The complexity of Policy Iteration is exponential for discounted Markov Decision Processes

Romain Hollanders, Jean-Charles Delvenne, Raphaël M. Jungers

***Abstract*— The question of knowing whether the Policy Iteration algorithm (PI) for solving stationary Markov Decision Processes (MDPs) has exponential or (strongly) polynomial complexity has attracted much attention in the last 25 years. Recently, an example on which PI requires an exponential number of iterations to converge was proposed for the total-reward and the average-reward criteria. On the other hand, it was shown that PI runs in strongly polynomial time on discounted-reward MDPs, yet only when the discount factor is fixed beforehand.**

In this work, we show that PI needs an exponential number of steps to converge on discounted-reward MDPs with a general discount factor.

I. Introduction

Markov Decision Processes (MDPs) are a popular and efficient tool to solve sequential decision problems under uncertainty. They are widely used to model stochastic optimization problems that appear in various engineering and industrial applications such as PageRank Optimization [5], [8], [13], Smart Grids [18] or Epidemics Management [4] for instance. See [19] for a survey of numerous applications in which MDP models play an essential role.

More specifically, an MDP describes the random process of an *agent* that evolves on a finite set of *states*. One of the states is chosen to be the initial state. At every time step, the *controller* of the process needs to choose one among the several *actions* available in the current state. The chosen action determines a *transition probability* distribution for the next state to reach as well as an immediate *reward* (or cost). The goal of the controller is to choose the right actions in each state in order to maximize the rewards collected by the agent over time according to some ad hoc optimization criterion. We call such a choice of actions a *policy*, the one maximizing the optimization criterion being the *optimal policy*.

In this work, we only consider infinite-horizon MDPs in which the agent is assumed to follow the process for an infinite number of steps. Furthermore, we consider two of the most important optimization criteria in practice, namely the *total-reward* and the *discounted-reward* criteria. In total-reward MDPs, rewards are summed up during the whole process. In that case, a necessary condition for the problem to be well defined is to have a reward-free absorbing state (or set of states), which we call the *target state*. Then, the goal of the controller is to reach that target state while maximizing the expected sum of rewards until there. In discounted-reward MDPs, the reward of an action at time t is multiplied by some factor λ^t, where $0 < \lambda < 1$ is the *discount factor*. This factor can be seen either as a deflation rate or as the probability for the process to stop at each time step. For an advanced study of MDPs and optimality criteria, see for example [16].

Markov Decision Processes can be solved in weakly polynomial time using *Linear Programming* (LP) [16]. However, a much more efficient way of solving these problems in practice is to use an appropriate iterative algorithm . Among them, *(greedy) Policy Iteration* (PI) is one of the most studied. It usually converges in a few iterations and is guaranteed to find the optimal solution in finite time. It can be viewed as a Simplex algorithm in which several pivoting steps are performed simultaneously. Unfortunately, the algorithmic complexity of PI is not well understood. Despite its practical efficiency, examples in which PI requires an exponential number of iterations to converge exist in a number of cases of practical importance.

There is a significant research effort for understanding the complexity of PI. For general MDPs, the best upper bound - $O(k^n/n)$ - is due to Mansour and Singh [14], where n and k are respectively the number of states and the maximum number of actions per state. For total- and average-reward MDPs, the largest known lower bound is also exponential and has recently been found by Fearnley through a carefully built example [6], based on a construction for parity games that was proposed by Friedmann [9]. This was a breakthrough after more than 25 years of research on the question of the complexity of PI [11], [15]. The story seems different though for discounted-reward MDPs for which a strongly polynomial upper bound has recently been found by Ye [20], yet only for a *fixed* discount factor; PI is shown to run in at most $\frac{n^2(k-1)}{1-\lambda} \cdot \log\left(\frac{n^2}{1-\lambda}\right)$ iterations in that case. This bound was later improved by a factor n by Hansen et al. [11] and adapted to two-player turn-based zero-sum games, a natural two-player extension of MDPs for which PI also applies. Thereby, they provided the first (strongly) polynomial time algorithm for this latter class of problems.

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 responsiblity rests with the authors.

The authors are with Department of Mathematical Engineering and ICTEAM at UCLouvain, 4, avenue G. Lemaitre, B-1348 Louvain-la-Neuve, Belgium. J.-C. D. is with CORE and NAXYS. R. M. J. is an F.R.S./FNRS fellow.

Corresponding author, romain.hollanders@uclouvain.be
jean-charles.delvenne@uclouvain.be
raphael.jungers@uclouvain.be

In this work, we show that if we do not assume a constant discount factor, then PI runs in exponential time. This question was mentioned as an open problem by both Ye [20] and Fearnley [7]. Our proof uses perturbation analysis to provide an adequate value of the discount factor such that adding discount to Fearnley's example does not change the choices made by PI. Hence, it takes the same number of steps with or without discount and therefore, it requires an exponential number of steps to converge in both cases. Our result combined with the ones of Ye [20] and Fearnley [6] completes the characterization of the worst-case complexity of PI for MDPs: it is strongly polynomial for discounted-reward MDPs with a fixed discount rate but exponential for total-reward, average-reward and discounted-reward MDPs in general.

It has to be mentioned that Andersson and Miltersen have already used the same kind of tools to show that discounted and undiscounted games are polynomial-time equivalent [1]. However, in their work, they do not focus on a particular algorithm—like PI—but rather on the general complexity of these problems. Hence, their result only has a small impact on MDPs, since those are already known to be solved in (weakly) polynomial-time when using linear programming methods, which is not the case for games.

The paper is organized as follows. In Section II, we properly define MDPs and some related concepts and we formulate the greedy Policy Iteration algorithm. We also give a number of preliminary results. In Section III, we state Fearnley's example and analyze some of its properties. Section IV lays the foundations for our main result by analyzing how the optimal solution of an MDP is modified when perturbing the problem instance. Section V applies the analysis of Section IV to discounted-reward MDPs and presents our main result.

II. Definitions and preliminaries

In this section, we properly define Markov Decision Processes and the Policy Iteration algorithm and we summarize the main features of the total-reward example on which Policy Iteration needs an exponential number of steps to converge.

A. Total-reward Markov Decision Processes

An instance of a *total-reward Markov Decision Process* is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{U}, \mathcal{P}, \mathcal{R})$ where

- $\mathcal{S} = \{1, ..., n, \tau\}$ is the finite set of *states* (τ is a reward-free absorbing state);
- $\mathcal{U}$ is the finite set of all *actions*, and $\mathcal{U}_s$ is the set of actions available to the agent in state s;
- $\mathcal{P} = \{ P^u_{s,s'} \mid s, s' \in \mathcal{S}, u \in \mathcal{U}_s \}$ is the set of *transition probabilities*. For any action $u \in \mathcal{U}_s$ available in s, $P^u_{s,s'}$ is the probability of going from state s to state s' when the action $u \in \mathcal{U}_s$ is chosen;
- $\mathcal{R} = \{ r^u_s \mid s \in \mathcal{S}, u \in \mathcal{U}_s \}$ is the set of *rewards* collected by the agent at any state s when using action $u \in \mathcal{U}_s$.

We define a *deterministic stationary policy* (or strategy) $\pi : \mathcal{S} \to \mathcal{U}$ as the deterministic choice of one action in each state. In this context, let $\overline{P}^\pi$ be the (row-stochastic) transition probability matrix obtained using action $\pi(s)$ in each state $s \in \mathcal{S}$ and let the substochastic matrix P^π be the same matrix as $\overline{P}^\pi$ but without the row and column corresponding to state τ. Similarly, let $\overline{r}^\pi$ be the reward vector obtained when using policy π and let r^π be the same vector as $\overline{r}^\pi$ without the entry corresponding to state τ.

We also define the *value* x^π_s of a policy π at state s as the expected total reward collected by the agent during its infinite walk starting in s and following policy π thereafter. Again, $\overline{x}^\pi$ is the vector containing entries x^π_s for all $s \in \mathcal{S}$ and x^π is the same vector as $\overline{x}^\pi$ but without its τ^{th} entry. The vector x^π is the solution of the following linear system:

$$(I - P^\pi)\, x^\pi = r^\pi, \tag{1}$$

where I designates the identity matrix. The controller's goal is to find the optimal policy π^* such that $x^{\pi^*}_s \geq x^\pi_s$ for every state s and every policy π. It can be shown that such a policy always exists [16, Theorem 7.1.9].

B. Policy Iteration

The *Policy Iteration* algorithm (PI) jumps from one policy to another until it converges to a global optimum. Every step is made through a comparison between policies, therefore we say that

- policy π *weakly dominates* π' ($\pi \succeq \pi'$) if $x^\pi_s \geq x^{\pi'}_s$ for every state s;
- π (strongly) *dominates* π' ($\pi \succ \pi'$) if the previous inequality is strict for at least one state;
- π and π' are *equivalent* ($\pi \approx \pi'$) if $x^\pi_s = x^{\pi'}_s$ for every state s.

We can also compare policies with respect to some fixed policy, say π. Therefore, for every state s, following Fearnley [6], we define the *appeal* of an action $u \in \mathcal{U}_s$ with respect to π as:

$$a^\pi_{s \to u} = r^u_s + \sum_{s' \in \mathcal{S}} P^u_{s,s'}\, x^\pi_{s'} \tag{2}$$

where x^π is solution of (1). At some state s, an action $u \in \mathcal{U}_s$ is said to be *appealing* with respect to π if $a^\pi_{s \to u} > x^\pi_s$. A formulation of the greedy version of Policy Iteration is given in Algorithm 1.

Algorithm 1 Greedy Policy Iteration

Require: An arbitrary policy π_0, $k = 0$.
Ensure: The optimal policy π^*.

1: **while** $\pi_k \neq \pi_{k-1}$ **do**
2: Evaluation step: compute x^{π_k}.
3: Greedy Improvement step:
$\pi_{k+1}(s) = \operatorname{argmax}_{u \in \mathcal{U}_s} a^{\pi_k}_{s \to u}$ for all states $s \in \mathcal{S}$.
4: $k \leftarrow k + 1$.
5: **end while**
6: **return** π_k.

The two following well known theorems are the main arguments behind the finite time convergence guarantees of PI. (Proofs can be found, e.g., in [2].)

Theorem 1: Let π and π' be two policies such that $a^{\pi}_{s\to\pi'(s)} \geq x^{\pi}_s$ for every state s and such that this inequality is strict for some states. Then $\pi' \succ \pi$.

Theorem 2: For any sub-optimal π, $\exists\,(s,u), u \in \mathcal{U}_s$ such that $a^{\pi}_{s\to\pi'(s)} > x^{\pi}_s$.

Theorem 1 essentially says that replacing any actions by more appealing ones strictly improves a policy, hence PI never considers the same policy twice. Theorem 2 says that if, for some policy, no appealing action exists, then this policy is optimal. Therefore, since there are only a finite number of policies, PI converges to a global optimum in a finite number of steps. However, Theorems 1 and 2 do not guarantee convergence in polynomial time.

III. Exponential complexity example for Policy Iteration

In [6], Fearnley provides an example of a total-reward MDP with n states on which PI requires an exponential number of steps to converge. He therefore implements a binary counter.

Example 1 (Fearnley, [6]): Let $\mathcal{M}_{\text{exp}}$ be the MDP instance with n states proposed in [6]. There exists an initial policy π_0 such that Policy Iteration needs to explore the sequence of policies

$$\{\,\pi_0, \pi_1, ..., \pi_K\,\} \tag{3}$$

to converge, with $K \geq 2^m - 1$, where m is the number of bits of the binary counter and where the number of states $n = 7m + 4$.

We now state three important properties of Example 1 that will be useful to our analysis. These features all follow from the developments made in [6].

Property 1: In Example 1, every step of Policy Iteration, starting at π_0, is made in a non-ambiguous way, i.e. at every step $k = 0, ..., K$ of sequence (3) and for every state s, $\underset{u\in\mathcal{U}_s}{\operatorname{argmax}}\, a^{\pi_k}_{s\to u}$ is unique.

Property 2: Every policy of Example 1 is proper, which means that whatever the chosen policy π and the starting state s, there exists a positive-probability path from s to the final state τ, i.e., for each π and s, there exists a sequence of states $\{s_0 = s, s_1, ..., s_k = \tau\}$ such that $P^{\pi}_{s_i,s_{i+1}} > 0$ for $i = 0, 1, ..., k-1$. Note that if such a sequence exists, then there exists a sequence of length at most $n + 1$.

Property 3: In Example 1, $P^{\pi} \in \mathbb{Q}^{n\times n}$ and $r^{\pi} \in \mathbb{Z}^{n\times n}$ for every policy π and there exist values $\delta(n) \in \mathbb{N},\ \delta(n) \leq (10m+4)2^m$ and $\kappa(n) \in \mathbb{N},\ \kappa(n) \leq (10m+4)2^m$, with $n = 7m+4$, such that $\delta(n)\cdot P^{\pi} \in \mathbb{N}^{n\times n}$ for all π and that $|r^{\pi}_s| \leq \kappa(n)$ for all s, π[1].

Property 2 makes sure that the value of the total-reward MDP from Example 1 is finite for any policy and any starting state, i.e., that the linear system (1) always has a unique solution [3]. Property 3 guarantees that the considered MDP has reasonable size.

Note that Properties 2 and 3 summarize the restrictive assumptions that are made in our results at the next Sections.

[1]To keep notations simple, we will write δ and κ and temporarily forget about the dependence in n

IV. Perturbed total-reward MDP

In this section, we study how a small perturbation of the total-reward MDP instance affects the value of the states. We expect that a small enough perturbation should not affect the behavior of Policy Iteration. This section builds the basis on which our main result will rely.

Let π be some policy in a total-reward MDP and let x be the value of all states when using policy π, which can be computed by solving the linear system (1). Our goal is to determine how much x is perturbed from a perturbation of the system's matrix.

Lemma 1: Let x be the solution of

$$A\,x = b \tag{4}$$

and $\tilde{x}$ be the solution of the perturbed system

$$\tilde{A}\,\tilde{x} = b, \tag{5}$$

where A is an invertible matrix and let $\Delta x \triangleq \tilde{x} - x$ and $\Delta A \triangleq \tilde{A} - A$. Then the following bound holds for any subordinate norm:

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|\cdot\|\Delta A\|}{1 - \|A^{-1}\|\cdot\|\Delta A\|}, \tag{6}$$

whenever

$$\|A^{-1}\|\cdot\|\Delta A\| < 1. \tag{7}$$

Proof: See, e.g., [12]. ∎

Note that (4) can be identified to (1) by taking $A = (I - P^{\pi})$ and $b = r^{\pi}$. We now bound the different norms that appear in (6) to obtain a usable bound on $\|\Delta x\|$. For that purpose, we will use the norm $\|\cdot\|_\infty$. Recall that for any matrix M, $\|M\|_\infty = \max_i \sum_j |M_{i,j}|$. Our analysis will make use of Hadamard's determinant inequality.

Theorem 3 (Hadamard): Let M be an n by n matrix such that $|M_{i,j}| \leq \beta$ for all i, j. Then:

$$|\det(M)| \leq \beta^n\, n^{n/2}.$$

The next lemma gives us an upper bound on $\|A^{-1}\|_\infty$.

Lemma 2: Let us assume that $A \in \mathbb{Q}^{n\times n}$ is an invertible matrix such that $|A_{i,j}| \leq 1$ and $\delta\cdot A_{i,j} \in \mathbb{Z}$ for all i, j. Then:

$$\|A^{-1}\|_\infty \leq \delta^n\, n^{(n+1)/2}.$$

Proof: Since A is invertible, we may use Cramer's rule and express A^{-1} as:

$$A^{-1} = \delta\,(\delta\cdot A)^{-1} = \delta\,\frac{\operatorname{adj}(\delta\cdot A)}{\det(\delta\cdot A)} \tag{8}$$

where $\operatorname{adj}(\delta\cdot A)$ is the adjugate matrix of $\delta\cdot A$ in which every entry is the determinant of an $(n-1)\times(n-1)$ submatrix of the integer matrix $\delta\cdot A$ (possibly with a minus sign). We know that every entry of $|\delta\cdot A|$ is less than δ and

that $|\det(\delta \cdot A)| \geq 1$. Hence, using (8) and Theorem 3, we have:

$$\begin{aligned}\|A^{-1}\|_\infty &\leq \delta \cdot \|\mathrm{adj}(\delta \cdot A)\|_\infty \\ &\leq \delta \cdot \max_{1\leq j\leq n} \sum_{i=1}^{n} \delta^{n-1} (n-1)^{(n-1)/2} \\ &\leq \delta^n \, n^{(n+1)/2}.\end{aligned}$$

■

It now remains to find an upper bound on $\|x\|_\infty$.

Lemma 3: Let x be the solution of (4), and assume that $A \in \mathbb{Q}^{n\times n}$ is an invertible matrix such that $|A_{i,j}| \leq 1$ and $\delta \cdot A_{i,j} \in \mathbb{N}$ for all i, j and that $b \in \mathbb{Z}^n$ with $|b_i| \leq \kappa$ for all i. Then:

$$\|x\|_\infty \leq \kappa \, \delta^n \, n^{(n+1)/2}.$$

Proof: Using Lemma 2, we have $\|x\|_\infty \leq \|A^{-1}\|_\infty \cdot \|b\|_\infty \leq \delta^n \, n^{(n+1)/2} \cdot \kappa$. ■

The next theorem uses the bounds from Lemmas 1 to 3 to obtain an upper bound on $\|\Delta x\|_\infty$.

Theorem 4: Let x be the solution of (4) and $\tilde{x}$ be the solution of the perturbed system (5), and let $\Delta x = \tilde{x} - x$ and $\Delta A = \tilde{A} - A$. Let us further assume that $A \in \mathbb{Q}^{n\times n}$ is an invertible matrix such that $|A_{i,j}| \leq 1$ and $\delta \cdot A_{i,j} \in \mathbb{N}$ for all i, j and that $b \in \mathbb{Z}^n$ with $|b_i| \leq \kappa$ for all i. Then, provided that $\|\Delta A\|_\infty$ satisfies:

$$\|\Delta A\|_\infty \leq 1/2 \cdot \delta^{-n} \, n^{-(n+1)/2}, \tag{9}$$

we have:

$$\|\Delta x\|_\infty \leq 2 \, \kappa \, \delta^{2n} \, n^{n+1} \cdot \|\Delta A\|_\infty.$$

Proof: We know from Lemma 2 that $\|A^{-1}\|_\infty \leq \delta^n \, n^{(n+1)/2}$. So if we impose $\|\Delta A\|_\infty$ to satisfy (9), then Assumption (7) from Lemma 1 is satisfied and the denominator in (6) is at least 1/2. Substituting the other available bounds from Lemmas 2 and 3 into (6) gives the result. ■

V. Discounted-reward MDPs as a perturbation of total-reward MDPs

In this section, we show that adding a discount factor $\lambda \triangleq 1 - \varepsilon$ close enough to 1 to the originally discount-free total-reward MDP of Example 1 does not change the behavior of Policy Iteration and thus that the latter requires an exponential number of steps to converge on discounted-reward MDPs. We show this by providing a value of ε such that the same choices are made by Algorithm 1 on both problems at each improvement step. We proceed in three steps:

1) first, we identify the minimum possible difference between the appeal of the best action and the appeal of the other actions in a state;
2) then, we characterize the perturbation induced by adding a discount factor λ;
3) finally, we provide a value for ε that induces a small enough perturbation so that the action with the best appeal does not change in each state, and this at every step of PI.

A. The minimum difference between the appeals of actions

First, let us observe that the value of the states of an MDP can be expressed as a fraction with bounded denominator.

Lemma 4: Let x be the solution of (4) and let us assume that $A \in \mathbb{Q}^{n\times n}$ is an invertible matrix such that $|A_{i,j}| \leq 1$ and $\delta \cdot A_{i,j} \in \mathbb{Z}$ for all i, j. Then the vector x can be expressed as:

$$x = \frac{v}{d}$$

where v is an integer vector of the same dimension as x and d is a positive integer satisfying $d \leq \delta^n \, n^{n/2}$.

Proof: The linear system (4) can be rewritten as $(\delta \cdot A)\, x = \delta \cdot b$, where both $\delta \cdot A$ and $\delta \cdot b$ are integer valued. Hence, using Cramer's rule for linear systems, x can be expressed as:

$$x = \frac{v}{|\det(\delta \cdot A)|}$$

where v is an integer vector with the same dimension as x and $|\det(\delta \cdot A)|$ is a positive integer, say d. Since $|\delta \cdot A_{i,j}| \leq \delta$ for every i, j, Theorem 3 enables us to conclude. ■

The following bound makes use of the fact that every step of PI is made in a non-ambiguous way.

Theorem 5: For any state s and any step k of Policy Iteration applied to Example 1, let $u^* = \operatorname*{argmax}_{u \in \mathcal{U}_s} a^{\pi_k}_{s\to u}$ and let $u' \neq u^*$ be any other action in $\mathcal{U}_s$. Then:

$$a^{\pi_k}_{s\to u^*} - a^{\pi_k}_{s\to u'} \geq \frac{1}{\delta^{n+1} \, n^{n/2}}.$$

Proof: From the definition of appeal (2) and from Property 1, we know that

$$\begin{aligned}a^{\pi_k}_{s\to u^*} &- a^{\pi_k}_{s\to u'} \\ &= (r^{u^*}_s - r^{u'}_s) + \sum_{s'\in\mathcal{S}} (P^{u^*}_{s,s'} - P^{u'}_{s,s'})\, x^{\pi_k}_{s'} \;>\; 0.\end{aligned}$$

Since Example 1 has Properties 2 and 3, $\delta \cdot (P^{u^*}_{s,s'} - P^{u'}_{s,s'})$ is an integer and we can use Lemma 4 and write:

$$a^{\pi_k}_{s\to u^*} - a^{\pi_k}_{s\to u'} = \frac{w}{\delta \cdot d},$$

where w is an integer strictly greater than 0 and d is less than $\delta^n \, n^{n/2}$. ■

B. The perturbation induced by the discount

Let $\mathcal{M}$ be a total-reward MDP and let $\mathcal{M}_\lambda$ be the corresponding discounted-reward MDP, i.e., the MDP with same states- and actions space, transition probabilities and rewards but with an additional discount factor $\lambda = 1 - \varepsilon$. The value $\tilde{x}^\pi$ of the states of $\mathcal{M}_\lambda$ under policy π is obtained by solving the following linear system:

$$(I - \lambda \, P^\pi)\, \tilde{x}^\pi = r^\pi. \tag{10}$$

We can identify this system to (5), where $\tilde{A} = I - \lambda \, P^\pi$. Furthermore, if we define $A = I - P^\pi$ as in (4), then $\tilde{A}$ can be expressed as a perturbation ΔA of A, namely $\Delta A = \tilde{A} - A = \varepsilon \, P^\pi$. Hence, $\|\Delta A\|_\infty \leq \varepsilon$. Given a state s in the discounted MDP $\mathcal{M}_\lambda$, we define the appeal $\tilde{a}^\pi_{s\to u}$ of an

action $u \in \mathcal{U}_s$ with respect to some policy π in a similar way as in (2):

$$\tilde{a}^{\pi}_{s\to u} \triangleq r^u_s + \sum_{s'\in\mathcal{S}} \lambda\, P^u_{s,s'}\, \tilde{x}^{\pi}_{s'}. \tag{11}$$

Let us now quantify the perturbation on the appeals incurred from the discount.

Theorem 6: For any state s, any action $u \in \mathcal{U}_s$ and any step k of Policy Iteration applied to Example 1, we have:

$$|\, a^{\pi_k}_{s\to u} - \tilde{a}^{\pi_k}_{s\to u} \,| \le 4\,\kappa\,\delta^{2n}\, n^{n+2}\,\varepsilon,$$

where $a^{\pi_k}_{s\to u}$ and $\tilde{a}^{\pi_k}_{s\to u}$ are defined by (2) and (11) respectively and $\lambda = 1-\varepsilon$ is the discount factor.

Proof: In the definition (11) of $\tilde{a}^{\pi_k}_{s\to u}$, we may write $\tilde{x}^{\pi_k}$ as a perturbation of x^{π_k} using the same notation as in Lemma 1: $\tilde{x}^{\pi_k} \triangleq x^{\pi_k} + \Delta x^{\pi_k}$. From (2) and (11), we have:

$$\begin{aligned}&|\, a^{\pi_k}_{s\to u} - \tilde{a}^{\pi_k}_{s\to u} \,| \\ &\qquad \le \sum_{s'\in\mathcal{S}} P^u_{s,s'}\, |\, \varepsilon\, x^{\pi_k}_{s'} - (1-\varepsilon)\,\Delta x^{\pi_k}_{s'} \,|\end{aligned}$$

Since $P^u_{s,s'} \le 1$ and $|v_i| \le \|v\|_\infty$ for any i and any vector v, we have:

$$\begin{aligned}&|\, a^{\pi_k}_{s\to u} - \tilde{a}^{\pi_k}_{s\to u} \,| \\ &\qquad \le n \cdot (\,\varepsilon\,\|x^{\pi_k}\|_\infty + (1-\varepsilon)\,\|\Delta x^{\pi_k}\|_\infty\,).\end{aligned}$$

Using the fact that $1-\varepsilon < 1$ and the bounds from Lemma 3 and Theorem 4 with a perturbation $\|\Delta A\|_\infty \le \varepsilon$ gives the result. ∎

C. Main result

Let us now combine the results from Sections V-A and V-B to show that the choices made by PI at every improvement step do not change when applied to the discounted or the undiscounted version of Example 1.

Theorem 7 (Main Theorem): There exists an infinite family of discounted-reward MDPs with a particular starting policy on which the number of iterations that Policy Iteration takes is lower bounded by an exponential function of the size n of the MDP.

Proof: Let $\mathcal{M}$ be the total-reward MDP from example 1 on which PI explores the exponential size sequence of policies (3) and let $\mathcal{M}_\lambda$ be the corresponding discounted-reward MDP with $\lambda = 1-\varepsilon$ defined in Section V-B. We show that PI also explores sequence (3) on $\mathcal{M}_\lambda$ provided ε is small enough.

Let $F(n,\delta,\kappa) \triangleq 4\,\kappa\,\delta^{2n}\,n^{n+2}$ and $G(n,\delta) \triangleq \delta^{n+1}\,n^{n/2}$. Theorem 6 tells us that

$$|\, a^{\pi_k}_{s\to u} - \tilde{a}^{\pi_k}_{s\to u} \,| \le F(n,\delta,\kappa)\cdot\varepsilon,$$

for every state s, action $u \in \mathcal{U}_s$ and policy π_k of sequence (3), where $a^{\pi_k}_{s\to u}$ and $\tilde{a}^{\pi_k}_{s\to u}$ are respectively defined by (2) and (11). Similarly, Theorem 5 tells us that for every state s and policy π_k from sequence (3),

$$a^{\pi_k}_{s\to u^*} - a^{\pi_k}_{s\to u} \ge \frac{1}{G(n,\delta)},$$

where $u^* = \operatorname{argmax}_{u'\in\mathcal{U}_s} a^{\pi_k}_{s\to u'}$ and u is any action in $\mathcal{U}_s$ different from u^*. Since $|y-x| \ge y-x \ge -|y-x|$ for any $x, y \in \mathbb{R}$, the following relations are true:

$$\begin{aligned} a^{\pi_k}_{s\to u} - \tilde{a}^{\pi_k}_{s\to u} &\ge -|a^{\pi_k}_{s\to u} - \tilde{a}^{\pi_k}_{s\to u}\,| \\ &\ge -F(n,\delta,\kappa)\,\varepsilon \qquad (12)\\ a^{\pi_k}_{s\to u^*} - \tilde{a}^{\pi_k}_{s\to u^*} &\le |a^{\pi_k}_{s\to u^*} - \tilde{a}^{\pi_k}_{s\to u^*}| \\ &\le F(n,\delta,\kappa)\,\varepsilon. \qquad (13)\end{aligned}$$

Substracting (13) to (12), we obtain

$$\begin{aligned}&\tilde{a}^{\pi_k}_{s\to u^*} - \tilde{a}^{\pi_k}_{s\to u} \\ &\qquad \ge a^{\pi_k}_{s\to u^*} - a^{\pi_k}_{s\to u} - 2\,F(n,\delta,\kappa)\,\varepsilon.\end{aligned}$$

From Theorem 5, we know that

$$\tilde{a}^{\pi_k}_{s\to u^*} - \tilde{a}^{\pi_k}_{s\to u} \ge \frac{1}{G(n,\delta)} - 2\,F(n,\delta,\kappa)\,\varepsilon.$$

If we take ε to satisfy

$$\varepsilon < \frac{1}{2\,F(n,\delta,\kappa)\,G(n,\delta)} = \frac{1}{8\,\kappa\,\delta^{3n+1}\,n^{3/2\,n+2}}, \tag{14}$$

we have $\tilde{a}^{\pi_k}_{s\to u^*} > \tilde{a}^{\pi_k}_{s\to u}$ for every $u \ne u^*$ and hence, $\operatorname{argmax}_{u\in\mathcal{U}_s} \tilde{a}^{\pi_k}_{s\to u} = u^* = \operatorname{argmax}_{u'\mathcal{U}_s} a^{\pi_k}_{s\to u'}$. Therefore, by induction, PI makes the same choice on both $\mathcal{M}$ and $\mathcal{M}_\lambda$ at every step k and sequence (3) is observed on both problems.

Note that $\mathcal{M}_\lambda$ has the same size as $\mathcal{M}$ and recall that $\delta(n) = \kappa(n) = (10m+4)2^m$, where $n = 7m+4 > m$. Hence, an ε that satisfies condition (14) can be written with a polynomial number of bits since:

$$\begin{aligned}\varepsilon &< \frac{1}{2^{3+\log_2\kappa+(3n+1)\,\log_2\delta+(1.5n+2)\,\log_2 n}} \\ &= \frac{1}{2^{3+m+(3n+1)n+(3n+2)\log_2(10n+4)+(1.5n+2)\,\log_2 n}} \\ &< \frac{1}{2^{q(n)}},\end{aligned}$$

where q is a suitable polynomial[2]. Finally, observe that condition (14) implies condition (9) from Theorem 4. The idea of the proof of Theorem 7 is sketched in Figure 1.

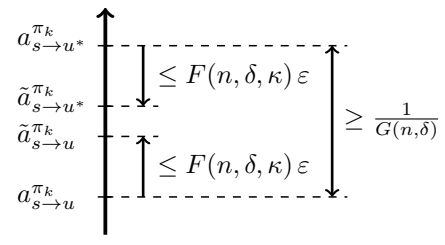

Fig. 1. The idea of the proof of Theorem 7 is to bound the parameter ε in order to make sure that $|a^{\pi_k}_{s\to u^*} - \tilde{a}^{\pi_k}_{s\to u^*}| + |a^{\pi_k}_{s\to u} - \tilde{a}^{\pi_k}_{s\to u}| \le a^{\pi_k}_{s\to u^*} - a^{\pi_k}_{s\to u}$.

∎

[2]In practice, one would already observe Theorem 7 with q linear in n.

VI. Conclusions and perspectives

In conclusion, Example 1 rules out hope for greedy Policy Iteration to be a strongly polynomial time algorithm to solve Markov Decision Processes, even though it was one of the best candidates. Nevertheless, Example 1 is artificial and is unlikely to be encountered in practical applications. In Figure 2, we attempt to challenge the robustness of Example 1 and we therefore perturb PI by making *all but one* (instead of *all*) improving switches at every step of the algorithm. We observe that the number of iterations seems to grow at a polynomial rate in that case.

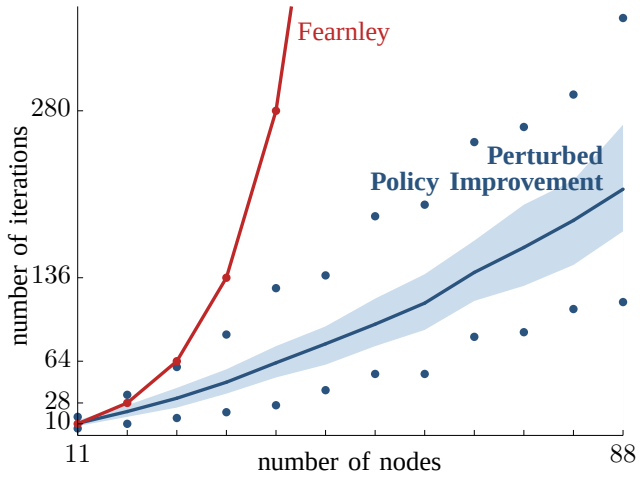

Fig. 2. (Red) Policy Iteration run on instances of Example 1 of increasing size. The exponential increase of the number of iterations is observed. (Blue) At each step of the algorithm, instead of making every improving switch, we choose one at random that is not switched. 200 trials have been made for every problem size and the number of iterations achieved has been recorded : the straight line is the average number of steps observed, the blue shadow contains $2/3$ of the points and the blue dots are the extreme values for each problem size. The exponential behavior seems to have disappeared.

Furthermore, Example 1 uses probabilistic actions and both positive and negative rewards. We would like to investigate whether Policy Iteration runs in polynomial time on deterministic MDPs or on MDPs with only positive rewards. Such situations appear in a number of practical application such as computing minimum mean-cost cycles [10] or optimizing the PageRank of nodes [5]. We would also like to apply smoothed analysis [17] on PI to explain its practical efficiency.

References

[1] D. Andersson and P. Miltersen. The Complexity of Solving Stochastic Games on Graphs. *Algorithms and Computation*, pages 112–121, 2009.

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, Massachusetts, 3rd edition, 2007.

[3] D. P. Bertsekas and J. N. Tsitsiklis. An Analysis of Stochastic Shortest Path Problems. *Mathematics of Operations Research*, 16(3):580-595, 1991.

[4] I. Chadès, T.G. Martin, S. Nicol, M.A. Burgman, H.P. Possingham, and Y.M. Buckley. General Rules for Managing and Surveying Networks of Pests, Diseases, and Endangered Species. *Proceedings of the National Academy of Sciences*, 108(20):8323, 2011.

[5] B. C. Csáji, R. M. Jungers, and V. D. Blondel. PageRank Optimization by Edge Selection. *To appear in: Discrete Applied Mathematics*, 2011.

[6] J. Fearnley. Exponential Lower Bounds for Policy Iteration. *CoRR, abs/1003.3418*, 2010.

[7] J. Fearnley. *Strategy Iteration Algorithms for Games and Markov Decision Processes*. PhD thesis, University of Warwick, 2010.

[8] O. Fercoq, M. Akian, M. Bouhtou, and S. Gaubert. Ergodic Control and Polyhedral approaches to PageRank Optimization. *Arxiv preprint arXiv:1011.2348*, 2010.

[9] O. Friedmann. An Exponential Lower Bound for the Parity Game Strategy Improvement Algorithm as we know it. *Proceedings of the 24th Annual IEEE Symposium on Logic In Computer Science*, pages 145–156, 2009.

[10] T. Hansen and U. Zwick. Lower Bounds for Howard's Algorithm for Finding Minimum Mean-Cost Cycles. *ISAAC'10: Proceedings of the 21st International Symposium on Algorithms and Computation*, pages 415–426, 2010.

[11] T. D. Hansen, P. B. Miltersen, and U. Zwick. Strategy Iteration is Strongly Polynomial for 2-Player Turn-Based Stochastic Games with a Constant Discount Factor. *Arxiv preprint arXiv:1008.0530*, 2010.

[12] N.J. Higham. *Accuracy and Stability of Numerical Algorithms*. Number 48. SIAM, 1996.

[13] H. Ishii and R. Tempo. Fragile Link Structure in PageRank Computation. *Proceedings of the 48th IEEE Conference on Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference*, pages 121–126, 2009.

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

[15] U. Meister and U. Holzbaur. A Polynomial Time Bound for Howard's Policy Improvement Algorithm. *OR Spectrum*, 8(1):37–40, 1986.

[16] M. L. Puterman. *Markov Decision Processes*. John Wiley & Sons, 1994.

[17] D.A. Spielman and S.H. Teng. Smoothed Analysis of Algorithms: Why the Simplex Algorithm Usually Takes Polynomial Time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

[18] K. Turitsyn, S. Backhaus, M. Ananyev, and M. Chertkov. Smart Finite State Devices: A Modeling Framework for Demand Response Technologies. *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, pages 7–14, 2011.

[19] D. J. White. Survey of Applications of Markov Decision Processes. *The Journal of the Operational Research Society*, 44(11):1073-1096, 1993.

[20] Y. Ye. The Simplex and Policy-Iteration Methods are Strongly Polynomial for the Markov Decision Problem with a Fixed Discount Rate. *Mathematics of Operations Research*, 2011.