QUASI-OPTIMAL LEARNING WITH CONTINUOUS TREATMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many real-world applications of reinforcement learning (RL) require making decisions in continuous action environments. In particular, determining the optimal dose level plays a vital role in developing medical treatment regimes. One challenge in adapting existing RL algorithms to medical applications, however, is that the popular infinite support stochastic policies, e.g., Gaussian policy, may assign riskily high dosages and harm patients seriously. Hence, it is important to induce a policy class whose support only contains near-optimal actions, and shrink the action-searching area for effectiveness and reliability. To achieve this, we develop a novel *quasi-optimal learning algorithm*, which can be easily optimized in off-policy settings with guaranteed convergence under general function approximations. Theoretically, we analyze the consistency, sample complexity, adaptability, and convergence of the proposed algorithm. We evaluate our algorithm with comprehensive simulated experiments and a dose suggestion real application to Ohio Type 1 diabetes dataset.

1 INTRODUCTION

Learning good strategies in a continuous action space is important for many real-world problems (Lillicrap et al., 2015), including precision medicine, autonomous driving, etc. In particular, when developing a new dynamic regime to guide the use of medical treatments, it is often necessary to decide the optimal dose level (Murphy, 2003; Laber et al., 2014; Chen et al., 2016; Zhou et al., 2021b). In infinite horizon sequential decision-making settings (Luckett et al., 2019; Shi et al., 2021), learning such a dynamic treatment regime falls into a reinforcement learning (RL) framework. Many RL algorithms (Mnih et al., 2013; Silver et al., 2017; Nachum et al., 2017; Chow et al., 2018; Hessel et al., 2018) have achieved considerable success when the action space is finite. A straightforward approach to adapting these methods to continuous domains is to use a discretization of the continuous action space. However, this strategy has several drawbacks. If the discretization is coarse, the estimated treatment effects do not preserve continuity and cause a large bias (Lee et al., 2018a; Cai et al., 2021). On the other hand, when the discretization is fine-grained, the number of elements in the resulting action space is intractably high. Hence, learning may be slow due to the curse of dimensionality (Chou et al., 2017).

There has been recent progress on model-free reinforcement learning in continuous action spaces without utilizing discretization. In policy-based methods (Williams, 1992; Sutton et al., 1999; Silver et al., 2014; Duan et al., 2016), a Gaussian distribution is used frequently for policy distribution representation, while its mean and variance are parameterized using function approximation and updated via policy gradient descent. In addition, many actor-critic based approaches, e.g., soft actor-critic (Haarnoja et al., 2018b), ensemble critic (Fujimoto et al., 2018) and Smoothie (Nachum et al., 2018a), have been developed to improve the performance in continuous action spaces. These works target to model a Gaussian policy for action allocations as well.

However, there are two less-investigated issues in the aforementioned approaches, especially under medical applications. First, existing methods that use an infinite support Gaussian policy as the treatment policy may assign arbitrarily high dose levels, which may potentially harm the patient. Hence, these approaches are not reliable in practice due to safety and ethical concerns. It would be more desirable to develop a policy class to identify the near-optimal, or at least safe, action regions, and reduce the optimal action search area for reliability and effectiveness. Those actions out of the

identified region are discriminated as non-optimal, and would be screened out with zero densities in the policy distribution. Second, for many real-world applications, the action spaces are bounded due to practical constraints. Examples include autonomous driving with a limited steering angle and dose assignment with a budget or safety constraint. In these scenarios, modeling an optimal policy by an infinite support probability distribution, e.g., Gaussian policy, would inevitably introduce a non-negligible off-support bias as shown in Figure 2. In consequence, the off-support bias damages the performance of policy learning and results in a biased decision-making procedure. Instead, constructing a policy class with finite but adjustable support might be one of the demanding solutions.

In this work, we take a substantial step towards solving the aforementioned issues by developing a novel *quasi-optimal learning* algorithm. Our development hinges upon a novel quasi-optimal Bellman operator and stationarity equation, which is solved via minimizing an unbiased kernel embedding loss. Quasi-optimal learning estimates an implicit stochastic policy distribution whose support region only contains near-optimal actions. In addition, our algorithm overcomes the difficulties of the non-smoothness learning issue and the double sampling issue (Baird, 1995), and can be easily optimized using sampled transitions in off-policy scenarios without training instability and divergence.

The main contribution of this paper can be summarized as follows:

- We construct a novel Bellman operator and develop a reliable stochastic policy class, which is able to identify quasi-optimal action regions in scenarios with a bounded or unbounded action space. This address the shortcomings of existing approaches relying on modeling an optimal policy with infinite support distributions.
- We formalize an unbiased learning framework for estimating the designed quasi-optimal policy. Our framework avoids the double sampling issue and can be optimized using sampled transitions, which is beneficial in offline policy optimization tasks.
- We thoroughly investigate the theoretical properties of the quasi-optimal learning algorithm, including the adaptability of the quasi-optimal policy class, the loss consistency, the finite-sample bound for performance error, and the convergence analysis of the algorithm.
- Empirical analyses are conducted with comprehensive numerical experiments and a real-world case study, to evaluate the model performance in practice.

2 PRELIMINARIES

Notations We first give an introduction to our notations. For two sequences $\{\Psi(m)\}_{m\geq 1}$ and $\{\Upsilon(m)\}_{m\geq 1}$, the notation $\{\Psi(m)\}_{m\geq 1} \lesssim \{\Upsilon(m)\}_{m\geq 1}$ means that there exists a sufficiently small constant $c \geq 0$ such that $\Psi(n) \leq c\Upsilon(n)$. $\|\cdot\|_{L^p}$ and $\|\cdot\|_\infty$ denote the L^p norm and supremum-norm, respectively. We define the set indicator function $\mathbb{1}_{\text{set}}(x) = 1$ if $x \in$ set or 0 otherwise. The notation $\mathbb{P}_n$ denotes the empirical measure i.e., $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n$. For two sets $\aleph_0$ and $\aleph_1$, the notation $\aleph_0 \setminus \aleph_1$ indicates that the set $\aleph_0$ excluding the elements in the set $\aleph_1$. We write $|\aleph_0|$ as the cardinality of the set $\aleph_0$. For any Borel set $\aleph_2$, we denote $\sigma(\aleph_2)$ as the Borel measure of $\aleph_2$. We use $\mathcal{O}$ as the convention, and denotes a probability simplex over a space $\mathcal{F}$ by $\Delta(\mathcal{F})$. And $\lfloor\cdot\rfloor$ is the floor function.

Background A Markov decision process (MDP) is defined as a tuple $< \mathcal{S}, \mathcal{A}, \mathbf{P}, R, \gamma >$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathbf{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the unknown transitional kernel, $R : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a bounded reward function, and $\gamma \in [0, 1)$ is the discounted factor. In this paper, we focus on the scenario of continuous action space. We assume the offline data consists of n i.i.d. trajectories, i.e., $\mathcal{D}_{1:n} = \{S_i^1, A_i^1, R_i^1, S_i^2, \ldots, S_i^{T_i}, A_i^T, R_i^T, S_i^{T+1}\}_{i=1}^n$, where the length of trajectory T is assumed to be non-random for simplicity. A policy π is a map from the state space to the action space $\pi : \mathcal{S} \to \mathcal{A}$. The learning goal is to search an optimal policy π^* which maximizes the expected discounted sum of rewards. $V_t^\pi(s) = \mathbb{E}_\pi\left[\sum_{k=1}^\infty \gamma^{k-1} R^{t+k} | S^t = s\right]$ is the value function under a policy π, where $\mathbb{E}_\pi$ is taken by assuming that the system follows a policy π, and the Q-function is defined as $Q_t^\pi(s, a) = \mathbb{E}_\pi\left[\sum_{k=1}^\infty \gamma^{k-1} R^{t+k} | S^t = s, A^t = a\right]$. In a time-homogenous Markov process (Puterman, 2014), $V_t^\pi(s)$ and $Q_t^\pi(s, a)$ do not depend on t. The optimal value function V^* is the unique fixed point of the Bellman operator $\mathcal{B}$, $\mathcal{B}V(s) := \max_a \mathbb{E}_{S^{t+1}\sim\mathbf{P}(s,a)}\left[R^t + \gamma V(S^{t+1}) | S^t = s, A^t = a\right]$. Then $\mathcal{B}V^*(s) = V^*(s)$ for any $s \in \mathcal{S}$. An optimal policy π^* can be obtained by taking the *greedy* action of $Q^*(s, a)$, that is $\pi^*(s) = \arg\max_a Q^*(s, a)$. For the rest of the paper, we use the short notation $\mathbb{E}_{s'|s,a}$ for the conditional expectation $\mathbb{E}_{s'\sim\mathbf{P}(s,a)}$; and $\mathbb{E}_{S^t,A^t,S^{t+1}}$ is short for $\mathbb{E}_{S^t\sim\upsilon, A^t\sim\pi_b(\cdot|S^t), S^{t+1}\sim\mathbf{P}(S^t,A^t)}$, where υ is a some fixed distribution and π_b is some behavior policy.

3 METHODOLOGY

To start with, we first revisit the Bellman optimality equation via a policy explicit view,

$$\mathcal{B}V^*(s) \coloneqq \max_{\pi} \mathbb{E}_{a\sim\pi(\cdot|s),\ S^{t+1}|s,a}\left[R(S^{t+1},s,a)+\gamma V^*(S^{t+1})\right] = V^*(s). \tag{1}$$

To obtain the optimal policy π^* and value function V^*, an optimization idea is to minimize the discrepancy between the two sides of the equation under a L^2 loss. Unfortunately, there are several major challenges when it comes to optimization: (1) **Non-smoothness**: the Bellman operator involves a non-smoothed hard-$\max$ operator, which leads to training instability; (2) **Policy class**: As discussed in Section 1, it is necessary to induce an optimal policy class whose support consists of quasi-optimal sub-regions for reliability, and avoids off-support bias in Figure 2; (3) **Double sampling**: the unknown conditional expectation $\mathbb{E}_{S^{t+1}|s,a}$ is required to be double sampled for obtaining an unbiased sample approximation for $\mathbb{E}_{S^{t+1}|s,a}$. However, this is usually infeasible in real-world environments; (4) **Off-policy data**: directly minimizing the Bellman error is not easy to incorporate off-policy data. To address these issues, we propose a quasi-optimal counterpart of the Bellman equation (1).

3.1 QUASI-OPTIMAL BELLMAN OPERATOR

In this subsection, we aim to tackle the first two challenges. We propose a quasi-optimal counterpart for the Bellman operator $\mathcal{B}$ that simultaneously circumvents the non-smoothness obstacles, and induce a novel policy class which can identify quasi-optimal sub-regions in continuous action spaces.

We leverage the Fenchel representation on the Bellman operator $\mathcal{B}$, that is,

$$\mathcal{B}V^*(s) = \max_{\pi\in\Delta_{\text{convex}}(\mathcal{A})} \int_{a\in\mathcal{A}} \left[Q^*(s,a)\pi(a|s) - \rho(\pi(a|s))\right] da, \tag{2}$$

where $\rho(\cdot)$ is a convex and continuous function. Upon Fenchel transformation (Hiriart-Urruty & Lemaréchal, 2012), for some strongly convex and continuous proximity function $\text{prox}(\pi) : \Delta_{\text{convex}}(\mathcal{A}) \to \mathbb{R}$, a quasi-optimal smoothed counterpart of $\mathcal{B}$ is defined as

$$\mathcal{B}_\mu V^*_\mu(s) = \max_{\pi\in\Delta_{\text{convex}}(\mathcal{A})} \int_{a\in\mathcal{A}} \left[Q^*_\mu(s,a)\pi(a|s) + \mu\text{prox}(\pi(a|s)) - \rho(\pi(a|s))\right] da, \tag{3}$$

where $Q^*_\mu(s,a) = \mathbb{E}_{S^{t+1}|s,a}[R(S^{t+1},s,a)+\gamma V^*_\mu(S^{t+1})]$, and $V^*_\mu(s)$ is the unique fixed point of the quasi-optimal Bellman operator $\mathcal{B}_\mu$. Note that, besides the smoothing purpose, we are also interested in constructing a stochastic optimal policy class that can screen out the non-optimal and sub-optimal actions. Therefore, we further define a special prox function class motivated by the rationale of q-logarithm as $\text{prox}(x) = \log_q(x) := \frac{x(1-x^{q-1})}{q-1}$, where $\int_{a\in\mathcal{A}} \text{prox}(\pi(a|s))da = \frac{1}{q-1}(1-\int_{a\in\mathcal{A}} \pi^q(a|s)da)$ essentially generalize Shannon's entropy (Martins et al., 2020). In this paper, we focus on the setting that $q=2$, and $\rho(\cdot)=0$ which is an unbiased Fenchel representation for $\mathcal{B}$ in (2) and satisfies the convexity and continuity conditions .

Assumption 3.1. *For any policy distribution $\pi \in \Delta_{convex}(\mathcal{A})$, its density is bounded above by a constant, i.e., $\pi(\cdot|s) \le \mathbf{C}$ for all $s \in \mathcal{S}$.*

This assumption avoids some extreme cases where a stochastic policy distribution degenerates to be deterministic. In the following, we disclose several nice properties of the proposed Bellman operator.

Proximal Approximation The operator $\mathcal{B}_\mu$ is a proximal approximation to $\mathcal{B}$. This delivers two messages: firstly, the approximation bias is upper bounded; secondly, the operator $\mathcal{B}_\mu$ is a smoothed substitute for $\mathcal{B}$. In particular, Theorem 3.1 demonstrates that the approximation bias can vanish to zero for small enough μ. In addition, the operator $\mathcal{B}_\mu$ has a differentiable and analytical form (4), which justifies that $\mathcal{B}_\mu$ is a smoothed counterpart of $\mathcal{B}$, see Corollary S.1 in Appendix for details.

Theorem 3.1 (Proximal bias)**.** *Under Assumption 3.1, for any $s \in \mathcal{S}$ and value function V, $\mathcal{B}_\mu V(s) - \mathcal{B}V(s) \in [\mu(1-\mathbf{C}), \mu]$.*

$$\mathcal{B}_\mu V^*_\mu(s) = \mu - \frac{1}{4\mu}\left(\frac{(\int_{a'\in\mathcal{W}_s} Q^*_\mu(s,a')da' - 2\mu)^2}{\sigma(\mathcal{W}_s)} - \int_{a\in\mathcal{W}_s} {Q^*_\mu}^2(s,a)da\right), \tag{4}$$

where $\mathcal{W}_s$ denotes the the support of π^*_μ in (5) for a given state s.

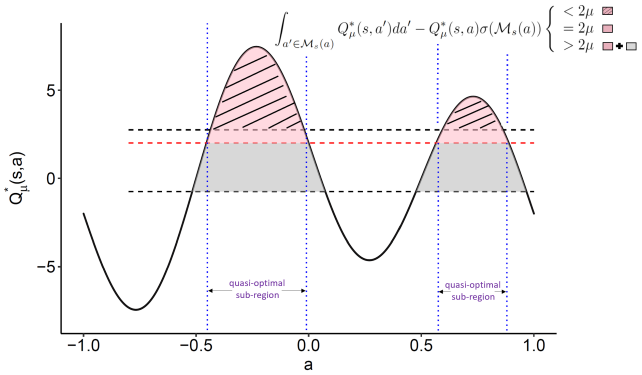

Figure 1: An illustrating example of the quasi-optimal sub-regions. The threshold parameter μ controls the margin between the smallest action-value (red dashed line) in the support set and the globally largest action-value. When μ decreases, the screening criterion is more restrictive and the quasi-optimal sub-regions become more compact.

Quasi-optimal Support Region In addition to the proximal approximation property, another unique and important property of $\mathcal{B}_\mu$ is inducing a policy π^*_μ whose support region contains all the actions with action-value higher than a certain threshold. The induced policy π^*_μ is bridged from the oracle Q-function:

$$\pi^*_\mu(a|s) = \left(\frac{Q^*_\mu(s,a)}{2\mu} - \frac{\int_{a \in \mathcal{W}_s} Q^*_\mu(s,a)da}{2\mu\sigma(\mathcal{W}_s)} + \frac{1}{\sigma(\mathcal{W}_s)} \right)^+, \tag{5}$$

where the support of π^*_μ, i.e., $\mathcal{W}_s := \bigcup_{a \in \mathcal{A}} a \mathbb{1}_{\texttt{screening set}}(a)$ with

$$\texttt{screening set} := \left\{ a \in \mathcal{A} : \int_{a' \in \mathcal{M}_s(a)} Q^*_\mu(s,a')da' - \sigma(\mathcal{M}_s(a)) Q^*_\mu(s,a) > 2\mu \right\}, \tag{6}$$

$$\mathcal{M}_s(a) := \bigcup_{a' \in \mathcal{A}} a' \mathbb{1}_{\{Q^*_\mu(s,a') > Q^*_\mu(s,a)\}}(a'). \tag{7}$$

This mechanism allows us to identify multiple sub-regions in the entire action space which only contains near-optimal actions, and weed out the sub-optimal and non-optimal support regions. Note that, the identified sub-region might not be joint in general, which is beneficial to the situation that the true Q-function has multiple modes. The screening set in (6) indicates that the threshold parameter μ not only controls the degree of smoothness, but also determines how the quasi-optimal region behaves and controls the screening intensity, as shown in Figure 1.

Remark 3.1. *As indicated in* (5)*, the induced policy is optimal with respect to the quasi-optimal Bellman operator, and it is bridged directly from $Q^*_\mu(s,a)$. This distinguishes our approach from other methods such as Haarnoja et al. (2018b); Lee et al. (2019), which use an independent pre-specified policy class without optimality guarantee, and model the policy and value function completely separate.*

3.2 q-Gaussian Policy Distribution

In this section, we bridge the induced policy distribution π^*_μ to an explainable q-Gaussian distribution. The q-Gaussian distribution is less favored for heavy tails, which makes it widely used in practice to model the effect of external stochasticity (d'Onofrio, 2013). In continuous actions problems, e.g., medical dose suggestion, the q-Gaussian distribution is a more suitable choice than the Gaussian distribution for policy modeling, since it can filter out non-optimal and risky dose levels, i.e., too high or too low dosage.

Motivated by the fact that the induced policy π^*_μ is feasible to identify quasi-optimal support sub-regions, and q-Gaussian policy distribution can realize bounded support, we conjectured that the q-Gaussian policy distribution might be recovered from the induced policy π^*_μ. Fortunately, the q-Gaussian policy distribution is indeed a special case of the induced policy if $Q^*_\mu(s,a)$ is a concavely quadratic function with respect to the action a. We illustrate this phenomenon in Theorem 3.2.

Theorem 3.2. *Suppose $Q^*_\mu(s,a)$ is a concavely quadratic function over $a \in \mathcal{A}$, i.e., $Q^*_\mu(s,a) = -\alpha_1(s)a^2 + \alpha_2(s)a + \alpha_3(s) := Q^{\mathcal{N}}_\mu(s,a)$ where $\alpha_1(s), \alpha_2(s), \alpha_3(s)$ are functions over $s \in \mathcal{S}$*

and $\alpha_1(s) > 0$ for all s, then the induced policy distribution $\pi^*_\mu(\cdot|s)$ would follow a q-Gaussian distribution with a density function

$$\pi^*_\mu(a|s) = \left(\frac{\alpha_1(s)}{2\mu}\left(a + \frac{\alpha_2(s)}{2\alpha_1(s)}\right)^2 - \frac{3}{2}\left(\frac{\alpha_1(s)}{12\mu}\right)^{\frac{1}{3}}\right)^+ := \pi^{\mathcal{N}}_\mu(a|s), \tag{8}$$

and a closed-form quasi-optimal support region

$$\mathcal{W}_s = \left[\frac{\alpha_2(s) - (12\alpha_1^2(s)\mu)^{\frac{1}{3}}}{2\alpha_1(s)}, \frac{\alpha_2(s) + (12\alpha_1^2(s)\mu)^{\frac{1}{3}}}{2\alpha_1(s)}\right] := \mathcal{W}^{\mathcal{N}}_s. \tag{9}$$

The policy distribution $\pi^{\mathcal{N}}_\mu(\cdot|s)$ behaves as a affine transformation of the standard q-Gaussian distribution with mean $-\frac{\alpha_2(s)}{2\alpha_1(s)}$, where the maximum action-value attains, i.e., $Q^{\mathcal{N}}_\mu(s, -\frac{\alpha_2(s)}{2\alpha_1(s)}) = \arg\max_{a\in\mathcal{A}} Q^{\mathcal{N}}_\mu(s,a)$. Note that the width of the quasi-optimal region is $\frac{(12\alpha_1^2(s)\mu)^{\frac{1}{3}}}{\alpha_1(s)}$ determined by the threshold parameter μ. The actions within the region $\mathbb{R} \setminus \mathcal{W}^{\mathcal{N}}_s$ are discriminated as the non-optimal and would be assigned with zero probability densities. For a small μ, i.e., strong screening intensity, a narrow region would be identified as the quasi-optimal, which yields a relatively conservative action recommendation. In contrast, with a large μ, more actions are included in the support. In an extreme case, $\mathcal{W}^{\mathcal{N}}_s$ degenerates to $\mathbb{R}$ as $\mu \to \infty$. In Theorem 4.1 of Section 4, we investigate how the intensity of μ affects the induced policy distribution is formally.

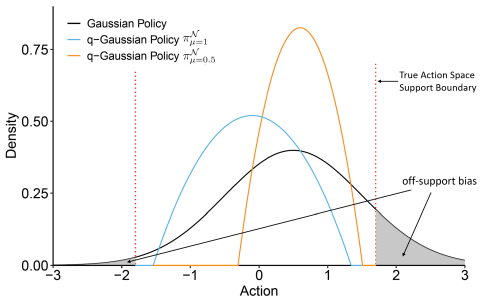

Figure 2: An illustrating example of bounded action space and q-Gaussian policy distribution. The Gaussian policy assigns non-zero probabilities density to all actions, even for those actions outside of the true action space support boundary. This causes the off-support bias. In contrast, the q-Gaussian policy relieves such off-support bias blessed by the boundedness of the quasi-optimal region.

So far, we have obtained the closed-form representations for the general policy $\pi^*_\mu(\cdot|s)$ and q-Gaussian policy $\pi^{\mathcal{N}}_\mu$. However, how to make a policy estimation remains unknown. Indicated by the challenges in Section 3, we need to address the double sampling issue and utilize off-policy data in optimization. Both challenges cannot be easily solved by minimizing the Bellman error. Fortunately, the kernel embedding helps us to bypass the difficulties.

3.3 Kernel Embedding on Quasi-optimal Error

In this subsection, we introduce the quasi-optimal learning framework for solving the induced policy π^*_μ. First, we establish a stationary equation in Theorem 3.3. This helps to incorporate off-policy data. Then we leverage the idea of the kernel embedding (Gretton et al., 2012) to obtain an unbiased empirical loss without the double sampling issue.

Theorem 3.3 (Stationarity equation)**.** *Let V^*_μ be a fixed point of the quasi-optimal Bellman operator $\mathcal{B}_\mu$, and π^*_μ is the induced policy in* (5)*. For any $s \in \mathcal{S}, a \in \mathcal{A}$, and $\mu \in (0,\infty)$, the pair (V^*_μ, π^*_μ) satisfies the following equation:*

$$\mathbb{E}_{S^{t+1}|s,a}\left[R(S^{t+1}, s, a) + \gamma V_\mu(S^{t+1})\right] - \mu\text{prox}^\circ(\pi_\mu(a|s)) - \eta(s) + \varpi(s,a) = V_\mu(s). \tag{10}$$

Here $\text{prox}^\circ(x) = 2x - 1$*, $\eta(s) : \mathcal{S} \to [-\mu\mathbf{C}, 0]$ and $\varpi(s,a) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^+$ are Lagrange multipliers in which $\varpi(s,a)\cdot\pi_\mu(a|s) = 0$. We refer the discrepancy between the two sides in* (10) *as "quasi-optimal error".*

The equation (10) connects the quasi-optimal value function V^*_μ and policy function π^*_μ along with any state-action pairs, implying it can be directly solved using off-policy data, i.e., the transition samples. This overcomes distribution shift and no adjustment for mismatched distribution is needed.

Min-max Optimization One way to solve the equation (10) is minimizing the quasi-optimal error under a L^2 loss function. Unfortunately, the double sampling issue would still appear if replacing the unknown $\mathbb{E}_{S^{t+1}|s,a}[R(S^{t+1}, s, a) + \gamma V_\mu(S^{t+1})]$ in the quasi-optimal error by its one-sample

Algorithm 1 Quasi-optimal Learning in Continuous Action Spaces

1: **Input** observed transition pairs data $\{(S_i^t, A_i^t, R_i^t, S_i^{t+1}) : t = 1, ..., T\}_{i=1}^n$.
2: **Initialize** the parameters of interests $\theta = (\theta_1^0, \theta_2^0)$, $\xi = (\xi_1^0, \xi_2^0)$, the mini-batch size n_0, the learning rate α_0, the prox parameter μ, the kernel bandwidth bw_0, and the stopping criterion ε.
3: **For iterations** $j = 1$ to k
4: Randomly sample a mini-batch $\{(S_i^t, A_i^t, R_i^t, S_i^{t+1}) : t = 1, ..., T\}_{i=1}^{n_0}$.
5: Decay the learning rate $\alpha_j = \mathcal{O}(j^{-1/2})$.
6: Compute stochastic gradients with respect to θ and ξ: $\bar{\nabla}_\theta = \mathbb{P}_{n_0}\widehat{\nabla}_\theta\widehat{\mathcal{L}_U}$ and $\bar{\nabla}_\xi = \mathbb{P}_{n_0}\widehat{\nabla}_\xi\widehat{\mathcal{L}_U}$.
7: Update the parameters of interest as $\theta^j \leftarrow \theta^{j-1} - \alpha_j\bar{\nabla}_\theta\widehat{\mathcal{L}_U},\ \xi^j \leftarrow \xi^{j-1} - \alpha_j\bar{\nabla}_\xi\widehat{\mathcal{L}_U}$.
8: Stop if $\|(\theta^j, \xi^j) - (\theta^{j-1}, \xi^{j-1})\| \leq \varepsilon$.
9: **Return** $\widehat{\theta} \leftarrow \theta^j, \widehat{\xi} \leftarrow \xi^j$.

bootstrapping counterpart $R^t + \gamma V_\mu(S^{t+1})$. Alternatively, inspired by the average Bellman error (Jiang et al., 2017), we propose to minimize a weighted average quasi-optimal error, and the unwanted conditional variance of the bootstrapping counterpart under L^2 loss could vanish. We define the loss $\mathcal{L}(V_\mu, \pi_\mu, \eta, \varpi, u)$ as

$$\mathbb{E}_{S^t, A^t, S^{t+1}}\left[u\left(S^t, A^t\right) \cdot \left(\mathcal{G}_{V_\mu,\pi_\mu}\left(S^t, A^t, S^{t+1}\right) - \eta(S^t) + \varpi(S^t, A^t) - V_\mu(S^t)\right)\right],$$

where $\mathcal{G}_{V_\mu,\pi_\mu}(s, a, s') := R(s', s, a) + \gamma V_\mu(s') - \mu\text{prox}^\circ(\pi_\mu(a|s))$ and $u(\cdot) : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$ is a bounded function in L^2 space $L^2(C_0) := \{u \in L^2 : \|u\|_{L^2} \leq C_0\}$. Essentially, the weight function u is to fit the discrepancy of (10) and promotes the sample points with large quasi-optimal errors.

As $\mathcal{L}(V_\mu^*, \pi_\mu^*, \eta, \varpi, u) = 0$ holds for any u function, this leads to a minimax optimization:

$$\min_{V_\mu,\pi_\mu,\eta,\varpi} \max_{h \in L^2(C_0)} \mathcal{L}^2(V_\mu, \pi_\mu, \eta, \varpi, u). \tag{11}$$

Kernel Representation Solving the minimax optimization problem (11) is unstable, and it is also intractable due to the difficulty for the representation of u in L^2 space. Therefore, we further consider modeling the weight function u in a bounded reproducing kernel Hilbert space (RKHS) associated with a positive define reproducing kernel K, i.e., $H_{\text{RKHS}}(C_0) := \{u \in H_{\text{RKHS}} : \|u\|_K \leq C_0\}$. This leads to a closed-form representation of the inner optimization maximizer similar to the maximum mean discrepancy (Gretton et al., 2012). The detailed derivation is provided in Theorem S.2 in Appendix. Upon this, the minimax optimization is reduced to only minimizing the loss

$$\mathcal{L}_U = \mathbb{E}_{S^t,\tilde{S}^t,A^t,\tilde{A}^t,S^{t+1},\tilde{S}^{t+1}}[\Lambda_{V_\mu,\pi_\mu}(S^t, A^t, S^{t+1})K(S^t, A^t; \tilde{S}^t, \tilde{A}^t)\Lambda_{V_\mu,\pi_\mu}(\tilde{S}^t, \tilde{A}^t, \tilde{S}^{t+1})] \tag{12}$$

where $\Lambda_{V_\mu,\pi_\mu}(s, a, s') := \mathcal{G}_{V_\mu,\pi_\mu}(s, a, s') - \eta(s) + \varpi(s, a) - V_\mu(s)$ and $(\tilde{S}^t, \tilde{A}^t, \tilde{S}^{t+1})$ is an independent copy of transition pair (S^t, A^t, S^{t+1}).

It observes that the loss $\mathcal{L}_U$ is symmetric and kernel represented. This motivates us to use an unbiased U-statistic estimator to obtain the sample loss. Given the observed data, $\mathcal{D}_{1:n}$, with n trajectories of length T, we can use a trajectory-based U-statistic estimator to capture the within-trajectory loss, thus the total loss $\mathcal{L}_U$ can be aggregated as the empirical mean of n i.i.d. within trajectory loss:

$$\min_{V_\mu,\pi_\mu,\eta,\varpi} \widehat{\mathcal{L}_U} = \mathbb{P}_n\binom{T}{2}\sum_{1\leq j\neq k\leq T}[\Lambda_{V_\mu,\pi_\mu}(S_i^j, A_i^j, S_i^{j+1})K(S^j, A^j; S^k, A^k)\Lambda_{V_\mu,\pi_\mu}(S_i^k, A_i^k, S_i^{k+1})]$$
$$\text{s.t.} \quad \varpi(a|s) \geq 0, \pi_\mu(a|s) \cdot \varpi(a|s) = 0 \ \text{ and } \eta(s) \in [-\mu\mathbf{C}, 0] \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}. \tag{13}$$

The sample loss $\widehat{\mathcal{L}_U}$ is unbiased and consistent with the population loss $\mathcal{L}_U$. The consistency is justified in Theorem 4.2 via examining the tail behavior of $\widehat{\mathcal{L}_U}$. In essence, solving the equation (13) is a computationally intensive non-linear programming problem. Alternatively, we convert the constrained problem to an unconstrained problem by restricting the Lagrange multipliers. It then can be solved by an unconstrained true gradient algorithm, i.e., Algorithm 1. See Appendix for details.

4 THEORY

In this section, we study the theoretical properties of the proposed method. First, we study some general properties of the proposed quasi-optimal Bellman operator, given in Proposition S.1 and S.2

of Appendix. In Theorem 4.1, we disclose the effect of the intensity of prox parameter μ on the induced optimal policy distribution. Moreover, a non-asymptotic concentration bound is established in Theorem 4.2, showing the consistency and measuring the rate of convergence of $\widehat{\mathcal{L}_U}$ to $\mathcal{L}_U$. Further, the overall performance error of the algorithm is given in Theorem 4.3, where the performance error is decomposed as the four sources. Finally, we show that the proposed quasi-optimal learning is a convergent algorithm. Before we present the theoretical results, we introduce some assumptions on the boundedness condition of the MDP and the sample trajectory properties, respectively.

Assumption 4.1. *The reward function $R(s', s, a)$ is uniformly bounded, i.e, $\|R(\cdot)\|_\infty \leq R_{\max}$.*

Assumption 4.2. *Suppose $\{S^t, A^t\}_{t\geq 1}$ is a strictly stationary and exponentially β-mixing sequence with a mixing coefficient $\beta(m) \lesssim \exp(-\delta_1 m)$ for $m \geq 1$. We further assume that the behavior policy π_b, which is used to collect the offline data $\mathcal{D}_{1:n}$, satisfies that $\min_{a\in\mathcal{A}, s\in\mathcal{S}} \pi_b(a|s) > 0$.*

Theorem 4.1 (Policy Adaptability)**.** *Under Assumption 4.1, for all $s \in \mathcal{S}$, the quasi-optimal policy distribution $\pi^*_\mu(\cdot|s)$ degenerates to a uniform distribution over $\Delta(\mathcal{A})$ as $\mu \to \infty$, and $\pi^*_\mu(\cdot|s)$ concentrates in a point mass as $\mu \to 0$ and $\mathbf{C} \to \infty$.*

Theorem 4.1 formally investigates the effect of μ on $\pi^*_\mu(\cdot|s)$. In an extreme case that $\mu \to 0, \mathbf{C} \to \infty$, only the action maximizing $Q^*_\mu(s, a)$ would be included in the quasi-optimal region. In the following, we establish a non-asymptotic concentration inequality for the empirical loss in the non-i.i.d. case.

Theorem 4.2. *For any $\mu \in (0, \infty)$ and $\epsilon > 0$, under Assumptions 4.1-4.2, we have ϵ-divergence of $|\widehat{\mathcal{L}_U} - \mathcal{L}_U|$ bounded in probability, i.e.,*

$$\mathbb{P}(|\widehat{\mathcal{L}_U} - \mathcal{L}_U| > \epsilon) \leq C_1 \exp\left(- \frac{\epsilon^2 T - C_2 \epsilon M_{\max}^2 \sqrt{T}}{M_{\max}^2 + (\frac{\epsilon}{2} - \frac{C_2 M_{\max}^2}{\sqrt{T}}) \log T \log\log(T)} \right) + C_3 \exp\left(\frac{-n\epsilon^2}{M_{\max}^4} \right),$$

where C_1, C_2 and C_3 are some constants depending on δ_1 respectively, and $M_{\max} = \frac{4}{1-\gamma} R_{\max} + \mu \mathbf{C}$.

Theorem 4.2 implies that $\widehat{\mathcal{L}_U}$ is a consistent estimator to $\mathcal{L}_U$, and thus avoiding the double sampling issue. Note that the concentration bound is sharper than the bound established in Chakrabortty & Kuchibhotla (2018) since we utilize a novel temporal correlatedness structure to decompose the U-statistic. We now analyze the performance error between the finite sample learner and true solution, which can be decomposed into four source errors.

Theorem 4.3. *Under Assumption 4.1-4.2, let $V_\mu^{\theta_1,k}$ be the optimizer from Algorithm 1 and V^* is the optimal value function and $\kappa_{\min}$ be the smallest eigenvalue corresponding to an orthonormal basis of $L^2(\mathcal{S} \times \mathcal{A})$ space. With probability $1 - \delta$, the performance error is upper bounded by*

$$\|\widehat{V}_\mu^{\theta_1,k} - V^*\|_{L^2}^2 \leq \underbrace{\frac{C_4}{\kappa_{\min}(1-\gamma)^2} \left(\sqrt{\frac{C_5 D_{\text{P-dim}} \log\left(\frac{8C_4}{\delta}\right)}{n}} + \sqrt{\frac{2\left(\frac{\bar{\Delta}}{\delta_1} \vee 1\right)\bar{\Delta}}{C_6 \lfloor T/2 \rfloor}} \right)}_{\text{generalization error}} +$$

$$\underbrace{C_7 \frac{\mu^2(\mathbf{C} + |1 - \mathbf{C}| \vee 1)^2}{(1-\gamma)^2}}_{\text{proximal bias}} + \underbrace{C_8 \left\| \widehat{V}_\mu^{\theta_1} - \widehat{V}_\mu^{\theta_1,k} \right\|_{L^2}^2}_{\text{optimization error}} + \epsilon_{\text{approximation error}},$$

where $\bar{\Delta} = \frac{D_{\text{P-dim}} \log\lfloor T/2 \rfloor}{2} + \log(\frac{e}{\delta}) + \log^+\left(\frac{C_5 C_6^{D_{\text{P-dim}}}}{2}\right)$, $D_{\text{P-dim}} = \text{P-dim}(\Theta_1) + \text{P-dim}(\Theta_2) + \text{P-dim}(\Xi_1) + \text{P-dim}(\Xi_2)$, and $C_4, ..., C_8$ are some constants. Here P-dim$(\cdot)$ denotes the pseudo-dimension operator (Györfi, 2010), and $\Theta_1, \Theta_2, \Xi_1$ and Ξ_2 are function spaces for V_μ, π_μ, ϖ and η, respectively. The error $\epsilon_{\text{approximation error}}$ is due to the parametrization $(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \varpi^{\xi_1}, \eta^{\xi_2})$ on $(V_\mu, \pi_\mu, \varpi, \eta)$.

The above sample complexity bound gives an insight into the performance error of the proposed algorithm. The generalization error $\varepsilon_{\text{gerr}} = \mathcal{O}(1/\sqrt{T})$ if n is as the same order of T, the proximal bias $\varepsilon_{\text{prox}} = \mathcal{O}(\mu^2)$ and the optimization error $\varepsilon_{\text{optim}} = \mathcal{O}(1/k)$ for k iterations. Although the prox function introduces a proximal bias in the quasi-optimal Bellman operator $\mathcal{B}_\mu$, it leads to a smoothed approximation for $\mathcal{B}$. There exists a trade-off between the proximal bias and approximation error. As the increase of μ, it enlarges the proximal bias but decreases the approximation error since true function space becomes more smoothed and easy for function approximation. On the other hand, a small μ leads to a small proximal bias but a relatively large approximation error.

Theorem 4.4. *Suppose $\widehat{\mathcal{L}_U}$ in Algorithm 1 is differentiable, but not necessarily convex, and its gradient $\nabla\widehat{\mathcal{L}_U}(\theta,\xi)$ is $M_{\mathcal{L}}$-Lipschitz and $\mathrm{Var}(\bar{\nabla}_\theta+\bar{\nabla}_\xi)\leq\sigma_0^2$. And suppose that the learning rate $\{\alpha_j\}$ are set to $\alpha_j=\min\left\{\frac{2}{M_{\mathcal{L}}},\frac{\Lambda}{\sigma_0\sqrt{j}}\right\}$ for some $\Lambda\geq 0$ and ε is sufficient small. Let $k=\widetilde{k}$ with $\mathbb{P}(\widetilde{k}=j)=\frac{\alpha_j(2-M_{\mathcal{L}}\alpha_j)}{\sum_{j=1}^{k}(\alpha_j(2-M_{\mathcal{L}}\alpha_j))}$ for $j=1,\ldots,k_\diamond$. Then, if $(\widehat{\theta},\widehat{\xi})$ is the optimization solution and (θ^1,ξ^1) is the first step solution, we have*

$$\left\|\nabla\widehat{\mathcal{L}_U}(\widehat{\theta},\widehat{\xi})\right\|_{L^2}^2\leq 2M_{\mathcal{L}}\Big(\widehat{\mathcal{L}_U}(\theta^1,\xi^1)-\min_{\theta,\xi}\widehat{\mathcal{L}_U}(\theta,\xi)\Big)\Big(\frac{M_{\mathcal{L}}}{k_\diamond}+\frac{\sigma_0}{M_{\mathcal{L}}\Lambda\sqrt{k_\diamond}}\Big)+\frac{\Lambda\sigma_0M_{\mathcal{L}}}{\sqrt{k_\diamond}},$$

Theorem 4.4 implies that the quasi-optimal learning algorithm is converges to a stationary point with a sub-linear rate $\mathcal{O}(1/\sqrt{k_\diamond})$ even if the empirical loss is non-convex. The property serves as a basis for applying non-linear function approximation still with convergent guarantees, e.g., neural nets or random forests. Theorem 4.4 is adapted from Corollary 2.2 in Ghadimi & Lan (2013) when a decay learning rate and a vanished Euclidean stopping criterion as in our algorithm. The convergence of Algorithm 1 is blessed by our unbiased stochastic gradient estimator.

5 Related Works

In this work, we propose a provably convergent and sample efficient policy optimization algorithm in off-policy settings. Broadly, our work is related to value function learning (Sutton & Barto, 2018). The domain approaches include fixed-point iterations (Szepesvári, 2010), Bellman residual minimization (Antos et al., 2008; Farahmand et al., 2016; Chen & Jiang, 2019) and projected Bellman error minimization (Maei et al., 2010; Macua et al., 2014; Dai et al., 2018).

Algorithmically, our work is also related to the entropy-regularized reinforcement learning algorithms (Rawlik et al., 2012; Haarnoja et al., 2017), but these works are fundamentally different from ours. Our formulation is motivated by constructing a proximal counterpart of the Bellman operator, which serves as a basis for the latter quasi-oracle learning algorithm. Besides, the major drawback of the existing algorithms (Lee et al., 2018b; Chow et al., 2018; Vieillard et al., 2020) is the lack of theoretical guarantees when accompanied by function approximation. It is not clear whether the algorithm is convergent, generalizable, and consistent. In contrast, our algorithm is thoroughly examined on both theoretical and empirical fronts. Nachum et al. (2017); Chow et al. (2018) exploit an analogous stationarity condition as in (3.3), and propose PCL algorithms to minimize the upper bound of the discrepancy errors, which is biased and encounters the double sampling issue. However, our work leverages the kernel embedding to solve a stationarity equation without requiring double sampling, and is provably consistent. Unlike our algorithm, the algorithms in continuous control problems, e.g., (Haarnoja et al., 2018b; Nachum et al., 2018b; Lee et al., 2019) do not check the policy optimality, but separately model a pre-specified policy class. This may introduce an additional bias if the pre-specified policy class is misspecified.

Our approach also exemplifies more recent efforts that aim to learn optimal policy with continuous actions (Lillicrap et al., 2015). One of our key innovations is to develop a policy class that can identify quasi-optimal sub-regions and the induced policy has a closed-form regarding value function. This distinguishes us from the approaches, e.g., (Silver et al., 2014; Mnih et al., 2016; Kumar et al., 2019; 2020). These methods typically require prior knowledge to determine pre-specified policy class and commonly use Gaussian family distribution, but unfortunately facing the risk from off-support bias.

6 Experiments

In this section, we evaluate our proposed method on simulated and real environments. We compare our method to the state-of-the-art methods including Deep Deterministic Policy Gradient (Lillicrap et al. (2015), DDPG), Soft Actor-Critic (Haarnoja et al. (2018a), SAC), Bootstrapping error Accumulation Reduction ((Kumar et al., 2019), BEAR), and Greedy-GQ learning (Ertefaie & Strawderman, 2018).

Numerical Experiments The four environments are simulated to mimic the real environments for continuous treatment applications. We consider bounded action spaces in Environment I and II, and unbounded action support in Environment III and IV. A unimodal reward function is used to mimic the true environment that the optimal dosage is unique and a high dosage leads to excessive toxicity while a lower dosage is ineffective (Zang et al., 2014). In addition, a more complex multimodal

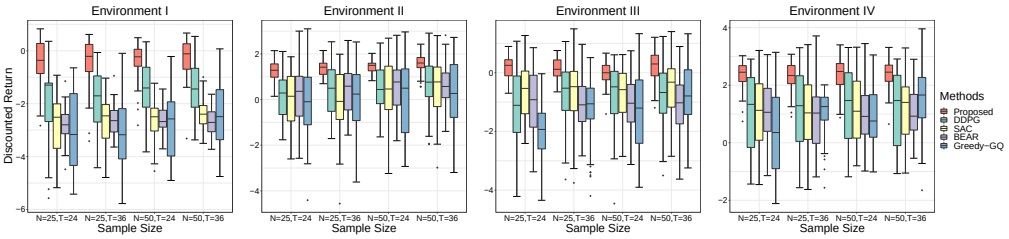

Figure 3: The boxplot of the discounted return over 50 repeated experiments.

reward function is used in Environment II and IV, respectively. The detailed data generating process and experiment details are deferred to Section C in Appendix.

Figure 3 shows that our proposed method outperforms competing methods with relatively small variance. This mainly benefits from the identification of the quasi-optimal region, which guarantees the suggested action is near-optimal, hence improving the performance. In comparison, SAC and BEAR use a Gaussian policy and assign non-negligible positive densities to all actions, even for the non-optimal ones, and thus the model performance is damaged. In addition, in Environment I and II with bounded action support, the competing methods are affected by an off-support bias which lowers their discounted return. Meanwhile, the two deterministic policy algorithms, i.e., DDPG and Greedy-GQ learning, suffer from high variance due to the sub-optimality issue of deterministic policy. Also, note that our algorithm achieves stable performance in small sample size settings, which is blessed by the smoothness and optimization-friendly of our algorithm. This is promising as the limited data is usual in medical applications.

Real Data: A Ohio Type 1 Diabetes Case Study Ohio type 1 diabetes (OhioT1DM) dataset (Marling & Bunescu, 2020), which contains 2 cohorts of patients with Type-1 diabetes, each patient with 8 weeks of life-event data including health status measurements and insulin injection dosage. Clinicians are pretty interested in adjusting the insulin injection dose level (Marling & Bunescu, 2020) based on personalized health status to maintain the glucose level within a certain range for safe dose suggestions. As each individual has dramatically distinctive glucose dynamics, We follow Zhu et al. (2020) to regard each patient data as an independent dataset, and the data from each day as a trajectory. The state variables are health status measurements, and the action space is a bounded insulin dose range. The glycemic index is regarded as a reward function to measure the goodness of dose suggestion. See more details of the experiment setup in Appendix. As shown in Table 1, the proposed method achieves the best performance among all methods, except for patient ID 567 and 570. The proposed method mitigates the off-support bias in this bounded dosage space and outperforms the competing methods. This finding is consistent with the results in the simulations and demonstrates the potential of our method in continuous action spaces.

Table 1: The discounted return for the policy improvement based on 50 repeated experiments.

Patient ID	Proposed	DDPG	SAC	BEAR	Greedy-GQ
540	$\mathbf{18.62 \pm 0.63}$	14.05 ± 2.34	14.22 ± 1.18	13.70 ± 0.86	15.54 ± 2.41
544	$\mathbf{10.99 \pm 0.72}$	7.51 ± 1.50	7.47 ± 2.51	5.94 ± 0.81	6.31 ± 2.94
552	$\mathbf{6.26 \pm 0.42}$	4.82 ± 0.52	5.69 ± 0.98	3.56 ± 0.60	4.13 ± 1.78
567	29.92 ± 1.51	$\mathbf{30.01} \pm 2.04$	27.34 ± 2.23	$29.63 \pm \mathbf{1.19}$	24.82 ± 3.80
584	$\mathbf{32.10 \pm 0.78}$	27.02 ± 2.03	23.29 ± 3.24	26.85 ± 1.30	17.78 ± 3.17
596	$\mathbf{5.52} \pm 1.12$	$4.13 \pm \mathbf{0.82}$	4.45 ± 0.95	2.74 ± 0.99	2.68 ± 1.75
559	$\mathbf{24.05} \pm 1.39$	20.14 ± 1.21	19.60 ± 1.23	$19.62 \pm \mathbf{0.75}$	17.30 ± 1.61
563	$\mathbf{11.55 \pm 0.56}$	8.44 ± 0.85	9.26 ± 0.71	8.42 ± 0.75	9.23 ± 1.46
570	$25.01 \pm \mathbf{0.79}$	24.53 ± 1.41	$\mathbf{26.05} \pm 0.81$	25.79 ± 0.83	22.79 ± 1.56
575	$\mathbf{15.55 \pm 1.02}$	10.39 ± 1.27	8.78 ± 1.41	10.21 ± 1.03	5.68 ± 2.84
588	$\mathbf{18.61 \pm 0.69}$	14.20 ± 1.32	13.52 ± 1.54	12.03 ± 0.85	10.03 ± 3.10
591	$\mathbf{15.41} \pm 1.04$	12.32 ± 0.62	$11.88 \pm \mathbf{0.59}$	12.75 ± 0.73	10.73 ± 1.70

7 Conclusions

We introduce a novel quasi-oracle learning algorithm for continuous action allocations, which is particularly useful in determining the dose level when developing medical treatment regimes. The quasi-optimal learning algorithm is provably convergent in off-policy cases, and a PAC bound is provided to analyze its sample complexity. The promising results arise some interesting directions for future works, including extending the framework to the online setting with streaming data.

8 REPRODUCIBILITY STATEMENT

We include the reproducible code for all the experiments and the guideline for access to the Ohio Type I Diabetes dataset in an anonymous GitHub link `https://anonymous.4open.science/r/Quasi-optimal-Learning-with-Continuous-Treatments-9B88`. The experiment details are provided in Appendix for reproducible purpose. All proofs of main theorems and addition theorems are included in Appendix.

9 ETHICS STATEMENT

The proposed method aims at finding optimal policy in continuous action space, with a special focus on medical applications. We believe the proposed work has potential applications in sequential treatment dose suggestion, e.g., managing diabetes through insulin injection. To the best of our knowledge, the proposed work would not lead to any significant negative social impact.

REFERENCES

András Antos, Csaba Szepesvári, and Rémi 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.

Miguel Angel Arcones and Bin Yu. Central limit theorems for empirical andu-processes of stationary mixing sequences. *Journal of Theoretical Probability*, 7(1):47–71, 1994.

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

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334, 1997.

Hengrui Cai, Chengchun Shi, Rui Song, and Wenbin Lu. Deep jump learning for off-policy evaluation in continuous treatment settings. *Advances in Neural Information Processing Systems*, 34:15285–15300, 2021.

Abhishek Chakrabortty and Arun Kumar Kuchibhotla. Tail bounds for canonical u-statistics and u-processes with unbounded kernels. 2018.

Guanhua Chen, Donglin Zeng, and Michael R Kosorok. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 111(516):1509–1521, 2016.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Po-Wei Chou, Daniel Maturana, and Sebastian Scherer. Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution. In *International conference on machine learning*, pp. 834–843. PMLR, 2017.

Yinlam Chow, Ofir Nachum, and Mohammad Ghavamzadeh. Path consistency learning in tsallis entropy regularized mdps. In *International Conference on Machine Learning*, pp. 979–988, 2018.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pp. 1125–1134. PMLR, 2018.

Alberto d'Onofrio. *Bounded noises in physics, biology, and engineering*. Springer, 2013.

Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2658–2667. PMLR, 2020.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pp. 1329–1338. PMLR, 2016.

Ashkan Ertefaie and Robert L Strawderman. Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika*, 105(4):963–977, 2018.

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

László Györfi. A distribution-free theory of nonparametric regression, 2010.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pp. 1352–1361. PMLR, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870. PMLR, 2018a.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018b.

Fang Han. An exponential inequality for u-statistics under mixing conditions. *Journal of Theoretical Probability*, 31(1):556–578, 2018.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer Science & Business Media, 2012.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pp. 409–426. Springer, 1994.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Eric B Laber, Daniel J Lizotte, Min Qian, William E Pelham, and Susan A Murphy. Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, 8(1): 1225, 2014.

Kyowoon Lee, Sol-A Kim, Jaesik Choi, and Seong-Whan Lee. Deep reinforcement learning in continuous action spaces: a case study in the game of simulated curling. In *International conference on machine learning*, pp. 2937–2946. PMLR, 2018a.

Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Sparse markov decision processes with causal sparse tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018b.

Kyungjae Lee, Sungyub Kim, Sungbin Lim, Sungjoon Choi, and Songhwai Oh. Tsallis reinforcement learning: A unified framework for maximum entropy reinforcement learning. *arXiv preprint arXiv:1902.00137*, 2019.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Daniel J Luckett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, and Michael R Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 2019.

Daniel J Luckett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, and Michael R Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 115(530):692–706, 2020.

Sergio Valcarcel Macua, Jianshu Chen, Santiago Zazo, and Ali H Sayed. Distributed policy evaluation under multiple behavior strategies. *IEEE Transactions on Automatic Control*, 60(5):1260–1274, 2014.

Hamid R Maei, Csaba Szepesvári, Shalabh Bhatnagar, and Richard S Sutton. Toward off-policy learning control with function approximation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 719–726, 2010.

Cindy Marling and Razvan Bunescu. The ohiot1dm dataset for blood glucose level prediction: Update 2020. *KHD@ IJCAI*, 2020.

André Martins, António Farinhas, Marcos Treviso, Vlad Niculae, Pedro Aguiar, and Mario Figueiredo. Sparse and continuous attention mechanisms. *Advances in Neural Information Processing Systems*, 33:20989–21001, 2020.

Florence Merlevède, Magda Peligrad, Emmanuel Rio, et al. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, pp. 273–292. Institute of Mathematical Statistics, 2009.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2775–2785, 2017.

Ofir Nachum, Mohammad Norouzi, George Tucker, and Dale Schuurmans. Smoothed action value functions for learning gaussian policies. In *International Conference on Machine Learning*, pp. 3692–3700. PMLR, 2018a.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Trust-pcl: An off-policy trust region method for continuous control. In *International Conference on Learning Representations*, 2018b.

David Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. *Proceedings of Robotics: Science and Systems VIII*, 2012.

Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Takuma Seno and Michita Imai. d3rlpy: An offline deep reinforcement learning library. *arXiv preprint arXiv:2111.03788*, 2021.

Chengchun Shi, Shengxing Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 2021.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. PMLR, 2014.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

Nino Vieillard, Olivier Pietquin, and Matthieu Geist. Munchausen reinforcement learning. *Advances in Neural Information Processing Systems*, 33:4235–4246, 2020.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Ken-ichi Yoshihara. Limiting behavior of u-statistics for stationary, absolutely regular processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 35(3):237–252, 1976.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pp. 94–116, 1994.

Yong Zang, J Jack Lee, and Ying Yuan. Adaptive designs for identifying optimal biological dose for molecularly targeted agents. *Clinical Trials*, 11(3):319–327, 2014.

Wenzhuo Zhou, Ruoqing Zhu, and Annie Qu. Estimating optimal infinite horizon dynamic treatment regimes via pt-learning. *arXiv preprint arXiv:2110.10719*, 2021a.

Wenzhuo Zhou, Ruoqing Zhu, and Donglin Zeng. A parsimonious personalized dose-finding model via dimension reduction. *Biometrika*, 108(3):643–659, 2021b.

Liangyu Zhu, Wenbin Lu, and Rui Song. Causal effect estimation and optimal dose suggestions in mobile health. In *International Conference on Machine Learning*, pp. 11588–11598. PMLR, 2020.

"Quasi-optimal Learning with Continuous Treatments"

Appendices

A Technical Proofs **15**

A.1 Proofs on Constructing Quasi-Optimal Bellman Operator 15

A.1.1 Proof of Theorem S.1 15

A.1.2 Proof of Corollary S.1 16

A.1.3 Proof of Theorem 3.1 16

A.1.4 Proof of Theorem 3.2 16

A.2 Proofs on Quasi-Optimal Staionarity Equation 17

A.2.1 Proof of Theorem 3.3 17

A.3 Proofs on Kernel Representation 18

A.3.1 Proof of Theorem S.2 18

A.4 Proofs on Generic Properties of Quasi-optimal Bellman Operator 19

A.4.1 Proof of Proposition S.1 19

A.4.2 Proof of Proposition S.2 20

A.5 Proof of Theorem 4.1 20

A.6 Proof of Lemma S.1 22

A.7 Proof of Theorem 4.2 23

A.8 Proof of Theorem 4.3 27

A.9 Proof of Theorem 4.4 34

B Practical Implementation **35**

C Experiment Details and Additional Results **36**

C.1 Details of Simulation Settings and Real Data Analysis 36

C.2 Additional Experiment Details and Results 36

A TECHNICAL PROOFS

A.1 PROOFS ON CONSTRUCTING QUASI-OPTIMAL BELLMAN OPERATOR

A.1.1 PROOF OF THEOREM S.1

Theorem S.1. *Assume the induced policy has density function $\pi^*_\mu(a|s) \leq \mathbf{C}$ for all a, s, where $\mathbf{C}$ is a given constant. Then the proximal Bellman operator $\mathcal{B}_\mu$ in equation* (3) *has a closed form equivalent:*

$$\mathcal{B}_\mu V^*_\mu(s) = \mu\left\{1 - \int_{a\in\mathcal{W}_{s,1}}\left[\left(\frac{\int_{a\in\mathcal{W}_{s,1}} Q^*_\mu(s,a)da}{2\mu\sigma(\mathcal{W}_{s,1})} - \frac{1}{\sigma(\mathcal{W}_{s,1})}\right)^2 - \left(\frac{Q^*_\mu(s,a)}{2\mu}\right)^2\right]da\right\}$$
$$+ \frac{\mathbf{C}\sigma(\mathcal{W}_{s,1})\int_{a\in\mathcal{W}_{s,2}} Q^*_\mu(s,a)da - \mathbf{C}\sigma(\mathcal{W}_{s,2})\int_{a\in\mathcal{W}_{s,1}} Q^*_\mu(s,a)da}{2\sigma(\mathcal{W}_{s,1})} - \frac{\mu\mathbf{C}^2\sigma(\mathcal{W}_{s,2})(\sigma(\mathcal{W}_{s,2})+\sigma(\mathcal{W}_{s,1}))}{\sigma(\mathcal{W}_{s,1})}, \tag{14}$$

where $\mathcal{W}_{s,1}$ refers to the set $\{a \in \mathcal{A} : \mathbf{C} > \pi^*_\mu(a|s) > 0\}$, $\mathcal{W}_{s,2}$ refers to the set $\{a \in \mathcal{A} : \pi^*_\mu(a|s) = \mathbf{C}\}$.

Proof: The proof is mainly to check the KKT conditions of the maximization. The Lagrangian function of the RHS of (3) can be expressed as follows:

$$L(\pi, \tilde{\eta}, \varpi_1, \varpi_2) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q_\mu(s,a) + \mu\text{prox}(\pi(a|s))\right] - \tilde{\eta}(s)\left(\int_{a\in\mathcal{A}}\pi(a|s)da - 1\right)$$
$$+ \varpi_1(s,a)\pi(a|s) - \varpi_2(s,a)(\pi(a|s) - \mathbf{C}).$$

The following KKT conditions are necessary for the maximizer π^*_μ in the equation:

- Primal: $\int_{a\in\mathcal{A}}\pi^*_\mu(a|s)da - 1 = 0$, $-\pi^*_\mu(a|s) \leq 0$, $\pi^*_\mu(a|s) \leq \mathbf{C}$.
- Duality: $\varpi_1(s,a) \geq 0$, $\varpi_2(s,a) \geq 0$.
- Complementary slackness: $\varpi_1(s,a)\pi^*_\mu(a|s) = 0$, $\varpi_2(s,a)(\pi^*_\mu(a|s) - \mathbf{C}) = 0$.
- Stationarity: $Q^*_\mu(s,a) + \mu(1 - 2\pi^*_\mu(a|s)) - \tilde{\eta}(s) + \varpi_1(s,a) - \varpi_2(s,a) = 0$.

We can obtain the equation for $\pi_\mu(a|s)$ from the stationary condition such that

$$\pi^*_\mu(a|s) = \frac{1}{2} - \frac{1}{2\mu}[\tilde{\eta}(s) - Q^*_\mu(s,a) - \varpi_1(s,a) + \varpi_2(s,a)].$$

Combined with complementary slackness condition,

- If $\pi^*_\mu(a|s) = 0$, then $\varpi_1(s,a) \geq 0$, $\varpi_2(s,a) = 0$, thus $Q^*_\mu(s,a) \leq \tilde{\eta}(s) - \mu$.
- If $\mathbf{C} > \pi^*_\mu(a|s) > 0$, then $\varpi_1(s,a) = \varpi_2(s,a) = 0$, thus $\tilde{\eta}(s) - \mu + 2\mu\mathbf{C} > Q^*_\mu(s,a) > \tilde{\eta}(s) - \mu$.
- If $\pi^*_\mu(a|s) = \mathbf{C}$, then $\varpi_1(s,a) = 0$, $\varpi_2(s,a) \leq 0$, thus $Q^*_\mu(s,a) \geq \tilde{\eta}(s) - \mu + 2\mu\mathbf{C}$.

Therefore, $\pi^*_\mu(s,a)$ can be expressed as:

$$\pi^*_\mu(a|s) = \begin{cases} 0 & \text{if } Q^*_\mu(s,a) \leq \tilde{\eta}(s) - \mu \\ \frac{1}{2} - \frac{1}{2\mu}\left(\tilde{\eta}(s) - Q^*_\mu(s,a)\right) & \text{if } \tilde{\eta}(s) - \mu + 2\mu\mathbf{C} > Q^*_\mu(s,a) > \tilde{\eta}(s) - \mu \\ \mathbf{C} & \text{if } Q^*_\mu(s,a) \geq \tilde{\eta}(s) - \mu + 2\mu\mathbf{C} \end{cases} \tag{15}$$

Meanwhile, notice that $\int_{a\in\mathcal{A}}\pi^*_\mu(s,a) = 1$, we can show that $\tilde{\eta}(s)$ has a closed form:

$$\tilde{\eta}(s) = \mu + \frac{\int_{a\in\mathcal{W}_{s,1}} Q^*_\mu(s,a)da - 2\mu + 2\mu\mathbf{C}\sigma(\mathcal{W}_{s,2})}{\sigma(\mathcal{W}_{s,1})},$$

where $\mathcal{W}_{s,1}$ refers to the set $\{a \in \mathcal{A} : C > \pi^*_\mu(a|s) > 0\}$, $\mathcal{W}_{s,2}$ refers to the set $\{a \in \mathcal{A} : \pi^*_\mu(a|s) = C\}$, and $\sigma(\mathcal{W}_{s,1})$, $\sigma(\mathcal{W}_{s,1})$ refers to the interval length of the corresponding set. We take $\tilde{\eta}(s)$ back to (15), we then have

$$\pi_\mu(a|s) = \begin{cases} 0 & \text{if } Q^*_\mu(s,a) \le \tilde{\eta}(s) - \mu \\ \frac{Q^*_\mu(s,a)}{2\mu} - \frac{\int_{a\in\mathcal{W}_{s,1}} Q^*_\mu(s,a)da}{2\mu\sigma(\mathcal{W}_{s,1})} + \frac{1-\mathbf{C}\sigma(\mathcal{W}_{s,2})}{\sigma(\mathcal{W}_{s,1})} & \text{if } \tilde{\eta}(s) - \mu + 2\mu\mathbf{C} > Q^*_\mu(s,a) > \tilde{\eta}(s) - \mu \\ \mathbf{C} & \text{if } Q^*_\mu(s,a) \ge \tilde{\eta}(s) - \mu + 2\mu\mathbf{C} \end{cases} \tag{16}$$

We finally plug in the closed form of $\pi^*_\mu(a|s)$ to (3), by some algebra, we have

$$\mathcal{B}_\mu V^*_\mu(s) = \mu\left\{1 - \int_{a\in\mathcal{W}_{s,1}}\left[\left(\frac{\int_{a\in\mathcal{W}_{s,1}} Q^*_\mu(s,a)da}{2\mu\sigma(\mathcal{W}_{s,1})} - \frac{1}{\sigma(\mathcal{W}_{s,1})}\right)^2 - \left(\frac{Q^*_\mu(s,a)}{2\mu}\right)^2\right]da\right\}$$
$$+ \frac{\mathbf{C}\sigma(\mathcal{W}_{s,1})\int_{a\in\mathcal{W}_{s,2}} Q^*_\mu(s,a)da - \mathbf{C}\sigma(\mathcal{W}_{s,2})\int_{a\in\mathcal{W}_{s,1}} Q^*_\mu(s,a)da}{2\sigma(\mathcal{W}_{s,1})} - \frac{\mu\mathbf{C}^2\sigma(\mathcal{W}_{s,2})(\sigma(\mathcal{W}_{s,2})+\sigma(\mathcal{W}_{s,1}))}{\sigma(\mathcal{W}_{s,1})}.$$

A.1.2 Proof of Corollary S.1

Corollary S.1. *When $\sigma(\mathcal{W}_{s,2}) = 0$, we denote $\mathcal{W}_1$ as $\mathcal{W}$, the closed form in* (14) *can be simplified as*

$$\mathcal{B}_\mu V^*_\mu(s) = \mu - \frac{1}{4\mu}\left(\frac{(\int_{a'\in\mathcal{W}_s} Q^*_\mu(s,a')da' - 2\mu)^2}{\sigma(\mathcal{W}_s)} - \int_{a\in\mathcal{W}_s} {Q^*_\mu}^2(s,a)da\right).$$

Proof: We plug in $\sigma(\mathcal{W}_{s,2}) = 0$ to (14), then could obtain the result.

A.1.3 Proof of Theorem 3.1

Proof of Theorem 3.1: For any generic value function $V(s)$ and the corresponding generic Q-function $Q(s,a)$, we first build the lower bound:

$$\begin{aligned}\mathcal{B}_\mu V(s) &= \max_{\pi\in\Delta_{\text{convex}}(\mathcal{A})} \mathbb{E}_{a\sim\pi(\cdot|s)}[Q(s,a) + \mu(1-\pi(a|s))] \\ &\ge \max_{\pi\in\Delta_{\text{convex}}(\mathcal{A})} \mathbb{E}_{a\sim\pi(\cdot|s)}[Q(s,a) + \mu - \mu\mathbf{C}] \\ &= \mathcal{B}V(s) + \mu(1-\mathbf{C}).\end{aligned}$$

For the upper bound:

$$\begin{aligned}\mathcal{B}_\mu V(s) &= \max_{\pi\in\Delta_{\text{convex}}(\mathcal{A})(\mathcal{A})} \mathbb{E}_{a\sim\pi(\cdot|s)}[Q(s,a) + \mu(1-\pi(a|s))] \\ &\le \max_{\pi\in\Delta_{\text{convex}}(\mathcal{A})(A)} \mathbb{E}_{a\sim\pi(\cdot|s)}[Q(s,a) + \mu] \\ &= \mathcal{B}V(s) + \mu.\end{aligned}$$

Therefore, we have $\mathcal{B}_\mu V(s) - \mathcal{B}V(s) \in [\mu(1-\mathbf{C}), \mu]$.

A.1.4 Proof of Theorem 3.2

Proof of Theorem 3.2: Suppose $Q^*_\mu(s,a) = -\alpha_1(s)a^2 + \alpha_2(s)a + \alpha_3(s)$ with $\alpha_1(s) > 0$. We assume the density won't reach its boundary value $\mathbf{C}$ for this theorem, and we proceed by simplifying $\alpha_i(s)$ as α_i for $i = 1,2,3$. By Equation (5), we have

$$\pi^*_\mu(a|s) = \left\{\frac{Q^*_\mu(s,a)}{2\mu} - \frac{\int_{a\in\mathcal{W}_s} Q^*_\mu(s,a)da}{2\mu\sigma(\mathcal{W}_s)} + \frac{1}{\sigma(\mathcal{W}_s)}\right\}^+.$$

We first try to find the support set of $\pi^*_\mu(a|s)$. Since Q^*_μ takes the maximum value at $y = \frac{\alpha_2}{2\alpha_1}$, by the symmetric property of quadratic function, the support set should be of the form $\mathcal{W}_s = [y-l, y+l](l > 0)$. Additionally, the boundary point of the support set should be the solution of

$$\frac{Q^*_\mu(s,a)}{2\mu} - \frac{\int_{a\in\mathcal{W}_s} Q^*_\mu(s,a)da}{2\mu\sigma(\mathcal{W}_s)} + \frac{1}{\sigma(\mathcal{W}_s)} = 0,$$

with respect to a. Thus, we can find the boundary point of the support set by solving the equation with respect to l:

$$-\alpha_1(y \pm l)^2 + \alpha_2(y \pm l) + \alpha_3 = \frac{1}{2l}\int_{y-l}^{y+l}(-\alpha_1 a^2 + \alpha_2 a + \alpha_3)da - \frac{\mu}{l}.$$

It turns out that $l = \frac{(12\alpha_1^2\mu)^{\frac{1}{3}}}{2\alpha_1}$. Thus, the support set has the closed-form

$$\mathcal{W}_s = \left\{ a : a \in \left[\frac{\alpha_2 - (12\alpha_1^2\mu)^{\frac{1}{3}}}{2\alpha_1}, \frac{\alpha_2 + (12\alpha_1^2\mu)^{\frac{1}{3}}}{2\alpha_1} \right] \right\}.$$

Therefore $\sigma(\mathcal{W}_s) = \frac{(12\alpha_1^2\mu)^{\frac{1}{3}}}{\alpha_1}$, and

$$\frac{\int_{a\in\mathcal{W}_s} Q_\mu^*(s,a)da}{2\mu\sigma(\mathcal{W}_s)} = -\frac{(12\alpha_1^2\mu)^{\frac{2}{3}} - 3\alpha_2^2}{24\mu\alpha_1} + \frac{\alpha_3}{2\mu}.$$

We plug in the result to the closed form of $\pi_\mu^*(a|s)$, and obtain the probability density function

$$\pi_\mu^*(a|s) = \left\{ \frac{\alpha_1}{2\mu}(a + \frac{\alpha_2}{2\alpha_1})^2 - \frac{3}{2}(\frac{\alpha_1}{12\mu})^{\frac{1}{3}} \right\}^+.$$

It is clear that the resulting distribution of $\pi_\mu^*(a|s)$ is of the exact form of q-Gaussian distribution with $q = 0, \beta = \frac{\alpha_1}{2\mu}$ and centered at $\frac{\alpha_2}{2\alpha_1}$.

A.2 Proofs on Quasi-Optimal Staionarity Equation

A.2.1 Proof of Theorem 3.3

Proof of Theorem 3.3: By the stationary condition from Theorem S.1 we have

$$Q_\mu^*(s,a) + \mu(1 - 2\pi_\mu^*(a|s)) - \tilde{\eta}(s) + \varpi_1(s,a) - \varpi_2(s,a) = 0,$$

therefore, by the definition of $Q_\mu^*(s,a)$, we have

$$\mathbb{E}_{S^{t+1}|s,a}[R(S^{t+1},s,a)] + \gamma\mathbb{E}_{S^{t+1}|s,a}[V_\mu^*(S^{t+1})] + \mu(1-2\pi_\mu^*(a|s)) - \tilde{\eta}(s) + \varpi_1(s,a) - \varpi_2(s,a) = 0. \tag{17}$$

Notice that $\mathbb{E}_{S^{t+1}|s,a}[R(S^{t+1},s,a)] = r(s,a)$, and we take expectation with respect to a following the policy distribution $\pi_\mu^*(a|s)$ from both sides of (17),

$$0 = \mathbb{E}_{a\sim\pi_\mu^*(a|s)}\left[r(s,a) + \gamma\mathbb{E}_{S^{t+1}|s,a}[V_\mu^*(S^{t+1})] + \mu(1 - 2\pi_\mu^*(a|s)) - \tilde{\eta}(s) + \varpi_1(s,a) - \varpi_2(s,a)\right],$$

$$0 = \int_{a\in\mathcal{A}} \pi_\mu^*(a|s)\left[r(s,a) + \gamma\mathbb{E}_{S^{t+1}|s,a}[V_\mu^*(S^{t+1})] + \mu(1 - 2\pi_\mu^*(a|s)) - \tilde{\eta}(s) + \varpi_1(s,a) - \varpi_2(s,a)\right]da.$$

According to the proximal Bellman optimality equation $\mathcal{B}_\mu V_\mu^*(s) = V_\mu^*(s)$, where $V_\mu^*(s)$ is the fixed point of $\mathcal{B}_\mu$. With the explicit definition of V_μ^*, we observe that

$$\begin{aligned}
0 =& \int_{a\in\mathcal{A}} \pi_\mu^*(a|s)\left[r(s,a) + \gamma\mathbb{E}_{S^{t+1}|s,a}[V_\mu^*(S^{t+1})] + \mu(1 - \pi_\mu^*(a|s))\right]da - \int_{a\in\mathcal{A}} \mu{\pi_\mu^*}^2(a|s)da \\
&- \int_{a\in\mathcal{A}} \pi_\mu^*(a|s)\tilde{\eta}(s)da + \int_{a\in\mathcal{A}} \pi_\mu^*(a|s)\varpi_1(s,a)da - \int_{a\in\mathcal{A}} \pi_\mu^*(a|s)\varpi_2(s,a)da \\
=& V_\mu^*(s) - \int_{a\in\mathcal{A}} \mu{\pi_\mu^*}^2(a|s)da - \int_{a\in\mathcal{A}} \pi_\mu^*(a|s)\tilde{\eta}(s)da + \int_{a\in\mathcal{A}} \pi_\mu^*(a|s)\varpi_1(s,a)da \\
&- \int_{a\in\mathcal{A}} \pi_\mu^*(a|s)\varpi_2(s,a)da
\end{aligned}$$

Meanwhile $\int_{a\in\mathcal{A}} \pi_\mu^*(a|s)\tilde{\eta}(s)da = \tilde{\eta}(s)\int_{a\in\mathcal{A}} \pi_\mu^*(a|s)da = \tilde{\eta}(s)$ by the property of density, $\int_{a\in\mathcal{A}} \pi_\mu^*(a|s)\varpi_1(s,a)da = 0$, and $\int_{a\in\mathcal{A}} \pi_\mu^*(a|s)\varpi_2(s,a)da = \mathbf{C}\int_{a\in\mathcal{A}} \varpi_2(s,a)da$ by complete slackness, we further have

$$V_\mu^*(s) - \mu\int_{a\in\mathcal{A}} {\pi_\mu^*}^2(a|s)da - \tilde{\eta}(s) - \mathbf{C}\int_{a\in\mathcal{A}} \varpi_2(s,a)da = 0.$$

Since $0 \leq \pi_\mu^*(a|s) \leq \mathbf{C}$, thus $\mu \int_{a\in\mathcal{A}} {\pi_\mu^*}^2(a|s)da = \mu\mathbb{E}\pi_\mu^*(a|s) \in [0, \mathbf{C}]$. Therefore,

$$\eta(s) := \tilde{\eta}(s) - V_\mu^*(s) \in [-\mu\mathbf{C} - \mathbf{C}\int_{a\in\mathcal{A}} \varpi_2(s,a)da, -\mathbf{C}\int_{a\in\mathcal{A}} \varpi_2(s,a)da].$$

The stationary condition can be reformulated as

$$\mathbb{E}_{S^{t+1}|s,a}\big[R(S^{t+1},s,a)+\gamma V_\mu^*(S^{t+1})\big]-\mu\text{prox}^\circ(\pi_\mu^*(a|s))-\eta(s)+\varpi_1(s,a)-\varpi_2(s,a)-V_\mu^*(s) = 0. \tag{18}$$

Obviously, (π_μ^*, V_μ^*) is a solution for the above equation for some $\eta(s), \varpi_1(s,a)$, and $\varpi_2(s,a)$, such that

$$\varpi_1(s,a) \geq 0, \varpi_2(s,a) \geq 0, \varpi_1(s,a)\cdot\pi_\mu(a|s) = 0, \varpi_2(s,a)\cdot(\mathbf{C} - \pi_\mu(a|s)) = 0$$
$$\text{and } \eta(s) \in \Big[-\mu\mathbf{C} - \mathbf{C}\int_{a\in\mathcal{A}} \varpi_2(s,a)da, -\mathbf{C}\int_{a\in\mathcal{A}} \varpi_2(s,a)da\Big].$$

When $\sigma(\mathcal{W}_{s,2}) = 0$, we have $\varpi_2(s,a) = 0$.
Plugging in to equation (18), and denote $\mathcal{W}_{s,1}$ as $\mathcal{W}_s$, we have the exact form of (10).

A.3 Proofs on Kernel Representation

A.3.1 Proof of Theorem S.2

Theorem S.2. *Suppose $u^* \in \mathcal{H}_\mathcal{K}^{C_0}$ is reproduced by a universal kernel $K(\cdot,\cdot)$, then the minimax optimizer* (11) *can be decoupled to a single-stage minimization problem as*

$$\min_{V_\mu,\pi_\mu,\eta,\varpi} \mathcal{L}_U = \mathbb{E}_{S^t,\widetilde{S}^t,A^t,\widetilde{A}^t,S^{t+1},\tilde{S}^{t+1}}\big[\big(\mathcal{G}_{V_\mu,\pi_\mu}\left(S^t,A^t,S^{t+1}\right) - \eta\left(S^t\right) + \varpi\left(A^t \mid S^t\right) - V_\mu\left(S^t\right)\big)$$
$$\cdot C_0 K\left(S^t,A^t;\widetilde{S}^t,\widetilde{A}^t\right)\left(\mathcal{G}_{V_\mu,\pi_\mu}(\widetilde{S}^t,\widetilde{A}^t,\widetilde{S}^{t+1}) - \eta(\widetilde{S}^t) + \varpi(\widetilde{A}^t \mid \widetilde{S}^t) - V_\mu(\widetilde{S}^t)\right)\Big],$$

where $(\widetilde{S}^t, \widetilde{A}^t, \widetilde{S}^{t+1})$ is an independent copy of the transition pair (S^t, A^t, S^{t+1}).

Proof: Let $\tilde{u} = \mathbb{E}_{S^t,A^t}\left[\left(\mathcal{G}_{V_\mu,\pi_m u}(S^t,A^t,S^{t+1}) - \eta(S^t) + \varpi(A^t|S^t) - V_\mu(S^t)\right)K(\cdot,\{S^t,A^t\})\right]$., and define the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_{\text{RKHS}}}$ in $\mathcal{H}_\mathcal{K}^{C_0}$. It follows from the definition of $\mathcal{L}(V_\mu,\pi_\mu,\eta,\varpi,u)$ and kernel reproducing property we have,

$$\min_{V_\mu,\pi_\mu,\eta,\varpi}\max_u \mathcal{L}^2\left(V_\mu,\pi_\mu,\eta,\varpi,u\right)$$
$$= \min_{V_\mu,\pi_\mu,\eta,\varpi}\max_u \left(\mathbb{E}_{S^t,A^t}\left[\left(\mathcal{G}_{V_\mu,\pi_\mu}\left(S^t,A^t,S^{t+1}\right) - \eta\left(S^t\right) + \varpi\left(A^t \mid S^t\right) - V_\mu\left(S^t\right)\right)u\left(S^t,A^t\right)\right]\right)^2$$
$$= \min_{V_\mu,\pi_\mu,\eta,\varpi}\max_u \Bigg(\mathbb{E}_{S^t,A^t}\Big[\Big\langle\big(\mathcal{G}_{V_\mu,\pi_\mu}\left(S^t,A^t,S^{t+1}\right) - \eta\left(S^t\right) + \varpi\left(A^t \mid S^t\right) - V_\mu\left(S^t\right)\big)\cdot$$
$$K\left(\cdot\,;S^t,A^t\right), u\left(S^t,A^t\right)\Big\rangle_{\mathcal{H}_{\text{RKHS}}}\Big]\Bigg)^2$$
$$= \min_{V_\mu,\pi_\mu,\eta,\varpi}\max_u \Bigg\langle\mathbb{E}_{S^t,A^t}\Big[\big(\mathcal{G}_{V_\mu,\pi_\mu}\left(S^t,A^t,S^{t+1}\right) - \eta\left(S^t\right) + \varpi\left(A^t \mid S^t\right) - V_\mu\left(S^t\right)\big)\cdot$$
$$K\left(\cdot\,;S^t,A^t\}\right)\Big], u\left(S^t,A^t\right)\Bigg\rangle^2_{\mathcal{H}_{\text{RKHS}}}$$
$$= \min_{V_\mu,\pi_\mu,\eta,\varpi}\Bigg\langle\mathbb{E}_{S^t,A^t}\Big[\big(\mathcal{G}_{V_\mu,\pi_\mu}\left(S^t,A^t,S^{t+1}\right) - \eta\left(S^t\right) + \varpi\left(A^t \mid S^t\right) - V_\mu\left(S^t\right)\big)\cdot$$
$$K\left(\cdot\,;S^t,A^t\right)\Big], \frac{\sqrt{C_0}\widetilde{u}}{\|\widetilde{u}\|_{\mathcal{H}_{\text{RKHS}}}}\Bigg\rangle^2_{\mathcal{H}_{\text{RKHS}}},$$

where the last equality holds because of the maximization of inner product between $\tilde{u}$ and $\mathbb{E}_{S^t,A^t}\left[\left(\mathcal{G}_{V_\mu,\pi_\mu}(S^t,A^t,S^{t+1})-\eta(S^t)+\varpi(A^t|S^t)-V_\mu(S^t)\right)K(\cdot\,;S^t,A^t)\right]$ should have the same direction. Then we have,

$$
\begin{aligned}
&\min_{V_\mu,\pi_\mu,\eta,\varpi}\Big\langle \mathbb{E}_{S^t,A^t}\left[\left(\mathcal{G}_{V_\mu,\pi_\mu}\left(S^t,A^t,S^{t+1}\right)-\eta\left(S^t\right)+\varpi\left(A^t\mid S^t\right)-V_\mu\left(S^t\right)\right)\cdot K\left(\cdot\,;S^t,A^t\right)\right],\\
&\qquad\qquad \sqrt{C_0}\tilde{u}/\|\widetilde{u}\|_{\mathcal{H}_{\mathrm{RKHS}}}\Big\rangle^2_{\mathcal{H}_{\mathrm{RKHS}}}\\
&=\min_{V_\mu,\pi_\mu,\eta,\varpi}\Big\langle \mathbb{E}_{S^t,A^t}\left[\left(\mathcal{G}_{V_\mu,\pi_\mu}\left(S^t,A^t,S^{t+1}\right)-\eta\left(S^t\right)+\varpi\left(A^t\mid S^t\right)-V_\mu\left(S^t\right)\right)\cdot K(\cdot\,;S^t,A^t)\right],\\
&\qquad \mathbb{E}_{S^t,A^t}\left[\left(\mathcal{G}_{V_\mu,\pi_\mu}\left(S^t,A^t,S^{t+1}\right)-\eta\left(S^t\right)+\varpi\left(A^t\mid S^t\right)-V_\mu\left(S^t\right)\right)\cdot K(\cdot\,;S^t,A^t)\right]\Big\rangle\\
&\qquad\qquad\cdot\left\langle\frac{\tilde{u}}{\|\widetilde{u}\|_{\mathcal{H}_{\mathrm{RKHS}}}},\frac{C_0\widetilde{u}}{\|\widetilde{u}\|_{\mathcal{H}_{\mathrm{RKHS}}}}\right\rangle_{\mathcal{H}_{\mathrm{RKHS}}}\\
&=\min_{V_\mu,\pi_\mu,\eta,\varpi}\Big\langle \mathbb{E}_{S^t,A^t}\left[\left(\mathcal{G}_{V_\mu,\pi_\mu}\left(S^t,A^t,S^{t+1}\right)-\eta\left(S^t\right)+\varpi\left(A^t\mid S^t\right)-V_\mu\left(S^t\right)\right)\cdot K\left(\cdot\,;S^t,A^t\right)\right],\\
&C_0\mathbb{E}_{\widetilde{S}^t,\widetilde{A}^t}\left[\left(\mathcal{G}_{V_\mu,\pi_\mu}\left(\widetilde{S}^t,\widetilde{A}^t,\widetilde{S}^{t+1}\right)-\eta\left(\widetilde{S}^t\right)+\varpi(\widetilde{A}^t\mid\widetilde{S}^t)-V_\mu(\widetilde{S}^t)\right)\cdot K\left(\cdot\,;\widetilde{S}^t,\widetilde{A}^t\right)\right]\Big\rangle_{\mathcal{H}_{\mathrm{RKHS}}},
\end{aligned}
$$

where the first equality is by the equality condition of Cauchy-Schwarz inequality, i.e. $\tilde{u}/\|\tilde{u}\|_{\mathcal{H}_{\mathrm{RKHS}}}$ is linear dependent of $\mathbb{E}_{S^t,A^t}\left[\left(\mathcal{G}_{V_\mu,\pi_\mu}(S^t,A^t,S^{t+1})-\eta(S^t)+\varpi(A^t|S^t)-V_\mu(S^t)\right)K(\cdot\,;S^t,A^t)\right]$. Then, by the reproducing property of $K(S^t,A^t;\tilde{S}^t,\tilde{A}^t)$, we have

$$
\begin{aligned}
&\min_{V_\mu,\pi_\mu,\eta,\varpi}\max_{u\in\mathcal{H}_{\mathcal{K}}^{C_0}}\quad \mathcal{L}^2(V_\mu,\pi_\mu,\eta,\varpi,u)\\
&=\min_{V_\mu,\pi_\mu,\eta,\varpi}\quad \mathbb{E}_{S^t,\widetilde{S}^t,A^t,\widetilde{A}^t}\Big[\left(\mathcal{G}_{V_\mu,\pi_\mu}(S^t,A^t,S^{t+1})-\eta(S^t)+\varpi(A^t|S^t)-V_\mu(S^t)\right)\\
&\qquad C_0\Big\langle K\left(S^t,A^t;\cdot\right),K\left(\widetilde{S}^t,\widetilde{A}^t;\cdot\right)\Big\rangle_{\mathcal{H}_{\mathrm{RKHS}}}\left(\mathcal{G}_{V_\mu,\pi_\mu}(\widetilde{S}^t,\widetilde{A}^t,\widetilde{S}^{t+1})-\eta(\widetilde{S}^t)+\varpi(\widetilde{A}^t|\widetilde{S}^t)-V_\mu(\widetilde{S}^t)\right)\Big]\\
&=\min_{V_\mu,\pi_\mu,\eta,\varpi}\quad \mathbb{E}_{S^t,\widetilde{S}^t,A^t,\widetilde{A}^t}C_0\Big[\left(\mathcal{G}_{V_\mu,\pi_\mu}(S^t,A^t,S^{t+1})-\eta(S^t)+\varpi(A^t|S^t)-V_\mu(S^t)\right)\\
&\qquad K\left(S^t,A^t;\widetilde{S}^t,\widetilde{A}^t\right)\left(\mathcal{G}_{V_\mu,\pi_\mu}(\widetilde{S}^t,\widetilde{A}^t,\widetilde{S}^{t+1})-\eta(\widetilde{S}^t)+\varpi(\widetilde{A}^t|\widetilde{S}^t)-V_\mu(\widetilde{S}^t)\right)\Big]\\
&=\min_{V_\mu,\pi_\mu,\eta,\varpi}\quad \mathbb{E}_{S^t,\widetilde{S}^t,A^t,\widetilde{A}^t,S^{t+1},\widetilde{S}^{t+1}}C_0\Big[\left(\widetilde{\mathcal{G}}_{V_\mu,\pi_\mu}(S^t,A^t,S^{t+1})-\eta(S^t)+\varpi(A^t|S^t)-V_\mu(S^t)\right)\\
&\qquad K\left(S^t,A^t;\widetilde{S}^t,\widetilde{A}^t\right)\left(\widetilde{\mathcal{G}}_{V_\mu,\pi_\mu}(\widetilde{S}^t,\widetilde{A}^t,\widetilde{S}^{t+1})-\eta(\widetilde{S^t})+\varpi(\widetilde{A}^t|\widetilde{S}^t)-V_\mu(\widetilde{S}^t)\right)\Big].
\end{aligned}
$$

Thus, we finish the proof.

A.4 Proofs on Generic Properties of Quasi-optimal Bellman Operator

A.4.1 Proof of Proposition S.1

Proposition S.1. *The quasi-optimal Bellman operator $\mathcal{B}_\mu$ is γ-contractive with respect to the supreme norm over $\mathcal{S}$. That is $\|\mathcal{B}_\mu V-\mathcal{B}_\mu V'\|_\infty\le\gamma\|V-V'\|_\infty$, for any generic value functions $\{V,V':\mathcal{S}\to\mathbb{R}\}$.*

Proposition S.1 justifies that there exists a unique fixed point of $\mathcal{B}_\mu$, i.e., V_μ^*, indicating that the quasi-optimal value function V_μ^* and the induced policy π_μ^* are well defined and unique.

Proof: By the definition of $\mathcal{B}_\mu$, the explicit form corresponding to V is as follows:

$$
\mathcal{B}_\mu V(s)=\max_\pi\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[\mathbb{E}_{S^{t+1}|s,a}[R(S^{t+1},s,a)+\gamma V(S^{t+1})]+\mu\mathrm{prox}^\circ(\pi(a|s))\Big].
$$

For any two arbitrary value functions V and V', we have

$$
\begin{aligned}
&\|\mathcal{B}_\mu V(s) - \mathcal{B}_\mu V'(s)\|_\infty \\
=&\max_{\pi_1} \mathbb{E}_{a\sim\pi(\cdot|s)}\Big[\mathbb{E}_{S^{t+1}|s,a}[R(S^{t+1},s,a)+\gamma V(S^{t+1})] + \mu \text{prox}^\circ(\pi_1(a|s))\Big] - \\
&\max_{\pi_2} \mathbb{E}_{a\sim\pi(\cdot|s)}\Big[\mathbb{E}_{S^{t+1}|s,a}[R(S^{t+1},s,a)+\gamma V'(S^{t+1})] + \mu \text{prox}^\circ(\pi_2(a|s))\Big] \\
&\leq \max_{\pi} \Bigg\{ \mathbb{E}_{a\sim\pi(\cdot|s)}\Big[\mathbb{E}_{S^{t+1}|s,a}[R(S^{t+1},s,a)+\gamma V(S^{t+1})] + \mu \text{prox}^\circ(\pi(a|s))\Big] - \\
&\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[\mathbb{E}_{S^{t+1}|s,a}[R(S^{t+1},s,a)+\gamma V'(S^{t+1})] + \mu \text{prox}^\circ(\pi(a|s))\Big] \Bigg\} \\
&= \max_{\pi} \gamma \mathbb{E}_{a\sim\pi(\cdot|s), S^{t+1}|s,a}\Big[\big(V(S^{t+1}) - V'(S^{t+1})\big)\Big] \\
&\leq \gamma \|V(s) - V'(s)\|_\infty.
\end{aligned}
$$

A.4.2 Proof of Proposition S.2

Proposition S.2. *For any $s \in \mathcal{S}$, the performance error between $V_\mu^*(s)$ and $V^*(s)$ satisfies*

$$\|V_\mu^* - V^*\|_\infty \leq \frac{\mu \cdot \max\{|1-\mathbf{C}|, 1\}}{1-\gamma},$$

where $\mathbf{C}$ is the upper bound for induced policy π_μ.

Proof of Proposition S.2:

$$
\begin{aligned}
\|V_\mu^* - V^*\|_\infty &= \|\mathcal{B}_\mu V_\mu^* - \mathcal{B} V^*\|_\infty \\
&\leq \|\mathcal{B}_\mu V_\mu^* - \mathcal{B}_\mu V^*\|_\infty + \|\mathcal{B}_\mu V^* - \mathcal{B} V^*\|_\infty.
\end{aligned}
$$

Notice that $\|\mathcal{B}_\mu V_\mu^* - \mathcal{B}_\mu V^*\|_\infty \leq \gamma \|V_\mu^* - V^*\|_\infty$ by Theorem S.1, and $\|\mathcal{B}_\mu V^* - \mathcal{B} V^*\|_\infty \leq \mu \cdot \max\{|1-\mathbf{C}|, 1\}$ by Proposition 3.1. Therefore,

$$(1-\gamma)\|V_\mu^* - V^*\|_\infty \leq \mu \cdot \max\{|1-\mathbf{C}|, 1\}.$$

We finish the proof.

A.5 Proof of Theorem 4.1

Proof of Theorem 4.1: We first prove that when $\mu \to \infty$, π_μ^* would degenerate to uniform distribution over $\mathcal{A}$. By (5), we only need to prove that for arbitrary small $\epsilon > 0$

$$\Big|\frac{Q_\mu^*(s,a)}{2\mu} - \frac{\int_{a\in\mathcal{W}_s} Q_\mu^*(s,a)da}{2\mu\sigma(\mathcal{W}_s)} + \frac{1}{\sigma(\mathcal{W}_s)} - \frac{1}{\sigma(\mathcal{A})}\Big| < \epsilon.$$

Lower bound:

$$
\begin{aligned}
\frac{Q_\mu^*(s,a)}{2\mu} - \frac{\int_{a\in\mathcal{W}_s} Q_\mu^*(s,a)}{2\mu\sigma(\mathcal{W}_s)} + \frac{1}{\sigma(\mathcal{W}_s)} &\geq \frac{Q_\mu^*(s,a)}{2\mu} - \frac{\sigma(\mathcal{W}_s)\max_{a'} Q_\mu^*(s,a')}{2\mu\sigma(\mathcal{W}_s)} + \frac{1}{\sigma(\mathcal{W}_s)} \qquad (19) \\
&\geq \frac{Q_\mu^*(s,a)}{2\mu} - \frac{\max_{a'} Q_\mu^*(s,a')}{2\mu} + \frac{1}{\sigma(\mathcal{A})} \qquad (20)
\end{aligned}
$$

Thus, we aim to prove that

$$\Big|\frac{Q_\mu^*(s,a) - \max_{a'} Q_\mu^*(s,a')}{2\mu}\Big| \to 0.$$

Let V^* be the unique fixed point of (2), and $H_{\max} = \max H(\pi)$, where

$$H(\pi) = \mathbb{E}_{a\sim\pi(\cdot|s)}[1 - \pi(a|s)].$$

Let $r(s,a) := \mathbb{E}_{S^{t+1}|s,a}[R(S^{t+1},s,a)]$, by the definition of Q^*_μ, we have

$$\frac{Q^*_\mu(s,a)}{2\mu} - \frac{\gamma \mathbb{E}_{S^{t+1}|s,a}\left[V^*_\mu\left(S^{t+1}\right)\right]}{2\mu} = \frac{r(s,a)}{2\mu}$$
$$\frac{Q^*_\mu(s,a)}{2\mu} - \frac{\gamma \mathbb{E}_{S^{t+1}|s,a}\left[V^*_\mu\left(S^{t+1}\right) - V^*\left(S^{t+1}\right)\right]}{2\mu} - \frac{\gamma \mathbb{E}_{S^{t+1}|s,a}\left[V^*\left(S^{t+1}\right)\right]}{2\mu} = \frac{r(s,a)}{2\mu}.$$

Therefore,

$$\begin{aligned}
&\frac{Q^*_\mu(s,a)}{2\mu} - \frac{\mu\gamma H_{\max}}{2(1-\gamma)} \leq \frac{r(s,a)}{2\mu} + \frac{\gamma \mathbb{E}_{s'|s,a}\left[V^*\left(s'\right)\right]}{2\mu}, \\
&\frac{Q^*_\mu(s,a)}{2\mu} - \frac{\mu\gamma H_{\max}}{2(1-\gamma)} \leq \frac{R_{\max}}{2(1-\gamma)\mu}.
\end{aligned} \tag{21}$$

Meanwhile, from another perspective, the proximal Bellman operator (3) can be treated as a new MDP with the immediate reward $r(s,a) + \mu H(\pi(\cdot|s))$ for given s, a. Combine with the fact that

$$\frac{\gamma\mu H_{\max}}{1-\gamma} = \max_\pi \mathbb{E}_\pi\Big[\sum_{t=2}^{\infty} \gamma^{t-1}(\mu - \mu\pi(A^t|S^t))|S^1 = s, A^1 = a\Big].$$

Let $\pi_H = \text{argmax}_\pi H(\pi(a|s))$, then

$$\begin{aligned}
\frac{Q^*_\mu(s,a)}{2\mu} - \frac{\mu\gamma H_{\max}}{2(1-\gamma)} &= \frac{Q^*_\mu(s,a)}{2\mu} - \max_\pi \mathbb{E}_\pi\left[\sum_{t=2}^{\infty} \gamma^{t-1}\left(\mu - \mu\pi\left(A^t \mid S^t\right)\right) \mid S^1 = s, A^1 = a\right] \\
&\geq \frac{Q^{\pi_H}_\mu(s,a)}{2\mu} - \mathbb{E}_{\pi_H}\left[\sum_{t=2}^{\infty} \gamma^{t-1}\left(\mu - \mu\pi_H\left(A^t \mid S^t\right)\right) \mid S^1 = s, A^1 = a\right] \\
&= \mathbb{E}_{\pi_H}\left[\sum_{t=1}^{\infty} \gamma^{t-1}\frac{r\left(S^t, A^t\right)}{2\mu} \mid S^1 = s, A^1 = a\right] \\
&\geq -\frac{R_{\max}}{2(1-\gamma)\mu}.
\end{aligned} \tag{22}$$

Based on (21) and (22), we have

$$\begin{aligned}
\frac{Q^*_\mu(s,a)}{2\mu} - \frac{\max_{a'} Q^*_\mu(s,a')}{2\mu} &= \frac{Q^*_\mu(s,a)}{2\mu} - \frac{\gamma H_{\max}}{2(1-\gamma)} + \frac{\gamma H_{\max}}{2(1-\gamma)} - \frac{\max_{a'} Q^*_\mu(s,a')}{2\mu} \\
&\geq -\frac{R_{\max}}{(1-\gamma)\mu}.
\end{aligned} \tag{23}$$

Similarly, we also have

$$\frac{Q^*_\mu(s,a)}{2\mu} - \frac{\max_{a'} Q^*_\mu(s,a')}{2\mu} \leq \frac{R_{\max}}{(1-\gamma)\mu}. \tag{24}$$

Therefore, we have the lower bound approaching to $\frac{1}{\sigma(\mathcal{A})}$.

For the upper bound, we have $\int_{a\in\mathcal{A}} \pi^*_\mu(a|s)da = 1$, thus

$$\begin{aligned}
&\int_{a\in\mathcal{A}} \left\{\frac{Q^*_\mu(s,a)}{2\mu} - \frac{\int_{a'\in\mathcal{W}_s} Q^*_\mu(s,a')da'}{2\mu\sigma(\mathcal{W}_s)} + \frac{1}{\sigma(\mathcal{W}_s)}\right\}^+ da \\
&\geq \int_{a\in\mathcal{A}} \left\{\frac{\min_{a''} Q^*_\mu(s,a'')}{2\mu} - \frac{\int_{a'\in\mathcal{W}_s} Q^*_\mu(s,a')da'}{2\mu\sigma(\mathcal{W}_s)} + \frac{1}{\sigma(\mathcal{W}_s)}\right\} da \\
&\frac{1}{\sigma(\mathcal{A})} \geq \frac{\min_{a''} Q^*_\mu(s,a'')}{2\mu} - \frac{\int_{a'\in\mathcal{W}_s)} Q^*_\mu(s,a')da'}{2\mu} + \frac{1}{\sigma(\mathcal{W}_s)}.
\end{aligned}$$

By (24), we then have

$$\frac{Q_\mu^*(s,a)}{2\mu} - \frac{\int_{a\in\mathcal{W}_s} Q_\mu^*(s,a)da}{2\mu\sigma(\mathcal{W}_s)} + \frac{1}{\sigma(\mathcal{W}_s)} = \frac{Q_\mu^*(s,a)}{2\mu} - \frac{\max_{a''} Q_\mu^*(s,a'')}{2\mu} \tag{25}$$

$$+ \frac{\max_{a''} Q_\mu^*(s,a'')}{2\mu} - \frac{\int_{a'\in\mathcal{W}_s} Q_\mu^*(s,a')da'}{2\mu} + \frac{1}{\sigma(\mathcal{W}_s)}$$

$$\leq \frac{1}{\sigma(\mathcal{A})} + \frac{R_{\max}}{(1-\gamma)\mu} \tag{26}$$

Therefore, by the lower bound and upper bound, we conclude that $\pi_\mu(a|s)$ will decay to the uniform distribution on $\mathcal{A}$ as $\mu \to \infty$.

For the case when $\mu \to 0$, we prove that π_μ would converge to the uniform distribution with the length of the support set equal to $\frac{1}{\mathbf{C}}$. Therefore, when $\mathbf{C} \to \infty$, it will converge to the point mass. According to (16), we only need to prove $\sigma(\mathcal{W}_{s,1}) \to 0$ as $\mu \to 0$. Meanwhile by Theorem (S.1), $a \in \mathcal{W}_{s,1}$, if

$$\sigma(\mathcal{W}_{s,1})Q_\mu^*(s,a) - \Big(\int_{a'\in\mathcal{W}_{s,1}} Q_\mu^*(s,a')da' - 2\mu + 2\mu\mathbf{C}\sigma(\mathcal{W}_{s,2})\Big) \in (0, 2\mu\mathbf{C}\sigma(\mathcal{W}_{s,1})).$$

As $\mu \to 0$, $(0, 2\mu\mathbf{C}\sigma(\mathcal{W}_{s,1})) \to 0$. Thus, by squeeze theorem, we have $\sigma(\mathcal{W}_{s,1})Q_\mu^*(s,a) - \Big(\int_{a'\in\mathcal{W}_{s,1}} Q_\mu^*(s,a')da' - 2\mu + 2\mu C\sigma(\mathcal{W}_{s,2})\Big) \to 0$ as $\mu \to 0$, which is equivalent to

$$\sigma(\mathcal{W}_{s,1})Q_\mu^*(s,a) - \int_{a'\in\mathcal{W}_{s,1}} Q_\mu^*(s,a')da' \to 0 \quad \text{for all } a \in \mathcal{W}_{s,1}.$$

Therefore, $\mathcal{W}_{s,1}$ could only include a with the same value of $Q_\mu^*(s,a)$, which should only be a series of points rather than an interval. Thus, $\sigma(\mathcal{W}_{s,1}) = 0$, and $\pi_\mu^*(a|s)$ would converge to uniform distribution with interval length $\frac{1}{\mathbf{C}}$.

A.6 Proof of Lemma S.1

Before we prove the main result, we first provide a helper lemma for studying the boundedness of the symmetric kernel in the U-statistic.

Lemma S.1. *Under Assumption 1, for any $s \in \mathcal{S}, a \in \mathcal{A}$ and $\mu \in (0,\infty)$, we have that*

$$\sup_{s\in\mathcal{S}, a\in\mathcal{A}} \Big|\mathcal{G}_{V_\mu,\pi_\mu}(s,a,s') - \eta(s) + \varpi(s,a) - V_\mu(s)\Big| \leq M_{\max},$$

where $M_{\max} = \frac{4}{1-\gamma}R_{\max} + \mu\mathbf{C}$.

Proof of Lemma S.1:

$$\begin{aligned}
&\mathcal{G}_{V_\mu,\pi_\mu}(s,a,s') - \eta(s) + \varpi(s,a) - V_\mu(s)\\
=&R(s',s,a) + \gamma V_\mu(s') + \mu - 2\mu\pi_\mu(a|s) - \eta(s) + \varpi(s,a) - V_\mu(s)\\
\leq&R_{\max} + \mu + \mu\mathbf{C} + \gamma V_\mu(s') - V_\mu(s) \underbrace{-2\mu\pi_\mu(a|s) + \varpi(s,a)}_{(a)}.
\end{aligned}$$

By checking the KKT conditions, we can further simplify the term (a). Specifically,

1. If $\pi_\mu = 0$, then $\varpi \geq 0$. By the stationarity equation (10), we have

$$\begin{aligned}
(a) &= \varpi(s,a)\\
&= \eta(s) - Q_\mu(s,a) - \mu + V_\mu(s)\\
&\leq R_{\max} + \gamma\frac{R_{\max} - \mu H}{1-\gamma} - \mu + \frac{R_{\max} + \mu H}{1-\gamma} \quad \Big(H := \mathbb{E}_{a\sim\pi_\mu(\cdot|s)}(1-\pi_\mu(a|s))\Big)\\
&\leq \frac{2}{1-\gamma}R_{\max} - \mu + \mu H\\
&\leq \frac{2}{1-\gamma}R_{\max}.
\end{aligned}$$

2. If $\pi_\mu \in (0, \mathbf{C}]$, then $\varpi = 0$

$$(a) = -2\mu\pi_\mu(a|s) < 0.$$

Therefore,

$$\begin{aligned}
&\mathcal{G}_{\pi_\mu}(s,a,s') - \eta(s) + \varpi(s,a) - V_\mu(s) \\
\leq & R_{\max} + \mu + \mu\mathbf{C} + \gamma V_\mu(s') - V_\mu(s) + \frac{2}{1-\gamma}R_{\max} \\
\leq & R_{\max} + \mu + \mu\mathbf{C} + \gamma\frac{R_{\max} + \mu H}{1-\gamma} - \frac{-R_{\max} + \mu H}{1-\gamma} + \frac{2}{1-\gamma}R_{\max} \\
\leq & \frac{4}{1-\gamma}R_{\max} + \mu\mathbf{C} + \mu - \mu H \\
\leq & \frac{4}{1-\gamma}R_{\max} + \mu\mathbf{C}.
\end{aligned}$$

Thus, we gain the upper bound. For the lower bound, the same technique is applied, and we can also gain that

$$\mathcal{G}_{V_\mu,\pi_\mu}(s,a,s') - \eta(s) + \varpi(s,a) - V_\mu(s) \geq -\frac{4}{1-\gamma}R_{\max} - \mu\mathbf{C}.$$

Therefore, this completes the proof.

A.7 Proof of Theorem 4.2

Proof of Theorem 4.2: We first define an operator $\mathcal{P}$ from $\mathcal{G}_{V_\mu,\pi_\mu}(S^k, A^k, S^{k+1})$ to $\mathcal{G}_{V_\mu,\pi_\mu}(S^k, A^k, S^{k+1}) - \eta(S^k) + \varpi(S^k, A^k)$ to simplify the expression, such that

$$\mathcal{P}\mathcal{G}_{V_\mu\pi_\mu}(S^k, A^k, S^{k+1}) := \mathcal{G}_{V_\mu,\pi_\mu}(S^k, A^k, S^{k+1}) - \eta(S^k) + \varpi(A^k|S^k),$$

We further define several other notations

$$\begin{aligned}
U_T :=& \binom{T}{2}^{-1} \sum_{1\leq j\neq k\leq T} K(S^j, A^j; S^k, A^k)\{\mathcal{P}\mathcal{G}_{V_\mu,\pi_\mu}(S^j, A^j, S^{j+1}) - V_\mu(S^j)\}\cdot \\
& \{\mathcal{P}\mathcal{G}_{V_\mu,\pi_\mu}(S^k, A^k, A^{k+1}) - V_\mu(S^k)\}
\end{aligned}$$

$$\begin{aligned}
&\tilde{K}\left(S^t, A^t, S^{t+1}; \widetilde{S}^t, \widetilde{A}^t, \widetilde{S}^{t+1}\right) \\
&:= K\left(S^t, A^t; \widetilde{S}^t, \widetilde{A}^t\right)\left\{\mathcal{P}\mathcal{G}_{V_\mu,\pi_\mu}\left(S^t, A^t, S^{t+1}\right) - V_\mu\left(S^t\right)\right\}\left\{\mathcal{P}\mathcal{G}_{V_\mu,\pi_\mu}\left(\widetilde{S}^t, \widetilde{A}^t, \widetilde{S}^{t+1}\right) - V_\mu\left(\widetilde{S}^t\right)\right\}.
\end{aligned}$$

Let the expectation with respect to stationary trajectory and i.i.d training set as $\mathbb{E}_T$ and $\mathbb{E}$ respectively. For any finite threshold parameter $\mu < \infty$ and any $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(\left|\widehat{\mathcal{L}_U} - \mathcal{L}_U\right| > \epsilon\right) &= \mathbb{P}\left(\left|\widehat{\mathcal{L}_U} - \mathbb{E}\left(U_T\right) + \mathbb{E}\left(U_T\right) - \mathcal{L}_U\right| > \epsilon\right) \\
&\leq \underbrace{\mathbb{P}\left(\left|\widehat{\mathcal{L}_U} - \mathbb{E}\left(U_T\right)\right| > \frac{\epsilon}{2}\right)}_{(i)} + \underbrace{\mathbb{P}\left(\left|\mathbb{E}\left(U_T\right) - \mathcal{L}_U\right| > \frac{\epsilon}{2}\right)}_{(ii)}.
\end{aligned}$$

For (i), since the Gaussian kernel satisfy that $|K(\cdot;\cdot)| \leq 1$, then by Lemma S.1, we have

$$\tilde{K}\left(s, a, s'; \tilde{s}, \tilde{a}, \tilde{s'}\right) \leq M_{\max}^2,$$

for any $s, \tilde{s}, a, \tilde{a}$. By Hoeffding's inequality, we have

$$(i) \leq 2\exp\Big\{-\frac{n\epsilon^2}{2M_{\max}^4}\Big\}. \tag{27}$$

For the term (ii), the expectation of U_T as $\mathbb{E}_T(U_T)$ can be calculated as follows:

$$\begin{aligned}
\mathbb{E}_T(U_T) = \binom{T}{2}^{-1} \sum_{1\leq j\neq k\leq T} \mathbb{E}_T\Big[&K(S^j, A^j; S^k, A^k)\{\mathcal{P}\mathcal{G}_{V_\mu,\pi_\mu}(S^j, A^j, S^{j+1}) - V_\mu(S^j)\}\cdot \\
&\{\mathcal{P}\mathcal{G}_{V_\mu,\pi_\mu}(S^k, A^k, S^{k+1}) - V_\mu(S^j)\}\Big].
\end{aligned}$$

If with-in trajectory samples are independent, then it is obvious that

$$\mathbb{E}_T(U_T) = \mathbb{E}_T\left[\tilde{K}\left(S^t, A^t, S^{t+1}; \widetilde{S}^t, \widetilde{A}^t, \widetilde{S}^{t+1}\right)\right] := U^*.$$

However, for weakly dependent data, dependency may introduce an additional bias term $\mathbb{E}_T(U_T) - U^*$, thus we further decompose the term (ii) as

$$(ii) = \mathbb{P}(\underbrace{|\mathbb{E}\left(U_T\right) - \mathbb{E}\left[\mathbb{E}_T\left(U_T\right)\right]|}_{(iii)} + \underbrace{|\ \mathbb{E}\left[\mathbb{E}_T\left(U_T\right)\right] - \mathbb{E}U^\star)\ |}_{(iv)} > \frac{\epsilon}{2}).$$

For the term (iii), we follow a similar idea to use a novel decomposition of the variance term of U-statistic from Han (2018). The idea is to break down the summation of U-statistic into numerous parts, where the current time is affected by randomness, and the historical time will be canceled out after conditioning on the future.

As $|\tilde{K}(\cdot\ ;\ \cdot)|$ is bounded by $M_{\max}^2$, under the mixing condition of Assumption 4.2, the exponential inequality from Merlevède et al. (2009) can be applied to to bound each decomposition part.

Then we follow the Theorem 3.1 from Han (2018) that for any ϵ_0,

$$\mathbb{P}(|\mathbb{E}\left(U_T\right) - \mathbb{E}\left[\mathbb{E}_T\left(U_T\right)\right]| > \epsilon_0) \leq 2\exp\left\{-\left(\frac{M_{\max}^4}{T\epsilon_0^2 C_1'} + \frac{M_{\max}^2 \log\log(4T)\log T}{T\epsilon_0 C_1'}\right)^{-1}\right\}, \quad (28)$$

where C_1' is some constant.

Then, we proceed to bound the term (iv). By Hoeffding decomposition of kernel function $\tilde{K}\left(S^t, A^t, S^{t+1}; \tilde{S}^t, \tilde{A}^t, \tilde{S}^{t+1}\right)$, there exist kernel functions $\tilde{K}_1(S^t, A^t, S^{t+1})$ and $\tilde{K}_2\left(S^t, A^t, S^{t+1}; \tilde{S}^t, \tilde{A}^t, \tilde{S}^{t+1}\right)$ such that

$$\begin{aligned}
\tilde{K}_1\left(s, a, s'\right) &= \mathbb{E}_T\tilde{K}\left(s, a, s'; \widetilde{S}^t, \widetilde{A}^t, \widetilde{S}^{t+1}\right) - U^*, \\
\tilde{K}_2\left(s, a, s'; \widetilde{s}, \widetilde{a}, \widetilde{s'}\right) &= \tilde{K}\left(s, a, s'; \widetilde{s}, \widetilde{a}, \widetilde{s'}\right) - \tilde{K}_1\left(s, a, s'\right) - \tilde{K}_1\left(\widetilde{s}, \widetilde{a}, \widetilde{s'}\right) - U^*,
\end{aligned}$$

and $\mathbb{E}_T\tilde{K}_1(S^t, A^t, S^{t+1}) = 0$, $\mathbb{E}_T\tilde{K}_2\left(S^t, A^t, S^{t+1}; \tilde{S}^t, \tilde{A}^t, \tilde{S}^{t+1}\right) = 0$. Then by Hoeffding decomposition of U_T, we have

$$U_T = U^* + \frac{2}{n}\sum_{t=1}^{T}\tilde{K}_1(S^t, A^t, S^{t+1}) + U_{\tilde{K}_2}.$$

Taking the expectation from both sides:

$$\begin{aligned}
\mathbb{E}_T[U_T] &= U^* + \frac{2}{n}\sum_{k=1}^{T}\mathbb{E}_T\tilde{K}_1(S^t, A^t, S^{t+1}) + \mathbb{E}_T[U_{\tilde{K}_2}] \\
&= U^* + \mathbb{E}_T[U_{\tilde{K}_2}]
\end{aligned}$$

Therefore, by Lyapunov inequality, we can bound the bias term

$$\begin{aligned}
&|\mathbb{E}_T\left[U_T\right] - U^\star| = \left|\mathbb{E}_T\left[U_{\tilde{K}_2}\right]\right| \leq \mathbb{E}_T\left[\left|U_{\tilde{K}_2}\right|\right] \leq \sqrt{\mathbb{E}_T\left[U_{\tilde{K}_2}^2\right]} \\
&= \sqrt{\sum_{1\leq h_1\leq l_1\leq T, 1\leq h_2\leq l_2\leq T}\mathbb{E}_T\Big[\tilde{K}_2\left(S^{h_1}, A^{h_1}, S^{h_1+1}; S^{l_1}, A^{l_1}, S^{l_1+1}\right)} \\
&\overline{\cdot\tilde{K}_2\left(S^{h_2}, A^{h_2}, S^{h_2+1}; S^{l_2}, A^{l_2}, S^{l_2+1}\right)\Big]\frac{4}{T^2(T-1)^2}}.
\end{aligned} \quad (29)$$

We proceed by the discussing the relationship between h_1, h_2, l_1, l_2.

Case 1.1: If $1 \leq h_1 \leq h_2 \leq l_1 \leq l_2 \leq T$ and $l_2 - l_1 \leq h_1 - h_2$.
Under the mixing condition assumption, and by Generalized Correlation inequality in Lemma 2 of, we have

$$\begin{aligned}&\left|\mathbb{E}_T\left[\tilde{K}_2\left(S^{h_1}, A^{h_1}, S^{h_1+1}; S^{l_1}, A^{l_1}, S^{l_1+1}\right)\tilde{K}_2\left(S^{h_2}, A^{h_2}, S^{h_2+1}; S^{l_2}, A^{l_2}, S^{l_2+1}\right)\right]\right| \\ \leq & 4\left(M_{\max}^{2r}\right)^{1/r}\beta^{1/s}\left(h_2 - h_1\right),\end{aligned}$$

where $1/r + 1/s = 1, s > -1$.

Case 1.2: If $1 \leq h_1 \leq h_2 \leq l_1 \leq l_2 \leq T$ and $h_1 - h_2 \leq l_2 - l_1$.
Similar as Case 1.1, we have

$$\begin{aligned}&\left|\mathbb{E}_T\left[\tilde{K}_2\left(S^{h_1}, A^{h_1}, S^{h_1+1}; S^{l_1}, A^{l_1}, S^{l_1+1}\right)\tilde{K}_2\left(S^{h_2}, A^{h_2}, S^{h_2+1}; S^{l_2}, A^{l_2}, S^{l_2+1}\right)\right]\right| \\ \leq & 4\left(M_{\max}^{2r}\right)^{1/r}\beta^{1/s}\left(l_2 - l_1\right).\end{aligned}$$

Combine Case 1.1 and Case 1.2, we apply the bounded inequalities (2.17-2.21) from Yoshihara (1976), and have the following result

$$\begin{aligned}&\Big|\sum_{1\leq h_1\leq h_2\leq l_1\leq l_2\leq T}\mathbb{E}_T\Big[\tilde{K}_2\left(S^{h_1}, A^{h_1}, S^{h_1+1}; S^{l_1}, A^{l_1}, S^{l_1+1}\right) \\ &\tilde{K}_2\left(S^{h_2}, A^{h_2}, S^{h_2+1}; S^{l_2}, A^{l_2}, S^{l_2+1}\right)\Big]\Big| \\ &\leq \sum_{\substack{l_2-l_1\leq h_2-h_1\\ 1\leq h_1\leq h_2\leq l_1\leq l_2\leq T}}\Big|\mathbb{E}_T\Big[\tilde{K}_2\left(S^{h_1}, A^{h_1}, S^{h_1+1}; S^{l_1}, A^{l_1}, S^{l_1+1}\right) \\ &\cdot\tilde{K}_2\left(S^{h_2}, A^{h_2}, S^{h_2+1}; S^{l_2}, A^{l_2}, S^{l_2+1}\right)\Big]\Big| + \\ &\sum_{\substack{h_2-h_1\leq l_2-l_2\\ 1\leq h_1\leq h_2\leq l_1\leq l_2\leq T}}\Big|\mathbb{E}_T\Big[\tilde{K}_2\left(S^{h_1}, A^{h_1}, S^{h_1+1}; S^{l_1}, A^{l_1}, S^{l_1+1}\right) \\ &\tilde{K}_2\left(S^{h_2}, A^{h_2}, S^{h_2+1}; S^{l_2}, A^{l_2}, S^{l_2+1}\right)\Big]\Big| \\ &\leq M_{\max}^2 T^2\sum_{j=1}^{T}(j+1)\beta^{1/s}(j) = \mathcal{O}\left(M_{\max}^2 T^{3-\tau}\right),\end{aligned}$$

where

$$\tau = \frac{\left(\frac{2}{s+1} - \frac{2}{1-\delta_1}\right)}{\left(\frac{1}{\delta_1-1}\right)\left(1+\frac{1}{s+1}\right)}. \tag{30}$$

Case 2: If $1 \leq h_1 \leq l_1 \leq h_2 \leq l_2 \leq T$.
Using similar technique as Case 1.1 and 1.2, we have

$$\begin{aligned}&\Big|\sum_{1\leq h_1\leq l_1\leq h_2\leq l_2\leq T}\mathbb{E}_T\Big[\tilde{K}_2\left(S^{h_1}, A^{h_1}, S^{h_1+1}; S^{l_1}, A^{l_1}, S^{l_1+1}\right) \\ &\tilde{K}_2\left(S^{h_2}, A^{h_2}, S^{h_2+1}; S^{l_2}, A^{l_2}, S^{l_2+1}\right)\Big]\Big| \\ &\leq \sum_{\substack{l_2-h_2\leq l_1-h_1\\ 1\leq h_1\leq l_1\leq h_2\leq l_2\leq T}}\Big|\mathbb{E}_T\Big[\tilde{K}_2\left(S^{h_1}, A^{h_1}, S^{h_1+1}; S^{l_1}, A^{l_1}, S^{l_1+1}\right) \\ &\cdot\tilde{K}_2\left(S^{h_2}, A^{h_2}, S^{h_2+1}; S^{l_2}, A^{l_2}, S^{l_2+1}\right)\Big]\Big| + \\ &\sum_{\substack{l_1-h_1\leq l_2-h_2\\ 1\leq h_1\leq l_1\leq h_1\leq l_2\leq T}}\Big|\mathbb{E}_T\Big[\tilde{K}_2\left(S^{h_1}, A^{h_1}, S^{h_1+1}; S^{l_1}, A^{l_1}, S^{l_1+1}\right) \\ &\tilde{K}_2\left(S^{h_2}, A^{h_2}, S^{h_2+1}; S^{l_2}, A^{l_2}, S^{l_2+1}\right)\Big]\Big| \\ &= \mathcal{O}\left(M_{\max}^2 T^{3-\tau}\right)\end{aligned}$$

Case 3: If $1 \leq h_1 \leq l_1 \leq T$ and $1 \leq h_2 = l_2 \leq T$.
Following the same technique, we have

$$\begin{aligned}
&\Big| \sum_{1\leq h_2=l_2\leq T} \sum_{1\leq h_1\leq l_1\leq T} \mathbb{E}_T \Big[\tilde{K}_2\left(S^{h_1}, A^{h_1}, S^{h_1+1}; S^{l_1}, A^{l_1}, S^{l_1+1}\right) \\
&\cdot \tilde{K}_2\left(S^{h_2}, A^{h_2}, S^{h_2+1}; S^{l_2}, A^{l_2}, S^{l_2+1}\right)\Big] \Big| \\
&\leq \sum_{1\leq h_1=l_1\leq T} \sum_{1\leq h_2=l_2\leq T} \Big|\mathbb{E}_T \Big[\tilde{K}_2\left(S^{h_1}, A^{h_1}, S^{h_1+1}; S^{l_1}, A^{l_1}, S^{l_1+1}\right) \\
&\tilde{K}_2\left(S^{h_2}, A^{h_2}, S^{h_2+1}; S^{l_2}, A^{l_2}, S^{l_2+1}\right)\Big] \Big| + \\
&2 \sum_{1\leq h_1<l_1\leq T} \sum_{1\leq h_2=l_2\leq T} \Big|\mathbb{E}_T \Big[\tilde{K}_2\left(S^{h_1}, A^{h_1}, S^{h_1+1}; \S^{l_1}, A^{l_1}, S^{l_1+1}\right) \\
&\cdot \tilde{K}_2\left(S^{h_2}, A^{h_2}, S^{h_2+1}; S^{l_2}, A^{l_2}, S^{l_2+1}\right)\Big] \Big| \\
&\leq U_{\max}^2 T^2 + M_{\max}^2 T^2 \sum_{j=1}^{T} \beta^{1/s}(j) = \mathcal{O}\left(M_{\max}^2 T^2\right).
\end{aligned}$$

Case 4: If $1 \leq h_1 = l_1 \leq T$ and $1 \leq h_2 \leq l_2 \leq T$.
Using the same technique, we can obtain the same rate as follows:

$$\begin{aligned}
&\Big| \sum_{1\leq h_1=l_1\leq T} \sum_{1\leq h_2\leq l_2\leq T} \mathbb{E}_T \Big[\tilde{K}_2\left(S^{h_1}, A^{h_1}, S^{h_1+1}; S^{l_1}, A^{l_1}, S^{l_1+1}\right) \\
&\cdot \tilde{K}_2\left(S^{h_2}, A^{h_2}, S^{h_2+1}; S^{l_2}, A^{l_2}, S^{l_2+1}\right)\Big] \Big| \\
&= \mathcal{O}\left(M_{\max}^2 T^2\right).
\end{aligned}$$

Combine Case 1-4 with the equation (29), we conclude that

$$|\mathbb{E}U_T - U^*| \leq C_0' M_{\max}^2 T^{-\frac{1+\tau}{2}} \quad a.s.$$

We further use the continuous mapping theorem to conclude that

$$\left|\mathbb{E}[\mathbb{E}_T(U_T)] - \mathbb{E}U^*\right| \leq C_0' M_{\max}^2 T^{-\frac{1+\tau}{2}} \quad a.s., \tag{31}$$

where τ is defined in (30) and C_0' is a constant.

As $\tau > 0$, we have $T^{-\frac{1+\tau}{2}} < T^{-\frac{1}{2}}$. Combine (28) and (31), for sufficiently large T, we have

$$\begin{aligned}
(ii) &= \mathbb{P}\left(|\mathbb{E}\left(U_T\right) - \mathbb{E}\left[\mathbb{E}_T\left(U_T\right)\right]| + |\,\mathbb{E}\left[\mathbb{E}_T\left(U_T\right)\right] - \mathbb{E}U^\star)\,| > \frac{\epsilon}{2}\right) \\
&\leq 2\exp\left(-\frac{TC_1'\left(\epsilon/2 - C_0' M_{\max}^2 T^{-(1+\tau)/2}\right)^2}{M_{\max}^4 + M_{\max}^2\left(\epsilon/2 - C_0' M_{\max}^2 T^{-(1+\tau)/2}\right)\log T \log\log 4T}\right) \\
&= 2\exp\left(-\frac{TC_1'\epsilon^2/4 - Tc_1\epsilon C_0' M_{\max}^2 T^{-(1+\tau)/2} + TC_1' {C_0'}^2 M_{\max}^4 T^{-(1+\tau)}}{M_{\max}^4 + M_{\max}^2\left(\epsilon/2 - C_0' M_{\max}^2 T^{-(1+\tau)/2}\right)\log T \log\log 4T}\right) \\
&= 2\exp\left(-\frac{Tc_1\epsilon^2/4 - TT^{-(1+\tau)/2} c_1\epsilon C_0' M_{\max}^2 + c_1 {C_0'}^2 M_{\max}^4 T^{-\tau}}{M_{\max}^4 + M_{\max}^2\left(\epsilon/2 - C_0' M_{\max}^2 T^{-(1+\tau)/2}\right)\log T \log\log 4T}\right)
\end{aligned} \tag{32}$$

Then by the monotonicity of $\exp(\cdot)$,

$$\begin{aligned}
&\frac{TT^{-(1+\tau)/2} C_1'\epsilon C_0' M_{\max}^2 - T^{-\tau} C_1' {C_0'}^2 M_{\max}^4 - TC_1'\epsilon^2/4}{M_{\max}^4 + \log T \log\log 4T M_{\max}^2 \epsilon/2 - T - (1+\tau)/2 \log T \log\log 4T C_0' M_{\max}^4} \\
\leq &- \frac{TC_1'\epsilon^2/4 - T^{1/2} C_1'\epsilon C_0' M_{\max}^2 + T^{-\tau} C_1' {C_0'}^2 M_{\max}^4}{M_{\max}^4 + \log T \log\log 4T M_{\max}^2 \epsilon/2 - T^{-1/2} \log T \log\log 4T C_0' M_{\max}^4} \\
\leq &- \frac{cC_1'\epsilon^2 T/4 - C_0' C_1'\epsilon M_{\max}^2 \sqrt{T}}{M_{\max}^2\left(\epsilon/2 - C_0' M_{\max}^2/\sqrt{T}\right)\log T \log\log 4T + M_{\max}^4}
\end{aligned} \tag{33}$$

where C_1' is a constant. Combine (27) and (33), we simplify the terms and then

$$\mathbb{P}(|\widehat{\mathcal{L}_U} - \mathcal{L}_U| > \epsilon) \leq C_1 \exp\left(- \frac{\epsilon^2 T - C_2 \epsilon M_{\max}^2 \sqrt{T}}{M_{\max}^2 + (\frac{\epsilon}{2} - \frac{C_2 M_{\max}^2}{\sqrt{T}}) \log T \log\log(T)} \right) + C_3 \exp\left(\frac{-n\epsilon^2}{M_{\max}^4} \right),$$

where C_1, C_2, C_3 are some constants depending on δ_1 respectively, and $M_{\max} = \frac{4}{1-\gamma} R_{\max} + \mu \mathbf{C}$.

A.8 Proof of Theorem 4.3

Proof of Theorem 4.3. To bound the performance error, we first decompose it as

$$\|\widehat{V}_\mu^{\theta_1,k} - V^*\|_{L^2}^2 \leq \|\widehat{V}_\mu^{\theta_1} - V_\mu^{\theta_1,k}\|_{L^2}^2 + \|\widehat{V}_\mu^{\theta_1} - V^*\|_{L^2}^2 + \epsilon_{\text{approximation error}}$$

where the first term is the optimization error and the last term is the approximation error. Then we proceed to bound

$$\|\widehat{V}_\mu^{\theta_1} - V^*\|_{L^2}^2 \leq \left| \| \underbrace{\widehat{V}_\mu^{\theta_1} - V_\mu^{\widetilde{\pi}_\mu}\|_{L^2}}_{\Delta_1} + \underbrace{\|V_\mu^{\widetilde{\pi}_\mu} - V^*\|_{L^2}}_{\Delta_2} \right|^2.$$

where $V_\mu^{\widetilde{\pi}_\mu}$ satisfying the stationarity equation (10) and V^* is the unique fixed point of $\mathcal{B}$. First, we move to bound Δ_1. Follow a similar kernel reproducing property and a eigen decomposition spirit in Bertsekas (1997); Sutton & Barto (2018); Zhou et al. (2021a), we have

$$\begin{aligned}
&\frac{2}{\kappa_{\min}} \big(\mathcal{L}_U(\widehat{V}_\mu^{\theta_1}, \widehat{\pi}_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) - \mathcal{L}_U(V_\mu^{\widetilde{\pi}_\mu}, \widetilde{\pi}_\mu, \eta, \varpi)\big) + \\
&2\| \big(\mu \text{prox}^\circ(\widehat{\pi}_\mu^{\theta_2}(A^t|S^t)) - \mu \text{prox}^\circ(\widetilde{\pi}_\mu(A^t|S^t))\big) - \big(\widehat{\eta}^{\xi_1}(S^t) - \eta(S^t)\big) + \big(\widehat{\varpi}^{\xi_2}(S^t, A^t) - \varpi(S^t, A^t)\big)\|_{L^2}^2 \\
\geq &\|\gamma\big(\mathbb{E}_{S^{t+1}|S^t,A^t}[\widehat{V}_\mu^{\theta_1}(S^{t+1})] - \mathbb{E}_{S^{t+1}|S^t,A^t}[V_\mu^{\widetilde{\pi}_\mu}(S^{t+1})]\big) - \big(\widehat{V}_\mu^{\theta_1}(S^t) - V_\mu^{\widetilde{\pi}_\mu}(S^t)\big)\|_{L^2}^2.
\end{aligned}$$

Then by

$$\|\mu \text{prox}^\circ(\widehat{\pi}_\mu^{\theta_2}(A^t|S^t)) - \mu \text{prox}^\circ(\widetilde{\pi}_\mu(A^t|S^t))\|_{L^2}^2 \leq \mu^2 \|\widehat{\pi}_\mu^{\theta_2}(A^t|S^t) - \widetilde{\pi}_\mu(A^t|S^t)\|_{L^2}^2 \leq \mathbf{C}\mu^2.$$

and the auxiliary functions $\eta^{\xi_1}(s) \in [-\mathbf{C}\mu, 0]$ for any $s \in \mathcal{S}$, then

$$\begin{aligned}
&\|\widehat{\eta}^{\xi_1}(S^t) - \eta(S^t)\|_{L^2}^2 \leq (\mathbf{C}\mu + \mathbf{C}\mu)^2 = (\mathbf{C}\mu)^2 \\
&\|\widehat{\eta}_1^{\xi_1}(S^t) - \eta_1(S^t)\|_{L^2}^2 \leq \frac{2}{\kappa_{\min}} \big(\mathcal{L}_U(\widehat{V}_\mu^{\theta_1}, \widehat{\pi}_\mu^{\theta_2}, \widehat{\eta}^{\xi_1}, \widehat{\varpi}^{\xi_2}) - \mathcal{L}_U(V_\mu^{\widetilde{\pi}_\mu}, \widetilde{\pi}_\mu, \eta, \varpi)\big)
\end{aligned}$$

Then we conclude that

$$\begin{aligned}
\|\widehat{V}_\mu^{\theta_1}(S^t) - V_\mu^{\widetilde{\pi}_\mu}(S^t)\|_{L^2}^2 \leq &\frac{C_5(\mathcal{L}_U(\widehat{V}_\mu^{\theta_1}, \widehat{\pi}_\mu^{\theta_2}, \widehat{\eta}^{\xi_1}, \widehat{\varpi}^{\xi_2}) - \mathcal{L}_U(V_\mu^{\widetilde{\pi}_\mu}, \widetilde{\pi}_\mu, \eta, \varpi))}{\kappa_{\min}(1-\gamma)^2} + \frac{C_6 \mu^2}{(1-\gamma)^2} \\
\leq &\frac{C_5(\mathcal{L}_U(\widehat{V}_\mu^{\theta_1}, \widehat{\pi}_\mu^{\theta_2}, \widehat{\eta}^{\xi_1}, \widehat{\varpi}^{\xi_2}) - \mathcal{L}_U^*}{\kappa_{\min}(1-\gamma)^2} + \frac{C_6 \mu^2}{(1-\gamma)^2}
\end{aligned}$$

where C_5 and C_6 are some constants, and

$$\mathcal{L}_U^* := \inf_{\{V_\mu, \pi_\mu, \eta, \varpi\}} \mathcal{L}_U(V_\mu, \pi_\mu, \eta, \varpi)$$

Now, we have the remainder term Δ_2 to bound.

$$\Delta_2 \leq \underbrace{\|V_\mu^{\widetilde{\pi}_\mu} - V_\mu^*\|_{L^2}}_{\Delta_2^1} + \underbrace{\|V_\mu^* - V^*\|_{L^2}}_{\Delta_2^2}$$

We first bound Δ_2^1. For any $s \in \mathcal{S}$, then we have that

$$\begin{aligned}
\mathcal{B}_\mu V_\mu^{\widetilde{\pi}_\mu}(s) = &\max_\pi \mathbb{E}_{a \sim \pi(\cdot|s),\ S^{t+1}|s,a} \Big[R(S^{t+1}, s, a) + \gamma V_\mu^{\widetilde{\pi}_\mu}(S^{t+1}) + \mu \text{prox}(\pi(a|s)) \Big] \\
= &\mathbb{E}_{a \sim \widetilde{\pi}_\mu(\cdot|s),\ S^{t+1}|s,a} \Big[R(S^{t+1}, s, a) + \gamma V_\mu^{\widetilde{\pi}_\mu}(S^{t+1}) + \mu \text{prox}(\widetilde{\pi}_\mu(a|s)) \Big] \\
= &\mathbb{E}_{a \sim \widetilde{\pi}_\mu(\cdot|s),\ S^{t+1}|s,a} \Big[R(S^{t+1}, s, a) + \gamma V_\mu^{\widetilde{\pi}_\mu}(S^{t+1}) + \mu(1 - \widetilde{\pi}_\mu(a|s)) \Big] \\
= &\mathbb{E}_{a \sim \widetilde{\pi}_\mu(\cdot|s),\ S^{t+1}|s,a} \Big[R(S^{t+1}, s, a) + \gamma V_\mu^{\widetilde{\pi}_\mu}(S^{t+1}) + \mu - \mu \widetilde{\pi}_\mu(a|s) \Big] + \\
&\mathbb{E}_{a \sim \widetilde{\pi}_\mu(\cdot|s)} \Big[\mu \widetilde{\pi}_\mu(a|s) \Big].
\end{aligned}$$

As $(V_\mu^{\widetilde{\pi}_\mu}, \widetilde{\pi}_\mu)$ is the solution of the stationarity equation,

$$\mathbb{E}_{a\sim\widetilde{\pi}_\mu(\cdot|s),\ S^{t+1}|s,a}\Big[R(S^{t+1},s,a)+\gamma V_\mu^{\widetilde{\pi}_\mu}(S^{t+1})+\mu-\mu\widetilde{\pi}_\mu(a|s)\Big]\leq V_\mu^{\widetilde{\pi}_\mu}(s)$$

and since $\mathbb{E}_{a\sim\widetilde{\pi}_\mu(\cdot|s)}\Big[\mu\widetilde{\pi}_\mu(a|s)\Big]\leq\mu$, then we have

$$\mathcal{B}_\mu V_\mu^{\widetilde{\pi}_\mu}(s)\leq V_\mu^{\widetilde{\pi}_\mu}(s)+\mu\mathbf{C}.$$

For the lower bound, as

$$\mathbb{E}_{a\sim\widetilde{\pi}_\mu(\cdot|s),\ S^{t+1}|s,a}\Big[R(S^{t+1},s,a)+\gamma V_\mu^{\widetilde{\pi}_\mu}(S^{t+1})+\mu-\mu\widetilde{\pi}_\mu(a|s)-V_\mu^{\widetilde{\pi}_\mu}(s)\mid S^t=s\Big]\geq-\mathbf{C}\mu$$

so similarly, we conclude that

$$\mathbf{C}\mu+\mathcal{B}_\mu V_\mu^{\widetilde{\pi}_\mu}(s)\geq V_\mu^{\widetilde{\pi}_\mu}(s).$$

If follows the definition of the proximal Bellman operator $\mathcal{B}_\mu$ and due to the monotonicity of the Bellman operator that $\mathcal{B}_\mu V_1(s)\geq\mathcal{B}_\mu V_2(s)$ for generic value functions $V_1(s)\geq V_2(s)$, and the $\mathcal{B}_\mu V(s)\geq\mathcal{B}_\mu^{\widetilde{\pi}_\mu}V(s)$ for any generic value function V, where $\mathcal{B}_\mu^{\widetilde{\pi}_\mu}$ is the proximal Bellman evaluation operator, i.e.,

$$\mathcal{B}_\mu^{\widetilde{\pi}_\mu}V(s):=\mathbb{E}_{a\sim\widetilde{\pi}_\mu(\cdot|s),\ S^{t+1}|s,a}\Big[R(S^{t+1},s,a)+\gamma V(S^{t+1})+\mu\text{prox}\big(\widetilde{\pi}_\mu(a|s)\big)\Big].$$

Note that, $V_\mu^{\widetilde{\pi}_\mu}$ is unique fixed point of the Bellman operator $\mathcal{B}_\mu^{\widetilde{\pi}_\mu}$, thus $\lim_{i\to\infty}(\mathcal{B}_\mu^{\widetilde{\pi}_\mu})^iV_\mu^{\widetilde{\pi}_\mu}(s)=V_\mu^{\widetilde{\pi}_\mu}(s)$, where $i\in\mathbb{Z}^+$. And for any initial value function. e.g., $V_\mu^{\widetilde{\pi}_\mu}$, $\lim_{i\to\infty}(\mathcal{B}_\mu)^iV_\mu^{\widetilde{\pi}_\mu}(s)=V_\mu^*(s)$ holds. Therefore the following inequality holds that

$$\begin{aligned}&V_\mu^{\widetilde{\pi}_\mu}(s)=\lim_{i\to\infty}(\mathcal{B}_\mu^{\widetilde{\pi}_\mu})^iV_\mu^{\widetilde{\pi}_\mu}(s)\leq\lim_{i\to\infty}(\mathcal{B}_\mu^{\widetilde{\pi}_\mu})^i\Big(V_\mu^{\widetilde{\pi}_\mu}+\mathbf{C}\mu\Big)(s)\leq\lim_{i\to\infty}(\mathcal{B}_\mu)^i\Big(V_\mu^{\widetilde{\pi}_\mu}+\mathbf{C}\mu\Big)(s)\\ \implies&V_\mu^{\widetilde{\pi}_\mu}(s)\leq\lim_{i\to\infty}(\mathcal{B}_\mu)^iV_\mu^{\widetilde{\pi}_\mu}(s)+\sum_{i=1}^{\infty}\mathbf{C}\mu\gamma^{i-1}\leq V_\mu^*(s)+\frac{\mathbf{C}\mu}{(1-\gamma)}.\end{aligned}\tag{34}$$

We repeatedly apply a similar procedure, without loss of generality. We first show one step that

$$\mathcal{B}_\mu(\mathcal{B}_\mu V_\mu^{\widetilde{\pi}_\mu}(s))\leq\mathcal{B}_\mu(V_\mu^{\widetilde{\pi}_\mu}(s)+\mathbf{C}\mu)=\mathcal{B}_\mu(V_\mu^{\widetilde{\pi}_\mu}(s))+\mathbf{C}\mu\gamma\leq V_\mu^{\widetilde{\pi}_\mu}(s)+\mathbf{C}\mu+\mathbf{C}\mu\gamma.$$

Then we apply infinite many time $\mathcal{B}_\mu$, then we can have that

$$V_\mu^*(s)=\lim_{i\to\infty}(\mathcal{B}_\mu)^iV_\mu^{\widetilde{\pi}_\mu}(s)\leq V_\mu^{\widetilde{\pi}_\mu}(s)+\sum_{i=1}^{\infty}\mathbf{C}\mu\gamma^{i-1}=V_\mu^{\widetilde{\pi}_\mu}(s)+\frac{\mathbf{C}\mu}{(1-\gamma)}.\tag{35}$$

Combine with the inequalities (34)-(35), we immediately have that

$$\|V_\mu^*-V_\mu^{\widetilde{\pi}_\mu}\|_{L^2}\leq\frac{\mathbf{C}\mu}{(1-\gamma)}$$

Next, by Proposition S.2, we have

$$\|V_\mu^*-V^*\|_\infty\leq\frac{\mu\cdot\max\{|1-\mathbf{C}|,1\}}{1-\gamma},$$

Now, we need to bound the excess risk. The excess risk can be decomposed into approximation error and estimation error, i.e.

$$\begin{aligned}\mathcal{L}_U(\widehat{V}_\mu^{\theta_1},\widehat{\pi}_\mu^{\theta_2},\widehat{\eta}^{\xi_1},\widehat{\varpi}^{\xi_2})-\mathcal{L}_U^*=&\underbrace{\left(\inf_{(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2})\in\Theta_1\times\Theta_2\times\Xi_1\times\Xi_2}\mathcal{L}_U(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2})-\mathcal{L}_U^*\right)}_{\Delta_{\text{approx}}}\\&+\underbrace{\left(\mathcal{L}_U\left(\widehat{V}_\mu^{\theta_1},\widehat{\pi}_\mu^{\theta_2},\widehat{\eta}^{\xi_1},\widehat{\varpi}^{\xi_2}\right)-\inf_{(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2})\in\Theta_1\times\Theta_2\times\Xi_1\times\Xi_2}\mathcal{L}_U(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2})\right)}_{\Delta_{\text{est}}},\end{aligned}$$

where Δ_{approx} is the approximation error and Δ_{est} is the estimation error. The approximation error is assumed to be zero in our proof for simplicity. At first, we consider to bound the estimation error.

$$
\begin{aligned}
&\mathcal{L}_U\left(\widehat{V}_\mu^{\theta_1}, \widehat{\pi}_\mu^{\theta_2}, \widehat{\eta}^{\xi_1}, \widehat{\varpi}^{\xi_2}\right) - \inf_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \mathcal{L}_U(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \\
&:= \mathcal{L}_U\left(\widehat{V}_\mu^{\theta_1}, \widehat{\pi}_\mu^{\theta_2}, \widehat{\eta}^{\xi_1}, \widehat{\varpi}^{\xi_2}\right) - \mathcal{L}_U(V_\mu^{\pi_\mu^\circ}, \pi_\mu^\circ, \eta^{\xi_1}, \varpi^{\xi_2}) \\
&\leq \mathcal{L}_U\left(\widehat{V}_\mu^{\theta_1}, \widehat{\pi}_\mu^{\theta_2}, \widehat{\eta}^{\xi_1}, \widehat{\varpi}^{\xi_2}\right) - \mathcal{L}_U(V_\mu^{\pi_\mu^\circ}, \pi_\mu^\circ, \eta^{\xi_1}, \varpi^{\xi_2}) + \widehat{\mathcal{L}_U}(V_\mu^{\pi_\mu^\circ}, \pi_\mu^\circ, \eta^{\xi_1}, \varpi^{\xi_2}) - \widehat{\mathcal{L}_U}\left(\widehat{V}_\mu^{\theta_1}, \widehat{\pi}_\mu^{\theta_2}, \widehat{\eta}^{\xi_1}, \widehat{\varpi}^{\xi_2}\right) \\
&\leq \left(\mathcal{L}_U\left(\widehat{V}_\mu^{\theta_1}, \widehat{\pi}_\mu^{\theta_2}, \widehat{\eta}^{\xi_1}, \widehat{\varpi}^{\xi_2}\right) - \widehat{\mathcal{L}_U}\left(\widehat{V}_\mu^{\theta_1}, \widehat{\pi}_\mu^{\theta_2}, \widehat{\eta}^{\xi_1}, \widehat{\varpi}^{\xi_2}\right)\right) - \left(\mathcal{L}_U(V_\mu^{\pi_\mu^\circ}, \pi_\mu^\circ, \eta^{\xi_1}, \varpi^{\xi_2}) - \widehat{\mathcal{L}_U}(V_\mu^{\pi_\mu^\circ}, \pi_\mu^\circ, \eta^{\xi_1}, \varpi^{\xi_2})\right) \\
&\leq 2 \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left|\mathcal{L}_U(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) - \widehat{\mathcal{L}_U}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})\right|.
\end{aligned}
$$

where $\eta^{\xi_1}, \varpi^{\xi_2}$ are Lagrange multipliers satisfying minimal Bayes risk associated with $V_\mu^{\pi_\mu^\circ}, \pi_\mu^\circ$ for the rest of this proof. Observe that the randomness of $\sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2 |\mathcal{L}_U(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) - \widehat{\mathcal{L}_U}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})|}$ can be decomposed into two parts, one is from the n number of i.i.d. trajectories and another one is from the dependent transition within each trajectory. For each single trajectory, we define the quantity

$$
U^\star(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) = \Lambda_{V_\mu^{\theta_1}, \pi_\mu^{\theta_2}}(S_i^t, A_i^t, S_i^{t+1}) K(S_i^t, A_i^t; \widetilde{S}_i^t, \widetilde{A}_i^t) \Lambda_{V_\mu^{\theta_1}, \pi_\mu^{\theta_2}}(\widetilde{S}_i^t, \widetilde{A}_i^t, V_\mu^{\theta_1}),
$$

where $\mathbb{E}_T$ is defined as taking expectation to single stationary trajectory and $\mathbb{E}$ is defined as taking expectation to i.i.d. trajectory random variable $\mathcal{D}_1$, respectively. Without loss of generality, we assume $C_0 = 1$. The U-statistic approximation for $\mathbb{E}_T(U^\star)$ is as follows:

$$
\begin{aligned}
&U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \\
&:= \frac{2}{T(T-1)} \sum_{1 \leq j \neq k \leq T} \left[\left(\Lambda_{V_\mu^{\theta_1}, \pi_\mu^{\theta_2}}(S_i^j, A_i^j, S_i^{j+1}) K\left(S_i^j, A_i^j; S_i^k, A_i^k\right)\left(\Lambda_{V_\mu^{\theta_1}, \pi_\mu^{\theta_2}}(S_i^k, A_i^k, V_\mu^{\theta_1})\right)\right].
\end{aligned}
$$

Then the uniform process is bounded by

$$
\begin{aligned}
&\sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left|\mathcal{L}_U(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) - \widehat{\mathcal{L}_U}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})\right| \\
&\leq \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left|\mathcal{L}_U(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) - \mathbb{P}_n^{(\mathcal{D}_{i:n})} \mathbb{E}_T[U^\star(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})]\right| \\
&+ \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left|\mathbb{P}_n^{(\mathcal{D}_{i:n})} \mathbb{E}_T[U^\star(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] - \mathbb{P}_n^{(\mathcal{D}_{i:n})} U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})]\right| \\
&\leq \underbrace{\sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left|\mathcal{L}_U(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) - \mathbb{P}_n^{(\mathcal{D}_{i:n})} \mathbb{E}_T[U^\star(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})]\right|}_{\Delta_1} \\
&+ \frac{1}{n} \sum_{i=1}^n \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left|\mathbb{E}_T[U^\star(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] - U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})]\right|,
\end{aligned}
$$

where $\mathbb{P}_n^{(\mathcal{D}_{i:n})}$ is the empirical measure with respect to $\mathcal{D}_{i:n} = \{\mathcal{D}_i\}_{i=1}^n$ and we simply denotes it as $\mathbb{P}_n$ in the following proof. The last term is the bound for uniform process w.r.t sum of trajectories. In this sense, it is necessary to bound

$$
\Delta_2 = \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left|\mathbb{E}_T[U^\star(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] - U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})]\right|,
$$

since the trajectories $\{\mathcal{D}_i\}_{i=1}^n$ are i.i.d. Now, we process to bound Δ_1. Δ_1 can be re-expressed as the empirical process of $\{\mathcal{D}_i\}_{i=1}^n$ w.r.t. the probability space $(\Omega_N, \mathcal{F}_N, \mathbb{P})$ equipped with empirical measure $\mathbb{P}_n$ such that

$$
\begin{aligned}
\Delta_1 &= \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left|\mathbb{E}(\mathbb{E}_T[U^\star(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})]) - \mathbb{P}_n \mathbb{E}_T[U^\star(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})]\right| \\
&= \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left|\mathbb{E} G(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i) - \mathbb{P}_n G(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i)\right|,
\end{aligned}
$$

where $G(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i)$ is the random function associated with random variable $\mathcal{D}_i$. To bound Δ_1, it is needed to calculate the covering number $\mathcal{N}\left(\epsilon, \mathcal{F}_{\theta,\xi}, \{\mathcal{D}_i\}_{i=1}^n\right)$ by Pollard's tail inequality (Pollard, 2012), where the function space is the composite space $\mathcal{F}_{\theta,\xi} = G(\Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2)$. Specifically, $G(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i) = \mathbb{E}_T\big[M(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i)K(\{S_i^t, A_i^t\}, \{\widetilde{S}_i^t, \widetilde{A}_i^t\})$ $\widetilde{M}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i)\big]$, where $M(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i) = \mu_{V_\mu^{\theta_1}, \pi_\mu^{\theta_2}}(S_i^t, A_i^t, S_i^{t+1})$ and $\widetilde{M}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i) = \Lambda_{V_\mu^{\theta_1}, \pi_\mu^{\theta_2}}(\widetilde{S}_i^t, \widetilde{A}_i^t, \widetilde{S}_i^{t+1})$. Next, we proceed to bound the distance in composite space $\mathcal{F}_{\theta,\xi}$. In particular, let $(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})$, $(V_\mu^{\theta_1\prime}, \pi_\mu^{\theta_2\prime}, \eta^{\xi_1\prime}, \varpi^{\xi_2\prime}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2$ are two arbitrary functions, then the empirical norm distance w.r.t. $\{\mathcal{D}_i\}_{i=1}^n$ for the two function can be upper bounded by

$$
\begin{aligned}
&\mathbb{P}_n\left|G(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i) - G(V_\mu^{\theta_1\prime}, \pi_\mu^{\theta_2\prime}, \eta^{\xi_1\prime}, \varpi^{\xi_2\prime}; \mathcal{D}_i)\right| \\
=&\mathbb{P}_n\Big|\mathbb{E}_T\big[M(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i)K(\{S^t, A^t\}, \{\widetilde{S}^t, \widetilde{A}^t\})\widetilde{M}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i)\big] - \\
&\qquad \mathbb{E}_T\big[M(V_\mu^{\theta_1\prime}, \pi_\mu^{\theta_2\prime}, \eta^{\xi_1\prime}, \varpi^{\xi_2\prime}; \mathcal{D}_i)K(S^t, A^t; \widetilde{S}^t, \widetilde{A}^t)\widetilde{M}(V_\mu^{\theta_1\prime}, \pi_\mu^{\theta_2\prime}, \eta^{\xi_1\prime}, \varpi^{\xi_2\prime}; \mathcal{D}_i)\big]\Big| \\
=&\mathbb{P}_n\Big|\mathbb{E}_T\Big[K(S^t, A^t; \widetilde{S}^t, \widetilde{A}^t)\big(M(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i)\widetilde{M}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i) - \\
&\qquad M(V_\mu^{\theta_1\prime}, \pi_\mu^{\theta_2\prime}, \eta^{\xi_1\prime}, \varpi^{\xi_2\prime}; \mathcal{D}_i)\widetilde{M}(V_\mu^{\theta_1\prime}, \pi_\mu^{\theta_2\prime}, \eta^{\xi_1\prime}, \varpi^{\xi_2\prime}; \mathcal{D}_i)\big)\Big]\Big| \\
=&\mathbb{P}_n\Big\{\mathbb{E}_T\Big|K(S^t, A^t; \widetilde{S}^t, \widetilde{A}^t)\big(M(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i)\widetilde{M}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i) - \\
&\qquad M(V_\mu^{\theta_1\prime}, \pi_\mu^{\theta_2\prime}, \eta^{\xi_1\prime}, \varpi^{\xi_2\prime}; \mathcal{D}_i)\widetilde{M}(V_\mu^{\theta_1\prime}, \pi_\mu^{\theta_2\prime}, \eta^{\xi_1\prime}, \varpi^{\xi_2\prime}; \mathcal{D}_i)\big)\Big|\Big\} \\
\leq&\mathbb{P}_n\Big\{\mathbb{E}_T\Big|K(S^t, A^t; \widetilde{S}^t, \widetilde{A}^t)\Big| \cdot \mathbb{E}_T\Big|\big(M(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i) + M(V_\mu^{\theta_1\prime}, \pi_\mu^{\theta_2\prime}, \eta^{\xi_1\prime}, \varpi^{\xi_2\prime}; \mathcal{D}_i)\Big| \cdot \\
&\qquad \mathbb{E}_T\Big|\big(\widetilde{M}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}; \mathcal{D}_i) - \widetilde{M}(V_\mu^{\theta_1\prime}, \pi_\mu^{\theta_2\prime}, \eta^{\xi_1\prime}, \varpi^{\xi_2\prime}; \mathcal{D}_i)\Big|\Big\} \\
\leq&M_{\max,1}\Big(\mathbb{P}_n\mathbb{E}_T|\eta^{\xi_1} - \eta^{\xi_1\prime}| + \mu\mathbb{P}_n\mathbb{E}_T|\text{prox}^\circ(\pi_\mu^{\theta_2}) - \text{prox}^\circ(\pi_\mu^{\theta_2\prime})| \\
&+ \mathbb{P}_n\mathbb{E}_T|(\gamma V_\mu^{\theta_1} - V_\mu^{\theta_1}) - (\gamma V_\mu^{\theta_1\prime} - V_\mu^{\theta_1\prime})| + \mathbb{P}_n\mathbb{E}_T|\psi^\theta - \varpi^{\xi_2\prime}|\Big) \\
=&M_{\max,1}\big(\mathbb{P}_n\mathbb{E}_T|\eta^{\xi_1} - \eta^{\xi_1\prime}| + \mu\mathbb{P}_n\mathbb{E}_T|\pi_\mu^{\theta_2} - \pi_\mu^{\theta_2\prime}| + (1+\gamma)\mathbb{P}_n\mathbb{E}_T|V_\mu^{\theta_1} - V_\mu^{\theta_1\prime}| + \mathbb{P}_n\mathbb{E}_T|\psi^\theta - \varpi^{\xi_2\prime}|\big) \\
\leq&M_{\max,1}\big(\mathbb{P}_n\|\eta^{\xi_1} - \eta^{\xi_1\prime}\|_\infty + \mu\mathbb{P}_n\|\pi_\mu^{\theta_2} - \pi_\mu^{\theta_2\prime}\|_\infty + (1+\gamma)\mathbb{P}_n\|V_\mu^{\theta_1} - V_\mu^{\theta_1\prime}\|_\infty + \mathbb{P}_n\|\varpi^{\xi_2} - \varpi^{\xi_2\prime}\|_\infty\big),
\end{aligned}
$$

where $M_{\max,1} = 2M_{\max}$. Therefore, as the proximal parameter $0 \leq \mu \leq \mu_{\max} < \infty$, for any $\varepsilon > 0$ the metric entropy $\log \mathcal{N}\left((\mu_{\max}+4)M_{\max,1}\varepsilon, \mathcal{F}_{\theta,\xi}, \{\mathcal{D}_i\}_{i=1}^n\right)$ can be bound with respect to separate metric entropy of $(\Theta_1, \Theta_2, \Xi_1, \Xi_2)$. Denote $\min(2(\mu_{\max}+4)M_{\max}, 1)$ as $\widetilde{C}$, then

$$
\begin{aligned}
&\mathcal{N}\Big((\mu_{\max}+4)M_{\max,1}\varepsilon, \mathcal{F}_{\theta,\xi}, \{\mathcal{D}_i\}_{i=1}^n\Big) \\
&\leq \mathcal{N}\left(\widetilde{C}\varepsilon, \Theta_1, \{\mathcal{D}_i\}_{i=1}^n\right)\mathcal{N}\left(\widetilde{C}\varepsilon, \Theta_2, \{\mathcal{D}_i\}_{i=1}^n\right)\mathcal{N}\left(\widetilde{C}\varepsilon, \Xi_1, \{\mathcal{D}_i\}_{i=1}^n\right)\mathcal{N}\left(\widetilde{C}\epsilon, \Xi_2, \{\mathcal{D}_i\}_{i=1}^n\right)
\end{aligned}
$$

To bound these factors, we first introduce a idea of pseudo-dimension , that is, for any set $\mathcal{X}$, any points $x^{1:N} \in \mathcal{X}^N$, any class $\mathcal{F}$ of functions on $\mathcal{X}$ taking values in $[0, C]$ with pseudo-dimension $D_\mathcal{F} < \infty$ and any $\epsilon > 0$, we have

$$
\mathcal{N}\left(\epsilon, \mathcal{F}, x^{1:N}\right) \leqslant e\left(D_\mathcal{F}+1\right)\left(\frac{2eC}{\epsilon}\right)^{D_\mathcal{F}}
$$

Therefore, we have

$$
\begin{aligned}
&\mathcal{N}\Big(2(\mu_{\max}+4)M_{\max}\epsilon, \mathcal{F}_{\theta,\xi}, \{\mathcal{D}_i\}_{i=1}^n\Big) \\
&\leq e^4\left(D_{\Theta_1}+1\right)\left(D_{\Theta_2}+1\right)\left(D_{\Xi 1}+1\right)\left(D_{\Xi_2}+1\right)\left(\frac{2eM_{\max}}{\widetilde{C}\epsilon}\right)^{D_{\Theta_1}+D_{\Theta_2}+D_{\Xi_1}+D_{\Xi_2}}
\end{aligned}
$$

which implies

$$\begin{aligned}
&\mathcal{N}\Big(\frac{\epsilon}{2}, \mathcal{F}_{\theta,\xi}, \{\mathcal{D}_i\}_{i=1}^n\Big) \\
&\leq e^4 \left(D_{\Theta_1}+1\right)\left(D_{\Theta_2}+1\right)\left(D_{\Xi_1}+1\right)\left(D_{\Xi_2}+1\right)\left(\frac{8(\mu_{\max}+4)M_{\max}^3 e}{\widetilde{C}\epsilon}\right)^{D_{\Theta_1}+D_{\Theta_2}+D_{\Xi_1}+D_{\Xi_2}} \\
&:= C_1\left(\frac{1}{\epsilon}\right)^{D_{\mathcal{F}_{\theta,\xi}}}
\end{aligned}$$

where $C_1 = e^4 \left(D_{\Theta_1}+1\right)\left(D_{\Theta_2}+1\right)\left(D_{\Xi_1}+1\right)\left(D_{\Xi_2}+1\right)\left(\frac{8(\mu_{\max}+4)M_{\max}^3 e}{\widetilde{C}}\right)^{D_{\Theta_1}+D_{\Theta_2}+D_{\Xi_1}+D_{\Xi_2}}$ and $D_{\mathcal{F}_{\theta,\xi}} = D_{\Theta_1}+D_{\Theta_2}+D_{\Xi_1}+D_{\Xi_2}$ i.e., the "effective" psuedo dimension.

Then we apply Pollard tail inequality, for any $n \geq 32/\epsilon^2$, we have

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2})\in\Theta_1\times\Theta_2\times\Xi_1\times\Xi_2} \left|\mathbb{E}G(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2};\mathcal{D}_i) - \mathbb{P}_n G(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2};\mathcal{D}_i)\right| \geq \frac{\epsilon}{2}\Big) \\
\leq & 8C_1\left(\frac{1}{\epsilon}\right)^{D_{\mathcal{F}_{\theta,\xi}}} \exp\left(-\frac{n\epsilon^2}{512M_{\max}^2}\right)
\end{aligned}$$

Then we can obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2})\in\Theta_1\times\Theta_2\times\Xi_1\times\Xi_2} \left|\mathbb{E}G(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2};\mathcal{D}_i) - \mathbb{P}_n G(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2};\mathcal{D}_i)\right|^2\right] \\
=&\int_0^\infty \mathbb{P}\Big(\sup_{(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2})\in\Theta_1\times\Theta_2\times\Xi_1\times\Xi_2} \left|\mathbb{E}G(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2};\mathcal{D}_i) - \mathbb{P}_n G(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2};\mathcal{D}_i)\right|^2 \geq t\Big)dt \\
=&\int_0 \mathbb{P}\Big(\sup_{(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2})\in\Theta_1\times\Theta_2\times\Xi_1\times\Xi_2} \left|\mathbb{E}G(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2};\mathcal{D}_i) - \mathbb{P}_n G(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2};\mathcal{D}_i)\right|^2 \geq t\Big)dt \\
+&\int_u^\infty \mathbb{P}\Big(\sup_{(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2})\in\Theta_1\times\Theta_2\times\Xi_1\times\Xi_2} \left|\mathbb{E}G(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2};\mathcal{D}_i) - \mathbb{P}_n G(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2};\mathcal{D}_i)\right|^2 \geq t\Big)dt \\
\leq& u + \int_u^\infty 8C_1\left(\frac{1}{t}\right)^{D_{\mathcal{F}_{\theta,\xi}}} \exp\left(-\frac{nt^2}{512M_{\max}^2}\right)dt \\
=& u + \frac{64C_1\left(\frac{1}{u}\right)^{D_{\mathcal{F}_{\theta,\xi}}}}{n}\exp\left(-\frac{nu}{512M_{\max}^2}\right)
\end{aligned}$$

With probability $1-\delta$, minimizing the RHS with respect to u, and plug the minimizer in, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2})} \left|\mathbb{E}G(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2};\mathcal{D}_i) - \mathbb{P}_n G(V_\mu^{\theta_1},\pi_\mu^{\theta_2},\eta^{\xi_1},\varpi^{\xi_2};\mathcal{D}_i)\right|^2\right] \\
\leq& \frac{64D_{\mathcal{F}_{\theta,\xi}}\log(8C_1\left(\frac{1}{\delta}\right))}{n},
\end{aligned}$$

where $C_2 = 8C_1$. Therefore, we conclude that, with probability $1-\delta$, we have

$$\Delta_1 \leq \sqrt{\frac{64D_{\mathcal{F}_{\theta,\xi}}\log\left(\frac{8C_1}{\delta}\right)}{n}} := \sqrt{\frac{C_3 D_{\mathcal{F}_{\theta,\xi}}\log\left(\frac{8C_1}{\delta}\right)}{n}}$$

Next, we proceed to bound Δ_2. To simply the notation, we denote the U-statistic kernel as

$$\bar{K}(S^t, A^t; \widetilde{S}^t, \widetilde{A}^t) := \Lambda_{V_\mu^{\theta_1},\pi_\mu^{\theta_2}}(S_i^t, A_i^t, S_i^{t+1}) K\big(S_i^t, A_i^t; \widetilde{S}_i^t, \widetilde{A}_i^t\big)\Lambda_{V_\mu^{\theta_1},\pi_\mu^{\theta_2}}(\widetilde{S}_i^t, \widetilde{A}_i^t, S_i^{t+1}).$$

By Hoeffding's decomposition of kernel function $\bar{K}(S^t, A^t; \widetilde{S}^t, \widetilde{A}^t)$, there exists kernel functions $\bar{K}_1(S^t, A^t)$ and $\bar{K}_2(S^t, A^t; \widetilde{S}^t, \widetilde{A}^t)$ that $\mathbb{E}_T\bar{K}_1(\widetilde{S}^t, \widetilde{A}^t) = 0$ and $\mathbb{E}_T\bar{K}_2(s, a; \widetilde{S}^t, \widetilde{A}^t) = 0$. The U-statistic U_T can be decomposed into

$$U_T = \mathbb{E}_T[U^\star] + \frac{2}{T}\sum_{t=1}^T \bar{K}_1\left(S^t, A^t\right) + U_{\bar{K}_2} \quad \text{and} \quad \mathbb{E}_T[U_T] = \mathbb{E}_T[U^\star] + \mathbb{E}_T[U_{\bar{K}_2}],$$

where $U_{\bar{K}_2} := U_{\bar{K}_2}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})$ is defined similarly as in the proof of Theorem 4.2. The details of the decomposition can be seen in the proof of Theorem 4.2. The term Δ_2 can be immediately decomposed as follows

$$\begin{aligned}
\Delta_2 = & \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left| U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] - \mathbb{E}_T[U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] \right| + \\
& \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left| \mathbb{E}_T[U^\star(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] - \mathbb{E}_T[U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] \right| \\
= & \underbrace{\sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left| U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) - \mathbb{E}_T[U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] \right|}_{\Delta_2^1} + \\
& \underbrace{\sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left| \mathbb{E}_T[U_{\bar{K}_2}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] \right|}_{\Delta_2^2}.
\end{aligned}$$

Note that the second term is not exactly zero since the samples are weakly dependent. But next, we will show that Δ_2^2 converges to zero. First, we check the conditions of Lemma 3.1 in Arcones & Yu (1994). Observe that $K(\cdot, \cdot) \leq 1$ and according to Lemma S.1, then

$$\sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} \left| \bar{K}\left(S^t, A^t; \widetilde{S}^t, \widetilde{A}^t\right) \right| \leq M_{\max}^2 K\left(S^t, A^t; \widetilde{S}^t, \widetilde{A}^t\right) \leq M_{\max}^2.$$

Therefore, the kernel $\bar{K}$ is a uniformly bounded function. Under Assumption 4.2 that $\beta(m) \lesssim m^{-\delta_1}$ for $\delta_1 > 1$. Therefore, $\beta(m) m^{\delta_1} \to 0$. By using a similar technique of calculating the metric entropy, for any $\epsilon > 0$, we have the covering number that

$$\mathcal{N}\left(\varepsilon, \mathcal{F}_{\theta,\xi}, \|\cdot\|_{L^2}\right) \leq \mathcal{N}\left(\varepsilon, \mathcal{F}_{\theta,\xi}, \{\mathcal{D}_i\}_{i=1}^n\right) < \infty$$

Then the conditions of Lemma 3.1 in Arcones & Yu (1994) are satisfied, we have

$$\begin{aligned}
& \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} |\sqrt{T} U_{\bar{K}_2}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})| = o_p(1) \\
\implies & \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}) \in \Theta_1 \times \Theta_2 \times \Xi_1 \times \Xi_2} |U_{\bar{K}_2}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})| = o_p(1). \qquad (36)
\end{aligned}$$

Since $U_{\bar{K}_2}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})$ is uniformly bounded, then

$$\sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} |U_{\bar{K}_2}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})| < \infty$$

As $x, T \to \infty$, then

$$\mathbb{E}\left[\sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} |U_{\bar{K}_2}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})| I\{ \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} |U_{\bar{K}_2}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})| > x \} \right] \to 0,$$

which means $\sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} |U_{\bar{K}_2}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})|$ is uniformly integrable. Combine with the weak convergence in (36), then as $T \to \infty$

$$\Delta_2^2 = \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} \mathbb{E}|U_{\bar{K}_2}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})| \leq \mathbb{E} \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} |U_{\bar{K}_2}(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})| \to 0.$$

Then we move to bound Δ_2^1. The U-statistic U_T is not degenerate, so we adopt Hoeffding's representation (Hoeffding, 1994) such that it reduces the problem to a "first-order" analysis. Specifically, let $\sigma(T)$ is the collection of all permutations of $\{1, 2, ..., T\}$, the U-statistic U_T can be re-expressed as

$$U_T\left(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2}\right) = \frac{1}{T!} \sum_{\sigma(T)} \frac{1}{T_0} \sum_{t=1}^{T_0} \bar{K}\left(X^{\sigma(t)}, X^{\sigma(T_0+t)}\right),$$

where $T_0 = \lfloor T/2 \rfloor$. By the trick, we have the following inequality

$$
\begin{aligned}
\Delta_2^1 &= \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} \left| \frac{2}{T(T-1)} \sum_{1 \le i \neq j \le T} \bar{K}\left(X^i, X^j\right) - \mathbb{E}_T[U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] \right| \\
&= \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} \left| \frac{1}{T!} \sum_{\sigma(T)} \frac{1}{T_0} \sum_{t=1}^{T_0} \bar{K}\left(X^{\sigma(t)}, X^{\sigma(T_0+t)}\right) - \mathbb{E}_T[U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] \right| \\
&= \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} \left| \frac{1}{T!} \sum_{\sigma(T)} \frac{1}{T_0} \sum_{i=t}^{T_0} \bar{K}\left(X^{\sigma(t)}, X^{\sigma(T_0+t)}\right) - \frac{1}{T!} \sum_{\sigma(T)} \mathbb{E}_T[U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] \right| \\
&\le \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} \frac{1}{T!} \sum_{\sigma(T)} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \bar{K}\left(X^{\sigma(t)}, X^{\sigma(T_0+t)}\right) - \mathbb{E}_T[U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] \right| \\
&\le \frac{1}{T!} \sum_{\sigma(T)} \mathbb{E}_T \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \bar{K}\left(X^{\sigma(t)}, X^{\sigma(T_0+t)}\right) - \mathbb{E}_T[U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] \right| \\
&= \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \bar{K}\left(X^{t}, X^{(T_0+t)}\right) - \mathbb{E}_T[U_T(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})] \right| \\
&= \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \bar{K}\left(X^{t}, X^{(T_0+t)}\right) - \mathbb{E}_T \left[\frac{1}{T_0} \sum_{t=1}^{T_0} \bar{K}\left(X^{t}, X^{(T_0+t)}\right) \right] \right| \\
&= \sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \bar{G}\left(\widetilde{X}^{t}\right) - \mathbb{E}_T \left[\frac{1}{T_0} \sum_{t=1}^{T_0} \bar{G}\left(\widetilde{X}^{t}\right) \right] \right|,
\end{aligned}
$$

where $\widetilde{X}^t = \left(X^t, X^{(T_0+t)}\right)$ which itself is a two-dimensional stationary sequences under mixing condition. Note that the last term is the expectation of the suprema of the empirical process $1/T_0 \sum_{t=1}^{T_0} \bar{G}(\widetilde{X}^t) - \mathbb{E}_T[1/T_0 \sum_{t=1}^{T_0} \bar{G}(\widetilde{X}^t)]$ on the space $\bar{\mathcal{G}}_{\theta,\xi}$. The distance in $\bar{\mathcal{G}}_{\theta,\xi}$ can be bounded by the following,

$$
\begin{aligned}
&\mathcal{N}\Big(\min\{(2\mu_{\max} + 4)M_{\max}, 1\}\varepsilon, \bar{\mathcal{G}}_{\theta,\xi}, \{\widetilde{X}^t\}_{t=1}^{T_0} \Big) \\
&\le \mathcal{N}\left(\widetilde{C}\varepsilon, \Theta_1, \{\mathcal{D}_i\}_{i=1}^n\right) \mathcal{N}\left(\widetilde{C}\varepsilon, \Theta_2, \{\mathcal{D}_i\}_{i=1}^n\right) \mathcal{N}\left(\widetilde{C}\varepsilon, \Xi_1, \{\mathcal{D}_i\}_{i=1}^n\right) \mathcal{N}\left(\widetilde{C}\epsilon, \Xi_2, \{\mathcal{D}_i\}_{i=1}^n\right) \\
&= e^4 \left(D_{\Theta_1} + 1\right)\left(D_{\Theta_2} + 1\right)\left(D_{\Xi_1} + 1\right)\left(D_{\Xi_2} + 1\right) \left(\frac{2eM_{\max}}{\widetilde{C}\epsilon}\right)^{D_{\Theta_1} + D_{\Theta_2} + D_{\Xi_1} + D_{\Xi_2}}
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\mathcal{N}\Big(\frac{\epsilon}{16}, \bar{\mathcal{G}}_{\theta,\xi}, \{\widetilde{X}^t\}_{t=1}^{T_0} \Big) \\
&\le e^4 \left(D_{\Theta_1} + 1\right)\left(D_{\Theta_2} + 1\right)\left(D_{\Xi_1} + 1\right)\left(D_{\Xi_2} + 1\right) \left(\frac{64(_{\max} + 32)U_{\max}^2 e}{\widetilde{C}\epsilon}\right)^{D_{\Theta_1} + D_{\Theta_2} + D_{\Xi_1} + D_{\Xi_2}} \\
&:= C_3 \left(\frac{1}{\epsilon}\right)^{D_{\bar{\mathcal{G}}_{\theta,\xi}}}
\end{aligned}
$$

where $D_{\bar{\mathcal{G}}_{\theta,\xi}} = D_{\Theta_1} + D_{\Theta_2} + D_{\Xi_1} + D_{\Xi_2}$. First, without loss of generality, let $T_0 = 2m_{T_0}k_{T_0}$ for appropriate positive integers $m_{T_0}k_{T_0}$ as in (Yu, 1994). Follow Lemma 5 in Antos et al. (2008), we obtain that

$$
\begin{aligned}
\mathbb{P}(\sup_{(V_\mu^{\theta_1}, \pi_\mu^{\theta_2}, \eta^{\xi_1}, \varpi^{\xi_2})} &\left| \frac{1}{T_0} \sum_{t=1}^{T_0} \bar{G}\left(\widetilde{X}^{t}\right) - \mathbb{E}_T \left[\frac{1}{T_0} \sum_{t=1}^{T_0} \bar{G}\left(\widetilde{X}^{t}\right) \right] \right| \ge \frac{\epsilon}{2}) \\
&\le C_3 \left(\frac{1}{\epsilon}\right)^{D_{\bar{\mathcal{G}}_{\theta,\xi}}} \exp\left(-4C_4 m_{T_0}\epsilon^2\right) + 2m_{T_0}\beta(k_{T_0})
\end{aligned}
$$

where $C_4 = \frac{1}{2}\left(\frac{1}{8M_{\max}^2}\right)^2$. If $D_{\bar{\mathcal{G}}} \geq 2$, and let $\beta(m) \lesssim \exp\left(-\delta_1 m\right), T \geq 1, m_T = \left\lceil \left(C_4 T_0 \epsilon^2/\delta_1\right)^{\frac{1}{2}} \right\rceil, m_{T_0} = T_0/\left(2k_{T_0}\right)$, where $D_{\bar{\mathcal{G}}_{\theta,\xi}} \geq 2, C_3, C_4, \delta_1$, we apply Lemma 14 in Antos et al. (2008), then

$$2m_{T_0}\beta_{k_{T_0}} + C_1\left(\frac{1}{\epsilon}\right)^{D_{\bar{\mathcal{G}}_{\theta,\xi}}} \exp\left(-4C_2 m_{T_0}\epsilon^2\right) \leq \delta$$

and we have, with probability $1-\delta$,

$$\Delta_2^1 \leq \sqrt{\frac{2\Delta(\Delta/\delta_1 \vee 1)}{C_4 T_0}}$$

$$\implies \Delta_2^1 \leq \sqrt{\frac{2\Delta(\Delta/\delta_1 \vee 1)}{C_4 \lfloor T/2 \rfloor}}$$

where

$$\Delta = (D_{\bar{\mathcal{G}}_{\theta,\xi}}/2)\log T_0 + \log(e/\delta) + \log^+\left(C_3 C_4^{D_{\bar{\mathcal{G}}_{\theta,\xi}}/2}\right)$$

$$\implies \Delta = (D_{\bar{\mathcal{G}}_{\theta,\xi}}/2)\log(T/2) + \log(e/\delta) + \log^+\left(C_3 C_4^{D_{\bar{\mathcal{G}}_{\theta,\xi}}/2}\right)$$

Now, we conclude that

$$\|\widehat{V}_\mu^{\theta_1,k} - V^*\|_{L^2}^2 \leq \frac{C_1}{\kappa_{\min}(1-\gamma)^2}\left(\sqrt{\frac{C_3 D \log\left(\frac{8C_1}{\delta}\right)}{n}} + \sqrt{\frac{2\Delta(\Delta/\delta_1 \vee 1)}{C_4\lfloor T/2\rfloor}}\right) + C_2\frac{\mu^2(\mathbf{C} + |1-\mathbf{C}| \vee 1)^2}{(1-\gamma)^2} + C_5\left\|\widehat{V}_\mu^{\theta_1} - \widehat{V}_\mu^{\theta_1,k}\right\|_{L^2}^2 + \epsilon_{\text{approximation error}}$$

where $\Delta = (D_{\bar{\mathcal{G}}_{\theta,\xi}}/2)\log(\lfloor T/2\rfloor) + \log(e/\delta) + \log^+\left(C_3 C_4^{D_{\bar{\mathcal{G}}_{\theta,\xi}}}/2\right)$, $D_{\bar{\mathcal{G}}_{\theta,\xi}} = \text{P-dim}(\Theta_1) + \text{P-dim}(\Theta_2) + \text{P-dim}(\Xi_1) + \text{P-dim}(\Xi_2)$, and $C_1, ..., C_5$ are some constants. Adapt the notations for the constants number from Theorem 4.2. By some algebra, we conclude that

$$\|\widehat{V}_\mu^{\theta_1,k} - V^{\pi^*}\|_{L^2}^2 \leq \underbrace{\frac{C_4}{\kappa_{\min}(1-\gamma)^2}\left(\sqrt{\frac{C_5 D_{\text{P-dim}} \log\left(\frac{8C_4}{\delta}\right)}{n}} + \sqrt{\frac{2\left(\frac{\bar{\Delta}}{\delta_1} \vee 1\right)\bar{\Delta}}{C_6\lfloor T/2\rfloor}}\right)}_{\text{generalization error}} + \underbrace{C_7\frac{\mu^2(\mathbf{C} + |1-\mathbf{C}| \vee 1)^2}{(1-\gamma)^2}}_{\text{proximal bias}} + \underbrace{C_8\left\|\widehat{V}_\mu^{\theta_1} - \widehat{V}_\mu^{\theta_1,k}\right\|_{L^2}^2}_{\text{optimization error}} + \epsilon_{\text{approximation error}}$$

where $\bar{\Delta} = \frac{D_{\text{P-dim}}\log(\lfloor T/2\rfloor)}{2} + \log(\frac{e}{\delta}) + \log^+\left(\frac{C_5 C_6^{D_{\text{P-dim}}}}{2}\right)$, $D_{\text{P-dim}} = \text{P-dim}(\Theta_1) + \text{P-dim}(\Theta_2) + \text{P-dim}(\Xi_1) + \text{P-dim}(\Xi_2)$, and $C_4, ..., C_8$ are some constants. □

A.9 Proof of Theorem 4.4

We note that SGD converges has a global convergence to a stationary point with a sublinear rate in the case of convexity. However, the resulting dose not typically holds for the non-convex analysis. The intuition behind the proof is that our quasi-optimal algorithm can be regarded as a special case of the randomized stochastic descent (RSD) algorithm for solving the non-convex minimization problem.

The convergence analysis of for randomized stochastic descent algorithm has been established in Corollary 2.2 of (Ghadimi & Lan, 2013). That is, RSD is provably convergent to a stationary point. Follow Theorem 3 in (Drori & Shamir, 2020), an unbiased SGD algorithm, i.e., the quasi-optimal

algorithm with diminishing learning rate and evaluated on Euclidean distance. Therefore, it suffices to show that the gradient of the loss is unbiased.

Now we show that the gradient is unbiased, as follows

$$\begin{aligned}
\nabla_{\theta_1}\mathcal{L}_U &= \mathbb{E}\Big[\nabla_{\theta_1}(\gamma V_\mu^{\theta_1}(S^{t+1}) - V_\mu^{\theta_1}(S^t))K(S^t,A^t;\tilde{S}^t,\tilde{A}^t)\Lambda_{V_\mu,\pi_\mu}(\tilde{S}^t,\tilde{A}^t,\tilde{S}^{t+1}) \\
&\quad + \Lambda_{V_\mu,\pi_\mu}(S^t,A^t,S^{t+1})K(S^t,A^t;\tilde{S}^t,\tilde{A}^t)\nabla_{\theta_1}(\gamma V_\mu^{\theta_1}(\tilde{S}^{t+1}) - V_\mu^{\theta_1}(\tilde{S}^t))\Big] \\
\nabla_{\theta_2}\mathcal{L}_U &= \mathbb{E}\Big[-2\mu(\nabla_{\theta_2}\pi_\mu^{\theta_2}(A^t|S^t)K(S^t,A^t;\tilde{S}^t,\tilde{A}^t)\Lambda_{V_\mu,\pi_\mu}(\tilde{S}^t,\tilde{A}^t,\tilde{S}^{t+1}) \\
&\quad - \Lambda_{V_\mu,\pi_\mu}(S^t,A^t,S^{t+1})K(S^t,A^t;\tilde{S}^t,\tilde{A}^t)(2\mu(\nabla_{\theta_2}\pi_\mu^{\theta_2}(\tilde{A}^t|S^t))\Big] \\
\nabla_{\xi_1}\mathcal{L}_U &= \mathbb{E}\Big[\nabla_{\xi_1}\varpi(S^t,A^t)K(S^t,A^t;\tilde{S}^t,\tilde{A}^t)\Lambda_{V_\mu,\pi_\mu}(\tilde{S}^t,\tilde{A}^t,\tilde{S}^{t+1}) \\
&\quad + \Lambda_{V_\mu,\pi_\mu}(S^t,A^t,S^{t+1})K(S^t,A^t;\tilde{S}^t,\tilde{A}^t)\nabla_{\xi_1}\varpi(\tilde{S}^t,\tilde{A}^t)\Big] \\
\nabla_{\xi_2}\mathcal{L}_U &= \mathbb{E}\Big[-\nabla_{\xi_2}\eta(S^t)K(S^t,A^t;\tilde{S}^t,\tilde{A}^t)\Lambda_{V_\mu,\pi_\mu}(\tilde{S}^t,\tilde{A}^t,\tilde{S}^{t+1}) \\
&\quad - \Lambda_{V_\mu,\pi_\mu}(S^t,A^t,S^{t+1})K(S^t,A^t;\tilde{S}^t,\tilde{A}^t)\nabla_{\xi_2}\eta(\tilde{S}^t)\Big]
\end{aligned}$$

We conclude that the gradient estimator is unbiased. Follow Theorem 3 in (Drori & Shamir, 2020), under the conditions stated in Theorem 4.4, we adapt Corollary 2.2 to our quasi-optimal algorithm, it completes the proof.

B Practical Implementation

In practice, $\{V_\mu^*, \pi_\mu^*, \eta, \varpi\}$ needs to be parameterized for practical implementation. However, noticing that V_μ^* and π_μ^* are both associated with Q_μ^* with closed-form expressions (4)(5). Thus, we propose to represent (V_μ^*, π_μ^*) by modeling Q_μ^*. Additionally, by modeling Q_μ^* as a quadratic function, the induced policy would follow a q-Gaussian distribution. Therefore, we model the coefficients associated with the quadratic form as a linear combination of basis function $\varphi(s)$ such that $Q_\mu^*(s,a;\theta) = -\exp\{\theta_1^T\varphi(s)\}a^2 + \theta_2^T\varphi(s)a + \theta_3^T\varphi(s)$, where $\varphi(s) = [\varphi_1(s), \varphi_2(s), ..., \varphi_m(s)]^T$ is the m-dimensional basis function, and $\theta = [\theta_1, \theta_2, \theta_3]^T$ is the $3m$-dimensional parameters we need to estimate. The advantage of such parametrization lies in that the parameter space could be reduced.

To solve the constrained optimization problem, we propose a computationally efficient algorithm by transforming the original constrained optimization problem into an unconstrained minimization problem. Specifically, we impose restrictions on the representation of Lagrangian multipliers $(\eta(s), \varpi(s,a))$ so that they satisfy their constraints automatically. Although such re-parametrization may sacrifice model flexibility, it gains great computational advantage as the unconstrained optimization problem would be much simpler. To be specific, we parametrize ϖ as

$$\varpi(s,a;\theta) = \max\Big(0, -\frac{Q_\mu^*(s,a;\theta)}{2\mu} + \frac{\int_{a\in\mathcal{W}_1(s)} Q_\mu^*(s,a;\theta)da}{2\mu\sigma(\mathcal{W}_s)} - \frac{1}{\sigma(\mathcal{W}_s)}\Big), \tag{37}$$

Therefore, $\varpi(S^t, A^t) \geq 0$ and $\pi_\mu^*(A^t|S^t)\cdot\varpi(S^t,A^t) = 0$ are automatically satisfied. Also, by specifying the expression of Lagrangian multipliers, $\varpi(s,a)$ share the same set of parameters θ as (V_μ^*, π_μ^*). We also define

$$\eta(s;\xi) = \Big\{\frac{-\mu\mathbf{C}}{1+\exp(-k_0(\xi^T s - b_0))}\Big\}, \tag{38}$$

where b_0 is the sigmoid's midpoint and k_0 is the logistic growth rate. By flipping the sigmoid function to parametrize $\eta(s;\xi)$, the constraint $\eta(s) \in [-\mu\mathbf{C}, 0]$ is also automatically satisfied.

C EXPERIMENT DETAILS AND ADDITIONAL RESULTS

For the reproducing purpose, we include our code for all the experiments and the guideline for access to the Ohio Type I Diabetes dataset in an anonymous GitHub link `https://anonymous.4open.science/r/Quasi-optimal-Learning-with-Continuous-Treatments-9B88`.

C.1 DETAILS OF SIMULATION SETTINGS AND REAL DATA ANALYSIS

The details of the data generative model of each environment in Section 6 are stated below:

Environment I: We consider a bounded action space where $\mathcal{A} = [0, 1]$, and a 2-dimensional state space. $A_i^t \overset{iid}{\sim} \text{Unif}(0, 1)$, the state transition function is defined as $S_{i,1}^{t+1} = \frac{1-\exp(-A_i^t)}{1+\exp(-A_i^t)} S_{i,1}^t + 0.25 S_{i,1}^t S_{i,2}^t + \epsilon_{i,1}^t$, $S_{i,2}^{t+1} = -\frac{1-\exp(-A_i^t)}{1+\exp(-A_i^t)} S_{i,2}^t + 0.25 S_{i,1}^t S_{i,2}^t + \epsilon_{i,2}^t$, where $\epsilon_{i,1}^t, \epsilon_{i,2}^t \overset{iid}{\sim} N(0, 0.5^2)$, and the reward function is defined as

$$R_i^t = 3\left(-\exp(S_{i,1}^{t+1} - S_{i,2}^{t+1})(A_i^t)^2 + (S_{i,1}^{t+1} + S_{i,2}^{t+1} + 0.5)A_i^t + S_{i,1}^{t+1} + S_{i,2}^{t+1}\right).$$

Environment II: We consider a bounded action space where $\mathcal{A} = [0, 1]$, and a 2-dimensional state space. $A_i^t \overset{iid}{\sim} \text{Unif}(0, 1)$, the state transition function is defined as $S_{i,1}^{t+1} = 0.75(2A_i^t - 1) \cdot S_{i,1}^t + 0.25 S_{i,1}^t S_{i,2}^t + \epsilon_{i,1}^t$, $S_{i,2}^{t+1} = 0.75(1 - 2A_i^t) S_{i,2}^t + 0.25 S_{i,1}^t S_{i,2}^t + \epsilon_{i,2}^t$. where $\epsilon_{i,1}^t, \epsilon_{i,2}^t \overset{i.i.d}{\sim} N(0, 0.5^2)$, and $R_i^t = 0.25(S_{i,1}^{t+1})^3 + 2S_{i,1}^{t+1} + 0.5(S_{i,2}^{t+1})^3 + S_{i,2}^{t+1} + 0.25(2A_i^t - 1)$.

Environment III: We consider an unbounded action space where $\mathcal{A} = (-\infty, \infty)$, and a 8-dimensional state space. We sampled action uniformly from a bounded space, $A_i^t \overset{iid}{\sim} \text{Unif}(-100, 100)$, while it is allowed to select actions on $\mathbb{R}$ for the learned policy. The state transition function is defined as, $S_i^{t+1} \sim N(\mu_i^{t+1}, \Sigma)$, where Σ is a pre-specified covariance matrix, and $\mu_i^t = [\mu_{i,1}^t, ..., \mu_{i,8}^t]$,

$$\mu_{i,j}^{t+1} = \frac{\exp(A_i^t/100 + \mu_{i,j}^t) - \exp(-(A_i^t/100 + \mu_{i,j}^t))}{\exp(A_i^t/100 + \mu_{i,j}^t) + \exp(-(A_i^t/100 + \mu_{i,j}^t))} \quad \text{for} \quad j = 1, 2, 3, 4,$$

$$\mu_{i,j}^{t+1} = \frac{\exp(-A_i^t/100 + \mu_{i,j}^t) - \exp(-(-A_i^t/100 + \mu_{i,j}^t))}{\exp(-A_i^t/100 + \mu_{i,j}^t) + \exp(-(-A_i^t/100 + \mu_{i,j}^t))} \quad \text{for} \quad j = 5, 6, 7, 8.$$

$R_i^t = -\exp(S_{i,1}^{t+1}/2 + S_{i,5}^{t+1}/2)(A_i^t/100)^2 + 2(S_{i,2}^{t+1} + S_{i,3}^{t+1} + S_{i,6}^{t+1} + S_{i,7}^{t+1} + 0.5)A_i^t/100 + S_{i,4}^{t+1} + S_{i,8}^{t+1}$.

Environment IV: This environment shares the same transition kernel as Environment III, the only difference is the reward function here is

$$R_i^t = (S_{i,1}^{t+1}/2)^3 + (S_{i,2}^{t+1}/2)^3 + S_{i,3}^{t+1} + S_{i,4}^{t+1} + 2[(S_{i,5}^{t+1}/2)^3 + (S_{i,6}^{t+1}/2)^3] + 0.5(S_{i,7}^{t+1} + S_{i,8}^{t+1}).$$

For all four environments, we consider different sample sizes where the number of trajectories $n = \{25, 50\}$, and the length of each trajectory $T = \{24, 36\}$. The discount factor γ is set to 0.9.

Ohio Type 1 Diabetes Dataset: For individuals in the first cohort, we treat glucose level , carbonhydrate intake, and acceleration level as state variables, i.e., $S_{i,1}^t$, $S_{i,2}^t$ and $S_{i,3}^t$. For individuals in the second cohort, heart rate is used instead of acceleration level as $S_{i,3}^t$. The reward function is defined as

$$R_i^t = -\frac{\mathbb{1}(S_{i,1}^t > 140)^{1.1} + \mathbb{1}(S_{i,1}^t < 80)(S_{i,1}^t - 80)^2}{30}.$$

C.2 ADDITIONAL EXPERIMENT DETAILS AND RESULTS

In our implementation, since the objective function, $\hat{\mathcal{L}}_U$ may not be convex with respect to (θ, ξ). We determine the initial point by randomly generating 200 initial values for all parameters and selecting the one with the smallest objective function value. For competing methods based on discrete action

space (Greedy-GQ), we discretize the original action space into twenty bins for implementation in simulation, and 14 bins for real data analysis. For other competing methods (DDPG, SAC and BEAR), we use the implementation from the offline deep reinforcement learning library d3rlpy (Seno & Imai, 2021), where we use multi-layer perceptron (MLP) with 2 hidden layers to approximate Q-function and policy distribution. This is a commonly used architecture in deep learning (Farrell et al., 2021).

To evaluate the policy obtained from the proposed method in simulation experiments, we generate 100 independent trajectories, each with a length of 100 based on the learned policy. We use rejection sampling (Robert et al., 1999) to randomly sample each action by the induced density $\pi_\mu(a|s)$ and calculate the discounted sum of reward for each trajectory. We compare the discounted return of each method. The boxplot of simulation results based on 50 runs is presented in Figure 3.

For real data analysis, since the data-generating process is unknown, we follow Luckett et al. (2020) to utilize the Monte Carlo approximation of the estimated V-function of the initial state of each trajectory to evaluate the performance of each method. To better evaluate the stability and performance of each method, we randomly select 10 or 20 trajectories from each individual based on available trajectories 50 times and apply all methods to the selected data. The baseline refers to the observed discounted return. The mean and standard deviation of the improvements on the Monto Carlo discounted returns are presented in Table 1.

We report all hyperparameters used in training and additional experiment results in this section. The value of μ is selected from the set $\{0.01, 0.05, 0.1, 0.2, 0.3, 0.5\}$. For simulation experiments, we conduct sensitivity analysis on the effect of μ, and select the μ that achieves the largest discounted return. For real data analysis, we conduct cross-validation, where we select μ with the largest fitted V-function value on the initial states of each trajectory, i.e., $\mathbb{P}_n \hat{V}_\mu(S_i^1) - (1-\gamma)^{-1}\mu$, where we mitigate the effect of the threshold parameter μ. In our implementation, we set $\mathbf{C} = 5$ for all simulations and real data analysis, and check that the induced policy π_μ never reaches the boundary value.

We set the learning rate α_j for the jth iteration is be $\frac{\alpha_0}{1+d\sqrt{j}}$, where α_0 is the learning rate of the initial iteration, and d is the decay rate of the learning rate. When $n = 25$, we set the batch size to be 5, and when $n = 50$, we set the batch size to be 7. We use the L_2 distance of iterative parameters as the stopping criterion for the SGD algorithm. The μ selected for each experiment, along with the learning rates and their descent rates, are shown in Table 23 and 4.

Table 2: Hyperparameters for each simulated environment

Hyperparameters	Environment I	Environment II	Environment III	Environment IV
μ	0.1	0.05	0.05	0.05
Learning Rate	0.002	0.0005	10^{-5}	10^{-5}
Descent Rate	10^{-4}	10^{-4}	10^{-4}	10^{-4}

Table 3: Hyperparameters for Ohio Type I Diabetes Analysis (Cohort I)

Patient ID	540	544	552	567	584	596
μ	0.1	0.1	0.1	0.05	0.05	0.1
Learning Rate	0.001	0.001	0.001	0.0005	0.001	0.02
Descent Rate	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	2×10^{-4}

Table 4: Hyperparameters for Ohio Type I Diabetes Analysis (Cohort II)

Patient ID	559	563	570	575	588	591
μ	0.2	0.1	0.2	0.05	0.05	0.05
Learning Rate	0.005	0.0001	0.005	0.0001	0.0001	0.0001
Descent Rate	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}

Table 5: The mean running time in seconds of each method over 50 simulation runs in Environment I. The simulations are conducted on a single `2.3 GHz Dual-Core Intel Core i5` CPU
.

n	T	Proposed	SAC	DDPG	BEAR	Greedy-GQ
25	24	23.11	22.17	14.31	35.42	11.39
	36	28.91	28.12	18.35	42.16	14.47
50	24	28.88	29.91	19.42	46.73	15.62
	36	45.23	44.46	36.82	63.54	24.81

Table 6: The mean running time in seconds of each method over 50 simulation runs in Environment II. The simulations are conducted on a single `2.3 GHz Dual-Core Intel Core i5` CPU
.

n	T	Proposed	SAC	DDPG	BEAR	Greedy-GQ
25	24	30.93	27.29	20.12	44.01	14.56
	36	39.34	36.91	26.43	52.86	19.43
50	24	41.12	42.25	28.42	55.17	21.56
	36	60.16	56.47	45.71	72.12	32.14