

---

# Model-Based Reinforcement Learning and the Eluder Dimension

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

We consider the problem of learning to optimize an unknown Markov decision process (MDP) whose reward and transition functions lie within the function classes  $\mathcal{R}$  and  $\mathcal{P}$ . For any functional families  $\mathcal{R}$  and  $\mathcal{P}$  with uniformly bounded means and  $\sigma$ -sub-Gaussian noise, we show the conceptually simple and computationally efficient algorithm *posterior sampling for reinforcement learning* (PSRL) has expected regret  $\tilde{O}\left(\sigma\sqrt{d_K(\mathcal{R})d_E(\mathcal{R}, T^{-1})T} + \sigma\mathbb{E}[K^*]\sqrt{d_K(\mathcal{P})d_E(\mathcal{P}, T^{-1})T}\right)$ . Here  $T$  is the time elapsed,  $K^*$  is a Lipschitz constant for the future value function,  $d_K$  is the Kolmogorov dimension and  $d_E$  is the eluder dimension. These bounds characterize model-based reinforcement learning in terms of the dimensionality, rather than cardinality, of the underlying MDP and provide new state of the art guarantees across several important domains.

## 1 Introduction

We consider an agent that interacts with an uncertain environment while trying to maximize its cumulative rewards through time [1]. We model this environment as a Markov decision process (MDP) where the agent is unsure of the true reward and transition but knows that they lie within the functional families  $\mathcal{R}$  and  $\mathcal{P}$ . Efficient reinforcement learning seeks algorithms which balance exploration with exploitation, so that the agent's performance is guaranteed to be close to the unknown optimal control in some sense [2, 3].

The main focus in this literature has been to develop algorithms for the *tabula rasa* setting, where little prior knowledge of the environment is assumed beyond the state and action spaces. Broadly, these algorithms can be separated as either model-based, which build a generative model of the environment or model-free, which do not. Algorithms of both type have been developed to provide PAC-MDP bounds polynomial in the number of states  $S$  and actions  $A$  [4, 5, 6], but only model-based algorithms have been able to provide efficient regret bounds so far. The strongest results establish algorithms with bounds on the cumulative regret at time  $T$ ,  $\tilde{O}(S\sqrt{AT})$  close to the lower bound  $\Omega(\sqrt{SAT})$  [7, 8].

In the *tabula rasa* setting MDP complexity grows with cardinality of the underlying state and action spaces; in many settings of interest this may be extremely large or even infinite. The standard approach to this problem is some discretization scheme with degradation to performance though imprecision, but even this falls victim to the curse of dimensionality. To achieve efficiency results that can exploit MDP structure, the existing literature usually presents algorithms tailored for specific forms of  $\mathcal{R}$  and  $\mathcal{P}$ .

The most widely-studied parameterization is the degenerate MDP with no transitions, the multi-armed bandit [9, 10, 11]. Another common assumption is that the transition func-

tion is linear in states and actions. Papers here have established sample complexity [12] and even regret bounds  $\tilde{O}(\sqrt{T})$  for linear quadratic control [13], but with constants that grow exponentially with dimension. Later works remove this dependence, but only under significant sparsity assumptions [14]. The most general previous analysis considers  $\mathcal{P}$  that is  $\alpha$ -Hölder in a  $d$ -dimensional space to establish regret bounds  $\tilde{O}(T^{(2d+\alpha)/(2d+2\alpha)})$  [15]. However, the proposed algorithm UCCRL is not computationally tractable and the bounds approach linearity in many settings.

We present the first regret bounds for learning in a general MDP with rewards in  $\mathcal{R}$  and transitions in  $\mathcal{P}$  for *any*  $\mathcal{R}$  and  $\mathcal{P}$  with bounded mean and  $\sigma$ -sub-Gaussian noise. To characterize the complexity of this learning problem we extend the definition of the eluder dimension, previously introduced for bandits, to capture the interdependence between state-action pairs [16]. These results provide a unified framework for the analysis of model-based reinforcement learning and provide new state of the art bounds in several important problem settings. What is more, these bounds are satisfied by one simple and intuitive algorithm *posterior sampling for reinforcement learning* (PSRL) [17, 18, 8] across all problem settings. PSRL can naturally encode complex problem knowledge with computational complexity generally no greater than solving a single known MDP.

## 2 Problem formulation

We consider the problem of learning to optimize a random finite horizon MDP  $M = (\mathcal{S}, \mathcal{A}, R^M, P^M, \tau, \rho)$  in repeated finite episodes of interaction.  $\mathcal{S}$  is the  $L^2$  state space,  $\mathcal{A}$  is the action space,  $R^M(s, a)$  is the reward distribution over  $\mathbb{R}$  and  $P^M(\cdot|s, a)$  is the transition distribution over  $\mathcal{S}$  when selecting action  $a$  in state  $s$ ,  $\tau$  is the time horizon, and  $\rho$  the initial state distribution. We define the MDP and all other random variables we will consider with respect to a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ .

A policy  $\mu$  is a function mapping each state  $s \in \mathcal{S}$  and  $i = 1, \dots, \tau$  to an action  $a \in \mathcal{A}$ . For each MDP  $M$  and policy  $\mu$ , we define a value function  $V$  and future value  $U$ :

$$V_{\mu,i}^M(s) := \mathbb{E}_{M,\mu} \left[ \sum_{j=i}^{\tau} \bar{r}^M(s_j, a_j) \middle| s_i = s \right]; U_i^M(Q) := \mathbb{E}_{M,\mu^M} \left[ V_{\mu^M,i+1}^M(s) \middle| s \sim Q \right] \quad (1)$$

where  $\bar{r}^M(s, a)$  denotes the expected reward realized when action  $a$  is selected while in state  $s$ , and the subscripts of the expectation operator indicate that  $a_j = \mu(s_j, j)$ , and  $s_{j+1} \sim P^M(\cdot|s_j, a_j)$  for  $j = i, \dots, \tau$ . A policy  $\mu$  is said to be optimal for MDP  $M$  if  $V_{\mu,i}^M(s) = \max_{\mu'} V_{\mu',i}^M(s)$  for all  $s \in \mathcal{S}$  and  $i = 1, \dots, \tau$ . We will associate with each MDP  $M$  a policy  $\mu^M$  that is optimal for  $M$ .

The reinforcement learning agent interacts with the MDP over episodes that begin at times  $t_k = (k-1)\tau + 1$ ,  $k = 1, 2, \dots$ . At each time  $t$ , the agent selects an action  $a_t$ , observes a scalar reward  $r_t$ , and then transitions to  $s_{t+1}$ . If an agent follows a policy  $\mu$  then when in state  $s$  at time  $t$  during episode  $k$ , it selects an action  $a_t = \mu(s, t - t_k)$ . Let  $H_t = (s_1, a_1, r_1, \dots, s_{t-1}, a_{t-1}, r_{t-1})$  denote the history of observations made *prior* to time  $t$ . A reinforcement learning algorithm is a deterministic sequence  $\{\pi_k | k = 1, 2, \dots\}$  of functions, each mapping  $H_{t_k}$  to a probability distribution  $\pi_k(H_{t_k})$  over policies which the agent will employ during the  $k$ th episode. We define the regret incurred by a reinforcement learning algorithm  $\pi$  up to time  $T$  to be

$$\text{Regret}(T, \pi, M^*) := \sum_{k=1}^{\lceil T/\tau \rceil} \Delta_k,$$

where  $\Delta_k$  denotes regret over the  $k$ th episode, defined with respect to the MDP  $M^*$  by

$$\Delta_k := \int_{s \in \mathcal{S}} \rho(s) \left( V_{\mu^*,1}^{M^*} - V_{\mu_k,1}^{M^*} \right) (s)$$

with  $\mu^* = \mu^{M^*}$  and  $\mu_k \sim \pi_k(H_{t_k})$ . Note that regret is not deterministic since it can depend on the random MDP  $M^*$ , the algorithm's internal random sampling and, through

the history  $H_{t_k}$ , on previous random transitions and random rewards. We will assess and compare algorithm performance in terms of regret and its expectation.

### 3 Main results

We now introduce the algorithm posterior sampling for reinforcement learning (PSRL) first proposed by Strens [18] and later was shown to satisfy efficient regret bounds in finite MDPs [8]. The algorithm begins with a prior distribution over MDPs with rewards in  $\mathcal{R}$  and transitions in  $\mathcal{P}$ , at the start of the  $k$ th episode, PSRL samples an MDP  $M_k$  from the posterior<sup>1</sup>. PSRL then follows the policy  $\mu_k = \mu^{M_k}$  which is optimal for this *sampled* MDP during episode  $k$ . This algorithm is a natural adaptation of Thompson sampling [17], one of the oldest and most effective algorithms for multi-armed bandits, to reinforcement learning.

---

#### Algorithm 1

Posterior Sampling for Reinforcement Learning (PSRL)

---

```

1: Input: Prior distribution  $\phi$  for  $M^*$ ,  $t=1$ 
2: for episodes  $k = 1, 2, \dots$  do
3:   sample  $M_k \sim \phi(\cdot | H_t)$ 
4:   compute  $\mu_k = \mu^{M_k}$ 
5:   for timesteps  $j = 1, \dots, \tau$  do
6:     apply  $a_t \sim \mu_k(s_t, j)$ 
7:     observe  $r_t$  and  $s_{t+1}$ 
8:     advance  $t = t + 1$ 
9:   end for
10: end for

```

---

We now introduce some notation that is developed in the remainder of the paper. For any set  $\mathcal{X}$  and  $L^2$  space  $(\mathcal{Y}, \|\cdot\|_2)$  let  $\mathcal{P}_{\mathcal{X}, \mathcal{Y}}^{C, \sigma}$  be the family of distributions from  $\mathcal{X}$  to  $\mathcal{Y}$  with mean norm bounded in  $[0, C]$  and additive  $\sigma$ -sub-Gaussian noise. For the family  $\mathcal{F}$  let  $N(\mathcal{F}, \alpha, \|\cdot\|_2)$  be the  $\alpha$ -covering number of  $\mathcal{F}$  with respect to the  $\|\cdot\|_2$ -norm and write  $n_{\mathcal{F}} = \log(8N(\mathcal{F}, 1/T^2, \|\cdot\|_2)T)$  for brevity. Finally we write  $d_{\mathcal{F}} = \dim_E(\mathcal{F}, T^{-1})$  for the eluder dimension of  $\mathcal{F}$ , a notion of dimension specialized to sequential measurements described in Section 4.

Our main result, Theorem 1, bounds the expected regret of PSRL at any time  $T$ .

**Theorem 1** (Expected regret for PSRL in parameterized MDPs).

Let  $M^*$  have rewards  $R^* \in \mathcal{R} \subseteq \mathcal{P}_{\mathcal{S} \times \mathcal{A}, \mathbb{R}}^{C_{\mathcal{R}}, \sigma_{\mathcal{R}}}$  and transitions  $P^* \in \mathcal{P} \subseteq \mathcal{P}_{\mathcal{S} \times \mathcal{A}, \mathcal{S}}^{C_{\mathcal{P}}, \sigma_{\mathcal{P}}}$ . If  $\phi$  is the distribution of  $M^*$  and  $K^*$  is the Lipschitz constant of the future value function then:

$$\begin{aligned} \mathbb{E}[\text{Regret}(T, \pi^{PS}, M^*)] \leq & \left[ C_{\mathcal{R}} + C_{\mathcal{P}} \right] + \left[ 1 + \tau C_{\mathcal{R}} d_{\mathcal{R}} + 8\sqrt{d_{\mathcal{R}}(4C_{\mathcal{R}} + \sqrt{2\sigma_{\mathcal{R}}^2 \log(32T^3)})} + 8\sqrt{2\sigma_{\mathcal{R}}^2 n_{\mathcal{R}} d_{\mathcal{R}} T} \right] \\ & + \mathbb{E}[K^*] \left( 1 + \frac{1}{T-1} \right) \left( 1 + \tau C_{\mathcal{P}} d_{\mathcal{P}} + 8\sqrt{d_{\mathcal{P}}(4C_{\mathcal{P}} + \sqrt{2\sigma_{\mathcal{P}}^2 \log(32T^3)})} + 8\sqrt{2\sigma_{\mathcal{P}}^2 n_{\mathcal{P}} d_{\mathcal{P}} T} \right) \end{aligned} \quad (2)$$

We note that, due to the effects of system noise, the future value function will often be Lipschitz even where the optimal value function is discontinuous. To clarify the asymptotic dependence of this bound we will make use of another classic measure of dimensionality, the Kolmogorov dimension of a function class  $\mathcal{F}$ ,  $\dim_K(\mathcal{F})$ . Using this definition in Theorem 1 we can obtain our Corollary.

**Corollary 1** (Asymptotic regret bounds for PSRL in parameterized MDPs).

Let  $M^*$  have rewards  $R^* \in \mathcal{R} \subseteq \mathcal{P}_{\mathcal{S} \times \mathcal{A}, \mathbb{R}}^{C_{\mathcal{R}}, \sigma_{\mathcal{R}}}$  and transitions  $P^* \in \mathcal{P} \subseteq \mathcal{P}_{\mathcal{S} \times \mathcal{A}, \mathcal{S}}^{C_{\mathcal{P}}, \sigma_{\mathcal{P}}}$ . If  $\phi$  is the distribution of  $M^*$  and  $K^*$  is the Lipschitz constant of the future value function then:

$$\mathbb{E}[\text{Regret}(T, \pi^{PS}, M^*)] = \tilde{O} \left( \sigma_{\mathcal{R}} \sqrt{\dim_K(\mathcal{R}) \dim_E(\mathcal{R}, T^{-1}) T} + \mathbb{E}[K^*] \sigma_{\mathcal{P}} \sqrt{\dim_K(\mathcal{P}) \dim_E(\mathcal{P}, T^{-1}) T} \right) \quad (3)$$

---

<sup>1</sup>Without conjugate priors this can be achieved by algorithms such as Metropolis-Hastings.

Where  $\tilde{O}(\cdot)$  ignores terms logarithmic in  $T$ .

Using the bounds on eluder dimension given in Section 4 we can provide concrete regret bounds in a number of canonical domains such as discrete MDPs, linear-quadratic control and even generalized linear systems. In all of these cases the eluder dimension scales comparably with more traditional notions of dimensionality. For clarity, we present bounds in the case of linear-quadratic control.

**Corollary 2** (Asymptotic regret bounds for PSRL in linear quadratic systems).

Let  $M^*$  be a linear-quadratic system with  $n$  – dimensional states and actions  $\|\cdot\|_2$ -bounded in  $[0, C]$  and  $\sigma$ -sub-Gaussian noise. If  $\phi$  is the distribution of  $M^*$ , then:

$$\mathbb{E}[\text{Regret}(T, \pi^{PS}, M^*)] = \tilde{O}\left(\sigma C \lambda_1 n^2 \sqrt{T}\right) \quad (4)$$

Where  $\lambda_1$  is the largest eigenvalue of the matrix  $Q$  given as the solution of the Ricatti equations for the unconstrained optimal value function  $V(s) = s^T Q s$  [19].

*Proof.* We simply apply the results of for eluder dimension in Section 4 to Corollary 1 and upper bound the Lipschitz constant of the constrained LQR by  $C\lambda_1$ .  $\square$

We note that it would also be possible to use the analysis in this paper to formally present an optimistic algorithm that satisfied similar regret bounds with high probability. We do not present these proofs in this paper to simplify the discussion, but also because the resultant algorithm UCRL-Eluder, outlined in Appendix D would be computationally intractable even when presented with an approximate MDP planner. Further, we believe that PSRL will generally be more statistically efficient than an optimistic variant with similar regret bounds since the algorithm is not affected by loose analysis [8].

## 4 Eluder dimension

We extend the existing notion of eluder dimension for real-valued functions [16] to those over any L2 space. Our definition remains the same as earlier treatments, but with  $|\cdot|$  replaced by  $\|\cdot\|_2$ . Let  $\mathcal{F} \subseteq \mathcal{P}_{\mathcal{X}, \mathcal{Y}}^{C, \sigma}$  with sequential measurements of  $f^*$  at  $x_1, x_2, \dots, x_t \in \mathcal{X}$  and realizations  $y_i = f^*(x_i) + \epsilon_i$  for  $i = 1, \dots, t$ . Intuitively, the eluder dimension of  $\mathcal{F}$  is the length  $d$  of the longest possible sequence  $x_1, \dots, x_d$  such that for all  $i$  knowing the function values of  $f(x_1), \dots, f(x_i)$  will not tell you  $f(x_{i+1})$ .

**Definition 1** ( $(\mathcal{F}, \epsilon)$  – dependence).

We will say that  $x \in \mathcal{X}$  is  $(\mathcal{F}, \epsilon)$ -dependent on  $\{x_1, \dots, x_n\} \subseteq \mathcal{X}$

$$\iff \forall f, \tilde{f} \in \mathcal{F}, \sum_{i=1}^n \|f(x_i) - \tilde{f}(x_i)\|_2^2 \leq \epsilon^2 \implies \|f(x) - \tilde{f}(x)\|_2 \leq \epsilon.$$

Similarly,  $x \in \mathcal{X}$  is  $(\epsilon, \mathcal{F})$ -independent of  $\{x_1, \dots, x_n\}$  iff it does not satisfy the definition for dependence.

**Definition 2** (Eluder Dimension).

The eluder dimension  $\text{dim}_E(\mathcal{F}, \epsilon)$  is the length of the longest possible sequence of elements in  $\mathcal{X}$  such that for some  $\epsilon' \geq \epsilon$  every element is  $(\mathcal{F}, \epsilon')$ -independent of its predecessors.

Traditional notions from supervised learning, such as the VC dimension, are not sufficient to characterize the complexity of reinforcement learning. In fact, a family learnable in constant time for supervised learning may require arbitrarily long to learn to optimize [16]. The Eluder dimension mirrors the linear dimension for vector spaces, which is the length of the longest sequence such that each element is linearly independent of its predecessors. We extend this notion of complexity to account for nonlinear and approximate dependencies.

#### 4.1 Eluder dimension for specific function classes

The eluder dimension is well-defined and straightforward notion for any  $\mathcal{F}, \epsilon$ . However, given  $\mathcal{F}, \epsilon$  calculating the eluder dimension may take some additional work. We now provide bounds on the eluder dimension for some common function classes in a similar approach to earlier work [11]. These proofs are available in Appendix C.

**Proposition 1** (Eluder dimension for finite  $\mathcal{X}$ ).

A counting argument shows that for  $|\mathcal{X}| = X$  finite, any  $\epsilon > 0$  and any function class  $\mathcal{F}$ :

$$\dim_E(\mathcal{F}, \epsilon) \leq X$$

This bound is tight in the case of independent measurements.

**Proposition 2** (Eluder dimension for linear functions).

Let  $\mathcal{F} = \{f \mid f(x) = \theta\phi(x) \text{ for } \theta \in \mathbb{R}^{n \times p}, \phi \in \mathbb{R}^p, \|\theta\|_2 \leq C_\theta, \|\phi\|_2 \leq C_\phi\}$  then  $\forall \mathcal{X}$ :

$$\dim_E(\mathcal{F}, \epsilon) \leq p(4n - 1) \frac{e}{e - 1} \log \left[ \left( 1 + \left( \frac{2C_\phi C_\theta}{\epsilon} \right)^2 \right) (4n - 1) \right] + 1 = \tilde{O}(np)$$

**Proposition 3** (Eluder dimension for quadratic functions).

Let  $\mathcal{F} = \{f \mid f(x) = \phi(x)^T \theta \phi(x) \text{ for } \theta \in \mathbb{R}^{p \times p}, \phi \in \mathbb{R}^p, \|\theta\|_2 \leq C_\theta, \|\phi\|_2 \leq C_\phi\}$  then  $\forall \mathcal{X}$ :

$$\dim_E(\mathcal{F}, \epsilon) \leq p(4p - 1) \frac{e}{e - 1} \log \left[ \left( 1 + \left( \frac{2pC_\phi^2 C_\theta}{\epsilon} \right)^2 \right) (4p - 1) \right] + 1 = \tilde{O}(p^2).$$

**Proposition 4** (Eluder dimension for generalized linear functions).

Let  $g(\cdot)$  be a component-wise independent function on  $\mathbb{R}^n$  with derivative in each component bounded  $\in [\underline{h}, \bar{h}]$  with  $\underline{h} > 0$ . Define  $r = \frac{\bar{h}}{\underline{h}} > 1$  to be the condition number. If  $\mathcal{F} = \{f \mid f(x) = g(\theta\phi(x)) \text{ for } \theta \in \mathbb{R}^{n \times p}, \phi \in \mathbb{R}^p, \|\theta\|_2 \leq C_\theta, \|\phi\|_2 \leq C_\phi\}$  then for any  $\mathcal{X}$ :

$$\dim_E(\mathcal{F}, \epsilon) \leq p(r^2(4n - 2) + 1) \frac{e}{e - 1} \left( \log \left[ (r^2(4n - 2) + 1) \left( 1 + \left( \frac{2C_\theta C_\phi}{\epsilon} \right)^2 \right) \right] \right) + 1 = \tilde{O}(r^2 np)$$

## 5 Confidence sets

We will now build confidence sets which contain the true function  $f^*$  with high probability. The confidence sets are centered around the least squares estimate  $\hat{f}_t^{LS} \in \arg \min_{f \in \mathcal{F}} L_{2,t}(f)$  where  $L_{2,t}(f) := \sum_{i=1}^{t-1} \|f(x_i) - y_i\|_2^2$  is the cumulative squared prediction error. The confidence sets are defined  $\mathcal{F}_t = \mathcal{F}_t(\beta_t) := \{f \in \mathcal{F} \mid \|f - \hat{f}_t^{LS}\|_{2,E_t} \leq \sqrt{\beta_t}\}$  where  $\beta_t$  is a parameter which controls the growth of the confidence set and the empirical 2-norm is defined  $\|g\|_{2,E_t}^2 := \sum_{i=1}^{t-1} \|g(x_i)\|_2^2$ .

For  $\mathcal{F} \subseteq \mathcal{P}_{\mathcal{X},\mathcal{Y}}^{C,\sigma}$ , we define the control parameter:

$$\beta_t^*(\mathcal{F}, \delta, \alpha) := 8\sigma^2 \log(N(\mathcal{F}, \alpha, \|\cdot\|_2)/\delta) + 2\alpha t \left( 8C + \sqrt{8\sigma^2 \log(4t^2/\delta)} \right) \quad (5)$$

This leads to confidence sets which contain the true function with high probability.

**Proposition 5** (Confidence sets with high probability).

For all  $\delta > 0$  and  $\alpha > 0$  and the confidence sets  $\mathcal{F}_t = \mathcal{F}_t(\beta_t^*(\mathcal{F}, \delta, \alpha))$  for all  $t \in \mathbb{N}$  then:

$$\mathbb{P} \left( f^* \in \bigcap_{t=1}^{\infty} \mathcal{F}_t \right) \geq 1 - 2\delta$$

*Proof.* Combine elementary martingale concentrations with a discretization argument. The argument is essentially the same as Proposition 6 in [11], but replaces statements about  $\mathbb{R}$  with more general  $L^2$  properties. A full derivation is available in the Appendix A.  $\square$

## 5.1 Bounding the sum of set widths

We now bound the deviation from  $f^*$  by the maximum deviation within the confidence set.

**Definition 3** (Set widths).

For any set of functions  $\mathcal{F}$  we define the width of the set at  $x$  to be the maximum L2 deviation between any two members of  $\mathcal{F}$  evaluated at  $x$ .

$$w_{\mathcal{F}}(x) := \sup_{\bar{f}, \underline{f} \in \mathcal{F}} \|\bar{f}(x) - \underline{f}(x)\|_2$$

We can bound for the number of large widths in terms of the eluder dimension.

**Lemma 1** (Bounding the number of large widths).

If  $\{\beta_t > 0 | t \in \mathbb{N}\}$  is a nondecreasing sequence with  $\mathcal{F}_t = \mathcal{F}_t(\beta_t)$  then

$$\sum_{k=1}^m \sum_{i=1}^{\tau} \mathbb{1}\{w_{\mathcal{F}_{t_k}}(x_{t_k+i}) > \epsilon\} \leq \left( \frac{4\beta_T}{\epsilon^2} + \tau \right) \dim_E(\mathcal{F}, \epsilon)$$

*Proof.* This result follows from proposition 8 in [11] but with a small adjustment to account for episodes. A full proof is given in Appendix B.  $\square$

We now use Lemma 1 to control the cumulative deviation through time.

**Proposition 6** (Bounding the sum of widths).

If  $\{\beta_t > 0 | t \in \mathbb{N}\}$  is a nondecreasing sequence with  $\mathcal{F}_t = \mathcal{F}_t(\beta_t)$  and  $\|f\|_2 \leq C$  for all  $f \in \mathcal{F}$  then:

$$\sum_{k=1}^m \sum_{i=1}^{\tau} w_{\mathcal{F}_{t_k}}(x_{t_k+i}) \leq 1 + \tau C \dim_E(\mathcal{F}, T^{-1}) + 4\sqrt{\beta_T \dim_E(\mathcal{F}, T^{-1})T} \quad (6)$$

*Proof.* Once again we follow the analysis of Russo [11] and streamline notation by letting  $w_t = w_{\mathcal{F}_{t_k}}(x_{t_k+i})$  and  $d = \dim_E(\mathcal{F}, T^{-1})$ . Reordering the sequence  $(w_1, \dots, w_T) \rightarrow (w_{i_1}, \dots, w_{i_T})$  such that  $w_{i_1} \geq \dots \geq w_{i_T}$  we have that:

$$\sum_{k=1}^m \sum_{i=1}^{\tau} w_{\mathcal{F}_{t_k}}(x_{t_k+i}) = \sum_{t=1}^T w_{i_t} \leq 1 + \sum_{i=1}^T w_{i_t} \mathbb{1}\{w_{i_t} \geq T^{-1}\}$$

By the reordering we know that  $w_{i_t} > \epsilon$  means that  $\sum_{k=1}^m \sum_{i=1}^{\tau} \mathbb{1}\{w_{\mathcal{F}_{t_k}}(x_{t_k+i}) > \epsilon\} \geq t$ .

From Lemma 1,  $\epsilon \leq \sqrt{\frac{4\beta_T d}{t - \tau d}}$ . So that if  $w_{i_t} > T^{-1}$  then  $w_{i_t} \leq \min\{C, \sqrt{\frac{4\beta_T d}{t - \tau d}}\}$ . Therefore,

$$\sum_{i=1}^T w_{i_t} \mathbb{1}\{w_{i_t} \geq T^{-1}\} \leq \tau C d + \sum_{t=\tau d+1}^T \sqrt{\frac{4\beta_T d}{t - \tau d}} \leq \tau C d + 2\sqrt{\beta_T} \int_0^T \sqrt{\frac{d}{t}} dt \leq \tau C d + 4\sqrt{\beta_T d T}$$

$\square$

## 6 Analysis

We will now follow the familiar line of argument that relates the regret of an optimistic or posterior sampling algorithm to the bellman error [7, 8]. The first step breaks down the regret by adding and subtracting the *imagined* optimal reward of  $\mu_k$  under the MDP  $M_k$ . We streamline our discussion of  $P^M, R^M, V_{\mu,i}^M, U_{\mu,i}^M$  and  $\mathcal{T}_{\mu}^M$  by simply writing  $*$  in place of  $M^*$  or  $\mu^*$  and  $k$  in place of  $M_k$  or  $\mu_k$  where appropriate; for example  $V_{k,i}^* := V_{\mu_k,i}^{M^*}$ .

$$\Delta_k = (V_{*,1}^* - V_{k,1}^*)(s_0) = (V_{*,1}^* - V_{k,1}^k)(s_0) + (V_{k,1}^k - V_{k,1}^*)(s_0) \quad (7)$$

Here  $s_0$  is a distinguished initial state, but moving to general  $\rho(s)$  poses no real challenge.

Algorithms based upon optimism bound  $(V_{*,1}^* - V_{k,1}^k) \leq 0$  with high probability. For PSRL we use Lemma 2 and the tower property to see that this is zero in expectation.

**Lemma 2** (Posterior sampling).

If  $\phi$  is the distribution of  $M^*$  then, for any  $\sigma(H_{t_k})$ -measurable function  $g$ ,

$$\mathbb{E}[g(M^*)|H_{t_k}] = \mathbb{E}[g(M_k)|H_{t_k}] \quad (8)$$

We introduce the Bellman operator  $\mathcal{T}_\mu^M$ , which for any MDP  $M = (\mathcal{S}, \mathcal{A}, R^M, P^M, \tau, \rho)$ , stationary policy  $\mu : \mathcal{S} \rightarrow \mathcal{A}$  and value function  $V : \mathcal{S} \rightarrow \mathbb{R}$ , is defined by

$$\mathcal{T}_\mu^M V(s) := \bar{r}^M(s, \mu(s)) + \int_{s' \in \mathcal{S}} P^M(s'|s, \mu(s)) V(s').$$

This returns the expected value of state  $s$  where we follow the policy  $\mu$  under the laws of  $M$ , for one time step. The following lemma gives a concise form for the dynamic programming paradigm in terms of the Bellman operator.

**Lemma 3** (Dynamic programming equation).

For any MDP  $M = (\mathcal{S}, \mathcal{A}, R^M, P^M, \tau, \rho)$  and policy  $\mu : \mathcal{S} \times \{1, \dots, \tau\} \rightarrow \mathcal{A}$ , the value functions  $V_\mu^M$  satisfy

$$V_{\mu,i}^M = \mathcal{T}_{\mu(\cdot,i)}^M V_{\mu,i+1}^M \quad (9)$$

for  $i = 1 \dots \tau$ , with  $V_{\mu,\tau+1}^M := 0$ .

Through repeated application of the dynamic programming operator and taking expectation of martingale differences we can mirror earlier analysis [8] to equate expected regret with the cumulative bellman error:

$$\mathbb{E}[\Delta_k] = \sum_{i=1}^{\tau} (\mathcal{T}_{k,i}^k - \mathcal{T}_{k,i}^*) V_{k,i+1}^k(s_{t_k+i}) \quad (10)$$

## 6.1 Lipschitz continuity

Efficient regret bounds for MDPs with an infinite number of states and actions require some regularity assumption. One natural notion is that the optimal value function of nearby states cannot differ too much. We express this through Lipschitz continuity.

**Definition 4** (Lipschitz continuity).

A function  $U : \mathcal{Y} \rightarrow \mathbb{R}$ , for  $(\mathcal{Y}, \|\cdot\|_2)$   $L^2$ , is said to be  $K$ -Lipschitz if for all  $y, y' \in \mathcal{Y}$ :

$$|U(y) - U(y')| \leq K \|y - y'\|_2.$$

In many cases of interest the optimal value function of the MDP will not be Lipschitz. For example, any discontinuous reward function will usually lead to discontinuous values. Instead we suppose that the *future* value is Lipschitz with respect to the norm  $\|\mathbb{E}[\cdot]\|_2$ . This is trivially satisfied if the underlying value function is Lipschitz, but is a strictly weaker assumption as the system noise helps to smooth out the expected value of nearby states.

Since  $\mathcal{P}$  has  $\sigma_P$ -sub-Gaussian noise we can write  $s_{t+1} = \bar{p}^M(s_t, a_t) + \epsilon_t^P$  in the natural way. We now use equation (10) to reduce regret to a sum of set widths. To reduce clutter and more closely follow the notation of Section 4 we will write  $x_{k,i} = (s_{t_k+i}, a_{t_k+i})$ .

$$\begin{aligned} \mathbb{E}[\Delta_k] &\leq \mathbb{E} \left[ \sum_{i=1}^{\tau} \{ \bar{r}^k(x_{k,i}) - \bar{r}^*(x_{k,i}) + U_i^k(P^k(x_{k,i})) - U_i^k(P^*(x_{k,i})) \} \right] \\ &\leq \mathbb{E} \left[ \sum_{i=1}^{\tau} \{ |\bar{r}^k(x_{k,i}) - \bar{r}^*(x_{k,i})| + K_k \|\bar{p}^k(x_{k,i}) - \bar{p}^*(x_{k,i})\|_2 \} \right] \end{aligned} \quad (11)$$

Where  $K_k$  is a global Lipschitz constant for the future value functions  $U_i^k$ ,  $i = 1, \dots, \tau$ .

We now use the results from Sections 4 and 5 to form the corresponding confidence sets  $\mathcal{R}_k := \mathcal{R}_{t_k}(\beta^*(\mathcal{R}, \delta, \alpha))$  and  $\mathcal{P}_k := \mathcal{P}_{t_k}(\beta^*(\mathcal{P}, \delta, \alpha))$  for the reward and transition functions

respectively. Let  $A = \{R^*, R_k \in \mathcal{R}_k \forall k\}$  and  $B = \{P^*, P_k \in \mathcal{P}_k \forall k\}$  and condition upon these events to give:

$$\begin{aligned} \mathbb{E}[\text{Regret}(T, \pi^{PS}, M^*)] &\leq \mathbb{E} \left[ \sum_{k=1}^m \sum_{i=1}^{\tau} \left\{ |\bar{r}^k(x_{k,i}) - \bar{r}^*(x_{k,i})| + K_k \|\bar{p}^k(x_{k,i}) - \bar{p}^*(x_{k,i})\|_2 \right\} \right] \\ &\leq \sum_{k=1}^m \sum_{i=1}^{\tau} \{w_{\mathcal{R}_k}(x_{k,i}) + \mathbb{E}[K_k | A, B] w_{\mathcal{P}_k}(x_{k,i}) + 8\delta(C_{\mathcal{R}} + C_{\mathcal{P}})\} \end{aligned} \quad (12)$$

Posterior sampling ensures that  $\mathbb{E}[K_k] = \mathbb{E}[K^*]$  so that  $\mathbb{E}[K_k | A, B] \leq \frac{\mathbb{E}[K^*]}{\mathbb{P}(A, B)} \leq \frac{\mathbb{E}[K^*]}{1 - \delta}$  by a union bound on  $\{A^c \cup B^c\}$ . We fix  $\delta = 1/8T$  to see that:

$$\mathbb{E}[\text{Regret}(T, \pi^{PS}, M^*)] \leq (C_{\mathcal{R}} + C_{\mathcal{P}}) + \sum_{k=1}^m \sum_{i=1}^{\tau} w_{\mathcal{R}_k}(x_{k,i}) + \mathbb{E}[K^*] \left(1 + \frac{1}{T-1}\right) \sum_{k=1}^m \sum_{i=1}^{\tau} w_{\mathcal{P}_k}(x_{k,i})$$

We now use equation (5) together with Proposition 6 to obtain our regret bounds. For ease of notation we will write  $d_{\mathcal{R}} = \dim_E(\mathcal{R}, T^{-1})$  and  $d_{\mathcal{P}} = \dim_E(\mathcal{P}, T^{-1})$ .

$$\begin{aligned} \mathbb{E}[\text{Regret}(T, \pi^{PS}, M^*)] &\leq 2 + (C_{\mathcal{R}} + C_{\mathcal{P}}) + \tau(C_{\mathcal{R}} d_{\mathcal{R}} + C_{\mathcal{P}} d_{\mathcal{P}}) + \\ &\quad 4\sqrt{\beta_T^*(\mathcal{R}, 1/8T, \alpha) d_{\mathcal{R}} T} + 4\sqrt{\beta_T^*(\mathcal{P}, 1/8T, \alpha) d_{\mathcal{P}} T} \end{aligned} \quad (13)$$

We let  $\alpha = 1/T^2$  and write  $n_{\mathcal{F}} = \log(8N(\mathcal{F}, 1/T^2, \|\cdot\|_2)T)$  for  $\mathcal{R}$  and  $\mathcal{P}$  to complete our proof of Theorem 1:

$$\begin{aligned} \mathbb{E}[\text{Regret}(T, \pi^{PS}, M^*)] &\leq [C_{\mathcal{R}} + C_{\mathcal{P}}] + \left[1 + \tau C_{\mathcal{R}} d_{\mathcal{R}} + 8\sqrt{d_{\mathcal{R}}(4C_{\mathcal{R}} + \sqrt{2\sigma_{\mathcal{R}}^2 \log(32T^3)})} + 8\sqrt{2\sigma_{\mathcal{R}}^2 n_{\mathcal{R}} d_{\mathcal{R}} T}\right] \\ &\quad + \mathbb{E}[K^*] \left(1 + \frac{1}{T-1}\right) \left[1 + \tau C_{\mathcal{P}} d_{\mathcal{P}} + 8\sqrt{d_{\mathcal{P}}(4C_{\mathcal{P}} + \sqrt{2\sigma_{\mathcal{P}}^2 \log(32T^3)})} + 8\sqrt{2\sigma_{\mathcal{P}}^2 n_{\mathcal{P}} d_{\mathcal{P}} T}\right] \end{aligned}$$

The first term  $[C_{\mathcal{R}} + C_{\mathcal{P}}]$  bounds the contribution from missed confidence sets. The second set of square brackets bounds the cost of learning the reward function  $R^*$ . In most problems the remaining contribution from transitions and lost future value will be dominant. To clarify the asymptotic dependence we will now introduce another common notion of dimension.

**Definition 5.** *The Kolmogorov dimension of a function class  $\mathcal{F}$  is given by:*

$$\dim_K(\mathcal{F}) := \limsup_{\alpha \downarrow 0} \frac{\log(N(\mathcal{F}, \alpha, \|\cdot\|_2))}{\log(1/\alpha)}.$$

Using this definition together with  $n_{\mathcal{R}}, n_{\mathcal{P}}$  we complete our proof of Corollary 1.

## 7 Conclusions

We present the first unified analysis for model-based reinforcement in terms of the dimensionality of the function classes of rewards  $\mathcal{R}$  and transitions  $\mathcal{P}$ . What is more, we show that the simple and computationally efficient algorithm PSRL satisfies these bounds. Our results provide new state of the guarantees when specialized to several important problem settings but in others, such as factored MDPs, alternative analysis can produce tighter bounds even for the same algorithm. Additionally, there is a fundamental concern for model-based reinforcement learning that even solving for the optimal policy in a given MDP may be intractable. In future work, we would like to examine whether it is possible to attain similar regret bounds with a model-free algorithm when the optimal value function  $V^*$  lies in  $\mathcal{V}$  to avoid this computation. We also wonder whether it is possible to extend our analysis to learning in MDPs without episodic resets.



## References

- [1] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.
- [2] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.
- [3] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *arXiv preprint cs/9605103*, 1996.
- [4] M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.
- [5] R. I. Brafman and M. Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *The Journal of Machine Learning Research*, 3:213–231, 2003.
- [6] Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888. ACM, 2006.
- [7] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *The Journal of Machine Learning Research*, 99:1563–1600, 2010.
- [8] Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) Efficient Reinforcement Learning via Posterior Sampling. *Advances in Neural Information Processing Systems*, 2013.
- [9] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 3:397–422, 2003.
- [10] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12:1587–1627, 2011.
- [11] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *CoRR*, abs/1301.2609, 2013.
- [12] Alexander L Strehl and Michael L Littman. Online linear regression and its application to model-based reinforcement learning. In *NIPS*, 2007.
- [13] Yassin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.
- [14] Morteza Ibrahimi, Adel Javanmard, and Benjamin Van Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. In *NIPS*, pages 2645–2653, 2012.
- [15] Ronald Ortner, Daniil Ryabko, et al. Online regret bounds for undiscounted continuous reinforcement learning. In *NIPS*, pages 1772–1780, 2012.
- [16] Dan 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.
- [17] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.
- [18] M. Strens. A Bayesian framework for reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, pages 943–950, 2000.
- [19] Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, MA, 1995.