Improved and Generalized Upper Bounds on the Complexity of Policy Iteration

Bruno Scherrer

Improved and Generalized Upper Bounds on the Complexity of Policy Iteration

Bruno Scherrer
INRIA Nancy Grand Est, Team MAIA
bruno.scherrer@inria.fr

February 10, 2016

Abstract

Given a Markov Decision Process (MDP) with n states and a total number m of actions, we study the number of iterations needed by Policy Iteration (PI) algorithms to converge to the optimal γ-discounted policy. We consider two variations of PI: Howard's PI that changes the actions in all states with a positive advantage, and Simplex-PI that only changes the action in the state with maximal advantage. We show that Howard's PI terminates after at most $O\left(\frac{m}{1-\gamma}\log\left(\frac{1}{1-\gamma}\right)\right)$ iterations, improving by a factor $O(\log n)$ a result by Hansen et al (2013), while Simplex-PI terminates after at most $O\left(\frac{nm}{1-\gamma}\log\left(\frac{1}{1-\gamma}\right)\right)$ iterations, improving by a factor $O(\log n)$ a result by Ye (2011). Under some structural properties of the MDP, we then consider bounds that are independent of the discount factor γ: quantities of interest are bounds τ_t and τ_r—uniform on all states and policies—respectively on the *expected time spent in transient states* and *the inverse of the frequency of visits in recurrent states* given that the process starts from the uniform distribution. Indeed, we show that Simplex-PI terminates after at most $\tilde{O}\left(n^3m^2\tau_t\tau_r\right)$ iterations. This extends a recent result for deterministic MDPs by Post & Ye (2013), in which $\tau_t \leq 1$ and $\tau_r \leq n$; in particular it shows that Simplex-PI is strongly polynomial for a much larger class of MDPs. We explain why similar results seem hard to derive for Howard's PI. Finally, under the additional (restrictive) assumption that the state space is partitioned in two sets, respectively states that are transient and recurrent for all policies, we show that both Howard's PI and Simplex-PI terminate after at most $\tilde{O}(m(n^2\tau_t + n\tau_r))$ iterations.

1 Introduction

We consider a discrete-time dynamic system whose state transition depends on a control, where the **state space** X is of finite size n. When at state $i \in \{1, .., n\}$, the action is chosen from a set of admissible actions $A_i \subset A$, where the **action space** A is of finite size m, such that $(A_i)_{1\leq i\leq n}$ form a partition of A. The action $a \in A_i$ specifies the **transition probability** $p_{ij}(a) = \mathbb{P}(i_{t+1} = j|i_t = i, a_t = a)$ to the next state j. At each transition, the system is given a reward $r(i, a, j) \in \mathbb{R}$ where r is the instantaneous **reward function**. In this context, we look for a stationary deterministic policy[1], that is a function $\pi : X \to A$ that maps states into admissible actions (for all i, $\pi(i) \in A_i$) that maximizes the expected discounted sum of rewards from any state i, called the **value of policy** π at state i:

$$v_\pi(i) := \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k r(i_k, a_k, i_{k+1})\middle| i_0 = i,\ \forall k \geq 0,\ a_k = \pi(i_k),\ i_{k+1} \sim \mathbb{P}(\cdot|i_k, a_k)\right],$$

where $\gamma \in (0, 1)$ is a discount factor. The tuple $\langle X, (A_i)_{i\in X}, p, r, \gamma\rangle$ is called a **Markov Decision Process (MDP)** (Puterman, 1994; Bertsekas and Tsitsiklis, 1996), and the associated problem is known as **stochastic optimal control.**

The **optimal value** starting from state i is defined as

$$v_*(i) := \max_\pi v_\pi(i).$$

[1]Restricting our attention to stationary deterministic policies is not a limitation. Indeed, for the optimality criterion to be defined soon, it can be shown that there exists at least one stationary deterministic policy that is optimal (Puterman, 1994).

For any policy π, we write P_π for the $n \times n$ stochastic matrix whose elements are $p_{ij}(\pi(i))$, and r_π for the vector whose components are $\sum_j p_{ij}(\pi(i))r(i, \pi(i), j)$. The value functions v_π and v_* can be seen as vectors on X. It is well known that v_π is the solution of the following Bellman equation:

$$v_\pi = r_\pi + \gamma P_\pi v_\pi,$$

that is v_π is a fixed point of the affine operator $T_\pi : v \mapsto r_\pi + \gamma P_\pi v$. It is also well known that v_* satisfies the following Bellman equation:

$$v_* = \max_\pi (r_\pi + \gamma P_\pi v_*) = \max_\pi T_\pi v_*$$

where the max operator is taken componentwise. In other words, v_* is a fixed point of the nonlinear operator $T : v \mapsto \max_\pi T_\pi v$. For any value vector v, we say that a policy π is **greedy with respect to the value** v if it satisfies:

$$\pi \in \arg\max_{\pi'} T_{\pi'} v$$

or equivalently $T_\pi v = Tv$. With some slight abuse of notation, we write $\mathcal{G}(v)$ for any policy that is greedy with respect to v. The notions of optimal value function and greedy policies are fundamental to optimal control because of the following property: any policy π_* that is greedy with respect to the optimal value v_* is an **optimal policy** and its value v_{π_*} is equal to v_*.

Let π be some policy. For any policy π', we consider the quantity

$$a_\pi^{\pi'} = T_{\pi'} v_\pi - v_\pi$$

that measures the difference in value resulting from switching the first action to π' with respect to always using π; we shall call it the **advantage of π' with respect to π**. Furthermore, we call **maximal advantage with respect to π** the componentwise best such advantage:

$$a_\pi = \max_{\pi'} a_\pi^{\pi'} = Tv_\pi - v_\pi,$$

where the second equality follows from the very definition of the Bellman operator T. While the advantage $a_\pi^{\pi'}$ may have negative values, the maximal advantage a_π has only non-negative values. We call the **set of switchable states of** π the set of states for which the maximal advantage with respect to π is positive:

$$S_\pi = \{i,\ a_\pi(i) > 0\}.$$

Assume now that π is non-optimal (this implies that S_π is a non-empty set). For any non-empty subset Y of S_π, we denote switch(π, Y) a policy satisfying:

$$\forall i,\ \text{switch}(\pi, Y)(i) = \begin{cases} \mathcal{G}(v_\pi)(i) & \text{if } i \in Y \\ \pi(i) & \text{if } i \notin Y. \end{cases}$$

The following result is well known (see for instance Puterman (1994)).

Lemma 1. *Let π be some non-optimal policy. If $\pi' = switch(\pi, Y)$ for some non-empty subset Y of S_π, then $v_{\pi'} \geq v_\pi$ and there exists at least one state i such that $v_{\pi'}(i) > v_\pi(i)$.*

This lemma is the foundation of the well-known iterative procedure, called Policy Iteration (PI), that generates a sequence of policies (π_k) as follows.

$$\pi_{k+1} \leftarrow \text{switch}(\pi_k, Y_k) \text{ for some set } Y_k \text{ such that } \emptyset \subsetneq Y_k \subseteq S_{\pi_k}.$$

The choice for the subsets Y_k leads to different variations of PI. In this paper we will focus on two of them:

- When for all iterations k, $Y_k = S_{\pi_k}$, that is one switches the actions in all states with positive advantage with respect to π_k, the above algorithm is known as Howard's PI; it can be seen then that $\pi_{k+1} \in \mathcal{G}(v_{\pi_k})$.

- When for all iterations k, Y_k is a singleton containing a state $i_k \in \arg\max_i a_{\pi_k}(i)$, that is if we only switch one action in the state with maximal advantage with respect to π_k, we will call it Simplex-PI[2].

Since it generates a sequence of policies with increasing values, any variation of PI converges to an optimal policy in a number of iterations that is smaller than the total number of policies. In practice, PI converges in very few iterations. On random MDP instances, convergence often occurs in time sub-linear in n. The aim of this paper is to discuss existing and provide new upper bounds on the number of iterations required by Howard's PI and Simplex-PI that are much sharper than m^n.

In the next sections, we describe some known results—see also Ye (2011) for a recent and comprehensive review—about the number of iterations required by Howard's PI and Simplex-PI, along with some of our original improvements and extensions. For clarity, all proofs are deferred to the later sections.

2 Bounds with respect to a fixed discount factor $\gamma < 1$

A key observation for both algorithms, that will be central to the results we are about to discuss, is that the sequences they generate satisfy some contraction property[3]. For any vector $u \in \mathbb{R}^n$, let $\|u\|_\infty = max_{1\leq i\leq n}|u(i)|$ be the max-norm of u. Let $\mathbb{1}$ be the vector of which all components are equal to 1.

Lemma 2 (e.g. Puterman (1994), proof in Section 5)**.** *The sequence* $(\|v_* - v_{\pi_k}\|_\infty)_{k\geq 0}$ *built by Howard's PI is contracting with coefficient* γ*.*

Lemma 3 ((Ye, 2011), proof in Section 6)**.** *The sequence* $(\mathbb{1}^T(v_* - v_{\pi_k}))_{k\geq 0}$ *built by Simplex-PI is contracting with coefficient* $1 - \frac{1-\gamma}{n}$*.*

Contraction is a widely known property for Howard's PI, and it was to our knowledge first proved by (Ye, 2011) for Simplex-PI; we provide simple proofs in this paper for the sake of completeness. While the first contraction property is based on the $\|\cdot\|_\infty$-norm, the second can be equivalently expressed in terms of the $\|\cdot\|_1$-norm defined by $\|u\|_1 = \sum_{i=1}^n |u(i)|$, since the vectors $v_* - v_{\pi_k}$ are non-negative and thus satisfy $\mathbb{1}^T(v_* - v_{\pi_k}) = \|v_* - v_{\pi_k}\|_1$. Contraction has the following immediate consequence[4].

Corollary 1. *Let* $V_{\max} = \frac{\max_\pi \|r_\pi\|_\infty}{1-\gamma}$ *be an upper bound on* $\|v_\pi\|_\infty$ *for all policies* π*. In order to get an* ϵ*-optimal policy, that is a policy* π_k *satisfying* $\|v_* - v_{\pi_k}\|_\infty \leq \epsilon$*, Howard's PI requires at most* $\left\lceil \frac{\log \frac{V_{\max}}{\epsilon}}{1-\gamma} \right\rceil$ *iterations, while Simplex-PI requires at most* $\left\lceil \frac{n \log \frac{nV_{\max}}{\epsilon}}{1-\gamma} \right\rceil$ *iterations.*

These bounds depend on the precision term ϵ, which means that Howard's PI and Simplex-PI are *weakly polynomial* for a fixed discount factor γ. An important breakthrough was recently achieved by Ye (2011) who proved that one can remove the dependency with respect to ϵ, and thus show that Howard's PI and Simplex-PI are *strongly polynomial* for a fixed discount factor γ.

Theorem 1 (Ye (2011))**.** *Simplex-PI and Howard's PI both terminate after at most*

$$(m-n)\left\lceil \frac{n}{1-\gamma}\log\left(\frac{n^2}{1-\gamma}\right)\right\rceil \;=\; O\left(\frac{mn}{1-\gamma}\log\frac{n}{1-\gamma}\right)$$

iterations.

The proof is based on the fact that PI corresponds to the simplex algorithm in a linear programming formulation of the MDP problem. Using a more direct proof—not based on linear programming arguments—Hansen *et al.* (2013) recently improved the result by a factor $O(n)$ for Howard's PI.

[2] In this case, PI is equivalent to running the simplex algorithm with the highest-pivot rule on a linear program version of the MDP problem (Ye, 2011).

[3] A sequence of non-negative numbers $(x_k)_{k\geq 0}$ is contracting with coefficient α if and only if for all $k \geq 0$, $x_{k+1} \leq \alpha x_k$.

[4] For Howard's PI, we have: $\|v_* - v_{\pi_k}\|_\infty \leq \gamma^k \|v_* - v_{\pi_0}\|_\infty \leq \gamma^k V_{\max}$. Thus, a sufficient condition for $\|v_* - v_{\pi_k}\|_\infty < \epsilon$ is $\gamma^k V_{\max} < \epsilon$, which is implied by $k \geq \frac{\log \frac{V_{\max}}{\epsilon}}{1-\gamma} > \frac{\log \frac{V_{\max}}{\epsilon}}{\log \frac{1}{\gamma}}$. For Simplex-PI, we have $\|v_* - v_{\pi_k}\|_\infty \leq \|v_* - v_{\pi_k}\|_1 \leq \left(1 - \frac{1-\gamma}{n}\right)^k \|v_* - v_{\pi_0}\|_1 \leq \left(1 - \frac{1-\gamma}{n}\right)^k nV_{\max}$, and the conclusion is similar to that for Howard's PI.

Theorem 2 (Hansen *et al.* (2013)). *Howard's PI terminates after at most*

$$(m+1)\left\lceil \frac{1}{1-\gamma}\log\left(\frac{n}{1-\gamma}\right)\right\rceil \;=\; O\left(\frac{m}{1-\gamma}\log\frac{n}{1-\gamma}\right)$$

iterations.

Our first results, that are consequences of the contraction property of Howard's PI (Lemma 2) are stated in the following theorems.

Theorem 3 (Proof in Section 7). *Howard's PI terminates after at most*

$$(m-n)\left\lceil \frac{1}{1-\gamma}\log\left(\frac{1}{1-\gamma}\right)\right\rceil \;=\; O\left(\frac{m}{1-\gamma}\log\frac{1}{1-\gamma}\right)$$

iterations.

Theorem 4 (Proof in Section 8). *Simplex-PI terminates after at most*

$$n(m-n)\left(1+\frac{2}{1-\gamma}\log\frac{1}{1-\gamma}\right) \;=\; O\left(\frac{mn}{1-\gamma}\log\frac{1}{1-\gamma}\right)$$

iterations.

Both results are a factor $O(\log n)$ better than the previously known results provided by Hansen *et al.* (2013) and Ye (2011). These improvements boil down to the use of the $\|\cdot\|_\infty$-norm instead of the $\|\cdot\|_1$-norm at various points of the previous analyses. For Howard's PI, the resulting arguments constitute a rather simple extension—the overall line of analysis ends up being very simple, and we consequently believe that it could be part of an elementary course on Policy Iteration; note that a similar improvement and analysis was discovered independently by Akian and Gaubert (2013) in a slightly more general setting. For Simplex-PI, however, the line of analysis is slightly trickier: it amounts to bound the improvement in value at individual states and requires a bit of bookkeeping; the technique we use is to our knowledge original.

The bound for Simplex-PI is a factor $O(n)$ larger than that for Howard's PI[5]. However, since one changes only one action per iteration, each iteration has a complexity that is in a worst-case sense lower by a factor n: the update of the value can be done in time $O(n^2)$ through the Sherman-Morrisson formula, though in general each iteration of Howard's PI, which amounts to compute the value of some policy that may be arbitrarily different from the previous policy, may require $O(n^3)$ time. Thus, it is remarkable that both algorithms seem to have a similar complexity.

The linear dependency of the bound for Howard's PI with respect to m is optimal (Hansen, 2012, Chapter 6.4). The linear dependency with respect to n *or* m (separately) is easy to prove for Simplex-PI; we conjecture that Simplex-PI's complexity is proportional to nm, and thus that our bound is tight for a fixed discount factor. The dependency with respect to the term $\frac{1}{1-\gamma}$ may be improved, but removing it is impossible for Howard's PI and very unlikely for Simplex-PI. Fearnley (2010) describes an MDP for which Howard's PI requires an exponential (in n) number of iterations for $\gamma = 1$ and Hollanders *et al.* (2012) argued that this holds also when γ is in the vicinity of 1. Though a similar result does not seem to exist for Simplex-PI in the literature, Melekopoglou and Condon (1994) consider four variations of PI that all switch one action per iteration, and show through specifically designed MDPs that they may require an exponential (in n) number of iterations when $\gamma = 1$.

3 Bounds for Simplex-PI that are independent of γ

In this section, we will describe some bounds that do not depend on γ but that will be based on some structural properties of the MDP. On this topic, Post and Ye (2013) recently showed the following result for deterministic MDPs.

Theorem 5 (Post and Ye (2013)). *If the MDP is deterministic, then Simplex-PI terminates after at most $O(n^3m^2\log^2 n)$ iterations.*

[5]Note that it was also the case in Corollary 1.

Given a policy π of a deterministic MDP, states are either on cycles or on paths induced by π. The core of the proof relies on the following lemmas that altogether show that cycles are created regularly and that significant progress is made every time a new cycle appears; in other words, significant progress is made regularly.

Lemma 4 (Post and Ye (2013, Lemma 3.4))**.** *If the MDP is deterministic, after $O(n^2 m \log n)$ iterations, either Simplex-PI finishes or a new cycle appears.*

Lemma 5 (Post and Ye (2013, Lemma 3.5))**.** *If the MDP is deterministic, when Simplex-PI moves from π to π' where π' involves a new cycle, we have*

$$\mathbb{1}^T(v_{\pi_*} - v_{\pi'}) \leq \left(1 - \frac{1}{n}\right) \mathbb{1}^T(v_{\pi_*} - v_\pi).$$

Indeed, these observations suffice to prove[6] that Simplex-PI terminates after $O(n^2 m^2 \log \frac{n}{1-\gamma})$. Completely removing the dependency with respect to the discount factor γ—the term in $O(\log \frac{1}{1-\gamma})$—requires a careful extra work described in Post and Ye (2013), which incurs an extra term of order $O(n \log(n))$.

The main result of this section is to show how these results can be extended to a more general setting. While Ye (2011) reason on states that belong to *paths* and *cycles* induced by policies on deterministic MDPs, we shall consider their natural generalization for stochastic MDPs: *transient states* and *recurrent classes* induced by policies. Precisely, we are going to consider bounds—uniform on all policies and states—of the average time 1) spent in transient states and 2) needed to revisit states in recurrent classes. For any policy π and state i, denote $\tau^\pi(i,t)$ the expected cumulative time spent in state i until time $t-1$ given than the process starts from the uniform distribution U on X and takes actions according to π:

$$\tau^\pi(i,t) = \mathbb{E}\left[\sum_{k=0}^{t-1} \mathbf{1}_{i_t=i} \mid i_0 \sim U,\ a_t = \pi(i_t)\right] = \sum_{k=0}^{t-1} \mathbb{P}(i_t = i \mid i_0 \sim U,\ a_t = \pi(i_t)),$$

where $\mathbf{1}$ denotes the indicator function. In addition, consider the vector μ^π on X providing the asymptotic frequency in all states given that policy π is used and that the process starts from the uniform distribution U:

$$\forall i,\ \mu^\pi(i) = \lim_{t\to\infty} \frac{1}{t} \tau^\pi(i,t).$$

When the Markov chain induced by π is ergodic, and thus admits a unique stationary distribution, μ^π is equal to this very stationary distribution. However, our definition is more general in that policies may induce Markov chains with aperiodicity and/or multiple recurrent classes. For any state i that is transient for the Markov chain induced by π, it is well known that $\lim_{t\to\infty} \tau^\pi(i,t) < \infty$ and $\mu^\pi(i) = 0$. However, for any recurrent state i, we know that $\lim_{t\to\infty} \tau^\pi(i,t) = \infty$ and $\mu^\pi(i) > 0$; in particular, if i belongs to some recurrent class $\mathcal{R}$, which is reached with probability q from the uniform distribution U, then $\frac{q}{\mu^\pi(i)}$ is the expected time between two visits of the state i.

We are now ready to express the structural properties with which we can provide an extension of the analysis of Post and Ye (2013).

Definition 1. *Let τ_t and τ_r be the smallest finite constants such that for all policies π and states i,*

$$\begin{aligned} &\textit{if } i \textit{ is transient for } \pi, \textit{ then} && \lim_{t\to\infty} \tau^\pi(i,t) \leq \tau_t \\ &\textit{else if } i \textit{ is recurrent for } \pi, \textit{ then} && \frac{1}{\mu^\pi(i)} \leq \tau_r. \end{aligned}$$

Note that for any finite MDP, these finite constants always exist. With Definition 1 in hand, we can generalize Lemmas 4-5 as follows.

Lemma 6. *After at most $(m-n)\lceil n^2\tau_t \log(n^2\tau_t)\rceil + n\lceil n^2\tau_t \log(n^2)\rceil$ iterations either Simplex-PI finishes or a new recurrent class appears.*

Lemma 7. *When Simplex-PI moves from π to π' where π' involves a new recurrent class, we have*

$$\mathbb{1}^T(v_{\pi_*} - v_{\pi'}) \leq \left(1 - \frac{1}{n\tau_r}\right) \mathbb{1}^T(v_{\pi_*} - v_\pi).$$

[6] This can be done by using arguments similar those for Theorem 1 (see Ye (2011) for details).

From these generalized observations, we can deduce the following original result.

Theorem 6 (Proof in Section 9)**.** *Simplex-PI terminates after at most*

$$\left[m\lceil n\tau_r \log(n^2\tau_r)\rceil + (m-n)\lceil n\tau_r \log(n^2\tau_t)\rceil\right]\left[(m-n)\lceil n^2\tau_t \log(n^2\tau_t)\rceil + n\lceil n^2\tau_t \log(n^2)\rceil\right] = \tilde{O}\left(n^3m^2\tau_t\tau_r\right)$$

iterations.

Remark 1. *This new result extends the result obtained for deterministic MDPs by Post and Ye (2013) recalled in Theorem 5. In the deterministic case, it is easy to see that $\tau_t = 1$ and $\tau_r \leq n$. Then, while Lemma 6 is a strict generalization of Lemma 4, Lemma 7 provides a contraction factor that is slightly weaker than that of Lemma 5—$\left(1 - \frac{1}{n^2}\right)$ instead of $\left(1 - \frac{1}{n}\right)$—, which makes the resulting bound provided in Theorem 6 a factor $O(n)$ worse than that of Theorem 5. This extra term in the bound is the price paid for making the constant τ_r (and the vector μ_π) independent of the discount factor γ, that is by presenting our result in a way that only depends on the dynamics of the underlying MDP. An analysis that would strictly generalizes that of Ye (2011) can be done under a variation of Definition 1 where the constants τ_t and τ_r depend on the discount factor[7] γ.*

An immediate consequence of the above result is that Simplex-PI is *strongly polynomial* for sets of MDPs that are much larger than the deterministic MDPs mentioned in Theorem 5.

Corollary 2. *For any family of MDPs indexed by n and m such that τ_t and τ_r are polynomial functions of n and m, Simplex-PI terminates after a number of steps that is polynomial in n and m.*

4 Similar results for Howard's PI?

One may then wonder whether similar results can be derived for Howard's PI. Unfortunately, and as briefly mentioned by Post and Ye (2013), the line of analysis developed for Simplex-PI does not seem to adapt easily to Howard's PI, because simultaneously switching several actions can interfere in a way such that the policy improvement turns out to be small. We can be more precise on what actually breaks in the approach we have described so far. On the one hand, it is possible to write counterparts of Lemmas 4 and 6 for Howard's PI (see Section 10 for proofs).

Lemma 8. *If the MDP is deterministic, after at most n iterations, either Howard's PI finishes or a new cycle appears.*

Lemma 9. *After at most $(m-n)\lceil n^2\tau_t \log(n^2\tau_t)\rceil + n\lceil n^2\tau_t \log(n^2)\rceil$ iterations, either Howard's PI finishes or a new recurrent class appears.*

On the other hand, we did not manage to adapt Lemma 5 nor Lemma 7. In fact, it is unlikely that a result similar to that of Lemma 5 will be shown to hold for Howard's PI. In a recent deterministic example due to Hansen and Zwick (2010) to show that Howard's PI may require at least $\Omega(n^2)$ iterations, new cycles are created every single iteration but the sequence of values satisfies[8] for all iterations $k < \frac{n^2}{4} + \frac{n}{4}$ and states i,

$$v_*(i) - v_{\pi_{k+1}}(i) \geq \left[1 - \left(\frac{2}{n}\right)^k\right](v_*(i) - v_{\pi_k}(i)).$$

[7] Define the following γ-discounted variation of $\tau^\pi(i,t)$: $\tau^\pi_\gamma(i,t) = \mathbb{E}\left[\sum_{k=0}^{t-1}\gamma^k \mathbf{1}_{i_t=i} \mid i_0 \sim U,\ a_t = \pi(i_t)\right] = \sum_{k=0}^{t-1}\gamma^k \mathbb{P}(i_t = i \mid i_0 \sim U,\ a_t = \pi(i_t))$ and $\tau^\pi_\gamma(i) = \lim_{t\to\infty}\tau^\pi_\gamma(i,t)$. Assume that we have constants τ^γ_t, and τ^γ_r such that for every policy π, $\tau^\pi_\gamma(i) \leq \tau^\gamma_t$ if i is a transient state for π, and $\frac{1}{(1-\gamma)\tau^\pi_\gamma(i)} \leq \tau^\gamma_r$ if i is recurrent for π. Then, one can derive a bound similar to that of Theorem 6 where τ_t and τ_r are respectively replaced by τ^γ_t and $\frac{\tau^\gamma_r}{n}$. At a more technical level, our analysis begins by removing the dependency with respect to γ: Lemma 11, page 11, shows that for every policy π, $\tau^\pi_\gamma(i) \leq \tau_t$ if i is a transient state for π, and $\frac{1}{(1-\gamma)\tau^\pi_\gamma(i)} \leq n\tau_r$ if i is recurrent for π (this is where we pay the $O(n)$ term because the upper bound is $n\tau_r$ instead of τ^γ_r); we then follow the line of arguments originally given by Post and Ye (2013), though our more general setting induces a few technicalities (in particular in the second part of the proof of Lemma 13 page 12).

[8] This MDP has an even number of states $n = 2p$. The goal is to minimize the long term expected cost. The optimal value function satisfies $v_*(i) = -p^N$ for all i, with $N = p^2 + p$. The policies generated by Howard's PI have values $v_{\pi_k}(i) \in (p^{N-k-1}, p^{N-k})$. We deduce that for all iterations k and states i, $\frac{v_*(i) - v_{\pi_{k+1}}(i)}{v_*(i) - v_{\pi_k}(i)} \geq \frac{1+p^{-k-2}}{1+p^{-k}} = 1 - \frac{p^{-k}-p^{-k-2}}{1+p^{-k}} \geq 1 - p^{-k}(1-p^{-2}) \geq 1 - p^{-k}$.

Contrary to Lemma 5, as k grows, the amount of contraction gets (exponentially) smaller and smaller. With respect to Simplex-PI, this suggests that Howard's PI may suffer from subtle specific pathologies. In fact, the problem of determining the number of iterations required by Howard's PI has been challenging for almost 30 years. It was originally identified as an open problem by Schmitz (1985). In the simplest—deterministic—case, the complexity is still an open problem: the currently best-known lower bound is $O(n^2)$ (Hansen and Zwick, 2010), while the best known upper bound is $O(\frac{m^n}{n})$ Mansour and Singh (1999); Hollanders *et al.* (2014).

On the positive side, an adaptation of the line of proof we have considered so far can be carried out under the following assumption.

Assumption 1. *The state space X can be partitioned in two sets $\mathcal{T}$ and $\mathcal{R}$ such that for all policies π, the states of $\mathcal{T}$ are transient and those of $\mathcal{R}$ are recurrent.*

Under this additional assumption, we can deduce the following original bounds.

Theorem 7 (Proof in Section 11)**.** *If the MDP satisfies Assumption 1, then Howard's PI and Simplex-PI terminate after at most*

$$(m-n)\left(\lceil n\tau_r \log n^2\tau_r\rceil + \lceil n^2\tau_t \log n^2\tau_t\rceil\right) \;=\; \tilde{O}(mn(n^2\tau_t + n\tau_r))$$

iterations.

It should however be noted that Assumption 1 is rather restrictive. It implies that the algorithms converge on the recurrent states independently of the transient states, and thus the analysis can be decomposed in two phases: 1) the convergence on recurrent states and then 2) the convergence on transient states (given that recurrent states do not change anymore). The analysis of the first phase (convergence on recurrent states) is greatly facilitated by the fact that in this case, a new recurrent class appears every single iteration (this is in contrast with Lemmas 4, 6, 8 and 9 that were designed to show under which conditions cycles and recurrent classes are created). Furthermore, the analysis of the second phase (convergence on transient states) is similar to that of the discounted case of Theorems 3 and 4. In other words, this last result sheds some light on the practical efficiency of Howard's PI and Simplex-PI, and a general analysis of Howard's PI is still largely open, and constitutes intriguing future work.

The following sections contains detailed proofs of Lemmas 2 and 3, Theorems 3, 4, and 6, Lemmas 8 and 9, and finally Theorem 7. Before we start, we provide a particularly useful identity relating the difference between the values of two policies π and π' and the relative advantage $a_\pi^{\pi'}$.

Lemma 10. *For all pairs of policies π and π',*

$$v_{\pi'} - v_\pi = (I - \gamma P_{\pi'})^{-1} a_\pi^{\pi'} = (I - \gamma P_\pi)^{-1}(-a_{\pi'}^{\pi}).$$

Proof. This first identity follows from simple linear algebra arguments:

$$\begin{aligned} v_{\pi'} - v_\pi &= (I - \gamma P_{\pi'})^{-1} r_{\pi'} - v_\pi && \{v_{\pi'} = T_{\pi'} v_{\pi'} \;\Leftrightarrow\; v_{\pi'} = (I - \gamma P_{\pi'})^{-1} r_{\pi'}\} \\ &= (I - \gamma P_{\pi'})^{-1}(r_{\pi'} + \gamma P_{\pi'} v_\pi - v_\pi) \\ &= (I - \gamma P_{\pi'})^{-1}(T_{\pi'} v_\pi - v_\pi). \end{aligned}$$

The second identity follows by symmetry. □

We will repeatedly use the following property: since for any policy π, the matrix $(1-\gamma)(I-\gamma P)^{-1} = (1-\gamma)\sum_{t=0}^{\infty}(\gamma P_\pi)^t$ is a stochastic matrix (as a mixture of stochastic matrices), then

$$\|(I - \gamma P)^{-1}\|_\infty = \frac{1}{1-\gamma},$$

where $\|\cdot\|_\infty$ is the natural induced max-norm on matrices. Finally, for any vector/matrix A and any number λ, we shall use the notation "$A \geq \lambda$" (respectively "$A \leq \lambda$") for denoting the fact that "all the coefficients of A are greater or equal to (respectively smaller or equal to) λ".

5 Contraction property for Howard's PI (Proof of Lemma 2)

For any k, we have

$$\begin{aligned} v_{\pi_*} - v_{\pi_k} &= T_{\pi_*} v_{\pi_*} - T_{\pi_*} v_{\pi_{k-1}} + T_{\pi_*} v_{\pi_{k-1}} - T_{\pi_k} v_{\pi_{k-1}} + T_{\pi_k} v_{\pi_{k-1}} - T_{\pi_k} v_{\pi_k} && \{\forall \pi,\ T_\pi v_\pi = v_\pi\} \\ &\leq \gamma P_{\pi_*}(v_{\pi_*} - v_{\pi_{k-1}}) + \gamma P_{\pi_k}(v_{\pi_{k-1}} - v_{\pi_k}) && \{T_{\pi_*} v_{\pi_{k-1}} \leq T_{\pi_k} v_{\pi_{k-1}}\} \\ &\leq \gamma P_{\pi_*}(v_{\pi_*} - v_{\pi_{k-1}}). && \{\text{Lemma 1 and } P_{\pi_k} \geq 0\} \end{aligned}$$

Since $v_{\pi_*} - v_{\pi_k}$ is non-negative, we can take the max-norm and get:

$$\|v_{\pi_*} - v_{\pi_k}\|_\infty \leq \gamma \|v_{\pi_*} - v_{\pi_{k-1}}\|_\infty.$$

6 Contraction property for Simplex-PI (Proof of Lemma 3)

The proof we provide here is very close to the one given by Ye (2011). We provide it here for completeness, and also because it resembles the proofs we will provide for the bounds that are independent of γ.

On the one hand, using Lemma 10, we have for any k:

$$\begin{aligned} v_{\pi_{k+1}} - v_{\pi_k} &= (I - \gamma P_{\pi_{k+1}})^{-1} a_{\pi_k}^{\pi_{k+1}} \\ &\geq a_{\pi_k}^{\pi_{k+1}}, && \{(I - \gamma P_{\pi_{k+1}})^{-1} - I \geq 0 \text{ and } a_{\pi_k}^{\pi_{k+1}} \geq 0\} \end{aligned}$$

which implies, by left multiplying by the vector $\mathbb{1}^T$, that

$$\mathbb{1}^T (v_{\pi_{k+1}} - v_{\pi_k}) \geq \mathbb{1}^T a_{\pi_k}^{\pi_{k+1}}. \tag{1}$$

On the other hand, we have:

$$\begin{aligned} v_{\pi_*} - v_{\pi_k} &= (I - \gamma P_{\pi_*})^{-1} a_{\pi_k}^{\pi_*} && \{\text{Lemma 10}\} \\ &\leq \frac{1}{1-\gamma} \max_s a_{\pi_k}^{\pi_{k+1}}(s) \quad \{\|(I - \gamma P_{\pi_*})^{-1}\|_\infty = \frac{1}{1-\gamma} \text{ and } \max_s a_{\pi_k}^{\pi_{k+1}}(s) = \max_{s,\pi} a_{\pi_k}^{\pi}(s) \geq 0\} \\ &\leq \frac{1}{1-\gamma} \mathbb{1}^T a_{\pi_k}^{\pi_{k+1}}, && \{\forall x \geq 0,\ \max_s x(s) \leq \mathbb{1}^T x\} \end{aligned}$$

which implies that

$$\begin{aligned} \mathbb{1}^T a_{\pi_k}^{\pi_{k+1}} &\geq (1-\gamma)\|v_{\pi_*} - v_{\pi_k}\|_\infty \\ &\geq \frac{1-\gamma}{n} \mathbb{1}^T (v_{\pi_*} - v_{\pi_k}). && \{\forall x,\ \mathbb{1}^T x \leq n \|x\|_\infty\} \end{aligned} \tag{2}$$

Combining Equations (1) and (2), we get:

$$\begin{aligned} \mathbb{1}^T (v_{\pi_*} - v_{\pi_{k+1}}) = \mathbb{1}^T (v_{\pi_*} - v_{\pi_k}) - \mathbb{1}^T (v_{\pi_{k+1}} - v_{\pi_k}) &\leq \mathbb{1}^T (v_{\pi_*} - v_{\pi_k}) - \frac{1-\gamma}{n} \mathbb{1}^T (v_{\pi_*} - v_{\pi_k}) \\ &= \left(1 - \frac{1-\gamma}{n}\right) \mathbb{1}^T (v_{\pi_*} - v_{\pi_k}). \end{aligned}$$

7 A bound for Howard's PI when $\gamma < 1$ (Proof of Theorem 3)

Although the overall line or arguments follows from those given originally by Ye (2011) and adapted by Hansen *et al.* (2013), our proof is slightly more direct and leads to a better result.

For any k, we have:

$$\begin{aligned} -a_{\pi_*}^{\pi_k} &= (I - \gamma P_{\pi_k})(v_* - v_{\pi_k}) && \{\text{Lemma 10}\} \\ &\leq v_* - v_{\pi_k}. && \{v_* - v_{\pi_k} \geq 0 \text{ and } P_{\pi_k} \geq 0\} \end{aligned}$$

By the optimality of π_*, $-a_{\pi_*}^{\pi_k}$ is non-negative, and we can take the max-norm:

$$\begin{aligned}
\|a_{\pi_*}^{\pi_k}\|_\infty &\leq \|v_* - v_{\pi_k}\|_\infty \\
&\leq \gamma^k \|v_{\pi_*} - v_{\pi_0}\|_\infty && \{\text{Lemma 2}\} \\
&= \gamma^k \|(I - \gamma P_{\pi_0})^{-1}(-a_{\pi_*}^{\pi_0})\|_\infty && \{\text{Lemma 10}\} \\
&\leq \frac{\gamma^k}{1-\gamma} \|a_{\pi_*}^{\pi_0}\|_\infty. && \{\|(I - \gamma P_{\pi_0})^{-1}\|_\infty = \frac{1}{1-\gamma}\}
\end{aligned}$$

By definition of the max-norm, and as $a_{\pi_*}^{\pi_0} \leq 0$ (using again the fact that π_* is optimal), there exists a state s_0 such that $-a_{\pi_*}^{\pi_0}(s_0) = \|a_{\pi_*}^{\pi_0}\|_\infty$. We deduce that for all k,

$$-a_{\pi_*}^{\pi_k}(s_0) \leq \|a_{\pi_*}^{\pi_k}\|_\infty \leq \frac{\gamma^k}{1-\gamma} \|a_{\pi_*}^{\pi_0}\|_\infty = \frac{\gamma^k}{1-\gamma}(-a_{\pi_*}^{\pi_0}(s_0)).$$

As a consequence, the action $\pi_k(s_0)$ must be different from $\pi_0(s_0)$ when $\frac{\gamma^k}{1-\gamma} < 1$, that is for all values of k satisfying

$$k \geq k^* = \left\lceil \frac{\log \frac{1}{1-\gamma}}{1-\gamma} \right\rceil > \left\lceil \frac{\log \frac{1}{1-\gamma}}{\log \frac{1}{\gamma}} \right\rceil.$$

In other words, if some policy π is not optimal, then one of its non-optimal actions will be eliminated *for good* after at most k^* iterations. By repeating this argument, one can eliminate all non-optimal actions (there are at most $n - m$ of them), and the result follows.

8 A bound for Simplex-PI when $\gamma < 1$ (Proof of Theorem 4)

At each iteration k, let s_k be the state in which an action is switched. We have (by definition of Simplex-PI):

$$a_{\pi_k}^{\pi_{k+1}}(s_k) = \max_{\pi, s} a_{\pi_k}^{\pi}(s).$$

Starting with arguments similar to those for the contraction property of Simplex-PI, we have on the one hand:

$$\begin{aligned}
v_{\pi_{k+1}} - v_{\pi_k} &= (I - \gamma P_{\pi_{k+1}})^{-1} a_{\pi_k}^{\pi_{k+1}} && \{\text{Lemma 10}\} \\
&\geq a_{\pi_k}^{\pi_{k+1}}, && \{(I - \gamma P_{\pi_{k+1}})^{-1} - I \geq 0 \text{ and } a_{\pi_k}^{\pi_{k+1}} \geq 0\}
\end{aligned}$$

which implies that

$$v_{\pi_{k+1}}(s_k) - v_{\pi_k}(s_k) \geq a_{\pi_k}^{\pi_{k+1}}(s_k). \tag{3}$$

On the other hand, we have:

$$\begin{aligned}
v_{\pi_*} - v_{\pi_k} &= (I - \gamma P_{\pi_*})^{-1} a_{\pi_k}^{\pi_*} && \{\text{Lemma 10}\} \\
&\leq \frac{1}{1-\gamma} a_{\pi_k}^{\pi_{k+1}}(s_k) && \{\|(I - \gamma P_{\pi_*})^{-1}\|_\infty = \frac{1}{1-\gamma} \text{ and } a_{\pi_k}^{\pi_{k+1}}(s_k) = \max_{s,\pi} a_{\pi_k}^{\pi}(s) \geq 0\}
\end{aligned}$$

which implies that

$$\|v_{\pi_*} - v_{\pi_k}\|_\infty \leq \frac{1}{1-\gamma} a_{\pi_k}^{\pi_{k+1}}(s_k). \tag{4}$$

Write $\Delta_k = v_{\pi_*} - v_{\pi_k}$. From Equations (3) and (4), we deduce that:

$$\Delta_{k+1}(s_k) \leq \Delta_k(s_k) - (1-\gamma)\|\Delta_k\|_\infty = \left(1 - (1-\gamma)\frac{\|\Delta_k\|_\infty}{\Delta_k(s_k)}\right) \Delta_k(s_k).$$

This implies—since $\Delta_k(s_k) \leq \|\Delta_k\|_\infty$—that

$$\Delta_{k+1}(s_k) \leq \gamma \Delta_k(s_k),$$

but also—since $\Delta_k(s_k)$ and $\Delta_{k+1}(s_k)$ are non-negative and thus $\left(1 - (1-\gamma)\frac{\|\Delta_k\|_\infty}{\Delta_k(s_k)}\right) \geq 0$—that

$$\|\Delta_k\|_\infty \leq \frac{1}{1-\gamma}\Delta_k(s_k).$$

Now, write n_k for the vector on the state space such that $n_k(s)$ is the number of times state s has been switched until iteration k (including k). Since by Lemma 1 the sequence $(\Delta_k)_{k\geq 0}$ is non-increasing, we have

$$\|\Delta_k\|_\infty \leq \frac{1}{1-\gamma}\Delta_k(s_k) \leq \frac{\gamma^{n_{k-1}(s_k)}}{1-\gamma}\Delta_0(s_k) \leq \frac{\gamma^{n_{k-1}(s_k)}}{1-\gamma}\|\Delta_0\|_\infty. \tag{5}$$

At any iteration k, let $s_k^* = \arg\max_s n_{k-1}(s)$ be the state in which actions have been switched the most. Since at each iteration k, one of the n components of n_k is increased by 1, we necessarily have

$$n_{k-1}(s_k^*) \geq \left\lfloor \frac{k-1}{n} \right\rfloor \geq \frac{k-n}{n}. \tag{6}$$

Write $k^* \leq k-1$ for the last iteration when the state s_k^* was updated, such that we have

$$n_{k-1}(s_k^*) = n_{k^*-1}(s_{k^*}). \tag{7}$$

Since $(\|\Delta_k\|_\infty)_{k\geq 0}$ is nonincreasing (using again Lemma 1), we have

$$\begin{aligned}
\|\Delta_k\|_\infty &\leq \|\Delta_{k^*}\|_\infty && \{k^* \leq k-1\}\\
&\leq \frac{\gamma^{n_{k^*-1}(s_{k^*})}}{1-\gamma}\|\Delta_0\|_\infty && \{\text{Equation (5)}\}\\
&= \frac{\gamma^{n_{k-1}(s_k^*)}}{1-\gamma}\|\Delta_0\|_\infty && \{\text{Equation (7)}\}\\
&\leq \frac{\gamma^{\frac{k-n}{n}}}{1-\gamma}\|\Delta_0\|_\infty. && \{\text{Equation (6) and } x \mapsto \gamma^x \text{ is decreasing}\}
\end{aligned}$$

We are now ready to finish the proof. By using arguments similar to those for Howard's PI, we have:

$$\|a_{\pi_*}^{\pi_k}\|_\infty \leq \|\Delta_k\|_\infty \leq \frac{\gamma^{\frac{k-n}{n}}}{1-\gamma}\|\Delta_0\|_\infty \leq \frac{\gamma^{\frac{k-n}{n}}}{(1-\gamma)^2}\|a_{\pi_*}^{\pi_0}\|_\infty.$$

In particular, we can deduce from the above relation that as soon as $\frac{\gamma^{\frac{k-n}{n}}}{(1-\gamma)^2} < 1$, that is for instance when $k > k^* = n\left(1 + \frac{2}{1-\gamma}\log\frac{1}{1-\gamma}\right)$, one of the non-optimal actions of π_0 cannot appear in π_k. Thus, every k^* iterations, a non-optimal action is eliminated for good, and the result follows from the fact that there are at most $n-m$ non-optimal actions.

9 A general bound for Simplex-PI (Proof of Theorem 6)

The proof we give here is strongly inspired by that for the deterministic case of Post and Ye (2013): the steps (a series of lemmas) are similar. There are mainly two differences. First, our arguments are *more direct* in the sense that we do not refer to linear programming, but only provide simple linear algebra arguments. Second, it is *more general*: for any policy π, we consider the set of transient states (respectively recurrent classes) instead of the set of path states (respectively cycles); it slightly complicates the arguments, the most complicated extension being the second part of the proof of the forthcoming Lemma 13.

Consider the vector $x_\pi = (I - \gamma P_\pi^T)^{-1}\mathbb{1}$ that provides a discounted measure of state visitations along the trajectories induced by a policy π starting from the uniform distribution U on the state space X:

$$\forall i \in X, \quad x_\pi(i) = n \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(i_t = i \mid i_0 \sim U,\ a_t = \pi(i_t)).$$

This vector plays a crucial role in the analysis. For any policy π and state i, we trivially have $x_\pi(i) \in \left(1, \frac{n}{1-\gamma}\right)$. In the case of deterministic MDPs, Post and Ye (2013)) exploits the fact that $x_\pi(i)$ belongs to the set $(1, n)$ when i is on path of π, while $x_\pi(i)$ belongs to the set $(\frac{1}{1-\gamma}, \frac{n}{1-\gamma})$ when i is on a cycle of π. Our extension of their result to the case of general (stochastic) MDPs will rely on the following result. For any policy π, we shall write $\mathcal{R}(\pi)$ for the set of states that are recurrent for π.

Lemma 11. *With the constants τ_t and τ_r of Definition 1, we have for every discount factor γ,*

$$\forall i \notin \mathcal{R}(\pi), \quad 1 \ \leq \ x_\pi(i) \ \leq \ n\tau_t \tag{8}$$

$$\forall i \in \mathcal{R}(\pi), \quad \frac{1}{\tau_r} \ \leq \ (1-\gamma)x_\pi(i) \ \leq \ n. \tag{9}$$

Proof. The fact that $x_\pi(i)$ belongs to $\left(1, \frac{n}{1-\gamma}\right)$ is obvious from the definition of x_π. The upper bound on x_π on the transient states i follows from the fact that for any policy π,

$$\begin{aligned}
\tau_t(i) &\geq \lim_{t\to\infty} \tau^\pi(i,t) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(i_t = i \mid i_0 \sim U,\ a_t = \pi(i_t)) \\
&\geq \sum_{k=0}^{\infty} \gamma^k \mathbb{P}(i_t = i \mid i_0 \sim U,\ a_t = \pi(i_t)) \\
&= \frac{1}{n} x_\pi(i).
\end{aligned}$$

Let us now consider the lower bound on $(1-\gamma)x_\pi(i)$ when i is a recurrent state of some policy π. In general, the asymptotic frequency μ^π of π does not necessarily satisfy ${\mu^\pi}^T P_\pi = P_\pi$ because P_π may correspond to an aperiodic or reducible chain. To deal with this issue, we consider the Cesàro mean

$$Q_\pi = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} (P_\pi)^k$$

that is well-defined (Stroock, 2005, Section 3.2). It can be shown (Fritz *et al.*, 1979, Proposition 3.5(a)) that $Q_\pi = Q_\pi P_\pi = P_\pi Q_\pi = Q_\pi Q_\pi$. This implies in particular that

$$(1-\gamma)Q_\pi(I - \gamma P_\pi)^{-1} = (1-\gamma)\sum_{k=0}^{\infty} \gamma^k Q_\pi (P_\pi)^k = (1-\gamma)\sum_{k=0}^{\infty} \gamma^k Q_\pi = Q_\pi. \tag{10}$$

Then, by using twice the fact that $\mu^\pi = \frac{1}{n}{Q_\pi}^T\mathbb{1}$, we can see that for all recurrent states i,

$$\begin{aligned}
\frac{1}{\tau_r} &\leq \mu^\pi(i) \\
&= \left[\frac{1}{n}{Q_\pi}^T\mathbb{1}\right](i) \\
&= \left[\frac{1}{n}(1-\gamma)(I - \gamma {P_\pi}^T)^{-1}{Q_\pi}^T\mathbb{1}\right](i) && \{\text{Equation (10)}\} \\
&= \left[(1-\gamma)(I - \gamma {P_\pi}^T)^{-1}\mu^\pi\right](i) \\
&\leq \left[(1-\gamma)(I - \gamma {P_\pi}^T)^{-1}\mathbb{1}\right](i) && \{\mu^\pi \leq \mathbb{1}\} \\
&= (1-\gamma)x_\pi(i).
\end{aligned}$$

□

Finally, a rewriting of Lemma 10 in terms of the vector x_π will be useful in the following proofs: for any pair of policies π and π',

$$\mathbb{1}^T(v_{\pi'} - v_\pi) = {x_{\pi'}}^T a_\pi^{\pi'} = {x_\pi}^T(-a_{\pi'}^\pi). \tag{11}$$

We are now ready to delve into the details of the arguments. As mentioned before, the proof is structured in two steps: first, we will show that recurrent classes are created often; then we will show that significant progress is made every time a new recurrent class appears.

9.1 Part 1: Recurrent classes are created often

Lemma 12. *Suppose one moves from policy π to policy π'* without creating any recurrent class. *Let $\pi_\dagger$ be the final policy before either a new recurrent class appears or Simplex-PI terminates. Then*

$$\mathbb{1}^T(v_{\pi_\dagger} - v_{\pi'}) \leq \left(1 - \frac{1}{n^2\tau_t}\right) \mathbb{1}^T(v_{\pi_\dagger} - v_\pi).$$

Proof. The arguments are similar to those for the proof of Theorem 4. On the one hand, we have:

$$\mathbb{1}^T(v_{\pi'} - v_\pi) \geq \mathbb{1}^T a_\pi^{\pi'}. \tag{12}$$

On the other hand, we have

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_\dagger} - v_\pi) &= x_{\pi_\dagger}^T a_\pi^{\pi_\dagger} && \{\text{Equation (11)}\} \\
&= \sum_{s \notin \mathcal{R}(\pi_\dagger)} x_{\pi_\dagger}(s) a_\pi^{\pi_\dagger}(s) + \sum_{s \in \mathcal{R}(\pi_\dagger)} x_{\pi_\dagger}(s) a_\pi^{\pi_\dagger}(s) \\
&\leq n^2\tau_t \max_{s \notin \mathcal{R}(\pi_\dagger)} a_\pi^{\pi_\dagger}(s) + \frac{n^2}{1-\gamma} \max_{s \in \mathcal{R}(\pi_\dagger)} a_\pi^{\pi_\dagger}(s). && \{\text{Equations (8)-(9)}\}
\end{aligned}$$

Since by assumption recurrent classes of $\pi_\dagger$ are also recurrent classes of π, we deduce that for all $s \in \mathcal{R}(\pi_\dagger)$, $\pi_\dagger(s) = \pi(s)$, so that $\max_{s \in \mathcal{R}(\pi_\dagger)} a_\pi^{\pi_\dagger}(s) = 0$. Thus, the second term of the above r.h.s. is null and

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_\dagger} - v_\pi) &\leq n^2\tau_t \max_s a_\pi^{\pi_\dagger}(s) \\
&\leq n^2\tau_t \max_s a_\pi^{\pi'}(s) && \{\max_s T_{\pi'} v_\pi(s) = \max_{s,\tilde{\pi}} T_{\tilde{\pi}} v_\pi(s)\} \\
&\leq n^2\tau_t \mathbb{1}^T a_\pi^{\pi'}. && \{\forall x \geq 0,\ \max_s x(s) \leq \mathbb{1}^T x\}
\end{aligned} \tag{13}$$

Combining Equations (12) and (13), we get:

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_\dagger} - v_{\pi'}) &= \mathbb{1}^T(v_{\pi_\dagger} - v_\pi) - \mathbb{1}^T(v_{\pi'} - v_\pi) \\
&\leq \left(1 - \frac{1}{n^2\tau_t}\right) \mathbb{1}^T(v_{\pi_\dagger} - v_\pi).
\end{aligned}$$

□

Lemma 13. *While Simplex-PI does not create any recurrent class nor finishes,*

- *either an action is eliminated from policies after at most $\lceil n^2\tau_t \log(n^2\tau_t) \rceil$ iterations,*
- *or a recurrent class is broken after at most $\lceil n^2\tau_t \log(n^2) \rceil$ iterations.*

Proof. Let π be the policy in some iteration. Let $\pi_\dagger$ be the last policy before a new recurrent class appears, and π' a policy generated after k iterations from π. We shall prove that one of the two events stated of the lemma must happen.
Since

$$\begin{aligned}
0 \leq \mathbb{1}^T(v_{\pi_\dagger} - v_\pi) & && \{v_{\pi_\dagger} \geq v_\pi\} \\
= {x_\pi}^T(-a_{\pi_\dagger}^\pi) & && \{\text{Equation (11)}\} \\
= \sum_{s \notin \mathcal{R}(\pi)} x_\pi(s)(-a_{\pi_\dagger}^\pi(s)) + \sum_{C \in \mathcal{R}(\pi)} \sum_{s \in C} x_\pi(s)(-a_{\pi_\dagger}^\pi(s))
\end{aligned}$$

there must exist either a state $s_0 \notin \mathcal{R}(\pi)$ such that

$$x_\pi(s_0)(-a^\pi_{\pi_\dagger}(s_0)) \geq \frac{1}{n} {x_\pi}^T(-a^\pi_{\pi_\dagger}) \geq 0. \tag{14}$$

or a recurrent class R_0 such that

$$\sum_{s\in R_0} x_\pi(s)(-a^\pi_{\pi_\dagger}(s)) \geq \frac{1}{n} {x_\pi}^T(-a^\pi_{\pi_\dagger}) \geq 0. \tag{15}$$

We consider these two cases separately below.

- **case 1:** Equation (14) holds for some $s_0 \notin \mathcal{R}(\pi)$. Let us prove by contradiction that for k sufficiently big, $\pi'(s_0) \neq \pi(s_0)$: let us assume that $\pi'(s_0) = \pi(s_0)$. Then

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_\dagger} - v_{\pi'}) &\geq v_{\pi_\dagger}(s_0) - v_{\pi'}(s_0) && \{v_{\pi_\dagger} \geq v_{\pi'}\} \\
&= v_{\pi_\dagger}(s_0) - T_{\pi'} v_{\pi'}(s_0) && \{v_{\pi'} = T_{\pi'} v_{\pi'}\} \\
&\geq v_{\pi_\dagger}(s_0) - T_{\pi'} v_{\pi_\dagger}(s_0) && \{v_{\pi_\dagger} \geq v_{\pi'}\} \\
&= -a^{\pi'}_{\pi_\dagger}(s_0) \\
&= -a^{\pi}_{\pi_\dagger}(s_0) && \{\pi(s_0) = \pi'(s_0)\} \\
&\geq \frac{1}{n\tau_t} x_\pi(s_0)(-a^\pi_{\pi_\dagger}(s_0)) && \{\text{Equation (8)}\} \\
&\geq \frac{1}{n^2\tau_t} {x_\pi}^T(-a^\pi_{\pi_\dagger}) && \{\text{Equation (14)}\} \\
&= \frac{1}{n^2\tau_t} \mathbb{1}^T(v_{\pi_\dagger} - v_\pi). && \{\text{Equation (11)}\}
\end{aligned}$$

  If there is no recurrent class creation, the contraction property given in Lemma 12 implies that if π' is obtained after $k = \lceil n^2\tau_t \log(n^2\tau_t)\rceil > \frac{\log(n^2\tau_t)}{\log\frac{1}{1-\frac{1}{n^2\tau_t}}}$ iterations, then

$$\mathbb{1}^T(v_{\pi_\dagger} - v_{\pi'}) < \frac{1}{n^2\tau_t} \mathbb{1}^T(v_{\pi_\dagger} - v_\pi),$$

  and we get a contradiction. As a conclusion, we necessarily have $\pi'(s_0) \neq \pi(s_0)$.

- **case 2:** Equation (15) holds for some R_0 that is a recurrent class of π. Let us prove by contradiction that for k sufficiently big, R_0 cannot be a recurrent class of π': let us thus assume that R_0 is a recurrent class of π'. Write $\mathcal{T}$ for the set of states that are transient for π (formally, $\mathcal{T} = X\backslash\mathcal{R}(\pi)$). For any subset Y of the state space X, write P^Y_π for the stochastic matrix of which the i^{th} row is equal to that of P_π if $i \in Y$, and is 0 otherwise, and write $\mathbb{1}_Y$ the vectors of which the i^{th} component is equal to 1 if $i \in Y$ and 0 otherwise.

  Using the fact that $P^{R_0}_\pi P^{\mathcal{T}}_\pi = 0$, one can first observe that

$$(I - \gamma P^{R_0}_\pi)(I - \gamma P^{\mathcal{T}}_\pi) = I - \gamma(P^{R_0}_\pi + P^{\mathcal{T}}_\pi),$$

  from which we can deduce that

$$\begin{aligned}
\forall s \in R_0,\ \left[\mathbb{1}_{\mathcal{T}\cup R_0}{}^T (I - \gamma P_\pi)^{-1}\right](s) &= \left[\mathbb{1}_{\mathcal{T}\cup R_0}{}^T (I - \gamma(P^{R_0}_\pi + P^{\mathcal{T}}_\pi))^{-1}\right](s) \\
&= \left[\mathbb{1}_{\mathcal{T}\cup R_0}{}^T (I - \gamma P^{\mathcal{T}}_\pi)^{-1}(I - \gamma P^{R_0}_\pi)^{-1}\right](s).
\end{aligned} \tag{16}$$

  Also, let s be an arbitrary state and s' be a state of R_0. Since $\mathbb{1}^T_s (P^{\mathcal{T}}_\pi)^k(s')$ is the probability that the chain starting in s reaches s' for the first time after k iterations, then

$$\mathbb{1}^T_s (I - \gamma P^{\mathcal{T}}_\pi)^{-1}(s') \leq \sum_{t=0}^{\infty} \mathbb{1}^T_s (P^{\mathcal{T}}_\pi)^i(s') \leq 1.$$

and therefore,

$$\forall s' \in R_0, \quad \mathbb{1}_{\mathcal{T}\cup R_0}{}^T (I - \gamma P_\pi^{\mathcal{T}})^{-1}(s') \leq n. \tag{17}$$

Writing δ for the vector that equals $-a^\pi_{\pi_\dagger}$ on R_0 and that is null everywhere else, we have

$$\begin{aligned}
&\sum_{s\in R_0} x_\pi(s)(-a^\pi_{\pi_\dagger}(s)) \\
&= \sum_{s\in R_0} [(I-\gamma P_\pi^T)^{-1}\mathbb{1}](s)\delta(s) \\
&= \sum_{s\in R_0} [(I-\gamma P_\pi^T)^{-1}\mathbb{1}_{\mathcal{T}\cup R_0}](s)\delta(s) && \left\{\forall s \in R_0,\ [(I-\gamma P_\pi^T)^{-1}\mathbb{1}_{X\setminus(\mathcal{T}\cup R_0)}](s) = 0\right\} \\
&= \sum_s [(I-\gamma P_\pi^T)^{-1}\mathbb{1}_{\mathcal{T}\cup R_0}](s)\delta(s) && \{\forall s \notin R_0,\ \delta(s) = 0\} \\
&= \mathbb{1}_{\mathcal{T}\cup R_0}{}^T (I-\gamma P_\pi)^{-1}\delta \\
&= \mathbb{1}_{\mathcal{T}\cup R_0}{}^T (I-\gamma P_\pi^{\mathcal{T}})^{-1}(I-\gamma P_\pi^{R_0})^{-1}\delta && \{\text{Equation (16)}\} \\
&= \sum_s [(I-\gamma {P_\pi^{\mathcal{T}}}^T)^{-1}\mathbb{1}_{\mathcal{T}\cup R_0}](s)[(I-\gamma P_\pi^{R_0})^{-1}\delta](s) \\
&= \sum_{s\in R_0} [(I-\gamma {P_\pi^{\mathcal{T}}}^T)^{-1}\mathbb{1}_{\mathcal{T}\cup R_0}](s)[(I-\gamma P_\pi^{R_0})^{-1}\delta](s) && \{\forall s \notin R_0,\ \delta(s) = 0\} \\
&= \sum_{s\in R_0} [(I-\gamma {P_\pi^{\mathcal{T}}}^T)^{-1}\mathbb{1}_{\mathcal{T}\cup R_0}](s)(v_{\pi_\dagger}(s) - v_\pi(s)) && \{\text{Lemma 10}\} \\
&\leq n\mathbb{1}_{R_0}{}^T(v_{\pi_\dagger} - v_\pi). && \{\text{Equation (17)}\}
\end{aligned} \tag{18}$$

We assumed that R_0 is also a recurrent class of π', which implies $\mathbb{1}_{R_0}{}^T v_\pi = \mathbb{1}_{R_0}{}^T v_{\pi'}$, and

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_\dagger} - v_{\pi'}) &\geq \mathbb{1}_{R_0}{}^T(v_{\pi_\dagger} - v_{\pi'}) && \{v_{\pi_\dagger} \geq v_{\pi'}\} \\
&= \mathbb{1}_{R_0}{}^T(v_{\pi_\dagger} - v_\pi) && \{\mathbb{1}_{R_0}{}^T v_\pi = \mathbb{1}_{R_0}{}^T v_{\pi'}\} \\
&\geq \frac{1}{n}\sum_{s\in R_0} x_\pi(s)(-a^\pi_{\pi_\dagger}(s)) && \{\text{Equation (18)}\} \\
&\geq \frac{1}{n^2} x_\pi{}^T(-a^\pi_{\pi_\dagger}) && \{\text{Equation (15)}\} \\
&= \frac{1}{n^2}\mathbb{1}^T(v_{\pi_\dagger} - v_\pi). && \{\text{Equation (11)}\}
\end{aligned}$$

If there is no recurrent class creation, the contraction property given in Lemma 12 implies that if π' is obtained after $k = \lceil n^2\tau_t \log(n^2)\rceil > \frac{\log(n^2)}{\log\frac{1}{1-\frac{1}{n^2\tau_t}}}$ iterations, then

$$\mathbb{1}^T(v_{\pi_\dagger} - v_{\pi'}) < \frac{1}{n^2}\mathbb{1}^T(v_{\pi_\dagger} - v_\pi),$$

and thus we get a contradiction. As a conclusion, R_0 cannot be a recurrent class of π'.

□

A direct consequence of the above result is Lemma 6 that we originally stated on page 5, and that we restate for clarity.

Lemma 6. *After at most* $(m-n)\lceil n^2\tau_t \log(n^2\tau_t)\rceil + n\lceil n^2\tau_t \log(n^2)\rceil$ *iterations, either Simplex-PI finishes or a new recurrent class appears.*

Proof. Before a recurrent class is created, at most n recurrent classes need to be broken and $(m-n)$ actions to be eliminated, and the time required by these events is bounded thanks to the previous lemma. □

9.2 Part 2: A new recurrent class implies a significant step towards the optimal value

We now proceed to the second part of the proof, and begin by proving Lemma 7 (originally stated page 5).

Lemma 7. *When Simplex-PI moves from* π *to* π' *where* π' *involves a new recurrent class, we have*

$$\mathbb{1}^T(v_{\pi_*} - v_{\pi'}) \leq \left(1 - \frac{1}{n\tau_r}\right)\mathbb{1}^T(v_{\pi_*} - v_\pi).$$

Proof. Let s_0 be the state such that $\pi'(s_0) \neq \pi(s_0)$. On the one hand, since π' contains a new recurrent class R (necessarily containing s_0), we have

$$\begin{aligned}
\mathbb{1}^T(v_{\pi'} - v_\pi) &= {x_{\pi'}}^T a_\pi^{\pi'} && \{\text{Equation (11)}\}\\
&= x_{\pi'}(s_0) a_\pi(s_0) && \{\text{Simplex-PI switches 1 action and } a_\pi(s_0) = a_\pi^{\pi'}(s_0)\}\\
&\geq \frac{1}{(1-\gamma)\tau_r} a_\pi(s_0). && \{\text{Equation (9) with } s_0 \in \mathcal{R}(\pi')\} \quad (19)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\forall s, \quad v_{\pi_*}(s) - v_\pi(s) &= [(I - \gamma P_{\pi_*})^{-1} a_\pi^{\pi_*}](s) && \{\text{Lemma 10}\}\\
&\leq \frac{1}{1-\gamma} a_\pi(s_0). && \{\|(I-\gamma P_{\pi_*})^{-1}\|_\infty \leq \frac{1}{1-\gamma} \text{ and } a_\pi(s_0) = \max_{s,\tilde{\pi}} a_\pi^{\tilde{\pi}}(s) \geq 0\} \quad (20)
\end{aligned}$$

Combining these two observations, we obtain

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_*} - v_{\pi'}) &= \mathbb{1}^T(v_{\pi_*} - v_\pi) - \mathbb{1}^T(v_{\pi'} - v_\pi)\\
&\leq \mathbb{1}^T(v_{\pi_*} - v_\pi) - \frac{1}{(1-\gamma)\tau_r} a_\pi(s_0) && \{\text{Equation (19)}\}\\
&\leq \mathbb{1}^T(v_{\pi_*} - v_\pi) - \frac{1}{\tau_r} \max_s v_{\pi*}(s) - v_{\pi'}(s) && \{\text{Equation (20)}\}\\
&\leq \left(1 - \frac{1}{n\tau_r}\right)\mathbb{1}^T(v_{\pi_*} - v_\pi). && \{\forall x,\ \frac{1}{n}\mathbb{1}^T x \leq \max_s x(s)\}
\end{aligned}$$

□

Lemma 14. *While Simplex-PI does not terminate,*

- *either some non-optimal action is eliminated from recurrent states after at most* $\lceil n\tau_r \log(n^2\tau_r)\rceil$ *recurrent class creations,*
- *or some non-optimal action is eliminated from policies after at most* $\lceil n\tau_r \log(n^2\tau_t)\rceil$ *recurrent class creations.*

Proof. Let π be the policy in some iteration and π' the policy generated after k iterations from π (without loss of generality we assume $\pi' \neq \pi_*$). Let $s_0 = \arg\max_s x_\pi(s)(-a^\pi_{\pi_*}(s))$. We have

$$\begin{aligned} x_\pi(s_0)(-a^\pi_{\pi_*}(s_0)) &\geq \frac{1}{n} {x_\pi}^T(-a^\pi_{\pi_*}) && \{\forall x,\ \mathbb{1}^T x \leq n \max_s x(s)\} \\ &= \frac{1}{n}\mathbb{1}^T(v_{\pi_*} - v_\pi). && \{\text{Equation (11)}\} \end{aligned} \tag{21}$$

We now consider two cases, respectively corresponding to $s_0 \notin \mathcal{R}(\pi)$ or $s_0 \in \mathcal{R}(\pi)$.

- **case 1:** $s_0 \notin \mathcal{R}(\pi)$. Let us prove by contradiction that $\pi'(s_0) \neq \pi(s_0)$ if k is sufficiently large: let us assume that $\pi'(s_0) = \pi(s_0)$. Then, by using repeatedly the fact that for all $\tilde{\pi}$, $a^{\tilde{\pi}}_{\pi_*} \leq 0$ (by definition of the optimal policy π_*), we have:

$$\begin{aligned} \mathbb{1}^T(v_{\pi_*} - v_{\pi'}) &= {x_{\pi'}}^T(-a^{\pi'}_{\pi_*}) && \{\text{Equation (11)}\} \\ &\geq x_{\pi'}(s_0)(-a^{\pi'}_{\pi_*}(s_0)) \\ &\geq -a^{\pi'}_{\pi_*}(s_0) && \{x_{\pi'}(s_0) \geq 1\} \\ &= -a^{\pi}_{\pi_*}(s_0) && \{\pi(s_0) = \pi'(s_0)\} \\ &\geq \frac{1}{n\tau_t} x_\pi(s_0)(-a^\pi_{\pi_*}(s_0)) && \{\text{Equation (8)}\} \\ &\geq \frac{1}{n^2\tau_t}\mathbb{1}^T(v_{\pi_*} - v_\pi). && \{\text{Equation (21)}\} \end{aligned}$$

  After $k = \lceil n\tau_r \log n^2\tau_t \rceil > \frac{\log n^2\tau_t}{\log \frac{1}{1-\frac{1}{n\tau_r}}}$ recurrent classes are created, we have by the contraction property of Lemma 7 that

$$\mathbb{1}^T(v_{\pi_*} - v_{\pi'}) < \frac{1}{n^2\tau_t}\mathbb{1}^T(v_{\pi_*} - v_\pi)$$

  and we get a contradiction. As a conclusion, we have $\pi'(s_0) \neq \pi(s_0)$.

- **case 2:** $s_0 \in \mathcal{R}(\pi)$. Let us prove by contradiction that $\pi'(s_0) \neq \pi(s_0)$ if s_0 is recurrent for π' and k is sufficiently large: let us assume that $\pi'(s_0) = \pi(s_0)$ and $s_0 \in \mathcal{R}(\pi')$. Then, by using again the fact that for all $\tilde{\pi}$, $a^{\tilde{\pi}}_{\pi_*} \leq 0$, we have:

$$\begin{aligned} \mathbb{1}^T(v_{\pi_*} - v_{\pi'}) &= {x_{\pi'}}^T(-a^{\pi'}_{\pi_*}) && \{\text{Equation (11)}\} \\ &= \sum_s x_{\pi'}(s)(-a^{\pi'}_{\pi_*}(s)) \\ &\geq \sum_{s \in R_0} x_{\pi'}(s)(-a^{\pi'}_{\pi_*}(s)) \\ &\geq \frac{1}{(1-\gamma)\tau_r} \sum_{s \in R_0} (-a^{\pi'}_{\pi_*}(s)) && \{\text{Equation (9)}\} \\ &\geq \frac{1}{(1-\gamma)\tau_r}(-a^{\pi'}_{\pi_*}(s_0)) \\ &= \frac{1}{(1-\gamma)\tau_r}(-a^{\pi}_{\pi_*}(s_0)) && \{\pi(s_0) = \pi'(s_0)\} \\ &\geq \frac{1}{n\tau_r} x_\pi(s_0)(-a^\pi_{\pi_*}(s_0)) && \{x_\pi(s_0) \leq \frac{n}{1-\gamma}\} \\ &\geq \frac{1}{n^2\tau_r}\mathbb{1}^T(v_{\pi_*} - v_\pi). && \{\text{Equation (21)}\} \end{aligned}$$

  After $k = \lceil n\tau_r \log n^2\tau_r \rceil > \frac{\log n^2\tau_r}{\log \frac{1}{1-\frac{1}{n\tau_r}}}$ new recurrent classes are created, we have by the contraction property of Lemma 7 that

$$\mathbb{1}^T(v_{\pi_*} - v_{\pi'}) < \frac{1}{n^2\tau_r}\mathbb{1}^T(v_{\pi_*} - v_\pi),$$

  and we get a contradiction. As a conclusion, we know that $\pi'(s_0) \neq \pi(s_0)$ if s_0 is recurrent for π'.

□

We are ready to conclude: At most, the $(m-n)$ non-optimal actions may need to be eliminated from all states; in addition, all actions may need to be eliminated from recurrent states (some optimal actions may only be used at transient states and thus also need to be eliminated from recurrent states). Overall, convergence can thus be obtained after at most a total of $m\lceil n\tau_r \log(n^2\tau_r)\rceil + (m-n)\lceil n\tau_r \log(n^2\tau_t)\rceil$ recurrent class creations. The result follows from the fact that each class creation requires at most $(m-n)\lceil n^2\tau_t \log(n^2\tau_t)\rceil + n\lceil n^2\tau_t \log(n^2)\rceil$ iterations (cf. Lemma 6).

10 Cycle and recurrent classes creations for Howard's PI (Proofs of Lemmas 8 and 9)

Lemma 8. *If the MDP is deterministic, after at most n iterations, either Howard's PI finishes or a new cycle appears.*

Proof. Consider a sequence of l generated policies $\pi_1, \cdots, \pi_l$ from an initial policy π_0 such that no new cycle appears. By induction, we have

$$\begin{aligned}
v_{\pi_l} - v_{\pi_k} &= T_{\pi_l} v_{\pi_l} - T_{\pi_l} v_{\pi_{k-1}} + T_{\pi_l} v_{\pi_{k-1}} - T_{\pi_k} v_{\pi_{k-1}} + T_{\pi_k} v_{\pi_{k-1}} - T_{\pi_k} v_{\pi_k} && \{\forall \pi,\ T_\pi v_\pi = v_\pi\} \\
&\leq \gamma P_{\pi_l}(v_{\pi_l} - v_{\pi_{k-1}}) + \gamma P_{\pi_k}(v_{\pi_{k-1}} - v_{\pi_k}) && \{T_{\pi_l} v_{\pi_{k-1}} \leq T_{\pi_k} v_{\pi_{k-1}}\} \\
&\leq \gamma P_{\pi_l}(v_{\pi_l} - v_{\pi_{k-1}}) && \{\text{Lemma 1 and } P_{\pi_k} \geq 0\} \\
&\leq (\gamma P_{\pi_l})^k (v_{\pi_l} - v_{\pi_0}). && \{\text{By induction on } k\}
\end{aligned} \tag{22}$$

Since the MDP is deterministic and has n states, $(P_{\pi_l})^n$ will only have non-zero values on columns that correspond to $\mathcal{R}(\pi_l)$. Furthermore, since no cycle is created, $\mathcal{R}(\pi_l) \subset \mathcal{R}(\pi_0)$, which implies that $v_{\pi_l}(s) - v_{\pi_0}(s) = 0$ for all $s \in \mathcal{R}(\pi_l)$. As a consequence, we have $(P_{\pi_l})^n (v_{\pi_l} - v_{\pi_0}) = 0$. By Equation (22), this implies that $v_{\pi_l} = v_{\pi_n}$. If $l > n$, then Howard's PI must have terminated. □

Lemma 9. *After at most* $(m-n)\lceil n^2\tau_t \log(n^2\tau_t)\rceil + n\lceil n^2\tau_t \log(n^2)\rceil$ *iterations, either Howard's PI finishes or a new recurrent class appears.*

Proof. A close examination of the proof of Lemma 6, originally designed for Simplex-PI, shows that it applies to Howard's PI without any modification. □

11 A bound for Howard's PI and Simplex-PI under Assumption 1 (Proof of Theorem 7)

We here consider that the state space is decomposed into 2 sets: $\mathcal{T}$ is the set of states that are transient under all policies, and $\mathcal{R}$ is the set of states that are recurrent under all policies. From this assumption, it can be seen that when running Howard's PI or Simplex-PI, the values and actions chosen on $\mathcal{T}$ have no influence on the evolution of the values and policies on $\mathcal{R}$. So we will study the convergence of both algorithms in two steps: we will first bound the number of iterations to converge on $\mathcal{R}$; we will then add the number of iterations for converging on $\mathcal{T}$ given that convergence has occurred on $\mathcal{R}$.

Convergence on the set $\mathcal{R}$ of recurrent states: Without loss of generality, we consider here that the state space is only made of the set of recurrent states.

First consider Simplex-PI. If all states are recurrent, new recurrent classes are created at every iteration, and Lemma 7 holds. Then, in a way similar to the proof of Lemma 14, it can be shown that every $\lceil n\tau_r \log n^2\tau_r \rceil$ iterations, a non-optimal action can be eliminated. As there are at most $(m-n)$

non-optimal actions, we deduce that Simplex-PI converges in at most $(m-n)\lceil n\tau_r \log n^2\tau_r\rceil$ iterations on $\mathcal{R}$.

Consider now Howard's PI. We can prove the following lemma.

Lemma 10. *If the MDP satisfies Assumption 1 and all states are recurrent under all policies, Howard's PI generates policies $(\pi_k)_{k\geq 0}$ that satisfy:*

$$\mathbb{1}^T(v_{\pi_*} - v_{\pi_{k+1}}) \leq \left(1 - \frac{1}{n\tau_r}\right)\mathbb{1}^T(v_{\pi_*} - v_{\pi_k}).$$

Proof. On the one hand, we have

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_{k+1}} - v_{\pi_k}) &= {x_{\pi_{k+1}}}^T a_{\pi_k}^{\pi_{k+1}} && \{\text{Equation (11)}\}\\
&= {x_{\pi_{k+1}}}^T a_{\pi_k} && \{a_{\pi_k}^{\pi_{k+1}} = a_{\pi_k}\}\\
&\geq \frac{1}{(1-\gamma)\tau_r}\mathbb{1}^T a_{\pi_k} && \{\text{Equation (9) and all states are recurrent}\}\\
&\geq \frac{1}{(1-\gamma)\tau_r}\|a_{\pi_k}\|_\infty. && \{\forall x \geq 0, \mathbb{1}^T x \geq \|x\|_\infty\} \qquad (23)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_*} - v_{\pi_k}) &= {x_{\pi_*}}^T a_{\pi_k}^{\pi_*} && \{\text{Equation (11)}\}\\
&\leq {x_{\pi_*}}^T a_{\pi_k} && \{a_{\pi_k} \geq a_{\pi_k}^{\pi_*}\}\\
&\leq \frac{n}{1-\gamma}\|a_{\pi_k}\|_\infty. && \{\sum_i x_{\pi_*}(i) \leq \frac{n}{1-\gamma} \text{ and } a_{\pi_k} \geq 0\} \qquad (24)
\end{aligned}$$

By combining Equations (23) and (24), we obtain:

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_*} - v_{\pi_{k+1}}) &= \mathbb{1}^T(v_{\pi_*} - v_{\pi_k}) - \mathbb{1}^T(v_{\pi_{k+1}} - v_{\pi_k})\\
&\leq \left(1 - \frac{1}{n\tau_r}\right)\mathbb{1}^T(v_{\pi_*} - v_{\pi_k}).
\end{aligned}$$

□

Then, similarly to Simplex-PI, we can prove that after every $\lceil n\tau_r \log n^2\tau_r\rceil$ iterations a non-optimal action must be eliminated. And as there are at most $(m-n)$ non-optimal actions, we deduce that Howard's PI converges in at most $(m-n)\lceil n\tau_r \log n^2\tau_r\rceil$ iterations on $\mathcal{R}$.

Convergence on the set $\mathcal{T}$ of transient states: Consider now that convergence has occurred on the recurrent states $\mathcal{R}$. A simple variation of the proof of Lemma 6/Lemma 9 (where we use the fact that we don't need to consider the events where recurrent classes are broken since recurrent classes do not evolve anymore) allows us to show that the extra number of iterations for both algorithms to converge on the transient states is at most $(m-n))\lceil n^2\tau_t \log n^2\tau_t\rceil$, and the result follows.

Acknowledgements.

I would like to thank Ian Post for exchanges about the proof in Post and Ye (2013), Thomas Dueholm Hansen for noticing a flaw in a claimed result for deterministic MDPs in an earlier version, Romain Azaïs for the reference on the Cesaro mean of stochastic matrices, and the reviewers and editor for their very careful feedback, who helped improve the paper overall, and the proof of Lemma 13 in particular.

References

Akian, M. and Gaubert, S. (2013). Policy iteration for perfect information stochastic mean payoff games with bounded first return times is strongly polynomial. Technical Report arxiv 1310.4953v1.

Bertsekas, D. and Tsitsiklis, J. (1996). *Neurodynamic Programming*. Athena Scientific.

Fearnley, J. (2010). Exponential lower bounds for policy iteration. In *Proceedings of the 37th international colloquium conference on Automata, languages and programming: Part II*, ICALP'10, pages 551–562, Berlin, Heidelberg. Springer-Verlag.

Fritz, F., Huppert, B., and Willems, W. (1979). *Stochastische Matrizen*. Springer, Berlin.

Hansen, T. (2012). *Worst-case Analysis of Strategy Iteration and the Simplex Method*. Ph.D. thesis, Department Office Computer Science, Aarhus University.

Hansen, T. and Zwick, U. (2010). Lower bounds for Howard's algorithm for finding minimum mean-cost cycles. In *ISAAC (1)*, pages 415–426.

Hansen, T., Miltersen, P., and Zwick, U. (2013). Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *J. ACM*, **60**(1), 1–16.

Hollanders, R., Delvenne, J., and Jungers, R. (2012). The complexity of policy iteration is exponential for discounted markov decision processes. In *51st IEEE conference on Decision and control (CDC'12)*.

Hollanders, R., Gerencsér, B., Delvenne, J., and Jungers, R. (2014). Improved bound on the worst case complexity of policy iteration. Technical Report arxiv 1410.7583v1.

Mansour, Y. and Singh, S. (1999). On the complexity of policy iteration. In *UAI*, pages 401–408.

Melekopoglou, M. and Condon, A. (1994). On the complexity of the policy improvement algorithm for Markov decision processes. *INFORMS Journal on Computing*, **6**(2), 188–192.

Post, I. and Ye, Y. (2013). The simplex method is strongly polynomial for deterministic Markov decision processes. In *24th ACM-SIAM Symposium on Discrete Algorithms*.

Puterman, M. (1994). *Markov Decision Processes*. Wiley, New York.

Schmitz, N. (1985). How good is Howard's policy improvement algorithm? *Zeitschrift für Operations Research*, **29**(7), 315–316.

Stroock, D. (2005). *An introduction to Markov processes*. Springer, Berlin.

Ye, Y. (2011). The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate. *Math. Oper. Res.*, **36**(4), 593–603.