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). We show that, if the MDP can be parameterized within some known function class, we can obtain regret bounds that scale with the dimensionality, rather than cardinality, of the system. We characterize this dependence explicitly as $\tilde{O}(\sqrt{d_K d_E T})$ where T is time elapsed, d_K is the Kolmogorov dimension and d_E is the *eluder dimension*. This represents the first unified framework for model-based reinforcement learning and provides state of the art guarantees in several important settings. Moreover, we present a simple and computationally efficient algorithm *posterior sampling for reinforcement learning* (PSRL) that satisfies these bounds.

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 does know that they lie within some 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]. However, model-free approaches often struggle to plan efficient exploration so that currently the most efficient regret bounds are given by model-based approaches. 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].

The problem with this *tabula rasa* setting is that 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 this falls victim to the curse of dimensionality. To achieve efficiency results that can exploit MDP structure, the existing literature usually presents algorithms tailored to some specific families of $\mathcal{R}$ and $\mathcal{P}$.

The most widely-studied parameterization is the degenerate MDP with no transitions, the mutli-armed bandit [9, 10, 11]. Another common assumption is that the transition function 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 exponential dependence, but only under significant sparsity assumptions [14]. The most general previous analysis considers rewards and transitions that are α-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 $\mathcal{P}$ in terms of the dimensionality, rather than the cardinality, of the MDP. To characterize the complexity of this learning problem we extend the definition of the eluder dimension, previously introduced for bandits, to capture the interedependence between state-action pairs [16]. Our 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 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, τ is the time horizon, and ρ the initial state distribution. We define the MDP and all other random variables we will consider with respect to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

For our unified analysis we will require that the state space $\mathcal{S}$ is represented as some L^2 space. In any MDP with states in $\mathbb{R}^n$ this is completely natural. For discrete MDPs over $\{1, .., S\}$, we will consider states $s_t \in \mathcal{S} := [0, 1]^S$. In this case, the traditional transition function is now given by the *expectation* of our transition distribution $\bar{p}^M(s, a) := \mathbb{E}\left[s'|s' \sim P^M(\cdot|s, a)\right]$.

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

$$V_{\mu,i}^M(s) := \mathbb{E}_{M,\mu}\left[\sum_{j=i}^{\tau} \bar{r}^M(s_j, a_j)\Big|s_i = s\right] \tag{1}$$

where $\bar{r}^M(s, a) := \mathbb{E}[r|r \sim R^M(s, a)]$ 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 $j = i, \ldots, \tau$. A policy μ 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 $i = 1, \ldots, \tau$. We will associate with each MDP M a policy μ^M that is optimal for M.

We also define the future value function U to be the expected value of the optimal policy following a single transition according to a distribution Q,

$$U_i^M(Q) := \mathbb{E}_{M,\mu^M}\left[V_{\mu^M,i+1}^M(s)\big|s \sim Q\right]. \tag{2}$$

In most settings of interest, similar transition distributions $Q, \tilde{Q}$ will have similar future values. We formalize this idea through the notion of Lipschitz continuity in the norm of expectation $\|\mathbb{E}[\cdot]\|_2$ over transition distributions such that for all $Q, \tilde{Q} \in \mathcal{P}$:

$$\|U_i^M(Q) - U_i^M(\tilde{Q})\|_2 \leq K_i^M(\mathcal{P})\|\mathbb{E}[Q - \tilde{Q}]\|_2. \tag{3}$$

We define $K^M := \max_i K_i^M(\mathcal{P})$ to be a global Lipschitz contant for the future value function.

The reinforcement learning agent interacts with the MDP over episodes that begin at times $t_k = (k-1)\tau + 1$, $k = 1, 2, \ldots$. 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 μ then when in state s at time t during episode k, it selects an action $a_t = \mu(s, t - t_k)$. Let $H_t = (s_1, a_1, r_1, \ldots, s_{t-1}, a_{t-1}, r_{t-1})$ denote the history of observations made *prior* to time t. A

reinforcement learning algorithm is a deterministic sequence $\{\pi_k | k = 1, 2, \ldots\}$ 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 kth episode. We define the regret incurred by a reinforcement learning algorithm π up to time T to be

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

where Δ_k denotes regret over the kth 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 kth episode, PSRL samples and MDP M_k from the posterior[1]. 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 ϕ for M^*, t=1
2: **for** episodes $k = 1, 2, ..$ **do**
3:     sample $M_k \sim \phi(\cdot | H_t)$
4:     compute $\mu_k = \mu^{M_k}$
5:     **for** timesteps $j = 1, .., \tau$ **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**

To state our results we first introduce some notation. 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 the distributions from $\mathcal{X}$ to $\mathcal{Y}$ with mean norm bounded in $[0, C]$ and additive σ-sub-Gaussian noise. For the family $\mathcal{F}$ let $N(\mathcal{F}, \alpha, \|\cdot\|_2)$ be the α-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 $\mathcal{F}$ at precision T^{-1}, 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^* be an MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, rewards $R^* \in \mathcal{R} \subseteq \mathcal{P}_{\mathcal{S} \times \mathcal{A}, \mathbb{R}}^{C_{\mathcal{R}}, \sigma_{\mathcal{R}}}$ and transitions $P^* \in \mathcal{P} \subseteq \mathcal{P}_{\mathcal{S} \times \mathcal{A}, \mathcal{S}}^{C_{\mathcal{P}}, \sigma_{\mathcal{P}}}$ for any $C_{\mathcal{R}}, C_{\mathcal{P}}, \sigma_{\mathcal{R}}, \sigma_{\mathcal{P}} > 0$. If ϕ is the distribution of M^*

---

[1]Without conjugate priors this can be acheived by algorithms such as Metropolis-Hastings.

and K^* is a global Lipschitz constant for the future value function then:

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

$$+ \mathbb{E}[K^*]\left(1 + \tfrac{1}{T-1}\right)\left(1 + \tau C_\mathcal{P} d_\mathcal{P} + 8\sqrt{d_\mathcal{P}(4C_\mathcal{P} + \sqrt{2\sigma_\mathcal{P}^2 \log(32T^3)})} + 8\sqrt{2\sigma_\mathcal{P}^2 n_\mathcal{P} d_\mathcal{P} T}\right) \quad (4)$$

To clarify the assymptotic 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 4 in Theorem 1 we can obtain our Corollary.

Corollary 1 (Assymptotic regret bounds for PSRL in parameterized MDPs)**.**
Let M^* be an MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, rewards $R^* \in \mathcal{R} \subseteq \mathcal{P}_{\mathcal{S} \times \mathcal{A}, \mathbb{R}}^{C_\mathcal{R}, \sigma_\mathcal{R}}$ and transitions $P^* \in \mathcal{P} \subseteq \mathcal{P}_{\mathcal{S} \times \mathcal{A}, \mathcal{S}}^{C_\mathcal{P}, \sigma_\mathcal{P}}$ for any $C_\mathcal{R}, C_\mathcal{P}, \sigma_\mathcal{R}, \sigma_\mathcal{P} > 0$. If ϕ is the distribution of M^* and K^* is a global Lipschitz constant for the future value function then:

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

$$(5)$$

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 (Assymptotic 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 σ-sub-Gaussian noise. If ϕ is the distribution of M^*, then:

$$\mathbb{E}[\text{Regret}(T, \pi^{PS}, M^*)] = \tilde{O}\left(\sigma C \lambda_1 n^2 \sqrt{T}\right) \quad (6)$$

Where λ_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 $C\lambda_1$. □

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 any L^2 space. For any $\mathcal{G} \subseteq \mathcal{P}_{\mathcal{X}, \mathcal{Y}}^{C, \sigma}$ we define $\mathcal{F} = \mathbb{E}[\mathcal{G}] := \{f | f = \mathbb{E}[G] \text{ for } G \in \mathcal{G}\}$. We consider sequential observations $y_i \sim G^*(x_i)$ which can equivalently be written $y_i = f^*(x_i) + \epsilon_i$ for $f^*(x_i) = \mathbb{E}[G^*(x_i)]$. Intuitively, the eluder dimension of $\mathcal{F}$ is the length d of the longest possible sequence $x_1, .., x_d$ such that for all i, knowing the function values of $f(x_1), .., f(x_i)$ will not reveal $f(x_{i+1})$.

Definition 1 (($\mathcal{F}, \epsilon) - dependence$)**.**
We will say that $x \in \mathcal{X}$ is $(\mathcal{F}, \epsilon)$-dependent on $\{x_1, ..., x_n\} \subseteq \mathcal{X}$

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

4

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

Definition 2 (Eluder Dimension).
The eluder dimension $\dim_E(\mathcal{F}, \epsilon)$ *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. We overload our notation to write $\dim_E(\mathcal{G}, \epsilon) := \dim_E(\mathbb{E}[\mathcal{G}], \epsilon)$ whenever $\mathcal{G} \subseteq \mathcal{P}_{\mathcal{X}, \mathcal{Y}}^{C, \sigma}$, which should be clear from the context.

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 analysis. 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 $\in [\underline{h}, \overline{h}]$ with $\underline{h} > 0$. Define $r = \frac{\overline{h}}{\underline{h}} > 1$ 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 $\mathcal{X}$:
$$\dim_E(\mathcal{F}, \epsilon) \leq p\left(r^2(4n-2)+1\right)\frac{e}{e-1}\left(\log\left[\left(r^2(4n-2)+1\right)\left(1 + \left(\frac{2C_\theta C_\phi}{\epsilon}\right)^2\right)\right]\right) + 1 = \tilde{O}(r^2 np)$$

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 $L_{2,t}(f) := \sum_{i=1}^{t-1} \|f(x_t) - y_t\|_2^2$ is the cumulative squared prediciton 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 β_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:
$$\beta_t^*(\mathcal{F}, \delta, \alpha) := 8\sigma^2 \log(N(\mathcal{F}, \alpha, \|\cdot\|_2)/\delta) + 2\alpha t\left(8C + \sqrt{8\sigma^2 \log(4t^2/\delta)}\right) \tag{7}$$

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. We combine standard martingale concentrations with a discretization scheme. 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. $\qquad\square$

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.

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

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. $\qquad\square$

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:

$$\sum_{k=1}^{m} \sum_{i=1}^{\tau} w_{\mathcal{F}_{t_k}}(x_{t_k+i}) \leq 1 + \tau C \dim_E(\mathcal{F}, T^{-1}) + 4\sqrt{\beta_T \dim_E(\mathcal{F}, T^{-1})T} \qquad (8)$$

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

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

$$\square$$

6

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]. We streamline our discussion of $P^M, R^M, V_{\mu,i}^M, U_i^M$ and $\mathcal{T}_\mu^M$ by simply writing $*$ in place of M^* or μ^* and k in place of M_k or μ_k where appropriate; for example $V_{k,i}^* := V_{\tilde{\mu}_k,i}^{M^*}$.

The first step breaks down the regret by adding and subtracting the *imagined* optimal reward of μ_k under the MDP M_k.

$$\Delta_k = \left(V_{*,1}^* - V_{k,1}^*\right)(s_0) = \left(V_{*,1}^* - V_{k,1}^k\right)(s_0) + \left(V_{k,1}^k - V_{k,1}^*\right)(s_0) \tag{9}$$

Here s_0 is a distinguished initial state, but moving to general $\rho(s)$ poses no real challenge. Algorithms based upon optimism bound $\left(V_{*,1}^* - V_{k,1}^k\right) \leq 0$ 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 ϕ is the distribution of M^* then, for any $\sigma(H_{t_k})$-measurable function g,

$$\mathbb{E}[g(M^*)|H_{t_k}] = \mathbb{E}[g(M_k)|H_{t_k}] \tag{10}$$

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 $\mu : \mathcal{S} \to \mathcal{A}$ and value function $V : \mathcal{S} \to \mathbb{R}$, 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 μ 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 $\mu : \mathcal{S} \times \{1, \ldots, \tau\} \to \mathcal{A}$, the value functions V_μ^M satisfy

$$V_{\mu,i}^M = \mathcal{T}_{\mu(\cdot,i)}^M V_{\mu,i+1}^M \tag{11}$$

for $i = 1 \ldots \tau$, 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:

$$\mathbb{E}[\Delta_k] = \sum_{i=1}^\tau (\mathcal{T}_{k,i}^k - \mathcal{T}_{k,i}^*)V_{k,i+1}^k(s_{t_k+i}) \tag{12}$$

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 nearby states might have similar optimal values, or that the optimal value function function might be Lipschitz. However, any discontinuous reward function will usually lead to discontious values functions so that this assumption is violated in many settings of interest.

Instead we suppose that the *future* value is Lipschitz in the sense of equation (3). This will will be satisfied whenever the underlying value function is Lipschitz, but is a strictly weaker requirement since the system noise helps to smooth future values.

Since $\mathcal{P}$ has σ_P-sub-Gaussian noise we write $s_{t+1} = \bar{p}^M(s_t, a_t) + \epsilon_t^P$ in the natural way. We now use equation (12) to reduce regret to a sum of set widths. To reduce clutter and more closely follow the notation of Section 4 we will write $x_{k,i} = (s_{t_k+i}, a_{t_k+i})$.

$$
\begin{aligned}
\mathbb{E}[\Delta_k] &\leq \mathbb{E}\left[\sum_{i=1}^\tau \left\{\bar{r}^k(x_{k,i}) - \bar{r}^*(x_{k,i}) + U_i^k(P^k(x_{k,i})) - U_i^*(P^*(x_{k,i}))\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^\tau \left\{|\bar{r}^k(x_{k,i}) - \bar{r}^*(x_{k,i})| + K^k\|\bar{p}^k(x_{k,i}) - \bar{p}^*(x_{k,i})\|_2\right\}\right] \tag{13}
\end{aligned}
$$

Where K^k is a global Lipschitz constant for the future value function of M_k.

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:

$$
\begin{aligned}
\mathbb{E}[\text{Regret}(T, \pi^{PS}, M^*)] &\leq \mathbb{E}\left[\sum_{k=1}^{m} \sum_{i=1}^{\tau} \left\{ |\bar{r}^k(x_{k,i}) - \bar{r}^*(x_{k,i})| + K^k \|\bar{p}^k(x_{k,i}) - \bar{p}^*(x_{k,i})\|_2 \right\} \right] \\
&\leq \sum_{k=1}^{m} \sum_{i=1}^{\tau} \left\{ w_{\mathcal{R}_k}(x_{k,i}) + \mathbb{E}[K^k | A, B] w_{\mathcal{P}_k}(x_{k,i}) + 8\delta(C_{\mathcal{R}} + C_{\mathcal{P}}) \right\} \quad (14)
\end{aligned}
$$

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

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

We now use equation (7) 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 $d_{\mathcal{P}} = \dim_E(\mathcal{P}, T^{-1})$.

$$
\begin{aligned}
\mathbb{E}[\text{Regret}(T, \pi^{PS}, M^*)] &\leq 2 + (C_{\mathcal{R}} + C_{\mathcal{P}}) + \tau(C_{\mathcal{R}} d_{\mathcal{R}} + C_{\mathcal{P}} d_{\mathcal{P}}) + \\
&\quad 4\sqrt{\beta_T^*(\mathcal{R}, 1/8T, \alpha) d_{\mathcal{R}} T} + 4\sqrt{\beta_T^*(\mathcal{P}, 1/8T, \alpha) d_{\mathcal{P}} T} \quad (15)
\end{aligned}
$$

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:

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

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 assymptotic dependence we will now introduce another common notion of dimension.

Definition 4. *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 [20]. We also wonder whether it is possible to extend our analysis to learning in MDPs without episodic resets. However, there is a more fundamental quagmire for model-based reinforcement learning in complex systems that even solving for the optimal policy in a *known* 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 $\mathcal{V}$ [21].

References

[1] Apostolos Burnetas and Michael Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

[2] Tze Leung Lai and Herbert 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] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

[5] Ronen Brafman and Moshe 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 Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael 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] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 3:397–422, 2003.

[10] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12:1587âĂŞ1627, 2011.

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

[17] William 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] Malcom Strens. A Bayesian framework for reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, pages 943–950, 2000.

[19] Dimitri Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, MA, 1995.

[20] Ian Osband and Benjamin Van Roy. Near-optimal regret bounds for reinforcement learning in factored MDPs. *arXiv preprint arXiv:1403.3741*, 2014.

[21] Benjamin Van Roy and Zheng Wen. Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*, 2014.

A   Confidence sets with high probability

In this appendix we will build up to a proof of Proposition 5, that the confidence sets defined by β^* in equation 7 hold with high probability. We begin with some elementary results from martingale theory.

Lemma 4 (Exponential Martingale).
Let $Z_i \in L^1$ be real-calued random variables adapted to $\mathcal{H}_i$. We define the conditional mean $\mu_i = \mathbb{E}[Z_i|\mathcal{H}_{i-1}]$ and conditional cumulant generating function $\psi_i(\lambda) = \log \mathbb{E}[\exp(\lambda(Z_i - \mu_i))|\mathcal{H}_{i-1}]$, then

$$M_n(\lambda) = \exp\left(\sum_1^n \lambda(Z_i - \mu_i) - \psi_i(\lambda)\right)$$

is a martingale with $\mathbb{E}[M_n(\lambda)] = 1$.

Lemma 5 (Concentration Guarantee).
For Z_i adapted real L^1 random variables adapted to $\mathcal{H}_i$. We define the conditional mean $\mu_i = \mathbb{E}[Z_i|\mathcal{H}_{i-1}]$ and conditional cumulant generating function $\psi_i(\lambda) = \log \mathbb{E}[\exp(\lambda(Z_i - \mu_i))|\mathcal{H}_{i-1}]$.

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty}\{\sum_1^n \lambda(Z_i - \mu_i) - \psi_i(\lambda) \geq x\}\right) \leq e^{-x}$$

Both of these lemmas are available in earlier discussion for real-valued variables [11]. We now specialize our discussion to the L^2-space $\mathcal{Y}$ where $<y,y> = \|y\|_2^2$. To simplify notation we will write $f_t^* := f^*(x_t)$ and $f_t = f(x_t)$ for arbitrary $f \in \mathcal{F}$. We now define

$$
\begin{aligned}
Z_t &= \|f_t^* - y_t\|_2 - \|f_t - y_t\|_2 \\
&= <f_t^* - y_t, f_t^* - y_t> - <f_t - y_t, f_t - y_t, f_t - y_t> \\
&= -<f_t - f_t^*, f_t - f_t^*> + 2<f_t - f_t^*, y_t - f_t^*> \\
&= -\|f_t - f_t^*\|_2 + 2<f_t - f_t^*, \epsilon_t>
\end{aligned}
$$

so that clearly $\mu_t = -\|f_t - f_t^*\|_2$. Now since we have said that the noise is σ-sub-Gaussian, $\mathbb{E}[\exp(<\phi,\epsilon>)] \leq \exp\left(\frac{\|\phi\|_2^2 \sigma^2}{2}\right)$ $\forall \phi \in \mathcal{Y}$. From here we can deduce that:

$$
\begin{aligned}
\psi_t(\lambda) &= \log \mathbb{E}[\exp(\lambda(Z_t - \mu_t))|\mathcal{H}_{t-1}] \\
&= \log \mathbb{E}[\exp(2\lambda <f_t - f_t^*, \epsilon_t>)] \\
&\leq \frac{\|2\lambda(f_t - f_t^*)\|_2^2 \sigma^2}{2}.
\end{aligned}
$$

We now write $\sum_{i=1}^{t-1} Z_i = L_{2,t}(f^*) - L_{2,t}(f)$ according to our earlier definition of $L_{2,t}$. We can apply Lemma 5 with $\lambda = 1/4\sigma^2$, $x = log(1/\delta)$ to obtain:

$$\mathbb{P}\{\left(L_{2,t}(f) \geq L_{2,t}(f^*) + \frac{1}{2}\|f - f^*\|_{2,E_t} - 4\sigma^2 \log(1/\delta)\right) \ \forall t\} \geq 1 - \delta$$

substituting $f = \hat{f}$ to be the least squares solution which minimizes $L_{2,t}(f)$ we can remove $L_{2,t}(\hat{f}) - L_{2,t}(f^*) \leq 0$. From here we use an α-cover discretization argument to complete the proof of Proposition 5.

Let $\mathcal{F}^{\alpha} \subset \mathcal{F}$ be an α-2 cover of $\mathcal{F}$ such that $\forall f \in \mathcal{F}$ there is some $\|f^{\alpha} - f\|_2 \leq \alpha$. We can use a union bound on $\mathcal{F}^{\alpha}$ so that $\forall f \in \mathcal{F}$:

$$L_{2,t}(f) - L_{2,t}(f^*) \geq \frac{1}{2}\|f - f^*\|_{2,E_t} - 4\sigma^2 \log(N(\mathcal{F}, \alpha, \|\cdot\|_2)/\delta) + DE(\alpha) \tag{16}$$

$$\text{For } DE(\alpha) = \min_{f^{\alpha} \in \mathcal{F}^{\alpha}} \left\{\frac{1}{2}\|f^{\alpha} - f^*\|_{2,E_t}^2 - \frac{1}{2}\|f - f^*\|_{2,E_t}^2 + L_{2,t}(f) - L_{2,t}(f^{\alpha})\right\}$$

We will now seek to bound this discretization error with high probability.

Lemma 6 (Bounding discretization error).
If $\|f^{\alpha}(x) - f(x)\|_2 \leq \alpha$ for all $x \in \mathcal{X}$ then with probability at least $1 - \delta$:

$$DE(\alpha) \leq \alpha t \left[8C + \sqrt{8\sigma^2 \log(4t^2/\delta)}\right]$$

Proof. For non-trivial bounds we will consider the case of $\alpha \leq C$ and note that via Cauchy-Schwarz:

$$\|f^\alpha(x)\|_2^2 - \|f(x)\|_2^2 \leq \max_{\|y\|_2 \leq \alpha} \|f(x) + y\|_2^2 - \|f\|_2^2 \leq 2C\alpha + \alpha^2.$$

From here we can say that

$$\|f^\alpha(x) - f^*(x)\|_2^2 - \|f(x) - f^*(x)\|_2^2 = \|f^\alpha(x)\|_2^2 - \|f(x)\|_2^2 + 2 < f^*(x), f(x) - f^\alpha(x) > \leq 4C\alpha$$

$$\|y - f(x)\|_2^2 - \|y - f^\alpha(x)\|_2^2 = 2 < y, f^\alpha(x) - f(x) > + \|f(x)\|_2^2 - \|f^\alpha(x)\|_2^2 \leq 2\alpha|y| + 2C\alpha + \alpha^2$$

Summing these expressions over time $i = 1, .., t-1$ and using sub-gaussian high probability bounds on $|y|$ gives our desired result. $\qquad\square$

Finally we apply Lemma 6 to equation 16 and use the fact that $\hat{f}_t^{LS}$ is the $L_{2,t}$ minimizer to obtain the result that with probability at least $1 - 2\delta$:

$$\|\hat{f}_t^{LS} - f^*\|_{2,E_t} \leq \sqrt{\beta_t^*(\mathcal{F}, \alpha, \delta)}$$

Which is our desired result.

B   Bounding the number of large widths

Lemma 1 (Bounding the number of large widths).
If $\{\beta_t > 0 \big| t \in \mathbb{N}\}$ 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. We first imagine that $w_{\mathcal{F}_t}(x_t) > \epsilon$ and is ϵ-dependent on K disjoint subsequences of $x_1, .., x_{t-1}$. If x_t is ϵ-dependent on K disjoint subsequences then there exist $\|\overline{f} - \underline{f}\|_{2,E_t} > K\epsilon^2$. By the triangle inequality $\|\overline{f} - \underline{f}\|_{2,E_t} \leq 2\sqrt{\beta_t} \leq 2\sqrt{\beta_T}$ so that $K < 4\beta_T/\epsilon^2$.

In the case without episodic delay, Russo went on to show that in any sequence of length l there is some element which is ϵ-dependent on at least $\frac{l}{\dim_E(\mathcal{F}, \epsilon)} - 1$ disjoint subsequences [11]. Our analysis follows similarly, but we may lose up to $\tau - 1$ proper subsequences due to the delay in updating the episode. This means that we can only say that $K \geq \frac{l}{\dim_E(\mathcal{F}, \epsilon)} - \tau$. Considering the subsequence $w_{\mathcal{F}_{t_k}}(x_{t_k+i}) > \epsilon$ we see that $l \leq \left(\frac{4\beta_T}{\epsilon^2} + \tau\right) \dim_E(\mathcal{F}, \epsilon)$ as required. $\qquad\square$

C   Eluder dimension for specific function classes

In this section of the appendix we will provide bounds upon the eluder dimension for some canonical function classes. Recalling Definition 2, $\dim_E(\mathcal{F}, \epsilon)$ is the length d of the longest sequence $x_1, .., x_d$ such that for some $\epsilon' \geq \epsilon$:

$$w_k = \sup\left\{\|(\overline{f} - \underline{f})(x_k)\|_2 \,\bigg|\, \|\overline{f} - \underline{f}\|_{2,E_t} \leq \epsilon'\right\} > \epsilon' \tag{17}$$

for each $k \leq d$.

C.1   Finite domain $\mathcal{X}$

Any $x \in \mathcal{X}$ is ϵ-dependent upon itself for all $\epsilon > 0$. Therefore for all $\epsilon > 0$ the eluder dimension of $\mathcal{F}$ is bounded by $|\mathcal{X}|$.

C.2   Linear functions $f(x) = \theta\phi(x)$

Let $\mathcal{F} = \{f \,|\, f(x) = \theta\phi(x)$ for $\theta \in \mathbb{R}^{n \times p}, \phi \in \mathbb{R}^p, \|\theta\|_2 \leq C_\theta, \|\phi\|_2 \leq C_\phi\}$. To simplify our notation we will write $\phi_k = \phi(x_k)$ and $\theta = \theta_1 - \theta_2$. From here, we may manipulate the expression

$$\|\theta\phi\|_2^2 = \phi_k^T \theta^T \theta \phi_k = Tr(\phi_k^T \theta^T \theta \phi_k) = Tr(\theta \phi_k \phi_k^T \theta)$$

$$\implies w_k = \sup_\theta \{\|\theta\phi_k\|_2 \,\big|\, Tr(\theta\Phi_k\theta^T) \leq \epsilon^2\} \text{ where } \Phi_k := \sum_{i=1}^{k-1} \phi_i \phi_i^T$$

We next require a lemma which gives an upper bound for trace constrained optimizations.

Lemma 7 (Bounding norms under trace constraints).
Let $\theta \in \mathbb{R}^{n \times p}, \phi \in \mathbb{R}^p$ and $V \in \mathbb{R}^{p \times p}_{++}$, the set of positive definite $p \times p$ matrices, then:

$$W^2 = \max_\theta \|\theta\phi\|_2^2 \text{ subject to } Tr(\theta V \theta^T) \leq \epsilon^2$$

is bounded above by $W^2 \leq (2n-1)\epsilon^2 \|\phi\|_{V^{-1}}^2$ where $\|\phi\|_A^2 := \phi^T A \phi$.

Proof. We first note that $\|\theta\phi\|_2^2 = Tr(\theta\phi\phi^T\theta^T) = \sum_1^n (\theta\phi)_i^2 \leq \left(\sum_1^n (\theta\phi)_i\right)^2$ by Jensen's inequality. We define $\tilde{\Phi} \in \mathbb{R}^{n \times p}$ such that each row of $\tilde{\Phi} = \phi^T$. Then this inequality can be expressed as:

$$W^2 = Tr(\theta\phi\phi^T\theta^T) \leq Sum(\theta \otimes \tilde{\Phi})^2$$

Where $A \otimes B = C$ for $C_{ij} = A_{ij}B_{ij}$ and $Sum(C) := \sum_{i,j} C_{ij}$ We can now obtain an upper bound for our original problem through this convex relaxation:

$$\max_\theta Sum(\theta \otimes \tilde{\Phi}) \text{ subject to } Tr(\theta V \theta^T) \leq \epsilon^2$$

We can now form the lagrangian $\mathcal{L}(\theta, \lambda) = -Sum(\theta \otimes \tilde{\Phi}) + \lambda(Tr(\theta V \theta^T) - \epsilon^2)$. Solving for first order optimality $\nabla_\theta \mathcal{L} = 0 \implies \theta^* = \frac{1}{2\lambda}\tilde{\Phi}V^{-1}$. From here we form the dual objective

$$g(\lambda) = -Sum(\frac{1}{2\lambda}\tilde{\Phi}V^{-1} \otimes \tilde{\Phi}) + Tr(\frac{1}{4\lambda}\tilde{\Phi}V^{-1}\tilde{\Phi}^T) - \lambda\epsilon^2$$

Here we solve for the dual-optimal λ^* $\nabla_\lambda g = 0 \implies \frac{1}{2\lambda^*}^2 Sum(\frac{1}{2\lambda}\tilde{\Phi}V^{-1} \otimes \tilde{\Phi}) - \frac{1}{4\lambda^*}^2 Tr(\frac{1}{4\lambda}\tilde{\Phi}V^{-1}\tilde{\Phi}^T) = \epsilon^2$. From the definition of $\tilde{\Phi}$, $Sum(\tilde{\Phi}V^{-1} \otimes \tilde{\Phi}) = n\phi^T V^{-1}\phi$ and $Tr(\tilde{\Phi}V^{-1}\tilde{\Phi}^T) = \phi^T V^{-1}\phi$. From this we can simplify our expression to conclude:

$$\frac{n}{2\lambda^{*2}}\phi^T V^{-1}\phi - \frac{1}{4\lambda^{*2}}\phi^T V^{-1}\phi = \epsilon^2 \implies \lambda^* = \sqrt{\frac{(n-1/2)}{2\epsilon^2}}\|\phi\|_{V^{-1}}$$

$$\implies g(\lambda^*) = -\frac{n}{2\lambda^*}\|\phi\|_{V^{-1}}^2 + \frac{1}{4\lambda^*}\|\phi\|_{V^{-1}}^2 - \lambda^*\epsilon$$

$$\text{strong duality} \implies f(\theta^*) = g(\lambda^*) = \sqrt{2n-1}\epsilon\|\phi\|_{V^{-1}}$$

From here we conclude that the optimal value of $W^2 \leq f(\theta^*)^2 \leq (2n-1)\epsilon^2 \|\phi\|_{V^{-1}}^2$. $\qquad\square$

Using this lemma, we will be able to address the eluder dimension for linear functions. Using the definition of w_k from equation 17 together with Φ_k we may rewrite:

$$w_k = \max_\theta \{\sqrt{Tr(\theta\phi_k\phi_k^T\theta)} \mid Tr(\theta\Phi_k\theta^T) \leq \epsilon^2\}.$$

Let $V_k := \Phi_k + \left(\frac{\epsilon}{2C_\theta}\right)^2 I$ so that $Tr(\theta\Phi_k\theta^T) \leq \epsilon^2 \implies Tr(\theta V_k \theta^T) \leq 2\epsilon^2$ through a triangle inequality. Now applying Lemma 7 we can say that $w_k \leq \epsilon\sqrt{4n-2}\|\phi_k\|_{V_k^{-1}}$. This means that if $w_k \geq \epsilon$ then $\|\phi_k\|_{V_k^{-1}}^2 > \frac{1}{4n-2} > 0$.

We now imagine that $w_i \geq \epsilon$ for each $i < k$. Then since $V_k = V_{k-1} + \phi_k\phi_k^T$ we can use the Matrix Determinant together with the above observation to say that:

$$det(V_k) = det(V_{k-1})(1 + \phi_k^T V_K^{-1}\phi_k) \geq det(V_{k-1})\left(1 + \frac{1}{4n-2}\right) \geq .. \geq \lambda^p \left(1 + \frac{1}{4n-2}\right)^{k-1} \quad (18)$$

for $\lambda := \left(\frac{\epsilon}{2C_\theta}\right)^2$. To get an upper bound on the determinant we note that $det(V_k)$ is maximized when all eigenvalues are equal or equivalently:

$$det(V_k) \leq \left(\frac{Tr(V_k)}{p}\right)^p \leq \left(\frac{C_\phi^2(k-1)}{p} + \lambda\right)^p \quad (19)$$

Now using equations 18 and 19 together we see that k must satistfy the inequality $\left(1 + \frac{1}{4n-2}\right)^{(k-1)/p} \leq \frac{C_\phi^2(k-1)}{\lambda p} + 1$. We now write $\zeta_0 = \frac{1}{4n-2}$ and $\alpha_0 = \frac{C_\phi^2}{\lambda} = \left(\frac{2C_\phi C_\theta}{\epsilon}\right)^2$ so that we can epress this as:

$$(1 + \zeta_0)^{\frac{k-1}{p}} \leq \alpha_0 \frac{k-1}{p} + 1$$

We now use the result that $B(x, \alpha) = \max\{B \mid (1+x)^B \leq \alpha B + 1\} \leq \frac{1+x}{x} \frac{e}{e-1} \{\log(1+\alpha) + \log(\frac{1+x}{x})\}$. We complete our proof of Proposition 2 through computing this upper bound at (ζ_0, α_0),

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

C.3 Quadratic functions $f(x) = \phi^T(x)\theta\phi(x)$

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

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

Where we have simply applied the linear result with $\tilde{\epsilon} = \frac{\epsilon}{pC_\mathcal{P}}$. This is valid since if we can identify the linear function $g(x) = \theta\phi(x)$ to within this tolerance then we will certainly know $f(x)$ as well.

C.4 Generalized linear models

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

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

This proof follows exactly as per the linear case, but first using a simple reduction on the form of equation (17).

$$
\begin{aligned}
w_k &= \sup\left\{\|(\overline{f} - \underline{f})(x_k)\|_2 \;\middle|\; \|\overline{f} - \underline{f}\|_{2, E_t} \leq \epsilon'\right\} \\[2mm]
&\leq \max_{\theta_1, \theta_2}\left\{\|g(\theta_1\phi_k) - g(\theta_2\phi_k)\|_2 \;\middle|\; \sum_{i=1}^{k-1}\|g(\theta_1\phi_i) - g(\theta_2\phi_i)\|_2^2 \leq \epsilon^2\right\} \\[2mm]
&\leq \max_\theta\left\{\overline{h}\|\theta\phi_k\|_2 \;\middle|\; \sum_{i=1}^{k-1}\underline{h}^2\|\theta\phi_i\|_2^2 \leq \epsilon^2\right\}
\end{aligned}
$$

To which we can now apply Lemma 7 with the ϵ rescaled by r. Following the same arguments as for linear functions now completes our proof.

D UCRL-Eluder

For completeness, we explicitly outline an optimistic algorithm which uses the confidence sets in our analysis of PSRL to guarantee similar regret bounds with high probability over all MDP M^*. The algorithm follows the style of UCRL2 [7] so that at the start of the kth episode the algorithm form $\mathcal{M}_k = \{M \mid P^M \in \mathcal{P}_k, R^M \in \mathcal{R}_k\}$ and then solves for the optimistic policy that attains the highest reward over any M in $\mathcal{M}_k$.

Algorithm 2
UCRL-Eluder

1: **Input:** Confidence parameter $\delta > 0$, t=1
2: **for** episodes $k = 1, 2, ..$ **do**
3:     form confidence sets $\mathcal{R}_k(\beta^*(\mathcal{R}, \delta, 1/k^2))$ and $\mathcal{P}_k(\beta^*(\mathcal{P}, \delta, 1/k^2))$
4:     compute μ_k optimistic policy over $\mathcal{M}_k = \{M \mid P^M \in \mathcal{P}_k, R^M \in \mathcal{R}_k\}$
5:     **for** timesteps $j = 1, .., \tau$ **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**

Generally, step 4 of this algorithm with not be computationally tractable even when solving for μ^M is possible for a given M.