Relational Reasoning via Set Transformers: Provable Efficiency and Applications to MARL

Fengzhuo Zhang[1] **Boyi Liu**[2] **Kaixin Wang**[1] **Vincent Y. F. Tan**[1] **Zhuoran Yang**[3]
Zhaoran Wang[2]
[1]National University of Singapore [2] Northwestern University [3]Yale University
{fzzhang, kaixin.wang}@u.nus.edu, boyiliu2018@u.northwestern.edu,
vtan@nus.edu.sg, zhuoranyang.work@gmail.com, zhaoranwang@gmail.com

Abstract

The cooperative Multi-Agent Reinforcement Learning (MARL) with permutation invariant agents framework has achieved tremendous empirical successes in real-world applications. Unfortunately, the theoretical understanding of this MARL problem is lacking due to the curse of many agents and the limited exploration of the relational reasoning in existing works. In this paper, we verify that the transformer implements complex relational reasoning, and we propose and analyze model-free and model-based offline MARL algorithms with the transformer approximators. We prove that the suboptimality gaps of the model-free and model-based algorithms are independent of and logarithmic in the number of agents respectively, which mitigates the curse of many agents. These results are consequences of a novel generalization error bound of the transformer and a novel analysis of the Maximum Likelihood Estimate (MLE) of the system dynamics with the transformer. Our model-based algorithm is the first provably efficient MARL algorithm that explicitly exploits the permutation invariance of the agents. Our improved generalization bound may be of independent interest and is applicable to other regression problems related to the transformer beyond MARL.

1 Introduction

Cooperative MARL algorithms have achieved tremendous successes across a wide range of real-world applications including robotics [1, 2], games [3, 4], and finance [5]. In most of these works, the *permutation invariance* of the agents is embedded into the problem setup, and the successes of these works hinge on leveraging this property. However, the theoretical understanding of why the permutation invariant MARL has been so successful is lacking due to the following two reasons. First, the size of the state-action space grows exponentially with the number of agents; this is known as "the curse of many agents" [6, 7]. The exponentially large state-action space prohibits the learning of value functions and policies due to the curse of dimensionality. Second, although the mean-field approximation is widely adopted to mitigate the curse of many agents [6, 8], this approximation fails to capture the complex interplay between the agents. In the mean-field approximation, the influence of all the other agents on a fixed agent is captured only through the empirical distribution of the local states and/or local actions [6, 8]. This induces a restricted class of function approximators, which nullifies the possibly complicated relational structure of the agents, and thus fails to incorporate the complex interaction between agents. Therefore, designing provably efficient MARL algorithms that incorporate the efficient relational reasoning and break the curse of many agents remains an interesting and meaningful question.

In this paper, we regard transformer networks as the representation learning module to incorporate relational reasoning among the agents. In particular, we focus on the offline MARL problem with

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

the transformer approximators in the cooperative setting. In this setting, all the agents learn policies *cooperatively* to maximize a common reward function. More specifically, in the offline setting, the learner only has access to a pre-collected dataset and cannot interact adaptively with the environment. Moreover, we assume that the underlying Markov Decision Process (MDP) is *homogeneous*, which means that the reward and the transition kernel are permutation invariant functions of the state-action pairs of the agents. Our goal is to learn an *optimal* policy that is also permutation invariant.

To design provably efficient offline MARL algorithms, we need to overcome three key challenges. (i) To estimate the action-value function and the system dynamics, the approximator function needs to implement efficient relational reasoning among the agents. However, the theoretically-grounded function structure that incorporates the complex relational reasoning needs to be carefully designed. (ii) To mitigate the curse of many agents, the generalization bound of the transformer should be independent of the number of agents. Existing results in [9] thus require rethinking and improvements. (iii) In offline Reinforcement Learning (RL), the mismatch between the sampling and visitation distributions induced by the optimal policy (i.e., "distribution shift") greatly restricts the application of the offline RL algorithm. Existing works adopt the *"pessimism"* principle to mitigate such a challenge. However, this requires the quantification of the uncertainty in the value function estimation and the estimation of the dynamics in the model-free and model-based methods respectively. The quantification of the estimation error with the transformer function class is a key open question.

We organize our work by addressing the abovementioned three challenges.

First, we theoretically identify the function class that can implement complex relational reasoning. We demonstrate the relational reasoning ability of the attention mechanism by showing that approximating the self-attention structure with the permutation invariant fully-connected neural networks (i.e., deep sets [10]) requires an *exponentially large* number of hidden nodes in the input dimension of each channel (Theorem 1). This result necessitates the self-attention structure in the set transformer.

Second, we design offline model-free and model-based RL algorithms with the transformer approximators. In the former, the transformer is adopted to estimate the action-value function of the policy. The *pessimism* is encoded in that we learn the policy according to the *minimal* estimate of the action-value function in the set of functions with bounded empirical Bellman error. In the model-based algorithm, we estimate the system dynamics with the transformer structure. The policy is learned pessimistically according to the estimate of the system dynamics in the confidence region that induces the conservative value function.

Finally, we analyze the suboptimality gaps of our proposed algorithms, which indicate that the proposed algorithms mitigate the curse of many agents. For the model-free algorithm, the suboptimality gap in Theorem 3 is independent of the number of agents, which is a consequence of the fact that the generalization bound of the transformer (Theorem 2) is independent of the number of channels. For the model-based algorithm, the bound on the suboptimality gap in Theorem 4 is logarithmic in the number of agents; this follows from the analysis of the MLE of the system dynamics in Proposition 3. We emphasize that our model-based algorithm is the first provably efficient MARL algorithm that exploits the permutation equivariance when estimating the dynamics.

Technical Novelties. In Theorem 2, we leverage a PAC-Bayesian framework to derive a generalization error bound of the transformer. Compared to [9, Theorem 4.6], the result is a significant improvement in the dependence on the number of channels N and the depth of neural network L. This result may be of independent interest for enhancing our theoretical understanding of the attention mechanism and is applicable to other regression problems related to the transformer. In Proposition 3, we derive the first estimation uncertainty quantification of the system dynamics with the transformer approximators, which can be also be used to analyze other RL algorithms with such approximators.

More Related Work. In this paper, we consider the offline RL problem, and the insufficient coverage lies at the core of this problem. With the global coverage assumption, a number of works have been proposed from both the model-free [11–15] and model-based [11, 16] perspectives. To weaken the global coverage assumption, we leverage the "pessimism" principle in the algorithms: the model-free algorithms impose additional penalty terms on the estimate of the value function [17, 18] or regard the function that attains the minimum in the confidence region as the estimate of the value function [19]; the model-based algorithms estimate the system dynamics by incorporating additional penalty terms [20] or minimizing in the region around MLE [21]. For the MARL setting, the offline MARL with the mean-field approximation has been studied in [8, 22].

The analysis of the MARL algorithm with the transformer approximators requires the generalization bound of the transformer. The transformer is an element of the group equi/invariant functions, whose benefit in terms of its generalization capabilities has attracted extensive recent attention. Generalization bounds have been successively improved by analyzing the cardinality of the "effective" input field and Lipschitz constants of functions [23, 24]. However, these methods result in loose generalization bounds when applied to deep neural networks [25]. Zhu, An, and Huang [26] empirically demonstrated the benefits of the invariance in the model by refining the covering number of the function class, but a unified theoretical understanding is still lacking. The covering number of the norm-bounded transformer was shown by [9] to be at most logarithmic in the number of channels. We show that this can be further improved using a PAC-Bayesian framework. In addition, we refer to the related concurrent work [27] for a Rademacher complexity-based generalization bound of the transformer that is independent of the length of the sequence for the tasks such as computer vision.

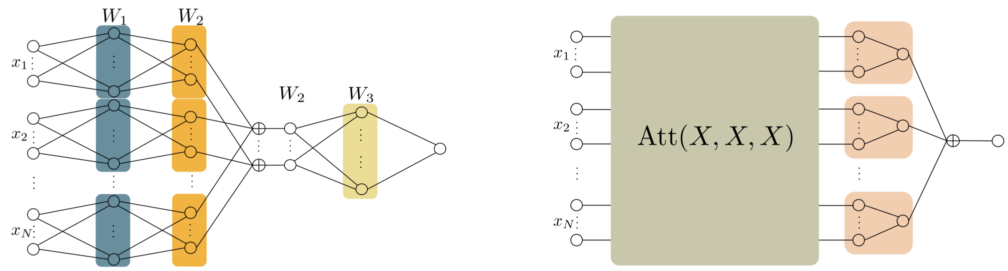
2 Preliminaries

Notation. Let $[n] = \{1, \ldots, n\}$. The i^{th} entry of the vector x is denoted as x_i or $[x]_i$. The i^{th} row and the i^{th} column of matrix X are denoted as $X_{i,:}$ and $X_{:,i}$ respectively. The *ℓ_p-norm* of the vector x is $\|x\|_p$. The *$\ell_{p,q}$-norm* of the matrix $X \in \mathbb{R}^{m\times n}$ is defined as $\|X\|_{p,q} = (\sum_{i=1}^n \|X_{:,i}\|_p^q)^{1/q}$, and the *Frobenius norm* of X is defined as $\|X\|_{\mathrm{F}} = \|X\|_{2,2}$. The *total variation distance* between two distributions P and Q on $\mathcal{A}$ is defined as $\mathrm{TV}(P, Q) = \sup_{A\subseteq\mathcal{A}} |P(A) - Q(A)|$. For a set $\mathcal{X}$, we use $\Delta(\mathcal{X})$ to denote the set of distributions on $\mathcal{X}$. For two conditional distributions $P, Q : \mathcal{X} \to \Delta(\mathcal{Y})$, the d_∞ distance between them is defined as $d_\infty(P, Q) = 2\sup_{x\in\mathcal{X}} \mathrm{TV}(P(\cdot\,|\,x), Q(\cdot\,|\,x))$. Given a metric space $(\mathcal{X}, \|\cdot\|)$, for a set $\mathcal{A} \subseteq \mathcal{X}$, an *$\varepsilon$-cover* of $\mathcal{A}$ is a finite set $\mathcal{C} \subseteq \mathcal{X}$ such that for any $a \in \mathcal{A}$, there exists $c \in \mathcal{C}$ and $\|c - a\| \leq \varepsilon$. The *$\varepsilon$-covering number* of $\mathcal{A}$ is the cardinality of the smallest ε-cover, which is denoted as $\mathcal{N}(\mathcal{A}, \varepsilon, \|\cdot\|)$.

Attention Mechanism and Transformers. The *attention mechanism* is a technique that mimics cognitive attention to process multi-channel inputs [28]. Compared with the Convolutional Neural Network (CNN), the transformer has been empirically shown to possess outstanding robustness against occlusions and preserve the global context due to its special relational structure [29]. Assume we have N query vectors that are in $\mathbb{R}^{d_Q}$. These vectors are stacked to form the matrix $Q \in \mathbb{R}^{N\times d_Q}$. With N_V *key vectors* in the matrix $K \in \mathbb{R}^{N_V\times d_Q}$ and N_V *value vectors* in the matrix $V \in \mathbb{R}^{N_V\times d_V}$, the attention mechanism maps the queries Q using the function $\mathrm{Att}(Q, K, V) = \mathrm{SM}(QK^\top)V$, where $\mathrm{SM}(\cdot)$ is the row-wise softmax operator that normalizes each row using the exponential function, i.e., for $x \in \mathbb{R}^d$, $[\mathrm{SM}(x)]_i = \exp(x_i)/\sum_{j=1}^d \exp(x_j)$ for $i \in [d]$. The product $QK^\top$ measures the similarity between the queries and the keys, which is then passed through the activation function $\mathrm{SM}(\cdot)$. Thus, $\mathrm{SM}(QK^\top)V$ essentially outputs a weighted sum of V where a value vector has greater weight if the corresponding query and key are more similar. The *self-attention mechanism* is defined as the attention that takes $Q = XW_Q$, $K = XW_K$ and $V = XW_V$ as inputs, where $X \in \mathbb{R}^{N\times d}$ is the input of self-attention, and $W_Q, W_K \in \mathbb{R}^{d\times d_Q}$ and $W_V \in \mathbb{R}^{d\times d_V}$ are the parameters. Intuitively, self-attention weighs the inputs with the correlations among N different channels. This mechanism demonstrates a special pattern of *relational reasoning* among the channels of X.

In addition, the self-attention mechanism is *permutation invariant* in the channels in X. This implies that for any row-wise permutation function $\psi(\cdot)$, which swaps the rows of the input matrix according to a given permutation of $[N]$, we have $\mathrm{Att}(\psi(X)W_Q, \psi(X)W_K, \psi(X)W_V) = \psi(\mathrm{Att}(XW_Q, XW_K, XW_V))$. The permutation equivariance of the self-attention renders it suitable for inference tasks where the output is equivariant with respect to the ordering of inputs. For example, in image segmentation, the result should be invariant to the permutation of the objects in the input image [30]. The resultant transformer structure combines the self-attention with multi-layer perceptrons and composes them to form deep neural networks. It remains permutation equi/invariant with respect to the order of the channels and has achieved excellent performance in many tasks [31–33].

Offline Cooperative MARL. In this paper, we consider the *cooperative* MARL problem, where all agents aim to maximize a *common* reward function. The corresponding MDP is characterized by the tuple $(\bar{S}_0, \bar{\mathcal{S}}, \bar{\mathcal{A}}, P^*, r, \gamma)$ and the number of agents is N. The state space $\bar{\mathcal{S}} = \mathcal{S}^N$ is the Cartesian product of the state spaces of each agent $\mathcal{S}$, and $\bar{S} = [s_1, \ldots, s_N]^\top$ is the state, where $s_i \in \mathbb{R}^{d_S}$ is the state of the i^{th} agent. The initial state is $\bar{S}_0$. The action space $\bar{\mathcal{A}} = \mathcal{A}^N$ is the Cartesian product of the action spaces $\mathcal{A}$ of each agent, and $\bar{A} = [a_1, \ldots, a_N]^\top$ is the action, where



(a) $\rho_{\mathrm{ReLU}}(\sum_{i=1}^{N} \phi_{\mathrm{ReLU}}(x_i))$ with ρ_{ReLU} and ψ_{ReLU} as single-hidden layer neural networks.

(b) Self-attention mechanism $\mathbb{I}_N^\top \mathrm{Att}(\mathrm{X}, \mathrm{X}, \mathrm{X})\mathrm{w}$.

Figure 1: The blocks with the same color share the same parameters. The left figure shows that $\rho_{\mathrm{ReLU}}(\sum_{i=1}^{N} \phi_{\mathrm{ReLU}}(x_i))$ first sums the outputs of $\phi_{\mathrm{ReLU}}(x_i)$, and it implements the relational reasoning only through the single-hidden layer network ρ_{ReLU}. In contrast, the self-attention block in the right figure captures the relationship among channels and then sums the outputs of each channel.

$a_i \in \mathbb{R}^{d_{\mathcal{A}}}$ is the action of the i^{th} agent. The transition kernel is $P^* : \mathcal{S}^N \times \mathcal{A}^N \to \Delta(\mathcal{S}^N)$, and $\gamma \in (0, 1)$ is the *discount factor*. Without loss of generality, we assume that the reward function r is deterministic and bounded, i.e., $r : \mathcal{S}^N \times \mathcal{A}^N \to [-R_{\max}, R_{\max}]$. We define the the *state-value function* $V_P^\pi : \mathcal{S}^N \to [-V_{\max}, V_{\max}]$, where $V_{\max} = R_{\max}/(1-\gamma)$, and the *action-value function* $Q_P^\pi : \mathcal{S}^N \times \mathcal{A}^N \to [-V_{\max}, V_{\max}]$ of a policy π and a transition kernel P as

$$V_P^\pi(\bar{S}) = \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t r(\bar{S}_t, \bar{A}_t) \,\middle|\, \bar{S}_0 = \bar{S}\right] \text{ and } Q_P^\pi(\bar{S}, \bar{A}) = \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t r(\bar{S}_t, \bar{A}_t) \,\middle|\, \bar{S}_0 = \bar{S}, \bar{A}_0 = \bar{A}\right],$$

respectively. Here, the expectation is taken with respect to the Markov process induced by the policy $\bar{A}_t \sim \pi(\cdot \,|\, \bar{S}_t)$ and the transition kernel P. The action-value function $Q_{P^*}^\pi$ is the unique fixed point of the operator $(\mathcal{T}^\pi f)(\bar{S}, \bar{A}) = r(\bar{S}, \bar{A}) + \gamma \mathbb{E}_{\bar{S}' \sim P^*(\cdot \,|\, \bar{S}, \bar{A})}[f(\bar{S}', \pi) \,|\, \bar{S}, \bar{A}]$, where the term in the expectation is defined as $f(\bar{S}, \pi) = \mathbb{E}_{\bar{A} \sim \pi(\cdot \,|\, \bar{S})}[f(\bar{S}, \bar{A})]$. We further define the *visitation measure* of the state and action pair induced the policy π and transition kernel P as $d_P^\pi(\bar{S}, \bar{A}) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t d_{P,t}^\pi$, where $d_{P,t}^\pi$ is the distribution of the state and the action at step t.

In offline RL, the learner only has access to a pre-collected dataset and cannot interact with the environment. The dataset $\mathcal{D} = \{(\bar{S}_i, \bar{A}_i, r_i, \bar{S}_i')\}_{i=1}^{n}$ is collected in an i.i.d. manner, i.e., $(\bar{S}_i, \bar{A}_i)$ is independently sampled from $\nu \in \Delta(\bar{\mathcal{S}} \times \bar{\mathcal{A}})$, and $\bar{S}_i' \sim P^*(\cdot \,|\, \bar{S}_i, \bar{A}_i)$. This i.i.d. assumption is made to simplify our theoretical results; see Appendix N.2 for extensions to the non i.i.d. case. Given a policy class Π, our goal is to find an optimal policy that maximizes the state-value function $\pi^* = \operatorname{argmax}_{\pi \in \Pi} V_{P^*}^\pi(\bar{S}_0)$. For any $\pi \in \Pi$, the *suboptimality gap* of π is defined as $V_{P^*}^{\pi^*}(\bar{S}_0) - V_{P^*}^\pi(\bar{S}_0)$.

3 Provable Efficiency of Transformer on Relational Reasoning

In this section, we provide the theoretical understanding of the outstanding relational reasoning ability of transformer. These theoretical results serves as a firm base for adopting set transformer to estimate the value function and system dynamics in RL algorithms in the following sections.

3.1 Relational Reasoning Superiority of Transformer Over MLP

The transformer neural network combines the self-attention mechanism and the fully-connected neural network, which includes the MultiLayer Perceptrons (MLP) function class as a subset. On the inverse direction, we show that permutation invariant MLP can not approximate transformer unless its width is exponential in the input dimension due to the poor relational reasoning ability of MLP.

Zaheer et al. [10, Theorem 2] showed that all permutation invariant functions take the form $\rho(\sum_{i=1}^{N} \phi(x_i))$ with $X = [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N \times d}$ as the input. Since the single-hidden layer ReLU neural network is an universal approximator for continuous functions [34], we set $\phi : \mathbb{R}^{N \times d} \to \mathbb{R}^{W_2}$ and $\rho : \mathbb{R}^{W_2} \to \mathbb{R}$ to be single-hidden layer neural networks with ReLU activation functions as shown in Figure 1(a), where W_2 is the dimension of the intermediate outputs. The widths of the hidden layers in ϕ_{ReLU} and ρ_{ReLU} are W_1 and W_3 respectively. For the formal definition of ϕ_{ReLU} and ρ_{ReLU}, please refer to Appendix A. Then the function class with ρ_{ReLU} and ϕ_{ReLU} as width-constrained ReLU networks is defined as

$$\mathcal{N}(W) = \left\{ f : \mathbb{R}^{N\times d} \to \mathbb{R} \;\middle|\; f(X) = \rho_{\text{ReLU}}\Big(\sum_{i=1}^{N} \phi_{\text{ReLU}}(x_i)\Big) \text{ with } \max_{i\in[3]} W_i \le W \right\}.$$

We would like to use functions in $\mathcal{N}(W)$ to approximate the self-attention function class

$$\mathcal{F} = \big\{ f : \mathbb{R}^{N\times d} \to \mathbb{R} \;\big|\; f(X) = \mathbb{I}_N^\top \text{Att}(X, X, X) w \text{ for some } w \in [0,1]^d \big\}.$$

Figure 1(a) shows that $\rho_{\text{ReLU}}(\sum_{i=1}^N \phi_{\text{ReLU}}(x_i))$ first processes each channel with ϕ_{ReLU}, and the relationship between channels is only reasoned with ρ_{ReLU}. The captured relationship in $\rho_{\text{ReLU}}(\sum_{i=1}^N \phi_{\text{ReLU}}(x_i))$ cannot be too complex due to the simple structure of ρ_{ReLU}. In contrast, the self-attention structure shown in Figure 1(b) first captures the relationship between channels with the self-attention structure and then weighs the results to derive the final output. Consequently, it is difficult to approximate the self-attention structure with $\rho_{\text{ReLU}}(\sum_{i=1}^N \phi_{\text{ReLU}}(x_i))$ due to its poor relational reasoning ability. This observation is formally quantified in the following theorem.

Theorem 1. *Let $W^*(\xi, d, \mathcal{F})$ be the smallest width of the neural network such that*

$$\forall\, f \in \mathcal{F},\ \exists\, g \in \mathcal{N}(W) \quad \textit{s.t.} \quad \sup_{X\in[0,1]^{N\times d}} \big| f(X) - g(X) \big| \le \xi.$$

With sufficient number of channels N, it holds that $W^*(\xi, d, \mathcal{F}) = \Omega(\exp(cd)\xi^{-1/4})$ for some $c > 0$.

Theorem 1 shows that the fully-connected neural network cannot approximate the relational reasoning process in the self-attention mechanism unless the width is exponential in the input dimension. This exponential lower bound of the width of the fully-connected neural network implies that the relational reasoning process embedded within the self-attention structure is complicated, and it further motivates us to explicitly incorporate the self-attention structure in the neural networks in order to reason the complex relationship among the channels.

3.2 Channel Number-independent Generalization Error Bound

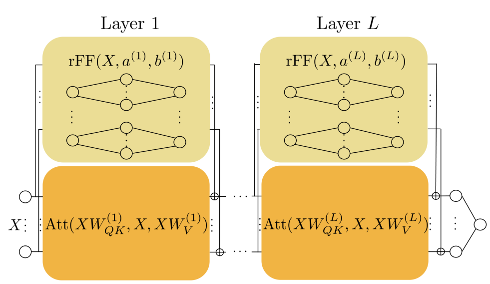


Figure 2: Structure of the transformer function class, where the row-wise feedforward function is specified as fully-connected networks.

In this section, we derive the generalization error bound of transformer. We take $X \in \mathbb{R}^{N\times d}$ as the input of the neural network. In the i^{th} layer, as shown in Figure 3.2, we combine the self-attention mechanism $\text{Att}(XW_{QK}^{(i)}, X, XW_V^{(i)})$ with the row-wise FeedForward (rFF) single-hidden layer neural network $\text{rFF}(X, a^{(i)}, b^{(i)})$ with width m. We combine $W_Q^{(i)}$ and $W_K^{(i)}$ to $W_{QK}^{(i)}$ for ease of calculation, and $b^{(i)}$ and $a^{(i)}$ are the parameters of the first and second layer of rFF. The output of each layer is normalized by the row-wise normalization function $\Pi_{\text{norm}}(\cdot)$, which projects each row of the input into the unit ℓ_p-ball (for some $p \ge 1$). For the last layer, we derive the scalar estimate of the action-value function by averaging the outputs of all the channels, and the "clipping" function $\Pi_V(x)$ is applied to normalize the output to $[-V, V]$. We note that such structures are also known as *set transformers* in [33]. For the formal definition of the transformer, please refer to Appendix B.

We consider a transformer with bounded parameters. For a pair of conjugate numbers $p, q \in \mathbb{R}$, i.e., $1/p + 1/q = 1$ and $p, q \ge 1$, the transformer function class with bounded parameters is defined as

$$\mathcal{F}_{\text{tf}}(B) = \Big\{ g_{\text{tf}}(X; W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}, w) \;\Big|\; \big|a_{kj}^{(i)}\big| < B_a, \big\|b_{kj}^{(i)}\big\|_q < B_b,$$
$$\big\|W_{QK}^{(i)\top}\big\|_{p,q} < B_{QK}, \big\|W_V^{(i)\top}\big\|_{p,q} < B_V, \|w\|_q < B_w \text{ for } i \in [L], j \in [m], k \in [d] \Big\},$$

where $B = [B_a, B_b, B_{QK}, B_V, B_w]$ are the parameters of the function class, and $W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}$ and $b^{1:L}$ are the stacked parameters in each layer. We only consider the non-trivial case where

$B_a, B_b, B_{QK}, B_V, B_w$ are larger than one, otherwise the norms of the outputs decrease exponentially with growing depth. For ease of notation, we denote $\mathcal{F}_{\mathrm{tf}}(B)$ as $\mathcal{F}_{\mathrm{tf}}$ when the parameters are clear.

Consider the regression problem where we aim to predict the value of the response variable $y \in \mathbb{R}$ from the observation matrix $X \in \mathbb{R}^{N\times d}$, where $(X, y) \sim \nu$, and $|y| \le V$. We derive our estimate $f : \mathbb{R}^{N\times d} \to \mathbb{R}$ from i.i.d. observations $\mathcal{D}_{\mathrm{reg}} = \{(X_i, y_i)\}_{i=1}^n$ generated from ν. The *risk* of using $f \in \mathcal{F}_{\mathrm{tf}}(B)$ as a regressor on sample (X, y) is defined as $(f(X) - y)^2$. Then the *excess risk* of functions in the transformer function class $\mathcal{F}_{\mathrm{tf}}$ can be bounded as in the following proposition.

Proposition 1. *Let $\bar{B} = B_V B_{QK} B_a B_b B_w$. For all $f \in \mathcal{F}_{\mathrm{tf}}$, with probability at least $1-\delta$, we have*

$$\begin{aligned}&\left|\mathbb{E}_\nu\Big[\big(f(X)-y\big)^2\Big] - \frac{1}{n}\sum_{i=1}^n \big(f(X_i)-y_i\big)^2\right|\\ &\le \frac{1}{2}\mathbb{E}_\nu\Big[\big(f(X)-y\big)^2\Big] + O\bigg(\frac{V^2}{n}\bigg[mL^2d^2\log\frac{mdL\bar{B}n}{V} + \log\frac{1}{\delta}\bigg]\bigg).\end{aligned}$$

Proposition 1 is a corollary of Theorem 2. We state it here since the generalization error bound of transformer may be interesting for other regression problems. We compare our generalization error bound in Proposition 1 with [9, Theorem 4.6]. For the dependence on the number of agents N, the result in [9, Theorem 4.6] shows that the logarithm of the covering number of the transformer function class is logarithmic in N. Combined with the use of the Dudley integral [35], [9, Theorem 4.6] implies that the generalization error bound is logarithmic in N. In contrast, our result is independent of N. This superiority is attributed to our use of the PAC-Bayesian framework, in which we measure the distance between functions using the KL divergence of the distributions on the function parameter space. For the transformer structure, the size of the parameter space is independent of the number of agents N, which helps us to remove the dependence on N.

Concerning the dependence on the depth L of the neural network, [9, Theorem 4.6] shows that the logarithm of the covering number of the transformer function class scales exponentially in L. In contrast, Proposition 1 shows that the generalization bound is *polynomial* in L. We note that Proposition 1 does not contradict the exponential dependence shown in [36, 37], since we implement the layer normalization to restrict the range of the output. As a byproduct, Proposition 1 shows that the invariant of the layer normalization adopted in our paper can greatly reduce the dependence of the generalization error on the depth of the neural network L. We note that our results can be generalized to the multi-head attention structure, and the extensions are provided in Appendix N.

4 Offline Multi-Agent Reinforcement Learning with Set Transformers

In this section, we apply the results in Section 3 to MARL. We implement efficient relational reasoning via the set transformer to obtain improved suboptimality bounds of the MARL problem. In particular, we consider the *homogeneous* MDP, where the transition kernel and the reward function are invariant to permutations of the agents, i.e., for any row-wise permutation function $\psi(\cdot)$, we have

$$P^*(\bar{S}' \,|\, \bar{S}, \bar{A}) = P^*\big(\psi(\bar{S}') \,\big|\, \psi(\bar{S}), \psi(\bar{A})\big) \quad \text{and} \quad r(\bar{S}, \bar{A}) = r\big(\psi(\bar{S}), \psi(\bar{A})\big)$$

for all $\bar{S}, \bar{S}' \in \mathcal{S}^N$ and $\bar{A} \in \mathcal{A}^N$. A key property of the homogeneous MDP is that there exists a permutation invariant optimal policy, and the corresponding state-value function and the action-value function are also permutation invariant [22].

Proposition 2. *For the cooperative homogeneous MDP, there exists an optimal policy that is permutation invariant. Also, for any permutation invariant policy π, the corresponding value function $V^\pi_{P^*}$ and action-value function $Q^\pi_{P^*}$ are permutation invariant.*

Thus, we restrict our attention to the class of permutation invariant policies Π, where $\pi(\bar{A} \,|\, \bar{S}) = \pi(\psi(\bar{A}) \,|\, \psi(\bar{S}))$ for all $\bar{A} \in \bar{\mathcal{A}}$, $\bar{S} \in \bar{\mathcal{S}}$, $\pi \in \Pi$ and all permutations ψ. For example, if $\pi(\bar{A} \,|\, \bar{S}) = \prod_{i=1}^N \mu(a_i \,|\, s_i)$ for some μ, then π is permutation invariant. An optimal policy is any $\pi^* \in \mathrm{argmax}_{\pi\in\Pi} V^\pi_{P^*}(\bar{S}_0)$.

4.1 Pessimistic Model-Free Offline Reinforcement Learning

In this subsection, we present a model-free algorithm, in which we adopt the transformer to estimate the action-value function. We also learn a policy based on such an estimate.

4.1.1 Algorithm

We modify the single-agent offline RL algorithm in [19] to be applicable to the multi-agent case with the transformer approximators, but the analysis is rather different from that in [19]. Given the dataset $\mathcal{D} = \{(\bar{S}_i, \bar{A}_i, r_i, \bar{S}_i')\}_{i=1}^n$, we define the mismatch between two functions f and $\tilde{f}$ on $\mathcal{D}$ for a fixed policy π as $\mathcal{L}(f, \tilde{f}, \pi; \mathcal{D}) = \frac{1}{n}\sum_{(\bar{S},\bar{A},\bar{r},\bar{S}')\in\mathcal{D}}(f(\bar{S},\bar{A}) - \bar{r} - \gamma\tilde{f}(\bar{S}',\pi))^2$. We adopt the transformer function class $\mathcal{F}_{\mathrm{tf}}(B)$ in Section 3.2 to estimate the action-value function and regard $X = [\bar{S}, \bar{A}] \in \mathbb{R}^{N\times d}$ as the input of the neural network. The dimension $d = d_{\mathcal{S}} + d_{\mathcal{A}}$ and each agent corresponds to a channel in X. The *Bellman error* of a function f with respect to the policy π is defined as $\mathcal{E}(f, \pi; \mathcal{D}) = \mathcal{L}(f, f, \pi; \mathcal{D}) - \inf_{\tilde{f}\in\mathcal{F}_{\mathrm{tf}}}\mathcal{L}(\tilde{f}, f, \pi; \mathcal{D})$.

For a fixed policy π, we construct the confidence region of the action-value function of π by selecting the functions in $\mathcal{F}_{\mathrm{tf}}$ with the ε-controlled Bellman error. We regard the function attaining the minimum in the confidence region as the estimate of the action-value function of the policy; this reflects the terminology "pessimism". Then the optimal policy is learned by maximizing the action-value function estimate. The algorithm can be written formally as

$$\hat{\pi} = \operatorname*{argmax}_{\pi\in\Pi} \min_{f\in\mathcal{F}(\pi,\varepsilon)} f(\bar{S}_0, \pi), \quad \text{where} \quad \mathcal{F}(\pi, \varepsilon) = \big\{f \in \mathcal{F}_{\mathrm{tf}}(B) \,\big|\, \mathcal{E}(f, \pi; \mathcal{D}) \le \varepsilon\big\}. \tag{1}$$

The motivation for the pessimism originates from the *distribution shift*, where the induced distribution of the learned policy is different from the sampling distribution ν. Such an issue is severe when there is no guarantee that the sampling distribution ν supports the visitation distribution $d_{P^*}^{\pi^*}$ induced by the optimal policy π^*. In fact, the algorithm in Eqn. (1) does not require the *global* coverage of the sampling distribution ν, where the global coverage means that $d_{P^*}^{\pi}(\bar{S},\bar{A})/\nu(\bar{S},\bar{A})$ is upper bounded by some constant for all $(\bar{S},\bar{A}) \in \bar{\mathcal{S}}\times\bar{\mathcal{A}}$ and all $\pi\in\Pi$. Instead, it only requires *partial* coverage, and the mismatch between the distribution induced by the optimal policy $d_{P^*}^{\pi^*}$ and the sampling distribution ν is captured by

$$C_{\mathcal{F}_{\mathrm{tf}}} = \max_{f\in\mathcal{F}_{\mathrm{tf}}} \mathbb{E}_{d_{P^*}^{\pi^*}}\big[\big(f(\bar{S},\bar{A}) - \mathcal{T}^{\pi^*}f(\bar{S},\bar{A})\big)^2\big]/\mathbb{E}_\nu\big[\big(f(\bar{S},\bar{A}) - \mathcal{T}^{\pi^*}f(\bar{S},\bar{A})\big)^2\big]. \tag{2}$$

We note that $C_{\mathcal{F}_{\mathrm{tf}}} \le \max_{(\bar{S},\bar{A})\in\bar{\mathcal{S}}\times\bar{\mathcal{A}}} d_{P^*}^{\pi^*}(\bar{S},\bar{A})/\nu(\bar{S},\bar{A})$, so the suboptimality bound involving $C_{\mathcal{F}_{\mathrm{tf}}}$ in Theorem 3 is tighter than the bound requiring global convergence [38]. Similar coefficients also appear in many existing works such as [19] and [39].

4.1.2 Bound on the Suboptimality Gap

Before stating the suboptimality bound, We require two assumptions on $\mathcal{F}_{\mathrm{tf}}$ and the sampling distribution ν. We first state the standard regularity assumption of the transformer function class.

Assumption 1. *For any $\pi\in\Pi$, we have $\inf_{f\in\mathcal{F}_{\mathrm{tf}}}\sup_{\mu\in d_\Pi}\mathbb{E}_\mu[(f(\bar{S},\bar{A}) - \mathcal{T}^\pi f(\bar{S},\bar{A}))^2] \le \varepsilon_{\mathcal{F}}$ and $\sup_{f\in\mathcal{F}_{\mathrm{tf}}}\inf_{\tilde{f}\in\mathcal{F}_{\mathrm{tf}}}\mathbb{E}_\nu[(\tilde{f}(\bar{S},\bar{A}) - \mathcal{T}^\pi f(\bar{S},\bar{A}))^2] \le \varepsilon_{\mathcal{F},\mathcal{F}}$, where $d_\Pi = \{\mu \mid \exists\, \pi\in\Pi$ s.t. $\mu = d_{P^*}^{\pi}\}$ is the set of distributions of the state and the action pair induced by any policy $\pi\in\Pi$.*

This assumption, including the *realizability* and the *completeness*, states that for any policy $\pi\in\Pi$ there is a function in the transformer function class $\mathcal{F}_{\mathrm{tf}}$ such that the Bellman error is controlled by $\varepsilon_{\mathcal{F}}$, and the transformer function class is approximately closed under the Bellman operator $\mathcal{T}^\pi$ for any $\pi\in\Pi$. In addition, we require that the mismatch between the sampling distribution and the visitation distribution of the optimal policy is bounded.

Assumption 2. *For the sampling distribution ν, the coefficient $C_{\mathcal{F}_{\mathrm{tf}}}$ defined in Eqn.* (2) *is finite.*

We note that similar assumptions also appear in many existing works [19, 39].

In the analysis of the algorithm in Eqn. (1), we first derive a generalization error bound of the estimate of the Bellman error using the PAC-Bayesian framework [40, 41].

Theorem 2. *Let $\bar{B} = B_V B_{QK} B_a B_b B_w$. For all $f, \tilde{f} \in \mathcal{F}_{\mathrm{tf}}(B)$ and all policies $\pi\in\Pi$, with probability at least $1-\delta$, we have*

$$\Big|\mathbb{E}_\nu\Big[\big(f(\bar{S},\bar{A}) - \mathcal{T}^\pi\tilde{f}(\bar{S},\bar{A})\big)^2\Big] - \mathcal{L}(f,\tilde{f},\pi;\mathcal{D}) + \mathcal{L}(\mathcal{T}^\pi\tilde{f},\tilde{f},\pi;\mathcal{D})\Big|$$
$$\le \frac{1}{2}\mathbb{E}_\nu\Big[\big(f(\bar{S},\bar{A}) - \mathcal{T}^\pi\tilde{f}(\bar{S},\bar{A})\big)^2\Big] + O\bigg(\frac{V_{\max}^2}{n}\Big[mL^2d^2\log\frac{mdL\bar{B}n}{V_{\max}} + \log\frac{\mathcal{N}(\Pi,1/n,d_\infty)}{\delta}\Big]\bigg).$$

For ease of notation, we define $e(\mathcal{F}_{\mathrm{tf}}, \Pi, \delta, n)$ to be n times the second term of the generalization error bound. We note that the generalization error bound in Theorem 2 is independent of the number of agents, which will help us to remove the dependence on the number of agents in the suboptimality of the learned policy. The suboptimality gap of the learned policy $\hat{\pi}$ can be upper bounded as the following.

Theorem 3. *If Assumptions 1 and 2 hold, and we take $\varepsilon = 3\varepsilon_{\mathcal{F}}/2 + 2e(\mathcal{F}_{\mathrm{tf}}, \Pi, \delta, n)/n$, then with probability at least $1-\delta$, the suboptimality gap of the policy derived in the algorithm shown in Eqn.* (1) *is upper bounded as*

$$V_{P^*}^{\pi^*}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0) \leq O\left(\frac{\sqrt{C_{\mathcal{F}_{\mathrm{tf}}}\tilde{\varepsilon}}}{1-\gamma} + \frac{V_{\max}\sqrt{C_{\mathcal{F}_{\mathrm{tf}}}}}{(1-\gamma)\sqrt{n}}\sqrt{mL^2d^2\log\frac{mdL\bar{B}n}{V_{\max}} + \log\frac{2\mathcal{N}(\Pi, 1/n, d_\infty)}{\delta}}\right),$$

where $d = d_{\mathcal{S}} + d_{\mathcal{A}}$, $\tilde{\varepsilon} = \varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}$, and $\bar{B}$ is defined in Proposition 2.

Theorem 3 shows that the upper bound of the suboptimality gap does not scale with the number of agents N, which demonstrates that the proposed model-free algorithm breaks the curse of many agents. We note that the model-free offline/batch MARL with homogeneous agents has been studied in [8] and [22], and the suboptimality upper bounds in [8, Theorem 1] and [22, Theorem 4.1] are also independent of N. However, these works adopt the mean-field approximation of the original MDP, in which the influence of all the other agents on a specific agent is only coarsely considered through the distribution of the state. The approximation error between the action-value function of the mean-field MDP and that of the original MDP is not analyzed therein. Thus, the independence of N in their works comes with the cost of the poor relational reasoning ability and the unspecified approximation error. In contrast, we analyze the suboptimality gap of the learned policy in the original MDP, and the interaction among agents is captured by the transformer network.

4.2 Pessimistic Model-based Offline Reinforcement Learning

In this subsection, we present the model-based algorithm, where we adopt the transformer to estimate the system dynamics and learn the policy based on such an estimate.

4.2.1 Neural Nonlinear Regulator

In this section, we consider the Neural Nonlinear Regulator (NNR), in which we use the transformer to estimate the system dynamics. The ground truth transition $P^*(\bar{S}' \mid \bar{S}, \bar{A})$ is defined as $\bar{S}' = F^*(\bar{S}, \bar{A}) + \bar{\varepsilon}$, where F^* is a nonlinear function, $\bar{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_N]^\top$ is the noise, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{d\times d})$ for $i \in [N]$ are independent random vectors. We note that the function F^* and the transition kernel P^* are equivalent, and we denote the transition kernel corresponding to the function F as P_F. Since the transition kernel $P^*(\bar{S}' \mid \bar{S}, \bar{A})$ is permutation invariant, F^* should be permutation equivariant, i.e., $F^*(\psi(\bar{S}), \psi(\bar{A})) = \psi(F^*(\bar{S}, \bar{A}))$ for all row-wise permutation functions $\psi(\cdot)$.

We take $X = [\bar{S}, \bar{A}] \in \mathbb{R}^{N\times d}$ as the input of the network and adopt a similar network structure as the transformer specified in Section 3.2. However, to predict the next state instead of the action-value function with the transformer, we remove the average aggregation module in the final layer of the structure in Section 3.2. Please refer to Appendix B for the formal definition. The permutation equivariance of the proposed transformer structure can be easily proved with the permutation equivariance of the self-attention mechanism. We consider the transformer function class with bounded parameters, which is defined as

$$\mathcal{M}_{\mathrm{tf}}(B') = \Big\{F_{\mathrm{tf}}(X; W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}) \,\Big|\, \big|a_{kj}^{(i)}\big| < B_a, \big\|b_{kj}^{(i)}\big\|_2 < B_b,$$
$$\big\|W_{QK}^{(i)\top}\big\|_{\mathrm{F}} < B_{QK}, \big\|W_V^{(i)\top}\big\|_{\mathrm{F}} < B_V \text{ for } i \in [L], j \in [m], k \in [d]\Big\},$$

where $B' = [B_a, B_b, B_{QK}, B_V]$ is the vector of parameters of the function class. We denote $\mathcal{M}_{\mathrm{tf}}(B')$ as $\mathcal{M}_{\mathrm{tf}}$ when the parameters are clear from the context.

4.2.2 Algorithm

Given the offline dataset $\mathcal{D} = \{(\bar{S}_i, \bar{A}_i, r_i, \bar{S}_i')\}_{i=1}^n$, we first derive the MLE of the system dynamics. Next, we learn the optimal policy according to the confidence region of the dynamics that are constructed around the MLE. The term "pessimism" is reflected in the procedure that we choose the system dynamics that induce the *smallest* value function, i.e.,

$$\hat{F}_{\mathrm{MLE}} = \operatorname*{argmin}_{F\in\mathcal{M}_{\mathrm{tf}}} \frac{1}{n}\sum_{i=1}^{n} \big\|\bar{S}_i' - F(\bar{S}_i, \bar{A}_i)\big\|_{\mathrm{F}}^2 \quad \text{and} \quad \hat{\pi} = \operatorname*{argmax}_{\pi\in\Pi} \min_{F\in\mathcal{M}_{\mathrm{MLE}}(\zeta)} V_{P_F}^{\pi}(\bar{S}_0), \tag{3}$$

where $\mathcal{M}_{\mathrm{MLE}}(\zeta) = \{F \in \mathcal{M}_{\mathrm{tf}}(B') \,|\, 1/n \cdot \sum_{i=1}^{n} \mathrm{TV}(P_F(\cdot \,|\, \bar{S}_i, \bar{A}_i), \hat{P}_{\mathrm{MLE}}(\cdot \,|\, \bar{S}_i, \bar{A}_i))^2 \le \zeta\}$ is the confidence region, which has a closed-form expression in terms of the difference between F and $\hat{F}_{\mathrm{MLE}}$ as stated in in Appendix C. The transition kernel induced by $\hat{F}_{\mathrm{MLE}}$ is denoted as $\hat{P}_{\mathrm{MLE}}$. The parameter ζ is used to measure the tolerance of estimation error of the system dynamics, and it is set according to the parameters of $\mathcal{M}_{\mathrm{tf}}(B')$ such that F^* belongs to $\mathcal{M}_{\mathrm{MLE}}(\zeta)$ with high probability.

Similar to the model-free algorithm, the model-based algorithm specified in Eqn. (3) does not require global coverage. Instead, the mismatch between the distribution induced by the optimal policy $d_{P^*}^{\pi^*}$ and the sampling distribution ν is captured by the constant

$$C_{\mathcal{M}_{\mathrm{tf}}} = \max_{F\in\mathcal{M}_{\mathrm{tf}}} \mathbb{E}_{d_{P^*}^{\pi^*}}\big[\mathrm{TV}\big(P_F(\cdot \,|\, \bar{S}, \bar{A}), P^*(\cdot \,|\, \bar{S}, \bar{A})\big)^2\big] \Big/ \mathbb{E}_{\nu}\big[\mathrm{TV}\big(P_F(\cdot \,|\, \bar{S}, \bar{A}), P^*(\cdot \,|\, \bar{S}, \bar{A})\big)^2\big]. \tag{4}$$

We note that $C_{\mathcal{M}_{\mathrm{tf}}} \le \max_{(\bar{S},\bar{A})\in\bar{\mathcal{S}}\times\bar{\mathcal{A}}} d_{P^*}^{\pi^*}(\bar{S}, \bar{A})/\nu(\bar{S}, \bar{A})$, so the suboptimality bound involving $C_{\mathcal{P}_{\mathcal{F}_{\mathrm{tf}}}}$ in Theorem 4 is tighter than the bound requiring global convergence. Similar coefficients also appear in many existing works such as [42] and [20].

4.2.3 Analysis of the Maximum Likelihood Estimate

Every $F \in \mathcal{M}_{\mathrm{MLE}}(\zeta)$ is near to the MLE in the total variation sense and thus well approximates the ground truth system dynamics. Therefore, to derive an upper bound of the suboptimality gap of the learned policy, we first analyze the convergence rate of the MLE $\hat{P}_{\mathrm{MLE}}$ to P^*.

Proposition 3. *Let $\tilde{B} = B_V B_{QK} B_a B_b$. For the maximum likelihood estimate $\hat{P}_{\mathrm{MLE}}$ in Eqn.* (3)*, the following inequality holds with probability at least $1-\delta$,*

$$\mathbb{E}_{\nu}\Big[\mathrm{TV}\big(P^*(\cdot \,|\, \bar{S}, \bar{A}), \hat{P}_{\mathrm{MLE}}(\cdot \,|\, \bar{S}, \bar{A})\big)^2\Big] \le O\bigg(\frac{1}{n} mL^2 d^2 \log\big(NLmd\tilde{B}n\big) + \frac{1}{n}\log\frac{1}{\delta}\bigg).$$

We define $e'(\mathcal{M}_{\mathrm{tf}}, n)$ to be n times the total variation bound. Proposition 3 shows that the total variation estimation error is polynomial in the depth of the neural network L. However, different from the model-free RL results in Section 4.1, the estimation error of MLE $\hat{P}_{\mathrm{MLE}}$ is logarithmic in the number of agents N. We note that this logarithm dependency on N comes from the fact that $\mathrm{TV}(P^*(\cdot \,|\, \bar{S}, \bar{A}), \hat{P}_{\mathrm{MLE}}(\cdot \,|\, \bar{S}, \bar{A}))$ measures the distance between two transition kernels that involves the states of N agents, different from the scalar estimate of the value function in Section 4.1. To prove the result, we adopt a PAC-Bayesian framework to analyze the convergence rate of MLE, which is inspired by the analysis of density estimation [43]; more details are presented in Appendix J.

4.2.4 Bound on the Suboptimality Gap

To analyze the error of the learned model, we make the following realizability assumption.

Assumption 3. *The nominal system dynamics belongs to the function class $\mathcal{M}_{\mathrm{tf}}$, i.e., $F^* \in \mathcal{M}_{\mathrm{tf}}(B')$.*

In addition, we require that the mismatch between the sampling distribution and the visitation distribution of the optimal policy is bounded.

Assumption 4. *For the sampling distribution ν, the coefficient $C_{\mathcal{M}_{\mathrm{tf}}}$ defined in* (4) *is finite.*

We note that these two assumptions are also made in many existing works, e.g., [20, 21].

Theorem 4. *If Assumptions 3 and 4 hold, and we take $\zeta = c_1 e'(\mathcal{M}_{\mathrm{tf}}, n)/n$ for some constant $c_1 > 0$, then with probability at least $1-\delta$, the suboptimality gap of the policy learned in the algorithm in Eqn.* (3) *is upper bounded as*

$$V_{P^*}^{\pi^*}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0) \le O\Bigg(\frac{V_{\max}}{(1-\gamma)^2}\sqrt{C_{\mathcal{M}_{\mathrm{tf}}}\bigg(\frac{1}{n} mL^2 d^2 \log\big(NLmd\tilde{B}n\big) + \frac{1}{n}\log\frac{1}{\delta}\bigg)}\Bigg),$$

where $d = d_{\mathcal{S}} + d_{\mathcal{A}}$, and $\tilde{B}$ is defined in Proposition 3.

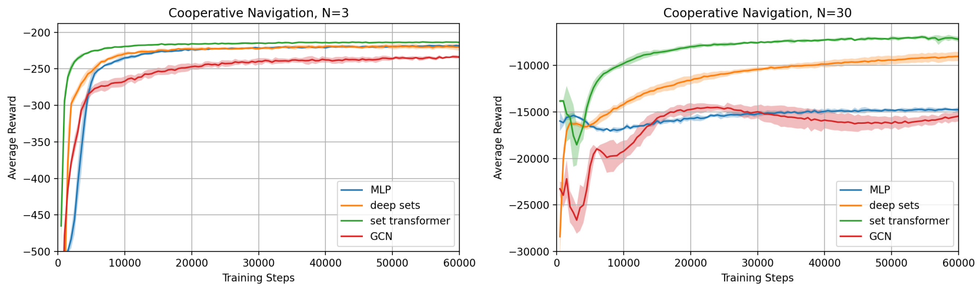


Figure 3: Average rewards of model-free RL algorithms with their standard deviations for $N = 3, 30$.

Theorem 4 presents an upper bound on the suboptimality gap of the offline model-based RL with the transformer approximators. The suboptimality gap depends on the number of agents only as $O(\sqrt{\log N})$, which shows that the proposed model-based MARL algorithm mitigates the curse of many agents. This weak dependence on N originates from measuring the distance between two system dynamics of N agents in the learning of the dynamics. To the best of our knowledge, there is no prior work on analyzing the model-based algorithm for the homogeneous MARL, even from the mean-field approximation perspective. The proof of Theorem 4 leverages novel analysis of the MLE in Proposition 3. For more details, please refer to Appendix H.

5 Experimental Results

We evaluate the performance of the algorithms on the Multiple Particle Environment (MPE) [44, 45]. We focus on the *cooperative navigation* task, where N agents move cooperatively to cover L landmarks in an environment. Given the positions of the N agents $x_i \in \mathbb{R}^2$ (for $i \in [N]$) and the positions of the L landmarks $y_j \in \mathbb{R}^2$ (for $j \in [L]$), the agents receive reward $r = -\sum_{j=1}^{L} \min_{i\in[N]} \|y_j - x_i\|_2$. This reward encourages the agents to move closer to the landmarks. We set the number of agents as $N = 3, 6, 15, 30$ and the number of landmarks as $L = N$. Here, we only present the result for $N = 3, 30$. Please refer to Appendix O for more numerical results. To collect an offline dataset, we learn a policy in the online setting. Then the offline dataset is collected from the induced stationary distribution of such a policy. We use MLP, deep sets, Graph Convolutional Network (GCN) [46], and set transformer to estimate the value function. We note that the deep sets, GCN, and set transformer are permutation invariant functions. For the implementation details, please refer to Appendix O.

Figure 3 shows that the performances of the MLP and deep sets are worse than that of the set transformer. This is due to the poor relational reasoning abilities of MLP and deep sets, which corroborates Theorem 1. Figure 3 indicates that when the number of agents N increases, the superiority of the algorithm with set transformer becomes more pronounced, which is strongly aligned with our theoretical result in Theorem 3.

6 Concluding remarks

In view of the tremendous empirical successes of cooperative MARL with permutation invariant agents, it is imperative to develop a firm theoretical understanding of this MARL problem because it will inspire the design of even more efficient algorithms. In this work, we design and analyze algorithms that break the curse of many agents and, at the same time, implement efficient relational reasoning. Our algorithms and analyses serve as a first step towards developing provably efficient MARL algorithms with permutation invariant approximators. We leave the extension of our results of the transformer to *general permutation invariant approximators* as future works.

Acknowledgments and Disclosure of Funding

Fengzhuo Zhang and Vincent Tan acknowledge funding by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-018) and by Singapore Ministry of Education (MOE) AcRF Tier 1 Grants (A-0009042-01-00 and A-8000189-01-00). Zhaoran Wang acknowledges the National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J. P. Morgan, and Two Sigma for their support.

References

[1] D. Wang, R. Walters, X. Zhu, and R. Platt. Equivariant Q Learning in Spatial Action Spaces. In *Conference on Robot Learning*, pages 1713–1723. PMLR, 2022.

[2] D. Wang, R. Walters, and R. Platt. $\mathrm{SO}(2)$ -equivariant reinforcement learning. *arXiv preprint arXiv:2203.04439*, 2022.

[3] Y. Tang and D. Ha. The sensory neuron as a transformer: Permutation-invariant neural networks for reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[5] K. Xu, Y. Zhang, D. Ye, P. Zhao, and M. Tan. Relation-aware transformer for portfolio policy learning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4647–4653, 2021.

[6] L. Wang, Z. Yang, and Z. Wang. Breaking the curse of many agents: Provable mean embedding q-iteration for mean-field reinforcement learning. In *International Conference on Machine Learning*, pages 10092–10103. PMLR, 2020.

[7] K. Menda, Y. Chen, J. Grana, J. W. Bono, B. D. Tracey, M. J. Kochenderfer, and D. Wolpert. Deep reinforcement learning for event-driven multi-agent decision processes. *IEEE Transactions on Intelligent Transportation Systems*, 20(4):1259–1268, 2018.

[8] M. Chen, Y. Li, E. Wang, Z. Yang, Z. Wang, and T. Zhao. Pessimism meets invariance: Provably efficient offline mean-field multi-agent RL. *Advances in Neural Information Processing Systems*, 34, 2021.

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

[10] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. *Advances in Neural Information Processing Systems*, 30, 2017.

[11] J. Chen and N. Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

[12] A. Antos, C. Szepesvári, and R. Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

[13] O. Nachum, B. Dai, I. Kostrikov, Y. Chow, L. Li, and D. Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

[14] Y. Duan, C. Jin, and Z. Li. Risk bounds and rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*, pages 2892–2902. PMLR, 2021.

[15] T. Xie and N. Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.

[16] S. Ross and J. A. Bagnell. Agnostic system identification for model-based reinforcement learning. *arXiv preprint arXiv:1203.1007*, 2012.

[17] Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[18] N. Rajaraman, L. Yang, J. Jiao, and K. Ramchandran. Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33:2914–2924, 2020.

[19] T. Xie, C. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[20] J. Chang, M. Uehara, D. Sreenivas, R. Kidambi, and W. Sun. Mitigating covariate shift in imitation learning via offline data with partial coverage. *Advances in Neural Information Processing Systems*, 34, 2021.

[21] M. Uehara and W. Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

[22] Y. Li, L. Wang, J. Yang, E. Wang, Z. Wang, T. Zhao, and H. Zha. Permutation invariant policy optimization for mean-field multi-agent reinforcement learning: A principled approach. *arXiv preprint arXiv:2105.08268*, 2021.

[23] J. Sokolic, R. Giryes, G. Sapiro, and M. Rodrigues. Generalization error of invariant classifiers. In *Artificial Intelligence and Statistics*, pages 1094–1103. PMLR, 2017.

[24] A. Sannai, M. Imaizumi, and M. Kawano. Improved generalization bounds of group invariant/equivariant deep networks via quotient feature spaces. In *Uncertainty in Artificial Intelligence*, pages 771–780. PMLR, 2021.

[25] D. Jakubovitz, R. Giryes, and M. R. D. Rodrigues. Generalization error in deep learning. In *Compressed Sensing and Its Applications*, pages 153–193. Springer, 2019.

[26] S. Zhu, B. An, and F. Huang. Understanding the generalization benefit of model invariance from a data perspective. *Advances in Neural Information Processing Systems*, 34, 2021.

[27] Anonymous. An analysis of attention via the lens of invariance: Approximation, generalization, and optimization. *Technical Report*, 2022.

[28] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[29] M. M. Naseer, K. Ranasinghe, S. H Khan, M. Hayat, F. Shahbaz Khan, and M. Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

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

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *In International Conference on Learning Representations*, 2021.

[32] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on Imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021.

[33] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.

[34] S. Sonoda and N. Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268, 2017.

[35] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT press, 2018.

[36] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

[37] B. Neyshabur, S. Bhojanapalli, and N. Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

[38] M. Uehara, J. Huang, and N. Jiang. Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

[39] M. Yin, Y. Duan, M. Wang, and Y.-X. Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

[40] D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

[41] D. A. McAllester. Simplified PAC-Bayesian margin bounds. In *Learning Theory and Kernel Machines*, pages 203–215. Springer, 2003.

[42] W. Sun, N. Jiang, A. Krishnamurthy, A. Agarwal, and J. Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pages 2898–2933. PMLR, 2019.

[43] T. Zhang. From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

[44] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, OpenAI Pieter A., and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[45] I. Mordatch and P. Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[46] I. Liu, R. A. Yeh, and A. G. Schwing. PIC: permutation invariant critic for multi-agent deep reinforcement learning. In *Conference on Robot Learning*, pages 590–602. PMLR, 2020.

[47] L. Devroye, A. Mehrabian, and T. Reddad. The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

[48] M. Telgarsky. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*, 2015.

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

[50] M. Tsagris, C. Beneki, and H. Hassani. On the folded normal distribution. *Mathematics*, 2(1):12–28, 2014.

[51] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540. PMLR, 2018.

[52] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

[53] B. Jiang, Q. Sun, and J. Fan. Bernstein's inequality for general Markov chains. *arXiv preprint arXiv:1805.10721*, 2018.

[54] D. Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1–32, 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary Materials for “Relational Reasoning via Set Transformers: Provable Efficiency and Applications to MARL”

A Formal Definition of the Fully-Connected Networks in Section 3

For a multi-channel input, the output is the sum of the output of each channel, i.e.,

$$[\phi_{\mathrm{ReLU}}]_i(X) = \sum_{k=1}^{N} [\phi_{\mathrm{ReLU}}]_i(X_{k,:}) \quad \text{for} \quad i \in [W_2],$$

where $X_{k,:}$ is the k^{th} row of X. The fully-connected neural network for each channel is defined as

$$[\phi_{\mathrm{ReLU}}]_i(x) = \sum_{j=1}^{W_1} c_{ij} \mathrm{ReLU}\big(a_j^\top x + b_j\big) + d_i \quad \text{for} \quad i \in [W_2],$$

where $a_j \in \mathbb{R}^d$ and $b_j, c_{ij}, d_i \in \mathbb{R}$ for $i \in [W_2], j \in [W_1]$ are the parameters of ϕ_{ReLU}. The network ρ_{ReLU} is defined as

$$\rho_{\mathrm{ReLU}}(y) = \sum_{k=1}^{W_3} g_k \mathrm{ReLU}\big(e_k^\top y + f_k\big) + h,$$

where $e_i \in \mathbb{R}^{W_2}$ and $f_k, g_k, h \in \mathbb{R}$ for $k \in [W_3]$ are the parameters of ρ.

B Formal Definition of the Transformer Structures in Sections 4.1 and 4.2

The transformer structure in Section 4.1. In each layer, we combine the self-attention mechanism with the Row-wise FeedForward (rFF) single-hidden layer neural network. rFF takes $X \in \mathbb{R}^{N\times d}$ as the input and outputs a matrix in the same dimension. It applies a single-hidden layer network in a row-wise manner. For the entry in the i^{th} row and the k^{th} column of the output, we have

$$\big[\mathrm{rFF}(X, a, b)\big]_{i,k} = \big[\mathrm{rFF}(X_{i,:}, a, b)\big]_k \text{ for } k \in [d], i \in [N],$$

where $X_{i,:} \in \mathbb{R}^d$ is the i^{th} row of X. For a d-dimensional vector input, the single-hidden layer outputs a vector in the same dimension as

$$\big[\mathrm{rFF}(x, a, b)\big]_k = \sum_{j=1}^{m} a_{kj} \mathrm{ReLU}(b_{kj}^\top x) \text{ for } k \in [d],$$

where $x \in \mathbb{R}^d$ is the input, m is the width of the network, and $a = [a_{11}, a_{12}, \ldots, a_{dm}] \in \mathbb{R}^{dm}$ and $b = [b_{11}, b_{12}, \ldots, b_{dm}] \in \mathbb{R}^{d\times dm}$ are the parameters of rFF.

Then for any layer $i \in [L-1]$, the layer output is

$$G_{\mathrm{tf}}^{(i+1)} = \Pi_{\mathrm{norm}}\Big[\mathrm{Att}\big(G_{\mathrm{tf}}^{(i)} W_{QK}^{(i+1)}, G_{\mathrm{tf}}^{(i)}, G_{\mathrm{tf}}^{(i)} W_V^{(i+1)}\big) + \mathrm{rFF}\big(G_{\mathrm{tf}}^{(i)}, a^{(i+1)}, b^{(i+1)}\big)\Big], \tag{B.1}$$

where

$$\begin{aligned} a^{1:i} &= [a^{(1)}, \ldots, a^{(i)}],\\ b^{1:i} &= [b^{(1)}, \ldots, b^{(i)}],\\ W_{QK}^{1:i} &= [W_{QK}^{(1)}, \ldots, W_{QK}^{(i)}],\\ W_V^{1:i} &= [W_V^{(1)}, \ldots, W_V^{(i)}], \end{aligned}$$

are the stacked parameters of the first i layers of the network, and $G_{\mathrm{tf}}^{(i)}$ is a shorthand for $G_{\mathrm{tf}}^{(i)}(X; W_{QK}^{1:i}, W_V^{1:i}, a^{1:i}, b^{1:i})$. $\Pi_{\mathrm{norm}}(X)$ is the row-wise normalization function, which projects each row of X into the ℓ_p-ball (where $p \geq 1$). We take $G_{\mathrm{tf}}^{(0)}(X) = \Pi_{\mathrm{norm}}\big(X\big)$ as the input of the first layer. For the last layer L, we derive the scalar estimate of the action-value function with the average aggregation among all the channels, i.e.,

$$g_{\mathrm{tf}}(X; W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}, w) = \Pi_{V_{\max}}\bigg(\frac{1}{N}\mathbb{I}_N G_{\mathrm{tf}}^{(L)}(X; W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L})w\bigg),$$

where $\Pi_{V_{\max}}(x)$ is the "clipping" function, which is defined as $\Pi_{V_{\max}}(x) = x$ if $|x| \leq V_{\max}$ and $\Pi_{V_{\max}}(x) = V_{\max}\mathrm{sign}(x)$ otherwise.

The transformer structure in Section 4.2. For the layer $i \in [L-2]$, we adopt the same neural network structure in Eqn. (B.1). For the final layer, we implement the structure that

$$\begin{aligned} &F_{\mathrm{tf}}(X; W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}) \\ &\quad = \mathrm{SM}\big(G_{\mathrm{tf}}^{(L-1)} W_{QK}^{(L)} G_{\mathrm{tf}}^{(L-1)\top}\big) G_{\mathrm{tf}}^{(L-1)} W_V^{(L)} + \mathrm{rFF}\big(G_{\mathrm{tf}}^{(L-1)}, a^{(L)}, b^{(L)}\big). \end{aligned}$$

C Equivalent Expression for the Model-based RL algorithm in Section 4.2

The algorithm in Eqn. (3) can be equivalently expressed in two forms.

Transition Function. The algorithm in Eqn. (3) can be expressed with the transition function F as

$$\hat{F}_{\mathrm{MLE}} = \operatorname*{argmin}_{F \in \mathcal{M}_{\mathrm{tf}}} \frac{1}{n}\sum_{i=1}^{n} \big\|\bar{S}_i' - F(\bar{S}_i, \bar{A}_i)\big\|_{\mathrm{F}}^2 \quad \text{and} \quad \hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \min_{F \in \mathcal{M}_{\mathrm{MLE}}(\zeta)} V_{P_F}^{\pi}(\bar{S}_0),$$

where the "confidence region" $\mathcal{M}_{\mathrm{MLE}}(\zeta)$ is the set of all $F \in \mathcal{M}_{\mathrm{tf}}$ such that [47]

$$\frac{1}{n}\sum_{i=1}^{n}\Bigg(2\Phi\Bigg(\sqrt{\frac{\big\|F(\bar{S}_i, \bar{A}_i) - \hat{F}_{\mathrm{MLE}}(\bar{S}_i, \bar{A}_i)\big\|_{\mathrm{F}}^2}{2\sigma^2}}\Bigg) - 1\Bigg)^2 \leq \zeta,$$

and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Transition Probability. The algorithm in Eqn. (3) can also be expressed with the transition probability P. Since the function F is equivalent to the transition kernel P_F, the transition kernel class can be correspondingly defined as

$$\mathcal{P}_{\mathrm{tf}}(B') = \big\{P \,\big|\, \exists\, F \in \mathcal{M}_{\mathrm{tf}}(B') \text{ s.t. } P = P_F\big\}.$$

Then the algorithm can be expressed as

$$\hat{P}_{\mathrm{MLE}} = \operatorname*{argmax}_{P \in \mathcal{P}_{\mathrm{tf}}} \sum_{i=1}^{n} \log P(\bar{S}_i' \,|\, \bar{S}_i, \bar{A}_i) \quad \text{and} \quad \hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \min_{P \in \mathcal{P}_{\mathrm{MLE}}(\zeta)} V_P^{\pi}(\bar{S}_0),$$

where the confidence region $\mathcal{P}_{\mathrm{MLE}}(\zeta)$ is defined as

$$\mathcal{P}_{\mathrm{MLE}}(\zeta) = \Bigg\{P \in \mathcal{P}_{\mathrm{tf}} \,\Bigg|\, \frac{1}{n}\sum_{i=1}^{n} \mathrm{TV}\big(P(\cdot \,|\, \bar{S}_i, \bar{A}_i), \hat{P}_{\mathrm{MLE}}(\cdot \,|\, \bar{S}_i, \bar{A}_i)\big)^2 \leq \zeta\Bigg\}.$$

D Proof of Propositions 2

Proof of Proposition 2. We denote any optimal policy as $\pi^* = \operatorname{argmax}_{\pi} V_{P^*}^{\pi}(\bar{S}_0)$. Note that the optimal policy may be not unique, and any policy that achieves the maximal value function is called an optimal policy. The corresponding action-value function is denoted as $Q_{P^*}^*$, which is defined as

$$Q_{P^*}^*(\bar{S}, \bar{A}) = \mathbb{E}_{\bar{S}' \sim P^*(\cdot \,|\, \bar{S}, \bar{A})}\big[r(\bar{S}, \bar{A}) + \max_{\bar{A}'} Q_{P^*}^*(\bar{S}', \bar{A}')\big]. \tag{D.1}$$

For any row-wise permutation function $\psi(\cdot)$, we have

$$\begin{aligned} Q_{P^*}^*\big(\psi(\bar{S}), \psi(\bar{A})\big) &= \mathbb{E}_{\psi(\bar{S}') \sim P^*(\cdot \,|\, \psi(\bar{S}), \psi(\bar{A}))}\Big[r\big(\psi(\bar{S}), \psi(\bar{A})\big) + \max_{\bar{A}'} Q_{P^*}^*\big(\psi(\bar{S}'), \psi(\bar{A}')\big)\Big] \\ &= \mathbb{E}_{\bar{S}' \sim P^*(\cdot \,|\, \bar{S}, \bar{A})}\Big[r(\bar{S}, \bar{A}) + \max_{\bar{A}'} Q_{P^*}^*\big(\psi(\bar{S}'), \psi(\bar{A}')\big)\Big], \end{aligned} \tag{D.2}$$

where Eqn. (D.2) follows from the homogeneity of the MDP. Since $Q^*_{P^*}$ is the unique solution of Eqn. (D.1), we have $Q^*_{P^*}(\bar{S}, \bar{A}) = Q^*_{P^*}(\psi(\bar{S}), \psi(\bar{A}))$ for all $\psi(\cdot)$. Thus, the permutation invariant policy $\pi(\bar{S}) = \mathrm{argmax}_{\bar{A}}\, Q^*_{P^*}(\bar{S}, \bar{A})$ is the optimal policy.

When the policy π is permutation invariant, we can show that the corresponding action-value function and the value function are permutation invariant following the similar argument as above. Therefore, we conclude the proof of Proposition 2. $\square$

E Proof of Proposition 1

Proof of Proposition 1. We note that Proposition 1 is a corollary of Theorem 2. Take $\tilde{f} = 0$ in Theorem 2, then we recover the result of Proposition 1. Thus, we only provide the proof of Theorem 2 in Appendix I. $\square$

F Proof of Theorem 1

Proof of Theorem 1. The functions in $\mathcal{N}(W)$ are the fully-connected networks with the ReLU activation, so they are piece-wise linear functions on $[0,1]^{N\times d}$, where the number of the linear pieces are polynomial in the width of the network. In contrast, the self-attention function is convex on some subset of $[0,1]^{N\times d}$. In the following proof procedures, we specify a line in $[0,1]^{N\times d}$ where the second derivative of the self-attention function is high enough such that $\mathcal{N}(W)$ should be exponentially wide to approximate the self-attention function on the longest linear piece of that line.

To specify a line in $[0,1]^{N\times d}$, we set the inputs of all but the first channels to be x, and set the input of the first channel to be a scaled version of x. Fix any $x \in [0,1]^d$ and $k \in \mathbb{R}$, we set $x_1 = kx$ and $x_i = x$ for all $i \in \{2, \ldots, N\}$. For $X = [x_1, \ldots, x_N]^\top$, $w \in [0,1]^d$ and $a \in \mathbb{R}$, we define

$$
\begin{aligned}
f(a, X, w) &= \mathbb{I}_N^\top \mathrm{Att}(aX, aX, aX)w \\
&= akx^\top w\bigg[\frac{e^{a^2k^2x^\top x}}{e^{a^2k^2x^\top x} + (N-1)e^{a^2kx^\top x}} + \frac{(N-1)e^{a^2kx^\top x}}{e^{a^2kx^\top x} + (N-1)e^{a^2x^\top x}}\bigg] \\
&\quad + akx^\top w(N-1)\bigg[\frac{e^{a^2kx^\top x}}{e^{a^2k^2x^\top x} + (N-1)e^{a^2kx^\top x}} + \frac{(N-1)e^{a^2x^\top x}}{e^{a^2kx^\top x} + (N-1)e^{a^2x^\top x}}\bigg],
\end{aligned}
$$

where $\mathbb{I}_N \in \mathbb{R}^N$ is the vector with all entries being equal to 1. The partial derivatives of $f(a, X, w)$ with respect to a can be derived as

$$
\frac{\partial f(a, X, w)}{\partial a} = \big[2a^2(k-1)x^\top x + 1\big]kx^\top w e^{a^2(k-1)x^\top x} + Nx^\top w + O\Big(\frac{1}{N}\Big), \tag{F.1}
$$

$$
\frac{\partial^2 f(a, X, w)}{\partial a^2} = 2x^\top w(k-1)^2 x^\top x a e^{a^2(k-1)x^\top x}\big[2a^2(k-1)x^\top x + 3\big] + O\Big(\frac{1}{N}\Big). \tag{F.2}
$$

We set $x = 2/3 \cdot \mathbb{I}_d$, $k = 1.1$, $w = x$, and define the function $g(a) = f(1, X + aX/3, x)$. Then Eqn. (F.1) and (F.2) show that $g(a)$ is a increasing convex function on $[-1, 1]$.

We can rearrange the weights in the first layer of ϕ_{ReLU} such that the input of the resultant network is a scalar $a \in [-1, 1]$; the width of the resultant network is same as the width of $\rho_{\mathrm{ReLU}}(\sum_{i=1}^N \phi_{\mathrm{ReLU}}(x_i))$; the resultant network represents the same function as

$$
h(a) = \rho_{\mathrm{ReLU}}\bigg(\sum_{i=1}^N \phi_{\mathrm{ReLU}}\Big(x_i + \frac{a}{3}x_i\Big)\bigg).
$$

Since $\rho_{\mathrm{ReLU}}(\sum_{i=1}^N \phi_{\mathrm{ReLU}}(x_i))$ can approximate $\mathbb{I}_N^\top \mathrm{Att}(X, X, X)w$, the modified network can approximate $g(a)$ in terms of the sup-norm on $[-1, 1]$.

Since ReLU is a 2-piece-wise linear function, $h(a)$ is also a piece-wise linear function, whose number of pieces is denoted as M. Lemma 2.1 of [48] shows that $M \le 2(2W)^2 = 8W^2$, where $(2W)^2$ follows from two ReLU layers, and the additional factor of 2 follows from that x_1 and x_i for $i \in \{2, \ldots, N\}$ take different values.

The pigeonhole principle implies that there is a piece-wise linear segment $[u, v] \subseteq [-1, 1]$ whose length is at least $2/M$. On this linear segment, the linear function $h(a)$ approximates $g(a)$ with error at most ξ. Eqn. (F.2) then implies that

$$\inf_{a \in [-1,1]} h''(a) \geq c_1 > 0,$$

where $c_1 = \Omega(d^2 e^{cd})$ for some $c > 0$. Denote the linear function on a linear piece $[u, v]$ and the approximation error as $\hat{h} : \mathbb{R} \to \mathbb{R}$ and $e = h - \hat{h}$, respectively. Since $\hat{h}$ is a linear function, we have

$$\max\big\{e(u), e(v)\big\} \geq e\Big(\frac{u+v}{2}\Big) + \frac{c_1}{2}\Big(\frac{v-u}{2}\Big)^2 \tag{F.3}$$

and

$$\xi \geq \frac{1}{2}\Bigg(\max\big\{e(u), e(v)\big\} - e\Bigg(\frac{u+v}{2}\Bigg)\Bigg). \tag{F.4}$$

Combining inequalities (F.3) and (F.4), we have

$$W \geq \left(\frac{c_1}{256\xi}\right)^{\frac{1}{4}}.$$

Thus, we have $W = \Omega(\exp(cd)\xi^{-1/4})$ for some constant $c > 0$, and this concludes the proof of Theorem 1. $\square$

G Proof of Theorem 3

Proof of Theorem 3. Recall the definition below Theorem 2

$$e(\mathcal{F}_{\mathrm{tf}}, \Pi, \delta, n) = 32V_{\max}^2\Bigg[2 + \gamma + 2(m+1)L^2d^2\log\Bigg(\frac{16mdLB_VB_{QK}B_aB_bn}{V_{\max}}\Bigg) \\ + 2(m+1)Ld^2\log B_w + \log\Bigg(\frac{2\mathcal{N}(\Pi, 1/n, d_\infty)}{\delta}\Bigg)\Bigg].$$

To simplify the proof, we define

$$f^*_{\pi^*} = \operatorname*{arginf}_{f \in \mathcal{F}_{\mathrm{tf}}} \sup_{\mu \in d_\Pi} \mathbb{E}_\mu\Big[\big(f(\bar{S}, \bar{A}) - \mathcal{T}^{\pi^*} f\big(\bar{S}, \bar{A}\big)\big)^2\Big],$$
$$\varepsilon = \frac{3}{2}\varepsilon_{\mathcal{F}} + \frac{2}{n}e(\mathcal{F}_{\mathrm{tf}}, \Pi, \delta, n).$$

Our proof can be decomposed into three main procedures.

- Since $f^*_{\pi^*}$ is the best approximation of action-value function of the optimal policy π^*, we expect that it should belong to the confidence region of the action-value functions $\mathcal{F}(\pi^*, \varepsilon)$ with high probability.
- For any $\pi \in \Pi$ and any $f \in \mathcal{F}(\pi, \varepsilon)$, since the empirical Bellman error is bounded $\mathcal{E}(f, \pi; \mathcal{D}) \leq \varepsilon$, we expect that the population Bellman error $\mathbb{E}_\nu[(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A}))^2]$ can be controlled with high probability, which implies that f is a reliable estimate of the action-value function of π.
- The suboptimality gap of the learned policy according to the reliable action-value function estimate can be bounded using the estimation error bound.

We lay out the proof by the three steps as stated in the proof sketch.

Step 1: Show that $f^*_{\pi^*} \in \mathcal{F}(\pi^*, \varepsilon)$ with high probability.

From the definition of $f^*_{\pi^*}$ and Assumption 1, we note that the population Bellman error of $f^*_{\pi^*}$ with respect to π^* is bounded by $\varepsilon_{\mathcal{F}}$. To bound the empirical Bellman error $\mathcal{E}(f^*_{\pi^*}, \pi^*; \mathcal{D})$ of $f^*_{\pi^*}$, we need the generalization error bound of the action-value function with the transformer function class.

Theorem 2. *Let $\bar{B} = B_V B_{QK} B_a B_b B_w$. For all $f, \tilde{f} \in \mathcal{F}_{\mathrm{tf}}(B)$ and all policies $\pi \in \Pi$, with probability at least $1-\delta$, we have*

$$\begin{aligned}&\left|\mathbb{E}_\nu\Big[\big(f(\bar{S},\bar{A}) - \mathcal{T}^\pi \tilde{f}(\bar{S},\bar{A})\big)^2\Big] - \mathcal{L}(f,\tilde{f},\pi;\mathcal{D}) + \mathcal{L}(\mathcal{T}^\pi \tilde{f},\tilde{f},\pi;\mathcal{D})\right| \\ &\leq \frac{1}{2}\mathbb{E}_\nu\Big[\big(f(\bar{S},\bar{A}) - \mathcal{T}^\pi \tilde{f}(\bar{S},\bar{A})\big)^2\Big] + O\bigg(\frac{V_{\max}^2}{n}\bigg[mL^2d^2\log\frac{mdL\bar{B}n}{V_{\max}} + \log\frac{\mathcal{N}(\Pi,1/n,d_\infty)}{\delta}\bigg]\bigg).\end{aligned}$$

Proof. See Appendix I for a detailed proof. □

We can decompose the empirical Bellman error $\mathcal{E}(f^*_{\pi^*},\pi^*;\mathcal{D})$ as the sum of the population Bellman error and the generalization error, where the population Bellman error can be controlled with $\varepsilon_{\mathcal{F}}$ according to Assumption 1, and the generalization error can be controlled with Theorem 2. Thus, we have the following lemma.

Lemma 4. *For any $\pi \in \Pi$, let $f^*_\pi = \mathrm{arginf}_{f\in\mathcal{F}_{\mathrm{tf}}} \sup_{\mu\in d_\Pi} \mathbb{E}_\mu[(f(\bar{S},\bar{A}) - \mathcal{T}^\pi f(\bar{S},\bar{A}))^2]$. If Assumption 1 holds, the following inequality holds with probability at least $1-\delta$,*

$$\mathcal{E}(f^*_\pi,\pi;\mathcal{D}) \leq \frac{3}{2}\varepsilon_{\mathcal{F}} + \frac{2e(\mathcal{F}_{\mathrm{tf}},\Pi,\delta,n)}{n}.$$

Proof. See Appendix L.1 for a detailed proof. □

Step 2: For any policy $\pi \in \Pi$ and $f \in \mathcal{F}(\pi,\varepsilon)$, show $\mathbb{E}_\nu[(f(\bar{S},\bar{A}) - \mathcal{T}^\pi f(\bar{S},\bar{A}))^2] \leq 2\varepsilon + 3\varepsilon_{\mathcal{F},\mathcal{F}} + 4e(\mathcal{F}_{\mathrm{tf}},\Pi,\delta,n)/n$ holds with high probability.

To prove the desired result, we relate the population Bellman error $\mathbb{E}_\nu[(f(\bar{S},\bar{A}) - \mathcal{T}^\pi f(\bar{S},\bar{A}))^2]$ with $\mathcal{E}(f,\pi;\mathcal{D})$ through Theorem 2, where we bound the population Bellman error as the difference between the empirical Bellman error and the generalization error. Thus, we have the following lemma.

Lemma 5. *For any $\pi \in \Pi$ and $f \in \mathcal{F}_{\mathrm{tf}}$, if $\mathcal{E}(f,\pi;\mathcal{D}) \leq \varepsilon$ for some $\varepsilon > 0$, and Assumption 1 holds, the following inequality holds with probability at least $1-\delta$,*

$$\mathbb{E}_\nu\Big[\big(f(\bar{S},\bar{A}) - \mathcal{T}^\pi f(\bar{S},\bar{A})\big)^2\Big] \leq 2\varepsilon + 3\varepsilon_{\mathcal{F},\mathcal{F}} + \frac{4e(\mathcal{F}_{\mathrm{tf}},\Pi,\delta,n)}{n}.$$

Proof. See Appendix L.2 for a detailed proof. □

Step 3: Bound the suboptimality gap of the learned policy with the population Bellman error bound in Step 2.

We define

$$\hat{f}_{\pi^*} = \operatorname*{argmax}_{f\in\mathcal{F}(\pi^*,\varepsilon)} f(\bar{S}_0,\pi^*) \quad \text{and} \quad \check{f}_{\pi^*} = \operatorname*{argmin}_{f\in\mathcal{F}(\pi^*,\varepsilon)} f(\bar{S}_0,\pi^*),$$

where $\hat{f}_{\pi^*}$ and $\check{f}_{\pi^*}$ are the maximal and minimal value functions in $\mathcal{F}(\pi^*,\varepsilon)$, respectively. Intuitively, since $f^*_{\pi^*} \in \mathcal{F}(\pi^*,\varepsilon)$ and that we learn the policy according to the pessimistic estimation of the action-value function in $\mathcal{F}(\hat{\pi},\varepsilon)$, we can upper bound the suboptimality gap by the difference between $\hat{f}_{\pi^*}$ and $\check{f}_{\pi^*}$.

Step 1 shows that with probability at least $1-\delta$, $f^*_{\pi^*} \in \mathcal{F}(\pi^*,\varepsilon)$. Then we have

$$\max_{f\in\mathcal{F}(\pi^*,\varepsilon)} f(\bar{S}_0,\pi) \geq f^*_{\pi^*} f(\bar{S}_0,\pi) = V^{\pi^*}_{P^*}(\bar{S}_0) + \frac{\mathbb{E}_{d^{\pi^*}_{P^*}}\big[f^*_{\pi^*} - \mathcal{T}^{\pi^*} f(\bar{S},\bar{A})\big]}{1-\gamma} \geq V^{\pi^*}_{P^*}(\bar{S}_0) - \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}, \tag{G.1}$$

where the equality follows from Lemma 14, and the last inequality follows from Assumption 1. Similarly, we can prove that

$$\min_{f\in\mathcal{F}(\hat{\pi},\varepsilon)} f(\bar{S}_0,\hat{\pi}) \leq V^{\hat{\pi}}_{P^*}(\bar{S}_0) + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}. \tag{G.2}$$

Combining inequalities (G.1) and (G.2), we have

$$
\begin{aligned}
&V_{P^*}^{\pi^*}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0) \\
&\quad \le \max_{f\in\mathcal{F}(\pi^*,\varepsilon)} f(\bar{S}_0,\pi) - \min_{f\in\mathcal{F}(\hat{\pi},\varepsilon)} f(\bar{S}_0,\hat{\pi}) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
&\quad \le \hat{f}_{\pi^*}(\bar{S}_0,\pi^*) - \check{f}_{\pi^*}(\bar{S}_0,\pi^*) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
&\quad = \hat{f}_{\pi^*}(\bar{S}_0,\pi^*) - V_{P^*}^{\pi^*}(\bar{S}_0) + V_{P^*}^{\pi^*}(\bar{S}_0) - \check{f}_{\pi^*}(\bar{S}_0,\pi^*) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma},
\end{aligned} \tag{G.3}
$$

where the first inequality follows from inequalities (G.1) and (G.2), the second inequality follows from Eqn. (1). Applying the suboptimality gap decomposition in Lemma 14 to inequality (G.3), we have

$$
\begin{aligned}
&V_{P^*}^{\pi^*}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0) \\
&\quad \le \frac{1}{1-\gamma}\Big\{\mathbb{E}_{d_{P^*}^{\pi^*}}\big[\hat{f}_{\pi^*}(\bar{S},\bar{A}) - \mathcal{T}^{\pi^*}\hat{f}_{\pi^*}(\bar{S},\bar{A})\big] \\
&\qquad - \mathbb{E}_{d_{P^*}^{\pi^*}}\big[\check{f}_{\pi^*}(\bar{S},\bar{A}) - \mathcal{T}^{\pi^*}\check{f}_{\pi^*}(\bar{S},\bar{A})\big]\Big\} + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
&\quad \le \frac{1}{1-\gamma}\bigg\{\sqrt{C_{\mathcal{F}_{\mathrm{tf}}}\mathbb{E}_{\nu}\Big[\big(\hat{f}_{\pi^*}(\bar{S},\bar{A}) - \mathcal{T}^{\pi^*}\hat{f}_{\pi^*}(\bar{S},\bar{A})\big)^2\Big]} \\
&\qquad + \sqrt{C_{\mathcal{F}_{\mathrm{tf}}}\mathbb{E}_{\nu}\Big[\big(\check{f}_{\pi^*}(\bar{S},\bar{A}) - \mathcal{T}^{\pi^*}\check{f}_{\pi^*}(\bar{S},\bar{A})\big)^2\Big]}\bigg\} + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma},
\end{aligned}
$$

where the first inequality follows from Lemma 14, and the second inequality follows from Jensen's inequality and the definition of $C_{\mathcal{F}_{\mathrm{tf}}}$. Combined with the result in step 2, we have

$$
\begin{aligned}
&V_{P^*}^{\pi^*}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0) \\
&\quad \le \frac{2\sqrt{C_{\mathcal{F}_{\mathrm{tf}}}}}{1-\gamma}\sqrt{2\varepsilon + 3\varepsilon_{\mathcal{F},\mathcal{F}} + \frac{4e(\mathcal{F}_{\mathrm{tf}},\Pi,\delta,n)}{n}} + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
&\quad \le O\bigg(\frac{\sqrt{C_{\mathcal{F}_{\mathrm{tf}}}(\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}})}}{1-\gamma} + \frac{\sqrt{C_{\mathcal{F}_{\mathrm{tf}}}}}{1-\gamma}\sqrt{\frac{e(\mathcal{F}_{\mathrm{tf}},\Pi,\delta,n)}{n}}\bigg).
\end{aligned}
$$

Therefore, we conclude the proof of Theorem 3. □

H Proof of Theorem 4

For ease of notation, we denote the parameters of the neural network as

$$\theta = [W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}].$$

The parameter space is

$$
\begin{aligned}
\Theta(B_a,B_b,B_{QK},B_V) = \Big\{\theta\,\Big|\,&|a_{kj}^{(i)}| < B_a, \|b_{kj}^{(i)}\|_2 < B_b, \|W_{QK}^{(i)\top}\|_{\mathrm{F}} < B_{QK}, \\
&\|W_V^{(i)\top}\|_{\mathrm{F}} < B_V \text{ for } i\in[L], j\in[m], k\in[d]\Big\}.
\end{aligned}
$$

Then we can denote the functions in $\mathcal{M}_{\mathrm{tf}}(B_a,B_b,B_{QK},B_V)$ as F_θ and the corresponding transition kernel in $\mathcal{P}_{\mathrm{tf}}(B_a,B_b,B_{QK},B_V)$ as P_θ, where $\theta\in\Theta$ is the parameter of the function.

From the perspective of the parameter space Θ, the algorithm in Eqn. (3) can be equivalently stated as

$$
\begin{aligned}
P_{\hat{\theta}_{\mathrm{MLE}}} &= \operatorname*{argmax}_{P\in\mathcal{P}_{\mathrm{tf}}} \frac{1}{n}\sum_{i=1}^n \log P(\bar{S}_i' \,|\, \bar{S}_i,\bar{A}_i), \\
\hat{\pi} &= \operatorname*{argmax}_{\pi\in\Pi}\min_{P\in\mathcal{P}(\zeta)} V_P^{\pi}(\bar{S}_0),
\end{aligned}
$$

where the confidence region of the dynamics is defined as

$$\mathcal{P}(\zeta) = \Big\{P\in\mathcal{P}_{\mathrm{tf}}\,\Big|\,\frac{1}{n}\sum_{i=1}^n \mathrm{TV}\big(P(\cdot\,|\,\bar{S}_i,\bar{A}_i), P_{\hat{\theta}_{\mathrm{MLE}}}(\cdot\,|\,\bar{S}_i,\bar{A}_i)\big)^2 \le \zeta\Big\}.$$

Proof of Theorem 4. For some constant $c_1 > 0$, we take

$$\zeta = c_1\bigg(\frac{1}{n}(m+1)L^2d^2\log\Big(4NLmdB_VB_{QK}B_aB_bn\Big) + \frac{1}{n}\log\frac{1}{\delta}\bigg).$$

Our proof can be decomposed into three main parts.

- Intuitively, the nominal transition kernel P^* should belong to the confidence region of the system dynamics set $\mathcal{P}(\zeta)$ with high probability.
- For any $P \in \mathcal{P}(\zeta)$, we expect that the population squared total variation between P and P^*, i.e., $\mathbb{E}_\nu[\mathrm{TV}(P(\cdot\,|\,\bar{S},\bar{A}), P^*(\cdot\,|\,\bar{S},\bar{A}))^2]$, can be controlled with high probability, which implies that any $P \in \mathcal{P}(\zeta)$ is a reliable estimate of the system dynamics.
- The suboptimality gap of the learned policy according to the reliable dynamic estimate can be bounded in terms of the total variation.

We lay out the proof by the three steps as stated in the proof sketch.

Step 1: Show that $P^* \in \mathcal{P}(\zeta)$ with probability at least $1-\delta$.

From the definition of $\mathcal{P}(\zeta)$, we need to bound the empirical total variation between the nominal transition kernel and the MLE estimate. Thus, we need an upper bound of the population total variation between P^* and $\hat{P}_{\mathrm{MLE}}$ and an accompanying generalization error bound. For the population error, we state the following proposition.

Proposition 3. *Let $\tilde{B} = B_VB_{QK}B_aB_b$. For the maximum likelihood estimate $\hat{P}_{\mathrm{MLE}}$ in Eqn.* (3)*, the following inequality holds with probability at least $1-\delta$,*

$$\mathbb{E}_\nu\Big[\mathrm{TV}\big(P^*(\cdot\,|\,\bar{S},\bar{A}), \hat{P}_{\mathrm{MLE}}(\cdot\,|\,\bar{S},\bar{A})\big)^2\Big] \le O\bigg(\frac{1}{n}mL^2d^2\log\big(NLmd\tilde{B}n\big) + \frac{1}{n}\log\frac{1}{\delta}\bigg).$$

Proof. See Appendix J for a detailed proof. □

Similar to Theorem 2, we can derive the generalization error bound in terms of the total variation distance.

Proposition 6. *For any $\theta \in \Theta$, with probability at least $1-\delta$, we have*

$$\begin{aligned}
&\bigg|\mathbb{E}_{\mathcal{D}}\Big[\mathrm{TV}\big(P^*(\cdot\,|\,\bar{S},\bar{A}), P_\theta(\cdot\,|\,\bar{S},\bar{A})\big)^2\Big] - \frac{1}{n}\sum_{i=1}^n \mathrm{TV}\big(P^*(\cdot\,|\,\bar{S}_i,\bar{A}_i), P_\theta(\cdot\,|\,\bar{S}_i,\bar{A}_i)\big)^2\bigg| \\
&\quad\le \frac{1}{2}\mathbb{E}_{\mathcal{D}}\Big[\mathrm{TV}\big(P^*(\cdot\,|\,\bar{S},\bar{A}), P_\theta(\cdot\,|\,\bar{S},\bar{A})\big)^2\Big] \\
&\qquad + O\bigg(\frac{1}{n}mL^2d^2\log(NLmdB_VB_{QK}B_aB_bn) + \frac{1}{n}\log\frac{1}{\delta}\bigg).
\end{aligned}$$

Proof. See Appendix K for a detailed proof. □

With Propositions 3 and 6, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathrm{TV}\big(P^*(\cdot\,|\,\bar{S}_i,\bar{A}_i), P_{\hat{\theta}_{\mathrm{MLE}}}(\cdot\,|\,\bar{S}_i,\bar{A}_i)\big)^2 \\
&\quad= \bigg\{\frac{1}{n}\sum_{i=1}^n \mathrm{TV}\Big(P^*(\cdot\,|\,\bar{S}_i,\bar{A}_i), P_{\hat{\theta}_{\mathrm{MLE}}}(\cdot\,|\,\bar{S}_i,\bar{A}_i)\Big)^2 - \frac{3}{2}\mathbb{E}_\nu\Big[\mathrm{TV}\big(P^*(\cdot\,|\,\bar{S},\bar{A}), P_{\hat{\theta}_{\mathrm{MLE}}}(\cdot\,|\,\bar{S},\bar{A})\big)^2\Big]\bigg\} \\
&\qquad + \frac{3}{2}\mathbb{E}_\nu\Big[\mathrm{TV}\big(P^*(\cdot\,|\,\bar{S},\bar{A}), P_{\hat{\theta}_{\mathrm{MLE}}}(\cdot\,|\,\bar{S},\bar{A})\big)^2\Big] \qquad \text{(H.1)} \\
&\quad\le O\Big(\frac{1}{n}(m+1)L^2d^2\log(4NLmdB_VB_{QK}B_aB_bn) + \frac{1}{n}\log\frac{1}{\delta}\Big), \qquad \text{(H.2)}
\end{aligned}$$

where the first term in Eqn. (H.1) is bounded with Proposition 6, and the second term in Eqn. (H.1) is bounded with Proposition 3.

Step 2: Show that for any $P \in \mathcal{P}(\zeta)$, the population total variation between P and P^* is bounded.

For the population total variation between P and P^*, we have

$$
\begin{aligned}
&\mathbb{E}_\nu\Big[\mathrm{TV}\big(P(\cdot \mid \bar{S}, \bar{A}), P^*(\cdot \mid \bar{S}, \bar{A})\big)^2\Big] \\
&\quad = \bigg\{\mathbb{E}_\nu\Big[\mathrm{TV}\big(P(\cdot \mid \bar{S}, \bar{A}), P^*(\cdot \mid \bar{S}, \bar{A})\big)^2\Big] - \frac{2}{n}\sum_{i=1}^n \mathrm{TV}\Big(P(\cdot \mid \bar{S}_i, \bar{A}_i), P^*(\cdot \mid \bar{S}_i, \bar{A}_i)\Big)^2\bigg\} \\
&\qquad + \frac{2}{n}\sum_{i=1}^n \mathrm{TV}\big(P(\cdot \mid \bar{S}_i, \bar{A}_i), P^*(\cdot \mid \bar{S}_i, \bar{A}_i)\big)^2 \\
&\quad \le O\Big(\frac{1}{n}(m+1)L^2 d^2 \log\Big(4NLmdB_V B_{QK} B_a B_b n\Big) + \frac{1}{n}\log\frac{1}{\delta}\Big) \\
&\qquad + \frac{4}{n}\sum_{i=1}^n \mathrm{TV}\Big(P(\cdot \mid \bar{S}_i, \bar{A}_i), P_{\hat{\theta}_{\mathrm{MLE}}}(\cdot \mid \bar{S}_i, \bar{A}_i)\Big)^2 \\
&\qquad + \frac{4}{n}\sum_{i=1}^n \mathrm{TV}\Big(P_{\hat{\theta}_{\mathrm{MLE}}}(\cdot \mid \bar{S}_i, \bar{A}_i), P^*(\cdot \mid \bar{S}_i, \bar{A}_i)\Big)^2 \\
&\quad \le O(\zeta),
\end{aligned} \tag{H.3}
$$

where the first inequality follows from Proposition 6 and triangle inequality, and the last inequality follows from inequality (H.2) and the fact that $P \in \mathcal{P}(\zeta)$.

Step 3: Bound the suboptimality gap of the learned policy with the total variation bound.

With the results in Step 1 and 2, we have that with probability at least $1-\delta$

$$
\begin{aligned}
V_{P^*}^{\pi^*}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0) &= V_{P^*}^{\pi^*}(\bar{S}_0) - \min_{P \in \mathcal{P}(\zeta)} V_P^{\pi^*}(\bar{S}_0) + \min_{P \in \mathcal{P}(\zeta)} V_P^{\pi^*}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0) \\
&\le V_{P^*}^{\pi^*}(\bar{S}_0) - \min_{P \in \mathcal{P}(\zeta)} V_P^{\pi^*}(\bar{S}_0) + \min_{P \in \mathcal{P}(\zeta)} V_P^{\hat{\pi}}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0) \\
&\le V_{P^*}^{\pi^*}(\bar{S}_0) - \min_{P \in \mathcal{P}(\zeta)} V_P^{\pi^*}(\bar{S}_0),
\end{aligned}
$$

where the first inequality follows from the fact that $\hat{\pi}$ maximizes $\min_{P \in \mathcal{P}(\zeta)} V_P^{\pi}(\bar{S}_0)$, and the last inequality follows from the fact that $P^* \in \mathcal{P}(\zeta)$. Define $\check{P} = \operatorname{argmin}_{P \in \mathcal{P}(\zeta)} V_P^{\pi^*}(\bar{S}_0)$. Then we have

$$
\begin{aligned}
V_{P^*}^{\pi^*}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0) &\le V_{P^*}^{\pi^*}(\bar{S}_0) - V_{\check{P}}^{\pi^*}(\bar{S}_0) \\
&\le \frac{V_{\max}}{(1-\gamma)^2}\mathbb{E}_{(\bar{S},\bar{A})\sim d_{P^*}^{\pi^*}}\Big[\mathrm{TV}\big(\check{P}(\cdot \mid \bar{S}, \bar{A}), P^*(\cdot \mid \bar{S}, \bar{A})\big)\Big],
\end{aligned}
$$

where the second inequality follows from Lemma 15. By the Jensen's inequality, it can be further bounded as

$$
\begin{aligned}
V_{P^*}^{\pi^*}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0) &\le \frac{V_{\max}}{(1-\gamma)^2}\sqrt{C_{\mathcal{M}_{\mathrm{tf}}}\mathbb{E}_{(\bar{S},\bar{A})\sim\nu}\Big[\mathrm{TV}\big(\check{P}(\cdot \mid \bar{S}, \bar{A}), P^*(\cdot \mid \bar{S}, \bar{A})\big)^2\Big]} \\
&\le O\bigg(\frac{V_{\max}}{(1-\gamma)^2}\sqrt{C_{\mathcal{M}_{\mathrm{tf}}}\zeta}\bigg),
\end{aligned}
$$

where the first inequality follows Jensen's inequality, and the last inequality follows from inequality (H.3). Therefore, we conclude the proof of Theorem 4. □

I Proof of Theorem 2

Proof of Theorem 2. We adopt a PAC-Bayesian framework to derive the generalization error bound of the Bellman error of the transformer functions, in which the generalization error is bounded by the

Kullback–Leibler divergence between the distributions of functions. Recall that the KL divergence between P and Q is defined as $\mathrm{KL}(P \,\|\, Q) = \int_{\mathcal{A}} \log(\mathrm{d}P/\mathrm{d}Q)\,\mathrm{d}P$ if $P \ll Q$, and $+\infty$ otherwise. We start with preliminary result.

Proposition 7. *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 $0 < \lambda \le 1/(2b)$, given a distribution P_0 on $\mathcal{F}$, with probability at least $1 - \delta$, we have

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

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 $\sigma^2(f) \le c\mu(f)$ for some constant $c \in \mathbb{R}$ for all $f \in \mathcal{F}$, we have

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

with probability at least $1 - \delta$.

Proof. See Appendix M.1 for a detailed proof. □

Our proof can be decomposed into four main parts.

- We verify that the Bellman error satisfies the conditions in Proposition 7 and apply it to the Bellman error.
- Since the desired result is a point-wise generalization error bound, we need to control he fluctuation of both sides of inequality (I.1) with respect to any pair of functions $(f, \tilde{f}) \in \mathcal{F}_{\mathrm{tf}} \times \mathcal{F}_{\mathrm{tf}}$.
- We specify two distributions Q and P_0 and calculate $\mathrm{KL}(Q \,\|\, P_0)$.
- We implement a standard covering argument to prove the result that holds for all the policies in Π.

Step 1: Verify the conditions in Proposition 7

Let $X = (\bar{S}, \bar{A}, \bar{S}')$ for all $f, \tilde{f} \in \mathcal{F}_{\mathrm{tf}}(B_a, B_b, B_{QK}, B_V, B_w)$. We define

$$l(f, \tilde{f}, \pi; X) = \big(f(\bar{S}, \bar{A}) - \bar{r}(\bar{S}, \bar{A}) - \gamma \tilde{f}(\bar{S}', \pi)\big)^2 - \big(\mathcal{T}^\pi \tilde{f}(\bar{S}, \bar{A}) - \bar{r}(\bar{S}, \bar{A}) - \gamma \tilde{f}(\bar{S}', \pi)\big)^2.$$

Then the term we consider in Theorem 2 can be expressed as

$$\mathcal{L}(f, \tilde{f}, \pi; \mathcal{D}) - \mathcal{L}(\mathcal{T}^\pi \tilde{f}, \tilde{f}, \pi; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^n l(f, \tilde{f}, \pi; X_i) \text{ and } \big|l(f, \tilde{f}, \pi; X)\big| \le 4V_{\max}^2.$$

Since $(\bar{S}_i, \bar{A}_i)$ is sampled from ν, and $\bar{S}_i' \sim \bar{P}^*(\cdot \,|\, \bar{S}_i, \bar{A}_i)$, we have $(\bar{S}_i, \bar{A}_i, \bar{S}_i') \sim \nu \times \bar{P}^*$, i.e., $X_i \sim \nu \times \bar{P}^*$ for $i \in [N]$. Then the expectation of $l(f, \tilde{f}, \pi; X)$ is

$$\begin{aligned}
&\mathbb{E}_{\nu \times \bar{P}^*}\big[l(f, \tilde{f}, \pi; X)\big] \\
&\quad = \mathbb{E}_{\nu \times \bar{P}^*}\Big[\big(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi \tilde{f}(\bar{S}, \bar{A})\big)\big(f(\bar{S}, \bar{A}) + \mathcal{T}^\pi \tilde{f}(\bar{S}, \bar{A}) - 2\bar{r} - 2\gamma \tilde{f}(\bar{S}', \pi)\big)\Big] \\
&\quad = \mathbb{E}_\nu\left[\mathbb{E}_{\bar{P}^*}\Big[\big(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi \tilde{f}(\bar{S}, \bar{A})\big)\big(f(\bar{S}, \bar{A}) + \mathcal{T}^\pi \tilde{f}(\bar{S}, \bar{A}) - 2\bar{r} - 2\gamma \tilde{f}(\bar{S}', \pi)\big) \,\Big|\, \bar{S}, \bar{A}\Big]\right] \\
&\quad = \mathbb{E}_\nu\Big[\big(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi \tilde{f}(\bar{S}, \bar{A})\big)^2\Big],
\end{aligned}$$

where the last equality follows from the definition of the Bellman operator. As a consequence, the variance of $l(f, \tilde{f}, \pi; X)$ can be bounded by its expectation as

$$
\begin{aligned}
&\operatorname{Var}\bigl(l(f, \tilde{f}, \pi; X)\bigr) \\
&\quad \le \mathbb{E}_{\nu \times \bar{P}^*}\Bigl[\bigl(l(f, \tilde{f}, \pi; X)\bigr)^2\Bigr] \\
&\quad = \mathbb{E}_{\nu}\biggl[\mathbb{E}_{P^*}\Bigl[\bigl(f(\bar{S}, \bar{A}) - \mathcal{T}^{\pi}\tilde{f}(\bar{S}, \bar{A})\bigr)^2\bigl(f(\bar{S}, \bar{A}) + \mathcal{T}^{\pi}\tilde{f}(\bar{S}, \bar{A}) - 2\bar{r} - 2\gamma\tilde{f}(\bar{S}', \pi)\bigr)^2 \Bigm| \bar{S}, \bar{A}\Bigr]\biggr] \\
&\quad \le 16V_{\max}^2\mathbb{E}_{\nu}\biggl[\mathbb{E}_{P^*}\Bigl[\bigl(f(\bar{S}, \bar{A}) - \mathcal{T}^{\pi}\tilde{f}(\bar{S}, \bar{A})\bigr)^2 \Bigm| \bar{S}, \bar{A}\Bigr]\biggr] \\
&\quad = 16V_{\max}^2\mathbb{E}_{\nu \times \bar{P}^*}\bigl[l(f, \tilde{f}, \pi; X)\bigr], \qquad \text{(I.2)}
\end{aligned}
$$

where the last inequality follows from the fact that f and $\tilde{f}$ is bounded by $V_{\max}$. Inequality (I.2) shows that $l(f, \tilde{f}, \pi; X)$ satisfies the condition in Proposition 7 with $b = 4V_{\max}^2$ and $c = 16V_{\max}^2$. In the following, we apply Proposition 7 to $l(f, \tilde{f}, \pi; X)$.

For ease of notation, we denote the parameters of the neural network as

$$\theta = [W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}, w].$$

The parameter space is

$$
\begin{aligned}
\Theta(B_a, B_b, B_{QK}, B_V, B_w) = \Bigl\{\theta \Bigm| |a_{kj}^{(i)}| < B_a, \|b_{kj}^{(i)}\|_q < B_b, \|W_{QK}^{(i)\top}\|_{p,q} < B_{QK}, \\
\|W_V^{(i)\top}\|_{p,q} < B_V, \|w\|_q < B_w \text{ for } i \in [L], j \in [m], k \in [d]\Bigr\}.
\end{aligned}
$$

We denote the functions in $\mathcal{F}_{\mathrm{tf}}(B_a, B_b, B_{QK}, B_V, B_w)$ equivalently as f_θ, where $\theta \in \Theta$ is the parameter of the function.

For a finite policy class $\tilde{\Pi}$ (which is set to be a cover of the original policy class Π in Step 4), Proposition 7 shows that: Given a distribution P_0 of (θ, θ') on $\Theta \times \Theta$, for all distribution Q on $\Theta \times \Theta$ and any policy $\pi \in \tilde{\Pi}$, with probability at least $1 - \delta$, we have

$$
\begin{aligned}
&\left|\mathbb{E}_Q\left[\mathbb{E}_{\nu \times \bar{P}^*}\bigl[l(f_\theta, f_{\theta'}, \pi; X)\bigr] - \frac{1}{n}\sum_{i=1}^n l(f_\theta, f_{\theta'}, \pi; X_i)\right]\right| \\
&\quad \le 16V_{\max}^2\lambda \cdot \mathbb{E}_{Q, \nu \times \bar{P}^*}\bigl[l(f_\theta, f_{\theta'}, \pi; X)\bigr] + \frac{1}{n\lambda}\left[\mathrm{KL}(Q \,\|\, P_0) + \log\frac{2|\tilde{\Pi}|}{\delta}\right], \qquad \text{(I.3)}
\end{aligned}
$$

where $\lambda \le 1/(8V_{\max}^2)$.

Step 2: Control the fluctuation of both sides of inequality (I.3) **introduced by** Q**.**

To derive a generalization error bound for any function pair (θ, θ') in $\mathcal{F}_{\mathrm{tf}} \times \mathcal{F}_{\mathrm{tf}}$, we set Q as the uniform distribution on a neighborhood area of (θ, θ') , P_0 as the uniform distribution $\Theta \times \Theta$, and control the fluctuation of the left-hand side of inequality (I.3) due to the averaging according to Q.

We define the difference between the functions of different parameter pairs $(\tilde{\theta}, \tilde{\theta}')$ and (θ, θ') as

$$e(\tilde{\theta}, \tilde{\theta}', \theta, \theta', X) = l(f_{\tilde{\theta}}, f_{\tilde{\theta}'}, \pi; X) - l(f_\theta, f_{\theta'}, \pi; X).$$

To control the fluctuation of the left-hand side of inequality (I.3) due to the average according to Q, we need to upper bound $e(\tilde{\theta}, \tilde{\theta}', \theta, \theta', X)$ for all $X \in \mathbb{R}^{N \times d}$, which can be achieved by the following result.

Proposition 8. *For any input $X \in \mathbb{R}^{N \times d}$, any functions $g_{\mathrm{tf}}(X; W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}, w)$ and $g_{\mathrm{tf}}(X; \tilde{W}_{QK}^{1:L}, \tilde{W}_V^{1:L}, \tilde{a}^{1:L}, \tilde{b}^{1:L}, \tilde{w}) \in \mathcal{F}_{\mathrm{tf}}(B_a, B_b, B_{QK}, B_V, B_w)$, and two positive conjugate numbers $p, q \in \mathbb{R}$, we have*

$$
\begin{aligned}
&\bigl|g_{\mathrm{tf}}(X; W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}, w) - g_{\mathrm{tf}}(X; \tilde{W}_{QK}^{1:L}, \tilde{W}_V^{1:L}, \tilde{a}^{1:L}, \tilde{b}^{1:L}, \tilde{w})\bigr| \\
&\quad \le \sum_{i=1}^L \alpha_i \cdot (\beta_i + \iota_i + \kappa_i + \rho_i) + \|w - \tilde{w}\|_q,
\end{aligned}
$$

where

$$\begin{aligned}
\alpha_i &= B_w\big[B_V(1+4c_{p,q}B_{QK}) + d^{\frac{1}{p}} m B_a B_b\big]^{L-i},\\
\beta_i &= 2c_{p,q}B_V\|W_{QK}^{(i)\top} - \tilde{W}_{QK}^{(i)\top}\|_{p,q},\\
\iota_i &= \|W_V^{(i)\top} - \tilde{W}_V^{(i)\top}\|_{p,q},\\
\kappa_i &= B_b\bigg[\sum_{k=1}^{d}\bigg(\sum_{j=1}^{m}\big|a_{kj}^{(i)} - \tilde{a}_{kj}^{(i)}\big|\bigg)^p\bigg]^{\frac{1}{p}},\\
\rho_i &= B_a\bigg[\sum_{k=1}^{d}\bigg(\sum_{j=1}^{m}\big\|b_{kj}^{(i)} - \tilde{b}_{kj}^{(i)}\big\|_q\bigg)^p\bigg]^{\frac{1}{p}},
\end{aligned}$$

for $i \in [L]$.

Proof. See Appendix M.2 for a detailed proof. □

Motivated by Proposition 8, we define the upper bound of the difference of functions in $\mathcal{F}_{\mathrm{tf}}$ with different parameters θ and $\tilde{\theta}$ as

$$\begin{aligned}
\Delta(\theta,\tilde{\theta}) = \sum_{i=1}^{L} B_w\big[B_V(1+4c_{p,q}B_{QK}) + d^{\frac{1}{p}} m B_a B_b\big]^{L-i}\bigg\{&2c_{p,q}B_V\|W_{QK}^{(i)\top} - \tilde{W}_{QK}^{(i)\top}\|_{p,q}\\
&+ \|W_V^{(i)\top} - \tilde{W}_V^{(i)\top}\|_{p,q} + B_b\bigg[\sum_{k=1}^{d}\bigg(\sum_{j=1}^{m}\big|a_{kj}^{(i)} - \tilde{a}_{kj}^{(i)}\big|\bigg)^p\bigg]^{\frac{1}{p}}\\
&+ B_a\bigg[\sum_{k=1}^{d}\bigg(\sum_{j=1}^{m}\big\|b_{kj}^{(i)} - \tilde{b}_{kj}^{(i)}\big\|_q\bigg)^p\bigg]^{\frac{1}{p}}\bigg\} + \|w-\tilde{w}\|_q.
\end{aligned}$$

Then we can upper bound the absolute value of $e(\tilde{\theta},\tilde{\theta}',\theta,\theta',X)$ as

$$\begin{aligned}
&\big|e(\tilde{\theta},\tilde{\theta}',\theta,\theta',X)\big|\\
&\quad\le \Big|\big(f_{\tilde{\theta}}(\bar{S},\bar{A}) - \bar{r}(\bar{S},\bar{A}) - \gamma f_{\tilde{\theta}'}(\bar{S}',\pi)\big)^2 - \big(f_{\theta}(\bar{S},\bar{A}) - \bar{r}(\bar{S},\bar{A}) - \gamma f_{\theta'}(\bar{S}',\pi)\big)^2\Big|\\
&\qquad + \Big|\big(\mathcal{T}^\pi f_{\tilde{\theta}'}(\bar{S},\bar{A}) - \bar{r}(\bar{S},\bar{A}) - \gamma f_{\tilde{\theta}'}(\bar{S}',\pi)\big)^2 - \big(\mathcal{T}^\pi f_{\theta'}(\bar{S},\bar{A}) - \bar{r}(\bar{S},\bar{A}) - \gamma f_{\theta'}(\bar{S}',\pi)\big)^2\Big|\\
&\quad\le 4V_{\max}\big(\Delta(\tilde{\theta},\theta) + 3\gamma\Delta(\tilde{\theta}',\theta')\big),
\end{aligned}\tag{I.4}$$

where the first inequality follows from the triangle inequality, and the second inequality follows from that $f_\theta \in [-V_{\max}, V_{\max}]$ and $r \in [-R_{\max}, R_{\max}]$. For any fixed pair of parameters (θ,θ'), using inequality (I.4), we can upper bound the generalization error for a fixed parameter pair (θ,θ') by the left-hand side of inequality (I.3) as

$$\begin{aligned}
&\bigg|\mathbb{E}_{\nu\times\bar{P}^*}\big[l(f_\theta,f_{\theta'},\pi;X)\big] - \frac{1}{n}\sum_{i=1}^{n} l(f_\theta,f_{\theta'},\pi;X_i)\bigg|\\
&\quad\le \bigg|\mathbb{E}_{(\tilde{\theta},\tilde{\theta}')\sim Q}\bigg[\mathbb{E}_{\nu\times\bar{P}^*}\big[l(f_{\tilde{\theta}},f_{\tilde{\theta}'},\pi;X)\big] - \frac{1}{n}\sum_{i=1}^{n} l(f_{\tilde{\theta}},f_{\tilde{\theta}'},\pi;X_i)\bigg]\bigg|\\
&\qquad + \bigg|\mathbb{E}_{(\tilde{\theta},\tilde{\theta}')\sim Q}\bigg[\mathbb{E}_{\nu\times\bar{P}^*}\big[e(\tilde{\theta},\tilde{\theta}',\theta,\theta',X)\big] - \frac{1}{n}\sum_{i=1}^{n} e(\tilde{\theta},\tilde{\theta}',\theta,\theta',X_i)\bigg]\bigg|\\
&\quad\le \bigg|\mathbb{E}_{(\tilde{\theta},\tilde{\theta}')\sim Q}\bigg[\mathbb{E}_{\nu\times\bar{P}^*}\big[l(f_{\tilde{\theta}},f_{\tilde{\theta}'},\pi;X)\big] - \frac{1}{n}\sum_{i=1}^{n} l(f_{\tilde{\theta}},f_{\tilde{\theta}'},\pi;X_i)\bigg]\bigg|\\
&\qquad + 8V_{\max}\mathbb{E}_{(\tilde{\theta},\tilde{\theta}')\sim Q}\big[\Delta(\tilde{\theta},\theta) + 3\gamma\Delta(\tilde{\theta}',\theta')\big].
\end{aligned}\tag{I.5}$$

Similarly, for a fixed parameter pair of parameters(θ,θ'), the first term in the right-hand side of inequality (I.3) can be upper bounded as

$$\begin{aligned}&\mathbb{E}_{(\tilde\theta,\tilde\theta')\sim Q}\Big[\mathbb{E}_{\nu\times\bar{P^*}}\big[l(f_{\tilde\theta},f_{\tilde\theta'},\pi;X)\big]\Big]\\&\qquad\le\mathbb{E}_{\nu\times\bar{P^*}}[l(f_\theta,f_{\theta'},\pi;X)]+4V_{\max}\mathbb{E}_{(\tilde\theta,\tilde\theta')\sim Q}\big[\Delta(\tilde\theta,\theta)+3\gamma\Delta(\tilde\theta',\theta')\big].\end{aligned}\tag{I.6}$$

Substituting Eqn. (I.5) and (I.6) into Eqn. (I.3), we derive that : Given a distribution P_0 of (θ,θ') on $\Theta\times\Theta$, for all distribution Q on $\Theta\times\Theta$, any policy $\pi\in\tilde\Pi$ and any $(\theta,\theta')\in\Theta\times\Theta$, with probability at least $1-\delta$, we have,

$$\begin{aligned}&\bigg|\mathbb{E}_{\nu\times\bar{P^*}}\big[l(f_\theta,f_{\theta'},\pi;X)\big]-\frac{1}{n}\sum_{i=1}^n l(f_\theta,f_{\theta'},\pi;X_i)\bigg|\\&\quad\le V_{\max}(64V_{\max}^2\lambda+8)\mathbb{E}_{(\tilde\theta,\tilde\theta')\sim Q}\big[\Delta(\tilde\theta,\theta)+3\gamma\Delta(\tilde\theta',\theta')\big]+16V_{\max}^2\lambda\mathbb{E}_{\nu\times\bar{P^*}}\big[l(f_\theta,f_{\theta'},\pi;X)\big]\\&\qquad+\frac{1}{n\lambda}\bigg[\mathrm{KL}(Q\,\|\,P_0)+\log\frac{2|\tilde\Pi|}{\delta}\bigg],\end{aligned}$$

where $\lambda\le 1/(8V_{\max}^2)$. We take $\lambda=1/(32V_{\max}^2)$, then

$$\begin{aligned}&\bigg|\mathbb{E}_{\nu\times\bar{P^*}}\big[l(f_\theta,f_{\theta'},\pi;X)\big]-\frac{1}{n}\sum_{i=1}^n l(f_\theta,f_{\theta'},\pi;X_i)\bigg|\\&\quad\le 10V_{\max}\mathbb{E}_{(\tilde\theta,\tilde\theta')\sim Q}\big[\Delta(\tilde\theta,\theta)+3\gamma\Delta(\tilde\theta',\theta')\big]+\frac{1}{2}\mathbb{E}_{\nu\times\bar{P^*}}\big[l(f_\theta,f_{\theta'},\pi;X)\big]\\&\qquad+\frac{32V_{\max}^2}{n}\bigg[\mathrm{KL}(Q\,\|\,P_0)+\log\frac{2|\tilde\Pi|}{\delta}\bigg].\end{aligned}\tag{I.7}$$

Step 3: Specify the distributions P_0 and Q on the function class $\mathcal{F}_{\mathrm{tf}}$.

For a fixed parameters pair (θ,θ'), we set P_0 as the product of the uniform distribution of each parameter on the whole space and Q as the product of the uniform distribution of each parameter on the neighborhood around (θ,θ'), i.e.,

$$\begin{aligned}P_0=&\bigg\{\mathrm{U}\big(\mathbb{B}(0,B_w,\|\cdot\|_q)\big)\cdot\prod_{i=1}^L\bigg[\mathrm{U}\big(\mathbb{B}(0,B_{QK},\|\cdot\|_{p,q})\big)\cdot\mathrm{U}\big(\mathbb{B}(0,B_V,\|\cdot\|_{p,q})\big)\\&\cdot\Big(\mathrm{U}\big(\mathbb{B}(0,B_a,|\cdot|)\big)\cdot\mathrm{U}\big(\mathbb{B}(0,B_b,\|\cdot\|_q)\big)\Big)^{md}\bigg]\bigg\}^2,\quad\text{and}\\Q=&\bigg\{\mathrm{U}\big(\mathbb{B}(w,\varepsilon_w,\|\cdot\|_q)\big)\cdot\prod_{i=1}^L\bigg[\mathrm{U}\big(\mathbb{B}(W_{QK}^{(i)\top},\varepsilon_{QK}^{(i)},\|\cdot\|_{p,q})\big)\cdot\mathrm{U}\big(\mathbb{B}(W_V^{(i)\top},\varepsilon_V^{(i)},\|\cdot\|_{p,q})\big)\\&\cdot\prod_{j\in[m],k\in[d]}\Big(\mathrm{U}\big(\mathbb{B}(a_{kj}^{(i)},\varepsilon_{a,kj}^{(i)},|\cdot|)\big)\cdot\mathrm{U}\big(\mathbb{B}(b_{kj}^{(i)},\varepsilon_{b,kj}^{(i)},\|\cdot\|_q)\big)\Big)\bigg]\bigg\}\\&\cdot\bigg\{\mathrm{U}\big(\mathbb{B}(w',\varepsilon_w,\|\cdot\|_q)\big)\cdot\prod_{i=1}^L\bigg[\mathrm{U}\big(\mathbb{B}(W_{QK}'^{(i)\top},\varepsilon_{QK}^{(i)},\|\cdot\|_{p,q})\big)\cdot\mathrm{U}\big(\mathbb{B}(W_V'^{(i)\top},\varepsilon_V^{(i)},\|\cdot\|_{p,q})\big)\\&\cdot\prod_{j\in[m],k\in[d]}\Big(\mathrm{U}\big(\mathbb{B}(a_{kj}^{(i)\prime},\varepsilon_{a,kj}^{(i)},|\cdot|)\big)\cdot\mathrm{U}\big(\mathbb{B}(b_{kj}^{(i)\prime},\varepsilon_{b,kj}^{(i)},\|\cdot\|_q)\big)\Big)\bigg]\bigg\}\end{aligned}$$

where $\mathbb{B}(x,r,\|\cdot\|)$ denotes the ball $\{y\,|\,\|y-x\|<r\}$ in some metric space $(\mathcal{X},\|\cdot\|)$, and $\mathrm{U}(\cdot)$ denotes the uniform distribution on some set. For a constant $C>0$, we define $\varepsilon_w=\Delta=C/[(1+3\gamma)(4L+1)n]$. For $i\in[L]$, $j\in[m]$ and $k\in[d]$, we set

$$\begin{aligned}\varepsilon_{QK}^{(i)}&=\big(2c_{p,q}B_VB_w\big)^{-1}\big[B_V(1+4c_{p,q}B_{QK})+d^{\frac{1}{p}}mB_aB_b\big]^{-L+i}\Delta,\\\varepsilon_V^{(i)}&=B_w^{-1}\big[B_V(1+4c_{p,q}B_{QK})+d^{\frac{1}{p}}mB_aB_b\big]^{-L+i}\Delta,\end{aligned}$$

$$\varepsilon_{a,kj}^{(i)} = d^{-\frac{1}{p}}\big(mB_bB_w\big)^{-1}\big[B_V(1+4c_{p,q}B_{QK}) + d^{\frac{1}{p}}mB_aB_b\big]^{-L+i}\Delta,$$
$$\varepsilon_{b,kj}^{(i)} = d^{-\frac{1}{p}}\big(mB_aB_w\big)^{-1}\big[B_V(1+4c_{p,q}B_{QK}) + d^{\frac{1}{p}}mB_aB_b\big]^{-L+i}\Delta.$$

By Proposition 8, we then have

$$\mathbb{E}_{(\tilde{\theta},\tilde{\theta}')\sim Q}\big[\Delta(\tilde{\theta},\theta) + 3\gamma\Delta(\tilde{\theta}',\theta')\big] \le \frac{C}{n}. \tag{I.8}$$

Since the distributions P_0 and Q are the products of the distributions of each parameters, $\mathrm{KL}(Q \,\|\, P_0)$ is the sum of the KL-divergences between the distributions of each parameters. For $i \in [L]$, the KL divergence between the distributions of W_{QK} can be upper bounded as

$$\begin{aligned}
&\mathrm{KL}\Big(\mathrm{U}\big(\mathbb{B}(W_{QK}^{(i)\top}, \varepsilon_{QK}^{(i)}, \|\cdot\|_{p,q})\big)\Big\|\mathrm{U}\big(\mathbb{B}(0, B_{QK}, \|\cdot\|_{p,q})\big)\Big) \\
&\quad = d^2 \log\bigg(\frac{B_{QK}}{\varepsilon_{QK}^{(i)}}\bigg) \\
&\quad \le 2(L-i)d^2 \log\bigg(\frac{4mdB_VB_{QK}B_aB_b}{\Delta}\bigg) + d^2 \log B_w,
\end{aligned}$$

where the equality follows from the fact that $W_{QK}^{(i)} \in \mathbb{R}^{d\times d}$ for all $i \in [L]$, in which the logarithm of the ratio between two $\ell_{p,q}$-norm balls is equal to d^2 times the logarithm of the ratio between the radiuses.

We note that the product $B_VB_{QK}B_aB_bB_w$ is defined as $\bar{B}$ in Theorem 2, which is adopted to simplify the result. Similar bounds for the KL divergence of the distributions of parameters $W_V^{(i)}$, $a_{kj}^{(i)}$, $b_{kj}^{(i)}$ and w for $i \in [L]$, $k \in [d]$ and $j \in [m]$ can be derived by replacing d^2 by the dimension of the parameter. Thus, we have

$$\mathrm{KL}(Q \,\|\, P_0) \le 2(m+1)L^2d^2 \log\bigg(\frac{4mdB_VB_{QK}B_aB_b}{\Delta}\bigg) + 2(m+1)Ld^2 \log B_w. \tag{I.9}$$

Substituting inequalities (I.8) and (I.9) into inequality (I.7), we derive that for any $(\theta,\theta') \in \Theta^2$, with probability at least $1-\delta$

$$\begin{aligned}
&\bigg|\mathbb{E}_{\nu\times\bar{P}^*}\big[l(f_\theta, f_{\theta'}, \pi; X)\big] - \frac{1}{n}\sum_{i=1}^n l(f_\theta, f_{\theta'}, \pi; X_i)\bigg| \\
&\quad \le 10V_{\max}\frac{C}{n} + \frac{1}{2}\mathbb{E}_{\nu\times\bar{P}^*}\big[l(f_\theta, f_{\theta'}, \pi; X)\big] + \frac{32V_{\max}^2}{n}\bigg[2(m+1)L^2d^2 \log\bigg(\frac{4mdB_VB_{QK}B_aB_b}{\Delta}\bigg) \\
&\qquad + 2(m+1)Ld^2 \log B_w + \log\frac{2|\tilde{\Pi}|}{\delta}\bigg].
\end{aligned} \tag{I.10}$$

Step 4: Cover the policy class Π.

Note that inequality (I.10) only applies to the situation where the policy class is finite. When the policy class is infinite, we consider the covering of the policy class with respect to $d_\infty(\cdot,\cdot)$. The ε-covering number of the policy class with respect to $d_\infty(\cdot,\cdot)$ is denoted as $\mathcal{N}(\Pi, \varepsilon, d_\infty)$, and the corresponding ε-cover is $\mathcal{C}(\Pi, \varepsilon, d_\infty)$, which is defined in Section 2. From the definition of $d_\infty(\cdot,\cdot)$, we have

$$d_\infty(\pi,\pi') = \sup_{\bar{S}\in\bar{\mathcal{S}}} \sum_{\bar{A}\in\bar{\mathcal{A}}} \big|\pi(\bar{A}\,|\,\bar{S}) - \pi'(\bar{A}\,|\,\bar{S})\big|$$

$$\begin{aligned}
\big|f(\bar{S},\pi) - f(\bar{S},\pi')\big| &= \bigg|\sum_{\bar{A}\in\bar{\mathcal{A}}}\big[\pi(\bar{A}\,|\,\bar{S}) - \pi'(\bar{A}\,|\,\bar{S})\big]f(\bar{S},\bar{A})\bigg| \\
&\le \sum_{\bar{A}\in\bar{\mathcal{A}}}\big|\pi(\bar{A}\,|\,\bar{S}) - \pi'(\bar{A}\,|\,\bar{S})\big| \cdot \big|f(\bar{S},\bar{A})\big| \\
&\le V_{\max}d_\infty(\pi,\pi')
\end{aligned} \tag{I.11}$$

$$\begin{aligned}|\mathcal{T}^{\pi} f(\bar{S},\bar{A}) - \mathcal{T}^{\pi'} f(\bar{S},\bar{A})| &= \gamma\Big|\mathbb{E}_{(\bar{S}')\sim\hat{\bar{P}}^*(\cdot\,|\,\bar{S},\bar{A})}\big[f(\bar{S}',\pi) - f(\bar{S}',\pi')\,\big|\,\bar{S},\bar{A}\big]\Big| \\ &\le \gamma V_{\max} d_\infty(\pi,\pi'). \end{aligned} \tag{I.12}$$

Thus, we can upper bound the difference between $l(f,\tilde{f},\pi;X)$ and $l(f,\tilde{f},\pi';X)$ by $d_\infty(\pi,\pi')$ as

$$\begin{aligned}&\big|l(f,\tilde{f},\pi;X) - l(f,\tilde{f},\pi';X)\big| \\ &\quad\le \Big|\big(f(\bar{S},\bar{A}) - \bar{r} - \gamma\tilde{f}(\bar{S},\pi)\big)^2 - \big(f(\bar{S},\bar{A}) - \bar{r} - \gamma\tilde{f}(\bar{S},\pi')\big)^2\Big| \\ &\qquad + \Big|\big(\mathcal{T}^{\pi}\tilde{f}(\bar{S},\bar{A}) - \bar{r} - \gamma\tilde{f}(\bar{S},\pi)\big)^2 - \big(\mathcal{T}^{\pi'}\tilde{f}(\bar{S},\bar{A}) - \bar{r} - \gamma\tilde{f}(\bar{S},\pi')\big)^2\Big| \\ &\quad= \Big|\big(\gamma\tilde{f}(\bar{S},\pi') - \gamma\tilde{f}(\bar{S},\pi)\big)\big(2f(\bar{S},\bar{A}) - 2\bar{r} - \gamma\tilde{f}(\bar{S},\pi') - \gamma\tilde{f}(\bar{S},\pi)\big)\Big|, \\ &\qquad + \Big|\big(\mathcal{T}^{\pi}\tilde{f}(\bar{S},\bar{A}) - \mathcal{T}^{\pi'}\tilde{f}(\bar{S},\bar{A}) - \gamma\tilde{f}(\bar{S},\pi) + \gamma\tilde{f}(\bar{S},\pi')\big) \\ &\qquad \cdot \Big(\mathcal{T}^{\pi'}\tilde{f}(\bar{S},\bar{A}) - 2\bar{r} - \gamma\tilde{f}\big(\bar{S},\pi' - \gamma\tilde{f}(\bar{S},\pi)\big)\Big)^2\Big|,\end{aligned}$$

where the inequality follows from the triangle inequality. Combined with inequalities (I.11) and (I.12), it can be further upper bounded as

$$\begin{aligned}&\big|l(f,\tilde{f},\pi;X) - l(f,\tilde{f},\pi';X)\big| \\ &\quad\le \gamma V_{\max} d_\infty(\pi,\pi')\cdot 4V_{\max} + 2\gamma V_{\max} d_\infty(\pi,\pi')\cdot 4V_{\max} \\ &\quad= 12\gamma V_{\max}^2 d_\infty(\pi,\pi').\end{aligned} \tag{I.13}$$

From the definition of the ε-cover and inequality (I.13), for any $\pi \in \Pi$, there exist a policy $\pi' \in \mathcal{C}(\Pi,\varepsilon,d_\infty)$ such that for any $f,\tilde{f} \in \mathcal{F}_{\mathrm{tf}}$,

$$\big|l(f,\tilde{f},\pi;X) - l(f,\tilde{f},\pi';X)\big| \le 12\gamma\varepsilon V_{\max}^2. \tag{I.14}$$

Substituting inequality (I.14) into the term involving $l(f,\tilde{f},\pi;X)$ in inequality (I.10), we have that for all $f_\theta, f_{\theta'} \in \mathcal{F}_{\mathrm{tf}}$ and all policy $\pi \in \Pi$, with probability at least $1-\delta$,

$$\begin{aligned}&\Big|\mathbb{E}_{\nu\times\bar{P}^*}[l(f_\theta,f_{\theta'},\pi;X)] - \frac{1}{n}\sum_{i=1}^{n} l(f_\theta,f_{\theta'},\pi;X_i)\Big| \\ &\quad\le 30\gamma V_{\max}^2\varepsilon + 10V_{\max}\frac{C}{n} + \frac{1}{2}\mathbb{E}_{\nu\times\bar{P}^*}\big[l(f_\theta,f_{\theta'},\pi;X)\big] \\ &\qquad + \frac{32V_{\max}^2}{n}\bigg[2(m+1)L^2d^2\log\Big(\frac{4mdB_VB_{QK}B_aB_b}{\Delta}\Big) \\ &\qquad + 2(m+1)Ld^2\log B_w + \log\frac{2\mathcal{N}(\Pi,\varepsilon,d_\infty)}{\delta}\bigg].\end{aligned}$$

Setting $\varepsilon = 1/n$ and $C = 5V_{\max}$, we obtain the desired result. Therefore, we conclude the proof of Proposition 2. □

J Proof of Proposition 3

Proof of Proposition 3. We adopt a Bayesian framework to prove the desired result. The total variation is first upper bounded through Pinsker's inequality. Then the derived upper bounded is further relaxed by the bounds related to the KL divergence. For ease of notation, we denote the parameters of the neural network as

$$\theta = [W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}].$$

Step 1: Bound the total variation distance with Pinsker's inequality.

From Pinsker's inequality, the total variation between two conditional distribution can be bounded as

Lemma 9 (Lemma 25 in [49]). *For any two conditional probability densities $P(\cdot \mid \bar{S}, \bar{A}), P'(\cdot \mid \bar{S}, \bar{A})$ and any distribution $\nu \in \Delta(\bar{\mathcal{S}} \times \bar{\mathcal{A}})$,we have*

$$\mathbb{E}_\nu\Big[\mathrm{TV}\big(P(\cdot \mid \bar{S}, \bar{A}), P'(\cdot \mid \bar{S}, \bar{A})\big)^2\Big] \leq -2\log\bigg(\mathbb{E}_{(\bar{S},\bar{A})\sim\nu, \bar{S}'\sim P(\cdot \mid \bar{S},\bar{A})}\bigg[\exp\bigg(-\frac{1}{2}\log\frac{P(\bar{S}' \mid \bar{S}, \bar{A})}{P'(\bar{S}' \mid \bar{S}, \bar{A})}\bigg)\bigg]\bigg).$$

Thus, we only need to upper bound the right-hand side of the inequality in Lemma 9. We adopt a Bayesian framework to relax this upper bound.

Lemma 10 (Lemma 2.1 in [43]). *Given a distribution P on Θ, for all $Q \gg P$ on Θ and all measurable real-valued function $L(\theta; \mathcal{D}) : \Theta \times (\bar{\mathcal{S}} \times \bar{\mathcal{A}})^n \to \mathbb{R}$, we have*

$$\mathbb{E}_{\mathcal{D}}\bigg[\exp\Big\{\mathbb{E}_Q\big[L(\theta; \mathcal{D}) - \log\mathbb{E}_{\mathcal{D}}[e^{L(\theta;\mathcal{D})}]\big] - \mathrm{KL}(Q \,\|\, P)\Big\}\bigg] \leq 1,$$

where $\mathbb{E}_{\mathcal{D}}[\,\cdot\,]$ is the expectation with respect to the underlying distribution of $\{(\bar{S}_i, \bar{A}_i, \bar{S}_i'\}_{i=1}^n$, i.e., $(\nu \times P^*)^n$.

By Lemma 10 and the Chernoff inequality, we have that with probability at least $1 - \delta/2$,

$$-\mathbb{E}_Q\big[\log\mathbb{E}_{\mathcal{D}}[e^{L(\theta;\mathcal{D})}]\big] \leq -\mathbb{E}_Q\big[L(\theta; \mathcal{D})\big] + \mathrm{KL}(Q \,\|\, Q_0) + \log\frac{2}{\delta}, \tag{J.1}$$

where $\mathbb{E}_{\mathcal{D}}[\,\cdot\,]$ is the expectation with respect to the underlying distribution of $\{(\bar{S}_i, \bar{A}_i, \bar{S}_i')\}_{i=1}^n$, i.e., $(\nu \times P^*)^n$, and Q and Q_0 are two distributions on Θ.

Take $L(\theta; \mathcal{D}) = -\frac{1}{4}\sum_{i=1}^n \log(P^*(\bar{S}_i' \mid \bar{S}_i, \bar{A}_i)/P_\theta(\bar{S}_i' \mid \bar{S}_i, \bar{A}_i))$, where $\mathcal{D} = \{(\bar{S}_i, \bar{A}_i, r_i, \bar{S}_i')\}_{i=1}^n$. Then the left-hand side of inequality (J.1) becomes

$$\begin{aligned}
-\mathbb{E}_Q\big[\log\mathbb{E}_{\mathcal{D}}[e^{L(\theta;\mathcal{D})}]\big] &= -\mathbb{E}_Q\bigg[\log\mathbb{E}_{\mathcal{D}}\bigg[\exp\Big(-\frac{1}{4}\sum_{i=1}^n\log\frac{P^*(\bar{S}_i' \mid \bar{S}_i, \bar{A}_i)}{P_\theta(\bar{S}_i' \mid \bar{S}_i, \bar{A}_i)}\Big)\bigg]\bigg] \\
&= -\mathbb{E}_Q\bigg[n\log\mathbb{E}_{(\bar{S},\bar{A},\bar{S}')\sim\nu\times P^*}\bigg[\exp\Big(-\frac{1}{4}\log\frac{P^*(\bar{S}' \mid \bar{S}, \bar{A})}{P_\theta(\bar{S}' \mid \bar{S}, \bar{A})}\Big)\bigg]\bigg].
\end{aligned}$$

Step 2: Control the fluctuation of the both sides of inequality (J.1) introduced by Q.

Since $\hat{P}_{\mathrm{MLE}}$ is a random variable, we want to derive an uniform bound for all $\theta \in \Theta$. Because the left-hand side of inequality (J.1) takes the expectation with respect to the distribution Q on Θ, which is chosen as the uniform distribution on the neighborhood around a fixed parameter $\theta \in \Theta$, we need to control the fluctuation of the left-hand side of inequality (J.1) due to the distribution Q around θ. For any two parameters θ and $\tilde{\theta}$, we define the logarithm of the ratio between the transition kernels induced by them as

$$\begin{aligned}
e(\theta, \tilde{\theta}; \bar{S}', \bar{S}, \bar{A}) &= \log\frac{P_{\tilde{\theta}}(\bar{S}' \mid \bar{S}, \bar{A})}{P_\theta(\bar{S}' \mid \bar{S}, \bar{A})} \\
&= \log\bigg(\exp\Big(-\frac{\big\|\bar{S}' - F_{\tilde{\theta}}(\bar{S}, \bar{A})\big\|_{\mathrm{F}}^2}{2\sigma^2}\Big)\Big/\exp\Big(-\frac{\big\|\bar{S}' - F_\theta(\bar{S}, \bar{A})\big\|_{\mathrm{F}}^2}{2\sigma^2}\Big)\bigg) \\
&= \frac{\big\|\bar{S}' - F_\theta(\bar{S}, \bar{A})\big\|_{\mathrm{F}}^2 - \big\|\bar{S}' - F_{\tilde{\theta}}(\bar{S}, \bar{A})\big\|_{\mathrm{F}}^2}{2\sigma^2}.
\end{aligned}$$

To upper bound the absolute value of $e(\theta, \tilde{\theta}; \bar{S}', \bar{S}, \bar{A})$, we need to bound the norm of the output of the neural network.

Proposition 11. *For any $X \in \mathbb{R}^{N\times d}$, any $W_{QK}, W_V \in \mathbb{R}^{d\times d}$, $a \in \mathbb{R}^{dm}$, $b \in \mathbb{R}^{d\times dm}$ and two positive conjugate numbers $p, q \in \mathbb{R}$, we have*

$$\begin{aligned}
&\Big\|\big(\mathrm{SM}(XW_{QK}X^\top)XW_V + \mathrm{rFF}(X, a, b)\big)^\top\Big\|_{p,\infty} \\
&\quad\leq \|W_V^\top\|_{p,q}\|X^\top\|_{p,\infty} + \bigg[\sum_{k=1}^d\bigg(\sum_{j=1}^m |a_{kj}|\|b_{kj}\|_q\|X^\top\|_{p,\infty}\bigg)^p\bigg]^{1/p}.
\end{aligned}$$

Proof. See Appendix M.5 for a detailed proof. $\square$

Proposition 11 shows that $\|F_{\tilde{\theta}}(\bar{S},\bar{A})\|_{\mathrm{F}} \leq \sqrt{N}B^*$ for all $\theta \in \Theta$, $\bar{S} \in \bar{\mathcal{S}}$ and $\bar{A} \in \bar{\mathcal{A}}$, where $B^* = B_V + md^{1/2}B_aB_b$. As a consequence, we have

$$\begin{aligned}\big|e(\theta,\tilde{\theta};\bar{S}',\bar{S},\bar{A})\big| &\leq \frac{1}{2\sigma^2}\Big(\big\|\bar{S}' - F_\theta(\bar{S},\bar{A})\big\|_{\mathrm{F}} + \big\|\bar{S}' - F_{\tilde{\theta}}(\bar{S},\bar{A})\big\|_{\mathrm{F}}\Big)\big\|F_\theta(\bar{S},\bar{A}) - F_{\tilde{\theta}}(\bar{S},\bar{A})\big\|_{\mathrm{F}} \\ &\leq \frac{1}{\sigma^2}\big(\|\bar{\varepsilon}\|_{\mathrm{F}} + \sqrt{N}B^*\big)\big\|F_\theta(\bar{S},\bar{A}) - F_{\tilde{\theta}}(\bar{S},\bar{A})\big\|_{\mathrm{F}},\end{aligned} \tag{J.2}$$

where these two inequalities follow from the triangle inequality. For two parameters θ and $\tilde{\theta}$, we define the upper bound of the difference between the dynamic functions induced by them as

$$\Delta(\theta,\tilde{\theta}) = \max_{(\bar{S},\bar{A})\in\bar{\mathcal{S}}\times\bar{\mathcal{A}}}\big\|F_\theta(\bar{S},\bar{A}) - F_{\tilde{\theta}}(\bar{S},\bar{A})\big\|_{\mathrm{F}}.$$

For a fixed parameter θ, the left-hand side of inequality (J.1) can be lower bounded as

$$\begin{aligned}&-\mathbb{E}_Q\big[\log\mathbb{E}_{\mathcal{D}}e^{L(\tilde{\theta};\mathcal{D})}\big] \\ &\quad= -\mathbb{E}_Q\bigg[n\log\mathbb{E}_{(\bar{S},\bar{A},\bar{S}')\sim\nu\times P^*}\bigg[\exp\bigg(-\frac{1}{4}e(\theta,\tilde{\theta};\bar{S}',\bar{S},\bar{A}) - \frac{1}{4}\log\frac{P^*(\bar{S}'\,|\,\bar{S},\bar{A})}{P_\theta(\bar{S}'\,|\,\bar{S},\bar{A})}\bigg)\bigg]\bigg] \\ &\quad\geq -\frac{n}{2}\log\mathbb{E}_{\nu\times P^*}\bigg[\exp\bigg(-\frac{1}{2}\log\frac{P^*(\bar{S}'\,|\,\bar{S},\bar{A})}{P_\theta(\bar{S}'\,|\,\bar{S},\bar{A})}\bigg)\bigg] \\ &\qquad\quad - \mathbb{E}_Q\bigg[\frac{n}{2}\log\mathbb{E}_{\nu\times P^*}\bigg[\exp\Big(-\frac{1}{2}e(\theta,\tilde{\theta};\bar{S}',\bar{S},\bar{A})\Big)\bigg]\bigg] \\ &\quad\geq \frac{n}{4}\mathbb{E}_\nu\Big[\mathrm{TV}\big(P^*(\cdot\,|\,\bar{S},\bar{A}),P_\theta(\cdot\,|\,\bar{S},\bar{A})\big)^2\Big] - \mathbb{E}_Q\bigg[\frac{n}{2}\log\mathbb{E}_{\nu\times P^*}\bigg[\exp\bigg(-\frac{1}{2}e(\theta,\tilde{\theta};\bar{S}',\bar{S},\bar{A})\bigg)\bigg]\bigg],\end{aligned} \tag{J.3}$$

where the first inequality follows from the Cauchy–Schwarz inequality, and the last inequality follows from Lemma 9. The second term of inequality (J.3) can be bounded as

$$\begin{aligned}&\log\mathbb{E}_{\nu\times P^*}\bigg[\exp\bigg(-\frac{1}{2}e(\theta,\tilde{\theta};\bar{S}',\bar{S},\bar{A})\bigg)\bigg] \\ &\quad\leq \log\mathbb{E}_{\nu\times P^*}\bigg[\exp\bigg(\frac{1}{2}\big|e(\theta,\tilde{\theta};\bar{S}',\bar{S},\bar{A})\big|\bigg)\bigg] \\ &\quad\leq \log\mathbb{E}_{\bar{\varepsilon}\sim\mathcal{N}(0,\sigma^2 I)}\bigg[\exp\bigg(\frac{1}{2\sigma^2}\big(\|\bar{\varepsilon}\|_{\mathrm{F}} + \sqrt{N}B^*\big)\Delta(\theta,\tilde{\theta})\bigg)\bigg] \\ &\quad= \sqrt{N}B^*\Delta(\theta,\tilde{\theta}) + \log\mathbb{E}_{\bar{\varepsilon}\sim\mathcal{N}(0,\sigma^2 I)}\bigg[\exp\bigg(\frac{1}{2\sigma^2}\|\bar{\varepsilon}\|_{\mathrm{F}}\Delta(\theta,\tilde{\theta})\bigg)\bigg],\end{aligned}$$

where the second inequality follows from inequality (J.2). Since Lemma 16 shows that $\|X\|_{\mathrm{F}} \leq \|X\|_{1,1}$, we further have

$$\begin{aligned}&\log\mathbb{E}_{\nu\times P^*}\bigg[\exp\bigg(-\frac{1}{2}e(\theta,\tilde{\theta};\bar{S}',\bar{S},\bar{A})\bigg)\bigg] \\ &\quad\leq \sqrt{N}B^*\Delta(\theta,\tilde{\theta}) + \log\mathbb{E}_{\bar{\varepsilon}\sim\mathcal{N}(0,\sigma^2 I)}\bigg[\exp\bigg(\frac{1}{2\sigma^2}\|\bar{\varepsilon}\|_{1,1}\Delta(\theta,\tilde{\theta})\bigg)\bigg] \\ &\quad= \sqrt{N}B^*\Delta(\theta,\tilde{\theta}) + Nd\log\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2)}\bigg[\exp\bigg(\frac{\Delta(\theta,\tilde{\theta})}{2\sigma^2}|\varepsilon|\bigg)\bigg],\end{aligned} \tag{J.4}$$

where the inequality follows from Lemma 16. The moment generating function of the folded normal distribution is (see [50])

$$\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2)}\Big[\exp\big(\lambda|\zeta|\big)\Big] = 2\exp(\sigma^2\lambda^2/2)\big[1 - \Phi(-\sigma\lambda)\big], \tag{J.5}$$

where $\Phi(\cdot)$ is the cumulative distribution function of $\mathcal{N}(0,1)$. From the Taylor expansion of $\Phi(\cdot)$, we have

$$2\big[1-\Phi(-\sigma\lambda)\big] \le 1+\sqrt{\frac{3}{\pi}}\sigma\lambda \tag{J.6}$$

for small enough λ. Since $\log(1+x)\le x$ for $x>0$, substituting inequalities (J.4), (J.5) and (J.6) into inequality (J.3), we have

$$\begin{aligned}
&-\mathbb{E}_Q\big[\log \mathbb{E}_{\mathcal{D}} e^{L(\tilde{\theta};\mathcal{D})}\big]\\
&\quad\ge \frac{n}{4}\mathbb{E}_\nu\Big[\mathrm{TV}\big(P^*(\cdot\,|\,\bar{S},\bar{A}),P_\theta(\cdot\,|\,\bar{S},\bar{A})\big)^2\Big]\\
&\qquad -\frac{n}{2}\mathbb{E}_{\tilde{\theta}\sim Q}\bigg[\sqrt{N}B^*\Delta(\theta,\tilde{\theta})+Nd\bigg(\frac{\Delta^2(\theta,\tilde{\theta})}{8\sigma^2}+\frac{1}{2\sigma}\sqrt{\frac{3}{\pi}}\Delta(\theta,\tilde{\theta})\bigg)\bigg],
\end{aligned} \tag{J.7}$$

for small enough $\Delta(\theta,\tilde{\theta})$, which is set to $O(1/n)$ later.

For the scaled right-hand side of inequality (J.1), we have

$$\begin{aligned}
&\frac{4}{n}\bigg\{-\mathbb{E}_Q\big[L(\theta;\mathcal{D})\big]+\mathrm{KL}(Q\,\|\,Q_0)+\log\frac{2}{\delta}\bigg\}\\
&\quad=\frac{1}{n}\sum_{i=1}^n\log\frac{P^*(\bar{S}_i'\,|\,\bar{S}_i,\bar{A}_i)}{P_\theta(\bar{S}_i'\,|\,\bar{S}_i,\bar{A}_i)}+\mathbb{E}_{\tilde{\theta}\sim Q}\bigg[\frac{1}{n}\sum_{i=1}^n e(\theta,\tilde{\theta};\bar{S}_i',\bar{S}_i,\bar{A}_i)\bigg]+\frac{4}{n}\bigg[\mathrm{KL}(Q\,\|\,Q_0)+\log\frac{2}{\delta}\bigg]\\
&\quad\le\frac{1}{n}\sum_{i=1}^n\log\frac{P^*(\bar{S}_i'\,|\,\bar{S}_i,\bar{A}_i)}{P_\theta(\bar{S}_i'\,|\,\bar{S}_i,\bar{A}_i)}+\mathbb{E}_{\tilde{\theta}\sim Q}\bigg[\frac{1}{n}\sum_{i=1}^n\frac{1}{\sigma^2}\big(\|\bar{\varepsilon}_i\|_{\mathrm{F}}+\sqrt{N}B^*\big)\Delta(\theta,\tilde{\theta})\bigg]\\
&\qquad+\frac{4}{n}\bigg[\mathrm{KL}(Q\,\|\,Q_0)+\log\frac{2}{\delta}\bigg],
\end{aligned} \tag{J.8}$$

where the last inequality follows from inequality (J.2) and the definition of $\Delta(\theta,\tilde{\theta})$. To upper bound the right-hand side of inequality (J.8), we need to upper bound $\|\bar{\varepsilon}\|_{\mathrm{F}}$, which can be achieved by combining the upper bound of the moment generating function of $\|\bar{\varepsilon}\|_{\mathrm{F}}$

$$\mathbb{E}\Big[\exp\big(\lambda\|\bar{\varepsilon}\|_{\mathrm{F}}\big)\Big]\le\mathbb{E}\Big[\exp\big(\lambda\|\bar{\varepsilon}\|_{1,1}\big)\Big]=\bigg(\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2)}\Big[\exp\big(\lambda|\varepsilon|\big)\Big]\bigg)^{Nd}\le\big(2\exp(\sigma^2\lambda^2/2)\big)^{Nd}$$

and the Chernoff inequality. Thus, with probability at least $1-\delta/2$, we have

$$\frac{1}{n}\sum_{i=1}^n\|\bar{\varepsilon}_i\|_{\mathrm{F}}\le\sqrt{2N^2d^2\sigma^2+\frac{2Nd\sigma^2}{n}\log\frac{2}{\delta}}. \tag{J.9}$$

Substituting inequalities (J.7), (J.8) and (J.9) into inequality (J.1), we have that for any $\theta\in\Theta$ and any two distributions Q and Q_0, the following inequality holds with probability at least $1-\delta$

$$\begin{aligned}
&\mathbb{E}_\nu\Big[\mathrm{TV}\big(P^*(\cdot\,|\,\bar{S},\bar{A}),P_\theta(\cdot\,|\,\bar{S},\bar{A})\big)^2\Big]\\
&\quad\le\frac{1}{n}\sum_{i=1}^n\log\frac{P^*(\bar{S}_i'\,|\,\bar{S}_i,\bar{A}_i)}{P_\theta(\bar{S}_i'\,|\,\bar{S}_i,\bar{A}_i)}+2\mathbb{E}_{\tilde{\theta}\sim Q}\bigg[\sqrt{N}B^*\Delta(\theta,\tilde{\theta})+Nd\bigg(\frac{\Delta^2(\theta,\tilde{\theta})}{8\sigma^2}+\frac{1}{2\sigma}\sqrt{\frac{3}{\pi}}\Delta(\theta,\tilde{\theta})\bigg)\bigg]\\
&\qquad+\bigg(\sqrt{2N^2d^2\sigma^2+\frac{2Nd\sigma^2}{n}\log\frac{2}{\delta}}+\sqrt{N}B^*\bigg)\mathbb{E}_{\tilde{\theta}\sim Q}\big[\Delta(\theta,\tilde{\theta})\big]+\frac{4}{n}\bigg[\mathrm{KL}(Q\,\|\,Q_0)+\log\frac{2}{\delta}\bigg].
\end{aligned} \tag{J.10}$$

For any fixed $\theta=[W_{QK}^{1:L},W_V^{1:L},a^{1:L},b^{1:L}]$, we set Q_0 as the product of the uniform distribution of each parameter on the whole space and Q as the product of the uniform distribution of each parameter on the neighborhood around θ, i.e.,

$$\begin{aligned}
Q_0=&\,\mathrm{U}\big(\mathbb{B}(0,B_w,\|\cdot\|_q)\big)\cdot\prod_{i=1}^L\bigg[\mathrm{U}\big(\mathbb{B}(0,B_{QK},\|\cdot\|_{p,q})\big)\cdot\mathrm{U}\big(\mathbb{B}(0,B_V,\|\cdot\|_{p,q})\big)\\
&\cdot\Big(\mathrm{U}\big(\mathbb{B}(0,B_a,|\cdot|)\big)\cdot\mathrm{U}\big(\mathbb{B}(0,B_b,\|\cdot\|_q)\big)\Big)^{md}\bigg],\quad\text{and}
\end{aligned}$$

$$
\begin{aligned}
Q =& \mathrm{U}\big(\mathbb{B}(w, \varepsilon_w, \|\cdot\|_q)\big) \cdot \prod_{i=1}^{L} \Big[\mathrm{U}\big(\mathbb{B}(W_{QK}^{(i)\top}, \varepsilon_{QK}^{(i)}, \|\cdot\|_{p,q})\big) \cdot \mathrm{U}\big(\mathbb{B}(W_V^{(i)\top}, \varepsilon_V^{(i)}, \|\cdot\|_{p,q})\big) \\
& \cdot \prod_{j\in[m],k\in[d]} \Big(\mathrm{U}\big(\mathbb{B}(a_{kj}^{(i)}, \varepsilon_{a,kj}^{(i)}, |\cdot|)\big) \cdot \mathrm{U}\big(\mathbb{B}(b_{kj}^{(i)}, \varepsilon_{b,kj}^{(i)}, \|\cdot\|_q)\big) \Big) \Big],
\end{aligned}
$$

For a constant $C > 0$, we define $\Delta = C/(4LnN^{3/2}dB^*)$. For $i \in [L]$, $j \in [m]$, and $k \in [d]$, we set

$$
\begin{aligned}
\varepsilon_{QK}^{(i)} &= \big(2B_V\big)^{-1} \big[B_V(1 + 4B_{QK}) + d^{\frac{1}{2}} mB_aB_b\big]^{-L+i} \Delta, \\
\varepsilon_V^{(i)} &= \big[B_V(1 + 4B_{QK}) + d^{\frac{1}{2}} mB_aB_b\big]^{-L+i} \Delta, \\
\varepsilon_{a,kj}^{(i)} &= d^{-\frac{1}{2}} \big(mB_b\big)^{-1} \big[B_V(1 + 4B_{QK}) + d^{\frac{1}{2}} mB_aB_b\big]^{-L+i} \Delta, \\
\varepsilon_{b,kj}^{(i)} &= d^{-\frac{1}{2}} \big(mB_a\big)^{-1} \big[B_V(1 + 4B_{QK}) + d^{\frac{1}{2}} mB_aB_b\big]^{-L+i} \Delta.
\end{aligned}
$$

By Proposition 8, we have

$$
\mathbb{E}_{\tilde{\theta}\sim Q}\big[\Delta(\theta, \tilde{\theta})\big] \le \frac{C}{nNdB^*}, \quad \mathbb{E}_{\tilde{\theta}\sim Q}\big[\Delta^2(\theta, \tilde{\theta})\big] \le \frac{C^2}{n^2N^2d^2B^{*2}}. \tag{J.11}
$$

Sine Q and Q_0 are product distributions, $\mathrm{KL}(Q \,\|\, Q_0)$ is the sum of the KL-divergences between each constituent distribution. For the KL-divergence between the distributions of W_{QK},

$$
\mathrm{KL}\Big(\mathrm{U}\big(\mathbb{B}(W_{QK}^{(i)\top}, \varepsilon_{QK}^{(i)}, \|\cdot\|_{p,q})\big) \,\Big\|\, \mathrm{U}\big(\mathbb{B}(0, B_{QK}, \|\cdot\|_{p,q})\big)\Big) \le 2(L - i)d^2 \log\bigg(\frac{4mdB_VB_{QK}B_aB_b}{\Delta}\bigg),
$$

for $i \in [L]$. Similar bounds for the KL divergence of the distributions of parameters $W_V^{(i)}$, $a_{kj}^{(i)}$, $b_{kj}^{(i)}$ and w for $i \in [L]$, $k \in [d]$ and $j \in [m]$ can be derived by replacing d^2 by the dimension of the parameter. Thus, we have

$$
\mathrm{KL}(Q \,\|\, Q_0) \le 2(m + 1)L^2d^2 \log\bigg(\frac{4mdB_VB_{QK}B_aB_b}{\Delta}\bigg). \tag{J.12}
$$

Substituting Eqn. (J.11) and (J.12) into Eqn. (J.10), we have that for any $\theta \in \Theta$ and any two distributions Q and Q_0, the following inequality holds with probability at least $1 - \delta$

$$
\begin{aligned}
&\mathbb{E}_\nu\Big[\mathrm{TV}\big(P^*(\cdot \,|\, \bar{S}, \bar{A}), P_\theta(\cdot \,|\, \bar{S}, \bar{A})\big)^2\Big] \\
&\quad \le \frac{1}{n}\sum_{i=1}^{n} \log \frac{P^*(\bar{S}_i' \,|\, \bar{S}_i, \bar{A}_i)}{P_\theta(\bar{S}_i' \,|\, \bar{S}_i, \bar{A}_i)} \\
&\qquad + O\bigg(\frac{C}{n} + \frac{1}{n}(m + 1)L^2d^2 \log\bigg(\frac{4NLmdB^*B_VB_{QK}B_aB_bn}{C}\bigg) + \frac{1}{n}\log\frac{1}{\delta}\bigg).
\end{aligned}
$$

Take $\theta = \hat{\theta}_{\mathrm{MLE}}$, which is the estimate derived in Eqn. (3). Since it is the maximum likelihood estimate, we have

$$
\frac{1}{n}\sum_{i=1}^{n} \log \frac{P^*(\bar{S}_i' \,|\, \bar{S}_i, \bar{A}_i)}{P_{\hat{\theta}_{\mathrm{MLE}}}(\bar{S}_i' \,|\, \bar{S}_i, \bar{A}_i)} \le 0,
$$

which proves the desired result. Therefore, this concludes the proof of Proposition 3. □

K Proof of Proposition 6

Proof of Proposition 6. We adopt the PAC-Bayes framework to prove the desired result. Define $l(\theta, \bar{S}, \bar{A}) = \mathrm{TV}(P^*(\cdot \,|\, \bar{S}, \bar{A}), P_\theta(\cdot \,|\, \bar{S}, \bar{A}))^2$. Then we have

$$
\mathrm{Var}\big(l(\theta, \bar{S}, \bar{A})\big) \le \mathbb{E}_{(\bar{S},\bar{A})\sim\nu}\big[l(\theta, \bar{S}, \bar{A})^2\big] \le \mathbb{E}_{(\bar{S},\bar{A})\sim\nu}\big[l(\theta, \bar{S}, \bar{A})\big],
$$

which implies that $l(\theta, \bar{S}, \bar{A})$ satisfies the conditions of Proposition 7 with $b = c = 1$. Thus, Proposition 7 shows that for any distributions Q and Q_0 on Θ, the following inequality holds with probability at least $1 - \delta$

$$\left| \mathbb{E}_Q \left[\mathbb{E}_{\mathcal{D}}\big[l(\theta, \bar{S}, \bar{A})\big] - \frac{1}{n} \sum_{i=1}^{n} l(\theta, \bar{S}_i, \bar{A}_i) \right] \right| \leq \lambda \mathbb{E}_{Q,\mathcal{D}}\big[l(\theta, \bar{S}, \bar{A})\big] + \frac{1}{n\lambda} \left[\mathrm{KL}(Q \,\|\, Q_0) + \log \frac{2}{\delta} \right], \tag{K.1}$$

for $0 < \lambda \leq 1/2$. Since we want to derive the generalization error bound for all $\theta \in \Theta$ uniformly, we set Q as the uniform distribution on the neighborhood of any fixed θ and Q_0 as the uniform distribution on Θ. To derive the uniform generalization bound for any $\theta \in \Theta$, we need to control the fluctuation of inequality (K.1) induced by Q.

With triangle inequality, for any $\theta, \tilde{\theta} \in \Theta$, we have

$$\begin{aligned}
&\mathrm{TV}\big(P^*(\cdot|\bar{S}, \bar{A}), P_{\tilde{\theta}}(\cdot \,|\, \bar{S}, \bar{A})\big)^2 \\
&\quad \leq \mathrm{TV}\big(P^*(\cdot \,|\, \bar{S}, \bar{A}), P_\theta(\cdot \,|\, \bar{S}, \bar{A})\big)^2 + \mathrm{TV}\big(P_{\tilde{\theta}}(\cdot \,|\, \bar{S}, \bar{A}), P_\theta(\cdot \,|\, \bar{S}, \bar{A})\big)^2 \\
&\qquad + 2\mathrm{TV}\big(P^*(\cdot \,|\, \bar{S}, \bar{A}), P_\theta(\cdot \,|\, \bar{S}, \bar{A})\big)\mathrm{TV}\big(P_{\tilde{\theta}}(\cdot \,|\, \bar{S}, \bar{A}), P_\theta(\cdot \,|\, \bar{S}, \bar{A})\big) &\text{(K.2)}\\
&\mathrm{TV}\big(P^*(\cdot \,|\, \bar{S}, \bar{A}), P_{\tilde{\theta}}(\cdot \,|\, \bar{S}, \bar{A})\big)^2 \\
&\quad \geq \mathrm{TV}\big(P^*(\cdot \,|\, \bar{S}, \bar{A}), P_\theta(\cdot \,|\, \bar{S}, \bar{A})\big)^2 + \mathrm{TV}\big(P_{\tilde{\theta}}(\cdot \,|\, \bar{S}, \bar{A}), P_\theta(\cdot \,|\, \bar{S}, \bar{A})\big)^2 \\
&\qquad - 2\mathrm{TV}\big(P^*(\cdot \,|\, \bar{S}, \bar{A}), P_\theta(\cdot \,|\, \bar{S}, \bar{A})\big)\mathrm{TV}\big(P_{\tilde{\theta}}(\cdot \,|\, \bar{S}, \bar{A}), P_\theta(\cdot \,|\, \bar{S}, \bar{A})\big). &\text{(K.3)}
\end{aligned}$$

For two parameters θ and $\tilde{\theta}$, we define the upper bound of the difference between the dynamic functions induced by them as $\Delta(\theta, \tilde{\theta}) = \max_{(\bar{S}, \bar{A}) \in \bar{\mathcal{S}} \times \bar{\mathcal{A}}} \|F_\theta(\bar{S}, \bar{A}) - F_{\tilde{\theta}}(\bar{S}, \bar{A})\|_{\mathrm{F}}$. By Pinsker's inequality, we then have

$$\begin{aligned}
\max_{(\bar{S}, \bar{A}) \in \bar{\mathcal{S}} \times \bar{\mathcal{A}}} \mathrm{TV}\big(P_{\tilde{\theta}}(\cdot \,|\, \bar{S}, \bar{A}), (P_\theta(\cdot \,|\, \bar{S}, \bar{A})\big) &\leq \max_{(\bar{S}, \bar{A}) \in \bar{\mathcal{S}} \times \bar{\mathcal{A}}} \sqrt{\mathrm{KL}\big(P_{\tilde{\theta}}(\cdot \,|\, \bar{S}, \bar{A}) \,\|\, P_\theta(\cdot \,|\, \bar{S}, \bar{A})\big)/2} \\
&= \frac{1}{2\sigma} \max_{(\bar{S}, \bar{A}) \in \bar{\mathcal{S}} \times \bar{\mathcal{A}}} \big\|F_\theta(\bar{S}, \bar{A}) - F_{\tilde{\theta}}(\bar{S}, \bar{A})\big\|_{\mathrm{F}} \\
&= \frac{\Delta(\theta, \tilde{\theta})}{2\sigma}, &\text{(K.4)}
\end{aligned}$$

where the first equality follows from the expression of the KL divergence between two Gaussian random vectors. Substituting inequalities (K.2), (K.3) and (K.4) into the left-hand side of inequality (K.1), for a fixed $\theta \in \Theta$ we have

$$\begin{aligned}
&\left| \mathbb{E}_Q \left[\mathbb{E}_{\mathcal{D}}\big[l(\tilde{\theta}, \bar{S}, \bar{A})\big] - \frac{1}{n} \sum_{i=1}^{n} l(\tilde{\theta}, \bar{S}_i, \bar{A}_i) \right] \right| \\
&\quad \geq \left| \mathbb{E}_{\mathcal{D}}\big[l(\theta, \bar{S}, \bar{A})\big] - \frac{1}{n} \sum_{i=1}^{n} l(\theta, \bar{S}_i, \bar{A}_i) \right| - \frac{5\mathbb{E}_Q\big[\Delta(\theta, \tilde{\theta})\big]}{2\sigma}. &\text{(K.5)}
\end{aligned}$$

Similarly, for the right-hand side of inequality (K.1), we have

$$\begin{aligned}
&\lambda \mathbb{E}_{Q,\mathcal{D}}\big[l(\tilde{\theta}, \bar{S}, \bar{A})\big] + \frac{1}{n\lambda} \left[\mathrm{KL}(Q \,\|\, Q_0) + \log \frac{2}{\delta} \right] \\
&\quad \leq \lambda \mathbb{E}_{\mathcal{D}}\big[l(\theta, \bar{S}, \bar{A})\big] + \frac{1}{n\lambda} \left[\mathrm{KL}(Q \,\|\, Q_0) + \log \frac{2}{\delta} \right] + \frac{3\lambda \mathbb{E}_Q\big[\Delta(\theta, \tilde{\theta})\big]}{2\sigma}. &\text{(K.6)}
\end{aligned}$$

Substituting inequalities (K.5) and (K.6) into inequality (K.1), we have that for any distributions Q and Q_0 on Θ

$$\begin{aligned}
&\left| \mathbb{E}_{\mathcal{D}}[l(\theta, \bar{S}, \bar{A})] - \frac{1}{n} \sum_{i=1}^{n} l(\theta, \bar{S}_i, \bar{A}_i) \right| \\
&\quad \leq \lambda \mathbb{E}_{\mathcal{D}}\big[l(\theta, \bar{S}, \bar{A})\big] + \frac{1}{n\lambda} \left[\mathrm{KL}(Q \,\|\, Q_0) + \log \frac{2}{\delta} \right] + (3\lambda + 5) \frac{\mathbb{E}_Q\big[\Delta(\theta, \tilde{\theta})\big]}{2\sigma} &\text{(K.7)}
\end{aligned}$$

holds with probability at least $1-\delta$.

For any fixed $\theta = [W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}]$, we set Q and Q_0 as

$$
\begin{aligned}
Q_0 = \mathrm{U}\big(\mathbb{B}(0, B_w, \|\cdot\|_q)\big) \cdot \prod_{i=1}^{L} \Big[& \mathrm{U}\big(\mathbb{B}(0, B_{QK}, \|\cdot\|_{p,q})\big) \cdot \mathrm{U}\big(\mathbb{B}(0, B_V, \|\cdot\|_{p,q})\big) \\
& \cdot \Big(\mathrm{U}\big(\mathbb{B}(0, B_a, |\cdot|)\big) \cdot \mathrm{U}\big(\mathbb{B}(0, B_b, \|\cdot\|_q)\big)\Big)^{md} \Big] \\
Q = \mathrm{U}\big(\mathbb{B}(w, \varepsilon_w, \|\cdot\|_q)\big) \cdot \prod_{i=1}^{L} \Big[& \mathrm{U}\big(\mathbb{B}(W_{QK}^{(i)\top}, \varepsilon_{QK}^{(i)}, \|\cdot\|_{p,q})\big) \cdot \mathrm{U}\big(\mathbb{B}(W_V^{(i)\top}, \varepsilon_V^{(i)}, \|\cdot\|_{p,q})\big) \\
& \cdot \prod_{j\in[m], k\in[d]} \Big(\mathrm{U}\big(\mathbb{B}(a_{kj}^{(i)}, \varepsilon_{a,kj}^{(i)}, |\cdot|)\big) \cdot \mathrm{U}\big(\mathbb{B}(b_{kj}^{(i)}, \varepsilon_{b,kj}^{(i)}, \|\cdot\|_q)\big)\Big)\Big],
\end{aligned}
$$

For a constant $C > 0$, we define $\Delta = C/(4LnN^{1/2})$. For $i \in [L]$, $j \in [m]$, and $k \in [d]$, we set

$$
\begin{aligned}
\varepsilon_{QK}^{(i)} &= \big(2c_{p,q}B_V\big)^{-1}\big[B_V(1+4c_{p,q}B_{QK}) + d^{\frac{1}{p}} m B_a B_b\big]^{-L+i}\Delta, \\
\varepsilon_{V}^{(i)} &= \big[B_V(1+4c_{p,q}B_{QK}) + d^{\frac{1}{p}} m B_a B_b\big]^{-L+i}\Delta, \\
\varepsilon_{a,kj}^{(i)} &= d^{-\frac{1}{p}}\big(mB_b\big)^{-1}\big[B_V(1+4c_{p,q}B_{QK}) + d^{\frac{1}{p}} m B_a B_b\big]^{-L+i}\Delta, \\
\varepsilon_{b,kj}^{(i)} &= d^{-\frac{1}{p}}\big(mB_a\big)^{-1}\big[B_V(1+4c_{p,q}B_{QK}) + d^{\frac{1}{p}} m B_a B_b\big]^{-L+i}\Delta.
\end{aligned}
$$

By Proposition 8, we then have

$$
\mathbb{E}_{\tilde\theta\sim Q}\big[\Delta(\theta, \tilde\theta)\big] \le \frac{C}{n}. \tag{K.8}
$$

Following the similar procedure in the proof of Proposition 3, we have

$$
\mathrm{KL}(Q \,\|\, Q_0) \le 2(m+1)L^2 d^2 \log\bigg(\frac{4mdB_V B_{QK} B_a B_b}{\Delta}\bigg). \tag{K.9}
$$

Substituting inequalities (K.8) and (K.9) into inequality (K.7), we derive that for any $\theta \in \Theta$, with probability at least $1-\delta$, the following inequality holds

$$
\begin{aligned}
&\bigg|\mathbb{E}_{\mathcal{D}}[l(\theta, \bar S, \bar A)] - \frac{1}{n}\sum_{i=1}^{n} l(\theta, \bar S_i, \bar A_i)\bigg| \\
&\quad \le \frac{1}{2}\mathbb{E}_{\mathcal{D}}\big[l(\theta, \bar S, \bar A)\big] + O\bigg(\frac{C}{n} + \frac{1}{n}(m+1)L^2 d^2 \log\bigg(\frac{4mdB_V B_{QK} B_a B_b}{\Delta}\bigg) + \frac{1}{n}\log\frac{1}{\delta}\bigg),
\end{aligned}
$$

where we take $\lambda = 1/2$. Therefore, this concludes the proof of Proposition 6. □

L Proof of Lemmas in Appendix G

L.1 Proof of Lemma 4

Proof of Lemma 4. Let $g^* = \operatorname{arginf}_{f\in\mathcal{F}_{\mathrm{tf}}} \mathcal{L}(f, f_\pi^*, \pi^*; \mathcal{D})$. Then the Bellman error of the best approximation $f_{\pi^*}^*$ can be decomposed as

$$
\begin{aligned}
\mathcal{E}(f_{\pi^*}^*, \pi^*; \mathcal{D}) &= \mathcal{L}(f_{\pi^*}^*, f_{\pi^*}^*, \pi^*; \mathcal{D}) - \mathcal{L}(g^*, f_{\pi^*}^*, \pi^*; \mathcal{D}) \\
&= \mathcal{L}(f_{\pi^*}^*, f_{\pi^*}^*, \pi^*; \mathcal{D}) - \mathcal{L}(\mathcal{T}^{\pi^*} f_{\pi^*}^*, f_{\pi^*}^*, \pi^*; \mathcal{D}) \\
&\qquad + \mathcal{L}(\mathcal{T}^{\pi^*} f_{\pi^*}^*, f_{\pi^*}^*, \pi^*; \mathcal{D}) - \mathcal{L}(g^*, f_{\pi^*}^*, \pi^*; \mathcal{D}).
\end{aligned} \tag{L.1}
$$

Note that the terms in inequality (L.1) can be bounded with their population version and the generalization error shown in Theorem 2. With probability at least $1-\delta$, we have

$$
\mathcal{L}(f_{\pi^*}^*, f_{\pi^*}^*, \pi^*; \mathcal{D}) - \mathcal{L}(\mathcal{T}^{\pi^*} f_{\pi^*}^*, f_{\pi^*}^*, \pi^*; \mathcal{D}) \le \frac{3}{2}\varepsilon_{\mathcal{F}} + \frac{e(\mathcal{F}_{\mathrm{tf}}, \pi^*, \delta, n)}{n}, \tag{L.2}
$$

$$\mathcal{L}(\mathcal{T}^{\pi^*} f^*_{\pi^*}, f^*_{\pi^*}, \pi^*; \mathcal{D}) - \mathcal{L}(g^*, f^*_{\pi^*}, \pi^*; \mathcal{D}) \le \frac{e(\mathcal{F}_{\mathrm{tf}}, \Pi, \delta, n)}{n}, \tag{L.3}$$

where inequality (L.2) follows from the definition of $f^*_{\pi^*}$, and inequality (L.3) follows from that $(g^*(\bar{S}, \bar{A}) - \mathcal{T}^{\pi^*} f^*_{\pi^*}(\bar{S}, \bar{A}))^2 \ge 0$. Substituting inequalities (L.2) and (L.3) into inequality (L.1), we have

$$\mathcal{E}(f^*_{\pi^*}, \pi^*; \mathcal{D}) \le \frac{3}{2}\varepsilon_{\mathcal{F}} + \frac{2e(\mathcal{F}_{\mathrm{tf}}, \Pi, \delta, n)}{n}.$$

This concludes the proof of Lemma 4. □

L.2 Proof of Lemma 5

Proof of Lemma 5. Let $h^*_\pi = \operatorname{arginf}_{g \in \mathcal{F}_{\mathrm{tf}}} \mathbb{E}_\nu[(g(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A}))^2]$, which is the best approximation of $\mathcal{T}^\pi f$. Then Assumption 1 implies that

$$\mathbb{E}_\nu\Big[\big(h^*_\pi(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A})\big)^2\Big] \le \varepsilon_{\mathcal{F},\mathcal{F}}. \tag{L.4}$$

For any $f \in \mathcal{F}(\pi, \varepsilon)$, the Bellman error of f with respect to the policy π can be decomposed as

$$\begin{aligned}
\mathcal{E}(f, \pi; \mathcal{D}) &= \mathcal{L}(f, f, \pi; \mathcal{D}) - \inf_{g \in \mathcal{F}_{\mathrm{tf}}} \mathcal{L}(g, f, \pi; \mathcal{D}) \\
&\ge \mathcal{L}(f, f, \pi; \mathcal{D}) - \mathcal{L}(h^*_\pi, f, \pi; \mathcal{D}) \\
&= \mathcal{L}(f, f, \pi; \mathcal{D}) - \mathcal{L}(\mathcal{T}^\pi f, f, \pi; \mathcal{D}) + \mathcal{L}(\mathcal{T}^\pi f, f, \pi; \mathcal{D}) - \mathcal{L}(h^*_\pi, f, \pi; \mathcal{D}).
\end{aligned} \tag{L.5}$$

Similar to Step 1, we bound the terms in inequality (L.5) with their population version and the generalization error bound in Theorem 2. With probability at least $1 - \delta$, we have

$$\mathcal{L}(f, f, \pi; \mathcal{D}) - \mathcal{L}(\mathcal{T}^\pi f, f, \pi; \mathcal{D}) \ge \frac{1}{2}\mathbb{E}_\nu\Big[\big(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A})\big)^2\Big] - \frac{e(\mathcal{F}_{\mathrm{tf}}, \Pi, \delta, n)}{n}, \quad \text{and} \tag{L.6}$$

$$\mathcal{L}(\mathcal{T}^\pi f, f, \pi; \mathcal{D}) - \mathcal{L}(h^*_\pi, f, \pi; \mathcal{D}) \ge -\frac{3}{2}\mathbb{E}_\nu\Big[\big(h^*_\pi(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A})\big)^2\Big] - \frac{e(\mathcal{F}_{\mathrm{tf}}, \Pi, \delta, n)}{n}. \tag{L.7}$$

Substituting inequalities (L.6) and (L.7) into inequality (L.5), we have

$$\begin{aligned}
\mathbb{E}_\nu\Big[\big(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A})\big)^2\Big] &\le 2\mathcal{E}(f, \pi; \mathcal{D}) + 4\frac{e(\mathcal{F}_{\mathrm{tf}}, \Pi, \delta, n)}{n} + 3\mathbb{E}_\nu\Big[\big(h^*_\pi(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A})\big)^2\Big] \\
&\le 2\mathcal{E}(f, \pi; \mathcal{D}) + 4\frac{e(\mathcal{F}_{\mathrm{tf}}, \Pi, \delta, n)}{n} + 3\varepsilon_{\mathcal{F},\mathcal{F}},
\end{aligned} \tag{L.8}$$

where inequality (L.8) follows from inequality (L.4). This concludes the proof of Lemma 5. □

M Proofs of Supporting Propositions

M.1 Proof of Proposition 7

To prove Proposition 7, we need the variational definition of the Kullback–Leibler divergence.

Theorem 5 (Donsker–Varadhan representation [51])**.** *Let P and Q be distributions on a common space $\mathcal{X}$. Then*

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

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

Proof of Proposition 7. Since $|f(X) - \mu(f)| \le b$ a.s., $f(X)$ is a bounded random variable. Then by [52], we have for $|\lambda| \le 1/(2b)$,

$$\mathbb{E}_X\Big[\exp\Big(\lambda\big(f(X) - \mu(f)\big)\Big)\Big] \le \exp\big(\lambda^2\sigma^2(f)\big).$$

Consequently, set $\varepsilon_n(f,\lambda) = \lambda\big[\mu(f) - \frac{1}{n}\sum_{i=1}^n f(X_i) - \lambda\sigma^2(f)\big]$, then we have

$$\mathbb{E}_{X_{1:n}}\Big[\exp\big(n\varepsilon_n(f,\lambda)\big)\Big] = \mathbb{E}_X\Big[\exp\Big(\lambda\big(\mu(f) - f(X)\big) - \lambda^2\sigma^2(f)\Big)\Big]^n \le 1$$

for all $f \in \mathcal{F}$ and $0 < \lambda \le \frac{1}{2b}$.

By Markov's inequality, we have that for any distribution P_0 on the function class $\mathcal{F}$, the random variable ε_n induced by random variables $\{X_i\}_{i=1}^n$ satisfies

$$\Pr\bigg(\mathbb{E}_{P_0}\Big[\exp\big(n\varepsilon_n(f,\lambda)\big)\Big] \ge \frac{2}{\delta}\bigg) \le \frac{\delta}{2}, \tag{M.1}$$

where the probability is taken with respect to the distribution of X_i for $i \in [n]$.

Setting $g(f) = n\varepsilon_n(f,\lambda)$ in Theorem 5, we have

$$\mathbb{E}_Q\big[n\varepsilon_n(f,\lambda)\big] \le \mathrm{KL}(Q\,\|\,P_0) + \log\mathbb{E}_{P_0}\Big[\exp\big(n\varepsilon_n(f,\lambda)\big)\Big]. \tag{M.2}$$

Combining inequalities (M.1) and (M.2), with prob at least $1-\frac{\delta}{2}$, for $0 < \lambda \le \frac{1}{2b}$, we have

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

for all Q. Similarly, setting $\varepsilon_n'(f,\lambda) = \lambda\big[\frac{1}{n}\sum_{i=1}^n f(X_i) - \mu(f) - \lambda\sigma^2(f)\big]$, we have

$$\mathbb{E}_Q\Big[\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_X[f(X)]\Big] \le \lambda\mathbb{E}_Q\big[\sigma^2(f)\big] + \frac{1}{n\lambda}\Big[\mathrm{KL}(Q\,\|\,P_0) + \log\frac{2}{\delta}\Big], \tag{M.3}$$

with probability at least $1-\frac{\delta}{2}$. The desired result can be proved using the union bound. When $\sigma^2(f) \le c\mu(f)$ for all $f \in \mathcal{F}$, the result follows from substituting this condition into inequality (M.3). Therefore, we conclude the proof of Proposition 7. $\square$

M.2 Proof of Proposition 8

Proof of Proposition 8. To prove the desired result, we first analyze the error propagation through each layer. Then we combine the error propagation of each layer to derive the error bound of the whole network.

Step 1: Bound the difference of each layer.

For $i \in [L-1]$, we can bound the difference of the output of the $(i+1)^{\mathrm{st}}$ as

$$\begin{aligned}
&\Big\|\big(G_{\mathrm{tf}}^{(i+1)}(X; W_{QK}^{1:i+1}, W_V^{1:i+1}, a^{1:i+1}, b^{1:i+1}) - G_{\mathrm{tf}}^{(i+1)}(X; \tilde{W}_{QK}^{1:i+1}, \tilde{W}_V^{1:i+1}, \tilde{a}^{1:i+1}, \tilde{b}^{1:i+1})\big)^\top\Big\|_{p,\infty} \\
&\quad= \bigg\|\Big(\mathrm{SM}\big(G_{\mathrm{tf}}^{(i)} W_{QK}^{(i+1)} G_{\mathrm{tf}}^{(i)\top}\big) G_{\mathrm{tf}}^{(i)} W_V^{(i+1)} + \mathrm{rFF}\big(G_{\mathrm{tf}}^{(i)}, a^{(i+1)}, b^{(i+1)}\big) \\
&\qquad\quad - \mathrm{SM}\big(\tilde{G}_{\mathrm{tf}}^{(i)} \tilde{W}_{QK}^{(i+1)} \tilde{G}_{\mathrm{tf}}^{(i)\top}\big) \tilde{G}_{\mathrm{tf}}^{(i)} \tilde{W}_V^{(i+1)} - \mathrm{rFF}\big(\tilde{G}_{\mathrm{tf}}^{(i)}, \tilde{a}^{(i+1)}, \tilde{b}^{(i+1)}\big)\Big)^\top\bigg\|_{p,\infty} \\
&\quad\le \Big\|\Big(\mathrm{SM}\big(G_{\mathrm{tf}}^{(i)} W_{QK}^{(i+1)} G_{\mathrm{tf}}^{(i)\top}\big) G_{\mathrm{tf}}^{(i)} W_V^{(i+1)} - \mathrm{SM}\big(\tilde{G}_{\mathrm{tf}}^{(i)} \tilde{W}_{QK}^{(i+1)} \tilde{G}_{\mathrm{tf}}^{(i)\top}\big) \tilde{G}_{\mathrm{tf}}^{(i)} \tilde{W}_V^{(i+1)}\Big)^\top\Big\|_{p,\infty} \\
&\qquad\quad + \Big\|\Big(\mathrm{rFF}\big(G_{\mathrm{tf}}^{(i)}, a^{(i+1)}, b^{(i+1)}\big) - \mathrm{rFF}\big(\tilde{G}_{\mathrm{tf}}^{(i)}, \tilde{a}^{(i+1)}, \tilde{b}^{(i+1)}\big)\Big)^\top\Big\|_{p,\infty},
\end{aligned} \tag{M.4}$$

where $G_{\mathrm{tf}}^{(i)}$ and $\tilde{G}_{\mathrm{tf}}^{(i)}$ are shorthands for $G_{\mathrm{tf}}^{(i)}(X; W_{QK}^{1:i}, W_V^{1:i}, a^{1:i}, b^{1:i})$ and $G_{\mathrm{tf}}^{(i)}(X; \tilde{W}_{QK}^{1:i}, \tilde{W}_V^{1:i}, \tilde{a}^{1:i}, \tilde{b}^{1:i})$, respectively, and inequality (M.4) follows from the triangle inequality.

Now we consider the first term in inequality (M.4). For $i \in [L-1]$, with the triangle inequality, we have

$$
\begin{aligned}
&\Big\|\Big(\mathrm{SM}\big(G_{\mathrm{tf}}^{(i)} W_{QK}^{(i+1)} G_{\mathrm{tf}}^{(i)\top}\big) G_{\mathrm{tf}}^{(i)} W_V^{(i+1)} - \mathrm{SM}\big(\tilde{G}_{\mathrm{tf}}^{(i)} \tilde{W}_{QK}^{(i+1)} \tilde{G}_{\mathrm{tf}}^{(i)\top}\big) \tilde{G}_{\mathrm{tf}}^{(i)} \tilde{W}_V^{(i+1)}\Big)^\top\Big\|_{p,\infty} \qquad \text{(M.5)}\\
&\quad \le \Big\|\Big(\mathrm{SM}\big(G_{\mathrm{tf}}^{(i)} W_{QK}^{(i+1)} G_{\mathrm{tf}}^{(i)\top}\big) G_{\mathrm{tf}}^{(i)} W_V^{(i+1)} - \mathrm{SM}\big(\tilde{G}_{\mathrm{tf}}^{(i)} W_{QK}^{(i+1)} \tilde{G}_{\mathrm{tf}}^{(i)\top}\big) \tilde{G}_{\mathrm{tf}}^{(i)} W_V^{(i+1)}\Big)^\top\Big\|_{p,\infty}\\
&\qquad + \Big\|\Big(\mathrm{SM}\big(\tilde{G}_{\mathrm{tf}}^{(i)} W_{QK}^{(i+1)} \tilde{G}_{\mathrm{tf}}^{(i)\top}\big) \tilde{G}_{\mathrm{tf}}^{(i)} W_V^{(i+1)} - \mathrm{SM}\big(\tilde{G}_{\mathrm{tf}}^{(i)} \tilde{W}_{QK}^{(i+1)} \tilde{G}_{\mathrm{tf}}^{(i)\top}\big) \tilde{G}_{\mathrm{tf}}^{(i)} \tilde{W}_V^{(i+1)}\Big)^\top\Big\|_{p,\infty}.
\end{aligned}
$$

Thus, we need the upper bounds of the two terms in the right-hand side of inequality (M.5), which are stated as following.

Proposition 12. *For any $X, \tilde{X} \in \mathbb{R}^{N\times d}$, any $W_V, W_{QK}, \tilde{W}_V, \tilde{W}_{QK} \in \mathbb{R}^{d\times d}$ and two positive conjugate numbers $p, q \in \mathbb{R}$, if $\|X^\top\|_{p,\infty}, \|\tilde{X}^\top\|_{p,\infty} \le B_X$, $\|W_{QK}^\top\|_{p,q} \le B_{QK}$, and $\|W_V^\top\|_{p,q} \le B_V$, then we have*

$$
\begin{aligned}
&\Big\|\big(\mathrm{SM}(XW_{QK}X^\top)XW_V - \mathrm{SM}(\tilde{X}W_{QK}\tilde{X}^\top)\tilde{X}W_V\big)^\top\Big\|_{p,\infty}\\
&\quad \le B_V\big(1 + 4c_{p,q}B_X^2 \cdot B_{QK}\big)\|X^\top - \tilde{X}^\top\|_{p,\infty}, \quad \textit{and}\\
&\Big\|\big(\mathrm{SM}(XW_{QK}X^\top)XW_V - \mathrm{SM}(X\tilde{W}_{QK}X^\top)X\tilde{W}_V\big)^\top\Big\|_{p,\infty}\\
&\quad \le 2c_{p,q}B_X^3 \cdot B_V \cdot \|W_{QK}^\top - \tilde{W}_{QK}^\top\|_{p,q} + B_X\big\|W_V^\top - \tilde{W}_V^\top\|_{p,q}.
\end{aligned}
$$

where $c_{p,q} = 1$ if $p \le q$, and $c_{p,q} = d^{1/q-1/p}$ otherwise.

Proof. See Appendix M.3 for a detailed proof. ☐

Thus, we have

$$
\begin{aligned}
&\Big\|\Big(\mathrm{SM}\big(G_{\mathrm{tf}}^{(i)} W_{QK}^{(i+1)} G_{\mathrm{tf}}^{(i)\top}\big) G_{\mathrm{tf}}^{(i)} W_V^{(i+1)} - \mathrm{SM}\big(\tilde{G}_{\mathrm{tf}}^{(i)} \tilde{W}_{QK}^{(i+1)} \tilde{G}_{\mathrm{tf}}^{(i)\top}\big) \tilde{G}_{\mathrm{tf}}^{(i)} \tilde{W}_V^{(i+1)}\Big)^\top\Big\|_{p,\infty}\\
&\quad \le B_V(1 + 4c_{p,q}B_{QK})\big\|G_{\mathrm{tf}}^{(i)\top} - \tilde{G}_{\mathrm{tf}}^{(i)\top}\big\|_{p,\infty} + 2c_{p,q}B_V\|W_{QK}^{(i+1)\top} - \tilde{W}_{QK}^{(i+1)\top}\|_{p,q}\\
&\qquad + \|W_V^{(i+1)\top} - \tilde{W}_V^{(i+1)\top}\|_{p,q}, \qquad \text{(M.6)}
\end{aligned}
$$

where the inequality follows from the fact that the radius of parameters are bounded and the norm of $\|\tilde{G}_{\mathrm{tf}}^{(i)\top}\|_{p,\infty}$ is bounded by 1 due to the normalization procedure.

Now we consider the second term in inequality (M.4). For $i \in [L-1]$, we have

$$
\begin{aligned}
&\Big\|\Big(\mathrm{rFF}\big(G_{\mathrm{tf}}^{(i)}, a^{(i+1)}, b^{(i+1)}\big) - \mathrm{rFF}\big(\tilde{G}_{\mathrm{tf}}^{(i)}, \tilde{a}^{(i+1)}, \tilde{b}^{(i+1)}\big)\Big)^\top\Big\|_{p,\infty}\\
&\quad \le \Big\|\Big(\mathrm{rFF}\big(G_{\mathrm{tf}}^{(i)}, a^{(i+1)}, b^{(i+1)}\big) - \mathrm{rFF}\big(\tilde{G}_{\mathrm{tf}}^{(i)}, a^{(i+1)}, b^{(i+1)}\big)\Big)^\top\Big\|_{p,\infty}\\
&\qquad + \Big\|\Big(\mathrm{rFF}\big(\tilde{G}_{\mathrm{tf}}^{(i)}, a^{(i+1)}, b^{(i+1)}\big) - \mathrm{rFF}\big(\tilde{G}_{\mathrm{tf}}^{(i)}, \tilde{a}^{(i+1)}, \tilde{b}^{(i+1)}\big)\Big)^\top\Big\|_{p,\infty} \qquad \text{(M.7)}
\end{aligned}
$$

Thus, we need to upper bound the two terms in the right-hand side of inequality (M.7). These upper bounds are stated as follows.

Proposition 13. *For any $X, \tilde{X} \in \mathbb{R}^{N\times d}$, $a, \tilde{a} \in \mathbb{R}^{dm}$, $b, \tilde{b} \in \mathbb{R}^{d\times dm}$ and two positive conjugate numbers $p, q \in \mathbb{R}$, if $\|X^\top\|_{p,\infty} \le B_X$, $|a_{kj}|, |\tilde{a}_{kj}| \le B_a$, and $\|b_{kj}\|_q, \|\tilde{b}_{kj}\|_q \le B_b$ for $k \in [d]$ and $j \in [m]$, then we have*

$$
\Big\|\big(\mathrm{rFF}(X, a, b) - \mathrm{rFF}(\tilde{X}, a, b)\big)^\top\Big\|_{p,\infty} \le d^{\frac{1}{p}} m B_a \cdot B_b \cdot \big\|X^\top - \tilde{X}^\top\big\|_{p,\infty}, \ \textit{and}
$$

$$
\begin{aligned}
&\Big\|\big(\mathrm{rFF}(X,a,b)-\mathrm{rFF}(X,\tilde{a},\tilde{b})\big)^\top\Big\|_{p,\infty}\\
&\quad\le B_b\cdot B_X\bigg[\sum_{k=1}^{d}\bigg(\sum_{j=1}^{m}|a_{kj}-\tilde{a}_{kj}|\bigg)^p\bigg]^{\frac{1}{p}}+B_a\cdot B_X\bigg[\sum_{k=1}^{d}\bigg(\sum_{j=1}^{m}\|b_{kj}-\tilde{b}_{kj}\|_q\bigg)^p\bigg]^{\frac{1}{p}}.
\end{aligned}
$$

Proof. See Appendix M.4 for a detailed proof. □

Thus, we have

$$
\begin{aligned}
&\bigg\|\Big(\mathrm{rFF}\big(G_{\mathrm{tf}}^{(i)},a^{(i+1)},b^{(i+1)}\big)-\mathrm{rFF}\big(\tilde{G}_{\mathrm{tf}}^{(i)},\tilde{a}^{(i+1)},\tilde{b}^{(i+1)}\big)\Big)^\top\bigg\|_{p,\infty}\\
&\quad\le d^{\frac{1}{p}}mB_aB_b\big\|G_{\mathrm{tf}}^{(i)\top}-\tilde{G}_{\mathrm{tf}}^{(i)\top}\big\|_{p,\infty}+B_b\bigg[\sum_{k=1}^{d}\bigg(\sum_{j=1}^{m}|a_{kj}^{(i+1)}-\tilde{a}_{kj}^{(i+1)}|\bigg)^p\bigg]^{\frac{1}{p}}\\
&\qquad+B_a\bigg[\sum_{k=1}^{d}\bigg(\sum_{j=1}^{m}\|b_{kj}^{(i+1)}-\tilde{b}_{kj}^{(i+1)}\|_q\bigg)^p\bigg]^{\frac{1}{p}}
\end{aligned}
\tag{M.8}
$$

where the inequality follows from the fact that the radius of parameters are bounded and the norm of $\|\tilde{G}_{\mathrm{tf}}^{(i)\top}\|_{p,\infty}$ is bounded by 1 due to the normalization procedure.

Substituting inequalities (M.6) and (M.8) into inequality (M.4), we have

$$
\begin{aligned}
&\Big\|\big(G_{\mathrm{tf}}^{(i+1)}(X;W_{QK}^{1:i+1},W_V^{1:i+1},a^{1:i+1},b^{1:i+1})-G_{\mathrm{tf}}^{(i+1)}(X;\tilde{W}_{QK}^{1:i+1},\tilde{W}_V^{1:i+1},\tilde{a}^{1:i+1},\tilde{b}^{1:i+1})\big)^\top\Big\|_{p,\infty}\\
&\quad\le\big[B_V(1+4c_{p,q}B_{QK})+d^{\frac{1}{p}}mB_aB_b\big]\big\|g_{tf}^{(i)\top}-\tilde{G}_{\mathrm{tf}}^{(i)\top}\big\|_{p,\infty}+2c_{p,q}B_V\|W_{QK}^{(i+1)\top}-\tilde{W}_{QK}^{(i+1)\top}\|_{p,q}\\
&\qquad+\|W_V^{(i+1)\top}-\tilde{W}_V^{(i+1)\top}\|_{p,q}+B_b\bigg[\sum_{k=1}^{d}\bigg(\sum_{j=1}^{m}|a_{kj}^{(i+1)}-\tilde{a}_{kj}^{(i+1)}|\bigg)^p\bigg]^{\frac{1}{p}}\\
&\qquad+B_a\bigg[\sum_{k=1}^{d}\bigg(\sum_{j=1}^{m}\|b_{kj}^{(i+1)}-\tilde{b}_{kj}^{(i+1)}\|_q\bigg)^p\bigg]^{\frac{1}{p}}.
\end{aligned}
\tag{M.9}
$$

Step 2: Combine the error bound of each layer in inequality (M.9)**.**

Repeating inequality (M.9) for $i\in[L-1]$, we derive

$$
\begin{aligned}
&\Big\|\big(G_{\mathrm{tf}}^{(L)}(X;W_{QK}^{1:L},W_V^{1:L},a^{1:L},b^{1:L})-G_{\mathrm{tf}}^{(L)}(X;\tilde{W}_{QK}^{1:L},\tilde{W}_V^{1:L},\tilde{a}^{1:L},\tilde{b}^{1:L})\big)^\top\Big\|_{p,\infty}\\
&\quad\le\sum_{i=1}^{L}\big[B_V(1+4c_{p,q}B_{QK})+d^{\frac{1}{p}}mB_aB_b\big]^{L-i}\bigg\{2c_{p,q}B_V\|W_{QK}^{(i)\top}-\tilde{W}_{QK}^{(i)\top}\|_{p,q}\\
&\qquad+\|W_V^{(i)\top}-\tilde{W}_V^{(i)\top}\|_{p,q}+B_b\bigg[\sum_{k=1}^{d}\bigg(\sum_{j=1}^{m}|a_{kj}^{(i)}-\tilde{a}_{kj}^{(i)}|\bigg)^p\bigg]^{\frac{1}{p}}\\
&\qquad+B_a\bigg[\sum_{k=1}^{d}\bigg(\sum_{j=1}^{m}\|b_{kj}^{(i)}-\tilde{b}_{kj}^{(i)}\|_q\bigg)^p\bigg]^{\frac{1}{p}}\bigg\}.
\end{aligned}
\tag{M.10}
$$

For the output of the neural network, we have

$$
\begin{aligned}
&\big|g_{\mathrm{tf}}(X;W_{QK}^{1:L},W_V^{1:L},a^{1:L},b^{1:L},w)-g_{\mathrm{tf}}(X;\tilde{W}_{QK}^{1:L},\tilde{W}_V^{1:L},\tilde{a}^{1:L},\tilde{b}^{1:L},\tilde{w})\big|\\
&\quad=\bigg|\Pi_{V_{\max}}\big(\frac{1}{N}\mathbb{I}_NG_{\mathrm{tf}}^{(L)}(X;W_{QK}^{1:L},W_V^{1:L},a^{1:L},b^{1:L})w\big)\\
&\qquad-\Pi_{V_{\max}}\big(\frac{1}{N}\mathbb{I}_NG_{\mathrm{tf}}^{(L)}(X;\tilde{W}_{QK}^{1:L},\tilde{W}_V^{1:L},\tilde{a}^{1:L},\tilde{b}^{1:L})\tilde{w}\big)\bigg|
\end{aligned}
$$

$$\leq \left| \frac{1}{N} \mathbb{I}_N G_{\mathrm{tf}}^{(L)}(X; W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}) w - \frac{1}{N} \mathbb{I}_N G_{\mathrm{tf}}^{(L)}(X; \tilde{W}_{QK}^{1:L}, \tilde{W}_V^{1:L}, \tilde{a}^{1:L}, \tilde{b}^{1:L}) \tilde{w} \right|,$$

where the inequality follows from the contraction property of the normalization function. It can be further upper bounded as

$$\begin{aligned}
&\left| g_{\mathrm{tf}}(X; W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}, w) - g_{\mathrm{tf}}(X; \tilde{W}_{QK}^{1:L}, \tilde{W}_V^{1:L}, \tilde{a}^{1:L}, \tilde{b}^{1:L}, \tilde{w}) \right| \\
&\quad \leq \left\| G_{\mathrm{tf}}^{(L)}(X; W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}) w - G_{\mathrm{tf}}^{(L)}(X; \tilde{W}_{QK}^{1:L}, \tilde{W}_V^{1:L}, \tilde{a}^{1:L}, \tilde{b}^{1:L}) \tilde{w} \right\|_\infty \\
&\quad \leq \left\| G_{\mathrm{tf}}^{(L)\top} - \tilde{G}_{\mathrm{tf}}^{(L)\top} \right\|_{p,\infty} \cdot \|w\|_q + \left\| \tilde{G}_{\mathrm{tf}}^{(L)\top} \right\|_{p,\infty} \cdot \|w - \tilde{w}\|_q \\
&\quad \leq B_w \left\| G_{\mathrm{tf}}^{(L)\top} - \tilde{G}_{\mathrm{tf}}^{(L)\top} \right\|_{p,\infty} + \|w - \tilde{w}\|_q,
\end{aligned} \tag{M.11}$$

where first inequality follows from Hölder's inequality, and the second inequality follows from Lemma 17 with $u = p$, $v = q$ and $p = \infty$.

Combining inequalities (M.10) and (M.11), we have

$$\begin{aligned}
&\left| g_{\mathrm{tf}}(X; W_{QK}^{1:L}, W_V^{1:L}, a^{1:L}, b^{1:L}, w) - g_{\mathrm{tf}}(X; \tilde{W}_{QK}^{1:L}, \tilde{W}_V^{1:L}, \tilde{a}^{1:L}, \tilde{b}^{1:L}, \tilde{w}) \right| \\
&\leq \|w - \tilde{w}\|_q + \sum_{i=1}^{L} B_w \left[B_V (1 + 4 c_{p,q} B_{QK}) + d^{\frac{1}{p}} m B_a B_b \right]^{L-i} \Big\{ 2 c_{p,q} B_V \| W_{QK}^{(i)\top} - \tilde{W}_{QK}^{(i)\top} \|_{p,q} \\
&\quad + \| W_V^{(i)\top} - \tilde{W}_V^{(i)\top} \|_{p,q} + B_b \left[\sum_{k=1}^{d} \left(\sum_{j=1}^{m} |a_{kj}^{(i)} - \tilde{a}_{kj}^{(i)}| \right)^p \right]^{\frac{1}{p}} + B_a \left[\sum_{k=1}^{d} \left(\sum_{j=1}^{m} \| b_{kj}^{(i)} - \tilde{b}_{kj}^{(i)} \|_q \right)^p \right]^{\frac{1}{p}} \Big\}.
\end{aligned}$$

This concludes the proof. □

M.3 Proof of Proposition 12

Proof of Proposition 12. Let $\tau \in [N]$ and $x_\tau^\top$ be the τ^{th} row of X. For the first inequality, we have

$$\begin{aligned}
&\left\| \left(\mathrm{SM}(X W_{QK} X^\top) X W_V - \mathrm{SM}(\tilde{X} W_{QK} \tilde{X}^\top) \tilde{X} W_V \right)^\top \right\|_{p,\infty} \\
&\quad = \max_{\tau \in [N]} \| \mathrm{SM}(x_\tau^\top W_{QK} X^\top) X W_V - \mathrm{SM}(\tilde{x}_\tau^\top W_{QK} \tilde{X}^\top) \tilde{X} W_V \|_p \\
&\quad \leq \max_{\tau \in [N]} \| \mathrm{SM}(x_\tau^\top W_{QK} X^\top) X W_V - \mathrm{SM}(x_\tau^\top W_{QK} X^\top) \tilde{X} W_V \|_p \\
&\qquad + \| \mathrm{SM}(x_\tau^\top W_{QK} X^\top) \tilde{X} W_V - \mathrm{SM}(\tilde{x}_\tau^\top W_{QK} \tilde{X}^\top) \tilde{X} W_V \|_p \\
&\quad = \max_{\tau \in [N]} \left\| W_V^\top (X^\top - \tilde{X}^\top) \left(\mathrm{SM}(x_\tau^\top W_{QK} X^\top) \right)^\top \right\|_p \\
&\qquad + \left\| W_V^\top \tilde{X}^\top \left(\mathrm{SM}(x_\tau^\top W_{QK} X^\top) - \mathrm{SM}(\tilde{x}_\tau^\top W_{QK} \tilde{X}^\top) \right)^\top \right\|_p,
\end{aligned}$$

where the inequality follows from the triangle inequality. We further upper bounded it as

$$\begin{aligned}
&\left\| \left(\mathrm{SM}(X W_{QK} X^\top) X W_V - \mathrm{SM}(\tilde{X} W_{QK} \tilde{X}^\top) \tilde{X} W_V \right)^\top \right\|_{p,\infty} \\
&\quad \leq \max_{\tau \in [N]} \| W_V^\top (\tilde{X}^\top - X^\top) \|_{p,\infty} + \| W_V^\top \tilde{X}^\top \|_{p,\infty} \cdot \| \mathrm{SM}(x_\tau^\top W_{QK} X^\top) - \mathrm{SM}(\tilde{x}_\tau^\top W_{QK} \tilde{X}^\top) \|_1 \\
&\quad \leq \max_{\tau \in [N]} 2 \| W_V^\top \|_{p,q} \cdot \| \tilde{X}^\top \|_{p,\infty} \cdot \| x_\tau^\top W_{QK} X^\top - \tilde{x}_\tau^\top W_{QK} \tilde{X}^\top \|_\infty \\
&\qquad + \| W_V^\top \|_{p,q} \cdot \| (\tilde{X}^\top - X^\top) \|_{p,\infty},
\end{aligned} \tag{M.12}$$

where the first inequality follows from Lemma 17 with $u = \infty$ and $v = 1$, and the last inequality follows from Lemma 18 and Lemma 19. Now we consider the second term of inequality (M.12), and we have

$$\begin{aligned}
&\| x_\tau^\top W_{QK} X^\top - \tilde{x}_\tau^\top W_{QK} \tilde{X}^\top \|_\infty \\
&\quad \leq \| x_\tau^\top W_{QK} X^\top - x_\tau^\top W_{QK} \tilde{X}^\top \|_\infty + \| x_\tau^\top W_{QK} \tilde{X}^\top - \tilde{x}_\tau^\top W_{QK} \tilde{X}^\top \|_\infty
\end{aligned}$$

$$= \left\|(X - \tilde{X}) W_{QK}^\top x_\tau\right\|_\infty + \left\|\tilde{X}(W_{QK}^\top x_\tau - W_{QK}^\top \tilde{x}_\tau)\right\|_\infty$$
$$\le \|X^\top - \tilde{X}^\top\|_{p,\infty} \cdot \|W_{QK}^\top x_\tau\|_q + \|\tilde{X}^\top\|_{p,\infty} \cdot \|W_{QK}^\top(x_\tau - \tilde{x}_\tau)\|_q,$$

where the last inequality follows from Lemma 17 with $u = p$, $v = q$ and $p = \infty$. We then bound the ℓ_q norm with the ℓ_p norm as

$$\begin{aligned}
&\|x_\tau^\top W_{QK} X^\top - \tilde{x}_\tau^\top W_{QK} \tilde{X}^\top\|_\infty \qquad \text{(M.13)}\\
&\quad\le c_{p,q}\Big[\|X^\top - \tilde{X}^\top\|_{p,\infty} \cdot \|W_{QK}^\top x_\tau\|_p + \|\tilde{X}^\top\|_{p,\infty} \cdot \|W_{QK}^\top(x_\tau - \tilde{x}_\tau)\|_p\Big]\\
&\quad\le c_{p,q}\Big[\|X^\top - \tilde{X}^\top\|_{p,\infty} \cdot \|W_{QK}^\top\|_{p,q} \cdot \|x_\tau\|_p + \|\tilde{X}^\top\|_{p,\infty} \cdot \|W_{QK}^\top\|_{p,q} \cdot \|x_\tau - \tilde{x}_\tau\|_p\Big]\\
&\quad\le c_{p,q}\Big[\|X^\top - \tilde{X}^\top\|_{p,\infty} \cdot \|W_{QK}^\top\|_{p,q} \cdot \|X^\top\|_{p,\infty} + \|\tilde{X}^\top\|_{p,\infty} \cdot \|W_{QK}^\top\|_{p,q} \cdot \|X^\top - \tilde{X}^\top\|_{p,\infty}\Big],
\end{aligned}$$

where $c_{p,q} = 1$ if $p \le q$, and $c_{p,q} = d^{1/q-1/p}$ otherwise, the first inequality follows from Lemma 16, and the second inequality follows from Lemma 17 with $u = q$ and $v = p$.

Substituting inequality (M.13) into inequality (M.12), we obtain

$$\begin{aligned}
&\|\mathrm{SM}(x_\tau^\top W_{QK} X^\top) X W_V - \mathrm{SM}(\tilde{x}_\tau^\top W_{QK} \tilde{X}^\top)\tilde{X} W_V\|_p\\
&\quad\le \|W_V^\top\|_{p,q}\Big(1 + 2c_{p,q}\|\tilde{X}^\top\|_{p,\infty} \cdot \|W_{QK}^\top\|_{p,q}\big(\|\tilde{X}^\top\|_{p,\infty} + \|X^\top\|_{p,\infty}\big)\Big)\|X^\top - \tilde{X}^\top\|_{p,\infty}
\end{aligned}$$

as desired.

For the second inequality, we have

$$\begin{aligned}
&\Big\|\big(\mathrm{SM}(X W_{QK} X^\top) X W_V - \mathrm{SM}(X \tilde{W}_{QK} X^\top) X \tilde{W}_V\big)^\top\Big\|_{p,\infty}\\
&\quad= \max_{\tau \in [N]} \|\mathrm{SM}(x_\tau^\top W_{QK} X^\top) X W_V - \mathrm{SM}(x_\tau^\top \tilde{W}_{QK} X^\top) X \tilde{W}_V\|_p\\
&\quad\le \max_{\tau \in [N]} \|\mathrm{SM}(x_\tau^\top W_{QK} X^\top) X W_V - \mathrm{SM}(x_\tau^\top \tilde{W}_{QK} X^\top) X W_V\|_p\\
&\qquad\quad + \|\mathrm{SM}(x_\tau^\top \tilde{W}_{QK} X^\top) X W_V - \mathrm{SM}(x_\tau^\top \tilde{W}_{QK} X^\top) X \tilde{W}_V\|_p\\
&\quad= \max_{\tau \in [N]} \Big\|W_V^\top X^\top\big(\mathrm{SM}(x_\tau^\top W_{QK} X^\top) - \mathrm{SM}(x_\tau^\top \tilde{W}_{QK} X^\top)\big)^\top\Big\|_p\\
&\qquad\quad + \big\|(W_V^\top X^\top - \tilde{W}_V^\top X^\top)\mathrm{SM}(x_\tau^\top \tilde{W}_{QK} X^\top)^\top\big\|_p,
\end{aligned}$$

where the inequality follows from the triangle inequality. It can be further upper bounded as

$$\begin{aligned}
&\Big\|\big(\mathrm{SM}(X W_{QK} X^\top) X W_V - \mathrm{SM}(X \tilde{W}_{QK} X^\top) X \tilde{W}_V\big)^\top\Big\|_{p,\infty} \qquad \text{(M.14)}\\
&\quad\le \max_{\tau \in [N]} \|W_V^\top X^\top\|_{p,\infty} \cdot \|\mathrm{SM}(x_\tau^\top W_{QK} X^\top) - \mathrm{SM}(x_\tau^\top \tilde{W}_{QK} X^\top)\|_1\\
&\qquad\quad + \big\|(W_V^\top - \tilde{W}_V^\top) X^\top\big\|_{p,\infty} \cdot \|\mathrm{SM}(x_\tau^\top \tilde{W}_{QK} X^\top)\|_1\\
&\quad\le \max_{\tau \in [N]} 2\|W_V^\top X^\top\|_{p,\infty} \cdot \|x_\tau^\top W_{QK} X^\top - x_\tau^\top \tilde{W}_{QK} X^\top\|_\infty + \big\|(W_V^\top - \tilde{W}_V^\top) X^\top\big\|_{p,\infty}\\
&\quad\le \max_{\tau \in [N]} 2\|W_V^\top\|_{p,q} \cdot \|X^\top\|_{p,\infty} \cdot \|x_\tau^\top W_{QK} X^\top - x_\tau^\top \tilde{W}_{QK} X^\top\|_\infty + \big\|W_V^\top - \tilde{W}_V^\top\|_{p,q} \cdot \|X^\top\big\|_{p,\infty},
\end{aligned}$$

where the first inequality follows from Lemma 17 with $u = \infty$ and $v = 1$, the second inequality follows from Lemma 19, and the last inequality follows from Lemma 18. Now we consider the first term of inequality (M.14) and have

$$\begin{aligned}
&\|x_\tau^\top W_{QK} X^\top - x_\tau^\top \tilde{W}_{QK} X^\top\|_\infty\\
&\quad\le \|X^\top\|_{p,\infty} \cdot \|x_\tau^\top W_{QK} - x_\tau^\top \tilde{W}_{QK}\|_q\\
&\quad\le c_{p,q}\|X^\top\|_{p,\infty} \cdot \|x_\tau^\top W_{QK} - x_\tau^\top \tilde{W}_{QK}\|_p\\
&\quad\le c_{p,q}\|X^\top\|_{p,\infty} \cdot \|W_{QK}^\top - \tilde{W}_{QK}^\top\|_{p,q} \cdot \|x_\tau\|_p\\
&\quad\le c_{p,q}\|X^\top\|_{p,\infty}^2 \cdot \|W_{QK}^\top - \tilde{W}_{QK}^\top\|_{p,q}, \qquad \text{(M.15)}
\end{aligned}$$

where $c_{p,q} = 1$ if $p \le q$, and $c_{p,q} = d^{1/q-1/p}$ otherwise, the first and third inequalities follows from Lemma 17, and the second inequality follows from Lemma 16.

Combining Eqn. (M.14) and (M.15), we have

$$\begin{aligned}
&\|\mathrm{SM}(x_\tau^\top W_{QK} X^\top) X W_V - \mathrm{SM}(x_\tau^\top \tilde{W}_{QK} X^\top) X \tilde{W}_V\|_p \\
&\quad \le 2c_{p,q} \|X^\top\|_{p,\infty}^3 \cdot \|W_V^\top\|_{p,q} \cdot \|W_{QK}^\top - \tilde{W}_{QK}^\top\|_{p,q} + \left\|W_V^\top - \tilde{W}_V^\top\|_{p,q} \cdot \|X^\top\right\|_{p,\infty}.
\end{aligned}$$

This concludes the proof. $\square$

M.4 Proof of Proposition 13

Proof of Proposition 13. Let $\tau \in [N]$ and $x_\tau^\top$ be the τ^{th} row of X. For the first inequality, we have

$$\begin{aligned}
&\left\|\left(\mathrm{rFF}(X, a, b) - \mathrm{rFF}(\tilde{X}, a, b)\right)^\top\right\|_{p,\infty} \\
&\quad = \max_{\tau \in [N]} \|\mathrm{rFF}(x_\tau, a, b) - \mathrm{rFF}(\tilde{x}_\tau, a, b)\|_p \\
&\quad = \max_{\tau \in [N]} \left[\sum_{k=1}^d \left|\sum_{j=1}^m a_{kj} \left[\mathrm{ReLU}(b_{kj}^\top x_\tau) - \mathrm{ReLU}(b_{kj}^\top \tilde{x}_\tau)\right]\right|^p\right]^{\frac{1}{p}},
\end{aligned}$$

which follows from the definition of the rFF network. It can be upper bounded as

$$\begin{aligned}
&\left\|\left(\mathrm{rFF}(X, a, b) - \mathrm{rFF}(\tilde{X}, a, b)\right)^\top\right\|_{p,\infty} \\
&\quad \le \max_{\tau \in [N]} \left[\sum_{k=1}^d \left(\sum_{j=1}^m |a_{kj}| \cdot \left|b_{kj}^\top x_\tau - b_{kj}^\top \tilde{x}_\tau\right|\right)^p\right]^{\frac{1}{p}} \\
&\quad \le \max_{\tau \in [N]} \left[\sum_{k=1}^d \left(\sum_{j=1}^m |a_{kj}| \cdot \|b_{kj}\|_q \cdot \|x_\tau - \tilde{x}_\tau\|_p\right)^p\right]^{\frac{1}{p}} \\
&\quad \le \left[\sum_{k=1}^d \left(\sum_{j=1}^m |a_{kj}| \cdot \|b_{kj}\|_q\right)^p\right]^{\frac{1}{p}} \left\|X^\top - \tilde{X}^\top\right\|_{p,\infty},
\end{aligned}$$

where the first inequality follows from the fact that $\mathrm{ReLU}(\cdot)$ is 1-Lipschitz, the second inequality follows from Hölder's inequality, and the last inequality follows from the definition of $\ell_{p,\infty}$ norm.

For the second inequality, we have

$$\begin{aligned}
&\left\|\left(\mathrm{rFF}(X, a, b) - \mathrm{rFF}(X, \tilde{a}, \tilde{b})\right)^\top\right\|_{p,\infty} \\
&\quad = \max_{\tau \in [N]} \|\mathrm{rFF}(x_\tau, a, b) - \mathrm{rFF}(x_\tau, \tilde{a}, \tilde{b})\|_p \\
&\quad = \max_{\tau \in [N]} \left[\sum_{k=1}^d \left|\sum_{j=1}^m a_{kj} \mathrm{ReLU}(b_{kj}^\top x_\tau) - \tilde{a}_{kj} \mathrm{ReLU}(\tilde{b}_{kj}^\top x_\tau)\right|^p\right]^{\frac{1}{p}} \\
&\quad \le \max_{\tau \in [N]} \left[\sum_{k=1}^d \left|\sum_{j=1}^m a_{kj} \mathrm{ReLU}(b_{kj}^\top x_\tau) - \tilde{a}_{kj} \mathrm{ReLU}(b_{kj}^\top x_\tau)\right|^p\right]^{\frac{1}{p}} \\
&\qquad + \left[\sum_{k=1}^d \left|\sum_{j=1}^m \tilde{a}_{kj} \mathrm{ReLU}(b_{kj}^\top x_\tau) - \tilde{a}_{kj} \mathrm{ReLU}(\tilde{b}_{kj}^\top x_\tau)\right|^p\right]^{\frac{1}{p}},
\end{aligned}$$

where the inequality follows from triangle inequality. Using the Lipschitz property of the ReLU function, it can be upper bounded as

$$\begin{aligned}
&\left\|\left(\mathrm{rFF}(X, a, b) - \mathrm{rFF}(X, \tilde{a}, \tilde{b})\right)^\top\right\|_{p,\infty} \\
&\quad \le \max_{\tau \in [N]} \left[\sum_{k=1}^d \left(\sum_{j=1}^m |a_{kj} - \tilde{a}_{kj}| \cdot |b_{kj}^\top x_\tau|\right)^p\right]^{\frac{1}{p}} + \left[\sum_{k=1}^d \left(\sum_{j=1}^m |\tilde{a}_{kj}| \cdot |b_{kj}^\top x_\tau - \tilde{b}_{kj}^\top x_\tau|\right)^p\right]^{\frac{1}{p}}
\end{aligned}$$

$$
\begin{aligned}
&\le \max_{\tau\in[N]} \Bigg[\sum_{k=1}^{d}\Big(\sum_{j=1}^{m}|a_{kj}-\tilde a_{kj}|\cdot\|b_{kj}\|_q\cdot\|x_\tau\|_p\Big)^p\Bigg]^{\frac{1}{p}} \\
&\quad + \Bigg[\sum_{k=1}^{d}\Big(\sum_{j=1}^{m}|\tilde a_{kj}|\cdot\|b_{kj}-\tilde b_{kj}\|_q\cdot\|x_\tau\|_p\Big)^p\Bigg]^{\frac{1}{p}} \\
&\le \Bigg[\sum_{k=1}^{d}\Big(\sum_{j=1}^{m}|a_{kj}-\tilde a_{kj}|\cdot\|b_{kj}\|_q\Big)^p\Bigg]^{\frac{1}{p}}\big\|X^\top\big\|_{p,\infty} \\
&\quad + \Bigg[\sum_{k=1}^{d}\Big(\sum_{j=1}^{m}|\tilde a_{kj}|\cdot\|b_{kj}-\tilde b_{kj}\|_q\Big)^p\Bigg]^{\frac{1}{p}}\big\|X^\top\big\|_{p,\infty},
\end{aligned}
$$

where the first inequality follows from the fact that $\mathrm{ReLU}(\cdot)$ is 1-Lipschitz, the second inequality follows from Hölder's inequality, and the last inequality follows from the definition of $\ell_{p,\infty}$ norm. This concludes the proof. □

M.5 Proof of Proposition 11

Proof of Proposition 11. With triangle inequality, we have

$$
\begin{aligned}
&\Big\|\Big(\mathrm{SM}\big(XW_{QK}X^\top\big)XW_V+\mathrm{rFF}\big(X,a,b\big)\Big)^\top\Big\|_{p,\infty} \\
&\le \Big\|\Big(\mathrm{SM}\big(XW_{QK}X^\top\big)XW_V\Big)^\top\Big\|_{p,\infty}+\Big\|\Big(\mathrm{rFF}\big(X,a,b\big)\Big)^\top\Big\|_{p,\infty}.
\end{aligned}
\tag{M.16}
$$

Let $\tau\in[N]$ and $x_\tau^\top$ be the τ^{th} row of X. Then the first term in the right-hand side of Eqn. (M.16) is

$$
\begin{aligned}
\Big\|\Big(\mathrm{SM}\big(XW_{QK}X^\top\big)XW_V\Big)^\top\Big\|_{p,\infty} &= \max_{\tau\in[N]}\Big\|\Big(\mathrm{SM}\big(x_\tau^\top W_{QK}X^\top\big)XW_V\Big)^\top\Big\|_p \\
&\le \max_{\tau\in[N]}\|W_V^\top X^\top\|_{p,\infty}\cdot\|\mathrm{SM}\big(x_\tau^\top W_{QK}X^\top\big)\|_1 \\
&\le \|W_V^\top\|_{p,q}\cdot\|X^\top\|_{p,\infty},
\end{aligned}
\tag{M.17}
$$

where the first inequality follows from Lemma 17 with $u=\infty$ and $v=1$, and the last inequality follows from Lemma 18. The second term in the right-hand side of inequality (M.16) is

$$
\begin{aligned}
\Big\|\Big(\mathrm{rFF}\big(X,a,b\big)\Big)^\top\Big\|_{p,\infty} &= \max_{\tau\in[N]}\Big\|\Big(\mathrm{rFF}\big(x_\tau,a,b\big)\Big)^\top\Big\|_p \\
&= \max_{\tau\in[N]}\Big[\sum_{k=1}^{d}\Big(\sum_{j=1}^{m}a_{kj}\mathrm{ReLU}(b_{kj}^\top x_\tau)\Big)^p\Big]^{1/p} \\
&\le \max_{\tau\in[N]}\Big[\sum_{k=1}^{d}\Big(\sum_{j=1}^{m}|a_{kj}|\cdot\|b_{kj}\|_q\cdot\|x_\tau\|_p)\Big)^p\Big]^{1/p} \\
&= \Big[\sum_{k=1}^{d}\Big(\sum_{j=1}^{m}|a_{kj}|\cdot\|b_{kj}\|_q\cdot\|X^\top\|_{p,\infty})\Big)^p\Big]^{1/p},
\end{aligned}
\tag{M.18}
$$

where the inequality follows from Hölder's inequality and that $\mathrm{ReLU}(\cdot)$ is 1-Lipchitz. Combining inequalities (M.17) and (M.18), we prove the desired result. □

M.6 Technical Lemmas

Lemma 14 (Lemma 1 in [15])**.** *For any policy $\pi\in\Pi$ and any function $f:\bar{\mathcal{S}}\times\bar{\mathcal{A}}\to\mathbb{R}$, we have*

$$
f(\bar S_0,\pi)-V_{P^*}^{\pi}(\bar S_0)=\frac{\mathbb{E}_{d_{P^*}^{\pi}}\big[f(\bar S,\bar A)-r(\bar S,\bar A)-f(\bar S,\pi)\big]}{1-\gamma}.
\tag{M.19}
$$

Lemma 15 (Lemma 10 in [42])**.** *For any two transition kernels P and P' and any policy $\pi \in \Pi$, we have*

$$\left|V_P^\pi(\bar{S}_0) - V_{P'}^\pi(\bar{S}_0)\right| \le \frac{1}{1-\gamma}\left|\mathbb{E}_{(\bar{S},\bar{A})\sim d_P^\pi}\left[\mathbb{E}_{\bar{S}'\sim P(\cdot\,|\,\bar{S},\bar{A})}V_{P'}^\pi(\bar{S}') - \mathbb{E}_{\bar{S}'\sim P'(\cdot\,|\,\bar{S},\bar{A})}V_{P'}^\pi(\bar{S}')\right]\right|$$
$$\le \frac{V_{\max}}{(1-\gamma)^2}\mathbb{E}_{(\bar{S},\bar{A})\sim d_P^\pi}\Big[\mathrm{TV}\big(\check{P}(\cdot\,|\,\bar{S},\bar{A}), P'(\cdot\,|\,\bar{S},\bar{A})\big)\Big].$$

Lemma 16. *For any $x \in \mathbb{R}^d$ and $0 < p < q$, $\|x\|_q \le \|x\|_p \le d^{1/p-1/q}\|x\|_q$.*

Proof of Lemma 16. $\|x\|_q \le \|x\|_p$ simply follows from Hölder's inequality. For the right inequality, when $q < \infty$, we have

$$\|x\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p} \le \left[\left(\sum_{i=1}^d \big(|x_i|^p\big)^{q/p}\right)^{p/q}\left(\sum_{i=1}^d 1^{q/(q-p)}\right)^{1-p/q}\right]^{1/p} = d^{1/p-1/q}\|x\|_q,$$

where the inequality follows from Hölder's inequality. When $q = \infty$, $\|x\|_p \le d^{1/p}\|x\|_\infty$. □

Lemma 17. *Given any two conjugate numbers $u, v \in [1,\infty]$, i.e., $\frac{1}{u}+\frac{1}{v}=1$, and $1 \le p \le \infty$, for any $A \in \mathbb{R}^{r\times c}$ and $x \in \mathbb{R}^c$, we have*

$$\|Ax\|_p \le \|A\|_{p,u}\|x\|_v \quad \textit{and} \quad \|Ax\|_p \le \|A^\top\|_{u,p}\|x\|_v$$

Proof of Lemma 17. To prove the first inequality, we write $A = [a_1 \dots a_c]$, where $a_i \in \mathbb{R}^r$ for $i \in [c]$. Then we have

$$\|Ax\|_p = \left\|\sum_{i=1}^c a_i x_i\right\|_p \overset{(a)}{\le} \sum_{i=1}^c |x_i|\|a_i\|_p \overset{(b)}{\le} \|A\|_{p,u}\|x\|_v,$$

where inequality (a) comes from the triangle inequality, and inequality (b) comes from Hölder's inequality.

To prove the second inequality, we write $A = [a_1^\top \dots a_r^\top]^\top$, where $a_i \in \mathbb{R}^c$ for $i \in [r]$. Then we have

$$\|Ax\|_p^p = \sum_{i=1}^r |a_i^\top x|^p \overset{(c)}{\le} \sum_{i=1}^r \|a_i\|_u^p\|x\|_v^p = \|A\|_{p,u}^p\|x\|_v^p,$$

for $1 \le p < \infty$, where inequality (c) follows from Hölder's inequality. When $p = \infty$, we have

$$\|Ax\|_\infty = \max_{i\in[r]}|a_i^\top x| \le \max_{i\in[r]}\|a_i\|_u\|x\|_v = \|A\|_{\infty,u}\|x\|_v.$$

□

Lemma 18. *Given any two conjugate numbers $p, q \in [1,\infty]$, i.e., $\frac{1}{p}+\frac{1}{q}=1$, for any $A \in \mathbb{R}^{r\times c}$ and $B \in \mathbb{R}^{c\times d}$, we have*

$$\|AB\|_{p,\infty} \le \|A\|_{p,q}\|B\|_{p,\infty}.$$

Proof of Lemma 18. To prove the result, we write $B = [b_1,\dots,b_d]$, where $b_i \in \mathbb{R}^c$ for $i \in [d]$.

$$\|AB\|_{p,\infty} = \max_{i\in[d]}\|Ab_i\|_p \le \max_{i\in[d]}\|A\|_{p,q}\|b_i\|_p = \|A\|_{p,q}\|B\|_{p,\infty},$$

where the inequality follows from Lemma 17. □

Lemma 19. *For any $x, y \in \mathbb{R}^d$, we have*

$$\|\mathrm{SM}(x) - \mathrm{SM}(y)\|_1 \le 2\|x-y\|_\infty.$$

Proof of Lemma 19. The Jacobian matrix of the softmax function is

$$\frac{\mathrm{dSM}(x)}{\mathrm{d}x} = \mathrm{diag}(\mathrm{SM(x)}) - \mathrm{SM(x)}\mathrm{SM(x)}^\top.$$

The $\ell_{1,1}$ norm of the Jacobian matrix can be bounded as

$$\begin{aligned}
\Big\|\frac{\mathrm{dSM}(x)}{\mathrm{d}x}\Big\|_{1,1} &= \sum_{i=1}^d\sum_{j=1}^d \Big|\big[\mathrm{SM}(x)\big]_i\Big(\mathbb{I}_{i=j} - \big[\mathrm{SM}(x)\big]_j\Big)\Big| \\
&= 2\sum_{i=1}^d \big[\mathrm{SM}(x)\big]_i\Big(1 - \big[\mathrm{SM}(x)\big]_i\Big) \\
&\leq 2. \qquad (\mathrm{M.20})
\end{aligned}$$

Then the ℓ_1-norm of the difference between $\mathrm{SM}(x)$ and $\mathrm{SM}(y)$ can be bounded as

$$\begin{aligned}
\|\mathrm{SM}(x) - \mathrm{SM}(y)\|_1 &= \Big\|\int_0^1 \frac{\mathrm{dSM}(z)}{\mathrm{d}z}|_{z=tx+(1-t)y}(y-x)\mathrm{d}t\Big\|_1 \\
&\leq \int_0^1 \Big\|\frac{\mathrm{dSM}(z)}{\mathrm{d}z}|_{z=tx+(1-t)y}(y-x)\Big\|_1 \mathrm{d}t \\
&\leq \int_0^1 \Big\|\frac{\mathrm{dSM}(z)}{\mathrm{d}z}|_{z=tx+(1-t)y}\Big\|_{1,1} \|y-x\|_\infty \mathrm{d}t \\
&\leq \int_0^1 2\|y-x\|_\infty \mathrm{d}t \\
&= 2\|y-x\|_\infty,
\end{aligned}$$

where the first inequality follows from triangle inequality, the second inequality follows from Lemma 17 by setting $p=1$, $u=1$ and $v=\infty$, and the last inequality follows from inequality (M.20). This concludes the proof. ∎

N Some Extensions

N.1 Extension to Multi-Head Attention

Our results in Theorem 2 can be extended to the neural network with multi-head attention, which is defined as

$$\begin{aligned}
f(X, W_{QK}, W_V) &= \mathrm{SM}\big(XW_{QK}X^\top\big)XW_V, \\
\mathrm{MHA}(X, W_{QK}^{1:h}, W_V^{1:h}, W_O^{1:h}) &= \sum_{i=1}^h f(X, W_{QK,i}, W_{V,i})W_{O,i},
\end{aligned}$$

where $W_{QK,i} \in \mathbb{R}^{d\times d}, W_{V,i} \in \mathbb{R}^{d\times\frac{d}{h}}, W_{O,i} \in \mathbb{R}^{\frac{d}{h}\times d}$ for $i \in [h]$. Note that we only need to reprove the results in Propositions 12 and 11 for the multi-head attention.

Proposition 20. *For any $X, \tilde{X} \in \mathbb{R}^{N\times d}$, and any $W_{QK,i} \in \mathbb{R}^{d\times d}, W_{V,i} \in \mathbb{R}^{d\times\frac{d}{h}}, W_{O,i} \in \mathbb{R}^{\frac{d}{h}\times d}$ for $i \in [h]$ and two positive conjugate numbers $p, q \in \mathbb{R}$, if $\|X^\top\|_{p,\infty}, \|\tilde{X}^\top\|_{p,\infty} \leq B_X$, $\|W_{QK,i}^\top\|_{p,q} \leq B_{QK}$, $\|W_{V,i}^\top\|_{p,q} \leq B_V$, and $\|W_{O,i}^\top\|_{p,q} \leq B_O$ for $i \in [h]$, then we have*

$$\begin{aligned}
&\Big\|\big(\mathrm{MHA}(X, W_{QK}^{1:h}, W_V^{1:h}, W_O^{1:h}) - \mathrm{MHA}(\tilde{X}, W_{QK}^{1:h}, W_V^{1:h}, W_O^{1:h})\big)^\top\Big\|_{p,\infty} \\
&\leq hB_O \cdot B_V\big(1 + 4c_{p,q}B_X^2 \cdot B_{QK}\big)\|X^\top - \tilde{X}^\top\|_{p,\infty}.
\end{aligned}$$

Proof of Proposition 20. For the difference between the outputs of the multi-head attention with different inputs, we have

$$\Big\|\big(\mathrm{MHA}(X, W_{QK}^{1:h}, W_V^{1:h}, W_O^{1:h}) - \mathrm{MHA}(\tilde{X}, W_{QK}^{1:h}, W_V^{1:h}, W_O^{1:h})\big)^\top\Big\|_{p,\infty}$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{h} \Big\| \Big(f(X, W_{QK,i}, W_{V,i}) W_{O,i} - f(\tilde{X}, W_{QK,i}, W_{V,i}) W_{O,i}) \Big)^{\top} \Big\|_{p,\infty} \\
&\leq \sum_{i=1}^{h} \|W_{O,i}^{\top}\|_{p,q} \cdot \Big\| \Big(f(X, W_{QK,i}, W_{V,i}) - f(\tilde{X}, W_{QK,i}, W_{V,i}) \Big)^{\top} \Big\|_{p,\infty} \\
&\leq \sum_{i=1}^{h} \|W_{O,i}^{\top}\|_{p,q} \cdot \|W_{V,i}^{\top}\|_{p,q} \Big(1 + 2c_{p,q} \|\tilde{X}^{\top}\|_{p,\infty} \cdot \|W_{QK,i}^{\top}\|_{p,q} \big(\|\tilde{X}^{\top}\|_{p,\infty} \\
&\qquad + \|X^{\top}\|_{p,\infty} \big) \Big) \|X^{\top} - \tilde{X}^{\top}\|_{p,\infty},
\end{aligned}
$$

where the first inequality follows from triangle inequality, the second inequality follows from Lemma 18, and the last inequality follows from Proposition 12. □

Proposition 21. *For any $X \in \mathbb{R}^{N \times d}$, and any $W_{QK,i}, \tilde{W}_{QK,i} \in \mathbb{R}^{d \times d}, W_{V,i}, \tilde{W}_{V,i} \in \mathbb{R}^{d \times \frac{d}{h}}, W_{O,i}, \tilde{W}_{O,i} \in \mathbb{R}^{\frac{d}{h} \times d}$ for $i \in [h]$ and two positive conjugate numbers $p, q \in \mathbb{R}$, if $\|X^{\top}\|_{p,\infty} \leq B_X$, $\|W_{V,i}^{\top}\|_{p,q}, \|\tilde{W}_{V,i}^{\top}\|_{p,q} \leq B_V$, and $\|W_{O,i}^{\top}\|_{p,q}, \|\tilde{W}_{O,i}^{\top}\|_{p,q} \leq B_O$ for $i \in [h]$, then we have*

$$
\begin{aligned}
&\Big\| \big(\mathrm{MHA}(X, \tilde{W}_{QK}^{1:h}, \tilde{W}_{V}^{1:h}, \tilde{W}_{O}^{1:h}) - \mathrm{MHA}(X, W_{QK}^{1:h}, W_{V}^{1:h}, W_{O}^{1:h}) \big)^{\top} \Big\|_{p,\infty} \\
&\quad \leq \sum_{i=1}^{h} B_V \cdot B_X \big\| (\tilde{W}_{O,i} - W_{O,i})^{\top} \big\|_{p,q} + B_O \cdot B_X \big\| W_{V,i}^{\top} - \tilde{W}_{V,i}^{\top} \|_{p,q} \\
&\qquad + 2c_{p,q} B_X^3 \cdot B_V \cdot B_O \|W_{QK,i}^{\top} - \tilde{W}_{QK,i}^{\top}\|_{p,q}
\end{aligned}
$$

Proof of Proposition 21. For the difference between the outputs of the multi-head attention with different parameters, we have

$$
\begin{aligned}
&\Big\| \big(\mathrm{MHA}(X, \tilde{W}_{QK}^{1:h}, \tilde{W}_{V}^{1:h}, \tilde{W}_{O}^{1:h}) - \mathrm{MHA}(X, W_{QK}^{1:h}, W_{V}^{1:h}, W_{O}^{1:h}) \big)^{\top} \Big\|_{p,\infty} \\
&\quad = \Big\| \Big(\sum_{i=1}^{h} f(X, \tilde{W}_{QK,i}, \tilde{W}_{V,i}) \tilde{W}_{O,i} - \sum_{i=1}^{h} f(X, W_{QK,i}, W_{V,i}) W_{O,i}) \Big)^{\top} \Big\|_{p,\infty} \\
&\quad \leq \Big\| \Big(\sum_{i=1}^{h} f(X, \tilde{W}_{QK,i}, \tilde{W}_{V,i}) \tilde{W}_{O,i} - \sum_{i=1}^{h} f(X, \tilde{W}_{QK,i}, \tilde{W}_{V,i}) W_{O,i}) \Big)^{\top} \Big\|_{p,\infty} \\
&\qquad + \Big\| \Big(\sum_{i=1}^{h} f(X, \tilde{W}_{QK,i}, \tilde{W}_{V,i}) W_{O,i} - \sum_{i=1}^{h} f(X, W_{QK,i}, W_{V,i}) W_{O,i}) \Big)^{\top} \Big\|_{p,\infty} \\
&\quad \leq \sum_{i=1}^{h} \Big\| \big(f(X, \tilde{W}_{QK,i}, \tilde{W}_{V,i}) \big)^{\top} \Big\|_{p,\infty} \cdot \big\| (\tilde{W}_{O,i} - W_{O,i})^{\top} \big\|_{p,q} \\
&\qquad + \sum_{i=1}^{h} \Big\| \Big(f(X, \tilde{W}_{QK,i}, \tilde{W}_{V,i}) - f(X, W_{QK,i}, W_{V,i}) \Big)^{\top} \Big\|_{p,\infty} \cdot \big\| W_{O,i}^{\top} \big\|_{p,q},
\end{aligned} \tag{N.1}
$$

where the first inequality follows from triangle inequality, and the second inequality follows from Lemma 18.

For the first term in inequality (N.1), let $\tau \in [N]$ and $x_\tau^{\top}$ be the τ^{th} row of X, then we have

$$
\begin{aligned}
\Big\| \big(f(X, \tilde{W}_{QK,i}, \tilde{W}_{V,i}) \big)^{\top} \Big\|_{p,\infty} &= \max_{\tau \in [N]} \|\mathrm{SM}\big(x_\tau^{\top} \tilde{W}_{QK,i} X^{\top} \big) X \tilde{W}_{V,i}\| \\
&\leq \max_{\tau \in [N]} \big\| \tilde{W}_{V,i}^{\top} X^{\top} \big\|_{p,\infty} \cdot \Big\| \mathrm{SM}\big(x_\tau^{\top} \tilde{W}_{QK,i} X^{\top} \big) \Big\|_1 \\
&\leq \big\| \tilde{W}_{V,i}^{\top} \big\|_{p,q} \cdot \|X^{\top}\|_{p,\infty},
\end{aligned} \tag{N.2}
$$

where the first inequality follows from Lemma 17. For the second term in inequality (N.1), recall Proposition 12, then we have

$$\begin{aligned}&\left\|\Big(f(X,\tilde{W}_{QK,i},\tilde{W}_{V,i})-f(X,W_{QK,i},W_{V,i})\Big)^\top\right\|_{p,\infty}\\&\quad\le 2c_{p,q}\|X^\top\|_{p,\infty}^3\cdot\|W_{V,i}^\top\|_{p,q}\cdot\|W_{QK,i}^\top-\tilde{W}_{QK,i}^\top\|_{p,q}+\big\|W_{V,i}^\top-\tilde{W}_{V,i}^\top\|_{p,q}\cdot\|X^\top\big\|_{p,\infty}.\end{aligned}\tag{N.3}$$

The desired result follows by substituting inequalities (N.2) and (N.3) into inequality (N.1). This concludes the proof. □

Proposition 22. *For any $X\in\mathbb{R}^{N\times d}$, and any $W_{QK,i}\in\mathbb{R}^{d\times d}, W_{V,i}\in\mathbb{R}^{d\times\frac{d}{h}}, W_{O,i}\in\mathbb{R}^{\frac{d}{h}\times d}$ for $i\in[h]$ and two positive conjugate numbers $p,q\in\mathbb{R}$, we have*

$$\left\|\big(\mathrm{MHA}(X,W_{QK}^{1:h},W_V^{1:h},W_O^{1:h})\big)^\top\right\|_{p,\infty}\le\sum_{i=1}^h\|W_{O,i}^\top\|_{p,q}\|W_{V,i}^\top\|_{p,q}\|X^\top\|_{p,\infty}.$$

Proof of Proposition 22. For the $\ell_{p,\infty}$-norm of the multi-head attention, we have

$$\begin{aligned}&\left\|\big(\mathrm{MHA}(X,W_{QK}^{1:h},W_V^{1:h},W_O^{1:h})\big)^\top\right\|_{p,\infty}\\&\quad\le\sum_{i=1}^h\left\|\Big(f(X,W_{QK,i},W_{V,i})W_{O,i}\Big)^\top\right\|_{p,\infty}\\&\quad\le\sum_{i=1}^h\|W_{O,i}^\top\|_{p,q}\cdot\left\|\big(f(X,W_{QK,i},W_{V,i})\big)^\top\right\|_{p,\infty}\\&\quad\le\sum_{i=1}^h\|W_{O,i}^\top\|_{p,q}\cdot\|W_{V,i}^\top\|_{p,q}\cdot\|X^\top\|_{p,\infty},\end{aligned}$$

where the first inequality follows from triangle inequality, the second inequality follows from Lemma 18, and the final inequality follows from inequality (M.17) in Proposition 11. □

N.2 Extension to Non-i.i.d. Sampling

The dataset $\mathcal{D}$ is collected in an i.i.d. manner in the main paper. In this this section, we extend our result to the non-i.i.d. case. Specifically, we collect the dataset $\mathcal{D}'=\{(\bar{S}_t,\bar{A}_t,r_t)\}_{t=0}^n$ by implementing a policy π_0, i.e., the action is taken as $\bar{A}_t\sim\pi_0(\cdot\,|\,\bar{S}_t)$, and the sequence of states is updated as $\bar{S}_{t+1}\sim P^*(\cdot\,|\,\bar{S}_t,\bar{A}_t)$ for $t\in[n]$. We assume that the initial state $\bar{S}_0$ is generated according to a distribution q_0, i.e., the initial state-action pair is distributed as $(\bar{S}_0,\bar{A}_0)\sim q_0\pi_0$. We denote the *stationary distribution* on the state-action pair of the Markov chain induced by the policy π_0 as $q_{P^*}^{\pi_0}(\bar{S},\bar{A})$. Note that the initial distribution $q_0\pi_0$ may not equal to the stationary distribution $q_{P^*}^{\pi_0}$. To distinguish these two different cases, we will use $P_{q_0\pi_0}$ and $P_{q_{P^*}^{\pi_0}}$ to denote the probability distributions with respect to the Markov chains with initial state distributed as $q_0\pi_0$ and $q_{P^*}^{\pi_0}$ respectively.

In such setting, we define the mismatch between two functions f and $\tilde{f}$ on $\mathcal{D}$ for a fixed policy π as $\mathcal{L}'(f,\tilde{f},\pi;\mathcal{D}')=\frac{1}{n}\sum_{t=0}^{n-1}(f(\bar{S}_t,\bar{A}_t)-\bar{r}_t-\gamma\tilde{f}(\bar{S}_{t+1},\pi))^2$, then the Bellman error of a function f with respect to the policy π is defined as $\mathcal{E}'(f,\pi;\mathcal{D}')=\mathcal{L}'(f,f,\pi;\mathcal{D}')-\inf_{\tilde{f}\in\mathcal{F}_{\mathrm{tf}}}\mathcal{L}'(\tilde{f},f,\pi;\mathcal{D}')$. The corresponding model-free algorithm can be written as

$$\hat{\pi}'=\operatorname*{argmax}_{\pi\in\Pi}\min_{f\in\mathcal{F}'(\pi,\varepsilon)}f(\bar{S}_0,\pi),\quad\text{where}\quad\mathcal{F}'(\pi,\varepsilon)=\big\{f\in\mathcal{F}_{\mathrm{tf}}(B)\,\big|\,\mathcal{E}'(f,\pi;\mathcal{D}')\le\varepsilon\big\}.\tag{N.4}$$

In the dataset $\mathcal{D}$ collected by implementing policy π_0, the mismatch between the distribution induced by the optimal policy $d_{P^*}^{\pi^*}$ and the stationary distribution $q_{P^*}^{\pi_0}$ is captured by

$$C'_{\mathcal{F}_{\mathrm{tf}}}(\pi_0)=\max_{f\in\mathcal{F}_{\mathrm{tf}}}\mathbb{E}_{d_{P^*}^{\pi^*}}\big[\big(f(\bar{S},\bar{A})-\mathcal{T}^{\pi^*}f(\bar{S},\bar{A})\big)^2\big]/\mathbb{E}_{q_{P^*}^{\pi_0}}\big[\big(f(\bar{S},\bar{A})-\mathcal{T}^{\pi^*}f(\bar{S},\bar{A})\big)^2\big],\tag{N.5}$$

where $\mathcal{F}_{\mathrm{tf}}$ is the transformer function class defined in Section 4.1.

To analyze the concentration behavior of the action-value function estimate under such sampling method, we need to define additional quantities to describe how fast the Markov chain approximates its stationary distribution. For a Markov chain with finite state space Ω and transition probability matrix P, we label the eigenvalues of P in decreasing order: $1 = \lambda_1 \geq \ldots \geq \lambda_{|\Omega|} \geq -1$. Define $\lambda^* = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } P \text{ and } \lambda \neq 1\}$. The *absolute spectral gap* of P is defined as $1 - \lambda^*$. The notion of the absolute spectral gap and our following results can also be generalized to the Markov chain with infinite state space by treating of transition kernel P as an operator of a Hilbert space. For two distributions p and q on Ω, we define

$$N(p, q) = \int_\Omega \frac{\mathrm{d}p}{\mathrm{d}q}(x)\, p(\mathrm{d}x).$$

Inspired by the ubiquitous change-of-measure technique, we will use $N(q_0, q)$ to capture the difference between the non-stationary Markov chain with initial distribution q_0 and the stationary Markov chain with stationary distribution q.

To analyze the algorithm in Eqn. (N.4), we first derive a generalization error bound of the estimate of the Bellman error using the PAC-Bayesian framework.

Proposition 23. *Consider the dataset $\mathcal{D}'$ collected by implementing a policy π_0. Let $\bar{B} = B_V B_{QK} B_a B_b B_w$. For all $f, \tilde{f} \in \mathcal{F}_{\mathrm{tf}}(B)$ and all policies $\pi \in \Pi$, with probability at least $1-\delta$, we have*

$$\begin{aligned}
&\left|\mathbb{E}_{q_{P^*}^{\pi_0}}\left[\left(f(\bar{S},\bar{A}) - \mathcal{T}^\pi \tilde{f}(\bar{S},\bar{A})\right)^2\right] - \mathcal{L}'(f,\tilde{f},\pi;\mathcal{D}') + \mathcal{L}'(\mathcal{T}^\pi\tilde{f},\tilde{f},\pi;\mathcal{D}')\right| \\
&\quad \leq \frac{C + (2-C)\lambda}{2}\mathbb{E}_{q_{P^*}^{\pi_0}}\left[\left(f(\bar{S},\bar{A}) - \mathcal{T}^\pi \tilde{f}(\bar{S},\bar{A})\right)^2\right] \\
&\qquad + O\left(\frac{V_{\max}^2}{(1-\lambda)n}\left[mL^2d^2\log\frac{mdL\bar{B}n}{V_{\max}} + \log\frac{N(q_0\pi_0, q_{P^*}^{\pi_0})\mathcal{N}(\Pi, 1/n, d_\infty)}{\delta}\right]\right), \qquad \text{(N.6)}
\end{aligned}$$

where $1-\lambda$ is the absolute spectral gap of the Markov chain $\{(\bar{S}_t, \bar{A}_t)\}_{t=0}^\infty$ induced by the policy π_0, and $0 < C < e^{1/10}$ is an absolute constant.

For ease of notation, we define $\tilde{e}(\mathcal{F}_{\mathrm{tf}}, \Pi, \pi_0, \delta, n)$ to be $(1-\lambda)n$ times the second term of the generalization error bound in (N.6). We note that Proposition 23 is a generalization of Theorem 2. When the dataset $\mathcal{D}$ consists of i.i.d. samples drawn according to μ, the dataset $\mathcal{D}$ can be treated as a Markov chain with $\lambda = 0$, and $N(\mu,\mu) = 1$. In this case, our result in Proposition 23 particularizes to the result in Theorem 2 up to a constant.

Before stating the suboptimality bound, we require two additional assumptions on the function class and the policy π_0. We first state the standard regularity assumption of the transformer function class. We assume that the collected dataset $\mathcal{D}'$ provides a good coverage of the optimal policy.

Assumption 5. *For the policy π_0, the coefficient $C'_{\mathcal{F}_{\mathrm{tf}}}(\pi_0)$ defined in Eqn.* (N.5) *is finite.*

Correspondingly, we slightly adjust the approximate realizability and complete assumption as follows:

Assumption 6. *For any $\pi \in \Pi$, we have $\inf_{f\in\mathcal{F}_{\mathrm{tf}}}\sup_{\mu\in q_\Pi}\mathbb{E}_\mu[(f(\bar{S},\bar{A}) - \mathcal{T}^\pi f(\bar{S},\bar{A}))^2] \leq \varepsilon'_{\mathcal{F}}$ and $\sup_{f\in\mathcal{F}_{\mathrm{tf}}}\inf_{\tilde{f}\in\mathcal{F}_{\mathrm{tf}}}\mathbb{E}_{q_{P^*}^{\pi_0}}[(\tilde{f}(\bar{S},\bar{A}) - \mathcal{T}^\pi f(\bar{S},\bar{A}))^2] \leq \varepsilon'_{\mathcal{F},\mathcal{F}}$, where $q_\Pi = \{\mu \mid \exists\, \pi \in \Pi \text{ s.t. } \mu = q_{P^*}^\pi\}$ is the set of stationary distributions of the state and the action pair induced by any policy $\pi \in \Pi$.*

Then the suboptimality gap of the learned policy can be upper bounded as follows.

Theorem 6. *If Assumptions 5 and 6 hold, and we take $\varepsilon = [2 + C + (2-C)\lambda]\varepsilon'_{\mathcal{F}}/2 + 2\tilde{e}(\mathcal{F}_{\mathrm{tf}}, \Pi, \pi_0, \delta, n)/[(1-\lambda)n]$, then with probability at least $1-\delta$, the suboptimality gap of the policy derived in the algorithm shown in Eqn.* (N.4) *is upper bounded as*

$$\begin{aligned}
V_{P^*}^{\pi^*}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0) \leq O\Bigg(&\sqrt{\frac{C'_{\mathcal{F}_{\mathrm{tf}}}(\pi_0)\tilde{\varepsilon}}{(1-\gamma)^2(1-\lambda)}} \\
&+ \frac{V_{\max}\sqrt{C'_{\mathcal{F}_{\mathrm{tf}}}(\pi_0)}}{(1-\gamma)(1-\lambda)\sqrt{n}}\sqrt{mL^2d^2\log\frac{mdL\bar{B}n}{V_{\max}} + \log\frac{2N(q_0\pi_0, q_{P^*}^{\pi_0})\mathcal{N}(\Pi, 1/n, d_\infty)}{\delta}}\Bigg),
\end{aligned}$$

where $d = d_{\mathcal{S}} + d_{\mathcal{A}}$, $\tilde{\varepsilon} = \varepsilon'_{\mathcal{F}} + \varepsilon'_{\mathcal{F},\mathcal{F}}$, $\bar{B}$ is defined in Proposition 23, $0 < C < e^{1/10}$ is an absolute constant, and $1-\lambda$ is the absolute spectral gap of the Markov chain $\{(\bar{S}_t, \bar{A}_t)\}_{t=0}^{\infty}$ induced by the policy π_0.

We note that Theorem 6 is a generalization of Theorem 3. Sampling in an i.i.d. manner according to μ can be regarded as a Markov chain with $\lambda = 0$, and $N(\mu, \mu) = 1$. In this case, our result in Theorem 6 particularizes to the result in Theorem 3.

Proof of Theorem 6. The proof follows along similar lines as that of Theorem 3. Recall the definition below Proposition 23, i.e.,

$$\tilde{e}(\mathcal{F}_{\mathrm{tf}}, \Pi, \pi_0, \delta, n) = C' V_{\max}^2 \left[mL^2 d^2 \log \frac{mdL\bar{B}n}{V_{\max}} + \log \frac{N(q_0\pi_0, q_{P^*}^{\pi_0}) \mathcal{N}(\Pi, 1/n, d_\infty)}{\delta} \right],$$

where $C' > 0$ is an absolute constant. To simplify the proof, we define

$$\begin{aligned} f_{\pi^*}^* &= \operatorname*{arginf}_{f \in \mathcal{F}_{\mathrm{tf}}} \sup_{\mu \in q_\Pi} \mathbb{E}_\mu \Big[\big(f(\bar{S}, \bar{A}) - \mathcal{T}^{\pi^*} f\big(\bar{S}, \bar{A}\big) \big)^2 \Big], \\ \varepsilon &= \frac{2 + C + (2-C)\lambda}{2} \varepsilon'_{\mathcal{F}} + \frac{2\tilde{e}(\mathcal{F}_{\mathrm{tf}}, \Pi, \pi_0, \delta, n)}{(1-\lambda)n}, \end{aligned}$$

where $0 < C < e^{1/10}$ is an absolute constant.

Our proof can be decomposed into three main parts.

- Since $f_{\pi^*}^*$ is the best approximation of action-value function of the optimal policy π^*, we expect that it should belong to the confidence region of the action-value functions $\mathcal{F}'(\pi^*, \varepsilon)$ with high probability. We show this in Step 1.
- For any $\pi \in \Pi$ and any $f \in \mathcal{F}'(\pi, \varepsilon)$, since the empirical Bellman error is bounded $\mathcal{E}'(f, \pi; \mathcal{D}') \leq \varepsilon$, we expect that the population Bellman error $\mathbb{E}_{q_{P^*}^{\pi_0}}[(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A}))^2]$ can be controlled with high probability, which implies that f is a reliable estimate of the action-value function of π. We show this in Step 2.
- The suboptimality gap of the learned policy according to the reliable action-value function estimate can be bounded using the estimation error bound. We do this in Step 3.

We lay out the proof by the three steps as stated in the above proof sketch.

Step 1: Show that $f_{\pi^*}^* \in \mathcal{F}'(\pi^*, \varepsilon)$ with high probability.

From the definition of $f_{\pi^*}^*$ and Assumption 6, we note that the population Bellman error of $f_{\pi^*}^*$ with respect to π^* is bounded by $\varepsilon'_{\mathcal{F}}$. To bound the empirical Bellman error $\mathcal{E}'(f_{\pi^*}^*, \pi^*; \mathcal{D}')$ of $f_{\pi^*}^*$, we utilize the generalization error bound of the action-value function with the transformer function class.

Proposition 23. *Consider the dataset $\mathcal{D}'$ collected by implementing a policy π_0. Let $\bar{B} = B_V B_{QK} B_a B_b B_w$. For all $f, \tilde{f} \in \mathcal{F}_{\mathrm{tf}}(B)$ and all policies $\pi \in \Pi$, with probability at least $1 - \delta$, we have*

$$\begin{aligned} &\left| \mathbb{E}_{q_{P^*}^{\pi_0}} \Big[\big(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi \tilde{f}(\bar{S}, \bar{A}) \big)^2 \Big] - \mathcal{L}'(f, \tilde{f}, \pi; \mathcal{D}') + \mathcal{L}'(\mathcal{T}^\pi \tilde{f}, \tilde{f}, \pi; \mathcal{D}') \right| \\ &\quad \leq \frac{C + (2-C)\lambda}{2} \mathbb{E}_{q_{P^*}^{\pi_0}} \Big[\big(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi \tilde{f}(\bar{S}, \bar{A}) \big)^2 \Big] \\ &\qquad + O\bigg(\frac{V_{\max}^2}{(1-\lambda)n} \bigg[mL^2 d^2 \log \frac{mdL\bar{B}n}{V_{\max}} + \log \frac{N(q_0\pi_0, q_{P^*}^{\pi_0}) \mathcal{N}(\Pi, 1/n, d_\infty)}{\delta} \bigg] \bigg), \end{aligned} \tag{N.6}$$

where $1 - \lambda$ is the absolute spectral gap of the Markov chain $\{(\bar{S}_t, \bar{A}_t)\}_{t=0}^{\infty}$ induced by the policy π_0, and $0 < C < e^{1/10}$ is an absolute constant.

Proof. See Appendix N.3.1 for a detailed proof. □

We can decompose the empirical Bellman error $\mathcal{E}'(f^*_{\pi^*}, \pi^*; \mathcal{D}')$ as the sum of the population Bellman error and the generalization error, where the population Bellman error can be controlled with $\varepsilon'_{\mathcal{F}}$ according to Assumption 6, and the generalization error can be controlled with Proposition 23. Thus, we have the following lemma.

Lemma 24. *For any $\pi \in \Pi$, let $f^*_\pi = \operatorname{arginf}_{f \in \mathcal{F}_{\text{tf}}} \sup_{\mu \in q_\Pi} \mathbb{E}_\mu[(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A}))^2]$. If Assumption 6 holds, the following inequality holds with probability at least $1-\delta$,*

$$\mathcal{E}'(f^*_\pi, \pi; \mathcal{D}') \leq \frac{2 + C + (2-C)\lambda}{2} \varepsilon'_{\mathcal{F}} + \frac{2\tilde{e}(\mathcal{F}_{\text{tf}}, \Pi, \pi_0, \delta, n)}{(1-\lambda)n}.$$

Proof. The proof is same as the proof of Lemma 4 except using the concentration inequality in Proposition 23. $\square$

Step 2: For any policy $\pi \in \Pi$ and $f \in \mathcal{F}'(\pi, \varepsilon)$, show $\mathbb{E}_{q^{\pi_0}_{P^*}}[(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A}))^2]$ is small with high probability.

To prove the desired result, we relate the population Bellman error $\mathbb{E}_{q^{\pi_0}_{P^*}}[(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A}))^2]$ with $\mathcal{E}'(f, \pi; \mathcal{D}')$ using Proposition 23, where we bound the population Bellman error as the difference between the empirical Bellman error and the generalization error. Thus, we have the following lemma.

Lemma 25. *For any $\pi \in \Pi$ and $f \in \mathcal{F}_{\text{tf}}$, if $\mathcal{E}'(f, \pi; \mathcal{D}') \leq \varepsilon$ for some $\varepsilon > 0$, and Assumption 6 holds, the following inequality holds with probability at least $1-\delta$,*

$$\begin{aligned}
&\mathbb{E}_{q^{\pi_0}_{P^*}}\left[\left(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A})\right)^2\right] \\
&\quad \leq \frac{2}{(2-C)(1-\lambda)} \varepsilon + \frac{2 + C + (2-C)\lambda}{(2-C)(1-\lambda)} \varepsilon'_{\mathcal{F}, \mathcal{F}} + \frac{4\tilde{e}(\mathcal{F}_{\text{tf}}, \Pi, \pi_0, \delta, n)}{(2-C)(1-\lambda)^2 n}.
\end{aligned}$$

Proof. The proof is same as the proof of Lemma 5 except using the cencentration inequality in Proposition 23. $\square$

Step 3: Bound the suboptimality gap of the learned policy with the population Bellman error bound in Step 2.

We define

$$\begin{aligned}
\hat{f}_{\pi^*} &= \underset{f \in \mathcal{F}'(\pi^*, \varepsilon)}{\operatorname{argmax}} \, f(\bar{S}_0, \pi^*), \\
\check{f}_{\pi^*} &= \underset{f \in \mathcal{F}'(\pi^*, \varepsilon)}{\operatorname{argmin}} \, f(\bar{S}_0, \pi^*),
\end{aligned}$$

Following the same procedures in step 3 of the proof of Theorem 3, we can show that

$$V^{\pi^*}_{P^*}(\bar{S}_0) - V^{\hat{\pi}}_{P^*}(\bar{S}_0) \leq \hat{f}_{\pi^*}(\bar{S}_0, \pi^*) - V^{\pi^*}_{P^*}(\bar{S}_0) + V^{\pi^*}_{P^*}(\bar{S}_0) - \check{f}_{\pi^*}(\bar{S}_0, \pi^*) + \frac{2\sqrt{\varepsilon'_{\mathcal{F}}}}{1-\gamma}. \tag{N.7}$$

Applying the suboptimality gap decomposition in Lemma 14 to inequality (N.7), we have

$$\begin{aligned}
&V^{\pi^*}_{P^*}(\bar{S}_0) - V^{\hat{\pi}}_{P^*}(\bar{S}_0) \\
&\quad \leq \frac{1}{1-\gamma} \Big\{ \mathbb{E}_{d^{\pi^*}_{P^*}}\left[\hat{f}_{\pi^*}(\bar{S}, \bar{A}) - \mathcal{T}^{\pi^*} \hat{f}_{\pi^*}(\bar{S}, \bar{A})\right] \\
&\qquad - \mathbb{E}_{d^{\pi^*}_{P^*}}\left[\check{f}_{\pi^*}(\bar{S}, \bar{A}) - \mathcal{T}^{\pi^*} \check{f}_{\pi^*}(\bar{S}, \bar{A})\right] \Big\} + \frac{2\sqrt{\varepsilon'_{\mathcal{F}}}}{1-\gamma} \\
&\quad \leq \frac{1}{1-\gamma} \Bigg\{ \sqrt{C'_{\mathcal{F}_{\text{tf}}}(\pi_0) \mathbb{E}_{q^{\pi_0}_{P^*}}\left[\left(\hat{f}_{\pi^*}(\bar{S}, \bar{A}) - \mathcal{T}^{\pi^*} \hat{f}_{\pi^*}(\bar{S}, \bar{A})\right)^2\right]} \\
&\qquad + \sqrt{C'_{\mathcal{F}_{\text{tf}}}(\pi_0) \mathbb{E}_{q^{\pi_0}_{P^*}}\left[\left(\check{f}_{\pi^*}(\bar{S}, \bar{A}) - \mathcal{T}^{\pi^*} \check{f}_{\pi^*}(\bar{S}, \bar{A})\right)^2\right]} \Bigg\} + \frac{2\sqrt{\varepsilon'_{\mathcal{F}}}}{1-\gamma},
\end{aligned}$$

where the first inequality follows from Lemma 14, and the second inequality follows from Jensen's inequality and the definition of $C'_{\mathcal{F}_{\mathrm{tf}}}(\pi_0)$. Combined with the result in Step 2, we have

$$
\begin{aligned}
&V_{P^*}^{\pi^*}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0)\\
&\quad\leq \frac{\sqrt{C'_{\mathcal{F}_{\mathrm{tf}}}(\pi_0)}}{1-\gamma}\sqrt{\frac{2}{(2-C)(1-\lambda)}\varepsilon + \frac{2+C+(2-C)\lambda}{(2-C)(1-\lambda)}\varepsilon'_{\mathcal{F},\mathcal{F}} + \frac{4\tilde{e}(\mathcal{F}_{\mathrm{tf}},\Pi,\pi_0,\delta,n)}{(2-C)(1-\lambda)^2 n}} + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}\\
&\quad\leq O\bigg(\sqrt{\frac{C'_{\mathcal{F}_{\mathrm{tf}}}(\pi_0)(\varepsilon'_{\mathcal{F}} + \varepsilon'_{\mathcal{F},\mathcal{F}})}{(1-\gamma)^2(1-\lambda)}} + \frac{\sqrt{C'_{\mathcal{F}_{\mathrm{tf}}}(\pi_0)}}{(1-\gamma)(1-\lambda)}\sqrt{\frac{\tilde{e}(\mathcal{F}_{\mathrm{tf}},\Pi,\pi_0,\delta,n)}{n}}\bigg).
\end{aligned}
$$

Therefore, we conclude the proof of Theorem 6. □

N.3 Proofs of Supporting Propositions in Section N.2

N.3.1 Proof of Proposition 23

Proof of Proposition 23. Similar to the proof of Theorem 2, we adopt a PAC-Bayesian framework to derive our desired generalization error bound. We first state a preliminary result.

Proposition 26. *Let $\{X_i\}_{i\geq 1}$ be a Markov chain with state space Ω, stationary distribution q, initial distribution $X_1 \sim q_0$, and absolute spectral gap $1-\lambda$. Set $\mathcal{F}$ be the collection of functions of $f : \Omega \to \mathbb{R}$. For any $f \in \mathcal{F}$, we define*

$$
q(f) = \mathbb{E}_q\big[f(X)\big], \quad \sigma^2(f) = \mathrm{Var}_q\big(f(X)\big),
$$

where the expectation is taken with respect to the stationary distribution q. Let Q be the distribution of the random function f. Assume that $|f(X)) - q(f)| \leq c$ almost surely with respect to Q for some constant $c > 0$. Then we have that with probability at least $1-\delta$, the following inequality holds.

$$
\bigg|\mathbb{E}_Q\bigg[q(f) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\bigg]\bigg| \leq \frac{C+(2-C)\lambda}{10c}\mathbb{E}_Q[\sigma^2(f)] + \frac{10c}{(1-\lambda)n}\bigg[\mathrm{KL}(Q\|P_0) + \log\frac{2N(q_0,q)}{\delta^2}\bigg], \tag{N.8}
$$

where C is an absolute constant such that $0 < C < e^{1/10}$.

Proof. See Appendix N.3.2. □

Our proof can be decomposed into two main parts.

- We verify that the Bellman error satisfies the conditions in Proposition 26 and apply it to the Bellman error.
- We adopt the similar procedure in the proof of Theorem 2 to control the fluctuation of both sides in inequality (N.8) and calculate $\mathrm{KL}(Q\|P_0)$.

Step 1: Verify the conditions in Proposition 23

We consider the Markov chain formed by $\{(\bar{S}_t, \bar{A}_t, \bar{S}_{t+1}, \bar{A}_{t+1})\}_{t=0}^{\infty}$. Note that this Markov chain shares the same absolute spectral gap with the Markov chain $\{(\bar{S}_t, \bar{A}_t)\}_{t=0}^{\infty}$ when $\mathcal{S}$ and $\mathcal{A}$ are finite.

Let $X_t = (\bar{S}_t, \bar{A}_t, \bar{S}_{t+1}, \bar{A}_{t+1})$ for all $f, \tilde{f} \in \mathcal{F}_{\mathrm{tf}}(B_a, B_b, B_{QK}, B_V, B_w)$. We define

$$
\begin{aligned}
l'(f,\tilde{f},\pi;X_t) = \big(&f(\bar{S}_t,\bar{A}_t) - \bar{r}(\bar{S}_t,\bar{A}_t) - \gamma\tilde{f}(\bar{S}_{t+1},\pi)\big)^2\\
&- \big(\mathcal{T}^\pi\tilde{f}(\bar{S}_t,\bar{A}_t) - \bar{r}(\bar{S}_t,\bar{A}_t) - \gamma\tilde{f}(\bar{S}_{t+1},\pi)\big)^2.
\end{aligned}
$$

Then the term we consider in Theorem 2 can be expressed as

$$
\mathcal{L}'(f,\tilde{f},\pi;\mathcal{D}') - \mathcal{L}'(\mathcal{T}^\pi\tilde{f},\tilde{f},\pi;\mathcal{D}') = \frac{1}{n}\sum_{i=1}^{n} l'(f,\tilde{f},\pi;X_i) \text{ and } \big|l'(f,\tilde{f},\pi;X)\big| \leq 4V_{\max}^2.
$$

Then the expectation of $l(f, \tilde{f}, \pi; X)$ with respect to the stationary distribution $(\bar{S}_t, \bar{A}_t, \bar{S}_{t+1}, \bar{A}_{t+1}) \sim q_{P^*}^{\pi_0} \times P^* \times \pi_0$ is

$$
\begin{aligned}
&\mathbb{E}_{q_{P^*}^{\pi_0} \times P^* \times \pi_0}\big[l'(f, \tilde{f}, \pi; X_t)\big] \\
&\quad = \mathbb{E}_{q_{P^*}^{\pi_0} \times \bar{P^*}}\Big[\big(f(\bar{S}_t, \bar{A}_t) - \mathcal{T}^\pi \tilde{f}(\bar{S}_t, \bar{A}_t)\big)\big(f(\bar{S}_t, \bar{A}_t) + \mathcal{T}^\pi \tilde{f}(\bar{S}_t, \bar{A}_t) - 2\bar{r} - 2\gamma \tilde{f}(\bar{S}_{t+1}, \pi)\big)\Big] \\
&\quad = \mathbb{E}_{q_{P^*}^{\pi_0}}\bigg[\mathbb{E}_{P^*}\Big[\big(f(\bar{S}_t, \bar{A}_t) - \mathcal{T}^\pi \tilde{f}(\bar{S}_t, \bar{A}_t)\big)\big(f(\bar{S}_t, \bar{A}_t) + \mathcal{T}^\pi \tilde{f}(\bar{S}_t, \bar{A}_t) \\
&\qquad\qquad\qquad - 2\bar{r} - 2\gamma \tilde{f}(\bar{S}_{t+1}, \pi)\big) \,\Big|\, \bar{S}_t, \bar{A}_t\Big]\bigg] \\
&\quad = \mathbb{E}_{q_{P^*}^{\pi_0}}\Big[\big(f(\bar{S}_t, \bar{A}_t) - \mathcal{T}^\pi \tilde{f}(\bar{S}_t, \bar{A}_t)\big)^2\Big], \qquad \text{(N.9)}
\end{aligned}
$$

where the last equality follows from the definition of the Bellman operator. As a consequence, the variance of $l'(f, \tilde{f}, \pi; X)$ can be bounded by its expectation as

$$
\begin{aligned}
&\mathrm{Var}_{q_{P^*}^{\pi_0} \times P^* \times \pi_0}\big(l'(f, \tilde{f}, \pi; X_t)\big) \\
&\quad \le \mathbb{E}_{q_{P^*}^{\pi_0} \times P^* \times \pi_0}\Big[\big(l'(f, \tilde{f}, \pi; X_t)\big)^2\Big] \\
&\quad = \mathbb{E}_{q_{P^*}^{\pi_0}}\bigg[\mathbb{E}_{P^*}\Big[\big(f(\bar{S}_t, \bar{A}_t) - \mathcal{T}^\pi \tilde{f}(\bar{S}_t, \bar{A}_t)\big)^2\big(f(\bar{S}_t, \bar{A}_t) + \mathcal{T}^\pi \tilde{f}(\bar{S}_t, \bar{A}_t) \\
&\qquad\qquad\qquad - 2\bar{r} - 2\gamma \tilde{f}(\bar{S}_{t+1}, \pi)\big)^2 \,\Big|\, \bar{S}_t, \bar{A}_t\Big]\bigg] \\
&\quad \le 16 V_{\max}^2 \mathbb{E}_{q_{P^*}^{\pi_0}}\Big[\big(f(\bar{S}_t, \bar{A}_t) - \mathcal{T}^\pi \tilde{f}(\bar{S}_t, \bar{A}_t)\big)^2\Big] \qquad \text{(N.10)}
\end{aligned}
$$

where the last inequality follows from the fact that f and $\tilde{f}$ is bounded by $V_{\max}$. Eq. (N.9) shows that $l'(f, \tilde{f}, \pi; X_t)$ satisfies the condition in Proposition 26 with $c = 4V_{\max}^2$. Applying Proposition 7 and inequality (N.10) to $l'(f, \tilde{f}, \pi; X_t)$, we have with probability at least $1 - \delta$,

$$
\begin{aligned}
&\bigg|\mathbb{E}_Q\bigg[\mathbb{E}_{q_{P^*}^{\pi_0}}\Big[\big(f(\bar{S}_t, \bar{A}_t) - \mathcal{T}^\pi \tilde{f}(\bar{S}_t, \bar{A}_t)\big)^2\Big] - \frac{1}{n}\sum_{t=0}^{n-1} l'(f, \tilde{f}, \pi; X_t)\bigg]\bigg| \\
&\quad \le \frac{C + (2 - C)\lambda}{2} \mathbb{E}_{Q, q_{P^*}^{\pi_0}}\Big[\big(f(\bar{S}_t, \bar{A}_t) - \mathcal{T}^\pi \tilde{f}(\bar{S}_t, \bar{A}_t)\big)^2\Big] \\
&\qquad + \frac{40 V_{\max}^2}{(1 - \lambda)n}\bigg[\mathrm{KL}(Q \| P_0) + \log \frac{2N(q_0\pi_0, q_{P^*}^{\pi_0})}{\delta^2}\bigg], \qquad \text{(N.11)}
\end{aligned}
$$

where $0 < C < e^{1/10}$ is an absolute constant.

Step 2: Control the fluctuation of both sides in inequality (N.11) **and calculate** $\mathrm{KL}(Q \| P_0)$

To control the fluctuation of both sides in inequality (N.11) and calculate $\mathrm{KL}(Q \| P_0)$, we take the same procedure in the steps 2, 3 and 4 in the proof of Theorem 2. We derive the uniform convergence result that for all $f, \tilde{f} \in \mathcal{F}_{\mathrm{tf}}(B)$ and all policies $\pi \in \Pi$, with probability at least $1 - \delta$, we have

$$
\begin{aligned}
&\Big|\mathbb{E}_{q_{P^*}^{\pi_0}}\Big[\big(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi \tilde{f}(\bar{S}, \bar{A})\big)^2\Big] - \mathcal{L}'(f, \tilde{f}, \pi; \mathcal{D}') + \mathcal{L}'(\mathcal{T}^\pi \tilde{f}, \tilde{f}, \pi; \mathcal{D}')\Big| \\
&\quad \le \frac{C + (2 - C)\lambda}{2} \mathbb{E}_{q_{P^*}^{\pi_0}}\Big[\big(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi \tilde{f}(\bar{S}, \bar{A})\big)^2\Big] \\
&\qquad + O\bigg(\frac{V_{\max}^2}{(1 - \lambda)n}\bigg[mL^2 d^2 \log \frac{mdL\bar{B}n}{V_{\max}} + \log \frac{N(q_0\pi_0, q_{P^*}^{\pi_0})\mathcal{N}(\Pi, 1/n, d_\infty)}{\delta}\bigg]\bigg).
\end{aligned}
$$

Therefore, we conclude the proof of Proposition 23 □

N.3.2 Proof of Proposition 26

Proof of Proposition 26. The proof consists of two main steps. First, we assume that the initial state is distributed as the stationary distribution q and derive the results under this stationary setting.

Second, we extend the result to the non-stationary Markov chain, i.e., the initial state is not distributed as q but q_0.

Step 1: Derive a concentration bound when the initial state's distribution is the stationary distribution

Under the stationary setting, we make use of the following concentration results in [53].

Proposition 27 (Theorem 1 in [53])**.** *Suppose $\{X_i\}_{i\geq 1}$ is a stationary Markov chain with invariant distribution q and non-zero absolute spectral gap $1-\lambda > 0$, and $f_i : x \to [-c, +c]$ is a sequence of functions with $q(f_i) = 0$. Let $\sigma^2 = 1/n \sum_{i=1}^n q(f_i^2)$. Then for any $0 \leq t < (1-\lambda)/5c$, we have*

$$\mathbb{E}_q\bigg[\exp\bigg(t\sum_{i=1}^n f_i(X_i)\bigg)\bigg] \leq \exp\bigg(\frac{n\sigma^2}{c^2}(e^{tc}-1-tc) + \frac{n\sigma^2\lambda t^2}{1-\lambda-5ct}\bigg).$$

Set $f_i(X_i) = f(X_i) - q(f) = g(X_i)$. Proposition 27 shows that for $0 \leq t < (1-\lambda)n/(5c)$,

$$\mathbb{E}_q\bigg[\exp\bigg(\frac{t}{n}\sum_{i=1}^n g(X_i)\bigg)\bigg] \leq \exp\bigg[\frac{n\sigma^2}{c^2}\Big(e^{ct/n}-1-\frac{ct}{n}\Big) + \frac{\lambda\sigma^2 t^2}{n(1-\lambda-5ct/n)}\bigg], \tag{N.12}$$

where $\sigma^2 = \sigma^2(f)$. We define

$$\varepsilon_n(f, X_1^n) = \frac{t}{n}\sum_{i=1}^n g(X_i) - \bigg[\frac{n\sigma^2}{c^2}\Big(e^{ct/n}-1-\frac{ct}{n}\Big) + \frac{\lambda\sigma^2 t^2}{n(1-\lambda-5ct/n)}\bigg],$$

By inequality (N.12) and Markov's inequality, we have that for any distribution P_0 on the function class $\mathcal{F}$, the random variable $\varepsilon_n(f, X_1^n)$ induced by the Markov chain $\{X_i\}_{i=1}^n$ satisfies

$$P_q\bigg(\mathbb{E}_{f\sim P_0}\Big[\exp\big(\varepsilon_n(f, X_1^n)\big) \geq \frac{2}{\delta}\Big]\bigg) \leq \frac{\delta}{2}, \tag{N.13}$$

where the probability is taken with respect to the Markov chain with initial distribution q.

Setting $g(f) = \varepsilon_n(f, X_1^n)$ in Theorem 5, we have

$$\mathbb{E}_Q[\varepsilon_n(f, X_1^n)] \leq \mathrm{KL}(Q\|P_0) + \log \mathbb{E}_{P_0}\Big[\exp\big(\varepsilon_n(f, X_1^n)\big)\Big]. \tag{N.14}$$

Substituting inequality (N.13) into inequality (N.14), we have that with probability at least $1-\delta/2$

$$\mathbb{E}_Q\bigg[\frac{t}{n}\sum_{i=1}^n g(X_i) - \bigg[\frac{n\sigma^2}{c^2}\Big(e^{ct/n}-1-\frac{ct}{n}\Big) + \frac{\lambda\sigma^2 t^2}{n(1-\lambda-5ct/n)}\bigg]\bigg] \leq \mathrm{KL}(Q\|P_0) + \log\frac{2}{\delta}. \tag{N.15}$$

Set $t/n = (1-\lambda)/(10c)$. Since $e^x - 1 - x \leq ax^2$ for all $x \in [0, \log 2a]$, the left-hand side of inequality (N.15) can be upper bounded as

$$\begin{aligned}
&\mathbb{E}_Q\bigg[\frac{1}{n}\sum_{i=1}^n g(X_i)\bigg] \\
&\leq \bigg[\frac{n}{c^2 t}\Big(e^{ct/n}-1-\frac{ct}{n}\Big) + \frac{\lambda t}{n(1-\lambda-5ct/n)}\bigg]\mathbb{E}_Q[\sigma^2(f)] + \frac{1}{t}\mathrm{KL}(Q\|P_0) + \frac{1}{t}\log\frac{2}{\delta} \\
&\leq \bigg[\frac{n}{c^2 t}\cdot C\frac{t^2}{n^2} + \frac{\lambda t^2}{n(1-\lambda-5ct/n)}\bigg]\mathbb{E}_Q[\sigma^2(f)] + \frac{1}{t}\mathrm{KL}(Q\|P_0) + \frac{1}{t}\log\frac{2}{\delta} \\
&= \frac{C+(2-C)\lambda}{10c}\mathbb{E}_Q[\sigma^2(f)] + \frac{10c}{(1-\lambda)n}\bigg[\mathrm{KL}(Q\|P_0) + \log\frac{2}{\delta}\bigg],
\end{aligned}$$

where the C in the second inequality is a constant that $C \leq e^{(1-\lambda)/10} < e^{1/10}$, the equality follows from substituting the value of t into the second inequality, and the expectation in $\mathbb{E}_Q[\sigma^2(f)]$ is taken with respect to the distribution Q on the set of function class $\mathcal{F}$. From symmetry, we can show that the with probability (taken with respect to the Markov chain initialized with the stationary distribution) at least $1-\delta$

$$\left|\mathbb{E}_Q\left[\frac{1}{n}\sum_{i=1}^n g(X_i)\right]\right| \le \frac{C+(2-C)\lambda}{10c}\mathbb{E}_Q[\sigma^2(f)] + \frac{10c}{(1-\lambda)n}\left[\mathrm{KL}(Q\|P_0)+\log\frac{2}{\delta}\right], \tag{N.16}$$

where $0<C<e^{1/10}$ is an absolute constant.

Step 2: Extend inequality (N.16) **to an arbitrarily initialized Markov chain.**

To extend the results to an arbitrarily initialized Markov chain, we make use of the following result in [54].

Proposition 28 (Proposition 3.15 in [54])**.** *Let $\{X_i\}_{i=1}^\infty$ be a time homogeneous Markov chain with state space Ω, and stationary distribution q. Suppose that $g:\Omega^n\to\mathbb{R}$ is a real-valued measurable function. Then*

$$P_{q_0}\big(g(X_1,\cdots,X_n)\ge t\big) \le N(q_0,q)^{1/2}\cdot\Big[P_q\big(g(X_1,\cdots,X_n)\ge t\big)\Big]^{1/2},$$

where q_0 is any distribution on Ω, and P_{q_0} and P_q are the probability measures with respect to the Markov chains with initial state $X_1\sim q_0$ and $X_1\sim q$ respectively.

Combining Proposition 28 and inequality (N.16), we have that with probability (taken with respect to the arbitrarily initialized Markov chain) at least $1-\delta$

$$\left|\mathbb{E}_Q\left[\frac{1}{n}\sum_{i=1}^n g(X_i)\right]\right| \le \frac{C+(2-C)\lambda}{10c}\mathbb{E}_Q[\sigma^2(f)] + \frac{10c}{(1-\lambda)n}\left[\mathrm{KL}(Q\|P_0)+\log\frac{2N(q_0,q)}{\delta^2}\right]. \tag{N.17}$$

This concludes the proof of Proposition 26. □

O Experiments

Although the main aim of this paper is primarily theoretical, we provide some experiments of the model-free algorithms to illustrate the superiority of the transformer in homogeneous MARL.

O.1 Simulation Environment

In the experiments, we evaluate the performance of the algorithms on the MPE [44, 45]. We focus on the *cooperative navigation* task, where N agents move cooperatively to cover L landmarks in the environment. Given N agent positions $x_i\in\mathbb{R}^2$ for $i\in[N]$ and L landmark positions $y_j\in\mathbb{R}^2$ for $j\in[L]$, the agents receive the reward

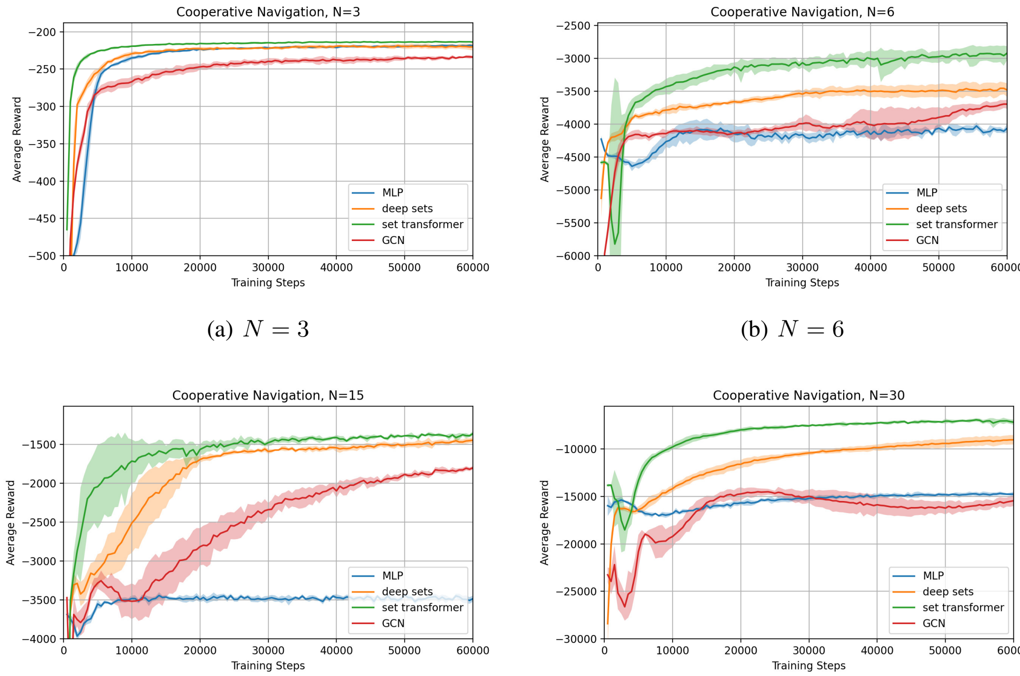
$$r = -\sum_{j=1}^L \min_{i\in[N]}\|y_j-x_i\|_2.$$

This reward encourages the agents to move closer to the landmarks. We set the number of agents as $N=3,6,15,30$ and the number of landmarks as $L=N$. To collect an offline dataset, we learn a policy in the online setting, and the dataset is collected from the induced stationary distribution of such policy.

In the training process, we use the Titan RTX and Intel(R) Core(TM) i7-6900K CPU @ 3.20GHz to train the neural networks. The size of the offline dataset is 60000×25, where we simulate 60000 episodes and implement 25 steps in each episode. The learning rate is set to 10^{-3}. The batch size is 1024. The discount factor is $\gamma=0.95$.

O.2 Simulation Results

We respectively adopt the MLP, deep sets, GCN [46] and set transformer to estimate the value function. We note that the deep sets, GCN, and set transformer are permutation invariant functions.



(a) $N = 3$

(b) $N = 6$

(c) $N = 15$

(d) $N = 30$

Figure 4: The average rewards of the model-free RL algorithms with their standard deviations.

We use the code in in [10] for the implementation of the deep sets and set transformer. To implement the model-free algorithm specified in Eqn. (1), we optimize the policy and the action-value function in an alternating fashion. In addition, instead of imposing the hard constraint on the Bellman error $\mathcal{E}(f, \pi; \mathcal{D})$, we added a Lagrangian multiplier to account for this inequality constraint.

In Figure 4, we plot the performances of the model-free RL algorithms that adopt different neural networks to estimate the action-value function. When the number of agents are small, as shown in Figure 4(a), the performances of different neural networks are similar. As shown in Theorem 1, relational reasoning abilities of the deep sets and the MLP are worse than that of the set transformer. As a consequence, when the number of agents increases, as shown in Figures 4(b) to 4(d), the superiority of the algorithm that adopts the set transformer to estimate the action-value function becomes obvious. This strongly corroborates our theoretical results in Theorems 1 and 3.