HAL
open science

Improved and Generalized Upper Bounds on the Complexity of Policy Iteration

Bruno Scherrer

To cite this version:

Bruno Scherrer. Improved and Generalized Upper Bounds on the Complexity of Policy Iteration. 2013. hal-00829532v1

HAL Id: hal-00829532

https://inria.hal.science/hal-00829532v1

Preprint submitted on 3 Jun 2013 (v1), last revised 10 Feb 2016 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Improved and Generalized Upper Bounds on the Complexity of Policy Iteration

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

June 3, 2013

Abstract

Given a Markov Decision Process (MDP) with n states and m actions per state, we study the number of iterations needed by Policy Iteration (PI) algorithms to converge. We consider two variations of PI: Howard's PI that changes all the actions with a positive advantage, and Simplex-PI that only changes one action with maximal advantage. We show that Howard's PI terminates after at most $n(m-1)\left\lceil \frac{1}{1-\gamma}\log\left(\frac{1}{1-\gamma}\right)\right\rceil$ iterations, improving by a factor $O(\log n)$ a result by Hansen *et al.* (2013), while Simplex-PI terminates after at most $n(m-1)\left\lceil \frac{n}{1-\gamma}\log\left(\frac{n}{1-\gamma}\right)\right\rceil$ iterations, improving by a factor 2 a result by Ye (2011). We then consider bounds that are independent of the discount factor γ. When the MDP is deterministic, we show that Simplex-PI terminates after at most $2n^2m(m-1)\lceil 2(n-1)\log n\rceil\lceil 2n\log n\rceil = O(n^4m^2\log^2 n)$ iterations, improving by a factor $O(n)$ a bound obtained by Post and Ye (2012). We generalize this result to general MDPs under some structural assumptions: given a measure of the maximal transient time τ_t and the maximal time τ_r to revisit states in recurrent classes under all policies, we show that Simplex-PI terminates after at most $n^2m(m-1)\left(\lceil\tau_r\log(n\tau_r)\rceil + \lceil\tau_r\log(n\tau_t)\rceil\right)\lceil\tau_t\log(n(\tau_t+1))\rceil = \tilde{O}(n^2\tau_t\tau_r m^2)$ iterations. We explain why similar results seem hard to derive for Howard's PI. Finally, under the additional (restrictive) assumption that the MDP is weakly-communicating, we show that Simplex-PI and Howard's PI terminate after at most $n(m-1)\left(\lceil\tau_t\log n\tau_t\rceil + \lceil\tau_r\log n\tau_r\rceil\right) = \tilde{O}(nm(\tau_t+\tau_r))$ iterations.

1 Introduction

We consider a discrete-time dynamic system whose state transition depends on a control. We assume that there is a **state space** X of finite size n. When at state $i \in \{1,..,n\}$, the control is chosen from a finite **control space** A of finite size[1] m. The control $a \in A$ 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 $r(i,a,j)$ where r is the instantaneous **reward function**. In this context, we look for a stationary deterministic policy (a function $\pi : X \to A$ that maps states into controls[2]) 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 $\langle X, A, p, r, \gamma\rangle$ is called a **Markov Decision Process (MDP)** (Puterman, 1994; Bertsekas and Tsitsiklis, 1996).

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

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

[1] In the works of Ye (2011); Post and Ye (2012); Hansen *et al.* (2013) that we reference, the integer "m" denotes the total number of actions, that is nm with our notation. When we restate their result, we do it with our own notation, that is we replace their $''m''$ by $''nm''$.

[2] 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 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 a 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 $T_\pi v := r_\pi + \gamma P_\pi v$. 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 componentwise. In other words, v_* is a fixed point of the nonlinear operator $Tv := \max_\pi T_\pi v$. 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 $T_\pi v = Tv$. 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 is an **optimal policy** and its value v_{π_*} is equal to v_*.

Let π be some policy. For any state i, we call the **advantage in state** i **with respect to** π the quantity:

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

We call the **set of switchable states of** π the following set

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

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 on 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 specific variations:

- When for all iteration k, $Y_k = S_{\pi_k}$, that is one changes all actions that have a positive advantage, 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 k, Y_k is a singleton containing a state $i_k \in \arg\max_i a_{\pi_k}(i)$, that is if we only change one action with a maximal advantage with respect to π_k, we will call it Simplex-PI[3].

Since it generates a sequence of policies with increasing values, and that the number of policies is m^n, any implementation of PI converges to the optimal value in a finite number of iterations smaller than m^n. In practice, PI converges in very few iterations. On random MDP instances, convergen often occur in time sublinear 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.

[3]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)

2 Results

In this section, we describe some known results—see (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 Appendix.

A key observtation for both algorithms, that will be central to the results we are about to discuss, is that the sequence they generate satisfies some contraction property[4]. For any vector u, let $\|u\|_\infty = max_{x \in X}|u(x)|$ be the max-norm of u. Let $\mathbb{1}$ be the vector of which all components are equal to 1.

Lemma 2 (Proof in Section 3)**.** *The sequence* $(\|v_* - v_{\pi_k}\|_\infty)_{k \geq 0}$ *built by Howard's PI is contracting with coefficient* γ*.*

Lemma 3 (Proof in Section 4)**.** *The sequence* $(\mathbb{1}^T(v_* - v_{\pi_k}))_{k \geq 0}$ *built by Simplex-PI is contracting with coefficient* $1 - \frac{1-\gamma}{n}$*.*

Though this observation is widely known for Howard's PI, it was to our knowledge never mentionned in the literature for Simplex-PI. These contraction properties have the following immediate consequence[5].

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* $n(m-1)\left\lceil \frac{n}{1-\gamma} \log\left(\frac{n^2}{1-\gamma}\right)\right\rceil$ *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, Hansen *et al.* (2013) recently improved the result by a factor $O(n)$ for Howard's PI.

Theorem 2 (Hansen *et al.* (2013))**.** *Howard's PI terminates after at most* $(nm+1)\left\lceil \frac{1}{1-\gamma} \log\left(\frac{n}{1-\gamma}\right)\right\rceil$ *iterations.*

Our first two results are stated in the following theorems.

Theorem 3 (Proof in Section 5)**.** *Howard's PI terminates after at most* $n(m-1)\left\lceil \frac{1}{1-\gamma} \log\left(\frac{1}{1-\gamma}\right)\right\rceil$ *iterations.*

Theorem 4 (Proof in Section 6)**.** *Simplex-PI terminates after at most* $n(m-1)\left\lceil \frac{n}{1-\gamma} \log\left(\frac{n}{1-\gamma}\right)\right\rceil$ *iterations.*

Our result for Howard's PI is a factor $O(\log n)$ better than the previous best result of Hansen *et al.* (2013). Our result for Simplex-PI is only very slightly better (by a factor 2) than that of Ye (2011), and uses a proof that is more direct. Compared to Howard's PI, the number of iterations of Simplex-PI is a factor $\tilde{O}(n)$ larger. However, since one changes only one action per iteration, each iteration may have a complexity 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 requires to compute the value of some policy, may require $O(n^3)$ time. Overall, both algorithms seem to have a similar complexity.

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

[5] 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 $\|v_* - v_{\pi_k}\|_\infty \leq \|v_* - v_{\pi_k}\|_1 \leq \gamma^k \|v_* - v_{\pi_0}\|_1 \leq \gamma^k n V_{\max}$, and the conclusion is similar to that for Howard's PI.

It is easy to see that the linear dependency of the bound for Howard's PI with respect to n is optimal. We conjecture that the linear dependency of both bounds with respect to m is also optimal. 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 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 $\gamma = 1$.

In the rest of this section, we will describe some bounds that do not depend on γ but that will be based on some structural assumptions of the MDPs. On this topic, Post and Ye (2012) recently showed the following result for deterministic MDPs.

Theorem 5 (Post and Ye (2012))**.** *If the MDP is deterministic, then Simplex-PI terminates after at most $O(n^5 m^2 \log^2 n)$ iterations.*

Once again, the arguments are based on the simplex algorithm applied to the the linear programming formulation of the MDP. We were able to improve this bound by a factor $O(n)$.

Theorem 6 (Proof in Section 7)**.** *If the MDP is deterministic, then Simplex-PI terminates after at most $2n^2 m(m-1)\lceil 2n \log n \rceil \lceil 2(n-1)\log n \rceil = O(n^4 m^2 \log^2 n)$ iterations.*

The proof follows the general structure of that of Post and Ye (2012), but is more direct (it does not involve linear programming arguments). Given a policy π of a deterministic MDP, states are either on cycles or on path induced by π. The core of the proofs relies on the following lemmas that altogether show that cycles are created regularly and that significant progress is made every time a cycle is created; in other words, significant progress is made regularly.

Lemma 4. *If the MDP is deterministic, after at most $nm\lceil 2(n-1)\log n \rceil$ iterations, either Simplex-PI finishes or a new cycle is created.*

Lemma 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 $O(n^3 m^2 \log \frac{n}{1-\gamma}) = \tilde{O}(n^3 m^2)$. Removing completely 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 Section 7 and very similar to the one initially done by Post and Ye (2012), which incurs an extra term of order $O(n \log(n))$. With respect to the result by Post and Ye (2012), the eventual $O(n)$ improvement of Theorem 6 lies in the fact that Lemma 4 is a factor $O(n)$ better than the equivalent result (Lemma 3.3) in (Post and Ye, 2012).

At a more technical level, our proof and that of Post and Ye (2012) critically rely on some properties of 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:

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

where U denotes the uniform law on the state space X. For any policy π and state i, we trivially have $x_\pi(i) \in \left(1, \frac{n}{1-\gamma}\right)$. Both proofs exploit the fact that $\sum_i$ on a path of $_\pi\, x_\pi(i)$ belongs to the set $(1, n-1)$, 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 rewriting and improvement of the proof of (Post and Ye, 2012) is done in a way that a generalization to stochastic MDPs is made straightforward. Given a policy π of a stochastic MDP, states are either *in recurrent subchains* or *transient* (these two categories respectively generalize those of cycles and paths). We will consider the following structural assumption.

[6] This can be done by using arguments similar to the proof of Theorem 4.

Assumption 1. *For all policies π, let $\tau_t \geq 1$ and $\tau_r \geq 1$ be the smallest constants such that for all states i that are transient for π,*

$$\sum_{i \text{ transient for } \pi} x_\pi(i) \leq \tau_t,$$

and for all states i that are recurrent for π,

$$\frac{n}{(1-\gamma)\tau_r} \leq x_\pi(i) \left(\leq \frac{n}{1-\gamma} \right).$$

The constant τ_t (resp. τ_r) can be seen as a measure of the time needed to leave transient states (resp. the time needed to revisit states in recurrent classes). In particular, when γ tends to 1, it can be seen that τ_t is related to the expectation of $\mathcal{L}$, the (random) time needed to leave the set of transient states, since for any policy π,

$$\begin{aligned}\lim_{\gamma \to 1} \sum_{i \text{ transient for } \pi} x_\pi(i) &= n \sum_{t=0}^{\infty} \mathbb{P}(i_t \text{ transient for } \pi \mid i_0 \sim U,\ a_t = \pi(i_t)) \\ &= n\mathbb{E}\left[\, \mathcal{L} \mid i_0 \sim U,\ a_t = \pi(i_t)\right].\end{aligned}$$

Similarly, when γ is in the vicinity of 1, $\frac{1}{\tau_r}$ is the minimal asymptotic frequency[7] in recurrent states given that one starts from a random uniform state, since for any policy π and recurrent state i:

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

With Assumption 1 in hand, we can generalize Lemmas 4-5 as follows.

Lemma 6. *If the MDP satisfies Assumption 1, after at most $n\left[(m-1)\lceil \tau_t \log(n(\tau_t+1))\rceil + \lceil \tau_t \log(n\tau_r)\rceil\right]$ iterations, either Simplex-PI finishes or a new recurrent subchain is created.*

Lemma 7. *If the MDP satisfies Assumption 1, when Simplex-PI moves from π to π' where π' involves a new recurrent subchain, we have*

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

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

Theorem 7 (Proof in appdx 8)**.** *If the MDP satisfies Assumption 1, then Simplex-PI terminates after at most*

$$n^2 m(m-1)\left(\lceil \tau_r \log(n\tau_r)\rceil + \lceil \tau_r \log(n\tau_t)\rceil\right)\lceil \tau_t \log(n(\tau_t+1))\rceil = \tilde{O}\left(n^2 \tau_t \tau_r m^2\right)$$

iterations.

Remark 1. *This new result is a strict generalization of the result for deterministic MDPs. Indeed, in the deterministic case, we have $\tau_t = n-1$ and $\tau_r = n$, and it is is easy to see that Lemmas 6, 7 and Theorem 7* exactly *reduce to Lemmas 4, 5 and Theorem 6.*

An immediate consequence of the above result is Simplex-PI is stronly 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.*

[7] If the MDP is aperiodic, and thus admits a stationary distribution ν_π for any policy π, one can see that

$$\frac{1}{\tau_r} = n \min_{\pi,\ i \text{ recurrent for } \pi} \nu_\pi(i).$$

One may then wonder whether a similar result can be derived for Howard's PI. Unfortunately, and as quickly mentionned by Post and Ye (2012), the line of analysis developped for Simplex-PI does not seem to adapt easily to Howard's PI, because simultaneously switching several actions can interfere in a way 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 Lemma 4 and 6 for Howard's PI (see Section 9).

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

Lemma 9. *If the MDP satisfies Assumption 1, after at most $nm\lceil \tau_t \log n\tau_t \rceil$ iterations, either Howard's PI finishes or a new recurrent subchain is created.*

However, 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 $O(n^2)$ 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,

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

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 a long time. It was originally identified as an open problem by Schmitz (1985). In the simplest—deterministic—case, the question is still open: the currently best known lower bound is the $O(n^2)$ bound by Hansen and Zwick (2010) we have just mentionned, while the best known upper bound is $O(\frac{m^n}{n})$ due to Mansour and Singh (1999).

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 2 (Weakly communicating MDP)**.** *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.*

Indeed, under this assumption, we can prove for Howard's PI a variation of Lemma 7 introduced for Simplex-PI.

Lemma 10. *For an MDP satisfying Assumptions 1-2, suppose Howard's PI moves from π to π' and that π' involves a new recurrent subchain. Then*

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

And we can deduce the following original bound (that also applies to Simplex-PI).

Theorem 8 (Proof in Section 10)**.** *If the MDP satisfies Assumptions 1-2, then Simplex-PI and Howard's PI terminate after at most $n(m-1)\left(\lceil \tau_t \log n\tau_t \rceil + \lceil \tau_r \log n\tau_r \rceil\right)$ iterations.*

It should however be noted that Assumption 2 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 subchain is created every single iteration (this is in contrast with Lemmas 4, 6, 8 and 9 that were

[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, $\frac{v_*(i) - v_{\pi_{k+1}}(i)}{v_*(i) - v_{\pi_k}(i)} \geq \frac{1+p^{-k-2}}{1+p^{-k}} = 1 - \frac{p^{-k} - p^{-k-2}}{1+p^{-k}} \geq 1 - p^{-k}(1-p^{-2}) \geq 1 - p^{-k}$.

designed to show under which conditions cycles and recurrent subchains 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, if this last result sheds some light on the practical efficiency of Howard's PI and Simplex-PI, the problem of understanding better the behavior of Howard's PI is still largely open, and constitutes our main future work.

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

For any k, using the notation "$A \succ 0$" for "A is positive definite", we have

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

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

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

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

We begin by proving a useful identity.

Lemma 11. *For all pairs of policies π and π',*

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

Proof. We have:

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

□

On the one hand, by using this lemma, we have for any k:

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

which implies that

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

On the other hand, we have:

$$\begin{aligned}
v_{\pi_*} - v_{\pi_k} &= (I - \gamma P_{\pi_*})^{-1}(T_{\pi_*} v_{\pi_k} - v_{\pi_k}) && \{\text{Lemma 11}\} \\
&\leq \frac{1}{1-\gamma} \max_s T_{\pi_*} v_{\pi_k}(s) - v_{\pi_k}(s) && \{\|(I - \gamma P_{\pi_*})^{-1}\|_\infty = \frac{1}{1-\gamma} \text{ and } (I - \gamma P_{\pi_*})^{-1} \succ 0\} \\
&\leq \frac{1}{1-\gamma} \max_s T_{\pi_{k+1}} v_{\pi_k}(s) - v_{\pi_k}(s) && \{\max_s T_{\pi_{k+1}} v_{\pi_k}(s) = \max_{s,\tilde{\pi}} T_{\tilde{\pi}} v_{\pi_k}(s)\} \\
&\leq \frac{1}{1-\gamma} \mathbb{1}^T(T_{\pi_{k+1}} v_{\pi_k} - v_{\pi_k}), && \{\forall x \geq 0,\ \max_s x(s) \leq 1^T x\}
\end{aligned}$$

which implies that

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

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

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

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

Though the overall line or arguments follows 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:

$$\begin{aligned}
v_* - T_{\pi_k}v_* &= (I - \gamma P_{\pi_k})(v_* - v_{\pi_k}) && \{\text{Lemma 11}\} \\
&\leq v_* - v_{\pi_k}. && \{v_* - v_{\pi_k} \geq 0 \text{ and } P_{\pi_k} \succ 0\}
\end{aligned}$$

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

$$\begin{aligned}
\|v_* - T_{\pi_k}v_*\|_\infty &\leq \|v_* - v_{\pi_k}\|_\infty \\
&\leq \gamma^k\|v_{\pi_*} - v_{\pi_0}\|_\infty && \{\text{Lemma 2}\} \\
&= \gamma^k\|(I - \gamma P_{\pi_0})^{-1}(v_* - T_{\pi_0}v_*)\|_\infty && \{\text{Lemma 11}\} \\
&\leq \frac{\gamma^k}{1-\gamma}\|v_* - T_{\pi_0}v_*\|_\infty. && \{\|(I - \gamma P_{\pi_0})^{-1}\|_\infty = \frac{1}{1-\gamma}\}
\end{aligned}$$

By definition of the max-norm, there exists a state s_0 such that $v_*(s_0) - [T_{\pi_0}v_*](s_0) = \|v_* - T_{\pi_0}v_*\|_\infty$. We deduce that for all k,

$$\begin{aligned}
v_*(s_0) - [T_{\pi_k}v_*](s_0) &\leq \|v_* - T_{\pi_k}v_*\|_\infty \\
&\leq \frac{\gamma^k}{1-\gamma}\|v_* - T_{\pi_0}v_*\|_\infty \\
&= \frac{\gamma^k}{1-\gamma}(v_*(s_0) - [T_{\pi_0}v_*](s_0))
\end{aligned}$$

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 (they are at most $n(m-1)$), and the result follows.

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

The overall line or arguments follows those given originally by Ye (2011), and is similar to that of the previous section. Still the result we get is slightly better. For any k, we have:

$$\begin{aligned}
\|v_{\pi_*} - T_{\pi_k} v_{\pi_*}\|_\infty &\leq \|v_{\pi_*} - v_{\pi_k}\|_\infty && \{\text{Lemma 11 and } P_{\pi_k} \succ 0\} \\
&\leq \mathbb{1}^T (v_{\pi_*} - v_{\pi_k}) && \{\forall x \geq 0,\ \|x\|_\infty \leq \mathbb{1}^T x\} \\
&\leq \left(1 - \frac{1-\gamma}{n}\right)^k \mathbb{1}^T (v_{\pi_*} - v_{\pi_0}) && \{\text{Lemma 3}\} \\
&\leq n \left(1 - \frac{1-\gamma}{n}\right)^k \|v_{\pi_*} - v_{\pi_0}\|_\infty && \{\forall x,\ \mathbb{1}^T x \leq n\|x\|_\infty\} \\
&= n \left(1 - \frac{1-\gamma}{n}\right)^k \|(I - \gamma P_{\pi_0})^{-1}(v_* - T_{\pi_0} v_*)\|_\infty && \{\text{Lemma 11}\} \\
&\leq \frac{n}{1-\gamma} \left(1 - \frac{1-\gamma}{n}\right)^k \|v_{\pi_*} - T_{\pi_0} v_{\pi_*}\|_\infty, && \{\|(I - \gamma P_{\pi_0})^{-1}\|_\infty = \frac{1}{1-\gamma}\}
\end{aligned}$$

Similarly to the proof for Howard's PI, we deduce that a non-optimal action is eliminated after at most

$$k^* = \left\lceil \frac{n}{1-\gamma} \log \frac{n}{1-\gamma} \right\rceil \geq \left\lceil \frac{\log \frac{n}{1-\gamma}}{\log\left(1 - \frac{1-\gamma}{n}\right)} \right\rceil,$$

and the overall number of iterations is obtained by noting that there are at most $n(m-1)$ non optimal actions to eliminate.

7 A bound for Simplex-PI for deterministic MDPs (Proof of Theorem 6)

The proof we give here is strongly inspired by that of Post and Ye (2012): the steps (a series of lemmas) are very similar, one of them being better (Lemma 12), which leads to an eventual improvement of a factor $O(n)$ and – more importantly – our proofs are more direct (they do not involve linear programming arguments).

For any policy π, write $\mathcal{C}(\pi)$ the set of cycles induced by policy π. Recall that $x_\pi = (I - \gamma {P_\pi}^T)^{-1}\mathbb{1}$. A useful corollary of Lemma 11 is that for any pair of policies π and π',

$$\mathbb{1}^T (v_{\pi'} - v_\pi) = {x_{\pi'}}^T (T_{\pi'} v_\pi - v_\pi). \tag{3}$$

We will write that $s \in \mathcal{C}(\pi)$ if there exists a cycle $C \in \mathcal{C}(\pi)$ that contains s. We will repeatedly exploit the (obvious) facts that

$$\forall s \in \mathcal{C}(\pi),\ \frac{1}{1-\gamma} \leq x_\pi(s) \leq \frac{n}{1-\gamma}, \tag{4}$$

$$\sum_{s \notin \mathcal{C}(\pi)} x_\pi(s) \leq n - 1, \tag{5}$$

$$\text{which implies that: } \forall s \notin \mathcal{C}(\pi),\ x_\pi(s) \leq n - 1. \tag{6}$$

7.1 Part 1: cycles are created rather often

Lemma 12. *Suppose one moves from policy π to policy π'* without creating any cycle. *Let $\pi_\dagger$ be the final policy before either a new cycle is created or the algorithm terminates. Then*

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

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

$$\mathbb{1}^T(v_{\pi'} - v_\pi) \geq \mathbb{1}^T(T_{\pi'}v_\pi - v_\pi). \tag{7}$$

On the other hand, we have

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_\dagger} - v_\pi) &= \mathbb{1}^T(I - \gamma P_{\pi_\dagger})^{-1}(T_{\pi_\dagger}v_\pi - v_\pi) \\
&= x_{\pi_\dagger}^T(T_{\pi_\dagger}v_\pi - v_\pi) \\
&= \sum_s x_{\pi_\dagger}(s)(T_{\pi_\dagger}v_\pi(s) - v_\pi(s)) \\
&\leq (n-1)\max_{s\notin\mathcal{C}(\pi_\dagger)} T_{\pi_\dagger}v_\pi(s) - v_\pi(s) + \frac{n}{1-\gamma}\max_{s\in\mathcal{C}(\pi_\dagger)} T_{\pi_\dagger}v_\pi(s) - v_\pi(s). \quad \{\text{Equations (4) and (6)}\}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_\dagger} - v_\pi) &\leq (n-1)\max_s T_{\pi_\dagger}v_\pi(s) - v_\pi(s) \\
&\leq (n-1)\max_s T_{\pi'}v_\pi(s) - v_\pi(s) \qquad \{\max_s T_{\pi'}v_\pi(s) = \max_{s,\tilde{\pi}} T_{\tilde{\pi}}v_\pi(s)\} \\
&= (n-1)\mathbb{1}^T(T_{\pi'}v_\pi - v_\pi).
\end{aligned} \tag{8}$$

Combining Equations (7) and (8), we get:

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

□

Lemma 13. *After at most $\lceil 2(n-1)\log n\rceil$ iterations, either the algorithm finishes, a new cycle is created, a cycle is broken, or some action never appears in a policy again before a new cycle is created.*

Proof. Let π be the policy in some iteration, $\pi_\dagger$ be the last policy before a new cycle is created, and π' any policy between π and $\pi_\dagger$. Since

$$\begin{aligned}
0 &\leq 1^T(v_{\pi_\dagger} - v_\pi) && \{v_{\pi_\dagger} \geq v_\pi\} \\
&= {x_\pi}^T(v_{\pi_\dagger} - T_\pi v_{\pi_\dagger}) && \{\text{Equation (3)}\} \\
&= \sum_{s\notin\mathcal{C}(\pi)} x_\pi(s)(v_{\pi_\dagger}(s) - T_\pi v_{\pi_\dagger}(s)) + \sum_{C\in\mathcal{C}(\pi)}\sum_{s\in C} x_\pi(s)(v_{\pi_\dagger}(s) - T_\pi v_{\pi_\dagger}(s)),
\end{aligned}$$

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

$$x_\pi(s_0)(v_{\pi_\dagger}(s_0) - T_\pi v_{\pi_\dagger}(s_0)) \geq \frac{1}{n}{x_\pi}^T(v_{\pi_\dagger} - T_\pi v_{\pi_\dagger}) \geq 0. \tag{9}$$

or a cycle C_0 such that

$$\sum_{s\in C_0} x_\pi(s)(v_{\pi_\dagger}(s) - T_\pi v_{\pi_\dagger}(s)) \geq \frac{1}{n}{x_\pi}^T(v_{\pi_\dagger} - T_\pi v_{\pi_\dagger}) \geq 0. \tag{10}$$

We consider these two cases separately below.

- **case 1:** Equation (9) holds for some $s_0 \notin \mathcal{C}(\pi)$.
  If $\pi'(s_0) = \pi(s_0)$, then

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_\dagger} - v_{\pi'}) &\geq v_{\pi_\dagger}(s_0) - v_{\pi'}(s_0) && \{v_{\pi_\dagger} \geq v_{\pi'}\} \\
&= v_{\pi_\dagger}(s_0) - T_{\pi'}v_{\pi'}(s_0) && \{v_{\pi'} = T_{\pi'}v_{\pi'}\} \\
&\geq v_{\pi_\dagger}(s_0) - T_{\pi'}v_{\pi_\dagger}(s_0) && \{v_{\pi_\dagger} \geq v_{\pi'}\} \\
&= v_{\pi_\dagger}(s_0) - T_{\pi}v_{\pi_\dagger}(s_0) && \{\pi(s_0) = \pi'(s_0)\} \\
&\geq \frac{1}{n}x_\pi(s_0)(v_{\pi_\dagger}(s_0) - T_\pi v_{\pi_\dagger}(s_0)) && \{\text{Equation (6)}\} \\
&\geq \frac{1}{n^2}{x_\pi}^T(v_{\pi_\dagger} - T_\pi v_{\pi_\dagger}) && \{\text{Equation (9)}\} \\
&= \frac{1}{n^2}\mathbb{1}^T(v_{\pi_\dagger} - v_\pi). && \{\text{Equation (3)}\}
\end{aligned}$$

- **case 2:** Equation (10) holds for some $C_0 \in \mathcal{C}(\pi)$.
  Let $\mathcal{P}$ be the set of states that are in a path of π (formally, $\mathcal{P} = X \backslash \mathcal{C}(\pi)$). For any subset Y of the state space X, write P_π^Y for the stochastic matrix of which the ith 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 ith component is equal to 1 if $i \in Y$ and 0 otherwise. Using the fact that $P_\pi^{C_0}P_\pi^{\mathcal{P}} = 0$, one can first observe that

$$(I - \gamma P_\pi^{c_0})(I - \gamma P_\pi^{\mathcal{P}}) = I - \gamma(P_\pi^{C_0} + P_\pi^{\mathcal{P}}),$$

  from which we can deduce that

$$\begin{aligned}
{\mathbb{1}_{\mathcal{P}\cup C_0}}^T(I - \gamma P)^{-1} &= {\mathbb{1}_{\mathcal{P}\cup C_0}}^T(I - \gamma(P_\pi^{C_0} + P_\pi^{\mathcal{P}}))^{-1} \\
&= {\mathbb{1}_{\mathcal{P}\cup C_0}}^T(I - \gamma P_\pi^{\mathcal{P}})^{-1}(I - \gamma P_\pi^{c_0})^{-1}. \qquad (11)
\end{aligned}$$

  Also, writing $h_{\mathcal{P}} = (I - \gamma {P_\pi^{\mathcal{P}}}^T)^{-1}\mathbb{1}_{\mathcal{P}}$, that satisfies

$$h_{\mathcal{P}} = \mathbb{1}_{\mathcal{P}} + \gamma {P_\pi^{\mathcal{P}}}^T h_{\mathcal{P}},$$

  we can see that:

$$\forall s \in C_0,\ h_{\mathcal{P}}(s) = \gamma \sum_{s' \in \mathcal{P}} p_{s's}(\pi(s'))h_{\mathcal{P}}(s'), \qquad \{s \in C_0 \ \Rightarrow\ \mathbb{1}_{\mathcal{P}}(s) = 0\} \qquad (12)$$

  and thus:

$$\begin{aligned}
(I - \gamma {P_\pi^{\mathcal{P}}}^T)^{-1}\mathbb{1}_{\mathcal{P}\cup C_0}(s) &= (I - \gamma {P_\pi^{\mathcal{P}}}^T)^{-1}\mathbb{1}_{\mathcal{P}}(s) + 1 && \{{P_\pi^{\mathcal{P}}}^T\mathbb{1}_{C_0} = 0\} \\
&= h_{\mathcal{P}}(s) + 1 \\
&\leq \gamma \sum_{s' \in \mathcal{P}} p_{s's}(\pi(s'))h_{\mathcal{P}}(s') + 1 && \{\text{Equation (12)}\} \\
&\leq \sum_{s' \in \mathcal{P}} h_{\mathcal{P}}(s') + 1 \\
&= \sum_{s' \in \mathcal{P}} x_\pi(s') + 1 && \{\forall s' \in \mathcal{P},\ h_{\mathcal{P}}(s') = x_\pi(s')\} \\
&\leq n && \{\text{Equation (5)}\} \qquad (13)
\end{aligned}$$

Writing δ the vector that equals $v_{\pi_\dagger} - T_\pi v_{\pi_\dagger}$ on C_0 and that is null everywhere else, we have

$$\begin{aligned}
&\sum_{s\in C_0} x_\pi(s)(v_{\pi_\dagger}(s) - T_\pi v_{\pi_\dagger}(s)) \\
&= \sum_{s\in C_0} [(I-\gamma P_\pi^T)^{-1}\mathbb{1}](s)\delta(s) \\
&= \sum_{s\in C_0} [(I-\gamma P_\pi^T)^{-1}\mathbb{1}_{\mathcal{P}\cup C_0}](s)\delta(s) && \{\forall s\in C_0,\ \forall s'\notin \mathcal{P}\cup C_0,\ [(I-\gamma P_\pi^T)^{-1}\mathbb{1}_{s'}](s)=0\} \\
&= \sum_{s} [(I-\gamma P_\pi^T)^{-1}\mathbb{1}_{\mathcal{P}\cup C_0}](s)\delta(s) && \{\forall s\notin C_0,\ \delta(s)=0\} \\
&= {\mathbb{1}_{\mathcal{P}\cup C_0}}^T (I-\gamma P_\pi)^{-1}\delta \\
&= {\mathbb{1}_{\mathcal{P}\cup C_0}}^T (I-\gamma P_\pi^{\mathcal{P}})^{-1}(I-\gamma P_\pi^{c_0})^{-1}\delta && \{\text{Equation (11)}\} \\
&= \sum_{s} [(I-\gamma P_\pi^{\mathcal{P}^T})^{-1}\mathbb{1}_{\mathcal{P}\cup C_0}](s)[(I-\gamma P_\pi^{c_0})^{-1}\delta](s) \\
&= \sum_{s\in C_0} [(I-\gamma P_\pi^{\mathcal{P}^T})^{-1}\mathbb{1}_{\mathcal{P}\cup C_0}](s)[(I-\gamma P_\pi^{c_0})^{-1}\delta](s) && \{\forall s\notin C_0,\ \delta(s)=0\} \\
&= \sum_{s\in C_0} [(I-\gamma P_\pi^{\mathcal{P}^T})^{-1}\mathbb{1}_{\mathcal{P}\cup C_0}](s)(v_{\pi_\dagger}(s) - v_\pi(s)) && \{\text{Lemma 11}\} \\
&\leq n\mathbb{1}_{C_0}(v_{\pi_\dagger} - v_\pi). && \{\text{Equation (13)}\}
\end{aligned}$$

Now, one can deduce from this that if C_0 is also a cycle of π', which implies ${\mathbb{1}_{C_0}}^T v_\pi = {\mathbb{1}_{C_0}}^T v_{\pi'}$, then

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_\dagger} - v_{\pi'}) &\geq {\mathbb{1}_{C_0}}^T(v_{\pi_\dagger} - v_{\pi'}) && \{v_{\pi_\dagger} \geq v_{\pi'}\} \\
&= {\mathbb{1}_{C_0}}^T(v_{\pi_\dagger} - v_\pi) && \{{\mathbb{1}_{C_0}}^T v_\pi = {\mathbb{1}_{C_0}}^T v_{\pi'}\} \\
&\geq \frac{1}{n}\sum_{s\in C_0} x_\pi(s)(v_{\pi_\dagger}(s) - T_\pi v_{\pi_\dagger}(s)) && \{\text{Equation (4) with } s\in C_0\subset\mathcal{C}(\pi)\} \\
&\geq \frac{1}{n^2}{x_\pi}^T(v_{\pi_\dagger} - T_\pi v_{\pi_\dagger}) && \{\text{Equation (10)}\} \\
&= \frac{1}{n^2}\mathbb{1}^T(v_{\pi_\dagger} - v_\pi). && \{\text{Equation (3)}\}
\end{aligned}$$

If there is no cycle creation, the contraction property given in Lemma 12 implies that after $k = \lceil 2(n-1)\log n\rceil > \frac{\log n^2}{\log\frac{1}{1-\frac{1}{n-1}}}$ iterations we have

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

In the first case considered above, this implies that $\pi'(s_0) \neq \pi(s_0)$. In the second case, this implies that C_0 cannot be a cycle of π'. □

A direct consequence of the above result is Lemma 4 that we originally stated page 4, and that we restate for clarity.

Lemma 4. *After at most $nm\lceil 2(n-1)\log n\rceil$ iterations, either Simplex-PI finishes or a new cycle is created.*

Proof. Before a cycle is created, at most n cycles need to be broken and $n(m-1)$ actions to be eliminated. Each of these $n + n(m-1) = nm$ events requires at most $\lceil 2(n-1)\log n\rceil$ iterations. □

7.2 Part 2: A new cycle implies a significant step towards the optimal value

We now proceed to the second part of the proof, and begin by proving Lemma 5 (originally stated page 4).

Lemma 5. *Suppose Simplex-PI moves from π to π' and that π' involves a new cycle. Then*

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

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

$$\begin{aligned}
\mathbb{1}^T(v_{\pi'} - v_\pi) &= {x_{\pi'}}^T(T_{\pi'}v_\pi - v_\pi) && \{\text{Equation (3)}\} \\
&= x_{\pi'}(s_0)(T_{\pi'}v_\pi(s_0) - v_\pi(s_0)) && \{\text{Simplex-PI switches 1 action}\} \\
&\geq \frac{1}{1-\gamma}(T_{\pi'}v_\pi(s_0) - v_\pi(s_0)). && \{\text{Equation 4 with } s_0 \in C \subset \mathcal{C}(\pi')\} \qquad (14)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
v_{\pi_*} - v_\pi &= (I - \gamma P_{\pi_*})^{-1}(T_{\pi_*}v_\pi - v_\pi) && \{\text{Lemma 11}\} \\
&\leq \frac{1}{1-\gamma}\max_s T_{\pi_*}v_\pi(s) - v_\pi(s) && \{\|(I-\gamma P_{\pi_*})^{-1}\|_\infty \leq \frac{1}{1-\gamma} \text{ and } (I-\gamma P_{\pi_*})^{-1} \succ 0\} \\
&\leq \frac{1}{1-\gamma}\max_s T_{\pi'}v_\pi(s) - v_\pi(s) && \{\max_s T_{\pi'}v_\pi(s) = \max_{s,\tilde{\pi}} T_{\tilde{\pi}}v_\pi(s)\} \\
&= \frac{1}{1-\gamma}(T_{\pi'}v_\pi(s_0) - v_\pi(s_0)). && \{\text{Simplex-PI switches 1 action}\} \quad (15)
\end{aligned}$$

Combining these two observations, we obtain:

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_*} - v_{\pi'}) &= \mathbb{1}^T(v_{\pi_*} - v_{\pi'}) - \mathbb{1}^T(v_{\pi'} - v_\pi) \\
&\leq \mathbb{1}^T(v_{\pi_*} - v_{\pi'}) - \frac{1}{1-\gamma}(T_{\pi'}v_\pi(s_0) - v_\pi(s_0)) && \{\text{Equation (14)}\} \\
&\leq \mathbb{1}^T(v_{\pi_*} - v_{\pi'}) - \max_s v_{\pi*}(s) - v_{\pi'}(s) && \{\text{Equation (15)}\} \\
&\leq \left(1 - \frac{1}{n}\right)\mathbb{1}^T(v_{\pi_*} - v_{\pi'}). && \{\forall x,\ \mathbb{1}^T x \leq \|x\|_\infty\}
\end{aligned}$$

□

Lemma 14. *After $\lceil 2n \log n \rceil$ new cycles are created then some non-optimal action is either eliminated from cycles or entirely eliminated from policies.*

Proof. Let π be the policy in some iteration and π' be any policy between π and π_*. Let $s_0 = \arg\max_s x_\pi(s)(v_{\pi_*}(s) - T_\pi v_{\pi_*}(s))$. We have

$$\begin{aligned}
x_\pi(s_0)(v_{\pi_*}(s_0) - T_\pi v_{\pi_*}(s_0)) &\geq \frac{1}{n}{x_\pi}^T(v_{\pi_*} - T_\pi v_{\pi_*}) && \{\forall x,\ \mathbb{1}^T x \leq n \max_s x(s)\} \\
&= \mathbb{1}^T(v_{\pi_*} - v_\pi). && \{\text{Equation (3)}\} \qquad (16)
\end{aligned}$$

We now consider two cases.

- **case 1:** $s_0 \notin \mathcal{C}(\pi)$.
  If $\pi'(s_0) = \pi(s_0)$, then

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_*} - v_{\pi'}) &= {x_{\pi'}}^T(v_{\pi_*} - T_{\pi'}v_{\pi_*}) && \{\text{Equation (3)}\} \\
&\geq x_{\pi'}(s_0)(v_{\pi_*}(s_0) - T_{\pi'}v_{\pi_*}(s_0)) && \{v_{\pi_*} \geq T_{\pi'}v_{\pi_*}\} \\
&\geq v_{\pi_*}(s_0) - T_{\pi'}v_{\pi_*}(s_0) && \{x_{\pi'}(s_0) \geq 1\} \\
&= v_{\pi_*}(s_0) - T_\pi v_{\pi_*}(s_0) && \{\pi(s_0) = \pi'(s_0)\} \\
&\geq \frac{1}{n}x_\pi(s_0)(v_{\pi_*}(s_0) - T_\pi v_{\pi_*}(s_0)) && \{\text{Equation (6)}\} \\
&\geq \frac{1}{n^2}\mathbb{1}^T(v_{\pi_*} - v_\pi). && \{\text{Equation (16)}\}
\end{aligned}$$

- **case 2:** $s_0 \in \mathcal{C}(\pi)$.
  If $\pi'(s_0) = \pi(s_0)$ and $s_0 \in \mathcal{C}(\pi')$, then

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_*} - v_{\pi'}) &= {x_{\pi'}}^T(v_{\pi_*} - T_{\pi'}v_{\pi_*}) && \{\text{Equation (3)}\}\\
&= \sum_s x_{\pi'}(s)(v_{\pi_*}(s) - T_{\pi'}v_{\pi_*}(s))\\
&\geq \sum_{s\in C_0} x_{\pi'}(s)(v_{\pi_*}(s) - T_{\pi'}v_{\pi_*}(s)) && \{v_{\pi_*} \geq T_{\pi'}v_{\pi_*}\}\\
&\geq \frac{1}{1-\gamma}\sum_{s\in C} v_{\pi_*}(s) - T_{\pi'}v_{\pi_*}(s) && \{\text{Equation 4}\}\\
&\geq \frac{1}{1-\gamma}v_{\pi_*}(s_0) - T_{\pi'}v_{\pi_*}(s_0) && \{v_{\pi_*} \geq T_{\pi'}v_{\pi_*}\}\\
&= \frac{1}{1-\gamma}v_{\pi_*}(s_0) - T_{\pi}v_{\pi_*}(s_0) && \{\pi(s_0) = \pi'(s_0)\}\\
&= \frac{1}{n}x_\pi(s_0)(v_{\pi_*}(s_0) - T_{\pi}v_{\pi_*}(s_0)) && \{x_\pi(s_0) \leq \frac{n}{1-\gamma}\}\\
&\geq \frac{1}{n^2}\mathbb{1}^T(v_{\pi_*} - v_\pi). && \{\text{Equation (16)}\}
\end{aligned}$$

After $k = \lceil 2n\log n\rceil > \frac{\log n^2}{\log\frac{1}{1-\frac{1}{n}}}$ new cycles are created, we have by the contraction property of Lemma 5 that

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

In the first case considered above, this implies that $\pi'(s_0) \neq \pi(s_0)$. In the second case, this implies that $\pi'(s_0) \neq \pi(s_0)$ in the cycles of π'. □

We are ready to conclude: At most, the $n(m-1)$ non-optimal actions may need to be eliminated from cycles and from paths, each such event requiring at most $\lceil 2n\log n\rceil$ cycle creations, that is a total of $2n(m-1)\lceil 2n\log n\rceil$ cycle creations. The result follows from the fact stated in Lemma 4 that a cycle creation requires at most $nm\lceil 2(n-1)\log n\rceil$ iterations.

8 A general bound for Simplex-PI (Proof of Theorem 7)

The proof follows closely the lines of that we have just described for the deterministic case. The only differences are that for any policy π, we need to consider the set of recurrent subchains $\mathcal{R}(\pi)$ instead of the set of cycles $\mathcal{C}(\pi)$, and that Equations (4), (5) and (6) are replaced by:

$$\begin{aligned}
&\forall s \in \mathcal{R}(\pi),\ \frac{1}{(1-\gamma)\tau_r} \leq x_\pi(s) \leq \frac{n}{1-\gamma}, \qquad (17)\\
&\sum_{s\notin\mathcal{R}(\pi)} x_\pi(s) \leq \tau_t,\\
&\text{which implies that: } \forall s \notin \mathcal{R}(\pi),\ x_\pi(s) \leq \tau_t.
\end{aligned}$$

The changes being straightforward, we only provide the resulting Lemmas without further arguments.

8.1 Part 1: Recurrent Subchains are created often

Lemma 15. *Suppose one moves from policy π to policy π'* without creating a new recurrent subchain. *Let $\pi_\dagger$ be the final policy before either a new recurrent subchain is created or the algorithm terminates. Then*

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

Lemma 16. *After at most $\lceil \tau_t \log(n\tau_t) \rceil$ iterations, either Simplex-PI finishes, a new recurrent subchain is created, a cycle is broken, or some action never appears in a policy before a new recurrent subchain is created.*

Lemma 6. *After at most $nm\lceil \tau_t \log(n(\tau_t+1)) \rceil$ iterations, either Simplex-PI finishes or a new recurrent subchain is created.*

8.2 Part 2: A new recurrent subchain implies a significant step towards the optimal value

Lemma 7. *Suppose Simplex-PI moves from π to π' and that π' involves a new recurrent subchain. Then*

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

Lemma 17. *While the algorithm runs, the following events happen:*

- *some non-optimal action is eliminated from recurrent states after $\lceil \tau_r \log(n\tau_r) \rceil$ recurrent subchain creations*
- *some non-optimal action is eliminated from policies after $\lceil \tau_r \log(n\tau_t) \rceil$ recurrent subchain creations*

We are ready to conclude: At most, the $n(m-1)$ non-optimal actions may need to be eliminated from recurrent and transient states, requiring at most a total of $n(m-1)(\lceil \tau_r \log(n\tau_r) \rceil + \lceil \tau_r \log(n\tau_t) \rceil)$ recurrent subchains creations. The result follows from the fact that each subchain creation requires at most $nm\lceil \tau_t \log(n(\tau_t+1)) \rceil$ iterations.

9 Cycle and Recurrent chain 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 is created.*

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

$$\begin{aligned}
v_{\pi_l} - v_{\pi_k} &= T_{\pi_l} v_{\pi_l} - T_{\pi_l} v_{\pi_{k-1}} + T_{\pi_l} v_{\pi_{k-1}} - T_{\pi_k} v_{\pi_{k-1}} + T_{\pi_k} v_{\pi_{k-1}} - T_{\pi_k} v_{\pi_k} && \{\forall \pi,\ T_\pi v_\pi = v_\pi\} \\
&\leq \gamma P_{\pi_l}(v_{\pi_l} - v_{\pi_{k-1}}) + \gamma P_{\pi_k}(v_{\pi_{k-1}} - v_{\pi_k}) && \{T_{\pi_l} v_{\pi_{k-1}} \leq T_{\pi_k} v_{\pi_{k-1}}\} \\
&\leq \gamma P_{\pi_l}(v_{\pi_l} - v_{\pi_{k-1}}). && \{\text{Lemma 1 and } P_{\pi_k} \succ 0\} \\
&\leq (\gamma P_{\pi_l})^k (v_{\pi_l} - v_{\pi_0}). && \{\text{by induction on } k\}
\end{aligned} \tag{18}$$

Since the MDP is deterministic and has n states, $(P_{\pi_l})^n$ will only have non-zero values on columns that correspond to $\mathcal{C}(\pi_l)$. Furthermore, since no cycle is created, $\mathcal{C}(\pi_l) \subset \mathcal{C}(\pi_0)$, which implies that $v_{\pi_l}(s) - v_{\pi_0}(s) = 0$ for all $s \in \mathcal{C}(\pi_l)$. As a consequence, we have $(P_{\pi_l})^n(v_{\pi_l} - v_{\pi_0}) = 0$. By Equation (18), this implies that $v_{\pi_l} = v_{\pi_n}$. If $l > n$, then the algorithm must have terminated. □

Lemma 9. *If the MDP satisfies Assumption 1, after at most $nm\lceil \tau_t \log n\tau_t \rceil$ iterations, either Howard's PI finishes or a new recurrent subchain is created.*

Proof. It can be seen that the proof of Lemma 6 also applies to Howard's PI. □

10 A bound for Howard's PI and Simplex-PI in weakly-communicating MDPs (proof of Theorem 8)

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 first 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 that the state space is only made of the set of recurrent states.

First consider Simplex-PI. If all states are recurrent, new recurrent subchains are created at every iteration, and Lemma 6 holds. Then, in a way similar to the proof of Lemma 17, it can be shown that every $\lceil \tau_r \log n\tau_r \rceil$ iterations, a non-optimal action can be eliminated. As there are at most $n(m-1)$ non-optimal actions, we deduce that Simplex-PI converges in at most $n(m-1)\lceil \tau_r \log n\tau_r \rceil$ iterations on $\mathcal{R}$.

Now consider Howard's PI. We can prove Lemma 10, that we restate for clarity.

Lemma 10. *For an MDP satisfying Assumptions 1-2, suppose Howard's PI moves from π to π' and that π' involves a new recurrent subchain. Then*

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

Proof. In the case we focus on the convergence on the set $\mathcal{R}$ of recurrent states, new recurrent subchains are created at every iteration. So we will prove that the inequality holds for every k. On the one hand, we have for all iterations k,

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_{k+1}} - v_{\pi_k}) &= {x_{\pi_{k+1}}}^T(T_{\pi_{k+1}} v_{\pi_k} - v_{\pi_k}) && \{\text{Equation (3)}\} \\
&\geq \frac{n}{(1-\gamma)\tau_r} \mathbb{1}^T(T_{\pi_{k+1}} v_{\pi_k} - v_{\pi_k} && \{\text{Equation (17)}\} \\
&\geq \frac{n}{(1-\gamma)\tau_r} \|T_{\pi_{k+1}} v_{\pi_k} - v_{\pi_k}\|_\infty. && \{\forall x \geq 0, \mathbb{1}^T x \geq \|x\|_\infty\} \qquad (19)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathbb{1}^T(v_{\pi_*} - v_{\pi_k}) &= {x_{\pi_*}}^T(T_{\pi_*} v_{\pi_k} - v_{\pi_k}) && \{\text{Equation (3)}\} \\
&\leq \frac{n}{1-\gamma} \|T_{\pi_*} v_{\pi_k} - v_{\pi_k})\|_\infty \\
&\leq \frac{n}{1-\gamma} \|T_{\pi_{k+1}} v_{\pi_k} - v_{\pi_k}\|_\infty. \qquad (20)
\end{aligned}$$

By combining Equations (19) and (20), we obtain:

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

□

Then, similary to Simplex-PI, we can prove that after every $\lceil \tau_r \log n\tau_r \rceil$ iterations a non-optimal action must be eliminated. And as there are at most $n(m-1)$ non-optimal actions, we deduce that Howard's PI converges in at most $n(m-1)\lceil \tau_r \log n\tau_r \rceil$ 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 cycles are broken since cycles do not evolve anymore) allows to show that the extra number of iterations for both algorithms to converge on the transient states is $n(m-1)\lceil \tau_t \log n\tau_t \rceil$, and the result follows.

References

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.

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), 1:1–1:16.

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)*.

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. (2012). The simplex method is strongly polynomial for deterministic markov decision processes. Technical report, arXiv:1208.5083v1.

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

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

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.