Open Problem: Order Optimal Regret Bounds for Kernel-Based Reinforcement Learning

Sattar Vakili　　　　　　　　　　　　　　　　　　　　　　　　　SV388@CORNELL.EDU
MediaTek Research

Editors: Shipra Agrawal and Aaron Roth

Abstract

Reinforcement Learning (RL) has shown great empirical success in various application domains. The theoretical aspects of the problem have been extensively studied over past decades, particularly under tabular and linear Markov Decision Process structures. Recently, non-linear function approximation using kernel-based prediction has gained traction. This approach is particularly interesting as it naturally extends the linear structure, and helps explain the behavior of neural-network-based models at their infinite width limit. The analytical results however do not adequately address the performance guarantees for this case. We will highlight this open problem, overview existing partial results, and discuss related challenges.
Keywords: Kernel methods, reinforcement learning, order optimal performance guarantees

1. Introduction

The analytical study of Reinforcement Learning (RL) faces a natural progression in the complexity of the Markov Decision Process (MDP) structure: Tabular $\rightarrow$ Linear $\rightarrow$ Kernel-based $\rightarrow$ Neural-network-based. Kernel-based models serve as natural intermediary between well-studied linear models and the less understood neural-network-based models (having the neural tangent kernel theory in mind; see, e.g., the discussions in Yang et al., 2020). Kernel-based models provide a rich representation capacity for nonlinear function approximation in RL, while still lending themselves to theoretical analysis, and have gained traction in recent years. However, some fundamental problems still remain open.

For a sharp and clear presentation of our open problem, we focus on an episodic MDP within an online framework under the regret performance measure. However, similar problems can be raised in other settings, including infinite horizon discounted or undiscounted MDPs within an online or offline framework. We specifically ask: *Is it possible to design order-optimal or, at the very least, no-regret learning algorithms under reasonable assumptions on an MDP with a kernel-based structure?* In this paper, we will formally present this open problem, provide an overview of existing work, and discuss some of the challenges involved.

2. Episodic MDP

An episodic MDP can be described by the tuple $M = (\mathcal{S}, \mathcal{A}, H, P, r)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, the integer H is the length of each episode, $r = \{r_h\}_{h=1}^{H}$ are the reward functions and $P = \{P_h\}_{h=1}^{H}$ are the transition probability distributions. We use the notation $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$ to denote the state-action space. For each $h \in [H]$, $r_h : \mathcal{Z} \rightarrow [0,1]$ is the reward function at step h, which is supposed to be deterministic and known for simplicity, and $P_h(\cdot|s,a)$ is the *unknown* transition probability distribution on $\mathcal{S}$ for the next state from state-action pair (s,a).

A policy $\pi = \{\pi_h\}_{h=1}^H$, at each step h, determines the (possibly random) action $\pi_h : \mathcal{S} \to \mathcal{A}$ taken by the agent at state s. At the beginning of each episode $t = 1, 2, \cdots$, the environment picks an arbitrary state $s_{1,t}$. The agent determines a policy $\pi_t = \{\pi_{h,t}\}_{h=1}^H$. Then, at each step $h \in [H]$, the agent observes the state $s_{h,t} \in \mathcal{S}$, and picks an action $a_{h,t} = \pi_{h,t}(s_{h,t})$. The new state $s_{h+1,t}$ then is drawn from the transition distribution $P_h(\cdot|s_{h,t}, a_{h,t})$. The episode ends when the agent receives the final reward $r_H(s_{H,t}, a_{H,t})$. We are interested in maximizing the expected total reward in the episode, starting at step h, i.e., the value function, defined as:

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})\bigg| s_h = s\right], \quad \forall s \in \mathcal{S}, h \in [H], \tag{1}$$

where the expectation is taken with respect to the randomness in the trajectory $\{(s_h, a_h)\}_{h=1}^H$ obtained by the policy π. It can be shown that under mild assumptions (e.g., continuity of P_h, compactness of $\mathcal{Z}$, and boundedness of r) there exists an optimal policy $\pi^\star$ which attains the maximum possible value of $V_h^\pi(s)$ at every step and at every state (e.g., see, Puterman, 2014). We use the notation $V_h^\star(s) = \max_\pi V_h^\pi(s)$, $\forall s \in \mathcal{S}, h \in [H]$. By definition $V_h^{\pi^\star} = V_h^\star$. The performance of a learning algorithm $\{\pi_t\}_{t \in [T]}$ is measured in terms of the total loss in the value function, referred to as *regret*, denoted by $\mathcal{R}(T)$ in the following definition:

$$\mathcal{R}(T) = \sum_{t=1}^T (V_1^\star(s_{1,t}) - V_1^{\pi_t}(s_{1,t})). \tag{2}$$

A learning algorithm with sublinear regret in T is often referred to as a *no-regret* algorithm, since the average regret over T tends toward zero as T increases. This implies that the value of the policy executed by the learning algorithm converges to that of the optimal policy over episodes.

For a value function $V : \mathcal{S} \to \mathbb{R}$, and a conditional distribution $P(s|z)$, $s \in \mathcal{S}, z \in \mathcal{Z}$, we define the notation $[PV](z) = \mathbb{E}_{s \sim P(\cdot|z)}[V(s)]$. The state-action value function $Q_h^\pi : \mathcal{Z} \to [0, H]$ is defined as follows: $Q_h^\pi(s, a) = \mathbb{E}_\pi\left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})\bigg| s_h = s, a_h = a\right]$, where the expectation is taken with respect to the randomness in the trajectory $\{(s_h, a_h)\}_{h=1}^H$ obtained by the policy π. The Bellman equation associated with a policy π then is represented as: $Q_h^\pi(s, a) = r_h(s, a) + [P_h V_{h+1}^\pi](s, a), V_h^\pi(s) = \max_{a \in \mathcal{A}} Q_h^\pi(s, a), V_{H+1}^\pi \equiv \mathbf{0}$.

3. Kernel-Based Modelling

Various structural complexities for MDPs have been considered, including the tabular model with small finite $\mathcal{S}$ and $\mathcal{A}$, where regret bounds of $\tilde{\mathcal{O}}(\sqrt{|\mathcal{S}||\mathcal{A}|H^3T})$ have been shown (see, e.g., Jin et al., 2018). In the linear setting, the transition probability model P_h is assumed to be representable using a linear feature mapping (Jin et al., 2020):

$$P_h(\cdot|s, a) = \boldsymbol{\theta}_h(\cdot)\boldsymbol{\phi}(s, a), \quad \forall s \in \mathcal{S}, a \in \mathcal{A},$$

where $\boldsymbol{\theta}_h(\cdot) \in \mathbb{R}^d$ are unknown measures over $\mathcal{S}$, and $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a d-dimensional feature map. This representation enables the use of linear function approximation for the expected value function $[P_h V]$ and facilitates the design of a policy based on an optimistic modification of

Least-Squares Value Iteration (LSVI) that achieves a regret bound of $\tilde{\mathcal{O}}(\sqrt{d^3 H^3 T})$, scaling with the dimension d of the feature map ϕ, rather than the size of $\mathcal{S}$ and $\mathcal{A}$.

A natural extension of the linear model is the kernel-based model, which is a linear model in the (possibly infinite-dimensional) feature space of a positive definite kernel. This approach provides rich representation capacity for nonlinear function approximation using kernel-ridge regression. An intuitive explanation of this approach can be provided by using Mercer's theorem. Let $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a known positive definite kernel. By Mercer's theorem, k can be expressed as $k(z, z') = \sum_{m=1}^{\infty} \lambda_m \varphi_m(z) \varphi_m(z')$, where $\lambda_m > 0$ and $\varphi_m : \mathcal{Z} \to \mathbb{R}$ are the eigenvalues and eigenfeatures corresponding to k, respectively. Additionally, the space defined by these functions is referred to as the reproducing kernel Hilbert space (RKHS) corresponding to k:

$$\mathcal{H}_k = \left\{ f : f = \sum_{m=1}^{\infty} \theta_m \lambda_m^{\frac{1}{2}} \varphi_m(z), \ \|\boldsymbol{\theta}\| < \infty \right\},$$

where $\|\boldsymbol{\theta}\|$ is the ℓ_2 norm of the weights θ_m. The RKHS norm of f is defined as $\|f\|_{\mathcal{H}_k} = \|\boldsymbol{\theta}\|$. I.e., $\mathcal{H}_k$ is the class of linear functions in the feature space of $\{\lambda_m^{\frac{1}{2}} \varphi_m\}_{m=1}^{\infty}$.

Analogous to the linear model, the following assumption is made for kernel-based models, with the linear model as a special case when using a linear kernel.

Assumption 1 *Let $\mathcal{H}_k$ be the RKHS corresponding to a positive definite kernel $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. We assume, $\forall h \in [H], \forall s' \in \mathcal{S}, P_h(s'|\cdot) \in \mathcal{H}_k$ and $\|P_h(s'|\cdot)\|_{\mathcal{H}_k} \leq u$, for some constant $u > 0$.*

The target function of interest here is $f = [PV]$ for some unknown transition probability distribution P satisfying Assumption 1, and some value function $V : \mathcal{S} \to \mathbb{R}$. For a dataset of n transitions $\{(z_i, s'_i)\}_{i=1}^{n}$, where $s'_i \sim P(\cdot|z_i)$, we have the following prediction and uncertainty estimate, respectively, for f, utilizing kernel-ridge regression:

$$\hat{f}_n(z) = \boldsymbol{k}_n^{\top}(z)(\boldsymbol{K}_n + \rho I)^{-1} \boldsymbol{v}_n, \quad \text{and} \quad \sigma_n^2(z) = \boldsymbol{k}_n^{\top}(z)(\boldsymbol{K}_n + \rho I)^{-1} \boldsymbol{k}_n(z), \tag{3}$$

where $\boldsymbol{k}_n = [k(z, z_1), k(z, z_2), \cdots, k(z, z_n)]$, $\boldsymbol{K}_n = [k(z_i, z_j)]_{i,j=1}^{n}$ is the kernel matrix, $\rho > 0$ is a free parameter, I is the identity matrix of dimensions n, and $\boldsymbol{v}_n = [V(s'_1), V(s'_2), \cdots, V(s'_n)]$ is the vector of observations. Note that $\mathbb{E}_{s' \sim P(\cdot|z)}[V(s')] = f(z)$ by definition; thus, $V(s'_i)$ are noisy realizations of $f(z_i)$.

4. Open Problem

The question we ask is as follows: *Consider the episodic MDP setting described in Section 2. Under Assumption 1, (a) Can a no-regret learning algorithm be designed? (b) What is the minimum regret growth rate with T (and also H)? And, can a learning algorithm be designed to achieve order-optimal (or near-optimal) regret performance, closely aligning with the established lower bound?*

5. Existing Results and Challenges

Recall the kernel ridge prediction $\hat{f}_n$ and the uncertainty estimate σ_n given in Equation (3). Confidence intervals based on these statistics are important building blocks in the analysis of function approximation in RL using kernel-based models. For a fixed $f \in \mathcal{H}_k$, with $\|f\|_{\mathcal{H}_k} \leq c_f$,

for some constant $c_f > 0$, assuming σ-sub-Gaussian observation noise, several confidence intervals of the form $|f(z) - \hat{f}_n(z)| \leq \beta_n(\delta)\sigma_n(s)$, at a $1 - \delta$ confidence level are established, where the confidence interval width multiplier $\beta_n(\delta)$ depends on δ, n, σ, and the kernel k (Abbasi-Yadkori, 2013; Chowdhury and Gopalan, 2017; Vakili et al., 2021; Whitehouse et al., 2024). In particular, in an offline setting where the observation points z_i are predetermined and independent of the observation noise, the confidence interval for a fixed $z \in \mathcal{Z}$ is given by $\beta_n(\delta) = c_f + \frac{\sigma}{\sqrt{\rho}}\sqrt{2\log(\frac{1}{\delta})}$ (Vakili et al., 2021). In an online setting, such as RL and bandits, where the observation points are adaptively selected based on prior observations, the application of self-normalized confidence bounds for vector-valued martingales (Abbasi-Yadkori, 2013) to the kernel setting results in a confidence interval with $\beta_n(\delta) = c_f + \frac{\sigma}{\sqrt{\rho}}\sqrt{2\log(\frac{1}{\delta}) + \gamma(n)}$ (Whitehouse et al., 2024), where $\gamma(n) = \sup_{z_1,\cdots,z_n \in \mathcal{Z}} \frac{1}{2}\log\det(I + \rho^{-1}\boldsymbol{K}_n)$ represents the maximum information gain, a kernel specific complexity term that intuitively reflects the *effective dimension* of the kernel model.

In the RL setting, function approximation is typically applied to $f_{h,n} = [P_h V_{h,n}]$, with proxies $V_{h,n} : \mathcal{S} \to \mathbb{R}$ for the value function—usually an upper confidence bound on the value function. Due to the Markovian nature of the temporal dynamics, these functions are not prefixed, making the previously mentioned confidence intervals inapplicable. This necessitates further considerations, leading to an increased $\beta_n(\delta)$. We here overview the most important approaches addressing this issue and the corresponding eventual regret bounds.

In Yang et al. (2020), the authors adopt a rigorous approach, considering the class of all proxy value functions that appear throughout their optimistic LSVI algorithm:

$$\mathcal{V} = \{V : \forall (s,a) \in \mathcal{Z}, V(s) = \max_{a \in \mathcal{A}} \min\{H, Q(s,a)\}, Q(z) = Q_0(z) + b\sigma_n(z)\},$$

where $\|Q_0\|_{\mathcal{H}_k} \leq c$, for some $c > 0$, and $b \in [0, B]$, Q_0 represents possible predictions and σ_n represents possible uncertainty estimates derived from a set of n observations using kernel ridge regression. They proceed by bounding the ϵ-covering number of $\mathcal{V}$. Specifically, they establish a bound on $\mathcal{N}(\mathcal{V}, \epsilon)$, the minimum number of functions needed to cover $\mathcal{V}$ up to an ϵ error in ℓ_∞ norm. Using this technique, they derive a confidence interval that is applicable to all $V \in \mathcal{V}$ and the corresponding $f = [PV]$, with $\beta_n(\delta) = \mathcal{O}(c_f + \sqrt{\log(\frac{1}{\delta}) + \gamma(n) + \log\mathcal{N}(\mathcal{V}, \frac{1}{n})})$ for a choice of $\epsilon = \mathcal{O}(\frac{1}{n})$. This leads to a regret bound of $\tilde{\mathcal{O}}(\beta_T(\delta)H^2\sqrt{T\gamma(T)})$, which does not ensure no-regret performance when $\beta_T(\delta)$ grows at least as fast as $\sqrt{T}$, a case common with several kernels of interest, including the Metérn and neural tangent kernels.

Chowdhury and Oliveira (2023) make a strong assumption, referred to as *optimistic closure*, that $\forall V \in \mathcal{V}, \|V\|_{\mathcal{H}_k} \leq c_v$, for some $c_v > 0$. Utilizing kernel mean embedding, they prove a tighter confidence interval with $\beta_n(\delta) = \tilde{\mathcal{O}}(c_f + \frac{\sigma}{\sqrt{\rho}}\sqrt{\log(\frac{1}{\delta}) + \gamma(n)})$. This improved bound potentially leads to no-regret guarantees using techniques from bandits (see, Whitehouse et al., 2024). However, the optimistic closure assumption seems unrealistic. One potential solution to rigorously obtain no-regret guarantees involves relaxing the optimistic closure assumption in their work.

A related problem is kernel bandits (also known as Bayesian optimization), a spacial case with $|\mathcal{S}| = 1, H = 1$, where no-regret guarantees are established for Upper Confidence Bound type algorithms (Whitehouse et al., 2024), and more sophisticated algorithms that perform sample or domain partitioning have been shown to achieve order-optimal regret bounds (see, e.g., Valko et al., 2013; Salgia et al., 2021; Li and Scarlett, 2022). While the ideas in these works are insightful, their applicability to general MDPs is unclear.

References

Yasin Abbasi-Yadkori. Online learning for linearly parametrized control problems. *PhD Thesis, University of Alberta*, 2013.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

Sayak Ray Chowdhury and Rafael Oliveira. Value function approximations via kernel embeddings for no-regret reinforcement learning. In *Asian Conference on Machine Learning*, pages 249–264. PMLR, 2023.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Zihan Li and Jonathan Scarlett. Gaussian process bandit optimization with few batches. In *International Conference on Artificial Intelligence and Statistics*, 2022.

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

Sudeep Salgia, Sattar Vakili, and Qing Zhao. A domain-shrinking based Bayesian optimization algorithm with order-optimal regret performance. *Conference on Neural Information Processing Systems*, 34, 2021.

Sattar Vakili, Nacime Bouziani, Sepehr Jalali, Alberto Bernacchia, and Da-shan Shiu. Optimal order simple regret for Gaussian process bandits. *Advances in Neural Information Processing Systems*, 34:21202–21215, 2021.

Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time analysis of kernelised contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, page 654–663, 2013.

Justin Whitehouse, Aaditya Ramdas, and Steven Z Wu. On the sublinear regret of GP-UCB. *Advances in Neural Information Processing Systems*, 36, 2024.

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*, 33:13903–13916, 2020.