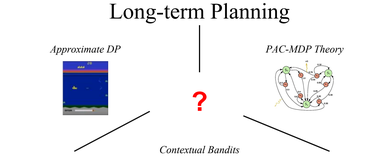
Contextual Decision Processes with Low Bellman Rank are PAC-Learnable

Nan Jiang[1,3], Akshay Krishnamurthy[2], Alekh Agarwal[3], John Langford[3], Robert E. Schapire[3]
[1]University of Michigan, Ann Arbor [2]University of Massachusetts, Amherst [3]Microsoft Research, NYC

Introduction: 3 challenges of RL



Our Answer:

- A new measure – Bellman rank
  - Captures a wide range of tractable RL problems
- A new algorithm – OLIVE
  - Polynomial sample complexity guarantee

Value-based RL in CDPs

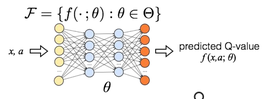
Contextual Decision Processes (CDPs): ***episodic RL with rich observations***

- Action space A, horizon H.
- Context space X. A context is ...
  - any function of history that expresses a good policy & value function
  - e.g., last 4 frames of images in Atari games
  - e.g., (state, time-step) for finite-horizon tabular MDPs

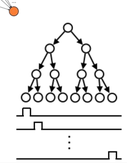
- An episode: $x_1, a_1, r_1, x_2, \ldots, x_H, a_H, r_H$
- Policy $\pi : X \to A$. Want to maximize $V^\pi = \mathbb{E}\left[\sum_{h=1}^H r_h \,\middle|\, a_{1:H} \sim \pi\right]$.

In general, $|X|$ is very large ⇒ Requires generalization!

Value-based PAC-RL in CDPs



- Input: a function space F which contains Q^*
- Output: π such that, w.p. ≥ 1-δ, $V^{\pi^*} - V^\pi \le \varepsilon$ after acquiring $poly(|A|, H, \log|F|, 1/\varepsilon, 1/\delta)$ trajectories.

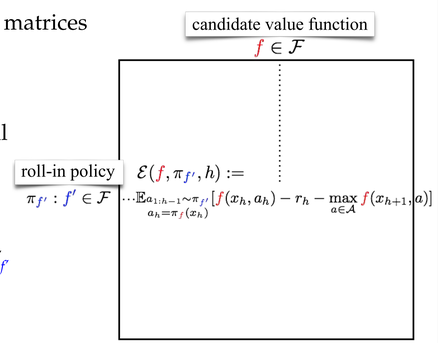


Need additional condition, otherwise exponential lower bound applies. [1]

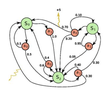
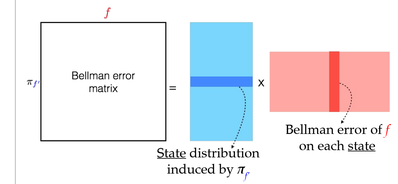
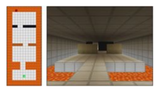
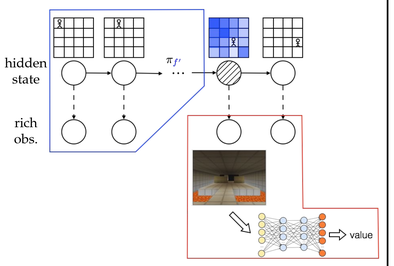
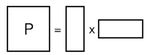
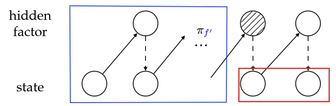
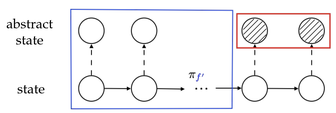
Bellman rank

Rank of average Bellman error matrices (maximum over h=1, …, H)

- Size $|F| \times |F|$
- Q^* has 0 Bellman error on all roll-in policies (col of 0's)
- Sample-efficient to evaluate a row at a time: generate trajectories using π_f until h, then random action + importance weighting



RL problems with low Bellman rank

Problem	Bound	Proof Sketch
 Tabular MDP (context = state) *PAC Learning*: known (e.g., [2])	Bellman rank ≤ # states	
 POMDP with rich obs. and reactive value function (context = current obs.) **Extends** [1]	Bellman rank ≤ # hidden states	
 Large MDP with low-rank transition (context = state) **New**	Bellman rank ≤ rank of transition matrix	
 Large MDP with Q^*-irrelevant abstraction (context = abstract state) Known [3]	Bellman rank ≤ *poly*(# abstract states, # actions)	
$P_{\mathcal{T}\|h}$ PSRs with rich obs. and reactive value function (context = current obs.) **New**	Bellman rank ≤ *poly*(system dim, # actions)	Expressing Bellman error matrix using a submatrix of the System Dynamics Matrix (naturally low-rank for PSRs). (histories = all (h-1)-long seq., tests = length 2 seq.)
 Linear Quadratic Regulators (context = state) Known [4]	Bellman rank ≤ *poly*(state space dim, action space dim)	• Need policy class + state-value function class representation (see Extensions). • Crucially depends on the choice of function classes: linear policies + quadratic value functions. • Algorithm does not apply as-is due to continuous action space.

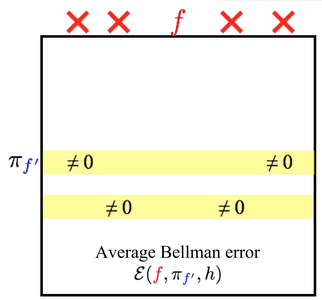
References
[1] Krishnamurthy, Agarwal, and Langford. PAC reinforcement learning with rich observations. NIPS 2016.
[2] Kearns and Singh. Near-Optimal Reinforcement Learning in Polynomial Time. ML 2000.
[3] Lihong Li. A unifying framework for computational reinforcement learning theory. PhD thesis, 2000.
[4] Osband and Van Roy. Model-based reinforcement learning and the eluder dimension. NIPS 2014.

OLIVE (**O**ptimism-**L**ed **I**terative **V**alue-function **E**limination)

Simplified Algorithm
(assuming no statistical errors)

- Generate trajectories using π_f.
- Eliminate all f with non-zero Bellman error.
- Choose a new π_f optimistically: f' is the maximizer of $V_f := \mathbb{E}[f(x_1, \pi_f(x_1))]$ among the surviving functions.
- Repeat until $V^{\pi_{f'}} \ge V_{f'}\ (\ge V^\star)$.

Full matrix view

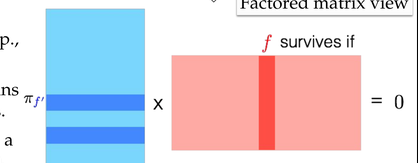


Factored matrix view

Analysis of iteration complexity

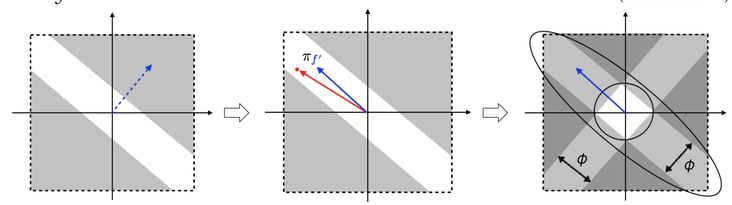
- If dark blue vectors are linearly indep., #iterations (for h) ≤ Bellman rank.
- Suffices to find a row that contains non-zero entry in surviving columns.
- Optimism finds the row with a non-zero diagonal entry (for some h).

$V_{f'} - V^{\pi_{f'}} = \sum_{h=1}^H \mathcal{E}(f', \pi_{f'}, h)$



Geometric view
(Bellman rank = 2)

Analysis that considers statistical errors



[Todd'82]: $|\langle \rightarrow, \rightarrow \rangle| = \Omega(\sqrt{M}\phi)$ => Significant reduction in ellipsoid volume

Sample complexity: $\tilde{O}\left(\frac{M^2 H^3 |\mathcal{A}|}{\epsilon^2} \log(|\mathcal{F}|/\delta)\right)$ M: Bellman rank

Extensions

- Can use doubling trick to guess unknown Bellman rank.
- Can compete with functions that have small non-zero Bellman errors.
- Can work with policy class + V-value function class (as opposed to Q).
  - Compete with the best (policy, V-value function) pair that respects *Bellman equation for policy evaluation*.
- Can accommodate infinite classes with bounded statistical complexity.
- Can handle approximately low-rank Bellman error matrices.