AofA 2014, Paris, France

A new upper bound for Policy Iteration[†]

Romain Hollanders[1][‡], Balázs Gerencsér[1], Jean-Charles Delvenne[1§] and Raphaël M. Jungers[1¶]

[1] *Department of Mathematical Engineering and ICTEAM at UCLouvain, 4, avenue G. Lemaitre, B-1348 Louvain-la-Neuve, Belgium*

Solving Markov Decision Processes (MDPs) is a recurrent task in engineering. Even though it is known that solutions for minimizing the infinite horizon expected cost can be found in polynomial time using LP techniques, iterative methods like the Policy Iteration algorithm (PI) remain usually the most efficient in practice. This method is guaranteed to converge in a finite number of steps. Unfortunately, it is known that it may require an exponential number of steps in the size of the problem to converge. On the other hand, many unknowns remain considering the actual worst case complexity. In this work, we provide the first improvement over the fifteen years old upper bound from Mansour and Singh (1999) by showing that PI requires at most $2 \cdot \frac{2^n}{n}$ iterations to converge.

Keywords: Complexity, Policy Iteration, Markov Decision Processes

Markov Decision Processes (MDPs) are a renowned tool to model decision problems. They are represented through the set of n states in which a system can be. When being in a state, the system must choose an available action in that state, each of which induces a cost and moves the system to another state according to given transition probabilities. An MDP can always be reduced to a form in which every state has exactly two available actions, hence we restrict ourselves to that case. A policy refers to the choice of one action in every state. Given any policy (there are 2^n of them), we can associate a value to each state of the MDP that corresponds to the infinite horizon expected cost of an agent starting in that state. This expected cost can be defined in several ways depending on the application. To solve an MDP, one should provide the optimal policy that minimizes the value of every state. Such a policy always exists.

One practically efficient way of finding the optimal policy for an MDP is to use the Policy Iteration algorithm (PI). Starting from an initial policy π_0, this simple iterative scheme improves the current policy until convergence to the optimal one π^*. More precisely, being at a policy π_i at step i, PI identifies all the states in which switching to the other action while keeping the actions of every other state unchanged improves the value of every state. We refer to these states as the improvement set T^{π_i} of π_i. Then, π_{i+1} is obtained from π_i by switching the actions of every state in T^{π_i}. The algorithm stops with the optimal

[†] This work was supported by the ARC grant 'Large Graphs and Networks' from the French Community of Belgium and by the IAP network 'Dysco' funded by the office of the Prime Minister of Belgium. The scientific responsibility rests with the authors.

[‡] Corresponding author, `romain.hollanders@uclouvain.be`

[§] Member of CORE and NAXYS

[¶] F.R.S./FNRS Research Associate

policy $\pi_K = \pi^*$ whenever T^{π_K} is empty. The fact that π_{i+1} improves the value of every state guarantees convergence in a finite number of steps.

There is a natural partial ordering on the policies of an MDP. We put a directed link from policy π to policy σ when σ gives values at least as good to every state as π. We may view the policies as the vertices of a cube of dimension n. The partial order gives an orientation to the edges of the cube exhibiting a particular structure which we call an Acyclic Unique Sink Orientation (AUSO). In AUSOs, the cube with its directed edges must satisfy two conditions: (1) the cube must be acyclic and (2) any sub-cube must have a unique sink, i.e., a unique vertex with only incoming links. With this structure, PI steps can be viewed as jumps in the cube, with the global sink corresponding to the optimal policy. Hence, bounding the number of steps of PI can be relaxed to bounding the number of jumps in an AUSO.

Policy Iteration has been shown to require $\Omega(2^{n/7})$ steps to converge in the worst case by Fearnley (2010). On the other hand, the best known upper bound to date was due to Mansour and Singh (1999) with a $6 \cdot \frac{2^n}{n}$ steps bound, also holding for AUSOs. In this work, we provide the first improvement in fifteen years over Mansour and Singh's bound, which is given in Theorem 1.

Theorem 1 *The number of iterations of Policy Iteration is bounded above by* $2 \cdot \frac{2^n}{n} + o\left(\frac{2^n}{n}\right)$.

Our proof uses the ingredients of Mansour and Singh's, and does not make use of the non-inclusion property of the improvement sets. This property states that for any two policies π_i and π_j explored by PI with $i < j$, then $T^{\pi_i} \not\subseteq T^{\pi_j}$. It was thought by Mansour and Singh to be a key ingredient to improve the upper bound. As a side result, we showed that it is not the case. To this end, we built a sequence of policies of size $2 \cdot \frac{2^n}{n}$ satisfying all the ingredients from Mansour and Singh's proof as well as the non-inclusion property. Thus our upper bound cannot be improved by the non-inclusion property alone.

Hansen (2012) and Zwick proposed a relaxation of the upper bound problem on AUSOs and found using exhaustive search that the number of steps of PI are bounded above for $n = 1, ..., 6$ by F_{n+2}, the $(n + 2)^{\text{nd}}$ Fibonacci number. Following this observation, they conjectured F_{n+2} to be a possible upper bound for the number of steps of PI. However, it is interesting to note that for $n = 3, ..., 6$, our $2 \cdot \frac{2^n}{n}$ bound also fits the Fibonacci numbers almost perfectly, as shown in the table below. Which bound is more likely to be the right fit is therefore unclear at the moment.

n	1	2	3	4	5	6	7
max PI	2	3	5	8	13	21	≥ 33
F_{n+2}	2	3	5	8	13	21	34
$2 \cdot \frac{2^n}{n}$	4	4	5.3	8	12.8	21.3	36.6

References

J. Fearnley. Exponential Lower Bounds for Policy Iteration. *Proceedings of the 37th International Colloquium on Automata, Languages and Programming (ICALP'10)*, pages 551–562, 2010.

T. Hansen. *Worst-case Analysis of Strategy Iteration and the Simplex Method.* PhD thesis, Aarhus University, Science and Technology, Department of Computer Science, 2012.

Y. Mansour and S. Singh. On the Complexity of Policy Iteration. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 1999.