Approximate Receding Horizon Approach for Markov Decision Processes: Average Reward Case

by Hyeong Soo Chang, Steven I. Marcus

TR 2001-46

APPROXIMATE RECEDING HORIZON APPROACH FOR MARKOV DECISION PROCESSES: AVERAGE REWARD CASE

Hyeong Soo Chang and Steven I. Marcus*
Institute for Systems Research
University of Maryland, College Park, MD 20742
E-mail: {hyeong,marcus}@isr.umd.edu

May 24, 2002

Abstract

We consider an approximation scheme for solving Markov Decision Processes (MDPs) with countable state space, finite action space, and bounded rewards that uses an *approximate* solution of a fixed finite-horizon sub-MDP of a given infinite-horizon MDP to create a stationary policy, which we call "approximate receding horizon control". We first analyze the performance of the approximate receding horizon control for infinite-horizon average reward under an ergodicity assumption, which also generalizes the result obtained by White [36]. We then study two examples of the approximate receding horizon control via lower bounds to the exact solution to the sub-MDP. The first control policy is based on a finite-horizon approximation of Howard's policy improvement of a single policy and the second policy is based on a generalization of the single policy improvement for multiple policies. Along the study, we also provide a simple alternative proof on the policy improvement for countable state space. We finally discuss practical implementations of these schemes via simulation.

Keywords: Markov decision process, receding horizon control, infinite horizon average reward, policy improvement, rollout, ergodicity

*This research was supported in part by the National Science Foundation under Grant DMI-9988867, and by the Air Force Office of Scientific Research under Grant F496200110161.

1 Introduction

We consider an approximation scheme for solving Markov Decision Processes (MDPs) with countable state space, finite action space, and bounded rewards. The scheme, which we call "approximate receding horizon control", uses an approximate solution of a fixed finite-horizon sub-MDP of a given infinite-horizon MDP to create a stationary policy to solve the infinite-horizon MDP.

The idea of receding horizon control has been applied to many interesting problems in various contexts to solve the problems in an "*on-line*" manner, where in this case we obtain an optimal exact solution with respect to a "small" moving horizon at each decision time and apply the solution to the system. For example, it has been applied to planning problems (e.g., inventory control) that can be modeled as linear programs [14] and that can be represented as a shortest path problem in an acyclic network (see [13] for example problems and references therein), a routing problem in a communication network by formulating the problem as a nonlinear optimal control problem [2], dynamic games [8], aircraft tracking [31], the stabilization of nonlinear time-varying systems [21, 26, 28] in the model predictive control literature, and macroplanning in economics [20], etc. The intuition behind the approach is that if the horizon is "long" enough to obtain a stationary behavior of the system, the moving horizon control would have good performance. Indeed, for MDPs, Hernández-Lerma and Lasserre [16] showed that the value of the receding horizon control converges geometrically to the optimal value, uniformly in the initial state, as the value of the moving horizon increases. For infinite-horizon discounted reward case, it converges geometrically fast with a given discounting factor in (0,1), and for infinite-horizon average reward case, it converges geometrically fast with a given "ergodicity coefficient" in (0,1). Furthermore, it has been shown that there always exists a *minimal finite horizon* H such that the receding H-horizon control prescribes exactly the same action as the policy that achieves the optimal infinite-horizon rewards at every state (see [6] for the discounted case, and [17] for the average reward case with ergodicity assumptions).

Unfortunately, a large state-space size makes it almost impossible to solve the MDPs in practice even with a relatively small receding horizon. Motivated by this, we first analyze the performance of the approximate receding horizon control for the infinite-horizon average reward. The analysis also generalizes the result obtained by White [36] for finite state space with a unichain assumption. We show that the infinite-horizon average reward obtained by following the approximate receding horizon control is bounded by the error due to the finite-horizon approximation that approaches to zero geometrically fast with a given ergodicity coefficient and the error due to the approximation of the optimal finite-horizon value so that if the receding horizon is "long" enough and the approximation of the optimal finite-horizon value is good, the performance bound will be relatively small.

We then study two examples of approximate receding horizon control via lower bounds to the exact solution of the sub-MDP problem of the given infinite-horizon MDP, where both examples can be implemented easily by Monte-Carlo simulation. The first control policy is based on a finite-horizon approximation of Howard's policy improvement of a single policy and the second policy is based on a generalization of the single policy improvement for multiple policies. In the study of the first policy, we provide a simple alternative proof of the policy improvement principle for countable state space, which is rather cumbersome to prove (see, e.g., Chapter 7 in [12] for a proof via the vanishing discount approach for finite state space or [27] for general state space). The Monte-Carlo simulation implementation of the first policy is an extension of an "on-line" simulation method, called "rollout", proposed by Bertsekas and Castanon [4] to solve MDPs for the total reward criterion.

The rollout approach is promising if we have a good base policy. Indeed, several recent works reported successful results in this direction (see Subsection 4.1 of the present paper for a brief survey). Suppose we have multiple base policies available instead of a single base policy. Because we cannot predict each policy's performance easily in advance, it is difficult to select which policy to rollout or to use to be improved upon. Furthermore, it is often true that the available policies are distinct in that each policy's performance is good in different sample paths, in which case one wish to combine the multiply available policies to create a single control policy. To this end, we consider a generalization of the single policy improvement for multiple policies and study its properties. One of the properties of the generalized policy improvement principle is that if there exists a "best" policy that achieves both the best bias and the best gain among the multiple policies, the generalized policy improvement method improves the infinite-horizon average reward of the best policy in the set. As in the rollout policy case, we also approximate the generalized policy improvement principle in a finite horizon sense, generalizing the rollout policy. We call the resulting policy as "parallel rollout". We analyze the performances of the two example policies relative to the policies being rolled out within the framework of the approximate receding horizon approach.

All of the analysis in this paper is based on an "ergodicity" assumption on a given MDP as in the work of Hernández-Lerma and Lasserre [16]. This assumption allows us to discuss the relationship between the value of the receding horizon and the performance of the approximate receding horizon approach. We note that analysis work along this line for the cases of infinite-horizon discounted reward and total reward are reported in [10].

This paper is organized as follows. In Section 2, we formally introduce Markov decision processes and in Section 3, we define the (approximate) receding horizon control and analyze its performance. We then provide two examples of the approximate receding horizon control and analyze their performances in Section 4. We conclude the present paper in Section 5.

2 Markov Decision Process

In this section, we present the essentials of the MDPs we consider and the properties we use in the present paper. For a more substantial introduction, see Puterman's book [33] or the survey paper by Arapostathis et al. [1]. We consider an MDP with a countable state set X, a finite action set A, a nonnegative and bounded reward function R such that $R : X \times A \to \mathcal{R}^+$, and a state transition function P that maps the state and action pair to a probability distribution over X. We will denote the probability of transitioning to state $y \in X$ from state $x \in X$ by taking an action $a \in A$ at x as $p(y|x,a)$. For simplicity, we assume that every action is admissible at each state.

Define a *stationary* policy π as a function $\pi : X \to A$ and denote Π as the set of all possible stationary policies. Given an initial state x, we define the infinite-horizon average reward of following a policy $\pi \in \Pi$ as

$$J_\infty^\pi(x) := \liminf_{H\to\infty} \frac{1}{H} E\left\{ \sum_{t=0}^{H-1} R(x_t, \pi(x_t)) \middle| x_0 = x \right\}, \tag{1}$$

where x_t is a random variable denoting the state at time t following the policy π and we use the subscript ∞ to emphasize the infinite horizon. We seek an optimal policy that achieves

$$J_\infty^*(x) = \sup_{\pi\in\Pi} J_\infty^\pi(x), x \in X.$$

Because there might not always exist such an optimal policy in Π that achieves $J_\infty^*(x)$ [1, 12, 33], we impose an ergodicity assumption throughout the present paper (see, e.g., the page 56 in [18] for stronger assumptions) stated as follows:

Assumption 2.1 *Define $K := \{(x,a)|x \in X, a \in A\}$ and $p(y|k) := p(y|x,a)$ for all $(x,a) \in K$. There exists a positive number $\alpha < 1$ such that*

$$\sup_{k,k'\in K} \sum_{y\in X} |p(y|k) - p(y|k')| \leq 2\alpha.$$

The above ergodicity assumption implies that there always exists an optimal policy π^* in Π, and that for any policy $\pi \in \Pi$, $J_\infty^\pi(x)$ is independent of the starting state x, from which we write J_∞^π omitting x, and that there exists a bounded measurable function h^π on X and a constant J_∞^π such that for all $x \in X$,

$$J_\infty^\pi + h^\pi(x) = R(x,\pi(x)) + \sum_{y\in X} p(y|x,\pi(x))h^\pi(y). \tag{2}$$

We refer to Eq. (2) as *the Poisson's equation with respect to π*.

Let $B(X)$ be the space of real-valued bounded measurable functions on X endowed with the supremum norm $\|V\| = \sup_x |V(x)|$ for $V \in B(X)$. We define an operator $T : B(X) \to B(X)$ as

$$T(V)(x) = \max_{a\in A}\left\{R(x,a) + \sum_{y\in X} p(y|x,a)V(y)\right\}, V \in B(X), x \in X \tag{3}$$

and let $\{V_n^*\}$ be the sequence of *value iteration functions* $V_n^* := T(V_{n-1}^*)$ where $n = 1, 2, ...$ and we assume that $V_0^*(x) = 0$ for all $x \in X$. V_n^* might not converge to a function in $B(X)$ as $n \to \infty$. However, an appropriate transformation of V_n^* does converge. We state this fact by the following theorem (see Theorem 4.8 (a) in [18]).

Theorem 2.1 *Assume that Assumption 2.1 holds. For all* $n \geq 0$*,*

$$\begin{aligned} -\frac{\|R\|}{1-\alpha}\cdot\alpha^n &\leq \inf_x |V_{n+1}^*(x) - V_n^*(x)| - J_\infty^* \\ &\leq \sup_x |V_{n+1}^*(x) - V_n^*(x)| - J_\infty^* \leq \frac{\|R\|}{1-\alpha}\cdot\alpha^n. \end{aligned}$$

Let us define an operator $T_\pi : B(X) \to B(X)$ for $\pi \in \Pi$ as

$$T_\pi(V^\pi)(x) = R(x,\pi(x)) + \sum_{y\in X} p(y|x,\pi(x))V^\pi(y), V^\pi \in B(X), x \in X$$

and let $\{V_n^\pi\}$ be the sequence of *value iteration functions with respect to* π, $V_n^\pi := T_\pi(V_{n-1}^\pi)$ where $n = 1, 2, ...$ and $V_0^\pi(x) = 0$ for all $x \in X$. We can see that V_n^π is the *total reward over horizon of length* n *following the policy* π, i.e, $V_n^\pi(x) = E\{\sum_{t=0}^{n-1} R(x_t, \pi(x_t))|x_0 = x\}$. The above theorem immediately implies the following corollary:

Corollary 2.1 *Assume that Assumption 2.1 holds. For all* $n \geq 0$*, and any* $\pi \in \Pi$*, and all* $x \in X$*,*

$$-\frac{\|R\|}{1-\alpha}\cdot\alpha^n \leq V_{n+1}^\pi(x) - V_n^\pi(x) - J_\infty^\pi \leq \frac{\|R\|}{1-\alpha}\cdot\alpha^n.$$

3 Receding Horizon Control

We define *the receding* H*-horizon control policy* $\pi_H \in \Pi$ with $H < \infty$ as a policy that satisfies for all $x \in X$,

$$T(V_{H-1}^*)(x) = T_{\pi_H}(V_{H-1}^*)(x).$$

It has been shown that there always exists a *minimal finite-horizon* H such that $\pi_H(x) = \pi^*(x)$ for all $x \in X$, where π^* is the policy that achieves $\sup_{\pi\in\Pi}(J_\infty^\pi)$ under Assumption 2.1 [17]. In addition to the existence of such a finite horizon, the paper [17] provides an algorithm (stopping rule) to detect such a horizon in a finite number of time steps. Furthermore, Hernández-Lerma and Lasserre [16] showed that

$$0 \leq J_\infty^* - J_\infty^{\pi_H} \leq \frac{\|R\|}{1-\alpha}\cdot\alpha^{H-1},$$

where we can see that the performance of the receding horizon control policy provides a good approximation for the optimal infinite-horizon average reward and the error approaches zero geometrically with α. Unfortunately, obtaining the true H-horizon optimal value is often difficult, e.g., due to the large state space (see, e.g., [29] for a discussion of the complexity of solving finite-horizon MDPs). Motivated by this, we study an approximate receding horizon control that uses an approximate value function as an approximate solution of V^*_{H-1} for some $H < \infty$.

3.1 Analysis of approximate receding horizon control

We start with a general result that we will use throughout the present paper.

Lemma 3.1 *Assume that Assumption 2.1 holds. Given $V \in B(X)$, consider a policy $\pi^V \in \Pi$ such that*

$$T(V)(x) = T_{\pi^V}(V)(x) \text{ for all } x \in X.$$

Then a stationary distribution P^{π^V} over X exists and $J^{\pi^V}_\infty = \sum_{y\in X}[T(V)(y) - V(y)]P^{\pi^V}(y)$.

Proof: The proof here is similar to that on the page 65 in [18]. We provide the proof of this lemma for completeness. From Remark 3.2 (b) in [18], for any stationary policy $\pi \in \Pi$, there is a unique stationary probability distribution P^π satisfying the invariance property of $P^\pi(x) = \sum_{y\in X} p(x|y,\pi(y))P^\pi(y)$. From the definition of π^V,

$$T(V)(x) = R(x,\pi^V(x)) + \sum_{y\in X} p(y|x,\pi^V(x))V(y).$$

Now summing both sides with respect to the stationary distribution P^{π^V},

$$\sum_{x\in X} T(V)(x)P^{\pi^V}(x) = \sum_{x\in X} R(x,\pi^V(x))P^{\pi^V}(x) + \sum_{x\in X}\sum_{y\in X} p(y|x,\pi^V(x))V(y)P^{\pi^V}(x).$$

The first term on the right side is equal to $J^{\pi^V}_\infty$ by Lemma 3.3 (b.ii) in [18], and the second term on the right side is equal to $\sum_{y\in X} V(y)P^{\pi^V}(y)$ from the invariance property. Rearranging terms yields the desired result. ∎

We define the approximate H-horizon control policy π^V as a policy such that for a given $V \in B(X)$ such that for some $n \geq 0$, $|V^*_n(x) - V(x)| \leq \epsilon$ for all $x \in X$, it satisfies

$$T(V)(x) = T_{\pi^V}(V)(x), x \in X.$$

We now state and prove one of our main theorems.

Theorem 3.1 *Assume that Assumption 2.1 holds. Given $V \in B(X)$ such that for some $n \geq 0$, $|V^*_n(x) - V(x)| \leq \epsilon$ for all x in X, consider a policy π^V such that for all $x \in X$, $T(V)(x) = T_{\pi^V}(V)(x)$. Then,*

$$0 \leq J^*_\infty - J^{\pi^V}_\infty \leq \frac{\|R\|}{1-\alpha}\cdot\alpha^n + 2\epsilon.$$

Proof: We first prove that if $|V_n^*(x) - V(x)| \leq \epsilon$ for all $x \in X$, then $|T(V_n^*)(x) - T(V)(x)| \leq \epsilon$ for all $x \in X$. This simply follows from

$$|T(V_n^*)(x) - T(V)(x)| \leq \max_{a \in A} \left| \sum_{y \in X} [V_n^*(y) - V(y)] p(y|x,a) \right| \leq \sup_x |V_n^*(x) - V(x)|,$$

where the first inequality follows from Hinderer's proposition (page 123 in [18]). Therefore, we have that

$$T(V_n^*)(x) - V_n^*(x) - 2\epsilon \leq T(V)(x) - V(x) \leq T(V_n^*)(x) - V_n^*(x) + 2\epsilon$$

for all $x \in X$ with simple algebra. Now by Lemma 3.1, $J_\infty^\pi = \sum_{y \in X} [T(V)(y) - V(y)] P^\pi(y)$. It follows that

$$\begin{aligned} J_\infty^\pi = \sum_{y \in X} [T(V)(y) - V(y)] P^\pi(y) &\leq \sum_{y \in X} [T(V_n^*)(y) - V_n^*(y)] P^\pi(y) + 2\epsilon \\ &\leq \sup_x |T(V_n^*)(x) - V_n^*(x)| + 2\epsilon \leq J_\infty^* + \frac{\|R\|}{1-\alpha} \cdot \alpha^n + 2\epsilon, \end{aligned}$$

where the last inequality is from Theorem 2.1. The lower bound is trivial by the definition of J_∞^*. ∎

We remark that the infinite-horizon average reward of following the approximate receding horizon control via V is bounded by a term due to the finite-horizon approximation and a term due to the approximation of V, so that if the receding horizon is "long" enough and the approximation by V is good, the performance bound will be relatively small. If $\epsilon = 0$, the result coincides with the one obtained in [16]. As $n \to \infty$, the error approaches 2ϵ, which coincides with the result obtained by White [36] for *finite* state space with a unichain assumption. Furthermore, the above result can be extended to general state space (*Borel* space) with appropriate measure-theoretic arguments.

4 Examples of Approximate Receding Horizon Control

4.1 Rollout—finite-horizon approximation of Howard's policy improvement

Given a policy $\pi \in \Pi$, suppose we solved the Poisson's equation with respect to π given by Eq. (2):

$$J_\infty^\pi + h^\pi(x) = T_\pi(h^\pi)(x), x \in X,$$

obtaining a function h^π and J_∞^π for π. Define a new policy $\bar{\pi}$ such that for all $x \in X$,

$$T_{\bar{\pi}}(h^\pi)(x) = T(h^\pi)(x).$$

That is, for all $x \in X$,

$$\bar{\pi}(x) \in \arg\max_{a \in A} \left(R(x,a) + \sum_{y \in X} p(y|x,a) h^\pi(y) \right). \tag{4}$$

It is well-known that $\bar{\pi}$ improves π in the sense that $J^{\bar{\pi}}_{\infty} \geq J^{\pi}_{\infty}$ under an appropriate condition for a given MDP, and it is called Howard's policy improvement. See, e.g., Chapter 7 in [12] for a proof with a finite state space under an irreducibility condition via the vanishing discount approach. For general state space, see, e.g., [27]. In general, the proof of the policy improvement for the average reward case is not straightforward. We provide here a simple alternative proof under Assumption 2.1:

$$\begin{aligned}
J^{\bar{\pi}}_{\infty} &= \sum_{x}[T(h^{\pi})(x) - h^{\pi}(x)]P^{\bar{\pi}}(x) \text{ from Lemma 3.1} \\
&\geq \sum_{x}[T_{\pi}(h^{\pi})(x) - h^{\pi}(x)]P^{\bar{\pi}}(x) \\
&= \sum_{x}[J^{\pi}_{\infty} + h^{\pi}(x) - h^{\pi}(x)]P^{\bar{\pi}}(x) \\
&= J^{\pi}_{\infty}.
\end{aligned}$$

Unfortunately, obtaining the h^{π}-function is often very difficult, so we consider an approximation scheme that is implementable in practice via Monte-Carlo simulation. As a finite approximation of the policy improvement scheme, we replace h^{π} by the finite-horizon value function of the policy π. The value of following the given policy π can be simply estimated by a sample mean over a set of sample-paths generated by Monte-Carlo simulation. The very idea of simulating a given (heuristic) policy to obtain an (approximately) improved policy originated from Tesauro's work in backgammon [35] and recently, Bertsekas and Castanon extended the idea into an on-line policy improvement scheme called "rollout" to solve finite-horizon MDPs with total reward criterion. It is an on-line scheme in the sense of "planning". That is, at each decision time t, we rollout the given base policy to estimate the utility (called Q-value) of taking an initial action at state x_t and take the action with the highest utility, which creates effectively an improved policy of the base policy in an on-line manner.

Formally, we define the H-horizon *rollout* policy $\pi_{ro,H} \in \Pi$ with a base policy π and $H < \infty$ as the policy:

$$\pi_{ro,H}(x) \in \arg\max_{a \in A} \left(R(x,a) + \sum_{y \in X} p(y|x,a) V^{\pi}_{H-1}(y) \right), x \in X. \tag{5}$$

Note that $V^{\pi}_{H-1}(x)$ is a lower bound to the $V^{*}_{H-1}(x)$ for all $x \in X$. From the result of Theorem 3.1, if V^{π}_{H-1} is a good approximation of V^{*}_{H-1}, the resulting performance will be close to that of the true receding horizon control policy. Note also that the finite-horizon approximation of the policy improvement does not use a function that approximates h^{π} directly but we use an approximation function for $V^{*}_{H-1}(x)$. We will discuss this issue in the next subsection in more detail. The question is how the H-horizon rollout policy performs relatively to the policy π that it rolls out in terms of infinite-horizon average reward.

Theorem 4.1 *Assume that Assumption 2.1 holds. Consider the H-horizon rollout policy $\pi_{ro,H}$ with a base policy π and $H < \infty$. Then*

$$J_\infty^{\pi_{ro,H}} \geq J_\infty^\pi - \frac{\|R\|}{1-\alpha} \cdot \alpha^{H-1}.$$

Proof:

$$\begin{aligned} J_\infty^{\pi_{ro,H}} &= \sum_x [T(V_{H-1}^\pi)(x) - V_{H-1}^\pi(x)] P^{\pi_{ro,H}}(x) \text{ from Lemma 3.1} \\ &\geq \inf_x (T(V_{H-1}^\pi)(x) - V_{H-1}^\pi(x)) \\ &\geq \inf_x (V_H^\pi(x) - V_{H-1}^\pi(x)) \text{ by the definition of } T\text{-operator} \\ &\geq J_\infty^\pi - \frac{\|R\|}{1-\alpha} \cdot \alpha^{H-1} \text{ from Corollary 2.1.} \end{aligned}$$

■

The above result immediately allows us to obtain the value of H to have a desired approximate on-line policy improvement performance. That is, given any $\epsilon > 0$, if we let $\alpha^{H-1} \cdot \frac{\|R\|}{1-\alpha} \leq \epsilon$ so that $H \geq 1 + \log_\alpha \frac{\epsilon(1-\alpha)}{\|R\|}$, then $J_\infty^{\pi_{ro,H}} \geq J_\infty^\pi - \epsilon$.

Several papers for MDP problems (with their related cost criteria) based on a one-step policy improvement idea have reported successful results. For example, Bertsekas and Castanon studied stochastic scheduling problems [4], Secomandi [34] applied the rollout technique combined with neuro-dynamic programming [5] to a vehicle routing problem, Ott and Krishnan [30] and Kolarov and Hui [22] studied network routing problems, Bhulai and Koole [7] consider a multi-server queueing problem, and Koole and Nain [24] consider a two-class single-server queueing model under a preemptive priority rule. In particular, [7] and [24] obtained explicit form expressions for the value function of a fixed threshold policy, which plays the role of the heuristic base policy, and showed numerically that the rollout policy generated from the threshold policy behaves almost optimally. Chang et al. [10] also empirically showed the rollout of a fixed threshold policy (Droptail) performs well for a buffer management problem. Koole [23] also derived the deviation matrix of the $M/M/1/\infty$ and $M/M/1/N$ queue, which is used for computing the bias vector for a particular choice of cost function and a certain base policy, from which the rollout policy of the base policy is generated. Even though the value function of a particular policy can be obtained explicitly for relatively simple cases from problem structure analysis, calculating the exact value function of a particular policy is in general very difficult in practice, in which case we apply the receding rollout policy via simulation. If we have a good heuristic policy, this approach often provides good performance, improving the performance of the given heuristic policy (see, e.g., [10] for queueing problems regarding the performance of the receding horizon rollout policy).

4.2 Parallel rollout

The rollout approach is promising if we have a good base policy, because the performance of the rollout policy is no worse than that of the base policy. Note that in practice, what we are really interested in is the *ranking* of actions, not the degree of approximation. Therefore, as long as the rollout policy preserves the true ranking of actions well, the resulting policy will perform fairly well. However, when we have multiple policies available, because we cannot predict the performance of each policy in advance, selecting a particular single base policy to be rolled out is not an easy task. Furthermore, for some cases, each base policy available is good for different system trajectories. (See, for example, a multiclass scheduling problem with deadlines discussed in [9] where the *static priority* policy and the *earliest deadline first* policy perform optimally for different paths of states.) When this is the case, we wish to combine these base policies *dynamically* in an on-line manner to generate a single policy which adapts automatically to different trajectories of the system, in addition to alleviating the difficulty of choosing a single base policy to be rolled out.

To this end, we first study a generalization of Howard's policy improvement scheme for *multiple* policies and then consider a finite-horizon approximation of the generalized scheme.

Given a finite set $\Lambda \subset \Pi$, suppose that for each $\pi \in \Lambda$, we solved the Poisson's equation with respect to π given by Eq. (2), obtaining a function h^π and J^π_∞ for π.

Note that the function h^π that satisfies the Poisson's equation with respect to π is not necessarily unique [1, 33]. Under Assumption 2.1, the following function known as the "relative value function"

$$\lim_{\gamma \to 1^-} (V^\pi_\gamma(x) - V^\pi_\gamma(0)),$$

where 0 is an arbitrarily fixed state in X and $V^\pi_\gamma(x)$ is the infinite-horizon discounted reward of following π, starting with state x, given by

$$E\left[\sum_{t=0}^{\infty} \gamma^t R(x_t, \pi(x_t)) \middle| x_0 = x\right],$$

solves the Poisson equation with respect to π [18]. Another important h^π that satisfies the Poisson's equation with respect to π is the *bias*. If there exists a state $0 \in X$ that is reachable from any state in X in a finite number of time steps by following any fixed stationary policy, then the function $g^\pi \in B(X)$ defined by

$$g^\pi(x) = \lim_{n \to \infty} (V^\pi_n(x) - nJ^\pi_\infty), x \in X,$$

satisfies the Poisson's equation with respect to π and is called the bias [27]. Therefore, g^π can be taken as h^π. If X is finite and the given MDP is unichain and if we add the condition of

$$\sum_{x \in X} P^\pi(x) h^\pi(x) = 0$$

to the Poisson's equation with respect to π, the bias is the *unique solution* to the Poisson's equation with respect to π (see, e.g., [1, 12, 25]). For the relationship between relative value function and bias and the computation of a bias-optimal policy, see, e.g., [25] for finite state and action spaces with a unichain assumption. Our discussion in this section will focus on the bias but can be extended to the relative value function.

Define a value function $\Phi \in B(X)$ such that

$$\Phi(x) = \max_{\pi \in \Lambda} h^{\pi}(x)$$

and define a new policy $\hat{\pi}$ such that for all $x \in X$,

$$T_{\hat{\pi}}(\Phi)(x) = T(\Phi)(x)$$

and call $\hat{\pi}$ a "parallel rollout" policy. That is, for all $x \in X$,

$$\hat{\pi}(x) \in \arg\max_{a \in A} \left(R(x,a) + \sum_{y \in X} p(y|x,a) \max_{\pi \in \Lambda} h^{\pi}(y) \right). \tag{6}$$

Let $\Delta = \arg\max_{\pi \in \Lambda} J^{\pi}_{\infty}$. We say that δ is a *gain-optimal policy in* Λ if $\delta \in \Delta$, and δ is a *bias-optimal policy in* Λ if $\delta \in \Delta$ and $h^{\delta}(x) \geq \max_{\pi \in \Lambda} h^{\pi}(x)$ for all $x \in X$.

Theorem 4.2 *Given a finite set $\Lambda \subset \Pi$, suppose that for each $\pi \in \Lambda$, h^{π} satisfies the Poisson's equation with respect to π given by Eq. (2):*

$$J^{\pi}_{\infty} + h^{\pi}(x) = T_{\pi}(h^{\pi})(x), x \in X.$$

Consider $\hat{\pi}$ defined in Eq. (6).

a) If there exists a bias-optimal policy in Λ, then

$$J^{\hat{\pi}}_{\infty} \geq \max_{\pi \in \Lambda} J^{\pi}_{\infty}.$$

For any gain-optimal policy δ in Λ,

$$J^{\hat{\pi}}_{\infty} \geq \max_{\pi \in \Lambda} J^{\pi}_{\infty} - \sup_{x \in X} \left(\max_{\pi \in \Lambda}(h^{\pi}(x)) - h^{\delta}(x) \right).$$

b)

$$J^{\hat{\pi}}_{\infty} \geq \sum_{x \in X} J^{\arg\max_{\pi \in \Lambda}(h^{\pi}(x))} P^{\hat{\pi}}(x).$$

We provide the proof of this theorem first before we discuss how these bounds can be interpreted.

Proof: Observe that for any $x \in X$,

$$\begin{aligned} T(\Phi)(x) &= \max_{a \in A}\left(R(x,a) + \sum_{y \in X} p(y|x,a)\Phi(y) \right) \\ &\geq R(x,\pi(x)) + \sum_{y \in X} p(y|x,\pi(x))\Phi(y) \text{ for any } \pi \in \Lambda \\ &\geq R(x,\pi(x)) + \sum_{y \in X} p(y|x,\pi(x))h^{\pi}(y) \text{ from the definition of } \Phi \\ &= J^{\pi}_{\infty} + h^{\pi}(x). \end{aligned}$$

Because the above inequality holds for any $\pi \in \Lambda$, for all $x \in X$,

$$T(\Phi)(x) \geq \max_{\pi \in \Lambda}(J^{\pi}_{\infty} + h^{\pi}(x)).$$

Now,

$$\begin{aligned} J^{\hat{\pi}}_{\infty} &= \sum_{x \in X}[T(\Phi)(x) - \Phi(x)]P^{\hat{\pi}}(x) \text{ from Lemma 3.1} \\ &\geq \sum_{x \in X}[\max_{\pi \in \Lambda}(J^{\pi}_{\infty} + h^{\pi}(x)) - \max_{\pi \in \Lambda} h^{\pi}(x)]P^{\hat{\pi}}(x) \text{ by the previous observation} \qquad (7) \end{aligned}$$

For part a), selecting a gain-optimal policy $\delta \in \Lambda$ that achieves $\max_{\pi \in \Lambda}(J^{\pi}_{\infty})$ yields

$$\begin{aligned} J^{\hat{\pi}}_{\infty} &\geq \sum_{x \in X}[J^{\delta}_{\infty} + h^{\delta}(x) - \max_{\pi \in \Lambda} h^{\pi}(x)]P^{\hat{\pi}}(x) \\ &\geq J^{\delta}_{\infty} + \sum_{x \in X}[h^{\delta}(x) - \max_{\pi \in \Lambda} h^{\pi}(x)]P^{\hat{\pi}}(x) \\ &\geq \max_{\pi \in \Lambda} J^{\pi}_{\infty} - \sup_{x}\left(\max_{\pi \in \Lambda} h^{\pi}(x) - h^{\delta}(x) \right). \end{aligned}$$

If δ is a bias-optimal policy in Λ, $\sum_{x \in X}[h^{\delta}(x) - \max_{\pi \in \Lambda} h^{\pi}(x)]P^{\hat{\pi}}(x) \geq 0$. Thus for this case,

$$J^{\hat{\pi}}_{\infty} \geq \max_{\pi \in \Lambda} J^{\pi}_{\infty}.$$

For part b), at each x, selecting a policy in Λ that achieves $\max_{\pi \in \Lambda}(h^{\pi}(x))$ in Eq. (7) yields the desired result with simple algebra. ∎

An interpretation of the above theorem is as follows. Part a) of Theorem 4.2 states that the parallel rollout policy improves the infinite-horizon average reward of a gain-optimal policy in Λ and the error is bounded by the maximal difference of the biases achieved by the gain-optimal policy and the policies in Λ, and it is guaranteed that the the parallel rollout policy improves the gain of any policy in Λ if Λ contains at least one bias-optimal policy. Part b) states that the gain of the parallel rollout policy is no worse than the average gain of the best policy that achieves the maximal bias value at each state, where the average is taken over the stationary distribution of $\hat{\pi}$ (it can be thought as an initial distribution over X). That is,

$$\sum_{x \in X} \left(J_{\infty}^{\hat{\pi}} - J_{\infty}^{\arg\max_{\pi \in \Lambda} h^{\pi}(x)} \right) P^{\hat{\pi}}(x) \geq 0.$$

We now consider a finite-horizon approximation of the parallel rollout control policy within the framework of the approximate receding horizon control. The direct generalization of the rollout policy defined in the previous subsection is to replace $\max_{\pi \in \Lambda} h^{\pi}(x), x \in X$ by the maximum of the values of the policies in Λ for a finite horizon at the state x. Formally, we define the H-horizon parallel rollout policy $\pi_{pr,H}$ with a finite set $\Lambda \subset \Pi$ of base policies in Π as

$$\pi_{pr,H}(x) \in \arg\max_{a \in A} \left(R(x,a) + \sum_{y \in X} p(y|x,a) \max_{\pi \in \Lambda} V_{H-1}^{\pi}(y) \right), x \in X. \quad (8)$$

We can first easily see that this is based on a more accurate lower bound of the optimal total reward value than that of the H-horizon rollout policy and if $\sup_{x \in X} |\max_{\pi \in \Lambda} V_{H-1}^{\pi}(x) - V_{H-1}^{*}(x)| \leq \epsilon$, by Theorem 3.1, the performance will be close to that of the true receding horizon approach. If we view h^{π} as the relative value function or the bias, in the definitions of the rollout and parallel rollout policies, we don't directly estimate h^{π}. One can use a finite-horizon approximation of h^{π} directly. For example, we could use $V_{H-1}^{\pi}(x) - V_{H-1}^{\pi}(0)$ with a fixed state $0 \in X$ instead of $V_{H-1}^{\pi}(x)$ for an approximate value of the relative value function. The result of Theorem 4.1 still holds with this replacement via Corollary 2.1 with the simple observation that

$$J_{\infty}^{\pi_{ro,H}} = \sum_{x} [T(V_{H-1}^{\pi})(x) - V_{H-1}^{\pi}(0) - V_{H-1}^{\pi}(x) + V_{H-1}^{\pi}(0)] P^{\pi_{ro,H}}(x).$$

The main reasons that we use the total reward value, not the relative total reward value or the bias, are twofold. First, we want our approximation scheme to be within our framework of the approximate receding horizon control, which allows us to compare the performance of the (parallel) rollout policy with the optimal infinite-horizon average reward. Second, we want to keep the spirit of the (parallel) rollout policy defined for the discounted reward criterion. (For the infinite-horizon discounted criterion, we simply replace the total reward of following a policy in the definition of the (parallel) rollout policy by the total discounted reward. See [9, 10].)

We analyze below the performance of the H-horizon parallel rollout policy compared with the infinite-horizon average rewards obtained by policies in Λ. To this end, for any $\pi \in \Pi$, define $J_n^{\pi}(x) = \frac{V_n^{\pi}(x)}{n}$ for all $x \in X$ and $n = 1, 2, ...$ That is, this is the n-horizon approximation of the infinite-horizon average reward. With a similar argument as Platzman's given in Section 3.3 in [32], we can show that $J_n^{\pi}(x)$ converges, uniformly in x, as $O(n^{-1})$, to J_{∞}^{π}, $n = 1, 2, ...$

Theorem 4.3 *Assume that Assumption 2.1 holds. Consider the H-horizon parallel rollout policy*

$\pi_{pr,H}$ *with a finite set* $\Lambda \subset \Pi$ *and* $H < \infty$*. Then*

$$J_\infty^{\pi_{pr,H}} \geq \sum_{x \in X} J_\infty^{\arg\max_{\pi \in \Lambda} J_{H-1}^\pi(x)} P^{\pi_{pr,H}}(x) - \frac{\|R\|}{1-\alpha} \cdot \alpha^{H-1}.$$

Proof: From Corollary 2.1, for all $x \in X$,

$$V_H^\pi(x) - V_{H-1}^\pi(x) \geq J_\infty^\pi - \frac{\|R\|}{1-\alpha} \cdot \alpha^{H-1}.$$

Define $\Phi \in B(X)$ such that $\Phi(x) = \max_{\pi \in \Lambda} V_{H-1}^\pi(x)$ for all $x \in X$. Observe that for all $x \in X$,

$$\begin{aligned} T(\Phi)(x) &= \max_{a \in A} \left(R(x,a) + \sum_{y \in X} p(y|x,a)\Phi(y) \right) \\ &\geq R(x,\pi(x)) + \sum_{y \in X} p(y|x,\pi(x))\Phi(y) \text{ for any } \pi \in \Lambda \\ &\geq R(x,\pi(x)) + \sum_{y \in X} p(y|x,\pi(x))V_{H-1}^\pi(y) = V_H^\pi(x). \end{aligned}$$

Therefore, for all $x \in X$,

$$T(\Phi)(x) \geq \max_{\pi \in \Lambda} V_H^\pi(x). \tag{9}$$

Now,

$$\begin{aligned} J_\infty^{\pi_{pr,H}} &= \sum_x [T(\Phi)(x) - \Phi(x)]P^{\pi_{pr,H}}(x) \text{ by Lemma 3.1} \\ &\geq \sum_x [\max_{\pi \in \Lambda} V_H^\pi(x) - \max_{\pi \in \Lambda} V_{H-1}^\pi(x)]P^{\pi_{pr,H}}(x) \text{ by Eq. (9)} \\ &\geq \sum_x [V_H^{\arg\max_{\pi \in \Lambda} J_{H-1}^\pi(x)}(x) - V_{H-1}^{\arg\max_{\pi \in \Lambda} J_{H-1}^\pi(x)}(x)]P^{\pi_{pr,H}}(x) \\ &\geq \sum_x J_\infty^{\arg\max_{\pi \in \Lambda} J_{H-1}^\pi(x)} P^{\pi_{pr,H}}(x) - \frac{\|R\|}{1-\alpha} \cdot \alpha^{H-1} \text{ by Corollary 2.1 .} \end{aligned}$$

∎

As we increase H, the term with α will decrease geometrically fast in α and $J_{H-1}^\pi(x), x \in X$ approaches J_∞^π in $O(H^{-1})$. An interpretation of this result is as follows. The infinite-horizon average reward by following the H-horizon parallel rollout policy is no worse than that by following the best policy in Λ that achieves the maximum H-horizon average reward associated with a starting state x among the policies in Λ, where the distribution of the starting state is given by the stationary distribution of following the parallel rollout policy.

We conclude this section with a brief discussion of some intuition behind the parallel rollout policy. Consider a policy ϕ_H that selects the action given by the policy π in Λ that has the highest $J_H^\pi(x)$ estimate at the current state x. That is, at state x, ϕ_H takes an action given by

$$\arg\max_{\pi \in \Lambda} (J_H^\pi(x))\,(x).$$

Note that this policy will converge to the policy $\arg\max_{\pi\in\Lambda} J^{\pi}_{\infty}$ in $O(H^{-1})$. However, this receding horizon policy, ϕ_H, selects only the action prescribed by $\pi \in \Lambda$. In other words, the policy does not give enough emphasis and freedom in the evaluation to the *initial* action (this drawback has been empirically shown in [10]). Therefore, we conjecture informally that this policy is generally suboptimal even though we can expect that ϕ_H is much more uniformly reasonable (across the state space) than any single base policy in Λ. On the other hand, the receding horizon rollout policy evaluates each possible initial action based on one-step lookahead relative to a base policy being improved. We can view the parallel rollout technique as a method of capturing the spirit of rolling out ϕ_H with a low cost (see, also [10] on the similar discussion for discounted reward criterion).

5 Concluding Remarks

When we simulate a base policy by Monte-Carlo simulation, using different sets of random number sequences (different sample-paths) across actions increases the variance in the utility (Q-value) measure. Therefore, we suggest using the same set of random number sequences across actions. This has the same flavor as the differential training method [3] and common random number simulation in the discrete event systems literature [19].

There are several papers in the literature regarding the simulation-based policy iteration method where the policy evaluation step is done via simulation. Rather than estimating a finite-horizon total-reward value of a policy, those papers consider approximating h^{π} directly. For example, Cooper et al. [11] use a sampling method called "coupling-from-the-past" that requires obtaining a sample from the stationary distribution of the (aperiodic) Markov chain generated by a fixed policy and He et al. [15] use a temporal-difference learning scheme in order to estimate the bias of the policy, where both papers are under the finite state and action space constraint and a unichain assumption. On the other hand, Bertsekas and Tsitsiklis [5] discuss estimating h^{π} defined as the relative value function via Monte-Carlo simulation.

Acknowledgements This research is partly based on the first author's Ph.D. work conducted at Purdue university in West Lafayette, Indiana. The first author is grateful to his advisors, Robert Givan and Edwin Chong, for their guidance and support throughout the Ph.D. work.

References

1. A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus, Discrete-time controlled Markov processes with average cost criterion: a survey, *SIAM J. Control and Optim.* **31** (1993), 282–344.

2. M. Baglietto, T. Parisini, and R. Zoppoli, Neural approximators and team theory for dynamic routing: a receding horizon approach, *Proc. IEEE CDC.* (1999), 3283–3288.

3. D. P. Bertsekas, Differential training of rollout policies, *Proc. 35th Allerton Conference on Communication, Control, and Computing* (1997).

4. D. P. Bertsekas and D. A. Castanon, Rollout algorithms for stochastic scheduling problems, *J. Heuristics* **5** (1999), 89–108.

5. D. P. Bertsekas and J. Tsitsiklis, "Neuro-Dynamic Programming," Athena Scientific, 1996.

6. C. Bes and J. B. Lasserre, An on-line procedure in discounted infinite-horizon stochastic optimal control, *J. Optim. Theory Appl.* **50** (1986), 61-67.

7. S. Bhulai and G. Koole, On the structure of value functions for threshold policies in queueing models, Technical Report 2001-4, Department of Stochastics, Vrije Universiteit Amsterdam, 2001.

8. W. A. van den Broek, Moving horizon control in dynamic games, *Computing in Economics and Finance* **122** (1999).

9. H. S. Chang, "On-line sampling-based control for network queueing problems," Ph.D. thesis, School of Electrical and Computer Engineering, Purdue University, 2001.

10. H. S. Chang, R. Givan, and E. K. P. Chong, Parallel rollout for on-line solution of partially observable Markov decision processes, *Discrete Event Dyn. Syst.* (2002), Revised.

11. W. Cooper, S. Henderson, and M. Lewis, Perfect simulation and the policy iteration algorithm, (2001), Submitted.

12. E. B. Dynkin and A. A. Yushkevich, "Controlled Markov Processes," Springer-Verlag, 1979.

13. A. Federgruen and M. Tzur, Detection of minimal forecast horizons in dynamic programs with multiple indicators of the future, *Naval Res. Logist.* **43** (1996), 169–189.

14. R. C. Grinold, Finite horizon approximations of infinite horizon linear programs, *Math. Program.* **12** (1997), 1–17.

15. Y. He, M. Fu, and S. Marcus, Simulation-based algorithms for average cost Markov decision processes, *Computing Tools for Modeling, Optimization and Simulation, Interfaces in Computer Science and Operations Research*, (2000), 161–182.

16. O. Hernández-Lerma and J. B. Lasserre, Error bounds for rolling horizon policies in discrete-time Markov control processes, *IEEE Trans. Automat. Control* **35** (1990), 1118–1124.

17. O. Hernández-Lerma and J. B. Lasserre, A forecast horizon and a stopping rule for general Markov decision processes, *J. Math. Anal. Appl.* **132** (1988), 388–400.

18. O. Hernández-Lerma, "Adaptive Markov Control Processes," Springer-Verlag, 1989.

19. Y. -C. Ho and X. -R. Cao, "Perturbation Analysis of Discrete Event Dynamic Systems," Norwell, Massachusetts: Kluwer Academic Publishers, 1991.

20. L. Johansen, "Lectures on Macroeconomic Planning," Amsterdam, The Netherlands: North-Holland, 1977.

21. S. S. Keerthi and E. G. Gilbert, Optimal, infinite horizon feedback laws for a general class of constrained discrete time systems: stability and moving-horizon approximations, *J. Optim. Theory Appl.* **57** (1988), 265–293.

22. A. Kolarov and J. Hui, On computing Markov decision theory-based cost for routing in circuit-switched broadband networks, *J. Network and Systems Management* **3** (1995), 405-425.

23. G. Koole, The deviation matrix of the $M/M/1/\infty$ and $M/M/1/N$ queue, with applications to controlled queueing models, *Proc. of the 37th IEEE CDC.* (1998), 56–59.

24. G. Koole and Philippe Nain, On the value function of a priority queue with an application to a controlled pollying model, *QUESTA*, to appear.

25. M. Lewis and M. Puterman, A probabilistic analysis of bias optimality in unichain Markov decision processes, *IEEE Trans. Automat. Control*, to appear.

26. D. Q. Mayne and H. Michalska, Receding horizon control of nonlinear system, *IEEE Trans. Automat. Control* **38** (1990), 814–824.

27. S. Meyn, The policy iteration algorithm for average reward Markov decision processes with general state space, *IEEE Trans. Automat. Control* **42** (1997), 1663–1680.

28. M. Morari and J. H. Lee, Model predictive control: past, present, and future, *Computers and Chemical Engineering* **23** (1999), 667–682.

29. M. Mundhenk, J. Goldsmith, C. Lusena, and E. Allender, Complexity of finite-horizon Markov decision process problems, *J. ACM* **47** (2000), 681–720.

30. T. J. Ott and K. R. Krishnan, Separable routing: a scheme for state-dependent routing of circuit switched telephone traffic, *Ann. Oper. Res.* **35** (1992), 43–68.

31. W. N. Patten and L. W. White, A sliding horizon feedback control problem with feedforward and disturbance, *J. Mathematical Systems, Estimation, and Control* **7** (1997), 1–33.

32. L. K. Platzman, Optimal infinite-horizon undiscounted control of finite probabilistic system, *SIAM J. Control and Optim.* **14** (1980), 362–380.

33. M. L. Puterman, "Markov Decision Processes: Discrete Stochastic Dynamic Programming," Wiley, New York, 1994.

34. N. Secomandi, Comparing neuro-dynamic programming algorithms for the vehicle routing problem with stochastic demands, *Comput. Oper. Res.* **27** (2000), 1201–1225.

35. G. Tesauro and G. R. Galperin, On-line policy improvement using Monte Carlo search, *Proc. of NIPS* (1997), 1068.

36. D. J. White, The determination of approximately optimal policies in Markov decision processes by the use of bounds, *J. Oper. Res. Soc.* **33** (1982), 253–259.