

Sample Complexity of Policy Search with Known Dynamics

*Summer 2007 Program on Challenges in Dynamic Treatment Regimes
and Multistage Decision-Making*

SAMSI, June 2007

Peter Bartlett

EECS and Statistics

UC Berkeley

Ambuj Tewari

EECS

UC Berkeley

{bartlett, ambuj} AT cs.berkeley.edu

Introduction

- Markov Decision Processes (MDPs) are often used to model sequential decision making under uncertainty
- At any time, decision maker (DM) is in a state $s_t \in S$ and has to choose an action $a_t \in A$
- The probability of moving to state s' after taking action a_t in state s_t is given by the *transition probability* $p(s'|s_t, a_t)$
- When in state s_t , DM collects *reward* drawn from a distribution $r(s_t)$ with support in $[0, R]$
- A (randomized) *policy* $\pi : S \rightarrow \mathcal{P}(A)$ is a rule used by DM to take actions

Performance Criteria

- Suppose policy π is followed starting from an initial state s_0 to generate a *trajectory*

$$s_0, s_1, s_2, \dots$$

- Several performance criteria can be considered (note: $\rho_t \sim r(s_t)$)

- *Finite Horizon*

$$V(\pi) := \mathbb{E} \left[\sum_{t=0}^{N-1} \rho_t \right]$$

- Infinite Horizon: *Discounted*

$$V(\pi) := \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \rho_t \right]$$

- Infinite Horizon: *Average*

$$V(\pi) := \lim_{N \rightarrow \infty} \frac{\mathbb{E} \left[\sum_{t=0}^{N-1} \rho_t \right]}{N}$$

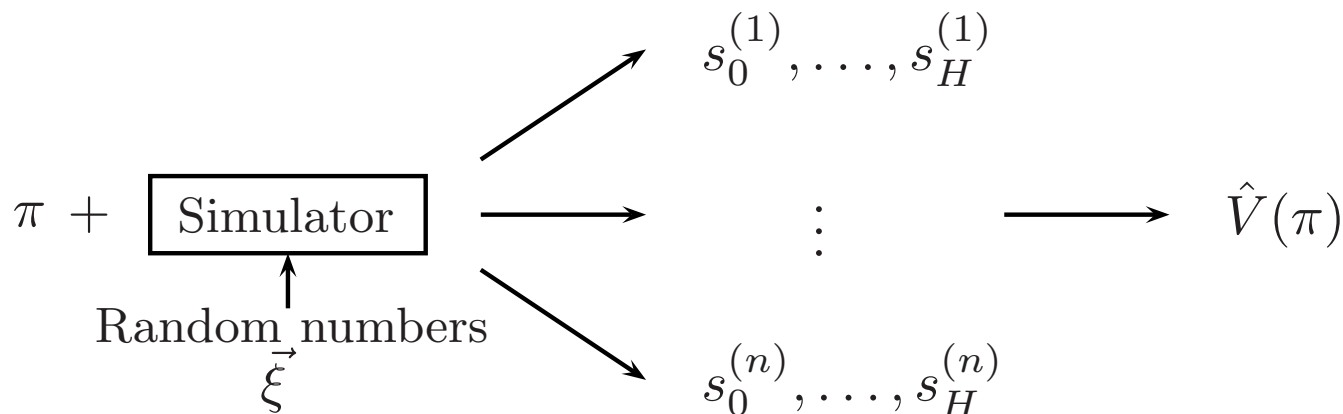
Outline

- Handling large state spaces with the help of *policy search* using *simulators*
- *Sample Complexity*: how many trajectories do we need to find a good policy?
- Reducing the problem to finding *uniform convergence* bounds using ideas from Learning Theory
- Why the problem is harder to analyze compared to function or concept learning?
- Final Bound

Our Setting

If we have a *good simulator* for a dynamical system (say, a discounted MDP) with a *large state space*, then we might do the following:

- Choose a class Π of policies.
- Use n trajectories to get an estimate $\hat{V}(\pi)$ of the true performance $V(\pi)$ of every policy π in the policy class.
- $\hat{\pi}_n :=$ policy in Π with maximum \hat{V}



Sample Complexity Question

- How many trajectories do we need so that

$$V(\hat{\pi}_n) \approx \max_{\pi \in \Pi} V(\pi) ?$$

- In other words want small regret where

$$\text{Reg}_{\Pi}(\hat{\pi}_n) := \max_{\pi \in \Pi} V(\pi) - V(\hat{\pi}_n) .$$

- Answer will depend on the nature of:
 - the dynamics implemented by the simulator
 - the policy class
 - the rewards.

Sample Complexity in Function/Concept Learning

Suppose we:

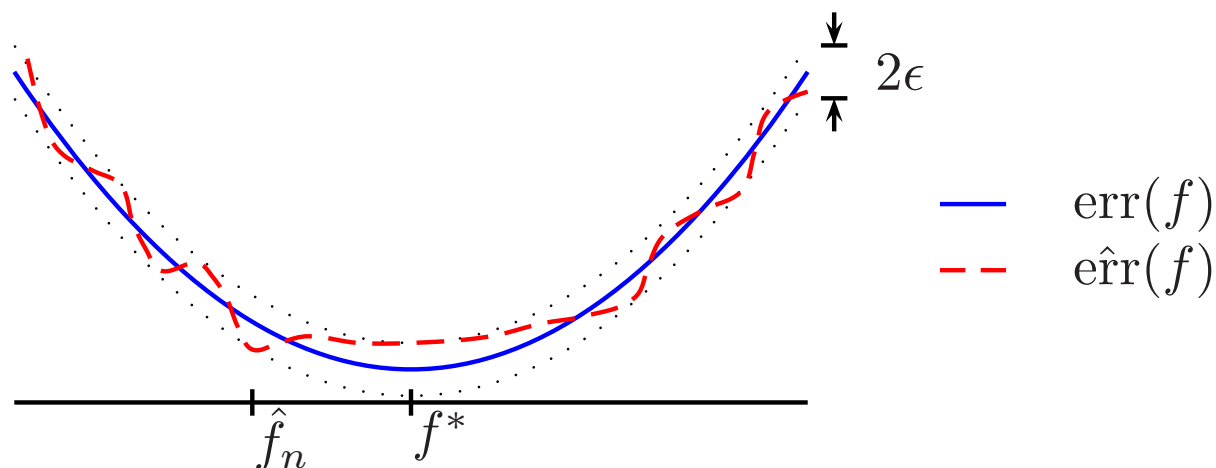
- want to find a functional dependence between X and Y , and
- have access to i.i.d. samples $(X_1, Y_1), \dots, (X_n, Y_n)$.

Then we can:

- choose some class \mathcal{F} of functions, and
- pick $\hat{f}_n \in \mathcal{F}$ to minimize $\hat{\text{err}}(f)$.

$$f \quad + \quad \begin{array}{c} (X_1, Y_1) \\ \vdots \\ (X_n, Y_n) \end{array} \quad \longrightarrow \quad \hat{\text{err}}(f) := \frac{1}{n} \sum_i \mathbf{1}[f(X_i) \neq Y_i]$$

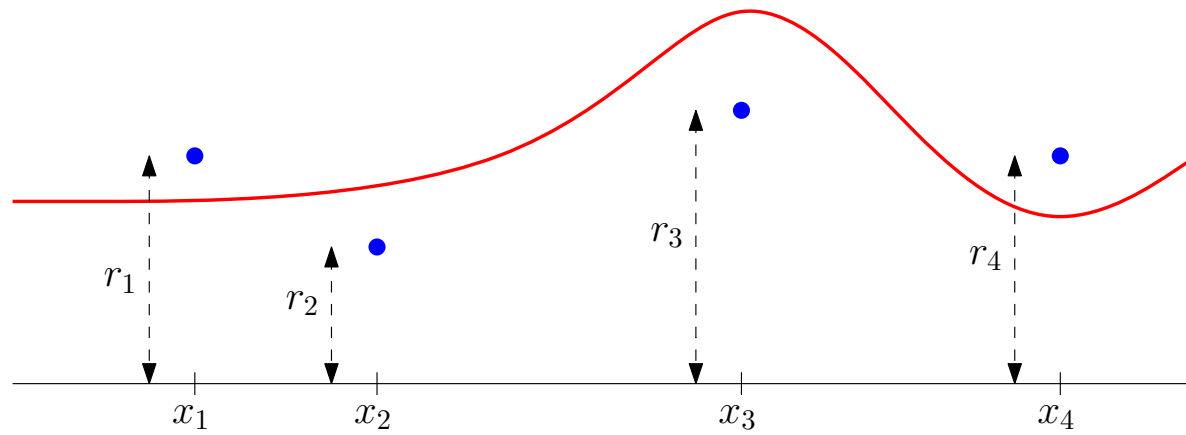
Uniform Convergence



$$\forall f \in \mathcal{F}, |\text{err}(f) - \hat{\text{err}}(f)| \leq \epsilon \quad \Rightarrow \quad \text{err}(\hat{f}_n) \leq \text{err}(f^*) + 2\epsilon$$

- Rate of convergence determined by combinatorial quantities associated with the class \mathcal{F} :
 - Pollard's pseudodimension: $\text{Pdim}(\mathcal{F})$
 - Fat shattering dimension: $\text{fat}_{\mathcal{F}}(\epsilon)$.

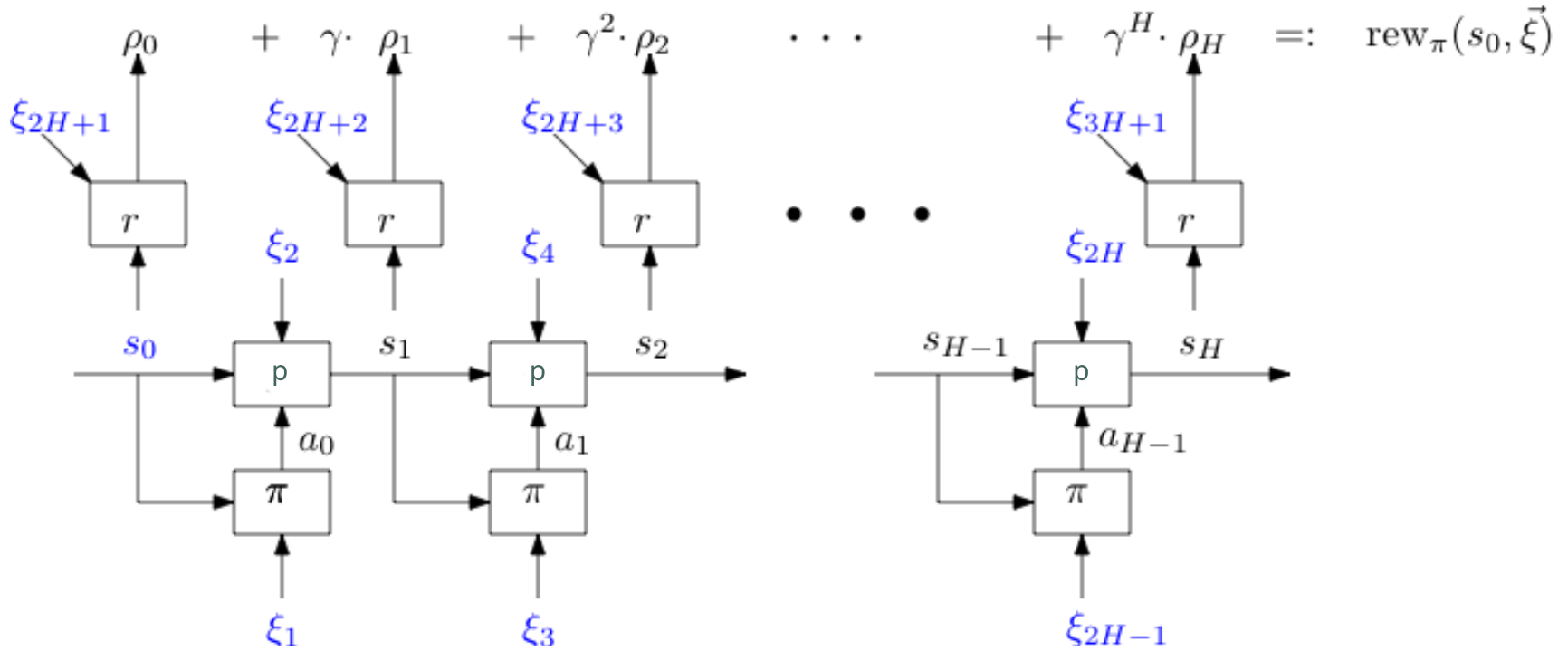
Two important dimensions



A function realizing the dichotomy 0, 1, 1, 0 for $X = \{x_1, x_2, x_3, x_4\}$.

- A function class \mathcal{F} *shatters* a set X of size n if there are functions in \mathcal{F} realizing all 2^n dichotomies of X .
- $\text{Pdim}(\mathcal{F}) :=$ size of largest shattered set.
- To *ϵ -shatter* a set, the gap must be at least ϵ .
- $\text{fat}_{\mathcal{F}}(\epsilon) :=$ size of largest ϵ -shattered set.

Iteration and Composition



$\vec{\xi}$ = random numbers

s_0 = initial state

Iteration and Composition

- The class $\mathcal{R} := \{(s_0, \vec{\xi}) \mapsto \text{rew}_\pi(s_0, \vec{\xi}) : \pi \in \Pi\}$ is obtained from Π via a series of compositions with the system dynamics p and reward mapping r .
- For \mathcal{R} to be “well-behaved” it:
 - does not suffice to assume combinatorial dimension of Π is bounded.
 - suffices to assume that the computation π, p and r requires a bounded number of steps in a model of real computation.

Boundedness of Standard Dimensions is not Enough

- We can show that there is an MDP and a policy class with

$$\text{fat}_{\Pi}(\epsilon) < \infty, \forall \epsilon > 0,$$

such that *any method* based on empirical estimates will “fail” to find a good policy.

- In fact, there is a policy class with $\text{Pdim}(\Pi) = 1$ for which the same is true!
- Therefore, more assumptions are needed to get any positive results.

Main Result

Suppose *policies are parameterized by* $\theta \in \mathbb{R}^d$. Consider the mappings:

$$\begin{aligned} (s_t, \theta) &\mapsto a_t && \text{the policy} \\ (s_t, a_t) &\mapsto s_{t+1} && \text{one-step system dynamics} \\ s_t &\mapsto \rho_t \in [0, R] && \text{reward mapping} \end{aligned}$$

If τ is an *upper bound on the number of elementary arithmetic* (i.e. $\{+, -, \times, /\}$) *and comparison* (i.e. $\{<, =, >\}$) *operations* needed to compute any of the mappings above, and the number of trajectories sampled is at least

$$\tilde{O} \left(\frac{R^2 d \tau}{(1 - \gamma)^3 \epsilon^2} \right)$$

then

$$\mathbb{E}[\text{Reg}_{\Pi}(\hat{\pi}_n)] \leq \epsilon .$$

Proof Idea

- *Key idea* Computation bounds interact nicely with iteration & composition.
 - If it takes τ_1, τ_2 steps to compute f_1, f_2 respectively, then it takes no more than $\tau_1 + \tau_2$ steps to compute $f_1 \circ f_2$.
- Using results of Goldberg and Jerrum^a, we can convert computation bounds into Pdim bounds:

$$\text{Pdim}(\mathcal{R}) = O(d\tau H) .$$

- Pdim bounds imply uniform convergence of $\hat{V}(\pi)$ to $V(\pi)$.
- Uniform convergence implies $\hat{\pi}_n$ has small regret.

^aP.W. Goldberg and M.R. Jerrum (1995) Bounding the Vapnik-Chervonenkis Dimension of Concept Classes Parameterized by Real Numbers, *Machine Learning* **18**:(2–3), pp. 131–148

Summary and Discussion

- Looked at a *policy search* method that is implemented using a simulator
- Obtained *sample complexity* bounds assuming that policy class, system dynamics and reward mappings are “nice”
- Sharing randomness was crucial in letting us use Learning Theory tools
- *Discussion*
 - What if randomness is not shared?
 - Extension to policies that depend on entire history (as opposed to just the current state)?
 - Is policy search using simulators relevant to dynamic treatment regimes?