MODEL-BASED REINFORCEMENT LEARNING AND THE ELUDER DIMENSION

IAN OSBAND AND BENJAMIN VAN ROY STANFORD UNIVERSITY

ABSTRACT

Any algorithm that applies to all MDPs will suffer $\Omega(\sqrt{|\mathcal{S}||\mathcal{A}|T})$ regret on some MDP. So what do we do when $|\mathcal{S}|, |\mathcal{A}|$ are extremely large or infinite? The curse of dimensionality means our only hope is to exploit some low-dimensional structure.

We show that if the MDP can be parameterized within some known function class, we obtain regret bounds that scale with the dimensionality, rather than cardinality, of the system. We characterize this dependence explicitly in terms of the eluder dimension. We also present a simple and computationally efficient algorithm (PSRL) that satisfies these bounds. These are the first regret bounds for general model-based learning.

PROBLEM FORMULATION

Learn to optimize a random finite horizon MDP M in repeated finite episodes of interaction.

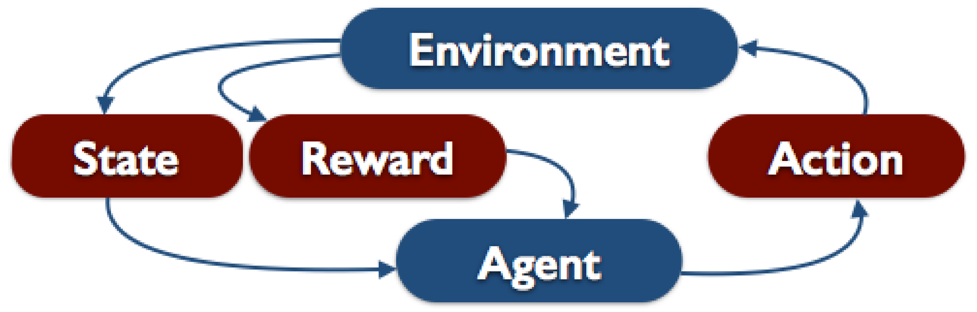


Figure 1: classic reinforcement learning setting

- State space $\mathcal{S}$, action space $\mathcal{A}$
- Rewards $r_t \sim R^M(s_t, a_t) \in \mathcal{R}$
- Transitions $s_{t+1} \sim P^M(s_t, a_t) \in \mathcal{P}$
- Epsiode length τ, define $t_k := (k-1)\tau + 1$

For MDP M and policy μ, define a value function

$$V^M_{\mu,i}(s) := \mathbb{E}_{M,\mu}\left[\sum_{j=i}^{\tau} \overline{R}^M(s_j, a_j)\Big| s_i = s\right],$$

Define the regret in episode k using μ_k on M^*

$$\Delta_k := \int_{\mathcal{S}} \rho(s)\Big(\underbrace{V^{M^*}_{\mu^*,1}(s)}_{\text{optimal value}} - \underbrace{V^{M^*}_{\mu_k,1}(s)}_{\text{actual value}}\Big)$$

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

Naive exploration such as Boltzman or ϵ-greedy can lead to exponential regret. Good performance requires balancing **exploration vs exploitation**.

ELUDER DIMENSION

Eluder principle: a measurement at x is independent of $\{x_1, .., x_n\}$ if functions that are similar at $\{x_1, .., x_n\}$ could differ significantly at x.

Definition 1 (($\mathcal{F}, \epsilon$) $-$ *dependence*).
We will say that $x \in \mathcal{X}$ is $(\mathcal{F}, \epsilon)$-dependent on $\{x_1, ..., x_n\} \subseteq \mathcal{X} \iff \forall f, \tilde{f} \in \mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}^n\}$

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

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

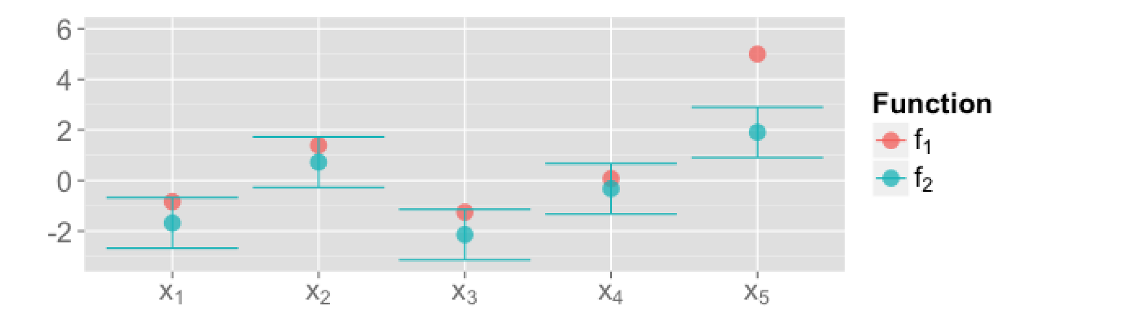


Figure 2: x_5 is $(\{f_1, f_2\}, 1)$-independent of $\{x_1, .., x_4\}$.

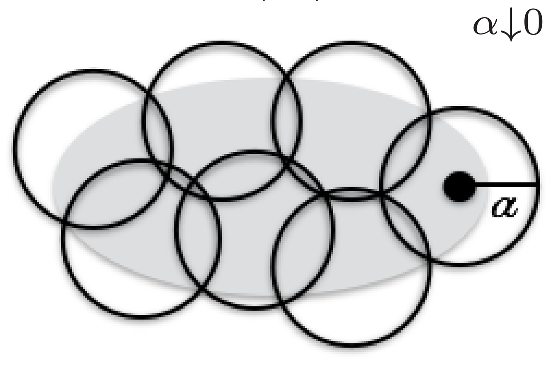
Definition 2 (Eluder Dimension $= \dim_E(\mathcal{F}, \epsilon)$).
The length of the longest possible sequence of elements in $\mathcal{X}$ such that for some $\epsilon' \geq \epsilon$ every element is $(\mathcal{F}, \epsilon')$-independent of its predecessors.

Examples
- $\mathcal{X}$ finite $\implies \dim_E(\mathcal{F}, \epsilon) \leq |\mathcal{X}|$.
- $\mathcal{F} \subseteq \{f : \mathbb{R}^n \to \mathbb{R}^p \text{ linear}\}$
  $\implies \dim_E(\mathcal{F}, \epsilon) = O(np\log(1/\epsilon))$

KOLMOGOROV DIMENSION

The Kolmogorov dimension of a function class $\mathcal{F}$:

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



In this diagram $N(\mathcal{F}, \alpha, \|\cdot\|_2) \leq 7$

Example: $\dim_K(\mathbb{R}^d) = d$

LIPSCHITZ SMOOTHNESS

Definition 3 (Future value function U_i^M).
For any distribution Φ over $\mathcal{S}$ we define:

$$U_i^M(\Phi) := \mathbb{E}_{M,\mu^M}\left[V^M_{\mu^M,i+1}(s) | s \sim \Phi\right]$$

as the value of the optimal policy, starting from Φ.

- Learning an infinite MDP requires regularity
- Assume $U_i^{M^*}$ is Lipschitz in $\mathbb{E}[s|s \sim \Phi]$ wrt $\|\cdot\|_2$
- Satisfied whenever $V^{M^*}_{\mu^*,i}$ Lipschitz in s wrt $\|\cdot\|_2$
- But this is a strictly weaker condition since system noise can help smooth future value.

POSTERIOR SAMPLING

For each episode k:

1. Sample an MDP from the posterior distribution for the true MDP: $M_k \sim \phi(\cdot|H_t)$.
2. Use policy $\mu_k \in \arg\max_{\mu} V_\mu^{M_k}$.

MAIN RESULTS

If M^* is an MDP with rewards $R^* \in \mathcal{R}$ and transitions $P^* \in \mathcal{P}$ with sub σ-Gaussian noise then the expected regret to time T of PSRL is bounded:

$$\tilde{O}\Big(\underbrace{\sigma_{\mathcal{R}}\sqrt{d_K(\mathcal{R})d_E(\mathcal{R})T}}_{\text{rewards}} + \underbrace{\mathbb{E}[K^*]}_{\text{Lipschitz}}\underbrace{\sigma_{\mathcal{P}}\sqrt{d_K(\mathcal{P})d_E(\mathcal{P})T}}_{\text{transitions}}\Big)$$

Notation:
- Kolmogorov dimension $d_K(\mathcal{F}) := \dim_K(\mathcal{F})$
- Eluder dimension $d_E(\mathcal{F}) := \dim_E(\mathcal{F}, T^{-1})$
- Lipschitz constant K^* for future value function

Corollary:
Let M^* be a linear-quadratic system in $\mathbb{R}^d$ with σ-sub-Gaussian noise mean-bounded by C then:

$$\mathbb{E}[\text{Regret}(T, \pi^{PS}, M^*)] = \underbrace{\tilde{O}\left(\sigma C d^2 \sqrt{T}\right)}_{\text{no exponential scaling in } d}.$$

REFERENCES

Please see the full paper:
http://arxiv.org/abs/1406.1853

PROOF SKETCH

We consider the regret in an episode k:

$$\begin{aligned}\Delta_k &= V^*_{*,1}(s) - V^*_{k,1}(s) \\ &= \underbrace{\left(V^k_{k,1}(s) - V^*_{k,1}(s)\right)}_{\text{Imagined - Actual}} + \underbrace{\left(V^*_{*,1}(s) - V^k_{k,1}(s)\right)}_{\mathbb{E}[\cdot]=0 \text{ by posterior}}\end{aligned}$$

We can decompose this into Bellman error:

$$V^k_{k,1} - V^*_{k,1} = \underbrace{\sum_{i=1}^{\tau}\left(\mathcal{T}^k_{k,i} - \mathcal{T}^*_{k,i}\right)V^k_{k,i+1}}_{B:=\text{Bellman error}} + \underbrace{\sum_{i=1}^{\tau} d_{t_k+1}}_{\mathbb{E}=0 \text{ martingale}}.$$

We can now use the Hölder inequality to bound:

$$B \leq \sum_{i=1}^{\tau}\Big\{\underbrace{|\overline{R}^k - \overline{R}^*|}_{\text{reward error}} + \underbrace{K^k}_{\text{Lipschitz}}\underbrace{\|P^k - P^*\|_2}_{\text{transition error}}\Big\}$$

We conclude the proof by upper bounding these deviations in terms of our estimation errors on R^* and P^*. We use concentration inequalities to express the error bounds for $\mathcal{R}$ and $\mathcal{P}$ in terms of the eluder dimension and Kolmogorov dimension.

Note a proof for a similar optimistic algorithm is possible, however this would require a generally intractable planning step. We believe that the sampling approach will also be more statistically efficient since it is not affected by loose analysis.

SO WHAT?

- **Practical reinforcement learning problems often have $|\mathcal{S}|$ and $|\mathcal{A}|$ very large or infinite.**
- **"Tabula rasa" learning will always require minimum $T = \Omega(|\mathcal{S}||\mathcal{A}|)$ for good guarantees $\implies$ must exploit low-dimensional structure.**
- **We produce a unified analysis for model-based RL in terms of the dimensionality, rather than the cardinality, of the system.**
- **Conceptually simple, computationally efficient algorithm PSRL satisfies these bounds.**

CONTACT INFORMATION

Web www.stanford.edu/~iosband
Email iosband@stanford.edu