Policy iteration for robust nonstationary Markov decision processes

Saumya Sinha Archis Ghate

September 17, 2015

Abstract

Policy iteration is a well-studied algorithm for solving stationary Markov decision processes (MDPs). It was recently extended to robust stationary MDPs. For robust nonstationary MDPs, however, an "as is" execution of this algorithm is not possible because it would call for an infinite amount of computation in each iteration. We therefore present a policy iteration algorithm for robust nonstationary MDPs, which performs finitely implementable approximate variants of policy evaluation and policy improvement in each iteration. We prove that the sequence of cost-to-go functions produced by this algorithm monotonically converges pointwise to the optimal cost-to-go function; the policies generated converge subsequentially to an optimal policy.

1 Introduction

Policy iteration is a classic algorithm for solving stationary Markov decision processes (MDPs)[1] [9]. The algorithm starts with an initial policy, and, in each iteration, implements two steps: (i) policy evaluation, where the current policy's cost-to-go function is calculated, and (ii) policy improvement, where the current policy is updated to a better policy that optimizes the Q-function of dynamic programming; calculation of this Q-function utilizes the current policy's value function from the first step. When the state- and action-spaces are finite, this algorithm discovers an optimal policy in a finite number of iterations (see Theorem 6.4.2 in [13]). In that case, policy iteration can be viewed as a block-pivoting variant of the simplex algorithm applied to the linear programming (LP) formulation of the MDP. A so-called *simple* version of policy iteration has also been studied in the literature, and in fact, it was recently proven in [16] to exhibit strongly polynomial complexity. In this variation, an action in only one state is updated in each iteration. This single state-action pair is chosen so as to maximize the resulting improvement in the Q-function. This choice is akin to Dantzig's steepest descent pivoting rule for the corresponding simplex method.

Nonstationary MDPs are a generalization of stationary MDPs, where the problem data are no longer assumed to be time-invariant [3, 4, 8]. An asymptotically convergent simple policy iteration algorithm for these MDPs was developed recently in [7]. That paper also analyzed in detail a close connection between this simple policy iteration and an infinite-dimensional simplex method.

In the above MDPs, the state transition probabilities are assumed to be known. Typically, these transition probabilities are estimated statistically from historical data. The resulting estimation errors are ignored in the above MDPs. *Robust* MDPs address this limitation by instead assuming that the transition probabilities are only known to reside in the so-called "uncertainty sets." Roughly

[1] Most MDPs discussed in this paper are finite-state, finite-action and infinite-horizon; we therefore omit such qualifiers for brevity throughout, unless they are essential for clarity.

speaking, the decision-maker then attempts to find a policy that optimizes the worst-case expected cost over all transition probabilities from these uncertainty sets. Detailed analytical treatments of robust MDPs are available in [2, 10, 12].

The classic policy iteration algorithm was extended to the robust case in [10]. For finite-state, finite-action, stationary MDPs, it discovers a robust optimal policy in a finite number of iterations. This result was proven in [10] by invoking Theorem 6.4.2 from [13]. In fact, the policy iteration algorithm in [10] was presented for robust *countable-state* stationary MDPs. Hence, it is, in principle, applicable to *nonstationary* MDPs because, as shown in [7], nonstationary MDPs can be viewed as a special case of countable-state stationary MDPs by appending the states with a time-index. An "as is" execution of this algorithm, however, is not possible for countable-state or for nonstationary MDPs because it would call for infinite computations in both the policy evaluation and policy improvement steps of every iteration. Specifically, an implementable and provably convergent version of policy iteration is currently not available for robust nonstationary MDPs. We develop such an algorithm in this paper.

The key idea in our approach is that it proposes finitely implementable approximations of policy evaluation and simple policy improvement with steepest descent. These approximations are designed adaptively such that the resulting sequence of policies has monotonically decreasing costs. Moreover, the cost-improvement in consecutive iterations is large enough to guarantee convergence to optimality (see [7] for a counterexample of a nonstationary MDP where simply guaranteeing a cost-improvement in each iteration is not enough for convergence to optimality). These statements are made precise in the next two sections. We focus on the simple version of policy iteration to keep notation at a minimum, but our algorithm and proof of convergence can be generalized to a full version without technical difficulty. The only change needed in this full version is that instead of choosing a single period-state pair for updating an action, we select each pair that provides a sufficient improvement.

2 Problem setup and algorithm

Consider a nonstationary MDP with decision epochs $n = 1, 2, \ldots$. At the beginning of each period n, the system occupies a state $s \in \mathcal{S}$, where $\mathcal{S} = \{1, 2, \ldots, S\}$ is a finite set. A decision-maker observes this state and chooses an action $a \in \mathcal{A}$, where $\mathcal{A} = \{1, 2, \ldots, A\}$ is also a finite set. Given that action a was chosen in state s in period n, the system makes a transition to state s' at the beginning of period $n + 1$ with probability $p_n(s'|s, a)$, incurring a nonnegative and bounded cost $0 \leq c_n(s, a, s') \leq c$ for some bound c. This process continues ad infinitum, starting the first period in some initial state $s_1 \in \mathcal{S}$. A (deterministic Markovian) policy π is a mapping that prescribes actions $\pi_n(s)$ in states $s \in \mathcal{S}$ in periods $n \in \mathbb{N}$. The decision-maker's objective is to find a policy that simultaneously (for all $s \in \mathcal{S}$ and all $n \in \mathbb{N}$) minimizes the infinite-horizon discounted expected cost incurred on starting period n in state s. The single-period discount factor is denoted by $0 \leq \lambda < 1$. We note, as an aside, that it is not possible in general to finitely describe the input data needed to completely specify a nonstationary MDP. It is therefore standard in the literature to assume the existence of a "forecast oracle" that, when queried by supplying a positive integer m, returns the cost and probability data for the first m periods. We work in this paper with nonstationary MDPs defined in this manner and refer the reader to [3, 5, 7] for detailed discussions of this issue. Following the language of robust optimization, we will call the problem described in this paragraph a *nominal* nonstationary MDP.

In the above nominal MDP, the transition probabilities $p_n(s'|s,a)$ are assumed to be known. *Robust* nonstationary MDPs account for estimation errors in these transition probabilities by instead assuming that for each state-action pair (s,a) in period n, the (conditional) probability mass function (pmf) $p_n(\cdot|s,a)$ of the next state is only known to lie in some nonempty compact set $\mathcal{P}^a_{n,s}$. This set is called the uncertainty set and it is a subset of the probability simplex $\mathcal{M}(\mathcal{S}) = \{q \in \mathbb{R}^S_+ \mid q_1 + \ldots + q_S = 1\}$. Specifically, robust nonstationary MPDs pursue an adversarial modeling approach where the adversary, also often called "nature", observes the state s in period n as well as the action a chosen there by the decision-maker and then selects a pmf $p_n(\cdot|s,a)$ from the uncertainty set $\mathcal{P}^a_{n,s}$. As per the standard "rectangularity assumption", nature's pmf selection in n,s,a is assumed to be independent of the history of previously visited states and actions and also of the actions chosen in other states (see [10, 12]). The decision-maker's objective is to find a policy that simultaneously (for all $s \in \mathcal{S}$ and all $n \in \mathbb{N}$) minimizes the "worst-case" (with respect to all possible adversarial choices) infinite-horizon discounted expected cost incurred on starting period n in state s.

This finite-state, finite-action robust nonstationary MDP can be equivalently viewed as a robust *stationary* MDP with the *countable* state-space $\mathcal{S} \times \mathbb{N}$ by appending states s with the time-index n. Let $v^*_n(s)$ denote the decision-maker's minimum worst-case cost, against all adversarial policies, on starting period $n \in \mathbb{N}$ in state $s \in \mathcal{S}$. The functions $v^*_n : \mathcal{S} \to \mathbb{R}_+$ are called robust optimal cost-to-go functions, and according to the theory of robust countable-state stationary MDPs from [10], they are unique solutions of the Bellman's equations

$$v^*_n(s) = \min_{a\in\mathcal{A}} \Bigg\{ \underbrace{\max_{p_n(\cdot|s,a)\in\mathcal{P}^a_{n,s}} \Bigg(\sum_{s'\in\mathcal{S}} p_n(s'|s,a) \left[c_n(s,a,s') + \lambda v^*_{n+1}(s') \right] \Bigg)}_{\text{inner problem}} \Bigg\}, \tag{1}$$

for $s \in \mathcal{S}$ and $n \in \mathbb{N}$. Actions that achieve the outer minima in the above equations define a robust optimal policy. Similarly, the infinite-horizon expected discounted cost incurred by implementing a policy π starting in state s in period n is denoted by $v^\pi_n(s)$. These costs-to-go are characterized by the infinite system of equations

$$v^\pi_n(s) = \max_{p_n(\cdot|s,\pi_n(s))\in\mathcal{P}^{\pi_n(s)}_{n,s}} \Bigg(\sum_{s'\in\mathcal{S}} p_n(s'|s,\pi_n(s)) \left[c_n(s,\pi_n(s),s') + \lambda v^\pi_{n+1}(s') \right] \Bigg), \ s \in \mathcal{S}, \ n \in \mathbb{N}. \tag{2}$$

For the robust nonstationary MDP described above, an "as is" execution of robust policy iteration from [10] would roughly amount to the following algorithm. Start with an initial policy π^1. In iteration $k \geq 1$, solve the infinite system of equations in (2) to obtain the cost-to-go function v^{π^k} of policy π^k. This is the policy evaluation step. Then, update policy π^k to a new policy π^{k+1} that prescribes an action from the set

$$\operatorname*{argmin}_{a\in\mathcal{A}} \left\{ \max_{p_n(\cdot|s,a)\in\mathcal{P}^a_{n,s}} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,a) \left[c_n(s,a,s') + \lambda v^{\pi^k}_{n+1}(s') \right] \right) \right\} \tag{3}$$

in each state $s \in \mathcal{S}$ in each period $n \in \mathbb{N}$. This is the policy improvement step. Unfortunately, both these steps require infinite computations, rendering this algorithm unimplementable.

We remedy the above situation by proposing approximate implementations of policy evaluation and simple policy improvement. Specifically, in the policy evaluation step of the kth iteration, the cost-to-go function of policy π^k is approximated by the cost-to-go function of an $m(k)$-horizon

truncation of that policy. In the simple policy improvement step of the kth iteration, an action is updated in state $s(k)$ in period $n(k)$ somewhere in the first $m(k)$-periods via the steepest descent rule applied to this cost-to-go function approximation. In order to guarantee that all actual infinite-horizon costs $v_n^{\pi^{k+1}}(s)$ of the resulting new policy π^{k+1} improve upon the actual infinite-horizon costs $v_n^{\pi^k}(s)$ of the old policy π^k, the truncation-length $m(k)$ is chosen adaptively via an iterative procedure such that the corresponding steepest improvement in the $m(k)$-horizon cost-approximations is large enough. In fact, the discussion in [7] and a counterexample in [6] show that even in the context of nominal nonstationary MDPs, it is not enough (for value convergence to optimality) to simply ensure that π^{k+1} improves upon π^k; it is essential to guarantee that the improvement is sufficiently large. As we shall see in Section 3, our choice of $m(k)$ also carefully handles this delicate issue. The details of this procedure are listed in Algorithm 1 below.

Algorithm 1 Simple policy iteration for robust nonstationary MDPs.

1: Initialize: Set iteration counter $k = 1$. Arbitrarily fix the initial policy π^1 to one that prescribes the first action in $\mathcal{A}$ in every state in every period. Let $n(0) = 1$.

2: **for** iterations $k = 1, 2, 3, \ldots,$ **do**

(a) Set $m = n(k-1)$. Let $m(k) = \infty$ and $\gamma^{k,\infty} = 0$.

Approximate policy evaluation:

(b) Compute the m-horizon approximation $v^{k,m}$ of the cost-to-go function v^{π^k} as

$$v_{m+1}^{k,m}(s) = 0, \ \forall s \in \mathcal{S}, \tag{4}$$

$$v_n^{k,m}(s) = \max_{p_n(\cdot|s,\pi_n^k(s)) \in \mathcal{P}_{n,s}^{\pi_n^k(s)}} \left(\sum_{s' \in \mathcal{S}} p_n(s'|s,\pi_n^k(s)) \left[c_n(s,\pi_n^k(s),s') + \lambda v_{n+1}^{k,m}(s') \right] \right), \ \forall s \in \mathcal{S}, \ n \le m. \tag{5}$$

Approximate simple policy improvement:

(c) Compute the approximate Q-function

$$Q_n^{k,m}(s,a) = \max_{p_n(\cdot|s,a) \in \mathcal{P}_{n,s}^a} \left(\sum_{s' \in \mathcal{S}} p_n(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^{k,m}(s') \right] \right), \ s \in \mathcal{S}, \ a \in \mathcal{A}, \ n \le m. \tag{6}$$

(d) Compute $\gamma_n^{k,m}(s,a) = \lambda^{n-1}\big(Q_n^{k,m}(s,a) - v_n^{k,m}(s)\big)$, for $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $n \le m$. Then calculate the amount of steepest descent

$$\gamma^{k,m} = \min_{\substack{s \in \mathcal{S} \\ a \in \mathcal{A}, a \neq \pi_n^k(s) \\ 1 \le n \le m}} \gamma_n^{k,m}(s,a). \tag{7}$$

(e) If $\gamma^{k,m} < -\lambda^m \frac{c}{1-\lambda}$, set $m(k) = m$, let $(n(k), s(k), a(k))$ be an argmin in (7), and update π^k to π^{k+1} by replacing $\pi_{n(k)}^k(s(k))$ with $a(k)$; else set $m = m + 1$ and go to Step 2(b) above.

3: **end for**

Note that although policy π^k in the kth iteration of this algorithm is "infinite-dimensional", it is described finitely because (i) π^1 is chosen such that it has a finite representation, and (ii) only a single component is changed in each iteration. Consequently, π^k can be stored on a computer.

In addition, we emphasize that each iteration of this algorithm performs only a finite amount of computations. We also make the minor observation that the value of m is initiated at $n(k-1)$ in Step 2(a) of our algorithm, whereas m was initiated at 1 in the simple policy iteration algorithm for nominal nonstationary MDPs in [7]. This initial value of $m = 1$ was inefficient (in the sense that it called for unnecessary additional computations) because $m(k)$ is bounded below by $n(k-1)$ in their nominal case as well as in our robust case. This holds because the steepest descent action in the kth iteration cannot be found for a horizon m shorter than $n(k-1)$ as policies π^{k-1} and π^k prescribe identical actions in the first $n(k-1)-1$ periods.

We prove in the next section that the sequence of costs $v_n^{\pi^k}(s)$ corresponding to the policies π^k produced by this algorithm monotonically converges pointwise to the optimal costs $v_n^*(s)$ as $k \to \infty$. We also establish subsequential convergence of the corresponding policies π^k to an optimal policy. The main ideas in our algorithm and proofs are similar to the aforementioned recent work on simple policy iteration for nominal nonstationary MDPs [7]; the details are modified to accommodate our robust counterpart. For instance, the proofs in [7] for the nominal case thoroughly exploited the close connection between simple policy iteration and an infinite-dimensional simplex algorithm with the steepest descent pivoting rule. We cannot pursue that approach here because robust MDPs do not have an equivalent LP formulation (see [10]).

3 Convergence results

Our two main convergence results in this paper appear toward the end of this section in Theorems 3.7 and 3.8. The proofs of these two theorems utilize several lemmas that we prove next.

The lemma below establishes a simple, fundamental property of Bellman's equations.

Lemma 3.1. *Suppose policy π is not optimal. Then there exist a state $s \in \mathcal{S}$, an action $a \in \mathcal{A}$, and a period $n \in \mathbb{N}$ such that*

$$Q_n^\pi(s,a) = \max_{p_n(\cdot|s,a)\in\mathcal{P}_{n,s}^a} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^\pi(s') \right] \right) < v_n^\pi(s). \tag{8}$$

Proof. Suppose not. Then, for each $n \in \mathbb{N}$ and each $s \in \mathcal{S}$, we have,

$$\max_{p_n(\cdot|s,a)\in\mathcal{P}_{n,s}^a} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^\pi(s') \right] \right) \geq v_n^\pi(s), \ \forall a \in \mathcal{A}.$$

Consequently, for each $n \in \mathbb{N}$ and each $s \in \mathcal{S}$, we obtain,

$$\begin{aligned} v_n^\pi(s) &= \max_{p_n(\cdot|s,\pi_n(s))\in\mathcal{P}_{n,s}^{\pi_n(s)}} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,\pi_n(s)) \left[c_n(s,\pi_n(s),s') + \lambda v_{n+1}^\pi(s') \right] \right) \\ &\geq \min_{a\in\mathcal{A}} \left\{ \max_{p_n(\cdot|s,a)\in\mathcal{P}_{n,s}^a} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^\pi(s') \right] \right) \right\} \geq v_n^\pi(s). \end{aligned}$$

This shows that, for each $n \in \mathbb{N}$ and each $s \in \mathcal{S}$,

$$v_n^\pi(s) = \min_{a\in\mathcal{A}} \left\{ \max_{p_n(\cdot|s,a)\in\mathcal{P}_{n,s}^a} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^\pi(s') \right] \right) \right\}.$$

This shows that the cost-to-go functions v_n^π satisfy Bellman's equations. Then π must be optimal. This is a contradiction. □

Within each iteration, our algorithm computes an m-horizon approximation $v^{k,m}$ to the true infinite-horizon cost-to-go function v^{π^k} of the policy π^k. The lemma below provides bounds for the quality of this approximation.

Lemma 3.2. *The approximation $v^{k,m}$ of v^{π^k} in Step 2(b) of Algorithm 1 satisfies*

$$v_n^{k,m}(s) \le v_n^{\pi^k}(s) \le v_n^{k,m}(s) + \lambda^{m+1-n}\frac{c}{1-\lambda},\ \forall s \in \mathcal{S},\ n = 1, 2, \ldots, m+1. \tag{9}$$

Proof. We prove the claim by backward induction on $n = m+1, m, \ldots, 1$.

Since the costs are nonnegative and bounded above by c, we know that v^{π^k} satisfies $0 \le v_{m+1}^{\pi^k}(s) \le c/(1-\lambda)$. Also, $v_{m+1}^{k,m}(s) = 0$ for all $s \in \mathcal{S}$ by (4). So, for $n = m+1$, we trivially have,

$$v_{m+1}^{k,m}(s) \le v_{m+1}^{\pi^k}(s) \le v_{m+1}^{k,m}(s) + \lambda^{m+1-n}\frac{c}{1-\lambda},\ \forall s \in \mathcal{S}.$$

Now, assume, as the inductive hypothesis, that the claim is true for $n+1$. That is,

$$v_{n+1}^{k,m}(s') \le v_{n+1}^{\pi^k}(s') \le v_{n+1}^{k,m}(s') + \lambda^{m-n}\frac{c}{1-\lambda},\ \forall s' \in \mathcal{S}.$$

After multiplying each term by λ and then adding $c_n(s,a,s')$ to all terms, this implies that

$$c_n(s,a,s') + \lambda v_{n+1}^{k,m}(s') \le c_n(s,a,s') + \lambda v_{n+1}^{\pi^k}(s') \le c_n(s,a,s') + \lambda v_{n+1}^{k,m}(s') + \lambda^{m+1-n}\frac{c}{1-\lambda},$$

for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. Consequently, for the specific actions $\pi_n^k(s)$ prescribed by the policy π^k, we have,

$$c_n(s,\pi_n^k(s),s') + \lambda v_{n+1}^{k,m}(s') \le c_n(s,\pi_n^k(s),s') + \lambda v_{n+1}^{\pi^k}(s') \le c_n(s,\pi_n^k(s),s') + \lambda v_{n+1}^{k,m}(s') + \lambda^{m+1-n}\frac{c}{1-\lambda},$$

for all $s, s' \in \mathcal{S}$. Now, for a fixed $s \in \mathcal{S}$, consider any pmf $p_n(\cdot|s,\pi_n^k(s)) \in \mathcal{P}_{n,s}^{\pi_n^k(s)}$. By multiplying the above inequalities with this pmf and then adding over all $s' \in \mathcal{S}$, we obtain,

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}} p_n(s'|s,\pi_n^k(s)) \left[c_n(s,\pi_n^k(s),s') + \lambda v_{n+1}^{k,m}(s')\right] \le \sum_{s'\in\mathcal{S}} p_n(s'|s,\pi_n^k(s)) \left[c_n(s,\pi_n^k(s),s') + \lambda v_{n+1}^{\pi^k}(s')\right] \\
&\le \sum_{s'\in\mathcal{S}} p_n(s'|s,\pi_n^k(s)) \left[c_n(s,\pi_n^k(s),s') + \lambda v_{n+1}^{k,m}(s')\right] + \lambda^{m+1-n}\frac{c}{1-\lambda}.
\end{aligned}$$

Then, by taking the maximum of each side of these inequalities over all such pmfs in $\mathcal{P}_{n,s}^{\pi_n^k(s)}$, we obtain,

$$\begin{aligned}
&\max_{p_n(\cdot|s,\pi_n^k(s))\in\mathcal{P}_{n,s}^{\pi_n^k(s)}} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,\pi_n^k(s)) \left[c_n(s,\pi_n^k(s),s') + \lambda v_{n+1}^{k,m}(s')\right]\right) \\
&\le \max_{p_n(\cdot|s,\pi_n^k(s))\in\mathcal{P}_{n,s}^{\pi_n^k(s)}} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,\pi_n^k(s)) \left[c_n(s,\pi_n^k(s),s') + \lambda v_{n+1}^{\pi^k}(s')\right]\right)
\end{aligned}$$

$$\leq \max_{p_n(\cdot|s,\pi_n^k(s))\in\mathcal{P}_{n,s}^{\pi_n^k(s)}} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,\pi_n^k(s)) \left[c_n(s,\pi_n^k(s),s') + \lambda v_{n+1}^{k,m}(s') \right] \right) + \lambda^{m+1-n} \frac{c}{1-\lambda}.$$

These maxima preserve the order of the earlier inequalities because of the following property: if one function is everywhere smaller than another function, then the maximum of the first function is smaller than the maximum of the second function provided that the maxima are taken over identical sets. Then, by (2) and (5), we have,

$$v_n^{k,m}(s) \leq v_n^{\pi^k}(s) \leq v_n^{k,m}(s) + \lambda^{m+1-n} \frac{c}{1-\lambda}, \ \forall s \in \mathcal{S}.$$

This restores the inductive hypothesis and completes the proof by induction. □

Step 2 of the algorithm iteratively increases the value of the approximating horizon's length m until a horizon that guarantees a large enough improvement is discovered. The algorithm thus runs the risk of being caught in an infinite loop, wherein it never discovers such a horizon. The next lemma tackles this issue.

Lemma 3.3. *Step 2 of Algorithm 1 terminates at a finite value of m if and only if policy π^k is not optimal.*

Proof. Suppose π^k is not optimal. Then, by Lemma 3.1, there exist a period $n \in \mathbb{N}$, $s \in \mathcal{S}$, and an action $a \in \mathcal{A}$ such that $Q_n^{\pi^k}(s,a) < v_n^{\pi^k}(s)$. Thus, let $\epsilon = v_n^{\pi^k}(s) - Q_n^{\pi^k}(s,a) > 0$. Then, by Lemma 3.2, we have,

$$\begin{aligned}
\epsilon &= v_n^{\pi^k}(s) - Q_n^{\pi^k}(s,a) = v_n^{\pi^k}(s) - \max_{p_n(\cdot|s,a)\in\mathcal{P}_{n,s}^a} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^{\pi^k}(s') \right] \right) \\
&\leq v_n^{k,m}(s) + \lambda^{m+1-n} \frac{c}{1-\lambda} - \max_{p_n(\cdot|s,a)\in\mathcal{P}_{n,s}^a} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^{k,m}(s') \right] \right) \\
&= v_n^{k,m}(s) + \lambda^{m+1-n} \frac{c}{1-\lambda} - Q_n^{k,m}(s,a) = \lambda^{1-n} \left(\lambda^m \frac{c}{1-\lambda} - \lambda^{n-1} \big(Q_n^{k,m}(s,a) - v_n^{k,m}(s) \big) \right) \\
&\leq \lambda^{1-n} \left(\lambda^m \frac{c}{1-\lambda} - \gamma^{k,m} \right).
\end{aligned}$$

where the last inequality holds by the definition of $\gamma^{k,m}$ in Step 2(d) of the algorithm. This inequality yields $\lambda^{1-n}\gamma^{k,m} \leq \lambda^{m+1-n}\frac{c}{1-\lambda} - \epsilon$. For a sufficiently large m, we have that $\lambda^{m+1-n}\frac{c}{1-\lambda} < \epsilon/2$ because $0 \leq \lambda < 1$. Therefore, for any such large m, we have, $\lambda^{m+1-n}\frac{c}{1-\lambda} - \epsilon < -\epsilon/2 < -\lambda^{m+1-n}\frac{c}{1-\lambda}$. Thus, we obtain, $\lambda^{1-n}\gamma^{k,m} < -\lambda^{m+1-n}\frac{c}{1-\lambda}$, that is, $\gamma^{k,m} < -\lambda^m\frac{c}{1-\lambda}$. Thus, if the policy π^k is not optimal, there exists a large enough m for which the stopping condition in Step 2(e) of the algorithm is satisfied, and Step 2 terminates finitely.

Conversely, suppose policy π^k is optimal, and Step 2 of the algorithm terminates for some $m(k)$. Then, $\gamma^{k,m(k)} + \lambda^{m(k)}\frac{c}{1-\lambda} < 0$. That is, $\gamma_{n(k)}^{k,m(k)}(s(k),a(k)) + \lambda^{m(k)}\frac{c}{1-\lambda} < 0$, where $(n(k),s(k),a(k))$ is the argmin in (7) with $a(k) \neq \pi_{n(k)}^k(s(k))$. That is, $\lambda^{1-n(k)}\gamma_{n(k)}^{k,m(k)}(s(k),a(k)) + \lambda^{m(k)+1-n(k)}\frac{c}{1-\lambda} < 0$. Then, by the definition of $\gamma_{n(k)}^{k,m(k)}(s(k),a(k))$, this implies that $Q_{n(k)}^{k,m(k)}(s(k),a(k)) - v_{n(k)}^{k,m(k)}(s(k)) + \lambda^{m(k)+1-n(k)}\frac{c}{1-\lambda} < 0$. Then, by using the definition of $Q_{n(k)}^{k,m(k)}(s(k),a(k))$ as in (6), we get,

$$\max_{p_{n(k)}(\cdot|s(k),a(k))\in\mathcal{P}_{n(k),s(k)}^{a(k)}} \left(\sum_{s'\in\mathcal{S}} p_{n(k)}(s'|s(k),a(k)) \left[c_n(s(k),a(k),s') + \lambda v_{n(k)+1}^{k,m(k)}(s') \right] \right) -$$

$$v_{n(k)}^{k,m(k)}(s(k)) + \lambda^{m(k)+1-n(k)}\frac{c}{1-\lambda} < 0.$$

By applying Lemma 3.2, the above strict inequality implies that $Q_{n(k)}^{\pi^k}(s(k), a(k)) < v_{n(k)}^{\pi^k}(s(k))$. In other words, the optimal cost-to-go function v^{π^k} violates Bellman's equation (1) in state $s(k)$ in period $n(k)$. But this contradicts the optimality of π^k. This completes the proof of the lemma. □

The lemma implies that if π^k is optimal, then Step 2 of the algorithm never terminates. One subtlety here is that the algorithm therefore cannot tell that it has discovered an optimal policy. As explained in more detail in [7], this, however, is not a limitation of our algorithm. Rather, it is rooted in a fundamental property of nonstationary sequential decision problems that optimality cannot be verified, in general, with finite computation.

The next lemma shows that, despite the approximations they employ, our policy evaluation and simple policy improvement steps produce a sequence of policies with nonincreasing cost-to-go functions.

Lemma 3.4. *Suppose policy π^k is not optimal. Then $v_n^{\pi^{k+1}}(s) \leq v_n^{\pi^k}(s)$ for all periods $n \in \mathbb{N}$ and all states $s \in \mathcal{S}$, with this inequality being strict when $n = n(k)$ and $s = s(k)$. Furthermore, $v_{n(k)}^{\pi^{k+1}}(s(k)) - v_{n(k)}^{\pi^k}(s(k)) \leq \lambda^{1-n(k)}\left(\lambda^{m(k)}\frac{c}{1-\lambda} + \gamma^{k,m(k)}\right)$.*

Proof. Since policy π^k is not optimal, Step 2 of the algorithm terminates at some $m(k)$ by Lemma 3.3. Also, policies π^{k+1} and π^k differ only in the actions they prescribe in period $n(k) \leq m(k)$ in state $s(k)$. Consequently, $v_n^{\pi^{k+1}}(s) = v_n^{\pi^k}(s)$ for all $s \in \mathcal{S}$ and all $n > n(k)$. Similarly, $v_{n(k)}^{\pi^{k+1}}(s) = v_{n(k)}^{\pi^k}(s)$ for all $s(k) \neq s \in \mathcal{S}$. Moreover, (2) implies that

$$\begin{aligned}
v_{n(k)}^{\pi^{k+1}}(s(k)) &= \max_{p_n(\cdot|s(k),a(k)) \in \mathcal{P}_{n(k),s(k)}^{a(k)}} \left(\sum_{s' \in \mathcal{S}} p_{n(k)}(s'|s(k),a(k)) \left[c_{n(k)}(s(k),a(k),s') + \lambda v_{n(k)+1}^{\pi^{k+1}}(s') \right] \right) \\
&= \max_{p_n(\cdot|s(k),a(k)) \in \mathcal{P}_{n(k),s(k)}^{a(k)}} \left(\sum_{s' \in \mathcal{S}} p_{n(k)}(s'|s(k),a(k)) \left[c_{n(k)}(s(k),a(k),s') + \lambda v_{n(k)+1}^{\pi^k}(s') \right] \right) \\
&\leq \max_{p_n(\cdot|s(k),a(k)) \in \mathcal{P}_{n(k),s(k)}^{a(k)}} \left(\sum_{s' \in \mathcal{S}} p_{n(k)}(s'|s(k),a(k)) \Bigg[c_{n(k)}(s(k),a(k),s') + \right. \\
&\left. \lambda \left(v_{n(k)+1}^{k,m(k)}(s') + \lambda^{m(k)-n(k)}\frac{c}{1-\lambda} \right) \Bigg] \right) \\
&= \max_{p_n(\cdot|s(k),a(k)) \in \mathcal{P}_{n(k),s(k)}^{a(k)}} \left(\sum_{s' \in \mathcal{S}} p_{n(k)}(s'|s(k),a(k)) \left[c_{n(k)}(s(k),a(k),s') + \lambda v_{n(k)+1}^{k,m(k)}(s') \right] \right) + \\
&\lambda^{m(k)+1-n(k)}\frac{c}{1-\lambda}, \\
&= Q_{n(k)}^{k,m(k)}(s(k),a(k)) + \lambda^{m(k)+1-n(k)}\frac{c}{1-\lambda} \\
&= \lambda^{1-n(k)}\gamma^{k,m(k)} + v_{n(k)}^{k,m(k)}(s(k)) + \lambda^{m(k)+1-n(k)}\frac{c}{1-\lambda} \\
&\leq \lambda^{1-n(k)}\gamma^{k,m(k)} + v_{n(k)}^{\pi^k}(s(k)) + \lambda^{m(k)+1-n(k)}\frac{c}{1-\lambda} < v_{n(k)}^{\pi^k}(s(k)).
\end{aligned}$$

Here, the first inequality follows by Lemma 3.2, the penultimate equality holds by the definition in (6) of $Q_{n(k)}^{k,m(k)}(s(k),a(k))$, the last equality holds by the definition of $\gamma^{k,m(k)}$, the penultimate

inequality holds by Lemma 3.2, and finally, the strict inequality follows by the stopping condition in Step 2(e). In summary, we have shown that $v^{\pi^{k+1}}_{n(k)}(s(k)) < v^{\pi^k}_{n(k)}(s(k))$ and that $v^{\pi^{k+1}}_{n(k)}(s(k)) - v^{\pi^k}_{n(k)}(s(k)) < \lambda^{1-n(k)}\left(\lambda^{m(k)}\frac{c}{1-\lambda} + \gamma^{k,m(k)}\right)$. Now we complete the rest of the proof by induction on $n = n(k), n(k)-1, \ldots, 1$. To start off this induction process, we note that the argument thus far has established that $v^{\pi^{k+1}}_n(s) \leq v^{\pi^k}_n(s)$, for all $s \in \mathcal{S}$ and $n = n(k)$. Now, as the inductive hypothesis, suppose that $v^{\pi^{k+1}}_n(s) \leq v^{\pi^k}_n(s)$, for all $s \in \mathcal{S}$ and some $n \leq n(k)$. Then, from (2), we have, for each $s \in \mathcal{S}$, that

$$\begin{aligned}
v^{\pi^{k+1}}_{n-1}(s) &= \max_{p_{n-1}(\cdot|s,\pi^{k+1}_{n-1}(s)) \in \mathcal{P}^{\pi^{k+1}_{n-1}(s)}_{n-1,s}} \left(\sum_{s' \in \mathcal{S}} p_{n-1}(s'|s,\pi^{k+1}_{n-1}(s)) \left[c_{n-1}(s,\pi^{k+1}_{n-1}(s),s') + \lambda v^{\pi^{k+1}}_n(s') \right] \right) \\
&= \max_{p_{n-1}(\cdot|s,\pi^{k}_{n-1}(s)) \in \mathcal{P}^{\pi^{k}_{n-1}(s)}_{n-1,s}} \left(\sum_{s' \in \mathcal{S}} p_{n-1}(s'|s,\pi^{k}_{n-1}(s)) \left[c_{n-1}(s,\pi^{k}_{n-1}(s),s') + \lambda v^{\pi^{k+1}}_n(s') \right] \right) \\
&\leq \max_{p_{n-1}(\cdot|s,\pi^{k}_{n-1}(s)) \in \mathcal{P}^{\pi^{k}_{n-1}(s)}_{n-1,s}} \left(\sum_{s' \in \mathcal{S}} p_{n-1}(s'|s,\pi^{k}_{n-1}(s)) \left[c_{n-1}(s,\pi^{k}_{n-1}(s),s') + \lambda v^{\pi^{k}}_n(s') \right] \right) \\
&= v^{\pi^k}_{n-1}(s).
\end{aligned}$$

Here, the second equality holds because π^{k+1} and π^k prescribe identical actions in all states in period $n-1$, the inequality holds by the inductive hypothesis, and the last equality follows by the definition of $v^{\pi^k}_{n-1}(s)$. This restores the inductive hypothesis and completes our proof. □

Lemma 3.5. *We have that* $\left[\lambda^{m(k)}\frac{c}{1-\lambda} + \gamma^{k,m(k)}\right] \to 0$ *as* $k \to \infty$.

Proof. Observe that if π^k is optimal for any k, then, by Lemma 3.3, Step 2 of the algorithm does not terminate finitely; hence $\left[\lambda^{m(k)}\frac{c}{1-\lambda} + \gamma^{k,m(k)}\right] = 0$ because the algorithm is initiated with $m(k) = \infty$ and $\gamma^{k,\infty} = 0$ and hence the claim holds. Now suppose that π^k is not optimal for any k. The algorithm thus produces a sequence of solutions v^{π^k}. Now define, for each k, $f^k = \sum_{n \in \mathbb{N}} \sum_{s \in \mathcal{S}} \lambda^{n-1} v^{\pi^k}_n(s)$ and let $\delta^k = f^{k+1} - f^k$. Then, since the sum f^k is finite for all k, we have,

$$\begin{aligned}
\delta^k &= \sum_{n \in \mathbb{N}} \sum_{s \in \mathcal{S}} \lambda^{n-1} \left[v^{\pi^{k+1}}_n(s) - v^{\pi^k}_n(s) \right] \leq \lambda^{n(k)-1} \left[v^{\pi^{k+1}}_{n(k)}(s(k)) - v^{\pi^k}_{n(k)}(s(k)) \right] \\
&\leq \gamma^{k,m(k)} + \lambda^{m(k)} \frac{c}{1-\lambda} < 0,
\end{aligned}$$

where the first inequality follows since every term in the sum is non-positive, the second inequality holds by the second claim in Lemma 3.4, and the last inequality follows from the stopping condition in Step 2(e) of the algorithm. That is, f^k is a nonnegative decreasing sequence of real numbers, hence it converges. This implies that $\delta^k \to 0$ as $k \to \infty$. Since $\delta^k \leq \gamma^{k,m(k)} + \lambda^{m(k)}\frac{c}{1-\lambda} < 0$ for all k, the second claim holds. □

The sequence of approximating horizons $m(k)$ in not monotonically increasing in k. The next lemma shows that it nevertheless diverges to infinity as $k \to \infty$. This also implies that the amount of steepest descent improvement converges to zero.

Lemma 3.6. *The sequence* $m(k) \to \infty$ *as* $k \to \infty$. *Also,* $\gamma^{k,m(k)} \to 0$ *as* $k \to \infty$.

Proof. Identical to the proof of Lemma 5.7 in [7] hence omitted. □

Theorem 3.7 (Value Convergence)**.** *The sequence of cost-to-go functions produced by Algorithm 1 converges pointwise to the optimal cost-to-go function. That is,*

$$\lim_{k\to\infty} v_n^{\pi^k}(s) = v_n^*(s) \quad \textit{for all} \quad n \in \mathbb{N}, s \in \mathcal{S}. \tag{10}$$

Proof. Policies for the nonstationary MDP lie in the strategy space $\Phi = \prod_{n=1}^{\infty} \mathcal{A}^S \subset \prod_{n=1}^{\infty} \mathbb{R}^S$, which is compact in the metrizable product topology by Tychonoff's product theorem (see Theorem 2.61 on page 52 of [1]). In fact, $\rho(\cdot,\cdot)$ defined by

$$\rho(\pi, \tilde{\pi}) = \sum_{n=1}^{\infty} \frac{1}{2^n} \left(\frac{d(\pi_n, \tilde{\pi}_n)}{1 + d(\pi_n, \tilde{\pi}_n)} \right)$$

where $d(\cdot,\cdot)$ is the Euclidean metric on $\mathbb{R}^S$, is an example of a metric which induces the product topology on $\prod_{n=1}^{\infty} \mathbb{R}^S$ (see Theorem 3.36 on page 89 of [1]). Further, the cost-to-go functions lie in the set $V = \left\{ v \in \prod_{n=1}^{\infty} \mathbb{R}^S : 0 \le v_n(s) \le \frac{c}{1-\lambda}, n \in \mathbb{N}, s \in \mathcal{S} \right\}$. Again, by Tychonoff's theorem, $V \subset \prod_{n=1}^{\infty} \mathbb{R}^S$ is compact in the metrizable product topology. Since Φ is compact, the sequence of policies π^k has a convergent subsequence π^{k_i}. Let $\bar{\pi}$ be the limit of this sequence. By the same reasoning, the corresponding sequence of cost-to-go functions $v^{\pi^{k_i}}$ has a convergent subsequence, $v^{\pi^{k_{i_j}}}$, whose limit is, say, $\bar{v}$.

We first show that $\bar{v} = v^{\bar{\pi}}$, that is, $\bar{v}$ is the cost-to-go function corresponding to the policy $\bar{\pi}$.

Consider any $n \in \mathbb{N}$ and $s \in \mathcal{S}$. Then, for any j,

$$v_n^{\pi^{k_{i_j}}}(s) - \max_{p(\cdot|s,\pi_n^{k_{i_j}}(s)) \in \mathcal{P}_{n,s}^{\pi_n^{k_{i_j}}(s)}} \sum_{s' \in \mathcal{S}} p(s'|s, \pi_n^{k_{i_j}}(s)) \left[c_n(s, \pi_n^{k_{i_j}}(s), s') + \lambda v_{n+1}^{k_{i_j}}(s') \right] = 0.$$

Now, since $\pi^{k_{i_j}} \to \bar{\pi}$ in the product topology, we have that $\pi_n^{k_{i_j}}(s) \to \bar{\pi}_n(s)$ as $j \to \infty$. Since $\mathcal{A}$ is a finite set, this implies that there exists a number $J(n,s)$ such that for all $j \ge J(n,s)$, $\pi_n^{k_{i_j}}(s) = \bar{\pi}_n(s)$. Hence, for all $j \ge J(n,s)$, the sets $\mathcal{P}_{n,s}^{\pi_n^{k_{i_j}}(s)}$ and $\mathcal{P}_{n,s}^{\bar{\pi}_n(s)}$ are identical, and we have,

$$v_n^{\pi^{k_{i_j}}}(s) - \max_{p(\cdot|s,\bar{\pi}_n(s)) \in \mathcal{P}_{n,s}^{\bar{\pi}_n(s)}} \sum_{s' \in \mathcal{S}} p(s'|s, \bar{\pi}_n(s)) \left[c_n(s, \bar{\pi}_n(s), s') + \lambda v_{n+1}^{k_{i_j}}(s') \right] = 0. \tag{11}$$

For each fixed $p(\cdot|s,\bar{\pi}_n(s)) \in \mathcal{P}_{n,s}^{\bar{\pi}_n(s)}$, we have,

$$v_n^{\pi^{k_{i_j}}}(s) - \sum_{s' \in \mathcal{S}} p(s'|s, \bar{\pi}_n(s)) \left[c_n(s, \bar{\pi}_n(s), s') + \lambda v_{n+1}^{k_{i_j}}(s') \right] \ge 0.$$

Taking limits as $j \to \infty$, this yields,

$$\bar{v}_n(s) - \sum_{s' \in \mathcal{S}} p(s'|s, \bar{\pi}_n(s)) \left[c_n(s, \bar{\pi}_n(s), s') + \lambda \bar{v}_{n+1}(s') \right] \ge 0,$$

for all $p(\cdot|s,\bar{\pi}_n(s)) \in \mathcal{P}_{n,s}^{\bar{\pi}_n(s)}$. This implies

$$\bar{v}_n(s) - \max_{p(\cdot|s,\bar{\pi}_n(s))\in\mathcal{P}_{n,s}^{\bar{\pi}_n(s)}} \sum_{s'\in\mathcal{S}} p(s'|s,\bar{\pi}_n(s)) \left[c_n(s,\bar{\pi}_n(s),s') + \lambda \bar{v}_{n+1}(s')\right] \geq 0. \tag{12}$$

Now, we show that inequality (12) cannot be strict. For each $j \geq J(n,s)$, let $p^{k_{i_j}}(\cdot|s,\bar{\pi}_n(s))$ be an argmax in (11). Then, we rewrite (11) as

$$v_n^{\pi\,k_{i_j}}(s) - \sum_{s'\in\mathcal{S}} p^{k_{i_j}}(s'|s,\bar{\pi}_n(s)) \left[c_n(s,\bar{\pi}_n(s),s') + \lambda v_{n+1}^{k_{i_j}}(s')\right] = 0.$$

Note that as $\mathcal{P}_{n,s}^{\bar{\pi}_n(s)}$ is a compact subset of $\mathbb{R}^S$, the sequence $\{p^{k_{i_j}}(\cdot|s,\bar{\pi}_n(s)),\ j \geq J(n,s)\}$ has a convergent subsequence $p^{k_{i_{j_l}}}$. Let $\bar{p}(\cdot|s,\bar{\pi}_n(s))$ be the limit of this subsequence. For each l, we have,

$$v_n^{k_{i_{j_l}}}(s) - \sum_{s'\in\mathcal{S}} p^{k_{i_{j_l}}}(s'|s,\bar{\pi}_n(s)) \left[c_n(s,\bar{\pi}_n(s),s') + \lambda v_{n+1}^{k_{i_{j_l}}}(s')\right] = 0.$$

Taking limits as $l \to \infty$, this gives us

$$\bar{v}_n(s) - \sum_{s'\in\mathcal{S}} \bar{p}(s'|s,\bar{\pi}_n(s)) \left[c_n(s,\bar{\pi}_n(s),s') + \lambda \bar{v}_{n+1}(s')\right] = 0.$$

Hence, the inequality in (12) must be an equality, and we have

$$\bar{v}_n(s) - \max_{p(\cdot|s,\bar{\pi}_n(s))\in\mathcal{P}_{n,s}^{\bar{\pi}_n(s)}} \sum_{s'\in\mathcal{S}} p(s'|s,\bar{\pi}_n(s)) \left[c_n(s,\bar{\pi}_n(s),s') + \lambda \bar{v}_{n+1}(s')\right] = 0. \tag{13}$$

Since the above is true for all (n,s), we have proved that the limiting cost-to-go function $\bar{v}$ is the evaluation of the limiting policy $\bar{\pi}$, and we denote it by $v^{\bar{\pi}}$.

We now show, by contradiction, that the limiting policy $\bar{\pi}$ must be optimal. Suppose $\bar{\pi}$ is not optimal. Then, by Lemma 3.1, there exists a period n, a state s and an action a such that

$$\begin{aligned} 0 < \epsilon &= v_n^{\bar{\pi}}(s) - Q_n^{\bar{\pi}}(s,a) \\ &= v_n^{\bar{\pi}}(s) - \max_{p_n(\cdot|s,a)\in\mathcal{P}_{n,s}^a} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^{\bar{\pi}}(s')\right] \right). \end{aligned} \tag{14}$$

For any j, let $p_n^{k_{i_j}}(\cdot|s,a)$ be the argmax in

$$Q_n^{\pi\,k_{i_j}}(s,a) = \max_{p_n(\cdot|s,a)\in\mathcal{P}_{n,s}^a} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^{\pi\,k_{i_j}}(s')\right] \right).$$

As before, the sequence $p_n^{k_{i_j}}(\cdot|s,a)$ has a convergent subsequence $p_n^{k_{i_{j_l}}}(\cdot|s,a)$. Let $\bar{p}_n(\cdot|s,a)$ be the limit of this subsequence. Then, we have,

$$\begin{aligned} \lim_{l\to\infty} \left(v_n^{\pi\,k_{i_{j_l}}}(s) - Q_n^{\pi\,k_{i_{j_l}}}(s,a) \right) &= \lim_{l\to\infty} \left(v_n^{k_{i_{j_l}}}(s) - \sum_{s'\in\mathcal{S}} p_n^{k_{i_{j_l}}}(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^{\pi\,k_{i_{j_l}}}(s')\right] \right) \\ &= v_n^{\bar{\pi}}(s) - \sum_{s'\in\mathcal{S}} \bar{p}_n(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^{\bar{\pi}}(s')\right] \end{aligned}$$

$$\geq v_n^{\bar{\pi}}(s) - \max_{p_n(\cdot|s,a)\in\mathcal{P}^a_{n,s}} \left(\sum_{s'\in\mathcal{S}} p_n(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^{\bar{\pi}}(s') \right] \right)$$
$$= \epsilon.$$

Then, there exists an integer L such that for $l \geq L$,

$$\epsilon/2 < v_n^{\pi^{k_{i_{j_l}}}}(s) - Q_n^{\pi^{k_{i_{j_l}}}}(s,a) = v_n^{\pi^{k_{i_{j_l}}}}(s) - \max_{p_n(\cdot|s,a)\in\mathcal{P}^a_{n,s}} \left(\sum_{s'\in\mathcal{S}} p_n^{k_{i_{j_l}}}(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^{\pi^{k_{i_{j_l}}}}(s') \right] \right).$$

Since $m(k) \to \infty$, we have that for large enough l, $m(k_{i_{j_l}}) \geq n$. Then, applying Lemma 3.2 gives us that

$$\begin{aligned}
\epsilon/2 &< v_n^{k_{i_{j_l}},m(k_{i_{j_l}})}(s) + \lambda^{m(k_{i_{j_l}})+1-n}\frac{c}{1-\lambda} \\
&\quad - \max_{p_n(\cdot|s,a)\in\mathcal{P}^a_{n,s}} \left(\sum_{s'\in\mathcal{S}} p_n^{k_{i_{j_l}}}(s'|s,a) \left[c_n(s,a,s') + \lambda v_{n+1}^{k_{i_{j_l}},m(k_{i_{j_l}})}(s') \right] \right) \\
&\leq \lambda^{m(k_{i_{j_l}})+1-n}\frac{c}{1-\lambda} - \lambda^{1-n}\gamma^{k_{i_{j_l}},m(k_{i_{j_l}})} = \lambda^{1-n}\left(\lambda^{m(k_{i_{j_l}})}\frac{c}{1-\lambda} - \gamma^{k_{i_{j_l}},m(k_{i_{j_l}})} \right).
\end{aligned}$$

But this contradicts the fact from Lemma 3.6 that both $\lambda^{m(k_{i_{j_l}})}\frac{c}{1-\lambda}$ and $\gamma^{k_{i_{j_l}},m(k_{i_{j_l}})}$ converge to zero as $l \to \infty$. Hence, our assumption is false and the limiting policy $\bar{\pi}$ must be optimal.

We remark that so far we have only proven that v^{π^k} converges subsequentially to the optimal value function v^*. But from Lemma 3.4, we know that each component $v_n^{\pi^k}(s)$ is a nonincreasing sequence of nonnegative real numbers, and therefore must converge. This combined with the subsequential convergence proves that $\lim_{k\to\infty} v_n^{\pi^k}(s) = v_n^*(s)$ for all $s \in \mathcal{S}$ and $n \in \mathbb{N}$. □

Theorem 3.8 (Policy Convergence)**.** *For any $\epsilon > 0$, there exists an iteration counter k_ϵ such that $\rho(\pi^k, \pi^{*k}) < \epsilon$ for some optimal policy π^{*k}, for all $k \geq k_\epsilon$. In fact, if the MDP has a unique optimal policy π^*, then $\lim_{k\to\infty} \pi^k = \pi^*$. Further, for every period n, there exists an iteration counter K_n such that for all $k \geq K_n$, actions $\pi_m^k(s)$ are optimal for the robust non-stationary MDP in all states $s \in \mathcal{S}$ and all periods $m \leq n$.*

Proof. We prove the first claim by contradiction. Suppose this is not true. Then, there exists an $\epsilon > 0$ and a subsequence π^{k_i} of π^k such that $\rho(\pi^{k_i}, \pi) > \epsilon$ for all optimal policies π, for all $i \in \mathbb{N}$. Since the space of all policies is compact, the sequence π^{k_i} has a convergent subsequence $\pi^{k_{i_j}}$, whose limit is, say, $\bar{\pi}$. Then, there exists an integer J such that $\rho(\pi^{k_{i_j}}, \bar{\pi}) < \epsilon$ for all $j \geq J$. Further, as in the proof of Theorem 3.7, $\bar{\pi}$ must be an optimal policy. This leads to a contradiction. Hence, the first claim is true.

Further, suppose that π^* is the unique optimal policy. Then, as shown above, for every $\epsilon > 0$, there exists an integer k_ϵ, such that $\rho(\pi^k, \pi^*) < \epsilon$ for all $k \geq k_\epsilon$. This implies that $\lim_{k\to\infty} \pi^k = \pi^*$.

Now, for the third claim, we note that the result is trivially true if π^k is optimal for some k. When this is not the case, we first claim that given $\epsilon > 0$ and any period n, there exists an iteration counter K_n such that for all $k \geq K_n$, $|\pi_m^k(s) - \pi_m^{k*}(s)| < \epsilon$, for all $m \leq n$ and for all $s \in \mathcal{S}$, for some optimal policy π^{k*}. Suppose not. Then, there exists a subsequence k_i, and for each i, a period $m_i \leq n$ and state $s_i \in \mathcal{S}$ such that $|\pi_{m_i}^{k_i}(s_i) - \pi_{m_i}^*(s_i)| \geq \epsilon$ for all i, for all optimal policies π^*. But

k_i has a further subsequence k_{i_j} such that $\pi^{k_{i_j}}$ converges to an optimal policy $\bar{\pi}$ as in the proof of Theorem 3.8. This leads to a contradiction. Now, fix $0 < \epsilon < 1$ and a period n, and consider any iteration $k \geq K_n$. For any $m \leq n$ and $s \in \mathcal{S}$, we have, $|\pi_m^k(s) - \pi_m^{k*}(s)| < \epsilon$ for some optimal action $\pi_m^{k^*}(s)$. Then, since $\epsilon < 1$ and $\pi_m^k(s), \pi_m^{k*}(s) \in \mathcal{A} = \{1, 2, \ldots, A\}$, we have $\pi_m^k(s) = \pi_m^{k*}(s)$. This proves that all actions up to period n are optimal for policies π^k with $k \geq K_n$. □

We comment that this type of subsequential convergence is the most one can obtain, in general without exploiting any problem-specific features, in infinite-horizon nonstationary sequential decision problems [14, 15].

In this paper, we did not consider the question of how to solve the inner maximization problems in (5) and (6) within Algorithm 1. These problems can be solved by following standard procedures from robust MDPs, and in particular, this can be done efficiently when the uncertainty sets $\mathcal{P}_{n,s}^a$ are convex. We refer the readers to [2, 10, 12] for detailed discussions of this issue.

As we stated in Section 1, nonstationary MDPs are a special case of countable-state stationary MDPs. The simple policy iteration algorithm for nonstationary MDPs in [7] was recently extended to countable-state stationary MDPs in [11]. Along similar lines, it may be possible in the future to extend our work in this paper to robust countable-state stationary MDPs.

4 Acknowledgments

Funded in part by the National Science Foundation through grant #CMMI 1333260.

References

[1] C D Aliprantis and K C Border. *Infinite-dimensional analysis: a hitchhiker's guide*. Springer-Verlag, Berlin, Germany, 1994.

[2] A Ben-Tal, L El Ghaoui, and A Nemirovski. *Robust Optimization*. Princeton University Press, Princeton, NJ, USA, 2009.

[3] T Cheevaprawatdomrong, I E Schochetman, R L Smith, and A Garcia. Solution and forecast horizons for infinite-horizon non-homogeneous Markov decision processes. *Mathematics of Operations Research*, 32(1):51–72, 2007.

[4] A Garcia and R L Smith. Solving nonstationary infinite horizon dynamic optimization problems. *Journal of Mathematical Analysis and Applications*, 244:304–317, 2000.

[5] A Ghate. Infinite Horizon Problems. Wiley Encyclopedia of Operations Research and Management Science, 2010.

[6] A Ghate, D Sharma, and R L Smith. A shadow simplex method for infinite linear programs, forthcoming. *Operations Research*, 58(4):865–877, 2010.

[7] A Ghate and R L Smith. A linear programming approach to nonstationary infinite-horizon Markov decision processes. *Operations Research*, 61(2):413–425, 2013.

[8] W J Hopp, J C Bean, and R L Smith. A new optimality criterion for non-homogeneous Markov decision processes. *Operations Research*, 35:875–883, 1987.

[9] R A Howard. *Dynamic programming and Markov processes*. PhD thesis, MIT, Cambridge, MA, USA, 1960.

[10] G N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[11] I Lee, M A Epelman, H E Romeijn, and R L Smith. Simplex algorithm for countable-state discounted Markov decision processes. `http://www.optimization-online.org/DB_HTML/2014/11/4645.html`, 2014.

[12] A Nilim and L El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[13] M L Puterman. *Markov decision processes : Discrete stochastic dynamic programming*. John Wiley and Sons, New York, NY, USA, 1994.

[14] I E Schochetman and R L Smith. Infinite horizon optimization. *Mathematics of Operations Research*, 14(3):559–574, 1989.

[15] I E Schochetman and R L Smith. Finite dimensional approximation in infinite dimensional mathematical programming. *Mathematical Programming*, 54(3):307–333, 1992.

[16] 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.