

TWO-PERSON ZERO-SUM MARKOV GAMES: RECEDING HORIZON APPROACH

Hyeong Soo Chang and Steven I. Marcus*

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” 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 [2], we then analyze error bounds when the minimizer plays the (approximate) receding horizon control and the maximizer plays the worst case policy. We finally discuss some state-space size independent methods to compute the value of subgame approximately for the approximate receding horizon control, along with heuristic receding horizon policies for the minimizer.

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

*Correspondence author.

S. Marcus can be reached by telephone at +1 (301) 405-3683 or by fax at +1 (301) 314-9920.

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

H. S. Chang is with Research Institute for Information and Communication Technology, Korea University, Seoul, Korea, and can be reached by e-mail at hyeongsoo@korea.ac.kr and S. I. Marcus is with Department of Electrical and Computer Engineering, Univ. of Maryland at College Park, and can be reached by e-mail at marcus@eng.umd.edu.

1 Introduction

This paper studies solving two-person zero-sum Markov games with infinite horizon discounted cost and infinite horizon average cost criteria via an *approximation* framework. 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 [24], model predictive control literature [26, 27, 28], and planning in MDPs [21, 1, 14, 11, 12], etc. In the game setting, Baglietto et al. [5] applied team theory [22] 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 Cruz et al. [15] studied a military operation problem formulated as a Markov game with one and two step receding horizon, and Van den Broek [10] considered a receding horizon control in non-zero sum differential games, specifically analyzing the performance of 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” 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 [21] obtained a similar result for MDPs [21]). 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 “equilibrium” policy, motivated from the worst case queueing control study by Altman [2], 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?

Unfortunately, solving the finite horizon Markov game or subgame exactly, or computing the equilibrium value of the game, 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 [13].

We discuss two previously published state-space size independent computational methods to compute the equilibrium value of the game approximately. The first approach [2] is based on solving a sequence of approximating games, where the scale of the games are small compared with the original game, and the second approach uses sampling [25]. We further discuss two heuristic receding horizon policies for the minimizer, called parallel rollout [12] and hindsight optimization [14], which can be implemented via Monte-Carlo simulation methods for on-line control.

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 two example methods for generating approximate receding horizon control in Section 4 and two heuristic receding horizon policies for the minimizer in Section 5. In Section 6, we conclude this paper with some remarks.

2 Markov Game

In this section, we formulate the two-person zero-sum Markov game introduced by Shapley [32] in a formal mathematical setting. For a substantial discussion on this topic, see, e.g., [16] [6] or [29]. 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

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

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 $\pi : X \rightarrow G(X)$ 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*

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

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. Similarly, we define the value of the game for the *infinite horizon average cost criterion* as

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

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., [29]) 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 $x \in X$,

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

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).

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 $x \in X$,

$$T(V)(x) = \inf_{g \in G(x)} \sup_{f \in F(x)} \left\{ C_x(g, f) + \gamma \sum_{y \in X} P_{xy}(g, f) V(y) \right\} \quad (1)$$

$$T_{\pi, \phi}(V)(x) = C_x(\pi(x), \phi(x)) + \gamma \sum_{y \in X} P_{xy}(\pi(x), \phi(x)) V(y) \quad (2)$$

$$T_\phi(V)(x) = \inf_{g \in G(x)} \left\{ C_x(g, \phi(x)) + \gamma \sum_{y \in X} P_{xy}(g, \phi(x)) V(y) \right\} \quad (3)$$

$$T_\pi(V)(x) = \sup_{f \in F(x)} \left\{ C_x(\pi(x), f) + \gamma \sum_{y \in X} P_{xy}(\pi(x), f) V(y) \right\} \quad (4)$$

It is well-known [16, 29] 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., [29]).

Let $\{V_n^*\}$ be the sequence of *value iteration functions* $V_n^* := T(V_{n-1}^*)$ where $n = 1, 2, \dots$ 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 $n \rightarrow \infty$, 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 $\tilde{\pi} = \{\pi_0, \pi_1, \dots\}$ 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., [2, 35]) and is given by

$$V_n^*(x) = \inf_{\tilde{\pi} \in \tilde{\Pi}} \sup_{\tilde{\phi} \in \tilde{\Phi}} E \left\{ \sum_{t=0}^{n-1} \gamma^t C_{x_t}(\pi_t(x_t), \phi_t(x_t)) + \gamma^n V_0^*(x_n) \middle| x_0 = x \right\} \quad (5)$$

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 [19] 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 condition.

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 $x \in X$,

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

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 [35]. 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^* .

We define several counterpart operators to the operators defined in Equations (1)-(4) by letting $\gamma = 1$ and denote each operator with bar on it. It is well-known (see, e.g., [16, 29, 35]) that each of these 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 $n = 1, 2, \dots$ and $\bar{V}_0^*$ is arbitrary function in $B(X)$. It has been shown [35] (under Assumption 2.1 on average Markov games) that as $n \rightarrow \infty$, $\bar{V}_n^* - \bar{V}_{n-1}^*$ converges to a constant J_∞^* and there exists 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 $x \in X$.

We will say throughout the present paper that two policies π and ϕ are *correlated* if for a given $V \in B(X)$, π and ϕ satisfy that $T_{\pi, \phi}(V)(x) = T(V)(x)$ for all $x \in X$, similarly for the $\bar{T}$ -operator and omit “correlated” if the context is clear.

3 Receding Horizon Control

3.1 Infinite horizon discounted cost

As we mentioned before, solving a large-scale Markov games for infinite horizon costs is often impractical. Therefore, we adopt receding horizon control as a finite-horizon approximation scheme for the infinite horizon problem. The receding horizon control is simply defined as follows. Given a finite horizon $H \geq 1$, we define the (correlated) 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 $x \in X$. 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 the page 181 in [16] 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 .

The minimizer will play the game by the receding horizon control based on assumption that the maximizer also plays the correlated 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 [34] 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 $H = 1, 2, \dots$, 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 $H = 1, \dots, k$, we prove that it holds for $H = k + 1$. For all $x \in X$,

$$V_{k+1}^*(x) = T(V_k^*)(x) = T(T(V_{k-1}^*))(x) \leq T(V_{k-1}^*)(x) = V_k^*(x),$$

where the inequality comes from the monotonicity of T and the assumption. This concludes the proof. $\blacksquare$

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 $n = 0, 1, \dots$,

$$\max_x |V_\infty^*(x) - V_n^*(x)| \leq \gamma^n \max_x |V_\infty^*(x) - V_0^*(x)| \leq \frac{2C_{\max}}{1 - \gamma} \cdot \gamma^n. \quad (6)$$

Combining the two inequalities, we have the desired result. $\blacksquare$

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 $\delta(x) \geq 0$, $0 < \sum_{x \in X} \delta(x) < 1$ 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 ergodicity 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 [21] via a method called “column reduction”. We can write P as a convex combination of the two distributions where one describes a new state transition structure and the other is independent of transitions, and this decomposition allows us to convert the original Markov game into an equivalent model with a new discount factor in the sense that for any finite horizon, the finite horizon subgames for two models have the same set of optimal policies (see [1] for a further discussion). Let $\beta = 1 - \sum_x \delta(x)$. 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 write the transition probability P by

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

We define the operator $T' : B(X) \rightarrow B(X)$ such that for any $V \in B(X)$,

$$T'(V)(x) = \inf_{g \in G(x)} \sup_{f \in F(x)} \left\{ C_x(g, f) + \gamma \beta \sum_{y \in X} P'_{xy}(g, f) V(y) \right\}$$

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 $n = 1, 2, \dots$ and set $V_0^*(x) = V'_0(x) = C_{\max}/(1 - \gamma)$ for all $x \in X$. By induction, we can show that

$$V_n^*(x) = V'_n(x) + C_n, x \in X. \quad (7)$$

where C_n is the constant given by

$$\begin{aligned}C_n &= \gamma(1 - \beta) \sum_{k=0}^{n-1} (\gamma\beta)^k \psi(V_{n-1-k}^*) \\ &= \gamma(1 - \beta) \sum_{k=0}^{\infty} (\gamma\beta)^k \psi(V_{n-1-k}^*)\end{aligned} \quad (8)$$

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 [21])

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

where $V'_\infty(x) = \lim_{n \rightarrow \infty} (T')^n(V'_0)(x)$ and $C(\gamma, \beta) = \gamma(1 - \beta)/(1 - \gamma\beta)$. 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 (7). 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 [35]. We begin with a modified version of Van der Wal's Corollary 13.2 in the page 230 in [35] within our context. For a function $v \in B(X)$, let span semi-norm of v be $\text{sp}(v) = \max_x v(x) - \min_x v(x)$.

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 $\phi' \in \Phi$,*

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

From now on, we will set $|X| = s$ (we assume that $s \geq 2$). 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 $x, y \in X$,

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

where $\mathcal{P}_{x,\pi,\phi}^{s-1}(z)$ 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 $n = 0, 1, \dots$, $\text{sp}(\bar{V}_{n+1}^* - \bar{V}_n^*) \leq 4\eta^{\lfloor \frac{n}{s-1} \rfloor} C_{\max}$*

Proof: Van der Wal showed that (see the page 235 in [35]) $\text{sp}(\bar{V}_{n+s}^* - \bar{V}_{n+s-1}^*) \leq \eta \cdot \text{sp}(\bar{V}_{n+1}^* - \bar{V}_n^*)$, $n = 0, 1, \dots$, which implies that

$$\text{sp}(\bar{V}_{n+1}^* - \bar{V}_n^*) \leq \eta^{\lfloor \frac{n}{s-1} \rfloor} \text{sp}(\bar{V}_{j+1}^* - \bar{V}_j^*) \text{ for } n = (s-1)i + j, i = 0, 1, 2, \dots, j = 0, 1, \dots, (s-2).$$

Since $\text{sp}(\bar{V}_{j+1}^* - \bar{V}_j^*) \leq 4C_{\max}$ for $j = 1, \dots, (s-2)$, 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,*

$$|J_{\infty}(\pi_H^*, \phi_H^*) - J_{\infty}^*| \leq 4\eta^{\lfloor \frac{H-1}{s-1} \rfloor} C_{\max}$$

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 .

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,*

$$0 \leq J_{\infty}(\pi_H^*, \phi^*) - J_{\infty}^* \leq 4\eta^{\lfloor \frac{H-1}{s-1} \rfloor} C_{\max}$$

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 new conditions to the transition probability matrix and the cost function 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 and K , with $0 \leq \alpha < 1$, $M \geq 0$,*

$$\sum_{y \in X} P_{xy}(\pi(x), \phi(x)) \mu(y) \leq \alpha \mu(x) \text{ and } |C_x(\pi(x), \phi(x))| \leq K \mu(x) \leq C_{\max}$$

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

We will refer to this assumption as the μ -contraction condition. The condition ensures the summability of the costs with no discounting and boundedness for any policy pair via contraction in α only.

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

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

It is well-known that under the μ -contraction 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., the page 199 in [35]) 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 $n = 1, 2, \dots$,

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

Because $\|\bar{V}_1^*\|_\mu \leq \frac{C_{\max}}{1-\alpha}$ (see, e.g., the page 199 in [35]), we have the following immediate result with μ -contraction 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 -contraction 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 proposition 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 $n = 0, 1, \dots$,*

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

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 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 ,

$$|T(V_n^*)(x) - T(V)(x)| \leq \gamma \cdot \max_x |V_n^*(x) - V(x)| \leq \gamma\epsilon \quad (9)$$

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

$$\max_x |V_\infty^*(x) - V_{n+1}^*(x)| \leq \gamma^{n+1} \max_x |V_\infty^*(x) - V_0^*(x)| \leq \frac{2C_{\max}}{1-\gamma} \cdot \gamma^{n+1}. \quad (10)$$

Therefore, from Equation (9) and (10) and $V_{n+1}^* = T(V_n^*)$ by definition, for all $x \in X$,

$$\begin{aligned} |V_\infty^*(x) - T(V)(x)| &\leq |V_\infty^*(x) - T(V_n^*)(x)| + |T(V_n^*)(x) - T(V)(x)| \\ &\leq \frac{2C_{\max}}{1-\gamma} \cdot \gamma^{n+1} + \gamma\epsilon. \end{aligned} \quad (11)$$

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 (11), 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 (9), we have that for all $x \in X$, by letting $w = \frac{\gamma^n}{1-\gamma} \cdot 2C_{\max}$,

$$V(x) \leq V_n^*(x) + \epsilon \leq V_{n+1}^*(x) + \epsilon + w \leq T(V)(x) + \gamma\epsilon + \epsilon + w. \quad (12)$$

Then for all $x \in X$,

$$\begin{aligned} T(V)(x) &= T_{\pi, \phi}(V)(x) = C_x(\pi(x), \phi(x)) + \gamma \sum_{y \in X} P_{xy}(\pi(x), \phi(x)) V(y) \text{ by definitions of } \pi \text{ and } \phi \text{ and } T \\ &\leq C_x(\pi(x), \phi(x)) + \gamma \sum_{y \in X} P_{xy}(\pi(x), \phi(x)) [T(V)(y) + \gamma\epsilon + \epsilon + w] \text{ by Equation (12)} \\ &= C_x(\pi(x), \phi(x)) + \gamma \sum_{y \in X} P_{xy}(\pi(x), \phi(x)) T(V)(y) + \gamma\epsilon(1+\gamma) + \gamma w \\ &= C_x(\pi(x), \phi(x)) + \gamma \sum_{y \in X} P_{xy}(\pi(x), \phi(x)) \left(C_y(\pi(y), \phi(y)) + \gamma \sum_{z \in X} P_{yz}(\pi(y), \phi(y)) V(z) \right) \\ &\quad + \gamma\epsilon(1+\gamma) + \gamma w \\ &= C_x(\pi(x), \phi(x)) + \gamma \sum_{y \in X} P_{xy}(\pi(x), \phi(x)) C_y(\pi(y), \phi(y)) \\ &\quad + \gamma^2 \sum_{y \in X} \sum_{z \in X} P_{xy}(\pi(x), \phi(x)) P_{yz}(\pi(y), \phi(y)) V(z) + \gamma\epsilon(1+\gamma) + \gamma w \\ &\leq C_x(\pi(x), \phi(x)) + \gamma \sum_{y \in X} P_{xy}(\pi(x), \phi(x)) C_y(\pi(y), \phi(y)) \\ &\quad + \gamma^2 \sum_{y \in X} \sum_{z \in X} P_{xy}(\pi(x), \phi(x)) P_{yz}(\pi(y), \phi(y)) T(V)(z) \\ &\quad + \gamma^2\epsilon(1+\gamma) + \gamma\epsilon(1+\gamma) + (\gamma^2 w + \gamma w) \end{aligned}$$

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

$$\begin{aligned} T(V)(x) &\leq E \left[\sum_{t=0}^k \gamma^t C_{x_t}(\pi(x_t), \phi(x_t)) | x_0 = x \right] + \gamma^{k+1} E[T(V)(x_{k+1}) | x_0 = x] \\ &\quad + \gamma\epsilon(1+\gamma) + \dots + \gamma^{k+1}\epsilon(1+\gamma) + (\gamma w + \dots + \gamma^{k+1}w), \end{aligned} \quad (13)$$

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 (13) converges to zero as $k \rightarrow \infty$ 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 (9), 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 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 (9) 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 $k = 0, 1, \dots$, and $x \in X$,

$$T(V)(x) \geq E \left[\sum_{t=0}^k \gamma^t C_{x_t}(\pi(x_t), \phi^*(x_t)) | x_0 = x \right] + \gamma^{k+1} E[T(V)(x_{k+1}) | x_0 = x] - [\gamma\epsilon(1 + \gamma) + \dots + \gamma^{k+1}\epsilon(1 + \gamma)], \quad (14)$$

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 (14) converges to zero as $k \rightarrow \infty$ 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 $\delta(x) \geq 0$, $0 < \sum_{x \in X} \delta(x) < 1$ 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 (8).

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 $x \in X$,*

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

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 and 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 -contraction condition (Assumption 3.1) to this case also so that we can eliminate the dependence on the state space size as we did previously.

3.4 Remarks on transient game and reducible average cost games

We give some remarks on “transient” games, specifically stochastic games with infinite horizon total cost criterion with cost-free absorbing states and no discounting, and on stochastic games that are not irreducible or do not exhibit strong aperiodicity with respect to the receding horizon control framework.

Consider a transient Markov game M with a state space $X = X_1 \cup X_2$ with $X_1 \cap X_2 = \emptyset$ where X_1 consists of non-absorbing states such that for each $x \in X_1$, there exists a constant $\gamma_x > 0$ such that

$$\sum_{y \in X_1} P_{xy}(g, f) = \gamma_x < 1, g \in G(x), f \in F(x),$$

and X_2 consists of absorbing states such that for each $x \in X_2$, $C_x(g, f) = 0, P_{xx}(g, f) = 1$ for all $g \in G(x)$ and $f \in F(x)$. Write the equilibrium value of the game M with respect to total cost criterion as W_∞^* and consider a Markov game M' defined with only X_1 by defining the state transition probability for M' as $P_{xy}(g, f)/\gamma_x$ for all $x, y \in X_1, g \in G(x)$ and $f \in F(x)$ and with the discount factor of $\gamma_{\max} = \max_{x \in X_1} \{\gamma_x\}$. Write the equilibrium value of the game M' with respect to the infinite horizon discounted criterion with the discount $\gamma_{\max}$ and initial state x as $V_{\infty, \gamma_{\max}}^*(x)$. We also consider a Markov game M'' defined similarly to M' except for the discount factor of $\gamma_{\min} = \min_{x \in X_1} \{\gamma_x\}$. It is left for the reader to check that the following inequality holds: for all $x \in X_1$,

$$V_{\infty, \gamma_{\min}}^*(x) \leq W_\infty^*(x) \leq V_{\infty, \gamma_{\max}}^*(x).$$

Therefore, we can apply the results obtained as above by considering the receding horizon control for the game M' and M'' to bound the value of the game M .

For a given (irreducible) Markov game, consider the following transformation of the Markov game into another game: we define the state transition probability $\hat{P}$ and the cost $\hat{C}$ of the new game as

$$\begin{aligned} \hat{P}_{xx}(g, f) &:= \rho + (1 - \rho)P_{xx}(g, f), \\ \hat{P}_{xy}(g, f) &:= (1 - \rho)P_{xy}(g, f), y \neq x, \\ \hat{C}_x(g, f) &:= (1 - \rho)C_x(g, f) \end{aligned}$$

for all $x \in X, g \in G(x)$ and $f \in F(x)$, where ρ is some constant with $0 < \rho < 1$. Observe that for this new model, besides irreducibility, the strong aperiodicity condition is satisfied. It turns out that the two models are equivalent in the sense that if $J_\infty(\pi, \phi)$ is the value of the original game for the infinite horizon average cost criterion following $\pi \in \Pi$ and $\phi \in \Phi$ that satisfies

$$\bar{T}_{\pi, \phi}(\bar{V}_\infty(\pi, \phi))(x) = J_\infty(\pi, \phi) + \bar{V}_\infty(\pi, \phi)(x) \text{ for all } x \in X,$$

for a function $\bar{V}_\infty(\pi, \phi) \in B(X)$, $(1 - \rho)J_\infty(\pi, \phi)$ is the value of the transformed game following π and ϕ and satisfies

$$\hat{T}_{\pi, \phi}(\bar{V}_\infty(\pi, \phi))(x) = (1 - \rho)J_\infty(\pi, \phi) + \bar{V}_\infty(\pi, \phi)(x) \text{ for all } x \in X,$$

where $\hat{T}$ is defined with $\hat{P}$ and $\hat{C}$, and if π^* is ϵ -optimal policy for the minimizer in the original game, π^* is $(1 - \rho)^{-1}\epsilon$ -optimal policy for the minimizer in the transformed game (see, e.g., the page 231 in [35] for a detailed discussion). Therefore, extending the results obtained as above into arbitrary (irreducible) games is straightforward.

To the best of the authors' knowledge, there is no prior work on successive approximation (like value iteration) for multichain Markov games (for average cost case) and we expect that drawing a result like Theorem 3.3 will not be easy. This will make an analysis of the receding horizon control for multichain Markov games even more difficult. We briefly give an idea on how to approach to this problem to generate a pair of approximate equilibrium policies and leave the analysis as a future research topic.

Extending the idea given in [31, 9], we assume first that for any pair of pure policies for the minimizer and the maximizer, the underlying Markov chain consists of a finite number of recurrent classes. Suppose that a Markov game is decomposed into recurrent classes, where for any pair of policies, both players cannot leave any class once they enter that class, and a set of transient states. Decomposing a Markov game with such classes and a set of transient states can be done simply by adapting the algorithm given in [30] for MDPs that extended the Fox-Landi algorithm [17] for decomposing a given Markov chain. Then we can observe that the pair of equilibrium policies restricted to any recurrent class can be obtained *independently* of any states outside of the class. We use the receding horizon control approximation to approximate the solution of each class with possibly different ergodicity coefficients. This will give a bound on the equilibrium value of each recurrent class. We “eliminate” then each recurrent class (all states in each class) with this bound, leaving only transient states and creating a new sort of Markov game. By doing this, a set of certain transient states that could enter a recurrent class can form a new recurrent class “effectively” by considering transitions out of the set to the recurrent class via the bound on the equilibrium value for the recurrent class. (In the context of MDP, this corresponds to solving an MDP when we know the bounds of the optimal values for certain states. The bound can be approximated with the average value of the upper and lower limits.) We decompose this new Markov game again and repeat the above step until we can obtain all the necessary approximate solutions for all states.

4 Example Methods for Generating Approximate Receding Horizon Control

If the state space of the underlying Markov game is large, the usual methods of solving Markov game exactly, e.g., value iteration, policy iteration, mathematical programming, etc., are impossible to be applied due to the curse of dimensionality, even for solving the receding horizon subgames unless structures of optimal policies and value functions can be analyzed. In this section, we briefly summarize two previously published approaches to approximating the value of a finite horizon Markov game, both of which work independently of the state space size of the game.

Given an H -horizon Markov game formulated as in Section 2, consider the following model: the state space $I = X \times \{0, 1, \dots, H-1\}$, the admissible action set $U(i)$ for the minimizer at a state $i = (x, h), x \in X, h \in \{0, 1, \dots, H-1\}$ is given such that $U(i) = G(x)$ and the action set $W(i)$ for the maximizer for the state i is given such that $W(i) = F(x)$, the cost function such that $C'_i(u, w) = C_x(u, w)$, the transition function such that for $i = (x, h), j = (y, h') \in I$, $P'_{ij}(u, w) = P_{xy}(u, w)$ if $h+1 = h' \leq H-1$ and 0 otherwise, and the same discount factor.

Let B be a given subset of I , and set $Z(i) = \{j | P'_{ij}(u, w) > 0 \text{ for some } u, w\}$. Define I_n such that $I_0 = B$, $I_{n+1} = \bigcup_{i \in I_n} Z(i) \cup I_n$. Define a sequence of the value functions Θ_n such that

$$\Theta_n(i) = \begin{cases} \inf_{u \in U(i)} \sup_{w \in W(i)} \{C'_i(u, w) + \gamma \sum_{j \in I_n} P'_{ij}(u, w) \Theta_n(j)\} & \text{if } i \in I_n \\ 0 & \text{if } i \notin I_n \end{cases}$$

Then for any $i = (x, h) \in B$ and $n = 0, 1, \dots$, Tidball and Altman [33] showed that $|\Theta_n(i) - V_H^*(x)| \leq 2C_{\max} \cdot \gamma^n / (1 - \gamma)$, where V_H^* is defined as in Equation (5) with the zero function assumption of V_0^* . Note that obtaining the function value Θ_n at each n corresponds to solving a matrix game with the set I_n .

Kearns et al. [25] proposes a sampling algorithm that generates a pair of nonstationary policies $\tilde{\pi} \in \tilde{\Pi}$ for the minimizer and $\tilde{\phi} \in \tilde{\Phi}$ for the maximizer, where both policies are *random* in the sense that for each run of the algorithm, it generates a different pair of policies. Let $\tilde{V}_H(\tilde{\pi}, \tilde{\phi})$ be the H -horizon (undiscounted) value of the given Markov game following the policies $\tilde{\pi}$ and $\tilde{\phi}$. They provide the sample size for the algorithm that guarantees a desired bound ϵ such that for any policy $\sigma \in \tilde{\Pi}$ and $\tau \in \tilde{\Phi}$,

$$E[\tilde{V}_H(\tilde{\pi}, \tau)] - \epsilon \leq E[\tilde{V}_H(\tilde{\pi}, \tilde{\phi})] \leq E[\tilde{V}_H(\sigma, \tilde{\phi})] + \epsilon.$$

That is, the expected performance of the policies generated by the algorithm makes ϵ -equilibrium (in the expected sense) for the Markov game. The bound can be made arbitrarily smaller as the sample size is increased.

We can use the sampling algorithm to generate an approximate receding horizon control and in particular, the expected performance of the worst-case scenario can be analyzed (by letting τ the worst-case policy for the maximizer). The drawback of Kearns et al.'s approach is that even

though the per-state running time of their algorithm is independent of the state space size, it is exponential in the horizon size.

Both approximation approaches will be useful when the game is *sparse* in that the one-step reachable state set from each given state is small. If the game is not sparse and the operation extraction of the min-max (randomized) actions is cumbersome, the above approaches will suffer from the relatively big time-complexity. In the next section, we present two heuristic receding horizon policies for the minimizer that can eliminate this problem.

5 Heuristic Receding Horizon Policies

Consider a finite set of multiple heuristic policies, where each policy is good for different system trajectories for the minimizer. 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.

Given a fixed policy ϕ for the maximizer, define 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 $T_{\pi_{pr,H},\phi}(\psi)(x) = T_{\phi}(\psi)(x)$ for all $x \in X$, where $\psi(x) = \min_{\pi \in \Lambda} V_{H-1}(\pi, \phi)(x)$, $x \in X$. It can be shown that if for each $\pi \in \Lambda$, $V_0(\pi, \phi)$ is selected such that for all $x \in X$, $T_{\pi,\phi}(V_0(\pi, \phi))(x) \leq V_0(\pi, \phi)(x)$, then given any $\epsilon > 0$, for $H \geq 1 + \log_{\gamma} \frac{\epsilon(1-\gamma)}{C_{\max}}$, for all $x \in X$, $V_{\infty}(\pi_{pr,H}, \phi)(x) \leq \min_{\pi \in \Lambda} V_{\infty}(\pi, \phi)(x) + \epsilon$, extending the proof given in [12]. The average cost case can be similarly defined and analyzed, extending the proof given in [13].

The 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

$$\phi_{\max}(x) \in \{\arg \max_{\phi \in \Omega} [\min_{\pi \in \Lambda} V_{\infty}(\pi, \phi)](x)\}, x \in X,$$

and uses the policy $\phi_{\max}$ as the fixed policy for the maximizer. See [13] for a brief survey on successful results of this approach for the MDP problems.

The next heuristic is based on the recently proposed approach called hindsight optimization [14]. Under the assumption that 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.

We first introduce an equivalent model description of Markov games. We can derive a function called the *next state function* $\tilde{P} : X \times G(X) \times F(X) \times [0, 1] \rightarrow X$ 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, $x_{t+1} = \tilde{P}(x_t, g_t, f_t, w_t)$ 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 [7]. We define a function Q such that

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

and for convenience, we will abbreviate this to $Q(x_0, \tilde{\pi}, \vec{w})$ in an obvious notation, where $\vec{w} = \langle w_0, \dots, w_{n-1} \rangle \in [0,1]^n$ and $\tilde{\pi} = \{\pi_0, \dots, \pi_{n-1}\}$. We then define the *hindsight optimal value* over n -horizon with an initial state x_0 as

$$\rho_{n,\phi}(x_0) := E_{\vec{w}}[\min_{\tilde{\pi} \in \tilde{\Pi}} Q(x_0, \tilde{\pi}, \vec{w})]. \quad (15)$$

Given a policy ϕ for the maximizer, we 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 (15), this corresponds to solving the deterministic sample-path problem. See, e.g., [14, 11] for case studies on some MDP problems with this approach. 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).

The main usefulness of the above two approaches are simpleness of implementation. The exponential dependence on the horizon size in Kearns et al's approach [25] can be alleviated by using the heuristic approaches we just discussed. We can simply use a Monte-Carlo simulation to estimate the relevant function values. The minimizer who uses the *H-horizon parallel rollout policy* simulates (at the current state at each decision time) the given heuristic base policies and the (guessed) fixed policy for the maximizer using sampling over a finite horizon $H - 1$, and the results of the simulation are used to choose the best current randomized action at the current state, assuming that there is a selection function available that extracts the randomized action that achieves the infimum/supremum. For the hindsight optimization approach, the minimizer generates a set of fixed horizon random disturbance sequences and solves the deterministic optimization problems with respect to the sequences to estimate the hindsight optimal value, from which he determines the current best action at the current state.

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 [18]. 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.

6 Concluding Remarks

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 two 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., [2, 4] and references therein). However, for nonzero-sum games, we need to use a different non-standard technique (see, e.g., [3]) to analyze the structure, which is quite cumbersome.

Finally, we can incorporate the idea of Neuro-Dynamic programming (NDP) [8] 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] J. M. Alden and R. L. Smith, "Rolling horizon procedures in nonhomogeneous Markov decision processes," *Operations Research*, vol. 40, Suppl. 2, S183–S194, 1992.
- [2] E. Altman, "Zero-sum Markov games and worst-case optimal control of queueing systems," *QUESTA*, vol. 21, pp. 415–447, 1995.
- [3] E. Altman, "Non zero-sum stochastic games in admission, service and routing control in queueing systems," *QUESTA*, vol. 23, pp. 259–279, 1996.
- [4] 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.
- [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 G. J. Olsder, *Dynamic Noncooperative Game Theory*, Academic Press, London/New York, 1995.

- [7] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Volumes 1 and 2*. Athena Scientific, 1995.
- [8] D. P. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.
- [9] C. Boutilier, T. Dean and S. Hanks, “Decision-theoretic planning: structural assumptions and computational leverage,” *J. of AI Research*, vol. 11, pp. 1–94, 1999.
- [10] W. A. van den Broek, “Moving horizon control in dynamic games,” *Computing in Economics and Finance*, vol. 122, 1999.
- [11] H. S. Chang, R. Givan, and E. K. P. Chong, “On-line scheduling via sampling,” in *Proc. 5th Int. Conf. on Artificial Intelligence Planning and Scheduling*, 2000, pp. 62–71.
- [12] H. S. Chang, R. Givan, and E. K. P. Chong, “Parallel rollout for online solution of partially observable Markov decision processes,” submitted to *Discrete Event Dynamic Systems: Theory and Application*, 2000.
- [13] H. S. Chang and S. Marcus, “Approximate receding horizon approach for Markov decision processes: average reward case,” TR 2001-46, ISR, Univ. of Maryland, 2001.
- [14] 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.
- [15] J. Cruz, M. Simaan, A. Gacic, and Y. Liu, “Moving Horizon Nash Strategies for a Military Air Operation,” *IEEE Trans. on Aerospace and Electronic Systems*, vol. 38, no. 3, 2002, to appear.
- [16] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*, Springer-Verlag, 1996.
- [17] B. L. Fox and D. M. Landi, “An algorithm for identifying the ergodic subchains and transient states of a stochastic matrix,” *Comm. of the ACM*, vol. 2, pp. 619–621, 1968.
- [18] M. L. Ginsberg, “GIB: steps toward an expert-level bridge-playing program,” in *Proc. of IJCAI*, 1999, pp. 584–589.
- [19] 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.
- [20] 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.
- [21] 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.

- [22] 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.
- [23] A. J. Hoffman and R. M. Karp, "On nonterminating stochastic games," *Management Science*, vol. 19, no. 5, pp. 359–370, 1966.
- [24] L. Johansen, *Lectures on Macroeconomic Planning*. Amsterdam, The Netherlands: North-Holland, 1977.
- [25] M. Kearns, Y. Mansour, and S. Singh, "Fast planning in stochastic games," in *Proc. of UAI*, 2000.
- [26] 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.
- [27] D. Q. Mayne and H. Michalska, "Receding horizon control of nonlinear system," *IEEE Trans. Auto. Contr.*, vol. 38, no. 7, pp. 814–824, 1990.
- [28] M. Morari and J. H. Lee, "Model predictive control: past, present, and future," *Computers and Chemical Engineering*, vol. 23, pp. 667–682, 1999.
- [29] 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.
- [30] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, New York, 1994.
- [31] K. W. Ross and R. Varadarajan, "Multichain Markov decision processes with a sample-path constraint: A decomposition approach," *Math. of Op. Res.*, vol. 16, no. 1, pp. 195–207, 1991.
- [32] L. Shapley, Stochastic games, in *Proc. of the National Academy of Sciences*, vol. 39, pp. 1095–1100, 1953.
- [33] M. Tidball and E. Altman, "Approximations in dynamic zero-sum games," *SIAM J. Control and Optimization*, vol. 34, no. 1, pp. 311–328, 1996.
- [34] 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.
- [35] 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.