STEEL: Singularity-aware Reinforcement Learning

Xiaohong Chen* Zhengling Qi† Runzhe Wan‡

May 15, 2023

Abstract

Batch reinforcement learning (RL) aims at leveraging pre-collected data to find an optimal policy that maximizes the expected total rewards in a dynamic environment. A fundamental challenge behind this task is the distributional mismatch between the batch data generating process and the distribution induced by target policies. Nearly all existing algorithms rely on the absolutely continuous assumption on the distribution induced by target policies with respect to the data distribution, so that the batch data can be used to calibrate target policies via the change of measure. However, the absolute continuity assumption could be violated in practice, especially when the state-action space is large or continuous. In this paper, we propose a new batch RL algorithm without requiring absolute continuity in the setting of an infinite-horizon Markov decision process with continuous states and actions. We call our algorithm STEEL: SingulariTy-awarE rEinforcement Learning. Our algorithm is motivated by a new error analysis on off-policy evaluation, where we use maximum mean discrepancy, together with distributionally robust optimization, to characterize the error of off-policy evaluation caused by the possible singularity and to enable model extrapolation. By leveraging the idea of pessimism and under some mild conditions, we derive a finite-sample regret guarantee for our proposed algorithm without imposing absolute continuity. Compared with existing algorithms, by requiring only minimal data-coverage assumption, STEEL significantly improves the applicability and robustness of batch RL. Extensive simulation studies and one real experiment on personalized pricing demonstrate the superior performance of our method in dealing with possible singularity in batch RL.

Author names are sorted alphabetically

*Cowles Foundation for Research in Economics, Yale University. `xiaohong.chen@yale.edu`

†Department of Decision Sciences, George Washington University. `qizhengling@gwu.edu`

‡Amazon. `runzhe.wan@gmail.com`. This work does not relate to the position at Amazon.

1 Introduction

Batch reinforcement learning (RL) aims to learn an optimal policy using pre-collected data without further interacting with the environment. Recently, there is an increasing interest in studying batch RL, which has shown great potentials in many applications such as welfare maximization (Manski, 2004; Kitagawa and Tetenov, 2018; Athey and Wager, 2021; Mbakop and Tabord-Meehan, 2021; Kitagawa et al., 2022), mobile health (Shi et al., 2021; Liao et al., 2022), robotics (Pinto and Gupta, 2016), digital marketing (Thomas et al., 2017), precision medicine (Kosorok and Laber, 2019), among many others.

Arguably, the biggest challenge in batch RL is the distributional mismatch between the batch data distribution and those induced by some candidate policies (Levine et al., 2020). It has been observed that, in practice, the distributional mismatch often results in an unsatisfactory performance of many existing algorithms, due to the insufficient coverage of the batch data (Fujimoto et al., 2019). From a theoretical standpoint, classical methods such as fitted q-iteration (Ernst et al., 2005) and policy iteration (e.g., Sutton et al., 1998) crucially rely on the full-coverage assumption for finding the optimal policy with the fixed dataset. The full-coverage assumption ensures that the distributions induced by all candidate policies can be well calibrated by the batch data generating process, using the change of measure, which could easily fail. To address this limitation, the recent literature (e.g., Kumar et al., 2019; Jin et al., 2021) has found that, by incorporating the idea of pessimism in the algorithm, the full-coverage requirement can be relaxed to only that the data cover the trajectories of an optimal policy, which is easier to satisfy. Here an optimal policy refers to either a globally optimal or an in-class optimal one.

Despite the recent progress, a key technical requirement for nearly all existing batch RL algorithms is the absolute continuity of the policy-induced distribution with respect to the batch data distribution. With this assumption, the change of measure is enabled and the concentration coefficient (Munos, 2003) is used to characterize the deviation between the data distribution and the one induced by some target policy. However, in real applications, this assumption is not known a priori and could be easily violated. For example, when the action space is continuous, and all candidate policies are deterministic but the behavior one used to collect the batch data is stochastic, absolute continuity fails to hold and the singularity issue arises. Meanwhile, the concentration coefficient is no longer well defined. Moreover, for many complex RL tasks such as robotic controls, the state and action spaces could be extremely large and high-dimensional. Due to the lack of interaction with the environment, some important state-action subspace induced by the optimal policy could be inevitably less explored by the behavior policy, which leads to an issue of a non-overlapping support. In this case, the requirement of absolute continuity cannot be satisfied either. Therefore, most existing batch RL algorithms cannot guarantee to obtain an optimal policy in both cases.

1.1 Our Contribution: A Provably Efficient RL Algorithm Allowing for Singularity

Motivated by the aforementioned issue, in this paper, we study batch RL with potential singularity. We consider an infinite-horizon Markov decision process (MDP) with continuous states and actions, despite that our idea is equally applicable to other setups. We propose an efficient policy-iteration-type algorithm, without requiring any form of absolute continuity. By saying an efficient algorithm, we refer to that the sample complexity requirement in finding an optimal policy is polynomial in terms of key parameters. We call our algorithm STEEL: SingulariTy-awarE rEinforcement Learning.

Our algorithm is motivated by a new error analysis on off-policy evaluation (OPE), which has been extensively studied in the recent literature. The goal of OPE is to use the batch data for evaluating the performance of other policies. By leveraging Lebesgue's Decomposition Theorem, we decompose the error of OPE into two parts: the absolutely continuous part and the singular one with respect to the data distribution. For the absolutely continuous part, which can be calibrated by the behavior policy using the change of measure, standard OPE methods can be applied for controlling the error. For the singular part, we use the maximum mean discrepancy and leverage distributionally robust optimization to characterize its worst-case error measured by the behavior policy. Once we understand these two sources of the OPE error induced by any given target policy, a new estimating method for OPE without requiring absolute continuity can be formulated, based on which a policy iteration algorithm with pessimism is proposed.

From theoretical perspective, we show that under some technical conditions and without requiring absolute continuity, the regret of our estimated policy converges to zero at a satisfactory rate in terms of total decision points in the batch data. This novel result demonstrates that our method can be more applicable in solving batch RL problems, in comparison to existing solutions. More specifically, when the distribution induced by the optimal policy is covered by our batch data generating process in the usual manner, we recover the existing theoretical results given by those pessimistic RL algorithms (e.g., Xie et al., 2021; Fu et al., 2022). In contrast, when absolute continuity fails, our algorithm is provable to find an optimal policy with a finite-sample regret warranty. To the best of our knowledge, this is the first finite-sample regret guarantee in batch RL without assuming any form of absolute continuity. Moreover, our work can provide theoretical guidance on the existing algorithms aiming to find a deterministic optimal policy in a continuous action space. For example, our STEEL method and the related theoretical results can be helpful for improving the understanding of the celebrated deterministic policy gradient method (Silver et al., 2014).

To further demonstrate the effectiveness of our algorithm, we consider a contextual bandit problem and conduct extensive simulation studies. We show that compared with two existing baseline methods, under a possible singularity, our algorithm has a significantly better finite-sample performance. In particular, we observe that our algorithm converges faster and is more robust when the singularity arises. Lastly, we apply our method to a personalized pricing application using the data from a US auto loan company, and find that our method outperforms the existing baseline methods.

1.2 Relation to Existing Work

Classical methods on batch RL mainly focus on either value iteration or policy iteration (Watkins and Dayan, 1992; Sutton et al., 1998; Antos et al., 2008*a*). As discussed before, these methods require the full-coverage assumption on the batch data for finding the optimal policy, which is hard to satisfy in practice due to the inability to further interact with the environment. Failure of satisfying this condition often leads to unstable performance such as the lack of convergence or error magnification (Wang et al., 2021).

Recently, significant efforts from the empirical perspective have been made trying to address the challenge from the insufficient data coverage (e.g., Kumar et al., 2020; Fujimoto et al., 2019). The key idea of these works is to restrict the policy class within the reach of the batch dataset, so as to relax the stringent full-coverage assumption. To achieve this, the underlying strategy is to adopt the principle of pessimism for modeling the state-value function, in order to discourage the exploration over state-action pairs less-seen in the batch data. See Liu et al. (2020); Rashidinejad et al. (2021); Jin et al. (2021); Xie et al. (2021); Zanette et al. (2021); Zhan et al. (2022); Fu et al. (2022). Thanks to these pessimistic-type algorithms, the full-coverage assumption can be relaxed to the partial coverage or the so-called single-trajectory concentration assumption, i.e., the distribution induced by the (in-class) optimal policy is absolutely continuous with respect to the one induced by the behavior policy. This enhances the applicability of batch RL algorithms to some extent. However, this partial coverage assumptions still cannot be verified and hardly holds, especially when the state-action space is large or when the imposed policy class to search from is very complex (e.g., neural networks) that leads to unavoidable over-exploration. Our STEEL algorithm can address this limitation without assuming any form of absolute continuity, and can find an (in-class) optimal policy with a finite-sample regret guarantee. Thus, STEEL could be more generally applicable than existing solutions. The price to pay for this appealing property is a slightly stronger modeling assumption, i.e., Bellman completeness, which provides a desirable extrapolation property so that the singular part incurred by the distributional mismatch can be properly controlled.

Our STEEL algorithm is particularly useful for policy learning with a continuous action space, where the singularity arises naturally in the batch setting. This fundamental issue hinders the theoretical understanding of many existing algorithms such as deterministic policy gradient algorithms and their variants (e.g., Silver et al., 2014; Lillicrap et al., 2015). To the best of our knowledge, Antos et al. (2008*a*) is the only work in the RL literature that tackles this problem. However, Antos et al. (2008*a*) imposes a strong regularity condition on the action space, which is hard to interpret and essentially implies that the L^∞-norm of any function over the action space is bounded by its L^1-norm (multiplied by some constant).

Moreover, Antos et al. (2008*a*) cannot address the possible singularity over the state space. In contrast, our STEEL algorithm can handle the singularity in both the state and action spaces with a strong theoretical guarantee. To the best of our knowledge, our paper is the first to address the aforementioned issue.

The rest of the paper is organized as follows. In Section 2, we introduce our setup, the discrete-time homogeneous MDP with continuous states and actions. We introduce related notations and also the problem formulation. Then, an illustrative example on the contextual bandit problem is presented in Section 3 for describing the challenge of policy learning without assuming absolute continuity. We also briefly introduce our solution in this section. In Section 4, we formally introduce the proposed method for finding an optimal policy when facing the possible singularity in batch RL. A comprehensive theoretical study of our algorithm is given in Section 5. Section 6 demonstrates the empirical performance of our methods using both simulation studies and a real data application. Section 7 concludes this paper. All technical proofs can be found in the appendix.

2 Preliminaries and Notations

In this section, we briefly introduce the framework of the discrete-time homogeneous MDP and the related notations. Consider a trajectory $\{S_t, A_t, R_t\}_{t\geq 0}$, where S_t denotes the state of the environment at the decision point t, A_t is the action and R_t is the immediate reward received from the environment. We use $\mathcal{S}$ and $\mathcal{A}$ to denote the state and action spaces, respectively. Throughout this paper, we assume both $\mathcal{S}$ and $\mathcal{A}$ are continuous, and $\mathcal{S} \times \mathcal{A} \subseteq \mathbb{R}^(d + 1)$ with $d \geq 1$. We focus on the setting where the action space is single-dimensional, while the extension to multi-dimension is nature. Our method can also be applied in the general state-action space. The following two standard assumptions are imposed on the trajectory $\{S_t, A_t, R_t\}_{t\geq 0}$.

Assumption 1 *There exists a time-invariant transition kernel P such that for every $t \geq 0$, $s \in \mathcal{S}$, $a \in \mathcal{A}$ and any set $F \in \mathcal{B}(\mathcal{S})$,*

$$\Pr(S_{t+1} \in F \,|\, S_t = s, A_t = a, \{S_j, A_j, R_j\}_{0\leq j<t}) = P(S_{t+1} \in F \,|\, S_t = s, A_t = a),$$

where $\mathcal{B}(\mathcal{S})$ *is the family of Borel subsets of* $\mathcal{S}$ *and* $\{S_j, A_j, R_j\}_{0\leq j<t} = \emptyset$ *if* $t = 0$. *In addition, assume that for each* $(s,a) \in \mathcal{S}\times\mathcal{A}$, $P(\bullet \mid s,a)$ *is absolutely continuous with respect to the Lebeque measure and thus there exists a probability density function* $q(\bullet \mid s,a)$ *associated with* $P(\bullet \mid s,a)$.

Assumption 2 *The immediate reward* R_t *is a known function of* (S_t, A_t, S_{t+1}), *i.e.,* $R_t = \widetilde{R}(S_t, A_t, S_{t+1})$ *for any* $t \geq 0$, *where* $\widetilde{R} : \mathbb{R}^{2d+1} \to \mathbb{R}$. *In addition, we assume* R_t *is uniformly bounded, i.e., there exists a constant* $R_{\max}$ *such that* $|R_t| \leq R_{\max}$ *for every* $t \geq 0$.

By Assumption 2, we define a reward function as $r(s,a) = \mathbb{E}[R_t \mid S_t = s, A_t = a]$ for $t \geq 0$. The uniformly bounded assumption on the immediate reward R_t is used to simplify the technical proofs and can be relaxed.

An essential goal of RL is to find an optimal policy that maximizes the expected discounted sum of rewards. A policy is a decision rule that an agent chooses her action A_t based on the environment state S_t at each decision point t. In this paper, we focus on the stationary policy π, which is a function mapping from the state space $\mathcal{S}$ into a probability distribution over the action space $\mathcal{A}$. For each policy π, one can define a Q-function to measure its performance starting from any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, denoted by

$$Q^\pi(s,a) = \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t R_t \mid S_0 = s, A_0 = a\right], \tag{1}$$

where $\mathbb{E}^\pi$ refers to the expectation that all actions along the trajectory follow the stationary policy π. We evaluate the overall performance of a policy by

$$\mathcal{V}(\pi) = (1-\gamma)\mathbb{E}_{S_0\sim\nu}[Q^\pi(S_0, \pi(S_0))], \tag{2}$$

where ν is some *known* reference distribution. Here, for any function Q defined over $\mathcal{S}\times\mathcal{A}$, we define $Q(s,\pi(s)) = \int_{a\in\mathcal{A}} Q(s,a)\pi(a \mid s)da$. Our goal in this paper is to search for an optimal *in-class* policy π^* such that

$$\pi^* \in \arg\max_{\pi\in\Pi} \mathcal{V}(\pi), \tag{3}$$

where Π is some pre-specified class of stationary policies. Some commonly used policy classes include linear decision functions, neural networks and decision trees. The sufficiency of focusing on the stationary policies is guaranteed by Assumptions 1 and 2. See Section 6.2 of Puterman (1994) for the justification. We also remark that stationary policy does not rule out the deterministic policy, which is a function mapping from the state space $\mathcal{S}$ into the action space $\mathcal{A}$.

In this work, we focus on the batch setting, where the observed data consist of N independent and identically distributed copies of $\{S_t, A_t, R_t\}_{t\geq 0}$ up to T decision points. For the i-th trajectory, where $1 \leq i \leq N$, the data can be represented by $\{S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1}\}_{0\leq t<T}$. We aim to leverage the batch data to find the in-class optimal policy π^* defined in (3). The following standard assumption is imposed on our batch data generating process.

Assumption 3 *The batch data $\mathcal{D}_N = \{(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1})\}_{0\leq t<T, 1\leq i\leq N}$ is generated by a stationary policy π^b.*

Next, we introduce the average visitation probability measure. Let $q_t^{\pi^b}$ be the marginal probability measure over $\mathcal{S} \times \mathcal{A}$ at the decision point t induced by the behavior policy π^b. Then the average visitation probability measure across T decision points is defined as

$$\bar{d}_T^{\pi^b} = \frac{1}{T}\sum_{t=0}^{T-1} q_t^{\pi^b}.$$

The corresponding expectation with respect to $\bar{d}_T^{\pi^b}$ is denoted by $\overline{\mathbb{E}}$. Similarly, we can define the discounted visitation probability measure over $\mathcal{S} \times \mathcal{A}$ induced by a policy π as

$$d_\gamma^\pi = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t q_t^\pi, \tag{4}$$

where q_t^π is the marginal probability measure of (S_t, A_t) induced by the policy π with the initial state distribution ν. For notation simplicity, we also use $\bar{d}_T^{\pi^b}$ and d_γ^π to denote their corresponding probability density function when there is no confusion.

Additional notations: For generic sequences $\{\varpi(N)\}$ and $\{\theta(N)\}$, the notation $\varpi(N) \gtrsim \theta(N)$ (resp. $\varpi(N) \lesssim \theta(N)$) means that there exists a sufficiently large constant (resp. small) constant $c_1 > 0$ (resp. $c_2 > 0$) such that $\varpi(N) \geq c_1\theta(N)$ (resp. $\varpi(N) \leq c_2\theta(N)$). We use $\varpi(N) \asymp \theta(N)$ when $\varpi(N) \gtrsim \theta(N)$ and $\varpi(N) \lesssim \theta(N)$. For matrix and vector norms, we use $\| \bullet \|_{\ell_q}$ to denote either the vector ℓ_q-norm or operator norm induced by the vector ℓ_q-norm, for $1 \le q < \infty$, when there is no confusion. For any random variable X, we use $L^q_{\mathbb{P}}(X)$ to denote the class of all measurable functions with finite q-th moments for $1 \le q \le \infty$, where the underlying probability measure is $\mathbb{P}$. We may also write it as $L^q_{\mathbb{P}}$ when there is no confusion about the underlying random vectors. Then the L^q-norm is denoted by $\| \bullet \|_{L^q_{\mathbb{P}}(X)}$. We use $\| \bullet \|_\infty$ to denote the sup-norm. In addition, we often use (S, A, R, S') or (S, A, S') to represent a generic transition tuple. We define the maximum mean discrepancy (MMD) between two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ as

$$\text{MMD}_k(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \le 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{X \sim \mathbb{Q}}[f(X)], \tag{5}$$

where $\mathcal{H}_k$ is a reproducing kernel Hilbert space (RKHS) with the kernel k and the corresponding norm is denoted by $\|\bullet\|_{\mathcal{H}_k}$. Lastly, we use $\mathbb{Q} \ll \mathbb{P}$ when $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$, and $\mathbb{Q} \perp \mathbb{P}$ when $\mathbb{Q}$ is singular to $\mathbb{P}$.

3 An Illustrative Example: Contextual Bandit without Absolute Continuity

In this section, we use the offline contextual bandit problem (Manski, 2004) as an example to illustrate the challenge of policy learning without absolute continuity and our main idea to address it. Note that the contextual bandit problem is a special case of batch RL (i.e., $T = 0$).

Suppose we have a batch dataset $\mathcal{D}_N^0 = \{S_{i,0}, A_{i,0}, R_{i,0}\}_{1 \le i \le N}$ which contains i.i.d. copies from (S_0, A_0, R_0), where $S_0 \sim \nu_0$ and ν_0 is some unknown distribution. The target is to find an optimal in-class policy π^* such that

$$\pi^* \in \arg\max_{\pi \in \Pi} \left\{ \mathcal{V}_0(\pi) \triangleq \mathbb{E}_{S_0 \sim \nu} \left[r(S_0, \pi(S_0)) \right] \right\}. \tag{6}$$

Recall that $r(s, a) = \mathbb{E}\left[R_0 \mid S_0 = s, A_0 = a\right]$.

Most existing methods for solving (6) rely on a key assumption that $\nu \times \pi \ll \nu_0 \times \pi^b$ for all $\pi \in \Pi$, under which one is able to use the batch data $\mathcal{D}_N^0$ to calibrate π and evaluate its performance via estimating $\mathcal{V}_0(\pi)$ for all $\pi \in \Pi$. For instance, the regression-based approaches (Qian and Murphy, 2011; Bhattacharya and Dupas, 2012) consider a discrete action space, and require (i) $\pi^b(a \mid s)$ is uniformly bounded away from 0 for all s and a, and (ii) $\nu = \nu_0$. In this case, the absolute continuity holds. Meanwhile, the popular classification-based approaches (e.g., Zhao et al., 2012; Athey and Wager, 2021), which crucially rely on the inverse propensity weighting formulation, also have the same requirement. However, due to the distributional shift, such absolutely continuous assumption may not hold in general. There is a recent streamline of research studying contextual bandits under the framework of distributionally robust optimization such as Mo et al. (2021); Qi et al. (2022); Adjaho and Christensen (2022) and the reference therein. These works investigate policy learning in the presence of distributional shifts in the covariates or rewards when deploying the policy in the future. In particular, Adjaho and Christensen (2022) considers to use Wasserstein distance for quantifying the distributional shift and can allow the singularity for such a shift. However, none of them study the policy learning when there is a singularity issue in terms of the policy during the training procedure. For example, when π^b refers to some stochastic policy used to collect the data and Π is a class of deterministic policies, π becomes singular to π^b, i.e., $\pi \perp \pi^b$ for all $\pi \in \Pi$. In this case, most existing methods may fail to find a desirable policy.

To address such singularity issue induced by the deterministic policy with respect to the stochastic one, existing solutions either adopt the kernel smoothing techniques on π in order to approximately estimate $\mathcal{V}_0(\pi)$ (e.g., Chen et al., 2016) or impose some structure assumption on the reward function r (e.g., Chernozhukov et al., 2019). The first type of approaches requires the selection of kernel and the tuning on the bandwidth for approximating all deterministic policies, while the latter one could suffer from the model mis-specification due to the strong (parametric) model assumption. It is also well-known that the performance of the kernel smoothing deteriorates when the dimension of the action space increases. In addition, these methods cannot handle general settings such as that ν is also singular to ν_0, i.e., there exists a *covariate shift* problem. This problem becomes more severe when considering the dynamic setting.

In the following, we present a brief description of our method for policy learning without requiring absolute continuity of the target distribution (i.e., $\nu \times \pi$) with respect to the data distribution (i.e., $\nu_0 \times \pi^b$) in this contextual bandit problem. The idea relies on the following observation. Consider a fix policy $\pi \in \Pi$. Let $\widetilde{r}$ be any estimator for the reward function. It can be easily seen that

$$\mathcal{V}_0(\pi) - \mathbb{E}_{S_0 \sim \nu}\left[\widetilde{r}(S_0, \pi(S_0))\right] = \mathbb{E}_{(S_0,A_0)\sim \nu \times \pi}\left[(r - \widetilde{r})(S_0, A_0)\right]. \tag{7}$$

By leveraging Lebesgue's Decomposition Theorem, we can always represent $\nu \times \pi$ as

$$\nu \times \pi = \widetilde{\lambda}_1^\pi + \widetilde{\lambda}_2^\pi,$$

where $\widetilde{\lambda}_1^\pi \ll (\nu_0 \times \pi^b)$ and $\widetilde{\lambda}_2^\pi \perp (\nu_0 \times \pi^b)$. Since $\widetilde{\lambda}_1^\pi$ and $\widetilde{\lambda}_2^\pi$ are finite measures, we normalize them and rewrite the above decomposition as

$$\nu \times \pi = \widetilde{\lambda}_1^\pi(\mathcal{S} \times \mathcal{A})\lambda_1^\pi + \widetilde{\lambda}_2^\pi(\mathcal{S} \times \mathcal{A})\lambda_2^\pi,$$

where λ_1^π and λ_2^π are two probability measures. In particular, if $\widetilde{\lambda}_i^\pi(\mathcal{S} \times \mathcal{A}) = 0$, the corresponding λ_i^π can be chosen as an arbitrary probability measure. Given this decomposition, one can further show that

$$\begin{aligned}
&\mathcal{V}_0(\pi) - \mathbb{E}_{S_0 \sim \nu}\left[\widetilde{r}(S_0, \pi(S_0))\right] \\
=&\widetilde{\lambda}_1^\pi(\mathcal{S} \times \mathcal{A}) \times \underbrace{\mathbb{E}_{(S_0,A_0)\sim \nu_0 \times \pi^b}\left[\frac{\lambda_1^\pi(S_0, A_0)}{(\nu_0 \times \pi^b)(S_0, A_0)}(r - \widetilde{r})(S_0, A_0)\right]}_{\text{absolutely continuous part}} \\
+&\widetilde{\lambda}_2^\pi(\mathcal{S} \times \mathcal{A}) \times \underbrace{\mathbb{E}_{(S_0,A_0)\sim \lambda_2^\pi}\left[(r - \widetilde{r})(S_0, A_0)\right]}_{\text{singular part}}.
\end{aligned}$$

In order to have an accurate estimation of $\mathcal{V}_0(\pi)$, one needs to find an $\widetilde{r}$ such that the absolutely continuous and singular parts in the above equality are minimized. For the absolutely continuous part, since λ_1^π is unknown, define $\mathcal{W}$ as some class of symmetric functions and assume $\lambda_1^\pi/(\nu_0 \times \pi^b) \in \mathcal{W}$, we can show that

$$\begin{aligned}
&\left|\mathbb{E}_{(S_0,A_0)\sim \nu_0 \times \pi^b}\left[\frac{\lambda_1^\pi(S_0, A_0)}{(\nu_0 \times \pi^b)(S_0, A_0)}(r - \widetilde{r})(S_0, A_0)\right]\right| \\
\leq& \sup_{w \in \mathcal{W}} \mathbb{E}_{(S_0,A_0)\sim \nu_0 \times \pi^b}\left[w(S_0, A_0)(R_0 - \widetilde{r}(S_0, A_0))\right].
\end{aligned}$$

The right-hand side of the above inequality can be properly controlled using the batch data because (S_0, A_0) now follows $\nu_0 \times \pi^b$. To handle the singular part, we adopt the idea of distributionally robust optimization (e.g., Rahimian and Mehrotra, 2019) with MMD. Suppose there exists a constant $\delta > 0$ such that $\mathrm{MMD}_k(\lambda_2^\pi, \nu_0 \times \pi^b) \leq \delta$, then we can show that

$$\left|\mathbb{E}_{(S_0,A_0)\sim\lambda_2^\pi}\left[(r-\widetilde{r})(S_0,A_0)\right]\right| \leq \sup_{\mathbb{P}:\mathrm{MMD}_k(\mathbb{P},\nu_0\times\pi^b)\leq\delta} \left|\mathbb{E}_{(S_0,A_0)\sim\mathbb{P}}\left[r(S_0,A_0)-\widetilde{r}(S_0,A_0)\right]\right|.$$

Summarizing the above derivations together, we can quantify the error of using $\widetilde{r}$ for estimating $\mathcal{V}_0(\pi)$ as

$$\begin{aligned}
&\left|\mathcal{V}_0(\pi)-\mathbb{E}_{S_0\sim\nu}\left[\widetilde{r}(S_0,\pi(S_0))\right]\right| \\
\leq&\widetilde{\lambda}_1^\pi(\mathcal{S}\times\mathcal{A})\sup_{w\in\mathcal{W}}\mathbb{E}_{(S_0,A_0)\sim\nu_0\times\pi^b}\left[w(S_0,A_0)(R_0-\widetilde{r}(S_0,A_0))\right] \\
&+\widetilde{\lambda}_2^\pi(\mathcal{S}\times\mathcal{A})\sup_{\mathbb{P}:\mathrm{MMD}_k(\mathbb{P},\nu_0\times\pi^b)\leq\delta}\left|\mathbb{E}_{(S_0,A_0)\sim\mathbb{P}}\left[r(S_0,A_0)-\widetilde{r}(S_0,A_0)\right]\right| \\
\leq&\widetilde{\lambda}_1^\pi(\mathcal{S}\times\mathcal{A})\sup_{w\in\mathcal{W}}\mathbb{E}_{(S_0,A_0)\sim\nu_0\times\pi^b}\left[w(S_0,A_0)(R_0-\widetilde{r}(S_0,A_0))\right] \\
&+\widetilde{\lambda}_2^\pi(\mathcal{S}\times\mathcal{A})\left|\mathbb{E}_{(S_0,A_0)\sim\nu_0\times\pi^b}\left[R_0-\widetilde{r}(S_0,A_0)\right]\right|+\delta\|\mathbb{E}\left[R_0-\widetilde{r}(S_0,A_0)\,|\,S_0=\bullet,A_0=\bullet\right]\|_{\mathcal{H}_k},
\end{aligned}$$

where the last inequality is given by Lemma 2 in the later section under some mild conditions. Moreover, it can be checked that if $\widetilde{r}=r$, the right-hand side of the above inequality vanishes. These facts motivate us to estimate the reward function r and later $\mathcal{V}_0(\pi)$ via

$$\begin{aligned}
\min_{\widetilde{r}}\Bigg\{&\sup_{w\in\mathcal{W}}\mathbb{E}_{(S_0,A_0)\sim\nu_0\times\pi^b}\left[w(S_0,A_0)(R_0-\widetilde{r}(S_0,A_0))\right] \\
&+\left|\mathbb{E}_{(S_0,A_0)\sim\nu_0\times\pi^b}\left[R_0-\widetilde{r}(S_0,A_0)\right]\right|+\delta\|\mathbb{E}\left[R_0-\widetilde{r}(S_0,A_0)\,|\,S_0=\bullet,A_0=\bullet\right]\|_{\mathcal{H}_k}\Bigg\}.
\end{aligned} \tag{8}$$

Once we are able to provide a valid estimation for $\mathcal{V}_0(\pi)$, an optimization rountine can be implemented to estimate π^*, where we incorporate the idea of pessimism. Specifically, for each given $\widetilde{r}$, we first implement the kernel ridge regression using $\mathcal{D}_N^0$ for estimating $\mathbb{E}\left[R_0-\widetilde{r}\,|\,S_0=\bullet,A_0=\bullet\right]$. Denote the resulting estimator by $\widehat{\mathbb{E}}\left[R_0-\widetilde{r}\,|\,S_0=\bullet,A_0=\bullet\right]$. Note that $\widetilde{r}$ is some candidate reward function. Then we propose to compute the optimal policy via solving the min-max problem

$$\begin{aligned}
\max_{\pi\in\Pi}\min_{\widetilde{r}\in\mathcal{R}}\quad&\mathbb{E}_{S_0\sim\nu}\left[\widetilde{r}(S_0,\pi(S_0))\right], \\
\text{subject to}\quad&\sup_{w\in\mathcal{W}}\mathbb{E}_N\left[w(S_0,A_0)(R_0-\widetilde{r}(S_0,A_0))\right]\leq\varepsilon_1 \\
&\|\widehat{\mathbb{E}}\left[R_0-\widetilde{r}\,|\,S_0=\bullet,A_0=\bullet\right]\|_{\mathcal{H}_k}\leq\varepsilon_2,
\end{aligned} \tag{9}$$

where $\mathcal{R}$ is some pre-specified class of functions for modeling the true reward function r and $\mathbb{E}_N$ refers to the empirical average over the batch data $\mathcal{D}_N^0$. Two constants $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ are used to quantify the uncertainty for estimating r and also control the degree of pessimism. Under some technical conditions, with properly chosen ε_1 and ε_2, we can show that the true reward function r always belongs to the feasible set of (9). Therefore, by implementing the above algorithm, we search a policy that maximizes the most pessimistic estimation of its value $\mathcal{V}_0(\pi)$ within the corresponding uncertainty set. The proposed algorithm will produce a valid policy with regret guaranteed and without assuming absolute continuity. The details of our contextual bandit algorithm can be found in Section 6.

4 Policy Learning in Batch RL

In this section, we consider policy learning in batch RL and generalize the idea in the last section with more details. Our proposed algorithm for obtaining π^* is motivated by off-policy evaluation (OPE). For any given policy π, the target of OPE is to use the batch data to estimate the policy value $\mathcal{V}(\pi)$ defined in (2). Note that by the definition of the discounted visitation probability measure, we can show that

$$\mathcal{V}(\pi) = \int_{\mathcal{S}\times\mathcal{A}} r(s,a) d_\gamma^\pi(\mathrm{d}s, \mathrm{d}a),$$

where $r(s,a) = \mathbb{E}\left[R_t \mid S_t = s, A_t = a\right]$ for any $t \geq 0$. A direct approach to perform OPE is via estimating Q^π defined in (1). Let $\widetilde{Q}$ be any estimator for Q^π and define the corresponding estimator for the policy value as $\widetilde{\mathcal{V}}(\pi) = (1-\gamma)\mathbb{E}_{S_0\sim\nu}[\widetilde{Q}(S_0, \pi(S_0))]$. Then we have the following lemma to characterize the estimation error of $\widetilde{\mathcal{V}}(\pi)$ to $\mathcal{V}(\pi)$.

Lemma 1 *Under Assumptions 1 and 2, we have*

$$\mathcal{V}(\pi) - \widetilde{\mathcal{V}}(\pi) = \mathbb{E}_{(S,A)\sim d_\gamma^\pi}\left[R + \gamma\widetilde{Q}(S', \pi(S')) - \widetilde{Q}(S,A)\right]. \tag{10}$$

Based on Lemma 1, if $d_\gamma^\pi \ll \bar{d}_T^{\pi^b}$, then by change of measure and assuming that $\frac{d_\gamma^\pi}{\bar{d}_T^{\pi^b}} \in \mathcal{W}$, where recall that $\mathcal{W}$ is a symmetric class of functions, we can obtain that

$$|\mathcal{V}(\pi) - \widetilde{\mathcal{V}}(\pi)| = \left| \overline{\mathbb{E}} \left[\frac{d_\gamma^\pi(S,A)}{\bar{d}_T^{\pi^b}(S,A)} \left(R + \gamma \widetilde{Q}(S', \pi(S')) - \widetilde{Q}(S,A) \right) \right] \right|$$
$$\leq \sup_{w \in \mathcal{W}} \overline{\mathbb{E}} \left[w(S,A) \left(R + \gamma \widetilde{Q}(S', \pi(S')) - \widetilde{Q}(S,A) \right) \right].$$

This naturally motivates a min-max estimating approach to learn Q^π and hence $\mathcal{V}(\pi)$, i.e.,

$$\min_{\widetilde{Q}} \sup_{w \in \mathcal{W}} \overline{\mathbb{E}} \left[w(S,A) \left(R + \gamma \widetilde{Q}(S', \pi(S')) - \widetilde{Q}(S,A) \right) \right],$$

which has been used in the literature of OPE (e.g., Jiang and Huang, 2020). As discussed before, due to the potentially large state-action space, it is quite usual that some values of the state-action pairs induced by the target policy π are not covered by the batch data generating process. In addition, there are many applications where $\pi \in \Pi$ is a deterministic policy but the behavior one is stochastic (e.g., Silver et al., 2014; Lillicrap et al., 2015). In either of these cases, $d_\gamma^\pi \ll \bar{d}_T^{\pi^b}$ fails to hold and the above min-max estimating approach may no longer be valid for OPE as it cannot upper bound the estimation error of $\mathcal{V}(\pi)$. Indeed, similar to the contextual bandit example discussed in the previous section, most of existing OPE (and policy learning) approaches in batch RL rely on this absolute continuity assumption, which could be illusive in practice.

Motivated by this gap, in the following, we *do not assume* the absolute continuity of d_γ^π with respect to $\bar{d}_T^{\pi^b}$, i.e., there exists a measurable subset $F \in \mathcal{B}(\mathcal{S} \times \mathcal{A})$ such that $\bar{d}_T^{\pi^b}(F) = 0$ but $d_\gamma^\pi(F) \neq 0$. Again, by Lebesgue's Decomposition Theorem, with some abuse of notations, there always exist two finite measures $\widetilde{\lambda}_1^\pi$ and $\widetilde{\lambda}_2^\pi$ over $\mathcal{B}(\mathcal{S} \times \mathcal{A})$ such that

$$d_\gamma^\pi = \widetilde{\lambda}_1^\pi + \widetilde{\lambda}_2^\pi,$$

where $\widetilde{\lambda}_1^\pi \ll \bar{d}_T^{\pi^b}$ and $\widetilde{\lambda}_2^\pi \perp \bar{d}_T^{\pi^b}$. By normalization, we can rewrite the above decomposition as

$$d_\gamma^\pi = \widetilde{\lambda}_1^\pi(\mathcal{S} \times \mathcal{A})\lambda_1^\pi + \widetilde{\lambda}_2^\pi(\mathcal{S} \times \mathcal{A})\lambda_2^\pi,$$

where λ_1^π and λ_2^π are two probability measures. Similar to the contextual bandit example, we can decompose the OPE error as

$$\mathcal{V}(\pi) - \widetilde{\mathcal{V}}(\pi) = \widetilde{\lambda}_1^\pi(\mathcal{S}\times\mathcal{A}) \underbrace{\mathbb{E}_{(S,A)\sim\lambda_1^\pi}\left[R + \gamma\widetilde{Q}(S', \pi(S')) - \widetilde{Q}(S,A)\right]}_{\text{absolute continuous part}}$$
$$+ \widetilde{\lambda}_2^\pi(\mathcal{S}\times\mathcal{A}) \underbrace{\mathbb{E}_{(S,A)\sim\lambda_2^\pi}\left[R + \gamma\widetilde{Q}(S', \pi(S')) - \widetilde{Q}(S,A)\right]}_{\text{singular part}}.$$

The absolutely continuous part can be properly controlled by using the above min-max formulation. However, the singularity of λ_2^π with respect to $\bar{d}_T^{\pi^b}$ is the major obstacle that makes most existing OPE approaches not applicable. To address this issue, One must rely on the extrapolation ability of Q^π. As the difficulty of OPE largely comes from the distributional mismatch between d_γ^π and $\bar{d}_T^{\pi^b}$, we leverage the kernel mean embedding approach to quantifying the difference between λ_2^π and $\bar{d}_T^{\pi^b}$ in order to control the singular part. See Lemma 6 in the appendix for the existence of kernel mean embeddings, which follows from Lemma 3 of Gretton et al. (2012).

Next, we present our formal result in bounding the OPE error of $\widetilde{\mathcal{V}}(\pi)$. Define the Bellman operator as

$$\mathcal{T}^\pi\widetilde{Q}(\bullet,\bullet) = \mathbb{E}[R + \gamma\widetilde{Q}(S',\pi(S')) - \widetilde{Q}(S,A) \,|\, S=\bullet, A=\bullet]$$

for any transition tuple (S, A, R, S'). Since $\lambda_1^\pi \ll \bar{d}_T^{\pi^b}$, by the Radon-Nikodym Theorem, we can define the Radon-Nikodym derivative as

$$\omega^\pi(s,a) = \frac{\lambda_1^\pi(s,a)}{\bar{d}_T^{\pi^b}(s,a)},$$

for every $(s,a)\in\mathcal{S}\times\mathcal{A}$. Recall that $\mathcal{W}$ is a symmetric class of functions defined over $\mathcal{S}\times\mathcal{A}$. We have the following key lemma for our proposal. For notation simplicity, let $Z = (S, A)$ and denote $\mathcal{Z} = \mathcal{S}\times\mathcal{A}$.

Lemma 2 *Let the kernel $k(\bullet,\bullet):\mathcal{Z}\times\mathcal{Z}\to\mathbb{R}$ be measurable with respect to both λ_2^π and $\bar{d}_T^{\pi^b}$, and $\max\{\overline{\mathbb{E}}[\sqrt{k(Z,Z)}], \mathbb{E}_{Z\sim\lambda_2^\pi}[\sqrt{k(Z,Z)}]\} < +\infty$. Let $\omega^\pi\in\mathcal{W}$ and $\mathcal{T}^\pi\widetilde{Q}\in\mathcal{H}_k$. In addition, for any policy π, suppose there exists a constant $\delta > 0$ such that $MMD_k(\bar{d}_T^{\pi^b}, \lambda_2^\pi) \leq$*

δ. *Then the following holds.*

$$
\begin{aligned}
|\mathcal{V}(\pi)-\widetilde{\mathcal{V}}(\pi)| \leq & \widetilde{\lambda}_1^{\pi}(\mathcal{S}\times\mathcal{A}) \times \sup_{w\in\mathcal{W}} \overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right)\right] \\
& + \widetilde{\lambda}_2(\mathcal{S}\times\mathcal{A})\left\{\left|\overline{\mathbb{E}}\left[\left(R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right)\right]\right| + \delta\|\mathcal{T}^{\pi}\widetilde{Q}\|_{\mathcal{H}_k}\right\}. \quad (11)
\end{aligned}
$$

The existence of δ is mild. For example, as long as the conditions stated in Lemma 6 of the appendix hold, δ always exists. Lemma 2 provides an upper bound for the estimation error of OPE using any estimator $\widetilde{Q}$ for Q^{π}. Moreover, it can be further checked that if $\widetilde{Q} = Q^{\pi}$, by Bellman equation,

$$
\begin{aligned}
& \widetilde{\lambda}_1^{\pi}(\mathcal{S}\times\mathcal{A}) \times \sup_{w\in\mathcal{W}} \overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma Q^{\pi}(S',\pi(S'))-Q^{\pi}(S,A)\right)\right] \\
+ & \widetilde{\lambda}_2(\mathcal{S}\times\mathcal{A})\left\{\left|\overline{\mathbb{E}}\left[\left(R+\gamma Q^{\pi}(S',\pi(S'))-Q^{\pi}(S,A)\right)\right]\right| + \delta\|\mathcal{T}^{\pi}Q^{\pi}\|_{\mathcal{H}_k}\right\} = 0.
\end{aligned}
$$

Based on the above analysis, we propose to estimating Q^{π} by solving the following min-max problem.

$$
\begin{aligned}
\min_{\widetilde{Q}} \sup_{w\in\mathcal{W}} \quad & \overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right)\right] \\
& + \left|\overline{\mathbb{E}}\left[\left(R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right)\right]\right| + \delta\|\mathcal{T}^{\pi}\widetilde{Q}\|_{\mathcal{H}_k}.
\end{aligned} \tag{12}
$$

Without loss of generality, we assume that $\{1,-1\}\in\mathcal{W}$. In this case, optimization problem (12) is equivalent to

$$
\min_{\widetilde{Q}} \sup_{w\in\mathcal{W}} \quad \overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right)\right] + \delta\|\mathcal{T}^{\pi}\widetilde{Q}\|_{\mathcal{H}_k} \tag{13}
$$

This method will serve as the foundation of our proposed algorithm developed below.

Recall that our goal is to leverage the batch data $\mathcal{D}_N$ to estimate the in-class optimal policy π^*. For any policy π, we can implement the empirical version of (13) for obtaining the estimation of Q^{π} and performing the corresponding OPE. This motivates us to develop a pessimistic policy iteration algorithm for finding π^*. Pessimism in batch RL serves as the main tool for conservative policy optimization by quantifying the uncertainty of the estimation and discouraging the exploration of the learned policy from visiting the less explored state-action pair in the batch data. The success of the pessimsitic-typed algorithms has been demonstrated in many applications (e.g., Kumar et al., 2019; Bai et al., 2022). In terms of the data coverage, instead of the full-coverage assumption (i.e., $\bar{d}_T^{\pi^b}$ is uniformly bounded away from 0) required by many classic RL algorithms, algorithms with a proper degree of pessimism only require that the in-class optimal policy π^* is covered by the behavior one, which is thus more desirable. Adapting the pessimism idea to our setting without assuming absolute continuity, we expect to develop an algorithm that finds the optimal policy by at most requiring the data coverage assumption such that $\omega^{\pi^*} \in \mathcal{W}$ and $\text{MMD}_k(\bar{d}_T^{\pi^b}, \lambda_2^{\pi^*}) \leq \delta$, This requirement is much weaker than those needed by all existing RL algorithms.

A key building block of our proposed algorithm is to construct the following two uncertainty sets for Q^π with $\pi \in \Pi$. Let $\mathcal{F}$ be some pre-specified class of functions over $\mathcal{S} \times \mathcal{A}$, which is used to model Q^π, and define the following two uncertainty sets

$$\Omega_1(\pi, \mathcal{F}, \mathcal{W}, \varepsilon_{NT}^{(1)}) = \left\{ Q \in \mathcal{F} \mid \sup_{w \in \mathcal{W}} \overline{\mathbb{E}}_{NT} \left[w(S, A) \left(R + \gamma Q(S', \pi(S')) - Q(S, A) \right) \right] \leq \varepsilon_{NT}^{(1)} \right\} \tag{14}$$

$$\Omega_2(\pi, \mathcal{F}, \varepsilon_{NT}^{(2)}) = \left\{ Q \in \mathcal{F} \mid \|\widehat{\mathcal{T}}^\pi Q\|_{\mathcal{H}_k} \leq \varepsilon_{NT}^{(2)} \right\},$$

where $\varepsilon_{NT}^{(1)} > 0$ and $\varepsilon_{NT}^{(2)} > 0$ are some constants depending on N and T, which will be specified later. Here

$$\overline{\mathbb{E}}_{NT} \left[f(S, A, R, S') \right] = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=0}^{T} f(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1}),$$

for any generic function f and

$$\widehat{\mathcal{T}}^\pi Q \in \arg\min_{f \in \mathcal{H}_k} \overline{\mathbb{E}}_{NT} \left[\left(R + \gamma Q(S', \pi(S')) - Q(S, A) - f(S, A) \right)^2 \right] + \zeta_{NT} \|f\|_{\mathcal{H}_k}^2, \tag{15}$$

with the regularization parameter $\zeta_{NT} > 0$. Denote

$$Y(\pi, Q) = (Y_{1,0}(\pi, Q), \cdots, Y_{1,T-1}(\pi, Q), Y_{2,0}, \cdots, Y_{N,T-1}(\pi, Q)) \in \mathbb{R}^{NT} \quad \text{with}$$
$$Y_{i,t}(\pi, Q) = R_{i,t} + \gamma Q(S_{i,t+1}, \pi(S_{i,t+1})) - Q(S_{i,t}, A_{i,t}), \quad \text{and let}$$
$$K \in \mathbb{R}^{NT \times NT} \quad \text{with} \quad K_{i,j} = k(Z_{\lfloor i/T \rfloor + 1, i \bmod T-1}, Z_{\lfloor j/T \rfloor + 1, j \bmod T-1}).$$

Thanks to the representer theorem, we can show that

$$\|\widehat{\mathcal{T}}^\pi Q\|_{\mathcal{H}_k}^2 = Y(\pi, Q)^\top (K + \zeta_{NT} I_{NT})^{-1} K (K + \zeta_{NT} I_{NT})^{-1} Y(\pi, Q), \tag{16}$$

where I_{NT} is an identity matrix with the dimension NT. Based on the defined uncertainty sets $\Omega_j, j = 1, 2$, we propose to estimate the in-class optimal policy π^* via

$$\max_{\pi\in\Pi} \min_{Q\in\Omega_1(\pi,\mathcal{F},\mathcal{W},\varepsilon_{NT}^{(1)})\cap\Omega_2(\pi,\mathcal{F},\varepsilon_{NT}^{(2)})} (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi(S_0))\right], \tag{17}$$

i.e., maximizing the worst case of the policy value within the uncertainty sets. Note that our optimization problem (17) is *free* of δ. Denote the resulting estimated policy as $\widehat{\pi}$. In Section 5, we provide a regret guarantee for $\widehat{\pi}$ in finding π^* under minimal data coverage assumption.

The policy optimization problem (17) may be difficult to solve because of the constraint set for Q, especially when $\mathcal{F}$ and $\mathcal{W}$ are highly complex such as neural networks. In the following, we propose to solve it via the dual formulation of the inner minimization problem in (17). Specifically, let $\rho = (\rho_1, \rho_2)$ be dual variables (i.e., Lagrange multipliers). Define a Lagrangian function as

$$\begin{aligned} L_{NT}(Q,\rho,\pi) = &(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi(S_0))\right] \\ &+ \rho_1 \times \left\{\sup_{w\in\mathcal{W}} \overline{\mathbb{E}}_{NT}\left[w(S,A)\left(R+\gamma Q(S',\pi(S')) - Q(S,A)\right)\right] - \varepsilon_{NT}^{(1)}\right\} \\ &+ \rho_2 \times \left\{\|\widehat{\mathcal{T}}^{\pi}Q\|_{\mathcal{H}_k} - \varepsilon_{NT}^{(2)}\right\}. \end{aligned} \tag{18}$$

Based on the formulation of $L_{NT}(Q,\rho,\pi)$, we consider an alternative way to estimating the optimal policy by solving the optimization problem

$$\max_{\pi\in\Pi,\rho\succeq 0} \min_{Q\in\mathcal{F}} L_{NT}(Q,\rho,\pi), \tag{19}$$

where $\succeq$ refers to the component-wise comparison. Note that compared with (17), Problem (19) can be solved in a more efficient manner as the optimization with respect to the two dual variables can be easily performed. In Section 6, we will provide more details on the computational aspect of this approach. Denote the resulting estimated policy given by (19) as $\widehat{\pi}_{\text{dual}}$. By the weak duality, it can always be shown that

$$\max_{\rho\succeq 0}\min_{Q\in\mathcal{F}} L_{NT}(Q,\rho,\pi) \le \min_{Q\in\Omega_1(\pi,\mathcal{F},\mathcal{W},\varepsilon_{NT}^{(1)})\cap\Omega_2(\pi,\mathcal{F},\varepsilon_{NT}^{(2)})} (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi(S_0))\right].$$

Therefore, solving Problem (19) can also be viewed as a policy optimization algorithm with pessimism. However, the strong duality for the primal and dual problems may not hold as the primal problem is not convex with respect to Q, which therefore leads to that $\widehat{\pi} \neq \widehat{\pi}_{\text{dual}}$ in general. More seriously, the existence of the duality gap will result in a non-negligible regret of $\widehat{\pi}_{\text{dual}}$ in finding π^*. Nevertheless, in Section 5 below, we show that under one additional mild assumption, $\widehat{\pi}_{\text{dual}}$ indeed can achieve the same regret guarantee as that of $\widehat{\pi}$.

5 Theoretical Results

The performance of a policy optimization algorithm is often measured by the difference between the value of the (in-class) optimal policy and that of the estimated one, which is referred to as the *regret*. In our case, we evaluate the performance of any estimated policy $\widetilde{\pi} \in \Pi$ via the regret defined as

$$\text{Regret}(\widetilde{\pi}) = \mathcal{V}(\pi^*) - \mathcal{V}(\widetilde{\pi}). \tag{20}$$

Clearly, $\text{Regret}(\widetilde{\pi}) \geq 0$. In this section, we aim to derive the finite-sample upper bounds for both $\text{Regret}(\widehat{\pi})$ and $\text{Regret}(\widehat{\pi}_{\text{dual}})$. To begin with, we list several technical assumptions and discuss their corresponding implications.

5.1 Technical Assumptions

Assumption 4 *The stochastic process $\{S_t, A_t\}_{t\geq 0}$ induced by the behavior policy π^b is stationary, exponentially $\boldsymbol{\beta}$-mixing. The $\boldsymbol{\beta}$-mixing coefficient at time lag j satisfies that $\beta(j) \leq \beta_0 \exp(-\beta_1 j)$ for $\beta_0 \geq 0$ and $\beta_1 > 0$. The induced stationary distribution is denoted by d^{π^b}.*

Our batch data consist of multiple trajectories, where observations in each trajectory follow a MDP and thus are dependent. Assumption 4 characterizes the dependency among those observations, instead of assuming transition tuples are all independent as in many previous works. This assumption has been used in recent works (e.g., Chen and Qi, 2022). The upper bound on the $\boldsymbol{\beta}$-mixing coefficient at time lag j indicates that the dependency between

$\{S_t, A_t\}_{t\leq k}$ and $\{S_t, A_t\}_{t\geq(k+j)}$ decays to 0 at least exponentially fast with respect to j. See Bradley (2005) for the exact definition of the exponentially $\boldsymbol{\beta}$-mixing. Recall that we have let $Z = (S, A)$ and denoted $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$.

Assumption 5 *The kernel $k(\bullet, \bullet) : \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}$ is positive definite, and $\sup_{Z\in\mathcal{Z}} \sqrt{k(Z, Z)} = \kappa < +\infty$. For any $\pi \in \Pi$, k is measurable with respect to both d^π_γ and λ^π_2.*

The bounded and measurable conditions on the kernel k in Assumption 5 are mild, which can be easily satisfied by many popular kernels such as the Gaussian kernel. Assumption 5 ensures that the conditions in Lemma 6 of the appendix hold so that the kernel mean embeddings exist. Assumption 5 also indicates that $\mathcal{H}_k$ is compactly embedded in $L^2_{\bar{d}^{\pi^b}_T}(\mathcal{Z})$, e.g., for any $f \in \mathcal{H}_k$, $f \in L^2_{\bar{d}^{\pi^b}_T}(\mathcal{Z})$. See Definition 2.1 and Lemma 2.3 of Steinwart and Scovel (2012) for more details. Define an integral operator $\mathcal{L}_k : L^2_{d^{\pi^b}}(\mathcal{Z}) \to L^2_{d^{\pi^b}}(\mathcal{Z})$ as

$$\mathcal{L}_k f(z) \triangleq \int_{\mathcal{Z}} k(z, \widetilde{z}) f(\widetilde{z}) d^{\pi^b}(\widetilde{z}) \mathrm{d}z, \quad \text{for } z \in \mathcal{Z} \text{ and } f \in L^2_{\bar{d}^{\pi^b}_T}(\mathcal{Z}). \tag{21}$$

Assumption 5 implies that the self-adjoint operator $\mathcal{L}_k$ is compact and enjoys a spectral representation such that for any $f \in L^2_{\bar{d}^{\pi^b}_T}(\mathcal{Z})$,

$$\mathcal{L}_k f = \sum_{i=1}^{+\infty} e_j \langle \phi_j, f\rangle_{L^2_{\bar{d}^{\pi^b}_T}(\mathcal{Z})} \phi_j, \tag{22}$$

where $\langle \bullet, \bullet\rangle_{L^2_{\bar{d}^{\pi^b}_T}(\mathcal{Z})}$ is referred to as the inner product in $L^2_{d^{\pi_b}}(\mathcal{Z})$, $\{e_j\}_{j\geq 1}$ is a non-increasing sequence of eigenvalues towards 0 and $\{\phi_j\}_{j\geq 1}$ are orthonormal bases of $L^2_{\bar{d}^{\pi^b}_T}(\mathcal{Z})$. See Theorem 2.11 of Steinwart and Scovel (2012) for a justification of such spectral decomposition. We will use $\mathcal{L}_k$ to characterize the bias/approximation error induced by the regularization (15). See Assumption 7 (b) and the comment afterwards. In the following, we quantify the complexities of function classes used in our policy optimization algorithm via the ϵ-covering number. See definition of the ϵ-covering number in C.1 of the appendix.

Assumption 6 *The following conditions hold:*

(a) For any $Q \in \mathcal{F}$, $\|Q\|_\infty \leq c_{\mathcal{F}} < +\infty$. For any $Q \in \mathcal{F}$, $s \in \mathcal{S}$, and $a_1, a_2 \in \mathcal{A}$,

$$|Q(s, a_1) - Q(s, a_2)| \lesssim |a_1 - a_2|. \tag{23}$$

In addition, for any $\epsilon > 0$, we have

$$\mathcal{N}(\epsilon, \mathcal{F}, \|\bullet\|_\infty) \lesssim \left(\frac{1}{\epsilon}\right)^{v(\mathcal{F})}, \tag{24}$$

where $v(\mathcal{F}) > 0$ is some constant.

(b) For any $w \in \mathcal{W}$, $\|w\|_\infty \leq c_{\mathcal{W}} < +\infty$. In addition, for any $\epsilon > 0$, we have

$$\mathcal{N}(\epsilon, \mathcal{W}, \|\bullet\|_\infty) \lesssim \left(\frac{1}{\epsilon}\right)^{v(\mathcal{W})}, \tag{25}$$

where $v(\mathcal{W}) > 0$ is some constant.

(c) The policy class Π is a class of deterministic policies, i.e., $\pi : \mathcal{S} \to \mathcal{A}$. The action space $\mathcal{A}$ is bounded. In addition, for any $\epsilon > 0$, we have

$$\mathcal{N}(\epsilon, \Pi, \|\bullet\|_\infty) \lesssim \left(\frac{1}{\epsilon}\right)^{v(\Pi)}, \tag{26}$$

where $v(\Pi) > 0$ is some constant.

Assumption 6 imposes metric entropy conditions on function classes $\mathcal{F}$, $\mathcal{W}$ and Π. For simplicity, we consider uniformly bounded classes for deriving the exponential inequalities (Van Der Vaart and Wellner, 1996). An example of these classes could be sparse neural networks studied in Schmidt-Hieber (2020). We remark that $v(\mathcal{F})$, $v(\mathcal{W})$ and $v(\Pi)$ could increase with respect to N and T. The Lipschitz-typed condition in (23) is imposed so that the ϵ-covering number of the following class of functions:

$$\widetilde{\mathcal{F}} = \{R + \gamma Q(S', \pi(S')) - Q(S, A) \mid Q \in \mathcal{F}, \pi \in \Pi\}$$

could be properly controlled by the ϵ-covering numbers of Π and $\mathcal{F}$ (Van Der Vaart and Wellner, 1996).

Assumption 7 *The following conditions hold.*

(a) For any $\pi \in \Pi$, $Q^\pi \in \mathcal{F}$.

(b) For any $\pi \in \Pi$, $Q \in \mathcal{F}$, and some constant $c \in (1/2, 3/2]$, there exists $g_{\pi,Q} \in L^2_{d^{\pi_b}}(\mathcal{Z})$ such that $\mathcal{L}_k^c g_{\pi,Q} = \mathcal{T}^\pi Q$ and $\sup_{\pi\in\Pi, Q\in\mathcal{F}} \|g_{\pi,Q}\| < +\infty$.

(c) For any $\pi_1, \pi_2 \in \Pi$ and $Q_1, Q_2 \in \mathcal{F}$, $\|g_{\pi_1,Q_1} - g_{\pi_2,Q_2}\|_{L^2_{d^{\pi_b}}} \lesssim \|\pi_1 - \pi_2\|_\infty + \|Q_1 - Q_2\|_\infty$.

Assumption 7 (a) is called the realizability of Q-function for all policies, which has been widely imposed in the literature of batch RL (e.g., Antos et al., 2008*b*). Realizability indicates that there is no model misspecification error for estimating Q^π for every $\pi \in \Pi$.

Assumption 7 (b) imposes a smoothness condition on $\mathcal{T}^\pi Q$. It basically states that for every $\pi \in \Pi$ and $Q \in \mathcal{F}$, $\mathcal{L}_k^{-c}(\mathcal{T}^\pi Q) = g_{\pi,Q} \in L^2_{d^{\pi_b}}(\mathcal{Z})$ for some $c \in (1/2, 3/2]$. Specifically, for $g_{\pi,Q} \in L^2_{d^{\pi_b}}(\mathcal{Z})$, we can always represent it as

$$g_{\pi,Q} = \sum_{i=1}^{\infty} e_{i,\pi,Q}\phi_i \quad \text{with} \quad \|\{e_{i,\pi,Q}\}_{i\geq 1}\|_{\ell_2} = \|g_{\pi,Q}\|_{L^2_{d^{\pi_b}}} < +\infty.$$

Then Assumption 7 (b) implies that one can always write $\mathcal{T}^\pi Q$ as

$$\mathcal{T}^\pi Q = \sum_{i=1}^{\infty} e_i^c e_{i,\pi,Q}\phi_j \quad \text{with} \quad \|\{e_{i,\pi,Q}\}_{i\geq 1}\|_{\ell_2} < +\infty,$$

or equivalently, $\mathcal{T}^\pi Q = \sum_{i=1}^{\infty} \widetilde{e}_{i,\pi,Q}\phi_j$ with $\|\{\widetilde{e}_{i,\pi,Q}/e_i^c\}_{i\geq 1}\|_{\ell_2} = \|g_{\pi,Q}\|_{L^2_{d^{\pi_b}}} < +\infty$, which indicates that the larger c is, the smoother the function $\mathcal{T}^\pi Q$ will be. Moreover, define a space $\mathcal{H}_k^{2c}$ as

$$\mathcal{H}_k^{2c} = \left\{ f = \sum_{i=1}^{\infty} \bar{e}_i\phi_i \mid \sum_{i=1}^{\infty} \frac{\bar{e}_i^2}{e_i^{2c}} < +\infty \right\}.$$

Then assumption 7 (b) can also be interpreted as that for every $\pi \in \Pi$ and $Q \in \mathcal{F}$, $\mathcal{T}^\pi Q \in \mathcal{H}_k^{2c}$. In the standard literature of the kernel ridge regression (e.g., Smale and Zhou, 2007; Caponnetto and De Vito, 2007), for fixed Q and π, a similar type of the smoothness assumption is also imposed but only requires $c \in (1/2, 1]$. Here we impose a slightly stronger condition for obtaining a faster rate in terms of $\|\bullet\|_{\mathcal{H}_k}$. In addition, we strengthen this smoothness assumption by considering for all $\pi \in \Pi$ and $Q \in \mathcal{F}$, which is used for controlling the bias uniformly over $\mathcal{F}$ and Π induced by the regularization in the kernel ridge regression (15). See more discussion related to this in Remark 2 of Theorem 2.

Assumptions 7 (a)-(b) also have a direct link with the Bellman completeness assumption used in the RL literature. Bellman completeness assumes that for any $Q \in \mathcal{U}$, $\mathcal{T}^\pi Q \in \mathcal{U}$ for some class of functions $\mathcal{U}$. Hence if we assume $\mathcal{H}_k^{2c} = \mathcal{F}$, then Assumption 7 (b) is equivalent to Bellman completeness. Furthermore, by the contraction property of Bellman operator, Bellman completeness implies the realizability of Q^π for all $\pi \in \Pi$. In this case, Assumption 7 (a) will automatically hold. Since Assumptions 7 (b) has a close relationship with Bellman completeness, to align and have an easy comparison with the existing literature, we make a stronger and alternative assumption in the following that implies Assumptions 7 (a)-(b).

Assumption 8 *For any $\pi \in \Pi$, $Q \in \mathcal{F}$, $\mathcal{T}^\pi Q \in \mathcal{F} \subseteq \mathcal{H}_k^{2c}$ for some $c \in (1/2, 3/2]$.*

Finally, Assumption 7 (c) is used for controlling the metric entropy of the class of vector-valued functions such as $\{\mathcal{L}_k^{-c}(\mathcal{T}^\pi Q) \mid Q \in \mathcal{F}, \pi \in \Pi\}$ by the ϵ-covering numbers of $\mathcal{F}$ and Π defined in Assumption 6.

5.2 Regret Bounds

To derive the regret bounds for $\widehat{\pi}$ and $\widehat{\pi}_{\text{dual}}$, we first show that for every $\pi \in \Pi$, Q^π is a feasible solution to the inner minimization problem of (17) with a high probability. Let

$$\Omega(\pi, \mathcal{F}, \mathcal{W}, \varepsilon_{NT}^{(1)}, \varepsilon_{NT}^{(2)}) \triangleq \Omega_1(\pi, \mathcal{F}, \mathcal{W}, \varepsilon_{NT}^{(1)}) \cap \Omega_2(\pi, \mathcal{F}, \varepsilon_{NT}^{(2)}).$$

Theorem 1 *Suppose Assumptions 1-7 are satisfied, by letting*

$$\varepsilon_{NT}^{(1)} \asymp \log(NT)\sqrt{\left(v(\Pi) + v(\mathcal{F}) + v(\mathcal{W})\right)/NT} \quad \text{and}$$
$$\varepsilon_{NT}^{(2)} \asymp \left(\log(NT)\sqrt{\left(v(\Pi) + v(\mathcal{F})\right)/NT}\right)^{\frac{2c-1}{2c+1}},$$

where c is defined in Assumption 7 (b), we have with probability at least $1 - 1/NT$, for every $\pi \in \Pi$,

$$Q^\pi \in \Omega(\pi, \mathcal{F}, \mathcal{W}, \varepsilon_{NT}^{(1)}, \varepsilon_{NT}^{(2)}). \tag{27}$$

Theorem 1 is the key for establishing the regret bounds of our proposed algorithms. By showing (27), we can guarantee that for every $\pi \in \Pi$, Q^π is bounded below by the optimal value of the inner minimization problem in (17) with a high probability. This verifies the use of pessimism in our algorithms, i.e., evaluating the value of each $\pi \in \Pi$ by a lower bound during the policy optimization procedure. Since we require $c \in (1/2, 3/2]$, the smallest value that $\varepsilon_{NT}^{(2)}$ can be set is of the order $\sqrt{\log(NT)}(NT)^{-1/4}$. Based on Theorem 1, we have the following regret warranty for $\widehat{\pi}$.

Theorem 2 (Regret warranty for $\widehat{\pi}$) *Under Assumptions 1-7, the following statements hold.*

(a) Suppose $\omega^{\pi^*} \in \mathcal{W}$ and $MMD_k(\bar{d}_T^{\pi^b}, \lambda_2^{\pi^*}) \leq \delta$, then with probability at least $1 - \frac{1}{NT}$,

$$\textit{Regret}(\widehat{\pi}) \lesssim \varepsilon_{NT}^{(1)} + \widetilde{\lambda}_2^{\pi^*}(\mathcal{S} \times \mathcal{A})\delta\varepsilon_{NT}^{(2)}, \tag{28}$$

where $\varepsilon_{NT}^{(1)}$ and $\varepsilon_{NT}^{(2)}$ are given in Theorem 1.

(b) Suppose $d_\gamma^{\pi^*} \ll d^{\pi^b}$ and $\omega^{\pi^*} \in \mathcal{W}$, then with probability at least $1 - \frac{1}{NT}$,

$$\textit{Regret}(\widehat{\pi}) \lesssim \varepsilon_{NT}^{(1)}. \tag{29}$$

(c) Suppose $d_\gamma^{\pi^*} \perp d^{\pi^b}$ and $MMD_k(\bar{d}_T^{\pi^b}, \lambda_2^{\pi^*}) \leq \delta$, then with probability at least $1 - \frac{1}{NT}$,

$$\textit{Regret}(\widehat{\pi}) \lesssim \varepsilon_{NT}^{(1)} + \delta\varepsilon_{NT}^{(2)}. \tag{30}$$

Finally, all statements above hold if Assumptions 7 (a)-(b) are replaced by Assumption 8.

Remark 1 *Among all assumptions we have made, (i) the Bellman completeness assumption (Assumptions 7 (b) or 8), (ii) realizability (Assumption 7 (a)), (iii) $\omega^{\pi^*} = \frac{\lambda_1^{\pi^*}}{d^{\pi^b}} \in \mathcal{W}$ with $\|\omega^{\pi^*}\|_\infty < +\infty$ (implicitly imposed by Assumption 6 (b)), and (iv) $MMD_k(\bar{d}_T^{\pi^b}, \lambda_2^{\pi^*}) \leq \delta$, are four main conditions for our regret guarantee stated in Theorem 2. The first two conditions are related to the modeling assumption, while the latter two are related to our batch data coverage induced by the behavior policy.*

In the existing literature, the weakest data coverage assumption for establishing the regret guarantee is $d_\gamma^{\pi^*} \ll d^{\pi^b}$ with $\|d_\gamma^{\pi^*}/d^{\pi^b}\|_\infty < +\infty$, i.e., absolute continuity with a uniformly bounded Radon–Nikodym derivative. See for example Xie et al. (2021) and Zhan et al. (2022). In contrast, our proposed method remains valid without requiring the absolute *continuity. For example, in Case (a) of Theorem 2, to achieve the regret guarantee of our proposed method, we only require the ratio of the absolutely continuous part of $d_\gamma^{\pi^*}$ with respect to d^{π^b} over d^{π^b}, i.e., $\frac{\lambda_1^{\pi^*}}{d^{\pi^b}}$, is uniformly bounded above, and the MMD between the singular part of $d_\gamma^{\pi^*}$ with respect to d^{π^b} and d^{π^b}, i.e., $MMD_k(\bar{d}_T^{\pi^b}, \lambda_2^{\pi^*})$, is bounded by a constant δ. Moreover, due to Assumption 5, by choosing $\delta = 2\kappa$, $MMD_k(\bar{d}_T^{\pi^b}, \lambda_2^{\pi^*}) \leq \delta$ is always satisfied. Therefore our method requires much weaker assumptions (essentially no assumptions) on the batch data coverage induced by the behavior policy, and is thus more applicable than existing algorithms from the perspective of the data coverage. However, our method requires a slightly stronger condition on modeling Q^π, which could be regarded as a trade-off. As discussed after Assumption 7, we require an additional Bellman completeness for modeling Q^π (i.e., Assumption 7 (b) or Assumption 8), while some existing pessimistic algorithms (e.g., Jiang and Huang, 2020) only require the realizability condition on modeling Q^π (i.e., Assumption 7 (a)).*

It is worth mentioning that, when $d_\gamma^{\pi^*} \ll d^{\pi^b}$ as discussed in the Case (b) of Theorem 2, ω^{π^*} becomes $d_\gamma^{\pi^*}/d^{\pi^b}$, with $\widetilde{\lambda}_1^{\pi^*}(\mathcal{S} \times \mathcal{A}) = 1$ and $\widetilde{\lambda}_2^{\pi^*}(\mathcal{S} \times \mathcal{A}) = 0$. In this case, we recover the existing results such as Jiang and Huang (2020) on the regret bound by running our algorithm. However, compared with their results, besides $\omega^{\pi^*} \in \mathcal{W}$ and realizability of all Q^π for $\pi \in \Pi$, we require one additional condition, i.e., Assumption 7 (b) for modeling Q^π. The main reason for requiring a stronger condition here is due to the unknown information that $d_\gamma^{\pi^*} \ll d^{\pi^b}$. When $d_\gamma^{\pi^*} \perp d^{\pi^b}$ as discussed in Case (c) of Theorem 2, it can be seen that $\widetilde{\lambda}_1^{\pi^*}(\mathcal{S} \times \mathcal{A}) = 0$ and $\widetilde{\lambda}_2^{\pi^*}(\mathcal{S} \times \mathcal{A}) = 1$. We showcase that the regret of our estimated policy can still converge to 0 in a satisfactory rate without any data coverage assumption. To achieve this, we rely on the Bellman completenss for enabling the extrapolation property. To the best of our knowledge, this is the first regret guarantee in batch RL without assuming the absolute continuity. See Table 1 for a summary of our main assumptions for obtaining the regret of the proposed algorithm compared with existing ones. We remark that for easy comparison, we use a stronger assumption (i.e., Assumption 8) to replace our two modeling assumptions (i.e., Assumptions 7 (a)-(b)) in Table 1.

Remark 2 *Our proposed algorithm is motivated by Lemma 2, which provides a careful*

Table 1: Comparison of the main assumptions behind our algorithm and several state-of-the-art methods. The operator $\mathcal{T}^*$ in the approximate value iteration method is referred to as the optimal Bellman operator (Sutton et al., 1998). Pessimistic RL refers to some batch RL algorithms using pessimism. V^π is the state-value function of a policy π and $\bar{\mathcal{V}}$ is the corresponding hypothesized class of functions.

Algorithm	Data Coverage	Modeling Assumption
Approximate Value Iteration	$\\|\frac{d_\gamma^\pi}{\bar{d}_T^{\pi^b}}\\|_\infty < +\infty, \forall\pi$	$\forall f \in \mathcal{F}, \mathcal{T}^* f \in \mathcal{F}$ (Munos and Szepesvári, 2008)
Approximate Policy Iteration		$\forall f \in \mathcal{F}$ and $\pi \in \Pi$, $\mathcal{T}^\pi f \in \mathcal{F}$ (Antos et al., 2008*c*)
Pessimistic RL	$\\|\frac{d_\gamma^{\pi^*}}{\bar{d}_T^{\pi^b}}\\|_\infty < +\infty$	$\forall f \in \mathcal{F}$ and $\pi \in \Pi$, $\mathcal{T}^\pi f \in \mathcal{F}$ (Xie et al., 2021)
		$\frac{d_\gamma^{\pi^*}}{\bar{d}_T^{\pi^b}} \in \mathcal{W}$ and $\forall\pi \in \Pi$, $Q^\pi \in \mathcal{F}$ (Jiang and Huang, 2020)
PRO-RL (Zhan et al., 2022) without regularization	$\frac{d_\gamma^\pi(s)}{\bar{d}_T^{\pi^b}(s)} \lesssim 1, \forall\pi, s \in \mathcal{S}$ $\frac{d_\gamma^{\pi^*}(s)}{\bar{d}_T^{\pi^b}(s)} \gtrsim 1, \forall s \in \mathcal{S}$	$\frac{d_\gamma^{\pi^*}}{\bar{d}_T^{\pi^b}} \in \mathcal{W}$ and $V^{\pi^*} \in \bar{\mathcal{V}}$
STEEL (general)	$d_\gamma^{\pi^*} = \widetilde{\lambda}_1^{\pi^*}(\mathcal{S}\times\mathcal{A})\lambda_1^{\pi^*} + \widetilde{\lambda}_2^{\pi^*}(\mathcal{S}\times\mathcal{A})\lambda_2^{\pi^*}$ $\\|\frac{\lambda_1^{\pi^*}}{\bar{d}_T^{\pi^b}}\\|_\infty < +\infty$ $\mathrm{MMD}_k(\bar{d}_T^{\pi^b}, \lambda_2^{\pi^*}) \leq \delta$	$\frac{\lambda_1^{\pi^*}}{\bar{d}_T^{\pi^b}} \in \mathcal{W}; \forall\pi \in \Pi$, $Q \in \mathcal{F}$, $\mathcal{T}^\pi Q \in \mathcal{F} \subseteq \mathcal{H}_k^{2c}$
STEEL ($d_\gamma^\pi \ll \bar{d}_T^{\pi^b}$)	$\\|\frac{d_\gamma^{\pi^*}}{\bar{d}_T^{\pi^b}}\\|_\infty < +\infty$	$\frac{\lambda_1^{\pi^*}}{\bar{d}_T^{\pi^b}} \in \mathcal{W}; \forall\pi \in \Pi$, $Q \in \mathcal{F}$, $\mathcal{T}^\pi Q \in \mathcal{F} \subseteq \mathcal{H}_k^{2c}$
STEEL ($d_\gamma^\pi \perp \bar{d}_T^{\pi^b}$)	$\mathrm{MMD}_k(\bar{d}_T^{\pi^b}, \lambda_2^{\pi^*}) \leq \delta$	$\frac{\lambda_1^{\pi^*}}{\bar{d}_T^{\pi^b}} \in \mathcal{W}; \forall\pi \in \Pi$, $Q \in \mathcal{F}$, $\mathcal{T}^\pi Q \in \mathcal{F} \subseteq \mathcal{H}_k^{2c}$

decomposition on the estimation error of OPE without relying on the absolute continuity. The key reason for the success of our algorithm in handling insufficient data coverage (i.e., the singular part) relies on the smoothness condition we imposed on modeling Q^π*, i.e., Assumption 7* (b)*, which provides a strong extrapolation property for our estimated* $\widehat{\mathcal{T}}^\pi Q$ *for* $Q \in \mathcal{F}$ *and* $\pi \in \Pi$*. Specifically, for any state-action pair* $(s, a) \in F$*, where* F *is a measurable set such that* $\bar{d}_T^{\pi^b}(F) = 0$ *but* $\bar{d}_T^{\pi^b}(F^c) = 1$*, by leveraging the convolution operator* $\mathcal{L}_k$*, the corresponding* $\mathcal{T}^\pi Q(s, a)$ *can be extrapolated by the kernel smoothing of* $\mathcal{T}^\pi Q(\widetilde{s}, \widetilde{a})$ *for all* $(\widetilde{s}, \widetilde{a}) \in F^c$*. Therefore,* $\widehat{\mathcal{T}}^\pi Q$ *can approximate* $\mathcal{T}^\pi Q$ *well in terms of the RKHS norm. Such an extrapolation property is inherited by Q-function due to the Bellman equation that* $\mathcal{T}^\pi Q^\pi = Q^\pi$*.*

Remark 3 *In Case (b) of Theorem 2, where* $d_\gamma^{\pi^*} \ll d^{\pi^b}$*, to achieve the desirable regret guarantee, by implementing our algorithm without using the constraint set* $\Omega_2(\pi, \mathcal{F}, \varepsilon_{NT}^{(2)})$*, we indeed do not require our batch data to uniquely identify* Q^π *or* $\mathcal{V}(\pi)$ *for* $\pi \neq \pi^*$ *but only* π^**. Specifically, for a fixed policy* $\pi \neq \pi^*$*, we allow that there exist two different* Q_1^π

and Q_2^π under the batch data distribution such that

$$\begin{aligned}
&\sup_{w\in\mathcal{W}} \overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma Q_1^\pi(S',\pi(S')) - Q_1^\pi(S,A)\right)\right] \\
=&\sup_{w\in\mathcal{W}} \overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma Q_2^\pi(S',\pi(S')) - Q_2^\pi(S,A)\right)\right] = 0.
\end{aligned}$$

In addition, because we do not require ω^π for $\pi \neq \pi^*$ modeled correctly or being uniformly bounded above, the above equality cannot ensure that

$$\mathcal{V}(\pi) = (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q_1^\pi(S_0,\pi(S_0))\right] = (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q_2^\pi(S_0,\pi(S_0))\right].$$

Thus $\mathcal{V}(\pi)$ for $\pi \neq \pi^*$ are not required to be identified by our batch data either if our proposed algorithm is implemented without incorporating $\Omega_2(\pi,\mathcal{F},\varepsilon_{NT}^{(2)})$. However, due to the unknown information on π^* (e.g., whether Case (b) happens or not), our algorithm has to include the constraint set $\Omega_2(\pi,\mathcal{F},\varepsilon_{NT}^{(2)})$ for handling the error induced by the possibly singular part, which implicitly imposes that Q^π and $\mathcal{V}(\pi)$ can be uniquely identified by our batch data for $\pi\in\Pi$ because of the strong RKHS norm used in Ω_2. This is another trade-off for addressing the issue of a possible singularity. Lastly, we would like to emphasize that two conditions that $\omega^{\pi^*}\in\mathcal{W}$ and $MMD_k(\bar{d}_T^{\pi^b},\lambda_2^{\pi^*})\leq\delta$ are imposed only on the in-class optimal policy π^* but not on all $\pi\in\Pi$.

Next, we develop a finite-sample regret bound for $\widehat{\pi}_{\text{dual}}$. Consider the following constrained optimization problem, which corresponds to the population counterpart of (28) with fixed $\varepsilon_{NT}^{(1)}$ and $\varepsilon_{NT}^{(2)}$

$$\min_{Q\in\widetilde{\Omega}(\pi,\mathcal{F},\varepsilon_{NT}^{(1)},\varepsilon_{NT}^{(2)})} \quad (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi(S_0))\right],$$

where

$$\begin{aligned}
&\widetilde{\Omega}(\pi,\mathcal{F},\varepsilon_{NT}^{(1)},\varepsilon_{NT}^{(2)}) \\
=&\left\{Q\in\mathcal{F} \mid \sup_{w\in\mathcal{W}} \overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma Q(S',\pi(S')) - Q(S,A)\right)\right] \leq 2\varepsilon_{NT}^{(1)}\right\} \\
\cap&\left\{Q\in\mathcal{F} \mid \|\mathcal{T}^\pi Q\|_{\mathcal{H}_k} \leq 2\varepsilon_{NT}^{(2)}\right\}.
\end{aligned}$$

Define the corresponding Lagrangian function as

$$\begin{aligned} L(Q,\rho,\pi) = {} & (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi(S_0))\right] \\ & + \rho_1 \times \left\{ \sup_{w\in\mathcal{W}} \overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma Q(S',\pi(S')) - Q(S,A)\right)\right] - 2\varepsilon_{NT}^{(1)} \right\} \\ & + \rho_2 \times \left\{ \|\mathcal{T}^{\pi} Q\|_{\mathcal{H}_k} - 2\varepsilon_{NT}^{(2)} \right\}, \end{aligned} \tag{31}$$

which could be roughly viewed as the population counterpart of $L_{NT}(Q,\rho,\pi)$. We ignore the dependency of $L(Q,\rho,\pi)$ on NT for notation simplicity. To proceed, we need one additional assumption.

Assumption 9 *$\mathcal{F}$ is a convex set.*

Assumption 9 is imposed so that together with Assumptions 1, 2 and 6 (a), the following strong duality holds so that

$$\min_{Q\in\widetilde{\Omega}(\pi,\mathcal{F},\varepsilon_{NT})} (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi^*(S_0))\right] = \max_{\rho\succeq 0}\min_{Q\in\mathcal{F}} L(Q,\rho,\pi).$$

This implies that there is no duality gap between $\max_{\rho\succeq 0}\min_{Q\in\mathcal{F}} L_{NT}(Q,\rho,\pi^*)$ and

$$\min_{Q\in\Omega_1(\pi,\mathcal{F},\mathcal{W},\varepsilon_{NT}^{(1)})\cap\Omega_2(\pi,\mathcal{F},\varepsilon_{NT}^{(2)})} (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi(S_0))\right]$$

asymptotically, which is essential for us to derive a finite-sample bound for the regret of $\widehat{\pi}_{\text{dual}}$ given below.

Theorem 3 *Under Assumptions 1-7, and 9, all statements in Theorem 2 hold for $\widehat{\pi}_{dual}$ as well.*

Remark 4 *The implications behind Theorem 3 are the same as those of Theorem 2. The benefit of considering $\widehat{\pi}_{dual}$ mainly comes from the computational efficiency. However, assuming $\mathcal{F}$ convex could be restrictive. Therefore it will be interesting to study the behavior of duality gap when $\mathcal{F}$ is modeled by a non-convex but "asymptotically" convex set. This will allow the use of some deep neural network architectures such as Zhang et al. (2019) in practice. We leave it for future work.*

6 Numerical Studies

6.1 Implementation Details

In this subsection, we present our computational algorithm for obtaining $\widehat{\pi}_{\text{dual}}$. We set $\mathcal{W} = \{w \mid \|w\|_{\mathcal{H}_k} \leq C\}$ for some constant $C > 0$. We remark that a different kernel can be chosen in $\mathcal{W}$. By the representer property, we can obtain a closed-form optimal value for

$$\max_{w \in \mathcal{W}} \overline{\mathbb{E}}_{NT}\left[w(S,A)\left(R + \gamma Q(S', \pi(S')) - Q(S,A)\right)\right],$$

which is given as

$$\frac{C}{NT}\sqrt{Y(\pi,Q)^\top K Y(\pi,Q)}.$$

Therefore, in order to compute $\widehat{\pi}_{\text{dual}}$, it is sufficient to solve

$$\begin{aligned}\max_{\pi \in \Pi, \rho \succeq 0} \min_{Q \in \mathcal{F}} \Big\{ &(1-\gamma)\,\mathbb{E}_{S_0 \sim \nu}\left[Q(S_0, \pi(S_0))\right] \\ &+\rho_1 \times \left\{\frac{1}{(NT)^2} Y(\pi,Q)^\top K Y(\pi,Q) - (\varepsilon_{NT}^{(1)})^2\right\} \\ &+\rho_2 \times \left\{Y(\pi,Q)^\top (K + \zeta_{NT} I_{NT})^{-1} K (K + \zeta_{NT} I_{NT})^{-1} Y(\pi,Q) - (\varepsilon_{NT}^{(2)})^2\right\}\Big\}.\end{aligned} \tag{32}$$

Lastly, if letting $\mathcal{F}$ and Π be any pre-specified functional classes such as neural network architectures, we can apply stochastic gradient decent to solve the primal-dual problem (32) and obtain $\widehat{\pi}_{\text{dual}}$. For the contextual bandit problem discussed in Section 3, we only need to let $\gamma = 0$ in (32) and input the data $\mathcal{D}_N^0$.

6.2 Simulation

In this section, we use Monte Carlo (MC) experiments to systematically study the proposed method in terms of the convergence of our algorithm, the effect of pessimism in finding the optimal policy, and the robustness to the covariate shift. For simplicity, we only consider the contextual bandit problem.

For all numerical studies, we compare our method with two popular continuous-action policy optimization methods. The regression-based approach first estimates the reward function $r(s,a)$ as $\hat{r}(s,a)$ by running a regression method, and then estimates the value of any given policy π with the plug-in estimator $\mathbb{E}_{S_0 \sim \nu}\left[\hat{r}(S_0, \pi(S_0))\right]$, where the expectation is approximated via MC sampling of ν. Recall that ν is known in advance. Given such a policy value estimator, we implement the off-the-shelf optimization algorithm to estimate the optimal policy within a given policy class. The kernel-based approach (Chen et al., 2016; Kallus and Zhou, 2018) is similar to the regression approach but first estimates the policy value using the inverse probability weighting with kernel smoothing. Specifically, we estimate the value of any given policy π by

$$(Nh)^{-1} \sum_{i=1}^{N} \frac{K((\pi(S_{i,0}) - A_{i,0})/h)}{\pi^b(A_{i,0} \mid S_{i,0})} \times R_{i,0},$$

where $K(\cdot)$ is a kernel function, h is the bandwidth parameter, and π^b is referred as to the generalized propensity score in this setting.

Next, we describe the data generating process in our simulation study. We consider the mean reward function $r(s,a) = \left(a - \boldsymbol{B}s\right)^\top \boldsymbol{C} \left(a - \boldsymbol{B}s\right)$, where the state-value s has dimension $d_s = 5$, the action-value a has dimension $d_a = 4$, the parameter matrix $\boldsymbol{B} \in \mathcal{R}^{d_a \times d_s}$, and $\boldsymbol{C}$ is a $d_a \times d_a$ negative definite matrix. We randomly sample every entry of $\boldsymbol{B}$ from Uniform$(0, 1)$. We construct $\boldsymbol{C} = -\boldsymbol{C}_0^\top \boldsymbol{C}_0$, where every entry of $\boldsymbol{C}_0$ is sampled from the standard normal. Our reward R_0 is generated following $R_0 = r(S_0, A_0) + \epsilon_0$, where the independent noise $\epsilon_0 \sim \mathcal{N}(0, 1)$. Therefore, the optimal policy $\pi^*(s) = s^\top \boldsymbol{B}$ for every $s \in \mathcal{S}$, regardless of the state distribution.

The training dataset $\{S_{i,0}, A_{i,0}, R_{i,0}\}_{1 \le i \le N}$ is generated as follows. The state $S_{i,0}$ is uniformly sampled from $[0, 2]^{d_s}$, which is the same as the reference distribution. Except for the case where we study the effect of covariate shift, values of learned policies in this section are evaluated over the same distribution as the training distribution. The action is sampled following a behavior policy π^b such that $A_{i,0} = \boldsymbol{B} S_{i,0} + \epsilon'$, where $\epsilon' \sim \mathcal{N}(0, \sigma_b^2 \boldsymbol{I}_{d_a})$. Therefore, a larger σ_b indicates that the behavior policy is more different from the optimal one and also implies that the action space is explored more in the training data.

For all three methods, we search the optimal policy within the class of linear deterministic policies $\Pi = \{\pi \mid \text{for every } s \in \mathcal{S}, \pi(s) = \tilde{\boldsymbol{B}} s \text{ for some } \tilde{\boldsymbol{B}} \in [-1, 1]^{d_s \times d_a}\}$. For the kernel-based method, we assume the true value of the generalized propensity score π^b is *known*, instead of plugging-in the estimated value. Although the latter is known to be

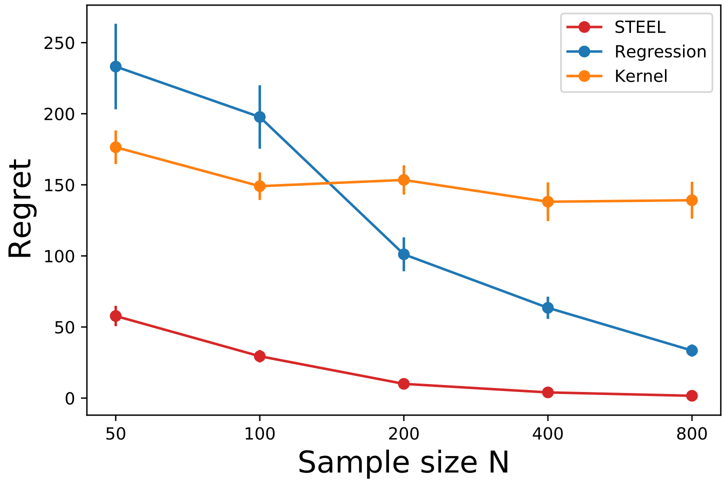


Figure 1: Convergence of different methods. The error bars indicate the standard errors of the averages. The x-axis is plotted on the logarithmic scale.

asymptotically more efficient in the causal inference literature, we observe that its empirical performance is close or sometimes worse than the oracle version in the studied setting, given the challenge to estimating a multi-dimensional conditional density function (See Appendix D). We tune the bandwidth for the kernel-based method and use the best one for it. To implement the optimization problems in two baseline methods, we adopt the widely used L-BFGS-B optimization algorithm following Liu and Nocedal (1989). Regarding estimating the reward function $r(s,a)$, we use a neural network of three hidden layers with 32 units for each layer and ReLU as the activation function. For **STEEL**, we use the Laplacian kernel $k(x,y) = \exp\big(-||x-y||_1/\gamma\big)$ in both the kernel ridge regression and the modeling of ω, where the bandwidth γ is picked based on the popular median heuristic, i.e., set as the median of $\{||(S_{i,0}, A_{i,0}) - (S_{j,0}, A_{j,0})||_1\}_{1 \leq i,j \leq N}$. For the other tuning parameters, we set $\zeta_{NT} = 0.001$, $(\varepsilon_{NT}^{(1)})^2 = 300$, and $(\varepsilon_{NT}^{(2)})^2 = 600$.

For each setting below, we run 100 repetitions and report the average regret of each method as well as its standard error. Except for when we study the convergence of three methods in terms of the sample size, we fix $N = 200$ data points for each repetition.

Convergence. We first study the convergence of all three methods, by increasing the training sample size N. The results when $\sigma_b = 0.5$ are presented in Figure 1. Overall, we observe that our method yields a desired convergence with the regret decaying to zero very

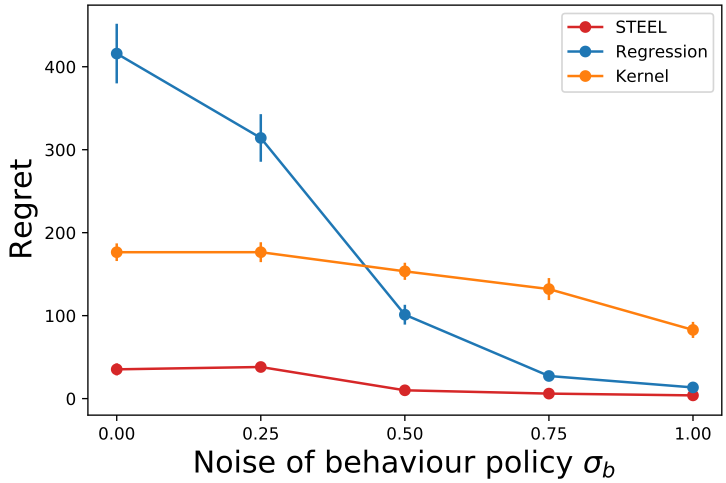


Figure 2: Effect of pessimism. The error bars indicate the standard errors of the averages.

quickly. The regression-based method also improves as the sample size increases, despite that the finite-sample performance is worse compare with ours. The kernel-based method suffers from the curse of dimensionality of the action space and the regret does not decay very significantly. Results under a few other settings are reported in the appendix, with similar findings.

Effect of pessimism. Next, recall that σ_b controls the degree of exploration in the training data. It is known that a well-explored action space in the training dataset is beneficial for policy learning, in the sense that we can accurately estimate the values of most candidate policies. In contrast, when σ_b is small (i.e., the behavior policy is close to the optimal one), although the value estimation of the optimal policy may be accurate, that of a sub-optimal policy suffers from a large uncertainty and by chance the resulting estimator could be very large, which leads to returning a sub-optimal policy. Our method is designed with addressing such a large uncertainty in mind, by utilizing the pessimistic mechanism.

In Figure 2, we empirically study the effect of σ_b on different methods. The results are based on $N = 200$ data points and different values of σ_b. It can be seen clearly that, our method is fairly robust, when the training data is either well explored or not. In contrast, the performance of the two baseline methods deteriorates significantly when σ_b is small.

Robustness to covariate shift. Recall that the singularity may be caused by

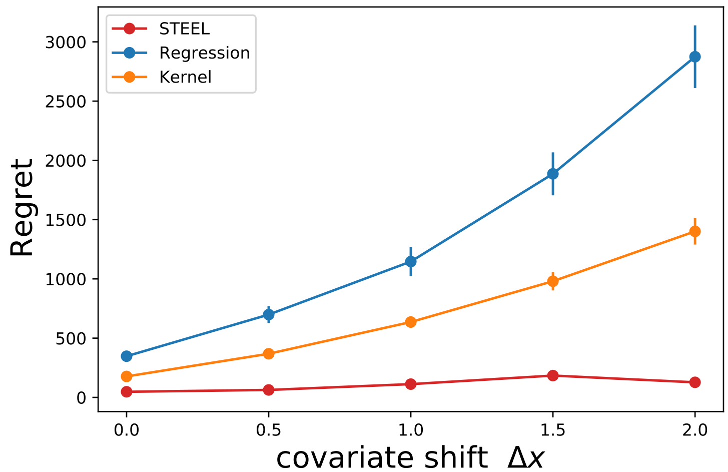


Figure 3: Robustness to covariate shift. The error bars indicate the standard errors of the averages.

the covariate shift as well. Motivated by this, we also study the robustness of different methods to such a shift. Specifically, the policy will be tested when the covaraites are sampled uniformly from $[\Delta x, 2 + \Delta x]^{d_s}$. A larger value of Δx hence indicates a larger degree of covariate shift.

The results when $\sigma_b = 0.25$ and $N = 200$ are presented in Figure 3. It can be seen clearly that our method shows a desired robustness. This is by design, as it accounts for such a shift explicitly. In contrast, the regret of the other two methods increase significantly when the testing state distribution of S_0 becomes more different from that in the training data.

6.3 Real Data

In this section, we apply our method to personalized pricing with a real dataset. The dataset is from an online auto loan lending company in the United States [1]. It was first studied by Phillips et al. (2015) and we follow similar setups as in subsequent studies (Ban and Keskin, 2021).

The dataset contains all auto loan records (208085 in total) in a major online auto loan lending company from July 2002 to November 2004. For each record, it contains

[1]The dataset is available at `https://business.columbia.edu/cprm` upon request.

some covariates (e.g., FIFO credit score, the term and amount of loan requested, etc.), the monthly payment offered by the company, and the decision of the applicant (accept the offer or not). We refer interested readers to Phillips et al. (2015) for more details of the dataset.

The monthly payment (together with the loan term) can be regarded as the offered price (*action*) and the binary decision of the applicant reflects the corresponding demand. Specifically, we follow Ban and Keskin (2021) and Bastani et al. (2022) to define the price as

$$\text{price} = \text{Monthly payment} \times \sum_{t=1}^{\text{Term}} (1 + \text{Prime rate})^{-t} - \text{amount of loan requested},$$

which considers the net present value of future payments. Here, the prime rate is the interest rate that the company itself needs to pay. We filter the dataset to exclude the outlier records (defined as having a price higher than \$10,000), and 99.49% data points remain in the final dataset.

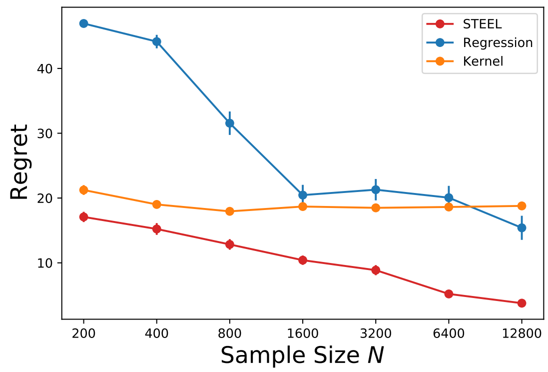
In our notations, the price is hence the action A_0. Let the binary decision of the applicant be D_0. The reward is then naturally computed as $R_0 = D_0 \times A_0$. We use the same set of features which are identified as significant in Ban and Keskin (2021) to construct our feature vector S_0: it contains the FICO credit score, the loan amount approved, the prime rate, the competitor's rate, and the loan term.

When running offline experiments for problems with *discrete* action spaces, to compared different learned policies, the standard procedure is to find those data points where the recorded actions are consistent with the recommendations from a policy π that we would like to evaluate. However, with a *continuous* action space, such a procedure is typically impossible, as it is difficult for two continuous actions to have exactly the same value. Therefore, it is well acknowledged that, offline experiments without any model assumption is infeasible. We closely follow Bastani et al. (2022) to design a semi-real experiment, where we fit a linear demand function from the real dataset and use it to evaluate the policies learned by different methods. The linear regression uses the concatenation of S_0 (which describes the baseline effect) and $S_0 A_0$ (which describes the interaction effect) as the covariate vector. To evaluate the statistical performance of the three methods considered in Section 6.2, we repeat for 100 random seeds, and for each random seed we sample N data points for policy optimization.

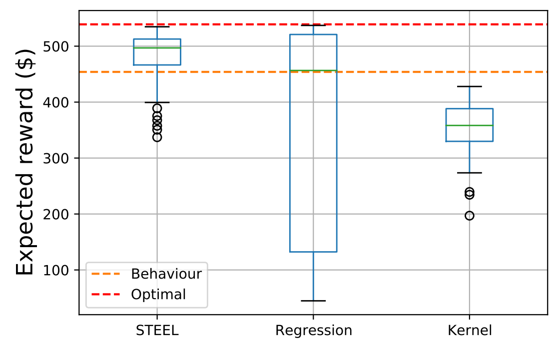
The implementation and tuning details of the three methods are almost the same with Section 6.2. The main modification is that, for the reward function class used in our method and the regression-based method, we use the neural network to model the demand function (instead of the expected reward itself), which multiplies with the action (price) gives the expected reward. Such a modification utilizes the problem-specific structure. All covariates are normalized and we always add an intercept term. For tuning parameters, we set $\zeta_{NT} = 0.001$, $(\varepsilon_{NT}^{(1)})^2 = 13$, and $(\varepsilon_{NT}^{(2)})^2 = 16$.

In Figure 4a, we increase N from 200 to 12800 and report the average regrets. We find that our method consistently outperforms the other two methods and its regret decays to zero. In Figure 4b, we zoom into the case where $N = 8000$, and report the values of the learned policies across the 100 random seeds. We observe that, although the regression-based method can outperform the behaviour policy in about half of the random replications, it can perform very poor in the remaining replicates. In contrast, our method has an impressive performance and the robustness of our methods is consistent with our methodology design. The kernel-based method does not perform well in most settings. A deep dive shows that this is because this method over-estimates the values of some sub-optimal policies that assign many actions near the boundaries where we have few data points. The over-estimation is due to its poor extrapolation ability and the lack of the pessimism mechanism (i.e., it is not aware of the uncertainty).

Finally, we randomly pick a policy learned by our method (with the first random seed we used) and report its coefficient to provide some insights. Recall that we use a linear policy, mapping from the five features to the recommended price. The coefficients are overall aligned with the intuition: the learned personalized pricing policy offers a higher price when the FICO score is lower, the loan amount approved is higher, the prime rate (the interest rate that the company itself faces) is higher, the competitor's rate is higher (implying a consensus on the high risk of this loan), and the term is longer.



(a) Regret trend with sample size. The error bars indicate the standard errors of the averages.



(b) Value distribution over 100 seeds when $N = 8000$.

Figure 4: Performance in personalized pricing with real data.

Table 2: Coefficients of the learned linear personalized pricing policy.

FICO score	Loan amount approved	Prime rate	Competitor's rate	Term
-0.135	0.041	0.180	0.027	0.514

7 Conclusion

In this paper, we study batch RL in the presence of singularities and propose a new policy learning algorithm to tackle this issue. Our proposed method finds an optimal policy without requiring absolute continuity on the distribution of target policies with respect to the data distribution. Both theoretical guarantees and numerical evidence are presented to illustrate the superior performance of our proposed method, compared with existing works. In the current proposal, we leverage the MMD, together with distributionally robust optimization, to enable the extrapolation. Future research could involve studying the algorithm under the Wasserstein distance and exploring model-based solutions for addressing the singularity issue in batch RL.

References

Adjaho, C. and Christensen, T. (2022), 'Externally valid treatment choice', *arXiv preprint arXiv:2205.05561* .

Antos, A., Szepesvári, C. and Munos, R. (2008*a*), Fitted q-iteration in continuous action-space mdps, *in* 'Advances in neural information processing systems', pp. 9–16.

Antos, A., Szepesvári, C. and Munos, R. (2008*b*), 'Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path', *Machine Learning* **71**(1), 89–129.
URL: *http://link.springer.com/10.1007/s10994-007-5038-2*

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

Athey, S. and Wager, S. (2021), 'Policy learning with observational data', *Econometrica* **89**(1), 133–161.

Bai, C., Wang, L., Yang, Z., Deng, Z., Garg, A., Liu, P. and Wang, Z. (2022), 'Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning', *arXiv preprint arXiv:2202.11566* .

Ban, G.-Y. and Keskin, N. B. (2021), 'Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity', *Management Science* **67**(9), 5549–5568.

Bastani, H., Simchi-Levi, D. and Zhu, R. (2022), 'Meta dynamic pricing: Transfer learning across experiments', *Management Science* **68**(3), 1865–1881.

Berbee, H. C. (1979), 'Random walks with stationary increments and renewal theory', *MC Tracts* .

Bhattacharya, D. and Dupas, P. (2012), 'Inferring welfare maximizing treatment assignment under budget constraints', *Journal of Econometrics* **167**(1), 168–196.

Bradley, R. C. (2005), 'Basic properties of strong mixing conditions. a survey and some open questions', *Probability Surveys* **2**, 107–144.

Caponnetto, A. and De Vito, E. (2007), 'Optimal rates for the regularized least-squares algorithm', *Foundations of Computational Mathematics* **7**(3), 331–368.

Chen, G., Zeng, D. and Kosorok, M. R. (2016), 'Personalized dose finding using outcome weighted learning', *Journal of the American Statistical Association* **111**(516), 1509–1521.

Chen, X. and Qi, Z. (2022), 'On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation', *arXiv preprint arXiv:2201.06169* .

Chernozhukov, V., Demirer, M., Lewis, G. and Syrgkanis, V. (2019), 'Semi-parametric efficient policy learning with continuous actions', *Advances in Neural Information Processing Systems* **32**.

Dedecker, J. and Louhichi, S. (2002), Maximal inequalities and empirical central limit theorems, *in* 'Empirical process techniques for dependent data', Springer, pp. 137–159.

Ernst, D., Geurts, P. and Wehenkel, L. (2005), 'Tree-based batch mode reinforcement learning', *Journal of Machine Learning Research* **6**, 503–556.

Fu, Z., Qi, Z., Wang, Z., Yang, Z., Xu, Y. and Kosorok, M. R. (2022), 'Offline reinforcement learning with instrumental variables in confounded markov decision processes'. Working Paper, Northwestern University, Evanston, IL.
URL: *https://arxiv.org/abs/2209.08666*

Fujimoto, S., Meger, D. and Precup, D. (2019), Off-policy deep reinforcement learning without exploration, *in* 'International conference on machine learning', PMLR, pp. 2052–2062.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012), 'A kernel two-sample test', *The Journal of Machine Learning Research* **13**(1), 723–773.

Jiang, N. and Huang, J. (2020), 'Minimax value interval for off-policy evaluation and policy optimization', *Advances in Neural Information Processing Systems* **33**, 2747–2758.

Jin, Y., Yang, Z. and Wang, Z. (2021), Is pessimism provably efficient for offline rl?, *in* 'International Conference on Machine Learning', PMLR, pp. 5084–5096.

Kallus, N. and Zhou, A. (2018), Policy evaluation and optimization with continuous treatments, *in* 'International conference on artificial intelligence and statistics', PMLR, pp. 1243–1251.

Kitagawa, T. and Tetenov, A. (2018), 'Who should be treated? empirical welfare maximization methods for treatment choice', *Econometrica* **86**(2), 591–616.

Kitagawa, T., Wang, W. and Xu, M. (2022), 'Policy choice in time series by empirical welfare maximization', *arXiv preprint arXiv:2205.03970* .

Kosorok, M. R. and Laber, E. B. (2019), 'Precision medicine', *Annual review of statistics and its application* **6**, 263–286.

Kumar, A., Fu, J., Soh, M., Tucker, G. and Levine, S. (2019), Stabilizing off-policy q-learning via bootstrapping error reduction, *in* 'Advances in Neural Information Processing Systems', pp. 11784–11794.

Kumar, A., Zhou, A., Tucker, G. and Levine, S. (2020), 'Conservative q-learning for offline reinforcement learning', *arXiv preprint arXiv:2006.04779* .

Levine, S., Kumar, A., Tucker, G. and Fu, J. (2020), 'Offline reinforcement learning: Tutorial, review, and perspectives on open problems', *arXiv preprint arXiv:2005.01643* .

Liao, P., Qi, Z., Wan, R., Klasnja, P. and Murphy, S. A. (2022), 'Batch policy learning in average reward markov decision processes', *The Annals of Statistics* **50**(6), 3364–3387.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D. and Wierstra, D. (2015), 'Continuous control with deep reinforcement learning', *arXiv preprint arXiv:1509.02971* .

Liu, D. C. and Nocedal, J. (1989), 'On the limited memory bfgs method for large scale optimization', *Mathematical programming* **45**(1-3), 503–528.

Liu, Y., Swaminathan, A., Agarwal, A. and Brunskill, E. (2020), 'Provably good batch off-policy reinforcement learning without great exploration', *Advances in neural information processing systems* **33**, 1264–1274.

Luenberger, D. G. (1997), *Optimization by vector space methods*, John Wiley & Sons.

Manski, C. F. (2004), 'Statistical treatment rules for heterogeneous populations', *Econometrica* **72**(4), 1221–1246.

Mbakop, E. and Tabord-Meehan, M. (2021), 'Model selection for treatment choice: Penalized welfare maximization', *Econometrica* **89**(2), 825–848.

Mo, W., Qi, Z. and Liu, Y. (2021), 'Learning optimal distributionally robust individualized treatment rules', *Journal of the American Statistical Association* **116**(534), 659–674.

Munos, R. (2003), Error bounds for approximate policy iteration, *in* 'ICML', Vol. 3, Citeseer, pp. 560–567.

Munos, R. and Szepesvári, C. (2008), 'Finite-time bounds for fitted value iteration', *Journal of Machine Learning Research* **9**(May), 815–857.

Park, J. and Muandet, K. (2022), 'Towards empirical process theory for vector-valued functions: Metric entropy of smooth function classes', *arXiv preprint arXiv:2202.04415* .

Phillips, R., Şimşek, A. S. and Van Ryzin, G. (2015), 'The effectiveness of field price discretion: Empirical evidence from auto lending', *Management Science* **61**(8), 1741–1759.

Pinto, L. and Gupta, A. (2016), Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours, *in* '2016 IEEE international conference on robotics and automation (ICRA)', IEEE, pp. 3406–3413.

Puterman, M. L. (1994), *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, Inc.

Qi, Z., Pang, J.-S. and Liu, Y. (2022), 'On robustness of individualized decision rules', *Journal of the American Statistical Association* pp. 1–15.

Qian, M. and Murphy, S. A. (2011), 'Performance guarantees for individualized treatment rules', *Annals of statistics* **39**(2), 1180.

Rahimian, H. and Mehrotra, S. (2019), 'Distributionally robust optimization: A review', *arXiv preprint arXiv:1908.05659* .

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J. and Russell, S. (2021), 'Bridging offline reinforcement learning and imitation learning: A tale of pessimism', *Advances in Neural Information Processing Systems* **34**, 11702–11716.

Rivasplata, O., Parrado-Hernández, E., Shawe-Taylor, J. S., Sun, S. and Szepesvári, C. (2018), 'Pac-bayes bounds for stable algorithms with instance-dependent priors', *Advances in Neural Information Processing Systems* **31**.

Schmidt-Hieber, J. (2020), 'Nonparametric regression using deep neural networks with relu activation function', *The Annals of Statistics* **48**(4), 1875–1897.

Shi, C., Wan, R., Chernozhukov, V. and Song, R. (2021), 'Deeply-debiased off-policy interval estimation', *arXiv preprint arXiv:2105.04646* .

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D. and Riedmiller, M. (2014), Deterministic policy gradient algorithms, *in* 'International conference on machine learning', PMLR, pp. 387–395.

Smale, S. and Zhou, D.-X. (2005), 'Shannon sampling ii: Connections to learning theory', *Applied and Computational Harmonic Analysis* **19**(3), 285–302.

Smale, S. and Zhou, D.-X. (2007), 'Learning theory estimates via integral operators and their approximations', *Constructive approximation* **26**(2), 153–172.

Steinwart, I. and Scovel, C. (2012), 'Mercer?s theorem on general domains: On the interaction between measures, kernels, and rkhss', *Constructive Approximation* **35**(3), 363–417.

Sutton, R. S., Barto, A. G. et al. (1998), *Introduction to reinforcement learning*, Vol. 135, MIT press Cambridge.

Thomas, P. S., Theocharous, G., Ghavamzadeh, M., Durugkar, I. and Brunskill, E. (2017), Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing, *in* 'Twenty-Ninth IAAI Conference'.

Van Der Vaart, A. W. and Wellner, J. A. (1996), Weak convergence, *in* 'Weak convergence and empirical processes', Springer, pp. 16–28.

Van Der Vaart, A. and Wellner, J. A. (2011), 'A local maximal inequality under uniform entropy', *Electronic Journal of Statistics* **5**(2011), 192.

Wang, R., Wu, Y., Salakhutdinov, R. and Kakade, S. (2021), Instabilities of offline rl with pre-trained neural representation, *in* 'International Conference on Machine Learning', PMLR, pp. 10948–10960.

Watkins, C. J. and Dayan, P. (1992), 'Q-learning', *Machine learning* **8**(3), 279–292.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P. and Agarwal, A. (2021), 'Bellman-consistent pessimism for offline reinforcement learning', *Advances in neural information processing systems* **34**, 6683–6694.

Zanette, A., Wainwright, M. J. and Brunskill, E. (2021), 'Provable benefits of actor-critic methods for offline reinforcement learning', *Advances in neural information processing systems* **34**, 13626–13640.

Zhan, W., Huang, B., Huang, A., Jiang, N. and Lee, J. (2022), Offline reinforcement learning with realizability and single-policy concentrability, *in* 'Conference on Learning Theory', PMLR, pp. 2730–2775.

Zhang, H., Shao, J. and Salakhutdinov, R. (2019), Deep neural networks with multi-branch architectures are intrinsically less non-convex, *in* 'The 22nd International Conference on Artificial Intelligence and Statistics', PMLR, pp. 1099–1109.

Zhao, Y., Zeng, D., Rush, A. J. and Kosorok, M. R. (2012), 'Estimating individualized treatment rules using outcome weighted learning', *Journal of the American Statistical Association* **107**(499), 1106–1118.

A Proof of Regret

Proof of Theorem 1 By summarizing the results of Lemmas 3-5, we can conclude our proof by noting that for any $\pi \in \Pi$, the upper bound in Lemma 4 is the same order of that in Lemma 3.

Proof of Theorem 2: By Theorem 1, with probability at least $1-1/(NT)$, the regret can be decomposed as

$$
\begin{aligned}
\text{Regret}(\widehat{\pi}) &= (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q^{\pi^*}(S_0,\pi^*(S_0))\right] - (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q^{\widehat{\pi}}(S_0,\widehat{\pi}(S_0))\right] \\
&\le (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q^{\pi^*}(S_0,\pi^*(S_0))\right] - \min_{Q\in\Omega(\widehat{\pi},\mathcal{F},\mathcal{W},\varepsilon_{NT}^{(1)},\varepsilon_{NT}^{(2)})}(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\widehat{\pi}(S_0))\right] \\
&\le (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q^{\pi^*}(S_0,\pi^*(S_0))\right] - \min_{Q\in\Omega(\pi^*,\mathcal{F},\mathcal{W},\varepsilon_{NT}^{(1)},\varepsilon_{NT}^{(2)})}(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi^*(S_0))\right] \\
&\le \max_{Q\in\Omega(\pi^*,\mathcal{F},\mathcal{W},\varepsilon_{NT}^{(1)},\varepsilon_{NT}^{(2)})}\left\{\mathcal{V}(\pi^*) - (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi^*(S_0))\right]\right\},
\end{aligned}
$$

where the first inequality is based on Theorem 1 and the second one relies on our policy optimization algorithm (17). The last line of the above inequalities transforms the regret of $\widehat{\pi}$ into the estimation error of OPE for Q^{π^*} using $Q \in \Omega(\pi^*,\mathcal{F},\mathcal{W},\varepsilon_{NT}^{(1)},\varepsilon_{NT}^{(2)})$. Lastly, by leveraging results in Lemma 2, as long as $\text{MMD}(\bar{d}_T^{\pi^b}, \lambda_2^{\pi^*}) \leq \delta$ and $\omega^{\pi^*} \in \mathcal{W}$, we obtain that

$$
\begin{aligned}
&\max_{Q \in \Omega_1(\pi^*, \mathcal{F}, \mathcal{W}, \varepsilon_{NT}^{(1)}) \cap \Omega_2(\pi^*, \mathcal{F}, \varepsilon_{NT}^{(2)})} \left\{\mathcal{V}(\pi^*) - (1-\gamma)\mathbb{E}_{S_0 \sim \nu}\left[Q(S_0, \pi^*(S_0))\right]\right\} \\
\leq & \widetilde{\lambda}_1^{\pi^*}(\mathcal{S} \times \mathcal{A}) \times \sup_{Q \in \Omega_1(\pi^*, \mathcal{F}, \mathcal{W}, \varepsilon_{NT}^{(1)})} \sup_{w \in \mathcal{W}} \overline{\mathbb{E}}\left[w(S,A)\left(R + \gamma Q(S', \pi^*(S')) - Q(S,A)\right)\right] \\
&+ \widetilde{\lambda}_2^{\pi^*}(\mathcal{S} \times \mathcal{A}) \times \max_{Q \in \Omega_2(\pi^*, \mathcal{F}, \varepsilon_{NT}^{(2)})} \left\{\left|\overline{\mathbb{E}}\left[\left(R + \gamma Q(S', \pi^*(S')) - Q(S,A)\right)\right]\right| + \delta \|\mathcal{T}^{\pi^*} Q\|_{\mathcal{H}_k}\right\} \\
\leq & \widetilde{\lambda}_1^{\pi^*}(\mathcal{S} \times \mathcal{A}) \times \sup_{Q \in \Omega_1(\pi^*, \mathcal{F}, \mathcal{W}, \varepsilon_{NT}^{(1)})} \sup_{w \in \mathcal{W}} \left\{\overline{\mathbb{E}}\left[w(S,A)\left(R + \gamma Q(S', \pi^*(S')) - Q(S,A)\right)\right]\right. \\
&\quad \left. - \overline{\mathbb{E}}_{NT}\left[w(S,A)\left(R + \gamma Q(S', \pi^*(S')) - Q(S,A)\right)\right]\right\} \\
&+ \widetilde{\lambda}_1^{\pi^*}(\mathcal{S} \times \mathcal{A}) \times \sup_{Q \in \Omega_1(\pi^*, \mathcal{F}, \mathcal{W}, \varepsilon_{NT}^{(1)})} \sup_{w \in \mathcal{W}} \overline{\mathbb{E}}_{NT}\left[w(S,A)\left(R + \gamma Q(S', \pi(S')) - Q(S,A)\right)\right] \\
&+ \widetilde{\lambda}_2^{\pi^*}(\mathcal{S} \times \mathcal{A}) \times \sup_{Q \in \Omega_2(\pi^*, \mathcal{F}, \varepsilon_{NT}^{(2)})} \left\{\left|\ \overline{\mathbb{E}}\left[\left(R + \gamma Q(S', \pi^*(S')) - Q(S,A)\right)\right]\right.\right. \\
&\quad \left. - \overline{\mathbb{E}}_{NT}\left[\left(R + \gamma Q(S', \pi^*(S')) - Q(S,A)\right)\right]\ \right| + \delta \|\mathcal{T}^{\pi^*} Q\|_{\mathcal{H}_k} - \delta \|\widehat{\mathcal{T}}^{\pi^*} Q\|_{\mathcal{H}_k}\Big\} \\
&\quad + \widetilde{\lambda}_2^{\pi^*}(\mathcal{S} \times \mathcal{A}) \times \sup_{Q \in \Omega_2(\pi^*, \mathcal{F}, \varepsilon_{NT}^{(2)})} \left\{\left|\overline{\mathbb{E}}_{NT}\left[\left(R + \gamma Q(S', \pi^*(S')) - Q(S,A)\right)\right]\right| + \delta \|\widehat{\mathcal{T}}^{\pi^*} Q\|_{\mathcal{H}_k}\right\} \\
\lesssim & \varepsilon_{NT}^{(1)} + \lambda_2^{\pi^*}(\mathcal{S} \times \mathcal{A})\delta\varepsilon_{NT}^{(2)} + \widetilde{\lambda}_2^{\pi^*}(\mathcal{S} \times \mathcal{A}) \times \sup_{Q \in \Omega_2(\pi^*, \mathcal{F}, \varepsilon_{NT}^{(2)})} \left\{\left|\overline{\mathbb{E}}_{NT}\left[\left(R + \gamma Q(S', \pi^*(S')) - Q(S,A)\right)\right]\right|\right\} \\
\lesssim & \varepsilon_{NT}^{(1)} + \lambda_2^{\pi^*}(\mathcal{S} \times \mathcal{A})\delta\varepsilon_{NT}^{(2)} + \widetilde{\lambda}_2^{\pi^*}(\mathcal{S} \times \mathcal{A}) \times \sup_{Q \in \mathcal{F}, w \in \mathcal{W}} \left\{\overline{\mathbb{E}}_{NT}\left[w(S,A)\left(R + \gamma Q(S', \pi^*(S')) - Q(S,A)\right)\right]\right\} \\
\lesssim & \varepsilon_{NT}^{(1)} + \lambda_2^{\pi^*}(\mathcal{S} \times \mathcal{A})\delta\varepsilon_{NT}^{(2)},
\end{aligned}
$$

where we use results in Lemmas 3-5 for the last inequality. In the last two line of above derivation, we use the assumption that $\{1, -1\} \in \mathcal{W}$. This concludes our proof.

Proof of Theorem 3: Denote the solution of

$$\max_{\pi \in \Pi, \rho \succeq 0} \min_{Q \in \mathcal{F}} L_{NT}(Q, \rho, \pi)$$

by $(\widehat{\pi}_{\text{dual}}, \widehat{\rho}, \widehat{Q})$, where $\widehat{\rho} \succeq 0$. As shown in Theorem 1, with probability at least $1 - 1/(NT)$, for every $\pi \in \Pi$,

$$Q^\pi \in \Omega(\pi, \mathcal{F}, \mathcal{W}, \varepsilon_{NT}^{(1)}, \varepsilon_{NT}^{(2)}).$$

Then with probability at least $1 - 1/(NT)$, we have the following regret decomposition for any constant $C > 0$.

$$
\begin{aligned}
\text{Regret}(\widehat{\pi}_{\text{dual}}) &= (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q^{\pi^*}(S_0,\pi^*(S_0))\right] - (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q^{\widehat{\pi}_{\text{dual}}}(S_0,\widehat{\pi}_{\text{dual}}(S_0))\right] \\
&\leq (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q^{\pi^*}(S_0,\pi^*(S_0))\right] - (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q^{\widehat{\pi}_{\text{dual}}}(S_0,\widehat{\pi}_{\text{dual}}(S_0))\right] \\
&- \widehat{\rho}_1 \times \left\{\sup_{w\in\mathcal{W}} \overline{\mathbb{E}}_{NT}\left[w(S,A)\left(R+\gamma Q^{\widehat{\pi}_{\text{dual}}}(S',\widehat{\pi}_{\text{dual}}(S')) - Q^{\widehat{\pi}_{\text{dual}}}(S,A)\right)\right] - \varepsilon_{NT}^{(1)}\right\} \\
&- \widehat{\rho}_2 \times \left\{\|\widehat{\mathcal{T}}^{\pi} Q^{\widehat{\pi}_{\text{dual}}}\|_{\mathcal{H}_k} - \varepsilon_{NT}^{(2)}\right\} \\
&\leq (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q^{\pi^*}(S_0,\pi^*(S_0))\right] - \min_{Q\in\mathcal{F}}\{(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\widehat{\pi}_{\text{dual}}(S_0))\right] \\
&-\widehat{\rho}_1 \times \left\{\sup_{w\in\mathcal{W}} \overline{\mathbb{E}}_{NT}\left[w(S,A)\left(R+\gamma Q(S',\widehat{\pi}_{\text{dual}}(S')) - Q(S,A)\right)\right] - \varepsilon_{NT}^{(1)}\right\} \\
&-\widehat{\rho}_2 \times \left\{\|\widehat{\mathcal{T}}^{\widehat{\pi}_{\text{dual}}} Q\|_{\mathcal{H}_k} - \varepsilon_{NT}^{(2)}\right\}\Big\} \\
&\leq (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q^{\pi^*}(S_0,\pi^*(S_0))\right] - \max_{\rho\succeq 0}\min_{Q\in\mathcal{F}}\{(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi^*(S_0))\right] \\
&+\rho_1 \times \left\{\sup_{w\in\mathcal{W}} \overline{\mathbb{E}}_{NT}\left[w(S,A)\left(R+\gamma Q(S',\pi^*(S')) - Q(S,A)\right)\right] - \varepsilon_{NT}^{(1)}\right\} \\
&+\rho_2 \times \left\{\|\widehat{\mathcal{T}}^{\pi^*} Q\|_{\mathcal{H}_k} - \varepsilon_{NT}^{(2)}\right\}\Big\} \\
&\leq (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q^{\pi^*}(S_0,\pi^*(S_0))\right] - \max_{0\preceq\rho\preceq C}\min_{Q\in\mathcal{F}}\{(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi^*(S_0))\right] \\
&+\rho_1 \times \left\{\sup_{w\in\mathcal{W}} \overline{\mathbb{E}}_{NT}\left[w(S,A)\left(R+\gamma Q(S',\pi^*(S')) - Q(S,A)\right)\right] - \varepsilon_{NT}^{(1)}\right\} \\
&+\rho_2 \times \left\{\|\widehat{\mathcal{T}}^{\pi^*} Q\|_{\mathcal{H}_k} - \varepsilon_{NT}^{(2)}\right\}\Big\},
\end{aligned}
$$

where the first inequality is due to $Q^{\pi} \in \Omega(\pi,\mathcal{F},\mathcal{W},\varepsilon_{NT}^{(1)},\varepsilon_{NT}^{(2)})$, the second inequality holds as we minimize over $Q \in \mathcal{F}$ and by Assumption 7 (a), $Q^{\widehat{\pi}_{\text{dual}}} \in \mathcal{F}$, the third inequality is based on the optimization property of (19), and the last one holds because of the restriction on the dual variable ρ. By results in Lemmas 3-5, we can further obtain that

$$
\begin{aligned}
\text{Regret}(\widehat{\pi}_{\text{dual}}) \leq (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q^{\pi^*}(S_0,\pi^*(S_0))\right] - \max_{0\preceq\rho\preceq C}\min_{Q\in\mathcal{F}}\{(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi^*(S_0))\right] \\
+\rho_1 \times \left\{\sup_{w\in\mathcal{W}} \overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma Q(S',\pi^*(S')) - Q(S,A)\right)\right] - 2\varepsilon_{NT}^{(1)}\right\} \\
+\rho_2 \times \left\{\|\mathcal{T}^{\pi^*} Q\|_{\mathcal{H}_k} - \varepsilon_{NT}^{(2)}\right\}\Big\},
\end{aligned}
$$

with probability at least $1 - O(1)/(NT)$. Consider the following constraint optimization problem.

$$\min_{Q\in\widetilde{\Omega}(\pi^*,\mathcal{F},\varepsilon_{NT}^{(1)},\varepsilon_{NT}^{(2)})} \quad (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi^*(S_0))\right],$$

where

$$\begin{aligned}
&\widetilde{\Omega}(\pi^*,\mathcal{F},\varepsilon_{NT}^{(1)},\varepsilon_{NT}^{(2)}) \\
=&\left\{Q\in\mathcal{F}\mid \sup_{w\in\mathcal{W}}\overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma Q(S',\pi^*(S'))-Q(S,A)\right)\right]\leq 2\varepsilon_{NT}^{(1)}\right\} \\
&\cap\left\{Q\in\mathcal{F}\mid \|\mathcal{T}^{\pi^*}Q\|_{\mathcal{H}_k}\leq\varepsilon_{NT}^{(2)}\right\}.
\end{aligned}$$

It can be verified that the objective function is convex functional with respect to Q and the constraint set is constructed by convex mapping of Q under the Assumption 9. Moreover, when $Q^{\pi^*}\in\mathcal{F}$ by Assumption 7 (a), the inequality is strictly satisfied due to the Bellman equation, which implies that Q^{π^*} is the interior point of $\widetilde{\Omega}(\pi^*,\mathcal{F},\varepsilon_{NT})$. Lastly, by Assumption 6 (a), the objective function is always bounded below. Then by Theorem 8.6.1 of Luenberger (1997), strong duality holds. Therefore

$$\begin{aligned}
&\min_{Q\in\widetilde{\Omega}(\pi^*,\mathcal{F},\varepsilon_{NT}^{(1)},\varepsilon_{NT}^{(2)})} \quad (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi^*(S_0))\right] \\
=&\max_{\rho\succeq 0}\min_{Q\in\mathcal{F}}\{(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[Q(S_0,\pi^*(S_0))\right] \\
&+\rho_1\times\left\{\sup_{w\in\mathcal{W}}\overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma Q(S',\pi^*(S'))-Q(S,A)\right)\right]-2\varepsilon_{NT}^{(1)}\right\} \\
&+\rho_2\times\left\{\|\mathcal{T}^{\pi^*}Q\|_{\mathcal{H}_k}-\varepsilon_{NT}^{(2)}\right\}\} \\
=&\max_{\rho\succeq 0}\min_{Q\in\mathcal{F}}L(Q,\rho,\pi).
\end{aligned}$$

In the following, we show that

$$\max_{\rho\succeq 0}\min_{Q\in\mathcal{F}}L(Q,\rho,\pi)=\max_{0\preceq\rho\preceq C}\min_{Q\in\mathcal{F}}L(Q,\rho,\pi)$$

for some $C>0$. For every $\pi\in\Pi$, let $\rho^*(\pi)$ be the optimal dual variables, i.e.,

$$\rho^*(\pi)\in\arg\max_{\rho\succeq 0}\min_{Q\in\mathcal{F}}\ L(Q,\rho,\pi). \tag{33}$$

By the strong duality and complementary slackness, one must have that

$$\rho_1^*(\pi^*)\times\left\{\sup_{w\in\mathcal{W}}\overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma Q^*(S',\pi^*(S'))-Q^*(S,A)\right)\right]-2\varepsilon_{NT}^{(1)}\right\}=0$$

$$\rho_2^*(\pi^*)\times\left\{\|\mathcal{T}^{\pi^*}Q^*\|_{\mathcal{H}_k}-\varepsilon_{NT}^{(2)}\right\}=0,$$

where Q^* is the optimal primal solution. Since the optimal value for the primal problem is always finite duo to Assumption 6 (a), we claim that $\rho_1^*(\pi^*), \rho_2^*(\pi^*)$ are finite. We can show this statement by contradiction. Without loss of generality, if $\rho_1^*(\pi^*) = \infty$ and $\rho_2^*(\pi^*)$ is finite, one must have

$$\sup_{w \in \mathcal{W}} \overline{\mathbb{E}} \left[w(S, A) \left(R + \gamma Q^*(S', \pi^*(S')) - Q^*(S, A) \right) \right] < 2\varepsilon_{NT}^{(1)}$$

so that Q^* is an optimal solution of

$$\begin{aligned}
&\min_{Q \in \mathcal{F}} \{ (1-\gamma) \mathbb{E}_{S_0 \sim \nu} \left[Q(S_0, \pi^*(S_0)) \right] \\
&+ \rho_1^* \times \left\{ \sup_{w \in \mathcal{W}} \overline{\mathbb{E}} \left[w(S, A) \left(R + \gamma Q(S', \pi^*(S')) - Q(S, A) \right) \right] - 2\varepsilon_{NT}^{(1)} \right\} \\
&+ \rho_2^* \times \left\{ \| \mathcal{T}^{\pi^*} Q \|_{\mathcal{H}_k} - \varepsilon_{NT}^{(2)} \right\} \right\},
\end{aligned}$$

with the optimal value $-\infty$. This violates the complementary slackness. Therefore, there always exists an optimal dual solution $\rho^*(\pi^*) = (\rho_1^*(\pi^*), \rho_2^*(\pi^*))$ such that $\max\{\rho_1^*(\pi^*), \rho_2^*(\pi^*)\} \leq C$ some constant $C > 0$ of the above dual problem. This further indicates that

$$\begin{aligned}
&\min_{Q \in \widetilde{\Omega}(\pi^*, \mathcal{F}, \varepsilon_{NT}^{(1)}, \varepsilon_{NT}^{(2)})} (1-\gamma) \mathbb{E}_{S_0 \sim \nu} \left[Q(S_0, \pi^*(S_0)) \right] \\
= &\max_{0 \preceq \rho \preceq C} \min_{Q \in \mathcal{F}} \{ (1-\gamma) \mathbb{E}_{S_0 \sim \nu} \left[Q(S_0, \pi^*(S_0)) \right] \\
&+ \rho_1 \times \left\{ \sup_{w \in \mathcal{W}} \overline{\mathbb{E}} \left[w(S, A) \left(R + \gamma Q(S', \pi^*(S')) - Q(S, A) \right) \right] - 2\varepsilon_{NT}^{(1)} \right\} \\
&+ \rho_2 \times \left\{ \| \mathcal{T}^{\pi^*} Q \|_{\mathcal{H}_k} - \varepsilon_{NT}^{(2)}) \right\} \right\}.
\end{aligned}$$

By leveraging the above equality, we obtain that

$$\begin{aligned}
\text{Regret}(\widehat{\pi}_{\text{dual}}) &\leq (1-\gamma) \mathbb{E}_{S_0 \sim \nu} \left[Q^{\pi^*}(S_0, \pi^*(S_0)) \right] - \min_{Q \in \widetilde{\Omega}(\pi^*, \mathcal{F}, \varepsilon_{NT}^{(1)}, \varepsilon_{NT}^{(2)})} (1-\gamma) \mathbb{E}_{S_0 \sim \nu} \left[Q(S_0, \pi^*(S_0)) \right] \\
&\leq \max_{Q \in \widetilde{\Omega}(\pi^*, \mathcal{F}, \varepsilon_{NT}^{(1)}, \varepsilon_{NT}^{(2)})} \left\{ (1-\gamma) \mathbb{E}_{S_0 \sim \nu} \left[Q^{\pi^*}(S_0, \pi^*(S_0)) \right] - (1-\gamma) \mathbb{E}_{S_0 \sim \nu} \left[Q(S_0, \pi^*(S_0)) \right] \right\} \\
&\lesssim \varepsilon_{NT}^{(1)} + \widetilde{\lambda}_2^{\pi^*}(\mathcal{S} \times \mathcal{A}) \varepsilon_{NT}^{(2)},
\end{aligned}$$

where the last inequality is given by Lemma 2 and the definition of $\widetilde{\Omega}(\pi^*, \mathcal{F}, \varepsilon_{NT}^{(1)}, \varepsilon_{NT}^{(2)})$. See a similar argument in the proof of Theorem 2. This concludes our proof.

B Supporting Lemmas

Proof of Lemma 1: Without loss of generality, assume d_γ^π is the probability density function of the discounted visitation probability measure over $\mathcal{S}\times\mathcal{A}$. Then by the backward Bellman equation, we can show that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$d_\gamma^\pi(s,a) = (1-\gamma)\nu(s)\pi(a\,|\,s) + \gamma\int_{\mathcal{S}\times\mathcal{A}} d_\gamma^\pi(s',a')q(s\,|\,s',a')\pi(a\,|\,s)ds'da',$$

Multiplying $Q^\pi(s,a)-\widetilde{Q}(s,a)$ and integrating out over $\mathcal{S}\times\mathcal{A}$ on both sides gives that

$$\mathbb{E}_{(S,A)\sim d_\gamma^\pi}\left[Q^\pi(S,A)-\widetilde{Q}(S,A)\right] = \mathcal{V}(\pi)-\widetilde{\mathcal{V}}(\pi)+\gamma\mathbb{E}_{(S,A)\sim d_\gamma^\pi}\left[Q^\pi(S',\pi(S'))-\widetilde{Q}(S',\pi(S'))\right]. \tag{34}$$

By using the Bellman equation for Q^π, i.e., for every $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$Q^\pi(s,a) = \mathbb{E}\left[R+\gamma Q^\pi(S',\pi(S'))\,|\,S=s,A=a\right], \tag{35}$$

we can conclude our proof by showing that Equation (34) can be simplified to that

$$\mathcal{V}(\pi)-\widetilde{\mathcal{V}}(\pi) = \mathbb{E}_{(S,A)\sim d_\gamma^\pi}\left[R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right].$$

Proof of Lemma 2: By Lebesgue's decomposition theorem, we can show that

$$\begin{aligned}
\mathcal{V}(\pi)-\widetilde{\mathcal{V}}(\pi) &= \mathbb{E}_{(S,A)\sim d_\gamma^\pi}\left[R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right]\\
&= \widetilde{\lambda}_1^\pi(\mathcal{S}\times\mathcal{A})\times\mathbb{E}_{(S,A)\sim\lambda_1^\pi}\left[R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right]\\
&+\widetilde{\lambda}_2^\pi(\mathcal{S}\times\mathcal{A})\times\mathbb{E}_{(S,A)\sim\lambda_2^\pi}\left[R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right]\\
&= \widetilde{\lambda}_1^\pi(\mathcal{S}\times\mathcal{A})\times\overline{\mathbb{E}}\left[\omega(S,A)\left(R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right)\right]\\
&+\widetilde{\lambda}_2^\pi(\mathcal{S}\times\mathcal{A})\times\mathbb{E}_{(S,A)\sim\lambda_2^\pi}\left[R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right],
\end{aligned}$$

where the last equality uses the change of measure. By the conditions that $\omega\in\mathcal{W}$ and $\mathcal{W}$ is symmetric, we can derive an upper bound for the first term in the above inequality, i.e.,

$$\begin{aligned}
&\overline{\mathbb{E}}\left[\omega(S,A)\left(R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right)\right]\\
\leq&\left|\overline{\mathbb{E}}\left[\omega(S,A)\left(R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right)\right]\right|\\
\leq&\sup_{w\in\mathcal{W}}\overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right)\right].
\end{aligned}$$

Furthermore, under the conditions in Lemma 6 with $\mathrm{MMD}_k(\bar{d}_T^{\pi^b}, \lambda_2^\pi) \leq \delta$, we have $\|\mu_{\pi^b} - \mu_\pi\|_{\mathcal{H}_k} \leq \delta$. Since $\mathcal{T}^\pi \widetilde{Q} \in \mathcal{H}_k$, we have that

$$\begin{aligned}
&\left|\mathbb{E}_{(S,A)\sim\lambda_2^\pi}\left[R + \gamma\widetilde{Q}(S', \pi(S')) - \widetilde{Q}(S,A)\right]\right| \\
=&\left|\langle \mathcal{T}^\pi \widetilde{Q}, \mu_\pi\rangle_{\mathcal{H}_k}\right| \\
=&\max\left\{\langle \mathcal{T}^\pi \widetilde{Q}, \mu_\pi\rangle_{\mathcal{H}_k}, \langle -\mathcal{T}^\pi \widetilde{Q}, \mu_\pi\rangle_{\mathcal{H}_k}\right\} \\
\leq&\max\left\{\sup_{\substack{\mu_{\mathbb{P}}\in\mathcal{H}_k \\ \|\mu_{\pi^b}-\mu_{\mathbb{P}}\|_{\mathcal{H}_k}\leq\delta}} \langle \mathcal{T}^\pi \widetilde{Q}, \mu_{\mathbb{P}}\rangle_{\mathcal{H}_k}, \sup_{\substack{\mu_{\mathbb{P}}\in\mathcal{H}_k \\ \|\mu_{\pi^b}-\mu_{\mathbb{P}}\|_{\mathcal{H}_k}\leq\delta}} \langle -\mathcal{T}^\pi \widetilde{Q}, \mu_{\mathbb{P}}\rangle_{\mathcal{H}_k}\right\},
\end{aligned}$$

where the first equality is based on Lemma 1 and the last inequality is due to that $\|\mu_{\pi^b} - \mu_\pi\|_{\mathcal{H}_k} \leq \delta$. Furthermore, by Cauchy–Schwarz inequality, we can show that

$$\begin{aligned}
\sup_{\substack{\mu_{\mathbb{P}}\in\mathcal{H}_k \\ \|\mu_{\pi^b}-\mu_{\mathbb{P}}\|_{\mathcal{H}_k}\leq\delta}} \langle \mathcal{T}^\pi \widetilde{Q}, \mu_{\mathbb{P}}\rangle_{\mathcal{H}_k} &= \langle \mathcal{T}^\pi \widetilde{Q}, \mu_{\pi^b}\rangle_{\mathcal{H}_k} + \sup_{\substack{\mu_{\mathbb{P}}\in\mathcal{H}_k \\ \|\mu_{\pi^b}-\mu_{\mathbb{P}}\|_{\mathcal{H}_k}\leq\delta}} \langle \mathcal{T}^\pi \widetilde{Q}, \mu_{\mathbb{P}} - \mu_{\pi^b}\rangle_{\mathcal{H}_k} \quad (36)\\
&\leq \langle \mathcal{T}^\pi \widetilde{Q}, \mu_{\pi^b}\rangle_{\mathcal{H}_k} + \delta\|\mathcal{T}^\pi \widetilde{Q}\|_{\mathcal{H}_k}, \quad (37)
\end{aligned}$$

where the equality in the last line holds if $\mu_{\mathbb{P}} = \mu_{\pi^b} + \delta\mathcal{T}^\pi \widetilde{Q}/\|\mathcal{T}^\pi \widetilde{Q}\|_{\mathcal{H}_k}$ for $\mathcal{T}^\pi \widetilde{Q} \neq 0$ or $\mathcal{T}^\pi \widetilde{Q} = 0$. Therefore we can conclude that

$$\sup_{\substack{\mu_{\mathbb{P}}\in\mathcal{H}_k \\ \|\mu_{\pi^b}-\mu_{\mathbb{P}}\|_{\mathcal{H}_k}\leq\delta}} \langle \mathcal{T}^\pi \widetilde{Q}, \mu_{\mathbb{P}}\rangle_{\mathcal{H}_k} = \langle \mathcal{T}^\pi \widetilde{Q}, \mu_{\pi^b}\rangle_{\mathcal{H}_k} + \delta\|\mathcal{T}^\pi \widetilde{Q}\|_{\mathcal{H}_k}.$$

By a similar argument, we can show that

$$\sup_{\substack{\mu_{\mathbb{P}}\in\mathcal{H}_k \\ \|\mu_{\pi^b}-\mu_{\mathbb{P}}\|_{\mathcal{H}_k}\leq\delta}} \langle -\mathcal{T}^\pi \widetilde{Q}, \mu_{\mathbb{P}}\rangle_{\mathcal{H}_k} = \langle -\mathcal{T}^\pi \widetilde{Q}, \mu_{\pi^b}\rangle_{\mathcal{H}_k} + \delta\|\mathcal{T}^\pi \widetilde{Q}\|_{\mathcal{H}_k},$$

and note that

$$\langle \mathcal{T}^\pi \widetilde{Q}, \mu_{\pi^b}\rangle_{\mathcal{H}_k} = \overline{\mathbb{E}}\left[\left(R + \gamma\widetilde{Q}(S', \pi(S')) - \widetilde{Q}(S,A)\right)\right].$$

Summarizing two upper bounds together gives that for every π

$$\begin{aligned}
&\left|\mathcal{V}(\pi)-\widetilde{\mathcal{V}}(\pi)\right| \\
\leq&\widetilde{\lambda}_1^\pi(\mathcal{S}\times\mathcal{A})\times\left|\overline{\mathbb{E}}\left[\omega(S,A)\left(R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right)\right]\right| \\
+&\widetilde{\lambda}_2^\pi(\mathcal{S}\times\mathcal{A})\times\left|\mathbb{E}_{(S,A)\sim\lambda_2^\pi}\left[R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right]\right| \\
\leq&\widetilde{\lambda}_1^\pi(\mathcal{S}\times\mathcal{A})\times\sup_{w\in\mathcal{W}}\overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right)\right] \\
+&\widetilde{\lambda}_2^\pi(\mathcal{S}\times\mathcal{A})\left\{\left|\overline{\mathbb{E}}\left[\left(R+\gamma\widetilde{Q}(S',\pi(S'))-\widetilde{Q}(S,A)\right)\right]\right|+\delta\|\mathcal{T}^\pi\widetilde{Q}\|_{\mathcal{H}_k}\right\}.
\end{aligned}$$

Lemma 3 *Under Assumptions 1-4 and 6, we have with probability at least* $1-\frac{1}{NT}$*, for every* $Q\in\mathcal{F}, w\in\mathcal{W}$ *and* $\pi\in\Pi$*,*

$$\begin{aligned}
&\left|\overline{\mathbb{E}}\left[w(S,A)\left(R+\gamma Q(S',\pi(S'))-Q(S,A)\right)\right]-\overline{\mathbb{E}}_{NT}\left[w(S,A)\left(R+\gamma Q(S',\pi(S'))-Q(S,A)\right)\right]\right| \\
\lesssim&\log(NT)\sqrt{\left(v(\Pi)+v(\mathcal{F})+v(\mathcal{W})\right)/NT}.
\end{aligned}$$

Proof. We apply Lemma 7 to show the statement. Define

$$\mathcal{G}_1=\left\{(s,a,s')\to w(s,a)\left(\widetilde{R}(s,a,s')+\gamma Q(s',\pi(s'))-Q(s,a)\right)\mid\pi\in\Pi,Q\in\mathcal{F},w\in\mathcal{W}\right\}.$$

By Assumption 6, we can show that

$$\log\mathcal{N}(\epsilon,\mathcal{G}_1,\|\bullet\|_\infty)\lesssim\left(v(\mathcal{W})+v(\Pi)+v(\mathcal{F})\right)\log(1/\epsilon).$$

Then by applying Lemma 7, we can show that with probability at least $1-1/(NT)$,

$$\begin{aligned}
&\left|\overline{\mathbb{E}}\left[\left(R+\gamma Q(S',\pi(S'))-Q(S,A)\right)\right]-\overline{\mathbb{E}}_{NT}\left[\left(R+\gamma Q(S',\pi(S'))-Q(S,A)\right)\right]\right| \\
\lesssim&\log(NT)\sqrt{\left(v(\mathcal{W})+v(\Pi)+v(\mathcal{F})\right)/NT}.
\end{aligned}$$

■

Lemma 4 *Under Assumptions 1-4 and 6 (a) and (c), we have with probability at least* $1-\frac{1}{NT}$*, for every* $Q\in\mathcal{F}$ *and* $\pi\in\Pi$*,*

$$\begin{aligned}
&\left|\overline{\mathbb{E}}\left[\left(R+\gamma Q(S',\pi(S'))-Q(S,A)\right)\right]-\overline{\mathbb{E}}_{NT}\left[\left(R+\gamma Q(S',\pi(S'))-Q(S,A)\right)\right]\right| \\
\lesssim&\log(NT)\sqrt{\left(v(\Pi)+v(\mathcal{F})\right)/NT}.
\end{aligned}$$

Proof. The proof is similar as those in Lemma 3, so we omit for brevity. ■

Lemma 5 *Under Assumptions 1-6 (a), and 7, we have with probability at least* $1 - \frac{1}{NT}$*, for every* $Q \in \mathcal{F}$ *and* $\pi \in \Pi$*,*

$$\|\mathcal{T}^\pi Q - \widehat{\mathcal{T}}^\pi Q\|_{\mathcal{H}_k} \lesssim \left(\log(NT)\sqrt{\left(v(\Pi) + v(\mathcal{F})\right)/NT}\right)^{\frac{2c-1}{2c+1}}.$$

.

Proof. Define an operator $T^\pi_{\zeta_{NT}} : \mathcal{F} \to \mathcal{H}_k$ such that

$$T^\pi_{\zeta_{NT}} Q = \arg\min_{f \in \mathcal{H}_k} \overline{\mathbb{E}}\left[\left(R + \gamma Q(S', \pi(S')) - Q(S, A) - f(S, A)\right)^2\right] + \zeta_{NT}\|f\|^2_{\mathcal{H}_k}, \tag{38}$$

for every $Q \in \mathcal{F}$. It can be seen that

$$T^\pi_{\zeta_{NT}} Q = \left(\mathcal{L}_k + \zeta_{NT}\mathcal{I}\right)^{-1} \mathcal{L}_k\left(\mathcal{T}^\pi Q\right),$$

where $\mathcal{I}$ is the identity operator. In order to bound $\sup_{\pi\in\Pi, Q\in\mathcal{F}} \|\widehat{\mathcal{T}}^\pi Q - \mathcal{T}^\pi Q\|_{\mathcal{H}_k}$, we first decompose it as

$$\begin{aligned}
&\sup_{\pi\in\Pi, Q\in\mathcal{F}} \|\widehat{\mathcal{T}}^\pi Q - \mathcal{T}^\pi Q\|_{\mathcal{H}_k} \\
\leq & \underbrace{\sup_{\pi\in\Pi, Q\in\mathcal{F}} \|\widehat{\mathcal{T}}^\pi Q - T^\pi_{\zeta_{NT}} Q\|_{\mathcal{H}_k}}_{\text{Term (I)}} + \underbrace{\sup_{\pi\in\Pi, Q\in\mathcal{F}} \|\mathcal{T}^\pi Q - T^\pi_{\zeta_{NT}} Q\|_{\mathcal{H}_k}}_{\text{Term (II)}}.
\end{aligned}$$

We then derive an upper bound for Term (II), which is the "bias" term for estimating $\mathcal{T}^\pi Q$. We follow the proof of Theorem 4 in Smale and Zhou (2005) and extend their pointwise results to the uniform covergence over Π and $\mathcal{F}$. Specifically, under Assumption 7 (b), for every $\pi \in \Pi$ and $Q \in \mathcal{F}$, there exists $g_{\pi,Q}$ such that $\mathcal{L}^c_k g_{\pi,Q} = \mathcal{T}^\pi Q$. Recall that

$$g_{\pi,Q} = \sum_{i=1}^\infty e_{i,\pi,Q}\phi_i \quad \text{with} \quad \|\{e_{i,\pi,Q}\}_{i\geq1}\|_{\ell_2} = \|g_{\pi,Q}\|_{L^2_{d^{\pi_b}}} < +\infty$$

and thus

$$\mathcal{T}^\pi Q = \sum_{i=1}^\infty e^c_i e_{i,\pi,Q}\phi_i \quad \text{with} \quad \|\{e_{i,\pi,Q}\}_{i\geq1}\|_{\ell_2} < +\infty.$$

By the definition of $T^{\pi}_{\zeta_{NT}}Q$, we can show that for every $\pi \in \Pi$ and $Q \in \mathcal{F}$,

$$\begin{aligned}\mathcal{T}^{\pi}_{\zeta_{NT}}Q - \mathcal{T}^{\pi}Q &= \left(\mathcal{L}_k + \zeta_{NT}\mathcal{I}\right)^{-1}\mathcal{L}_k\left(\mathcal{T}^{\pi}Q\right) - \mathcal{T}^{\pi}Q \\ &= -\sum_{i=1}^{+\infty}\frac{\zeta_{NT}}{\zeta_{NT}+e_i}e_i^c e_{i,\pi,Q}\phi_i.\end{aligned}$$

Then for $1/2 < c \leq 3/2$, we have

$$\begin{aligned}\left\|\mathcal{T}^{\pi}_{\zeta_{NT}}Q - \mathcal{T}^{\pi}Q\right\|^2_{\mathcal{H}_k} &= \sum_{i=1}^{+\infty}\left(\frac{\zeta_{NT}}{\zeta_{NT}+e_i}e_i^{c-1/2}e_{i,\pi,Q}\right)^2 \\ &= \zeta_{NT}^{2c-1}\sum_{i=1}^{+\infty}\left(\frac{\zeta_{NT}}{\zeta_{NT}+e_i}\right)^{3-2c}\left(\frac{e_i}{\zeta_{NT}+e_i}\right)^{2c-1}e^2_{i,\pi,Q} \\ &\leq \zeta_{NT}^{2c-1}\sum_{i=1}^{+\infty}e^2_{i,\pi,Q},\end{aligned}$$

which implies that for any $\pi \in \Pi$ and $Q \in \mathcal{F}$,

$$\|\mathcal{T}^{\pi}Q - T^{\pi}_{\zeta_{NT}}Q\|_{\mathcal{H}_k} \leq \zeta_{NT}^{c-1/2}\|\mathcal{L}_k^{-c}\mathcal{T}^{\pi}Q\|_{L^2_{d^{\pi_b}}},$$

which further gives that

$$\text{Term (II)} \leq \zeta_{NT}^{c-1/2}\sup_{\pi\in\Pi, Q\in\mathcal{F}}\|\mathcal{L}_k^{-c}\mathcal{T}^{\pi}Q\|_{L^2_{d^{\pi_b}}}.$$

Next, we derive an upper bound for the "variance" of $\widehat{\mathcal{T}}^{\pi}Q$, i.e., Term (I). Following the result of Smale and Zhou (2007), we define the sampling operator $U_{\boldsymbol{z}} : \mathcal{H}_k \to \mathbb{R}^{NT}$ with batch data $\{Z_{i,t} = (S_{i,t}, A_{i,t})\}_{1\leq i\leq N, 0\leq t<T}$ over $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$ as

$$U_{\boldsymbol{z}}(f) = \{f(Z_{i,t})\}_{1\leq i\leq N, 0\leq t<T} \in \mathbb{R}^{NT}$$

for any function f defined over $\mathcal{Z}$. The adjoint of the sampling operator is defined as $U_{\boldsymbol{z}}^{\top} : \mathbb{R}^{NT} \to \mathcal{H}_k$ such that

$$U_{\boldsymbol{z}}^{\top}\theta = \sum_{i=1}^{N}\sum_{t=0}^{T-1}\theta_{i,t}k(Z_{i,t}, \bullet), \quad \text{with} \quad \theta \in \mathbb{R}^{NT}.$$

Then we have

$$\widehat{\mathcal{T}}^{\pi}Q = \left(\frac{1}{NT}U_{\boldsymbol{z}}^{\top}U_{\boldsymbol{z}} + \zeta_{NT}I_{NT}\right)^{-1}\frac{1}{NT}U_{\boldsymbol{z}}^{\top}Y_{Q,\pi},$$

where

$$Y_{Q,\pi} = \{Y_{i,t}(Q,\pi) = R_{i,t} + \gamma Q(S_{i,t+1}, \pi(S_{i,t+1})) - Q(S_{i,t}, A_{i,t})\}_{1\leq i\leq N, 0\leq t<T} \in \mathbb{R}^{NT},$$

and I_{NT} is an identity matrix with the dimension NT. It can be seen that for any $Q \in \mathcal{F}$ and $\pi \in \Pi$,

$$\begin{aligned}
&\widehat{\mathcal{T}}^{\pi} Q - \mathcal{T}^{\pi}_{\zeta_{NT}} Q \\
=&\left(\frac{1}{NT} U_{\boldsymbol{z}}^{\top} U_{\boldsymbol{z}} + \zeta_{NT} I_{NT}\right)^{-1} \left(\frac{1}{NT} U_{\boldsymbol{z}}^{\top} Y_{Q,\pi} - \frac{1}{NT} U_{\boldsymbol{z}}^{\top} U_{\boldsymbol{z}} \left(\mathcal{T}^{\pi}_{\zeta_{NT}} Q\right) - \zeta_{NT} \mathcal{T}^{\pi}_{\zeta_{NT}} Q\right) \\
=&\left(\frac{1}{NT} U_{\boldsymbol{z}}^{\top} U_{\boldsymbol{z}} + \zeta_{NT} I_{NT}\right)^{-1} \left(\overline{\mathbb{E}}_{NT}\left[\left(Y_{Q,\pi} - \mathcal{T}^{\pi}_{\zeta_{NT}} Q(S,A)\right) k(Z,\bullet)\right] - \mathcal{L}_k(\mathcal{T}^{\pi} Q - \mathcal{T}^{\pi}_{\zeta_{NT}} Q)\right),
\end{aligned}$$

where the last equation is based on the definition of the sampling operator, its adjoint and the closed form of $\mathcal{T}^{\pi}_{\zeta_{NT}} Q$. Now we can show that

$$\begin{aligned}
&\|\widehat{\mathcal{T}}^{\pi} Q - \mathcal{T}^{\pi}_{\zeta_{NT}} Q\|_{\mathcal{H}_k} \\
\leq&\frac{1}{\zeta_{NT}} \|\overline{\mathbb{E}}_{NT}\left[\left(Y_{Q,\pi} - \mathcal{T}^{\pi}_{\zeta_{NT}} Q(S,A)\right) k(Z,\bullet)\right] - \mathcal{L}_k(\mathcal{T}^{\pi} Q - \mathcal{T}^{\pi}_{\zeta_{NT}} Q)\|_{\mathcal{H}_k}.
\end{aligned}$$

Note that for any $(T-1) \geq t \geq 0$ and $1 \leq i \leq N$, by Assumption 4,

$$\mathbb{E}\left[\left(Y_{i,t}(Q,\pi) - \mathcal{T}^{\pi}_{\zeta_{NT}} Q(S_{i,t}, A_{i,t})\right) k(Z_{i,t},\bullet)\right] = \mathcal{L}_k(\mathcal{T}^{\pi} Q - \mathcal{T}^{\pi}_{\zeta_{NT}} Q).$$

To bound the above empirical process for vector-valued functions, we leverage the result developed in Lemma 8. First, we define the following class of vector-valued functions as

$$\begin{aligned}
\mathcal{G} = &\left\{g : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathcal{H}_k \;\middle|\; g(s,a,s') = \left(Y(Q,\pi,s,a,s') - \mathcal{T}^{\pi}_{\zeta_{NT}} Q(s,a)\right) k((s,a),\bullet)\right. \\
&\left.\text{with} \quad Q \in \mathcal{F}, \pi \in \Pi\right\},
\end{aligned}$$

where $Y(Q,\pi,s,a,s') = \widetilde{R}(s,a,s') + \gamma Q(s',\pi(s')) - Q(s,a)$. Next, we show that for any $g \in \mathcal{G}$, $\sup_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}} \sup_{g\in\mathcal{G}} \|g(s,a,s')\|_{\mathcal{H}_k} < +\infty$, i.e., the envelop function of $\mathcal{G}$ is uniformly bounded above. Since for any $Q \in \mathcal{F}$, $\|Q\|_\infty \leq c_{\mathcal{F}}$ by Assumption 6 (a), we can show that for any $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,

$$\begin{aligned}
\sup_{g\in\mathcal{G}} \|g(s,a,s')\|^2_{\mathcal{H}_k} \leq& \sup_{Q\in\mathcal{F},\pi\in\Pi} \left(Y(Q,\pi,s,a,s') - \mathcal{T}^{\pi}_{\zeta_{NT}} Q(s,a)\right)^2 K((s,a),(s,a)) \\
\leq& 2\left(R_{\max} + (1+\gamma) c_{\mathcal{F}}\right)^2 \kappa^2 \\
&+2\left(\sup_{\pi\in\Pi, Q\in\mathcal{F}} \|\mathcal{T}^{\pi} Q\|_{\mathcal{H}_k} + \sup_{\pi\in\Pi, Q\in\mathcal{F}} \|\mathcal{L}_k^{-c} \mathcal{T}^{\pi} Q\|_{L^2_{d^{\pi_b}}}\right)^2 \kappa^3 \triangleq c^2_{\mathcal{G}},
\end{aligned}$$

where the second inequality is based on Assumption 2 and the following argument.

$$\|\mathcal{T}^{\pi}_{\zeta_{NT}}Q\|_\infty \le \kappa\|\mathcal{T}^{\pi}_{\zeta_{NT}}Q\|_{\mathcal{H}_k} \le \kappa \sup_{\pi\in\Pi, Q\in\mathcal{F}} \|\mathcal{T}^{\pi}Q\|_{\mathcal{H}_k} + \kappa\zeta_{NT}^{c-1/2} \sup_{\pi\in\Pi, Q\in\mathcal{F}} \|\mathcal{L}_k^{-c}\mathcal{T}^{\pi}Q\|_{L^2_{d^{\pi_b}}}$$
$$\le \kappa\left(\sup_{\pi\in\Pi, Q\in\mathcal{F}} \|\mathcal{T}^{\pi}Q\|_{\mathcal{H}_k} + \sup_{\pi\in\Pi, Q\in\mathcal{F}} \|\mathcal{L}_k^{-c}\mathcal{T}^{\pi}Q\|_{L^2_{d^{\pi_b}}}\right),$$

where the second inequality is based on the condition that $\zeta_{NT} \lesssim 1$ and $c > 1/2$. In the following, we calculate the uniform entropy of $\mathcal{G}$ with respect to any empirical probability measure, i.e., $\sup_{\mathbb{P}} \log(\mathcal{N}(\epsilon c_{\mathcal{G}}, \mathcal{G}, L^2(\mathbb{P})))$, where $\mathbb{P}$ refers to the discrete probability measure over $\mathcal{D}_N$. In particular, the metric related to the covering number $\mathcal{N}(\epsilon c_{\mathcal{G}}, \mathcal{G}, L^2(\mathbb{P}))$ is define as

$$\left(\mathbb{E}_{\mathbb{P}}(\|g_1(S,A,S') - g_2(S,A,S')\|^2_{\mathcal{H}_k})\right)^{1/2},$$

for any $g_1, g_2 \in \mathcal{G}$. Then, for any $\epsilon > 0$, it can be seen that for any $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, $\pi_1, \pi_2 \in \Pi$ and $Q_1, Q_2 \in \mathcal{F}$,

$$\begin{aligned}
&\|\left(Y(Q_1,\pi_1,s,a,s') - \mathcal{T}^{\pi_1}_{\zeta_{NT}}Q_1(s,a)\right)k((s,a),\bullet) - \left(Y(Q_2,\pi_2,s,a,s') - \mathcal{T}^{\pi_2}_{\zeta_{NT}}Q_2(s,a)\right)k((s,a),\bullet)\|_{\mathcal{H}_k}\\
&\le \kappa|Y(Q_1,\pi_1,s,a,s') - Y(Q_2,\pi_2,s,a,s')| + \kappa|\mathcal{T}^{\pi_1}_{\zeta_{NT}}Q_1(s,a) - \mathcal{T}^{\pi_2}_{\zeta_{NT}}Q_2(s,a)|\\
&\lesssim \kappa\left(\|Q_1-Q_2\|_\infty + \|\pi_1-\pi_2\|_\infty\right) + \kappa|\mathcal{T}^{\pi_1}_{\zeta_{NT}}Q_1(s,a) - \mathcal{T}^{\pi_1}_{\zeta_{NT}}Q_2(s,a)| + \kappa|\mathcal{T}^{\pi_1}_{\zeta_{NT}}Q_2(s,a) - \mathcal{T}^{\pi_2}_{\zeta_{NT}}Q_2(s,a)|\\
&\le \kappa\left(\|Q_1-Q_2\|_\infty + \|\pi_1-\pi_2\|_\infty\right) + \kappa\left|\mathcal{T}^{\pi_1}_{\zeta_{NT}}(Q_1-Q_2)(s,a)\right| + \kappa|\mathcal{T}^{\pi_1}_{\zeta_{NT}}Q_2(s,a) - \mathcal{T}^{\pi_2}_{\zeta_{NT}}Q_2(s,a)|,
\end{aligned}$$

where the first inequality is based on Assumption 5, the second one is given by Lipschitz condition stated in Assumption 6 (a), and last one is due to the linearity of the operator $\mathcal{T}^{\pi}_{\zeta_{NT}}$. For the second term in the above inequality, we can show that

$$\begin{aligned}
\left|\mathcal{T}^{\pi_1}_{\zeta_{NT}}(Q_1-Q_2)(s,a)\right| &\le \kappa\|\mathcal{T}^{\pi_1}_{\zeta_{NT}}(Q_1-Q_2)\|_{\mathcal{H}_k}\\
&\le \kappa\|\mathcal{T}^{\pi_1}(Q_1-Q_2)\|_{\mathcal{H}_k}\\
&= \|\mathcal{L}_k^{c-1/2}(g_{\pi,Q_1} - g_{\pi,Q_2})\|_2\\
&\lesssim \|g_{\pi,Q_1} - g_{\pi,Q_2})\|_2\\
&\lesssim \|Q_1-Q_2\|_\infty,
\end{aligned}$$

where the first inequality is based on Assumption 5. The second and third lines hold because of Assumption 7 (b) and the spectral decomposition of $\mathcal{L}_k$. The last inequality holds because of the Lipschitz condition in Assumption 7 (c). The third inequality is based on the following argument.

$$\|\mathcal{L}_k^{c-1/2}\left(g_{\pi,Q_1}-g_{\pi,Q_2}\right)\|_2^2=\sum_{i=1}^{+\infty}e_i^{2c-1}e_{i,\pi,Q_1,Q_2}^2 \lesssim \sum_{i=1}^{+\infty}e_{i,\pi,Q_1,Q_2}^2=\|g_{\pi,Q_1}-g_{\pi,Q_2}\|_2,$$

where the inequality holds as $\{e_i\}_{i\geq 1}$ is non-decreasing towards 0 and $c>1/2$. Now we focus on the last term $|\mathcal{T}_{\zeta_{NT}}^{\pi_1}Q_2(s,a)-\mathcal{T}_{\zeta_{NT}}^{\pi_2}Q_2(s,a)|$. Similar as before, we can show that

$$\begin{aligned}\left|\mathcal{T}_{\zeta_{NT}}^{\pi_1}Q_2(s,a)-\mathcal{T}_{\zeta_{NT}}^{\pi_2}Q_2(s,a)\right| &\leq \kappa\|(\mathcal{L}_k+\zeta_{NT}I)^{-1}\mathcal{L}_k(\mathcal{T}^{\pi_1}Q_2-\mathcal{T}^{\pi_2}Q_2)\|_{\mathcal{H}_k}\\ &\lesssim \|\mathcal{T}^{\pi_1}Q_2-\mathcal{T}^{\pi_2}Q_2\|_{\mathcal{H}_k}\\ &=\|\mathcal{L}_k^{c-1/2}g_{\pi_1,Q_2}-\mathcal{L}_k^{c-1/2}g_{\pi_2,Q_2}\|_2\\ &\lesssim \|g_{\pi_1,Q_2}-g_{\pi_2,Q_2}\|_2\\ &\lesssim \|\pi_1-\pi_2\|_\infty,\end{aligned}$$

where we use Assumptions 7 (b)-(c) again.

Therefore, we can show that

$$\sup_{\mathbb{P}}\log(\mathcal{N}(\epsilon c_{\mathcal{G}},\mathcal{G},L^2(\mathbb{P})))\lesssim \log\mathcal{N}(\epsilon,\mathcal{F},\|\bullet\|_\infty)+\log\mathcal{N}(\epsilon,\Pi,\|\bullet\|_\infty)\asymp(v(\Pi)+v(\mathcal{F}))\log(1/\epsilon).$$

Then by Lemma 8, we can show that with probability at least $1-\frac{1}{NT}$,

$$\begin{aligned}&\|\left(\mathbb{E}_{NT}\left[\left(Y_{Q,\pi}-\mathcal{T}_{\zeta_{NT}}^{\pi}Q(S,A)\right)k(Z,\bullet)\right]-\mathcal{L}_k(\mathcal{T}^\pi Q-\mathcal{T}_{\zeta_{NT}}^{\pi}Q)\right)\|_{\mathcal{H}_k}\\ \lesssim&\log(NT)\sqrt{\left(v(\Pi)+v(\mathcal{F})\right)/NT}.\end{aligned}$$

Summarizing Terms (I) and (II) gives that

$$\begin{aligned}&\sup_{\pi\in\Pi,Q\in\mathcal{F}}\|\widehat{\mathcal{T}}^\pi Q-\mathcal{T}^\pi Q\|_{\mathcal{H}_k}\\ \lesssim&\zeta_{NT}^{-1}\log(NT)\sqrt{\left(v(\Pi)+v(\mathcal{F})\right)/NT}+\zeta_{NT}^{c-1/2}\sup_{\pi\in\Pi,Q\in\mathcal{F}}\|\mathcal{L}_k^{-c}\mathcal{T}^\pi Q\|_{L^2_{d^{\pi_b}}}.\end{aligned}$$

By choosing $\zeta_{NT}\asymp\left(\log(NT)\sqrt{\left(v(\Pi)+v(\mathcal{F})\right)/NT}\right)^{2/(2c+1)}$, we optimize the right-hand-side of the above inequality and obtain that

$$\sup_{\pi\in\Pi,Q\in\mathcal{F}}\|\widehat{\mathcal{T}}^\pi Q-\mathcal{T}^\pi Q\|_{\mathcal{H}_k}\lesssim\sup_{\pi\in\Pi,Q\in\mathcal{F}}\|\mathcal{L}_k^{-c}\mathcal{T}^\pi Q\|_{L^2_{d^{\pi_b}}}^{2/(2c+1)}\left(\log(NT)\sqrt{\left(v(\Pi)+v(\mathcal{F})\right)/NT}\right)^{\frac{2c-1}{2c+1}}.$$

■

C Additional Definitions and Supporting Lemmas

We define the ϵ-covering number below, which is used in the main text.

Definition C.1 (ϵ-covering number) *An ϵ-cover of a set Θ with respect to some semi-metric $\widetilde{d}$ is a set of finite elements $\{\theta_i\}_{i\geq 1} \subseteq \Theta$ such that for every $\theta \in \Theta$, there exists θ_j such that $\widetilde{d}(\theta_j, \theta) \leq \epsilon$. An ϵ-covering number of a set Θ denoted by $\mathcal{N}(\epsilon, \Theta, \widetilde{d})$ is the infimum of the cardinality of ϵ-cover of Θ.*

The following lemma provides a sufficient condition for the existence of mean embeddings for both λ_2^π and $\bar{d}_T^{\pi^b}$. For notation simplicity, let $Z = (S, A)$ and denote $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$.

Lemma 6 *If the kernel $k(\bullet, \bullet) : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is measurable with respect to both λ_2^π and $\bar{d}_T^{\pi^b}$ and $\max\{\overline{\mathbb{E}}[\sqrt{k(Z,Z)}], \mathbb{E}_{Z\sim\lambda_2^\pi}[\sqrt{k(Z,Z)}]\} < +\infty$, then there exist mean embeddings $\mu_{\pi^b}, \mu_\pi \in \mathcal{H}_k$ such that for any $f \in \mathcal{H}_k$*

$$\overline{\mathbb{E}}\left[f(Z)\right] = \langle \mu_{\pi^b}, f\rangle \quad \textit{and} \quad \mathbb{E}_{Z\sim\lambda_2^\pi}\left[f(Z)\right] = \langle \mu_\pi, f\rangle.$$

The proof of Lemma 6 can be found in Lemma 3 of Gretton et al. (2012). Note that if there exists a positive constant δ such that $\text{MMD}_k(\bar{d}_T^{\pi^b}, \lambda_2^\pi) \leq \delta$, then under the conditions in Lemma 6, we can show that the mean embeddings μ_{π^b} and μ_π must satisfy that

$$\|\mu_{\pi^b} - \mu_\pi\|_{\mathcal{H}_k} \leq \delta. \tag{39}$$

See Lemma 4 of Gretton et al. (2012) for more details. Equation (39) motivates us to bound the singular part of the OPE error $|\mathcal{V}(\pi) - \widetilde{\mathcal{V}}(\pi)|$ by its worst-case performance over all the mean embeddings that are within δ-distance to μ_{π^b}.

Lemma 7 *Let $\{\{Z_{i,t}\}_{0\leq t<T}\}_{1\leq i\leq N}$ be i.i.d. copies of stochastic process $\{Z_t\}_{t\geq 0}$. Suppose $\{Z_t\}_{t\geq 0}$ is a stationary and exponential β-mixing process with β-mixing coefficient $\beta(q) \leq \beta_0 \exp(-\beta_1 q)$ for some $\beta_0 \geq 0$ and $\beta_1 > 0$. Let $\mathcal{G}$ be a class of measurable functions that take Z_t as input. For any $g \in \mathcal{G}$, assume $\mathbb{E}[g(Z_t)] = 0$ for any $t \geq 0$. Suppose the envelop function of $\mathcal{G}$ is uniformly bounded by some constant $C > 0$. In addition, if $\mathcal{G}$ belongs to* *the class of VC-typed functions such that* $\mathcal{N}(\epsilon, \mathcal{G}, \|\bullet\|_\infty) \lesssim (1/\epsilon)^\alpha$. *Then with probability at least* $1 - 1/(NT)$,

$$\sup_{g\in\mathcal{G}} |\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=0}^{T-1} g(Z_{i,t})| \lesssim \log(NT)\sqrt{\frac{\alpha}{NT}}.$$

Proof. To prove the lemma, it is sufficient to bound $\sup_{g\in\mathcal{G}} |\sum_{i=1}^{N}\sum_{t=0}^{T-1} g(Z_{i,t})|$. In the following, we apply Berbee's coupling lemma (Berbee, 1979) and follow the remark below Lemma 4.1 of Dedecker and Louhichi (2002). Specifically, Let q be some positive integer. We can always construct a sequence $\{\widetilde{Z}_{i,t}\}_{t\geq 0}$ such that with probability at least $1 - (NT\beta(q))/q$,

$$\sup_{g\in\mathcal{G}} |\sum_{i=1}^{N}\sum_{t=0}^{T-1} g(Z_{i,t})| = \sup_{g\in\mathcal{G}} |\sum_{i=1}^{N}\sum_{t=0}^{T-1} g(\widetilde{Z}_{i,t})|,$$

and meanwhile the block sequence $\widetilde{X}_{i,k}(g) = \{g(\widetilde{Z}_{i,(k-1)q+j})\}_{0\leq j<q}$ are identically distributed for $k \geq 1$ and $i \geq 1$. In addition, the sequence $\{\widetilde{X}_{i,k}(g) \mid k = 2\omega, \omega \geq 1\}$ are independent and so are the sequence $\{\widetilde{X}_{i,k}(g) \mid k = 2\omega+1, \omega \geq 0\}$. Let $I_r = \{\lfloor T/q\rfloor q, \cdots, T-1\}$ with $\text{Card}(I_r) < q$. Then we can show that with probability at least $1 - (NT\beta(q))/q$,

$$\begin{aligned} &\sup_{g\in\mathcal{G}} |\sum_{i=1}^{N}\sum_{t=0}^{T-1} g(Z_{i,t})| \\ \leq &\sup_{g\in\mathcal{G}} |\sum_{i=1}^{N}\sum_{t=0}^{q\lfloor T/q\rfloor-1} g(\widetilde{Z}_{i,t})| + \sup_{g\in\mathcal{G}} |\sum_{i=1}^{N}\sum_{t\in I_r} g(Z_{i,t})|. \end{aligned}$$

In the following, assuming that the above inequality holds, we bound each of the above two terms separately. First of all, without loss of generality, we assume $\lfloor T/q\rfloor$ is an even number. Then for the first term, we have

$$\begin{aligned} &\sup_{g\in\mathcal{G}} |\sum_{i=1}^{N}\sum_{t=0}^{q\lfloor T/q\rfloor-1} g(\widetilde{Z}_{i,t})| \\ \leq &\sum_{j=0}^{2q-1}\sup_{g\in\mathcal{G}} |\sum_{i=1}^{N}\sum_{k=0}^{\lfloor T/q\rfloor/2-1} g(\widetilde{Z}_{i,2kq+j})|. \end{aligned}$$

By the construction, $\sup_{g\in\mathcal{G}} |\sum_{i=1}^{N}\sum_{k=0}^{\lfloor T/q\rfloor/2-1} g(\widetilde{Z}_{i,2kq+j})|$ is a suprema empirical process of i.i.d. sequences. Then by conditions in Lemma 7 and Mcdiarmid's inequality, we have with probability at least $1-\varepsilon$,

$$\sup_{g\in\mathcal{G}}|\sum_{i=1}^{N}\sum_{k=0}^{\lfloor T/q\rfloor/2-1} g(\widetilde{Z}_{i,2kq+j})| \lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{N}\sum_{k=0}^{\lfloor T/q\rfloor/2-1} g(\widetilde{Z}_{i,2kq+j})\right|\right] + \sqrt{\frac{NT\log(1/\varepsilon)}{q}}.$$

Given the condition that

$$\mathcal{N}(\epsilon,\mathcal{G},\|\bullet\|_\infty) \lesssim (1/\epsilon)^\alpha.$$

By a standard maximal inequality using uniform entropy integral (e.g., Van Der Vaart and Wellner, 2011), we can show that with probability at least $1-\varepsilon$,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}|\sum_{i=1}^{N}\sum_{k=0}^{\lfloor T/q\rfloor/2-1} g(\widetilde{Z}_{i,2kq+j})|\right] \lesssim \sqrt{\frac{\alpha NT}{q}}.$$

By letting $\varepsilon = 1/(NT)$, we can show that with probability at least $1-1/(NT)$,

$$\sup_{g\in\mathcal{G}}|\sum_{i=1}^{N}\sum_{k=0}^{\lfloor T/q\rfloor/2-1} g(\widetilde{Z}_{i,2kq+j})| \lesssim \sqrt{\frac{\alpha NT}{q}} + \sqrt{\frac{NT\log(NT)}{q}}.$$

Next, we bound $\sup_{g\in\mathcal{G}}|\sum_{i=1}^N\sum_{t\in I_r} g(Z_{i,t})|$. Since for $i\geq 1$, $\sum_{t\in I_r} g(Z_{i,t})$ are i.i.d. sequences. Then by a similar argument as before, we can show that with probability at least $1-1/(NT)$

$$\sup_{g\in\mathcal{G}}|\sum_{i=1}^{N}\sum_{t\in I_r} g(Z_{i,t})| \lesssim q\left(\sqrt{\alpha N} + \sqrt{N\log(NT)}\right).$$

Without loss of generality, we assume that $\log(NT)\leq T$. Otherwise, we can apply a standard maximal inequality on $\sup_{g\in\mathcal{G}}|\frac{1}{N}\sum_{i=1}^N(\sum_{t=0}^{T-1} g(Z_{i,t})/T)|$ by treating it as N i.i.d. sequences for obtaining the desirable result. Summarizing together and by letting $q\asymp \log(NT)$, we can show that with probability at least $1-1/(NT)$,

$$\begin{aligned}
&\sup_{g\in\mathcal{G}}|\sum_{i=1}^{N}\sum_{t=0}^{T-1} g(Z_{i,t})| \\
\lesssim&\log(NT)\sqrt{\frac{\alpha NT}{\log(NT)}} + \log(NT)\sqrt{\frac{NT\log(NT)}{\log(NT)}} + \log(NT)\left(\sqrt{\alpha N} + \sqrt{N\log(NT)}\right) \\
\lesssim&\log(NT)\sqrt{NT\alpha},
\end{aligned}$$

where the last inequality uses $\log(NT) \le T$. This concludes our proof via dividing both sides by NT. ■

Before presenting Lemma 8, let $\mathcal{G}$ be a class of vector-valued functions that take values in $\mathcal{Z}$ as input and output an element in $\mathcal{H}_k$. Then the metric related to the covering number $\mathcal{N}(\epsilon c_{\mathcal{G}}, \mathcal{G}, L^2(\mathbb{P}))$ is defined as

$$\left(\mathbb{E}_{\mathbb{P}}(\|g_1(Z) - g_2(Z)\|^2_{\mathcal{H}_k})\right)^{1/2}.$$

Here we suppose that the envelop function of $\mathcal{G}$, defined as $G(z) = \sup_{g\in\mathcal{G}} \|g(z)\|_{\mathcal{H}_k}$, is uniformly bounded by some constant $c_{\mathcal{H}_k}$, i.e., $\|G\|_\infty \le c_{\mathcal{H}_k}$.

Lemma 8 *Let $\{\{Z_{i,t}\}_{0\le t<T}\}_{1\le i\le N}$ be i.i.d. copies of stochastic process $\{Z_t\}_{t\ge 0}$. Suppose $\{Z_t\}_{t\ge 0}$ is a stationary and exponential β-mixing process with β-mixing coefficient $\beta(q) \le \beta_0 \exp(-\beta_1 q)$ for some $\beta_0 \ge 0$ and $\beta_1 > 0$. Let $\mathcal{G}$ be a class of vector-valued functions that take Z_t as input and output an element in $\mathcal{H}_k$. For any $g \in \mathcal{G}$, assume $\mathbb{E}[g(Z_t)] = 0$ for any $t \ge 0$. Suppose the envelop function of $\mathcal{G}$, defined as $G(z) = \sup_{g\in\mathcal{G}} \|g(z)\|_{\mathcal{H}_k}$, is uniformly bounded by some constant $c_{\mathcal{H}_k}$, i.e., $\|G\|_\infty \le c_{\mathcal{H}_k}$. In addition, if $\mathcal{G}$ belongs to the class of VC-typed functions such that $\sup_{\mathbb{P}} \mathcal{N}(\epsilon c_{\mathcal{H}_k}, \mathcal{G}, L^2(\mathbb{P})) \lesssim (1/\epsilon)^\alpha$. Then with probability at least $1 - 1/(NT)$,*

$$\sup_{g\in\mathcal{G}} \left\| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=0}^{T-1} g(Z_{i,t}) \right\|_{\mathcal{H}_k} \lesssim \log(NT) \sqrt{\frac{\alpha}{NT}}.$$

Proof. We use the same strategy as Lemma 7 to prove this lemma. It is sufficient to bound $\sup_{g\in\mathcal{G}} \| \sum_{i=1}^N \sum_{t=0}^{T-1} g(Z_{i,t})\|_{\mathcal{H}_k}$. In the following, we apply Berbee's coupling lemma Berbee (1979) and follow the remark below Lemma 4.1 of Dedecker and Louhichi (2002). Specifically, Let q be some positive integer. We can always construct a sequence $\{\widetilde{Z}_{i,t}\}_{t\ge 0}$ such that with probability at least $1 - (NT\beta(q))/q$,

$$\sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sum_{t=0}^{T-1} g(Z_{i,t}) \|_{\mathcal{H}_k} = \sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sum_{t=0}^{T-1} g(\widetilde{Z}_{i,t}) \|_{\mathcal{H}_k},$$

and meanwhile the block sequence $\widetilde{X}_{i,k}(g) = \{g(\widetilde{Z}_{i,(k-1)q+j})\}_{0\le j<q}$ are identically distributed for $k \ge 1$. In addition, the sequence $\{\widetilde{X}_{i,k}(g) \mid k = 2\omega, \omega \ge 1s\}$ are independent and so

are the sequence $\{\widetilde{X}_{i,k}(g) \mid k = 2\omega + 1, \omega \geq 0\}$. Let $I_r = \{\lfloor T/q \rfloor q, \cdots, T-1\}$ with $\mathrm{Card}(I_r) < q$. Then we can show that with probability at least $1 - (NT\beta(q))/q$,

$$
\begin{aligned}
&\sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sum_{t=0}^{T-1} g(Z_{i,t}) \|_{\mathcal{H}_k} \\
\leq &\sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sum_{t=0}^{q\lfloor T/q \rfloor - 1} g(\widetilde{Z}_{i,t}) \|_{\mathcal{H}_k} + \sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sum_{t\in I_r} g(Z_{i,t}) \|_{\mathcal{H}_k}.
\end{aligned}
$$

In the following, assuming that the above inequality holds, we bound each of the above two terms separately. First of all, without loss of generality, we assume $\lfloor T/q \rfloor$ is an even number. Then for the first term, we have

$$
\begin{aligned}
&\sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sum_{t=0}^{q\lfloor T/q \rfloor - 1} g(\widetilde{Z}_{i,t}) \|_{\mathcal{H}_k} \\
\leq &\sum_{j=0}^{2q-1} \sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sum_{k=0}^{\lfloor T/q \rfloor/2-1} g(\widetilde{Z}_{i,2kq+j}) \|_{\mathcal{H}_k}.
\end{aligned}
$$

By the construction, $\sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sum_{k=0}^{\lfloor T/q \rfloor/2-1} g(\widetilde{Z}_{i,2kq+j}) \|_{\mathcal{H}_k}$ is a suprema empirical process of i.i.d. vector-valued functions. Then we apply McDiarmid's inequality for vector-valued functions (See Theorem 7 of Rivasplata et al. (2018)). It can be seen that the bounded difference of $\sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sum_{k=0}^{\lfloor T/q \rfloor/2-1} g(\widetilde{Z}_{i,2kq+j}) \|_{\mathcal{H}_k}$ is $c_{\mathcal{H}_k}$. Then with probability at least $1-\varepsilon$, we have

$$
\left| \sup_{g\in\mathcal{G}} \left\| \sum_{i=1}^{N} \sum_{k=0}^{\lfloor T/q \rfloor/2-1} g(\widetilde{Z}_{i,2kq+j}) \right\|_{\mathcal{H}_k} - \mathbb{E}\left[\sup_{g\in\mathcal{G}} \left\| \sum_{i=1}^{N} \sum_{k=0}^{\lfloor T/q \rfloor/2-1} g(\widetilde{Z}_{i,2kq+j}) \right\|_{\mathcal{H}_k} \right] \right| \lesssim \sqrt{NT/q} + \sqrt{NT/q \log(1/\varepsilon)}.
$$

Next, we derive the upper bound for $\mathbb{E}\left[\sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sum_{k=0}^{\lfloor T/q \rfloor/2-1} g(\widetilde{Z}_{i,2kq+j}) \|_{\mathcal{H}_k} \right]$. By the symmetric argument for vector-valued functions (e.g., Lemmas C.1 and C.2 of Park and Muandet (2022)), we can show that

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sum_{k=0}^{\lfloor T/q \rfloor/2-1} g(\widetilde{Z}_{i,2kq+j}) \|_{\mathcal{H}_k} \right] \\
\leq &2\mathbb{E}\left[\sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sum_{k=0}^{\lfloor T/q \rfloor/2-1} \sigma_{i,2kq+j} g(\widetilde{Z}_{i,2kq+j}) \|_{\mathcal{H}_k} \right],
\end{aligned}
$$

where $\{\sigma_{i,2kq+j}\}_{1\le i\le N, 0\le k\le \lfloor T/q\rfloor/2-1}$ are i.i.d. Rademacher variables. Define the empirical Rademacher complexity as

$$\widehat{\mathcal{R}}(\mathcal{G}) = \frac{1}{N\lfloor T/q\rfloor/2}\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\|\sum_{i=1}^{N}\sum_{k=0}^{\lfloor T/q\rfloor/2-1}\sigma_{i,2kq+j}g(\widetilde{Z}_{i,2kq+j})\right\|_{\mathcal{H}_k} \mid \left\{\widetilde{Z}_{i,2kq+j}\right\}_{1\le i\le N, 0\le k\le \lfloor T/q\rfloor/2-1}\right].$$

Notice that the entropy condition in the lemma gives that

$$\sup_{\mathbb{P}}\log(\mathcal{N}(\epsilon c_{\mathcal{H}_k}, \mathcal{G}, L^2(\mathbb{P}))) \lesssim \alpha\log(\frac{1}{\epsilon}).$$

Then by Theorem C.12 of Park and Muandet (2022) with some slight modification, we can show that

$$\begin{aligned}\widehat{\mathcal{R}}(\mathcal{G}) &\lesssim \frac{c_{\mathcal{H}_k}}{\sqrt{N\lfloor T/q\rfloor/2}}\int_0^1\sup_{\mathbb{P}}\sqrt{\log(\mathcal{N}(\epsilon c_{\mathcal{H}_k}, \mathcal{G}, L^2(\mathbb{P})))}d\epsilon\\ &\lesssim \frac{c_{\mathcal{H}_k}\sqrt{\alpha}}{\sqrt{N\lfloor T/q\rfloor/2}}.\end{aligned}$$

This implies that with probability at least $1-\varepsilon$,

$$\begin{aligned}&\sup_{g\in\mathcal{G}}\|\sum_{i=1}^{N}\sum_{k=0}^{\lfloor T/q\rfloor/2} g(\widetilde{Z}_{i,2kq+j})\|_{\mathcal{H}_k}\\ \lesssim& c_{\mathcal{H}_k}\sqrt{\alpha N\lfloor T/q\rfloor/2} + \sqrt{NT/q} + \sqrt{NT/q\log(1/\varepsilon)}.\end{aligned}$$

Next, we bound $\sup_{g\in\mathcal{G}}\|\sum_{i=1}^{N}\sum_{t\in I_r} g(Z_{i,t})\|_{\mathcal{H}_k}$. Since for $i\ge 1$, $\sum_{t\in I_r} g(Z_{i,t})$ are i.i.d. sequences. Then by a similar argument as before, we can show that with probability at least $1-\varepsilon$,

$$\sup_{g\in\mathcal{G}}\|\sum_{i=1}^{N}\sum_{t\in I_r} g(Z_{i,t})\|_{\mathcal{H}_k} \lesssim q\left(\sqrt{\alpha N} + \sqrt{N\log(1/\varepsilon)}\right).$$

Summarizing together and by letting $q = 2\log(NT)$ and $\varepsilon = 1/(NT)$, we can show that with probability at least $1-1/(NT)$,

$$\begin{aligned}&\sup_{g\in\mathcal{G}}\|\sum_{i=1}^{N}\sum_{t=0}^{T-1} g(Z_{i,t})\|_{\mathcal{H}_k}\\ \lesssim&\log(NT)\sqrt{NT\alpha}.\end{aligned}$$

This concludes our proof by dividing both sides by NT. ■

D Additional Numerical Details and Results

In Figure 5, we compare the performance of the kernel-based method with the true propensity scores and that with the estimated propensity scores. It is observed that the latter yields higher regrets consistently.

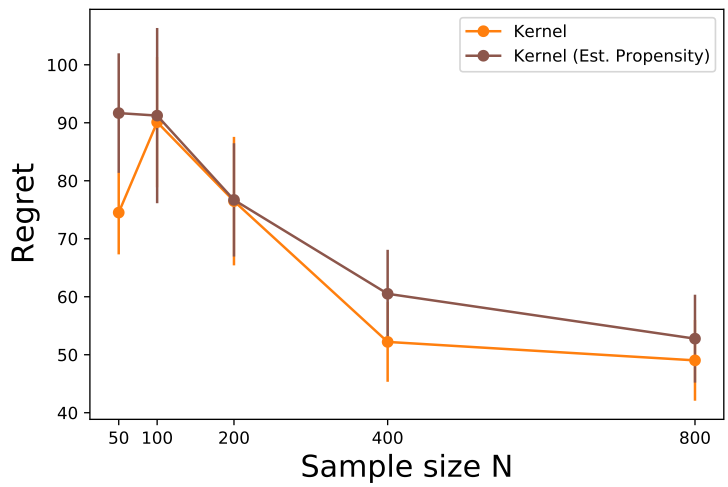


Figure 5: Performance of the kernel-based method: comparison between the variant with the true propensity scores and that with the estimated propensity scores.