Value Function Approximations via Kernel Embeddings for No-Regret Reinforcement Learning

Sayak Ray Chowdhury T-SAYAKR@MICROSOFT.COM
Microsoft Research, India

Rafael Oliveira RAFAEL.OLIVEIRA@SYDNEY.EDU.AU
The University of Sydney, Australia

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

We consider the regret minimization problem in reinforcement learning (RL) in the episodic setting. In many real-world RL environments, the state and action spaces are continuous or very large. Existing approaches establish regret guarantees by either a low-dimensional representation of the stochastic transition model or an approximation of the Q-functions. However, the understanding of function approximation schemes for state-value functions largely remains missing. In this paper, we propose an online model-based RL algorithm, namely the CME-RL, that learns embeddings of the state-transition distribution in a reproducing kernel Hilbert space while carefully balancing the exploitation-exploration tradeoff. We demonstrate the efficiency of our algorithm by proving a frequentist (worst-case) regret bound that is of order $\tilde{O}\big(H\gamma_N\sqrt{N}\big)$[1], where H is the episode length, N is the total number of time steps and γ_N is an information theoretic quantity relating the effective dimension of the state-action feature space. Our method bypasses the need for estimating transition probabilities and applies to any domain on which kernels can be defined. It also brings new insights into the general theory of kernel methods for approximate inference and RL regret minimization.

Keywords: Model-based RL; Value function approximation; Kernel mean embeddings.

1. Introduction

Reinforcement learning (RL) is concerned with learning to take actions to maximize rewards, by trial and error, in environments that can evolve in response to actions. A Markov decision process (MDP) (Puterman, 2014) is a popular framework to model decision making in RL environments. In the MDP, starting from an initial observed state, an agent repeatedly (a) takes an action, (b) receives a reward, and (c) observes the next state of the MDP. The traditional RL objective is a *search* goal – find a *policy* (a rule to select an action for each state) with high total reward using as few interactions with the environment as possible, also known as the sample complexity of RL (Strehl et al., 2009). This is, however, quite different from the corresponding *optimization* goal, where the learner seeks to maximize the total reward earned from all its decisions, or equivalently, minimize the *regret* or shortfall in total reward compared to that of an optimal policy (Jaksch et al., 2010). This objective is relevant in many practical sequential decision-making settings in which every decision that is taken carries utility or value – recommendation systems, sequential investment and

1. $\tilde{O}(\cdot)$ hides only absolute constant and poly-logarithmic factors.

portfolio allocation, dynamic resource allocation in communication systems etc. In such *online* optimization settings, there is no separate budget or time devoted to purely exploring the unknown environment; rather, exploration and exploitation must be carefully balanced.

1.1. Related work

Several studies have considered the task of regret minimization in *tabular* MDPs, in which the state and action spaces are finite, and the value function is represented by a table (Jaksch et al., 2010; Osband et al., 2013; Gheshlaghi Azar et al., 2017; Dann et al., 2017; Jin et al., 2018; Efroni et al., 2019; Zanette and Brunskill, 2019). The regret bound achieved by these works essentially is proportional to $\sqrt{SAN}$, where S and A denote the numbers of states and actions, respectively, and N the total number of steps. In many practical applications, however, the number of states and actions is enormous. For example, the game of Go has a state space with size 3^{361}, and the state and action spaces of certain robotics applications can even be continuous. These continuous state and action spaces make RL a challenging task, especially in terms of generalizing learnt knowledge across unseen states and actions. In such cases, the tabular model suffers from the "curse of dimensionality" problem. To tackle this issue, the popular "optimism in the face of uncertainty" principle from Jaksch et al. (2010) has been extended to handle continuous MDPs, when assuming some Lipschitz-like smoothness or regularity on the rewards and dynamics (Ortner and Ryabko, 2012; Domingues et al., 2020).

Another line of work considers *function approximation*, i.e., they use features to parameterize reward and transition models, with the hope that the features can capture leading structures of the MDP (Osband and Van Roy, 2014; Chowdhury and Gopalan, 2019). The *model-based* algorithms developed in these works assume oracle access to an optimistic planner to facilitate the learning. The optimistic planning step is quite prohibitive and often becomes computationally intractable for continuous state and action spaces. Yang and Wang (2019) consider a low-rank bilinear transition model bypassing the complicated planning step; however, their algorithm potentially needs to compute the *value function* across all states. This suffers an $\Omega(S)$ computational complexity and as a consequence cannot directly handle continuous state spaces. Ayoub et al. (2020) consider linear-mixture transition structure that includes the bilinear model as a special case. However, their algorithm too suffers the $\Omega(S)$ computational complexity. To alleviate the computational burden intrinsic to these model-based approaches, a recent body of work parameterizes the *value functions* directly, using d-dimensional state-action feature maps, and develop *model-free* algorithms bypassing the need for fully learning the reward and transition models (Jin et al., 2019; Wang et al., 2019; Zanette et al., 2020a). Under the assumption that the (action-)value function can be approximated by a linear or a generalized linear function of the feature vectors, these papers develop algorithms with regret bound proportional to $\text{poly}(d)\sqrt{T}$, which is independent of the size of the state and action spaces. Wang et al. (2020) generalizes this approach by designing an algorithm that works with general (non-linear) value function approximators and prove a similar regret guarantee that depends on the eluder dimension (Russo and Van Roy, 2013) and log-covering number of the underlying function class.

A few recent works have proposed kernel-based value function approximation algorithms. Yang et al. (2020) consider kernel and neural function approximations and designed algo-

rithms with regret characterized by intrinsic complexity of the function classes. More closely related to our work, Domingues et al. (2021) recently proposed a kernel-based RL algorithm via value function approximation. Their main assumption relies on Lipschitz continuity of the reward functions and the state transition kernels. In contrast to their work, we are able to obtain tighter regret bounds by applying typical assumptions in the kernel embeddings literature, which we show are satisfied for a variety of practical systems. Nevertheless, there is a lack of theoretical understanding in designing provably efficient model-based RL algorithms with (non-linear) value function approximation, which we aim to address.

1.2. Contributions

In this work, we revisit function approximation in RL by modeling the value functions as elements of a reproducing kernel Hilbert space (RKHS) (Schölkopf and Smola, 2002) compatible with a (possibly infinite dimensional) state feature map. The main motivation behind this formulation is that the conditional expectations of any function in the RKHS become a linear operation, via the RKHS inner product with an appropriate distribution embedding, known as the *conditional mean embedding* (Muandet et al., 2016). In recent years, conditional mean embeddings (CMEs) have found extensive applications in many machine learning tasks (Song et al., 2009, 2010a,b, 2013; Fukumizu et al., 2008, 2009; Hsu and Ramos, 2019; Chowdhury et al., 2020). The foremost advantage of CMEs in our setup is that one can directly compute conditional expectations of the value functions based only on the observed data, since the alternative approach of estimating the transition probabilities as an intermediate step scales poorly with the dimension of the state space (Grünewälder et al., 2012). The convergence of conditional mean estimates to the true embeddings in the RKHS norm has been established by Grünewälder et al. (2012) assuming access to *independent and identically distributed* (i.i.d.) transition samples (the "simulator" setting). However, in the online RL environment like the one considered in this work, one collects data based on past observations, and hence the existing result fails to remain useful. Against this backdrop, we make the following contributions:

- In online RL environment, we derive a concentration inequality for mean embedding estimates of the transition distribution around the true embeddings as a function of the uncertainties around these estimates (Theorem 1). This bound serves as a key tool in designing our model-based RL algorithm, while also being of independent interest.
- Focusing on the value function approximation in the RKHS setting, we present the first model-based RL algorithm, namely the *Conditional Mean Embedding RL* (CME-RL), that is provably efficient in regret performance and does not require any additional oracle access or stronger computational assumptions (Algorithm 1). Concretely, in the general episodic MDP setting, CME-RL enjoys a regret bound of $\tilde{O}(H\gamma_N\sqrt{N})$, where H is the length of each episode, γ_N is a complexity measure relating the effective dimension of the RKHS compatible with the state-action features (Theorem 2).
- Our approach is also robust to the RKHS modelling assumption: when the value functions are not elements of the RKHS, but ζ-close to some RKHS element in the ℓ_∞ norm, then (a modified version of) CME-RL achieves a $\tilde{O}(H\gamma_N\sqrt{N}+\zeta N)$ regret, where the linear regret term arises due to the function class misspecification (Theorem 3).

2. Preliminaries

Notations We begin by introducing some notations. Let $\mathcal{H}$ be an arbitrary Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and corresponding norm $\|\cdot\|_{\mathcal{H}}$. When $\mathcal{G}$ is another Hilbert space, we denote by $\mathcal{L}(\mathcal{G}, \mathcal{H})$ the Banach space of linear operators $\mathrm{A} : \mathcal{G} \to \mathcal{H}$ with bounded operator norm $\|\mathrm{A}\| := \sup_{\|g\|_{\mathcal{G}}=1} \|\mathrm{A}\, g\|_{\mathcal{H}}$. We let $\mathrm{HS}(\mathcal{G}, \mathcal{H})$ denote the subspace of operators in $\mathcal{L}(\mathcal{G}, \mathcal{H})$ with bounded Hilbert-Schmidt norm, defined for $\mathrm{A} \in \mathrm{HS}(\mathcal{G}, \mathcal{H})$ as $\|\mathrm{A}\|_{\mathrm{HS}} := \left(\sum_{i,j=1}^{\infty} \langle f_i, \mathrm{A}\, g_j \rangle_{\mathcal{H}}^2 \right)^{1/2}$, where the f_i's form a complete orthonormal system (CONS) for $\mathcal{H}$ and the g_j's form a CONS for $\mathcal{G}$. In the case $\mathcal{G} = \mathcal{H}$, we set $\mathcal{L}(\mathcal{H}) := \mathcal{L}(\mathcal{H}, \mathcal{H})$. We denote by $\mathcal{L}_+(\mathcal{H})$ the set of all bounded, positive-definite linear operators on $\mathcal{H}$, i.e., $\mathrm{A} \in \mathcal{L}_+(\mathcal{H})$ if, for any non-zero $h \in \mathcal{H}$, $\langle h, \mathrm{A}\, h \rangle_{\mathcal{H}} > 0$.

Regret minimization in finite-horizon episodic MDPs We consider episodic reinforcement learning in a finite-horizon Markov decision process (MDP) of episode length H with (possibly infinite) state and action spaces $\mathcal{S}$ and $\mathcal{A}$, respectively, reward function $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$, and transition probability measure $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ denotes the probability simplex on $\mathcal{S}$. The learning agent interacts with the MDP in episodes and, at each episode t, a trajectory $(s_1^t, a_1^t, r_1^t, \ldots, s_H^t, a_H^t, r_H^t, s_{H+1}^t)$ is generated. Here a_h^t denotes the action taken at state s_h^t, $r_h^t := R(s_h^t, a_h^t)$ denotes the immediate reward, and $s_{h+1}^t \sim P(\cdot | s_h^t, a_h^t)$ denotes the random next state. The initial state s_1^t is assumed to be fixed and history independent, and can even be possibly chosen by an adversary. The episode terminates when s_{H+1}^t is reached, where the agent cannot take any action and hence receives no reward. The actions are chosen following some policy $\pi = (\pi_1, \ldots, \pi_H)$, where each π_h is a mapping from the state space $\mathcal{S}$ into the action space $\mathcal{A}$. The agent would like to find a policy π that maximizes the long-term expected cumulative reward starting from every state $s \in \mathcal{S}$ and every step $h \in [H]$, defined as:

$$V_h^{\pi}(s) := \mathbb{E}\left[\sum_{j=h}^{H} R\left(s_j, \pi_j(s_j)\right) \;\middle|\; s_h = s \right] .$$

We call $V_h^{\pi} : \mathcal{S} \to \mathbb{R}$ the value function of policy π at step h. Accordingly, we also define the action-value function, or Q-function, $Q_h^{\pi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as:

$$Q_h^{\pi}(s, a) := R(s, a) + \mathbb{E}\left[\sum_{j=h+1}^{H} R\left(s_j, \pi_j(s_j)\right) \;\middle|\; s_h = s, a_h = a \right] ,$$

which gives the expected value of cumulative rewards starting from a state-action pair at the h-th step and following the policy π afterwards. Note that $V_h^{\pi}(s) = Q_h^{\pi}(s, \pi_h(s))$ and it satisfies the Bellman equation:

$$V_h^{\pi}(s) = R(s, \pi_h(s)) + \mathbb{E}_{X \sim P(\cdot | s, \pi_h(s))}\left[V_{h+1}^{\pi}(X) \right] , \quad \forall h \in [H] , \tag{1}$$

with $V_{H+1}^{\pi}(s) = 0$ for all $s \in \mathcal{S}$. We denote by $\pi^{\star}$ an optimal policy satisfying:

$$V_h^{\pi^{\star}}(s) = \max_{\pi \in \Pi} V_h^{\pi}(s), \quad \forall s \in \mathcal{S},\ \forall h \in [H],$$

where Π is the set of all non-stationary policies. Since the episode length is finite, such a policy exists when the action space $\mathcal{A}$ is large but finite (Puterman, 2014). We denote the optimal value function by $V_h^{\star}(s) := V_h^{\pi^{\star}}(s)$. We also denote the optimal action-value function (or Q-function) as $Q_h^{\star}(s, a) = \max_{\pi} Q_h^{\pi}(s, a)$. It is easily shown that the optimal

action-value function satisfies the Bellman optimality equation:

$$Q_h^\star(s,a) := R(s,a) + \mathbb{E}_{X\sim P(\cdot|s,a)}\left[V_{h+1}^\star(X)\right], \quad \forall h \in [H], \tag{2}$$

with $V_h^\star(s) = \max_{a\in\mathcal{A}} Q_h^\star(s,a)$. This implies that the optimal policy is the greedy policy with respect to the optimal action-value functions. Thus, to find the optimal policy $\pi^\star$, it suffices to estimate the optimal action-value functions $(Q_h^\star)_{h\in[H]}$.

The agent aims to learn the optimal policy by interacting with the environment during a set of episodes. We measure performance of the agent by the cumulative (pseudo) regret accumulated over T episodes, defined as:

$$\mathcal{R}(N) := \sum_{t=1}^{T}\left[V_1^\star(s_1^t) - V_1^{\pi^t}(s_1^t)\right],$$

where π^t is the policy chosen by the agent at episode t and $N = TH$ is the total number of steps. The regret measures the quantum of reward that the learner gives up by not knowing the MDP in advance and applying the optimal policy $\pi^\star$ from the start. We seek algorithms that attain sublinear regret $\mathcal{R}(N) = o(N)$ in the number of steps they face, since, for instance, an algorithm that does not adapt its policy selection behavior depending on past experience can easily be seen to achieve linear ($\Omega(N)$) regret (Lai and Robbins, 1985).

Value function approximation in episodic MDPs A very large or possibly infinite state and action space makes reinforcement learning a challenging task. To obtain sub-linear regret guarantees, it is necessary to posit some regularity assumptions on the underlying function class. In this paper, we use reproducing kernel Hilbert spaces to model the value functions. Let $\mathcal{H}_\psi$ and $\mathcal{H}_\varphi$ be two RKHSs with continuous positive semi-definite kernel functions $k_\psi : \mathcal{S}\times\mathcal{S} \to \mathbb{R}_+$ and $k_\varphi : (\mathcal{S}\times\mathcal{A})\times(\mathcal{S}\times\mathcal{A}) \to \mathbb{R}_+$, with corresponding inner products $\langle\cdot,\cdot\rangle_{\mathcal{H}_\psi}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}_\varphi}$, respectively. There exist feature maps $\psi : \mathcal{S} \to \mathcal{H}_\psi$ and $\varphi : \mathcal{S}\times\mathcal{A} \to \mathcal{H}_\varphi$ such that $k_\psi(\cdot,\cdot) = \langle\psi(\cdot),\psi(\cdot)\rangle_{\mathcal{H}_\psi}$ and $k_\varphi(\cdot,\cdot) = \langle\varphi(\cdot),\varphi(\cdot)\rangle_{\mathcal{H}_\varphi}$, respectively (Steinwart and Christmann, 2008).

The weakest assumption one can pose on the value functions is realizability, which posits that the optimal value functions $(V_h^\star)_{h\in[H]}$ lie in the RKHS $\mathcal{H}_\psi$, or at least are well-approximated by $\mathcal{H}_\psi$. For stateless MDPs or multi-armed bandits where $H = 1$, realizability alone suffices for provably efficient algorithms (Abbasi-Yadkori et al., 2011; Chowdhury and Gopalan, 2017). But it does not seem to be sufficient when $H > 1$, and in these settings it is common to make stronger assumptions (Jin et al., 2019; Wang et al., 2019, 2020). Following these works, our main assumption is a closure property for all value functions in the following class:

$$\mathcal{V} := \left\{ s \mapsto \min\left\{ H, \max_{a\in\mathcal{A}}\left\{ R(s,a) + \langle\varphi(s,a),\mu\rangle_{\mathcal{H}_\varphi} + \eta\sqrt{\langle\varphi(s,a),\Sigma^{-1}\varphi(s,a)\rangle_{\mathcal{H}_\varphi}}\right\}\right\}\right\}, \tag{3}$$

where $0 < \eta < \infty$, $\mu \in \mathcal{H}_\varphi$ and $\Sigma \in \mathcal{L}_+(\mathcal{H}_\varphi)$ are the parameters of the function class.

Assumption 1 (Optimistic closure) *For any $V \in \mathcal{V}$ (cf. Equation 3), we have $V \in \mathcal{H}_\psi$. Furthermore, for a positive constant B_V, we have $\|V\|_{\mathcal{H}_\psi} \leqslant B_V$.*

While this property seems quite strong, we note that related closure-type assumptions are common in the literature. We will relax this assumption later in Section 4.3. In addition, our results do not require explicit knowledge of $\mathcal{H}_\psi$ nor its kernel k_ψ, as we will only interact with elements of $\mathcal{V}$ via point evaluations and RKHS norm bounds.

3. RKHS embeddings of transition distribution

In order to find an estimate of the optimal value function, it is imperative to estimate the conditional expectations of the form $\mathbb{E}_{X\sim P(\cdot|s,a)}[f(X)]$. In the model-based approach considered in this work, we do so by estimating the mean embedding of the conditional distribution $P(\cdot|s,a)$, which is the focus of this section. For a bounded kernel[2] k_ψ on the state space $\mathcal{S}$, the mean embedding of the conditional distribution $P(\cdot|s,a)$ in $\mathcal{H}_\psi$ is an element $\vartheta_P^{(s,a)} \in \mathcal{H}_\psi$ such that:

$$\forall f \in \mathcal{H}_\psi, \quad \mathbb{E}_{X\sim P(\cdot|s,a)}[f(X)] = \left\langle f, \vartheta_P^{(s,a)} \right\rangle_{\mathcal{H}_\psi} . \tag{4}$$

The mean embedding can be explicitly expressed as a function:

$$\vartheta_P^{(s,a)}(y) = \mathbb{E}_{X\sim P(\cdot|s,a)}\left[k_\psi(X,y)\right],$$

for all $y \in \mathcal{S}$. If the kernel k_ψ is characteristic, such as a stationary kernel, then the mapping $P(\cdot|s,a) \mapsto \vartheta_P^{(s,a)}$ is injective, defining a one-to-one relationship between transition distributions and elements of $\mathcal{H}_\psi$ (Sriperumbudur et al., 2011). Following existing works (Song et al., 2009; Grünewälder et al., 2012), we now make a smoothness assumption on the transition distribution.

Assumption 2 *For any $f \in \mathcal{H}_\psi$, the function $(s,a) \mapsto \mathbb{E}_{X\sim P(\cdot|s,a)}\left[f(X)\right]$ lies in $\mathcal{H}_\varphi$.*

Under Assumption 2, the mean embeddings admit a linear representation in state-action features via the *conditional embedding operator* $\Theta_P \in \mathcal{L}(\mathcal{H}_\varphi, \mathcal{H}_\psi)$ such that:

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \quad \vartheta_P^{(s,a)} = \Theta_P \varphi(s,a) . \tag{5}$$

Assumption 2 always holds for finite domains with characteristic kernels. Though it is not necessarily true for continuous domains, we note that the CMEs for classical linear (Abbasi-Yadkori and Szepesvári, 2011) and non-linear (Kakade et al., 2020) dynamical systems satisfy this assumption.

3.1. Sample estimate of conditional mean embedding

At the beginning of each episode t, given the observations $\mathcal{D}_t := (s_h^\tau, a_h^\tau, s_{h+1}^\tau)_{\tau<t, h\leqslant H}$ until episode $t-1$, we consider a sample based estimate of the conditional embedding operator. This is achieved by solving the following ridge-regression problem:

$$\min_{\Theta\in \mathrm{HS}(\mathcal{H}_\varphi,\mathcal{H}_\psi)} \sum\nolimits_{\tau<t,h\leqslant H} \left\|\psi(s_{h+1}^\tau) - \Theta\varphi(s_h^\tau, a_h^\tau)\right\|_{\mathcal{H}_\psi}^2 + \lambda \left\|\Theta\right\|_{\mathrm{HS}}^2 , \tag{6}$$

where $\lambda > 0$ is a regularising constant. The solution of Equation 6 is given by:

$$\widehat{\Theta}_t = \sum\nolimits_{\tau<t,h\leqslant H} \psi(s_{h+1}^\tau) \otimes \varphi(s_h^\tau, a_h^\tau) \left(\widehat{\mathrm{C}}_{\varphi,t} + \lambda\, \mathrm{I}\right)^{-1} , \tag{7}$$

where $\widehat{\mathrm{C}}_{\varphi,t} := \sum_{\tau<t,h\leqslant H} \varphi(s_h^\tau, a_h^\tau) \otimes \varphi(s_h^\tau, a_h^\tau)$ and $\otimes$ denotes the tensor product of elements in a Hilbert space. To simplify notations, we now let $n = (t-1)H$ denote the total number of steps completed at the beginning of episode t. We denote a vector $k_{\varphi,t}(s,a) \in \mathbb{R}^n$ and a matrix $\mathrm{K}_{\varphi,t} \in \mathbb{R}^{n\times n}$ by:

$$k_{\varphi,t}(s,a) := \left[k_\varphi\left((s_h^\tau, a_h^\tau), (s,a)\right)\right]_{\tau<t,h\leqslant H} , \quad \mathrm{K}_{\varphi,t} := \left[k_\varphi((s_h^\tau, a_h^\tau), (s_{h'}^{\tau'}, a_{h'}^{\tau'}))\right]_{\tau,\tau'<t,h,h'\leqslant H} .$$

2. Boundedness of a kernel holds for any stationary kernel, e.g., the *squared exponential* kernel and the *Matérn* kernel (Rasmussen and Williams, 2006).

Then, via Equation 7, the conditional mean embeddings can be estimated as

$$\widehat{\vartheta}_t^{(s,a)} = \widehat{\Theta}_t \varphi(s,a) = \sum\nolimits_{\tau<t, h\leqslant H} [\alpha_t(s,a)]_{(\tau,h)}\, \psi(s_{h+1}^{\tau}) \ , \tag{8}$$

where we define the weight vector $\alpha_t(s,a) := (\mathrm{K}_{\varphi,t} + \lambda \mathrm{I})^{-1} k_{\varphi,t}(s,a)$.

3.2. Concentration of mean embedding estimates

In this section, we show that for any state-action pair (s,a), the CME estimates $\widehat{\vartheta}_t^{(s,a)}$ lies within a high-probability confidence region around the true embedding $\vartheta_P^{(s,a)}$. This eventually translates, via Equation 4, to a concentration property of $\left\langle f, \widehat{\vartheta}_t^{(s,a)} \right\rangle_{\mathcal{H}_\psi}$ around $\mathbb{E}_{X\sim P(\cdot|s,a)}[f(X)]$ for any $f \in \mathcal{H}_\psi$. The uncertainty of CME estimates can be characterized by the variance estimate $\sigma^2_{\varphi,t}(s,a) := \lambda \left\langle \varphi(s,a), \mathrm{M}_t^{-1}\, \varphi(s,a) \right\rangle_{\mathcal{H}_\varphi}$, where $\mathrm{M}_t := \widehat{\mathrm{C}}_{\varphi,t} + \lambda \mathrm{I}$. To see this, note that an application of Sherman-Morrison formula yields:

$$\sigma^2_{\varphi,t}(s,a) := k_\varphi((s,a),(s,a)) - k_{\varphi,t}(s,a)^\top (\mathrm{K}_{\varphi,t} + \lambda \mathrm{I})^{-1} k_{\varphi,t}(s,a) \ , \tag{9}$$

which is equivalent to the predictive variance of a Gaussian process (GP) (Rasmussen and Williams, 2006). Although a sample from a GP is usually not an element of the RKHS defined by its kernel (Lukic and Beder, 2001), the following result allows us to use $\sigma^2_{\varphi,t}(s,a)$ as an error measure.

Theorem 1 (Concentration of the conditional embedding operator) *Suppose that* $\sup_{s\in\mathcal{S}} \sqrt{k_\psi(s,s)} \leqslant B_\psi$. *Then, under Assumption 2, for any* $\lambda > 0$ *and* $\delta \in (0,1]$,

$$\mathbb{P}\left[\forall t \in \mathbb{N}, \ \left\| \left(\Theta_P - \widehat{\Theta}_t\right) \mathrm{M}_t^{1/2} \right\| \leqslant \beta_t(\delta) \right] \geqslant 1 - \delta \ ,$$

where $\beta_t(\delta) := \sqrt{2\lambda B_P^2 + 256(1+\lambda^{-1}) \log(\det(\mathrm{I} + \lambda^{-1}\mathrm{K}_{\varphi,t})^{1/2}) \log(2t^2 H/\delta)}$, $B_P \geqslant \|\Theta_P\|_{\mathrm{HS}}$.

Theorem 1 implies a concentration inequality for the CME estimates, since, for all $t \geqslant 1$:

$$\left\| \vartheta_P^{(s,a)} - \widehat{\vartheta}_t^{(s,a)} \right\|_{\mathcal{H}_\psi} \leqslant \left\| \left(\Theta_P - \widehat{\Theta}_t\right) \mathrm{M}_t^{1/2} \right\| \|\varphi(s,a)\|_{\mathrm{M}_t^{-1}} \leqslant \beta_t(\delta) \lambda^{-1/2} \sigma_{\varphi,t}(s,a), \ \forall (s,a) \in \mathcal{S}\times\mathcal{A},$$

with probability at least $1-\delta$. This forms the core of our value function approximations.[3]

Remark 1 *Considering the simulation setting, Grünewälder et al. (2012) assume access to a sample* $(s_i, a_i, s_i')_{i=1}^m$, *drawn i.i.d. from a joint distribution* P_0 *such that the conditional probabilities satisfy* $P_0(s_i'|s_i,a_i) = P(s_i'|s_i,a_i), \forall i$. *Under Assumption 2, they establish the convergence of CME estimates* $\widehat{\vartheta}_t^{(s,a)}$ *to the true CMEs* $\vartheta_P^{(s,a)}$ *in* P_0*-probability. This guarantee, however, does not apply to our setting, since we do not assume any simulator access.*

Proof sketch of Theorem 1 To derive this result, we note that the sequence evaluation noise $\varepsilon_h^t := \psi(s_{h+1}^t) - \Theta_P \varphi(s_h^t, a_h^t)$ at each step h of episode t forms a martingale difference sequence, with each element having a bounded RKHS norm. We overload notation to define, for each pair (t,h), the operator $\mathrm{M}_{t,h} = \mathrm{M}_t + \sum_{j\leqslant h} \varphi(s_j^t, a_j^t) \otimes \varphi(s_j^\tau, a_j^\tau)$, and the estimate

$$\widehat{\Theta}_{t,h} = \left(\sum\nolimits_{\tau<t, h\leqslant H} \psi(s_{h+1}^\tau) \otimes \varphi(s_h^\tau, a_h^\tau) + \sum\nolimits_{j\leqslant h} \psi(s_{j+1}^t) \otimes \varphi(s_j^t, a_j^t) \right) \mathrm{M}_{t,h}^{-1} \ .$$

3. Deshmukh et al. (2017) employ a variant of kernel ridge regression to learn the mean reward function of a contextual bandit problem. Their concentration bound holds only for finite action space. In contrast, Theorem 1 holds for infinite state-action spaces, and hence, can be seen as a generalization of their result.

Now, we consider the random variable $z_{t,h} = \left\|(\widehat{\Theta}_{t,h} - \Theta_P)\,\mathrm{M}_{t,h}^{1/2}\right\|_{\mathrm{HS}}^2$, and prove a high-probability upper bound on it using Azuma-Hoeffding's inequality for martingales. In fact, we show that $z_{t,h} \leqslant \beta_{t,h}^2(\delta)$ uniformly over all pair (t,h) with probability at least $1-\delta$, where $\beta_{t,h}(\delta)$ is defined similarly to $\beta_t(\delta)$ with only $\mathrm{K}_{\varphi,t}$ being replaced by $\mathrm{K}_{\varphi,t,h} := [k_\varphi((s_j^\tau, a_j^\tau), (s_{j'}^{\tau'}, a_{j'}^{\tau'}))]_{(\tau,j),(\tau',j')\leqslant(t,h)}$ – the gram-matrix at step h of episode t. The proof then follows by noting that $\left\|(\widehat{\Theta}_t - \Theta_P)\,\mathrm{M}_t^{1/2}\right\|_{\mathrm{HS}} = z_{t-1,H}^{1/2} \leqslant \beta_{t-1,H}(\delta) \leqslant \beta_t(\delta)$. The complete proof is given in the supplementary material.

4. RL exploration using RKHS embeddings

In this section, we aim to develop an online RL algorithm using the conditional mean embedding estimates that balances exploration and exploitation (near) optimally. We realize this, at a high level, by following the Upper-Confidence Bound (UCB) principle and thus our algorithm falls in a similar framework as in Jaksch et al. (2010); Gheshlaghi Azar et al. (2017); Yang and Wang (2019).

4.1. The Conditional Mean Embedding RL (CME-RL) algorithm

At a high level, each episode t consists of two passes over all steps. In the first pass, we maintain the Q-function estimates via dynamic programming. To balance the exploration-exploitation trade-off, we first define a confidence set $\mathcal{C}_t$ that contains the set of conditional embedding operators that are deemed to be consistent with all the data that has been collected in the past. Specifically, for any $\delta \in (0,1]$, $\lambda > 0$ and constants B_P and B_ψ, Theorem 1 governs us to define the confidence set

$$\mathcal{C}_t := \left\{\Theta \in \mathcal{L}(\mathcal{H}_\varphi, \mathcal{H}_\psi) : \left\|(\Theta - \widehat{\Theta}_t)\,\mathrm{M}_t^{1/2}\right\| \leqslant \beta_t(\delta/2)\right\}\ , \tag{10}$$

where $\beta_t(\cdot)$ governs the exploration-exploitation trade-off. This confidence set is then used to compute the optimistic Q-estimates, starting with $V_{H+1}^t(s) = 0$, and setting:

$$\text{for } h = H, H-1, \ldots, 1, \quad V_h^t(s) = \min\left\{H, \max_{a\in\mathcal{A}} Q_h^t(s,a)\right\}, \tag{11}$$

$$Q_h^t(s,a) = R(s,a) + \max_{\Theta_{P'}\in\mathcal{C}_t} \mathbb{E}_{X\sim P'(\cdot|s,a)}\left[V_{h+1}^t(X)\right]\ . \tag{12}$$

We note here that we only require an optimistic estimate of the optimal Q-function. Hence, it is not necessary to solve the maximization problem in Equation 12 explicitly. In fact, we can use a closed-form expression instead of searching for the optimal embedding operator $\Theta_{P'}$ in the confidence set $\mathcal{C}_t$. If the value estimate V_{h+1}^t lies in the RKHS $\mathcal{H}_\psi$, we then have from Equation 4 that $\mathbb{E}_{X\sim P'(\cdot|s,a)}\left[V_{h+1}^t(X)\right] = \left\langle V_{h+1}^t, \vartheta_{P'}^{s,a}\right\rangle_{\mathcal{H}_\psi}$, and from Equation 8 that:

$$\left\langle V_{h+1}^t, \widehat{\vartheta}_t^{(s,a)}\right\rangle_{\mathcal{H}_\psi} = \alpha_t(s,a)^\top v_{h+1}^t = k_{\varphi,t}(s,a)^\top(\mathrm{K}_{\varphi,t} + \lambda\,\mathrm{I})^{-1} v_{h+1}^t\ ,$$

where we define the vector $v_{h+1}^t := [V_{h+1}^t(s_{h'+1}^\tau)]_{\tau<t,h'\leqslant H}$. Now, since the confidence set $\mathcal{C}_t$ is convex, the Q-updates given by Equation 12 admit the closed-form expression:

$$Q_h^t(s,a) = R(s,a) + k_{\varphi,t}(s,a)^\top(\mathrm{K}_{\varphi,t} + \lambda\,\mathrm{I})^{-1} v_{h+1}^t + \left\|V_{h+1}^t\right\|_{\mathcal{H}_\psi} \beta_t(\delta/2)\lambda^{-1/2}\sigma_{\varphi,t}(s,a)\ . \tag{13}$$

We now note that, by the optimistic closure property (Assumption 1), the value estimate V_h^t given by Equation 11 lies in the RKHS $\mathcal{H}_\psi$, rendering the closed-form expression in Equation 13 valid.

In the second pass, we execute the greedy policy with respect to the Q-function estimates obtained in the first pass. Specifically, at each step h, we chose the action:

$$a_h^t = \pi_h^t(s_h^t) \in \operatorname{argmax}_{a \in \mathcal{A}} Q_h^t(s_h^t, a) \ . \tag{14}$$

The pseudo-code of CME-RL is given in Algorithm 1. Note that, in order to implement CME-RL, we do not need to know the kernel k_ψ; only the knowledge of the upper bound B_V over the RKHS norm of V_{h+1}^t suffices our purpose. For simplicity of representation, we assume that the agent, while not knowing the conditional mean embedding operator Θ_P, knows the reward function R. When R is unknown but an element of the RKHS $\mathcal{H}_\varphi$, our algorithm can be extended naturally with an optimistic reward estimation step at each episode, similar to the contextual bandit setting (Chowdhury and Gopalan, 2017).

Algorithm 1: Conditional Mean Embedding RL (CME-RL)

Input: Kernel k_φ, constants B_P, B_V and B_ψ, parameters $\eta > 0$ and $\delta \in (0, 1]$
for *episode* $t = 1, \ldots, T$ **do**
  Receive the initial state s_1^t and set $V_{H+1}^t(\cdot) = 0$
  for *step* $h = H, \ldots, 1$ **do** `// Update value function estimates`
    $Q_h^t(\cdot, \cdot) = R(\cdot, \cdot) + k_{\varphi,t}(\cdot, \cdot)^\top (\mathrm{K}_{\varphi,t} + \lambda \mathrm{I})^{-1} v_{h+1}^t + B_V \beta_t(\delta/2) \lambda^{-1/2} \sigma_{\varphi,t}(\cdot, \cdot)$
    $V_h^t(\cdot) = \min\left\{H, \max_{a \in \mathcal{A}} Q_h^t(\cdot, a)\right\}$
  for *step* $h = 1, \ldots, H$ **do** `// Run episode`
    Take action $a_h^t \in \operatorname{argmax}_{a \in \mathcal{A}} Q_h^t(s_h^t, a)$ and observe next state $s_{h+1}^t \sim P(\cdot | s_h^t, a_h^t)$

Computational complexity of CME-RL The dominant cost is evaluating the Q-function estimates Q_h^t (Equation 13). As typical in kernel methods (Schölkopf and Smola, 2002), it involves inversion of $tH \times tH$ matrices, which take $O(t^3 H^3)$ time. In the policy execution phase (Equation 14), we do not need to compute the entire Q-function as the algorithm only queries Q-values at visited states. Hence, assuming a constant cost of optimizing over the actions, the per-episode running time is $O(t^3 H^4)$. However, using standard sketching techniques like the Nyström approximation (Drineas and Mahoney, 2005) or the random Fourier features approximation (Rahimi and Recht, 2007), and by using the Sherman-Morrison formula to amortize matrix inversions, per-epsiode running cost can be reduced to $O(m^2 H)$, where m is the dimension of feature approximations.

4.2. Regret bound for CME-RL

In this section, we present the regret guarantee of our algorithm. We first define

$$\gamma_N \equiv \gamma_{\varphi,\lambda,N} := \sup_{\mathcal{X} \subset \mathcal{S} \times \mathcal{A} : |\mathcal{X}| = N} \frac{1}{2} \log\det(\mathrm{I} + \lambda^{-1} \mathrm{K}_{\varphi,\mathcal{X}}) \ ,$$

where $\mathcal{X} = \{(s_i, a_i)\}_{i \in [N]}$ and $\mathrm{K}_{\varphi,\mathcal{X}} = [k_\varphi((s_i, a_i), (s_j, a_j))]_{i,j \in [N]}$ is the gram matrix over the data set $\mathcal{X}$. γ_N denotes the *maximum information gain* about a (random) function f sampled from a zero-mean GP with covariance function k_φ after N noisy observations, obtained by passing f through an i.i.d. Gaussian channel $\mathcal{N}(0, \lambda)$. Consider the case when k_φ is a squared exponential kernel on $\mathbb{R}^d$. Then it can be verified that $\gamma_N = O\left((\log N)^{d+1}\right)$ (Srinivas et al., 2009).

Theorem 2 (Cumulative regret of CME-RL) *Under assumptions 1 and 2, after interacting with the environment for $N = TH$ steps, with probability at least $1 - \delta$,* CME-RL *(Algorithm 1) achieves the regret bound*

$$\mathcal{R}(N) \leqslant 2B_V \alpha_{N,\delta} \sqrt{2(1 + \lambda^{-1} B_\varphi^2 H) N \gamma_N} + 2H \sqrt{2N \log(2/\delta)} \ ,$$

where $B_\varphi \geqslant \sup_{s,a} \sqrt{k_\varphi((s,a),(s,a))}$, and $\alpha_{N,\delta} := \sqrt{2\lambda B_P^2 + 256(1+\lambda^{-1})\gamma_N \log(4N^2/\delta)}$.

Theorem 2 yields a $\tilde{O}(H\gamma_N\sqrt{N})$ regret bound for CME-RL. Comparing to the minimax regret in tabular setting, $\Theta(H\sqrt{SAN})$ (Gheshlaghi Azar et al., 2017), our bound replaces the sublinear dependency on the number of state-action pairs by a linear dependency on the intrinsic complexity measure, γ_N, of the feature space $\mathcal{H}_\varphi$, which is crucial in the large state-action space setting that entails function approximation. Additionally, in the kernelized bandit setting ($H = 1$), our bound matches the best known upper bound $O(\gamma_N\sqrt{N})$ (Chowdhury and Gopalan, 2017). We note, however, that while an MDP has state transitions, the bandits do not, and a naive adaptation of existing kernelized bandit algorithms to this setting would give a regret exponential in episode length H. Furthermore, due to the Markov transition structure, the lower bound for kernelized bandits (Scarlett et al., 2017) does not directly apply here. Hence, it remains an interesting future direction to determine the optimal dependency on γ_N.

Conversion to PAC guarantee Similarly to the discussion in Jin et al. (2019), our regret bound directly translates to a sample complexity or probably approximately correct (PAC) guarantee in the following sense. Assuming a fixed initial state $s_1^t = s$ for each episode t, with at least a constant probability, we can learn an ε-optimal policy π that satisfies $V_1^\star(s) - V_1^\pi(s) \leqslant \varepsilon$ by running CME-RL for $T = O(d_{\text{eff}}^2 H^2/\varepsilon^2)$ episodes, where d_{eff} is a known upper bound over γ_N, and then output the greedy policy according to the Q-function at t-th episode, where t is sampled uniformly from $[T]$. Here d_{eff} effectively captures the number of significant dimensions of $\mathcal{H}_\varphi$.

Remark 2 *Yang and Wang (2019) assumes the model $P(s'|s,a) = \langle \psi(s'), \Theta_P \varphi(s,a) \rangle_{\mathcal{H}_\psi}$, and propose an algorithm with regret $\tilde{O}(H^2\gamma_N\sqrt{N})$. In comparison, we get an $O(H)$ factor improvement thanks to a tighter control over the sum of predictive variances. Furthermore, their algorithm can't be implemented exactly as they need to apply random sampling to approximate the estimate $\widehat{\Theta}_t$. We overcome this implementational bottleneck by virtue of our novel confidence set construction using the CME estimates (Theorem 1). Moreover, in contrast to Yang and Wang (2019), our regret guarantee is anytime, i.e., we don't need to know the value of N before the algorithm runs.*

Remark 3 *Considering linear function approximation ($\mathcal{H}_\varphi = \mathbb{R}^d$), Jin et al. (2019) assumes that for any $V \in \mathcal{V}$ (Equation 3), the map $(s,a) \mapsto \mathbb{E}_{X \sim P(\cdot|s,a)}[V(X)]$ lies in $\mathcal{H}_\varphi$, and propose a model-free algorithm with regret $\tilde{O}(\sqrt{H^3 d^3 N})$. For linear kernels, it can be verified that $\gamma_N = O(d \log N)$ and thus our regret (Theorem 2) is of the order $\tilde{O}(Hd\sqrt{N})$. We note that this apparent improvement in our bound is a consequence of slightly stronger assumptions 1 and 2. While they obtain the bound by proving a uniform concentration result over the set $\mathcal{V}$, our result uses a novel concentration property of CME estimates (Theorem 1).*

Proof sketch of Theorem 2 A control on the Q-function estimates Q_h^t leads to the regret bound, as our policy is based on Q_h^t. We prove that as long as Θ_P lies in the confidence set $\mathcal{C}_t$, the Q-updates are optimistic estimates of the optimal Q-values, i.e., $Q_h^\star(s,a) \leqslant Q_h^t(s,a)$ for all (s,a), and thus, allow us to pick an optimistic action while sufficiently exploring the state space. This implies $V_1^\star(s_1^t) \leqslant V_1^t(s_1^t)$, so that the regret $\mathcal{R}(N) \leqslant \sum_{t=1}^T \big(V_1^t(s_1^t) - V_1^{\pi^t}(s_1^t)\big)$. Letting $g_1^t(s_1^t) := V_1^t(s_1^t) - V_1^{\pi^t}(s_1^t)$ denote the gap between the most optimistic value and the actual value obtained at episode t, we then have

$$g_1^t(s_1^t) \leqslant \sum\nolimits_{h=1}^H \big(Q_h^t(s_h^t,a_h^t) - \big(R(s_h^t,a_h^t) + \mathbb{E}_{X\sim P(\cdot|s_h^t,a_h^t)}\left[V_{h+1}^t(X)\right]\big) + m_h^t\big),$$

where $(m_h^t)_{t,h}$ denotes a martingale difference sequence. We control this via the Azuma-Hoeffding inequality as $\sum_{t,h} m_{t,h} = O(H\sqrt{N})$. The rest of the terms inside the summation can be controlled, by Theorem 1 and by design of the confidence set $\mathcal{C}_t$, using the predictive variances $\sigma_{\varphi,t}^2(s_h^t,a_h^t)$. In fact, for $\Theta_P \in \mathcal{C}_t$, it holds that

$$Q_h^t(s,a) - \big(R(s,a) + \mathbb{E}_{X\sim P(\cdot|s,a)}\left[V_{h+1}^t(X)\right]\big) \leqslant 2B_V\beta_t(\delta/2)\lambda^{-1/2}\sigma_{\varphi,t}(s,a)\ .$$

Now, the proof can be completed by showing that $\sum_{t,h}\sigma_{\varphi,t}(s_h^t,a_h^t) = O\big(\sqrt{HN\gamma_N}\big)$. Complete proof of this result is given in the supplementary material.

4.3. Robustness to model misspecification

Theorem 2 hinges on the fact that any optimistic estimate of the value function can be specified as an element in $\mathcal{H}_\psi$. In this section, we study the case when there is a misspecification error. Formally, we consider the following assumption.

Assumption 3 (Approximate optimistic closure) *There exists constants $\zeta > 0$ and $B_V > 0$, such that for any $V \in \mathcal{V}$ (Equation 3), there exists a function $\tilde{V} \in \mathcal{H}_\psi$ which satisfies $\|V - \tilde{V}\|_\infty \leqslant \zeta$ and $\|\tilde{V}\|_{\mathcal{H}_\psi} \leqslant B_V$. We call ζ the misspecification error.*

The quality of this approximation will further depend upon how well any $V \in \mathcal{V}$ can be approximated by a low-norm function in $\mathcal{H}_\psi$. One specialization is to the case when $\mathcal{V} \in \mathcal{C}_b(\mathcal{S})$, the vector space of continuous and bounded functions on $\mathcal{S}$, and k_ψ is a $\mathcal{C}_b(\mathcal{S})$-universal kernel (Steinwart and Christmann, 2008). In this case, we can choose $\tilde{V}$ such that $\|V-\tilde{V}\|_\infty$ is arbitrarily small. For technical reasons, we also make the following assumption.

Assumption 4 *The RKHS $\mathcal{H}_\psi$ contains the constant functions.*[4]

The following theorem states that our algorithm is in fact robust to a small model misspecification. To achieve this, we only need to adopt a different exploration term in Equation 13 to account for the misspecification error ζ. To this end, define the Q-function updates as

$$Q_h^t(s,a) := R(s,a) + k_{\varphi,t}(s,a)^\top(\mathrm{K}_{\varphi,t} + \lambda\,\mathrm{I})^{-1}v_{h+1}^t + \Big(B_V + \zeta\,\|1\|_{\mathcal{H}_\psi}\Big)\,\beta_t(\delta/2)\sigma_{\varphi,t}(s,a)\ , \quad (15)$$

where $\|1\|_{\mathcal{H}_\psi}$ denotes the norm of the all-one function $s \mapsto 1$ in $\mathcal{H}_\psi$.

Theorem 3 (Cumulative regret under misspecification) *Under assumptions 2, 3 and 4, with probability at least $1-\delta$,* CME-RL *achieves the regret bound*

$$\mathcal{R}(N) \leqslant 2\left(B_V + \zeta\,\|1\|_{\mathcal{H}_\psi}\right)\alpha_{N,\delta}\sqrt{2(1+\lambda^{-1}B_\varphi^2 H)N\gamma_N} + 4\zeta N + 2H\sqrt{2N\log(2/\delta)}\ ,$$

4. This is a mild assumption. For any RKHS $\mathcal{H}_\psi$, the direct sum $\mathcal{H}_\psi + \mathbb{R}$, where $\mathbb{R}$ denotes the RKHS associated with the kernel $k(s,s') = 1$, is again a RKHS with kernel $k_{\text{new}}(s,s') := k_\psi(s,s') + 1$.

where B_φ and $\alpha_{N,\delta}$ are as given in Theorem 2.

In comparison with Theorem 2, Theorem 3 asserts that CME-RL will incur at most an additional $O(\zeta\gamma_N\sqrt{HN}+\zeta N)$ regret when the model is misspecified. This additional term is linear in N due to the intrinsic bias introduced by the approximation. This linear dependency is standard in the literature, e.g., it is present even in the easier setting of linear function approximation (Jin et al., 2019). When ζ is sufficiently small (as is typical for universal kernels k_ψ), our algorithm will still enjoy good theoretical guarantees.

Conversion to PAC guarantee Similar to Theorem 2, we can also convert Theorem 3 to a PAC guarantee. Assuming a fixed initial state s, with at least a constant probability, we can learn an ε-optimal policy π that satisfies $V_1^\star(s) - V_1^\pi(s) \leqslant \varepsilon + \zeta\gamma_N H^{3/2}$ by running CME-RL for $T = O(d_{\text{eff}}^2 H^2/\varepsilon^2)$ episodes.

Remark 4 (Regret under unknown misspecification error) *When the misspecification error ζ is unknown to the agent apriori, one can invoke the dynamic regret balancing scheme of Cutkosky et al. (2021) to get essentially a similar bound as Theorem 3 (albeit with a polylog factor blow-up). In fact, Cutkosky et al. (2021) gives a bound for the linear MDP model of Jin et al. (2019). Similar techniques can be incorporated to derive a regret bound with unknown ζ in our setting also.*

Proof sketch of Theorem 3 Similar to the proof of Theorem 2, we control the Q-function estimates $Q_h^t(s,a)$ (cf. Equation 15), but with necessary modifications taking the effect of the misspecification error ζ into account. Specifically, we show, for $\Theta_P \in \mathcal{C}_t$, that

$$Q_h^t(s,a) - \left(R(s,a) + \mathbb{E}_{X\sim P(\cdot|s,a)}\left[V_{h+1}^t(X)\right]\right) \leqslant 2\left(B_V + \zeta\left\|1\right\|_{\mathcal{H}_\psi}\right)\lambda^{-1/2}\beta_t(\delta/2)\sigma_{\varphi,t}(s,a) + 2\zeta\ .$$

With the result above, we can derive an upper bound on the optimal value $Q_h^\star$ as $Q_h^\star(s,a) \leqslant Q_h^t(s,a) + 2(H-h)\zeta$, which allows us to pick an optimistic action. The proof then follows similar steps of Theorem 2 via control of predictive variances and Azuma's inequality. Complete proof is given in the supplementary material.

5. Conclusion

In this paper, we have presented a novel model-based RL algorithm with sub-linear regret guarantees under an optimistic RKHS-closure assumption on the value functions, without requiring a "simulator" access. The algorithm essentially performs an optimistic value iteration step, which is derived from a novel concentration inequality for the mean embeddings of the transition distribution. We have also shown robustness of our algorithm to small model misspecifications.

As future work, it remains an open research direction to relax the strong optimistic closure assumption to a milder one, as in Zanette et al. (2020b) and Domingues et al. (2021), without sacrificing on the computational and regret performances. In terms of computational complexity, Vial et al. (2022) proposed an algorithm for misspecified linear MDPs with bounded per-iteration computational complexity. Although our method has computational complexity growing with the number of data points, we highlight that constant cost per iteration can be achieved with kernel-based approximations by means of low-rank decompositions (Gijsberts and Metta, 2013), which is another avenue for future work.

Acknowledgments

The authors would like to thank Sho Takemori and an anonymous reviewer for pointing out a bug in the proof of Theorem 1 in an earlier version of this work.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin F Yang. Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*, 2020.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 844–853. JMLR. org, 2017.

Sayak Ray Chowdhury and Aditya Gopalan. Online learning in kernelized markov decision processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3197–3205, 2019.

Sayak Ray Chowdhury, Rafael Oliveira, and Fabio Ramos. Active learning of conditional mean embeddings via bayesian optimisation. In *Conference on Uncertainty in Artificial Intelligence*, pages 1119–1128. PMLR, 2020.

Ashok Cutkosky, Christoph Dann, Abhimanyu Das, Claudio Gentile, Aldo Pacchiano, and Manish Purohit. Dynamic balancing for model selection in bandits and rl. In *International Conference on Machine Learning*, pages 2276–2285. PMLR, 2021.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 5711–5721, 2017.

Aniket Anand Deshmukh, Urun Dogan, and Clay Scott. Multi-task learning for contextual bandits. *Advances in neural information processing systems*, 30, 2017.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Regret bounds for kernel-based reinforcement learning. *arXiv preprint arXiv:2004.05599*, 2020.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Kernel-based reinforcement learning: A finite-time analysis. In *International Conference on Machine Learning*, pages 2783–2792. PMLR, 2021.

Petros Drineas and Michael W Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pages 12203–12213, 2019.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pages 489–496, 2008.

Kenji Fukumizu, Francis R Bach, Michael I Jordan, et al. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 263–272, 2017.

Arjan Gijsberts and Giorgio Metta. Real-time model learning using Incremental Sparse Spectrum Gaussian Process Regression. *Neural Networks*, 41:59–69, 2013.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Massimilano Pontil, and Arthur Gretton. Modelling transition dynamics in mdps with rkhs embeddings. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1603–1610, 2012.

Kelvin Hsu and Fabio Ramos. Bayesian Learning of Conditional Kernel Mean Embeddings for Automatic Likelihood-Free Inference. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Naha, Okinawa, Japan, 2019.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *arXiv preprint arXiv:2006.12466*, 2020.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

M. N. Lukic and J. H. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969, 2001.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel Mean Embedding of Distributions: A Review and Beyond. *arXiv*, 2016.

Ronald Ortner and Daniil Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 1763–1771, 2012.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 1466–1474, 2014.

Ian Osband, Dan Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 3003–3011, 2013.

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

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Proceedings of the 21st conference on advances in Neural Information Processing Systems*, NIPS '07, Vancouver, British Columbia, Canada, dec 2007. MIT Press.

C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

Jonathan Scarlett, Ilijia Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy gaussian process bandit optimization. *arXiv preprint arXiv:1706.00090*, 2017.

Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge, Mass, 2002.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

Le Song, Byron Boots, Sajid M Siddiqi, Geoffrey Gordon, and Alex Smola. Hilbert space embeddings of hidden markov models. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 991–998, 2010a.

Le Song, Arthur Gretton, and Carlos Guestrin. Nonparametric tree graphical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 765–772, 2010b.

Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research (JMLR)*, 12:2389–2410, 2011.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite mdps: PAC analysis. *Journal of Machine Learning Research*, 10(Nov):2413–2444, 2009.

Daniel Vial, Advait Parulekar, Sanjay Shakkottai, and R Srikant. Improved algorithms for misspecified linear markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 4723–4746. PMLR, 2022.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Provably efficient reinforcement learning with general value function approximation. *arXiv preprint arXiv:2005.10804*, 2020.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Lin F Yang and Mengdi Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael Jordan. Provably efficient reinforcement learning with kernel and neural function approximations. *Advances in Neural Information Processing Systems*, 33, 2020.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964, 2020a.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. *arXiv preprint arXiv:2003.00153*, 2020b.