Improved bound on the worst case complexity of Policy Iteration

Romain Hollanders[a,*], Balázs Gerencsér[a], Jean-Charles Delvenne[a,1], Raphaël M. Jungers[a,2]

[a]*Euler Building, Avenue G. Lemaître 4, 1348 Louvain-la-Neuve, Belgium.*

Abstract

Solving Markov Decision Processes is a recurrent task in engineering which can be performed efficiently in practice using the Policy Iteration algorithm. Regarding its complexity, both lower and upper bounds are known to be exponential (but far apart) in the size of the problem. In this work, we provide the first improvement over the now standard upper bound from Mansour and Singh (1999). We also show that this bound is tight for a natural relaxation of the problem.

Keywords: Policy Iteration, Complexity, Markov Decision Process, Acyclic Unique Sink Orientation

1. Introduction

Markov Decision Processes (MDPs) have been found to be a powerful modeling tool for the decision problems that arise daily in various domains of engineering such as control, finance, queuing systems, PageRank optimization, and many more (see [1] for a more exhaustive list).

MDPs are described from a set of n states in which a system can be. When being in a state, the controller of the system must choose an available action in that state, each of which induces a reward and moves the system to another state according to given transition probabilities. In this work, we assume that the number of actions per state is bounded by a constant k. A policy refers to the stationary choice of one action in every state. Choosing a policy implies fixing a dynamics that corresponds to a Markov chain. Given any policy (there are at most k^n of them), we can associate a value to each state of the MDP that corresponds to the infinite-horizon expected reward of an agent starting in that state. By solving an MDP, we mean providing an optimal policy that maximizes the value of every state. Depending on the application, a total-, discounted- or average-reward criterion may be best suited to define the value function. In every case, an optimal policy always exists. See, e.g., [2] for an in-depth study of MDPs.

A practically efficient way of finding the optimal policy for an MDP is to use Policy Iteration (PI). Starting from an initial policy π_0, $i = 0$, this simple iterative scheme repeatedly computes the value of π_i at every state and greedily modifies this policy using its evaluation to obtain the next iterate π_{i+1}. The modification always ensures that the value of π_{i+1} improves on that of π_i at every state. The process is then repeated until convergence to the optimal policy π^* in a finite number of steps (obviously at most k^n steps—the maximum number of policies). We refer to the ordered set of explored policies as the PI-sequence. A more precise statement of the algorithm as well as some important properties are described in Section 2.

Every iteration of the algorithm can be performed in polynomial time and its number of steps has been shown by Ye to be strongly polynomial in the important particular case of discounted-reward MDPs with a fixed discount rate [3] (the bound in this result was later improved in [4] and [5]). Building on this result, similar conclusions were obtained for other special cases of MDPs [6, 7, 8, 9]. Ye's result does however not extend to *Value Iteration* and *Modified Policy Iteration*, the two standard and closely related competitors of PI [10, 11].

In contrast to these positive results, the number of iterations of PI can be exponentially large in general. Based on the work of Friedmann on Parity Games [12], PI has been shown to require at least $\Omega(2^{n/7})$ steps to converge in the worst case for the total- and average-reward criteria [13] and for the discounted-reward criterion [14]. Friedmann's result was also a major milestone for the study of the Simplex algorithm for Linear Programming as it led to exponential lower bounds for some critical pivoting rules [15, 16]. On the other hand, the best known upper bound for PI to date was due to Mansour and Singh with a $13 \cdot \frac{k^n}{n}$ steps bound [17]. In Theorem 1, Section 3, we provide the first improvement in fifteen years to this bound, namely $\frac{k}{k-1} \cdot \frac{k^n}{n} + o\big(\frac{k^n}{n}\big)$.

To obtain our bound, we use a number of properties of PI-sequences. It is of natural interest to explore which of these properties could be further exploited to improve the bound and which ones cannot. It turns out that the properties we actually use to obtain our upper bound cannot lead to further improvements, that is, they are "fully exploited". To formally prove this fact, we introduce in Section 2 the notion of *pseudo-PI-sequence* to describe any sequence of policies satisfying only the properties that

*Corresponding author, +32(0)10/47.80.10

Email addresses: romain.hollanders@gmail.com (Romain Hollanders), balazs.gerencser@uclouvain.be (Balázs Gerencsér), jean-charles.delvenne@uclouvain.be (Jean-Charles Delvenne), raphael.jungers@uclouvain.be (Raphaël M. Jungers)

[1]CORE and NAXYS fellow.

[2]F.R.S./FNRS Research Associate.

we use to obtain our bound from Theorem 1. We then show in Theorem 2, Section 3, that there always exists a pseudo-PI-sequence whose size matches the upper bound of Theorem 1. This confirms that the bound is sharp for pseudo-PI-sequences. Therefore, obtaining new bounds on PI-sequences would require exploiting stronger properties.

An attempt in that direction—based on the so-called *Order-Regular* matrices—has been proposed in [18] and developed in [19]. Based on numerical evidence, Hansen and Zwick conjectured that the number of iterations of PI for $k = 2$ should be bounded by F_{n+2} $(= O(1.618^n))$, the $(n+2)^{\text{nd}}$ Fibonacci number. If true, this bound would significantly improve ours.

As a final remark, note that our analysis also fits in the frameworks of the *Strategy Iteration* algorithm to solve 2-Player Turn-Based Stochastic Games [20]—a 2-player generalization of MDPs—and of the *Bottom Antipodal* algorithm to find the sink of an Acyclic Unique Sink Orientation of a grid [21, 22]. Our bound can also be adapted for these algorithms. It is to be noted that no polynomial-time algorithm is known for either case[3], which is an additional incentive to improve the exponential bounds.

2. Problem statement and preliminary results

Definition 1 (Markov Decision Process)**.** Let $\mathcal{S} = \{1, \ldots, n\}$ be a set of n *states* and $\mathcal{A}_s$ be a set of k *actions* available for state $s \in \mathcal{S}$. To each choice of an action corresponds a *transition probability* distribution for the next state to visit as well as a *reward*. For simplicity, we use a common numbering for the actions, that is, $\mathcal{A}_s \triangleq \mathcal{A} = \{1, \ldots, k\}$ for all $s \in \mathcal{S}$. With this notation, for every pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, the transition probability and reward functions are uniquely defined. Let a *policy* $\pi \in \{1, \ldots, k\}^n$ be the stationary choice of one action for every state. A policy induces a transition probability matrix P^π corresponding to some Markov chain and a reward vector r^π. We may ask how rewarding a policy π is in the long run. This is represented by its *value* vector $v^\pi \in \mathbb{R}^n$ whose s^{th} entry corresponds to the long term reward obtained from starting in state s and following the policy π thereafter. It can be computed by solving a system whose definition depends on the problem studied. For instance, for the standard *infinite-horizon discounted-reward criterion* where the aim is to maximize the discounted sum of rewards, v^π is obtained by:

$$v^\pi = \sum_{i=0}^{\infty} (\gamma P^\pi)^i \;\; r^\pi = (I - \gamma P^\pi)^{-1} r^\pi,$$

where $0 \leq \gamma < 1$ is the discount factor that ensures that $(I - \gamma P^\pi)$ is non-singular. However, in this work, the bounds that we derive hold for the three classical reward criteria, namely the discounted-, total- and average-reward criteria. For the total-reward criterion, we assume the existence of a terminal—reward free—state that we assume to be reachable with any policy, from any starting state. For the average-reward criterion, we need to extend the notion of value vectors to *valuations* as defined in [18, Section 2.3] or [2, Section 8.2.1]. By *solving* an MDP, we mean finding the optimal policy π^* such that for any other policy π, $v^{\pi^*} \geq v^\pi$, that is, $v^{\pi^*}(s) \geq v^\pi(s)$ for all states s. The existence of such a policy is guaranteed [23].

Definition 2 (Domination)**.** Given two policies π and π', if $v^{\pi'}(s) \geq v^\pi(s)$ for all states $s \in \mathcal{S}$, then we say that π' *dominates* π and we write $\pi' \succeq \pi$. If moreover $v^{\pi'}(s) > v^\pi(s)$ for at least one state, then the domination is strict and we write $\pi' \succ \pi$. An analoguous definition of domination can be obtained for valuations with the average reward criterion [18, Section 2.3].

Definition 3 (Switching)**.** Let U be a collection of state-action pairs (s, a). We say that U is *well-defined* if it contains every state $s \in \mathcal{S}$ at most once. In that case, we define $\pi' = \pi \oplus U$ to be the policy obtained from π by *switching* the action $\pi(s)$ to a for each (s, a)-pair in U.

Definition 4 (Improvement set)**.** We define the *improvement set* of a policy π as:

$$T^\pi = \big\{(s, a) \mid \pi \oplus \{(s, a)\} \succ \pi\big\},$$

and the set of *improvement states* S^π of π as the set of states that appear in T^π.

Proposition 1. *Let π be a policy and $U \neq \emptyset$ be any well-defined subset of its improvement set T^π. Then $\pi \oplus U \succ \pi$.*

Proposition 2. *For a given policy π, if $T^\pi = \emptyset$, then π is optimal.*

Proofs of Propositions 1 and 2 can be found, e.g., in [18]; see Theorems 2.2.12, 2.3.9 and 2.4.6 for the discounted-, average- and total-reward criteria, respectively. Alternative statements can also be found in [2] and [23]. Based on Propositions 1 and 2 we may define the *Policy Iteration* algorithm to find the optimal policy. In the rest of our analysis, we only assume that these two propositions hold.

Definition 5 (Policy Iteration)**.** Algorithm 2 describes *Policy Iteration* (PI). The standard way of choosing $U_i \subseteq T^{\pi_i}$ is the greedy update rule, namely choose any U_i with maximal cardinality $|S^{\pi_i}|$. We refer to the corresponding algorithm as *Greedy PI*, which is the focus of our work.

```
Initialization: π_0, i = 0
while T^{π_i} ≠ ∅ do
    Select a non-empty and well-defined U_i ⊆ T^{π_i}
    π_{i+1} = π_i ⊕ U_i
    i ← i + 1
end
return π_i
```

Algorithm 1: Policy Iteration

[3] Although the strongly polynomial time bound from Ye extends to 2TBSGs as well when a fixed discount factor is chosen [4].

Definition 6 (Comparable)**.** We say that two policies π and π' are *comparable* if either $\pi \preceq \pi'$ or $\pi \succeq \pi'$. We call two policies *neighbors* if they differ in only one state. Neighbors are always comparable (Lemma 3 in [17]).

Definition 7 (Partial order)**.** For a given MDP, we consider the *partial order* PO of the policies defined by the domination relation. A set of policies $\pi^{(1)}, \ldots, \pi^{(k)}$ is called a *sequence* if $\pi^{(1)} \preceq \cdots \preceq \pi^{(k)}$.

Definition 8 (PI-sequence)**.** We refer to the sequence of policies $\pi_0, \ldots, \pi_{m-1}$ explored by greedy PI as a *PI-sequence* of length m.

We aim to solve the following problem.

Problem 1. Find the longest possible PI-sequence.

Lemma 3 (Lemma 4 in [17])**.** *For any policies π, π' such that $\pi'(s) = \pi(s)$ for all states $s \in S^\pi$, we have $\pi' \preceq \pi$.*

Proof. See Appendix A. □

The next property indicates how the improvement set of a policy is constrained by the dominated policies and by their own improvement sets.

Proposition 4 (Extended from Lemma 12 in [17])**.** *For any two policies $\pi \prec \pi'$, there exists an improvement state $s \in S^\pi$ such that $\pi(s) \neq \pi'(s)$ and $(s, \pi(s)) \notin T^{\pi'}$.*

Proof. See Appendix B. □

Note that for $k = 2$, the statement of Proposition 4 can be simplified and implies that for any two policies $\pi \prec \pi'$, it holds that $S^\pi \not\subseteq S^{\pi'}$.

When performing a PI step, we jump from the current policy to some policy that can be quite different (in terms of number of different entries). However, we now show that there always exists a path of small steps in the partial order connecting the two, that is, from neighbor to neighbor.

Proposition 5 (Extended from Lemma 6 in [17])**.** *Let π and π' be two policies such that $\pi' = \pi \oplus U$ for some well-defined $U \subseteq T^\pi$ of cardinality d. Then there exist d distinct policies $\pi^{(1)}, \ldots, \pi^{(d)}$ such that $\pi \prec \pi^{(1)} \preceq \cdots \preceq \pi^{(d)} = \pi'$ and such that $\pi^{(i)}$ and $\pi^{(i+1)}$ are neighbors for all $1 \leq i < d$.*

Proof. See Appendix C. □

Definition 9 (Subsequence and supersequence)**.** Let O be a sequence. We call *subsequence* of O any ordered subset of elements of O. We call *supersequence* of O any sequence that contains O as a subsequence.

The following property is perhaps the most important consequence of Proposition 5.

Corollary 6 (Jumping)**.** *Let π_i be a policy of a PI-sequence. Then the partial order of policies contains a supersequence of the PI-sequence with at least $|S^{\pi_i}|$ different policies between π_i and π_{i+1}, that is, $|S^{\pi_i}|$ policies π such that $\pi_i \prec \pi \preceq \pi_{i+1}$. When we step from π_i to π_{i+1}, we say that we* jump *$|S^{\pi_i}|$ policies of the supersequence.*

Proof. The result is a direct consequence of Proposition 5. Recall that with Greedy PI, $|U_i|$ always equals $|S^{\pi_i}|$. □

We now introduce an object that is similar to a PI-sequence in that it describes a sequence of policies embedded into a partial order. However, we will forget about some of the structure that originates from MDPs and only require Proposition 4 and Corollary 6 to be ensured by the sequence and the partial order. This will allow us to show that these two properties—that will be the two milestones in the proof of our upper bound in Theorem 1—can actually not help to further improve the bound.

Definition 10 (Pseudo-PI-sequence)**.** We call *pseudo-PI-sequence* of size m a triple $(\Pi, O, \mathcal{T})$ where:

- $\Pi = \pi_0, \pi_1, \ldots, \pi_{m-1}$ is a sequence of policies. We define the abstract ordering $\prec$ on the elements of the sequence Π by the ordering of their indices.
- O is a sequence of policies of $\{1, \ldots, k\}^n$ that is a supersequence of Π.
- $\mathcal{T}$ is a collection of abstract improvement sets T^π for every policy π appearing in O.

We require the claim from Proposition 4 to hold for O and we require Π to satisfy Corollary 6 as a subsequence of O.

Definition 10 leads to a relaxation of Problem 1, and therefore any upper bound on the size of pseudo-PI-sequences also holds for PI-sequences. Note that there is a natural way of constructing a pseudo-PI-sequence from any PI-sequence. Of course, Proposition 4 and Corollary 6, which are the key results towards our upper bound in Theorem 1, still hold for pseudo-PI-sequences by design. Furthermore, as we will show in Theorem 2, our upper bound is tight for the relaxation.

Problem 2. Find the longest possible pseudo-PI-sequence.

3. Main result: a better upper bound on PI that is tight for Problem 2

In order to precisely solve Problem 2, we need to both provide a lower and an upper bound on the length of pseudo-PI-sequences. We start by showing the upper bound, which also holds for Problem 1 and therefore provides a new upper bound on the complexity of Policy Iteration in general.

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

Before we proceed to the proof of Theorem 1, we need to formulate two additional properties. First, we derive the following lemma from Proposition 4.

Lemma 7 (Adapted from Lemma 4 in [17])**.** *Let $(\Pi, O, \mathcal{T})$ be a pseudo-PI-sequence. For any two policies $\pi \prec \pi'$ of O and any $U \subseteq T^{\pi'}$, we have $\pi \neq \pi' \oplus U$.*

Proof. See Appendix D. □

When $k = 2$, it is easy to see using Proposition 4 that two policies with exactly the same improvement states cannot exist. When $k > 2$, this is no longer the case. However, using Lemma 3, Mansour and Singh showed that there cannot be more than k^d policies with the same d improvement states in a PI-sequence (see Corollary 13 in [17]). In the following proposition, we use Proposition 4 to improve this bound to $(k-1)^d$. Note that this improvement is a crucial and nontrivial step that allowed us to achieve tight bounds for Problem 2.

Proposition 8. *Given a pseudo-PI-sequence $(\Pi, O, \mathcal{T})$ and a set of states $S \subseteq \mathcal{S}$ of cardinality d, it holds that O contains at most $(k-1)^d$ policies π with $S^\pi = S$.*

Proof. Given the supersequence O of the pseudo-PI-sequence, we consider its subsequence $\pi^{(1)} \preceq \cdots \preceq \pi^{(K)}$ such that $S^{\pi^{(i)}} = S \triangleq \{s_1, \ldots, s_d\}$ for all $1 \le i \le K$. We show that if the subsequence satisfies Proposition 4, then $K \le (k-1)^d$. To this end, we first claim that the improvement sets of the policies of the subsequence can be assumed to be all well-defined. Indeed, for any policy of the subsequence $\pi^{(i)}$, we can simplify its improvement set $T^{\pi^{(i)}}$ by keeping only a single (s, a) pair for every $s \in S^{\pi^{(i)}}$. This does not modify $S^{\pi^{(i)}}$ (i.e., $\pi^{(i)}$ remains in the subsequence), nor does it imply the violation of Proposition 4. Therefore, given a policy $\pi^{(i)}$ of the subsequence and a state $s \in S$, we can assume that there is exactly one action a such that $(s, a) \in T^{\pi^{(i)}}$, which we refer to as $T^{\pi^{(i)}}(s)$.

We represent an action $i \in \mathcal{A}$ as a k-dimensional base vector $f_a(i) \triangleq e_i$ of $V = \mathbb{R}^k$, where $e_i(j) = 1$ if $i = j$, 0 otherwise. Similarly, we represent policies as base vectors of the space $W = V^{\otimes d}$ of dimension k^d through the application:

$$f_p : \pi \longmapsto f_a(\pi(s_1)) \otimes \cdots \otimes f_a(\pi(s_d)),$$

where $\otimes$ stands for the Kronecker product. Finally, we represent pairs of policies and their improvement sets in a similar way in W through the application:

$$\begin{aligned} f_c : (\pi, T^\pi) \longmapsto & \Big[f_a(\pi(s_1)) - f_a(T^\pi(s_1))\Big] \otimes \ldots \\ & \otimes \Big[f_a(\pi(s_d)) - f_a(T^\pi(s_d))\Big], \\ = f_p(\pi) \quad & + \sum_{\substack{U \subseteq T^\pi \\ U \neq \emptyset}} (-1)^{|U|} \cdot f_p(\pi \oplus U). \end{aligned}$$

We claim that the vectors $f_c\big(\pi^{(i)}, T^{\pi^{(i)}}\big)$ are linearly independent. Assume on the contrary that we have:

$$\sum_{i=1}^{K} \lambda_i \, f_c\left(\pi^{(i)}, T^{\pi^{(i)}}\right) = 0, \tag{1}$$

with not all λ_i being 0. Choose the first index i with non-zero λ_i. The corresponding term gives a non-zero coefficient to the base vector $f_p\big(\pi^{(i)}\big)$. But from Lemma 7, for all $j > i$ and all $U \subseteq T^{\pi^{(j)}}$, $\pi^{(i)} \neq \pi^{(j)} \oplus U$. Thus the base vector $f_p\big(\pi^{(i)}\big)$ never appears later in the series in (1) which can therefore not be null.

Additionally, the coordinates of $f_a(\pi(s_i)) - f_a(T^\pi(s_i)) \in V$ sum to 0 (in the standard base) for all $1 \le i \le d$ which means they lie in a subspace V_0 of V of dimension $k-1$. As a result,

$$f_c\left(\pi^{(i)}, T^{\pi^{(i)}}\right) \in W_0 = V_0^{\otimes d}.$$

The dimension of W_0 is $(k-1)^d$ implying this is the maximum number of linearly independent vectors $f_c\big(\pi^{(i)}, T^{\pi^{(i)}}\big)$. This translates to the desired upper bound. □

Of course, the above result also holds for usual PI-sequences.

Proof of Theorem 1. The proof proceeds in two steps. First, we consider "small" improvement sets and show that there are at most $o\left(\frac{k^n}{n}\right)$ of them. Then we consider "large" improvement sets and show that PI explores at most $\frac{k}{k-1} \cdot \frac{k^n}{n} + o\left(\frac{k^n}{n}\right)$ of them because they jump many policies on the way.

Policies with small improvement sets. We consider the improvement sets T^π such that $|S^\pi| \le \frac{k-1}{k} \cdot n - f(n)$ with $f(n) \triangleq \sqrt{n \log n}$.

From Proposition 8, policies with the same set of improvement states S of cardinality d can appear at most $(k-1)^d$ times in a (pseudo-)PI-sequence, hence the number of small improvement sets can be expressed as follows:

$$\begin{aligned} & \sum_{d=0}^{\left\lfloor \frac{k-1}{k} \cdot n - f(n) \right\rfloor} \binom{n}{d} (k-1)^d \\ & \qquad = k^n \sum_{d=0}^{\left\lfloor \frac{k-1}{k} \cdot n - f(n) \right\rfloor} \binom{n}{d} \left(\frac{k-1}{k}\right)^d \left(\frac{1}{k}\right)^{n-d}, \\ & \qquad = k^n \cdot P\left[X \le \frac{k-1}{k} \cdot n - f(n)\right], \end{aligned}$$

where $X \sim \mathrm{Bin}\left(n, \frac{k-1}{k}\right)$ follows a binomial distribution. Using Hoeffding's inequality [24], we have:

$$P\left[X \le n \cdot \left(\frac{k-1}{k} - \frac{f(n)}{n}\right)\right] \le e^{-2 \cdot \left(\frac{f(n)}{n}\right)^2 \cdot n} = \frac{1}{n^2}.$$

Therefore we have:

$$\sum_{d=0}^{\left\lfloor \frac{k-1}{k} \cdot n - f(n) \right\rfloor} \binom{n}{d} (k-1)^d \le k^n \cdot \frac{1}{n^2} = o\left(\frac{k^n}{n}\right).$$

Policies with large improvement sets. We now consider the improvement sets T^π with the set of improvement states satisfying $|S^\pi| > \frac{k-1}{k} \cdot n - f(n)$. We show that these sets jump many policies on the way and hence we cannot have many of them in the (pseudo-)PI-sequence. Suppose that we have K such improvement sets

in the sequence. Then, from Corollary 6, we jump at least $K \cdot \left(\frac{k-1}{k} \cdot n - f(n)\right)$ distinct policies. Since we cannot jump more that k^n policies, we have the following condition on K:

$$K \leq \frac{k^n}{\frac{k-1}{k}n - f(n)} = \frac{k}{k-1} \cdot \frac{k^n}{n} \cdot \frac{1}{1 - \frac{k-1}{k}\sqrt{\frac{\log n}{n}}}.$$

Hence, $K \leq \frac{k}{k-1} \cdot \frac{k^n}{n} \cdot \left(1 + o(1)\right)$. □

The following theorem shows that the upper bound from Theorem 1 is tight for Problem 2.

Theorem 2. *There exists a pseudo-PI-sequence of size* $\frac{k}{k-1} \cdot \frac{k^n}{n} + \omega\left(\frac{k^n}{n}\right)$.

Proof. We first build a sequence containing all the k^n policies that will play the role of the supersequence O for the pseudo-PI-sequence. Preliminarily, given any policy π of O, we define its (well-defined) improvement set T^π such that $(s, a) \in T^\pi$ iff $\pi(s) \neq k$ and $a = k$. Here action k can be thought of as some special action. Let $\mathcal{T}^d$ be the set of all policies π such that $|T^\pi| = d$. By definition, $\mathcal{T}^d$ contains all policies π such that $\pi(s) \neq k$ for exactly d different states s, hence $\binom{n}{d} \cdot (k-1)^d$ elements. We now order all k^n policies as a sequence by decreasing order of cardinality of their improvement sets, hence the policies in $\mathcal{T}^d$-sets with a large d come first in the sequence. The (total) ordering inside a given $\mathcal{T}^d$-set can be arbitrarily chosen. Given this ordering, notice that for any $\pi \prec \pi'$, if $S^\pi \subseteq S^{\pi'}$, then $S^\pi = S^{\pi'}$.

The sequence O obtained with the above construction satisfies the claim of Proposition 4. Indeed, let us choose any two policies of the sequence $\pi \prec \pi'$. First assume that $S^\pi \setminus S^{\pi'} \neq \emptyset$ and let $t \in S^\pi \setminus S^{\pi'}$. Then by construction, $\pi(t) \neq k = \pi'(t)$ and $(t, \pi(t)) \notin T^{\pi'}$ since $t \notin S^{\pi'}$, hence Proposition 4 is true in that case. If now $S^\pi \setminus S^{\pi'} = \emptyset$, then the ordering of the policies imposes that $S^\pi = S^{\pi'}$, as observed above. In that case, by construction $\pi(s) \neq k$ for all $s \in S^\pi$ and $\pi(s) = \pi'(s) = k$ for all $s \notin S^\pi$. Since $\pi \neq \pi'$, there must exist some state $t \in S^\pi$ such that $\pi(t) \neq \pi'(t)$. Furthermore by definition of $T^{\pi'}$, $(t, \pi(t)) \notin T^{\pi'}$ because $\pi(t) \neq k$, and the claim of Proposition 4 is true again.

At this point, we have built a supersequence for our PI-sequence that satisfies the claim of Proposition 4. Let

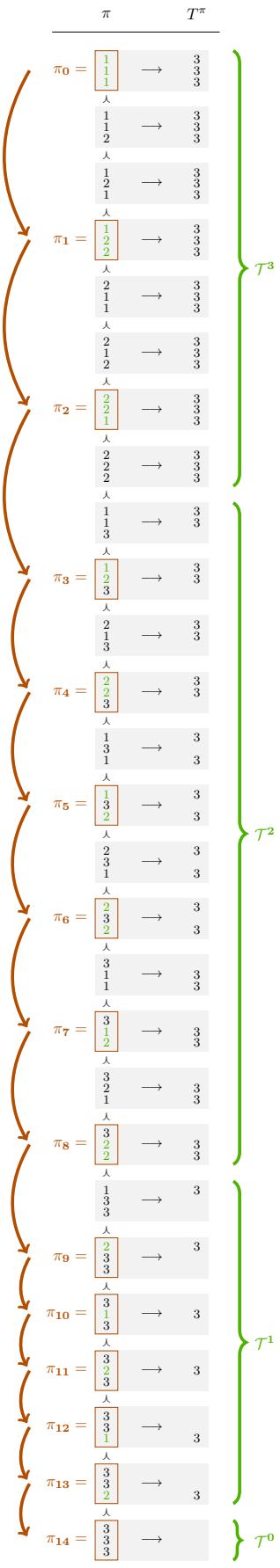

Figure 1: An example of a pseudo-PI-Sequence of size $\frac{k}{k-1} \cdot \frac{k^n}{n} + o\left(\frac{k^n}{n}\right)$ with its supersequence O for $n = k = 3$. Each gray box corresponds to a policy of the supersequence. We represent the improvement sets only through the prospective improving action for each state (action 3 for state s if $\pi(s) \neq 3$ or nothing, according to the construction). The red policies are the ones from the sequence Π from Definition 10. It can be checked that if some policy π_i is in $\mathcal{T}^d$, then d policies of the supersequence are jumped from π_i to π_{i+1} and it can be observed that the supersequence contains k^n elements and satisfies the claim of Proposition 4.

us now select a subsequence Π of O while ensuring Corollary 6 as follows: we start from the first policy of the supersequence π_0, $i = 0$. Then at each step i, we jump $|T^{\pi_i}|$ elements in the sequence to select π_{i+1}. With this greedy procedure, we clearly ensure Corollary 6 and we pick at least $\frac{1}{d+1}|\mathcal{T}^d|$ policies from each $\mathcal{T}^d$-set, for a total number of hypothetical PI-steps of at least:

$$
\begin{aligned}
&\sum_{d=0}^{n} \frac{1}{d+1}|\mathcal{T}^d|, \\
&= \sum_{d=0}^{n} \frac{1}{d+1}\binom{n}{d}(k-1)^d, \\
&= \frac{1}{n+1}\cdot\sum_{d=0}^{n}\binom{n+1}{d+1}\cdot(k-1)^d\cdot 1^{n-d}, \\
&= \frac{1}{k-1}\cdot\frac{1}{n+1}\cdot\Bigg[\underbrace{\sum_{d=0}^{n+1}\binom{n+1}{d}\cdot(k-1)^d\cdot 1^{(n+1)-d}}_{=k^{n+1}} - 1\Bigg], \\
&= \frac{k}{k-1}\cdot\frac{k^n}{n}\cdot\left(1-\frac{1}{k^{n+1}}\right)\cdot\left(1-\frac{1}{n+1}\right), \\
&= \frac{k}{k-1}\cdot\frac{k^n}{n}\cdot(1+\omega(1)),
\end{aligned}
$$

which corresponds to our claim and matches the upper bound from Theorem 1. An example of a pseudo-PI-sequence constructed from the above procedure with $n = k = 3$ is given in Figure 3. □

Of course, the lower bound from Theorem 2 only holds for pseudo-PI-sequences which are less constrained than usual PI-sequences. Indeed, it can for instance be observed that the pseudo-PI-sequence constructed in Figure 3 cannot correspond to a real PI-run since for instance its supersequence does not satisfy Proposition 1. Therefore, obtaining better bounds than the one from Theorem 1 will require a more advanced analysis.

4. Acknowledgments

This work was supported by an ARC grant from the French Community of Belgium and by the IAP network "Dysco" funded by the office of the Prime Minister of Belgium. The scientific responsibility rests with the authors.

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

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

[3] Y. Ye, The Simplex and Policy-Iteration Methods are Strongly Polynomial for the Markov Decision Problem with a Fixed Discount Rate, Mathematics of Operations Research 36(4) (2011) 593–603.

[4] T. D. Hansen, P. B. Miltersen, U. Zwick, Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor, Journal of the ACM, JACM 60 (1) (2013) 1.

[5] B. Scherrer, Improved and generalized upper bounds on the complexity of Policy Iteration, In Proceedings of the 27th Conference on Advances in Neural Information Processing Systems, NIPS (2013) 386–394.

[6] I. Post, Y. Ye, The simplex method is strongly polynomial for deterministic markov decision processes, Mathematics of Operations Research.

[7] E. A. Feinberg, J. Huang, Strong polynomiality of policy iterations for average-cost MDPs modeling replacement and maintenance problems, Operations Research Letters 41 (3) (2013) 249–251.

[8] M. Akian, S. Gaubert, Policy iteration for perfect information stochastic mean payoff games with bounded first return times is strongly polynomial, arXiv preprint arXiv:1310.4953.

[9] E. A. Feinberg, J. Huang, On the reduction of total-cost and average-cost MDPs to discounted MDPs, arXiv preprint arXiv:1507.00664.

[10] E. A. Feinberg, J. Huang, The value iteration algorithm is not strongly polynomial for discounted dynamic programming, Operations Research Letters 42 (2) (2014) 130–131.

[11] E. A. Feinberg, J. Huang, B. Scherrer, Modified policy iteration algorithms are not strongly polynomial for discounted dynamic programming, Operations Research Letters 42 (6) (2014) 429–431.

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

[13] J. Fearnley, Exponential Lower Bounds for Policy Iteration, In Proceedings of the 37th International Colloquium on Automata, Languages and Programming, ICALP (2010) 551–562.

[14] R. Hollanders, J.-C. Delvenne, R. M. Jungers, The Complexity of Policy Iteration is Exponential for Discounted Markov Decision Processes., In Proceedings of the 51st IEEE Conference on Decision and Control, CDC (2012) 5997–6002.

[15] O. Friedmann, A Subexponential Lower Bound for Zadehs Pivoting Rule for Solving Linear Programs and Games, In Poceedings of the 14th Conference on Integer Programming and Combinatoral Optimization, IPCO (2011) 192–206.

[16] O. Friedmann, T. D. Hansen, U. Zwick, Subexponential Lower Bounds for Randomized Pivoting Rules for the Simplex Algorithm, In Poceedings of the 43rd ACM Symposium on Theory of Computing, STOC 11 (2011) 283–292.

[17] Y. Mansour, S. Singh, On the Complexity of Policy Iteration, In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, UAI (1999) 401–408.

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

[19] B. Gerencsér, R. Hollanders, J.-C. Delvenne, R. M. Jungers, A complexity analysis of policy iteration through combinatorial matrices arising from unique sink orientations, arXiv preprint arXiv:1407.4293.

[20] T. D. Hansen, M. Paterson, U. Zwick, Improved upper bounds for random-edge and random-jump on abstract cubes, In Proceedings of the 25th Symposium On Discrete Algorithms (SODA) (2014) 874–881.

[21] T. Szabó, E. Welzl, Unique Sink Orientations of Cubes, In Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science, FOCS (2001) 547–555.

[22] B. Gärtner, D. Walter Jr, L. Rüst, Unique sink orientations of grids, Algorithmica 51 (2) (2008) 200–235.

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

[24] W. Hoeffding, Probability inequalities for sums of bounded random variables, Journal of the American statistical association 58 (301) (1963) 13–30.

Appendix A. Proof of Proposition 3

Lemma 3 (Lemma 4 in [17]). *For any policies π, π' such that $\pi'(s) = \pi(s)$ for all states $s \in S^\pi$, we have $\pi' \preceq \pi$.*

Proof. Consider an MDP M' where the only action that is available in the states $s \in S^\pi$ is $\pi(s)$. Clearly π and π' are valid policies for M' and their value does not change. On the other hand, the improvement set of π is empty in M', so by Proposition 2, π is optimal for M' and we must have $\pi' \preceq \pi$. □

Appendix B. Proof of Proposition 4

Proposition 4 (Extended from Lemma 12 in [17]). *For any two policies $\pi \prec \pi'$, there exists an improvement state $s \in S^\pi$ such that $\pi(s) \neq \pi'(s)$ and $(s, \pi(s)) \notin T^{\pi'}$.*

Proof. Suppose on the contrary that it is not the case. Then for each state $s \in S^\pi$, either $\pi(s) = \pi'(s)$ or $(s, \pi(s)) \in T^{\pi'}$. Let $U \triangleq \big\{ (s, \pi(s)) : s \in S^\pi \cap S^{\pi'} \text{ and } \pi(s) \neq \pi'(s) \big\}$, then we have $U \subseteq T^{\pi'}$. Therefore, Proposition 1 tells us that $\pi'' \triangleq \pi' \oplus U \succeq \pi'$.

Now, let us consider any $s \in S^\pi$. If $\pi'(s) = \pi(s)$, then for any $a \in \mathcal{A}$, we have $(s, a) \notin U$ and $\pi''(s) = \pi(s)$. On the other hand, if $\pi'(s) \neq \pi(s)$, then $s \in S^{\pi'}$, hence $(s, \pi(s)) \in U$ and $\pi''(s) = \pi(s)$ again. Therefore $\pi''(s) = \pi(s)$ for all $s \in S^\pi$ and from Lemma 3, $\pi'' \preceq \pi$ $(\prec \pi')$ which is a contradiction. □

Appendix C. Proof of Proposition 5

Proposition 5 (Extended from Lemma 6 in [17]). *Let π and π' be two policies such that $\pi' = \pi \oplus U$ for some well-defined $U \subseteq T^\pi$ of cardinality d. Then there exist d distinct policies $\pi^{(1)}, \ldots, \pi^{(d)}$ such that $\pi \prec \pi^{(1)} \preceq \cdots \preceq \pi^{(d)} = \pi'$ and such that $\pi^{(i)}$ and $\pi^{(i+1)}$ are neighbors for all $1 \leq i < d$.*

Proof. If $d = 1$, simply take $\pi^{(d)} = \pi'$. Suppose that the result is true for $d - 1 \geq 1$ and let us show it for d. From Proposition 4, there exists a state $s \in S^\pi$ such that $(s, \pi(s)) \notin T^{\pi'}$, that is, such that $\pi' \oplus (s, \pi(s)) \not\succ \pi'$. Since neighbors are always comparable, it means that $\pi'' \triangleq \pi' \oplus (s, \pi(s)) \preceq \pi'$. By definition of π', we have $(s, \pi'(s)) \in U$ and $U' \triangleq U \setminus (s, \pi'(s)) \subseteq U \subseteq T^\pi$. We can then recursively apply the statement of Proposition 5 with:

$$\pi' \longmapsto \pi'' = \pi' \oplus (s, \pi(s)),$$
$$U \longmapsto U' = U \setminus (s, \pi'(s)),$$

since $\pi'' = \pi \oplus U'$ and $|U'| = d - 1$. In that case, $\pi^{(d-1)} = \pi''$, and we can choose $\pi^{(d)} = \pi'$ which is indeed a neighbor of $\pi^{(d-1)}$. Moreover, the policies $\pi^{(1)}, ..., \pi^{(d)}$ are all distinct from each other. Indeed, let $\text{dist}(\mu, \mu') = |\{s : \mu(s) \neq \mu'(s)\}|$ be the number of different entries of two policies μ and μ'. Clearly, $\text{dist}(\pi, \pi^{(1)}) = \text{dist}(\pi^{(i)}, \pi^{(i+1)}) = 1$ for all $1 \leq i < d$ and $\text{dist}(\pi, \pi') = d$. This implies that $\text{dist}(\pi, \pi^{(i)}) = i$ for all $1 \leq i < d$ and the policies $\pi^{(i)}$ must therefore all be distinct from each other. □

Note that the above proof simplifies Mansour and Singh's original argument.

Appendix D. Proof of Lemma 7

Lemma 7 (Adapted from Lemma 4 in [17]). *Let $(\Pi, O, \mathcal{T})$ be a pseudo-PI-sequence. For any two policies $\pi \prec \pi'$ of O and any $U \subseteq T^{\pi'}$, we have $\pi \neq \pi' \oplus U$.*

Proof. Let $s \in S^\pi$ such that $\pi'(s) \neq \pi(s)$ and $(s, \pi(s)) \notin T^{\pi'}$ whose existence is guaranteed by Proposition 4. It is impossible to switch from $\pi'(s)$ to $\pi(s)$ hence the result. □