Policy Iteration on Markov Decision Process

Bingran Li 122020087

March 28, 2025

1 Introduction

2 Background

Four open questions:

1. Is the policy iteration method strongly polynomial for the deterministic MDP?
2. Is there a polynomial time method for MGP (logarithmic dependence of $1/(1-\gamma)$, using IPM by the Leader)?
3. Is there a strongly-polynomial time method for the deterministic MGP (independent of γ, extension from PY 16)?
4. Is there a strongly polynomial-time algorithm for MDP regardless of the discount factor?

An infinite-horizon discounted MDP can be formulated as a Linear Program D'Epenoux, 1963.

$$\begin{aligned} \max \quad & \sum_{j \in \Omega_S} v_j \\ \text{s.t.} \quad & v_j \leq c_i^k + \gamma \sum_{j \in \Omega_S} p_{ij}^k v_j, \quad \forall i \in \Omega_S, \forall k \in \Omega_A \end{aligned} \tag{1}$$

2.1 Complexity Results for MDPs

Table 1 summarizes the complexity results for various types of Markov Decision Processes.

Table 1: Complexity Results for Markov Decision Processes

Problem Type	Algorithm	Complexity	Reference
Discounted (general/stochastic) MDP	Simplex (most-negative-reduced-cost rule)	$O\left(\frac{mn}{1-\gamma}\log\frac{n}{1-\gamma}\right)$ iterations	[Ye11]
Discounted MDP	Policy Iteration, Value Iteration	$O\left(\frac{m}{1-\gamma}\log\frac{n}{1-\gamma}\right)$ iterations	[HMZ13]

Continued on next page

Table 1 – continued from previous page

Problem Type	**Algorithm**	**Complexity**	**Reference**
Two-player turn-based stochastic games	Strategy Iteration	$O\left(\frac{m}{1-\gamma}\log\frac{n}{1-\gamma}\right)$ iterations	[HMZ13]
Deterministic MDP (uniform discount)	Simplex (highest-gain pivot rule)	$O(n^3m^2\log^2 n)$ iterations	[PY15]
Deterministic MDP (nonuniform discounts)	Simplex (highest-gain pivot rule)	$O(n^5m^3\log^2 n)$ iterations	[PY15]
Deterministic MDP (uniform discount)	Minimum Mean Cycle Algorithm	$O(mn)$ time	[MTZ10]
General MDP	Specialized Interior-Point Method	Strongly polynomial in all parameters except discount factor	[Ye05]
General MDP	Policy Iteration	Exponential lower bound	[Fea10]
General MDP	Randomized simplex pivoting rules	Sub-exponential lower bound	[FHZ11]

Note: n = number of states, m = number of actions, γ = discount factor, T = time horizon

3 Summary and Discussion

References

[Fea10] John Fearnley. Exponential lower bounds for policy iteration. In *Automata, Languages and Programming: 37th International Colloquium, ICALP 2010, Bordeaux, France, July 6-10, 2010, Proceedings, Part II 37*, pages 551–562. Springer, 2010.

[FHZ11] Oliver Friedmann, Thomas Dueholm Hansen, and Uri Zwick. Subexponential lower bounds for randomized pivoting rules for the simplex algorithm. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 283–292, 2011.

[HMZ13] Thomas Dueholm Hansen, Peter Bro Miltersen, and Uri Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1–16, 2013.

[MTZ10] Omid Madani, Mikkel Thorup, and Uri Zwick. Discounted deterministic markov decision processes and discounted all-pairs shortest paths. *ACM Transactions on Algorithms (TALG)*, 6(2):1–25, 2010.

[PY15] Ian Post and Yinyu Ye. The simplex method is strongly polynomial for deterministic markov decision processes. *Mathematics of Operations Research*, 40(4):859–868, 2015.

[Ye05] Yinyu Ye. A new complexity result on solving the markov decision problem. *Mathematics of Operations Research*, 30(3):733–749, 2005.

[Ye11] Yinyu Ye. The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36(4):593–603, 2011.