Improved bound on the worst case complexity of Policy Iteration

Romain Hollanders, Balázs Gerencsér, Jean-Charles Delvenne and Raphaël M. Jungers*

Abstract

Solving Markov Decision Processes (MDPs) is a recurrent task in engineering. Even though it is known that solutions for minimizing the infinite horizon expected reward can be found in polynomial time using Linear Programming techniques, iterative methods like the Policy Iteration algorithm (PI) remain usually the most efficient in practice. This method is guaranteed to converge in a finite number of steps. Unfortunately, it is known that it may require an exponential number of steps in the size of the problem to converge. On the other hand, many open questions remain considering the actual worst case complexity. In this work, we provide the first improvement over the fifteen years old upper bound from Mansour & Singh (1999) by showing that PI requires at most $\frac{k}{k-1} \cdot \frac{k^n}{n} + o\big(\frac{k^n}{n}\big)$ iterations to converge, where n is the number of states of the MDP and k is the maximum number of actions per state. Perhaps more importantly, we also show that this bound is optimal for an important relaxation of the problem.

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 [Ber07], finance [BR11], communication networks [Alt02], queuing systems [Mey08], PageRank optimization [CJB14], and many more (see [Whi93] 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. Note that in every case, an optimal policy always exists. See e.g. [Ber07] and [Put94] for a comprehensive and in-depth study of MDPs.

One practically efficient way of finding the optimal policy for an MDP is to use the Policy Iteration algorithm (PI). Starting from an initial policy π_0, $i = 0$, 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 to be strongly polynomial in some important particular cases such as discounted-reward MDPs

*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 responsiblity rests with the authors. J.-C. D. is with CORE and NAXYS. R. M. J. is an F.R.S./FNRS Research Associate.

with a fixed discount rate [Tse90, Ye11] or deterministic MDPs [PY13] (the bounds in these results were later improved in [HMZ13] and [Sch13]). However, in the general case the number of iterations of PI can be exponentially large. Based on the work of Friedmann on Parity Games [Fri09], 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 [Fea10] and for the discounted-reward criterion [HDJ12]. Friedmann's result was also a major milestone for the study of the Simplex algorithm for Linear Programming as it lead to exponential lower bounds for some critical pivoting rules [Fri11, FHZ11]. On the other hand, the best known upper bound for PI to date was due to Mansour & Singh with a $13 \cdot \frac{k^n}{n}$ steps bound [MS99]. In Section 4, we provide the first improvement in fifteen years over Mansour & Singh's bound, namely $\frac{k}{k-1} \cdot \frac{k^n}{n} + o\big(\frac{k^n}{n}\big)$. PI-sequences need to verify several combinatorial conditions that have been identified over the years [MS99, Mad98, SW01, HK99]. However to obtain our bound, we only exploit a subset of these conditions. In Section 3, we define the notion of Pseudo-PI-sequence that describes any sequence of policies satisfying this subset of conditions. We then prove in Section 4 that the above upper bound holds for both PI- and Pseudo-PI-sequences. As it turns out, our bound is tight for Pseudo-PI-sequences. Indeed, in Section 5 we provide a construction of a Pseudo-PI-sequence of length $\frac{k}{k-1} \cdot \frac{k^n}{n} + o\big(\frac{k^n}{n}\big)$. We believe that this construction is important in that it shows that additional properties of PI-Sequences must be exploited if a tighter bound is to be obtained, see Sect 6.

2 Problem statement

Definition 1 (Markov Decision Process)**.** Let $\mathcal{S} = \{1, \dots, n\}$ 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 these actions 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, $\mathcal{A}_s \triangleq \mathcal{A} = \{1, \dots, k\}$ 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 us call a *policy* $\pi \in \{1, \dots, k\}^n$ the stationary choice of one action for every state. Every policy induces a given 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 average reward criterion* where the aim is to maximize the average reward at each step, v^π is obtained by:

$$v^\pi = \limsup_{N \to \infty} \frac{1}{N} \sum_{i=0}^{N-1} (P^\pi)^i \ r^\pi.$$

However, in this work, the bounds that we derive do not depend on the chosen reward criterion. 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 [Ber07].

Definition 2 (Domination)**.** Given two policies π and π', if $v^{\pi'}(s) \geq v^\pi(s)$ for all states $s \in \mathcal{S}$, 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 $\pi' \succ \pi$.

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 (Proposition 1.3.4 in [Ber07], Volume 2)**.** *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 (Proposition 1.3.4 in [Ber07], Volume 2)**.** *For a given policy π, if $T^\pi = \emptyset$, then π is optimal.*

Based on Propositions 1 and 2 we may define the *Policy Iteration* algorithm to find the optimal policy.

```
Initialization: $\pi_0$, $i = 0$
while $T^{\pi_i} \neq \emptyset$ do
    Select a non-empty and well-defined $U_i \subseteq T^{\pi_i}$
    $\pi_{i+1} = \pi_i \oplus U_i$
    $i \leftarrow i + 1$
end
return $\pi_i$
```

Algorithm 1: Policy Iteration

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.

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 [MS99]).

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 $\pi^{(1)}, \dots, \pi^{(k)}$ is called a *sequence* if $\pi^{(1)} \preceq \cdots \preceq \pi^{(k)}$.

Definition 8 (PI-sequence)**.** We refer to the sequence of policies $\pi_0, \dots, \pi_{m-1}$ 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 1 (Lemma 4 in [MS99])**.** *For any two policies π, π' such that $\pi'(s) = \pi(s)$ for all improvement states $s \in S^\pi$, we have $\pi' \preceq \pi$.*

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

Proposition 3. *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 all states $s \in S^\pi$, either $\pi(s) = \pi'(s)$ or $(s, \pi(s)) \in T^{\pi'}$. Let $U \triangleq \left\{ (s, \pi(s)) : s \in S^\pi \cap S^{\pi'} \text{ and } \pi(s) \neq \pi'(s) \right\}$, 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 1, $\pi'' \preceq \pi$ $(\prec \pi')$ which is a contradiction. □

Note that for $k = 2$, the statement of Proposition 3 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 4. *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 at least d 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 3, 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 4 with:

$$\begin{aligned} \pi' &\longmapsto \pi'' = \pi' \oplus (s, \pi(s)), \\ U &\longmapsto U' = U \setminus (s, \pi'(s)), \end{aligned}$$

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 $\pi^{(d-1)}$. □

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 4.

Corollary 1 (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 $\pi_i \prec \pi \prec \pi_{i+1}$. 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 4. Recall that with Greedy PI, $|U_i|$ always equals $|S^{\pi_i}|$. □

3 A relaxation of the problem

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 3 and Corollary 1 to be ensured by the sequence and the partial order.

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 $\{0, 1\}^n$ 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 3 to hold for O and we require Π to satisfy Corollary 1 as a subsequence of O.

Definition 10 leads to a relaxation of Problem 1. Note that there is a natural way of constructing a pseudo-PI-sequence from any PI-sequence. Of course, Proposition 3 and Corollary 1, that 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.

Relaxation 1. Find the longest possible pseudo-PI-sequence.

In the rest of this paper, we only consider pseudo-PI-sequences and we now derive some of their properties. The following lemma is a direct consequence of Proposition 3.

Lemma 2. *Let* $(\Pi, O, \mathcal{T})$ *be a pseudo-PI-sequence. Then 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 3. It is impossible to from switch $\pi'(s)$ to $\pi(s)$ hence the result. □

When $k = 2$, it is easy to see using Proposition 3 that two policies with exactly the same improvement states cannot exist. When $k > 2$, this is no longer the case. However, using Lemma 1, 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 [MS99]). In the following proposition, we use Proposition 3 to improve this bound to $(k-1)^d$.

Proposition 5. *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 $1 \leq i \leq K$. We show that if the subsequence satisfies Proposition 3, then $K \leq (k-1)^d$. 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 3. 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:

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

Finally, we represent pairs of policies and their improvement sets in a similar way in W through the application:

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

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 2, 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 $1 \leq i \leq d$ 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.

4 Main result: a better upper bound on PI

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

Proof. 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.

Small improvement sets. We consider the small improvement sets T^π such that $|S^\pi| \leq \frac{k-1}{k} \cdot n - f(n)$ with:

$$f(n) \triangleq \sqrt{n \log n}.$$

From Proposition 5, 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:

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

where $X \sim \text{Bin}\left(n, \frac{k-1}{k}\right)$ follows a binomial distribution. Using Hoeffding's inequality [Hoe63], we have:

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

Therefore we have:

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

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

5 The bound is tight for the relaxation

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

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

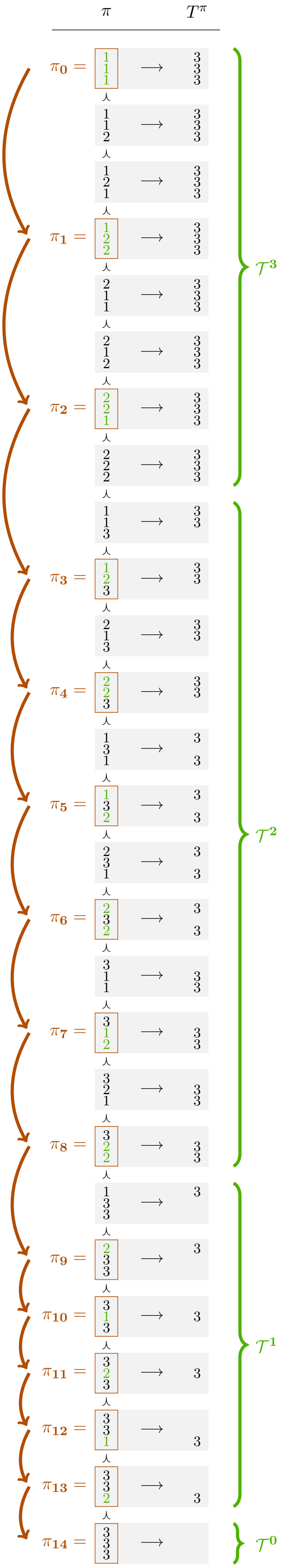

Figure 1: An example of a pseudo-PI-Sequence of size $\frac{k}{k-1} \cdot \frac{k^n}{n} + o\left(\frac{k^n}{n}\right)$ with its supersequence O for $n = k = 3$. Each gray box corresponds to a policy of the supersequence. We represent the improvement sets only through the prospective improving action 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 3.

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 3. 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 3 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 3 is true again.

At this point, we have built a supersequence for our PI-sequence that satisfies the claim of Proposition 3. Let us now select a subsequence Π of O while ensuring Corollary 1 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 1 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:

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

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 1. □

Of course, the lower bound from Theorem 2 only holds for pseudo-PI-sequences which are much less constrained than usual PI-sequences. Indeed, it can be observed that the pseudo-PI-sequence constructed above 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 as discussed in the next section.

6 Alternative approaches

Theorem 2 revealed that future improvements of our bound will require to take into account more of the combinatorial structure of PI-sequences. In this section, we describe two advanced approaches that could lead to new results.

The idea of the first approach is to represent the partial order of the policies of an MDP as an oriented graph whose nodes—the policies—are embedded in an n-dimensional grid and whose edges—that translate the domination relation—only connect neighboring policies (that is, that differ in only one state). In this framework, the structure of the partial order is best described by the Acyclic Unique Sink Orientation of a Grid[1] (Grid AUSOs), an object introduced by Gärtner et al [GMR05] as a generalization of Acyclic Unique Sink Orientations of Cubes [SW01] when $k > 2$. Grid AUSOs accurately characterize the structure of the partial order of any MDP and essentially all necessary conditions we know on PI-sequences originate from this framework. More precisely, it can be described as follows: take a Cartesian grid of dimension n, the number of states of the MDP. A policy can be represented by its action at each state as a vector in $\{1, ..., k\}^n$ and it thereby corresponds to a vertex of the grid. For every neighboring policies π, π', we draw a directed edge from π to π' if $\pi \prec \pi'$ (recall that neighboring policies are always comparable). Thereby, we obtain a directed graph on the grid that is guaranteed to be acyclic and unique sink, i.e. any sub-grid of dimension $d \leq n$ contains a unique vertex of maximum in-degree d [GMR05].

With this structure, PI-steps can be viewed as jumps in the grid as follows: from a policy π_i of the PI-sequence, the out-going links at the corresponding vertex span a sub-grid. In general, the next vertex π_{i+1} chosen by PI can be any vertex of this sub-grid, but in the greedy version, some antipodal vertex to π_i is chosen. This algorithm is also known as the Bottom-Antipodal method in the AUSO framework. Note that it is possible to design Cube AUSOs for which PI takes $\Omega\big(\sqrt{2}^n\big)$ steps [SS05] but to the best of our knowledge, this lower bound cannot be adapted for MDPs.

For $k = 2$, another promising approach was proposed by Hansen & Zwick [Han12] through a relaxation of the AUSO structure. Their idea is to record the policies visited by PI in a binary matrix $\Pi \in \{0, 1\}^{m \times n}$ whose columns correspond to the states of the MDP and whose $(i + 1)^{\text{th}}$ row corresponds to the policy π_i of a PI-sequence. They then formulate the following combinatorial condition on this matrix: for every rows i, j of $\Pi, i < j$, there must exist a column k such that:

$$\Pi_{i,k} \neq \Pi_{i+1,k} = \Pi_{j,k} = \Pi_{j+1,k}. \tag{2}$$

In case $j + 1 > m$, we use the convention that $\Pi_{m+1,k} = \Pi_{m,k}$. Furthermore, the last two rows (labeled $m - 1$ and m) are required to be distinct. Intuitively, at each step $i < m$, at least one change is made to the policy (otherwise we have convergence). Then, at any later step j, one of the changes made at step i must still be there and accepted for the next step. An upper bound on the number of rows of such matrices would immediately translate in a bound on the length of PI-sequences.

0	0	0
1	1	1
0	0	1
0	1	1
0	1	0

0	0	0	0
1	1	1	1
0	0	0	1
0	1	1	1
0	0	1	0
0	1	1	0
0	1	0	0
1	1	0	0

Table 1: Examples of extremal matrices satisfying condition (2) for 3 and 4 columns.

Simulations hint towards the Fibonacci sequence as a possible upper bound for this relaxed problem: for $n \leq 6$, extremal instances achieve $m = F_{n+2}$, the $(n + 2)^{\text{nd}}$ Fibonacci number. If true, this bound would be a significant improvement to ours in the case where $k = 2$. Note that it is possible to build matrices with $m = \Omega\big(\sqrt{2}^n\big)$ rows using similar constructions as for AUSOs. Improving these lower bounds is an interesting challenge in itself.

[1] One could strengthen even a bit further by requiring the Holt-Klee condition as well [HK99, GMR05].

7 Summary

Our contributions can be summarized as follows. First in Theorem 1, we show that Policy Iteration cannot take more than $\frac{k}{k-1} \cdot \frac{k^n}{n} + o\big(\frac{k^n}{n}\big)$ steps to converge, independently of the chosen reward criterion for the MDP. We thereby improve Mansour and Singh's fifteen years old bound. Then in Theorem 2, we show that our bound is optimal for some natural relaxation of the problem. Finally in Section 6, we survey two advanced combinatorial approaches that still could lead to an improvement to our bound.

References

[Alt02] E. Altman. *Applications of Markov Decision Processes in communication networks.* Springer, 2002.

[Ber07] D. P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, Belmont, Massachusetts, 3rd edition, 2007.

[BR11] N. Bäuerle and U. Rieder. *Markov Decision Processes with Applications to Finance.* Springer, 2011.

[CJB14] B. C. Csáji, R. M. Jungers, and V. D. Blondel. Pagerank optimization by edge selection. *Discrete Applied Mathematics*, 169:73–87, 2014.

[Fea10] J. Fearnley. Exponential Lower Bounds for Policy Iteration. *In Proceedings of the 37th International Colloquium on Automata, Languages and Programming, ICALP*, pages 551–562, 2010.

[FHZ11] O. Friedmann, T.D. Hansen, and 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*, 11:283–292, 2011.

[Fri09] 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*, pages 145–156, 2009.

[Fri11] O. Friedmann. A Subexponential Lower Bound for Zadeh's Pivoting Rule for Solving Linear Programs and Games. *In Poceedings of the 14th Conference on Integer Programming and Combinatoral Optimization, IPCO*, pages 192–206, 2011.

[GMR05] B. Gärtner, W. D. Morris, and L. Rüst. *Unique sink orientations of grids.* Springer, 2005.

[Han12] T.D. Hansen. *Worst-case Analysis of Strategy Iteration and the Simplex Method.* PhD thesis, Aarhus University, Science and Technology, Department of Computer Science, 2012.

[HDJ12] R. Hollanders, J.-C. Delvenne, and 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*, pages 5997–6002, 2012.

[HK99] F. Holt and V. Klee. A proof of the strict monotone 4-step conjecture. *Contemporary Mathematics*, 223:201–216, 1999.

[HMZ13] T. D. Hansen, P. B. Miltersen, and U. Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM, JACM*, 60(1):1, 2013.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[Mad98] O. Madani. On constraints on the search path of policy iteration. Technical report, Citeseer, 1998.

[Mey08] S. P. Meyn. *Control techniques for complex networks.* Cambridge University Press, 2008.

[MS99] Y. Mansour and S. Singh. On the Complexity of Policy Iteration. *In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, UAI*, pages 401–408, 1999.

[Put94] M. L. Puterman. *Markov Decision Processes.* John Wiley & Sons, 1994.

[PY13] I. Post and Y. Ye. The simplex method is strongly polynomial for deterministic Markov Decision Processes. *In Proceedings of the 24th ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 1465–1473, 2013.

[Sch13] 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*, pages 386–394, 2013.

[SS05] I. Schurr and T. Szabó. Jumping doesn't help in abstract cubes. *In Proceedings of the 8th Integer Programming and Combinatorial Optimization, IPCO*, pages 225–235, 2005.

[SW01] T. Szabó and E. Welzl. Unique Sink Orientations of Cubes. *In Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science, FOCS*, pages 547–555, 2001.

[Tse90] P. Tseng. Solving H-horizon, Stationary Markov Decision Problems in Time Proportional to log(h). *Operations Research Letters*, 9(4):287-297, 1990.

[Whi93] D. J. White. Survey of Applications of Markov Decision Processes. *The Journal of the Operational Research Society*, 44(11):1073-1096, 1993.

[Ye11] 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):593–603, 2011.