Optimal Approximation of Average Reward Markov Decision Processes

Yuri Sapronov, Nikita Yudin

Moscow Institute of Physics and Technology (National Research University)

Contact Information:
Moscow Institute of Physics and Technology (National Research University)
9 Institutskiy Pereulok, Dolgoprudny, Moscow Region, Russia
Email: sapronov.iuf@phystech.edu
Email: iudin.ne@phystech.edu

Abstract

We continue to develop the concept of studying the ε-optimal policy for Average Reward Markov Decision Processes (AMDP) by reducing it to Discounted Markov Decision Processes (DMDP). Existing research often stipulates that the discount factor must not fall below a certain threshold. Typically, this threshold is close to one, and as is well-known, iterative methods used to find the optimal policy for DMDP become less effective as the discount factor approaches this value.

Our work distinguishes itself from existing studies by allowing for inaccuracies in solving the empirical Bellman equation. Despite this, we have managed to maintain the sample complexity that aligns with the latest results. We have succeeded in separating the contributions from the inaccuracy of approximating the transition matrix and the residuals in solving the Bellman equation in the upper estimate so that our findings enable us to determine the total complexity of the epsilon-optimal policy analysis for DMDP across any method with a theoretical foundation in iterative complexity.

Reference	Sample Complexity	Takes Into Account Inaccuracy in Solution
[1]	$\tilde{O}\left(\frac{\lvert S\rvert\lvert A\rvert}{(1-\gamma)^3\epsilon^2}\right)$	✓
[2]	$\tilde{O}\left(\frac{\lvert S\rvert\lvert A\rvert t_{\text{minorize}}}{(1-\gamma)^2\epsilon^2}\right)$	×
[3]	$\tilde{O}\left(\frac{\lvert S\rvert\lvert A\rvert H}{(1-\gamma)^2\epsilon^2}\right)$	×
This paper	$\tilde{O}\left(\frac{\lvert S\rvert\lvert A\rvert H}{(1-\gamma)^2\epsilon^2}\right)$	✓

Table 1: Comparison of algorithms based on sample complexity.

S – set of possible states, A – set of possible actions, H – the span of the bias function of the optimal policy, t_{minorize} – minorization time for MDP.

Environment

Suppose we have a square grid where we can move in 4 directions: On square grid we have the following reward function:

$$r(s,a) = \begin{cases} 1, & \text{if } s = (20,19) \text{ and } a = \rightarrow; \\ 1, & \text{if } s = (19,20) \text{ and } a = \downarrow; \\ 0, & \text{otherwise.} \end{cases}$$

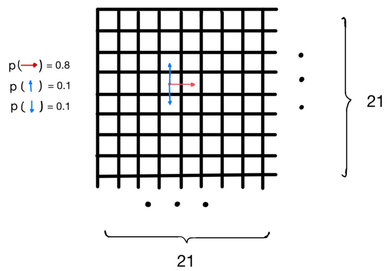


Figure 1: The grid environment used for testing.

Main Result

Algorithm 1 Perturbed Model-Based Planning

Input: Parameter $\eta \in (0,1)$, sample size per state-action pair $n \geq \frac{500H}{(1-\gamma)^2\varepsilon^2\eta^4}\beta$, target error $\varepsilon \in \left(0, \frac{1-\eta}{\frac{1}{5}+(2-\eta)\sqrt{\frac{|S|}{500H}}}\right]$, discount factor γ.

1: **for each** state-action pair $(s,a) \in S \times A$ **do**
2: Collect n samples $s_1^{s,a}, \ldots, s_n^{s,a}$ from $P(\cdot|s,a)$.
3: Form the empirical transition kernel $\hat{P}(s'|s,a) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{s_i^{s,a} = s'\}$, for all $s' \in S$.
4: **end for**
5: Set perturbation level $\xi = \frac{(1-\gamma)\varepsilon\eta}{4}$.
6: Form perturbed reward $\tilde{r} = r + Z$ where $Z(s,a) \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,\xi)$.
7: Compute a greedy policy π_T.
8: **return** π_T.

Theorem: The policy obtained by the algorithm is ε-optimal:

$$\|V^* - V^{\pi_t}\|_\infty \leq \varepsilon + \frac{1}{(1-\gamma)\eta}\|\hat{V}_p^* - V_t\|_\infty,$$

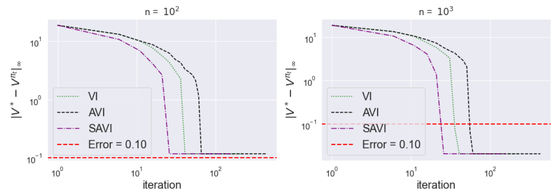
where the number of samples n satisfies:

$$n \geq \frac{500H}{(1-\gamma)^2\varepsilon^2\eta^4}\beta.$$

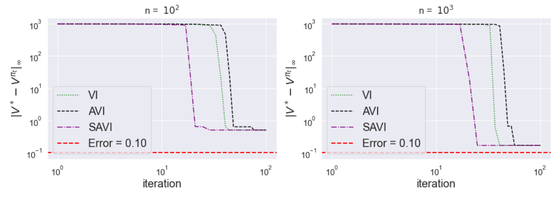
V^{π_t} – value function for MDP determined by policy π_t. V^* – optimal value function for such MDP. $\hat{V}_p^*$ – optimal value function for perturbed empirical MDP. V_t – current value function estimate, $\beta = 2\log\left(\frac{2|S||A|\log\left(\frac{e}{1-\gamma}\right)}{\delta}\right), \delta \in (0,1)$ – mismatch probability in estimates.

Experiment

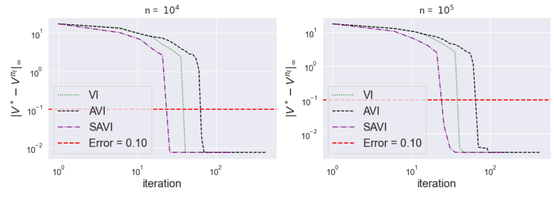
We conducted tests in a grid world with stochastic actions. The results show convergence to near-optimal policies as the number of samples increases. For high discount factors, the required sample size grows, but the method remains efficient. Value function solvers used: VI – value iteration, AVI – Nesterov accelerated value iteration, SAVI – safe (monotone) version of AVI [4].



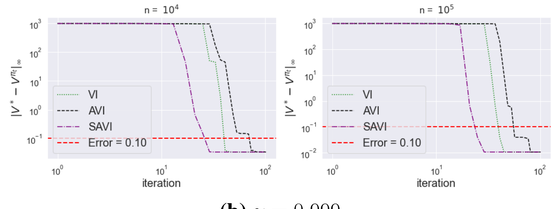
(a) $\gamma = 0.95$



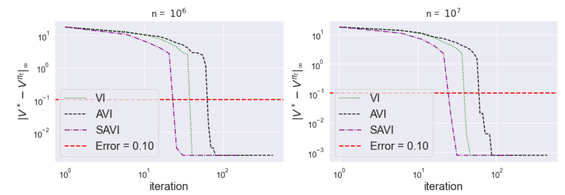
(b) $\gamma = 0.999$



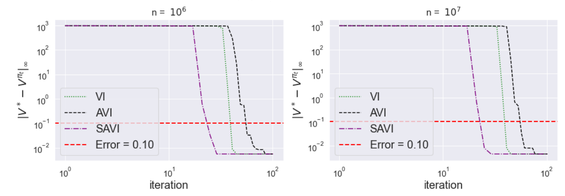
(a) $\gamma = 0.95$



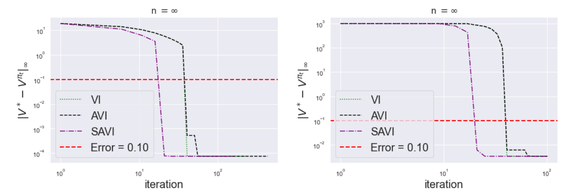
(b) $\gamma = 0.999$



(a) $\gamma = 0.95$



(b) $\gamma = 0.999$



(a) $\gamma = 0.95$ **(b)** $\gamma = 0.999$

References

[1] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91:325–349, 2013.

[2] Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity for average reward markov decision processes. *arXiv preprint arXiv:2310.08833*, 2023.

[3] Matthew Zurek and Yudong Chen. Span-based optimal sample complexity for average reward mdps. *arXiv preprint arXiv:2311.13469*, 2023.

[4] Vineet Goyal and Julien Grand-Clement. A first-order approach to accelerated value iteration. *Operations Research*, 71(2):517–535, 2023.