Markov Games: Receding Horizon Approach

by Hyeong Soo Chang, Steven I. Marcus

TR 2001-48

MARKOV GAMES: RECEDING HORIZON APPROACH

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

December 15, 2001

Abstract

We consider a receding horizon approach as an approximate solution to two-person zero-sum Markov games with infinite horizon discounted cost and average cost criteria. We first present error bounds from the optimal equilibrium value of the game when both players take correlated equilibrium receding horizon policies that are based on *exact* or *approximate* solutions of receding finite horizon subgames. Motivated by the worst-case optimal control of queueing systems by Altman [1], we then analyze error bounds when the minimizer plays the (approximate) receding horizon control and the maximizer plays the worst case policy. We give three heuristic examples of the approximate receding horizon control. We extend "rollout" by Bertsekas and Castanon [9] and "parallel rollout" and "hindsight optimization" by Chang *et al.* [13, 16] into the Markov game setting within the framework of the approximate receding horizon approach and analyze their performances. From the rollout/parallel rollout approaches, the minimizing player seeks to improve the performance of a single heuristic policy it rolls out or to combine dynamically multiple heuristic policies in a set to improve the performances of all of the heuristic policies simultaneously under the guess that the maximizing player has chosen a fixed worst-case policy. Given $\epsilon > 0$, we give the value of the receding horizon which guarantees that the parallel rollout policy with the horizon played by the minimizer *dominates* any heuristic policy in the set by ϵ. From the hindsight optimization approach, the minimizing player makes a decision based on his expected optimal hindsight performance over a finite horizon. We finally discuss practical implementations of the receding horizon approaches via simulation.

Keywords: Markov game, receding horizon control, infinite horizon cost, rollout, hindsight optimization

[*] This research was supported in part by the National Science Foundation under Grant DMI-9988867, by the Air Force Office of Scientific R esearch under Grant F496200110161, and by the Department of Defense under Contract MDA 90499C2521.

1 Introduction

Game Theory has been used to model dynamic sequential decision making problems in a wide variety of situations where multiple decision makers compete or cooperate to optimize their cost functionals. In this paper, we consider games with two players where one player (the minimizer) wishes to minimize his cost that will be paid to the other player (the maximizer). Both players take underlying decisions simultaneously at each state, with the complete knowledge of the state of the system but without knowing each other's current action being taken. We can view the maximizer as *nature* which controls the disturbances that are unknown to the minimizer [39]. The minimizer then tries to get the best performance under the worst possible dynamic choice of the unknown disturbance parameters controlled by nature. That is, the minimizer seeks to design a robust controller that works well under the worst case scenario [6]. This gives rise to two-person zero-sum Markov games.

Recently, Markov games have received an attention in the queueing system literature in order to solve interesting telecommunication network problems, for example, admission control, routing, flow control, etc. (see, e.g., Altman's paper [1] and the references therein and [24]). However, even though the worst-case scenario[1] that will be played by the maximizer can be analyzed for some problems, it is often quite difficult to obtain such a policy exactly. In that case, the natural step for the minimizer is to "guess" the seemingly worst possible play of the maximizer and then try to optimize his performance. If the minimizer assumes that the maximizer will play a fixed policy, to the minimizer the problem becomes solving a Markov decision process (MDP) [42]. It is well-known that solving MDPs in general (for infinite horizon cost) is often impractical if the state space is large even though exact solution techniques are available, e.g., value iteration or policy iteration, etc. (see, e.g., [42] for a substantial discussion). This means that even with the minimizer's guess on the opponent's play, getting the best performance for him is difficult.

With this motivation, we focus on solving two-person zero-sum Markov games with infinite horizon discounted cost and infinite horizon average cost criteria via an *approximation* framework in the context of "planning". We adopt a receding horizon control approach. The idea is to obtain an optimal solution with respect to a "small" moving horizon at each decision time and apply the solution to the system. In fact, this approach has been studied in several contexts in various fields, e.g., planning in economics [27], model predictive control literature [29, 34, 35], and planning in MDPs [23, 13], etc. In the game setting, Baglietto *et al.* [5] applied team theory [25] empirically with a receding horizon control to solve a routing problem in a communication network by formulating the problem as a nonlinear optimal control problem, and Van den Broek [12] considered a receding horizon control in non-zero sum differential games [12], specifically analyzing the performance of

[1] What we mean by the worst case scenario is the case when the maximizer plays his optimal equilibrium policy that we define later.

linear quadratic games. The receding horizon control he employed is somewhat different from what we do here. In his case, at any decision time, the players base their actions on a finite horizon but at each decision time, the horizon size increases. This paper focuses on a *fixed* receding horizon size.

At each state, the minimizing player selects a small but typical horizon and solves the given Markov game with the finite horizon (called the subgame) under the guess that the maximizing player makes his decision based on his best performance for the subgame. The minimizing player then takes a randomized action based on the solution to the subgame. The intuition is that if the horizon is "long" enough to get a stationary behavior of the game, this moving horizon control would have a good performance. Indeed, we first show that the value of the game played by the receding horizon control from *both players* converges geometrically fast, with given discount factor in (0,1) for infinite horizon discounted cost and with given "ergodicity coefficient" in (0,1) for infinite horizon average cost, to the optimal equilibrium value of the game, uniformly in the initial state, as the value of the moving horizon increases (Hernández-Lerma and Lasserre [23] obtained a similar result for MDPs [23]). We mention here that this error analysis assumes that the maximizing player also plays his respective half of a *common* copy of the approach, resulting in a so-called *correlated equilibrium*. In other words, the maximizer also plays the receding horizon control like the minimizer. We then present an error bound between the optimal equilibrium value and the value of the game in which the minimizer plays the receding horizon control and the maximizer plays the worst case scenario (playing the equilibrium policy), which also vanishes to zero as the size of the receding horizon goes to infinity. This also answers an important question that arises in the Markov game literature: what size of the planning horizon should the minimizer use to achieve a good approximate value of the equilibrium value?

However, as we mentioned before, solving the finite horizon Markov game or subgame is also troublesome if the state space is large. So we consider an approximate receding horizon control. Rather than solving the finite horizon subgames exactly at each decision time, at each state, the minimizer will make his decision based on the *approximate* solution for the subgame. We also analyze the performance of this approach as previously done for the receding horizon control in MDP contexts [15].

We then shift our attention to some examples of the approximate receding horizon control for the minimizer. These are all heuristics where the minimizer guesses the maximizer's worst case scenario and approximates the solution of the subgame and takes a decision based on this approximate solution, and all of these heuristics can be implemented via a simple Monte-Carlo simulation. The first two approaches that will be taken by the minimizer aim at improving a given heuristic policy (or a set of multiple heuristic policies) that is available to the minimizer, based on policy improvement arguments. We first consider an adaptation of the rollout approach by Bertsekas and Castanon [9] into the Markov game setting. In this approach, the minimizer guesses the maximizer's worst possible play and uses a single heuristic policy to rollout to generate a new policy whose performance in terms of the value of the game will be no worse than the given heuristic policy if the maximizer indeed plays the policy the minimizer guessed. To the minimizer, it will be often true that he has more than one heuristic policy available such that a particular heuristic policy's performance is near-otimal for the particular sample paths of the system. He may well try to combine these policies dynamically. The next approach, called parallel rollout [13], is a generalization of the rollout approach. It also appeals to the policy improvement principle and can show that for any fixed policy taken by the maximizer, the minimizer will improve the performances of all heuristic policies simultaneously if the minimizer plays the parallel rollout with respect to the fixed policy of the maximizer. In other words, the parallel rollout is a formal method of generating a policy that *dominates* all heuristic policies available. Based on the analysis of the approximate receding horizon control we will present in this paper, we can say that if the minimizer's guess on the opponent's play is good and the resulting approximate value of the subgame is also good, the two approaches will yield a reasonable performance to the minimizer. Furthermore, given $\epsilon > 0$, we provide the value of the receding horizon which guarantees that for any fixed policy played by the maximizer, the parallel rollout policy with the finite horizon played by the minimizer with respect to the fixed policy of the maximizer yields a value of the game no larger than that of the game played by any policy among heuristic policies by the minimizer and by the fixed policy chosen by the maximizer plus ϵ.

The final example approach is motivated by hindsight optimization proposed in [13, 16]. By this approach, at each state, the minimizer evaluates his candidate randomized actions based on the analysis of the expected optimal hindsight performance over a finite horizon under the assumption that the maximizer plays the worst-case fixed policy that chosen by the minimizer. This approach has a flavor of heuristically adapting the hindsight optimal solutions into on-line solutions (via on-line simulation).

This paper is organized as follows. In Section 2, we formalize mathematically the Markov games we consider. We then introduce the (approximate) receding horizon control in Section 3 and analyze performances. We then discuss three heuristics for the approximate receding horizon control in Section 4. In Section 5, we discuss implementation issues and other research directions.

2 Markov Game

In this section, we formulate the two-person zero-sum Markov game introduced by Shapley [44] in a formal mathematical setting. For a substantial discussion on this topic, see, e.g., [18] [7] or [40]. Let X denote a finite state space and for $x \in X$, $N(x)$ and $M(x)$ denote the finite sets of actions for the minimizing player (minimizer) and the maximizing player (maximizer), respectively. Both players play underlying actions simultaneously at each state, with the complete knowledge of the state of the system but without knowing each other's current action being taken. At each state x, each player will consider choosing an action to take according to a probability distribution over the available actions. For each $x \in X$, we define the players' "admissible *randomized action* sets" as $G(x)$ and $F(x)$ such that

$$\begin{aligned} G(x) &= \{g \in R^{|N(x)|} | \sum_{i \in N(x)} g_i = 1, \text{ and } \forall i, g_i \geq 0\} \\ F(x) &= \{f \in R^{|M(x)|} | \sum_{i \in M(x)} f_i = 1, \text{ and } \forall i, f_i \geq 0\} \end{aligned}$$

Once the actions $n \in N(x)$ and $m \in M(x)$ at state x are taken by both players, the state transitions probabilistically to next state y according to the probability $p(y|x, n, m)$. From this, we induce the probability $P_{xy}(g, f)$ denoting the probability of transitioning from state x to state y under the randomized actions $g \in G(x)$ and $f \in F(x)$:

$$P_{xy}(g, f) = \sum_{n \in N(x)} \sum_{m \in M(x)} g_n f_m p(y|x, n, m).$$

If the minimizer takes a randomized action $g \in G(x)$ and the maximizer takes $f \in F(x)$ at state x, then the minimizer gets the expected payoff (cost) of $C_x(g, f)$, which is given by

$$C_x(g, f) = \sum_{y \in X} \sum_{n \in N(x)} \sum_{m \in M(x)} c(x, y, n, m) p(y|x, n, m) g_n f_m,$$

where $c(x, y, n, m)$ is the immediate payoff to the minimizer (the negative of this will be incurred to the maximizer) associated with a current state and the next state pair (x, y) after taking the action $n \in N(x)$ if action m is taken by the maximizer. We assume that $|C_x(g, f)| \leq C_{\max} < \infty$ for any x, g and f. We now define a *stationary* policy π or strategy of the minimizer to be a function $\pi : X \to G(X)$ and denote Π as the set of all possible policies, and similarly a policy ϕ and the set Φ are defined for the maximizer. We will say that a stationary policy is *pure*, if the randomized action selected by the policy at every state yields a non-randomized action choice, i.e., an action is selected with probability one.

In this paper, we consider two objective function criteria: infinite horizon discounted cost and average cost. Given a policy π selected by the minimizer and a policy ϕ selected by the maximizer, we define *the value of the game played with π and ϕ by the minimizer and the maximizer, respectively, with a starting state x* as

$$V_\infty(\pi, \phi)(x) := E\{\sum_{t=0}^{\infty} \gamma^t C_{x_t}(\pi(x_t), \phi(x_t)) | x_0 = x\}$$

for the *infinite horizon discounted cost criterion*, where x_t is a random variable denoting the state at time t following the policies π and ϕ, and $\gamma \in (0, 1)$ is a given discount factor. The discount factor can be interpreted as the probability that the game will be allowed to continue after the current decisions made by both players. Similarly, we define the value of the game for the *infinite horizon average cost criterion* as

$$J_\infty(\pi,\phi)(x) := \lim_{H\to\infty}\frac{1}{H}E\{\sum_{t=0}^{H-1} C_{x_t}(\pi(x_t),\phi(x_t))|x_0 = x\}$$

with given policies π and ϕ.

The goal of the minimizer (the maximizer) is to find a policy $\pi \in \Pi$ ($\phi \in \Phi$) which minimizes (maximizes) the value of the game. Throughout this paper, V_∞ always refers to the value of the game with the infinite horizon discounted cost criterion and J_∞ refers to the value of the game with the infinite horizon average cost criterion, so that we will omit which criterion we mention at any point if the context is clear.

2.1 Some preliminaries

2.1.1 Infinite horizon discounted cost

It is well-known (see, e.g., [40]) that there exists an *optimal equilibrium policy pair* $\pi^* \in \Pi$ and $\phi^* \in \Phi$ such that for all $\pi \in \Pi$ and $\phi \in \Phi$ and $x \in X$,

$$V_\infty(\pi^*,\phi)(x) \le V_\infty(\pi^*,\phi^*)(x) \le V_\infty(\pi,\phi^*)(x).$$

We will refer to the value $V_\infty(\pi^*,\phi^*)(x)$ as the *equilibrium value of the game* associated with state x and to π^* and ϕ^* as the *equilibrium policies* for the minimizer and the maximizer, respectively. We will write $V_\infty(\pi^*,\phi^*)$ as V_∞^* and focus on finding or approximating the policy π^* (note that the content of this paper can be interpreted for the maximizer case by changing the role of the minimizer and the maximizer). A primitive but important notion that arises in game theory is that of *dominance* (see, e.g., [19]). We will say that a policy $\pi_1 \in \Pi$ (weakly) *dominates* $\pi_2 \in \Pi$ if and only if for any $\phi \in \Phi$, $V_\infty(\pi_1,\phi) \le V_\infty(\pi_2,\phi)$.

Now 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 several operators that map a function in $B(X)$ to a function in $B(X)$: for all $\pi \in \Pi, \phi \in \Phi, V \in B(X)$, and $x \in X$,

$$\begin{aligned}
T(V)(x) &= \inf_{g\in G(x)}\sup_{f\in F(x)}\left\{C_x(g,f) + \gamma\sum_{y\in X}P_{xy}(g,f)V(y)\right\}\\
T_{\pi,\phi}(V)(x) &= C_x(\pi(x),\phi(x)) + \gamma\sum_{y\in X}P_{xy}(\pi(x),\phi(x))V(y)\\
T_\phi(V)(x) &= \inf_{g\in G(x)}\left\{C_x(g,\phi(x)) + \gamma\sum_{y\in X}P_{xy}(g,\phi(x))V(y)\right\}
\end{aligned}$$

$$T_\pi(V)(x) \quad = \quad \sup_{f\in F(x)} \left\{ C_x(\pi(x), f) + \gamma \sum_{y\in X} P_{xy}(\pi(x), f)V(y) \right\}$$

It is well-known [18, 40] that each of the above operators is a *contraction mapping* in $B(X)$, that is, for the case of T, for any V_1 and V_2 in $B(X)$, $\|T(V_1) - T(V_2)\| \leq \gamma\|V_1 - V_2\|$, and that each operator has a *monotonicity* property, that is, if $V_1(x) \leq V_2(x)$ for all $x \in X$, $T(V_1)(x) \leq T(V_2)(x)$ for all $x \in X$ for the case of T. Furthermore, there exist a unique fixed point $v \in B(X)$ such that $T(v) = v$ and v is equal to $V_\infty(\pi^*, \phi^*)$, and a unique fixed point $u \in B(X)$ such that $T_{\pi,\phi}(u) = u$ and $u = V_\infty(\pi, \phi)$. We finally remark that for all $x \in X$, the infimum and supremum in the definitions of the operators T, T_ϕ, and T_π are achieved by elements in $G(x)$ and $F(x)$ (see, e.g., Section 3 in [38] or [40]).

Let $\{V_n^*\}$ be the sequence of *value iteration functions* $V_n^* := T(V_{n-1}^*)$ where $n = 1, 2, ...$ and let V_0^* be an arbitrary function in $B(X)$, but we assume that $\max_x |V_0^*(x)| \leq C_{\max}/(1-\gamma)$ for a technical reason. It is straightforward to show that as $n \to \infty$, V_n^* converges to $V_\infty(\pi^*, \phi^*)$ geometrically fast in γ by the contraction mapping property and the Banach fixed point theorem. Furthermore, V_n^* is the *equilibrium value of the finite n-horizon game.* We introduce a nonstationary or time-dependent policy for the minimizer $\tilde{\pi} = \{\pi_0, \pi_1, ...,\}$ where $\pi_i \in \Pi$ and denote the set of all possible nonstationary policies as $\tilde{\Pi}$ and similarly define for the maximizer. Then $V_n^*(x)$ is the value of the game when starting in state x, both players play their own equilibrium nonstationary policies for the n-horizon game with the terminal cost of V_0^* (see, e.g., [1, 46]) and is given by

$$V_n^*(x) = \inf_{\tilde{\pi}\in\tilde{\Pi}} \sup_{\tilde{\phi}\in\tilde{\Phi}} E\{\sum_{t=0}^{n-1} \gamma^t C_{x_t}(\pi_t(x_t), \phi_t(x_t)) + \gamma^n V_0^*(x_n)|x_0 = x\}$$

for $x \in X$.

2.1.2 Infinite horizon average cost

Unlike the discounted cost case, it is not true that there always exists an equilibrium value for average cost Markov games [21] in general. We make following assumption:

Assumption 2.1 *The Markov chain associated with each pair of any pure policies is irreducible and there exists $\rho > 0$ such that for any $\pi \in \Phi$ and $\phi \in \Phi$ and $x \in X$,*

$$P_{xx}(\pi(x), \phi(x)) \geq \rho.$$

The first assumption implies that the underlying Markov chain is a recurrent unichain and the second assumption is the strong aperiodicity[2] condition.

[2] This aperiodicity assumption is not a serious assumption (see, e.g., page 231 in [46]).

Under Assumption 2.1, there exists an *optimal equilibrium policy pair* $\pi^* \in \Pi$ and $\phi^* \in \Phi$ such that for all $\pi \in \Pi$ and $\phi \in \Phi$ and $x \in X$,

$$J_\infty(\pi^*, \phi)(x) \le J_\infty(\pi^*, \phi^*)(x) \le J_\infty(\pi, \phi^*), (x).$$

and in fact, each term in the above inequalities is independent of the initial state x so that we can omit x in each term above [46]. Furthermore, π^* (ϕ^*) here will be a different policy in general from the equilibrium policy for the discounted cost case. We will abuse the notation for our convenience and what we refer to will be clear from our presentation. We will refer to the value $J_\infty(\pi^*, \phi^*)$ as the *equilibrium value of the game* and to π^* and ϕ^* as the *equilibrium policies* for the minimizer and the maximizer, respectively, similar to the discounted case. We will write $J_\infty(\pi^*, \phi^*)$ as J^*_∞ and focus on finding or approximating the policy π^*. The dominance notion is also similarly defined: we will say that a policy $\pi_1 \in \Pi$ (weakly) *dominates* $\pi_2 \in \Pi$ if and only if for any $\phi \in \Phi$, $J_\infty(\pi_1, \phi) \le J_\infty(\pi_2, \phi)$.

We define several operators that map a function in $B(X)$ to a function in $B(X)$: for all $\pi \in \Pi, \phi \in \Phi, V \in B(X)$, and $x \in X$,

$$\begin{aligned}
\bar{T}(V)(x) &= \inf_{g \in G(x)} \sup_{f \in F(x)} \left\{ C_x(g,f) + \sum_{y \in X} P_{xy}(g,f) V(y) \right\} \\
\bar{T}_{\pi,\phi}(V)(x) &= C_x(\pi(x), \phi(x)) + \sum_{y \in X} P_{xy}(\pi(x), \phi(x)) V(y) \\
\bar{T}_{\phi}(V)(x) &= \inf_{g \in G(x)} \left\{ C_x(g, \phi(x)) + \sum_{y \in X} P_{xy}(g, \phi(x)) V(y) \right\} \\
\bar{T}_{\pi}(V)(x) &= \sup_{f \in F(x)} \left\{ C_x(\pi(x), f) + \sum_{y \in X} P_{xy}(\pi(x), f) V(y) \right\}
\end{aligned}$$

It is well-known (see, e.g., [18, 40, 46]) that each of the above operators has a *monotonicity* property and the infimum and supremum in the definitions of the operators $\bar{T}$, $\bar{T}_\phi$, and $\bar{T}_\pi$ are achieved by elements in $G(x)$ and $F(x)$.

Let $\{\bar{V}^*_n\}$ be the sequence of *value iteration functions* with respect to $\bar{T}$, $\bar{V}^*_n := \bar{T}(\bar{V}^*_{n-1})$ where $n = 1, 2, ...$ and $\bar{V}^*_0$ is arbitrary function in $B(X)$. It has been shown [46] (under Assumption 2.1 on average Markov games) that as $n \to \infty$, $\bar{V}^*_n$ converges to a function $\bar{V}^*_\infty \in B(X)$ that satisfies

$$\bar{T}(\bar{V}^*_\infty)(x) = J^*_\infty + \bar{V}^*_\infty(x) \text{ for all } x \in X.$$

Furthermore, $\bar{V}^*_n$ is the *equilibrium value of the finite n-horizon game without discount.* In this paper, we will assume that $\bar{V}^*_0(x) = 0$ for all $x \in X$.

3 Receding Horizon Control

3.1 Infinite horizon discounted cost

As we mentioned before, solving a large-state space Markov games for infinite horizon costs is often impractical. Therefore, we adopt a finite-horizon approximation scheme for the infinite horizon problem. We select a small but typical horizon and solve for the Markov game with the finite horizon (in our case, we are interested in only the optimal current or initial randomized actions for the minimizer and the maximizer). That is, we solve the Markov game with the total discounted cost criterion at each decision time. The intuition is that if the fixed horizon is "long" enough to get a stationary behavior of the system, this moving horizon control would have a good performance. Indeed, we show that the value of the game of the receding horizon control converges geometrically to the equilibrium value, uniformly in the initial state, as the value of the moving horizon increases.

The receding horizon control is simply defined as follows. Given a finite horizon $H \geq 1$, we define the receding H-horizon control as a policy $\pi^*_H \in \Pi$ for the minimizer and a policy $\phi^*_H \in \Phi$ for the maximizer such that $T_{\pi^*_H,\phi^*_H}(V^*_{H-1})(x) = T(V^*_{H-1})(x)$ for all $x \in X$. Note that the receding H-horizon control policy is a stationary policy. We have the following bound on the performance error.

Theorem 3.1 *For all $x \in X$,*

$$|V^*_\infty(x) - V_\infty(\pi^*_H, \phi^*_H)(x)| \leq \frac{\gamma^H(2-\gamma)}{(1-\gamma)^2} \cdot 2C_{\max}$$

Proof: See the proof of Theorem 3.6 below with $\epsilon = 0$ and $n = H - 1$. ∎

We remark that the same result can be obtained alternatively from Lemma 4.3.5 in page 181 in [18] via a simple algebraic manipulation.

From the theorem above, we can see that the receding horizon control gives a good approximation for the infinite horizon equilibrium policy for each player, and the value of the game using these policies approaches to the equilibrium performance for the infinite horizon cost geometrically in γ. Furthermore, by letting $\frac{\gamma^H(2-\gamma)}{(1-\gamma)^2} \cdot 2C_{\max} = \epsilon$, we can obtain the necessary value of the planning horizon which guarantees that the performance of the receding horizon control will be within ϵ of the equilibrium value.

The minimizer will play the game by the receding horizon control based on correlated equilibrium. That is, he assumes that the maximizer also plays the common copy of the receding horizon control. We need to analyze the error bound when the maximizer's play is the true worst case scenario, ϕ^*. We begin with a lemma regarding the monotonicity property of the $T_{\pi,\phi}$-operator.

Lemma 3.1 *For any $\pi \in \Pi$ and $\phi \in \Phi$, suppose there exists $\psi \in B(X)$ for which $T_{\pi,\phi}(\psi)(x) \leq \psi(x)$ for all $x \in X$; then $V_\infty(\pi,\phi)(x) \leq \psi(x)$ for all $x \in X$.*

The above lemma can be easily proven by the monotonicity property of the operator $T_{\pi,\phi}$ and the convergence to the unique fixed point of $V_\infty(\pi,\phi)$ from successive applications of the operator. The next lemma states that the function V_n^* is non-increasing in n under a suitable initial condition and is a simplified version of Lemma 3.1 in [45] in our context. We provide the proof for completeness.

Lemma 3.2 *Suppose V_0^* is selected such that $T(V_0^*)(x) \leq V_0^*(x)$ for all $x \in X$. Then, for $H = 1, 2, ...,$ and for all $x \in X$, $V_H^*(x) \leq V_{H-1}^*(x)$.*

Proof: The proof is by induction on H. For $H = 1$, since $V_1^* = T(V_0^*)$, we have $V_1^*(x) \leq V_0^*(x)$ for all $x \in X$ from the assumption.

Assuming that the assertion is true for $H = 1, ..., k$, we prove that it holds for $H = k+1$. For all $x \in X$,

$$\begin{aligned} V_{k+1}^*(x) &= T(V_k^*)(x) \\ &= T(T(V_{k-1}^*))(x) \\ &\leq T(V_{k-1}^*)(x) \text{ from the monotonicity of } T \text{ and the assumption} \\ &= V_k^*(x), \end{aligned}$$

which proves the claim. ∎

We remark that one such V_0^* can be simply given by $V_0^*(x) = C_{\max}/(1-\gamma)$ for all $x \in X$.

Theorem 3.2 *Suppose V_0^* is selected such that for all $x \in X$, $T(V_0^*)(x) \leq V_0^*(x)$. Then, for all $x \in X$,*

$$0 \leq V_\infty(\pi_H^*, \phi^*)(x) - V_\infty^*(x) \leq \frac{\gamma^H}{1-\gamma} \cdot 2C_{\max}$$

Proof: The lower bound is trivially true so that we prove the upper bound case.

$$\begin{aligned} T_{\pi_H^*,\phi^*}(V_H^*)(x) &= C_x(\pi_H^*(x), \phi^*(x)) + \gamma \sum_{y \in X} P_{xy}(\pi_H^*(x), \phi^*(x)) V_H^*(y) \\ &\leq C_x(\pi_H^*(x), \phi^*(x)) + \gamma \sum_{y \in X} P_{xy}(\pi_H^*(x), \phi^*(x)) V_{H-1}^*(y) \text{ by Lemma 3.2} \\ &\leq \sup_{f \in F(x)} \left\{ C_x(\pi_H^*(x), f) + \gamma \sum_{y \in X} P_{xy}(\pi_H^*(x), f) V_{H-1}^*(y) \right\} \\ &= C_x(\pi_H^*(x), \phi_H^*(x)) + \gamma \sum_{y \in X} P_{xy}(\pi_H^*(x), \phi_H^*(x)) V_{H-1}^*(y) \text{ by definition of } \phi_H^* \\ &= T_{\pi_H^*,\phi_H^*}(V_{H-1}^*)(x) = T(V_{H-1}^*)(x) = V_H^*(x) \end{aligned}$$

Therefore, by Lemma 3.1, $V_\infty(\pi^*_H, \phi^*)(x) \leq V^*_H(x)$ for all $x \in X$. It follows that for all $x \in X$,

$$V_\infty(\pi^*_H, \phi^*)(x) - V^*_\infty(x) \leq V^*_H(x) - V^*_\infty(x).$$

Observe that $\max_x |V^*_n(x)| \leq \frac{C_{\max}}{1-\gamma}$ for all $n \geq 0$ under the assumption of V^*_0. Therefore, for $n = 0, 1, ...,$

$$\max_x |V^*_\infty(x) - V^*_n(x)| \leq \gamma^n \max_x |V^*_\infty(x) - V^*_0(x)| \leq \frac{2C_{\max}}{1-\gamma} \cdot \gamma^n. \tag{1}$$

Combining the two inequalities, we have the desired result. ∎

As we expected, the error bound vanishes to zero as the size of the horizon increases to infinity geometrically fast with a given discount factor.

Consider the following condition: there exists a function δ defined on X such that $0 < \sum_{x \in X} \delta(x) < 1$ and $P_{xy}(f, g) \geq \delta(y)$ for all x, y, g, f. It turns out that if the given Markov game meets this condition, the error bounds in the above theorems can be improved by a factor $(1 - \sum_x \delta(x))^H$ as in the MDP case [23]. Let $\beta = 1 - \sum_x \delta(x)$. Now define two probability distributions P' and ψ such that

$$\begin{aligned} \psi(x) &= \frac{1}{1-\beta}\delta(x), x \in X. \\ P'_{xy}(f,g) &= \frac{1}{\beta}[P_{xy}(f,g) - (1-\beta)\psi(y)]. \end{aligned}$$

Then, we can define a transition probability P by

$$P_{xy}(f,g) = \beta P'_{xy}(f,g) + (1-\beta)\psi(y).$$

We further define the operator $T' : B(X) \rightarrow B(X)$ as in T except that we use P' instead of P. Then, for any function $v \in B(X)$, the T and T' operators are related by

$$T(v)(x) = T'(v)(x) + \gamma(1-\beta)\psi(v), x \in X.$$

where $\psi(v) = \sum_x \psi(x)v(x)$.

Let $\{V'_n\}$ be the sequence of value iteration functions with respect to T', $V'_n := T'(V'_{n-1})$ where $n = 1, 2, ...$ and set $V^*_0(x) = V'_0(x) = C_{\max}/(1-\gamma)$ for all $x \in X$. By induction, we can show that

$$V^*_n(x) = V'_n(x) + C_n, x \in X. \tag{2}$$

where C_n is the constant given by

$$\begin{aligned} C_n &= \gamma(1-\beta)\sum_{k=0}^{n-1}(\gamma\beta)^k \psi(V^*_{n-1-k}) \\ &= \gamma(1-\beta)\sum_{k=0}^{\infty}(\gamma\beta)^k \psi(V^*_{n-1-k}) \end{aligned} \tag{3}$$

setting V_k^* to the zero function for $k < 0$ if $n \geq 1$ and $C_0 = 0$. From this, we can conclude that (c.f., Lemma 4.1 in [23])

$$V_\infty^*(x) = V_\infty'(x) + C(\gamma, \beta)\psi(V_\infty^*), x \in X,$$

where $V_\infty' = \lim_{n\to\infty} T^n(V_0')$ and $C(\gamma,\beta) = \gamma(1-\beta)/(1-\gamma\beta)$. Observe that V_∞' is the optimal equilibrium value function for the underlying Markov game replaced with P' and discount factor $\gamma\beta$. By the same arguments, we can show that for any $\pi \in \Pi$ and $\phi \in \Phi$,

$$V_\infty(\pi,\phi)(x) = V_\infty'(\pi,\phi)(x) + C(\gamma,\beta)\psi(V_\infty(\pi,\phi)), x \in X.$$

This immediately implies that

$$V_\infty^*(x) - V_\infty(\pi_H^*, \phi_H^*)(x) = V_\infty'(x) - V_\infty'(\pi_H^*, \phi_H^*)(x) + C(\gamma,\beta)\psi(V_\infty^* - V_\infty(\pi_H^*, \phi_H^*)).$$

Observe that a policy pair $\pi \in \Pi$ and $\phi \in \Phi$ such that $T'_{\pi,\phi}(V'_{H-1})(x) = T'(V'_{H-1})(x)$ for all $x \in X$ prescribes the same randomized action choice as π_H^* and ϕ_H^* from the relationship given by Equation (2). Now, by majorization of $V_\infty^*(x) - V_\infty(\pi_H^*, \phi_H^*)(x)$ and from Theorem 3.1 with the observation just made, it follows that

$$\max_x[V_\infty^*(x) - V_\infty(\pi_H^*, \phi_H^*)(x)] \leq \frac{(\gamma\beta)^H(2-\gamma\beta)}{(1-\gamma\beta)^2}\cdot 2C_{\max} + C(\gamma,\beta)\max_x[V_\infty^*(x) - V_\infty(\pi_H^*, \phi_H^*)(x)], x \in X.$$

We can also minorize $V_\infty^*(x) - V_\infty(\pi_H^*, \phi_H^*)(x)$, from which we conclude that for all $x \in X$,

$$|V_\infty^*(x) - V_\infty(\pi_H^*, \phi_H^*)(x)| \leq [1 - C(\gamma,\beta)]^{-1} \cdot \frac{(\gamma\beta)^H(2-\gamma\beta)}{(1-\gamma\beta)^2} \cdot 2C_{\max}.$$

The upper bound on Theorem 3.2 can also be improved by a factor of β^H with the same arguments.

3.2 Infinite horizon average cost

The receding horizon control is defined as follows. Given a finite horizon $H \geq 1$, we define the receding H-horizon control as a policy $\pi_H^* \in \Pi$ for the minimizer and a policy $\phi_H^* \in \Phi$ for the maximizer such that $\bar{T}_{\pi_H^*, \phi_H^*}(\bar{V}_{H-1}^*)(x) = \bar{T}(\bar{V}_{H-1}^*)(x)$ for all $x \in X$. We now present the performance error of the receding horizon control in terms of the infinite horizon average cost comparing with the equilibrium value under our assumptions on Markov games. The analysis primarily builds on the work by Van der Wal [46]. We begin with a modified version of Van der Wal's Corollary 13.2 in page 230 in [46] within our context. For a function $v \in B(X)$, let span semi-norm of v be $\mathrm{sp}(v) = \max_x v(x) - \min_x v(x)$.

Theorem 3.3 *Assume that Assumption 2.1 holds. For any $V \in B(X)$, consider two policies $\pi \in \Pi$ and $\phi \in \Phi$ such that $\bar{T}_{\pi,\phi}(V)(x) = \bar{T}(V)(x)$ for all $x \in X$. Then, for any $\pi' \in \Pi$ and $\phi' \in \Phi$,*

$$\begin{aligned} J_\infty(\pi, \phi') &\leq J_\infty^* + \mathrm{sp}(T(V) - V) \\ J_\infty(\pi', \phi) &\geq J_\infty^* - \mathrm{sp}(T(V) - V) \end{aligned}$$

From now on, we will set $|X| = s$ (we naturally assume that $s \geq 1$). Under the aperiodicity assumption (the second part in Assumption 2.1), there exists a constant η, with $0 \leq \eta < 1$, such that the following scrambling condition holds: for any $\pi, \pi' \in \Pi$ and any $\phi, \phi' \in \Phi$ and for all $x, y \in X$,

$$\sum_{z \in X} \min\{\mathcal{P}^{s-1}_{x,\pi,\phi}(z), \mathcal{P}^{s-1}_{y,\pi',\phi'}(z)\} \geq 1 - \eta,$$

where $\mathcal{P}^{s-1}_{x,\pi,\phi}(y)$ denotes the probability that the initial state x will reach the state z in $s-1$ time steps under the policies π and ϕ. We will refer to η as an *ergodicity coefficient.*

Lemma 3.3 *Assume that Assumption 2.1 holds. For $n = 0, 1, ...,$* $\mathrm{sp}(\bar{V}^*_{n+1} - \bar{V}^*_n) \leq 2\eta^{\frac{n}{s-1}} C_{\max}$

Proof: Van der Wal showed that (see page 235 in [46]) $\mathrm{sp}(\bar{V}^*_{n+s} - \bar{V}^*_{n+s-1}) \leq \eta \cdot \mathrm{sp}(\bar{V}^*_{n+1} - \bar{V}^*_n)$, $n = 0, 1, ...,$ which implies that

$$\mathrm{sp}(\bar{V}^*_{n+1} - \bar{V}^*_n) \leq \eta^{\frac{n}{s-1}} \mathrm{sp}(\bar{V}^*_1 - \bar{V}^*_0) \text{ for } n = 0, 1, ...$$

Since $\mathrm{sp}(\bar{V}^*_1 - \bar{V}^*_0) \leq 2C_{\max}$ with $\bar{V}^*_0 = 0$, we have the desired result. ∎

The theorem and the lemma above yield immediately the following result.

Theorem 3.4 *Assume that Assumption 2.1 holds. Consider the receding H-horizon control $\pi^*_H \in \Pi$ for the minimizer and $\phi^*_H \in \Phi$ for the maximizer such that $\bar{T}_{\pi^*_H, \phi^*_H}(\bar{V}^*_{H-1})(x) = \bar{T}(\bar{V}^*_{H-1})(x)$ for all $x \in X$. Then,*

$$|J_\infty(\pi^*_H, \phi^*_H) - J^*_\infty| \leq 2\eta^{\frac{H-1}{s-1}} C_{\max}$$

We can see again that the receding horizon control for the average cost case also gives a good approximation for the infinite horizon equilibrium policy for each player and the value of the game by the policies approaches to the equilibrium performance for the infinite horizon average cost geometrically in the ergodicity coefficient η. Furthermore, by letting $2\eta^{\frac{H-1}{s-1}} C_{\max} = \epsilon$, we can obtain the necessary value of the planning horizon which guarantees that the performance of the receding horizon control will be within ϵ from the equilibrium value.

An error bound when the maximizer's play is the true worst cast scenario ϕ^* is also obtained directly from Theorem 3.3.

Theorem 3.5 *Assume that Assumption 2.1 holds. Consider the receding H-horizon control, $\pi^*_H \in \Pi$ for the minimizer such that $\bar{T}_{\pi^*_H}(\bar{V}^*_{H-1})(x) = \bar{T}(\bar{V}^*_{H-1})(x)$ for all $x \in X$. Then,*

$$0 \leq J_\infty(\pi^*_H, \phi^*) - J^*_\infty \leq 2\eta^{\frac{H-1}{s-1}} C_{\max}$$

The error bounds we presented above vanish geometrically to zero as the size of the horizon increases to infinity. However, it depends on the size of the state space. Therefore, if s is a huge number, the error bound will be large with relatively small H. We now add a new condition to the transition probability matrix so that we can eliminate the dependence on the size of the state space.

Assumption 3.1 *There exists a nonnegative function $\mu \in B(X)$ such that for some constant α, with $0 \leq \alpha < 1$,*

$$\sum_{y \in X} P_{xy}(\pi(y), \phi(y))\mu(y) \leq \alpha\mu(x)$$

for all $x \in X$, $\pi \in \Phi$, and $\phi \in \Phi$.

We will refer to this assumption as the μ-recurrent condition [17].

We define the μ-norm of a function $v \in B(X)$, $\|v\|_\mu$ given by

$$\|v\|_\mu = \inf\{c \in \mathcal{R} |\ |v(x)| \leq c\mu(x), \forall x \in X\}.$$

It is well-known that under the recurrent condition, $\bar{T}$ is a contraction mapping with respect to μ-norm. That is, for any $v, w \in B(X)$,

$$\|\bar{T}(v) - \bar{T}(w)\|_\mu \leq \alpha\|v - w\|_\mu.$$

Furthermore, it can then be easily proven (see, e.g., page 199 in [46]) that for any $x \in X$ and for any $v, w \in B(X)$,

$$-\alpha\mu(x)\|v - w\|_\mu \leq \bar{T}(v)(x) - \bar{T}(w)(x) \leq \alpha\mu(x)\|v - w\|_\mu.$$

It follows that for $n = 1, 2, ...$,

$$\mathrm{sp}(\bar{V}^*_{n+1} - \bar{V}^*_n) \leq 2\alpha \max_x \mu(x)\|\bar{V}^*_n - \bar{V}^*_{n-1}\|_\mu \leq 2\alpha^n \max_x \mu(x)\|\bar{V}^*_1 - \bar{V}^*_0\|_\mu = 2\alpha^n \max_x \mu(x)\|\bar{V}^*_1\|_\mu.$$

Because $\|\bar{V}^*_1\|_\mu \leq \frac{C_{\max}}{1-\alpha}$ (see, e.g., page 199 in [46]), we have the following immediate result with μ-recurrent condition.

Proposition 3.1 *Assume that Assumptions 2.1 and 3.1 hold. Consider the receding H-horizon control, $\pi^*_H \in \Pi$ for the minimizer and $\phi^*_H \in \Phi$ for the maximizer such that $\bar{T}_{\pi^*_H, \phi^*_H}(\bar{V}^*_{H-1})(x) = \bar{T}(\bar{V}^*_{H-1})(x)$ for all $x \in X$. Then under the μ-recurrent condition,*

$$\begin{aligned} |J_\infty(\pi^*_H, \phi^*_H) - J^*_\infty| &\leq \frac{\alpha^{H-1}}{1-\alpha} \cdot 2C_{\max} \max_x \mu(x) \\ 0 \leq J_\infty(\pi^*_H, \phi^*) - J^*_\infty &\leq \frac{\alpha^{H-1}}{1-\alpha} \cdot 2C_{\max} \max_x \mu(x). \end{aligned}$$

Therefore, the above theorem establishes the geometric convergence of the receding horizon control independently of the state space size. To apply the receding horizon control, we need to know the exact value of the finite horizon subgames. However, in practice, getting the true $(H-1)$-horizon equilibrium value, in order for the minimizer to get the receding H-horizon control policy, is also troublesome if the state-space size is huge. Motivated by this, we now analyze the *approximate* receding horizon control.

3.3 Analysis of approximate receding horizon control

3.3.1 Infinite horizon discounted cost

We start with lemmas to state our main result for the approximate receding horizon control.

Lemma 3.4 *For all $x \in X$ and $n = 0, 1, ...,$*

$$|V^*_{n+1}(x) - V^*_n(x)| \leq \frac{\gamma^n}{1-\gamma} \cdot 2C_{\max}$$

Proof: This is directly obtained from the contraction mapping property. ∎

The theorem below states an error bound from the equilibrium value of the game when both the minimizer and the maximizer play the receding horizon control based on the same approximate value, i.e., correlated equilibrium policies.

Theorem 3.6 *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 for the minimizer and ϕ for the maximizer such that for all $x \in X$, $T_{\pi,\phi}(V)(x) = T(V)(x)$. Then, for all $x \in X$,*

$$|V^*_\infty(x) - V_\infty(\pi, \phi)(x)| \leq \frac{\gamma^{n+1}(2-\gamma)}{(1-\gamma)^2} \cdot 2C_{\max} + \frac{2\gamma\epsilon}{1-\gamma}$$

Proof: From the contraction mapping property of the T operator, for all x in X,

$$|T(V^*_n)(x) - T(V)(x)| \leq \gamma \cdot \max_x |V^*_n(x) - V(x)| \leq \gamma\epsilon \tag{4}$$

and from $T(V^*_\infty) = V^*_\infty$ and a successive application of the contraction property we have

$$\max_x |V^*_\infty(x) - V^*_{n+1}(x)| \leq \gamma^{n+1} \max_x |V^*_\infty(x) - V^*_0(x)| \leq \frac{2C_{\max}}{1-\gamma} \cdot \gamma^{n+1}. \tag{5}$$

Therefore, from Equation (4) and (5) and $V^*_{n+1} = T(V^*_n)$ by definition, for all $x \in X$,

$$\begin{aligned} |V^*_\infty(x) - T(V)(x)| &\leq |V^*_\infty(x) - T(V^*_n)(x)| + |T(V^*_n)(x) - T(V)(x)| \\ &\leq \frac{2C_{\max}}{1-\gamma} \cdot \gamma^{n+1} + \gamma\epsilon. \end{aligned} \tag{6}$$

Below we show that $|T(V)(x) - V_\infty(\pi,\phi)(x)| \leq \frac{\gamma\epsilon(1+\gamma)}{1-\gamma} + \frac{2C_{\max}\gamma^{n+1}}{(1-\gamma)^2}$ for all $x \in X$. It then follows that from Equation (6), for all $x \in X$,

$$\begin{aligned} |V^*_\infty(x) - V_\infty(\pi,\phi)(x)| &\leq |V^*_\infty(x) - T(V)(x)| + |T(V)(x) - V\infty(\pi,\phi)(x)| \\ &\leq \frac{2C_{\max}}{1-\gamma}\cdot\gamma^{n+1} + \gamma\epsilon + \frac{\gamma\epsilon(1+\gamma)}{1-\gamma} + \frac{2C_{\max}\gamma^{n+1}}{(1-\gamma)^2} \\ &= \frac{\gamma^{n+1}(2-\gamma)}{(1-\gamma)^2}\cdot 2C_{\max} + \frac{2\gamma\epsilon}{1-\gamma}, \end{aligned}$$

which gives the desired result.

From Lemma 3.4 and Equation (4), we have that for all $x \in X$, by letting $w = \frac{\gamma^n}{1-\gamma}\cdot 2C_{\max}$,

$$V(x) \leq V^*_n(x) + \epsilon \leq V^*_{n+1}(x) + \epsilon + w \leq T(V)(x) + \gamma\epsilon + \epsilon + w. \tag{7}$$

Then for all $x \in X$,

$$\begin{aligned} T(V)(x) &= T_{\pi,\phi}(V)(x) = C_x(\pi(x),\phi(x)) + \gamma\sum_{y\in X} P_{xy}(\pi(x),\phi(x))V(y) \text{ by definitions of } \pi \text{ and } \phi \text{ and } T \\ &\leq C_x(\pi(x),\phi(x)) + \gamma\sum_{y\in X} P_{xy}(\pi(x),\phi(x))[T(V)(y) + \gamma\epsilon + \epsilon + w] \text{ by Equation (7)} \\ &= C_x(\pi(x),\phi(x)) + \gamma\sum_{y\in X} P_{xy}(\pi(x),\phi(x))T(V)(y) + \gamma\epsilon(1+\gamma) + \gamma w \\ &= C_x(\pi(x),\phi(x)) + \gamma\sum_{y\in X} P_{xy}(\pi(x),\phi(x))\left(C_y(\pi(y),\phi(y)) + \gamma\sum_{z\in X} P_{yz}(\pi(y),\phi(y))V(z)\right) \\ &\quad +\gamma\epsilon(1+\gamma) + \gamma w \\ &= C_x(\pi(x),\phi(x)) + \gamma\sum_{y\in X} P_{xy}(\pi(x),\phi(x))C_y(\pi(y),\phi(y)) \\ &\quad +\gamma^2\sum_{y\in X}\sum_{z\in X} P_{xy}(\pi(x),\phi(x))P_{yz}(\pi(y),\phi(y))V(z) + \gamma\epsilon(1+\gamma) + \gamma w \\ &\leq C_x(\pi(x),\phi(x)) + \gamma\sum_{y\in X} P_{xy}(\pi(x),\phi(x))C_y(\pi(y),\phi(y)) \\ &\quad +\gamma^2\sum_{y\in X}\sum_{z\in X} P_{xy}(\pi(x),\phi(x))P_{yz}(\pi(y),\phi(y))T(V)(z) \\ &\quad +\gamma^2\epsilon(1+\gamma) + \gamma\epsilon(1+\gamma) + (\gamma^2 w + \gamma w) \end{aligned}$$

Keep iterating (under the sum sign) this way, we have that for all $k = 0, 1, ...,$ and $x \in X$,

$$\begin{aligned} T(V)(x) \leq E\left[\sum_{t=0}^{k}\gamma^t C_{x_t}(\pi(x_t),\phi(x_t))|x_0 = x\right] &+ \gamma^{k+1}E[T(V)(x_{k+1})|x_0 = x] \\ &+\gamma\epsilon(1+\gamma) + \cdots + \gamma^{k+1}\epsilon(1+\gamma) + (\gamma w + \cdots + \gamma^{k+1}w), \end{aligned} \tag{8}$$

where x_t is the random variable representing the state at time t under π and ϕ. Since $T(V)$ is bounded, the second term on the r.h.s. of Equation (8) converges to zero as $k \to \infty$ and the first term becomes $V_\infty(\pi,\phi)(x)$. Therefore it follows that $T(V)(x) - V_\infty(\pi,\phi)(x) \leq \frac{\gamma\epsilon(1+\gamma)}{1-\gamma} + \frac{\gamma w}{1-\gamma}$. Therefore, $T(V)(x) - V_\infty(\pi,\phi)(x) \leq \frac{\gamma\epsilon(1+\gamma)}{1-\gamma} + \frac{2C_{\max}\gamma^{n+1}}{(1-\gamma)^2}$ for all $x \in X$.

Similarly, we can show that $T(V)(x) - V_\infty(\pi,\phi)(x) \geq -\frac{\gamma\epsilon(1+\gamma)}{1-\gamma} - \frac{2C_{\max}\gamma^{n+1}}{(1-\gamma)^2}$ for all $x \in X$ by the observation that from the assumption and Equation (4), we have that for all $x \in X$,

$$V(x) \geq V_n^*(x) - \epsilon \geq V_{n+1}^*(x) - \epsilon - w \geq T(V)(x) - \gamma\epsilon - \epsilon - w.$$

∎

From the approximate receding horizon control framework, given an approximate function V, the minimizer will play the policy π such that $T_{\pi,\phi} = T(V)$ at each $x \in X$. That is, he will assume that the maximizer will play the correlated equilibrium policy with respect to V. We now present the game of value when the maximizer *actually* plays the worst-case scenario.

Theorem 3.7 *Suppose V_0^* is selected such that for all $x \in X$, $T(V_0^*)(x) \leq V_0^*(x)$. 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 for the minimizer such that for all $x \in X$, $T_\pi(V)(x) = T(V)(x)$. Then, for all $x \in X$,*

$$0 \leq V_\infty(\pi,\phi^*)(x) - V_\infty^*(x) \leq \frac{\gamma^{n+1}}{1-\gamma} \cdot 2C_{\max} + \frac{2\gamma\epsilon}{1-\gamma}$$

Before we provide a proof of this theorem, we mention here that setting $\epsilon = 0$ with $n = H-1$ gives exactly the bound of Theorem 3.2. Even though we could have obtained the result for Theorem 3.2 by setting $\epsilon = 0$ with $n = H-1$ here, we wanted to show that there is an alternate but simpler proof than the proof below.

Proof: The lower bound is trivially true so we prove the upper bound. The proof technique is quite similar to the proof of the previous theorem.

For all $x \in X$, $V_\infty(\pi,\phi^*)(x) - V_\infty^*(x) = V_\infty(\pi,\phi^*)(x) - T(V)(x) + T(V)(x) - V_\infty^*(x)$. We have that $T(V)(x) - V_\infty^*(x) \leq \gamma\epsilon + \frac{\gamma^{n+1}2C_{\max}}{1-\gamma}$ (see the proof of the previous theorem). It remains to show that $V_\infty(\pi,\phi^*) - T(V)(x) \leq \frac{\gamma\epsilon(1+\gamma)}{1-\gamma}$.

Now, for all $x \in X$, $-\gamma\epsilon + T(V)(x) \leq V_{n+1}^*(x) \leq V_n^*(x) \leq V(x) + \epsilon$, where the first inequality is from Equation (4) and the second inequality is from Lemma 3.2 and the third inequality is from the assumption. It follows that

$$\begin{aligned} T(V)(x) &= T_\pi(V)(x) = \sup_{f \in F(x)} \left\{ C_x(\pi(x), f) + \gamma \sum_{y \in X} P_{xy}(\pi(x), f) V(y) \right\} \\ &\geq C_x(\pi(x), \phi^*(x)) + \gamma \sum_{y \in X} P_{xy}(\pi(x), \phi^*(x)) V(y) \\ &\geq C_x(\pi(x), \phi^*(x)) + \gamma \sum_{y \in X} P_{xy}(\pi(x), \phi^*(x)) [T(V)(y) - \epsilon(1+\gamma)] \end{aligned}$$

Keep iterating (under the sum sign) this way, we have that for all $k = 0, 1, ...,$ and $x \in X$,

$$T(V)(x) \geq E\left[\sum_{t=0}^{k} \gamma^t C_{x_t}(\pi(x_t), \phi^*(x_t))|x_0 = x\right] + \gamma^{k+1} E[T(V)(x_{k+1})|x_0 = x] \\ -[\gamma\epsilon(1+\gamma) + \cdots + \gamma^{k+1}\epsilon(1+\gamma)], \quad (9)$$

where x_t is the random variable representing the state at time t under π and ϕ^*. Since $T(V)$ is bounded, the second term on the r.h.s. of Equation (9) converges to zero as $k \to \infty$ and the first term becomes $V_\infty(\pi, \phi)(x)$. Therefore it follows that $T(V)(x) - V(\pi, \phi^*)(x) \geq -\frac{\gamma\epsilon(1+\gamma)}{1-\gamma}$. ∎

As we have studied in subsection 3.1, if there exists a function δ defined on X such that $0 < \sum_{x\in X} \delta(x) < 1$ and $P_{xy}(f, g) \geq \delta(y)$ for all x, y, g, f, the error bounds above can be improved. Let $\beta = 1 - \sum_x \delta(x)$ again. We only present the upper bound case of Theorem 3.6 as an example. With the same arguments given in subsection 3.1,

$$\max_x [V_\infty^*(x) - V_\infty(\pi, \phi)(x)] \leq [1 - C(\gamma, \beta)]^{-1} \max_x [V'_\infty(x) - V'_\infty(\pi, \phi)(x)].$$

We have

$$\max_x [V'_\infty(x) - V'_\infty(\pi, \phi)(x)] \leq \frac{(\gamma\beta)^{n+1}(2-\gamma\beta)}{(1-\gamma\beta)^2} \cdot 2C_{\max} + \frac{2\gamma\beta\epsilon'}{1-\gamma\beta}$$

if $V(x) - V'_n(x) \leq \epsilon'$. But $\epsilon' = \epsilon + C_n$ where C_n is given in Equation (3).

3.3.2 Infinite horizon average cost

Theorem 3.8 *Assume that Assumption 2.1 holds. Given $V \in B(X)$ such that for some $n \geq 0$, $|\bar{V}_n^*(x) - V(x)| \leq \epsilon$ for all x in X, consider a policy π for the minimizer and ϕ for the maximizer such that for all $x \in X$, $\bar{T}_{\pi,\phi}(V)(x) = \bar{T}(V)(x)$. Then, for all $x \in X$,*

$$\begin{aligned} |J_\infty^*(\pi, \phi) - J_\infty^*| &\leq 2\eta^{\frac{n}{s-1}} C_{\max} + 4\epsilon. \\ 0 \leq J_\infty^*(\pi, \phi^*) - J_\infty^* &\leq 2\eta^{\frac{n}{s-1}} C_{\max} + 4\epsilon. \end{aligned}$$

Proof: From the assumption, $-\epsilon \leq \bar{V}_n^*(x) - V(x) \leq \epsilon$ for all $x \in X$. Applying the $\bar{T}$-operator to each side and using the monotonicity property, we have $-\epsilon \leq \bar{T}(V_n^*)(x) - \bar{T}(V)(x) \leq \epsilon$ for all $x \in X$. Therefore we have that

$$\text{sp}(\bar{T}(V) - V) \leq \text{sp}(\bar{V}_{n+1}^* - \bar{V}_n^*) + 4\epsilon.$$

Applying Theorem 3.3, we have the result. The error bound on the value of the game when the maximizer *actually* plays the worst-case scenario is also directly obtained from Theorem 3.3. ∎

We remark that we can add the μ-recurrent condition (Assumption 3.1) to this case also so that we can eliminate the dependence on the state space size as we did previously.

4 Examples of Approximate Receding Horizon Control

In this section, we introduce three approaches as examples of approximate receding horizon control for the Markov games. These are heuristics for the minimizer who seeks to optimize his performance under the guess of the worst-case scenario from the opponent's play. The first two approaches (to the minimizer) aim at improving a given heuristic policy (or a set of multiple heuristic policies) that is available to the minimizer, based on the policy improvement arguments. The final approach is motivated by hindsight optimization proposed in [13, 16]. By this approach, at each state, the minimizer evaluates his candidate randomized actions based on the analysis of the expected optimal hindsight performance over a finite horizon under the guess that the maximizer plays the worst-case fixed policy chosen by the minimizer.

4.1 Rollout algorithm

Our discussion in this subsection will focus on the discounted case first and then consider the average case. To the minimizer, obtaining an equilibrium policy for him is often quite difficult due to *the curse of dimensionality.* One approach to take when a heuristic policy is available to the minimizer is to assume that the maximizer has chosen a fixed policy $\phi \in \Phi$ to play the given Markov game and then to try to improve the heuristic policy of the minimizer. Because it is also difficult for the minimizer to get the worst case policy (the equilibrium policy for the maximizer), the minimizer will need to choose a heuristic worst case policy for the maximizer. For some cases, we can actually get ϕ^* (see, e.g., [1] and the references therein). If we fix the maximizer's policy, the resulting game becomes a Markov decision process to the minimizer. It is well-known from the policy improvement principle that given a policy π, if we define a new policy π_{ro} such that $T_{\pi_{ro},\phi}(V_\infty(\pi,\phi))(x) = T_\phi(V_\infty(\pi,\phi))(x)$ for all $x \in X$, the new policy π_{ro} improves the policy π in terms of the infinite horizon discounted cost. That is, $V_\infty(\pi_{ro},\phi) \leq V_\infty(\pi,\phi)$. Because this holds for arbitrary $\phi \in \Phi$, π_{ro} dominates π.

Several works for MDP problems (with their related cost function) in this respect have reported successful results. For example, Bertsekas and Castanon consider stochastic scheduling problems [9], Secomandi [43] studied a vehicle routing problem, Ott and Krishnan [37] and Kolarov and Hui [30] studied network routing problems, Bhulai and Koole [11] consider a multi-server queueing problem, and Koole and Nain [32] consider a two-class single-server queueing model under a preemptive priority rule. In particular, [11] and [32] obtain explicit expressions for the value function of a fixed threshold policy, which plays the role of a heuristic base policy, and showed numerically that the rollout of the policy behaves almost optimally. Chang *et al.* [14] also empirically showed the rollout of a fixed threshold policy (Droptail) works well for a buffer management problem. Koole [31] 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. Note that in queueing systems viewed as Markov games, we can consider a worst-case arrival process and then analyze the value function of a certain fixed policy with the worst-case arrival process, from which we generate a rollout policy for the minimizer.

As a receding horizon approach for this improvement scheme, we replace $V_\infty(\pi,\phi)$ by the value of the game when the policies π and ϕ are followed over a finite horizon. Formally, we define the H-horizon rollout policy $\pi_{ro,H}$ with a base policy π to be a policy $\pi_{ro,H}$ that satisfies $T_{\pi_{ro,H},\phi}(V_{H-1}(\pi,\phi))(x) = T_\phi(V_{H-1}(\pi,\phi))(x)$ for all $x \in X$ where $V_{H-1}(\pi,\phi) := E\{\sum_{t=0}^{H-2}\gamma^t C_{x_t}(\pi(x_t),\phi(x_t))|x_0 = x\}$.

We present the result regarding the H-horizon rollout policy adapted from [13] and provide the proof for completeness. We first begin with a lemma similar to Lemma 3.2.

Lemma 4.1 *Suppose $V_0(\pi,\phi)$ is selected such that for all $x \in X$, $T_{\pi,\phi}(V_0(\pi,\phi))(x) \le V_0(\pi,\phi)(x)$. Then, for $H = 1, 2, ...,$ and for all $x \in X$, $V_H(\pi,\phi)(x) \le V_{H-1}(\pi,\phi)(x)$.*

Proof: The statement can be proven by induction on H as in the proof of Lemma 3.2. ∎

Proposition 4.1 *Given a fixed policy $\phi \in \Phi$ and a base policy $\pi \in \Pi$ for the minimizer, suppose $V_0(\pi,\phi)$ is selected such that $T_{\pi,\phi}(V_0(\pi,\phi))(x) \le V_0(\pi,\phi)(x)$ for all $x \in X$. For any $\epsilon > 0$, if $H \ge 1 + \log_\gamma \frac{\epsilon(1-\gamma)}{C_{\max}}$, then for all $x \in X$, $V_\infty(\pi_{ro,H},\phi)(x) \le V_\infty(\pi,\phi)(x) + \epsilon$.*

Proof: Define $\psi = V_{H-1}(\pi,\phi)$. By definition of the rollout policy,

$$\begin{aligned} T_{\pi_{ro,H},\phi}(\psi)(x) &= T_\phi(\psi)(x) = C_x(\pi_{ro,H}(x),\phi(x)) + \gamma\sum_{y\in X} P_{xy}(\pi_{ro,H}(x),\phi(x))\psi(y) \\ &\le C_x(\pi(x),\phi(x)) + \gamma\sum_{y\in X} P_{xy}(\pi(x),\phi(x))\psi(y) = V_H(\pi,\phi)(x) \le \psi(x), \end{aligned}$$

where the last inequality follows from Lemma 4.1. Therefore, for all $x \in X$, we have $V_\infty(\pi_{ro,H},\phi)(x) \le V_{H-1}(\pi,\phi)(x)$ by Lemma 3.1. Now we can write for all $x \in X$, $V_\infty(\pi,\phi)(x) = V_{H-1}(\pi,\phi)(x) + \gamma^{H-1}E[V_\infty(\pi,\phi)(x_{H-1})|x_0 = x]$. We know that $\min_x[V_\infty(\pi,\phi)(x)] \ge -\frac{C_{\max}}{1-\gamma}$. This implies that $V_\infty(\pi_{ro,H},\phi)(x) \le V_\infty(\pi,\phi)(x) + \frac{C_{\max}}{1-\gamma}\cdot\gamma^{H-1}$. Letting $\frac{C_{\max}}{1-\gamma}\cdot\gamma^{H-1} \le \epsilon$ yields the desired result. ∎

We note again that the minimizer is *assuming* the maximizer's play. If the minimizer's guess on the worst-cast scenario is good in the sense that $\max_x |V_{H-1}(\pi,\phi)(x) - V^*_{H-1}(x)| \le \epsilon$ with a relative small value, the resulting performance will be bounded by Theorem 3.7 from the optimal equilibrium performance.

The average case is similar to the discounted case except that we define the rollout policy with respect to "$\bar{T}$"-operators — the rollout policy is defined as a policy such that

$\bar{T}_{\pi_{ro,H},\phi}(\bar{V}_{H-1}(\pi,\phi))(x) = \bar{T}_{\phi}(\bar{V}_{H-1}(\pi,\phi))(x)$ for all $x \in X$ where $\bar{V}_{H-1}$ is obtained with $\gamma = 1$ in V_{H-1} and we assume that $\bar{V}_0(\pi,\phi)$ is zero function. The principle behind this is also the policy improvement scheme (see, e.g., [26]) under the assumptions we made for the average Markov games, i.e., aperiodicity and irreducibility.

Proposition 4.2 *Assume that Assumption 2.1 holds. Consider the H-horizon rollout policy $\pi_{ro,H}$ with a base policy π with respect to $\phi \in \Phi$. Then*

$$J_\infty(\pi_{ro,H},\phi) \le J_\infty(\pi,\phi) + 2\eta^{\frac{H-1}{s-1}} C_{\max}.$$

To prove the above proposition, we start with a lemma, which can be proven by the invariance property [23] of the stationary distribution of the underlying Markov chain (see, e.g., [15]). Note that under our assumptions, there exists a stationary distribution over X under any policy pair.

Lemma 4.2 *For any $\pi \in \Pi$ and $\phi \in \Phi$, a stationary distribution $P^{\pi,\phi}$ over X exists, and for all $n = 0, 1, ...,$*

$$J_\infty(\pi,\phi) = \sum_{y\in X} [\bar{V}_{n+1}(\pi,\phi)(y) - \bar{V}_n(\pi,\phi)(y)] P^{\pi,\phi}(y).$$

In particular, given $V \in B(X)$ and $\phi \in \Phi$, if π is defined such that $T_{\pi,\phi}(V) = T_\phi(V)(x)$ for all $x \in X$, then

$$J_\infty(\pi,\phi) = \sum_{y\in X} [T_\phi(V)(y) - V(y)] P^{\pi,\phi}(y).$$

Lemma 4.3 *For $n = 0, 1, ...,,$ and any $\pi \in \Pi$ and $\phi \in \Phi$,*

$$\max_x [\bar{V}_{n+1}(\pi,\phi)(x) - \bar{V}_n(\pi,\phi)(x)] \le J_\infty(\pi,\phi) + 2\eta^{\frac{n}{s-1}} C_{\max}$$

Proof: As in the statement of Lemma 3.3, we can show that for $n = 0, 1, ...,$ $\text{sp}(\bar{V}_{n+1}(\pi,\phi) - \bar{V}_n(\pi,\phi)) \le 2\eta^{\frac{n}{s-1}} C_{\max}$ by the similar reasoning to that given in page 234–235 in [46]. By Lemma 4.2,

$$\min_x [\bar{V}_{n+1}(\pi,\phi)(x) - \bar{V}_n(\pi,\phi)(x)] \le J_\infty(\pi,\phi) \le \max_x [\bar{V}_{n+1}(\pi,\phi)(x) - \bar{V}_n(\pi,\phi)(x)].$$

It follows that $\max_x [\bar{V}_{n+1}(\pi,\phi)(x) - \bar{V}_n(\pi,\phi)(x)] - J_\infty(\pi,\phi) \le \text{sp}(\bar{V}_{n+1}(\pi,\phi) - \bar{V}_n(\pi,\phi))$. Therefore the result follows. ∎

We are now ready to prove the proposition above.

Proof:

$$\begin{aligned}
J_\infty(\pi_{ro,H},\phi) &= \sum_x [\bar{T}_\phi(\bar{V}_{H-1}(\pi,\phi))(x) - \bar{V}_{H-1}(\pi,\phi)(x)]P^{\pi_{ro,H},\phi}(x) \text{ from Lemma 4.2} \\
&\leq \max_x(\bar{T}_\phi(\bar{V}_{H-1}(\pi,\phi))(x) - \bar{V}_{H-1}(\pi,\phi)(x)) \\
&\leq \max_x(\bar{V}_H(\pi,\phi)(x) - \bar{V}_{H-1}(\pi,\phi)(x)) \\
&\leq J_\infty(\pi,\phi) + 2\eta^{\frac{H-1}{s-1}} C_{\max} \text{ from Lemma 4.3}
\end{aligned}$$

■

Therefore, if $H \geq 1 + (s-1)\log_\eta \frac{\epsilon}{C_{\max}}$, the rollout policy dominates the heuristic base policy by ϵ. By adding the μ-recurrent condition, the similar result can be obtained.

4.2 Parallel rollout

When a good heuristic policy is available to the minimizer and a fixed worst-case policy can be assumed for the maximizer, the performance of the rollout policy played by the minimizer will be promising because it will improve the performance of the heuristic policy for the minimizer. However, often getting a good heuristic policy to roll out is very difficult. This will be particularly true for the case where for some trajectories of the states, a heuristic policy is good and for other trajectories of the states, another heuristic policy is good, etc. As a simple example, for a multiclass scheduling problem where the cost is a function of the delay and the (importance) weight of the class, the performances of the static priority policy and the earliest deadline first policy depend on the system trajectories (see [13] for a detailed discussion).

As a generalization of the rollout approach, we consider a finite set of multiple heuristic policies. The minimizing player seeks to combine dynamically the given heuristic policies in the set to adapt to the different trajectories of the system to improve the performance of all policies in the set under the assumption that the maximizing player plays a fixed worst-case policy chosen by the minimizer. As in the rollout algorithm discussion, we first study the discounted cost case and then the average cost case.

As we mentioned before, if we fix the maximizer's policy, the resulting game becomes a Markov decision process to the minimizer. Consider a finite set $\Lambda \subset \Pi$. It has been shown in [13] that if we define a new policy π_{pr} such that $T_{\pi_{pr},\phi}(\min_{\pi\in\Lambda} V_\infty(\pi,\phi))(x) = T_\phi(\min_{\pi\in\Lambda} V_\infty(\pi,\phi))(x)$ for all $x \in X$, where min is defined componentwise on X, the new policy π_{pr} improves all of the policies in Λ in terms of the infinite horizon discounted cost (to see this, we simply show that $T_{\pi_{pr},\phi}(\min_{\pi\in\Lambda} V_\infty(\pi,\phi))(x) \leq \min_{\pi\in\Lambda} V_\infty(\pi,\phi)(x)$ for all $x \in X$). That is, for all $x \in X$, $V_\infty(\pi_{pr},\phi)(x) \leq \min_{\pi\in\Lambda} V_\infty(\pi,\phi)(x)$. Therefore, π_{pr} dominates any policy $\pi \in \Lambda$.

As we have done for the rollout approach, we define formally the H-horizon parallel rollout policy $\pi_{pr,H}$ with a finite set Λ of base policies $\pi \in \Pi$ to be a policy such that $T_{\pi_{pr,H},\phi}(\min_{\pi\in\Lambda} V_{H-1}(\pi,\phi))(x) = T_\phi(\min_{\pi\in\Lambda} V_{H-1}(\pi,\phi))(x)$ for all $x \in X$.

We now give the main result regarding the H-horizon parallel rollout policy. It states that the parallel rollout policy dominates any policy in Λ by a small error, which is determined by the receding horizon size.

Proposition 4.3 *Let $\Lambda \subset \Pi$ be a nonempty finite set of stationary policies. Given a fixed policy $\phi \in \Phi$ for the maximizer, suppose for each $\pi \in \Lambda$, $V_0(\pi,\phi)$ is selected such that for all $x \in X$, $T_{\pi,\phi}(V_0(\pi,\phi))(x) \le V_0(\pi,\phi)(x)$. For $\pi_{pr,H}$ defined on Λ and played by the minimizer, given any $\epsilon > 0$, if $H \ge 1 + \log_\gamma \frac{\epsilon(1-\gamma)}{C_{\max}}$, then for all $x \in X$, $V_\infty(\pi_{pr,H},\phi)(x) \le \min_{\pi\in\Lambda} V_\infty(\pi,\phi)(x) + \epsilon$.*

Proof: The idea of the proof is similar to that of Proposition 4.1. We define $\psi(x) = \min_{\pi\in\Lambda} V_{H-1}(\pi,\phi)(x)$ for all $x \in X$.

$$\begin{aligned} T_{\pi_{pr,H},\phi}(\psi)(x) &= T_\phi(\psi)(x) = C_x(\pi_{pr,H}(x),\phi(x)) + \gamma \sum_{y\in X} P_{xy}(\pi_{pr,H}(x),\phi(x))\psi(y) \\ &\le C_x(\pi(x),\phi(x)) + \gamma \sum_{y\in X} P_{xy}(\pi(x),\phi(x)) V_{H-1}(\pi,\phi)(y) \\ &\qquad\qquad \text{for any } \pi \in \Lambda \text{ from the definition of } \pi_{pr,H} \\ &= V_H(\pi,\phi)(x) \le V_{H-1}(\pi,\phi)(x) \text{ by the given assumption and Lemma 3.2} \end{aligned}$$

It follows that $T_{\pi_{pr,H},\phi}(\psi)(x) \le \psi(x)$ for all $x \in X$. Therefore, for all $x \in X$, we have $V_\infty(\pi_{pr,H},\phi)(x) \le \min_{\pi\in\Lambda} V_{H-1}(\pi,\phi)(x)$ by Lemma 3.1. We know that $V_\infty(\pi_{pr,H},\phi)(x) \le \min_{\pi\in\Lambda} V_\infty(\pi,\phi)(x) + \frac{C_{\max}}{1-\gamma}\cdot\gamma^{H-1}$ (c.f., Proposition 4.1). Letting $\frac{C_{\max}}{1-\gamma}\cdot\gamma^{H-1} \le \epsilon$ yields the desired result. ▮

For the average cost case, the definition of the parallel rollout policy in the discounted case is replaced with "$\bar{T}$"-operator and $\bar{V}_{H-1}$. That is, the H-horizon parallel rollout policy $\pi_{pr,H}$ with a finite set Λ of base policies $\pi \in \Pi$ with respect to a policy $\phi \in \Phi$ is defined as a policy such that $\bar{T}_{\pi_{pr,H},\phi}(\min_{\pi\in\Lambda} \bar{V}_{H-1}(\pi,\phi))(x) = \bar{T}_\phi(\min_{\pi\in\Lambda} \bar{V}_{H-1}(\pi,\phi))(x)$ for all $x \in X$.

We first analyze the performance of the H-horizon parallel rollout policy compared with those obtained by policies in Λ. For this purpose, for any $\pi \in \Pi$ and $\phi \in \Phi$, define $J_n^{\pi,\phi}(x) = \frac{\bar{V}_n(\pi,\phi)(x)}{n}$ for all $x \in X$ and $n = 1, 2, ...$ That is, this is the n-horizon approximation of the value of the game for the average cost when the minimizer plays π and the maximizer plays ϕ. With similar arguments as Platzman's given in Section 3.3 in [41], we can show that $J_n^{\pi,\phi}(x)$ converges, uniformly in x, as $O(n^{-1})$, to $J_\infty(\pi,\phi)$, $n = 1, 2, ...$

Theorem 4.1 *Assume that Assumption 2.1 holds. Consider the H-horizon parallel rollout policy*

$\pi_{ro,H}$ *with a finite set* $\Lambda \subset \Pi$ *with respect to a policy* $\phi \in \Phi$. *Then*

$$J_\infty(\pi_{pr,H}, \phi) \le \sum_x J_\infty(\arg\min_{\pi\in\Lambda} J^{\pi,\phi}_{H-1}(x), \phi) P^{\pi_{pr,H},\phi}(x) + 2\eta^{\frac{H-1}{s-1}} C_{\max}.$$

Proof: We first observe that

$$\begin{aligned}
\bar{T}_\phi(\min_{\pi\in\Lambda} \bar{V}_{H-1}(\pi,\phi))(x) &= \bar{T}_{\pi_{pr,H},\phi}(\min_{\pi\in\Lambda} \bar{V}_{H-1}(\pi,\phi))(x) \text{ by definition of } \pi_{pr,H} \\
&= C_x(\pi_{pr,H}(x), \phi(x)) + \sum_{y\in X} P_{xy}(\pi_{pr,H}(x), \phi(x)) \min_{\pi\in\Lambda} \bar{V}_{H-1}(\pi,\phi)(y) \\
&\le C_x(\pi(x), \phi(x)) + \sum_{y\in X} P_{xy}(\pi(x), \phi(x)) \bar{V}_{H-1}(\pi,\phi)(y) \text{ for any } \pi \in \Lambda \\
&= \bar{V}_H(\pi,\phi)(x).
\end{aligned}$$

Therefore, for all $x \in X$, $\bar{T}_\phi(\min_{\pi\in\Lambda} \bar{V}_{H-1}(\pi,\phi))(x) \le \min_{\pi\in\Lambda} \bar{V}_H(\pi,\phi)(x)$. Now,

$$\begin{aligned}
J_\infty(\pi_{pr,H}, \phi) &= \sum_x [\bar{T}_\phi(\min_{\pi\in\Lambda} \bar{V}_{H-1}(\pi,\phi))(x) - \min_{\pi\in\Lambda} \bar{V}_{H-1}(\pi,\phi)(x)] P^{\pi_{pr,H},\phi}(x) \text{ by Lemma 4.2} \\
&\le \sum_x [\min_{\pi\in\Lambda} \bar{V}_H(\pi,\phi)(x) - \min_{\pi\in\Lambda} \bar{V}_{H-1}(\pi,\phi)(x)] P^{\pi_{pr,H},\phi}(x) \\
&\le \sum_x [\bar{V}_H(\arg\min_{\pi\in\Lambda} J^{\pi,\phi}_{H-1}(x), \phi)(x) - \bar{V}_{H-1}(\arg\min_{\pi\in\Lambda} J^{\pi}_{H-1}(x), \phi)(x)] P^{\pi_{pr,H},\phi}(x) \\
&\le \sum_x J_\infty(\arg\min_{\pi\in\Lambda} J^{\pi,\phi}_{H-1}(x), \phi) P^{\pi_{pr,H},\phi}(x) + 2\eta^{\frac{H-1}{s-1}} C_{\max} \text{ by Lemma 4.3 }.
\end{aligned}$$

■

From the result given in the above theorem, we can now discuss the convergence rate of the H-horizon parallel rollout policy. The second error term will approach zero geometrically in η as $H \to \infty$ and $\arg\min_{\pi\in\Lambda} J^{\pi,\phi}_{H-1}(x)$ will approach to the policy $\arg\min_{\pi\in\Lambda} J^{\pi,\phi}_\infty$ in $O(H^{-1})$. In the limit, the parallel rollout policy will improve all policies in Λ. We remark that if for each $\pi \in \Lambda$, $\bar{V}_0(\pi,\phi)$ is selected such that for all $x \in X$, $\bar{T}_{\pi,\phi}(\bar{V}_0(\pi,\phi))(x) \le \bar{V}_0(\pi,\phi)(x)$, then we can write the result of the above theorem as follows:

$$J_\infty(\pi_{pr,H}, \phi) \le \min_{\pi\in\Lambda} J_\infty(\pi,\phi) + 2\eta^{\frac{H-1}{s-1}} C_{\max}.$$

We conclude the discussion of the (parallel) rollout with a remark on the minimizer's guess of the maximizer's play. The above parallel rollout approach for the minimizer naturally gives a way of guessing a worst-case scenario of the maximizer to the minimizer. Suppose the minimizer can guess the best response from the maximizer when he plays a given heuristic policy $\pi \in \Lambda$. In this case, the minimizer considers a finite set $\Omega \subset \Phi$ of multiple heuristic policies for the maximizer and defines a policy $\phi_{\max}(x) = \arg\max_{\phi\in\Omega}[\min_{\pi\in\Lambda} V_\infty(\pi,\phi)](x)$ for all $x \in X$, and uses the policy $\phi_{\max}$ as the fixed policy for the maximizer.

4.3 Hindsight optimization

The recently proposed approach called hindsight optimization [13] to solving Markov decision processes can be also extended to solve Markov games if we fix a policy for the maximizer. Under the assumption that the opponent (the maximizer) plays his best policy (chosen by the minimizer), the hindsight optimizing minimizer plays the game at each state based on his analysis on the expected optimal "retroactive" performance.

Given a policy $\phi \in \Phi$, define a function $\rho_{n,\phi} \in B(X)$ such that

$$\rho_{n,\phi}(x) = E\left\{\min_{g_0,\dots,g_{n-1}} \sum_{t=0}^{n-1} \gamma^t C_{x_t}(g_t, \phi(x_t)) + \gamma^n V_0^*(x_n) | x_0 = x\right\}, g_t \in G(x_t) \text{ for all } t \qquad (10)$$

and call this the "hindsight optimal" value of state x because it stands for the (expected) value of taking (randomized) actions that the minimizer wishes to take if he encounters the particular random trace of the game. For the average cost case, we simply set $\gamma = 1$ and refer to the value as $\bar{\rho}_{n,\phi}(x)$.

Given a policy ϕ for the maximizer, we formally define the H-horizon hindsight optimization policy as a policy $\pi_{ho,H}$ such that for all $x \in X$, $T_{\pi_{ho,H},\phi}(\rho_{H-1,\phi})(x) = T_\phi(\rho_{H-1,\phi})(x)$. The average case is defined with "$\bar{T}$"-operator with $\bar{\rho}_{H-1,\phi}$. Because the minimization over the sequence of the randomized actions is inside the expectation in Equation 10, this corresponds to solving the sample-path problem, which is deterministic. The hindsight optimal value of state x is a lower bound to the equilibrium value if we set $\phi = \phi^*$ because by Jensen's inequality, $\rho_{n,\phi}(x) \leq V_n(\tilde{\pi}, \phi)(x)$ for any $\tilde{\pi} \in \tilde{\Pi}$ (for discounted case) and also $\bar{\rho}_{n,\phi}(x) \leq \bar{V}_n(\tilde{\pi}, \phi)(x)$ for any $\tilde{\pi} \in \tilde{\Pi}$ (for average case).

It is quite difficult to give a bound on the hindsight optimal value without restrictive conditions on the game. However, we believe that studying this issue is important. For this purpose, we introduce an equivalent model description of Markov games. We can derive a function called the *next state function* $\tilde{P} : X \times G(X) \times F(X) \times [0,1] \to X$ from the transition function P. In other words, given a policy pair π and ϕ and the current state x, a random number w selected uniformly from [0,1] can be mapped to $P_{xy}(\pi(x), \phi(x))$ for some $y \in X$. That is, $x_{t+1} = \tilde{P}(x_t, a_t, w_t)$ with so-called random disturbance $w_t \in [0,1]$. The average payoff function C is also newly defined by $\tilde{C}$ such that $C_x(\pi(x), \phi(x)) = E_w(\tilde{C}_x(\pi(x), \phi(x), w))$. See Bertsekas' book of definitions on MDP [8] or Ng's deterministic (partially observable) MDP model for a related construction [36].

Now we define a function Q such that

$$Q(x_0, \pi_0, ..., \pi_{n-1}, w_0, ..., w_{n-1}) = \sum_{t=0}^{n-1} \gamma^t \tilde{C}_{x_t}(\pi_t(x_t), \phi(x_t), w_t) + \gamma^n V_0^*(x_n)$$

and for convenience, we will abbreviate this to $Q(x_0, \tilde{\pi}, \vec{w})$ in an obvious notation, where $\vec{w} =< w_0, ..., w_{n-1} >\in [0,1]^n$ and $\tilde{\pi} = \{\pi_0, ..., \pi_{n-1}\}$. Then,

$$\rho_{n,\phi}(x_0) = E_{\vec{w}}[\min_{\tilde{\pi} \in \tilde{\Pi}} Q(x_0, \tilde{\pi}, \vec{w})]$$

because the minimization over nonstationary policy is equivalent to the minimization over the (randomized) action sequences given $\vec{w}$.

Proposition 4.4 *Suppose $\tilde{C}_x(g, f, w)$ is convex as a function of g and w jointly for every fixed $x \in X$ and $f \in F(x)$ and V_0^* is convex as a function of x. Then, for all $x \in X$,*

$$0 \leq \inf_{\tilde{\pi} \in \tilde{\Pi}} V_n(\tilde{\pi}, \phi)(x) - \rho_{n,\phi}(x) \leq Q(x, \tilde{\pi}_{0.5}, \vec{0.5}) - E_{\vec{w}}[Q(x, \tilde{\pi}_{0.5}, \vec{w})]$$

where $\vec{0.5}$ is a vector of size n with every entry 0.5 and $\tilde{\pi}_{0.5}$ solves $\inf_{\tilde{\pi} \in \tilde{\Pi}}[Q(x, \tilde{\pi}, \vec{0.5})]$.

Proof: First, under our assumptions, the function Q is convex in the space of $\vec{w}$ and $\tilde{\pi}$ ($[0, 1]^n$ and a cartesian product of polyhedral sets respectively), whose cartesian product space is a convex set. Therefore, we can directly apply Avriel and Williams' theorem on the Jensen's inequality on expected value of perfect information [22]. ∎

The same result holds for the average cost case (with $\gamma = 1$) and in particular if $\phi = \phi^*$, the proposition above gives a bound between the hindsight optimal value and the n-horizon equilibrium value.

We remark that the hindsight-optimization based approach appeals to the game-theoretic framework so that this is different from the simulation-based approach used in the computer bridge game player (GIB) in [20]. The approach taken there can be viewed as follows in the context of our discussion: many sample paths are drawn and for each sample path, the optimal solution with respect to the sample path is analyzed after taking each deterministic candidate action, and one *counts* the number of times that a particular deterministic action achieves the minimum cost sum, and takes a deterministic action by *voting*. It would be interesting to compare two approaches in practical applications.

5 Implementation and Research Directions

In this subsection, we briefly discuss how we can implement the (approximate) receding horizon approaches we discussed before in practice and discuss some issues and directions for the future research.

There is previous work done by Kearns *et al.* [28] that presents an algorithm that uses samples to estimate $\bar{V}_n^*$ (the undiscounted finite horizon value of game) within a given error bound, which can be easily adapted to the discounted setting. They analyzed the necessary number of sampling to obtain a desired accuracy. The per-state running time of their algorithm is independent of the state space size but exponential in the horizon size. Note that finite horizon value iteration's computation complexity depends on the state space size, even though it depends on the horizon size linearly, so that applying it for a game with a very large state space is difficult.

The exponential dependence on the horizon size can be alleviated by using the three heuristic approaches we discussed. We can simply use a Monte-Carlo simulation to estimate the relevant function values. For example, the minimizer who uses the H-horizon rollout policy simulates the given heuristic base policy and the fixed policy for the maximizer using sampling over a finite horizon $H-1$, and the results of the simulation are used to "select" the (apparently) best current randomized action at the current state. We assume that there is a selection function available that extracts the randomized action that achieves the infimum/supremum. The randomized action selected is the randomized action with the highest "utility" at the current state, as estimated by sampling. Of course, we can use various sampling techniques (see, e.g., [33]), such as importance sampling, to improve the estimation procedure. Therefore, the rollout/parallel rollout approach is practically viable. On the other hand, the hindsight optimization approach needs to have a fast hindsight problem solver.

Extending the receding horizon framework to the N-person ($N \geq 3$) case and analyzing the performance will be difficult, because no iteration algorithm based on a contraction mapping is available to the authors' knowledge. However, each player can heuristically use the rollout/parallel rollout and the hindsight optimization for his policy choice.

We can also consider applying the three heuristics to nonzero-sum stochastic games. Analyzing the structure of equilibrium policies, in this case, is often more difficult than for zero-sum games. For zero-sum games, a standard technique, e.g., value iteration, can be used (see, e.g., [1, 3] and references therein). However, for nonzero-sum games, we need to use a different non-standard technique (see, e.g., [2]) to analyze the structure, which is quite cumbersome.

Finally, we can incorporate the idea of Neuro-Dynamic programming (NDP) [10] into the approximate receding horizon control framework. That is, the feature-based approximations in NDP can be applied when we estimate the value of the underlying subgame, although how to extract good features is a difficult problem in general.

References

[1] E. Altman, "Zero-sum Markov games and worst-cast optimal control of queueing systems," *QUESTA*, vol. 21, pp. 415–447, 1995.

[2] E. Altman, "Non zero-sum stochastic games in admission, service and routing control in queueing systems," *QUESTA*, vol. 23, pp. 259–279, 1996.

[3] E. Altman, "Monotonicity of optimal policies in a zero sum game: a flow control model," *Advances of Dynamic Games and Applications*, pp. 269–286, 1994.

[4] E. Altman, "Contraction conditions for average and α-discount optimality in countable state Markov games with unbounded rewards," *Math. of Oper. Research*, vol. 22, no. 3, pp. 588–618, 1997.

[5] M. Baglietto, T. Parisini, and R. Zoppoli, "Neural approximators and team theory for dynamic routing: a receding horizon approach," in *Proc. IEEE CDC*, 1999, pp. 3283–3288.

[6] T. Basar and P. R. Kumar, "On worst case design strategies," *Comput. Math. Applic.*, vol. 13, pp. 239–245, 1987.

[7] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*, Academic Press, London/New York, 1995.

[8] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Volumes 1 and 2.* Athena Scientific, 1995.

[9] D. P. Bertsekas and D. A. Castanon, "Rollout algorithms for stochastic scheduling problems," *J. of Heuristics,* vol. 5, pp. 89–108, 1999.

[10] D. P. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming.* Athena Scientific, 1996.

[11] 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.

[12] W. A. van den Broek, "Moving horizon control in dynamic games," *Computing in Economics and Finance*, vol. 122, 1999.

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

[14] H. S. Chang, R. Givan, and E. K. P. Chong, "Model-based random early packet dropping using rollout policies," submitted to *Discrete Event Dynamic Systems: Theory and Application*, 2000.

[15] H. S. Chang and S. Marcus, "On approximate receding horizon approach for Markov decision processes: average reward case," submitted Automatica (TR 2001-46, ISR, Univ. of Maryland, 2001).

[16] E. K. P. Chong, R. Givan, and H. S. Chang, "A framework for simulation-based network control via hindsight optimization," in *Proc. 39th IEEE CDC*, 2000, pp. 1433–1438.

[17] R. Dekker and A. Hordijk, "Average, sensitive and Blackwell optimality in denumerable state Markov decision chains with unbounded rewards," in *Math. Operat. Res.*, vol. 17, pp. 271–290, 1988.

[18] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*, Springer-Verlag, 1996.

[19] D. Fudenberg and D. Levine, *The Theory of Learning in Games*, MIT Press, 1998.

[20] M. L. Ginsberg, "GIB: steps toward an expert-level bridge-playing program," in *Proc. of IJCAI*, 1999, pp. 584–589.

[21] D. Gillette, "Stochastic games with zero stop probabilities," in *Contributions to the theory of games*, Vol. III, M. Dresher, A. Tucker and P. Wolfe (Eds.), Princeton Univ. Press, Princeton, New Jersey, pp. 179–187.

[22] D. B. Hausch and W. T. Ziemba, "Bounds on the value of information in uncertain decision problems II," *Stochastics*, vol. 10, pp. 181–217, 1983.

[23] O. Hernández-Lerma and J. B. Lasserre, "Error bounds for rolling horizon policies in discrete-time Markov control processes," *IEEE Transactions on Automatic Control*, vol. 35, no. 10, pp. 1118–1124, 1990.

[24] J. Hespanha and S. Bohacek, "Preliminary results in routing games," in *Proc. ACC*, 2001.

[25] Y. -C. Ho and K. -C. Chu, "Team decision theory and information structures in optimal control problems-part I.," *IEEE Transactions on Automatic Control*, vol. 17, pp. 15–22, 1972.

[26] A. J. Hoffman and R. M. Karp, "On nonterminating stochastic games," *Management Science*, vol. 19, no. 5, pp. 359–370, 1966.

[27] L. Johansen, *Lectures on Macroeconomic Planning.* Amsterdam, The Netherlands: North-Holland, 1977.

[28] M. Kearns, Y. Mansour, and S. Singh, "Fast planning in stochastic games," in *Proc. of UAI*, 2000.

[29] 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. of Optimization Theory Appl.*, vol. 57, pp. 265–293, 1988.

[30] A. Kolarov and J. Hui, "On computing Markov decision theory-based cost for routing in circuit-switched broadband networks," *J. of Network and Systems Management*, vol. 3, no. 4, pp. 405-425, 1995.

[31] G. Koole, "The deviation matrix of the $M/M/1/\infty$ and $M/M/1/N$ queue, with applications to controlled queueing models," in *Proc. of the 37th IEEE CDC*, 1998, pp. 56–59.

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

[33] P. L'Ecuyer, "Efficiency improvement and variance reduction," in *Proc. Winter Simulation conf.*, 1994.

[34] D. Q. Mayne and H. Michalska, "Receding horizon control of nonlinear system," *IEEE Trans. Auto. Contr.*, vol. 38, no. 7, pp. 814–824, 1990.

[35] M. Morari and J. H. Lee, "Model predictive control: past, present, and future," *Computers and Chemical Engineering*, vol. 23, pp. 667–682, 1999.

[36] Y. Ng and M. Jordan, "PEGASUS: A policy search method for large MDPs and POMDPs," in *Proc. of UAI*, 2000, pp. 405–415.

[37] T. J. Ott and K. R. Krishnan, "Separable routing: a scheme for state-dependent routing of circuit switched telephone traffic," *Annals of Operations Research*, vol. 35, pp. 43–68, 1992.

[38] S. D. Patek and D. Bertsekas, "Stochastic shortest path games," *SIAM J. on Control and Optimization*, vol. 37, no. 3, pp. 804–824, 1999.

[39] C. H. Papadimitriou, "Games against nature," *J. of Computer and System Sciences*, vol. 31, pp. 288–301, 1985.

[40] T. Parthasarathy and M. Stern, "Markov Games - a survey," *Differential games and control theory* II, E. Roxin, P. Liu, and R. Sternberg (eds.), Dekker, pp. 1–46, 1977.

[41] L. K. Platzman, "Optimal infinite-horizon undiscounted control of finite probabilistic system," *SIAM J. Cont. and Opt.*, vol. 14, no. 4, pp. 362–380, 1980.

[42] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, New York, 1994.

[43] N. Secomandi, "Comparing neuro-dynamic programming algorithms for the vehicle routing problem with stochastic demands," *Computers and Operations Research*, vol. 27, pp. 1201–1225, 2000.

[44] L. Shapley, Stochastic games, in *Proc. of the National Academy of Sciences*, vol. 39, pp. 1095–1100, 1953.

[45] J. Van Der Wal, "Discounted Markov games: generalized policy iteration method," *J. of Optimization Theory and Applications*, vol. 25, no. 1, pp. 125–138, 1978.

[46] J. Van Der Wal, *Stochastic Dynamic Programming: successive approximations and nearly optimal strategies for Markov decision processes and Markov games*, Ph.D. Thesis, Eindhoven, 1980.