Online Learning in Kernelized Markov Decision Processes

Sayak Ray Chowdhury
Indian Institute of Science

Aditya Gopalan
Indian Institute of Science

Abstract

We consider online learning for minimizing regret in unknown, episodic Markov decision processes (MDPs) with continuous states and actions. We develop variants of the UCRL and posterior sampling algorithms that employ nonparametric Gaussian process priors to generalize across the state and action spaces. When the transition and reward functions of the true MDP are members of the associated Reproducing Kernel Hilbert Spaces of functions induced by symmetric psd kernels, we show that the algorithms enjoy sublinear regret bounds. The bounds are in terms of explicit structural parameters of the kernels, namely a novel generalization of the information gain metric from kernelized bandit, and highlight the influence of transition and reward function structure on the learning performance. Our results are applicable to multi-dimensional state and action spaces with composite kernel structures, and generalize results from the literature on kernelized bandits, and the adaptive control of parametric linear dynamical systems with quadratic costs.

1 INTRODUCTION

The goal of reinforcement learning (RL) is to learn optimal behavior by repeated interaction with an unknown environment, usually modelled as a Markov Decision Process (MDP). Performance is typically measured by the amount of interaction, in terms of episodes or rounds, needed to arrive at an optimal (or near-optimal) policy; this is also known as the sample complexity of RL [Strehl et al., 2009]. The sample complexity objective encourages efficient exploration across states and actions, but, at the same time, is indifferent to the reward earned during the learning phase.

A related, but different, goal in RL is the *online* one, i.e., to learn to gather high cumulative reward, or to equivalently keep the learner's *regret* (the gap between its and the optimal policy's net reward) as low as possible. This is preferable in settings where experimentation comes at a premium and the reward earned in each round is of direct value, e.g., recommender systems (in which rewards correspond to clickthrough events and ultimately translate to revenue), dynamic pricing – in general, control of unknown dynamical systems with instantaneous costs.

A primary challenge in RL is to learn efficiently across complex (very large or infinite) state and action spaces. In the most general *tabula rasa* MDP setting, the learner must explore each state-action transition before developing a reasonably clear understanding of the environment, which is prohibitive for large problems. Real-world domains, though, possess more structure: transition and reward behavior often varies smoothly over states and actions, making it possible to generalize via inductive inference – observing a state transition or reward is informative of other, similar transitions or rewards. Scaling RL to large, complex, real-world domains requires exploiting regularity structure in the environment, which has typically been carried out via the use of parametric MDP models in model-based approaches, e.g., Osband and Van Roy [2014].

This paper takes a step in developing theory and algorithms for online RL in environments with smooth transition and reward structure. We specifically consider the episodic online learning problem in the nonparametric, kernelizable MDP setting, i.e., of minimizing regret (relative to an optimal finite-horizon policy) in MDPs with continuous state and action spaces, whose transition and reward functions exhibit smoothness over states and actions compatible with the structure of a reproducing kernel. We develop variants of the well-known UCRL and posterior sampling algorithms for MDPs with continuous state and action spaces, and show that they enjoy sublinear, finite-time regret bounds when the mean transition and reward functions are assumed to belong to the associated Reproducing Kernel Hilbert Space (RKHS) of functions.

Our results bound the regret of the algorithms in terms of a novel generalization of the information gain of the state transition and reward function kernels, from the memoryless kernel bandit setting [Srinivas et al., 2009] to the

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

state-based kernel MDP setting, and help shed light on how the choice of kernel model influences regret performance. We also leverage recent concentration of measure results for RKHS-valued martingales, developed originally for the kernelized bandit setting [Chowdhury and Gopalan, 2017, Durand et al., 2017], to prove the results in the paper. To the best of our knowledge, these are the first concrete regret bounds for RL in the kernelizable setting, explicitly showing the dependence of regret on kernel structure.

Our results represent a generalisation of several streams of work. We generalise online learning in the kernelized bandit setting [Valko et al., 2013, Chowdhury and Gopalan, 2017] to kernelized MDPs, and *tabula rasa* online learning approaches for MDPs such as Upper Confidence Bound for Reinforcement Learning (UCRL) [Jaksch et al., 2010] and Posterior Sampling for Reinforcement Learning (PSRL) [Osband et al., 2013, Ouyang et al., 2017] to MDPs with kernel structure. We also generalize regret minimization for an episodic variant of the well-known parametric Linear Quadratic Regulator (LQR) problem [Abbasi-Yadkori and Szepesvári, 2011, 2015, Ibrahimi et al., 2012, Abeille and Lazaric, 2017] to its nonlinear, nonparametric, infinite-dimensional, kernelizable counterpart.

Overview of Main Results

Our first main result gives an algorithm for learning MDPs with mean transition dynamics and reward structure assumed to belong to appropriate Reproducing Kernel Hilbert Spaces (RKHSs). This result is, to our knowledge, the first frequentist regret guarantee for general kernelized MDPs.

Result 1 (Frequentist regret in kernelized MDPs, informal) *Consider episodic learning under the unknown dynamics $s_{t+1} = \overline{P}_M(s_t, a_t) + \text{Noise} \in \mathbb{R}^m$, and rewards $r_t = \overline{R}_M(s_t, a_t) + \text{Noise}$, where $\overline{P}_M$ and $\overline{R}_M$ are fixed RKHS functions with bounded norms. The regret of GP-UCRL (Algorithm 1) is, with high probability*[1]*, $\tilde{O}\Big(\big(\gamma_T(R) + \gamma_{mT}(P)\big)\sqrt{T}\Big)$.*

Here, $\gamma_t(P)$ (resp. $\gamma_t(R)$) roughly represents the maximum information gain about the unknown dynamics (resp. rewards) after t rounds, which, for instance is polylog(t) for the squared exponential kernel.

To put this in the perspective of existing work, Osband and Van Roy [2014] also consider learning under dynamics and rewards coming from general function classes, and show (Bayesian) regret bounds depending on the eluder dimensions of the classes. However, when applied to RKHS function classes as we consider here, these dimensions can be infinitely large. In contrast, our results show that the maximum information gain is a suitable measure of complexity of the function class that serves to bound regret.

[1] $\tilde{O}$ suppresses logarithmic factors.

An important corollary results when this is applied to the LQR problem, with a linear kernel structure for state transitions and a quadratic kernel structure for rewards:

Result 2 (Frequentist regret for LQR, informal) *Consider episodic learning under unknown linear dynamics $s_{t+1} = As_t + Ba_t + \text{Noise}$, and quadratic rewards $r_t = s_t^T P s_t + a_t^T Q a_t + \text{Noise}$. GP-UCRL (Algorithm 1) instantiated with a linear transition kernel and quadratic reward kernel enjoys, with high probability, regret $\tilde{O}\Big(\big(m^2+n^2+m(m+n)\big)\sqrt{T}\Big)$, where m and n are the state space and action space dimensions, respectively.*

This recovers the bound of Osband and Van Roy [2014] for the same bounded LQR problem. However, while they derive this via the eluder dimension approach, we arrive at this by a different bounding technique that applies more generally to any kernelized dynamics. The result also matches (order-wise) the bound of Abbasi-Yadkori and Szepesvári [2011] restricted to the bounded LQR problem.

We also have the following Bayesian regret analogue for PSRL.

Result 3 (Bayesian regret in kernelized MDPs, informal) *Under dynamics as in Result 1 but drawn according to a known prior, the Bayes regret of PSRL (Algorithm 2) is $\tilde{O}\Big(\big(\gamma_T(R) + \gamma_{mT}(P)\big)\sqrt{T}\Big)$. Consequently, if the dynamics are of the LQR form (Result 2), then PSRL instantiated with a linear transition kernel and quadratic reward kernel enjoys Bayes regret $\tilde{O}\Big(\big(m^2+n^2+m(m+n)\big)\sqrt{T}\Big)$.*

Note: All the above results are stated assuming that the episode duration $H = O(\ln T)$ for clarity; the explicit dependence on H can be found in the theorem statements that follow.

Related Work Regret minimization has been studied with parametric MDPs [Jaksch et al., 2010, Osband et al., 2013, Gopalan and Mannor, 2015, Agrawal and Jia, 2017]. For online regret minimization in complex MDPs, apart from the work of Osband and Van Roy [2014], Ortner and Ryabko [2012] and Lakshmanan et al. [2015] consider continuous state spaces with Lipschitz transition dynamics but unstructured, finite action spaces. Another important line of work considers kernel structures for safe exploration in MDPs [Turchetta et al., 2016, Berkenkamp et al., 2017]. We, however, seek to demonstrate algorithms with provable regret guarantees in the kernelized MDP setting, which to our knowledge are the first of their kind.

2 PROBLEM STATEMENT

We consider the problem of learning to optimize reward in an unknown finite-horizon MDP, $M_\star = \{\mathcal{S}, \mathcal{A}, R_\star, P_\star, H\}$, over repeated episodes of interaction.

Here, $\mathcal{S} \subset \mathbb{R}^m$ represents the state space, $\mathcal{A} \subset \mathbb{R}^n$ the action space, H the episode length, $R_\star(s,a)$ the reward distribution over $\mathbb{R}$, and $P_\star(s,a)$ the transition distribution over $\mathcal{S}$. At each period $h = 1, 2, \ldots, H$ within an episode, an agent observes a state $s_h \in \mathcal{S}$, takes an action $a_h \in \mathcal{A}$, observes a reward $r_h \sim R_\star(s_h, a_h)$, and causes the MDP to transition to a next state $s_{h+1} \sim P_\star(s_h, a_h)$. We assume that the agent, while not possessing knowledge of the reward and transition distribution $R_\star, P_\star$ of the unknown MDP $M_\star$, knows $\mathcal{S}$, $\mathcal{A}$ and H.

A policy $\pi : \mathcal{S} \times \{1, 2, \ldots, H\} \to \mathcal{A}$ is defined to be a mapping from a state $s \in \mathcal{S}$ and a period $1 \le h \le H$ to an action $a \in \mathcal{A}$. For any MDP $M = \{\mathcal{S}, \mathcal{A}, R_M, P_M, H\}$ and policy π, the finite horizon, undiscounted, value function for every state $s \in \mathcal{S}$ and every period $1 \le h \le H$ is defined as $V^M_{\pi,h}(s) := \mathbb{E}_{M,\pi}\Big[\sum_{j=h}^{H} \overline{R}_M(s_j, a_j) \;\Big|\; s_h = s\Big]$, where the subscript π indicates the application of the learning policy π, i.e., $a_j = \pi(s_j, j)$, and the subscript M explicitly references the MDP environment M, i.e., $s_{j+1} \sim P_M(s_j, a_j)$, for all $j = h, \ldots, H$.

We use $\overline{R}_M(s,a) = \mathbb{E}\left[r \;\middle|\; r \sim R_M(s,a)\right]$ to denote the mean of the reward distribution $R_M(s,a)$ that corresponds to playing action a at state s in the MDP M. We can view a sample r from the reward distribution $R_M(s,a)$ as $r = \overline{R}_M(s,a) + \varepsilon_R$, where ε_R denotes a sample of zero-mean, real-valued additive noise. Similarly, the transition distribution $P_M(s,a)$ can also be decomposed as a mean value $\overline{P}_M(s,a)$ in $\mathbb{R}^m$ plus a zero-mean additive noise ε_P in $\mathbb{R}^m$ so that $s' = \overline{P}_M(s,a) + \varepsilon_P$ lies in[2] $\mathcal{S} \subset \mathbb{R}^m$. A policy π_M is said to be optimal for the MDP M if $V^M_{\pi_M,h}(s) = \max_\pi V^M_{\pi,h}(s)$ for all $s \in \mathcal{S}$ and $h = 1, \ldots, H$.

For an MDP M, a distribution φ over $\mathcal{S}$ and period $1 \le h \le H$, we define the one step future value function as the expected value of the optimal policy π_M, with the next state distributed according to φ, i.e. $U^M_h(\varphi) := \mathbb{E}_{s' \sim \varphi}\Big[V^M_{\pi_M,h+1}(s')\Big]$. We assume the following regularity condition on the future value function of any MDP (also made by Osband and Van Roy [2014]).

Assumption (A1) For any two single-step transition distributions φ_1, φ_2 over $\mathcal{S}$, and $1 \le h \le H$,

$$\left|U^M_h(\varphi_1) - U^M_h(\varphi_2)\right| \le L_M \left\|\overline{\varphi}_1 - \overline{\varphi}_2\right\|_2, \qquad (1)$$

where $\overline{\varphi} := \mathbb{E}_{s' \sim \varphi}[s'] \in \mathcal{S}$ denotes the mean of the distribution φ. In other words, the one-step future value functions for each period h are Lipschitz continuous with respect to the $\|\cdot\|_2$-norm of the mean[3], with global Lipschitz constant L_M. We also assume that there is a known constant L such that $L_\star := L_{M_\star} \le L$.

Regret At the beginning of each episode l, an RL algorithm chooses a policy π_l depending upon the observed state-action-reward sequences upto episode $l-1$, denoted by the history $\mathcal{H}_{l-1} := \{s_{j,k}, a_{j,k}, r_{j,k}, s_{j,k+1}\}_{1 \le j \le l-1, 1 \le k \le H}$, and executes it for the entire duration of the episode. In other words, at each period h of the l-th episode, the learning algorithm chooses action $a_{l,h} = \pi_l(s_{l,h}, h)$, receives reward $r_{l,h} = \overline{R}_\star(s_{l,h}, a_{l,h}) + \varepsilon_{R,l,h}$ and observes the next state $s_{l,h+1} = \overline{P}_\star(s_{l,h}, a_{l,h}) + \varepsilon_{P,l,h}$. The goal of an episodic online RL algorithm is to maximize its cumulative reward across episodes, or, equivalently, minimize its cumulative *regret*: the loss incurred in terms of the value function due to not knowing the optimal policy $\pi_\star := \pi_{M_\star}$ of the unknown MDP $M_\star$ beforehand and instead using the policy π_l for each episode l, $l = 1, 2, \ldots$. The cumulative (expected) regret of an RL algorithm $\pi = \{\pi_1, \pi_2, \ldots\}$ upto time horizon $T = \tau H$ is defined as $\text{Regret}(T) = \sum_{l=1}^{\tau}\Big[V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\Big]$, where the initial states $s_{l,1}, l \ge 1$ are assumed to be fixed.

Notations For the rest of the paper, unless otherwise specified, we define $z := (s,a)$, $z' := (s',a')$ and $z_{l,h} := (s_{l,h}, a_{l,h})$ for all $l \ge 1$ and $1 \le h \le H$.

3 ALGORITHMS

3.1 Representing Uncertainty

The algorithms we design represent uncertainty in the reward and transition distribution $R_\star, P_\star$ by maintaining Gaussian process (GP) priors over the mean reward function $\overline{R}_\star : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and the mean transition function $\overline{P}_\star : \mathcal{S} \times \mathcal{A} \times \{1, \ldots, m\} \to \mathbb{R}$ of the unknown MDP $M_\star$. (We denote $\overline{P}_\star(s,a) := [\overline{P}_\star(s,a,1) \ldots \overline{P}_\star(s,a,m)]^T$.) A Gaussian Process over $\mathcal{X}$, denoted by $GP_{\mathcal{X}}(\mu(\cdot), k(\cdot,\cdot))$, is a collection of random variables $(f(x))_{x \in \mathcal{X}}$, one for each $x \in \mathcal{X}$, such that every finite sub-collection of random variables $(f(x_i))_{i=1}^m$ is jointly Gaussian with mean $\mathbb{E}[f(x_i)] = \mu(x_i)$ and covariance $\mathbb{E}[(f(x_i) - \mu(x_i))(f(x_j) - \mu(x_j))] = k(x_i, x_j)$, $1 \le i, j \le m$, $m \in \mathbb{N}$. We use $GP_{\mathcal{Z}}(0, k_R)$ and $GP_{\tilde{\mathcal{Z}}}(0, k_P)$ as the initial prior distributions over $\overline{R}_\star$ and $\overline{P}_\star$, with positive semi-definite covariance (kernel) functions k_R and k_P respectively, where $\mathcal{Z} := \mathcal{S} \times \mathcal{A}$, $\tilde{\mathcal{Z}} := \mathcal{Z} \times \{1, \ldots, m\}$. We also assume that the noise variables $\varepsilon_{R,l,h}$ and $\varepsilon_{P,l,h}$ are drawn independently, across l and h, from $\mathcal{N}(0, \lambda_R)$ and $\mathcal{N}(0, \lambda_P I)$ respectively, with $\lambda_R, \lambda_P \ge 0$. Then, by standard properties of GPs [Rasmussen and Williams, 2006], conditioned on the history $\mathcal{H}_l$, the posterior distribution over $\overline{R}_\star$ is also a Gaussian process, $GP_{\mathcal{Z}}(\mu_{R,l}, k_{R,l})$, with mean and kernel functions

$$\begin{aligned} \mu_{R,l}(z) &= k_{R,l}(z)^T (K_{R,l} + \lambda_R I)^{-1} R_l, \\ k_{R,l}(z,z') &= k_R(z,z') - k_{R,l}(z)^T (K_{R,l} + \lambda_R I)^{-1} k_{R,l}(z'), \\ \sigma^2_{R,l}(z) &= k_{R,l}(z,z). \end{aligned} \qquad (2)$$

[2] Osband and Van Roy [2014] argue that the assumption $\mathcal{S} \subset \mathbb{R}^m$ is not restrictive for most practical settings.

[3] Assumption (1) is essentially equivalent to assuming knowledge of the centered state transition noise distributions, since it implies that any two transition distributions with the same means are identical.

Here $R_l := [r_{1,1}, \dots, r_{l,H}]^T$ denotes the vector of rewards observed at $\mathcal{Z}_l := \{z_{j,k}\}_{1 \le j \le l, 1 \le k \le H} = \{z_{1,1}, \dots, z_{l,H}\}$, the set of all state-action pairs available at the end of episode l. $k_{R,l}(z) := [k_R(z_{1,1}, z), \dots, k_R(z_{l,H}, z)]^T$ denotes the vector of kernel evaluations between z and elements of the set $\mathcal{Z}_l$ and $K_{R,l} := [k_R(u,v)]_{u,v \in \mathcal{Z}_l}$ denotes the kernel matrix computed at $\mathcal{Z}_l$.

Similarly, conditioned on $\mathcal{H}_l$, the posterior distribution over $\overline{P}_\star$ is $GP_{\tilde{\mathcal{Z}}}(\mu_{P,l}, k_{P,l})$ with mean and kernel functions

$$\begin{aligned} \mu_{P,l}(z,i) &= k_{P,l}(z,i)^T (K_{P,l} + \lambda_P I)^{-1} S_l, \\ k_{P,l}\big((z,i),(z',j)\big) &= k_P\big((z,i),(z,i)\big) \\ &\quad - k_{P,l}(z,i)^T (K_{P,l} + \lambda_P I)^{-1} k_{P,l}(z',j), \\ \sigma^2_{P,l}(z,i) &= k_{P,l}\big((z,i),(z,i)\big). \end{aligned} \tag{3}$$

Here $S_l := [s_{1,2}^T, \dots, s_{l,H+1}^T]^T$ denotes the vector of state transitions at $\mathcal{Z}_l = \{z_{1,1}, \dots, z_{l,H}\}$, the set of all state-action pairs available at the end of episode l. $k_{P,l}(z,i) := \Big[k_P\big((z_{1,1},1),(z,i)\big), \dots, k_P\big((z_{l,H},m),(z,i)\big)\Big]^T$ denotes the vector of kernel evaluations between (z,i) and elements of the set $\tilde{\mathcal{Z}}_l := \big\{(z_{j,k},i)\big\}_{1 \le j \le l, 1 \le k \le H, 1 \le i \le m} = \big\{(z_{1,1},1), \dots, (z_{l,H},m)\big\}$ and $K_{P,l} := \big[k_P(u,v)\big]_{u,v \in \tilde{\mathcal{Z}}_l}$ denotes the kernel matrix computed at $\tilde{\mathcal{Z}}_l$.

Thus, at the end of episode l, conditioned on the history $\mathcal{H}_l$, the posterior distributions over $\overline{R}_\star(z)$ and $\overline{P}_\star(z,i)$ is updated and maintained as $\mathcal{N}\big(\mu_{R,l}(z), \sigma^2_{R,l}(z)\big)$ and $\mathcal{N}\big(\mu_{P,l}(z,i), \sigma^2_{P,l}(z,i)\big)$ respectively, for every $z \in \mathcal{Z}$ and $i = 1, \dots, m$. This representation not only permits generalization via inductive inference across state and action spaces, but also allows for tractable updates. We now present our online algorithms GP-UCRL and PSRL for kernelized MDPs.

3.2 GP-UCRL Algorithm

GP-UCRL (Algorithm 1) is an optimistic algorithm based on the Upper Confidence Bound principle, which adapts the confidence sets of UCRL2 [Jaksch et al., 2010] to exploit the kernel structure. At the start of every episode l, GP-UCRL constructs confidence sets $\mathcal{C}_{R,l}$ and $\mathcal{C}_{P,l}$ for the mean reward function and transition function, respectively, using the parameters of GP posteriors as given in Section 3.1. The exact forms of the confidence sets appear in the theoretical result later, e.g., (8) and (9). It then builds the set $\mathcal{M}_l$ of all plausible MDPs M with the mean reward function $\overline{R}_M \in \mathcal{C}_{R,l}$, the mean transition function $\overline{P}_M \in \mathcal{C}_{P,l}$ and the global Lipschitz constant (1) L_M of future value functions upper bounded by a known constant L, where $L_\star \le L$. It then selects an optimistic policy π_l for the family of MDPs $\mathcal{M}_l$ in the sense that $V^{M_l}_{\pi_l,1}(s_{l,1}) = \max_\pi \max_{M \in \mathcal{M}_l} V^M_{\pi,1}(s_{l,1})$, where $s_{l,1}$ is the initial state and M_l is the most optimistic realization from $\mathcal{M}_l$, and executes π_l for the entire episode. The pseudo-code of GP-UCRL is given in Algorithm 1. Even though GP-UCRL is described using the language of GP priors/posteriors, it can also be understood as kernelized regression with appropriately designed confidence sets.

Algorithm 1 GP-UCRL

Input: Kernel functions k_R and k_P.
Set $\mu_{R,0}(z) = \mu_{P,0}(z,i) = 0$, $\sigma^2_{R,0}(z) = k_R(z,z)$, $\sigma^2_{P,0}(z,i) = k_P\big((z,i),(z,i)\big) \forall z \in \mathcal{Z}, \forall i = 1, \dots, m$.
for episode $l = 1, 2, 3, \dots$ **do**
  Construct confidence sets $\mathcal{C}_{R,l}$ and $\mathcal{C}_{P,l}$.
  Construct the set of all plausible MDPs $\mathcal{M}_l = \{M : L_M \le L, \overline{R}_M \in \mathcal{C}_{R,l}, \overline{P}_M \in \mathcal{C}_{P,l}\}$.
  Choose policy π_l such that $V^{M_l}_{\pi_l,1}(s_{l,1}) = \max_\pi \max_{M \in \mathcal{M}_l} V^M_{\pi,1}(s_{l,1})$.

  for period $h = 1, 2, 3, \dots, H$ **do**
    Choose action $a_{l,h} = \pi_l(s_{l,h}, h)$.
    Observe reward $r_{l,h} = \overline{R}_\star(z_{l,h}) + \varepsilon_{R,l,h}$.
    Observe next state $s_{l,h+1} = \overline{P}_\star(z_{l,h}) + \varepsilon_{P,l,h}$.
  end for
  Update $\mu_{R,l}$, $\sigma_{R,l}$ using (2) and $\mu_{P,l}$, $\sigma_{P,l}$ using (3).
end for

Optimizing for an optimistic policy is not computationally tractable in general, even though planning for the optimal policy is possible for a given MDP. A popular approach to overcome this difficulty is to sample a random MDP at every episode and solve for its optimal policy, called posterior sampling [Osband and Van Roy, 2016].

3.3 PSRL Algorithm

PSRL (Algorithm 2), in its most general form, starts with a prior distribution $\Phi \equiv (\Phi_R, \Phi_P)$ over MDPs, where Φ_R and Φ_P are priors over reward and transition distributions respectively. At the beginning of episode l, using the history of observations $\mathcal{H}_{l-1}$, it updates the posterior $\Phi_l \equiv (\Phi_{R,l}, \Phi_{P,l})$ and samples an MDP M_l from it [4] ($\Phi_{R,l}$ and $\Phi_{P,l}$ are posteriors over reward and transition distributions respectively). It then selects an optimal policy π_l of the sampled MDP M_l, in the sense that $V^{M_l}_{\pi_l,h}(s) = \max_\pi V^{M_l}_{\pi,h}(s)$ for all $s \in \mathcal{S}$ and for all $h = 1, 2, \dots, H$, and executes π_l for the entire episode.

For example, if Φ_R and Φ_P are specified by GPs $GP_{\mathcal{Z}}(0, k_R)$ and $GP_{\tilde{\mathcal{Z}}}(0, k_P)$ respectively with Gaussian observation model, then the posteriors $\Phi_{R,l}$ and $\Phi_{P,l}$ are given by GP posteriors as discussed in Section 3.1. Here at every episode l, PSRL samples an MDP M_l with mean reward function $\overline{R}_{M_l} \sim GP_{\mathcal{Z}}(\mu_{R,l-1}, k_{R,l-1})$ and mean transition function $\overline{P}_{M_l} \sim GP_{\tilde{\mathcal{Z}}}(\mu_{P,l-1}, k_{P,l-1})$.

[4] Sampling can be done using MCMC methods even if Φ_l doesn't admit any closed form.

Algorithm 2 PSRL

Input: Prior Φ.
Set $\Phi_1 = \Phi$.
for episode $l = 1, 2, 3, \ldots$ **do**
  Sample $M_l \sim \Phi_l$.
  Choose policy π_l such that $V^{M_l}_{\pi_l,h}(s) = \max_\pi V^{M_l}_{\pi,h}(s)\ \forall s \in \mathcal{S}, \forall h = 1, 2, \ldots, H$.

  for period $h = 1, 2, 3, \ldots, H$ **do**
    Choose action $a_{l,h} = \pi_l(s_{l,h}, h)$.
    Observe reward $r_{l,h} = \overline{R}_\star(z_{l,h}) + \varepsilon_{R,l,h}$.
    Observe next state $s_{l,h+1} = \overline{P}_\star(z_{l,h}) + \varepsilon_{P,l,h}$.
  end for
  Update Φ_l to Φ_{l+1}, using $\{s_{l,h}, a_{l,h}, s_{l,h+1}\}_{1\le h\le H}$.
end for

Computational issues Optimal planning may be computationally intractable even for a given MDP, so it is common in the literature to assume access to an approximate MDP planner $\Gamma(M, \varepsilon)$ which returns an ε-optimal policy for M. Given such a planner Γ, if it is possible to obtain (through extended value iteration [Jaksch et al., 2010] or otherwise) an efficient planner $\tilde{\Gamma}(\mathcal{M}, \varepsilon)$ which returns an ε-optimal policy for the most optimistic MDP from a family $\mathcal{M}$, then we modify PSRL and GP-UCRL to choose $\pi_l = \Gamma(M_l, \sqrt{H/l})$ and $\pi_l = \tilde{\Gamma}(\mathcal{M}_l, \sqrt{H/l})$ respectively at every episode l. It follows that this adds only an $O(\sqrt{T})$ factor in the respective regret bounds. The design of such approximate planners for continuous state and action spaces remains a subject of active research, whereas our focus in this work is on the statistical efficiency of the online learning problem.

4 MAIN RESULTS

In this section, we provide our main theoretical upper bounds on the cumulative regret. All the proofs are deferred to the appendix for lack of space.

4.1 Preliminaries and assumptions

Maximum Information Gain (MIG) Let $f : \mathcal{X} \to \mathbb{R}$ be a (possibly random) real-valued function defined on a domain $\mathcal{X}$. For each $A \subset \mathcal{X}$, let $f_A := [f(x)]_{x\in A}$ denote a vector containing f's evaluations at each point in A and Y_A denote a noisy version of f_A obtained by passing f_A through a channel $\mathbb{P}[Y_A|f_A]$. The *Maximum Information Gain (MIG)* about f after t noisy observations is defined as $\gamma_t(f, \mathcal{X}) := \max_{A\subset\mathcal{X}:|A|=t} I(f_A; Y_A)$, where $I(X;Y)$ denotes the Shannon mutual information between two jointly distributed random variables X, Y. If $f \sim GP_{\mathcal{X}}(0, k)$ and the channel is iid Gaussian $\mathcal{N}(0, \lambda)$, then $\gamma_t(f, \mathcal{X})$ depends only on $k, \mathcal{X}, \lambda$ Srinivas et al. [2009]. But the dependency on λ is only of $\tilde{O}(1/\lambda)$ and hence in this setting we denote MIG as $\gamma_t(k, \mathcal{X})$ to indicate the dependencies on k and $\mathcal{X}$ explicitly. If $\mathcal{X} \subset \mathbb{R}^d$ is compact and convex, then $\gamma_t(k, \mathcal{X})$ is sublinear in t for different classes of kernels; e.g. for the linear kernel[5] $\gamma_t(k, \mathcal{X}) = \tilde{O}(d \ln t)$ and for the Squared Exponential (SE) kernel[6], $\gamma_t(k, \mathcal{X}) = \tilde{O}\left((\ln t)^d\right)$.

Composite kernels Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$. A composite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be constructed by using individual kernels $k_1 : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}$ and $k_2 : \mathcal{X}_2 \times \mathcal{X}_2 \to \mathbb{R}$. For instance, a *product kernel* $k = k_1 \otimes k_2$ is obtained by setting $(k_1 \otimes k_2)\big((x_1, x_2), (x_1', x_2')\big) := k_1(x_1, x_1')k_2(x_2, x_2')$. Another example is that of an *additive kernel* $k = k_1 \oplus k_2$ by setting $(k_1 \oplus k_2)\big((x_1, x_2), (x_1', x_2')\big) = k_1(x_1, x_1') + k_2(x_2, x_2')$. Krause and Ong [2011] bound the MIG for additive and product kernels in terms of the MIG for individual kernels as

$$\gamma_t(k_1 \oplus k_2, \mathcal{X}) \le \gamma_t(k_1, \mathcal{X}_1) + \gamma_t(k_2, \mathcal{X}_2) + 2\ln t, \quad (4)$$

and, if k_2 has rank at most d (i.e. all kernel matrices over any finite subset of $\mathcal{X}_2$ have rank at most d), as

$$\gamma_t(k_1 \otimes k_2, \mathcal{X}) \le d\gamma_t(k_1, \mathcal{X}_1) + d\ln t. \quad (5)$$

Therefore, if the MIGs for individual kernels are sublinear in t, then the same is true for their products and additions. For example, the MIG for the product of a d_1-dimensional linear kernel and a d_2-dimensional SE kernel is $\tilde{O}\big(d_1(\ln t)^{d_2}\big)$.

Regularity assumptions (A2) Each of our results in this section will assume that $\overline{R}_\star$ and $\overline{P}_\star$ have small norms in the Reproducing Kernel Hilbert Spaces (RKHSs) associated with kernels k_R and k_P respectively. An RKHS of real-valued functions $\mathcal{X} \to \mathbb{R}$, denoted by $\mathcal{H}_k(\mathcal{X})$, is completely specified by its kernel function $k(\cdot,\cdot)$ and vice-versa, with an inner product $\langle\cdot,\cdot\rangle_k$ obeying the reproducing property $f(x) = \langle f, k(x,\cdot)\rangle_k$ for all $f \in \mathcal{H}_k(\mathcal{X})$. The induced RKHS norm $\|f\|_k = \sqrt{\langle f, f\rangle_k}$ is a measure of smoothness of f with respect to the kernel function k. We assume known bounds on the RKHS norms of the mean reward and mean transition functions: $\overline{R}_\star \in \mathcal{H}_{k_R}(\mathcal{Z}), \|\overline{R}_\star\|_{k_R} \le B_R$ and $\overline{P}_\star \in \mathcal{H}_{k_P}(\tilde{\mathcal{Z}}), \|\overline{P}_\star\|_{k_P} \le B_P$, where $\mathcal{Z} := \mathcal{S} \times \mathcal{A}$ and $\tilde{\mathcal{Z}} := \mathcal{Z} \times \{1, \ldots, m\}$.

Noise assumptions (A3) For the purpose of this section, we assume that the noise sequence $\{\varepsilon_{R,l,h}\}_{l\ge1,1\le h\le H}$ is conditionally σ_R-sub-Gaussian, i.e., there exists a known $\sigma_R \ge 0$ such that for any $\eta \in \mathbb{R}$,

$$\mathbb{E}\left[\exp(\eta\,\varepsilon_{R,l,h}) \,\middle|\, \mathcal{F}_{R,l,h-1}\right] \le \exp\left(\eta^2\sigma_R^2/2\right), \quad (6)$$

where $\mathcal{F}_{R,l,h-1}$ is the sigma algebra generated by the random variables $\{s_{j,k}, a_{j,k}, \varepsilon_{R,j,k}\}_{1\le j\le l-1, 1\le k\le H}$, $\{s_{l,k}, a_{l,k}, \varepsilon_{R,l,k}\}_{1\le k\le h-1}$, $s_{l,h}$ and $a_{l,h}$. Similarly,

[5] $k(x, x') = x^T x'$.
[6] $k(x, x') = \exp\left(-\|x - x'\|_2^2 / 2l^2\right), l > 0$.

the noise sequence $\{\varepsilon_{P,l,h}\}_{l\geq 1, 1\leq h\leq H}$ is assumed to be conditionally component-wise independent and σ_P-sub-Gaussian, in the sense that there exists a known $\sigma_P \geq 0$ such that for any $\eta \in \mathbb{R}$ and $1 \leq i \leq m$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\eta\varepsilon_{P,l,h}(i)\right) \mid \mathcal{F}_{P,l,h-1}\right] &\leq \exp\left(\eta^2\sigma_P^2/2\right),\\ \mathbb{E}\left[\varepsilon_{P,l,h}\varepsilon_{P,l,h}^T \mid \mathcal{F}_{P,l,h-1}\right] &= I,\end{aligned} \tag{7}$$

where $\mathcal{F}_{P,l,h-1}$ is the sigma algebra generated by the random variables $\{s_{j,k}, a_{j,k}, \varepsilon_{P,j,k}\}_{1\leq j\leq l-1, 1\leq k\leq H}$, $\{s_{l,k}, a_{l,k}, \varepsilon_{P,l,k}\}_{1\leq k\leq h-1}$, $s_{l,h}$ and $a_{l,h}$.

4.2 Regret Bound for GP-UCRL in Kernelized MDPs

We run GP-UCRL (Algorithm 1) using GP priors and Gaussian noise models as given in Section 3.2. Note that, in this section, though the algorithm relies on GP priors, the setting under which it is analyzed is 'agnostic', i.e., under a *fixed* but unknown true MDP environment.

Choice of confidence sets At the beginning of each episode l, GP-UCRL constructs the confidence set $\mathcal{C}_{R,l}$ as

$$\mathcal{C}_{R,l} = \left\{f : |f(z) - \mu_{R,l-1}(z)| \leq \beta_{R,l}\sigma_{R,l-1}(z) \forall z \in \mathcal{Z}\right\}, \tag{8}$$

where $\mu_{R,l}(z)$ and $\sigma_{R,l}(z)$ are as defined in (2) with $\lambda_R = H$. $\beta_{R,l} := B_R + \frac{\sigma_R}{\sqrt{H}}\sqrt{2\big(\ln(3/\delta) + \gamma_{(l-1)H}(R)\big)}$, where $\gamma_t(R) \equiv \gamma_t(k_R, \mathcal{Z})$ denotes the maximum information gain (or an upper bound on the maximum information gain) about any $f \sim GP_{\mathcal{Z}}(0, k_R)$ after t noisy observations obtained by passing f through an iid Gaussian channel $\mathcal{N}(0, H)$.

Similarly, GP-UCRL constructs the confidence set $\mathcal{C}_{P,l}$ as

$$\mathcal{C}_{P,l} = \left\{f : \|f(z) - \mu_{P,l-1}(z)\|_2 \leq \beta_{P,l}\,\|\sigma_{P,l-1}(z)\|_2 \,\forall z \in \mathcal{Z}\right\}. \tag{9}$$

Here, $\mu_{P,l}(z) := [\mu_{P,l}(z,1), \ldots, \mu_{P,l}(z,m)]^T$ and $\sigma_{P,l}(z) := [\sigma_{P,l}(z,1), \ldots, \sigma_{P,l}(z,m)]^T$, where $\mu_{P,l}(z,i)$ and $\sigma_{P,l}(z,i)$ are as defined in (3) with $\lambda_P = mH$. $\beta_{P,l} := B_P + \frac{\sigma_P}{\sqrt{mH}}\sqrt{2\big(\ln(3/\delta) + \gamma_{m(l-1)H}(P)\big)}$, where $\gamma_t(P) \equiv \gamma_t(k_P, \tilde{\mathcal{Z}})$ denotes the maximum information gain about any $f \sim GP_{\tilde{\mathcal{Z}}}(0, k_P)$ after t noisy observations obtained by passing f through an iid Gaussian channel $\mathcal{N}(0, mH)$.

Theorem 1 (Frequentist regret bound for GP-UCRL) *Let assumptions (A1) - (A3) hold, $k_R(z,z) \leq 1$ and $k_P((z,i),(z,i)) \leq 1$ for all $z \in \mathcal{Z}$ and $1 \leq i \leq m$[7]. Then for any $0 \leq \delta \leq 1$, GP-UCRL, with confidence sets (8) and (9), enjoys, with probability at least $1-\delta$, the regret bound*

$$\begin{aligned}Regret(T) \leq 2\beta_{R,\tau}\sqrt{2eH\gamma_T(R)T} + 2L\beta_{P,\tau}\sqrt{2emH\gamma_{mT}(P)T}\\ +(LD + 2B_RH)\sqrt{2T\ln(3/\delta)},\end{aligned}$$

where $T := \tau H$ is the total time in τ episodes, $\beta_{R,\tau} = B_R + \frac{\sigma_R}{\sqrt{H}}\sqrt{2\big(\ln(3/\delta) + \gamma_{(\tau-1)H}(R)\big)}$ and $\beta_{P,\tau} = B_P + \frac{\sigma_P}{\sqrt{mH}}\sqrt{2\big(\ln(3/\delta) + \gamma_{m(\tau-1)H}(P)\big)}$, L is a known upper bound over the global Lipschitz constant (1) $L_\star$ for the future value function of $M_\star$ and $D := \max_{s,s'\in\mathcal{S}}\|s - s'\|_2$ denotes the diameter of $\mathcal{S}$.

Interpretation of the bound As MIG increases with the number of observations, $\beta_{R,l}$ and $\beta_{P,l}$ increase with l. Hence $\beta_{R,\tau} = \tilde{O}\Big(B_R + \frac{\sigma_R}{\sqrt{H}}\sqrt{\gamma_T(R)}\Big)$ and $\beta_{P,\tau} = \tilde{O}\Big(B_P + \frac{\sigma_P}{\sqrt{mH}}\sqrt{\gamma_{mT}(P)}\Big)$. Thus, Theorem 1 implies that the cumulative regret of GP-UCRL after T timesteps is $\tilde{O}\Big(\big(\sqrt{H\gamma_T(R)} + \gamma_T(R)\big)\sqrt{T} + L\big(\sqrt{mH\gamma_{mT}(P)} + \gamma_{mT}(P)\big)\sqrt{T} + H\sqrt{T}\Big)$ with high probability. Hence, we see that the cumulative regret of GP-UCRL scales linearly with $\gamma_T(R)$ and $\gamma_{mT}(P)$. As $\gamma_T(R)$ and $\gamma_{mT}(P)$ grow sublinearly with T for most popular kernels (eg. Squared Exponential (SE), polynomial), the cumulative regret of GP-UCRL can grow sublinearly with T. We illustrate this with the following concrete examples:

(a) Example bound on $\gamma_T(R)$: Recall that $\gamma_T(R) \equiv \gamma_T(k_R, \mathcal{Z})$, where the kernel k_R is defined on the product space $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$. If k_1 and k_2 are kernels on the state space $\mathcal{S} \subset \mathbb{R}^m$ and the action space $\mathcal{A} \subset \mathbb{R}^n$, respectively, and k_R is an additive kernel of k_1 and k_2, then (4) implies that $\gamma_T(k_R, \mathcal{Z}) \leq \gamma_T(k_1, \mathcal{S}) + \gamma_T(k_2, \mathcal{A}) + 2\ln T$. Further, if both $\mathcal{S}$, $\mathcal{A}$ are compact and convex, and both k_1, k_2 are Squared Exponential (SE) kernels, then $\gamma_T(k_1, \mathcal{S}) = \tilde{O}\big((\ln T)^m\big)$ and $\gamma_T(k_2, \mathcal{A}) = \tilde{O}\big((\ln T)^n\big)$. Hence in this case $\gamma_T(R) = \tilde{O}\big((\ln T)^{\max\{m,n\}}\big)$.

(b) Example bound on $\gamma_{mT}(P)$: Recall that $\gamma_{mT}(P) \equiv \gamma_{mT}(k_P, \tilde{\mathcal{Z}})$, where the kernel k_P is defined on the product space $\tilde{\mathcal{Z}} = \mathcal{Z} \times \{1, \ldots, m\}$. If k_3 and k_4 are kernels on the product space $\mathcal{Z}$ and the index set $\{1, \ldots, m\}$, respectively, and k_P is a product kernel of k_3 and k_4, then (5) implies that $\gamma_{mT}(k_P, \tilde{\mathcal{Z}}) \leq m\gamma_{mT}(k_3, \mathcal{Z}) + m\ln(mT)$, since all kernel matrices over any subset of $\{1, \ldots, m\}$ have rank at most m. Further, if k_5 is a SE kernel on the state space $\mathcal{S}$, k_6 is a linear kernel on the action space $\mathcal{A}$ and k_3 is a product kernel, then (5) implies that $\gamma_{mT}(k_3, \mathcal{Z}) = \tilde{O}\big(n\big(\ln(mT)\big)^m\big)$, as the rank of an n-dimensional linear kernel is at most n. Hence, in this case, $\gamma_{mT}(P) = \tilde{O}\big(mn\big(\ln(mT)\big)^m\big)$.

Proof Sketch for Theorem 1 First, see that when $\overline{R}_\star \in \mathcal{C}_{R,l}$ and $\overline{P}_\star \in \mathcal{C}_{P,l}$, then the following are true:

(a) $M_\star$ lies in $\mathcal{M}_l$, the family of MDPs constructed by GP-UCRL. Hence $V^{M_l}_{\pi_l,1}(s_{l,1}) \geq V^{M_\star}_{\pi_\star,1}(s_{l,1})$, where M_l is the most optimistic realization from $\mathcal{M}_l$, and thus $\text{Regret}(T) \leq \sum_{l=1}^{\tau}\big(V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\big)$.

[7] This is called the bounded variance property of kernels and it holds for most of the common kernels (e.g. Squared Exponential).

(b) Optimistic rewards/transitions do not deviate too much: $\left|\overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h})\right| \leq 2\beta_{R,l}\ \sigma_{R,l-1}(z_{l,h})$ and $\left\|\overline{P}_{M_l}(z_{l,h}) - \overline{P}_\star(z_{l,h})\right\|_2 \leq 2\beta_{P,l}\left\|\sigma_{P,l-1}(z_{l,h})\right\|_2$, since by construction $\overline{R}_{M_l} \in \mathcal{C}_{R,l}$ and $\overline{P}_{M_l} \in \mathcal{C}_{P,l}$.

Further it can be shown that:

(c) Cumulative predictive variances are bounded: $\sum_{l=1}^{\tau}\sum_{h=1}^{H}\sigma_{R,l-1}(z_{l,h}) \leq \sqrt{2eH\gamma_T(R)T}$ and $\sum_{l=1}^{\tau}\sum_{h=1}^{H}\left\|\sigma_{P,l-1}(z_{l,h})\right\|_2 \leq \sqrt{2emH\gamma_{mT}(P)T}$.

(d) Bounds on deviations of rewards and transitions imply bounds on deviation of the value function: $\sum_{l=1}^{\tau}\left(V_{\pi_l,1}^{M_l}(s_{l,1}) - V_{\pi_l,1}^{M_\star}(s_{l,1})\right) \leq \sum_{l=1}^{\tau}\sum_{h=1}^{H}\left|\overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h})\right| + L\sum_{l=1}^{\tau}\sum_{h=1}^{H}\left\|\overline{P}_{M_l}(z_{l,h}) - \overline{P}_\star(z_{l,h})\right\|_2 + (LD + 2B_R H)\sqrt{2\tau H\ln(3/\delta)}$, with probability at least $1-\delta/3$.

The proof now follows by combining (a), (b), (c) and (d), and showing the confidence set properties $\mathbb{P}\left[\overline{R}_\star \in \mathcal{C}_{R,l}\right] \geq 1-\delta/3$ and $\mathbb{P}\left[\overline{P}_\star \in \mathcal{C}_{P,l}\right] \geq 1-\delta/3$.

4.3 Regret Bound for PSRL

Osband and Van Roy [2016] show that if we have a frequentist regret bound for UCRL in hand, then we can obtain a similar bound (upto a constant factor) on the *Bayes regret* (defined as the *expected regret under the prior distribution* Φ) of PSRL. We use this idea to obtain a sublinear bound on the Bayes regret of PSRL for kernelized MDPs.

Theorem 2 (Bayes regret of PSRL under RKHS prior) *Let assumptions (A1) - (A3) hold, $k_R(z,z) \leq 1$ and $k_P((z,i),(z,i)) \leq 1$ for all $z \in \mathcal{Z}$ and $1 \leq i \leq m$. Let Φ be a (known) prior distribution over MDPs $M_\star$. Then, the Bayes regret of PSRL satisfies*

$$\mathbb{E}\left[Regret(T)\right] \leq 2\alpha_{R,\tau}\sqrt{2eH\gamma_T(R)T} + 3\mathbb{E}\left[L_\star\right]\alpha_{P,\tau}\sqrt{2emH\gamma_{mT}(P)T} + 3B_R,$$

where $T = \tau H$ is the total time in τ episodes, $L_\star$ is the global Lipschitz constant for the future value function (1) of $M_\star$, $\alpha_{R,\tau} = B_R + \frac{\sigma_R}{\sqrt{H}}\sqrt{2\big(\ln(3T) + \gamma_{(\tau-1)H}(R)\big)}$ and $\alpha_{P,\tau} = B_P + \frac{\sigma_P}{\sqrt{mH}}\sqrt{2\big(\ln(3T) + \gamma_{m(\tau-1)H}(P)\big)}$.

Theorem 2 implies that the Bayes regret of PSRL after T timesteps is $\tilde{O}\Big(\big(\sqrt{H\gamma_T(R)} + \gamma_T(R)\big)\sqrt{T} + \mathbb{E}\left[L_\star\right]\big(\sqrt{mH\gamma_{mT}(P)} + \gamma_{mT}(P)\big)\sqrt{T} + H\sqrt{T}\Big)$, and thus has the same scaling as the bound for GP-UCRL.

Remark. Observe that when $H \leq \gamma_T(R)$ and $mH \leq \gamma_{mT}(P)$[8], then the regret of GP-UCRL is $\tilde{O}\Big(\big(\gamma_T(R) + L\gamma_{mT}(P)\big)\sqrt{T}\Big)$ with high probability and the Bayes regret of PSRL is $\tilde{O}\Big(\big(\gamma_T(R) + \mathbb{E}\left[L_\star\right]\gamma_{mT}(P)\big)\sqrt{T}\Big)$, where both L (an upper bound on $L_\star$) and $\mathbb{E}\left[L_\star\right]$ basically measure the connectedness of the MDP $M_\star$.

Comparison with the eluder dimension results Osband and Van Roy [2014] assume that $\overline{R}_\star$ and $\overline{P}_\star$ are elements from two function classes $\mathcal{R}$ and $\mathcal{P}$, respectively, with bounded $\|\cdot\|_2$-norm, and show that PSRL obtains Bayes regret $\tilde{O}\Big(\big(\sqrt{d_K(\mathcal{R})d_E(\mathcal{R})} + \mathbb{E}\left[L_\star\right]\sqrt{d_K(\mathcal{P})d_E(\mathcal{P})}\big)\sqrt{T}\Big)$, where $d_K(\mathcal{F})$ (Kolmogorov dimension) and $d_E(\mathcal{F})$ (eluder dimension) measure the "complexity" of a function class $\mathcal{F}$. As a special case, if both $\overline{R}_\star$ and $\overline{P}_\star$ are linear functions in finite dimension d, then they show that $d_E(\mathcal{R}), d_K(\mathcal{R}) = \tilde{O}(d)$ and $d_E(\mathcal{P}), d_K(\mathcal{P}) = \tilde{O}(md)$. In our setting, $\overline{R}_\star$ and $\overline{P}_\star$ are RKHS functions, and hence, by the reproducing property, they are linear functionals in (possibly) infinite dimension. From this viewpoint, all of $d_E(\mathcal{R})$, $d_K(\mathcal{R})$, $d_E(\mathcal{P})$ and $d_K(\mathcal{P})$ can blow upto infinity yielding trivial bounds. Therefore, we need a suitable measure of complexity of the RKHS spaces, and a single information-theoretic quantity, namely the Maximum Information Gain (MIG), is seen to serve this purpose.

To the best of our knowledge, Theorem 1 is the first frequentist regret bound and Theorem 2 is the first Bayesian regret bound in the kernelized MDP setting (i.e., when the MDP model is from an RKHS class). We see that both algorithms achieve similar regret bounds in terms of dependencies on time, MDP connectedness and Maximum Information Gain. However, Theorem 1 is a stronger probabilistic guarantee than Theorem 2 since it holds with high probability for any MDP $M_\star$ and not just in expectation over the draw from the prior distribution.

As special cases of our results, we now derive regret bounds for two representative RL domains, namely tabular MDPs and linear quadratic control systems.

Tabula-rasa MDPs In this case, both $\mathcal{S}$ and $\mathcal{A}$ are finite and expressed as $\mathcal{S} = \{1,\ldots,|\mathcal{S}|\}$ and $\mathcal{A} = \{1,\ldots,|\mathcal{A}|\}$. This corresponds to taking k_R and k_P as product kernels, i.e., $k_R((i,j),(i',j')) = k_P((i,j),(i',j')) = k_1(i,i')k_2(j,j')$ for all $1 \leq i,i' \leq |\mathcal{S}|$ and $1 \leq j,j' \leq |\mathcal{A}|$, where both k_1 and k_2 are linear kernels with dimensions $|\mathcal{S}|$ and $|\mathcal{A}|$, respectively, such that $k_1(i,i') = \mathbb{1}_{\{i=i'\}}$ and $k_2(j,j') = \mathbb{1}_{\{j=j'\}}$. Hence (5) implies that $\gamma_T(k_R,\mathcal{Z}) \leq |\mathcal{A}|\,\gamma_T(k_1,\mathcal{S}) + |\mathcal{A}|\ln T$, as the rank of k_2 is at most $|\mathcal{A}|$. Further, as k_1 is a linear kernel, $\gamma_T(k_1,\mathcal{S}) = \tilde{O}(|\mathcal{S}|\ln T)$, hence $\gamma_T(R) \equiv \gamma_T(k_R,\mathcal{Z}) = \tilde{O}(|\mathcal{S}|\,|\mathcal{A}|\ln T)$. Similarly, $\gamma_{mT}(P) = \tilde{O}(|\mathcal{S}|\,|\mathcal{A}|\ln T)$ as in this case $m = 1$. Plugging these into our bounds with the Lipschitz constant $L_\star = O(H)$, we see that both GP-UCRL and PSRL suffer regret $\tilde{O}(H\,|\mathcal{S}|\,|\mathcal{A}|\sqrt{T})$ for tabula-rasa MDPs.

[8] Both conditions hold, for instance, if $H = O(\ln T)$ with a polynomial or SE kernel.

Control of Bounded Linear Quadratic Systems Consider learning under the standard discrete-time, episodic, linear quadratic regulator (LQR) model: at period h of episode l,

$$\begin{aligned} s_{l,h+1} &= As_{l,h} + Ba_{l,h} + \varepsilon_{P,l,h}, \\ r_{l,h} &= s_{l,h}^T P s_{l,h} + a_{l,h}^T Q a_{l,h} + \varepsilon_{R,l,h}, \end{aligned} \tag{10}$$

where $r_{l,h} \in \mathbb{R}$ is the reward obtained by executing action $a_{l,h} \in \mathcal{A} \subset \mathbb{R}^n$ at state $s_{l,h} \in \mathcal{S} \subset \mathbb{R}^m$ and $s_{l,h+1}$ is the next state. $P \in \mathbb{R}^{m\times m}$, $Q \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{m\times n}$ are unknown matrices with P and Q assumed positive-definite, and $\varepsilon_{R,l,h}$, $\varepsilon_{P,l,h}$ follow a sub-Gaussian noise model as per 6 and 7, respectively.

Corollary 1 (Regret of GP-UCRL for LQR) *Let $M_\star$ be a linear quadratic system (10), and both $\mathcal{S}$ and $\mathcal{A}$ be compact and convex. Let $H \leq \min\{(m^2+n^2)\ln T, (m+n)\ln(mT)\}$[9]. Then, for any $0 < \delta \leq 1$, GP-UCRL enjoys, with probability at least $1-\delta$, the regret bound*

$$Regret(T) = \tilde{O}\Big(\big(B_R(m^2{+}n^2){+}LB_P m(m{+}n)\big)\sqrt{T\ln(1/\delta)}\Big).$$

Here, $B_R = \big(\|P\|_F^2 + \|Q\|_F^2\big)^{1/2}$ and $B_P = \big(\|A\|_F^2 + \|B\|_F^2\big)^{1/2}$, L is a known upper bound over $D\lambda_1$, $D = \max_{s,s'\in\mathcal{S}} \|s-s'\|_2$ is the diameter of $\mathcal{S}$ and λ_1 is the largest eigenvalue of the positive definite matrix G, which is a unique solution to the Riccati equations Lancaster and Rodman [1995] for the unconstrained optimal value function $V(s) = s^T G s$.

Proof The proof uses composite kernels, based on linear and quadratic kernels, to represent the LQR model. First, note that the mean reward function is $\overline{R}_\star(s,a) = s^T P s + a^T Q a$ and the mean transition function is $\overline{P}_\star(s,a) = As + Ba$. Now recall our notation $z = (s,a)$, $z' = (s',a')$, $\mathcal{Z} = \mathcal{S}\times\mathcal{A}$ and $\tilde{\mathcal{Z}} = \mathcal{Z}\times\{1,\ldots,m\}$. Then defining $\overline{P}_\star = [A_1,\ldots,A_m,B_1,\ldots,B_m]^T$, where $A_i, 1\leq i\leq m$ and $B_i, 1\leq i\leq m$ are the rows A and B respectively, we see that $\overline{P}_\star$ lies in the RKHS $\mathcal{H}_{k_P}(\tilde{\mathcal{Z}})$ with the kernel $k_P\big((z,i),(z',j)\big) = k_1(z,z')k_2(i,j)$, where $k_1(z,z') = s^T s' + a^T a'$ and $k_2(i,j) = \mathbb{1}_{\{i=j\}}, 1\leq i,j\leq m$. Since k_P is a product of the kernels k_1 and k_2, (5) implies that

$$\gamma_t(P) \equiv \gamma_t(k_P, \tilde{\mathcal{Z}}) \leq m\gamma_t(k_1,\mathcal{Z}) + m\ln t, \tag{11}$$

as the rank of k_2 is atmost m. Further k_1 is a sum of two linear kernels, defined over $\mathcal{S}$ and $\mathcal{A}$ respectively. Hence (4) implies $\gamma_t(k_1,\mathcal{Z}) \leq \tilde{O}(m\ln t) + \tilde{O}(n\ln t) + 2\ln t = \tilde{O}\big((m+n)\ln t\big)$, since the MIG of a d-dimensional linear kernel is $\tilde{O}(d\ln t)$. Hence, by (11), we have $\gamma_t(P) = \tilde{O}\big(m(m+n)\ln t\big)$.

Similarly defining $\overline{R}_\star = [P_1,\ldots,P_m,Q_1,\ldots,Q_n]^T$, where $P_i, 1\leq i\leq m$ and $Q_i, 1\leq i\leq n$ are the rows P and Q respectively, we see that $\overline{R}_\star$ lies in the RKHS $\mathcal{H}_{k_R}(\mathcal{Z})$ with the quadratic kernel $k_R(z,z') = (s^T s')^2 + (a^T a')^2$. Since k_R is an additive kernel, (4) implies that

$$\gamma_t(R) = \gamma_t(k_R,\mathcal{Z}) \leq \gamma_t(k_3,\mathcal{S}) + \gamma_t(k_3,\mathcal{A}) + 2\ln t, \tag{12}$$

where $k_3(x,x') := (x^T x')^2 = (x^T x')(x^T x')$ is a quadratic kernel and thus a product of two linear kernels. Hence, (5) implies that $\gamma_t(k_3,\mathcal{S}) \leq m\ \tilde{O}(m\ln t) + m\ln t = \tilde{O}(m^2\ln t)$, since the rank of an m-dimensional linear kernel is at most m. Similarly $\gamma_t(k_3,\mathcal{A}) = \tilde{O}(n^2\ln t)$. Hence from (12), we have $\gamma_t(R) = \tilde{O}\big((m^2+n^2)\ln t\big)$. Now, following a similar argument as by Osband and Van Roy [2014, Corollary 2], we can show that the Lipschitz constant $L_\star = D\lambda_1$. Further, in this setting, we take $B_R = \big\|\overline{R}_\star\big\|_{k_R} = \big(\|P\|_F^2 + \|Q\|_F^2\big)^{1/2}$ and $B_P = \big\|\overline{P}_\star\big\|_{k_P} = \big(\|A\|_F^2 + \|B\|_F^2\big)^{1/2}$. Now the result follows from Theorem 1 using $H \leq \min\{(m^2+n^2)\ln T, (m+n)\ln(mT)\}$. ■

Corollary 2 (Bayes regret of PSRL for LQR) *Let $M_\star$ be a linear quadratic system defined as per (10), Φ be the (known) distribution of $M_\star$ and both $\mathcal{S}$ and $\mathcal{A}$ be compact and convex. Let $H \leq \min\{(m^2+n^2)\ln T, (m+n)\ln(mT)\}$. Then the Bayes regret of PSRL satisfies*

$$\mathbb{E}\left[Regret(T)\right] = \tilde{O}\Big(\big(B_R(m^2{+}n^2){+}D\lambda_1 B_P m(m{+}n)\big)\sqrt{T}\Big),$$

where B_R, B_P, D and λ_1 are as given in Corollary 1.

Proof Using the similar arguments as above and noting that $\mathbb{E}\left[L_\star\right] = D\lambda_1$, the result follows from Theorem 2. ■

Remark. Corollary 2 matches the bound given in Osband and Van Roy [2014] for the same bounded LQR problem. But the analysis technique is different here, and this result is derived as a special case of more general kernelized dynamics. Corollary 1 (order-wise) matches the bound given in Abbasi-Yadkori and Szepesvári [2011] if we restrict their result to the bounded LQR problem.

5 DISCUSSION

We have derived the first regret bounds for RL in the kernelized MDP setup with continuous state and action spaces, with explicit dependence of the bounds on the maximum information gains of the transition and reward function classes. In Appendix C, we have also developed the Bayesian RL analogue of Gaussian process bandits Srinivas et al. [2009], i.e., learning under the assumption that MDP dynamics and reward behavior are sampled according to Gaussian process priors. We have proved Bayesian regret bounds for GP-UCRL and PSRL under GP priors. We only have a (weak) Bayes regret bound for PSRL in kernelized MDPs, and would like to examine if a frequentist bound also holds. Another concrete direction is to examine if similar guarantees can be attained in the model-free setup, which may obviate complicated planning in the model-based setup here.

[9] This assumption naturally holds in most settings and is used here only for brevity.

Acknowledgements

The authors are grateful to the anonymous reviewers for their valuable comments. S. R Chowdhury is supported by Google India PhD fellowship grant. A. Gopalan is grateful for support from the DST INSPIRE faculty grant IFA13-ENG-69.

References

Y. Abbasi-Yadkori and C. 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.

Y. Abbasi-Yadkori and C. Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *UAI*, pages 1–11. Citeseer, 2015.

M. Abeille and A. Lazaric. Thompson sampling for linear-quadratic control problems. *arXiv preprint arXiv:1703.08972*, 2017.

S. Agrawal and R. Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.

P. L. Bartlett and A. Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42. AUAI Press, 2009.

F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems*, pages 908–919, 2017.

S. R. Chowdhury and A. Gopalan. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pages 844–853, 2017.

T. Desautels, A. Krause, and J. W. Burdick. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *The Journal of Machine Learning Research*, 15(1):3873–3923, 2014.

A. Durand, O.-A. Maillard, and J. Pineau. Streaming kernel regression with provably adaptive mean, variance, and regularization. *arXiv preprint arXiv:1708.00768*, 2017.

A. Gopalan and S. Mannor. Thompson sampling for learning parameterized markov decision processes. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 861–898, 2015.

M. Ibrahimi, A. Javanmard, and B. V. Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances in Neural Information Processing Systems*, pages 2636–2644, 2012.

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

K. Kandasamy, A. Krishnamurthy, J. Schneider, and B. Póczos. Parallelised bayesian optimisation via thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 133–142, 2018.

A. Krause and C. S. Ong. Contextual gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, pages 2447–2455, 2011.

K. Lakshmanan, R. Ortner, and D. Ryabko. Improved regret bounds for undiscounted continuous reinforcement learning. In *International Conference on Machine Learning*, pages 524–532, 2015.

P. Lancaster and L. Rodman. *Algebraic riccati equations*. Clarendon press, 1995.

R. Ortner and D. Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1763–1771, 2012.

I. Osband and B. Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474, 2014.

I. Osband and B. Van Roy. Why is posterior sampling better than optimism for reinforcement learning? *arXiv preprint arXiv:1607.00215*, 2016.

I. Osband, D. Russo, and B. Van Roy. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

Y. Ouyang, M. Gagrani, A. Nayyar, and R. Jain. Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems*, pages 1333–1342, 2017.

C. E. Rasmussen and C. K. I. Williams. Gaussian processes for machine learning. 2006.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

A. L. Strehl, L. Li, and M. L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *J. Mach. Learn. Res.*, 10:2413–2444, Dec. 2009.

M. Turchetta, F. Berkenkamp, and A. Krause. Safe exploration in finite markov decision processes with gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4312–4320, 2016.

M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*, 2013.