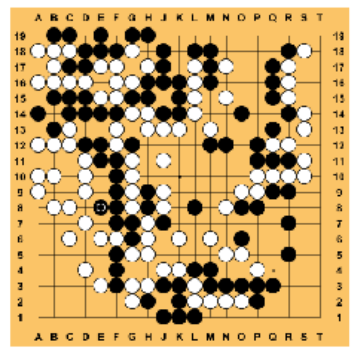
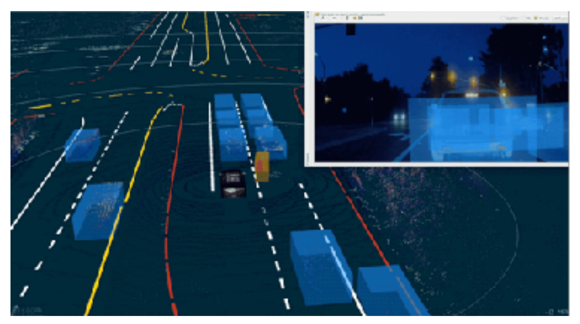
Model-based RL in Contextual Decision Process: PAC Bounds and Exponential Improvements over Model-free Approaches

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford

Motivations





1. Difference between model-based & model-free RL beyond tabular setting
2. Global exploration in large-scale MDPs w/ function approximation

Contextual Decision Processes

In this work, we consider MDPs with an extremely large state space $\mathcal{X}$ (hence $\text{poly}(|\mathcal{X}|)$ is intractable)

- Finite number of actions, horizon H;
- Context/State space $\mathcal{X}$
- Policy: $\pi : \mathcal{X} \to \Delta(\mathcal{A})$
- Transition $P^\star : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$
- reward $r^\star : \mathcal{X} \times \mathcal{A} \to [0,1]$

Model-based RL setting

A model is a pair of transition & reward: $M \triangleq (P, r)$

Given: a class of models $\mathcal{M}$, with $(P^\star, r^\star) \in \mathcal{M}$

Goal: learn a near-optimal policy $V^\pi \geq V^\star - \epsilon$ w/ # of sample:

$$\text{poly}(H, |\mathcal{A}|, 1/\epsilon, \log(|\mathcal{M}|))$$

(i.e., no explicit poly dependency on # of states)

Note: realizability itself is not enough to achieve the goal

Definition of Model-free Algorithms

Model-free Alg takes a function class $\mathcal{G} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ as input, accesses state x via G-profile: $\Phi_\mathcal{G} = \{g(x,a)\}_{g\in\mathcal{G}, a\in\mathcal{A}}$

- When $\mathcal{G}$ = policy class: policy gradient (e.g., REINFORCE, gradient can be computed from finite differencing)
- When $\mathcal{G}$ = Q-function class: (Delayed) Q-learning, OLIVE
- When $\mathcal{G}$ = Q-function + Policy: Actor-Critic methods

Intuition of the Definition

G-profile could obfuscate the context, leading to information loss in function approximation setting (but not in tabular setting)

Why Model-based RL

Formalize the Inputs:

- ***Optimal Planning oracle:*** $\text{OP}(M) = (\pi^M, Q^M)$
- Model-based methods take $\mathcal{M}$ as input
- Model-free methods take $\mathcal{G} \triangleq \text{OP}(\mathcal{M})$ as input

Informal Statement (Theorem 2)

There exists a family of MDPs, where a model-based alg can learn in poly sample complexity, while any model-free alg suffers an exponential sample complexity $\Omega(2^H)$

Remark:

Our lower bound does not hold when:
(1) model-free algs take some $\mathcal{G} \neq \text{OP}(\mathcal{M})$ as input
(2) when $\mathcal{M}$ is "over-parameterized" s.t. G-profile reveals state

Witness Rank

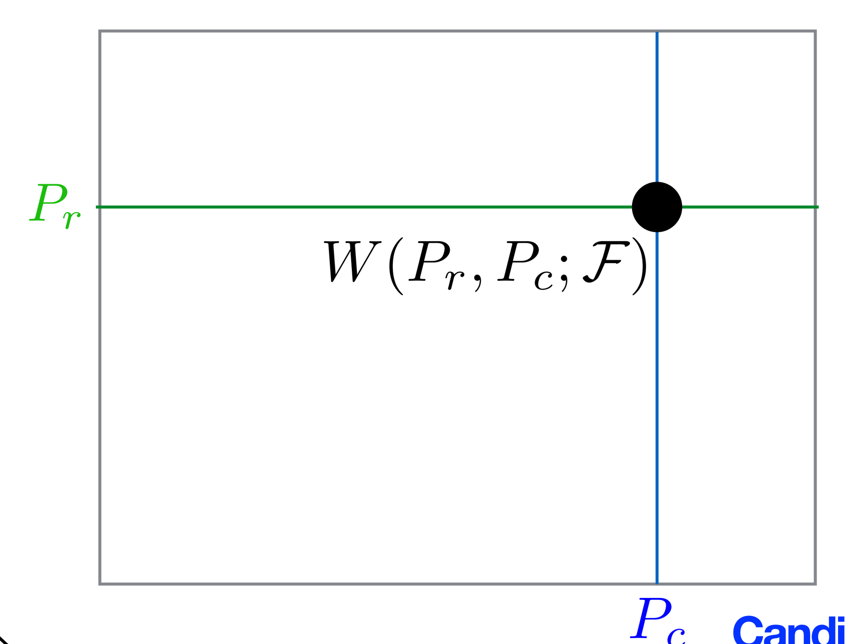
(for simplicity, from now on, we assume reward is known, model class just contains transitions)

Introduce a Witness function class:

$$\mathcal{F} = \{f : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}\}$$

Misfit Matrix:

Provides Conditional State-action Distribution



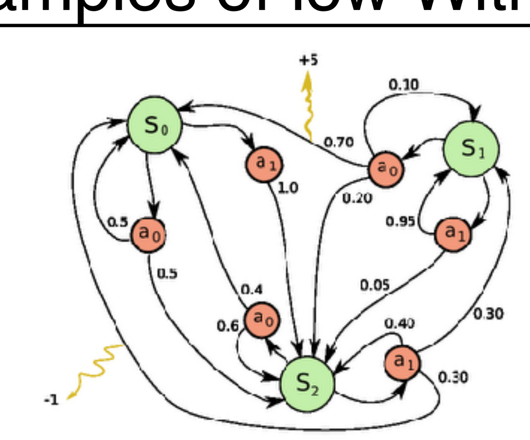
$$W(P_r, P_c; \mathcal{F}) = \max_{f\in\mathcal{F}} \mathbb{E}_{x\sim\pi_{P_r}, a\sim U}\left[\mathbb{E}_{x'\sim P_c} f(x,a,x') - \mathbb{E}_{x'\sim P^\star} f(x,a,x')\right]$$

Imaginary | Real

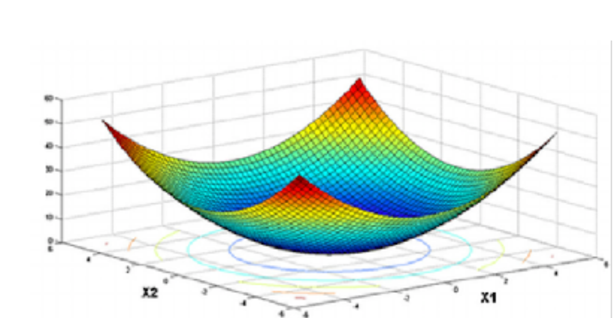
This is an Integral Probability Metric (IPM)
(Discriminators try to tell how real a transition from P_c is)

Witness Rank is defined as the rank of this misfit matrix

Examples of low Witness Rank:



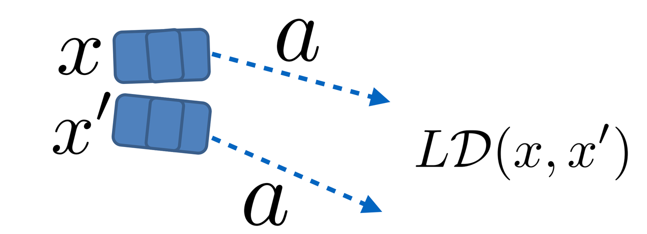
Small Discrete MDP
Rank <= # of state



Linear Quadratic Regulator
Rank <= O(d^2)

Examples (continued):

Factored MDPs
[Guestrin et.al, 03; Osband & Van Roy,13]
Rank <= exp(in-degree)



Lipschitz Continuous MDPs
[Kearn, Langford, Kakade, 03]
Rank <= Covering number of state space

…and any MDPs with low Bellman Rank (Jiang et al. 17)

Algorithm & Analysis

Optimistic model selection:
$$P_n = \arg\max_{P\in\mathcal{P}_n} V_P$$

Samples from real world via $\pi_n = \text{OP}(P_n)$
$$\{x, a, x'\}_N$$
$$x \sim \pi_n, a \sim U, x' \sim P^\star$$

Estimate Model Misfit
$$W(P_n, P; \mathcal{F}), \forall P \in \mathcal{P}_n$$

Eliminate models whose misfits are large
$$P : W(P_n, P; \mathcal{F}) \geq \delta$$

Terminate if $|V_{P_n} - V^{\pi_n}|$ is small, i.e., the current policy has similar values under P_n & $P^\star$

Sample Complexity:

Witness Rank

$$\tilde{O}\left(\frac{H^3 R^2 |\mathcal{A}|^2}{\epsilon^2} \log\left(\frac{|\mathcal{F}||\mathcal{P}|}{\delta}\right)\right)$$

Extensions

1. Refine analysis w/ conditional Scheffe Estimator to handle Total Variation distance (i.e., $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$), and sample complexity is reduced to:
$$\tilde{O}\left(\frac{H^3 R^2 |\mathcal{A}|}{\epsilon^2} \log\left(\frac{|\mathcal{P}|}{\delta}\right)\right)$$
2. Doubling trick to deal with unknown model rank
3. Refine witness rank to ensure it's never larger than Bellman rank (Jiang et al, 17)
4. Can handle approximate low-rank misfit matrices