\big) - r - \gamma \cdot \widehat{Q}\big(x'; \breve{\theta}^{(m)}(k)\big)\Big) \cdot \nabla_\theta \sigma\big(x; \breve{\theta}_i(k)\big)\Big] \\ &= \breve{\theta}_i(k) + \eta\epsilon \cdot \widehat{g}\big(\breve{\theta}_i(k); \breve{\theta}^{(m)}(k)\big). \qquad \text{(B.2.4)}\end{aligned}$$

- **Continuous-time temporal-difference (CTTD).** We consider the following continuous-time TD dynamics $\widetilde{\theta}^{(m)}(t)$, where $t \in \mathbb{R}_+$, with $\widetilde{\theta}_i(0) = \theta_i(0)$ ($i \in [m]$) as its initialization,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\theta}_i(t) &= -\eta \cdot \alpha \cdot \mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\Big(\widehat{Q}\big(x; \widetilde{\theta}^{(m)}(t)\big) - r - \gamma \cdot \widehat{Q}\big(x'; \widetilde{\theta}^{(m)}(t)\big)\Big) \cdot \nabla_\theta \sigma\big(x; \widetilde{\theta}_i(t)\big)\Big] \\ &= \eta \cdot \widehat{g}\big(\widetilde{\theta}_i(t); \widetilde{\theta}^{(m)}(t)\big). \qquad \text{(B.2.5)}\end{aligned}$$

- **Ideal particle (IP).** We consider the following ideal particle dynamics $\bar{\theta}^{(m)}(t)$, where $t \in \mathbb{R}_+$, with $\bar{\theta}_i(0) = \theta_i(0)$ ($i \in [m]$) as its initialization,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\bar{\theta}_i(t) &= -\eta \cdot \alpha \cdot \mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\big(Q(x; \rho_t) - r - \gamma \cdot Q(x'; \rho_t)\big) \cdot \nabla_\theta \sigma\big(x; \bar{\theta}_i(t)\big)\Big] \\ &= \eta \cdot g\big(\bar{\theta}_i(t); \rho_t\big), \qquad \text{(B.2.6)}\end{aligned}$$

where ρ_t is the PDE solution in (2.3.4).

We aim to prove that $\widehat{\rho}_k = m^{-1} \cdot \sum_{i=1}^m \delta_{\theta_i(k)}$ weakly converges to $\rho_{k\epsilon}$. For any continuous function $f : \mathbb{R}^D \to \mathbb{R}$ such that $\|f\|_\infty \leq 1$ and $\mathrm{Lip}(f) \leq 1$, we use the IP, CTTD, and ETD dynamics as the interpolating dynamics,

$$
\begin{aligned}
&\overbrace{\left|\int f(\theta)\,\mathrm{d}\rho_{k\epsilon}(\theta) - \int f(\theta)\,\mathrm{d}\widehat{\rho}_k(\theta)\right|}^{\mathrm{PDE-TD}} \\
&\quad \leq \left|\int f(\theta)\,\mathrm{d}\rho_{k\epsilon}(\theta) - m^{-1} \cdot \sum_{i=1}^m f\bigl(\bar{\theta}_i(k\epsilon)\bigr)\right| + \left|m^{-1} \cdot \sum_{i=1}^m f\bigl(\bar{\theta}_i(k\epsilon)\bigr) - m^{-1} \cdot \sum_{i=1}^m f\bigl(\widetilde{\theta}_i(k\epsilon)\bigr)\right| \\
&\qquad + \left|m^{-1} \cdot \sum_{i=1}^m f\bigl(\widetilde{\theta}_i(k\epsilon)\bigr) - m^{-1} \cdot \sum_{i=1}^m f\bigl(\breve{\theta}_i(k)\bigr)\right| \\
&\qquad + \left|m^{-1} \cdot \sum_{i=1}^m f\bigl(\breve{\theta}_i(k)\bigr) - m^{-1} \cdot \sum_{i=1}^m f\bigl(\theta_i(k)\bigr)\right| \\
&\quad \leq \underbrace{\left|\int f(\theta)\,\mathrm{d}\rho_{k\epsilon}(\theta) - m^{-1} \cdot \sum_{i=1}^m f\bigl(\bar{\theta}_i(k\epsilon)\bigr)\right|}_{\mathrm{PDE-IP}} + \underbrace{\bigl\|\bar{\theta}^{(m)}(k\epsilon) - \widetilde{\theta}^{(m)}(k\epsilon)\bigr\|_{(m)}}_{\mathrm{IP-CTTD}}
\end{aligned}
$$

(B.2.7)

$$
\qquad + \underbrace{\bigl\|\widetilde{\theta}^{(m)}(k\epsilon) - \breve{\theta}^{(m)}(k)\bigr\|_{(m)}}_{\mathrm{CTTD-ETD}} + \underbrace{\bigl\|\breve{\theta}^{(m)}(k) - \theta^{(m)}(k)\bigr\|_{(m)}}_{\mathrm{ETD-TD}},
$$

where the last inequality follows from the the fact that $\mathrm{Lip}(f) \leq 1$. Here the norm $\|\cdot\|_{(m)}$ of $\theta^{(m)} = \{\theta_i\}_{i=1}^m$ is defined as follows,

$$
\|\theta^{(m)}\|_{(m)} = \sup_{i\in[m]} \|\theta_i\|. \tag{B.2.8}
$$

In what follows, we define $B > 0$ as a constant that depends on α, η, γ, B_r, and B_j ($j \in \{0, 1, 2\}$), whose value varies from line to line. We establish the following lemmas to upper bound the terms on the right-hand side of (B.2.8).

Lemma B.2.2 (Upper Bound of PDE – IP)**.** Let f be any continuous function such that $\|f\|_\infty \le 1$ and $\mathrm{Lip}(f) \le 1$. Under Assumptions 2.4.1 and 2.4.2, it holds for any f that

$$\sup_{t\in[0,T]} \left| \int f(\theta) \,\mathrm{d}\rho_t(\theta) - m^{-1} \cdot \sum_{i=1}^{m} f\big(\bar{\theta}_i(t)\big) \right| \le B \cdot \sqrt{\log(mT/\delta)/m}$$

with probability at least $1 - \delta$.

Proof. See §B.2.1.1 for a detailed proof. □

Lemma B.2.3 (Upper Bound of IP – CTTD)**.** Under Assumptions 2.4.1 and 2.4.2, it holds that

$$\sup_{t\in[0,T]} \big\|\bar{\theta}^{(m)}(t) - \widetilde{\theta}^{(m)}(t)\big\|_{(m)} \le B \cdot e^{BT} \cdot \sqrt{\log(m/\delta)/m}$$

with probability at least $1 - \delta$.

Proof. See §B.2.1.2 for a detailed proof. □

Lemma B.2.4 (Upper Bound of CTTD – ETD)**.** Under Assumptions 2.4.1 and 2.4.2, it holds that

$$\sup_{\substack{k\le T/\epsilon \\ (k\in\mathbb{N})}} \big\|\widetilde{\theta}^{(m)}(k\epsilon) - \breve{\theta}^{(m)}(k)\big\|_{(m)} \le B \cdot e^{BT} \cdot \epsilon.$$

Proof. See §B.2.1.3 for a detailed proof. □

Lemma B.2.5 (Upper Bound of ETD – TD)**.** Under Assumptions 2.4.1 and 2.4.2, it holds that

$$\sup_{\substack{k\leq T/\epsilon\\(k\in\mathbb{N})}} \big\|\breve{\theta}^{(m)}(k)-\theta^{(m)}(k)\big\|_{(m)} \leq B\cdot e^{BT}\cdot\sqrt{\epsilon\cdot\big(D+\log(m/\delta)\big)}$$

with probability at least $1-\delta$

Proof. See §B.2.1.4 for a detailed proof. □

We are now ready to present the proof of Proposition B.2.1.

Proof. Plugging Lemmas B.2.2-B.2.5 into (B.2.7), we have that

$$\sup_{\substack{k\leq T/\epsilon\\(k\in\mathbb{N})}} \bigg|\int f(\theta)\,\mathrm{d}\rho_{k\epsilon}(\theta)-\int f(\theta)\,\mathrm{d}\widehat{\rho}_k(\theta)\bigg|$$

$$\leq B\cdot e^{BT}\cdot\Big(\sqrt{\log(m/\delta)/m}+\sqrt{\epsilon\cdot\big(D+\log(m/\delta)\big)}\Big)$$

with probability at least $1-\delta$. Thus, we complete the proof of Proposition B.2.1. □

B.2.1. Proofs of Lemmas B.2.2-B.2.5

In this section, we present the proofs of Lemmas B.2.2-B.2.5, which are based on Mei et al. (2018, 2019); Araújo et al. (2019). We include the required technical lemmas in §B.2.3. Recall that $B>0$ is a constant that depends on α, η, γ, B_r, and B_j ($j\in\{0,1,2\}$), whose value varies from line to line.

B.2.1.1. Proof of Lemma B.2.2.

Proof. For the IP dynamics in (B.2.6), it holds that $\bar{\theta}_i(t)\sim\rho_t$ ($i\in[m]$) (Proposition 8.1.8 in Ambrosio et al. (2008)). Furthermore, since the randomness of $\bar{\theta}_i(t)$ comes from

$\theta_i(0)$ while $\theta_i(0)$ $(i \in [m])$ are independent, we have that $\bar{\theta}_i(t) \overset{\text{i.i.d.}}{\sim} \rho_t$ $(i \in [m])$. Thus, we have that

$$\mathbb{E}_{\rho_t}\Big[m^{-1} \cdot \sum_{i=1}^m f\big(\bar{\theta}_i(t)\big)\Big] = \int f(\theta)\,\mathrm{d}\rho_t(\theta).$$

Let $\theta^{1,(m)} = \{\theta_1, \ldots, \theta_i^1, \ldots, \theta_m\}$ and $\theta^{2,(m)} = \{\theta_1, \ldots, \theta_i^2, \ldots, \theta_m\}$ be two sets that only differ in the i-th element. Then, by the condition of Lemma B.2.2 that $\|f\|_\infty \leq 1$, we have that

$$\Big|m^{-1} \cdot \sum_{j=1}^m f(\theta_j^1) - m^{-1} \cdot \sum_{j=1}^m f(\theta_j^2)\Big| = m^{-1} \cdot \big|f(\theta_i^1) - f(\theta_i^2)\big| \leq 2/m.$$

Applying McDiarmid's inequality (Wainwright, 2019), we have for a fixed $t \in [0, T]$ that

$$\mathbb{P}\bigg(\Big|m^{-1} \cdot \sum_{i=1}^m f\big(\bar{\theta}_i(t)\big) - \int f(\theta)\,\mathrm{d}\rho_t(\theta)\Big| \geq p\bigg) \leq \exp(-mp^2/4). \tag{B.2.9}$$

Moreover, we have for any $s, t \in [0, T]$ that

$$\begin{aligned}
&\bigg|\Big|m^{-1} \cdot \sum_{i=1}^m f\big(\bar{\theta}_i(t)\big) - \int f(\theta)\,\mathrm{d}\rho_t(\theta)\Big| - \Big|m^{-1} \cdot \sum_{i=1}^m f\big(\bar{\theta}_i(s)\big) - \int f(\theta)\,\mathrm{d}\rho_s(\theta)\Big|\bigg| \\
&\quad \leq \Big|m^{-1} \cdot \sum_{i=1}^m f\big(\bar{\theta}_i(t)\big) - m^{-1} \cdot \sum_{i=1}^m f\big(\bar{\theta}_i(s)\big)\Big| + \Big|\int f(\theta)\,\mathrm{d}\rho_t(\theta) - \int f(\theta)\,\mathrm{d}\rho_s(\theta)\Big| \\
&\quad \leq \big\|\bar{\theta}^{(m)}(t) - \bar{\theta}^{(m)}(s)\big\|_{(m)} + \mathcal{W}_1(\rho_t, \rho_s) \\
&\quad \leq \big\|\bar{\theta}^{(m)}(t) - \bar{\theta}^{(m)}(s)\big\|_{(m)} + \mathcal{W}_2(\rho_t, \rho_s),
\end{aligned}$$

where the second inequality follows from the fact that $\mathrm{Lip}(f) \le 1$ and (B.1.9). Applying (B.2.38) and (B.2.40) of Lemma B.2.7, we have for any $s, t \in [0, T]$ that

$$\biggl| \Bigl| m^{-1} \cdot \sum_{i=1}^{m} f\bigl(\bar{\theta}_i(t)\bigr) - \int f(\theta) \,\mathrm{d}\rho_t(\theta) \Bigr| - \Bigl| m^{-1} \cdot \sum_{i=1}^{m} f\bigl(\bar{\theta}_i(s)\bigr) - \int f(\theta) \,\mathrm{d}\rho_s(\theta) \Bigr| \biggr| \le B \cdot |t - s|.$$

Applying the union bound to (B.2.9) for $t \in \iota \cdot \{0, 1, \ldots, \lfloor T/\iota \rfloor\}$, we have that

$$\mathbb{P}\biggl(\sup_{t \in [0,T]} \Bigl| m^{-1} \cdot \sum_{i=1}^{m} f\bigl(\bar{\theta}_i(t)\bigr) - \int f(\theta) \,\mathrm{d}\rho_t(\theta) \Bigr| \ge p + B \cdot \iota \biggr) \le (T/\iota + 1) \cdot \exp(-mp^2/4).$$

Setting $\iota = m^{-1/2}$ and $p = B \cdot \sqrt{\log(mT/\delta)/m}$, we have that

$$\sup_{t \in [0,T]} \Bigl| m^{-1} \cdot \sum_{i=1}^{m} f\bigl(\bar{\theta}_i(t)\bigr) - \int f(\theta) \,\mathrm{d}\rho_t(\theta) \Bigr| \le B \cdot \sqrt{\log(mT/\delta)/m}$$

with probability at least $1 - \delta$. Thus, we complete the proof of Lemma B.2.2. □

B.2.1.2. Proof of Lemma B.2.3.

Proof. Recall that g and $\widehat{g}$ are defined in (2.3.5) and (B.2.1), respectively, that is,

$$g(\theta; \rho) = -\alpha \cdot \mathbb{E}_{\widetilde{\mathcal{D}}}\Bigl[\bigl(Q(x; \rho) - r - \gamma \cdot Q(x'; \rho)\bigr) \cdot \nabla_\theta \sigma(x; \theta)\Bigr],$$

$$\widehat{g}(\theta; \theta^{(m)}) = -\alpha \cdot \mathbb{E}_{\widetilde{\mathcal{D}}}\Bigl[\bigl(\widehat{Q}(x; \theta^{(m)}) - r - \gamma \cdot \widehat{Q}(x'; \theta^{(m)})\bigr) \cdot \nabla_\theta \sigma(x; \theta)\Bigr].$$

Following from the definition of $\widetilde{\theta}_i(t)$ and $\bar{\theta}_i(t)$ in (B.2.5) and (B.2.6), respectively, we have for any $i \in [m]$ and $t \in [0,T]$ that

$$\begin{aligned}
&\big\|\bar{\theta}_i(t) - \widetilde{\theta}_i(t)\big\| \\
&\quad \le \int_0^t \bigg\| \frac{\mathrm{d}\widetilde{\theta}_i(s)}{\mathrm{d}s} - \frac{\mathrm{d}\bar{\theta}_i(s)}{\mathrm{d}s} \bigg\| \,\mathrm{d}s \\
&\quad = \eta \cdot \int_0^t \Big\| \widehat{g}\big(\widetilde{\theta}_i(s); \widetilde{\theta}^{(m)}(s)\big) - g\big(\bar{\theta}_i(s); \rho_s\big) \Big\| \,\mathrm{d}s \\
&\quad \le \eta \cdot \int_0^t \Big\| \widehat{g}\big(\widetilde{\theta}_i(s); \widetilde{\theta}^{(m)}(s)\big) - \widehat{g}\big(\bar{\theta}_i(s); \bar{\theta}^{(m)}(s)\big) \Big\| \,\mathrm{d}s \\
&\qquad + \eta \cdot \int_0^t \Big\| \widehat{g}\big(\bar{\theta}_i(s); \bar{\theta}^{(m)}(s)\big) - g\big(\bar{\theta}_i(s); \rho_s\big) \Big\| \,\mathrm{d}s
\end{aligned}$$

(B.2.10)

$$\le B \cdot \int_0^t \big\|\widetilde{\theta}^{(m)}(s) - \bar{\theta}^{(m)}(s)\big\|_{(m)} \,\mathrm{d}s + \eta \cdot \int_0^t \Big\| \widehat{g}\big(\bar{\theta}_i(s); \bar{\theta}^{(m)}(s)\big) - g\big(\bar{\theta}_i(s); \rho_s\big) \Big\| \,\mathrm{d}s,$$

where the last inequality follows from (B.2.35) of Lemma B.2.6. We now upper bound the second term on the right-hand side of (B.2.10). Following from the definition of $\widehat{Q}$, Q, and $\widehat{g}$ in (2.3.1), (2.3.2), and (B.2.1), respectively, we have for any $s \in [0,T]$ and $i \in [m]$ that

$$\Big\| \widehat{g}\big(\bar{\theta}_i(s); \bar{\theta}^{(m)}(s)\big) - g\big(\bar{\theta}_i(s); \rho_s\big) \Big\| = \alpha^2 \cdot \Big\| m^{-1} \cdot \sum_{j=1}^m Z_i^j(s) \Big\|, \tag{B.2.11}$$

where

$$\begin{aligned}
Z_i^j(s) = \mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\Big(& \sigma\big(x; \bar{\theta}_j(s)\big) - \int \sigma(x;\theta)\,\mathrm{d}\rho_s(\theta) - \gamma \cdot \sigma\big(x'; \bar{\theta}_j(s)\big) \\
&+ \gamma \cdot \int \sigma(x';\theta)\,\mathrm{d}\rho_s(\theta) \Big) \cdot \nabla_\theta \sigma\big(x; \bar{\theta}_i(s)\big) \Big].
\end{aligned}$$

Following from Assumptions 2.4.1 and 2.4.2, we have that $\|Z_i^j(s)\| \leq B$. When $i \neq j$, following from the fact that $\bar{\theta}_i(s) \overset{\text{i.i.d.}}{\sim} \rho_s$ ($i \in [m]$), it holds that $\mathbb{E}[Z_i^j(s) \,|\, \bar{\theta}_i(s)] = 0$. Following from Lemma B.2.8, we have for fixed $s \in [0, T]$ and $i \in [m]$ that

$$\begin{aligned} &\mathbb{P}\biggl(\Bigl\|m^{-1} \cdot \sum_{j \neq i} Z_i^j(s)\Bigr\| \geq B \cdot (m^{-1/2} + p)\biggr) \\ &\qquad = \mathbb{E}\biggl[\mathbb{P}\biggl(\Bigl\|m^{-1} \cdot \sum_{j \neq i} Z_i^j(s)\Bigr\| \geq B \cdot (m^{-1/2} + p) \,\Big|\, \bar{\theta}_i(s)\biggr)\biggr] \\ &\qquad \leq \exp(-mp^2). \end{aligned} \tag{B.2.12}$$

By (B.1.9), we have that

$$\sup_{x \in \mathcal{X}} \Bigl| \int \sigma(x; \theta) \,\mathrm{d}\rho_s(\theta) - \int \sigma(x; \theta) \,\mathrm{d}\rho_t(\theta) \Bigr| \leq B \cdot \mathcal{W}_1(\rho_s, \rho_t) \leq B \cdot \mathcal{W}_2(\rho_s, \rho_t) \leq B \cdot |s - t|,$$

where the last inequality follows from (B.2.40) of Lemma B.2.7. Thus, following from Assumptions 2.4.1 and 2.4.2, Lemma B.2.7, and the fact that $\mathrm{Lip}(fg) \leq \|f\|_\infty \cdot \mathrm{Lip}(g) + \|g\|_\infty \cdot \mathrm{Lip}(f)$ for any functions f and g, we have for any $s, t \in [0, T]$ that

$$\biggl| \Bigl\|m^{-1} \cdot \sum_{j \neq i} Z_i^j(s)\Bigr\| - \Bigl\|m^{-1} \cdot \sum_{j \neq i} Z_i^j(t)\Bigr\| \biggr| \leq B \cdot |t - s|.$$

Applying the union bound to (B.2.12) for $i \in [m]$ and $t \in \iota \cdot \{0, 1, \ldots, \lfloor T/\iota \rfloor\}$, we have that

$$\mathbb{P}\biggl(\sup_{\substack{i \in [m], \\ s \in [0,T]}} \Bigl\|m^{-1} \cdot \sum_{j \neq i} Z_i^j(s)\Bigr\| \geq B \cdot (m^{-1/2} + p) + B\iota\biggr) \leq m \cdot (T/\iota + 1) \cdot \exp(-mp^2).$$

Setting $\iota = m^{-1/2}$ and $p = B \cdot \sqrt{\log(mT/\delta)/m}$, we have that

$$\sup_{\substack{i\in[m],\\ s\in[0,T]}} \Big\| m^{-1} \cdot \sum_{j\neq i} Z_i^j(s) \Big\| \le B \cdot \sqrt{\log(mT/\delta)/m} \tag{B.2.13}$$

with probability at least $1-\delta$. When $i = j$, it holds that $\|m^{-1} \cdot Z_i^i(s)\| \le B/m$ in (B.2.11), which follows from Assumptions 2.4.1 and 2.4.2. Thus, plugging (B.2.13) into (B.2.11), we have that

$$\begin{aligned}\sup_{\substack{i\in[m],\\ s\in[0,T]}} \Big\| \widehat{g}\big(\bar{\theta}_i(s); \bar{\theta}^{(m)}(s)\big) - g\big(\bar{\theta}_i(s); \rho_s\big) \Big\| &\le \sup_{\substack{i\in[m],\\ s\in[0,T]}} \alpha^2 \cdot \bigg(\big\| m^{-1} \cdot Z_i^i(s) \big\| + \Big\| m^{-1} \cdot \sum_{j\neq i} Z_i^j(s) \Big\| \bigg) \\ &\le B \cdot \sqrt{\log(mT/\delta)/m} \qquad \text{(B.2.14)}\end{aligned}$$

with probability at least $1-\delta$.

Conditioning on the event in (B.2.14), we obtain from (B.2.10) that

$$\big\| \widetilde{\theta}^{(m)}(t) - \bar{\theta}^{(m)}(t) \big\|_{(m)} \le B \cdot \int_0^t \big\| \widetilde{\theta}^{(m)}(s) - \bar{\theta}^{(m)}(s) \big\|_{(m)} \, \mathrm{d}s + BT \cdot \sqrt{\log(mT/\delta)/m}$$

for any $t \in [0,T]$. Following from Gronwall's Lemma (Holte, 2009), we have that

$$\begin{aligned}\big\| \widetilde{\theta}^{(m)}(t) - \bar{\theta}^{(m)}(t) \big\|_{(m)} &\le B \cdot e^{Bt} \cdot BT \cdot \sqrt{\log(mT/\delta)/m} \\ &\le B \cdot e^{BT} \cdot \sqrt{\log(m/\delta)/m}, \qquad \forall t \in [0,T]\end{aligned}$$

with probability at least $1-\delta$. Here the last inequality holds since we allow the value of B to vary from line to line. Thus, we complete the proof of Lemma B.2.3 □

B.2.1.3. Proof of Lemma B.2.4.

Proof. By the definition of $\widehat{g}$, $\breve{\theta}_i(t)$, and $\widetilde{\theta}_i(t)$ in (B.2.1), (B.2.4), and (B.2.5), respectively, it holds that

$$\begin{aligned}\big\|\widetilde{\theta}_i(k\epsilon) - \breve{\theta}_i(k)\big\| &\le \eta \cdot \int_0^{k\epsilon} \Big\|\widehat{g}\big(\widetilde{\theta}_i(s); \widetilde{\theta}^{(m)}(s)\big) - \widehat{g}\big(\breve{\theta}_i(\lfloor s/\epsilon \rfloor); \breve{\theta}^{(m)}(\lfloor s/\epsilon \rfloor)\big)\Big\| \, \mathrm{d}s \\ &\le \eta \cdot \int_0^{k\epsilon} \Big\|\widehat{g}\big(\widetilde{\theta}_i(s); \widetilde{\theta}^{(m)}(s)\big) - \widehat{g}\big(\widetilde{\theta}_i(\lfloor s/\epsilon \rfloor \cdot \epsilon); \widetilde{\theta}^{(m)}(\lfloor s/\epsilon \rfloor \cdot \epsilon)\big)\Big\| \, \mathrm{d}s \\ &\quad + \eta \cdot \sum_{\ell=0}^{k-1} \Big\|\widehat{g}\big(\widetilde{\theta}_i(\ell\epsilon); \widetilde{\theta}^{(m)}(\ell\epsilon)\big) - \widehat{g}\big(\breve{\theta}_i(\ell); \breve{\theta}^{(m)}(\ell)\big)\Big\| \\ &\le B \cdot k \cdot \epsilon^2 + B \cdot \sum_{\ell=0}^{k-1} \big\|\widetilde{\theta}^{(m)}(\ell\epsilon) - \breve{\theta}^{(m)}(\ell)\big\|_{(m)},\end{aligned}$$

where the last inequality follows from (B.2.35) of Lemma B.2.6 and (B.2.39) of Lemma B.2.7. Following from the definition of $\|\cdot\|_{(m)}$ in (B.2.8), it holds for any $k \le T/\epsilon$ ($k \in \mathbb{N}$) that

$$\big\|\widetilde{\theta}^{(m)}(k\epsilon) - \breve{\theta}^{(m)}(k)\big\|_{(m)} \le B \cdot T \cdot \epsilon + B \cdot \sum_{\ell=0}^{k-1} \big\|\widetilde{\theta}^{(m)}(\ell\epsilon) - \breve{\theta}^{(m)}(\ell)\big\|_{(m)}.$$

Following from the discrete Gronwall's lemma (Holte, 2009), we have that

$$\sup_{\substack{k \le T/\epsilon \\ (k \in \mathbb{N})}} \big\|\widetilde{\theta}^{(m)}(k\epsilon) - \breve{\theta}^{(m)}(k)\big\|_{(m)} \le B^2 \cdot T \cdot \epsilon \cdot e^{BT} \le B \cdot e^{BT} \cdot \epsilon,$$

where the last inequality holds since we allow the value of B to vary from line to line. Thus, we complete the proof of Lemma B.2.4. □

B.2.1.4. Proof of Lemma B.2.5.

Proof. Let $\mathcal{G}_k = \sigma(\theta^{(m)}(0), z_0, \ldots, z_k)$ be the σ-algebra generated by $\theta^{(m)}(0)$ and $z_\ell = (x_\ell, r_\ell, x'_\ell)$ ($\ell \le k$). Recall that $\widehat{g}$ and $\widehat{G}_k$ are defined in (B.2.1) and (B.2.2), respectively.

We have for any $i \in [m]$ and $k \in \mathbb{N}_+$ that

$$\mathbb{E}\Big[\widehat{G}_k\big(\theta_i(k); \theta^{(m)}(k)\big) \,\Big|\, \mathcal{G}_{k-1}\Big] = \widehat{g}\big(\theta_i(k); \theta^{(m)}(k)\big).$$

Recall that $\theta^{(m)}(k)$ and $\breve{\theta}^{(m)}(k)$ are the TD and ETD dynamics defined in (B.2.3) and (B.2.4), respectively. Thus, we have for any $i \in [m]$ and $k \in \mathbb{N}_+$ that

$$\begin{aligned}
\big\|\breve{\theta}_i(k) - \theta_i(k)\big\| &= \eta\epsilon \cdot \Big\|\sum_{\ell=0}^{k-1} \widehat{G}_\ell\big(\theta_i(\ell); \theta^{(m)}(\ell)\big) - \sum_{\ell=0}^{k-1} \widehat{g}\big(\breve{\theta}_i(\ell); \breve{\theta}^{(m)}(\ell)\big)\Big\| \\
&\le \eta\epsilon \cdot \Big\|\sum_{\ell=0}^{k-1} X_i(\ell)\Big\| + \eta\epsilon \cdot \sum_{\ell=0}^{k-1} \Big\|\widehat{g}\big(\breve{\theta}_i(\ell); \breve{\theta}^{(m)}(\ell)\big) - \widehat{g}\big(\theta_i(\ell); \theta^{(m)}(\ell)\big)\Big\| \\
&\le \eta\epsilon \cdot \big\|A_i(k)\big\| + B\epsilon \cdot \sum_{\ell=0}^{k-1} \big\|\breve{\theta}^{(m)}(\ell) - \theta^{(m)}(\ell)\big\|_{(m)}, \qquad \text{(B.2.15)}
\end{aligned}$$

where the last inequality follows from (B.2.35) of Lemma B.2.6, and $X_i(\ell)$ and $A_i(k)$ are defined as

$$\begin{aligned}
&X_i(0) = 0, \\
&X_i(\ell) = \widehat{G}_\ell\big(\theta_i(\ell); \theta^{(m)}(\ell)\big) - \mathbb{E}\Big[\widehat{G}_\ell\big(\theta_i(\ell); \theta^{(m)}(\ell)\big) \,\Big|\, \mathcal{G}_{\ell-1}\Big] \quad \forall \ell \ge 1, \\
&A_i(k) = \sum_{\ell=0}^{k-1} X_i(\ell).
\end{aligned}$$

Following from (B.2.32) of Lemma B.2.6, we have that $\|X_i(\ell)\| \le B$. Thus, the stochastic process $\{A_i(k)\}_{k\in\mathbb{N}_+}$ is a martingale with $\|A_i(k) - A_i(k-1)\| \le B$. Applying Lemma B.2.9, we have that

$$\mathbb{P}\Big(\max_{\substack{k \le T/\epsilon \\ (k\in\mathbb{N}_+)}} \big\|A_i(k)\big\| \ge B \cdot \sqrt{T/\epsilon} \cdot (\sqrt{D} + p)\Big) \le \exp(-p^2). \qquad \text{(B.2.16)}$$

Applying the union bound to (B.2.16) for $i \in [m]$, we have that

$$\mathbb{P}\Big(\max_{\substack{i\in[m],\\ k\le T/\epsilon\ (k\in\mathbb{N}_+)}} \big\|A_i(k)\big\| \ge B \cdot \sqrt{T/\epsilon} \cdot (\sqrt{D} + p)\Big) \le m \cdot \exp(-p^2).$$

By setting $p = \sqrt{\log(m/\delta)}$, we have that

$$\big\|A_i(k)\big\| \le B \cdot \sqrt{T/\epsilon} \cdot \big(\sqrt{D} + \sqrt{\log(m/\delta)}\big), \quad \forall i \in [m], k \le T/\epsilon\ (k \in \mathbb{N}_+) \tag{B.2.17}$$

with probability at least $1-\delta$. By (B.2.15) and (B.2.17), we have that

$$\begin{aligned}&\big\|\breve{\theta}^{(m)}(k) - \theta^{(m)}(k)\big\|_{(m)}\\ &\quad \le B \cdot \sqrt{T\epsilon} \cdot (\sqrt{D} + \sqrt{\log(m/\delta)}) + B\epsilon \cdot \sum_{\ell=0}^{k-1} \big\|\breve{\theta}^{(m)}(\ell) - \theta^{(m)}(\ell)\big\|_{(m)},\end{aligned}$$

for any $k \le T/\epsilon$ $(k \in \mathbb{N})$ with probability at least $1-\delta$. Applying the discrete Gronwall's Lemma (Holte, 2009), we have that

$$\begin{aligned}\big\|\breve{\theta}^{(m)}(k) - \theta^{(m)}(k)\big\|_{(m)} &\le B \cdot e^{BT} \cdot B \cdot \sqrt{T\epsilon} \cdot \big(\sqrt{D} + \sqrt{\log(m/\delta)}\big)\\ &\le B \cdot e^{BT} \cdot \sqrt{\epsilon \cdot \big(D + \log(m/\delta)\big)}, \quad \forall k \le T/\epsilon\ (k \in \mathbb{N})\end{aligned}$$

with probability at least $1-\delta$. Here the last inequality holds since we allow the value of B to vary from line to line. Thus, we complete the proof of Lemma B.2.5. □

B.2.2. Proof of Lemma 2.4.4

Proof. Recall that $\widehat{Q}$ and $Q(\cdot;\rho)$ are defined in (2.3.1) and (2.3.2), respectively. For notational simplicity, we denote the optimality gaps for $\theta^{(m)} = \{\theta_i\}_{i=1}^m$ and $\rho \in \mathscr{P}_2(\mathbb{R}^D)$

by

$$L(\theta^{(m)}) = \mathbb{E}_{\mathcal{D}}\Big[\big(\widehat{Q}(x;\theta^{(m)}) - Q^*(x)\big)^2\Big], \tag{B.2.18}$$

$$\bar{L}(\rho) = \mathbb{E}_{\mathcal{D}}\Big[\big(Q(x;\rho) - Q^*(x)\big)^2\Big]. \tag{B.2.19}$$

Recall that $\theta^{(m)}(k)$, $\bar{\theta}^{(m)}(k\epsilon)$, and ρ_t are the TD dynamics, the IP dynamics, and the PDE solution defined in (B.2.3), (B.2.6), and (2.3.4), respectively. It holds for any $k \in \mathbb{N}$ that

$$\Big|L\big(\theta^{(m)}(k)\big) - \bar{L}(\rho_{k\epsilon})\Big| \le \underbrace{\Big|L\big(\theta^{(m)}(k)\big) - L\big(\bar{\theta}^{(m)}(k\epsilon)\big)\Big|}_{\text{(i)}} + \underbrace{\Big|L\big(\bar{\theta}^{(m)}(k\epsilon)\big) - \bar{L}(\rho_{k\epsilon})\Big|}_{\text{(ii)}}. \tag{B.2.20}$$

In what follows, we upper bound the two terms on the right-hand side of (B.2.20).

Upper bounding term (i) of (B.2.20)**.** Following from the definition of L in (B.2.18), it holds for any $k \in \mathbb{N}$ that

$$\begin{aligned}&\Big|L\big(\theta^{(m)}(k)\big) - L\big(\bar{\theta}^{(m)}(k\epsilon)\big)\Big|\\ &\quad = \bigg|\mathbb{E}_{\mathcal{D}}\Big[\Big(\widehat{Q}\big(x;\theta^{(m)}(k)\big) + \widehat{Q}\big(x;\bar{\theta}_i(k\epsilon)\big) - 2Q^*(x)\Big)\cdot\Big(\widehat{Q}\big(x;\theta^{(m)}(k)\big) - \widehat{Q}\big(x;\bar{\theta}_i(k\epsilon)\big)\Big)\Big]\bigg|.\end{aligned} \tag{B.2.21}$$

Following from (B.2.30), (B.2.31), and (B.2.36) of Lemma B.2.6, we have for any $k \in \mathbb{N}$ that

$$\sup_{x\in\mathcal{X}}\Big|\widehat{Q}\big(x;\theta^{(m)}(k)\big) + \widehat{Q}\big(x;\bar{\theta}_i(k\epsilon)\big) - 2Q^*(x)\Big| \le B, \tag{B.2.22}$$

$$\sup_{x\in\mathcal{X}}\Big|\widehat{Q}\big(x;\theta^{(m)}(k)\big) - \widehat{Q}\big(x;\bar{\theta}_i(k\epsilon)\big)\Big| \le B\cdot\big\|\theta^{(m)}(k) - \bar{\theta}^{(m)}(k\epsilon)\big\|_{(m)}. \tag{B.2.23}$$

Thus, we have that

$$
\begin{aligned}
&\Big|L\big(\theta^{(m)}(k)\big) - L\big(\bar{\theta}^{(m)}(k\epsilon)\big)\Big| \\
&\qquad \le B \cdot \big\|\theta^{(m)}(k) - \bar{\theta}^{(m)}(k\epsilon)\big\|_{(m)} \\
&\qquad \le B \cdot e^{BT} \cdot \Big(\sqrt{\log(m/\delta)/m} + \sqrt{\epsilon \cdot \big(D + \log(m/\delta)\big)}\Big), \quad \forall k \le T/\epsilon \ (k \in \mathbb{N}) \qquad \text{(B.2.24)}
\end{aligned}
$$

with probability at least $1 - \delta$. Here the last inequality follows from Lemmas B.2.3-B.2.5.

Upper bounding term (ii) of (B.2.20)**.** Let $t = k\epsilon$. It holds for any $t \in [0, T]$ that

(B.2.25)

$$
\Big|L\big(\bar{\theta}^{(m)}(t)\big) - \bar{L}(\rho_t)\Big| \le \bigg|L\big(\bar{\theta}^{(m)}(t)\big) - \mathbb{E}_{\rho_t}\Big[L\big(\bar{\theta}^{(m)}(t)\big)\Big]\bigg| + \bigg|\mathbb{E}_{\rho_t}\Big[L\big(\bar{\theta}^{(m)}(t)\big)\Big] - \bar{L}(\rho_t)\bigg|,
$$

where the expectation is with respect to $\bar{\theta}_i(t) \overset{\text{i.i.d.}}{\sim} \rho_t$ $(i \in [m])$. For the second term on the right-hand side of (B.2.25), following from the fact that $\mathbb{E}_{\rho_t}[\widehat{Q}(x; \bar{\theta}^{(m)}(t))] = Q(x; \rho_t)$ for any $x \in \mathcal{X}$, we have that

$$
\begin{aligned}
\bigg|\mathbb{E}_{\rho_t}\Big[L\big(\bar{\theta}^{(m)}(t)\big)\Big] - \bar{L}(\rho_t)\bigg| &= \bigg|\int \mathbb{E}_{\rho_t}\Big[\widehat{Q}\big(x; \bar{\theta}^{(m)}(t)\big)^2 - Q(x; \rho_t)^2\Big]\, \mathrm{d}\mathcal{D}(x)\bigg| \\
&= \bigg|\int \mathrm{Var}_{\rho_t}\Big[\widehat{Q}\big(x; \bar{\theta}^{(m)}(t)\big)\Big]\, \mathrm{d}\mathcal{D}(x)\bigg| \\
&\le B/m, \qquad \text{(B.2.26)}
\end{aligned}
$$

where the inequality follows from the fact that $\|\sigma\| \le B$ in Assumption 2.4.2 and the independence of $\bar{\theta}_i(t)$ $(i \in [m])$. Let $\theta^{1,(m)} = \{\theta_1, \ldots, \theta_i^1, \ldots, \theta_m\}$ and $\theta^{2,(m)} =$

$\{\theta_1, \ldots, \theta_i^2, \ldots, \theta_m\}$ be two sets that only differ in the i-th element. It holds that

$$\big|L(\theta^{1,(m)}) - L(\theta^{2,(m)})\big| \le B \cdot m^{-1} \cdot \mathbb{E}_{\mathcal{D}}\Big[\big|\sigma(x;\theta_i^1) - \sigma(x;\theta_i^2)\big|\Big] \le B/m,$$

where the first inequality follows from (B.2.21) and (B.2.22) and the second inequality follows from Assumption 2.4.2. Applying McDiarmid's inequality (Wainwright, 2019), we have for a fixed $t \in [0,T]$ that

$$\mathbb{P}\left(\Big|L\big(\bar\theta^{(m)}(t)\big) - \mathbb{E}_{\rho_t}\Big[L\big(\bar\theta^{(m)}(t)\big)\Big]\Big| \ge p\right) \le \exp(-mp^2/B). \tag{B.2.27}$$

It holds for any $s, t \in [0,T]$ that

$$\begin{aligned}&\bigg|\Big|L\big(\bar\theta^{(m)}(t)\big) - \mathbb{E}_{\rho_t}\Big[L\big(\bar\theta^{(m)}(t)\big)\Big]\Big| - \Big|L\big(\bar\theta^{(m)}(s)\big) - \mathbb{E}_{\rho_t}\Big[L\big(\bar\theta^{(m)}(s)\big)\Big]\Big|\bigg|\\ &\quad\le B \cdot \big\|\bar\theta^{(m)}(t) - \bar\theta^{(m)}(s)\big\|_{(m)} \le B \cdot |t-s|,\end{aligned}$$

where the first inequality follows from (B.2.21), (B.2.22), and (B.2.23) and the second inequality follows from (B.2.38) of Lemma B.2.7. Applying the union bound to (B.2.27) for $t \in \iota \cdot \{0, 1, \ldots, \lfloor T/\iota \rfloor\}$, we have that

$$\mathbb{P}\left(\sup_{t\in[0,T]}\Big|L\big(\bar\theta^{(m)}(t)\big) - \mathbb{E}_{\rho_t}\Big[L\big(\bar\theta^{(m)}(t)\big)\Big]\Big| \ge p + B\iota\right) \le (T/\iota + 1) \cdot \exp(-mp^2/B),$$

Setting $\iota = m^{-1/2}$ and $p = B \cdot \sqrt{\log(mT\delta)/m}$, we have that

$$\sup_{t\in[0,T]}\Big|L\big(\bar\theta^{(m)}(t)\big) - \mathbb{E}_{\rho_t}\Big[L\big(\bar\theta^{(m)}(t)\big)\Big]\Big| \le B \cdot \sqrt{\log(mT\delta)/m} \tag{B.2.28}$$

with probability at least $1-\delta$. Plugging (B.2.26) and (B.2.28) into (B.2.25), noting that $t = k\epsilon$, we have that

$$\Big|L\big(\bar{\theta}^{(m)}(k\epsilon)\big) - \bar{L}(\rho_{k\epsilon})\Big| \le B \cdot \sqrt{\log(mT\delta)/m}, \quad \forall k \le T/\epsilon \ (k \in \mathbb{N}) \tag{B.2.29}$$

with probability at least $1-\delta$.

Plugging (B.2.24) and (B.2.29) into (B.2.20), we have that

$$\Big|L\big(\theta^{(m)}(k)\big) - \bar{L}(\rho_{k\epsilon})\Big| \le B \cdot e^{BT} \cdot \Big(\sqrt{\log(m/\delta)/m} + \sqrt{\epsilon \cdot \big(D + \log(m/\delta)\big)}\Big), \quad \forall k \le T/\epsilon \ (k \in \mathbb{N})$$

with probability at least $1-\delta$. Thus, we complete the proof of Lemma 2.4.4. □

B.2.3. Technical Lemmas for §B.2

In what follows, we present the technical lemmas used in §B.2. Recall that $\widehat{Q}$, $\widehat{g}$, and $\widehat{G}_k$ are defined in (2.3.1), (B.2.1), and (B.2.2), respectively. Let $B > 0$ be a constant depending on α, η, γ, B_r, and B_j ($j \in \{0, 1, 2\}$), whose value varies from line to line.

Lemma B.2.6. Under Assumptions 2.4.1 and 2.4.2, it holds for any $\theta^{(m)} = \{\theta_i\}_{i=1}^m$ and $\underline{\theta}^{(m)} = \{\underline{\theta}_i\}_{i=1}^m$ that

$$\sup_{x\in\mathcal{X}} \big|\widehat{Q}(x;\theta^{(m)})\big| \le B, \tag{B.2.30}$$

$$\sup_{x\in\mathcal{X}} \big|\widehat{Q}(x;\theta^{(m)}) - \widehat{Q}(x;\underline{\theta}^{(m)})\big| \le B\cdot \|\theta^{(m)} - \underline{\theta}^{(m)}\|_{(m)}, \tag{B.2.31}$$

$$\big\|\widehat{G}_k(\theta_i;\theta^{(m)})\big\| \le B, \tag{B.2.32}$$

$$\big\|\widehat{G}_k(\theta_i;\theta^{(m)}) - \widehat{G}_k(\underline{\theta}_i;\underline{\theta}^{(m)})\big\| \le B\cdot \|\theta^{(m)} - \underline{\theta}^{(m)}\|_{(m)}, \quad \forall k\in\mathbb{N}, \tag{B.2.33}$$

$$\big\|\widehat{g}(\theta_i;\theta^{(m)})\big\| \le B, \tag{B.2.34}$$

$$\big\|\widehat{g}(\theta_i;\theta^{(m)}) - \widehat{g}(\underline{\theta}_i;\underline{\theta}^{(m)})\big\| \le B\cdot \|\theta^{(m)} - \underline{\theta}^{(m)}\|_{(m)}. \tag{B.2.35}$$

Meanwhile, for any $Q\in\mathcal{F}$, it holds that

$$\sup_{x\in\mathcal{X}} \big\|Q(x)\big\| \le B. \tag{B.2.36}$$

For any $\rho\in\mathscr{P}_2(\mathbb{R}^D)$, it holds that

$$\big\|g(\theta;\rho)\big\| \le B. \tag{B.2.37}$$

Proof. For (B.2.30) and (B.2.31) of Lemma B.2.6, following from Assumptions 2.4.1 and 2.4.2 and the definition of $\widehat{Q}$ in (2.3.1), we have for any $x\in\mathcal{X}$, $\theta^{(m)}$, and $\underline{\theta}^{(m)}$ that

$$\begin{aligned}
&\big|\widehat{Q}(x;\theta^{(m)})\big| \le \alpha\cdot m^{-1}\sum_{i=1}^m \big|\sigma(x;\theta_i)\big| \le B,\\
&\big|\widehat{Q}(x;\theta^{(m)}) - \widehat{Q}(x;\underline{\theta}^{(m)})\big| \le \alpha\cdot m^{-1}\sum_{i=1}^m \big|\sigma(x;\theta_i) - \sigma(x;\underline{\theta}_i)\big| \le B\cdot\|\theta^{(m)} - \underline{\theta}^{(m)}\|_{(m)}.
\end{aligned}$$

For (B.2.32) and (B.2.33) of Lemma B.2.6, following from the definition of $\widehat{G}_k$ in (B.2.2), we have for any $\theta^{(m)}$ and $\underline{\theta}^{(m)}$ that

$$\begin{aligned}
&\big\|\widehat{G}_k(\theta_i;\theta^{(m)})\big\| = \alpha\cdot\big|\widehat{Q}(x_k;\theta^{(m)}) - r_k - \gamma\cdot\widehat{Q}(x_k';\theta^{(m)})\big|\cdot\big\|\nabla_\theta\sigma(x_k;\theta_i)\big\| \le B,\\
&\big\|\widehat{G}_k(\theta_i;\theta^{(m)}) - \widehat{G}_k(\underline{\theta}_i;\underline{\theta}^{(m)})\big\|\\
&\quad\le \alpha\cdot\sup_{\theta^{(m)}}\big|\widehat{Q}(x_k;\theta^{(m)}) - r_k - \gamma\cdot\widehat{Q}(x_k';\theta^{(m)})\big|\cdot\big\|\nabla_\theta\sigma(x_k;\theta_i) - \nabla_\theta\sigma(x_k;\underline{\theta}_i)\big\|\\
&\qquad + \alpha\big|\widehat{Q}(x_k;\theta^{(m)}) - \gamma\cdot\widehat{Q}(x_k';\theta^{(m)}) - \widehat{Q}(x_k;\underline{\theta}^{(m)}) + \gamma\cdot\widehat{Q}(x_k';\theta^{(m)})\big|\sup_{\theta_i\in\mathbb{R}^D}\big\|\nabla_\theta\sigma(x_k;\theta_i)\big\|\\
&\quad\le B\cdot\|\theta^{(m)} - \underline{\theta}^{(m)}\|_{(m)}.
\end{aligned}$$

The inequalities in (B.2.34) and (B.2.35) of Lemma B.2.6 for $\widehat{g}$ follow from the fact that

$$\widehat{g}(\theta_i;\theta^{(m)}) = \mathbb{E}_{(x_k,r_k,x_k')\sim\widetilde{\mathcal{D}}}\big[G_k(\theta_i;\theta^{(m)})\big].$$

The inequalities in (B.2.36) and (B.2.37) follow from the definition of $\mathcal{F}$ and g in (2.4.3) and (2.3.5), respectively. Thus, we complete the proof of Lemma B.2.6. □

Recall that ρ_t is the PDE solution in (2.3.4) and $\widetilde{\theta}^{(m)}(t)$ and $\bar{\theta}^{(m)}(t)$ are the CTTD and IP dynamics defined in (B.2.5) and (B.2.6), respectively.

Lemma B.2.7. Under Assumptions 2.4.1 and 2.4.2, it holds for any $s,t\in[0,T]$ that

(B.2.38) $$\big\|\bar{\theta}^{(m)}(t) - \bar{\theta}^{(m)}(s)\big\|_{(m)} \le B\cdot|t-s|,$$

(B.2.39) $$\big\|\widetilde{\theta}^{(m)}(t) - \widetilde{\theta}^{(m)}(s)\big\|_{(m)} \le B\cdot|t-s|,$$

(B.2.40) $$\mathcal{W}_2(\rho_t,\rho_s) \le B\cdot|t-s|.$$

Proof. For (B.2.38) of Lemma B.2.7, by the definition of $\bar{\theta}_i(t)$ in (B.2.6) and (B.2.37) of Lemma B.2.6, we have for any $s, t \in [0, T]$ and $i \in [m]$ that

$$\big\|\bar{\theta}_i(t) - \bar{\theta}_i(s)\big\| = \eta \cdot \int_s^t \Big\|g\big(\bar{\theta}_i(\tau); \rho_\tau\big)\Big\| \,\mathrm{d}\tau \leq B \cdot |t - s|.$$

Similarly, for (B.2.39) of Lemma B.2.7, by the definition of $\widetilde{\theta}_i(t)$ in (B.2.5) and (B.2.34) of Lemma B.2.6, we have for any $i \in [m]$ and $s, t \in [0, T]$ that $\|\widetilde{\theta}_i(t) - \widetilde{\theta}_i(s)\| \leq B \cdot |t - s|$.

For (B.2.40) of Lemma B.2.7, following from the fact that $\bar{\theta}_i(t) \overset{\text{i.i.d.}}{\sim} \rho_t$ $(i \in [m])$ and the definition of $\mathcal{W}_2$ in (2.2.4), it holds for any $s, t \in [0, T]$ that

$$\mathcal{W}_2(\rho_t, \rho_s) \leq \mathbb{E}\Big[\big\|\bar{\theta}_i(t) - \bar{\theta}_i(s)\big\|^2\Big]^{1/2} \leq B \cdot |t - s|.$$

Thus, we complete the proof of Lemma B.2.7. □

Lemma B.2.8 (Lemma 30 in Mei et al. (2019))**.** Let $\{X_i\}_{i=1}^m$ be i.i.d. random variables with $\|X_i\| \leq \xi$ and $\mathbb{E}[X_i] = 0$. Then, it holds for any $p > 0$ that

$$\mathbb{P}\bigg(\Big\|m^{-1} \cdot \sum_{i=1}^m X_i\Big\| \geq C\xi \cdot (m^{-1/2} + p)\bigg) \leq \exp(-mp^2),$$

where $C > 0$ is an absolute constant.

Lemma B.2.9 (Lemma 31 in Mei et al. (2019) and Lemma A.3 in Araújo et al. (2019))**.** Let $X_k \in \mathbb{R}^D$ $(k \in \mathbb{N})$ be a martingale with respect to the filtration $\mathcal{G}_k$ $(k \geq 0)$ with $X_0 = 0$. We assume for $\xi > 0$ and any $\lambda \in \mathbb{R}^D$ that

$$\mathbb{E}\Big[\exp\big(\langle \lambda, X_k - X_{k-1}\rangle\big) \,\Big|\, \mathcal{G}_{k-1}\Big] \leq \exp\big(\xi^2 \cdot \|\lambda\|^2/2\big).$$

Then, it holds that

$$\mathbb{P}\Big(\max_{\substack{k\leq n\\(k\in\mathbb{N})}}\|X_k\| \geq C\xi\cdot\sqrt{n}\cdot(\sqrt{D}+p)\Big) \leq \exp(-p^2),$$

where $C > 0$ is an absolute constant.

B.3. Auxiliary Lemmas

We use the definition of absolutely continuous curves in $\mathscr{P}_2(\mathbb{R}^D)$ in Ambrosio et al. (2008).

Definition B.3.1 (Absolutely Continuous Curve)**.** Let $\beta : [a, b] \to \mathscr{P}_2(\mathbb{R}^D)$ be a curve. Then, we say β is an absolutely continuous curve if there exists a square-integrable function $f : [a, b] \to \mathbb{R}$ such that

$$\mathcal{W}_2(\beta_s, \beta_t) \leq \int_s^t f(\tau)\,\mathrm{d}\tau$$

for any $a \leq s < t \leq b$.

Then, we have the following first variation formula.

Lemma B.3.2 (First Variation Formula, Theorem 8.4.7 in Ambrosio et al. (2008))**.** Given $\nu \in \mathscr{P}_2(\mathbb{R}^D)$ and an absolutely continuous curve $\mu : [0, T] \to \mathscr{P}_2(\mathbb{R}^D)$, let $\beta : [0, 1] \to \mathscr{P}_2(\mathbb{R}^D)$ be the geodesic connecting μ_t and ν. It holds that

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\mu_t, \nu)^2}{2} = -\langle \mu'_t, \beta'_0\rangle_{\mu_t},$$

where $\mu'_t = \partial_t\mu_t$, $\beta'_0 = \partial_t\beta_t\,|_{\,t=0}$, and the inner product is defined in (2.2.5).

Lemma B.3.3 (Talagrand's Inequality, Corollary 2.1 in Otto and Villani (2000))**.** Let ν be $N(0, \kappa \cdot I_D)$. It holds for any $\mu \in \mathscr{P}_2(\mathbb{R}^D)$ that

$$\mathcal{W}_2(\mu, \nu)^2 \leq 2D_{\mathrm{KL}}(\mu \,\|\, \nu)/\kappa.$$

Lemma B.3.4 (Eulerian Representation of Geodesics, Proposition 5.38 in Villani (2003))**.** Let $\beta : [0, 1] \to \mathscr{P}_2(\mathbb{R}^D)$ be a geodesic and v be the corresponding vector field such that $\partial_t \beta_t = -\operatorname{div}(\beta_t \cdot v_t)$. It holds that

$$\partial_t(\beta_t \cdot v_t) = -\operatorname{div}(\beta_t \cdot v_t \otimes v_t).$$

APPENDIX C

An Analysis of Attention via the Lens of Exchangeability and Latent Variable Models

C.1. Conditional Mean Embedding

We introduce the conditional mean embedding (Song et al., 2009), which embeds a conditional distribution to an element in an RKHS. Let $\mathcal{H}_x$ and $\mathcal{H}_y$ be the two RKHSs over the spaces $\mathfrak{X}$ and $\mathfrak{Y}$ with the kernels $\mathfrak{K}$ and $\mathfrak{L}$, respectively. We denote by $\phi : \mathfrak{X} \to \ell_2$ and $\varphi : \mathfrak{Y} \to \ell_2$ the feature mappings associated with $\mathcal{H}_x$ and $\mathcal{H}_y$, respectively. In other words, it holds for any $x, x' \in \mathfrak{X}$ and $y, y' \in \mathfrak{Y}$ that

$$\phi(x)^\top \phi(x') = \mathfrak{K}(x, x'), \qquad \varphi(y)^\top \varphi(y) = \mathfrak{L}(y, y'). \tag{C.1.1}$$

Let $\mathbb{P}_{\mathcal{X},\mathcal{Y}}$ be the joint distribution of the two random variables $\mathcal{X}$ and $\mathcal{Y}$ taking values in $\mathfrak{X}$ and $\mathfrak{Y}$, respectively. The conditional mean embedding $\texttt{CME}(x, \mathbb{P}_{\mathcal{X},\mathcal{Y}}) \in \mathcal{H}_y$ of the conditional distribution $\mathbb{P}_{\mathcal{Y}\,|\,\mathcal{X}}$ is defined as

$$\texttt{CME}(x, \mathbb{P}_{\mathcal{X},\mathcal{Y}}) = \mathbb{E}\big[\mathfrak{L}(\mathcal{Y}, \cdot) \,\big|\, \mathcal{X} = x\big].$$

By the reproducing property, it holds that

$$\mathbb{E}\big[g(\mathcal{Y}) \,\big|\, \mathcal{X} = x\big] = \big\langle g, \texttt{CME}(x, \mathbb{P}_{\mathcal{X},\mathcal{Y}})\big\rangle_{\mathcal{H}_y}, \quad \forall g \in \mathcal{H}_y, x \in \mathfrak{X}.$$

Correspondingly, the conditional mean embedding operator $\mathcal{C}_{\mathcal{Y}\,|\,\mathcal{X}} : \mathcal{H}_x \to \mathcal{H}_y$ is a linear operator such that

$$\mathcal{C}_{\mathcal{Y}\,|\,\mathcal{X}} \mathfrak{K}(x, \cdot) = \texttt{CME}(x, \mathbb{P}_{\mathcal{X},\mathcal{Y}}),$$

for any $x \in \mathfrak{X}$. We define the (uncentered) covariance operator $\mathcal{C}_{\mathcal{X}\mathcal{X}} : \mathcal{H}_x \to \mathcal{H}_x$ and the (uncentered) cross-covariance operator $\mathcal{C}_{\mathcal{Y}\mathcal{X}} : \mathcal{H}_x \to \mathcal{H}_y$ as follows,

$$\mathcal{C}_{\mathcal{X}\mathcal{X}} = \mathbb{E}\big[\mathfrak{K}(\mathcal{X}, \cdot) \otimes \mathfrak{K}(\mathcal{X}, \cdot)\big], \qquad \mathcal{C}_{\mathcal{Y}\mathcal{X}} = \mathbb{E}\big[\mathfrak{L}(\mathcal{Y}, \cdot) \otimes \mathfrak{K}(\mathcal{X}, \cdot)\big].$$

Here $\otimes$ is the tensor product. As shown in Song et al. (2009), it holds that $\mathcal{C}_{\mathcal{Y}\,|\,\mathcal{X}} = \mathcal{C}_{\mathcal{Y}\mathcal{X}}\mathcal{C}_{\mathcal{X}\mathcal{X}}^{-1}$. Thus, we have that

$$\texttt{CME}(x, \mathbb{P}_{\mathcal{X},\mathcal{Y}}) = \mathcal{C}_{\mathcal{Y}\mathcal{X}}\mathcal{C}_{\mathcal{X}\mathcal{X}}^{-1}\mathfrak{K}(x, \cdot). \tag{C.1.2}$$

To derive the empirical estimation of $\mathcal{C}_{\mathcal{Y}\,|\,\mathcal{X}}$, we consider the following regularized least-squares problem,

$$\min_{\mathcal{C}:\mathcal{H}_x \to \mathcal{H}_y} \widehat{\mathcal{E}}(\mathcal{C}) = \sum_{\ell=1}^{L} \big\| \mathfrak{L}(y^\ell, \cdot) - \mathcal{C}\mathfrak{K}(x^\ell, \cdot) \big\|_{\mathcal{H}_y}^2 + \lambda \cdot \|\mathcal{C}\|_{\mathrm{HS}}^2, \tag{C.1.3}$$

where $\{(x^\ell, y^\ell)\}_{\ell \in [L]}$ are independently and identically sampled from $\mathbb{P}_{\mathcal{X},\mathcal{Y}}$, $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert-Schmidt norm, and $\lambda > 0$ is the regularization parameter. Recall from (C.1.1) that ϕ and φ are the feature mappings associated with the RKHSs $\mathcal{H}_x$ and $\mathcal{H}_y$. To ease the presentation, we view the space ℓ_2 as an (infinite-dimensional) vector space and consider the feature mappings $\phi : \mathfrak{X} \to \mathbb{R}^{d_\phi}$ and $\varphi : \mathfrak{Y} \to \mathbb{R}^{d_\varphi}$, where d_ϕ and d_φ can be infinity. We write $\phi(X) = (\phi(x^1), \ldots, \phi(x^L))^\top \in \mathbb{R}^{L \times d_\phi}$ and $\varphi(Y) = (\phi(y^1), \ldots, \phi(y^L))^\top \in \mathbb{R}^{L \times d_\varphi}$. Also, we define the the (uncentered) empirical covariance operator $\widehat{\mathcal{C}}_{\mathcal{X}\mathcal{X}}$ and (uncentered) empirical cross-covariance operator $\widehat{\mathcal{C}}_{\mathcal{Y}\mathcal{X}}$ as follows,

$$\widehat{\mathcal{C}}_{\mathcal{X}\mathcal{X}} = L^{-1}\sum_{\ell=1}^{L}\phi(x^\ell)\phi(x^\ell)^\top = L^{-1}\phi(X)^\top\phi(X) \in \mathbb{R}^{d_\phi \times d_\phi}$$

$$\widehat{\mathcal{C}}_{\mathcal{Y}\mathcal{X}} = L^{-1}\sum_{\ell=1}^{L}\varphi(y^\ell)\varphi(x^\ell)^\top = L^{-1}\varphi(Y)\phi(X)^\top \in \mathbb{R}^{d_\varphi \times d_\phi}. \tag{C.1.4}$$

Then, the solution to (C.1.3) is

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

We denote by $\widehat{\texttt{CME}}_\lambda(x, \mathbb{P}_{\mathcal{X},\mathcal{Y}}) = \widehat{\mathcal{C}}_{\mathcal{Y}\,|\,\mathcal{X}}\phi(x) \in \mathbb{R}^{d_\varphi}$ the empirical conditional mean embedding. Note that

$$\phi(X)\big(\phi(X)^\top\phi(X) + \lambda\mathcal{I}\big)^{-1} = \big(\phi(X)\phi(X)^\top + \lambda I\big)^{-1}\phi(X).$$

Thus, it holds that

$$\begin{aligned}
\widehat{\texttt{CME}}_\lambda(x, \mathbb{P}_{\mathcal{X},\mathcal{Y}}) &= \widehat{\mathcal{C}}^{\lambda}_{\mathcal{Y}\,|\,\mathcal{X}}\phi(x) \\
&= \widehat{\mathcal{C}}_{\mathcal{Y}\mathcal{X}}(\widehat{\mathcal{C}}_{\mathcal{X}\mathcal{X}} + L^{-1}\lambda\mathcal{I})^{-1}\mathfrak{K}(x,\cdot) \\
&= \varphi(Y)^\top\phi(X)\big(\phi(X)^\top\phi(X) + \lambda\mathcal{I}\big)^{-1}\phi(x) \\
&= \varphi(Y)^\top\big(\phi(X)\phi(X)^\top + \lambda I\big)^{-1}\phi(X)\phi(x) \\
&= \varphi(Y)^\top(\mathfrak{K}(X,X) + \lambda I)^{-1}\mathfrak{K}(X,x). \qquad \text{(C.1.5)}
\end{aligned}$$

Here $\mathfrak{K}(X,X) = \phi(X)\phi(X)^\top = (\mathfrak{K}(x^i, x^j))_{i,j\in[L]} \in \mathbb{R}^{L\times L}$ is the Gram matrix and $\mathfrak{K}(X,x) = \phi(X)\phi(x) = (\mathfrak{K}(x^1,x),\dots,\mathfrak{K}(x^L,x)) \in \mathbb{R}^L$.

C.2. Attention Recovers Latent Posterior

C.2.1. Gaussian Process Regression

Gaussian Process Regression. We say that f follows a Gaussian process $\mathtt{GP}(\mu, \mathfrak{K})$ on $\mathbb{R}^d$ if for any $x^1, \ldots, x^L$, $(f(x^1), \ldots, f(x^L))$ follows a Gaussian distribution with mean $(\mu(x^1), \ldots, \mu(x^L))$ and covariance $(\mathfrak{K}(x^i, x^j))_{i,j\in[L]}$. Here $\mu(x) = \mathbb{E}[f(x)]$ is the mean function and $\mathfrak{K}(x, x') = \mathbb{E}[(f(x) - \mu(x))(f(x') - \mu(x'))]$ is the covariance (or kernel) function, where f is random. We take $\mathtt{GP}(0, \mathfrak{K})$ as the prior of f. Given a dataset $\mathcal{D} = \{(x^\ell, y^\ell)\}_{\ell\in[L]}$ from the regression model $y^\ell = f(x^\ell) + \epsilon^\ell$ with $\epsilon^\ell \sim N(0, \lambda I)$, the posterior of f is a Gaussian process with mean $\mu_{\mathcal{D}}(x)$ and covariance $\mathfrak{K}_{\mathcal{D}}(x, x')$ (Schulz et al., 2018) as follows,

$$\mu_{\mathcal{D}}(x) = \mathfrak{K}(x, X)\big(\mathfrak{K}(X, X) + \lambda I\big)^{-1} Y,$$

$$\mathfrak{K}_{\mathcal{D}}(x, x') = \mathfrak{K}(x, x') - \mathfrak{K}(x, X)\big(\mathfrak{K}(X, X) + \lambda I\big)^{-1}\mathfrak{K}(X, x').$$

Here $\mathfrak{K}(x, X) = (\mathfrak{K}(x, x^\ell))_{\ell\in[L]}^\top \in \mathbb{R}^{1\times L}$, $\mathfrak{K}(X, X) = (\mathfrak{K}(x^i, x^j))_{i,j\in[L]} \in \mathbb{R}^{L\times L}$, $\mathfrak{K}(X, x') = (\mathfrak{K}(x^\ell, x'))_{\ell\in[L]} \in \mathbb{R}^L$, and $Y = (y^1, \ldots, y^L) \in \mathbb{R}^L$

Rigorous Characterization of Latent Variable Model. We provide a rigorous characterization of the advanced infinite-dimensional example of the latent variable model in §3.4.1.We consider the following model,

$$r^\ell = f(c^\ell) + \epsilon^\ell, \qquad r^{\mathtt{msk}} = f(c^{\mathtt{msk}}) + \epsilon. \tag{C.2.1}$$

Here $f = (f_1, \dots, f_d)$ with $f_i \sim \mathtt{GP}(0, \mathfrak{K}(\cdot, \cdot))$ for any $i \in [d]$ and ϵ^ℓ and ϵ are independent Gaussian noises drawn from $N(0, \lambda I)$. Then, following the Gaussian process regression, we recover (3.4.3) as the mean of the posterior of the Gaussian process.

C.2.2. Implication of Convergence with $L \to \infty$

Necessity of Multiple Heads. Based on the definition of the attention mechanism $\mathtt{attn}$ in (3.2.1), we define the multihead attention as

$$\mathtt{mha}(q, X; W) = \sum_{i=1}^{h} \mathtt{head}_i \in \mathbb{R}^d. \tag{C.2.2}$$

Here $h \in \mathbb{N}_+$ is the head number, $W = \{(W_i^{\mathrm{q}}, W_i^{\mathrm{k}}, W_i^{\mathrm{v}})\}_{i=1}^h$ with $W_i^{\mathrm{q}} \in \mathbb{R}^{d \times d_{\mathrm{p}}}$, $W_i^{\mathrm{k}} \in \mathbb{R}^{d \times d_{\mathrm{p}}}$, and $W_i^{\mathrm{v}} \in \mathbb{R}^{d \times d}$ is the learnable parameter, and

$$\mathtt{head}_i = \mathtt{attn}(q, K_i, V_i) \in \mathbb{R}^d, \qquad \text{where} \quad K_i = XW_i^{\mathrm{k}} \in \mathbb{R}^{L \times d_{\mathrm{p}}}, \quad V_i = XW_i^{\mathrm{v}} \in \mathbb{R}^{L \times d}.$$

In the multihead attention, we set $d = d_{\mathrm{p}} \cdot h$, where h is the head number and d_{p} is the dimension of the key and the query. We remark that the multihead attention defined in (C.2.2) is written in the summation form, which is equivalent to the concatenation form (Vaswani et al., 2017). To see this, we consider the concatenation form of the multihead attention,

$$\widetilde{\mathtt{mha}}(q, X; \widetilde{W}) = \big((W_1^{\mathrm{o}})^\top, \dots, (W_h^{\mathrm{o}})^\top\big) \begin{pmatrix} \widetilde{\mathtt{head}}_1 \\ \dots \\ \widetilde{\mathtt{head}}_h \end{pmatrix} = \sum_{i=1}^{h} (W_i^{\mathrm{o}})^\top \widetilde{\mathtt{head}}_i, \tag{C.2.3}$$

where $W^{\mathrm{o}} \in \mathbb{R}^{d_{\mathrm{p}} \times d}$ is a learnable parameter with the i-th block W_i^{o} and the i-th head $\widetilde{\texttt{head}}_i$ is obtained via

$$\widetilde{\texttt{head}}_i = \texttt{attn}(q, K_i, \widetilde{V}_i) \in \mathbb{R}^{d_{\mathrm{p}}}, \qquad \text{where} \quad K_i = X W_i^{\mathrm{k}} \in \mathbb{R}^{L \times d_{\mathrm{p}}}, \quad V_i = X \widetilde{W}_i^{\mathrm{v}} \in \mathbb{R}^{L \times d_{\mathrm{p}}}.$$

Here $\widetilde{W}_i^{\mathrm{v}} \in \mathbb{R}^{d \times d_{\mathrm{p}}}$. We see that the (C.2.2) and (C.2.3) are equivalent when $\texttt{head}_i = (W_i^{\mathrm{o}})^\top \widetilde{\texttt{head}}_i$, which holds when $W_i^{\mathrm{v}} = \widetilde{W}_i^{\mathrm{v}} W_i^{\mathrm{o}}$.

We use $\mathbb{E}[\mathcal{V} \,|\, \mathcal{K} = q]$ to demonstrate the necessity of using multiple heads in the multihead attention. Note that the key and value are obtained by the following mappings,

$$k^\ell = (W^{\mathrm{k}})^\top x^\ell, \qquad v^\ell = (W^{\mathrm{v}})^\top x^\ell,$$

where $x^\ell \in \mathbb{R}^d$ is the input token and $W^{\mathrm{k}} \in \mathbb{R}^{d \times d_{\mathrm{p}}}$, $W^{\mathrm{v}} \in \mathbb{R}^{d \times d}$ are the learnable parameters. We consider a single-head attention, where $h = 1$, $d_{\mathrm{p}} = d$, and $W^{\mathrm{k}} \in \mathbb{R}^{d \times d}$ is invertible. We denote by $\mathcal{K}$, $\mathcal{V}$, and $\mathcal{X}$ the random variable with the same distribution as k^ℓ, v^ℓ, and x^ℓ, respectively. By Propositions 3.4.1 and 3.4.2, we have

$$\texttt{attn}(q, K, V) \approx \mathbb{E}[\mathcal{V} \,|\, \mathcal{K} = q] = \mathbb{E}\big[(W^{\mathrm{v}})^\top \mathcal{X} \,\big|\, (W^{\mathrm{k}})^\top \mathcal{X} = q\big] = \big((W^{\mathrm{k}})^{-1} W^{\mathrm{v}}\big)^\top q,$$

which is a linear mapping and fails to capture the nonlinear interaction query q and the input sequence X. In other words, the single-head attention becomes a linear mapping in the limit with $L \to \infty$. In contrast, when $h > 1$, we have $d_{\mathrm{p}} = d/h < d$, which implies that the matrix $W^{\mathrm{k}} \in \mathbb{R}^{d \times d}$ is not invertible. Thus, using multiple heads avoid the degenerating issue.

Connection to Instrumental Variable. We draw a connection from the attention mechanism to the instrumental variable model. Instrumental variable regression estimates the causal relationship between the input $\mathcal{X}$ and the output $\mathcal{Y}$. Specifically, when $(\mathcal{X}, \mathcal{Y})$ is confouneded, an instrumental variable $\mathcal{W}$ is effective in identifying the causal relationship between $\mathcal{X}$ and $\mathcal{Y}$. Intuitively, $\mathcal{W}$ is an instrumental variable if it influences $\mathcal{Y}$ only through $\mathcal{X}$ which is formalized as follows.

Assumption C.2.1 (Instrumental Variable Model)**.** Let $(\mathcal{X}, \mathcal{Y}, \mathcal{W})$ be a random variable on the space $\mathfrak{X} \times \mathfrak{Y} \times \mathfrak{W}$ with joint distribution $\mathbb{P}_{\mathcal{X},\mathcal{Y},\mathcal{W}}$. We assume that

(i) $\mathcal{Y} = g(\mathcal{X}) + \epsilon$ and $\mathbb{E}[\epsilon \,|\, \mathcal{W} = w] = 0$ for any $w \in \mathfrak{W}$, and

(ii) $\mathbb{P}_{\mathcal{X}\,|\,\mathcal{W}}(x \,|\, w)$ does not remain when w varies.

Under Assumption C.2.1, $\mathcal{W}$ is an instrumental variable. Specifically, (i) of Assumption C.2.1 is the exclusion restriction, where function g is the structural function of interest and ϵ is the confounding noise. Also, (ii) of Assumption C.2.1 is the relevance condition, which ensures that $\mathcal{W}$ is informative in the sense that it depends on w in a nontrivial manner. We remark that the instrumental variable model generalizes the standard regression model. To see this, when $\mathcal{X} = \mathcal{W}$, the estimation of g reduces to standard regression of unconfounded inputs and it holds that $g(\cdot) = \mathbb{E}[\mathcal{Y} \,|\, \mathcal{X} = \cdot]$. In particular, the instrumental variable model allows that $\mathcal{X}$ and ϵ are confounded, i.e., $\mathcal{X}$ and ϵ are dependent. By Assumption C.2.1, we have the following estimation equation

$$\mathbb{E}[\mathcal{Y} \,|\, \mathcal{W} = w] = \mathbb{E}\big[g(\mathcal{X}) \,|\, \mathcal{W} = w\big]. \tag{C.2.4}$$

The right-hand side of (C.2.4) provides a two-stage method for estimating the function g. At the first stage, we estimate the conditional mean mean embedding of $\mathbb{P}_{\mathcal{X}\,|\,\mathcal{W}}$. Then, at the second stage, we estimate the function g via regressing $\mathcal{Y}$ on the empirical conditional mean mean embedding of $\mathbb{P}_{\mathcal{X}\,|\,\mathcal{W}}$ (Singh et al., 2019).

To ease the presentation, we consider the following mapping,

$$g_\theta \circ \texttt{attn}(q, K, V) \in \mathbb{R}^d,$$

where g_θ is a function approximator with a learnable parameter θ. For example, g_θ is a linear or kernel function. By Proposition 3.4.1 and Proposition 3.4.2, it holds that

$$g_\theta \circ \texttt{attn}(q, K, V) \approx g_\theta\big(\mathbb{E}[\mathcal{V}\,|\,\mathcal{K} = q]\big), \qquad \text{as} \quad L \to \infty.$$

Let the target variable be y. Then, the learning objective takes the following form,

$$\min_\theta \widehat{\mathbb{E}}\Big[\big\|y - g_\theta(\mathbb{E}[\mathcal{V}\,|\,\mathcal{K} = q])\big\|_2^2\Big],$$

which corresponds to the second stage of estimating the instrumental variable model. Note that $\mathbb{E}[\mathcal{V}\,|\,\mathcal{K} = q]$ is the conditional mean embedding of $\mathbb{P}_{\mathcal{V}\,|\,\mathcal{K}}$. Thus, the key $\mathcal{K}$ can be viewed as the instrumental variable (Pearl, 2009), which handles the endogeneity. We provide an alternative view on how the attention mechanism performs relational reasoning as a causal inference procedure.

C.2.3. Proof of Lemma 3.3.2

Proof. First, we prove the statement that $b_z(X) = \mathbb{P}(z = \cdot \,|\, X)$ is a minimal sufficient statistic of X for z. To see the sufficiency of $b_z(X)$ for z, note that

$$\mathbb{P}(z \,|\, X) = \mathbb{P}\big(z \,\big|\, \mathbb{P}(z = \cdot \,|\, X)\big) = \mathbb{P}\big(z \,\big|\, b_z(X)\big).$$

To see $b_z(X)$ is the minimal sufficient statistic, let $U(X)$ be another sufficient statistic of X for z. Then, we have

$$\mathbb{P}(z \,|\, X) = \mathbb{P}\big(z \,\big|\, U(X)\big),$$

which implies that $b_z(X)$ is a function of $U(X)$. Thus, $b_z(X)$ is minimal.

Second, we prove the statement that $b_z(X)$ is a minimal sufficient statistics of X for y. To see the sufficiency of $b_z(X)$ for y, note that

$$\mathbb{P}(y \,|\, X) = \int \mathbb{P}(y \,|\, z) \cdot \mathbb{P}(z \,|\, X) \mathrm{d}z,$$

which implies that $\mathbb{P}(y \,|\, X) = \mathbb{P}(y \,|\, b_z(X))$ since $\mathbb{P}(y \,|\, X)$ only depends on X through $b_z(X)$. Suppose that $U(X)$ is a sufficient statistic of X for y. We have

$$\int \mathbb{P}(y \,|\, z) \cdot \mathbb{P}(z \,|\, X) \mathrm{d}z = \mathbb{P}(y \,|\, X) = \mathbb{P}\big(y \,\big|\, U(X)\big) = \int \mathbb{P}(y \,|\, z) \cdot \mathbb{P}\big(z \,\big|\, U(X)\big) \mathrm{d}z.$$

By the definition of $\mathcal{T}$ in (3.3.1), we then have that

$$b_z(X) = \mathbb{P}(z = \cdot \,|\, X) = \mathcal{T}^{-1}\Big(\int \mathbb{P}(y = \cdot \,|\, z) \cdot \mathbb{P}\big(z \,\big|\, U(X)\big) \mathrm{d}z\Big),$$

which implies that $b_z(X)$ is a function of $U(X)$. Thus, $b_z(X)$ is minimal. □

C.2.4. Proof of Proposition 3.4.1

Proof. For notational simplicity, we denote by $\|\cdot\|$ the RKHS norm for elements in an RKHS and the operator norm for operators between two RKHSs. Also, we denote by $\mathcal{H}_k$ and $\mathcal{H}_v$ the RKHSs for the key and the value with the kernel functions $\mathfrak{K}$ and $\mathfrak{L}$, respectively. Note that we consider the Euclidean kernel $\mathfrak{L}(v, v') = v^\top v'$ for the value, which uses the identity mapping φ as the feature mapping. Recall the definition of the empirical covariance operator and the empirical cross-covariance operator in (C.1.4). Correspondingly, we write

$$\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}} = L^{-1}\phi(K)^\top\phi(K) \in \mathbb{R}^{d_\phi \times d_\phi}$$

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

$$\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{V}} = L^{-1}\varphi(V)^\top\varphi(V) \in \mathbb{R}^{d_\varphi \times d_\varphi}.$$

Here $\phi(K) = (\phi(k^1), \ldots, \phi(k^L))^\top \in \mathbb{R}^{L\times d_\phi}$ and $\varphi(V) = (\phi(v^1), \ldots, \phi(v^L))^\top \in \mathbb{R}^{L\times d_\varphi}$ By the definition of the CME attention in (3.4.5) and the equality in (C.1.5), we have that

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

which implies that $\texttt{attn}_{\texttt{CME}}$ recovers the empirical conditional mean embedding. By (C.1.2), it holds that

$$\begin{aligned}
&\left\|\texttt{attn}(q, K, V) - \texttt{CME}(q, \mathbb{P}_{\mathcal{K},\mathcal{V}})\right\| \\
&\quad \leq \underbrace{\left\|\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{K}}(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\phi(q) - \mathcal{C}_{\mathcal{V}\mathcal{K}}(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\phi(q)\right\|}_{\text{(i)}} \\
&\qquad + \underbrace{\left\|\mathcal{C}_{\mathcal{V}\mathcal{K}}(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\mathfrak{K}(q, \cdot) - \mathcal{C}_{\mathcal{V}\mathcal{K}}\mathcal{C}_{\mathcal{K}\mathcal{K}}^{-1}\mathfrak{K}(q, \cdot)\right\|}_{\text{(ii)}}.
\end{aligned} \tag{C.2.5}$$

Upper bounding term (i) of (C.2.5)**.** We adapt the proof from Song et al. (2009). It suffices to upper bound $\|\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{K}}(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} - \mathcal{C}_{\mathcal{V}\mathcal{K}}(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\|$. It holds that

(C.2.6)

$$\begin{aligned}
&\left\|\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{K}}(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} - \mathcal{C}_{\mathcal{V}\mathcal{K}}(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\right\| \\
&\quad \leq \left\|\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{K}}\big((\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} - (\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big)\right\| + \left\|(\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{K}} - \mathcal{C}_{\mathcal{V}\mathcal{K}})(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\right\| \\
&\quad = \left\|\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{K}}(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}} - \mathcal{C}_{\mathcal{K}\mathcal{K}})(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\right\| \\
&\qquad + \left\|(\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{K}} - \mathcal{C}_{\mathcal{V}\mathcal{K}})(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\right\|.
\end{aligned}$$

For the first term on the right-hand side of (C.2.6), we have the operator decomposition that $\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{K}} = \widehat{\mathcal{C}}_{\mathcal{V}\mathcal{V}}^{1/2}\mathcal{W}\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}}^{1/2}$ for $\mathcal{W}$ such that $\|\mathcal{W}\| \le 1$. Then, we have that

$$\begin{aligned}
&\big\|\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{K}}(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}} - \mathcal{C}_{\mathcal{K}\mathcal{K}})(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\| \\
&\le \|\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{V}}\|^{1/2}\big\|\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}}^{1/2}(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}}+L^{-1}\lambda\mathcal{I})^{-1/2}\big\|\big\|(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}}+L^{-1}\lambda\mathcal{I})^{-1/2}\big\|\big\|(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}}-\mathcal{C}_{\mathcal{K}\mathcal{K}})(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\| \\
&\le (L^{-1}\lambda)^{-1/2} \cdot \big\|(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}}-\mathcal{C}_{\mathcal{K}\mathcal{K}})(\mathcal{C}_{\mathcal{K}\mathcal{K}}+L^{-1}\lambda\mathcal{I})^{-1}\big\|,
\end{aligned} \tag{C.2.7}$$

where the last inequality follows from

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

Plugging (C.2.7) into (C.2.6), we have

$$\begin{aligned}
&\big\|\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{K}}(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} - \mathcal{C}_{\mathcal{V}\mathcal{K}}(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\| \\
&\quad \le (L^{-1}\lambda)^{-1/2} \cdot \big\|(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}} - \mathcal{C}_{\mathcal{K}\mathcal{K}})(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\| + \big\|(\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{K}} - \mathcal{C}_{\mathcal{V}\mathcal{K}})(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\|.
\end{aligned} \tag{C.2.8}$$

In what follows, we upper bound the second term on the right-hand side of (C.2.8) using Lemma C.7.2. We define $\xi : \mathbb{R}^{d_{\mathrm{p}}} \times \mathbb{R}^d \to \mathcal{H}_k \otimes \mathcal{H}_v$ as follows,

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

Since $\big\|(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\| \le (L^{-1}\lambda)^{-1}$, we have that

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

where $C > 0$ is an absolute constant. In addition, we have that

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

Using the trace operator, we have

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

Here $\Gamma(L^{-1}\lambda)$ is the effective dimension of $\mathcal{C}_{\mathcal{K}\mathcal{K}}$, which is defined as follows,

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

Applying Lemma C.7.2 with $B = C(L^{-1}\lambda)^{-1}$ and $\sigma^2 = (L^{-1}\lambda)^{-1} \cdot \Gamma(L^{-1}\lambda)$, we have with probability at least $1 - \delta$ that

$$\big\|\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{K}}(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} - \mathcal{C}_{\mathcal{V}\mathcal{K}}(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\| \leq C \cdot \left(\frac{2}{\lambda} + \sqrt{\frac{\Gamma(L^{-1}\lambda)}{\lambda}}\right) \log\frac{2}{\delta}, \tag{C.2.9}$$

where $C > 0$ is an absolute constant. Similarly, we have with probability at least $1 - \delta$ that

$$\big\|\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}}(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} - \mathcal{C}_{\mathcal{K}\mathcal{K}}(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\| \leq C' \cdot \left(\frac{2}{\lambda} + \sqrt{\frac{\Gamma(L^{-1}\lambda)}{\lambda}}\right) \log\frac{2}{\delta}. \tag{C.2.10}$$

Here $C' > 0$ is an absolute constant. Plugging (C.2.9) and (C.2.10) into (C.2.8), we have with probability at least $1 - \delta$ that

$$\begin{aligned} &\big\|\widehat{\mathcal{C}}_{\mathcal{V}\mathcal{K}}(\widehat{\mathcal{C}}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1} - \mathcal{C}_{\mathcal{V}\mathcal{K}}(\mathcal{C}_{\mathcal{K}\mathcal{K}} + L^{-1}\lambda\mathcal{I})^{-1}\big\| \\ &\quad \leq C'' \cdot \sqrt{\frac{L}{\lambda}} \cdot \left(\frac{2}{\lambda} + \sqrt{\frac{\Gamma(L^{-1}\lambda)}{\lambda}}\right) \log\frac{2}{\delta}. \end{aligned} \tag{C.2.11}$$

Upper bounding term (ii) of (C.2.5)**.** We adapt the proof from Fukumizu (2015). For any $g \in \mathcal{H}_k$, it holds that

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

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

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

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

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

Note that $\mathbb{E}[\mathfrak{L}(v,\bar{v})\,|\,k=\cdot,\bar{k}=\dagger] \in \mathcal{H}_k\otimes\mathcal{H}_k$ is in the range of $\mathcal{C}_{\mathcal{K}\mathcal{K}}\otimes\mathcal{C}_{\mathcal{K}\mathcal{K}}$. We define $\widetilde{\mathcal{C}} \in \mathcal{H}_k\times\mathcal{H}_k$ such that $(\mathcal{C}_{\mathcal{K}\mathcal{K}}\otimes\mathcal{C}_{\mathcal{K}\mathcal{K}})\widetilde{\mathcal{C}} = \mathbb{E}[\mathfrak{L}(v,\bar{v})\,|\,k=\cdot,\bar{k}=\dagger]$. Let $\{\lambda_i\}_{i=1}^{\infty}$ and $\{\varphi_i\}_{i=1}^{\infty}$

be the eigenvalues and eigenvectors of $\mathcal{C}_{\mathcal{K}\mathcal{K}}$, respectively. Then, we have that

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

Thus, we have

$$\big\|\mathcal{C}_{\mathcal{V}\mathcal{K}}(\mathcal{C}_{\mathcal{K}\mathcal{K}}+\lambda\mathcal{I})^{-1}\mathfrak{K}(q,\cdot)-\mathcal{C}_{\mathcal{V}\mathcal{K}}\mathcal{C}_{\mathcal{K}\mathcal{K}}^{-1}\mathfrak{K}(q,\cdot)\big\|_2\le C\cdot\lambda L^{-1},\tag{C.2.12}$$

where $C>0$ is an absolute constant.

Plugging (C.2.11) and (C.2.12) into (C.2.5), we have with probability at least $1-\delta$ that

$$\big\|\texttt{attn}(q,K,V)-\texttt{CME}(q,\mathbb{P}_{\mathcal{K},\mathcal{V}})\big\|\le\mathcal{O}\bigg(\sqrt{\frac{L}{\lambda}}\cdot\bigg(\frac{2}{\lambda}+\sqrt{\frac{\Gamma(L^{-1}\lambda)}{\lambda}}\bigg)\log\frac{1}{\delta}+\lambda L^{-1}\bigg).$$

Thus, we complete the proof of Proposition 3.4.1. □

C.2.5. Proof of Proposition 3.4.2

Proof. 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

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

Moreover, it holds that

$$\begin{aligned}\int v\widehat{\mathbb{P}}^{\mathfrak{K}}_{\mathcal{V}\,|\,\mathcal{K}}(v\,|\,q)\mathrm{d}v &= \iota \cdot \int_{\mathbb{S}^{d-1}} v \cdot \frac{\sum_{\ell=1}^{L} \mathfrak{K}(k^\ell, q) \cdot \mathfrak{K}(v^\ell, v)}{\sum_{\ell=1}^{L} \mathfrak{K}(k^\ell, q)}\mathrm{d}v \\ &= \frac{\iota \cdot \sum_{\ell=1}^{L} \mathfrak{K}(k^\ell, q) \cdot \int_{\mathbb{S}^{d-1}} v \cdot \mathfrak{K}(v^\ell, v)\mathrm{d}v}{\sum_{\ell=1}^{L} \mathfrak{K}(k^\ell, q)}, \qquad \text{(C.2.13)}\end{aligned}$$

where $\mathbb{S}^{d-1}$ is the $(d-1)$-dimensional unit sphere. 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 utilize the following lemma.

Lemma C.2.2. 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 §C.2.5.1 for a detailed proof. □

By Lemma C.2.2, it holds for the right-hand side of (C.2.13) 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 in (3.4.10). By setting $C = \iota \cdot C_1$, we complete the proof of Proposition 3.4.2. $\square$

C.2.5.1. Proof of Lemma C.2.2.

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

$$c = (a^\top b) \cdot b - \big(a - (a^\top b) \cdot b\big) \in \mathbb{S}^{d-1}. \tag{C.2.14}$$

By direct calculation, we have the following property of c defined in (C.2.14),

$$c^\top b = (a^\top b) \cdot \|b\|_2^2 - a^\top b + (a^\top b) \cdot \|b\|_2^2 = a^\top b. \tag{C.2.15}$$

By (C.2.14) and (C.2.15), we have that

$$a + c = 2(a^\top b) \cdot b = 2(c^\top b) \cdot b = (a^\top b) \cdot b + (c^\top b) \cdot b. \tag{C.2.16}$$

We now calculate the desired integration. Note that

(C.2.17)
$$\int_{\mathbb{S}^{d-1}} a \cdot \exp(a^\top b) \mathrm{d}a = b \cdot \int_{\mathbb{S}^{d-1}} (a^\top b) \exp(a^\top b) \mathrm{d}a + \int_{\mathbb{S}^{d-1}} \big(a - (a^\top b) \cdot b\big) \cdot \exp(a^\top b) \mathrm{d}a.$$

For the second term on the right-hand side of (C.2.17), it follows from (C.2.14) and (C.2.15) and (C.2.16) that

$$
\begin{aligned}
\int_{\mathbb{S}^{d-1}} \big(a - (a^\top b)\cdot b\big)\cdot \exp(a^\top b)\mathrm{d}a &= -\int_{\mathbb{S}^{d-1}} \big(c - (c^\top b)\cdot b\big)\cdot \exp(c^\top b)\mathrm{d}a \\
&= -\int_{\mathbb{S}^{d-1}} \big(c - (c^\top b)\cdot b\big)\cdot \exp(c^\top b)\mathrm{d}c,
\end{aligned}
\tag{C.2.18}
$$

where the second 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 (C.2.18), we have

$$
\int_{\mathbb{S}^{d-1}} \big(a - (a^\top b)\cdot b\big)\cdot \exp(a^\top b)\mathrm{d}a = -\int_{\mathbb{S}^{d-1}} \big(a - (a^\top b)\cdot b\big)\cdot \exp(a^\top b)\mathrm{d}a = 0
\tag{C.2.19}
$$

Finally, by plugging (C.2.19) into (C.2.17), 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 Lemma C.2.2. Note that here C_1 is an absolute constant that does not depend on b due to the symmetry on the unit sphere. □

C.3. Generalization Error Analysis

In this section, we analyze the generalization error of the complete setup of the transformer architecture, which involves multiple layers, skip connections, and multihead attentions. We collect the notations used throughout this section as follows.

Notations. For two positive reals r and s such that $1/r + 1/s = 1$, we call (r, s) a conjugate pair. We denote by $\|\cdot\|_r$ the vector ℓ_r-norm when it operates on a vector. Let $M = (m_1, \ldots, m_{d_2}) \in \mathbb{R}^{d_1 \times d_2}$, where $m_i \in \mathbb{R}^{d_1}$ with $i \in [d_2]$. We define the matrix (r, s)-norm as $\|M\|_{r,s} = (\sum_{i=1}^{d_2} \|m_i\|_r^s)^{1/s}$. We define the (r, s)-operator norm as $\|M\|_{r \to s} = \sup_{u \in \mathbb{R}^{d_2}} \|Mu\|_s / \|u\|_r$. We write $\|\cdot\|_r = \|\cdot\|_{r \to r}$ when the (r, r)-operator norm operates on a matrix.

C.3.1. Complete Setup of Transformer Architecture

In what follows, we specify the complete setup of a T-layer transformer parameterized by $\theta = (\bar{\theta}, \theta^{(0)}, \ldots, \theta^{(T-1)})$, where the t-th layer $(t = 0, \ldots, T-1)$ is parameterized by $\theta^{(t)} \in \Theta^{(t)}$ and the aggregation layer is parameterized by $\bar{\theta} \in \overline{\Theta}$. Here $\Theta^{(t)}$ and $\overline{\Theta}$ are the parameter spaces for the t-th layer and the aggregation layer, respectively. We define a two-layer feedforward neural network (FFN) with a skip connection (and no bias term) as follows,

$$\texttt{ffn}(X; A) = \texttt{ReLU}(XA^{\mathrm{x}})A^{\sigma} + X \in \mathbb{R}^{L \times d}, \tag{C.3.1}$$

which is parameterized by $A = (A^{\mathrm{x}}, A^{\sigma})$. Here $X \in \mathbb{R}^{L \times d}$,$A^{\mathrm{x}} \in \mathbb{R}^{d \times d_{\sigma}}$, $A^{\sigma} \in \mathbb{R}^{d_{\sigma} \times d}$, and $\texttt{ReLU}(\cdot)$ is the rectified linear unit (ReLU) that operates elementwise. Corresponding to (3.4.10), we define the sequence-to-sequence counterpart of the softmax attention as follows,

$$\texttt{attn}_{\texttt{SM}}(Q, K, V) = \Big(V^\top \texttt{norm}_{\texttt{SM}}\big(\mathfrak{K}_{\texttt{RBF}}(K, q^\ell)\big)\Big)^\top_{\ell\in[L]} \in \mathbb{R}^{L\times d}. \tag{C.3.2}$$

Here $Q = (q^\ell)^\top_{\ell\in[L]} \in \mathbb{R}^{L\times d_{\rm p}}$, $K = (k^\ell)^\top_{\ell\in[L]} \in \mathbb{R}^{L\times d_{\rm p}}$, $V \in \mathbb{R}^{L\times d}$, and $\mathfrak{K}_{\texttt{RBF}}(K, q^\ell) = (\mathfrak{K}_{\texttt{RBF}}(q^\ell, k^{\ell'}))^\top_{\ell'\in[L]} \in \mathbb{R}^L$ is specified in Assumption C.3.1. Recall that h is the head number of the multihead attention defined in (C.2.2) and $d = d_{\rm p}\cdot h$. With a slight abuse of notations, we define the sequence-to-sequence counterpart of the multihead attention (MHA) as follows,

$$\texttt{mha}(X; W) = \sum_{i=1}^h \texttt{head}_i = \sum_{i=1}^h \texttt{attn}_{\texttt{SM}}(Q_i, K_i, V_i) \in \mathbb{R}^{L\times d}, \tag{C.3.3}$$

which is parameterized by $W = \{(W_i^{\rm q}, W_i^{\rm k}, W_i^{\rm v})\}_{i\in[h]}$. Here $Q_i = XW_i^{\rm q} \in \mathbb{R}^{L\times d_{\rm p}}$, $K_i = XW_i^{\rm k} \in \mathbb{R}^{L\times d_{\rm p}}$, and $V_i = XW_i^{\rm v} \in \mathbb{R}^{L\times d}$ for the attention head $i \in [h]$, where $W_i^{\rm q}, W_i^{\rm k} \in \mathbb{R}^{d\times d_{\rm p}}$ and $W_i^{\rm v} \in \mathbb{R}^{d\times d}$.

With $X_\star^{(0)} = X$, let $X_\star^{(t)} \in \mathbb{R}^{L\times d}$ be the intermediate input of the t-th layer ($t = 0, \ldots, T-1$) of the transformer architecture, which is defined as follows,

$$\begin{aligned} X^{(t)} &= \texttt{ffn}(X_\star^{(t)}; A^{(t)}), & A^{(t)} &= (A^{{\rm x},(t)}, A^{\sigma,(t)}), \\ X_\star^{(t+1)} &= \texttt{mha}(X^{(t)}; W^{(t)}) + X^{(t)}, & W^{(t)} &= \big\{(W_i^{{\rm q},(t)}, W_i^{{\rm k},(t)}, W_i^{{\rm v},(t)})\big\}_{i\in[h]}, \end{aligned} \tag{C.3.4}$$

Here $\theta^{(t)} = (A^{(t)}, W^{(t)})$ is the learnable parameter. We compute the output as follows,

$$\widehat{y} = \overline{\texttt{agg}}_{\bar\theta}(X_\star^{(T)}) \in \mathbb{R}^{d_{\rm y}}, \tag{C.3.5}$$

where the aggregation layer $\overline{\mathtt{agg}}_{\bar{\theta}} : \mathbb{R}^{L\times d} \to \mathfrak{Y}$ is parameterized by $\bar{\theta}$. Here $\mathfrak{Y}$ is defined in (3.5.2). Note that $\overline{\mathtt{agg}}_{\bar{\theta}}$ combines the aggregation layer $\mathtt{agg}_{\theta_0}$ defined in (3.5.1) and the input mask `msk`. For example, in the complete setup of ViT (Dosovitskiy et al., 2020), the aggregation layer is the function composition $\overline{\mathtt{agg}}_{\bar{\theta}}(X_\star^{(T)}) = \mathtt{agg}_{\theta_0}(\mathtt{mha}(q_W(\mathtt{msk}), X_\star^{(T)}; W))$ with $\bar{\theta} = (\theta_0, W)$, where `msk` corresponds to the class encoding. Here the multihead attention $\mathtt{mha}(q_W(\mathtt{msk}), X_\star^{(T)}; W)$ follows the definition in (C.2.2).

Empirical Image Class. In what follows, we formalize the function class of the transformer architecture and the empirical image class for each layer. We define the base function class as

$$\mathcal{F}_{\mathtt{mha}}^{L,(0)} = \big\{X_\star^{(0)}(X)\big\},$$

which is the function class that only contains the identity mapping. Here we use $X_\star^{(0)}(X) = X$ to denote the identity mapping since $X_\star^{(0)} = X$. In the following, we use $X_\star^{(t)}(X)$ and $X^{(t-1)}(X)$ to denote the functions that map X to $X_\star^{(t)}$ and $X^{(t)}$ for $t = 1, \ldots, T-1$, respectively. We define the intermediate function classes recursively as follows,

$$\mathcal{F}_{\mathtt{ffn}}^{L,(t)} = \Big\{\mathtt{ffn}\big(X_\star^{(t)}(X); A^{(t)}\big) : A^{(t)} \in \mathfrak{A}^{(t)}, X_\star^{(t)}(X) \in \mathcal{F}_{\mathtt{mha}}^{L,(t)}\Big\},$$
$$\mathcal{F}_{\mathtt{mha}}^{L,(t+1)} = \Big\{\mathtt{mha}\big(X^{(t)}(X); W^{(t)}\big) + X^{(t)} : W^{(t)} \in \mathfrak{W}^{(t)}, X^{(t)}(X) \in \mathcal{F}_{\mathtt{ffn}}^{L,(t)}\Big\},$$

where $0 \le t \le T-1$. Here $\Theta^{(t)} = \mathfrak{A}^{(t)} \times \mathfrak{W}^{(t)}$ is the parameter space of the t-th layer of the transformer architecture. Correspondingly, we define the function class of the T-layer transformer as follows,

$$\mathcal{F}^L = \Big\{ \overline{\mathtt{agg}}_{\bar\theta}\big(X_\star^{(T)}(X)\big) : \bar\theta \in \overline{\Theta}, X_\star^{(T)}(X) \in \mathcal{F}_{\mathtt{mha}}^{L,(T)} \Big\}. \tag{C.3.6}$$

Correspondingly, we define the empirical image classes as follows,

$$\begin{aligned} \mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{ffn}}^{L,(t)}) &= \Big\{ \big(f(X_i)^\top\big)_{i\in[n]} \in \mathbb{R}^{d\times nL} : f \in \mathcal{F}_{\mathtt{ffn}}^{L,(t)} \Big\}, \\ \mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{mha}}^{L,(t+1)}) &= \Big\{ \big(f(X_i)^\top\big)_{i\in[n]} \in \mathbb{R}^{d\times nL} : f \in \mathcal{F}_{\mathtt{mha}}^{L,(t+1)} \Big\}, \\ \mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}^L) &= \Big\{ \big(f(X_i)\big)_{i\in[n]} \in \mathbb{R}^{d_{\mathrm{y}}\times n} : f \in \mathcal{F}^L \Big\}, \end{aligned} \tag{C.3.7}$$

where $0 \le t \le T-1$.

C.3.2. Generalization Error Analysis for Complete Setup

In what follows, we present a general version of Theorem 3.5.3, which allows for the complete setup of the transformer architecture. By specializing it to the single-layer transformer equipped with singlehead attention mechanism and no skip connection, we obtain Theorem 3.5.3.

In parallel to (3.5.4), we make the following assumption on the Gaussian RBF kernel $\mathfrak{K}_{\mathtt{RBF}}(q,k)$, which induces the multihead attention $\mathtt{mha}(X;W)$ defined in (C.3.2) and (C.3.3).

Assumption C.3.1 (Gaussian RBF Kernel)**.** Let $s > 0$. We assume that the multihead attention $\mathtt{mha}(X;W)$ adopts the Gaussian RBF kernel $\mathfrak{K}_{\mathtt{RBF}}(q,k) = \exp(-\|q-k\|_2^2/2\sigma^2)$ with $\sigma = (2d_{\mathrm{p}})^{1/s}$.

Note that the kernel function $\mathfrak{K}_{\mathtt{RBF}}(q,k)$ in Assumption C.3.1 is a general version of the Gaussian RBF kernel defined in (3.5.4), which corresponds to the special case where $s = 2$.

Recall that the output range $\mathfrak{Y}$ is defined in (3.5.2). We make the following assumption on the aggregation layer $\overline{\mathsf{agg}}_{\bar{\theta}}$ defined in (C.3.5).

Assumption C.3.2 (Aggregation Layer)**.** We assume that the aggregation layer $\overline{\mathsf{agg}}_{\bar{\theta}} : \mathbb{R}^{L\times d} \to \mathbb{R}^{d_{\mathrm{y}}}$ has the output range $\mathfrak{Y}$. Let $\overline{\mathsf{agg}}_{\bar{\theta},j} : \mathbb{R}^{L\times d} \to \mathbb{R}$ be the j-th entry ($j \in [d_{\mathrm{y}}]$) of the aggregation function $\overline{\mathsf{agg}}_{\bar{\theta}}$. We assume that for any $\bar{\theta} \in \overline{\Theta}$, $X_\star, \widetilde{X}_\star \in \mathbb{R}^{L\times d}$, and $j \in [d_{\mathrm{y}}]$, it holds that

$$\left|\overline{\mathsf{agg}}_{\bar{\theta},j}(X_\star) - \overline{\mathsf{agg}}_{\bar{\theta},j}(\widetilde{X}_\star)\right| \leq \|X_\star^\top - \widetilde{X}_\star^\top\|_{r,\infty}.$$

Recall that $\|\cdot\|_r$ denotes the (r,r)-operator norm when it operates on a matrix and $\|\cdot\|_{r,s}$ is the matrix (r,s)-norm. In parallel to Assumption 3.5.1, we make the following assumption on the parameter space for each layer of the transformer architecture.

Assumption C.3.3 (Parameter Space)**.** Let (r,s) be a conjugate pair. For $t = 0,\dots,T-1$, we assume that the parameter space $\Theta^{(t)} = \mathfrak{A}^{(t)} \times \mathfrak{W}^{(t)}$ of the t-th layer of the transformer architecture satisfy

(C.3.8)

$$\begin{aligned}
\mathfrak{A}^{(t)} = \Big\{&(A^{\mathrm{x},(t)}, A^{\sigma,(t)}) \in \mathbb{R}^{d \times d_\sigma} \times \mathbb{R}^{d_\sigma \times d} : \\
&\big\|(A^{\mathrm{x},(t)})^\top\big\|_r \le \alpha^{\mathrm{x},(t)}, \big\|(A^{\mathrm{x},(t)})^\top\big\|_{r,s} \le R^{\mathrm{x},(t)}, \\
&\big\|(A^{\sigma,(t)})^\top\big\|_r \le \alpha^{\sigma,(t)}, \big\|(A^{\sigma,(t)})^\top\big\|_{r,s} \le R^{\sigma,(t)}\Big\}, \\
\mathfrak{W}^{(t)} = \Big\{&\big\{(W_i^{\mathrm{q},(t)}, W_i^{\mathrm{k},(t)}, W_i^{\mathrm{v},(t)})\big\}_{i\in[h]} : (W_i^{\mathrm{q},(t)}, W_i^{\mathrm{k},(t)}, W_i^{\mathrm{v},(t)}) \in \mathbb{R}^{d\times d_\mathrm{p}} \times \mathbb{R}^{d\times d_\mathrm{p}} \times \mathbb{R}^{d\times d},
\end{aligned}$$

(C.3.9)

$$\begin{aligned}
&\big\|(W_i^{\mathrm{q},(t)})^\top\big\|_r \le \omega_i^{\mathrm{q},(t)}, \big\|(W_i^{\mathrm{k},(t)})^\top\big\|_r \le \omega_i^{\mathrm{k},(t)}, \big\|(W_i^{\mathrm{v},(t)})^\top\big\|_r \le \omega_i^{\mathrm{v},(t)}, \\
&\big\|(W_i^{\mathrm{q},(t)})^\top\big\|_{r,s} \le R_i^{\mathrm{q},(t)}, \big\|(W_i^{\mathrm{k},(t)})^\top\big\|_{r,s} \le R_i^{\mathrm{k},(t)}, \big\|(W_i^{\mathrm{v},(t)})^\top\big\|_{r,s} \le R_i^{\mathrm{v},(t)}\Big\}.
\end{aligned}$$

Also, we assume that the parameter space $\overline{\Theta}$ of the aggregation layer takes the form of $\overline{\Theta} = \{\bar\theta \in \mathbb{R}^{d_\mathrm{agg}} : \|\bar\theta\|_r \le 1\}$. Here $\alpha^{\mathrm{x},(t)}, R^{\mathrm{x},(t)}, \alpha^{\sigma,(t)}, R^{\sigma,(t)}, \omega_i^{\mathrm{q},(t)}, \omega_i^{\mathrm{k},(t)}, \omega_i^{\mathrm{v},(t)}, R_i^{\mathrm{q},(t)}, R_i^{\mathrm{k},(t)}, R_i^{\mathrm{v},(t)} > 0$ with $i \in [h]$ and $t = 0, \ldots, T-1$.

To ease the presentation, we define the following quantities that combine the parameter bounds across the h heads within the t-th layer of the transformer architecture,

$$\begin{aligned}
&\omega^{\mathrm{v},(t)} = \sum_{i=1}^h \omega_i^{\mathrm{v},(t)}, \qquad \overline{\omega}^{\mathrm{qk},(t)} = \max_{i\in[h]}\{\omega_i^{\mathrm{q},(t)} + \omega_i^{\mathrm{k},(t)}\}, \\
&R^{\mathrm{v},(t)} = \sum_{i=1}^h R_i^{\mathrm{v},(t)}, \qquad R^{\mathrm{qk},(t)} = \sum_{i=1}^h (R_i^{\mathrm{q},(t)} + R_i^{\mathrm{k},(t)}),
\end{aligned} \tag{C.3.10}$$

where $t = 0, \ldots, T-1$. Let

$$\widetilde{\alpha}^{(t)} = 1 + \alpha^{\mathrm{x},(t)}\alpha^{\sigma,(t)}, \qquad \widetilde{\omega}^{\mathrm{v},(t)} = 1 + \omega^{\mathrm{v},(t)}, \qquad \gamma^{(t)} = \max\{\widetilde{\alpha}^{(t)}, \widetilde{\omega}^{\mathrm{v},(t)}\}. \tag{C.3.11}$$

Also, let

(C.3.12)
$$\kappa^{(t)} = \max\biggl\{\frac{\alpha^{\mathrm{x},(t)}R^{\sigma,(t)} + \alpha^{\sigma,(t)}R^{\mathrm{x},(t)}}{\widetilde{\alpha}^{(t)}}, \frac{R^{\mathrm{v},(t)}}{\widetilde{\omega}^{\mathrm{v},(t)}}, \frac{R^{\mathrm{qk},(t)}}{\overline{\omega}^{\mathrm{qk},(t)}\omega^{\mathrm{v},(t)}}\biggr\}, \qquad \zeta^{(t)} = \frac{(\overline{\omega}^{\mathrm{qk},(t)})^2 R^{\mathrm{v},(t)}}{\widetilde{\omega}^{\mathrm{v},(t)}}.$$

Recall that the generalization error $\mathcal{E}_{\mathrm{gen}}$ is defined in (3.5.3). The following theorem characterizes the generalization error of the transformer.

Theorem C.3.4 (Generalization Error of Transformer)**.** Let $D = \max\{d, d_\sigma, d_{\mathrm{p}}, d_{\mathrm{y}}\}$. Suppose that Assumptions C.3.1-C.3.3 and Assumption 3.5.2 hold. Then, for any $\delta > 0$, it holds with probability at least $1 - \delta$ that,

$$\mathcal{E}_{\mathrm{gen}} = O\biggl(\frac{D^2}{\sqrt{n}}\Bigl[T\sqrt{\log(1+\gamma)} + \sqrt{T}\sqrt{\log(1+\zeta R)} + \sqrt{\log(1+\kappa/\zeta)}\Bigr]\sqrt{hT} + \sqrt{\frac{\log(1/\delta)}{n}}\biggr),$$

where R is defined in (3.5.2). Here

(C.3.13)
$$\gamma = \max_{0\le t\le T-1}\gamma^{(t)}, \qquad \kappa = \max_{0\le t\le T-1}\kappa^{(t)}, \qquad \zeta = \max_{0\le t\le T-1}\zeta^{(t)},$$

where $\gamma^{(t)}$, $\kappa^{(t)}$, and $\zeta^{(t)}$ are defined in (C.3.11)-(C.3.12).

Proof. See §C.3.4 for a detailed proof. □

Highlight. In comparison with Edelman et al. (2021), we exploit the invariance/equivariance property of the transformer architecture in a fine grained manner. The key observation is that, due to the invariance/equivariance property, the dimensions of the learnable parameters $W^{(t)}$ and $A^{(t)}$ are independent of the sequence length L. By such an observation, we characterize the covering number of the function class with the covering numbers of the parameter spaces and propagate them through the T layers. As a consequence, the covering number of the function class is independent of L. In contrast, the generalization error in Edelman et al. (2021) has a logarithmic dependency on L.

Interpretation. We interpret the generalization error in Theorem C.3.4 as follows. On the one hand, the $O(1/\sqrt{n}$ dependencies in $O(\sqrt{\log(1/\delta)/n})$ and $O(D^2/\sqrt{n})$ over the sample size n are standard in the literature. On the other hand, the $O(D^2h^{1/2}T^{3/2})$ scaling implies that the transformer architecture requires more training data points to generalize as the dimension D of the parameter space, the number T of layers, and the head number h grow. Also, the generalization error in Theorem C.3.4 only scales logarithmically in γ, κ, and ζ, which implies that the generalization error remains polynomial order as long as γ, κ, and ζ do not scale doubly exponentially with D, h, or T.

Implication. Theorem C.3.4 demonstrates that γ, κ, ζ and R play a crucial role in the generalization error of the transformer. Specifically, we observe that (i) skip connections allow all layers to resemble the identity mapping (Bartlett et al., 2018b,a; Hardt and Ma, 2016), which helps reducing γ, κ and ζ, and (ii) layer normalizations helps controlling the scaling of the intermediate inputs $\{X_\star^{(t)}\}_{0\leq t\leq T-1}$, which reduces the covering number of the function class.

Simplification of Theorem C.3.4 to Theorem 3.5.3. Theorem C.3.4 characterizes the generalization error of the complete setup of the transformer architecture, which includes Theorem 3.5.3 as a special case. In what follows, we specialize Theorem C.3.4 to obtain Theorem 3.5.3.

- Single-layer transformer with single-head attention: We set $h = 1$ and $T = 1$, which implies that

- (C.3.10) becomes

$$\omega^{\mathrm{v},(0)} = \omega_1^{\mathrm{v},(0)}, \qquad \bar{\omega}^{\mathrm{qk},(0)} = \omega_1^{\mathrm{q},(0)} + \omega_1^{\mathrm{k},(0)},$$

$$R^{\mathrm{v},(0)} = R_1^{\mathrm{v},(0)}, \qquad R^{\mathrm{qk},(0)} = R_1^{\mathrm{q},(0)} + R_1^{\mathrm{k},(0)},$$

- (C.3.13) becomes

$$\gamma = \gamma^{(0)}, \qquad \kappa = \kappa^{(0)}, \qquad \zeta = \zeta^{(0)},$$

  which are defined in (C.3.12) but will be redefined in our subsequent simplification, and

- the generalization error in Theorem C.3.4 becomes

$$\mathcal{E}_{\mathrm{gen}} \leq O\biggl(\frac{D^2}{\sqrt{n}} \cdot \Bigl[\sqrt{\log(1+\gamma)} + \sqrt{\log(1+\zeta R)} + \sqrt{\log(1+\kappa/\zeta)}\Bigr] + \sqrt{\frac{\log(1/\delta)}{n}}\biggr).$$

- Skip connections: Since we have no skip connections, we set

$$\widetilde{\alpha}^{(0)} = \alpha^{\mathrm{x},(0)}\alpha^{\sigma,(0)}, \qquad \widetilde{\omega}^{\mathrm{v},(0)} = \omega^{\mathrm{v},(0)},$$

  which appear in the definitions of $\gamma^{(0)}$, $\kappa^{(0)}$, and $\zeta^{(0)}$ in (C.3.11) and (C.3.12).

- Feedforward neural network: Since there is no linear transformation for the second layer of $\texttt{nn}(X; A) = \texttt{ReLU}(XA)$. Specifically, we
  - set $\alpha^{\sigma,(0)} = 1$ and $R^{\sigma,(0)} = 0$,
  - set $d_\sigma = d$ for the intermediate output, and
  - set $\alpha^{\texttt{nn}} = \alpha^{\mathrm{x},(0)}$ and $R^{\texttt{nn}} = R^{\mathrm{x},(0)}$ in Assumption 3.5.1.

As a consequence, we obtain

$$\gamma = \max\{\alpha^{\mathrm{nn}}, \omega^{\mathrm{v},(0)}\}, \qquad \kappa = \max\bigg\{ \frac{R^{\mathrm{nn}}}{\alpha^{\mathrm{nn}}}, \frac{R^{\mathrm{v},(0)}}{\omega^{\mathrm{v},(0)}}, \frac{R^{\mathrm{qk},(0)}}{\overline{\omega}^{\mathrm{qk},(0)}\omega^{\mathrm{v},(0)}} \bigg\}.$$

- Softmax attention: Since we have only one head, we set
  - $(\omega^{\mathrm{q}}, \omega^{\mathrm{k}}, \omega^{\mathrm{v}}) = (\omega^{\mathrm{q},(0)}, \omega^{\mathrm{k},(0)}, \omega^{\mathrm{v},(0)})$ and
  - $(R^{\mathrm{q}}, R^{\mathrm{k}}, R^{\mathrm{v}}) = (R^{\mathrm{q},(0)}, R^{\mathrm{k},(0)}, R^{\mathrm{v},(0)})$.

  As a consequence, we obtain $\overline{\omega}^{\mathrm{qk},(0)} = \omega^{\mathrm{q}} + \omega^{\mathrm{k}}$, $R^{\mathrm{qk},(0)} = R^{\mathrm{q}} + R^{\mathrm{k}}$, and

$$\gamma = \max\{\alpha^{\mathrm{nn}}, \omega^{\mathrm{v}}\}, \qquad \kappa = \max\bigg\{ \frac{R^{\mathrm{nn}}}{\alpha^{\mathrm{nn}}}, \frac{R^{\mathrm{v}}}{\omega^{\mathrm{v}}}, \frac{R^{\mathrm{k}} + R^{\mathrm{q}}}{(\omega^{\mathrm{q}} + \omega^{\mathrm{k}}) \cdot \omega^{\mathrm{v}}} \bigg\}, \qquad \zeta = \frac{(\omega^{\mathrm{q}} + \omega^{\mathrm{k}})^2 \cdot R^{\mathrm{v}}}{\omega^{\mathrm{v}}}.$$

- Spectral norm and Frobenius norm: We use the conjugate pair $(r, s) = (2, 2)$, which implies that
  - $\| \cdot \|_r = \| \cdot \|_2$ is the spectral norm of matrices and $\| \cdot \|_{r,s} = \| \cdot \|_{2,2} = \| \cdot \|_{\mathrm{F}}$ is the Frobenius norm, which correspond to Assumption 3.5.1, and
  - the Gaussian RBF kernel $\mathfrak{K}_{\mathtt{RBF}}(q, k)$ in Assumption C.3.1 is normalized by $\sigma = (2d_{\mathrm{p}})^{1/s} = (2d_{\mathrm{p}})^{1/2}$, which corresponds to (3.5.4).

Therefore, we obtain Theorem 3.5.3.

Proof Sketch. We organize the proof of Theorem C.3.4 as follows.

(§C.3.3) We review how to analyze the generalization error through the Rademacher complexity, which requires a covering number of the function class.

(§C.3.4) We provide a covering number of the function class and sketch the proof of Theorem C.3.4.

(§C.3.5) We characterize the covering number of the function class by (i) analyzing the covering number of each MHA and FFN and (ii) analyzing how the covering numbers propagate through the T layers of the transformer architecture.

(§C.6) We leave the detailed proofs of the intermediate lemmas to §C.6.

C.3.3. Preliminary of Generalization

In this section, we introduce the building blocks for analyzing the generalization error of the transformer architecture.

C.3.3.1. Rademacher Complexity. Suppose that the dataset $\mathcal{D}_n = \{(X_i, y_i)\}_{i=1}^n$ is drawn independently and identically from the data distribution $\mathcal{D}$. Recall that $\mathcal{F}^L$ is defined in (C.3.6). Let $\mathcal{L}((X, y), f)$ be a fixed learning objective. We define the function class $\mathcal{L} \circ \mathcal{F}^L$ as follows,

$$\mathcal{L} \circ \mathcal{F}^L = \Big\{ \mathcal{L}\big((X, y), f\big) : f \in \mathcal{F}^L \Big\}, \tag{C.3.14}$$

which contains the function compositions of the learning objective $\mathcal{L}$ and the transformer function $f \in \mathcal{F}^L$. We define the empirical Rademacher complexity of the function class $\mathcal{L} \circ \mathcal{F}^L$ as follows,

$$\mathcal{R}_{\mathcal{D}_n}(\mathcal{L} \circ \mathcal{F}^L) = \mathbb{E}\bigg[\sup_{f \in \mathcal{F}^L} \frac{1}{n} \sum_{i=1}^n \epsilon_i \cdot \mathcal{L}\big((X_i, y_i), f\big) \bigg], \tag{C.3.15}$$

where the expectation is taken over the independent Rademacher sequence $\{\epsilon_i\}_{i \in [n]}$. The following lemma characterizes the generalization error for learning with the function class $\mathcal{F}^L$.

Lemma C.3.5 (Generalization Error via Rademacher Complexity, Mohri et al. (2018))**.** Suppose that the function class $\mathcal{L} \circ \mathcal{F}^L$ defined in (C.3.14) has the output range $[0, 1]$. For any $\delta > 0$, with probability at least $1 - \delta$ over the independent and identical draw of the dataset $\mathcal{D}_n = \{(X_i, y_i)\}_{i \in [n]}$ from the data distribution $\mathcal{D}$, it holds for all $f \in \mathcal{F}^L$ that,

$$\left| \mathbb{E}\Big[\mathcal{L}\big((X, y), f\big)\Big] - \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X, y), f\big)\Big] \right| \le 2\mathcal{R}_{\mathcal{D}_n}(\mathcal{L} \circ \mathcal{F}^L) + 3\sqrt{\frac{\log(2/\delta)}{2n}},$$

where $\widehat{\mathbb{E}}[\,\cdot\,]$ is the empirical expectation taken over the dataset $\mathcal{D}_n$ and the empirical Rademacher complexity $\mathcal{R}_{\mathcal{D}_n}(\mathcal{L} \circ \mathcal{F}^L)$ is defined in (C.3.15).

We define the product function class as follows,

$$\prod_{j=1}^{d_{\mathrm{y}}} \mathcal{F}_j^L = \Big\{ \big(f_j(X)\big)_{j \in [d_{\mathrm{y}}]}^\top : f_j \in \mathcal{F}_j^L, j \in [d_{\mathrm{y}}] \Big\},$$

where $\mathcal{F}_j^L$ is the function class of the j-th entry ($j \in [d_{\mathrm{y}}]$) of $f = (f_1, \ldots, f_{d_{\mathrm{y}}})^\top \in \mathcal{F}^L$. The following lemma characterizes the empirical Rademacher complexity $\mathcal{R}_{\mathcal{D}_n}(\mathcal{L} \circ \mathcal{F}^L)$.

Lemma C.3.6 (Vector Contraction Inequality, Maurer (2016))**.** Let $\mathcal{F}$ be the function class of $f : \mathbb{R}^{L \times d} \to \mathfrak{Y} \subseteq \mathbb{R}^{d_{\mathrm{y}}}$ and let $\mathcal{L}_i : \mathfrak{Y} \to \mathbb{R}$ be a 1-Lipschitz function with respect to the vector ℓ_2-norm, where $i \in [n]$. Then, we have

$$\mathbb{E}\bigg[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i \cdot \mathcal{L}_i\big(f(X_i)\big)\bigg] \le \sqrt{2} \cdot \mathbb{E}\bigg[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \sum_{j=1}^{d_{\mathrm{y}}} \epsilon_{ij} \cdot f_j(X_i)\bigg],$$

where $\{\epsilon_i\}_{i \in [n]}$ is an independent Rademacher sequence, $\{\epsilon_{ij}\}_{i \in [n], j \in [d]}$ is an independent Rademacher sequence that is doubly indexed, $f_j(X_i)$ is the j-th entry of $f(X_i)$, and the expectations are taken over the independent Rademacher sequences.

Lemma C.3.6 generalizes the Ledoux-Talagrand contraction inequality (Ledoux and Talagrand, 1991) to the multivariate setting. Note that $\mathcal{F}^L \subseteq \prod_{j=1}^{d_{\mathrm{y}}} \mathcal{F}_j^L$ since all entries of $f \in \mathcal{F}^L$ share the same parameters. By setting $\mathcal{L}_i(f(X_i)) = \mathcal{L}((X_i, y_i), f)$ in Lemma C.3.6, we have

$$\begin{aligned} \mathcal{R}_{\mathcal{D}_n}(\mathcal{L} \circ \mathcal{F}^L) &\leq \sqrt{2} \cdot \mathbb{E}\biggl[\sup_{f \in \mathcal{F}^L} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{y}}} \epsilon_{ij} \cdot f_j(X_i)\biggr] \\ &\leq \sqrt{2} \cdot \mathbb{E}\biggl[\sup_{\{f_j \in \mathcal{F}_j^L\}_{j \in [d_{\mathrm{y}}]}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{y}}} \epsilon_{ij} \cdot f_j(X_i)\biggr] \\ &\leq \sqrt{2} \cdot \sum_{j=1}^{d_{\mathrm{y}}} \mathbb{E}\biggl[\sup_{f_j \in \mathcal{F}_j} \frac{1}{n} \sum_{i=1}^{n} \epsilon_{ij} \cdot f_j(X_i)\biggr] \\ &= \sqrt{2} \cdot \sum_{j=1}^{d_{\mathrm{y}}} \mathcal{R}_{\mathcal{D}_n}(\mathcal{F}_j^L), \end{aligned} \tag{C.3.16}$$

which implies that it remains to characterize the empirical Rademacher complexity $\mathcal{R}_{\mathcal{D}_n}(\mathcal{F}_j^L)$, where $j \in [d_{\mathrm{y}}]$.

C.3.3.2. Rademacher Complexity via Covering Number. In what follows, we connect the Rademacher complexity $\mathcal{R}_{\mathcal{D}_n}(\mathcal{F}_j^L)$ to the covering number of the empirical image class $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j^L)$, which is defined as

$$\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j^L) = \Bigl\{\bigl(f_j(X_i)\bigr)_{i \in [n]} \in \mathbb{R}^{1 \times n} : f_j \in \mathcal{F}_j^L\Bigr\}. \tag{C.3.17}$$

We define the proper covering number as follows.

Definition C.3.7 (Proper Covering Number)**.** Let $N(\mathcal{S}, \varepsilon, \|\cdot\|)$ be the least cardinality of any subset $\mathcal{T} \subseteq \mathcal{S}$ that covers the set $\mathcal{S}$ at the resolution ε with respect to the norm

$\|\cdot\|$, that is,

$$N\big(\mathcal{S}, \varepsilon, \|\cdot\|\big) = \inf\Big\{\mathrm{card}(\mathcal{T}) : \sup_{S\in\mathcal{S}} \inf_{T\in\mathcal{T}} \|S - T\| \le \varepsilon, \mathcal{T} \subseteq \mathcal{S}\Big\},$$

where $\mathrm{card}(\mathcal{T})$ is the cardinality of the set $\mathcal{T}$.

To characterize $\mathcal{R}_{\mathcal{D}_n}(\mathcal{F}_j)$, we use the following version of the Dudley entropy integral lemma for the matrix (r, ∞)-norm.

Lemma C.3.8 (Dudley Entropy Integral for $\|\cdot\|_{r,\infty}$-Covering, Mohri et al. (2018))**.** Let $\mathcal{F}_j^L$ be a function class with the output range $[0, 1/2]$. Suppose that $0 \in \mathcal{F}_j^L$. Then, we have

$$\mathcal{R}_{\mathcal{D}_n}(\mathcal{F}_j^L) \le \inf_{\xi\in(0,1/2)} \bigg(4\xi + \frac{12}{\sqrt{n}} \int_{\xi}^{1/2} \sqrt{\log N\big(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j^L), \varepsilon, \|\cdot\|_{r,\infty}\big)\mathrm{d}\varepsilon}\bigg).$$

Proof. See §C.6.1 for a detailed proof. □

By Lemmas C.3.5-C.3.8, we see that it remains to characterize the covering number of the empirical image class $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j^L)$ with respect to the matrix (r, ∞)-norm.

C.3.4. Proof of Theorem C.3.4

Recall that the parameter spaces $\mathfrak{A}^{(t)}$ and $\mathfrak{W}^{(t)}$ are defined in (C.3.8) and (C.3.9), respectively. Also, recall that $\widetilde{\alpha}^{(t)}$, $\widetilde{\omega}^{\mathrm{v},(t)}$, $\overline{\omega}^{\mathrm{qk},(t)}$, and $R^{\mathrm{qk},(t)}$ are defined in (C.3.10)-(C.3.11). For the dataset $\mathcal{D}_n = \{(X_i, y_i)\}_{i\in[n]}$, we define

$$R^{(0)} = \max_{i\in[n]} \|X_i^\top\|_{r,\infty}, \qquad R^{(t)} = R^{(0)} \cdot \prod_{\tau=0}^{t-1} \widetilde{\omega}^{\mathrm{v},(\tau)}\widetilde{\alpha}^{(\tau)}, \tag{C.3.18}$$

which characterize the scaling of the intermediate input $X_\star^{(t)}$ for the t-th layer of the transformer architecture, where $t = 0, \ldots, T-1$. The following lemma characterizes the covering number of the empirical image class $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j^L)$.

Lemma C.3.9 (Covering Number of Transformer Architecture)**.** Let $D = \max\{d, d_\sigma, d_{\mathrm{p}}, d_{\mathrm{y}}\}$. Under Assumption C.3.3, we have for any $j \in [d_{\mathrm{y}}]$ that

$$\log N\big(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j^L), \varepsilon, \|\cdot\|_{r,\infty}\big) \le (4+h)D^2 T \cdot \log\biggl(1 + \frac{2R^{(T)} R_{\mathtt{trans}}}{\varepsilon}\biggr).$$

Here $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j^L)$ is defined in (C.3.17), $R^{(t)}$ is defined in (C.3.18), and

$$R_{\mathtt{trans}} = \sum_{t=0}^{T-2}\biggl(\frac{R_{\mathtt{mha}}^{(t)}}{\widetilde{\rho}^{(t)}} \cdot \prod_{\tau=t+1}^{T-1} \frac{\widetilde{\rho}^{(\tau)}}{\widetilde{\omega}^{\mathrm{v},(\tau)}}\biggr) + \sum_{t=0}^{T-1}\biggl(\frac{\alpha^{\mathrm{x},(t)} R^{\sigma,(t)} + \alpha^{\sigma,(t)} R^{\mathrm{x},(t)}}{\widetilde{\alpha}^{(t)}} \cdot \prod_{\tau=t+1}^{T-1} \frac{\widetilde{\rho}^{(\tau)}}{\widetilde{\omega}^{\mathrm{v},(\tau)}}\biggr), \tag{C.3.19}$$

where $\alpha^{\mathrm{x},(t)}$, $\alpha^{\sigma,(t)}$, $R^{\mathrm{x},(t)}$, and $R^{\sigma,(t)}$ are defined in Assumption C.3.3, and

$$\widetilde{\rho}^{(t)} = \widetilde{\omega}^{\mathrm{v},(t)} + (\bar{\omega}^{\mathrm{qk},(t)})^2 \omega^{\mathrm{v},(t)} \cdot (R^{(t)})^2, \qquad R_{\mathtt{mha}}^{(t)} = R^{\mathrm{v},(t)} + \bar{\omega}^{\mathrm{qk},(t)} R^{\mathrm{qk},(t)} \cdot (R^{(t)})^2. \tag{C.3.20}$$

Proof. See §C.3.5.3 for a detailed proof. □

Proof of Theorem C.3.4. For any $a > 0$, we have

$$\begin{aligned}
\int_\xi^{1/2} \sqrt{\log(1 + a/\varepsilon)}\mathrm{d}\varepsilon &\le \int_\xi^{1/2} \bigl[\sqrt{\log(1+a)} + \sqrt{\log(1+1/\varepsilon)}\bigr]\mathrm{d}\varepsilon \\
&\le (1/2 - \xi) \cdot \sqrt{\log(1+a)} + \int_\xi^{1/2} 1/\sqrt{\varepsilon}\mathrm{d}\varepsilon \\
&= (1/2 - \xi) \cdot \sqrt{\log(1+a)} + \sqrt{2} - 2\sqrt{\xi}, \qquad \text{(C.3.21)}
\end{aligned}$$

where the first inequality follows from the fact that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for any $a, b \ge 0$ and the second inequality follows from the fact that $\log(1+a) \le a$ for any $a > -1$. By Lemma C.3.8, we have for any $j \in [d_{\mathrm{y}}]$ that

(C.3.22)

$$\begin{aligned}
&\mathcal{R}_{\mathcal{D}_n}(\mathcal{F}_j^L) \\
&\quad \le \inf_{\xi \in (0,1/2)} \biggl(4\xi + \frac{12}{\sqrt{n}} \int_{\xi}^{1/2} \sqrt{\log N\bigl(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j^L), \varepsilon, \|\cdot\|_{r,\infty}\bigr)} \mathrm{d}\varepsilon \biggr) \\
&\quad \le \inf_{\xi \in (0,1/2)} \biggl(4\xi + \frac{12D\sqrt{(4+h)T}}{\sqrt{n}} \int_{\xi}^{1/2} \sqrt{\log\biggl(1 + \frac{2R^{(T)}R_{\mathtt{trans}}}{\varepsilon}\biggr)} \mathrm{d}\varepsilon \biggr) \\
&\quad \le 12D\sqrt{(4+h)T} \cdot \frac{2\sqrt{2} + \sqrt{\log(1 + 2R^{(T)}R_{\mathtt{trans}})}}{2\sqrt{n}} \\
&\quad + \inf_{\xi \in (0,1/2)} \biggl[\biggl(4 - \frac{12D\sqrt{(4+h)T}\sqrt{\log(1+2R^{(T)}R_{\mathtt{trans}})}}{\sqrt{n}} \biggr) \cdot \xi - \frac{24D\sqrt{(4+h)T}}{\sqrt{n}} \sqrt{\xi} \biggr],
\end{aligned}$$

where the second inequality follows from Lemma C.3.9 and the last inequality follows from (C.3.21). On the right-hand side of (C.3.22), we set $\xi = 0^+$ to obtain for any $j \in [d_{\mathrm{y}}]$ that

$$\begin{aligned}
\mathcal{R}_{\mathcal{D}_n}(\mathcal{F}_j^L) &\le 12D\sqrt{(4+h)T} \cdot \frac{2\sqrt{2} + \sqrt{\log(1+2R^{(T)}R_{\mathtt{trans}})}}{2\sqrt{n}} \\
&= O\biggl(D\sqrt{hT} \cdot \frac{1 + \sqrt{\log(1 + R^{(T)}R_{\mathtt{trans}})}}{\sqrt{n}} \biggr). \qquad \text{(C.3.23)}
\end{aligned}$$

Recall that $\mathcal{F}^L$ is defined in (C.3.6). Taking (C.3.23) into Lemma C.3.5, Lemma C.3.6, and (C.3.16), we obtain with probability at least $1-\delta$ over the independent and identical draw of the dataset $\mathcal{D}_n = \{(X_i, y_i)\}_{i\in[n]}$, it holds for any $f \in \mathcal{F}^L$ that

$$\begin{aligned}
&\Big|\mathbb{E}\Big[\mathcal{L}\big((X,y),f\big)\Big] - \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),f\big)\Big]\Big| \\
&\quad\le 2\sqrt{2}\cdot\sum_{j=1}^{d_y}\mathcal{R}_{\mathcal{D}_n}(\mathcal{F}_j^L) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \\
&\quad= O\biggl(D^2\sqrt{hT}\cdot\frac{1+\sqrt{\log(1+R^{(T)}R_{\texttt{trans}})}}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\biggr). \qquad \text{(C.3.24)}
\end{aligned}$$

By simplifying the first term on the right-hand side of (C.3.24) using Lemma C.7.5, we have with probability at least $1-\delta$ over the independent and identical draw of the dataset $\mathcal{D}_n = \{(X_i, y_i)\}_{i\in[n]}$ that

(C.3.25)

$$\begin{aligned}
&\Big|\mathbb{E}\Big[\mathcal{L}\big((X,y),f\big)\Big] - \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),f\big)\Big]\Big| \\
&\quad= O\biggl(\frac{D^2}{\sqrt{n}}\cdot\Big[T\cdot\sqrt{\log(1+\gamma)} + \sqrt{T}\cdot\sqrt{\log(1+\zeta R^{(0)})} + \sqrt{\log(1+\kappa/\zeta)}\Big]\cdot\sqrt{hT} \\
&\qquad + \sqrt{\frac{\log(1/\delta)}{n}}\biggr), \\
&\quad= O\biggl(\frac{D^2}{\sqrt{n}}\cdot\Big[T\cdot\sqrt{\log(1+\gamma)} + \sqrt{T}\cdot\sqrt{\log(1+\zeta R)} + \sqrt{\log(1+\kappa/\zeta)}\Big]\cdot\sqrt{hT} \\
&\qquad + \sqrt{\frac{\log(1/\delta)}{n}}\biggr)
\end{aligned}$$

holds for any $f \in \mathcal{F}^L$, where the last line follows from $R \ge R^{(0)}$ defined in Assumption 3.5.2. Here $R^{(0)}$ is defined in (C.3.18). Recall that $\widetilde{f} = \operatorname{argmin}_{f\in\mathcal{F}^L}\widehat{\mathbb{E}}[\mathcal{L}((X,y),f)]$ is the empirical risk minimizer, where $\mathcal{F}^L$ is the T-layer version of $\mathcal{F}_{\texttt{attn}}$ defined in (3.5.1). Let

$\overline{f} = \operatorname{argmin}_{f \in \mathcal{F}^L} \mathbb{E}[\mathcal{L}((X,y),f)]$ be the population risk minimizer. We have

$$
\begin{aligned}
&\widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),\widetilde{f}\big)\Big] - \min_{f\in\mathcal{F}^L} \mathbb{E}\Big[\mathcal{L}\big((X,y),f\big)\Big] \\
&\quad = \min_{f\in\mathcal{F}^L} \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),f\big)\Big] - \mathbb{E}\Big[\mathcal{L}\big((X,y),\overline{f}\big)\Big] \\
&\quad = \min_{f\in\mathcal{F}^L} \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),f\big)\Big] - \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),\overline{f}\big)\Big] + \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),\overline{f}\big)\Big] - \mathbb{E}\Big[\mathcal{L}\big((X,y),\overline{f}\big)\Big] \\
&\quad \le \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),\overline{f}\big)\Big] - \mathbb{E}\Big[\mathcal{L}\big((X,y),\overline{f}\big)\Big].
\end{aligned}
$$

By the definition of the generalization error $\mathcal{E}_{\text{gen}}$ in (3.5.3), we have with probability at least $1-\delta$ over the independent and identical draw of the dataset $\mathcal{D}_n = \{(X_i, y_i)\}_{i\in[n]}$ that

$$
\begin{aligned}
\mathcal{E}_{\text{gen}} &= \mathbb{E}\Big[\mathcal{L}\big((X,y),\widehat{f}\big)\Big] - \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),\widehat{f}\big)\Big] + \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),\widetilde{f}\big)\Big] - \min_{f\in\mathcal{F}^L} \mathbb{E}\Big[\mathcal{L}\big((X,y),f\big)\Big] \\
&\le \mathbb{E}\Big[\mathcal{L}\big((X,y),\widehat{f}\big)\Big] - \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),\widehat{f}\big)\Big] + \widehat{\mathbb{E}}\Big[\mathcal{L}\big((X,y),\overline{f}\big)\Big] - \mathbb{E}\Big[\mathcal{L}\big((X,y),\overline{f}\big)\Big] \\
&= O\bigg(\frac{D^2}{\sqrt{n}} \cdot \Big[T \cdot \sqrt{\log(1+\gamma)} + \sqrt{T} \cdot \sqrt{\log(1+\zeta R)} + \sqrt{\log(1+\kappa/\zeta)}\Big] \cdot \sqrt{hT} \\
&\qquad + \sqrt{\frac{\log(1/\delta)}{n}}\bigg),
\end{aligned}
$$

where the last line follows from (C.3.25). Therefore, we conclude the proof of Theorem C.3.4. □

C.3.5. Proof of Lemma C.3.9

We organize the proof of Lemma C.3.9 as follows.

(§C.3.5.1) We analyze the covering numbers of the empirical image classes of MHA and FFN, respectively, which serve as the building blocks for covering the empirical image class of the transformer architecture.

(§C.3.5.2) We propagate the covering numbers of MHA and FFN through the T layers of the transformer architecture.

(§C.3.5.3) We combine §C.3.5.1 and §C.3.5.2 to characterize the covering number of the empirical image class of the transformer architecture.

C.3.5.1. Covering Numbers of FFN and MHA. In what follows, we characterize the covering numbers of the empirical image classes of FFN and MHA. The following lemma characterizes the covering number of a set of matrices with respect to the matrix (r,s)-norm, which serves as the building block for analyzing the covering number of the empirical image classes of FFN and MHA.

Lemma C.3.10 (Covering Number of Matrix Set)**.** Let (r,s) be a conjugate pair. We have

$$\log N\Big(\big\{M^\top \in \mathbb{R}^{d_2\times d_1} : \|M^\top\|_{r,s} \le R_M\big\}, \varepsilon, \|\cdot\|_{r,s}\Big) \le d_1 d_2 \cdot \log\bigg(1 + \frac{2R_M}{\varepsilon}\bigg),$$

where $R_M, \varepsilon > 0$.

Proof. See §C.6.1.1 for a detailed proof. □

Empirical Image Class of FFN. In the sequel, we characterize the covering number of the empirical image class of FFN. In parallel to the parameter space $\mathfrak{A}^{(t)}$ defined in

(C.3.8), we define

$$
\begin{aligned}
&\mathfrak{A} = \Big\{(A^{\mathrm{x}}, A^{\sigma}) \in \mathbb{R}^{d\times d_\sigma} \times \mathbb{R}^{d_\sigma \times d} : \\
&\qquad \big\|(A^{\mathrm{x}})^\top\big\|_r \le \alpha^{\mathrm{x}}, \big\|(A^{\mathrm{x}})^\top\big\|_{r,s} \le R^{\mathrm{x}}, \big\|(A^{\sigma})^\top\big\|_r \le \alpha^{\sigma}, \big\|(A^{\sigma})^\top\big\|_{r,s} \le R^{\sigma}\Big\}, \qquad \text{(C.3.26)}
\end{aligned}
$$

where $\alpha^{\mathrm{x}}, \alpha^{\sigma}, R^{\mathrm{x}}, R^{\sigma} > 0$, while d and d_σ are positive integers. Correspondingly, we define the function class of FFN and the empirical image class of FFN as follows,

$$
\begin{aligned}
&\mathcal{F}_{\mathtt{ffn}} = \big\{\mathtt{ffn}(X_\star; A) : A \in \mathfrak{A}\big\}, \\
&\mathfrak{I}_{\widetilde{\mathcal{D}}_{n\star}}(\mathcal{F}_{\mathtt{ffn}}) = \Big\{\big(f(\widetilde{X}_{i\star})^\top\big)_{i\in[n]} \in \mathbb{R}^{d\times nL} : f \in \mathcal{F}_{\mathtt{ffn}}\Big\}. \qquad \text{(C.3.27)}
\end{aligned}
$$

Here $\widetilde{\mathcal{D}}_{n\star} = \{\widetilde{X}_{i\star}\}_{i\in[n]}$ is the input set of FFN, where $\widetilde{X}_{i\star} \in \mathbb{R}^{L\times d}$. In the following, we characterize the covering number of the empirical image class $\mathfrak{I}_{\widetilde{\mathcal{D}}_{n\star}}(\mathcal{F}_{\mathtt{ffn}})$.

Lemma C.3.11 (Covering Number of FFN)**.** Let $\varepsilon > 0$. Suppose that the input set $\widetilde{D}_{n\star} = \{\widetilde{X}_{i\star}\}_{i\in[n]} \subseteq \mathbb{R}^{L\times d}$ of FFN satisfies $\max_{i\in[n]} \|\widetilde{X}_{i\star}^\top\|_{r,\infty} \le \widetilde{R}_\star$. Then, we have

$$
\log N\big(\mathfrak{I}_{\widetilde{\mathcal{D}}_{n\star}}(\mathcal{F}_{\mathtt{ffn}}), \varepsilon, \|\cdot\|_{r,\infty}\big) \le 2dd_\sigma \cdot \log\bigg(1 + \frac{2(\alpha^{\mathrm{x}} R^{\sigma} + \alpha^{\sigma} R^{\mathrm{x}}) \cdot \widetilde{R}_\star}{\varepsilon}\bigg),
$$

where $\alpha^{\mathrm{x}}, \alpha^{\sigma}, R^{\mathrm{x}}, R^{\sigma}$ are defined in (C.3.26).

Proof. For any $A = (A^{\mathrm{x}}, A^{\sigma}) \in \mathfrak{A}$, suppose that $\widehat{A} = (\widehat{A}^{\mathrm{x}}, \widehat{A}^{\sigma}) \in \mathfrak{A}$ satisfy

$$
\big\|(A^{\mathrm{x}})^\top - (\widehat{A}^{\mathrm{x}})^\top\big\|_{r,s} \le \varepsilon^{\mathrm{x}}, \qquad \big\|(A^{\sigma})^\top - (\widehat{A}^{\sigma})^\top\big\|_{r,s} \le \varepsilon^{\sigma}, \qquad \text{(C.3.28)}
$$

where $\varepsilon^{\mathrm{x}}, \varepsilon^{\sigma} > 0$. For any $i \in [n]$, we have

$$\begin{aligned}
&\big\|\mathtt{ffn}(\widetilde{X}_{i\star}; A)^{\top} - \mathtt{ffn}(\widetilde{X}_{i\star}; \widehat{A})^{\top}\big\|_{r,\infty} \\
&\quad = \Big\|\big(\mathtt{ReLU}(\widetilde{X}_{i\star} A^{\mathrm{x}}) A^{\sigma}\big)^{\top} + \widetilde{X}_{i\star}^{\top} - \big(\mathtt{ReLU}(\widetilde{X}_{i\star} \widehat{A}^{\mathrm{x}}) \widehat{A}^{\sigma}\big)^{\top} - \widetilde{X}_{i\star}^{\top}\Big\|_{r,\infty} \\
&\quad \leq \Big\|\big((A^{\sigma})^{\top} - (\widehat{A}^{\sigma})^{\top}\big)\mathtt{ReLU}(\widetilde{X}_{i\star} A^{\mathrm{x}})^{\top}\Big\|_{r,\infty} + \Big\|(\widehat{A}^{\sigma})^{\top}\big(\mathtt{ReLU}(\widetilde{X}_{i\star} \widehat{A}^{\mathrm{x}}) - \mathtt{ReLU}(\widetilde{X}_{i\star} A^{\mathrm{x}})\big)^{\top}\Big\|_{r,\infty} \\
&\quad \leq \big\|(A^{\sigma})^{\top} - (\widehat{A}^{\sigma})^{\top}\big\|_{r,s} \cdot \big\|(A^{\mathrm{x}})^{\top}\big\|_{r} \cdot \|\widetilde{X}_{i\star}^{\top}\|_{r,\infty} + \big\|(\widehat{A}^{\sigma})^{\top}\big\|_{r} \cdot \big\|(\widehat{A}^{\mathrm{x}} - A^{\mathrm{x}})^{\top}\big\|_{r,s} \cdot \|\widetilde{X}_{i\star}^{\top}\|_{r,\infty} \\
&\quad \leq \varepsilon^{\sigma} \cdot \big\|(A^{\mathrm{x}})^{\top}\big\|_{r} \cdot \|\widetilde{X}_{i\star}^{\top}\|_{r,\infty} + \varepsilon^{\mathrm{x}} \cdot \big\|(\widehat{A}^{\sigma})^{\top}\big\|_{r} \cdot \|\widetilde{X}_{i\star}^{\top}\|_{r,\infty} \\
&\quad \leq (\varepsilon^{\sigma}\alpha^{\mathrm{x}} + \varepsilon^{\mathrm{x}}\alpha^{\sigma}) \cdot \widetilde{R}_{\star},
\end{aligned}$$

where the second line follows from the definition of FFN in (C.3.1), the fourth line follows from Lemma C.7.3, the fifth line follows from (C.3.28), and the last line follows from the definition of $\mathcal{F}_{\mathtt{ffn}}$ in (C.3.27) and the fact that $\max_{i\in[n]} \|\widetilde{X}_{i\star}^{\top}\|_{r,\infty} \leq \widetilde{R}_{\star}$. Setting

$$\varepsilon^{\mathrm{x}} = \varepsilon \cdot \frac{R^{\mathrm{x}}}{(\alpha^{\mathrm{x}} R^{\sigma} + \alpha^{\sigma} R^{\mathrm{x}}) \cdot \widetilde{R}_{\star}}, \qquad \varepsilon^{\sigma} = \varepsilon \cdot \frac{R^{\sigma}}{(\alpha^{\mathrm{x}} R^{\sigma} + \alpha^{\sigma} R^{\mathrm{x}}) \cdot \widetilde{R}_{\star}}, \tag{C.3.29}$$

we obtain $\|\mathtt{ffn}(\widetilde{X}_{i\star}; A)^{\top} - \mathtt{ffn}(\widetilde{X}_{i\star}; \widehat{A})^{\top}\|_{r,\infty} \leq \varepsilon$ for any $i \in [n]$, which implies

$$\Big\|\big(\mathtt{ffn}(\widetilde{X}_{i\star}; A)^{\top}\big)_{i\in[n]}^{\top} - \big(\mathtt{ffn}(\widetilde{X}_{i\star}; \widehat{A})^{\top}\big)_{i\in[n]}^{\top}\Big\|_{r,\infty} = \max_{i\in[n]} \big\|\mathtt{ffn}(\widetilde{X}_{i\star}; A)^{\top} - \mathtt{ffn}(\widetilde{X}_{i\star}; \widehat{A})^{\top}\big\|_{r,\infty} \leq \varepsilon.$$

To cover the empirical image class $\mathfrak{I}_{\widetilde{\mathcal{D}}_{n\star}}(\mathcal{F}_{\texttt{ffn}})$ at the resolution ε, it remains to cover the parameter spaces of A^{x} and A^{σ} at the resolutions ε^{x} and ε^{σ}, respectively, that is,

$$\begin{aligned}
&\log N\big(\mathfrak{I}_{\widetilde{\mathcal{D}}_{n\star}}(\mathcal{F}_{\texttt{ffn}}), \varepsilon, \|\cdot\|_{r,\infty}\big)\\
&\quad\le \log N\bigg(\Big\{(A^{\mathrm{x}})^\top \in \mathbb{R}^{d_\sigma\times d} : \big\|(A^{\mathrm{x}})^\top\big\|_{r,s} \le R^{\mathrm{x}}, \big\|(A^{\mathrm{x}})^\top\big\|_{r} \le \alpha^{\mathrm{x}}\Big\}, \varepsilon^{\mathrm{x}}, \|\cdot\|_{r,\infty}\bigg)\\
&\qquad\quad + \log N\bigg(\Big\{(A^{\sigma})^\top \in \mathbb{R}^{d\times d_\sigma} : \big\|(A^{\sigma})^\top\big\|_{r,s} \le R^{\sigma}, \big\|(A^{\sigma})^\top\big\|_{r} \le \alpha^{\sigma}\Big\}, \varepsilon^{\sigma}, \|\cdot\|_{r,\infty}\bigg)\\
&\quad\le \log N\bigg(\Big\{(A^{\mathrm{x}})^\top \in \mathbb{R}^{d_\sigma\times d} : \big\|(A^{\mathrm{x}})^\top\big\|_{r,s} \le R^{\mathrm{x}}\Big\}, \varepsilon^{\mathrm{x}}, \|\cdot\|_{r,\infty}\bigg)\\
&\qquad\quad + \log N\bigg(\Big\{(A^{\sigma})^\top \in \mathbb{R}^{d\times d_\sigma} : \big\|(A^{\sigma})^\top\big\|_{r,s} \le R^{\sigma}\Big\}, \varepsilon^{\sigma}, \|\cdot\|_{r,\infty}\bigg)\\
&\quad\le dd_\sigma\cdot\log\bigg(1+\frac{2R^{\mathrm{x}}}{\varepsilon^{\mathrm{x}}}\bigg) + dd_\sigma\cdot\log\bigg(1+\frac{2R^{\sigma}}{\varepsilon^{\sigma}}\bigg)\\
&\quad= 2dd_\sigma\cdot\log\bigg(1+\frac{2(\alpha^{\mathrm{x}}R^{\sigma}+\alpha^{\sigma}R^{\mathrm{x}})\cdot\widetilde{R}_\star}{\varepsilon}\bigg),
\end{aligned}$$

where the third inequality follows from Lemma C.3.10 and the equality follows from (C.3.29). Therefore, we conclude the proof of Lemma C.3.11. □

Empirical Image Class of MHA. In the sequel, we characterize the covering number of the empirical image class of MHA. In parallel to the parameter space $\mathfrak{W}^{(t)}$ defined in (C.3.9), we define

$$\begin{aligned}
\text{(C.3.30)}\qquad \mathfrak{W} = \Big\{&\big\{(W_i^{\mathrm{q}}, W_i^{\mathrm{k}}, W_i^{\mathrm{v}})\big\}_{i\in[h]} : (W_i^{\mathrm{q}}, W_i^{\mathrm{k}}, W_i^{\mathrm{v}}) \in \mathbb{R}^{d\times d_{\mathrm{p}}}\times\mathbb{R}^{d\times d_{\mathrm{p}}}\times\mathbb{R}^{d\times d},\\
&\big\|(W_i^{\mathrm{q}})^\top\big\|_r \le \omega_i^{\mathrm{q}}, \big\|(W_i^{\mathrm{k}})^\top\big\|_r \le \omega_i^{\mathrm{k}}, \big\|(W_i^{\mathrm{v}})^\top\big\|_r \le \omega_i^{\mathrm{v}},\\
&\big\|(W_i^{\mathrm{q}})^\top\big\|_{r,s} \le R_i^{\mathrm{q}}, \big\|(W_i^{\mathrm{k}})^\top\big\|_{r,s} \le R_i^{\mathrm{k}}, \big\|(W_i^{\mathrm{v}})^\top\big\|_{r,s} \le R_i^{\mathrm{v}}\Big\},
\end{aligned}$$

where $\omega_i^{\mathrm{q}}, \omega_i^{\mathrm{k}}, \omega_i^{\mathrm{v}}, R_i^{\mathrm{q}}, R_i^{\mathrm{k}}, R_i^{\mathrm{v}} > 0$. Correspondingly, we define the function class of MHA (with a skip connection) and the empirical image class of MHA as follows,

$$
\begin{aligned}
&\mathcal{F}_{\texttt{mha}} = \big\{\texttt{mha}(X; W) + X : W \in \mathfrak{W}\big\}, \\
&\mathfrak{I}_{\widetilde{\mathcal{D}}_n}(\mathcal{F}_{\texttt{mha}}) = \Big\{\big(f(\widetilde{X}_i)^\top\big)_{i\in[n]} \in \mathbb{R}^{d\times nL} : f \in \mathcal{F}_{\texttt{mha}}\Big\}.
\end{aligned} \tag{C.3.31}
$$

Here $\widetilde{\mathcal{D}}_n = \{\widetilde{X}_i\}_{i\in[n]}$ is the input set of MHA, where $\widetilde{X}_i \in \mathbb{R}^{L\times d}$. Recall that h is the head number of MHA. The following lemma characterizes the Lipschitz continuity of MHA with respect to the parameter in $W = \big\{(W_i^{\mathrm{q}}, W_i^{\mathrm{k}}, W_i^{\mathrm{v}})\big\}_{i\in[h]} \in \mathfrak{W}$.

Lemma C.3.12 (Parameter Lipschitz Continuity of MHA)**.** Let (r, s) be a conjugate pair. Suppose that $\widetilde{X} \in \mathbb{R}^{L\times d}$ satisfies $\|\widetilde{X}^\top\|_{r,\infty} \le \widetilde{R}$. Given any $W = \{(W_i^{\mathrm{q}}, W_i^{\mathrm{k}}, W_i^{\mathrm{v}})\}_{i\in[h]} \in \mathfrak{W}$, suppose that $\widehat{W} = \{(\widehat{W}_i^{\mathrm{q}}, \widehat{W}_i^{\mathrm{k}}, \widehat{W}_i^{\mathrm{v}})\}_{i\in[h]} \in \mathfrak{W}$ satisfies

$$
\big\|(W_i^{\mathrm{q}})^\top - (\widehat{W}_i^{\mathrm{q}})^\top\big\|_{r,s} \le \varepsilon_i^{\mathrm{q}}, \quad \big\|(W_i^{\mathrm{k}})^\top - (\widehat{W}_i^{\mathrm{k}})^\top\big\|_{r,s} \le \varepsilon_i^{\mathrm{k}}, \quad \big\|(W_i^{\mathrm{v}})^\top - (\widehat{W}_i^{\mathrm{v}})^\top\big\|_{r,s} \le \varepsilon_i^{\mathrm{v}}
$$

for any $i \in [h]$. Then, we have

$$
\big\|\texttt{mha}(\widetilde{X}; W)^\top - \texttt{mha}(\widetilde{X}; \widehat{W})^\top\big\|_{r,\infty} \le \widetilde{R} \cdot \sum_{i=1}^h \varepsilon_i^{\mathrm{v}} + \widetilde{R}^3 \cdot \sum_{i=1}^h (\omega_i^{\mathrm{q}} + \omega_i^{\mathrm{k}}) \cdot (\varepsilon_i^{\mathrm{q}} + \varepsilon_i^{\mathrm{k}}),
$$

where ω_i^{q}, ω_i^{k}, and ω_i^{v} are defined in (C.3.30).

Proof. See §C.6.2.1 for a detailed proof. □

The following lemma characterizes the covering number of the empirical image class $\mathfrak{I}_{\widetilde{\mathcal{D}}_n}(\mathcal{F}_{\texttt{mha}})$ defined in (C.3.31).

Lemma C.3.13 (Covering Number of MHA)**.** Let $\varepsilon > 0$. Suppose that the input set $\widetilde{\mathcal{D}}_n = \{\widetilde{X}_i\}_{i\in[n]} \subseteq \mathbb{R}^{L\times d}$ of MHA satisfies $\max_{i\in[n]} \|\widetilde{X}_i^\top\|_{r,\infty} \le \widetilde{R}$. Then, we have

$$\log N\big(\mathfrak{I}_{\widetilde{\mathcal{D}}_n}(\mathcal{F}_{\texttt{mha}}), \varepsilon, \|\cdot\|_{r,\infty}\big) \le (2+h)\cdot d^2 \cdot \log\Bigg(1 + \frac{2\widetilde{R}\cdot R_{\texttt{mha}}(\mathfrak{W})}{\varepsilon}\Bigg).$$

Here

$$R_{\texttt{mha}}(\mathfrak{W}) = \sum_{i=1}^h R_i^{\mathrm{v}} + \widetilde{R}^2 \cdot \sum_{i=1}^h (\omega_i^{\mathrm{q}} + \omega_i^{\mathrm{k}})(R_i^{\mathrm{q}} + R_i^{\mathrm{k}}), \tag{C.3.32}$$

where ω_i^{q}, ω_i^{k}, R_i^{q}, R_i^{k}, and R_i^{v} are defined in (C.3.30).

Proof. Throughout the following proof, we set $\varepsilon_i^{\mathrm{q}}, \varepsilon_i^{\mathrm{k}}, \varepsilon_i^{\mathrm{v}} > 0$ such that

$$\frac{R_i^{\mathrm{q}}}{\varepsilon_i^{\mathrm{q}}} = \frac{R_i^{\mathrm{k}}}{\varepsilon_i^{\mathrm{k}}} = \frac{R_i^{\mathrm{v}}}{\varepsilon_i^{\mathrm{v}}} = \frac{\widetilde{R}\cdot R_{\texttt{mha}}(\mathfrak{W})}{\varepsilon} \tag{C.3.33}$$

for any $i \in [h]$, where $R_{\texttt{mha}}(\mathfrak{W})$ is defined in (C.3.32). By Lemma C.3.12, we have for any $i \in [n]$ that

$$\begin{aligned}
&\big\|\texttt{mha}(\widetilde{X}_i; W)^\top + \widetilde{X}_i^\top - \texttt{mha}(\widetilde{X}_i; \widehat{W})^\top - \widetilde{X}_i^\top\big\|_{r,\infty} \\
&\quad \le \widetilde{R}\cdot \sum_{i=1}^h \varepsilon_i^{\mathrm{v}} + \widetilde{R}^3 \cdot \sum_{i=1}^h (\omega_i^{\mathrm{q}} + \omega_i^{\mathrm{k}}) \cdot (\varepsilon_i^{\mathrm{q}} + \varepsilon_i^{\mathrm{k}}) \\
&\quad = \widetilde{R}\cdot \sum_{i=1}^h \frac{R_i^{\mathrm{v}}\cdot\varepsilon}{\widetilde{R}\cdot R_{\texttt{mha}}(\mathfrak{W})} + \widetilde{R}^3 \cdot \sum_{i=1}^h (\omega_i^{\mathrm{q}} + \omega_i^{\mathrm{k}}) \cdot \Bigg(\frac{R_i^{\mathrm{q}}\cdot\varepsilon}{\widetilde{R}\cdot R_{\texttt{mha}}(\mathfrak{W})} + \frac{R_i^{\mathrm{k}}\cdot\varepsilon}{\widetilde{R}\cdot R_{\texttt{mha}}(\mathfrak{W})}\Bigg) \\
&\quad = \varepsilon \cdot \frac{\sum_{i=1}^h R_i^{\mathrm{v}} + \widetilde{R}^2 \cdot \sum_{i=1}^h (\omega_i^{\mathrm{q}} + \omega_i^{\mathrm{k}}) \cdot (R_i^{\mathrm{q}} + R_i^{\mathrm{k}})}{\widetilde{R}\cdot R_{\texttt{mha}}(\mathfrak{W})} = \varepsilon,
\end{aligned}$$

where the third line follows from (C.3.33) and the last line follows from the definition of $R_{\texttt{mha}}(\mathfrak{W})$ in (C.3.32). Hence, we have

$$\begin{aligned}
&\Big\| \big(\texttt{mha}(\widetilde{X}_i; W)^\top + \widetilde{X}_i^\top\big)_{i\in[n]}^\top - \big(\texttt{mha}(\widetilde{X}_i; \widehat{W})^\top + \widetilde{X}_i^\top\big)_{i\in[n]}^\top \Big\|_{r,\infty} \\
&\quad = \max_{i\in[n]} \big\| \texttt{mha}(\widetilde{X}_i; W)^\top + \widetilde{X}_i^\top - \texttt{mha}(\widetilde{X}_i; \widehat{W})^\top - \widetilde{X}_i^\top \big\|_{r,\infty} \le \varepsilon.
\end{aligned}$$

To cover the empirical image class $\mathfrak{I}_{\widetilde{\mathcal{D}}_n}(\mathcal{F}_{\texttt{mha}})$ at the resolution ε, it remains to cover the parameter spaces of W_i^{q}, W_i^{k}, and W_i^{v} at the resolutions $\varepsilon_i^{\mathrm{q}}$, $\varepsilon_i^{\mathrm{k}}$, and $\varepsilon_i^{\mathrm{v}}$, respectively, for any $i \in [n]$, that is,

$$\begin{aligned}
&\log N\Big(\mathfrak{I}_{\widetilde{\mathcal{D}}_n}(\mathcal{F}_{\texttt{mha}}), \varepsilon, \|\cdot\|_{r,\infty}\Big) \\
&\quad \le \sum_{i=1}^{h} \Bigg(\log N\bigg(\Big\{ (W_i^{\mathrm{q}})^\top \in \mathbb{R}^{d_{\mathrm{p}}\times d} : \big\|(W_i^{\mathrm{q}})^\top\big\|_{r,s} \le R_i^{\mathrm{q}}, \big\|(W_i^{\mathrm{q}})^\top\big\|_{r} \le \omega_i^{\mathrm{q}} \Big\}, \varepsilon_i^{\mathrm{q}}, \|\cdot\|_{r,\infty} \bigg) \\
&\qquad + \log N\bigg(\Big\{ (W_i^{\mathrm{k}})^\top \in \mathbb{R}^{d_{\mathrm{p}}\times d} : \big\|(W_i^{\mathrm{k}})^\top\big\|_{r,s} \le R_i^{\mathrm{k}}, \big\|(W_i^{\mathrm{k}})^\top\big\|_{r} \le \omega_i^{\mathrm{k}} \Big\}, \varepsilon_i^{\mathrm{k}}, \|\cdot\|_{r,\infty} \bigg) \\
&\qquad + \log N\bigg(\Big\{ (W_i^{\mathrm{v}})^\top \in \mathbb{R}^{d\times d} : \big\|(W_i^{\mathrm{v}})^\top\big\|_{r,s} \le R_i^{\mathrm{v}}, \big\|(W_i^{\mathrm{v}})^\top\big\|_{r} \le \omega_i^{\mathrm{v}} \Big\}, \varepsilon_i^{\mathrm{v}}, \|\cdot\|_{r,\infty} \bigg) \Bigg) \\
&\quad \le \sum_{i=1}^{h} \Bigg(\log N\bigg(\Big\{ (W_i^{\mathrm{q}})^\top \in \mathbb{R}^{d_{\mathrm{p}}\times d} : \big\|(W_i^{\mathrm{q}})^\top\big\|_{r,s} \le R_i^{\mathrm{q}} \Big\}, \varepsilon_i^{\mathrm{q}}, \|\cdot\|_{r,\infty} \bigg) \\
&\qquad + \log N\bigg(\Big\{ (W_i^{\mathrm{k}})^\top \in \mathbb{R}^{d_{\mathrm{p}}\times d} : \big\|(W_i^{\mathrm{k}})^\top\big\|_{r,s} \le R_i^{\mathrm{k}} \Big\}, \varepsilon_i^{\mathrm{k}}, \|\cdot\|_{r,\infty} \bigg) \\
&\qquad + \log N\bigg(\Big\{ (W_i^{\mathrm{v}})^\top \in \mathbb{R}^{d\times d} : \big\|(W_i^{\mathrm{v}})^\top\big\|_{r,s} \le R_i^{\mathrm{v}} \Big\}, \varepsilon_i^{\mathrm{v}}, \|\cdot\|_{r,\infty} \bigg) \Bigg) \\
&\quad \le d d_{\mathrm{p}} \cdot \sum_{i=1}^{h} \Bigg(\log\bigg(1 + \frac{2R_i^{\mathrm{q}}}{\varepsilon_i^{\mathrm{q}}}\bigg) + \log\bigg(1 + \frac{2R_i^{\mathrm{k}}}{\varepsilon_i^{\mathrm{k}}}\bigg) \Bigg) + d^2 \cdot \sum_{i=1}^{h} \log\bigg(1 + \frac{2R_i^{\mathrm{v}}}{\varepsilon_i^{\mathrm{v}}}\bigg) \\
&\quad = (2+h) \cdot d^2 \cdot \log\bigg(1 + \frac{2\widetilde{R} R_{\texttt{mha}}(\mathfrak{W})}{\varepsilon}\bigg),
\end{aligned}$$

where the third inequality follows from Lemma C.3.10 and the equality follows from (C.3.33) and the fact that $d = d_{\mathrm{p}} \cdot h$. Therefore, we conclude the proof of Lemma C.3.13. □

C.3.5.2. Propagation of Covering Numbers. Recall that $\mathfrak{W}$ is defined in (C.3.30). The following lemma characterizes the Lipschitz continuity of MHA in the input X.

Lemma C.3.14 (Input Lipschitz Continuity of MHA)**.** Let (r, s) be a conjugate pair. Suppose that $X \in \mathbb{R}^{L\times d}$ and $\widehat{X} \in \mathbb{R}^{L\times d}$ satisfy $\|X^\top\|_{r,\infty} \le \widetilde{R}$ and $\|\widehat{X}^\top\|_{r,\infty} \le \widetilde{R}$, respectively. Then for any $W = \{(W_i^{\mathrm{q}}, W_i^{\mathrm{k}}, W_i^{\mathrm{v}})\}_{i\in[n]} \in \mathfrak{W}$, we have

$$\big\|\mathtt{mha}(X; W)^\top - \mathtt{mha}(\widehat{X}; W)^\top\big\|_{r,\infty} \le \rho(\mathfrak{W}) \cdot \|X^\top - \widehat{X}^\top\|_{r,\infty}.$$

Here

$$\rho(\mathfrak{W}) = \sum_{i=1}^{h} \omega_i^{\mathrm{v}} + \widetilde{R}^2 \cdot \sum_{i=1}^{h} (\omega_i^{\mathrm{q}} + \omega_i^{\mathrm{k}})^2 \omega_i^{\mathrm{v}}, \tag{C.3.34}$$

where ω_i^{q}, ω_i^{k}, and ω_i^{v} are defined in (C.3.30).

Proof. See §C.6.2.2 for a detailed proof. □

Recall that the empirical image classes $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{ffn}}^{L,(t)})$, $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{mha}}^{L,(t+1)})$, and $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j^L)$ are defined in (C.3.7) and (C.3.17). Also, recall that the parameter space $\Theta^{(t)}$ is specified in Assumption C.3.3. The following lemma characterizes the propagation of the covering numbers of the empirical image classes of FFN and MHA through the T layers of the transformer architecture.

Lemma C.3.15 (Propagation of Covering Number)**.** Suppose that Assumption C.3.3 holds. For any $j \in [d_{\mathrm{y}}]$, we have

$$\log N\big(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j^L), \varepsilon, \|\cdot\|_{r,\infty}\big) \leq \sum_{t=0}^{T-2} \sup_{\{\theta^{(\tau)} \in \Theta^{(\tau)}\}_{0\leq\tau\leq t}} \log N\big(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{mha}}^{L,(t+1)}), \varepsilon_{\mathtt{mha}}^{(t)}, \|\cdot\|_{r,\infty}\big) + \sum_{t=0}^{T-1} \sup_{\{\theta^{(\tau)} \in \Theta^{(\tau)}\}_{0\leq\tau\leq t}} \log N\big(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{ffn}}^{L,(t)}), \varepsilon_{\mathtt{ffn}}^{(t)}, \|\cdot\|_{r,\infty}\big).$$

With the conventions $\prod_{\tau=T-1}^{T-1} \cdot \equiv 1$ and $\prod_{\tau=T}^{T-1} \cdot \equiv 1$, the covering resolution ε is defined as follows,

$$\varepsilon = \sum_{t=0}^{T-1}\bigg(\widetilde{\rho}^{(t)}\varepsilon_{\mathtt{ffn}}^{(t)} \cdot \prod_{\tau=t+1}^{T-1} \widetilde{\rho}^{(\tau)}\widetilde{\alpha}^{(\tau)}\bigg) + \sum_{t=0}^{T-2}\bigg(\varepsilon_{\mathtt{mha}}^{(t)} \cdot \prod_{\tau=t+1}^{T-1} \widetilde{\rho}^{(\tau)}\widetilde{\alpha}^{(\tau)}\bigg), \tag{C.3.35}$$

where $\widetilde{\alpha}^{(t)}$ and $\widetilde{\rho}^{(t)}$ are defined in (C.3.11) and (C.3.20), respectively.

Proof. Throughout the following proof, we fix the dataset $\mathcal{D}_n = \{(X_i, y_i)\}_{i\in[n]}$ and the parameters $\bar{\theta}$ and $\{\theta^{(t)} = (W^{(t)}, A^{(t)})\}_{0\leq t\leq T-1}$. By (C.3.4) and (C.3.5), the intermediate inputs $\{X_i^{(t)}\}_{i\in[n]}$ and $\{X_{i\star}^{(t)}\}_{i\in[n]}$ and the outputs $\{\widehat{y}_i = (\widehat{y}_{i,j})_{j\in[d_{\mathrm{y}}]}^\top\}_{i\in[n]}$ are fixed.

Perturbed Intermediate Inputs. For all $t = 0, \ldots, T-1$, we denote by $\mathfrak{N}_{\mathtt{ffn}}^{(t)}$ and $\mathfrak{N}_{\mathtt{mha}}^{(t+1)}$ the covering sets of the empirical image classes $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{ffn}}^{L,(t)})$ and $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{mha}}^{L,(t+1)})$ at the resolutions $\varepsilon_{\mathtt{ffn}}^{(t)}$ and $\varepsilon_{\mathtt{mha}}^{(t)}$ with respect to the matrix (r,∞)-norm, respectively. Starting from $((X_{i\star}^{(0)})^\top)_{i\in[n]} = ((\widetilde{X}_{i\star}^{(0)})^\top)_{i\in[n]} = (X_i^\top)_{i\in[n]}$, we construct the perturbed intermediate inputs in a recursive manner as follows,

(C.3.36)

$$\begin{aligned}
\big((\widetilde{X}_i^{(t)})^\top\big)_{i\in[n]} &\in \Big\{(\widetilde{X}_i^\top)_{i\in[n]} \in \mathfrak{N}^{(t)}_{\mathtt{ffn}} : \big\|\mathtt{ffn}(\widetilde{X}_{i\star}^{(t)}; A^{(t)})^\top - \widetilde{X}_i^\top\big\|_{r,\infty} \le \varepsilon^{(t)}_{\mathtt{ffn}}\Big\},\\
\big((\widetilde{X}_{i\star}^{(t+1)})^\top\big)_{i\in[n]} &\in \Big\{(\widetilde{X}_{i\star}^\top)_{i\in[n]} \in \mathfrak{N}^{(t+1)}_{\mathtt{mha}} : \big\|\mathtt{mha}(\widetilde{X}_i^{(t)}; W^{(t)})^\top + (\widetilde{X}_i^{(t)})^\top - \widetilde{X}_{i\star}^\top\big\|_{r,\infty} \le \varepsilon^{(t)}_{\mathtt{mha}}\Big\},\\
\big((\widetilde{X}_i^{(T-1)})^\top\big)_{i\in[n]} &\in \Big\{(\widetilde{X}_i^\top)_{i\in[n]} \in \mathfrak{N}^{(T-1)}_{\mathtt{ffn}} : \big\|\mathtt{ffn}(\widetilde{X}_{i\star}^{(T-1)}; A^{(T-1)})^\top - \widetilde{X}_i^\top\big\|_{r,\infty} \le \varepsilon^{(T-1)}_{\mathtt{ffn}}\Big\},
\end{aligned}$$

where $t = 0, \ldots, T-2$. For any $i \in [n]$ and any $j \in [d_{\mathrm{y}}]$, let

(C.3.37)

$$\begin{aligned}
\widetilde{X}_{i\star}^{(T)} &= \mathtt{mha}(\widetilde{X}_i^{(T-1)}; W^{(T-1)}) + \widetilde{X}_i^{(T-1)},\\
\widetilde{y}_{i,j} &= \overline{\mathtt{agg}}_{\bar\theta,j}(\widetilde{X}_{i\star}^{(T)}),
\end{aligned}$$

which implies

$$|\widehat{y}_{i,j} - \widetilde{y}_{i,j}| = \big|\overline{\mathtt{agg}}_{\bar\theta,j}(X_{i\star}^{(T)}) - \overline{\mathtt{agg}}_{\bar\theta,j}(\widetilde{X}_{i\star}^{(T)})\big| \le \big\|(X_{i\star}^{(T)})^\top - (\widetilde{X}_{i\star}^{(T)})^\top\big\|_{r,\infty},$$

where the inequality follows from Assumption C.3.2. Hence, to cover the empirical image class $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j^L)$ at the resolution ε, it remains to cover the empirical image class $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{mha}}^{L,(T)})$ at the resolution ε.

Propagation of Covering Resolutions. For the recursive constructions in (C.3.36), it holds for any $i \in [n]$ that

$$
\begin{aligned}
&\big\|(X_i^{(t)})^\top - (\widetilde{X}_i^{(t)})^\top\big\|_{r,\infty} \\
&\quad = \big\|\mathtt{ffn}(X_{i\star}^{(t)}; A^{(t)})^\top - (\widetilde{X}_i^{(t)})^\top\big\|_{r,\infty} \\
&\quad \leq \big\|\mathtt{ffn}(X_{i\star}^{(t)}; A^{(t)})^\top - \mathtt{ffn}(\widetilde{X}_{i\star}^{(t)}; A^{(t)})^\top\big\|_{r,\infty} + \big\|\mathtt{ffn}(\widetilde{X}_{i\star}^{(t)}; A^{(t)})^\top - (\widetilde{X}_i^{(t)})^\top\big\|_{r,\infty} \\
&\quad \leq \big\|\mathtt{ffn}(X_{i\star}^{(t)}; A^{(t)})^\top - \mathtt{ffn}(\widetilde{X}_{i\star}^{(t)}; A^{(t)})^\top\big\|_{r,\infty} + \varepsilon_{\mathtt{ffn}}^{(t)},
\end{aligned}
\tag{C.3.38}
$$

where the first line follows from (C.3.4) and the last line follows from the definition of $\widetilde{X}_i^{(t)}$ in (C.3.36). For the first term on the right-hand side of (C.3.38), it holds for all $t = 0, \ldots, T-1$ and any $i \in [n]$ that

$$
\begin{aligned}
&\big\|\mathtt{ffn}(X_{i\star}^{(t)}; A^{(t)})^\top - \mathtt{ffn}(\widetilde{X}_{i\star}^{(t)}; A^{(t)})^\top\big\|_{r,\infty} \\
&\quad = \big\|(A^{\sigma,(t)})^\top \mathtt{ReLU}(X_{i\star}^{(t)} A^{\mathrm{x},(t)})^\top + (X_{i\star}^{(t)})^\top - (A^{\sigma,(t)})^\top \mathtt{ReLU}(\widetilde{X}_{i\star}^{(t)} A^{\mathrm{x},(t)})^\top - (\widetilde{X}_{i\star}^{(t)})^\top\big\|_{r,\infty} \\
&\quad \leq \big\|(A^{\sigma,(t)})^\top\big\|_r \cdot \big\|\mathtt{ReLU}(X_{i\star}^{(t)} A^{\mathrm{x},(t)})^\top - \mathtt{ReLU}(\widetilde{X}_{i\star}^{(t)} A^{\mathrm{x},(t)})^\top\big\|_{r,\infty} + \big\|(X_{i\star}^{(t)})^\top - (\widetilde{X}_{i\star}^{(t)})^\top\big\|_{r,\infty} \\
&\quad \leq \Big(1 + \big\|(A^{\mathrm{x},(t)})^\top\big\|_r \cdot \big\|(A^{\sigma,(t)})^\top\big\|_r\Big) \cdot \big\|(X_{i\star}^{(t)})^\top - (\widetilde{X}_{i\star}^{(t)})^\top\big\|_{r,\infty} \\
&\quad \leq \widetilde{\alpha}^{(t)} \cdot \big\|(X_{i\star}^{(t)})^\top - (\widetilde{X}_{i\star}^{(t)})^\top\big\|_{r,\infty},
\end{aligned}
\tag{C.3.39}
$$

where the third and fourth lines follow from Lemma C.7.3 and the last line follows from the requirement in Assumption C.3.3 and the definition of $\widetilde{\alpha}^{(t)}$ in (C.3.11). Hence, it holds for all $t = 0, \ldots, T-2$ and any $i \in [n]$ that

$$\begin{aligned}
\big\|(X_{i\star}^{(t+1)})^\top - (\widetilde{X}_{i\star}^{(t+1)})^\top\big\|_{r,\infty} &\le \big\|\texttt{mha}(X_i^{(t)}; A^{(t)})^\top + (X_i^{(t)})^\top - \texttt{mha}(\widetilde{X}_i^{(t)}; A^{(t)})^\top - (\widetilde{X}_i^{(t)})^\top\big\|_{r,\infty} \\
&\quad + \big\|\texttt{mha}(\widetilde{X}_i^{(t)}; A^{(t)})^\top + (\widetilde{X}_i^{(t)})^\top - (\widetilde{X}_{i\star}^{(t+1)})^\top\big\|_{r,\infty} \\
&\le \big(\rho(\mathfrak{W}^{(t)}) + 1\big) \cdot \big\|(X_i^{(t)})^\top - (\widetilde{X}_i^{(t)})^\top\big\|_{r,\infty} + \varepsilon_{\texttt{mha}}^{(t)} \\
&\le \widetilde{\rho}^{(t)} \cdot \big\|(X_i^{(t)})^\top - (\widetilde{X}_i^{(t)})^\top\big\|_{r,\infty} + \varepsilon_{\texttt{mha}}^{(t)}, \qquad \text{(C.3.40)}
\end{aligned}$$

where the second inequality follows from Lemma C.3.14 and the definition of $\widetilde{X}_{i\star}^{(t+1)}$ in (C.3.36) and the last inequality follows from Lemma C.7.4. Taking (C.3.39) into (C.3.38) and (C.3.38) into (C.3.40), we obtain for any $i \in [n]$ and $t = 0, \ldots, T-2$ that

$$\big\|(X_{i\star}^{(t+1)})^\top - (\widetilde{X}_{i\star}^{(t+1)})^\top\big\|_{r,\infty} \le \widetilde{\rho}^{(t)}\widetilde{\alpha}^{(t)} \cdot \big\|(X_{i\star}^{(t)})^\top - (\widetilde{X}_{i\star}^{(t)})^\top\|_{r,\infty} + \widetilde{\rho}^{(t)}\varepsilon_{\texttt{ffn}}^{(t)} + \varepsilon_{\texttt{mha}}^{(t)}. \qquad \text{(C.3.41)}$$

Recursively applying (C.3.41), we have

$$\big\|(X_{i\star}^{(T-1)})^\top - (\widetilde{X}_{i\star}^{(T-1)})^\top\big\|_{r,\infty} \le \sum_{t=0}^{T-2}\Big[(\widetilde{\rho}^{(t)}\varepsilon_{\texttt{ffn}}^{(t)} + \varepsilon_{\texttt{mha}}^{(t)}) \cdot \prod_{\tau=t+1}^{T-2} \widetilde{\rho}^{(\tau)}\widetilde{\alpha}^{(\tau)}\Big],$$

which implies

$$
\begin{aligned}
&\big\|(X_{i\star}^{(T)})^\top - (\widetilde{X}_{i\star}^{(T)})^\top\big\|_{r,\infty} \\
&\quad = \big\|\mathtt{mha}(X_i^{(T-1)}; W^{(T-1)}) + X_i^{(T-1)} - \mathtt{mha}(\widetilde{X}_i^{(T-1)}; W^{(T-1)}) - \widetilde{X}_i^{(T-1)}\big\|_{r,\infty} \\
&\quad \le \widetilde{\rho}^{(T-1)} \cdot \big\|(X_i^{(T-1)})^\top - (\widetilde{X}_i^{(T-1)})^\top\big\|_{r,\infty} \\
&\quad \le \widetilde{\rho}^{(T-1)}\widetilde{\alpha}^{(T-1)} \cdot \big\|(X_{i\star}^{(t)})^\top - (\widetilde{X}_{i\star}^{(t)})^\top\big\|_{r,\infty} + \widetilde{\rho}^{(T-1)}\varepsilon_{\mathtt{ffn}}^{(T-1)} \\
&\quad \le \sum_{t=0}^{T-2}\Big[(\widetilde{\rho}^{(t)}\varepsilon_{\mathtt{ffn}}^{(t)} + \varepsilon_{\mathtt{mha}}^{(t)}) \cdot \prod_{\tau=t+1}^{T-1}\widetilde{\rho}^{(\tau)}\widetilde{\alpha}^{(\tau)}\Big] + \widetilde{\rho}^{(T-1)}\varepsilon_{\mathtt{ffn}}^{(T-1)} = \varepsilon,
\end{aligned}
$$

where the second line follows from (C.3.37), the third line follows from (C.3.40), the fourth line follows from (C.3.38) and (C.3.39), and the last line follows from the definition of ε in (C.3.35). To cover the empirical image class $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{mha}}^{L,(T)})$ at the resolution ε, it suffices to cover (i) the empirical image class $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{mha}}^{L,(t+1)})$ at the resolution $\varepsilon_{\mathtt{mha}}^{(t)}$ for all $t = 0, \ldots, T-2$, and (ii) the empirical image class $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{ffn}}^{L,(t)})$ at the resolution $\varepsilon_{\mathtt{ffn}}^{(t)}$ for all $t = 0, \ldots, T-1$. Therefore, we conclude the proof of Lemma C.3.15. □

C.3.5.3. Proof of Lemma C.3.9.

Proof. By Lemma C.3.15, we have

$$
\begin{aligned}
\log N\big(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j^L), \varepsilon, \|\cdot\|_{r,\infty}\big) \le \sum_{t=0}^{T-2} \sup_{\{\theta^{(\tau)}\in\Theta^{(\tau)}\}_{0\le\tau\le t}} \log N\big(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{mha}}^{L,(t+1)}), \varepsilon_{\mathtt{mha}}^{(t)}, \|\cdot\|_{r,\infty}\big) \\
+ \sum_{t=0}^{T-1} \sup_{\{\theta^{(\tau)}\in\Theta^{(\tau)}\}_{0\le\tau\le t}} \log N\big(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_{\mathtt{ffn}}^{L,(t)}), \varepsilon_{\mathtt{ffn}}^{(t)}, \|\cdot\|_{r,\infty}\big),
\end{aligned}
$$

where ε is defined in (C.3.35). In what follows, we set

$$\varepsilon^{(t)}_{\texttt{ffn}} = \varepsilon^{(t)} \cdot \frac{\alpha^{\mathrm{x},(t)} R^{\sigma,(t)} + \alpha^{\sigma,(t)} R^{\mathrm{x},(t)}}{\widetilde{\omega}^{\mathrm{v},(t)} \widetilde{\alpha}^{(t)}}, \qquad \varepsilon^{(t)}_{\texttt{mha}} = \varepsilon^{(t)} \cdot \frac{R^{(t)}_{\texttt{mha}}}{\widetilde{\omega}^{\mathrm{v},(t)}}. \tag{C.3.42}$$

By Lemma C.7.6, the intermediate inputs $\{X_i^{(t)}\}_{i\in[n]}$ and $\{X_{i\star}^{(t)}\}_{i\in[n]}$ satisfy

$$\max_{i\in[n]} \big\| (X_i^{(t)})^\top \big\|_{r,\infty} \le \widetilde{\alpha}^{(t)} R^{(t)}, \qquad \max_{i\in[n]} \big\| (X_{i\star}^{(t)})^\top \big\|_{r,\infty} \le R^{(t)},$$

where $R^{(t)}$ is defined in (C.3.18). By Lemma C.3.11, it holds for all $t = 0, \ldots, T-1$ that

$$\begin{aligned} &\log N\big(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}^{L,(t)}_{\texttt{ffn}}), \varepsilon^{(t)}_{\texttt{ffn}}, \|\cdot\|_{r,\infty}\big) \\ &\quad \le 2dd_\sigma \cdot \log\bigg(1 + \frac{2(\alpha^{\mathrm{x},(t)} R^{\sigma,(t)} + \alpha^{\sigma,(t)} R^{\mathrm{x},(t)}) \cdot R^{(t)}}{\varepsilon^{(t)}_{\texttt{ffn}}}\bigg) \le 2D^2 \cdot \log\bigg(1 + \frac{2R^{(t+1)}}{\varepsilon^{(t)}}\bigg), \end{aligned} \tag{C.3.43}$$

where the second inequality follows from (C.3.42), the fact that $D = \max\{d, d_{\mathrm{p}}, d_\sigma, d_{\mathrm{y}}\}$, and the definition of $R^{(t)}$ in (C.3.18). By Lemmas C.3.13 and C.7.4, it holds for all $t = 0, \ldots, T-2$ that

$$\begin{aligned} &\log N\big(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}^{L,(t+1)}_{\texttt{mha}}), \varepsilon^{(t)}_{\texttt{mha}}, \|\cdot\|_{r,\infty}\big) \\ &\quad \le (2+h)d^2 \cdot \log\bigg(1 + \frac{2\widetilde{\alpha}^{(t)} R^{(t)} \cdot R^{(t)}_{\texttt{mha}}}{\varepsilon^{(t)}_{\texttt{mha}}}\bigg) \le (2+h)D^2 \cdot \log\bigg(1 + \frac{2R^{(t+1)}}{\varepsilon^{(t)}}\bigg), \end{aligned} \tag{C.3.44}$$

where the second inequality follows from (C.3.42), the fact that $D = \max\{d, d_{\mathrm{p}}, d_\sigma, d_{\mathrm{y}}\}$, and the definition of $R^{(t)}$ in (C.3.18). It remains to choose the resolutions $\{\varepsilon^{(t)}\}_{0\le t\le T-1}$

that satisfy (C.3.35), which is

$$
\begin{aligned}
\varepsilon = &\sum_{t=0}^{T-2}\biggl(\varepsilon^{(t)} \cdot \frac{R_{\mathtt{mha}}^{(t)}}{\widetilde{\omega}^{\mathrm{v},(t)}} \cdot \prod_{\tau=t+1}^{T-1} \widetilde{\rho}^{(\tau)}\widetilde{\alpha}^{(\tau)}\biggr) \\
&+ \sum_{t=0}^{T-1}\biggl(\varepsilon^{(t)} \cdot \frac{\alpha^{\mathrm{x},(t)}R^{\sigma,(t)} + \alpha^{\sigma,(t)}R^{\mathrm{x},(t)}}{\widetilde{\alpha}^{(t)}} \cdot \prod_{\tau=t+1}^{T-1} \widetilde{\rho}^{(\tau)}\widetilde{\alpha}^{(\tau)}\biggr). \qquad \text{(C.3.45)}
\end{aligned}
$$

Recall that $R_{\mathtt{trans}}$ is defined in (C.3.19). For all $t = 0, \ldots, T-1$, we set

$$
\varepsilon^{(t)} = \frac{\varepsilon}{R_{\mathtt{trans}} \cdot \prod_{\tau=t+1}^{T-1} \widetilde{\omega}^{\mathrm{v},(\tau)}\widetilde{\alpha}^{(\tau)}}, \qquad \text{(C.3.46)}
$$

which satisfies (C.3.45). Note that, by the definition of $R^{(t)}$ in (C.3.18), it holds that $R^{(t+1)} \cdot \prod_{\tau=t+1}^{T-1} \widetilde{\omega}^{\mathrm{v},(\tau)}\widetilde{\alpha}^{(\tau)} = R^{(T)}$ for all $t = 0, \ldots, T-1$. Combining Lemma C.3.15, (C.3.43), (C.3.44), and the choices of $\{\varepsilon^{(t)}\}_{0 \le t \le T-1}$ in (C.3.46), we obtain

$$
\begin{aligned}
\log N\bigl(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j^L), \varepsilon, \|\cdot\|_{r,\infty}\bigr) &\le \bigl[(4+h)T - h - 2\bigr]D^2 \cdot \log\biggl(1 + \frac{2R^{(T)}R_{\mathtt{trans}}}{\varepsilon}\biggr) \\
&\le (4+h)D^2 T \cdot \log\biggl(1 + \frac{2R^{(T)}R_{\mathtt{trans}}}{\varepsilon}\biggr).
\end{aligned}
$$

Therefore, we conclude the proof of Lemma C.3.9. □

C.4. Optimization Error Analysis

Proof of Proposition 3.5.6. Let $\widehat{\mathcal{L}}(f_\theta) = \widehat{\mathbb{E}}[\mathcal{L}((X,y), f_\theta)]$. By (3.5.12), it holds for the stationary point $\widehat{\theta}$ that,

$$
0 \le \bigl\langle \nabla_\theta \widehat{\mathcal{L}}(f_{\widehat{\theta}}), \theta - \widehat{\theta}\bigr\rangle = \widehat{\mathbb{E}}\Bigl[\nabla_f \mathcal{L}\bigl((X,y), f_{\widehat{\theta}}\bigr)\nabla_\theta f_{\widehat{\theta}}(X)^\top(\theta - \widehat{\theta})\Bigr].
$$

Since the objective function $\mathcal{L}((X,y),f) = \|y - f(X)\|_2^2$ is convex with respect to $f(X)$, we have

$$0 \le \widehat{\mathbb{E}}\Big[\nabla_f \mathcal{L}\big((X,y), f_{\theta^*}\big)^\top (f - f_{\theta^*})(X)\Big], \tag{C.4.1}$$

where $\theta^* = \operatorname{argmin}_{\theta\in\Theta} \widehat{\mathcal{L}}(f_\theta)$. By definition of the objective function $\mathcal{L}((X,y),f)$, we have

$$\Big\|\nabla_f \mathcal{L}\big((X,y), f_{\widehat{\theta}}\big)\Big\|_2 = 2\big\|y - f_{\widehat{\theta}}(X)\big\|_2 \le 2\|y\|_2 + 2\big\|f_{\widehat{\theta}}(X)\big\|_2 \le 2, \tag{C.4.2}$$

where the last inequality follows from Assumption 3.5.2 and that the aggregation layer $\mathtt{agg}_{\theta_0} : \mathbb{R}^{d_{\mathrm{p}}} \to \mathbb{R}^{d_{\mathrm{y}}}$ outputs within $\mathfrak{Y}$. For any $\theta \in \Theta$, it holds that,

$$\begin{aligned}
&\widehat{\mathcal{L}}(f_{\widehat{\theta}}) - \widehat{\mathcal{L}}(f_{\theta^*}) \\
&\le \widehat{\mathbb{E}}\Big[\nabla_f \mathcal{L}\big((X,y), f_{\widehat{\theta}}\big)^\top (f_{\widehat{\theta}} - f_{\theta^*})(X)\Big] \\
&\le \widehat{\mathbb{E}}\Big[\nabla_f \mathcal{L}\big((X,y), f_{\widehat{\theta}}\big)^\top (f_{\widehat{\theta}} - f_{\theta^*})(X)\Big] + \widehat{\mathbb{E}}\Big[\nabla_f \mathcal{L}\big((X,y), f_{\widehat{\theta}}\big)^\top \nabla_\theta f_{\widehat{\theta}}(x)^\top (\theta - \widehat{\theta})\Big] \\
&= \widehat{\mathbb{E}}\Big[\nabla_f \mathcal{L}\big((X,y), f_{\widehat{\theta}}\big)^\top \big(f_{\widehat{\theta}}(X) + \nabla_\theta f_{\widehat{\theta}}(X)^\top(\theta - \widehat{\theta}) - f_{\theta^*}(X)\big)\Big] \\
&\le \widehat{\mathbb{E}}\Big[\Big\|\nabla_f \mathcal{L}\big((X,y), f_{\widehat{\theta}}\big)\Big\|_2 \cdot \big\|f_{\widehat{\theta}}(X) + \nabla_\theta f_{\widehat{\theta}}(X)^\top(\theta - \widehat{\theta}) - f_{\theta^*}(X)\big\|_2\Big] \\
&\le 2 \cdot \big\|f_{\widehat{\theta}}(X) + \nabla_\theta f_{\widehat{\theta}}(X)^\top(\theta - \widehat{\theta}) - f_{\theta^*}(X)\big\|_2, \qquad \text{(C.4.3)}
\end{aligned}$$

where the second line follows from (C.4.1), the fourth line follows from the Cauchy-Schwartz inequality, and the last line follows from (C.4.2). Since (C.4.3) holds for any $\theta \in \Theta$, we have

$$\widehat{\mathcal{L}}(f_{\widehat{\theta}}) - \widehat{\mathcal{L}}(f_{\theta^*}) \le 2 \cdot \min_{\theta\in\Theta} \widehat{\mathbb{E}}\Big[\big\|f_{\widehat{\theta}}(X) + \nabla_\theta f_{\widehat{\theta}}(X)^\top(\theta - \widehat{\theta}) - f_{\theta^*}(X)\big\|_2\Big].$$

Therefore, we conclude the proof of Proposition 3.5.6. □

C.5. Approximation Error Analysis

C.5.1. Latent-to-Value RKHS

In what follows, we cast the function class $\mathcal{G}_i^\dagger$ defined in (3.5.10) as the RKHS $\mathcal{H}_{\mathtt{LTV}}$, which plays a key role in our subsequent analysis of the approximation error. Recall that the latent-to-value mapping $\psi(z; \mathtt{msk})$ is defined in (3.5.7), which induces the kernel function $\mathfrak{K}_{\mathtt{LTV}}(z, z'; \mathtt{msk}) = \psi(z; \mathtt{msk})^\top \psi(z'; \mathtt{msk})$ and the following RKHS,

$$\mathcal{H}_{\mathtt{LTV}} = \biggl\{ g_\alpha(z; \mathtt{msk}) = \int \alpha(z') \mathfrak{K}_{\mathtt{LTV}}(z', z; \mathtt{msk}) \mathrm{d}z' : \bigl\| g_\alpha(\cdot; \mathtt{msk}) \bigr\|_{\mathcal{H}_{\mathtt{LTV}}} < \infty \biggr\}, \tag{C.5.1}$$

which is equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\mathtt{LTV}}}$. By the definition of the kernel function $\mathfrak{K}_{\mathtt{LTV}}(\cdot, \cdot; \mathtt{msk})$, we have for any $g_\alpha(\cdot; \mathtt{msk}) \in \mathcal{H}_{\mathtt{LTV}}$ that

$$\begin{aligned} g_\alpha(z; \mathtt{msk}) &= \int \alpha(z') \mathfrak{K}_{\mathtt{LTV}}(z', z; \mathtt{msk}) \mathrm{d}z' \\ &= \underbrace{\biggl(\int \alpha(z') \psi(z'; \mathtt{msk}) \mathrm{d}z' \biggr)}_{w_\alpha \in \mathbb{R}^d}^\top \psi(z; \mathtt{msk}) = w_\alpha^\top \psi(z; \mathtt{msk}). \end{aligned} \tag{C.5.2}$$

Here w_α corresponds to the parameter vector $w_i \in \mathbb{R}^d$ in the function class $\mathcal{G}_i^\dagger$. On the other hand, we have

$$
\begin{aligned}
\big\|g_\alpha(\cdot;\texttt{msk})\big\|_{\mathcal{H}_{\texttt{LTV}}}^2 &= \big\langle g_\alpha(\cdot;\texttt{msk}), g_\alpha(\cdot;\texttt{msk})\big\rangle_{\mathcal{H}_{\texttt{LTV}}} \\
&= \int \alpha(z')\mathfrak{K}_{\texttt{LTV}}(z', z;\texttt{msk})\alpha(z)\mathrm{d}z\mathrm{d}z' \\
&= \bigg(\int \alpha(z')\psi(z';\texttt{msk})\mathrm{d}z'\bigg)^\top \bigg(\int \alpha(z)\psi(z;\texttt{msk})\mathrm{d}z\bigg) = \|w_\alpha\|_2^2, \qquad \text{(C.5.3)}
\end{aligned}
$$

where the third equality follows from the definition of $\mathfrak{K}_{\texttt{LTV}}(\cdot,\cdot;\texttt{msk})$ and the last equality follows from the definition of w_α in (C.5.2). Combining (C.5.2), (C.5.3), and the definition of $\mathcal{H}_{\texttt{LTV}}$ in (C.5.1), we have

$$
\mathcal{H}_{\texttt{LTV}} = \big\{w_\alpha^\top \psi(z;\texttt{msk}) : w_\alpha \in \mathbb{R}^d, \|w_\alpha\|_2 < \infty\big\} = \mathcal{G}_i^\dagger.
$$

Thus, the function class $\mathcal{G}_i^\dagger$, which correspondes to the i-th entry of the function class $\mathcal{G}$, is the RKHS $\mathcal{H}_{\texttt{LTV}}$. Here the function class $\mathcal{G}^\dagger$ is defined in (3.5.9), which contains the latent-to-target function $g_W^\dagger(z;\texttt{msk}) = W^\top\psi(z;\texttt{msk})$ within the reweighted CME attention $f_W^\dagger(X;\texttt{msk})$ defined in (3.5.6).

C.5.2. Supervised Learning

Proof of Theorem 3.5.5. Suppose $f_\theta \in \mathcal{F}_{\texttt{attn}}$ and $\epsilon_{\texttt{attn}} \in [0, +\infty)$ satisfy (3.5.11). By the definition of the approximation error $\mathcal{E}_{\text{approx}}$ in (3.5.3), we have

$$\begin{aligned}
\mathcal{E}_{\text{approx}} &= \min_{f \in \mathcal{F}_{\texttt{attn}}} \mathbb{E}\Big[\mathcal{L}\big((X, y), f\big)\Big] - \mathbb{E}\Big[\mathcal{L}\big((X, y), f^*\big)\Big] \\
&\leq \mathbb{E}\Big[\mathcal{L}\big((X, y), f_\theta\big)\Big] - \mathbb{E}\Big[\mathcal{L}\big((X, y), f^*\big)\Big] \\
&= \mathbb{E}\Big[\big\|f_\theta(X; \texttt{msk}) - f^*(X)\big\|_2^2\Big] \\
&\leq 2\mathbb{E}\Big[\big\|f_\theta(X; \texttt{msk}) - f_W^\dagger(X; \texttt{msk})\big\|_2^2\Big] + 2\mathbb{E}\Big[\big\|f_W^\dagger(X; \texttt{msk}) - f^*(X)\big\|_2^2\Big] \\
&\leq 2\epsilon_{\texttt{attn}}^2 + 2\mathbb{E}\Big[\big\|f_W^\dagger(X; \texttt{msk}) - f^*(X)\big\|_2^2\Big], \qquad \text{(C.5.4)}
\end{aligned}$$

where the second line follows from the fact that $f_\theta \in \mathcal{F}_{\texttt{attn}}$, the third line follows from the fact that $\mathcal{L}((X, y), f) = \|y - f(X)\|_2^2$ and the definition of the regression function $f^*(X) = \mathbb{E}[y \,|\, X]$, and the last line follows from (3.5.11) and the definition of $f_W^\dagger(X; \texttt{msk})$ in (3.5.6).

In what follows, we characterize the gap between the regression function $f^*(X)$ and the reweighted CME attention in $f_W^\dagger(X; \texttt{msk})$, which is used as a surrogate function for approximating $f^*(X)$. By (3.5.5), we have

$$\begin{aligned}
f^*(X) &= \mathbb{E}_{z \,|\, X}\big[g^*(z; \texttt{msk})\big] \\
&= \mathbb{E}_{z \,|\, X}\big[g_W^\dagger(z; \texttt{msk})\big] + \mathbb{E}_{z \,|\, X}\big[g^*(z; \texttt{msk}) - g_W^\dagger(z; \texttt{msk})\big] \\
&= f_W^\dagger(X; \texttt{msk}) + \mathbb{E}_{z \,|\, X}\big[g^*(z; \texttt{msk}) - g_W^\dagger(z; \texttt{msk})\big],
\end{aligned}$$

where the last line follows from (3.5.8). Hence, it holds for $f_W^\dagger(X;\texttt{msk})$ that

$$\mathbb{E}\Big[\big\|f^*(X) - f_W^\dagger(X;\texttt{msk})\big\|_2^2\Big] = \mathbb{E}\Big[\Big\|\mathbb{E}_{z\,|\,X}\big[g^*(z;\texttt{msk}) - g_W^\dagger(z;\texttt{msk})\big]\Big\|_2^2\Big]. \tag{C.5.5}$$

By Assumption 3.5.4, we have

$$\begin{aligned}\Big\|\mathbb{E}_{z\,|\,X}\big[g^*(z;\texttt{msk}) - g_W^\dagger(z;\texttt{msk})\big]\Big\|_2^2 &= \sum_{i=1}^{d_{\mathrm{y}}} \mathbb{E}_{z\,|\,X}\big[g_i^*(z;\texttt{msk}) - g_{W,i}^\dagger(z;\texttt{msk})\big]^2 \\ &\le \sum_{i=1}^{d_{\mathrm{y}}} \big\|g_i^*(\cdot;\texttt{msk}) - g_{W,i}^\dagger(\cdot;\texttt{msk})\big\|_\infty^2 \le \epsilon_g^2(\texttt{msk}),\end{aligned} \tag{C.5.6}$$

where the ℓ_∞-norm is taken over the latent variable z. Taking (C.5.6) into (C.5.5), we obtain

$$\mathbb{E}\Big[\big\|f^*(X) - f_W^\dagger(X;\texttt{msk})\big\|_2^2\Big] \le \epsilon_g^2(\texttt{msk}). \tag{C.5.7}$$

Taking (C.5.7) into (C.5.4), we obtain

$$\mathcal{E}_{\text{approx}} \le 2\epsilon_{\texttt{attn}}^2 + 2\epsilon_g^2(\texttt{msk}),$$

which concludes the proof of Theorem 3.5.5. □

C.5.3. Self-Supervised Learning

Proof of Theorem 3.6.3. Suppose that $f_{\texttt{DS}}(\overline{X};\texttt{msk}_{\texttt{DS}})$ attains the infimum on the right-hand side of (3.6.11). Recall that $B = W_{\texttt{DS}}^\top(W_{\texttt{SSL}}W_{\texttt{SSL}}^\top)^{-1}W_{\texttt{SSL}}$ is defined in (3.6.10).

We define a surrogate function as follows,

$$\widetilde{f}_{\texttt{PT}}(\overline{X};\texttt{msk}_{\texttt{PT}}) = B\widehat{f}_{\texttt{PT}}(\overline{X};\texttt{msk}_{\texttt{PT}}).$$

Here $\widehat{f}_{\texttt{PT}}(\overline{X};\texttt{msk}_{\texttt{PT}})$ is the attention neural network obtained from the pretraining process. Recall that the regression function $f^*_{\texttt{DS}}(\overline{X})$ for the downstream task is defined in (3.6.4) and $f^\dagger_{W_{\texttt{DS}}}(\overline{X};\texttt{msk}_{\texttt{DS}})$ is defined in (3.6.5). For the approximation error $\mathcal{E}_{\text{approx}}$ defined in (3.6.13), we have

(C.5.8)

$$\begin{aligned}
\mathcal{E}_{\text{approx}} &\le \mathbb{E}\Big[\mathcal{L}\big((\overline{X}, y_{\texttt{DS}}), f_{\texttt{DS}}\big)\Big] - \mathbb{E}\Big[\mathcal{L}\big((\overline{X}, y_{\texttt{DS}}), f^*_{\texttt{DS}}\big)\Big] \\
&= \mathbb{E}\Big[\big\|f_{\texttt{DS}}(\overline{X};\texttt{msk}_{\texttt{DS}}) - f^*_{\texttt{DS}}(\overline{X})\big\|_2^2\Big] \\
&= \mathbb{E}\Big[\big\|f_{\texttt{DS}}(\overline{X};\texttt{msk}_{\texttt{DS}}) - \widetilde{f}_{\texttt{PT}}(\overline{X};\texttt{msk}_{\texttt{PT}}) + \widetilde{f}_{\texttt{PT}}(\overline{X};\texttt{msk}_{\texttt{PT}}) - f^\dagger_{W_{\texttt{DS}}}(\overline{X};\texttt{msk}_{\texttt{DS}}) \\
&\qquad + f^\dagger_{W_{\texttt{DS}}}(\overline{X};\texttt{msk}_{\texttt{DS}}) - f^*_{\texttt{DS}}(\overline{X})\big\|_2^2\Big] \\
&\le 3\mathbb{E}\Big[\big\|f_{\texttt{DS}}(\overline{X};\texttt{msk}_{\texttt{DS}}) - \widetilde{f}_{\texttt{PT}}(\overline{X};\texttt{msk}_{\texttt{PT}})\big\|_2^2\Big] + 3\mathbb{E}\Big[\big\|\widetilde{f}_{\texttt{PT}}(\overline{X};\texttt{msk}_{\texttt{PT}}) - f^\dagger_{W_{\texttt{DS}}}(\overline{X};\texttt{msk}_{\texttt{DS}})\big\|_2^2\Big] \\
&\qquad + 3\mathbb{E}\Big[\big\|f^\dagger_{W_{\texttt{DS}}}(\overline{X};\texttt{msk}_{\texttt{DS}}) - f^*_{\texttt{DS}}(\overline{X})\big\|_2^2\Big] \\
&\le 3\epsilon^2_{\texttt{agg}}(B) + 3\,\underbrace{\mathbb{E}\Big[\big\|\widetilde{f}_{W_{\texttt{DS}}}(\overline{X};\texttt{msk}_{\texttt{DS}}) - f^\dagger_{W_{\texttt{DS}}}(\overline{X};\texttt{msk}_{\texttt{DS}})\big\|_2^2\Big]}_{\text{(i)}} \\
&\qquad + 3\,\underbrace{\mathbb{E}\Big[\big\|f^\dagger_{W_{\texttt{DS}}}(\overline{X};\texttt{msk}_{\texttt{DS}}) - f^*_{\texttt{DS}}(\overline{X})\big\|_2^2\Big]}_{\text{(ii)}},
\end{aligned}$$

where the second line follows from the definition of the regression function $f^*_{\mathtt{DS}}(\overline{X}) = \mathbb{E}[y_{\mathtt{DS}} \,|\, \overline{X}]$ and the last line follows from (3.6.11). In what follows, we characterize terms (i) and (ii).

Term (i). Recall that the regression function $f^*_{\mathtt{PT}}(\overline{X})$ for the pretraining process is defined in (3.6.3). For any truncated input sequence $\overline{X}$, it holds that

$$\begin{aligned}
&\big\|\widetilde{f}_{\mathtt{PT}}(\overline{X}; \mathtt{msk}_{\mathtt{PT}}) - f^\dagger_{W_{\mathtt{DS}}}(\overline{X}; \mathtt{msk}_{\mathtt{DS}})\big\|_2^2 \\
&\quad = \Big\|B\widehat{f}_{\mathtt{PT}}(z; \mathtt{msk}_{\mathtt{PT}}) - W_{\mathtt{DS}}^\top \mathbb{E}_{z \,|\, \overline{X}}\big[\psi_{\mathtt{DS}}(z; \mathtt{msk}_{\mathtt{DS}})\big]\Big\|_2^2 \\
&\quad = \Big\|B\big(\widehat{f}_{\mathtt{PT}}(\overline{X}; \mathtt{msk}_{\mathtt{PT}}) - f^*_{\mathtt{PT}}(\overline{X})\big) + \Big(Bf^*_{\mathtt{PT}}(\overline{X}) - \mathbb{E}_{z \,|\, \overline{X}}\big[W_{\mathtt{DS}}^\top \psi_{\mathtt{DS}}(z; \mathtt{msk}_{\mathtt{DS}})\big]\Big)\Big\|_2^2
\end{aligned}$$

(C.5.9)

$$\leq 2\underbrace{\Big\|B\big(\widehat{f}_{\mathtt{PT}}(\overline{X}; \mathtt{msk}_{\mathtt{PT}}) - f^*_{\mathtt{PT}}(\overline{X})\big)\Big\|_2^2}_{\text{(i.a)}} + 2\underbrace{\Big\|\Big(Bf^*_{\mathtt{PT}}(\overline{X}) - \mathbb{E}_{z \,|\, \overline{X}}\big[W_{\mathtt{DS}}^\top \psi_{\mathtt{DS}}(z; \mathtt{msk}_{\mathtt{DS}})\big]\Big)\Big\|_2^2}_{\text{(i.b)}},$$

where the second line follows from the definition of $f^\dagger_{W_{\mathtt{DS}}}(\overline{X}; \mathtt{msk}_{\mathtt{DS}})$ in (3.6.6). In the sequel, we characterize terms (i.a) and (i.b). By Assumption 3.6.2, we have

$$\text{(i.a)} \leq \|B\|_2^2 \cdot \big\|\widehat{f}_{\mathtt{PT}}(\overline{X}; \mathtt{msk}_{\mathtt{PT}}) - f^*_{\mathtt{PT}}(\overline{X})\big\|_2^2 \leq \mu \cdot \big\|\widehat{f}_{\mathtt{PT}}(\overline{X}; \mathtt{msk}_{\mathtt{PT}}) - f^*_{\mathtt{PT}}(\overline{X})\big\|_2^2. \tag{C.5.10}$$

Recall that $g^{\dagger}_{W_{\mathtt{SSL}}}(z;\mathtt{msk}_{\mathtt{DS}})$ is defined in Assumption 3.6.1. Since $BW_{\mathtt{SSL}}^{\top} = W_{\mathtt{DS}}^{\top}$, we have

$$\begin{aligned}
\text{(i.b)} &= \Big\| B\mathbb{E}_{z \,|\, \overline{X}}\big[g^{*}_{\mathtt{PT}}(z;\mathtt{msk}_{\mathtt{PT}}) - W_{\mathtt{SSL}}^{\top}\psi_{\mathtt{DS}}(z;\mathtt{msk}_{\mathtt{DS}})\big]\Big\|_2^2 \\
&\le \|B\|_2^2 \cdot \Big\| \mathbb{E}_{z \,|\, \overline{X}}\big[g^{*}_{\mathtt{PT}}(z;\mathtt{msk}_{\mathtt{PT}}) - g^{\dagger}_{W_{\mathtt{SSL}}}(z;\mathtt{msk}_{\mathtt{DS}})\big]\Big\|_2^2 \\
&\le \mu \cdot \sum_{i=1}^{d} \mathbb{E}_{z \,|\, \overline{X}}\big[g^{*}_{\mathtt{PT},i}(z;\mathtt{msk}_{\mathtt{PT}}) - g^{\dagger}_{W_{\mathtt{SSL}},i}(z;\mathtt{msk}_{\mathtt{DS}})\big]^2 \\
&\le \mu \cdot \sum_{i=1}^{d} \big\| g^{*}_{\mathtt{PT},i}(\cdot;\mathtt{msk}_{\mathtt{PT}}) - g^{\dagger}_{W_{\mathtt{SSL}},i}(\cdot;\mathtt{msk}_{\mathtt{DS}})\big\|_{\infty}^2 \\
&\le \mu \cdot \epsilon_{\mathtt{SSL}}^2(\mathtt{msk}_{\mathtt{PT}}, \mathtt{msk}_{\mathtt{DS}}),
\end{aligned} \tag{C.5.11}$$

where the third line follows from Assumption 3.6.2 and the last line follows from Assumption 3.6.1. Taking (C.5.10) and (C.5.11) into (C.5.9), we obtain

$$\begin{aligned}
\text{(i)} &= \mathbb{E}\Big[\big\|\widetilde{f}_{\mathtt{PT}}(\overline{X};\mathtt{msk}_{\mathtt{PT}}) - f^{\dagger}_{W_{\mathtt{DS}}}(\overline{X};\mathtt{msk}_{\mathtt{DS}})\big\|_2^2\Big] \\
&\le \mathbb{E}\Big[2\mu \cdot \big\|\widehat{f}_{\mathtt{PT}}(\overline{X};\mathtt{msk}_{\mathtt{PT}}) - f^{*}_{\mathtt{PT}}(\overline{X})\big\|_2^2 + 2\mu \cdot \epsilon_{\mathtt{SSL}}(\mathtt{msk}_{\mathtt{PT}}, \mathtt{msk}_{\mathtt{DS}})\Big] \\
&\le 2\mu \cdot \mathcal{E}^{\mathtt{PT}}_{\text{approx}} + \mu \cdot \mathbb{E}\Big[\epsilon_{\mathtt{SSL}}^2(\mathtt{msk}_{\mathtt{PT}}, \mathtt{msk}_{\mathtt{DS}})\Big] \\
&= 2\mu \cdot \big(\mathcal{E}^{\mathtt{PT}}_{\text{approx}} + \epsilon_{\mathtt{SSL}}^2(\mathtt{msk}_{\mathtt{PT}}, \mathtt{msk}_{\mathtt{DS}})\big),
\end{aligned} \tag{C.5.12}$$

where the third line follows from the definition of the regression function $f^{*}_{\mathtt{PT}}(\overline{X}) = \mathbb{E}[y_{\mathtt{PT}} \,|\, \overline{X}]$ for the pretraining process and the definition of $\mathcal{E}^{\mathtt{PT}}_{\text{approx}}$ in (3.6.12).

Term (ii). By the same argument for (C.5.5), we have

$$\text{(ii)} = \mathbb{E}\Big[\Big\|\mathbb{E}_{z \,|\, \overline{X}}\big[g^{*}_{\mathtt{DS}}(z;\mathtt{msk}_{\mathtt{DS}}) - g^{\dagger}_{W_{\mathtt{DS}}}(z;\mathtt{msk}_{\mathtt{DS}})\big]\Big\|_2^2\Big]. \tag{C.5.13}$$

By Assumption 3.6.1, we have

$$\begin{aligned}
&\Big\|\mathbb{E}_{z\,|\,\overline{X}}\big[g^*_{\mathtt{DS}}(z;\mathtt{msk}_{\mathtt{DS}}) - g^{\dagger}_{W_{\mathtt{DS}}}(z;\mathtt{msk}_{\mathtt{DS}})\big]\Big\|_2^2 \\
&\quad = \sum_{i=1}^{d_{\mathrm{y}}} \mathbb{E}_{z\,|\,\overline{X}}\big[g^*_{\mathtt{DS},i}(z;\mathtt{msk}_{\mathtt{DS}}) - g^{\dagger}_{W_{\mathtt{DS}},i}(z;\mathtt{msk}_{\mathtt{DS}})\big]^2 \\
&\quad \leq \sum_{i=1}^{d_{\mathrm{y}}} \big\|g^*_{\mathtt{DS},i}(\cdot;\mathtt{msk}_{\mathtt{DS}}) - g^{\dagger}_{W_{\mathtt{DS}},i}(\cdot;\mathtt{msk}_{\mathtt{DS}})\big\|_\infty^2 \leq \epsilon_g^2(\mathtt{msk}_{\mathtt{DS}}).
\end{aligned} \tag{C.5.14}$$

Taking (C.5.14) into (C.5.13), we obtain

$$\text{(ii)} \leq \epsilon_g^2(\mathtt{msk}_{\mathtt{DS}}). \tag{C.5.15}$$

Taking (C.5.12) and (C.5.15) into (C.5.8), we conclude the proof of Theorem 3.6.3. □

C.6. Auxiliary Proofs for Generalization

C.6.1. Proof of Lemma C.3.8

Proof. Throughout this proof, we consider a fixed dataset $\mathcal{D}_n = \{(X_i, y_i)\}_{i\in[n]}$. Let $\varepsilon_m = 2^{-m}$ with $m \in [M+2]$, where M is a positive integer. We denote by $\mathfrak{N}_m$ the covering of the empirical image class $\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j)$ that achieves the covering number $N(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j), \varepsilon_m, \|\cdot\|_{r,\infty})$. In other words, for any $f_j \in \mathcal{F}_j$, let $\widehat{f}^m[f_j] = (\widehat{f}^m[f_{j,i}])^\top_{i\in[n]} \in \mathfrak{N}_m$ be the nearest element of $f_j(X_i)$ in $\mathfrak{N}_m$, which implies that

$$\max_{i\in[n]}\big|f_j(X_i) - \widehat{f}^m[f_{j,i}]\big| \leq \varepsilon_m.$$

We have

(C.6.1)

$$\begin{aligned}
\mathcal{R}_{\mathcal{D}_n}(\mathcal{F}_j) &= \mathbb{E}\bigg[\sup_{f_j \in \mathcal{F}_j} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot f_j(X_i)\bigg] \\
&= \mathbb{E}\bigg[\sup_{f_j \in \mathcal{F}_j} \bigg\{\frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot \big(f_j(X_i) - \widehat{f}^M[f_{j,i}]\big) + \frac{1}{n}\sum_{m=1}^{M-1}\sum_{i=1}^{n} \epsilon_i \cdot (\widehat{f}^m[f_{j,i}] - \widehat{f}^{m+1}[f_{j,i}]) \\
&\qquad - \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot \widehat{f}^1[f_{j,i}]\bigg\}\bigg] \\
&\le \mathbb{E}\bigg[\sup_{f_j \in \mathcal{F}_j} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot \big(f_j(X_i) - \widehat{f}^M[f_{j,i}]\big)\bigg] \\
&\qquad + \sum_{m=1}^{M-1} \mathbb{E}\bigg[\sup_{f_j \in \mathcal{F}_j} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot (\widehat{f}^m[f_{j,i}] - \widehat{f}^{m+1}[f_{j,i}])\bigg] + \mathbb{E}\bigg[\sup_{f_j \in \mathcal{F}_j} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot \widehat{f}^1[f_{j,i}]\bigg] \\
&\le \underbrace{\mathbb{E}\bigg[\sup_{f_j \in \mathcal{F}_j} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot \big(f_j(X_i) - \widehat{f}^M[f_{j,i}]\big)\bigg]}_{\text{(i)}} \\
&\qquad + \sum_{m=1}^{M-1} \underbrace{\mathbb{E}\bigg[\sup_{f_j \in \mathcal{F}_j} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot (\widehat{f}^m[f_{j,i}] - \widehat{f}^{m+1}[f_{j,i}])\bigg]}_{\text{(ii)}},
\end{aligned}$$

where the last line follows from the choice $\mathfrak{N}_1 = \{0\}$ and the fact that $f_j(X_i) \in [0, 1/2]$ for any $i \in [n]$. In what follows, we analyze terms (i) and (ii).

Term (i). We have

(C.6.2)
$$\text{(i)} \le \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^{n} |\epsilon_i|\bigg] \cdot \sup_{f_j \in \mathcal{F}_j} \max_{i \in [n]} \big|f_j(X_i) - \widehat{f}^M[f_{j,i}]\big| \le n \cdot \varepsilon_M.$$

Term (ii). Let $f_{j,\mathcal{D}_n} = (f_j(X_i))_{i\in[n]} \in \mathbb{R}^{1\times n}$. We have

$$\begin{aligned}
\sup_{f_j\in\mathcal{F}_j} \big\|\widehat{f}^m[f_j] - \widehat{f}^{m+1}[f_j]\big\|_2 &\le \sup_{f_j\in\mathcal{F}_j} \big\|\widehat{f}^m[f_j] - f_{j,\mathcal{D}_n}\big\|_2 + \sup_{f_j\in\mathcal{F}_j} \big\|f_{j,\mathcal{D}_n} - \widehat{f}^{m+1}[f_j]\big\|_2 \\
&\le \sqrt{n}\cdot \sup_{f_j\in\mathcal{F}_j} \big\|\widehat{f}^m[f_j] - f_{j,\mathcal{D}_n}\big\|_\infty + \sqrt{n}\cdot \sup_{f_j\in\mathcal{F}_j} \big\|f_{j,\mathcal{D}_n} - \widehat{f}^{m+1}[f_j]\big\|_\infty \\
&\le \sqrt{n}\cdot\varepsilon_m + \sqrt{n}\cdot\varepsilon_{m+1} = 3\sqrt{n}\cdot\varepsilon_{m+1}. \qquad \text{(C.6.3)}
\end{aligned}$$

Combining (C.6.3) with the Massart's finite class lemma (Mohri et al., 2018), we obtain

$$\begin{aligned}
\mathbb{E}\bigg[\sup_{f_j\in\mathcal{F}_j} \sum_{i=1}^n \epsilon_i \cdot \big(\widehat{f}^m[f_{j,i}] - \widehat{f}^{m+1}[f_{j,i}]\big)\bigg] &\le 3\sqrt{n}\cdot\varepsilon_{m+1}\cdot\sqrt{2\log\big(|\mathfrak{N}_m|\cdot|\mathfrak{N}_{m+1}|\big)} \\
&\le 6\sqrt{n}\cdot\varepsilon_{m+1}\cdot\sqrt{\log|\mathfrak{N}_{m+1}|}, \qquad \text{(C.6.4)}
\end{aligned}$$

where the second line follows from the fact that $|\mathfrak{N}_{m+1}| \ge |\mathfrak{N}_m|$. Taking (C.6.2) and (C.6.4) into (C.6.1), we obtain

$$\begin{aligned}
\mathcal{R}_{\mathcal{D}_n}(\mathcal{F}_j) &\le \varepsilon_M + \frac{6}{\sqrt{n}}\cdot\sum_{m=1}^{M-1} \varepsilon_{m+1}\cdot\sqrt{\log|\mathfrak{N}_{m+1}|} \\
&\le \varepsilon_M + \frac{12}{\sqrt{n}}\cdot\sum_{m=1}^{M} (\varepsilon_m - \varepsilon_{m+1})\cdot\sqrt{\log|\mathfrak{N}_m|} \\
&\le \varepsilon_M + \frac{12}{\sqrt{n}}\int_{\varepsilon_{M+1}}^{1/2} \sqrt{\log N\big(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j), \varepsilon, \|\cdot\|_{r,\infty}\big)}\mathrm{d}\varepsilon \\
&\le 4\xi + \frac{12}{\sqrt{n}}\int_{\xi}^{1/2} \sqrt{\log N\big(\mathfrak{I}_{\mathcal{D}_n}(\mathcal{F}_j), \varepsilon, \|\cdot\|_{r,\infty}\big)}\mathrm{d}\varepsilon,
\end{aligned}$$

where the last inequality holds for any $0 < \xi < 1$ and the smallest M such that $\xi \le \varepsilon_{M+1}$, which implies that $\varepsilon_M = 2\varepsilon_{M+1} < 4\xi$. Therefore, we conclude the proof of Lemma C.3.8. □

C.6.1.1. Proof of Matrix Ball Covering Lemma C.3.10.

Proof of Lemma C.3.10. Let $M^\top = (m_1, \ldots, m_{d_1}) \in \mathbb{R}^{d_2 \times d_1}$, where $m_j \in \mathbb{R}^{d_1}$ with $j \in [d_2]$. We define the vectorization of the matrix $M \in \mathbb{R}^{d_1 \times d_2}$ as $\mathrm{vec}(M) = (m_j^\top)_{j\in[d_2]}^\top \in \mathbb{R}^{d_1 d_2}$. We define the sectional norm for the vector $\mathrm{vec}(M) \in \mathbb{R}^{d_1 d_2}$ as follows,

$$\big\|\mathrm{vec}(M)\big\|_{r(d_2),s(d_1)} = \|M^\top\|_{r,s},$$

which can be verified to be a proper norm. In Lemma C.7.1, setting

$$\mathbb{B} = \mathbb{B}_* = \big\{m \in \mathbb{R}^{d_1 d_2} : \|m\|_{r(d_2),s(d_1)} \le 1\big\} = \big\{M^\top \in \mathbb{R}^{d_2 \times d_1} : \|M^\top\|_{r,s} \le 1\big\},$$

and $\|\cdot\| = \|\cdot\|_* = \|\cdot\|_{r(d_2),s(d_1)}$, we obtain

$$\begin{aligned}
&\log N\Big(\big\{M^\top \in \mathbb{R}^{d_2 \times d_1} : \|M^\top\|_{r,s} \le R^{\mathrm{m}}\big\}, \varepsilon, \|\cdot\|_{r,s}\Big)\\
&\quad = \log N\Big(\big\{m \in \mathbb{R}^{d_1 d_2} : \|m\|_{r(d_2),s(d_1)} \le 1\big\}, \varepsilon/R^M, \|\cdot\|_{r,s}\Big)\\
&\quad \le \frac{\mathrm{vol}(2R_M/\varepsilon \cdot \mathbb{B}_* + \mathbb{B})}{\mathrm{vol}(\mathbb{B})} = d_1 d_2 \cdot \log\bigg(1 + \frac{2R_M}{\varepsilon}\bigg).
\end{aligned}$$

Therefore, we conclude the proof of Lemma C.3.10. □

C.6.2. Lipschitz Continuity of Multihead Attention

Lemma C.6.1 (Lipschitz Continuous Softmax)**.** Let (r, s) be a conjugate pair. Under Assumption C.3.1, it holds for any $q, \widehat{q} \in \mathbb{R}^{d_{\mathrm{p}}}$ and $K = (k^\ell)_{\ell\in[L]}^\top, \widehat{K} = (\widehat{k}^\ell)_{\ell\in[L]}^\top \in \mathbb{R}^{L \times d_{\mathrm{p}}}$

that

(C.6.5)
$$\Big\|\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(K,q)\big)-\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(\widehat{K},q)\big)\Big\|_1\le\big(\|q\|_r+\|K^\top\|_{r,\infty}\big)\cdot\|K^\top-\widehat{K}^\top\|_{r,\infty},$$

(C.6.6)
$$\Big\|\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(K,q)\big)-\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(K,\widehat{q})\big)\Big\|_1\le\big(\|q\|_r+\|K^\top\|_{r,\infty}\big)\cdot\|q-\widehat{q}\|_r.$$

Proof. Let $P=\mathrm{diag}(p)-pp^\top\in\mathbb{R}^{L\times L}$ with $p=(p_\ell)_{\ell\in[L]}=\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(K,q)\big)\in\mathbb{R}^L$. We have

$$p_\ell\propto\exp\big\{-\|q-k^\ell\|_2^2/2\sigma^2\big\}.$$

We define $g_\ell=-\|q-k^\ell\|_2^2/2\sigma^2$ and $g=(g_\ell)_{\ell\in[L]}^\top\in\mathbb{R}^L$. Let the Jacobian of $p\in\mathbb{R}^L$ with respect to $k^\ell\in\mathbb{R}^{d_{\mathrm{p}}}$ be $J_\ell\in\mathbb{R}^{L\times d_{\mathrm{p}}}$. We have

$$J_\ell=\frac{\partial p}{\partial k^\ell}=\frac{\partial p}{\partial g}\cdot\frac{\partial g}{\partial k^\ell}=P\frac{\partial g}{\partial k^\ell},$$

where $\partial g/\partial k^\ell=(e_\ell q^\top-E_{\ell,\ell}K)/\sigma^2$. Here $E_{\ell,\ell'}\in\mathbb{R}^{L\times L}$ is the unit matrix whose (ℓ,ℓ')-th entry is one and all other entries are zero. Note that

$$\Big\|\sum_{\ell=1}^L J_\ell\Delta_\ell\Big\|_1\le\sum_{\ell=1}^L\|J_\ell\Delta_\ell\|_1\le\sum_{\ell=1}^L\|J_\ell\|_{r\to1}\cdot\|\Delta_\ell\|_r\le\|\Delta\|_{r,\infty}\cdot\sum_{\ell=1}^L\|J_\ell\|_{r\to1},$$

where $\Delta=(\Delta_\ell^\top)_{\ell\in[L]}$. Thus, the Lipschitz continuity constant of $\mathtt{softmax}(q,K)$ is bounded by $\sum_{\ell=1}^L\|J_\ell\|_{r\to1}$. Let $e_\ell\in\mathbb{R}^L$ be the ℓ-th one-hot vector with $\ell\in[L]$. For any $\ell\in[L]$, we have

$$
\begin{aligned}
\|J_\ell\|_{r\to 1} &\le \frac{d_{\mathrm{p}}^{1-1/r}}{\sigma^2}\cdot \big\|P(e_\ell q^\top - E_{\ell,\ell}K)\big\|_1 \\
&= \frac{d_{\mathrm{p}}^{1/s}}{\sigma^2}\cdot p_\ell \cdot \big\|(e_\ell - p)(q-k^\ell)^\top\big\|_1 \\
&= \frac{d_{\mathrm{p}}^{1/s}p_\ell}{\sigma^2}\cdot \|e_\ell - p\|_s\cdot \|q-k^\ell\|_r, \qquad \text{(C.6.7)}
\end{aligned}
$$

where the equalities follow from $1/r+1/s=1$. Summing up (C.6.7) for all $\ell\in[L]$, we obtain

$$
\begin{aligned}
\sum_{\ell\in[L]}\|J_\ell\|_{r\to 1} &\le \frac{d_{\mathrm{p}}^{1/s}}{\sigma^2}\cdot\sum_{\ell\in[L]} p_\ell\cdot\|e_\ell-p\|_s\cdot\|q-k^\ell\|_r \\
&\le \frac{d_{\mathrm{p}}^{1/s}}{\sigma^2}\cdot\big(\|q\|_r+\|K^\top\|_{r,\infty}\big)\cdot\sum_{\ell\in[L]} p_\ell\cdot\|e_\ell-p\|_s. \qquad \text{(C.6.8)}
\end{aligned}
$$

On the other hand, we have

$$
\begin{aligned}
\sum_{\ell\in[L]} p_\ell\cdot\|e_\ell-p\|_s &= \sum_{\ell\in[L]}\bigg\{p_\ell\cdot\bigg[\sum_{\ell'\neq l} p_{\ell'}^s + (1-p_\ell)^s\bigg]^{1/s}\bigg\} \\
&\le \sum_{\ell\in[L]} p_\ell\cdot\big[2(1-p_\ell)^s\big]^{1/s}\le 2^{1/s}. \qquad \text{(C.6.9)}
\end{aligned}
$$

Combining (C.6.8) and (C.6.9), we have for $\sigma=(2d_{\mathrm{p}})^{1/2s}$ that

$$
\sum_{\ell\in[L]}\|J_\ell\|_{r\to 1}\le \frac{(2d_{\mathrm{p}})^{1/s}}{\sigma^2}\cdot\big(\|q\|_r+\|K^\top\|_{r,\infty}\big).
$$

Thus, $\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(K,q)\big)$ is $(\|q\|_r + \|K^\top\|_{r,\infty})$-Lipschitz in $K^\top$ with respect to $\|\cdot\|_{r,\infty}$, which concludes the proof of (C.6.5). Since $\mathfrak{K}_{\mathtt{RBF}}(q,k) = \mathfrak{K}_{\mathtt{RBF}}(k,q)$, we also have (C.6.6) by the same arguments for (C.6.5). Therefore, we conclude the proof of Lemma C.6.1. □

C.6.2.1. Proof of Lemma C.3.12.

Proof. For notational simplicity, we write

$$\mathtt{head}_i = \mathtt{attn}_{\mathtt{SM}}(XW_i^{\mathrm{q}}, XW_i^{\mathrm{k}}, XW_i^{\mathrm{v}}), \quad \widehat{\mathtt{head}}_i = \mathtt{attn}_{\mathtt{SM}}(X\widehat{W}_i^{\mathrm{q}}, X\widehat{W}_i^{\mathrm{k}}, X\widehat{W}_i^{\mathrm{v}}).$$

By the definition of sequence-to-sequence multihehead attention in (C.3.3), we have

$$\begin{aligned}\big\|\mathtt{mha}(X;W)^\top - \mathtt{mha}(X;\widehat{W})^\top\big\|_{r,\infty} &\le \bigg\|\sum_{i=1}^{h}\mathtt{head}_i^\top - \sum_{i=1}^{h}\widehat{\mathtt{head}}_i^\top\bigg\|_{r,\infty} \\ &\le \sum_{i=1}^{h}\big\|(\mathtt{head}_i - \widehat{\mathtt{head}}_i)^\top\big\|_{r,\infty}. \qquad \text{(C.6.10)}\end{aligned}$$

Also, we have

$$\begin{aligned}&\big\|(\mathtt{head}_i - \widehat{\mathtt{head}}_i)^\top\big\|_{r,\infty} \qquad \text{(C.6.11)}\\ &= \big\|\mathtt{attn}_{\mathtt{SM}}(XW_i^{\mathrm{q}}, XW_i^{\mathrm{k}}, XW_i^{\mathrm{v}})^\top - \mathtt{attn}_{\mathtt{SM}}(X\widehat{W}_i^{\mathrm{q}}, X\widehat{W}_i^{\mathrm{k}}, X\widehat{W}_i^{\mathrm{v}})^\top\big\|_{r,\infty} \\ &= \bigg\|\Big((XW_i^{\mathrm{v}})^\top\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(XW_i^{\mathrm{k}}, x^\ell W_i^{\mathrm{q}})\big)\Big)_{\ell\in[L]} \\ &\quad - \Big((X\widehat{W}_i^{\mathrm{v}})^\top\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(X\widehat{W}_i^{\mathrm{k}}, x^\ell \widehat{W}_i^{\mathrm{q}})\big)\Big)_{\ell\in[L]}\bigg\|_{r,\infty} \\ &= \max_{\ell\in[L]}\Big\|(XW_i^{\mathrm{v}})^\top\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(XW_i^{\mathrm{k}}, x^\ell W_i^{\mathrm{q}})\big) \\ &\quad - (X\widehat{W}_i^{\mathrm{v}})^\top\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(X\widehat{W}_i^{\mathrm{k}}, x^\ell \widehat{W}_i^{\mathrm{q}})\big)\Big\|_{r}.\end{aligned}$$

Note that

(C.6.12)

$$\begin{aligned}\|x^{\ell}W_i^{\mathrm{q}}\|_r + \big\|(XW_i^{\mathrm{k}})^\top\big\|_{r,\infty} &\le \big\|(XW_i^{\mathrm{q}})^\top\big\|_{r,\infty} + \big\|(XW_i^{\mathrm{k}})^\top\big\|_{r,\infty} \\ &\le \Big(\big\|(W_i^{\mathrm{q}})^\top\big\|_r + \big\|(W_i^{\mathrm{k}})^\top\big\|_r\Big)\cdot\|X^\top\|_{r,\infty} \le (\omega_i^{\mathrm{q}}+\omega_i^{\mathrm{k}})\cdot R.\end{aligned}$$

Then for any $\ell \in [L]$, we have

$$\begin{aligned}&\Big\|(XW_i^{\mathrm{v}})^\top\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(XW_i^{\mathrm{k}}, x^{\ell}W_i^{\mathrm{q}})\big) - (X\widehat{W}_i^{\mathrm{v}})^\top\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(X\widehat{W}_i^{\mathrm{k}}, x^{\ell}\widehat{W}_i^{\mathrm{q}})\big)\Big\|_r \\ &\quad\le \Big\|\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(X\widehat{W}_i^{\mathrm{k}}, x^{\ell}\widehat{W}_i^{\mathrm{q}})\big)^\top X(W_i^{\mathrm{v}} - \widehat{W}_i^{\mathrm{v}})\Big\|_r \\ &\qquad\quad + \Big\|\Big(\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(XW_i^{\mathrm{k}}, x^{\ell}W_i^{\mathrm{q}})\big) - \mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(X\widehat{W}_i^{\mathrm{k}}, x^{\ell}\widehat{W}_i^{\mathrm{q}})\big)\Big)^\top XW_i^{\mathrm{v}}\Big\|_r \\ &\quad\le \big\|(W_i^{\mathrm{v}} - \widehat{W}_i^{\mathrm{v}})^\top\big\|_{r,s}\cdot\|X^\top\|_{r,\infty} \\ &\qquad\quad + \Big\|\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(XW_i^{\mathrm{k}}, x^{\ell}W_i^{\mathrm{q}})\big) - \mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(X\widehat{W}_i^{\mathrm{k}}, x^{\ell}\widehat{W}_i^{\mathrm{q}})\big)\Big\|_1\big\|(W_i^{\mathrm{v}})^\top\big\|_r\|X^\top\|_{r,\infty} \\ &\quad\le R\cdot\varepsilon_i^{\mathrm{v}} + \Big(\|x^{\ell}W_i^{\mathrm{q}}\|_r + \big\|(XW_i^{\mathrm{k}})^\top\big\|_{r,\infty}\Big)\cdot\omega_i^{\mathrm{v}}R^2(\varepsilon_i^{\mathrm{q}}+\varepsilon_i^{\mathrm{k}})\end{aligned}$$

(C.6.13)

$$\le R\cdot\varepsilon_i^{\mathrm{v}} + (\omega_i^{\mathrm{q}}+\omega_i^{\mathrm{k}})\omega_i^{\mathrm{v}}\cdot R^3\cdot(\varepsilon_i^{\mathrm{q}}+\varepsilon_i^{\mathrm{k}}),$$

where the third inequality follows from Lemma C.7.3, the fourth inequality follows from Lemma C.6.1, and the last inequality follows from (C.6.12). Taking (C.6.13) into (C.6.11), we have

(C.6.14)
$$\big\|(\mathtt{head}_i - \widehat{\mathtt{head}}_i)^\top\big\|_{r,\infty} \le R\cdot\varepsilon_i^{\mathrm{v}} + (\omega_i^{\mathrm{q}}+\omega_i^{\mathrm{k}})\omega_i^{\mathrm{v}}\cdot R^3\cdot(\varepsilon_i^{\mathrm{q}}+\varepsilon_i^{\mathrm{k}}).$$

Taking (C.6.14) into (C.6.10), we obtain

$$\left\|\mathtt{mha}(X;W)^\top - \mathtt{mha}(X;\widehat{W})^\top\right\|_{r,\infty} \le R \cdot \sum_{i=1}^{h} \varepsilon_i^{\mathrm{v}} + R^3 \cdot \sum_{i=1}^{h} (\omega_i^{\mathrm{q}} + \omega_i^{\mathrm{k}}) \cdot (\varepsilon_i^{\mathrm{q}} + \varepsilon_i^{\mathrm{k}}).$$

Therefore, we conclude the proof of Lemma C.3.12. □

C.6.2.2. Proof of Lemma C.3.14.

Proof. In this proof, with a slight abuse of notations, we write

$$\mathtt{head}_i = \mathtt{attn}_{\mathtt{SM}}(XW_i^{\mathrm{q}}, XW_i^{\mathrm{k}}, XW_i^{\mathrm{v}}), \quad \widehat{\mathtt{head}}_i = \mathtt{attn}_{\mathtt{SM}}(\widehat{X}W_i^{\mathrm{q}}, \widehat{X}W_i^{\mathrm{k}}, \widehat{X}W_i^{\mathrm{v}}).$$

Similar to (C.6.10), we have

$$\left\|\mathtt{mha}(X;W)^\top - \mathtt{mha}(\widehat{X};W)^\top\right\|_{r,\infty} \le \sum_{i=1}^{h} \left\|(\mathtt{head}_i - \widehat{\mathtt{head}}_i)^\top\right\|_{r,\infty}.$$

For any fixed $\ell \in [L]$, we have

$$
\begin{aligned}
&\big\|(\mathtt{head}_i - \widehat{\mathtt{head}}_i)^\top\big\|_{r,\infty} \\
&\quad = \bigg\|\Big((XW_i^{\mathrm{v}})^\top \mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(XW_i^{\mathrm{k}}, x^\ell W_i^{\mathrm{q}})\big) - (\widehat{X}W_i^{\mathrm{v}})^\top \mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(\widehat{X}W_i^{\mathrm{k}}, \widehat{x}^\ell W_i^{\mathrm{q}})\big)\Big)_{\ell\in[L]}\bigg\|_{r,\infty} \\
&\quad = \max_{\ell\in[L]}\Big\|(XW_i^{\mathrm{v}})^\top \mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(XW_i^{\mathrm{k}}, x^\ell W_i^{\mathrm{q}})\big) - (\widehat{X}W_i^{\mathrm{v}})^\top \mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(\widehat{X}W_i^{\mathrm{k}}, \widehat{x}^\ell W_i^{\mathrm{q}})\big)\Big\|_r \\
&\quad \le \max_{\ell\in[L]}\Big\|\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(XW_i^{\mathrm{k}}, x^\ell W_i^{\mathrm{q}})\big)\Big\|_1 \cdot \big\|(W_i^{\mathrm{v}})^\top\big\|_r \cdot \|X^\top - \widehat{X}^\top\|_{r,\infty} \\
&\qquad + \max_{\ell\in[L]}\Big\|\mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(XW_i^{\mathrm{k}}, x^\ell W_i^{\mathrm{q}})\big) - \mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(\widehat{X}W_i^{\mathrm{k}}, \widehat{x}^\ell W_i^{\mathrm{q}})\big)\Big\|_1 \big\|(W_i^{\mathrm{v}})^\top\big\|_r \|X^\top\|_{r,\infty} \\
&\quad \le \big\|(W_i^{\mathrm{v}})^\top\big\|_r \cdot \|X^\top - \widehat{X}^\top\|_{r,\infty} \\
&\qquad + \max_{\ell\in[L]}\big\|(W_i^{\mathrm{v}})^\top\big\|_r R^2 \cdot (\omega_i^{\mathrm{q}} + \omega_i^{\mathrm{k}})\Big(\|x^\ell W_i^{\mathrm{q}}\|_r + \big\|(XW_i^{\mathrm{k}})^\top\big\|_{r,\infty}\Big) \\
&\quad \le \omega_i^{\mathrm{v}} \cdot \big[1 + R^2 \cdot (\omega_i^{\mathrm{q}} + \omega_i^{\mathrm{k}})^2\big] \cdot \|X^\top - \widehat{X}^\top\|_{r,\infty}, \qquad \text{(C.6.15)}
\end{aligned}
$$

where the second inequality follows from Lemma C.6.1 and (C.6.12), the third inequality follows from Lemma C.6.1, and the last inequality follows from (C.6.12). Summing up (C.6.15), we obtain

$$
\begin{aligned}
\big\|\mathtt{mha}(X;W)^\top - \mathtt{mha}(\widehat{X};W)^\top\big\|_{r,\infty} &\le \left\{\sum_{i=1}^h \omega_i^{\mathrm{v}} \cdot \big[1 + R^2 \cdot (\omega_i^{\mathrm{q}} + \omega_i^{\mathrm{k}})^2\big]\right\} \cdot \|X^\top - \widehat{X}^\top\|_{r,\infty} \\
&= \left[\sum_{i=1}^h \omega_i^{\mathrm{v}} + R^2 \cdot \sum_{i=1}^h (\omega_i^{\mathrm{q}} + \omega_i^{\mathrm{k}})^2 \omega_i^{\mathrm{v}}\right] \cdot \|X^\top - \widehat{X}^\top\|_{r,\infty}.
\end{aligned}
$$

Therefore, we conclude the proof of Lemma C.3.14. □

C.7. Auxiliary Lemmas

Lemma C.7.1 (Volume Ratios and Metric Entropy, Wainwright (2019))**.** Consider a pair of norms $\|\cdot\|$ and $\|\cdot\|_*$ on $\mathbb{R}^d$, and let $\mathbb{B}$ and $\mathbb{B}_*$ be their corresponding unit balls. Then the ε-covering number of $\mathbb{B}_*$ in the $\|\cdot\|$-norm obeys the bounds

$$\varepsilon^{-d} \cdot \frac{\mathrm{vol}(\mathbb{B}_*)}{\mathrm{vol}(\mathbb{B})} \leq N\big(\mathbb{B}_*, \varepsilon, \|\cdot\|\big) \leq \frac{\mathrm{vol}(2/\varepsilon \cdot \mathbb{B}_* + \mathbb{B})}{\mathrm{vol}(\mathbb{B})}.$$

Lemma C.7.2 (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

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

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

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

Lemma C.7.3. Let (r, s) be a conjugate pair. For any $M \in \mathbb{R}^{d_1\times d_2}$, $u \in \mathbb{R}^{d_2\times 1}$, and $U \in \mathbb{R}^{d_2\times d_3}$, we have

$$\|Mu\|_r \leq \|M\|_{r,\infty} \cdot \|u\|_1,$$

$$\|MU\|_{r,\infty} \leq \|M\|_{r,s} \cdot \|U\|_{r,\infty},$$

$$\|MU\|_{r,\infty} \leq \|M\|_{r} \cdot \|U\|_{r,\infty}.$$

Proof. Let $M = (m_i)_{i \in [d_2]}$ and $b = (b_i)_{i \in [d_2]}^\top$. We have

$$\|Mu\|_r = \left\| \sum_{j=1}^{d_2} u_j \cdot m_j \right\|_r \le \left(\sum_{j=1}^{d_2} |u_j| \right) \cdot \max_{j \in [d_2]} \|m_j\|_r = \|M\|_{r,\infty} \cdot \|u\|_1.$$

Also, we have

$$\|Mu\|_r = \left\| \sum_{j=1}^{d_2} u_j \cdot m_j \right\|_r \le \left(\sum_{j=1}^{d_2} |u_j| \cdot \|m_j\|_r \right) \le \|M\|_{r,s} \cdot \|u\|_r.$$

As a consequence, with $U = (u_i)_{i \in [d_3]}$, we have

$$\|MU\|_{r,\infty} = \max_{j \in [d_3]} \|Mu_j\|_r \le \max_{j \in [d_3]} \|M\|_{r,s} \cdot \|u_j\|_r = \|M\|_{r,s} \cdot \|U\|_{r,\infty}.$$

On the other hand, by the definition of the matrix operator norm, we obtain

$$\|MU\|_{r,\infty} = \max_{j \in [d_3]} \|Mu_j\|_r \le \max_{j \in [d_3]} \|M\|_r \cdot \|u_j\|_r = \|M\|_r \cdot \|U\|_{r,\infty}.$$

Therefore, we conclude the proof of Lemma C.7.3. □

Lemma C.7.4 (Covering Coefficient Bounds)**.** We have for all $t = 0, \ldots, T-1$ that

$$1 + \rho(\mathfrak{W}^{(t)}) \le \widetilde{\omega}^{\mathrm{v},(t)} + (\overline{\omega}^{\mathrm{qk},(t)})^2 \omega^{\mathrm{v},(t)} \cdot (R^{(t)})^2 = \widetilde{\rho}^{(t)},$$

$$R_{\texttt{mha}}(\mathfrak{W}^{(t)}) \le R^{\mathrm{v},(t)} + \overline{\omega}^{\mathrm{qk},(t)} R^{\mathrm{qk},(t)} \cdot (R^{(t)})^2 = R_{\texttt{mha}}^{(t)},$$

where $\widetilde{\rho}^{(t)}$ and $R_{\texttt{mha}}^{(t)}$ are defined in (C.3.20).

Proof. By the definition of $R_{\mathtt{mha}}(\mathfrak{W})$ in (C.3.32), we have

$$
\begin{aligned}
R_{\mathtt{mha}}(\mathfrak{W}^{(t)}) &= \sum_{i=1}^{h} R_i^{\mathrm{v},(t)} + (R^{\mathrm{x},(t)})^2 \cdot \sum_{i=1}^{h} (\omega_i^{\mathrm{q},(t)} + \omega_i^{\mathrm{k},(t)})(R_i^{\mathrm{q},(t)} + R_i^{\mathrm{k},(t)}) \\
&\leq R^{\mathrm{v},(t)} + (R^{(t)})^2 \cdot \max_{i\in[h]} \{\omega_i^{\mathrm{q},(t)} + \omega_i^{\mathrm{k},(t)}\} \cdot \sum_{i=1}^{h} (R_i^{\mathrm{q},(t)} + R_i^{\mathrm{k},(t)}) \\
&= R^{\mathrm{v},(t)} + \overline{\omega}^{\mathrm{qk},(t)} R^{\mathrm{qk},(t)} \cdot (R^{(t)})^2 = R_{\mathtt{mha}}^{(t)}.
\end{aligned}
$$

Also, by the definition of $\rho(\mathfrak{W})$ in (C.3.34), we have

$$
\begin{aligned}
1 + \rho(\mathfrak{W}^{(t)}) &= 1 + \sum_{i=1}^{h} \omega_i^{\mathrm{v},(t)} + (R^{\mathrm{x},(t)})^2 \cdot \sum_{i=1}^{h} (\omega_i^{\mathrm{q},(t)} + \omega_i^{\mathrm{k},(t)})^2 \omega_i^{\mathrm{v},(t)} \\
&\leq \widetilde{\omega}^{\mathrm{v},(t)} + (R^{(t)})^2 \cdot \max_{i\in[h]} \{\omega_i^{\mathrm{q},(t)} + \omega_i^{\mathrm{k},(t)}\}^2 \cdot \sum_{i=1}^{h} \omega_i^{\mathrm{v},(t)} \\
&= \widetilde{\omega}^{\mathrm{v},(t)} + (\overline{\omega}^{\mathrm{qk},(t)})^2 \omega^{\mathrm{v},(t)} \cdot (R^{(t)})^2 = \widetilde{\rho}^{(t)}.
\end{aligned}
$$

Therefore, we conclude the proof of Lemma C.7.4. □

Lemma C.7.5 (Simplified Covering Coefficient)**.** It holds that

$$
\sqrt{\log(1 + R^{(T)} R_{\mathtt{trans}})} = O\Big(T\sqrt{\log(1+\gamma)} + \sqrt{T}\sqrt{\log(1 + \zeta R^{(0)})} + \sqrt{\log(1 + \kappa/\zeta)}\Big),
$$

where γ, ζ, and κ are defined in (C.3.13). Here $R^{(t)}$ and $R_{\mathtt{trans}}$ are defined in (C.3.18) and (C.3.19), respectively.

Proof. By the definitions of $\kappa^{(t)}$ and $\zeta^{(t)}$ in (C.3.12), we have

$$\frac{R_{\texttt{mha}}^{(t)}}{\widetilde{\rho}^{(t)}} = \frac{R^{\mathrm{v},(t)} + \overline{\omega}^{\mathrm{qk},(t)} R^{\mathrm{qk},(t)} \cdot (R^{(t)})^2}{\widetilde{\omega}^{\mathrm{v},(t)} + (\overline{\omega}^{\mathrm{qk},(t)})^2 \omega^{\mathrm{v},(t)} \cdot (R^{(t)})^2} \leq \max\biggl\{ \frac{R^{\mathrm{v},(t)}}{\widetilde{\omega}^{\mathrm{v},(t)}}, \frac{R^{\mathrm{qk},(t)}}{\overline{\omega}^{\mathrm{qk},(t)} \omega^{\mathrm{v},(t)}} \biggr\} \leq \kappa^{(t)},$$

$$\frac{\widetilde{\rho}^{(t)}}{\widetilde{\omega}^{\mathrm{v},(t)}} = 1 + \frac{(\overline{\omega}^{\mathrm{qk},(t)})^2 \omega^{\mathrm{v},(t)}}{\widetilde{\omega}^{\mathrm{v},(t)}} \cdot (R^{(t)})^2 = 1 + \zeta^{(t)} \cdot (R^{(t)})^2,$$

which implies

$$\begin{aligned} R_{\texttt{trans}} &= \sum_{t=0}^{T-2} \biggl(\frac{R_{\texttt{mha}}^{(t)}}{\widetilde{\rho}^{(t)}} \cdot \prod_{\tau=t+1}^{T-1} \frac{\widetilde{\rho}^{(\tau)}}{\widetilde{\omega}^{\mathrm{v},(\tau)}} \biggr) + \sum_{t=0}^{T-1} \biggl(\frac{\alpha^{\mathrm{x},(t)} R^{\sigma,(t)} + \alpha^{\sigma,(t)} R^{\mathrm{x},(t)}}{\widetilde{\alpha}^{(t)}} \cdot \prod_{\tau=t+1}^{T-1} \frac{\widetilde{\rho}^{(\tau)}}{\widetilde{\omega}^{\mathrm{v},(\tau)}} \biggr) \\ &\leq \sum_{t=0}^{T-1} \biggl\{ (\kappa^{(t)} + \kappa^{(t)}) \cdot \prod_{\tau=t+1}^{T-1} \bigl[1 + \zeta^{(t)} \cdot (R^{(t)})^2 \bigr] \biggr\} \\ &\leq 2\kappa \cdot \sum_{t=0}^{T-1} \biggl\{ \prod_{\tau=t+1}^{T-1} \bigl[1 + \zeta \cdot (R^{(t)})^2 \bigr] \biggr\}, \end{aligned}$$

where the last line follows from the definition of κ and ζ in (C.3.13). By the definition of $R^{(t)}$ in (C.3.18) and the definition of γ in (C.3.13), we have

$$R^{(t)} = R^{(0)} \cdot \prod_{\tau=0}^{t-1} \widetilde{\omega}^{\mathrm{v},(\tau)} \widetilde{\alpha}^{(\tau)} \leq R^{(0)} \cdot (1+\gamma)^{2t}.$$

As a consequence, we obtain

$$
\begin{aligned}
R^{(T)} R_{\texttt{trans}} &\leq 2\kappa R^{(0)} \cdot (1+\gamma)^{2T} \cdot \sum_{t=0}^{T-1} \prod_{\tau=t+1}^{T-1} \left[1 + \zeta R^{(0)} \cdot (1+\gamma)^{4\tau}\right] \\
&\leq 2\kappa R^{(0)} \cdot (1+\gamma)^{2T} \cdot \sum_{t=0}^{T-1} \left[1 + \zeta R^{(0)} \cdot (1+\gamma)^{4T}\right]^{T-t+2} \\
&= 2\kappa \cdot \left[1 + \zeta R^{(0)} \cdot (1+\gamma)^{4T}\right]^{3} \cdot \frac{\left[1 + \zeta R^{(0)} \cdot (1+\gamma)^{4T}\right]^{T} - 1}{\zeta \cdot (1+\gamma)^{2T}} \\
&\leq \frac{2\kappa}{\zeta} \cdot \left[1 + \zeta R^{(0)} \cdot (1+\gamma)^{4T}\right]^{T+3}.
\end{aligned}
$$

Thus, using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $\log(1+ab) \leq \log(1+a) + \log(1+b)$, we further obtain

$$
\begin{aligned}
\sqrt{\log(1 + R^{(T)} R_{\texttt{trans}})} &= O\Big(\sqrt{T^2 \log(1+\gamma) + T \log(1 + \zeta R^{(0)}) + \log(1 + \kappa/\zeta)}\Big) \\
&= O\Big(T\sqrt{\log(1+\gamma)} + \sqrt{T}\sqrt{\log(1+\zeta R^{(0)})} + \sqrt{\log(1+\kappa/\zeta)}\Big).
\end{aligned}
$$

Therefore, we conclude the proof of Lemma C.7.5. □

Lemma C.7.6 (Inter-Layer Magnitude)**.** Under Assumptions 3.5.2 and C.3.3, it holds that

$$
\left\|(X^{(t)})^\top\right\|_{r,\infty} \leq \widetilde{\alpha}^{(t)} R^{(t)}, \qquad \left\|(X_\star^{(t)})^\top\right\|_{r,\infty} \leq R^{(t)},
$$

where $R^{(t)}$ is defined in (C.3.18).

Proof. Setting $\widehat{X} = 0^{L\times d}$ in (C.6.10), we have

$$
\left\|\texttt{mha}(X; W)^\top\right\|_{r,\infty} \leq \sum_{i=1}^{h} \|\texttt{head}_i^\top\|_{r,\infty}.
$$

We have for all $i \in [h]$ that

$$\begin{aligned}\|\mathtt{head}_i^\top\|_{r,\infty} &= \left\| \Big((XW_i^{\mathrm{v}})^\top \mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(XW_i^{\mathrm{k}}, x^\ell W_i^{\mathrm{q}})\big) \Big)_{\ell\in[L]} \right\|_{r,\infty} \\ &= \max_{l\in[L]} \Big\| (XW_i^{\mathrm{v}})^\top \mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(XW_i^{\mathrm{k}}, x^\ell W_i^{\mathrm{q}})\big) \Big\|_r \\ &\leq \max_{l\in[L]} \Big\| \mathtt{norm}_{\mathtt{SM}}\big(\mathfrak{K}_{\mathtt{RBF}}(XW_i^{\mathrm{k}}, x^\ell W_i^{\mathrm{q}})\big) \Big\|_1 \cdot \big\|(W_i^{\mathrm{v}})^\top\big\|_r \cdot \|X^\top\|_{r,\infty} \leq \omega_i^{\mathrm{v}} \cdot \|X^\top\|_{r,\infty},\end{aligned}$$

which implies that

$$\big\|\mathtt{mha}(X;W)^\top + X^\top\big\|_{r,\infty} \leq \|X^\top\|_{r,\infty} + \left(\sum_{i=1}^h \omega_i^{\mathrm{v}}\right) \cdot \|X^\top\|_{r,\infty} = (1+\omega^{\mathrm{v}}) \cdot \|X^\top\|_{r,\infty}.$$

As a consequence, we have

$$\big\|(X_\star^{(t)})^\top\big\|_{r,\infty} \leq (1+\omega^{\mathrm{v},(t)}) \cdot \big\|(X^{(t)})^\top\big\|_{r,\infty} = \widetilde{\omega}^{\mathrm{v},(t)} \cdot \big\|(X^{(t)})^\top\big\|_{r,\infty}. \tag{C.7.1}$$

On the other hand, setting $\widehat{X}_\star^{(t)} = 0^{L\times d}$ in (C.3.39), we have

$$\big\|(X^{(t+1)})^\top\big\|_{r,\infty} = \big\|\mathtt{ffn}(X_\star^{(t)}; A^{(t)})^\top\big\|_{r,\infty} \leq \widetilde{\alpha}^{(t)} \cdot \big\|(X_\star^{(t)})^\top\big\|_{r,\infty}. \tag{C.7.2}$$

Recursively applying (C.7.1) and (C.7.2), we conclude the proof of Lemma C.7.6. □

APPENDIX D

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

D.1. More Related Works

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.

D.2. Proofs for Section 4.4.1

D.2.1. Proof of Theorem 4.4.1

Proof. By (4.4.1), we have that

$$\begin{aligned}\mathbb{P}(r_{t+1}\,|\,\texttt{prompt}_t) &= \int \mathbb{P}(r_{t+1}\,|\,c_{t+1}, h_{t+1})\mathbb{P}(h_{t+1}\,|\,S_t)\mathrm{d}h_{t+1} \\ &= \int \mathbb{P}(r_{t+1}\,|\,c_{t+1}, h_{t+1})\mathbb{P}(h_{t+1}\,|\,S_t, z)\mathbb{P}(z\,|\,S_t)\mathrm{d}h_{t+1}\mathrm{d}z \\ &= \int \mathbb{P}(r_{t+1}\,|\,c_{t+1}, S_t, z)\mathbb{P}(z\,|\,S_t)\mathrm{d}z, \qquad \text{(D.2.1)}\end{aligned}$$

where the first and the last equalities results from model (4.4.2), and the second equality results from Bayes' theorem. □

D.2.2. Proof of Theorem 4.4.2

Proof. Note that

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

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

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

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

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

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

D.2.3. Proof of Proposition 4.4.3

Proof. The result follows from Propositions 3.4.1 and 3.4.2.

□

D.3. Appendix for Section 4.5

D.3.1. 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

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

D.3.2. Proof of Theorem 4.5.3

Proof of Theorem 4.5.3. 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 $x_{t+1}^n \sim \mathbb{P}(\cdot \mid S_t^n)$, 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 $\widetilde{x}_{t+1}^n \sim \mathbb{P}(\cdot \mid \widetilde{S}_t^n)$ independently from $\mathcal{D}$. We define function $g(\theta) = L(\theta, D) - \log \mathbb{E}_{\widetilde{\mathcal{D}}}[\exp(L(\theta, \widetilde{\mathcal{D}})) \mid \mathcal{D}]$, where

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

For distributions $Q, P \in \Delta(\Theta)$, where P can potentially depends on $\mathcal{D}$, Lemma D.6.2 shows that

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

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

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

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

$$-\mathbb{E}_{\theta\sim P}\Big[\log \mathbb{E}_{\widetilde{\mathcal{D}}}\big[\exp\big(L(\theta,\widetilde{\mathcal{D}})\big)\,|\,\mathcal{D}\big]\Big] \leq -\mathbb{E}_P\big[L(\theta,\mathcal{D})\big] + \mathrm{KL}(P\,\|\,Q) + \log\frac{1}{\delta}. \tag{D.3.1}$$

We first cope with the left-hand side of (D.3.1).

$$
\begin{aligned}
& - \mathbb{E}_P\Big[\log \mathbb{E}_{\widetilde{\mathcal{D}}}\big[\exp\big(L(\theta, \widetilde{\mathcal{D}})\big) \,|\, \mathcal{D}\big]\Big] \\
& \geq -\frac{1}{2} \log \mathbb{E}_{\widetilde{\mathcal{D}}}\bigg[\exp\bigg(-\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T} \log \frac{\mathbb{P}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n \,|\, S_t^n)}\bigg) \,\bigg|\, \mathcal{D}\bigg] \\
& \qquad - \frac{1}{2}\mathbb{E}_{\theta\sim P}\bigg[\log \mathbb{E}_{\widetilde{\mathcal{D}}}\bigg[\exp\bigg(-\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T} \log \frac{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta}(x_{t+1}^n \,|\, S_t^n)}\bigg) \,\bigg|\, \mathcal{D}\bigg]\bigg] \\
& = -\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T} \log \mathbb{E}_{x_{t+1}^n \sim \mathbb{P}(\cdot \,|\, S_t^n)}\bigg[\exp\bigg(-\frac{1}{2} \log \frac{\mathbb{P}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n \,|\, S_t^n)}\bigg) \,\bigg|\, \mathcal{D}\bigg] \\
& \qquad - \frac{1}{2}\mathbb{E}_{\theta\sim P}\bigg[\log \mathbb{E}_{\widetilde{\mathcal{D}}}\bigg[\exp\bigg(-\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T} \log \frac{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta}(x_{t+1}^n \,|\, S_t^n)}\bigg) \,\bigg|\, \mathcal{D}\bigg]\bigg] \\
& \geq \frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T} \mathrm{TV}\left(\mathbb{P}(\cdot \,|\, S_t^n), \mathbb{P}_{\widehat{\theta}}(\cdot \,|\, S_t^n)\right)^2 \\
& \qquad - \frac{1}{2}\mathbb{E}_{\theta\sim P}\bigg[\log \mathbb{E}_{\widetilde{\mathcal{D}}}\bigg[\exp\bigg(-\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T} \log \frac{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta}(x_{t+1}^n \,|\, S_t^n)}\bigg) \,\bigg|\, \mathcal{D}\bigg]\bigg], \qquad \text{(D.3.2)}
\end{aligned}
$$

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 x_{t+1}^n are independent given $\mathcal{D}$, and the last inequality results from Lemma D.6.4. The second term in the right-hand side of (D.3.2) 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 (D.3.1). For any $\theta^* \in \Theta$, we can decompose it as

$$
\begin{aligned}
&-\mathbb{E}_P\big[L(\theta,\mathcal{D})\big]\\
&\quad=\mathbb{E}_P\Bigg[\frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T}\log\frac{\mathbb{P}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta^*}(x_{t+1}^n \,|\, S_t^n)}+\log\frac{\mathbb{P}_{\theta^*}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n \,|\, S_t^n)}+\log\frac{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta}(x_{t+1}^n \,|\, S_t^n)}\Bigg]\\
&\quad\leq\frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T}\log\frac{\mathbb{P}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta^*}(x_{t+1}^n \,|\, S_t^n)}+\frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathbb{E}_P\Bigg[\log\frac{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta}(x_{t+1}^n \,|\, S_t^n)}\Bigg],
\end{aligned}
\tag{D.3.3}
$$

where the inequality results from the fact that $\widehat{\theta}$ maximizes the likelihood. We will choose θ^* as the projection of $\mathbb{P}$ onto $\{\mathbb{P}_\theta \,|\, \theta \in \Theta\}$. Thus, the first term in the right-hand side of (D.3.3) is the approximation error. The second term in the right-hand side of (D.3.3) can be controlled in the same way as the second term in the right-hand side of (D.3.2). Combining inequalities (D.3.1), (D.3.2), and (D.3.3), we have that

$$
\begin{aligned}
&\frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathrm{TV}\left(\mathbb{P}(\cdot \,|\, S_t^n),\mathbb{P}_{\widehat{\theta}}(\cdot \,|\, S_t^n)\right)^2\\
&\quad\leq\underbrace{\frac{1}{2}\mathbb{E}_P\Bigg[\log\mathbb{E}_{\widetilde{\mathcal{D}}}\Bigg[\exp\Bigg(-\frac{1}{2}\sum_{n,t=1}^{N,T}\log\frac{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta}(x_{t+1}^n \,|\, S_t^n)}\Bigg)\,\Bigg|\,\mathcal{D}\Bigg]\Bigg]+\frac{1}{4}\sum_{n,t=1}^{N,T}\mathbb{E}_P\Bigg[\log\frac{\mathbb{P}_{\widehat{\theta}}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta}(x_{t+1}^n \,|\, S_t^n)}\Bigg]}_{\text{(I)}}\\
&\qquad+\underbrace{\frac{1}{4}\sum_{n=1}^{N}\sum_{t=1}^{T}\log\frac{\mathbb{P}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta^*}(x_{t+1}^n \,|\, S_t^n)}}_{\text{(II)}}+\underbrace{\mathrm{KL}(P\,\|\,Q)}_{\text{(III)}}+\log\frac{1}{\delta},
\end{aligned}
\tag{D.3.4}
$$

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). We need to quantify the fluctuation of $\mathbb{P}_\theta$ when θ is changing.

Proposition D.3.1. For any input $X \in \mathbb{R}^{L\times d}$ and $\theta, \widetilde{\theta} \in \Theta$, we have that

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

where

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

for all $t \in [D]$.

Proof of Proposition D.3.1 . See Appendix D.5.2. □

With the help of Proposition D.3.1, we set the distribution P as

$$
\begin{aligned}
&\text{(D.3.5)} \qquad P = \prod_{t=1}^{D+1} \mathcal{L}_P\big(\theta^{(t)}\big) \\
&\mathcal{L}_P\big(\theta^{(D+1)}\big) = \mathrm{Unif}\Big(\mathbb{B}\big(\widehat{A}^{(D+1)}, r^{(D+1)}, \|\cdot\|_{1,2}\big)\Big) \\
&\mathcal{L}_P\big(\theta^{(t)}\big) = \mathrm{Unif}\Big(\mathbb{B}\big(\widehat{\gamma}_1^{(t)}, r_{\gamma,1}^{(t)}, |\cdot|\big)\Big) \cdot \mathrm{Unif}\Big(\mathbb{B}\big(\widehat{\gamma}_2^{(t)}, r_{\gamma,2}^{(t)}, |\cdot|\big)\Big) \cdot \mathcal{L}_P(A^{(t)}) \cdot \mathcal{L}_P(W^{(t)}) \\
&\mathcal{L}_P(A^{(t)}) = \mathrm{Unif}\Big(\mathbb{B}\big(\widehat{A}_1^{(t)}, r_{A,1}^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \mathrm{Unif}\Big(\mathbb{B}\big(\widehat{A}_2^{(t)}, r_{A,2}^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big) \\
&\mathcal{L}_P(W^{(t)}) = \prod_{i=1}^{h} \mathrm{Unif}\Big(\mathbb{B}\big(\widehat{W}_i^{Q,(t)}, r_Q^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \mathrm{Unif}\Big(\mathbb{B}\big(\widehat{W}_i^{K,(t)}, r_K^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big) \\
&\qquad\qquad \cdot \mathrm{Unif}\Big(\mathbb{B}\big(\widehat{W}_i^{V,(t)}, r_V^{(t)}, \|\cdot\|_{\mathrm{F}}\big)\Big)
\end{aligned}
$$

for $t \in [D]$, where Unif denotes the uniform distribution on the set, $\mathbb{B}(a, r, \|\cdot\|) = \{x \,|\, \|x - a\| \leq r\}$ 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

$$
\text{(D.3.6)} \qquad \mathbb{P}_{\widehat{\theta}}(x \,|\, S) \geq b_y = (1 + d_y \exp(B_A/\tau))^{-1}
$$

for any S and x. This results from the fact that

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

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

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

Thus, we have

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

Based on this, we conclude that

$$(\mathrm{I}) = \mathcal{O}(1). \tag{D.3.7}$$

Next, we control term (III) in (D.3.4). We take Q as

$$\begin{aligned} Q &= \prod_{t=1}^{D+1} \mathcal{L}_Q\big(\theta^{(t)}\big) \qquad \text{(D.3.8)} \\ \mathcal{L}_Q\big(\theta^{(D+1)}\big) &= \mathrm{Unif}\Big(\mathbb{B}\big(0, B_A, \|\cdot\|_{1,2}\big)\Big) \\ \mathcal{L}_Q\big(\theta^{(t)}\big) &= \mathrm{Unif}\Big(\mathbb{B}\big(1/2, 1/2, |\cdot|\big)\Big) \cdot \mathrm{Unif}\Big(\mathbb{B}\big(1/2, 1/2, |\cdot|\big)\Big) \cdot \mathcal{L}_Q(A^{(t)}) \cdot \mathcal{L}_Q(W^{(t)}) \\ \mathcal{L}_Q(A^{(t)}) &= \mathrm{Unif}\Big(\mathbb{B}\big(0, B_{A,1}, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \mathrm{Unif}\Big(\mathbb{B}\big(0, B_{A,2}, \|\cdot\|_{\mathrm{F}}\big)\Big) \\ \mathcal{L}_Q(W^{(t)}) &= \prod_{i=1}^{h} \mathrm{Unif}\Big(\mathbb{B}\big(0, B_Q, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \mathrm{Unif}\Big(\mathbb{B}\big(0, B_K, \|\cdot\|_{\mathrm{F}}\big)\Big) \cdot \mathrm{Unif}\Big(\mathbb{B}\big(0, B_V, \|\cdot\|_{\mathrm{F}}\big)\Big). \end{aligned}$$

Then the KL divergence between P and Q is

$$\mathrm{KL}(P \,\|\, Q)$$

$$= \mathcal{O}\Big((D^2 \cdot d \cdot (d_F + d_h + d) + d \cdot d_y) \cdot \log\big(1 + NT\tau^{-1} R h B_A B_{A,1} B_{A,2} B_Q B_K B_V\big)\Big). \tag{D.3.9}$$

Finally, we control term (II) in (D.3.4). This term can be controlled as

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

The first two terms in the right-hand side of the equality is the generalization error, which can be bounded with Lemma D.6.3. With Assumption 4.5.2, we note that

$$\left|\log\frac{\mathbb{P}(x \,|\, S)}{\mathbb{P}_{\theta^*}(x \,|\, S)}\right| \le b^* = \log\max\{c_0^{-1}, b_y^{-1}\}, \tag{D.3.10}$$

so the function satisfies the condition in Lemma D.6.3 with $c_i = 2b^*$. Using the moment generating function bound in Lemma D.6.3 and Chernoff bound, we have that

$$\begin{aligned}
&\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\log\frac{\mathbb{P}(x_{t+1}^n \,|\, S_t^n)}{\mathbb{P}_{\theta^*}(x_{t+1}^n \,|\, S_t^n)} - \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathbb{E}_{S_t^n}\mathrm{KL}\big(\mathbb{P}(\cdot \,|\, S_t^n) \,\|\, \mathbb{P}_{\theta^*}(\cdot \,|\, S_t^n)\big) \\
&\quad \le \sqrt{\frac{t_{\min} b^{*,2}}{2NT}\log\frac{1}{\delta}}
\end{aligned} \tag{D.3.11}$$

with probability at least $1 - \delta$.

Step 3: Conclude the proof.

Combining inequalities (D.3.4), (D.3.7), (D.3.9), and (D.3.11), we have that

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

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 D.3.2. Let $\mathcal{F}$ be the collection of functions of $f : \mathbb{R}^n \to \mathbb{R}$, and we assume that $|f| \le b$ 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 $t_{\mathrm{mix}}(\varepsilon)$. Given a distribution $\mathbb{P}_0$ on $\mathcal{F}$, with probability at least $1-\delta$, we have

$$
\Big|\mathbb{E}_P\Big[\mathbb{E}_X\big[f(X)\big] - f(X)\Big]\Big| \le \sqrt{\frac{b^2\cdot t_{\min}}{2\log 2N}}\left[\mathrm{KL}(P \,\|\, \mathbb{P}_0) + \log\frac{4}{\delta}\right],
$$

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

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

Proof of Proposition D.3.2. See Appendix D.5.1. □

We note that Proposition D.3.2 is indeed an uniform convergence bound, since it holds simultaneously for all P. Thus, we can set P and $\mathbb{P}_0$ as those in equalities (D.3.5) and (D.3.8), then we have that

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

Thus, we have that

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

We conclude the proof of Theorem 4.5.3.

□

D.3.3. Formal Statement and Proof of Proposition 4.5.4

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

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

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

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

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

Assumption D.3.3. 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 D.3.4. Under Assumptions 4.5.2 and D.3.3, if $d_F \geq 16d_y$, $B_{A,1} \geq 16Rd_y$, $B_{A,2} \geq d_F$ $B_A \geq \sqrt{d_y}$, and $B_V \geq \sqrt{d}$, then

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

for some constant $C > 0$.

Proof of Proposition D.3.4. Our proof mainly involves three steps.

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

Step 1: Build the high-level transformer approximator for g^*

Without loss of generality, we assume that $B > 1$ in Assumption D.3.7. To approximate ϕ^*, we 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 now is

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

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 ρ_i^* to approximate ρ^*, 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 D.6.8 to construct the networks. In the proof of Lemma D.6.8, 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 ρ_i^* with fully-connected networks Ψ_{ϕ^*} and $\Psi_{\rho_i^*}$ for $i \in [d_y]$ as

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

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

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

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

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

$$\begin{aligned}
&\left\| \rho^* \left(\frac{1}{L} \sum_{i=1}^{L} \phi^*(x_i) \right) - \Psi_{\rho^*} \left(\frac{1}{L} \sum_{i=1}^{L} \Psi_{\phi^*}(x_i) \right) \right\|_1 \\
&\quad \leq \left\| \rho^* \left(\frac{1}{L} \sum_{i=1}^{L} \phi^*(x_i) \right) - \Psi_{\rho^*} \left(\frac{1}{L} \sum_{i=1}^{L} \phi^*(x_i) \right) \right\|_1 \\
&\qquad + \left\| \Psi_{\rho^*} \left(\frac{1}{L} \sum_{i=1}^{L} \phi^*(x_i) \right) - \Psi_{\rho^*} \left(\frac{1}{L} \sum_{i=1}^{L} \Psi_{\phi^*}(x_i) \right) \right\|_1 \\
&\quad \leq d_y \varepsilon_{\phi} + d_y \cdot (B_{A,1})^{D''} \cdot \varepsilon_{\phi},
\end{aligned}$$

where 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. For the whole network, we have that

$$D' + D'' \leq D.$$

We take that $D' = D/2 + D^{3/4}$ and $D'' = \sqrt{D}/(\sqrt{C \cdot B} \log B_{A,1})$ for the constant C in Lemma D.6.8. Then for $D > 3$, we have that

$$\left\| \rho^* \bigg(\frac{1}{L} \sum_{i=1}^{L} \phi^*(x_i) \bigg) - \Psi_{\rho^*} \bigg(\frac{1}{L} \sum_{i=1}^{L} \Psi_{\phi^*}(x_i) \bigg) \right\|_1 = \mathcal{O}\bigg(d_y \exp\bigg(- \frac{D^{1/4}}{\sqrt{C^2 B^2 \log B_{A,1}}} \bigg) \bigg).$$

Step 3: Conclude the proof.

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

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

□

D.3.4. Pretraining Results for ℓ_2 Loss

D.3.4.1. 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 4.5.1, we consider the transformer with bounded weights. The set of parameters is

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

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

$$\widehat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \big\|x^i - f_\theta(S^i)\big\|_2^2 \tag{D.3.12}$$

From the population version of (D.3.12), 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^*.

D.3.4.2. Performance Guarantee for Pretraining with ℓ_2 Loss. We first state the assumptions for the pretraining setting.

Assumption D.3.5. There exists a constant $R > 0$ such that for $(S, x) \sim \mu$, we have $\|S^\top\|_{2,\infty} \le R$ and $\|x\|_2 \le B_x$ almost surely.

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

Theorem D.3.6. 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 D.3.5 holds, the pretrained model $f_{\widehat{\theta}}$ by the algorithm in (D.3.12) satisfies

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

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 4.5.3, 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

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

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

Assumption D.3.7. 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 D.3.8. Under D.3.7, 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

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

for some constant $C > 0$.

D.3.4.3. Proof of Theorem D.3.6.

Proof of Theorem D.3.6. 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

$$
\begin{aligned}
&\mathcal{L}(f_{\widehat{\theta}}) - \mathcal{L}(f^*) \\
&\quad = \mathcal{L}(f_{\widehat{\theta}}) - \mathcal{L}(f^*) - 2\big(\widehat{\mathcal{L}}(f_{\widehat{\theta}}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D})\big) + 2\big(\widehat{\mathcal{L}}(f_{\widehat{\theta}}, \mathcal{D}) - \widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D})\big) \\
&\qquad + 2\big(\widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D})\big) \\
&\quad \leq \underbrace{\mathcal{L}(f_{\widehat{\theta}}) - \mathcal{L}(f^*) - 2\big(\widehat{\mathcal{L}}(f_{\widehat{\theta}}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D})\big)}_{\text{generalization error}} + \underbrace{2\big(\widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D})\big)}_{\text{approximation error}},
\end{aligned} \tag{D.3.13}
$$

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 D.6.1 to bound it. Define the function

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

To verify the conditions in Lemma D.6.1, we notice that $|g(S, x, \theta)| \leq (B_x + B_f)^2$ and that

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

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 D.6.1 shows that for a distribution $Q \in \Delta(\Theta)$ and $0 < \lambda \leq 1/(2(B_x + B_f)^2)$, the following holds with probability at least $1-\delta$ simultaneously for all $P \in \Delta(\Theta)$

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

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

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

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 D.3.9. For any input $X \in \mathbb{R}^{L\times d}$ and $\theta, \widetilde{\theta} \in \Theta$, we have that

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

where

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

for all $t\in[D]$.

Proof of Proposition D.3.9 . See Appendix D.5.3. □

With the help of Proposition D.3.9, we set the distribution P as

$$P=\prod_{t=1}^{D+1}\mathcal{L}_P\big(\theta^{(t)}\big) \tag{D.3.14}$$

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

for $t \in [D]$, where Unif denotes the uniform distribution on the set, $\mathbb{B}(a, r, \|\cdot\|) = \{x \,|\, \|x - a\| \leq r\}$ 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

$$\begin{aligned}
&\left| \mathbb{E}_{\theta \sim P}\Big[\big\|x - f_\theta(S)\big\|_2^2 - \big\|x - f_{\widehat{\theta}}(S)\big\|_2^2 \Big] \right| \\
&\qquad \leq 2(B_x + B_f) \left| \mathbb{E}_{\theta \sim P}\Big[\big\| f_\theta(S) - f_{\widehat{\theta}}(S) \big\|_2 \Big] \right| = \mathcal{O}\bigg(\frac{B_x + B_f}{N} \bigg),
\end{aligned}$$

where the inequality results from Cauchy-Schwarz inequality, and the equality results from Proposition D.3.9. Thus, we have that

$$\begin{aligned}
&\mathcal{L}(f_{\widehat{\theta}}) - \mathcal{L}(f^*) - \big(\widehat{\mathcal{L}}(f_{\widehat{\theta}}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D}) \big) \\
&\quad \leq \frac{1}{2} \big(\mathcal{L}(f_{\widehat{\theta}}) - \mathcal{L}(f^*) \big) + \mathcal{O}\bigg(\frac{B_x + B_f}{N} \bigg) + \frac{2(3B_x + B_f)^2}{N} \bigg[\mathrm{KL}(P \,\|\, Q) + \log \frac{2}{\delta} \bigg]. \qquad \text{(D.3.15)}
\end{aligned}$$

To access to the value of $\mathrm{KL}(P \,\|\, Q)$, we take Q as the distribution in (D.3.8) except that

$$\mathcal{L}_Q\big(\theta^{(D+1)}\big) = \mathrm{Unif}\Big(\mathbb{B}\big(0, B_A, \|\cdot\|_{\mathrm{F}}\big)\Big). \tag{D.3.16}$$

Then the KL divergence between P and Q is

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

Combining this equality with (D.3.15), we have that with probability at least $1-\delta$, the generalization error can be bounded as

$$\mathcal{L}(f_{\widehat{\theta}}) - \mathcal{L}(f^*) - 2\big(\widehat{\mathcal{L}}(f_{\widehat{\theta}}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D})\big) = \mathcal{O}\bigg(\frac{B_x^2}{N}\bigg[\bar{D}\log(1+N\bar{B}) + \log\frac{2}{\delta}\bigg]\bigg). \tag{D.3.17}$$

Next we control the approximation error in (D.3.13).

$$\begin{aligned}
&\widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D}) \\
&\quad = \widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D}) - \frac{3}{2}\big(\mathcal{L}(f_{\theta^*}) - \mathcal{L}(f^*)\big) + \frac{3}{2}\big(\mathcal{L}(f_{\theta^*}) - \mathcal{L}(f^*)\big) \\
&\quad = \widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D}) - \frac{3}{2}\big(\mathcal{L}(f_{\theta^*}) - \mathcal{L}(f^*)\big) + \frac{3}{2}\mathbb{E}\Big[\big\|f^*(S) - f_{\theta^*}(S)\big\|_2^2\Big],
\end{aligned} \tag{D.3.18}$$

where the second equality results from the definition of f^*. To bound the first two terms in the right-hand side of (D.3.18), we use Lemma D.6.1 and take P and Q as (D.3.14) and (D.3.16), replacing $\widehat{\theta}$ by θ^*. Then we have that

(D.3.19)
$$\widehat{\mathcal{L}}(f_{\theta^*}, \mathcal{D}) - \widehat{\mathcal{L}}(f^*, \mathcal{D}) - \frac{3}{2}\big(\mathcal{L}(f_{\theta^*}) - \mathcal{L}(f^*)\big) = \mathcal{O}\bigg(\frac{B_x^2}{N}\bigg[\bar{D}\log(1+N\bar{B}) + \log\frac{2}{\delta}\bigg]\bigg).$$

Step 3: Conclude the proof.

Combining inequalities (D.3.13), (D.3.17), (D.3.18), and (D.3.19), we have that

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

Thus, we conclude the proof of Theorem D.3.6.

□

D.3.4.4. Proof of Proposition D.3.8.

Proof of Proposition D.3.8. 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 D.3.4 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

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

Thus, we conclude the proof of Proposition D.3.8. □

D.4. Proofs for §4.6

D.4.1. Proof of Theorem 4.6.2

Proof. By Theorem 4.4.2 and the fact that $\log(1/p_0(z_*)) \leq \beta$, we have that

$$T^{-1} \cdot \mathbb{E}_{\mathcal{D}_{\text{ICL}}}\Big[\sum_{t=1}^{T} \log \mathbb{P}(r_t \,|\, z^*, \texttt{prompt}_{t-1}) - \sum_{t=1}^{T} \log \mathbb{P}(r_t \,|\, \texttt{prompt}_{t-1})\Big] \leq \beta/T. \tag{D.4.1}$$

In addition, we have that

$$\begin{aligned} &T^{-1} \cdot \mathbb{E}_{\mathcal{D}_{\text{ICL}}}\Big[\sum_{t=1}^{T} \log \mathbb{P}(r_t \,|\, \texttt{prompt}_{t-1}) - \sum_{t=1}^{T} \log \mathbb{P}_{\widehat{\theta}}(r_t \,|\, \texttt{prompt}_{t-1})\Big] \\ &\quad = \mathbb{E}_{\mathcal{D}_{\text{ICL}}}\Big[\mathrm{KL}\big(\mathbb{P}(\cdot \,|\, \texttt{prompt}) \,\Big\|\, \mathbb{P}_{\widehat{\theta}}(\cdot \,|\, \texttt{prompt})\big)\Big]. \end{aligned} \tag{D.4.2}$$

Similar to (D.3.10), we have that

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

By Lemma D.6.9, we have that

$$\mathrm{KL}\big(\mathbb{P}(\cdot \,|\, \texttt{prompt}) \,\|\, \mathbb{P}_{\widehat{\theta}}(\cdot \,|\, \texttt{prompt})\big) \leq (3 + b^*)/2 \cdot \mathrm{TV}\left(\mathbb{P}(\cdot \,|\, \texttt{prompt}), \mathbb{P}_{\widehat{\theta}}.(\cdot \,|\, \texttt{prompt})\right). \tag{D.4.3}$$

By Assumption 4.6.5, we have that $\mathbb{P}_{\mathcal{D}_{\text{ICL}}}(\texttt{prompt}) \leq \kappa \mathbb{P}_{\mathcal{D}}(\texttt{prompt})$. Thus, by Theorem 4.5.3, we have with probability at least $1 - \delta$ that

$$\begin{aligned} &\mathbb{E}_{\mathcal{D}_{\text{ICL}}}\Big[\mathrm{KL}\big(\mathbb{P}(\cdot \,|\, \texttt{prompt}) \,\|\, \mathbb{P}_{\widehat{\theta}}(\cdot \,|\, \texttt{prompt})\big)\Big] \\ &\quad \leq C \cdot b^* \cdot \kappa \cdot \mathbb{E}_{S \sim \mathcal{D}}\Big[\mathrm{TV}\left(\mathbb{P}(\cdot \,|\, S), \mathbb{P}_{\widehat{\theta}}.(\cdot \,|\, S)\right)\Big] \leq C \cdot b^* \cdot \kappa \cdot \Delta_{\text{pre}}(N, T, \delta). \end{aligned} \tag{D.4.4}$$

Combining (D.4.4), (D.4.1), and (D.4.2), we have with probability at least $1-\delta$ that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{\text{ICL}}}\Big[T^{-1}\cdot\sum_{t=1}^{T}\log\mathbb{P}(r_t\,|\,z^*,\texttt{prompt}_{t-1}) - T^{-1}\cdot\sum_{t=1}^{T}\log\mathbb{P}_{\widehat{\theta}}(r_t\,|\,\texttt{prompt}_{t-1})\Big]\\
&\qquad\le \beta/T + \mathbb{E}_{S\sim\mathcal{D}}\Big[\mathrm{KL}\big(\mathbb{P}(\cdot\,|\,S)\,\|\,\mathbb{P}_{\widehat{\theta}}(\cdot\,|\,S)\big)\Big]\\
&\qquad\le \mathcal{O}\big(\beta/T + b^*\cdot\kappa\cdot\Delta_{\text{pre}}(N,T,\delta)\big),
\end{aligned}\tag{D.4.5}$$

which completes the proof of Theorem 4.6.2. □

D.4.2. Proof of Proposition 4.6.7

Proof of Proposition 4.6.7. From Bayesian model averaging, the output distribution is

$$\begin{aligned}
&\mathbb{P}(r_{t+1}\,|\,S_t',\widetilde{c}_{t+1})\\
&\quad=\sum_{z\in\mathfrak{Z}}\mathbb{P}(r_{t+1}\,|\,\widetilde{c}_{t+1},z)\cdot\mathbb{P}_{\mathcal{Z}}(z\,|\,S_t')\\
&\quad=\mathbb{P}(r_{t+1}\,|\,\widetilde{c}_{t+1},z^*) + \sum_{z\neq z^*}\big(\mathbb{P}(r_{t+1}\,|\,\widetilde{c}_{t+1},z)-\mathbb{P}(r_{t+1}\,|\,\widetilde{c}_{t+1},z^*)\big)\cdot\mathbb{P}_{\mathcal{Z}}(z\,|\,S_t')\\
&\quad=\mathbb{P}(r_{t+1}\,|\,\widetilde{c}_{t+1},z^*)\\
&\qquad+\sum_{z\neq z^*}\big(\mathbb{P}(r_{t+1}\,|\,\widetilde{c}_{t+1},z)-\mathbb{P}(r_{t+1}\,|\,\widetilde{c}_{t+1},z^*)\big)\cdot\mathbb{P}_{\mathcal{Z}}(z^*\,|\,S_t')\cdot\frac{\mathbb{P}_{\mathcal{Z}}(z)\mathbb{P}(S_t'\,|\,z)}{\mathbb{P}_{\mathcal{Z}}(z^*)\mathbb{P}(S_t'\,|\,z^*)},
\end{aligned}\tag{D.4.6}$$

where the first equality results from Bayesian model averaging, the last equality results from Bayes' theorem. Next, we upperbound the ratio $\mathbb{P}(S_t'\,|\,z)/\mathbb{P}(S_t'\,|\,z^*)$ in the right-hand side of Eqn. (D.4.6). We have that

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

where the first inequality results from Assumption 4.6.3, and the second inequality results from Assumption 4.5.2, which also implies that $|\log\mathbb{P}((\widetilde{c}_i, r_i)\,|\,z)/\mathbb{P}((\widetilde{c}_i, r_i)\,|\,z^*)| \le (1+l)\log 1/c_0$. Hoeffding inequality shows that with probability at least $1-\delta$, we have

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

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

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

Combining this inequality with Eqn. (D.4.6), we have that

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

Let $\mathbb{E}_{\texttt{prompt}'\sim\mathcal{D}}[\mathrm{TV}(\mathbb{P}(\cdot\,|\,S_t', \widetilde{c}_{t+1}), \mathbb{P}_{\widehat{\theta}}(\cdot\,|\,S_t', \widetilde{c}_{t+1}))] \leq \Delta_{\mathrm{pre}}$, where Δ_{pre} is the bound in Theorem 4.5.3. Then we have that

$$
\begin{aligned}
&\mathbb{E}_{\texttt{prompt}'\sim\mathbb{P}'}\Big[\mathrm{KL}\big(\mathbb{P}(\cdot\,|\,\widetilde{c}_{t+1}, z^*)\|\mathbb{P}_{\widehat{\theta}}(\cdot\,|\,S_t', \widetilde{c}_{t+1})\big)\Big] \\
&\quad \leq \mathcal{O}\Big(\mathbb{E}_{\texttt{prompt}'\sim\mathbb{P}'}\Big[\,\mathrm{TV}\left(\mathbb{P}_{\widehat{\theta}}(\cdot\,|\,S_t', \widetilde{c}_{t+1}), \mathbb{P}(\cdot\,|\,\widetilde{c}_{t+1}, z^*)\right)\Big]\Big) \\
&\quad = \mathcal{O}\bigg(c_2\Delta_{\mathrm{pre}} + \exp\bigg(-t\bigg(\min_{z\neq z^*}\mathrm{KL}_{\mathrm{pair}}\big(\mathbb{P}(\cdot\,|\,z^*)\,\|\,\mathbb{P}(\cdot\,|\,z)\big) \\
&\qquad + 2\log c_0 - \frac{(1+l)}{\sqrt{t}}\log\frac{1}{c_0}\cdot\log\frac{|\mathfrak{Z}|}{\delta}\bigg)\bigg)\bigg).
\end{aligned}
$$

Thus, we conclude the proof of Proposition 4.6.7. □

D.5. Proof of Supporting Propositions

D.5.1. Proof of Proposition D.3.2

Proof of Proposition D.3.2. We note that $f(X)$ satisfies the condition in Lemma D.6.3 with $c_i = 2b/N$ for $i \in [N]$. Then Lemma D.6.3 shows that

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

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

$$
P\bigg(\mathbb{E}_{f\sim P_0}\Big(\exp\big[\lambda(f(X) - \mathbb{E}f(X))\big]\Big) \geq \frac{2}{\delta}\bigg) \leq \delta
$$

for any $0 < \delta < 1$. We note that this probability inequality does not involve P. Take the function g in Lemma D.6.2 as $g(f) = \lambda(f(X) - \mathbb{E}f(X))$, then it shows that

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

for any P simultaneously. Combining these inequalities, we have

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

for any distribution P on $\mathcal{F}$ simultaneously with probability at least $1 - \delta$. Thus, we conclude the proof of Proposition D.3.2. □

D.5.2. Proof of Proposition D.3.1

Proof of Proposition D.3.1 . 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

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

where the first inequality results from Lemma D.6.5, and the second inequality results from Lemma D.6.6 and that $\|X^{(D),\top}\|_{2,\infty} \leq 1$ due to the layer normalization. In the following, we build the recursion relationship between $\|X^{(t),\top} - \widetilde{X}^{(t),\top}\|_{2,\infty}$ for $t \in [D]$.

$$\begin{aligned}
&\|X^{(t+1),\top} - \widetilde{X}^{(t+1),\top}\|_{2,\infty} \\
&\quad \le \left\|\mathtt{ffn}(Y^{(t+1)}, A^{(t+1)})^\top - \mathtt{ffn}(\widetilde{Y}^{(t+1)}, \widetilde{A}^{(t+1)})^\top\right\|_{2,\infty} \\
&\qquad + |\gamma_2^{(t+1)} - \widetilde{\gamma}_2^{(t+1)}| + \left\|Y^{(t+1),\top} - \widetilde{Y}^{(t+1),\top}\right\|_{2,\infty} \\
&\quad \le |\gamma_2^{(t+1)} - \widetilde{\gamma}_2^{(t+1)}| + \left\|Y^{(t+1),\top} - \widetilde{Y}^{(t+1),\top}\right\|_{2,\infty} + B_{A,1} \cdot B_{A,2} \cdot \|Y^{(t+1),\top} - \widetilde{Y}^{(t+1),\top}\|_{2,\infty}
\end{aligned}$$

(D.5.1)

$$+ B_{A,2} \cdot \|A_1^{(t+1)} - \widetilde{A}_1^{(t+1)}\|_{\mathrm{F}} + B_{A,1} \cdot \|A_2^{(t+1)} - \widetilde{A}_2^{(t+1)}\|_{\mathrm{F}},$$

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 D.5.1. 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

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

Proof of Proposition D.5.1. See Appendix D.5.4. □

Next, we build the relationship between $\|Y^{(t+1),\top} - \widetilde{Y}^{(t+1),\top}\|_{2,\infty}$ in the right-hand side of inequality (D.5.1) and $\|X^{(t),\top} - \widetilde{X}^{(t),\top}\|_{2,\infty}$.

$$\begin{aligned}
&\|Y^{(t+1),\top} - \widetilde{Y}^{(t+1),\top}\|_{2,\infty} \\
&\quad \le \left\|\mathtt{mha}(X^{(t)}, W^{(t+1)})^\top - \mathtt{mha}(\widetilde{X}^{(t)}, \widetilde{W}^{(t+1)})^\top\right\|_{2,\infty} \\
&\qquad + |\gamma_1^{(t+1)} - \widetilde{\gamma}_1^{(t+1)}| + \left\|X^{(t),\top} - \widetilde{X}^{(t),\top}\right\|_{2,\infty} \\
&\quad \le |\gamma_1^{(t+1)} - \widetilde{\gamma}_1^{(t+1)}| + \left\|X^{(t),\top} - \widetilde{X}^{(t),\top}\right\|_{2,\infty} \\
&\qquad\quad + h \cdot B_V\big(1 + 4B_Q B_K\big)\|X^{(t),\top} - \widetilde{X}^{(t),\top}\|_{2,\infty} + \sum_{i=1}^{h} \big\|W_i^{V,(t+1)} - \widetilde{W}_i^{V,(t+1)}\|_{\mathrm{F}} \\
&\qquad\quad + 2B_V \cdot B_K \sum_{i=1}^{h} \|W_i^{Q,(t+1)} - \widetilde{W}_i^{Q,(t+1)}\|_{\mathrm{F}} + 2B_V \cdot B_Q \sum_{i=1}^{h} \|W_i^{K,(t+1)} - \widetilde{W}_i^{K,(t+1)}\|_{\mathrm{F}},
\end{aligned} \tag{D.5.2}$$

where the first inequality results from the triangle inequality, and the second inequality results from Lemma D.6.7. Combining inequalities (D.5.1) and (D.5.2), we derive that

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

This concludes the proof of Proposition D.3.1. □

D.5.3. Proof of Proposition D.3.9

Proof of Proposition D.3.9 . 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 D.6.6 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 D.3.1, 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 D.3.9. □

D.5.4. Proof of Proposition D.5.1

Proof of Proposition D.5.1. We have that

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

where the first inequality results from the triangle inequality, the second and the last inequalities result from Lemma D.6.6 and that `ReLU` is not expansive. Thus, we conclude the proof of Proposition D.5.1. □

D.6. Technical Lemmas

Lemma D.6.1 (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 $|f(X) - \mu(f)| \le b$ 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

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

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

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

with probability at least $1-\delta$.

Lemma D.6.2 (Donsker–Varadhan representation in Belghazi et al. (2018))**.** Let P and Q be distributions on a common space $\mathcal{X}$. Then

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

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

Lemma D.6.3 (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 D.6.4 (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

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

Lemma D.6.5 (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 D.6.6 (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 D.6.7 (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

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

Lemma D.6.8 (Lemma A.6 in Elbrächter et al. (2021))**.** For $a,b\in\mathbb{R}$ with $a<b$, let

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

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

$$\|f-\Psi_f\|_\infty\le\varepsilon,$$

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

Lemma D.6.9. Let $b=\sup_x\log(p(x)/q(x))$. We have that

$$\mathrm{KL}(p\,\|\,q)\le 2(3+b)\cdot\mathrm{TV}(p,q).\tag{D.6.1}$$

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

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

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