FINITE-STATE AVERAGE COST STOCHASTIC GAMES WITH COMPACT CONSTRAINT SETS AND A RECURRENCE CONDITION*

STEPHEN D. PATEK† AND DIMITRI P. BERTSEKAS‡

Abstract. We characterize and establish the existence of stationary equilibrium solutions for a class of finite-state average cost games. We assume that two players choose actions at each stage from compact constraint sets, enforcing some relatively mild assumptions on the transition probability and cost functions. We also assume that there is a distinguished state which is recurrent under each pair of stationary policies and that the corresponding Markov chains have a single recurrent class. In the second half of the paper, we establish the convergence of several dynamic programming algorithms.

Key words. game theory, average cost stochastic games, optimization, dynamic programming, stochastic shortest paths

AMS subject classifications. 90D15, 93E05, 49L20

1. Introduction. In this paper, we consider two player Markov decision processes where one player seeks to minimize average cost in controlling a finite state system and the other player seeks to maximize cost. The players choose actions at each stage from compact constraint sets, knowing only the state of the system. In addition to some regularity assumptions on the state transition probabilities and cost functions, we assume that the Markov chain associated with each pair of stationary policies is unichain and that there is a single state which is recurrent under all pairs of stationary policies. We will show that zero-sum games of this type have a unique equilibrium average cost which is independent of the initial state and is characterized by the essentially unique solution of Bellman's equation. We also show the convergence of several dynamic programming algorithms, including a new one: ϵ-policy iteration. Our results generalize earlier results about two-player stochastic games since our analysis applies to games where the action sets are compact (and not necessarily simplicial).

To provide a formal mathematical setting, let $S = \{1, \ldots, n\}$ denote a finite set of states. Let $U(i)$ and $V(i)$ denote the sets of actions available to the players at state i. Let

$$M = \left\{ \mu : S \mapsto \bigcup_{i \in S} U(i) \quad | \quad \mu(i) \in U(i), \quad \forall\, i \in S \right\}$$

and

$$\bar{M} = \left\{ \{\mu_0, \mu_1, \ldots\} \quad | \quad \mu_k \in M, \quad \forall\, k \right\}$$

be the sets of allowable one-stage (stationary) policies and nonstationary policies for the minimizer, respectively. Let N and $\bar{N}$ be the similarly defined sets of policies for

*Supported by the National Science Foundation Grant 9300494-DMI and an Office of Naval Research fellowship.

†Department of Systems Engineering, School of Engineering and Applied Science, University of Virginia, Olsson Hall, Charlottesville, Virginia, 22903 (sdp5f@virginia.edu)

‡Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Room 35-310, Massachusetts Institute of Technology, Cambridge, MA, 02139 (dimitrib@mit.edu)

the maximizer. The probability of transitioning from $i \in S$ to $j \in S$ under $u \in U(i)$ and $v \in V(i)$ is denoted $p_{ij}(u,v)$. The expected cost to the minimizer of transitioning from $i \in S$ under $u \in U(i)$ and $v \in V(i)$ is denoted $c_i(u,v)$. Given $\mu \in M$ and $\nu \in N$,

$$P(\mu,\nu) = \begin{bmatrix} p_{11}(\mu(1),\nu(1)) & \cdots & p_{1n}(\mu(1),\nu(1)) \\ \vdots & \vdots & \vdots \\ p_{n1}(\mu(n),\nu(n)) & \cdots & p_{nn}(\mu(n),\nu(n)) \end{bmatrix}$$

and

$$c(\mu,\nu) = \begin{pmatrix} c_1(\mu(1),\nu(1)) \\ \vdots \\ c_n(\mu(n),\nu(n)) \end{pmatrix}$$

are the corresponding transition probability matrix and expected transition cost vector, respectively. Throughout this paper we shall use $J(i)$ to denote the i-th component of a vector $J \in \Re^n$. This simplifies some of the notation in the sequel. Let $\mathbf{1} \in \Re^n$ be the vector whose components are all ones. Also, given $J, \bar{J} \in \Re^n$, we say $J \leq \bar{J}$ if $J(i) \leq \bar{J}(i)$ for every $i = 1, \ldots, n$. We now define the "dynamic programming" operators which operate on $\Re^n$:

$$T_{\mu\nu} J = c(\mu,\nu) + P(\mu,\nu)J, \qquad \mu \in M,\ \nu \in N;$$

$$T_\mu J = \sup_{\nu \in N} \left[c(\mu,\nu) + P(\mu,\nu)J \right], \qquad \mu \in M;$$

$$TJ = \inf_{\mu \in M} \sup_{\nu \in N} \left[c(\mu,\nu) + P(\mu,\nu)J \right];$$

$$\tilde{T}_\nu J = \inf_{\mu \in M} \left[c(\mu,\nu) + P(\mu,\nu)J \right], \qquad \nu \in N;$$

$$\tilde{T} J = \sup_{\nu \in N} \inf_{\mu \in M} \left[c(\mu,\nu) + P(\mu,\nu)J \right].$$

The suprema and infima in the above are taken componentwise. For example,

$$(TJ)(i) = \inf_{u \in U(i)} \sup_{v \in V(i)} \left[c_i(u,v) + \sum_{j=1}^{n} p_{ij}(u,v) J(j) \right].$$

We use the notation $T_{\mu\nu}^t J$ to denote the t-fold composition of $T_{\mu\nu}$ applied to J. Similar definitions hold for $T_\mu^t J$, $T^t J$, $\tilde{T}_\nu^t J$, and $\tilde{T}^t J$ (whenever they are well-defined).

Given a pair of nonstationary policies $\pi_M \in \bar{M}$ and $\pi_N \in \bar{N}$, the average cost to the cost-minimizing player from state i is defined to be

$$\bar{J}_{\pi_M,\pi_N}(i) = \liminf_{t \to \infty} \ \frac{1}{t+1}\, h^t_{\pi_M,\pi_N}(i), \tag{1.1}$$

where $h^t_{\pi_M,\pi_N}(i)$ is the expected $(t+1)$-stage cost from i under (π_M, π_N), defined by

$$h^t_{\pi_M,\pi_N}(i) \triangleq \left[c(\mu_0,\nu_0) + \sum_{k=1}^{t} [P(\mu_0,\nu_0)P(\mu_1,\nu_1)\cdots P(\mu_{k-1},\nu_{k-1})] c(\mu_k,\nu_k) \right](i).$$

We make the following assumptions.

Assumption RS (Recurrent State) The following are true:

1. *The Markov chain associated with each pair of stationary policies (μ, ν) is unichain.*
2. *The state $n \in S$ is recurrent under every pair of stationary policies.*

Assumption $\bar{R}$ (Regularity) The following are true:

1. *For each $i \in S$, the control constraint sets $U(i)$ and $V(i)$ are compact subsets of metric spaces.*
2. *The functions $p_{ij}(u, v)$ and $c_i(u, v)$ are continuous with respect to $(u, v) \in U(i) \times V(i)$. (In light of Proposition 7.32 in [4], this implies that the outer extrema in the operators T and $\tilde{T}$ are achieved by elements of M and N, respectively. That is, for every $H \in \Re^n$, there exists $\mu \in M$ and $\nu \in N$ such that $TH = T_\mu H \in \Re^n$ and $\tilde{T}H = \tilde{T}_\nu H \in \Re^n$.)*
3. *For every $H \in \Re^n$, we have $TH = \tilde{T}H$.*

We now make some remarks on these assumptions. Note that part 4 of Assumption $\bar{\mathrm{R}}$ is satisfied under conditions for which a minimax theorem can be used to interchange "inf" and "sup". In particular, this part, as well as the entire Assumption $\bar{\mathrm{R}}$, is satisfied if:

1. the sets $U(i)$ and $V(i)$ are nonempty, convex, and compact subsets of Euclidean spaces,
2. the functions $p_{ij}(u, v)$ are bilinear of the form $u'Q_{ij}v$, where Q_{ij} is a real matrix of dimension commensurate with $U(i)$ and $V(i)$,
3. the functions $c_i(u, v)$ are
   (a) convex and lower semi-continuous as functions of $u \in U(i)$ with v fixed in $V(i)$, and
   (b) concave and upper semi-continuous as functions of $v \in V(i)$ with u fixed in $U(i)$.

This follows from the Sion-Kakutani theorem (see [18], p.232 or [13], p. 397). The fact that the state n is recurrent (from Assumption RS) allows us to relate average cost games to stochastic shortest path games [12], a relationship we use throughout this paper. These assumptions together define the class of games that we shall refer to as *recurrent-state average cost games.*

Games with average cost objectives have been studied for a long time, starting with Gillette [7] in 1957. In [10], Liggett and Lippman used the existence of Blackwell optimal policies in one-player Markov decision problems (with finite action sets) to establish that sequential games have equilibria in pure stationary policies. In [8], after proving a result about the continuity of linear programming, Hoffman and Karp established the existence of stationary equilibrium policies in irreducible games. They also established the convergence of an average cost version of policy iteration in irreducible games. Later on, Federgruen [6] and van der Wal [19] gave successive approximation (value iteration) algorithms for these and slightly more general average cost stochastic games. In the more general context of nonzero-sum games, Stern [17] used a dynamic programming approach to show that stationary equilibrium policies exist in games where the Markov chain associated with each pair of pure policies is unichain and there is a special state which is recurrent under all pairs of pure policies. [This is our Assumption RS. Games of this type are to be distinguished from the smaller class of irreducible (recurrent) games which Gillette originally studied.] Equilibria in Stern's games are characterized (but not uniquely) by solutions to a generalized form of Bellman's equation.

All of the results cited above make use of Gillette's original assumption that the

players are optimizing with respect to mixed strategies over finite sets of actions. [In fact, we are unaware of *any* literature on average cost games (aside from the dissertation upon which this paper is based) where this assumption is relaxed.] Thus, one purpose of this paper is to show that Gillette's assumption is not essential. In general, it is not necessary to require the constraint sets $U(i)$ and $V(i)$ be simplicial and the functions $c_i(u,v)$ and $p_{ij}(u,v)$ to be bilinear. Rather, at least for some classes of games, it is sufficient to impose less restrictive topological assumptions. This, unfortunately, complicates the analysis.

In Section 2, we review the main results from [12] and establish a formal relationship between recurrent state average cost games and stochastic shortest path games. In Section 3, we use this relationship to characterize and prove the existence of equilibrium solutions for recurrent state average cost games. In Section 4, we discuss the convergence properties of several dynamic programming algorithms, including a new one called ϵ-policy iteration. In Section 5, we end the paper with a few general remarks concerning the relationship between recurrent state average cost games and stochastic shortest path games.

2. Relation to Stochastic Shortest Path Games. With Assumption RS in place we can view the recurrent state n as a terminal state which is inevitably reached in an infinite sequence of stochastic shortest path games. [12]

2.1. Stochastic Shortest Path Games: Review. Stochastic shortest path games are finite state additive cost games with compact constraint sets, where one of the states Ω is absorbing and zero-cost. (Throughout this paper Ω will be treated as an extra state, not included in $S = \{1,\ldots,n\}$. We did not observe this convention in [12].) Stochastic shortest path games are such that the minimizer wishes to drive the system to termination along a minimum expected cost path, and the maximizer seeks to maximize the cost of reaching termination. Formally, the players seek an equilibrium for the objective function

$$J_{\pi_M,\pi_N}(i) = \liminf_{t\to\infty} \; h^t_{\pi_M,\pi_N}(i), \tag{2.1}$$

where $\pi_M = \{\mu_0,\mu_1,\ldots\} \in \bar{M}$ and $\pi_N = \{\nu_0,\nu_1,\ldots\} \in \bar{N}$. A stationary policy for the minimizer which, for any policy of the maximizer, forces termination with probability one is called *proper*. A pair of policies, one for the minimizer and the other for the maximizer, which does not lead to termination with probability one is called *prolonging*. The following assumptions formally define stochastic shortest path games.

Assumption SSP The following are true:

1. *There exists at least one proper policy for the minimizer.*
2. *If a pair of policies (π_M,π_N) is prolonging, then the expected cost to the minimizer is infinite for at least one initial state. That is, there is a state i for which $\lim_{t\to\infty} h^t_{\pi_M,\pi_N}(i) = \infty$.*

Assumption R (Regularity for SSPs) The following are true:

1. *The control constraint sets are compact. That is, for each $i \in S$, $U(i)$ and $V(i)$ are compact subsets of metric spaces. (This implies that M and N are compact.)*
2. *The functions $p_{ij}(u,v)$ are continuous with respect to $(u,v) \in U(i)\times V(i)$, and the functions $c_i(u,v)$ are*
   (a) lower-semicontinuous with respect to $u \in U(i)$ (with $v \in V(i)$ fixed) and
   (b) upper–semicontinuous with respect to $v \in V(i)$ (with $u \in U(i)$ fixed).

(The Weierstrass theorem implies that the supremum and infimum in the definitions of the operators T_μ and $\tilde{T}_\nu$ are always achieved by elements of N and M, respectively. That is, for every $J \in \Re^n$, there exists $\nu \in N$ such that $T_\mu J = T_{\mu\nu} J \in \Re^n$. Similarly, for every $J \in \Re^n$, there exists $\mu \in M$ such that $\tilde{T}_\nu J = T_{\mu\nu} J \in \Re^n$.)

3. *For all $J \in \Re^n$, the infimum and supremum in the definitions of the operators T and $\tilde{T}$ are achieved by elements of M and N. That is, for every $J \in \Re^n$, there exists $\mu \in M$ and $\nu \in N$ such that $TJ = T_\mu J \in \Re^n$ and $\tilde{T}J = \tilde{T}_\nu J \in \Re^n$.*
4. *For each $J \in \Re^n$, we have $TJ = \tilde{T}J$.*

Note that Assumption R is slightly less restrictive than Assumption $\bar{\text{R}}$, and the earlier condition based on the Sion-Kakutani theorem still applies.

The following results were obtained in [12].

LEMMA 2.1. *Assume that all stationary policies for the minimizer are proper. The operator T is a contraction mapping on $\Re^n$ with respect to a weighted sup-norm $\|\cdot\|_\infty^w$, where w is a positive vector in $\Re^m$ and*

$$\|J\|_\infty^w = \max_{i \in S} |J(i)|/w(i). \tag{2.2}$$

Moreover, if $\mu \in M$ is proper, then T_μ is a contraction mapping with respect to a weighted sup-norm. [1]

PROPOSITION 2.2. *The operator T has a unique fixed point J^* on $\Re^n$.*

PROPOSITION 2.3. *The unique fixed point $J^* = TJ^*$ is the equilibrium cost of the stochastic shortest path game. There exist stationary policies $\mu^* \in M$ and $\nu^* \in N$ which achieve the equilibrium. Moreover, if $J \in \Re^n$, $\mu \in M$, and $\nu \in N$ are such that $J = TJ = T_\mu J = \tilde{T}_\nu J$, then*

1. *$J = J_{\mu,\nu}$*
2. *$J_{\pi_M,\nu} \geq J_{\mu,\nu}, \quad \forall\, \pi_M \in \bar{M}$,*
3. *$J_{\mu,\pi_N} \leq J_{\mu,\nu}, \quad \forall\, \pi_N \in \bar{N}$.*

PROPOSITION 2.4. *For every $J \in \Re^n$, there holds,*

$$\lim_{t \to \infty} T^t J = J^*, \tag{2.3}$$

where J^* is the unique equilibrium cost vector.

PROPOSITION 2.5. *Given a proper stationary policy $\mu^0 \in M$, we have that*

$$J_{\mu^k} \stackrel{\triangle}{=} \sup_{\nu \in N} J_{\mu^k,\nu} \to J^*,$$

where J^* is the unique equilibrium cost vector and $\{\mu^k\}$ is a sequence of policies (generated by policy iteration) such that $TJ_{\mu^k} = T_{\mu^{k+1}} J_{\mu^k}$ for all k.

2.2. The Relationship. Our results in [12] help us to establish the existence of equilibrium solutions for recurrent state average cost games, along with the convergence of some dynamic programming algorithms. It is useful to define for each recurrent state average cost game, along with an estimate its equilibrium average cost, the *associated stochastic shortest path game* (λ-SSPG). This the stochastic shortest path game with transition probabilities $\bar{p}_{ij}(u,v)$ and costs $\bar{c}_i(u,v)$ obtained by

[1] While not explicit in this statement of the lemma, there is a positive vector $w \in \Re^n$ and a scalar $\beta \in (0,1)$ such that $T_{\mu,\nu}, T_\mu, T, \tilde{T}_\nu$, and $\tilde{T}$ are all contractions with respect to $\|\cdot\|_\infty^w$ with modulus β. We may assume without loss of generality that the weighting on state n is unity.

1. setting $\bar{p}_{ij}(u,v) = p_{ij}(u,v)$ for all $i,j \in S$ with $j \neq n$,
2. setting $\bar{p}_{in}(u,v) = 0$ for all $i \in S$,
3. introducing an artificial terminal state Ω to which the system transitions from state i with probability $\bar{p}_{i,\Omega}(u,v) = p_{i,n}(u,v)$ for all $i \in S$, and
4. setting $\bar{c}_i(u,v) = c_i(u,v) - \lambda$ for all $i \in S$.

The definitions and observations of the following paragraphs will be useful in the sequel.

Let $J_{\lambda,\mu,\nu}(i)$ denote the cost of starting from i under the stationary policies $\mu \in M$ and $\nu \in N$ in the λ-SSPG. Let $J_{\lambda,\mu}(i) = \max_{\nu \in N} J_{\lambda,\mu,\nu}(i)$ denote the worst case cost of starting from i under μ. Let $J_\lambda(i) = \min_{\mu \in M} \max_{\nu \in N} J_{\lambda,\mu,\nu}(i)$ be the equilibrium cost of starting from i. (Note that these functions are well defined because Assumptions SSP and R are satisfied in the associated stochastic shortest path game. [12])

Note that the dynamic programming operators for the associated stochastic shortest path game are contractions with respect to a weighted sup-norm $\|\cdot\|_\infty^w$ (cf. Lemma 2.1 in this paper). Throughout the rest of this paper, we use $\|\cdot\|$ to denote such a "contractive" weighted sup-norm, whereas $\|\cdot\|_\infty$ will denote the usual sup-norm.

It is useful to relate the dynamic programming operators for average cost games and their associated stochastic shortest path games. Suppose $H \in \Re^n$ is such that $H(n) = 0$. Then, for all $i = 1, \ldots, n$,

$$\begin{aligned}(TH)(i) &= \inf_{u \in U(i)} \sup_{v \in V(i)} \left[c_i(u,v) + \sum_{j=1}^{n-1} p_{ij}(u,v) H(j) \right] \\ &= \inf_{u \in U(i)} \sup_{v \in V(i)} \left[c_i(u,v) + \sum_{j=1}^{n-1} \bar{p}_{ij}(u,v) H(j) \right].\end{aligned}$$

Thus, T applied to H in the context of an average cost game is equivalent to T applied to the equilibrium cost function estimate H in an associated stochastic shortest path game. As a result, T is a contraction on $\{H \in \Re^n \mid H(n) = 0\}$. The same is true of the other dynamic programming operators.

Let $N_{\mu,\nu}(i)$ denote the expected number of stages required to reach n in the original average cost game under the policies μ and ν starting from i. Define

$$\begin{aligned}N_{min} &= \min_{\mu \in M, \nu \in N} \min_{i=1,\ldots,n} N_{\mu,\nu}(i), \\ N_{max} &= \max_{\mu \in M, \nu \in N} \max_{i=1,\ldots,n} N_{\mu,\nu}(i).\end{aligned}$$

(Again, the maximum and minimum exist because Assumptions SSP and R are satisfied in an associated stochastic shortest path problem.) It is clear that $N_{min} \geq 1$.

LEMMA 2.6. *The following are statements are true for recurrent state average cost games.*

1. *For all $\mu \in M$, $\nu \in N$, λ, and λ'; we have*

$$J_{\lambda,\mu,\nu}(i) = J_{\lambda',\mu,\nu}(i) + (\lambda' - \lambda) N_{\mu,\nu}(i), \quad i = 1, \ldots, n. \tag{2.4}$$

2. *For all $\mu \in M$, the functions $J_{\lambda,\mu}(i)$ are continuous and decreasing as functions of λ and satisfy all $i = 1, \ldots, n$*

$$
\begin{aligned}
&J_{\lambda',\mu}(i) + N_{min}(\lambda' - \lambda) \le J_{\lambda,\mu}(i) \le J_{\lambda',\mu}(i) + N_{max}(\lambda' - \lambda), \quad \textit{if } \lambda' \ge \lambda, \\
&J_{\lambda',\mu}(i) + N_{max}(\lambda' - \lambda) \le J_{\lambda,\mu}(i) \le J_{\lambda',\mu}(i) + N_{min}(\lambda' - \lambda), \quad \textit{if } \lambda' \le \lambda.
\end{aligned} \tag{2.5}
$$

3. *The functions $J_\lambda(i)$ are continuous and decreasing as functions of λ and satisfy for all $i = 1, \ldots, n$*

$$
\begin{aligned}
&J_{\lambda'}(i) + N_{min}(\lambda' - \lambda) \le J_{\lambda}(i) \le J_{\lambda'}(i) + N_{max}(\lambda' - \lambda), \quad \textit{if } \lambda' \ge \lambda, \\
&J_{\lambda'}(i) + N_{max}(\lambda' - \lambda) \le J_{\lambda}(i) \le J_{\lambda'}(i) + N_{min}(\lambda' - \lambda), \quad \textit{if } \lambda' \le \lambda.
\end{aligned} \tag{2.6}
$$

Proof:. To prove statement 1, note that the second term on the right hand side of (2.4) is the expected cost differential associated with λ'-SSPG relative to λ-SSPG.

To prove statement 2, note that the continuity of the functions $J_{\lambda,\mu}(i)$ follows from Proposition 7.32 in [4] and the joint continuity of $J_{\lambda,\mu,\nu}(i)$ with respect to λ, μ, and ν. To see that the $J_{\lambda,\mu}(i)$ are decreasing, let $\lambda_1 < \lambda_2$ be given. For some $\bar{\nu} \in N$ we have

$$
\begin{aligned}
J_{\lambda_2,\mu}(i) &= J_{\lambda_2,\mu,\bar{\nu}}(i) \\
&= J_{\lambda_1,\mu,\bar{\nu}}(i) + (\lambda_1 - \lambda_2) N_{\mu,\bar{\nu}}(i) \\
&< J_{\lambda_1,\mu,\bar{\nu}}(i) \\
&\le J_{\lambda_1,\mu}(i).
\end{aligned}
$$

Finally, to see (2.5), let $\lambda' \ge \lambda$ be given; then, for all $\nu \in N$ we have $J_{\lambda,\mu,\nu}(i) = J_{\lambda',\mu,\nu}(i) + (\lambda' - \lambda) N_{\mu,\nu}(i) \ge J_{\lambda',\mu,\nu}(i) + (\lambda' - \lambda) N_{min}$. The right-most expression is maximized by some $\bar{\nu} \in N$. Thus,

$$
\begin{aligned}
J_{\lambda,\mu}(i) &\ge J_{\lambda,\mu,\bar{\nu}} \\
&\ge J_{\lambda',\mu,\bar{\nu}} + (\lambda' - \lambda) N_{min} \\
&= J_{\lambda',\mu} + (\lambda' - \lambda) N_{min}.
\end{aligned}
$$

The remaining inequalities of (2.5) follow similarly.

To prove statement 3, note that the continuity of $J_\lambda(i)$ follows from Proposition 7.32 in [4] and the joint continuity of $J_{\lambda,\mu}(i)$ with respect to λ and μ. To see that the $J_\lambda(i)$ are decreasing, let $\lambda_1 < \lambda_2$ be given; then, for some $\bar{\mu} \in M$ we have

$$
\begin{aligned}
J_{\lambda_1}(i) &= J_{\lambda_1,\bar{\mu}}(i) \\
&> J_{\lambda_2,\bar{\mu}}(i) \\
&\ge J_{\lambda_2,\mu}(i) \\
&\ge J_{\lambda_2}(i).
\end{aligned}
$$

Finally, we obtain (2.6) from (2.5) and similar arguments. **Q.E.D.**

It can be shown that the functions $J_{\lambda,\mu}(i)$ are convex with respect to λ. However, the functions $J_\lambda(i)$ are generally neither convex nor concave; they are only strictly decreasing as stated above.

3. Existence and Characterization of Equilibrium Solutions. We now establish the existence of stationary equilibrium solutions in recurrent state average cost games. We characterize the equilibrium value function as the effectively unique solution to a form of Bellman's equation. The results of this section can be viewed

as a generalization of Stern's results in [17] (cf. Chapter 2, restricted to the zero-sum case).

PROPOSITION 3.1. *The following statements are true for recurrent state average cost games.*

1. *There is a unique equilibrium average cost from each state. The equilibrium average cost is the same for each state and is denoted λ^*. There is a function $H^* \in \Re^n$ which, along with λ^*, satisfies Bellman's equation*

$$\lambda^* \mathbf{1} + H^* = TH^*. \tag{3.1}$$

   Furthermore, if $\mu \in M$ achieves the minimum in TH^* and $\nu \in N$ achieves the maximum in $\tilde{T}H^*$, then (μ, ν) forms an equilibrium solution for the average cost game. Out of all solutions (λ, H) to (3.1), there is a unique solution for which $H(n) = 0$.
2. *If a scalar λ and a function $H \in \Re^n$ satisfy (3.1), then λ is exactly the equilibrium average cost for each initial state.*
3. *Given a stationary policy $\mu \in M$, the corresponding worst-case average cost λ_μ, along with a unique function $H_\mu \in \Re^n$ such that $H_\mu(n) = 0$, satisfy*

$$\lambda_\mu \mathbf{1} + H_\mu = T_\mu H_\mu.$$

4. *Given a stationary policy $\nu \in N$, the corresponding worst-case average cost λ_ν, along with a unique function $H_\nu \in \Re^n$ such that $H_\nu(n) = 0$, satisfy*

$$\lambda_\nu \mathbf{1} + H_\nu = \tilde{T}_\nu H_\nu.$$

5. *Given stationary policies $\mu \in M$ and $\nu \in N$, the corresponding average cost $\lambda_{\mu\nu}$, along with a unique function $H_{\mu\nu} \in \Re^n$ such that $H_{\mu\nu}(n) = 0$, satisfy*

$$\lambda_{\mu\nu} \mathbf{1} + H_{\mu\nu} = T_{\mu\nu} H_{\mu\nu}.$$

Proof:. We first prove part 3. Let $C_{\mu,\nu}(n)$ denote the expected cost starting from n up to the first return to n under the policies $\mu \in M$ and $\nu \in N$ in the average cost game. Let $N_{\mu,\nu}(n)$ denote the expected number of stages to return to n starting from n, as defined earlier. Considering the 0-SSPG, we know from our results about stochastic shortest path games and Assumptions $\bar{\mathrm{R}}$ and RS, that $C_{\mu,\nu}(n)$ and $N_{\mu,\nu}(n)$ are bounded and continuous on the compact product space $M \times N$. Since $N_{\mu,\nu}(n) \geq 2$ for all μ and ν, the quotient $C_{\mu,\nu}(n)/N_{\mu,\nu}(n)$ is also continuous. As a result, with $\mu \in M$ fixed, there is a policy $\nu_\mu \in N$ which achieves the supremum in

$$\tilde{\lambda}_\mu \stackrel{\triangle}{=} \sup_{\nu \in N} \frac{C_{\mu,\nu}(n)}{N_{\mu,\nu}(n)}.$$

Thus,

$$\phi_\mu(\nu) \stackrel{\triangle}{=} \left\{ \frac{C_{\mu,\nu}(n) - \tilde{\lambda}_\mu N_{\mu,\nu}(n)}{N_{\mu,\nu}(n)} \right\} \leq 0.$$

Moreover, since $N_{\mu,\nu}(n)$ is bounded and greater than or equal to one, the following are true:

1. $C_{\mu,\nu}(n) - \tilde{\lambda}_\mu N_{\mu,\nu}(n) \leq 0$ for all $\nu \in N$, and
2. $\phi_\mu(\nu) = 0$ if and only if $C_{\mu,\nu}(n) - \tilde{\lambda}_\mu N_{\mu,\nu}(n) = 0$.

Since $\phi_\mu(\nu_\mu) = 0$, we have that ν_μ maximizes $C_{\mu,\nu}(n) - \tilde{\lambda}_\mu N_{\mu,\nu}(n)$. The rest of the proof for part 3 follows from arguments similar to those for Proposition 4.1 in Chapter 7 of [1]. Parts 4 and 5 follow similarly.

To show part 1, note that for each $\mu \in M$ there is a policy $\nu_\mu \in N$ which achieves the supremum in $\sup_{\nu\in N} C_{\mu,\nu}(n)/N_{\mu,\nu}(n)$. From Proposition 7.32 in [4], the function $C_{\mu,\nu}(n)/N_{\mu,\nu}(n)$ is continuous as a function of $\mu \in M$. Thus, there exists a minimax optimal policy $\tilde{\mu}$ which achieves the infimum in

$$\tilde{\lambda} \stackrel{\triangle}{=} \inf_{\mu\in M} \sup_{\nu\in N} \frac{C_{\mu,\nu}(n)}{N_{\mu,\nu}(n)}.$$

Observe that for all $\mu \in M$

$$\phi(\mu) \stackrel{\triangle}{=} \sup_{\nu\in N} \left\{ \frac{C_{\mu,\nu}(n) - \tilde{\lambda} N_{\mu,\nu}(n)}{N_{\mu,\nu}(n)} \right\} \geq 0.$$

Moreover, since $N_{\mu,\nu}(n)$ is bounded and greater than or equal to one, the following are true:

1. $\sup_{\nu\in N} \left\{ C_{\mu,\nu}(n) - \tilde{\lambda} N_{\mu,\nu}(n) \right\} \geq 0$ for all $\mu \in M$, and
2. $\phi(\mu) = 0$ if and only if

$$\sup_{\nu\in N} \left\{ C_{\mu,\nu}(n) - \tilde{\lambda} N_{\mu,\nu}(n) \right\} = 0.$$

Since $\phi(\tilde{\mu}) = 0$, we have that $\tilde{\mu}$ minimizes $\sup_{\nu\in N} \left\{ C_{\mu,\nu}(n) - \tilde{\lambda} N_{\mu,\nu}(n) \right\}$.

Now consider the associated stochastic shortest path game, $\tilde{\lambda}$-SSPG. Since Assumptions $\bar{\text{R}}$ and RS are in effect, the $\tilde{\lambda}$-SSPG satisfies Assumptions SSP and R. As a result there exists a unique function $H^* \in \Re^n$ (equal to $J_{\tilde{\lambda}}$) such that

$$H^*(i) = \min_{u\in U(i)} \max_{v\in V(i)} \left[c_i(u,v) - \tilde{\lambda} + \sum_{j=1}^{n-1} p_{ij}(u,v) H^*(j) \right], \quad i \in \{1,\ldots,n\},$$

where we have used the fact that $\bar{p}_{in}(u,v) = 0$. In fact, H^* represents the equilibrium cost-to-go function for the associated stochastic shortest path game. An equilibrium policy $\mu^* \in M$ minimizes $\sup_{\nu\in N} \left\{ C_{\mu,\nu}(n) - \tilde{\lambda} N_{\mu,\nu}(n) \right\}$, reducing it to zero [given our previous observation about $\tilde{\mu}$]. Thus, $H^*(n) = J_{\tilde{\lambda}}(n) = 0$ and

$$H^*(i) + \tilde{\lambda} = \min_{u\in U(i)} \max_{v\in V(i)} \left[c_i(u,v) + \sum_{j=1}^{n} p_{ij}(u,v) H^*(j) \right], \quad i \in \{1,\ldots,n\}.$$

Moreover, by Assumption $\bar{\text{R}}$,

$$H^*(i) + \tilde{\lambda} = \max_{v\in V(i)} \min_{u\in U(i)} \left[c_i(u,v) + \sum_{j=1}^{n} p_{ij}(u,v) H^*(j) \right], \quad i \in \{1,\ldots,n\}. \tag{3.2}$$

Because we have found an equilibrium of the associated stochastic shortest path game a policy $\nu^* \in N$ which achieves the maximum in (3.2) for all states $i \in S$ maximizes $\inf_{\mu\in M} \left\{ C_{\mu,\nu}(n) - \tilde{\lambda} N_{\mu,\nu}(n) \right\}$. (Such a policy exists thanks to Assumption $\bar{\text{R}}$.)

Now consider the one-player average cost problem which results when the minimizer announces the use of μ^*. The maximizer is left with a unichain average cost problem for which the state n is recurrent under all stationary policies. From part 3, the Bellman equations above characterize the average cost of this problem, resulting in the fact that for all states $i \in S$

$$\tilde{\lambda} = \sup_{\pi_N \in \bar{N}} \bar{J}_{\mu^*, \pi_N}(i).$$

Similarly, if the maximizer announces ν^* then we have that for all states i

$$\tilde{\lambda} = \inf_{\pi_M \in \bar{M}} \bar{J}_{\pi_M, \nu^*}(i).$$

Combining these observations, we obtain

$$\inf_{\pi_M \in \bar{M}} \sup_{\pi_N \in \bar{N}} \bar{J}_{\pi_M, \pi_N} \leq \sup_{\pi_N \in \bar{N}} \inf_{\pi_M \in \bar{M}} \bar{J}_{\pi_M, \pi_N}.$$

This, along with the usual minimax inequality, implies that equality holds and a constant-valued equilibrium average cost $\lambda^* = \tilde{\lambda}$ exists. It is apparent that μ^* and ν^* form an equilibrium solution for the average cost game.

Part 2 follows from similar arguments. **Q.E.D.**

COROLLARY 3.2. *The following are true regarding recurrent state average cost games.*

1. $J_{\lambda,\mu,\nu}(n) = 0$ *if and only if* $\lambda = \lambda_{\mu\nu}$, *where* $\lambda_{\mu\nu}$ *is the average cost associated with* $\mu \in M$ *and* $\nu \in N$.
2. $J_{\lambda,\mu}(n) = 0$ *if and only if* $\lambda = \lambda_\mu$, *where* λ_μ *is the worst case average cost associated with* $\mu \in M$.
3. $J_\lambda(n) = 0$ *if and only if* $\lambda = \lambda^*$, *where* λ^* *is the equilibrium average cost of the game.*

For single-player problems with finite action sets, it is possible to exploit the connection with stochastic shortest path problems to analyze the full class of unichain average cost problems (where it is not required that there be a special state which is always recurrent). In particular, it is possible (as in [2]) to use the existence of Blackwell optimal policies[2] to show that if every policy which is optimal within the class of stationary policies is unichain, then there exists a solution to Bellman's equation and the optimal average cost is independent of the initial state. If we allow the constraint sets to be arbitrary compact subsets of metric spaces, then the existence of Blackwell optimal policies is not clear. As a result, the analysis of [2] cannot be generalized easily to prove the existence of a solution to Bellman's equation for single-player unichain average cost problems with compact constraint sets.[3] Similarly, the analysis of [2] cannot be generalized easily to prove the existence of solutions to (3.1) in unichain games satisfying Assumption $\bar{\text{R}}$.[4]

4. Dynamic Programming Algorithms for Recurrent State Average Cost Games. In this section we state and discuss the convergence properties of several dynamic programming algorithms.

[2] A policy is Blackwell optimal if it is optimal for all discount factors α in a neighborhood of 1.

[3] For single-player unichain problems, the existence of solutions to Bellman's equation under was established in [15] by other methods.

[4] In [16], Sobel established the existence of stationary equilibrium solutions in N-player, nonzero-sum games where, for each profile of pure stationary policies, there is a single class of communicating states. Rogers [14] obtained similar results using a different technique.

4.1. Value Iteration. As shown in the following proposition, given any terminal cost function $J \in \Re^n$, the k-horizon equilibrium cost divided by k approaches the equilibrium average cost of the game.

PROPOSITION 4.1. *For every $J \in \Re^n$,*

$$\lim_{k\to\infty} \frac{1}{k} T^k J = \lambda^* \mathbf{1},$$

where λ^* is the equilibrium average cost of the recurrent state average cost game.

Proof:. The proof is nearly identical to an argument in [2] (cf. pages 318-319). The only difference lies in the fact that our T operator involves a minimax operation. Since T remains nonexpansive, the proof goes through with the same algebraic manipulations. **Q.E.D.**

4.2. Relative Value Iteration. An important practical difficulty of the value iteration method is that $|(T^k J)(i)|$ may approach infinity for some states i. Moreover, the method does not produce an estimate of the equilibrium differential cost function H^*. The relative value iteration method presented here is designed to address these issues. Unfortunately, to assure convergence, extra assumptions must be satisfied.

ALGORITHM 4.1. *(Relative Value Iteration)*

1. *Choose $\tau \in (0,1]$, $t \in S$, and an initial $H_0 \in \Re^n$.*
2. *Given H_k, compute*

$$H_{k+1} = (1-\tau)H_k + T(\tau H_k) - T(\tau H_k)(t)\mathbf{1}.$$

If this algorithm converges, say to $\bar{H}$, then the limit satisfies

$$\tau\bar{H} + T(\tau\bar{H})(t)\mathbf{1} = T(\tau\bar{H}).$$

Consequently, $T(\tau H_k)(t)$ converges to the equilibrium average cost λ^*, and it is true that $\bar{H} = (1/\tau)H^*$, where H^* is the unique solution to $TH = H + \lambda^*\mathbf{1}$ with $H^*(t) = 0$. Unfortunately, convergence is not clear without imposing extra conditions, as in the following proposition.[5]

PROPOSITION 4.2. *In addition to Assumptions $\bar{R}$ and RS, assume that there exists a positive integer m such that for every pair of admissible policies $\pi_M = \{\mu_0, \mu_1, \ldots\} \in \bar{M}$ and $\pi_N = \{\nu_0, \nu_1, \ldots\} \in \bar{N}$, there exists an $\epsilon > 0$ such that*

$$[P(\mu_m, \nu_m)P(\mu_{m-1}, \nu_{m-1}) \ldots P(\mu_1, \nu_1)]_{in} \geq \epsilon, \qquad i = 1, \ldots, n,$$
$$[P(\mu_{m-1}, \nu_{m-1})P(\mu_{m-2}, \nu_{m-2}) \ldots P(\mu_0, \nu_0)]_{in} \geq \epsilon, \qquad i = 1, \ldots, n,$$

where $[\cdot]_{in}$ denotes the element of the ith row and nth column of the corresponding matrix. Then, setting $t = n$ in relative value iteration, the sequence H_k converges to a vector H such that $(TH)(n)\mathbf{1} + H = TH$. (This implies $(TH)(n)$ is equal to the equilibrium average cost of the game.)

[5] If we set t to be the recurrent state n and we choose the initial cost function H_0 such that $H_0(n) = 0$, then we have $H_k(n) = 0$ for every $k \geq 1$. Thus, every time the T operator is applied in relative value iteration, it acts like a contraction mapping. Unfortunately, this does not seem to be of much help in establishing the convergence of the method.

Proof:. Let μ_k be such that $TH_k = T_{\mu_k} H_k$ and define $\lambda_k = (TH_k)(n)$, for every k. We have

$$H_{k+1} = T_{\mu_k} H_k - \lambda_k \mathbf{1} \leq T_{\mu_{k-1}} H_k - \lambda_k \mathbf{1}$$
$$H_k = T_{\mu_{k-1}} H_{k-1} - \lambda_{k-1} \mathbf{1} \leq T_{\mu_k} H_{k-1} - \lambda_{k-1} \mathbf{1}.$$

Defining $q_k = H_{k+1} - H_k$, we obtain from the above inequalities

$$q_k \geq T_{\mu_k} H_k - T_{\mu_k} H_{k-1} + (\lambda_{k-1} - \lambda_k)\mathbf{1}$$
$$q_k \leq T_{\mu_{k-1}} H_k - T_{\mu_{k-1}} H_{k-1} + (\lambda_{k-1} - \lambda_k)\mathbf{1}.$$

Let $\underline{\nu}_k$ be such that $T_{\mu_k} H_{k-1} = T_{\mu_k \underline{\nu}_k} H_{k-1}$ and similarly let $\bar{\nu}_k$ be such that

$$T_{\mu_{k-1}} H_k = T_{\mu_{k-1} \bar{\nu}_k} H_k,$$

for every k. Consequently,

$$q_k \geq P(\mu_k, \underline{\nu}_k) q_{k-1} + (\lambda_{k-1} - \lambda_k)\mathbf{1}$$
$$q_k \leq P(\mu_{k-1}, \bar{\nu}_k) q_{k-1} + (\lambda_{k-1} - \lambda_k)\mathbf{1}.$$

Since relations like this hold for all $k \geq 1$, we obtain

$$q_k \geq [P(\mu_k, \underline{\nu}_k) \dots P(\mu_{k-m+1}, \underline{\nu}_{k-m+1})] q_{k-1} + (\lambda_{k-m} - \lambda_k)\mathbf{1} \tag{4.1}$$
$$q_k \leq [P(\mu_{k-1}, \bar{\nu}_k) \dots P(\mu_{k-m}, \bar{\nu}_{k-m+1})] q_{k-1} + (\lambda_{k-m} - \lambda_k)\mathbf{1}. \tag{4.2}$$

By our assumption about the recurrent state n, there are two scalars $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that

$$\left[P(\mu_k, \underline{\nu}_k) \dots P(\mu_{k-m+1}, \underline{\nu}_{k-m+1})\right]_{in} \geq \epsilon_1, \qquad i = 1, \dots, n,$$
$$\left[P(\mu_{k-1}, \bar{\nu}_k) \dots P(\mu_{k-m}, \bar{\nu}_{k-m+1})\right]_{in} \geq \epsilon_2, \qquad i = 1, \dots, n.$$

From (4.2), we obtain

$$q_k(i) \leq (1 - \epsilon) \max_j q_{j-m}(j) + \lambda_{k-m} - \lambda_k, \quad i = 1, \dots, n,$$

where $\epsilon = \min\{\epsilon_1, \epsilon_2\}$. Thus,

$$\max_j q_k(j) \leq (1 - \epsilon) \max_j q_{j-m}(j) + \lambda_{k-m} - \lambda_k.$$

Similarly, from (4.1), we obtain

$$\min_j q_k(j) \geq (1 - \epsilon) \min_j q_{j-m}(j) + \lambda_{k-m} - \lambda_k.$$

Subtracting the last two inequalities, we get

$$\max_j q_k(j) - \min_j q_k(j) \leq (1 - \epsilon) \left(\max_j q_{k-m}(j) - \min_j q_{k-m}(j) \right),$$

and the rest of the argument follows the proof of Proposition 3.1 in Chapter 4 of [2]. **Q.E.D.**

As described in [2] for single-player problems, it is possible to extend this result to the case where $t \neq n$. Moreover, if the number of policies available to the respective players is finite, then setting $\tau < 1$ can be viewed as a data transformation which gives rise to a game with the aperiodic structure required in the hypothesis of the proposition. Proposition 4.2 generalizes the earlier result of Federgruen in [5] (Cf. Theorem 2, part (a)), where relative value iteration is shown to converge under the data transformation above for unichain games with mixed strategies over finite action sets.

4.3. Contracting Value Iteration. The next method we describe is a new type of value iteration for recurrent-state average cost games. It generalizes a similar algorithm for single-player problems described in [3] and is motivated by the connection with stochastic shortest path games.

ALGORITHM 4.2. *(Contracting Value Iteration)*

1. *Start with an initial estimate* (λ_0, H_0) *of a solution to Bellman's equation (3.1).*
2. *Given* (λ_k, H_k),
   (a) *first compute* $H_{k+1} = -\lambda_k \mathbf{1} + TH_k$, *and then*
   (b) *compute* $\lambda_{k+1} = \lambda_k + \gamma_k H_{k+1}(n)$.

PROPOSITION 4.3. *For every recurrent state average cost game, there exists a positive stepsize* $\bar{\gamma}$ *such that if*

$$\underline{\gamma} \leq \gamma_k \leq \bar{\gamma}$$

for some minimal positive stepsize $\underline{\gamma}$ *and all* k, *the sequence* (λ_k, H_k) *generated by contracting value iteration converges linearly to the unique solution* (λ^*, H^*) *of Bellman's equation (3.1) with* $H^*(n) = 0$.

Proof:. The proof uses Lemma 2.6 and Corollary 3.2 and closely follows the proof of Proposition 1 in [3]. To see this, associate $J_{\lambda,\mu}(i)$ with $h_{\lambda,\mu}(i)$ and $J_\lambda(i)$ with $h_\lambda(i)$. What is important is that these functions are

1. continuous and decreasing with bounded slope, and
2. the upper bound on their slopes is strictly less than zero. (The slopes of these functions lie between $-N_{\max}$ and $-N_{min}$.)

Q.E.D.

4.4. Policy Iteration. We now examine the policy iteration algorithm of Hoffman and Karp [8].

ALGORITHM 4.3. *(Policy Iteration)*

1. *Choose an initial stationary policy* $\mu_0 \in M$.
2. *Given* $\mu_k \in M$:
   (a) *(Policy Evaluation) Compute the unique solution* $(\lambda_{\mu_k}, H_{\mu_k})$ *to the equations*

$$\lambda \mathbf{1} + H = T_{\mu_k} H,$$
$$H(n) = 0.$$

   (b) *(Policy Improvement) Compute* $\mu_{k+1} \in M$ *such that* $TH_{\mu_k} = T_{\mu_{k+1}} H_{\mu_k}$.

This algorithm is known to converge [8] when both

1. $U(i)$ and $V(i)$ represent mixed strategies over finite sets of actions, and
2. the Markov chain associated with each pair of pure policies is irreducible.

The following proposition gives a monotonicity result under the less restrictive conditions of Assumption $\bar{\text{R}}$ and RS. Unfortunately, it falls short of actually proving convergence to a solution of Bellman's equation.

PROPOSITION 4.4. *For each* k *in policy iteration applied to a recurrent state average cost game, we either have*

$$\lambda_{\mu_{k+1}} < \lambda_{\mu_k}$$

or

$$\lambda_{\mu_{k+1}} = \lambda_{\mu_k}, \quad H_{\mu_{k+1}} \leq H_{\mu_k}.$$

If equality prevails in the latter, then both μ_k and μ_{k+1} are stationary equilibrium policies for the minimizer.

Proof:. Let $\{\mu_k\}$ be a sequence of stationary policies generated by policy iteration. Consider μ_k; we will show that either $\lambda_{\mu_{k+1}} < \lambda_{\mu_k}$ or else $\lambda_{\mu_{k+1}} = \lambda_{\mu_k}$ and $H_{\mu_{k+1}} \leq H_{\mu_k}$. Set $J_0 = H_{\mu_k}$, and define

$$J_m = T_{\mu_{k+1}} J_{m-1}.$$

Note that J_m is the m-stage worst-case (additive) cost function associated with the minimizer's policy μ_{k+1} when the terminal cost function is H_{μ_k}. Thanks to Proposition 4.1 we have that for every $i \in S$

$$\lambda_{\mu_{k+1}} = \lim_{m\to\infty} \frac{1}{m} J_m(i).$$

By Proposition 3.1 and the definition μ_{k+1} and J_0,

$$J_1 = TJ_0 = T_{\mu_{k+1}} J_0 \leq T_{\mu_k} J_0 = \lambda_{\mu_k} \mathbf{1} + J_0.$$

Consequently,

$$\begin{aligned} J_2 &= T_{\mu_{k+1}} J_1 \\ &\leq T_{\mu_{k+1}} (\lambda_{\mu_k} \mathbf{1} + J_0) \\ &= \lambda_{\mu_k} \mathbf{1} + T_{\mu_{k+1}} J_0 \\ &\leq 2\lambda_{\mu_k} \mathbf{1} + J_0, \end{aligned}$$

where the second equality follows from the fact that there is no terminal state in our formulation of average cost games. Proceeding inductively, we obtain

$$J_m \leq m\lambda_{\mu_k} \mathbf{1} + J_0.$$

Thus,

$$\frac{1}{m} J_m \leq \lambda_{\mu_k} \mathbf{1} + \frac{1}{m} J_0$$

and by taking the limit as $m \to \infty$ we obtain $\lambda_{\mu_{k+1}} \leq \lambda_{\mu_k}$.

If $\lambda_{\mu_{k+1}} = \lambda_{\mu_k}$, then we can interpret $H_{\mu_{k+1}}$ as the worst case cost of μ_{k+1} produced by a policy iteration step in the associated stochastic shortest path game λ_{μ_k}-SSPG. From the monotonicity of policy iteration for stochastic shortest path games, it follows that $H_{\mu_{k+1}} \leq H_{\mu_k}$.

If $\lambda_{\mu_{k+1}} = \lambda_{\mu_k}$ and $H_{\mu_{k+1}} = H_{\mu_k}$, then

$$\begin{aligned} \lambda_{\mu_k} \mathbf{1} + H_{\mu_k} &= \lambda_{\mu_{k+1}} \mathbf{1} + H_{\mu_{k+1}} \\ &= T_{\mu_{k+1}} H_{\mu_{k+1}} \\ &= T_{\mu_{k+1}} H_{\mu_k} \\ &= T H_{\mu_k}. \end{aligned}$$

Thus, λ_{μ_k} and H_{μ_k} satisfy Bellman's equation, and Proposition 3.1 implies that both μ_k and μ_{k+1} are equilibrium policies for the minimizer. **Q.E.D.**

COROLLARY 4.5. *If the minimizer has only finitely many policies, then policy iteration converges in a finite number of iterations.*

Convergence of policy iteration in the more general case [where $U(i)$ and $V(i)$ are compact subsets of metric spaces] is not clear. The possibility exists that λ_{μ_k} will converge to some $\bar{\lambda} > \lambda^*$ with $\lambda_{\mu_k} < \lambda_{\mu_{k+1}}$ for every k.[6]

4.5. ϵ-Policy Iteration. In this subsection we describe a variation of policy iteration which yields policies that are arbitrarily close to equilibrium. The basic idea is to implement conventional policy iteration (as in Algorithm 4.3) as long as the corresonding improvments in average cost are greater in magnitude than some fixed $\epsilon > 0$. If at stage k a conventional policy iteration step does not result in this much of an improvement, then the prevailing estimate λ_{μ_k} of the equilibrium average cost is held fixed while policy iteration for the λ_{μ_k}-SSPG is implemented. The inequalities of Lemma 2.6 give rise to a stopping criterion for the inner loop so that termination results in an improvement in average cost that is bounded away from zero.

ALGORITHM 4.4. *(ϵ-Policy Iteration)*

1. *Choose $\epsilon > 0$ and an initial policy $\mu_0 \in M$. Compute the unique solution $(\lambda_{\mu_0}, H_{\mu_0})$ to the equations*

$$T_{\mu_0} H = H + \lambda \mathbf{1},$$
$$H(n) = 0.$$

2. *Given $(\mu_k, \lambda_{\mu_k}, H_{\mu_k})$,*
   (a) Choose $\bar{\mu}$ such that

$$T H_{\mu_k} = T_{\bar{\mu}} H_{\mu_k}$$

and compute the unique solution $(\lambda_{\bar{\mu}}, H_{\bar{\mu}})$ to the equations

$$T_{\bar{\mu}} H = H + \lambda \mathbf{1},$$
$$H(n) = 0.$$

(b) If $\lambda_{\bar{\mu}} < \lambda_{\mu_k} - \epsilon$, then set

$$(\mu_{k+1},\, \lambda_{\mu_{k+1}},\, H_{\mu_{k+1}}) = (\bar{\mu},\, \lambda_{\bar{\mu}},\, H_{\bar{\mu}}).$$

Otherwise, set $\tilde{\mu}_0 = \bar{\mu}$ and iterate as follows. Given $\tilde{\mu}_j$,

i. Compute the unique solution $\tilde{H}_{\tilde{\mu}_j}$ to the equation

$$T_{\tilde{\mu}_j} \tilde{H} = \tilde{H} + \lambda_{\mu_k} \mathbf{1}.$$

ii. If $\tilde{H}_{\tilde{\mu}_j}(n) < -\epsilon$, then stop this inner loop; set $\mu_{k+1} = \tilde{\mu}_j$ and compute the unique solution $(\lambda_{\mu_{k+1}}, H_{\mu_{k+1}})$ to the equations

$$T_{\mu_{k+1}} H = H + \lambda \mathbf{1},$$
$$H(n) = 0.$$

Otherwise, continue the inner loop by choosing $\tilde{\mu}_{j+1}$ such that

$$T \tilde{H}_{\tilde{\mu}_j} = T_{\tilde{\mu}_{j+1}} \tilde{H}_{\tilde{\mu}_j}.$$

[6] Unfortunately, we cannot pursue the type of analysis of Hordijk and Puterman in [9] which relies upon a Newton's method interpretation of (single-player) policy iteration which does not generalize to the two-player case.

The following observations are useful in interpreting this algorithm.

1. The process of computing the unique solution (λ_μ, H_μ) to the equations $T_\mu H = H + \lambda \mathbf{1}$ and $H(n) = 0$ is equivalent to computing the maximal average cost in the single-player Markov decision problem which prevails when the minimizer specifies μ. By Corollary 3.2, λ_μ is the unique scalar such that $J_{\lambda_\mu,\mu}(n) = 0$.
2. Given μ and λ [where λ is possibly not equal to λ_μ (the worst-case average cost of μ)], the process of computing the unique solution $\bar{H}$ such that $T_\mu H = H + \lambda \mathbf{1}$ is equivalent to the computing the worst case cost of μ in the λ-SSPG. Thus, $\bar{H} = J_{\lambda,\mu}$. If $\lambda = \lambda_\mu$, then $\bar{H}(n) = J_{\lambda_\mu,\mu}(n) = 0$. Moreover, if μ' is such that $T\bar{H} = T_{\mu'}\bar{H}$, then μ' is the policy that results from one policy iteration step in the λ-SSPG, and $J_{\lambda,\mu'} \leq \bar{H}$.

PROPOSITION 4.6. *After a finite number of global iterations, the ϵ-policy iteration method will keep executing (get stuck in) the inner loop of step 2.(b)ii, and the μ_k which prevails is such that $\lambda_{\mu_k} - \lambda^* < \epsilon$.*

Proof:. Since λ_{μ_k} is the worst case average cost associated with μ_k, we have that $\lambda_{\mu_k} \geq \lambda^*$ for all k. Consider the global update where we start with $(\mu_k, \lambda_{\mu_k}, H_{\mu_k})$. If the resulting $\bar{\mu}$ is such that $\lambda_{\bar{\mu}} < \lambda_{\mu_k} - \epsilon$, then because we choose $\mu_{k+1} = \bar{\mu}$ the resulting improvement in worst case cost is at least ϵ/N_{max}. Otherwise, there are two cases to consider.

1. If $J_{\lambda_{\mu_k}}(n) < -\epsilon$, then, because policy iteration for the λ_{μ_k}-SSPG converges, the inner loop will terminate with some $\tilde{\mu}_j$ for which $J_{\lambda_{\mu_k},\bar{\mu}_j}(n) < -\epsilon$. Since $J_{\lambda,\bar{\mu}_j}(n)$ is strictly decreasing as a function of λ, it is true that $\lambda_{\bar{\mu}_j} < \lambda_{\mu_k}$. Moreover, from (2.5), associating λ' with λ_{μ_k}, λ with $\lambda_{\bar{\mu}_j}$, and μ with $\tilde{\mu}_j$, we have that

$$0 = J_{\lambda_{\tilde{\mu}_j},\bar{\mu}_j}(n) \leq J_{\lambda_{\mu_k},\bar{\mu}_j}(n) + N_{max}(\lambda_{\mu_k} - \lambda_{\bar{\mu}_j}),$$

   which implies that $\lambda_{\mu_k} - \lambda_{\bar{\mu}_j} \geq \epsilon/N_{max}$. Since we choose $\mu_{k+1} = \tilde{\mu}_j$, the resulting global update results in an improvement of at least ϵ/N_{max}.
2. If $J_{\lambda_{\mu_k}}(n) \geq -\epsilon$, then the inner loop of the algorithm will never terminate. From (2.6), associating λ' with λ_{μ_k} and λ with λ^*, we see that

$$J_{\lambda_{\mu_k}}(n) + N_{min}(\lambda_{\mu_k} - \lambda^*) \leq J_{\lambda^*}(n) = 0.$$

   Thus, $\lambda_{\mu_k} - \lambda^* \leq \epsilon/N_{min} \leq \epsilon$.

Since there can be only finitely many improvements of ϵ/N_{max}, the algorithm must eventually get stuck in step 2.(b)ii. **Q.E.D.**

5. Conclusion. An implicit purpose of this paper is to illustrate connections between stochastic shortest path games and average cost stochastic games. It is appropriate to search for such connections since both can be viewed generally as "undiscounted" games. It turns out (cf. [11]) that previously existing theory for average cost games can be used to prove a subset of the results established in [12] about stochastic shortest path games. Unfortunately, this line of reasoning does not apply to the general case where $U(i)$ and $V(i)$ are compact subsets of metric spaces. On the other hand, as we have shown in this paper, the results of [12] can be used to extend the theory of recurrent-state average cost games, namely the existence of an equilibrium value for recurrent state games when $U(i)$ and $V(i)$ are arbitrary compact subsets of metric spaces and appropriate regularity assumptions are imposed. The equilibrium value along with an equilibrium differential cost vector is characterized as the

essentially unique solution to Bellman's equation and can be achieved by stationary policies for the opposing players. We also examined several dynamic programming algorithms for recurrent-state average cost games. One important conclusion to be drawn from this paper is that it is not necessary to assume finite underlying action sets and mixed strategies to obtain powerful results for a broad class of average cost games: many of the usual results hold for the case of compact constraint sets, at least under the recurrent state assumption.

REFERENCES

[1] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1, Athena Scientific, Belmont, MA, 1995.

[2] ———, *Dynamic Programming and Optimal Control*, vol. 2, Athena Scientific, Belmont, MA, 1995.

[3] ———, *A New Value Iteration Method for the Average Cost Dynamic Programming Problem*, SIAM Journal on Control and Optimization, 36 (1998), pp. 742–759.

[4] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[5] A. Federgruen, *Successive Approximation Methods in Undiscounted Stochastic Games*, Operations Research, 28 (1980), pp. 794–809.

[6] ———, *Markovian Control Problems: Functional Equations and Algorithms*, Mathematical Centre Tract 97, Mathematisch Centrum, Amsterdam, 1983. (A repreint of A. Federgruen's 1978 doctoral dissertation, Department of Operations Research, Mathematical Centre, Amsterdam).

[7] D. Gillette, *Stochastic Games with Zero Stop Probabilities*, in Contributions to the Theory of Games III, A. W. Tucker, M. Dresher, and P. Wolfe, eds., Princeton University Press, Princeton, 1957, pp. 179–187.

[8] A. K. Hoffman and R. M. Karp, *On Nonterminating Stochastic Games*, Management Science, 12 (1966), pp. 359–370.

[9] A. Hordijk and M. Puterman, *On the Convergence of Policy Iteration in Finite State Undiscounted Markov Desicion Processes: the Unichain Case*, Mathematics of Operations Research, 12 (1969), pp. 163–176.

[10] T. M. Liggett and S. A. Lippman, *Stochastic Games with Perfect Information and Time Average Payoff*, SIAM Review, 11 (1969), pp. 604–607.

[11] S. D. Patek, *Stochastic Shortest Path Games: Theory and Algorithms*, PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, September 1997.

[12] S. D. Patek and D. P. Bertsekas, *Stochastic Shortest Path Games*, SIAM Journal on Control and Optimization, (1997). Accepted.

[13] T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[14] P. D. Rogers, *Nonzero-Sum Stochastic Games*, PhD thesis, Engineering Science, Graduate Division, Univerity of California, Berkeley, CA, June 1969. (Also referenced as: Ph.D. Dissertation Report ORC 69-8, Operations Research Center, University of California, Berkeley.).

[15] P. J. Schweitzer, *A Brouwer Fixed-Point Mapping Approach to Communicating Markov Decision Processes*, Journal of Mathematical Analysis and Applications, 123 (1987), pp. 117–130.

[16] M. J. Sobel, *Noncooperative Stochastic Games*, The Annals of Mathematical Statistics, 42 (1971), pp. 1930–1935.

[17] M. A. Stern, *On Stochastic Games with Limiting Average Payoff*, PhD thesis, University of Illinois at Chicago Circle, Chicago, June 1975.

[18] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions I*, Springer-Verlag, New York, 1970.

[19] J. van der Wal, *Stochastic Dynamic Programming*, Mathematical Centre Tracts 139, Mathematisch Centrum, Amsterdam, 1981.