

STOCHASTIC SHORTEST PATH GAMES*

STEPHEN D. PATEK[†] AND DIMITRI P. BERTSEKAS[‡]

Abstract. We consider dynamic, two-player, zero-sum games where the “minimizing” player seeks to drive an underlying finite-state dynamic system to a special terminal state along a least expected cost path. The “maximizer” seeks to interfere with the minimizer’s progress so as to maximize the expected total cost. We consider, for the first time, undiscounted finite-state problems, with compact action spaces, and transition costs that are *not* strictly positive. We admit that there are policies for the minimizer which permit the maximizer to prolong the game indefinitely. Under assumptions which generalize deterministic shortest path problems, we establish (i) the existence of a real-valued equilibrium cost vector achievable with stationary policies for the opposing players and (ii) the convergence of value iteration and policy iteration to the unique solution of Bellman’s equation.

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

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

1. Introduction. This paper develops basic theory relating to stochastic shortest path games. These are two-player, zero-sum, games where the minimizing player seeks to drive an underlying finite-state dynamic system to a special terminal state along a least expected cost path. The maximizer seeks to interfere with the minimizer’s progress so as to maximize the expected total cost. In actual play, the players implement actions simultaneously at each stage, with full knowledge of the state of the system but *without* knowledge of each other’s current decision.

Games of this type have been studied for some time. The field was initiated by Shapley in his classical paper [7]. In Shapley’s games, two players are successively faced with matrix-games (in mixed strategies) where both the immediate cost and transition probabilities to new matrix-games are influenced by the stagewise decisions of the players. In this formulation, the state of the system is the matrix-game currently being played. It is assumed that this set of states is finite and that there is a non-zero minimal probability that, at any stage, the game will transition to a terminal state, ending the sequence of rewards and payoffs. It turns out that this is equivalent to an infinite-horizon game with *discounted* additive cost. The analysis of such games is straightforward, the main results being (i) the existence and characterization of a unique real-valued equilibrium cost vector achievable in stationary randomized policies and (ii) the convergence of value iteration and policy iteration to the equilibrium cost.

Since Shapley’s work, game theorists have actively studied extensions to the discounted-cost model. In [4], Kushner and Chamberlain consider undiscounted, pursuit/evasion, stochastic games where there is a terminal state corresponding to the evader being “caught.” The state space is assumed to be finite (with $n + 1$ elements, one of which is the terminal state). Making some regularity assumptions on the transition probabilities and cost functions, they consider pure strategies over compact action spaces. In addition, they assume that either,

*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)

[‡]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, Room 35-210, Cambridge, MA 02139(dimitrib@mit.edu)

1. The n -stage probability transition matrix $[P(\mu, \nu)]^n$ (from non-terminal states to non-terminal states) is a “uniform contraction” in the stationary policy pairs (μ, ν) of the two players. (That is, for some $\epsilon > 0$, $[P(\mu, \nu)]^n$ has row-sums less than $1 - \epsilon$ for all stationary policy pairs (μ, ν) .) Or,
2. The transition costs (to the pursuer) are uniformly bounded below by $\delta > 0$ and there exists a stationary policy $\tilde{\mu}$ for the pursuer that makes $[P(\tilde{\mu}, \nu)]^n$ a uniform contraction under all stationary policies for the evader.

They show that there exists an equilibrium cost vector for the game which can be found through value iteration. In [10], van der Wal considers a special case of Kushner and Chamberlain’s games. Under more restrictive assumptions about the pursuer’s ability to catch the evader, he gives error bounds for the updates in value iteration.

In [3], Kumar and Shiau give a detailed analysis of stochastic games with very mild assumptions about the state space and control constraint sets. For the case of nonnegative additive cost (with no discounting), they establish the existence of an *extended* real equilibrium cost vector in non-Markov randomized policies (where for both players the best mixed action can depend on all of the past states and controls, as well as the current state). They show that the minimizing player can achieve the equilibrium using a stationary Markov randomized policy and that, in case the state space is finite, the maximizing player can play ϵ -optimally using stationary randomized policies.

Other researchers have studied so-called “non-terminating” stochastic games (also sometimes called “undiscounted” stochastic games), where the costs are not discounted but are averaged instead. Such *average-cost* games have a rich mathematical structure which has been extensively covered in the literature [13, 5].

In this paper, we consider undiscounted additive cost games without averaging. We admit that there are policies for the minimizer which allow the maximizer to prolong the game indefinitely *at infinite cost to the minimizer*. We do not assume nonnegativity of cost, as in [4] and [3]. We make alternative assumptions which guarantee that, at least under optimal policies, the terminal state is reached with probability one. Our results imply the results of Shapley [7], as well as those of Kushner and Chamberlain [4]. Because of our assumptions relating to termination, we are able to derive stronger conclusions than those made by Kumar and Shiau [3] for the case of a finite state space. Note that because we do not assume nonnegativity of the costs, the analysis is much more complicated than the corresponding analysis of Kushner and Chamberlain [4]. Our formal assumptions generalize (to the case of two-players) those for stochastic shortest path problems [2]. Because of this, we refer to our class of games as “stochastic shortest path games.” Our games are characterized by either (i) inevitable termination (under all policies) or (ii) an incentive for the minimizer to drive the system to termination in a finite expected number of stages. We shall see that the results of [2] are essential in developing our present theory.

In Section 2 we give a precise mathematical formulation for stochastic shortest path games. In Section 3, we relate our general formulation to Shapley’s original games [7]. We develop our main results in Section 4. This is where we show that stochastic shortest path games have an equilibrium solution which can be characterized by the unique solution to Bellman’s equation. We also prove the convergence of value iteration and policy iteration to the equilibrium cost. In Section 5, we give an example of pursuit and evasion, illustrating our main results. Finally, in the Appendix we collect some well known results about dynamic games which are crucial to our development.

2. Mathematical Formulation. Let S denote a finite state space, with elements labeled $i = 1, \dots, n$. For each $i \in S$, define $U(i)$ and $V(i)$ to be the sets of actions available to the minimizer and maximizer at state i , respectively. These are collectively referred to as *control constraint sets*. 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 $c_i(u, v)$.

We denote the sets of one-stage policies for the minimizing and maximizing players as M and N respectively, where

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

$$N = \left\{ \nu : S \mapsto \bigcup_{i \in S} V(i) \mid \nu(i) \in V(i), \quad \forall i \in S \right\}.$$

The sets of policies for the minimizing and maximizing players are denoted by \bar{M} and \bar{N} , where

$$\pi_M = \{\mu^0, \mu^1, \dots\} \in \bar{M} \iff \mu^k \in M, \quad \forall k,$$

$$\pi_N = \{\nu^0, \nu^1, \dots\} \in \bar{N} \iff \nu^k \in N, \quad \forall k.$$

Given $\mu \in M$ and $\nu \in N$, let $P(\mu, \nu)$ denote the transition probability matrix that results when μ and ν are in effect. That is

$$[P(\mu, \nu)]_{ij} = p_{ij}(\mu(i), \nu(i)).$$

Let $c(\mu, \nu)$ denote the vector whose components are $c_i(\mu(i), \nu(i))$. That is

$$[c(\mu, \nu)]_i = c_i(\mu(i), \nu(i)).$$

Given two allowable opposing policies, $\pi_M = \{\mu^0, \mu^1, \dots\} \in \bar{M}$ and $\pi_N = \{\nu^0, \nu^1, \dots\} \in \bar{N}$, we formally define the resulting cost (to the minimizer) to be

$$(2.1) \quad x(\pi_M, \pi_N) = \liminf_{t \rightarrow \infty} h_{\pi_M, \pi_N}^t,$$

where

$$(2.2) \quad h_{\pi_M, \pi_N}^t \triangleq \left\{ c(\mu^0, \nu^0) + \sum_{k=1}^t [P(\mu^0, \nu^0) \cdots P(\mu^{k-1}, \nu^{k-1})] c(\mu^k, \nu^k) \right\}.$$

Note that h_{π_M, π_N}^t can be interpreted loosely as the expected t -stage cost vector under the policies π_M and π_N .

In establishing our main results the definitions and assumptions in the following paragraphs will be helpful. We say that a policy $\pi_M = \{\mu^0, \mu^1, \dots\}$ for the minimizer is *stationary* if $\mu^k = \mu$, for all k . When this is the case and no confusion can arise, we use μ to denote the corresponding policy π_M , and we refer to π_M as *the stationary policy μ* . Similar definitions hold for stationary policies of the maximizer.

The state $1 \in S$ has special importance. We shall refer to it as the *terminal state*. This state is assumed to be absorbing and cost-free, that is $p_{11}(u, v) = 1$ and $c_1(u, v) = 0$ for all $u \in U(1)$ and $v \in V(1)$. Let $\pi_M = \{\mu^0, \mu^1, \dots\} \in \bar{M}$ and

$\pi_N = \{\nu^0, \nu^1, \dots\} \in \bar{N}$ be an arbitrary pair of policies. We say that the corresponding Markov chain *terminates with probability one* if the following limit satisfies

$$(2.3) \quad \lim_{t \rightarrow \infty} [P(\mu^0, \nu^0)P(\mu^1, \nu^1) \cdots P(\mu^t, \nu^t)]_{i1} = 1, \quad \forall i \in S.$$

(The limit above exists because the sequence (for each $i \in S$) is monotonically non-decreasing and bounded above.) We shall refer to a pair of policies (π_M, π_N) as *terminating with probability one* if the corresponding Markov chain terminates with probability one; otherwise, we refer to the pair as *prolonging*.

A stationary policy $\mu \in M$ for the minimizer is said to be *proper* if the pair (μ, π_N) is terminating with probability one for all $\pi_N \in \bar{N}$. A stationary policy μ is *improper* if it is not proper. If μ is improper then there is a policy for the maximizer $\pi_\mu \in \bar{N}$ under which there is a positive probability that the game will never end from some initial state. The designation of proper (or improper) applies only to stationary policies for the minimizer.

It is convenient to define the set $X = \{x \in R^n \mid x_1 = 0\}$. This is the space (of cost vectors) over which our main results hold. We denote by $\mathbf{0}$ the zero vector in X . Let $\mathbf{1}_X$ denote the vector $(0, 1, 1, \dots, 1)' \in X$. It is useful to define the following operators on X .

$$(2.4) \quad T_{\mu\nu}(x) = c(\mu, \nu) + P(\mu, \nu)x; \quad \mu \in M, \nu \in N.$$

$$(2.5) \quad T_\mu(x) = \sup_{\nu \in N} [c(\mu, \nu) + P(\mu, \nu)x]; \quad \mu \in M,$$

$$(2.6) \quad T(x) = \inf_{\mu \in M} \sup_{\nu \in N} [c(\mu, \nu) + P(\mu, \nu)x].$$

$$(2.7) \quad \tilde{T}_\nu(x) = \inf_{\mu \in M} [c(\mu, \nu) + P(\mu, \nu)x]; \quad \nu \in N,$$

$$(2.8) \quad \tilde{T}(x) = \sup_{\nu \in N} \inf_{\mu \in M} [c(\mu, \nu) + P(\mu, \nu)x].$$

The suprema and infima in the above are taken componentwise. We use the notation $T_{\mu\nu}^t(x)$ to denote the t -fold composition of $T_{\mu\nu}$ applied to $x \in X$. Similar definitions hold for $T_\mu^t(x)$, $T^t(x)$, $\tilde{T}_\nu^t(x)$, and $\tilde{T}^t(x)$. In the Appendix, we collect (and prove for completeness) some well-known results about these “ T ”-operators: monotonicity (Lemma A.1) and continuity (Lemma A.3).

The following are our standing assumptions.

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 $\lim_{t \rightarrow \infty} [h_{\pi_M, \pi_N}^t]_i = \infty$.*

Assumption R (Regularity) *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}_ν are always achieved by elements of N and M , respectively. That is, for every $x \in X$, there exists $\nu \in N$ such that $T_\mu(x) = T_{\mu\nu}(x) \in X$. Similarly, for every $x \in X$, there exists $\mu \in M$ such that $\tilde{T}_\nu(x) = T_{\mu\nu}(x) \in X$.)

3. For all $x \in X$, 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 $x \in X$, there exists $\mu \in M$ and $\nu \in N$ such that $T(x) = T_\mu(x) \in X$ and $\tilde{T}(x) = \tilde{T}_\nu(x) \in X$.
4. For each $x \in X$, we have $T(x) = \tilde{T}(x)$.

Note that part 4 of Assumption 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 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 [8], p.232 or [6], p. 397). We will show in Section 3 that dynamic games with “mixed” strategies over finite underlying action spaces satisfy this assumption.

To verify that a stationary policy $\mu \in M$ is proper, we need only check whether (μ, ν) is terminating with probability one for all *stationary* policies $\nu \in N$ for the maximizer. Furthermore, if $\mu \in M$ is improper, then we can always find a *stationary* policy $\nu \in N$ for the maximizer which is prolonging when paired with μ . This is shown in the following lemma:

LEMMA 2.1. *If $\mu \in M$ is such that (μ, ν) terminates with probability one for all $\nu \in N$, then μ is proper.*

Proof. The proof uses the analysis of [2]. Let $\mu \in M$ be a fixed policy for the minimizer, and suppose that the pair (μ, ν) is terminating with probability one for all stationary policies of the maximizer $\nu \in N$. With μ fixed, the maximizer is faced with a stochastic shortest path problem of the type considered in [2]. (The maximizer has no improper policies (against μ).) Now modify the problem such that the costs of transitioning from nonterminal states are all set to one but all of the transition probabilities are left unchanged. The assumptions of [2] remain satisfied, so the optimal expected cost for the maximizer in the new problem is bounded, even over nonstationary policies. Thus, the maximum expected number of stages to termination under μ is finite. This is true for both the modified problem and the original version of the game. This implies that μ is proper. **Q.E.D.**

One of the objectives of this paper is to show that under Assumptions SSP and R there exist policies $\pi_M^* \in \bar{M}$ and $\pi_N^* \in \bar{N}$ such that

$$\begin{aligned} x(\pi_M^*, \pi_N^*) &\geq x(\pi_M^*, \pi_N^*), & \forall \pi_M \in \bar{M}, \\ x(\pi_M^*, \pi_N^*) &\leq x(\pi_M^*, \pi_N^*), & \forall \pi_N \in \bar{N}. \end{aligned}$$

Such a cost vector $x^* \triangleq x(\pi_M^*, \pi_N^*)$ is called the *equilibrium cost vector* (or *value*) of the stochastic shortest path game. The policies π_M^* and π_N^* form an *equilibrium*

solution. Since this is a zero-sum game, we know that the equilibrium cost (if it exists) is unique. Another objective of this paper is to show that the equilibrium cost vector is characterized as the unique solution to Bellman's equation, with *stationary* equilibrium policies for the opposing players. After these results are established, we proceed to show that value iteration and policy iteration converge to the unique solution of Bellman's equation.

3. Connection to Shapley's Stochastic Games. The mathematical formulation of the preceding section includes as a special case the stochastic games of Shapley. To see this, assume that the number of actions available to either player at any state is finite. As before, the players implement underlying actions simultaneously at each stage, with full knowledge of the state of the system but *without* knowledge of each other's current decision. However, the players are now allowed to randomize their decisions in formulating a policy so as to keep their opponents from adapting to a deterministic policy. That is, in considering what to do at each state, the players choose *probability distributions* over underlying control sets rather than specific underlying control actions. In other words, the players use randomized or "mixed" policies.

For each $i \in S$, define $A(i)$ and $B(i)$ to be the *finite* sets of underlying actions to the minimizer and maximizer, respectively. These are the physical controls the players may ultimately implement at state i . Let $|A(i)|$ and $|B(i)|$ denote the numbers of elements in each set of actions. We define the players' "control constraint sets" for the game as

$$(3.1) \quad U(i) = \left\{ u \in R^{|A(i)|} \mid \sum_{j \in A(i)} u_j = 1; \quad u_j \geq 0 \right\},$$

$$(3.2) \quad V(i) = \left\{ v \in R^{|B(i)|} \mid \sum_{j \in B(i)} v_j = 1; \quad v_j \geq 0 \right\}.$$

Thus, $U(i)$ is the set of probability distributions over control actions $A(i)$ available to the minimizer from state $i \in S$. Similarly, $V(i)$ is the set of probability distributions over underlying control actions $B(i)$ available to the maximizer from state $i \in S$. Here the functions $p_{ij}(u, v)$ and $c_i(u, v)$ are respectively of the form

$$(3.3) \quad p_{ij}(u, v) = \sum_{k \in A(i)} \sum_{l \in B(i)} \underline{p}_{ij}(k, l) u_k v_l,$$

$$(3.4) \quad c_i(u, v) = \sum_{j \in S} \sum_{k \in A(i)} \sum_{l \in B(i)} \underline{g}_{ij}(k, l) \underline{p}_{ij}(k, l) u_k v_l,$$

where the functions \underline{p}_{ij} and \underline{g}_{ij} denote the transition probabilities and costs of the underlying two-player Markov Decision Process. Since the sets $U(i)$ and $V(i)$ are polyhedral and the functions $c_i(u, v)$ and $p_{ij}(u, v)$ are bilinear for all i and j (and continuous) as functions of $(u, v) \in U(i) \times V(i)$, it is clear that Assumption R is satisfied. (Parts 3 and 4 are satisfied thanks to the Minimax Theorem of von Neumann.)

4. Main Results. We now develop our main results; namely, the existence and characterization of a unique equilibrium cost vector, the convergence of value iteration, and the convergence of policy iteration. In Sections 4.1 and 4.2, we characterize optimal solutions for the maximizer and minimizer, respectively, for the case where the opposing player fixes a policy. After we lay this groundwork, we consider the game proper in Section 4.3.

4.1. The Case Where the Minimizer's Policy is Fixed. Consider the policy $\pi_M = \{\mu^0, \mu^1, \dots\} \in \bar{M}$. The cost of π_M is defined by

$$(4.1) \quad x(\pi_M) = \liminf_{t \rightarrow \infty} \max_{\pi_N \in \bar{N}} h_{\pi_M, \pi_N}^t.$$

The Appendix shows that, with our assumptions on c and P , the maximum in (4.1) is attained for every t (see Lemma A.5). The cost of a stationary policy μ for the minimizer is denoted $x(\mu)$ and is computed according to equation (4.1) where $\pi_M = \{\mu, \mu, \dots\}$.

Given a vector $w \in R^n$ whose elements are positive, the corresponding weighted sup-norm, denoted $\|\cdot\|_\infty^w$, is defined by

$$\|x\|_\infty^w = \max_{i=1, \dots, n} x_i/w_i, \quad \forall x \in R^n.$$

The next lemma follows from the theory of one-player stochastic shortest paths.

LEMMA 4.1. *Assume that all stationary policies for the minimizer are proper. The operator T is a contraction mapping on the set $X = \{x \in R^n \mid x_1 = 0\}$ with respect to a weighted sup-norm. Moreover, if $\mu \in M$ is proper, then T_μ is a contraction mapping with respect to a weighted sup-norm.*

Proof. We will show first the result about T for the case that all stationary policies are proper. Our strategy is to identify a vector of weights w and to show that this set of weights is one for which T is a contraction with respect to $\|\cdot\|_\infty^w$.

Let us define a new one-player stochastic shortest path problem of the type considered in [2]. This problem is defined such that the transition probabilities remain unchanged and the transition costs are all set equal to -1 for all states other than the terminal state. The important difference is that the maximizer and minimizer “work together” in the sense that the decision space (for the single player of the new problem) is over $\bar{M} \times \bar{N}$. This is a stochastic shortest path problem where all stationary policies are proper. Using the results of [2], there is an optimal cost vector $\tilde{x} \in X$ which can be achieved using a stationary policy $(\tilde{\mu}, \tilde{\nu}) \in \bar{M} \times \bar{N}$. Note that, since the transition costs from all non-terminal states are set to -1 in the new stochastic shortest path problem, we have $\tilde{x}_i \leq -1$ for all $i \neq 1$. Moreover, from Bellman's equation we have

$$\tilde{x} = -\mathbf{1}_X + P(\tilde{\mu}, \tilde{\nu})\tilde{x},$$

where $\mathbf{1}_X = (0, 1, 1, \dots, 1)' \in X$. Also, for all $\mu \in M$ and $\nu \in N$

$$\tilde{x} \leq -\mathbf{1}_X + P(\mu, \nu)\tilde{x}.$$

Thus, for all $\mu \in M$, $\nu \in N$, and for all $i \neq 1$

$$(4.2) \quad \sum_{j=2}^n p_{ij}(\mu(i), \nu(i)) \cdot (-\tilde{x}_j) \leq -\tilde{x}_i - 1 \leq -\tilde{x}_i \gamma,$$

where $\gamma = \max_{i \neq 1} (\tilde{x}_i + 1)/\tilde{x}_i$. Since the $\tilde{x}_i \leq -1$ for all $i \neq 1$, we have that $\gamma \in [0, 1)$. Now define $w = -\tilde{x} + (1, 0, 0, \dots, 0)'$. Note that w is a vector in R^n whose elements are all strictly positive.

Let us now resume consideration of the original stochastic shortest path game. Let x and \bar{x} be any two elements of X such that $\|x - \bar{x}\|_\infty^w = c$. Let $\mu \in M$ be such that $T_\mu(x) = T(x)$, and let $\nu \in N$ be such that $T_\mu(\bar{x}) = T_{\mu\nu}(\bar{x})$. Then,

$$\begin{aligned} T(\bar{x}) - T(x) &= T(\bar{x}) - T_\mu(x) \\ &\leq T_\mu(\bar{x}) - T_\mu(x) \\ &= T_{\mu\nu}(\bar{x}) - T_\mu(x) \\ &\leq T_{\mu\nu}(\bar{x}) - T_{\mu\nu}(x). \end{aligned}$$

Thus,

$$[T(\bar{x})]_i - [T(x)]_i \leq \sum_{j=2}^n [P(\mu, \nu)]_{ij} (\bar{x}_j - x_j).$$

Using this, we see that for all $i \neq 1$

$$\begin{aligned} \frac{[T(\bar{x}) - T(x)]_i}{cw_i} &\leq \frac{1}{cw_i} \sum_{j=2}^n p_{ij}(\mu(i), \nu(i)) (\bar{x}_j - x_j) \\ &\leq \frac{1}{w_i} \sum_{j=2}^n p_{ij}(\mu(i), \nu(i)) w_j \\ &= \frac{1}{-\tilde{x}_i} \sum_{j=2}^n p_{ij}(\mu(i), \nu(i)) (-\tilde{x}_j) \\ &\leq \frac{1}{-\tilde{x}_i} (-\tilde{x}_i) \gamma = \gamma, \end{aligned}$$

where the last inequality follows from (4.2). Thus, we get

$$\frac{[T(\bar{x})]_i - [T(x)]_i}{w_i} \leq c\gamma, \quad \forall i \neq 1.$$

Since $[T(\bar{x})]_1 - [T(x)]_1 = 0$,

$$\frac{[T(\bar{x})]_i - [T(x)]_i}{w_i} \leq c\gamma, \quad \forall i.$$

Using similar arguments, we may show that,

$$\frac{[T(x)]_i - [T(\bar{x})]_i}{w_i} \leq c\gamma, \quad \forall i.$$

Combining the preceding inequalities, we see that $\|T(x) - T(\bar{x})\|_\infty^w \leq c\gamma$. Since $0 \leq \gamma < 1$, we have that T is a contraction over X with respect to $\|\cdot\|_\infty^w$.

Now suppose $\mu \in M$ is proper. By viewing T_μ as the “ T ”-operator in a new game where $U(i) \equiv \{\mu(i)\}$, we have the desired result. **Q.E.D.**

LEMMA 4.2. *Given a proper policy μ , the following are true.*

1. *The cost $x(\mu)$ of μ is the unique fixed point of T_μ in $X = \{x \in R^n \mid x_1 = 0\}$.*
2. *$x(\mu) = \sup_{\pi_N \in \bar{N}} x(\mu, \pi_N)$.*
3. *We have $T_\mu^t(x) \rightarrow x(\mu)$ for all $x \in X$, with linear convergence.*

Proof. An induction argument (cf. Appendix Lemma A.5) easily shows that

$$T_\mu^{t+1}(\mathbf{0}) = \max_{\{\nu^0, \dots, \nu^t\}} \left\{ c(\mu, \nu^0) + \sum_{k=1}^t [P(\mu, \nu^0)P(\mu, \nu^1) \cdots P(\mu, \nu^{k-1})]c(\mu, \nu^k) \right\},$$

where $\mathbf{0}$ is the zero vector in X . Thus, using preceding lemma and the definition of $x(\mu)$, we have

$$x(\mu) = \lim_{t \rightarrow \infty} T_\mu^{t+1}(\mathbf{0}) = \tilde{x}_\mu,$$

where \tilde{x}_μ is the unique fixed point of the contraction mapping T_μ within X , proving statement 1.

Consider the following infinite-horizon stochastic shortest path problem for the maximizer:

$$\sup_{\pi_N \in \bar{N}} \liminf_{t \rightarrow \infty} \left\{ c(\mu, \nu^0) + \sum_{k=1}^t [P(\mu, \nu^0) \cdots P(\mu, \nu^{k-1})]c(\mu, \nu^k) \right\}.$$

This problem is covered by the theory developed in [2] since the fact that μ is proper implies that termination is inevitable under all policies in the maximizer's problem. The optimal cost of this problem is $\sup_{\pi_N \in \bar{N}} x(\mu, \pi_N)$, and according to the theory of [2], it is equal to the limit of the successive approximation method applied to this problem, which is $\lim_{t \rightarrow \infty} T_\mu^{t+1}(\mathbf{0})$ and is also the unique fixed point of T_μ within X . This proves statement 2.

Finally, the linear convergence of $T_\mu^{t+1}(\mathbf{0})$ follows from the contraction property of T_μ . **Q.E.D.**

LEMMA 4.3. *If $x \geq T_\mu(x)$ for some $x \in X$, then μ is proper.*

Proof. To reach a contradiction, suppose μ is improper. According to Assumption SSP and Lemma 2.1, there exists a stationary maximizer's policy $\bar{\nu} \in N$ such that $(\mu, \bar{\nu})$ is prolonging and results in unbounded expected cost from some initial state when played against μ .

Let x be an element in X such that $x \geq T_\mu(x)$. Then, applying T_μ to x , we have that

$$x \geq T_\mu(x) \geq c(\mu, \bar{\nu}) + P(\mu, \bar{\nu})x,$$

where the second inequality follows from the definition of T_μ . From the monotonicity of T_μ , we get

$$\begin{aligned} x \geq T_\mu(x) &\geq T_\mu^2(x) \geq T_\mu(c(\mu, \bar{\nu}) + P(\mu, \bar{\nu})x) \\ &\geq P(\mu, \bar{\nu})P(\mu, \bar{\nu})x + [c(\mu, \bar{\nu}) + P(\mu, \bar{\nu})c(\mu, \bar{\nu})], \end{aligned}$$

where the last inequality follows again from the definition of T_μ . Proceeding inductively, using the same steps, we have that for all t

$$x \geq T_\mu^t(x) \geq P(\mu, \bar{\nu})^{t+1}x + \sum_{k=0}^t P(\mu, \bar{\nu})^k c(\mu, \bar{\nu}).$$

On the other hand, because the policy $\bar{\nu}$ results in infinite expected cost (from some initial state) when played against μ , some subsequence of $\sum_{k=0}^t [P(\mu, \bar{\nu})]^k c(\mu, \bar{\nu})$ must have a coordinate that tends to infinity. (The term involving x remains bounded because it is just x multiplied by the product of stochastic matrices.) This contradicts the above inequality. Thus, μ must be proper. **Q.E.D.**

4.2. The Case Where the Maximizer's Policy is Fixed. By Assumption SSP there exists a proper policy for the minimizer. Thus, it is impossible that a single policy for the maximizer prolongs the game for *all* policies of the minimizer. Let us define $\tilde{x}(\pi_N)$ to be *the cost of the policy* $\pi_N \in \bar{N}$:

$$(4.3) \quad \tilde{x}(\pi_N) = \liminf_{t \rightarrow \infty} \min_{\pi_M \in \bar{M}} h_{\pi_M, \pi_N}^t,$$

where $\pi_N = \{\nu^0, \nu^1, \dots\}$. The cost of a stationary policy ν for the minimizer is denoted $\tilde{x}(\nu)$ and is computed according to equation (4.3) where $\pi_N = \{\nu, \nu, \dots\}$.

LEMMA 4.4. *For any $\nu \in N$, the following are true.*

1. *The cost $\tilde{x}(\nu)$ of ν is the unique fixed point of \tilde{T}_ν in $X = \{x \in R^n \mid x_1 = 0\}$.*
2. *$\tilde{x}(\nu) = \inf_{\pi_M \in \bar{M}} x(\pi_M, \nu)$.*
3. *We have $\tilde{T}_\nu^t(x) \rightarrow \tilde{x}(\nu)$ for all $x \in X$. If for all $\mu \in M$, the pair (μ, ν) terminates with probability one, then the convergence is linear.*

Proof. This follows directly from the theory of (one-player) stochastic shortest path problems. **Q.E.D.**

4.3. Main Results for the Game. We now establish the main results of the paper: the existence and characterization of a unique equilibrium solution, the convergence of value iteration, and the convergence of policy iteration.

PROPOSITION 4.5. *The operator T has a unique fixed point x^* on X .*

Proof. We begin by showing that T has at most one fixed point in X . Suppose x and x' are both fixed points of T in X . We can select $\mu \in M$ and $\mu' \in M$ such that $x = T(x) = T_\mu(x)$ and $x' = T(x') = T_{\mu'}(x')$. By Lemma 4.3, we have that μ and μ' are proper. Lemma 4.2 implies that $x = x(\mu)$ and $x' = x(\mu')$. Since μ' isn't necessarily optimal with respect to x in applying the T operator, we have from the monotonicity of T that $x = T^t(x) \leq T_{\mu'}^t(x)$ for all $t > 0$. Thus, by Lemma 4.2, we have that $x \leq \lim_{t \rightarrow \infty} T_{\mu'}^t(x) = x(\mu') = x'$. Similarly, $x' \leq x$, which implies that $x = x'$ and that T has at most one fixed point in X .

To establish the existence of a fixed point, fix a proper policy $\mu \in M$ for the minimizer. (One exists thanks to Assumption SSP.) By Lemma 4.2, we have that $x(\mu) = T_\mu(x(\mu))$. Thus, $x(\mu) \geq T(x(\mu))$. Similarly, by fixing a stationary policy $\nu \in N$ for the maximizer, we obtain from Lemma 4.4 that $\tilde{x}(\nu) = \tilde{T}_\nu(\tilde{x}(\nu))$. Thus, $\tilde{x}(\nu) \leq \tilde{T}(\tilde{x}(\nu)) = T(\tilde{x}(\nu))$. We now claim that $\tilde{x}(\nu) \leq x(\mu)$. To see this, note that, for every $\pi_M \in \bar{M}$, $\pi_N \in \bar{N}$, and $t > 0$,

$$h_{\pi_M, \pi_N}^t \leq \max_{\bar{\pi}_N \in \bar{N}} h_{\pi_M, \bar{\pi}_N}^t,$$

and

$$h_{\pi_M, \pi_N}^t \geq \min_{\bar{\pi}_M \in \bar{M}} h_{\bar{\pi}_M, \pi_N}^t,$$

where we have used the notation defined in (2.2). Thus, for any $\pi_N \in \bar{N}$ and for any $\pi_M \in \bar{M}$

$$\min_{\bar{\pi}_M \in \bar{M}} h_{\bar{\pi}_M, \pi_N}^t \leq \max_{\bar{\pi}_N \in \bar{N}} h_{\pi_M, \bar{\pi}_N}^t.$$

By taking the limit inferior of both sides with respect to t , we see that $\tilde{x}(\pi_N) \leq x(\pi_M)$ for all $\pi_N \in \bar{N}$ and $\pi_M \in \bar{M}$. In particular, $\tilde{x}(\nu) \leq x(\mu)$.

Using the monotonicity of T we have that

$$\tilde{x}(\nu) \leq T(\tilde{x}(\nu)) \leq T(x(\mu)) \leq x(\mu).$$

Again from the monotonicity of T , we obtain for all $t > 1$ that

$$\tilde{x}(\nu) \leq T^{t-1}(\tilde{x}(\nu)) \leq T^t(\tilde{x}(\nu)) \leq x(\mu).$$

Thus, the sequence $\{T^t(\tilde{x}(\nu))\}$ converges to a vector $x^\infty \in X$. From the continuity of T , we have that $x^\infty = T(x^\infty)$. Thus, T has a fixed point in X . **Q.E.D.**

PROPOSITION 4.6. *The unique fixed point $x^* = T(x^*)$ 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 $x \in X$, $\mu \in M$, and $\nu \in N$ are such that $x = T(x) = T_\mu(x) = \tilde{T}_\nu(x)$, then*

1. $x = x(\mu, \nu)$,
2. $x(\pi_M, \nu) \geq x(\mu, \nu), \quad \forall \pi_M \in \bar{M}$,
3. $x(\mu, \pi_N) \leq x(\mu, \nu), \quad \forall \pi_N \in \bar{N}$.

Proof. That there exists a unique fixed point $x^* = T(x^*)$ follows from the preceding proposition. Let $\mu^* \in M$ be such that $x^* = T(x^*) = T_{\mu^*}(x^*)$, and let $\nu^* \in N$ be such that $x^* = T(x^*) = \tilde{T}(x^*) = \tilde{T}_{\nu^*}(x^*)$. (Such policies exist thanks to Assumption R.) By Lemma 4.3, we have that μ^* is proper. Thus, by Lemma 4.2, we have that $x^* = x(\mu^*) = \sup_{\pi_N \in \bar{N}} x(\mu^*, \pi_N)$. Similarly, by Lemma 4.4, we have that $x^* = \tilde{x}(\nu^*) = \inf_{\pi_M \in \bar{M}} x(\pi_M, \nu^*)$. Combining these results we obtain

$$\inf_{\pi_M \in \bar{M}} \sup_{\pi_N \in \bar{N}} x(\pi_M, \pi_N) \leq x^* \leq \sup_{\pi_N \in \bar{N}} \inf_{\pi_M \in \bar{M}} x(\pi_M, \pi_N).$$

Since in general we have

$$\inf_{\pi_M \in \bar{M}} \sup_{\pi_N \in \bar{N}} x(\pi_M, \pi_N) \geq \sup_{\pi_N \in \bar{N}} \inf_{\pi_M \in \bar{M}} x(\pi_M, \pi_N)$$

(a statement of the Minimax Inequality), we obtain the desired result:

$$\inf_{\pi_M \in \bar{M}} \sup_{\pi_N \in \bar{N}} x(\pi_M, \pi_N) = x^* = \sup_{\pi_N \in \bar{N}} \inf_{\pi_M \in \bar{M}} x(\pi_M, \pi_N)$$

Q.E.D.

Lemma 4.1 implies that, when all stationary policies for the minimizer are proper, the iteration $x^{t+1} = T(x^t)$ converges linearly to the equilibrium cost x^* for all $\mathbf{0} \in X$. This follows from the contraction mapping principle. In the following proposition, we extend this result to the case where not all stationary policies for the minimizer are proper.

PROPOSITION 4.7. *For every $x \in X$, there holds,*

$$(4.4) \quad \lim_{t \rightarrow \infty} T^t(x) = x^*,$$

where x^* is the unique equilibrium cost vector.

Proof. The uniqueness and existence of a fixed point for T was established in Proposition 4.5. Let x^* be the unique fixed point, and let $\mu^* \in M$ (proper) be such that $T(x^*) = T_{\mu^*}(x^*)$. Our objective is to show that $T^t(x) \rightarrow x^*$ for all $x \in X$. Let Δ be the vector with coordinates,

$$(4.5) \quad \Delta_i = \begin{cases} 0, & \text{if } i = 1 \\ \delta, & \text{if } i \neq 1 \end{cases},$$

where δ is some scalar. Let x^Δ be the unique vector in X satisfying $T_{\mu^*}(x^\Delta) = x^\Delta - \Delta$. (Such a vector exists because μ^* is proper, making the operator $T_{\mu^*}(\cdot) + \Delta$ a contraction.) Note that

$$\begin{aligned} x^\Delta &= T_{\mu^*}(x^\Delta) + \Delta \\ &= \max_{\nu \in N} [c(\mu^*, \nu) + P(\mu^*, \nu)x^\Delta] + \Delta \\ &= \max_{\nu \in N} [c(\mu^*, \nu) + \Delta + P(\mu^*, \nu)x^\Delta]. \end{aligned}$$

Thus, x^Δ is the minimax cost of the fixed policy μ^* with the immediate transition cost $c(\mu^*, \cdot)$ replaced with $c(\mu^*, \cdot) + \Delta$. We have that

$$x^\Delta = T_{\mu^*}(x^\Delta) + \Delta \geq T_{\mu^*}(x^\Delta).$$

Thus, from the monotonicity of T_{μ^*} we have that for all $t > 0$

$$T_{\mu^*}^t(x^\Delta) \leq x^\Delta.$$

By taking the limit as $t \rightarrow \infty$, we see that $x(\mu^*) \leq x^\Delta$. (This is also implied by our interpretation of x^Δ above.)

Now using the monotonicity of T and the fact that $x^* = x(\mu^*)$, we get

$$x^* = T(x^*) \leq T(x^\Delta) \leq T_{\mu^*}(x^\Delta) = x^\Delta - \Delta \leq x^\Delta,$$

Proceeding inductively, we get

$$x^* \leq T^t(x^\Delta) \leq T^{t-1}(x^\Delta) \leq x^\Delta.$$

Hence, $\{T^t(x^\Delta)\}$ is a monotonically decreasing sequence bounded below which converges to some $\tilde{x} \in X$. By continuity of the operator T , we must have that $\tilde{x} = T(\tilde{x})$. By the uniqueness of the fixed point of T , we have that $\tilde{x} = x^*$.

We now examine the convergence of the operator T^t applied to $x^* - \Delta$. Note that,

$$x^* - \Delta = T(x^*) - \Delta \leq T(x^* - \Delta) \leq T(x^*) = x^*,$$

where the first inequality follows from the fact that $P(\mu, \nu)\Delta \leq \Delta$ for all $\mu \in M$ and $\nu \in N$. Once again monotonicity of T prevails, implying that $T^t(x^* - \Delta)$ is monotonically increasing and bounded above. From the continuity of T we have that $\lim_{t \rightarrow \infty} T^t(x^* - \Delta) = x^*$.

We saw earlier that $x^\Delta = T_{\mu^*}(x^\Delta) + \Delta$ and that $x^\Delta \geq x^*$. Then, from the monotonicity of T_{μ^*}

$$x^\Delta \geq T_{\mu^*}(x^*) + \Delta = x^* + \Delta.$$

Thus, for any $x \in X$ we can find $\delta > 0$ such that $x^* - \Delta \leq x \leq x^\Delta$. By the monotonicity of T , we then have,

$$T^t(x^* - \Delta) \leq T^t(x) \leq T^t(x^\Delta), \quad \forall t \geq 1.$$

Taking limits, we see that $\lim_{t \rightarrow \infty} T^t(x) = x^*$. **Q.E.D.**

PROPOSITION 4.8. *Given a proper stationary policy $\mu^0 \in M$, we have that*

$$x(\mu^k) \rightarrow x^*,$$

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

Proof. Choose $\mu^1 \in M$ such that $T_{\mu^1}(x(\mu^0)) = T(x(\mu^0))$. (Assumption SSP implies that such an initial proper policy μ^0 exists.) We have $T_{\mu^1}(x(\mu^0)) = T(x(\mu^0)) \leq T_{\mu^0}(x(\mu^0)) = x(\mu^0)$. By Lemma 4.3, μ^1 is proper. By the monotonicity of T_{μ^1} and Lemma 4.2, we have that for all t

$$x(\mu^0) \geq T(x(\mu^0)) \geq T_{\mu^1}^{t-1}(x(\mu^0)) \geq T_{\mu^1}^t(x(\mu^0)).$$

Thus,

$$x(\mu^0) \geq T(x(\mu^0)) \geq \lim_{t \rightarrow \infty} T_{\mu^1}^t(x(\mu^0)) = x(\mu^1).$$

Applying this argument iteratively, we construct a sequence $\{\mu^k\}$ of proper policies such that,

$$(4.6) \quad x(\mu^k) \geq T(x(\mu^k)) \geq x(\mu^{k+1}) \geq x^*, \quad \forall k = 0, 1, \dots$$

Since $\{x(\mu^k)\}$ is monotonically decreasing and bounded below by x^* , we have that the entire sequence converges to some vector x^∞ . From (4.6) and the continuity of T , we have that $x^\infty = T(x^\infty)$. Since x^* is the unique fixed point of T on X , we have that $x(\mu^k) \rightarrow x^*$. **Q.E.D.**

5. An Example of Pursuit and Evasion. Consider the following two-player game, played around a table with four corners. One player, the pursuer (who is actually the minimizer), is attempting to “catch” in minimum time the other player, the evader (who is the maximizer). The game evolves in stages where, in each stage, both players implement actions simultaneously. When the players are across from one another, they each decide (independently) whether to stay where they are, move one corner clockwise, or move one corner counter-clockwise. When the two players are adjacent to one another, they each decide (independently) whether to stay where they are, move toward the other’s current location, or move away from the other’s current location. The pursuer catches the evader only by arranging to land on the same corner of the table as the evader. (The possibility exists that, when they are adjacent, they can both move toward each other’s current location. This does not result in the evader being caught “in mid-air”.) The evader is slower than the pursuer in the sense that, when the evader decides to change location, he succeeds in doing so only with probability $p \in (0, 1)$. (With probability $1 - p$, the evader will wind up not moving at all.) Thus, the pursuer can ultimately catch the evader, provided he implements an appropriate policy (such as “always move toward the present location of the evader”). On the other hand, there exist policies for the pursuer (such as “always stay put”) which allow the maximizer to prolong the game indefinitely. This results in infinite cost (i.e. infinite capture time) to the pursuer.

This game fits into our framework for stochastic shortest path games. As described above there are three states: evader caught (state 1), players adjacent to one another (state 2), and players across from one another (state 3). Thus, $S = \{1, 2, 3\}$. Once the evader is caught, the game is over, so state 1 serves as the terminal state, which is zero cost and absorbing.

In state two, when the players are adjacent, the players may move toward the other’s location (action 1), stay where they are (action 2), or move away from the other’s location (action 3). Thus, $A(2) = B(2) = \{1, 2, 3\}$. From the description of the problem given above, it is not hard to see that

$$p_{21}(u, v) = u_1[(v_1 + v_3)(1 - p) + v_2] + u_2v_1p,$$

$$\begin{aligned} p_{22}(u, v) &= (u_1 + u_3)(v_1 + v_3)p + u_2[(v_1 + v_3)(1 - p) + v_2], \\ p_{23}(u, v) &= u_2v_3p + u_3[(v_1 + v_3)(1 - p) + v_2]. \end{aligned}$$

The expected transition cost functions $c_2(u, v)$ take on the value of one for all $u \in U(2)$ and $v \in V(2)$.

In state three (when the players are on opposite corners of the table), the players may move clockwise toward the other's current location (action 1), stay where they are (action 2), or move counter-clockwise toward the other's location (action 3). Thus, $A(3) = B(3) = \{1, 2, 3\}$. It is not hard to see that

$$\begin{aligned} p_{31}(u, v) &= u_1v_3p + u_3v_1p, \\ p_{32}(u, v) &= (u_1 + u_3)[(v_1 + v_3)(1 - p) + v_2] + u_2(v_1 + v_3)p, \\ p_{33}(u, v) &= u_1v_1p + u_2[(v_1 + v_3)(1 - p) + v_2] + u_3v_3p. \end{aligned}$$

The expected transition cost functions $c_3(u, v)$ take on the value of one for all $u \in U(3)$ and $v \in V(3)$.

We will show that the equilibrium value of this stochastic shortest path game is given by

$$x^* = \left(0, \frac{1}{1-p}, \frac{2-p}{1-p}\right)',$$

and that equilibrium randomized strategies for the two players are given by $\mu^* \in M$ and $\nu^* \in N$ such that

$$\begin{aligned} \mu^*(2) &= (1, 0, 0)' \\ \nu^*(2) &= (v_1, 0, v_3)' \\ \mu^*(3) &= (u_1, 0, u_3)' \\ \nu^*(3) &= (0, 1, 0)', \end{aligned}$$

where v_1, v_3, u_1 , and u_3 are nonnegative, and $v_1 + v_3 = 1$ and $u_1 + u_3 = 1$. Thus, any probability vector $v \in V(2)$ such that $v_2 = 0$ forms an equilibrium strategy for the evader. In other words, as long as the evader chooses not to remain at his current location (when the two players are adjacent), any mixed decision (at state 2) for the evader is optimal. The pursuer does not have the same flexibility; his optimal mixed decision is deterministic: always move toward the evader. On the other hand, any probability vector $u \in U(3)$ such that $u_2 = 0$ forms an equilibrium strategy for the pursuer. In other words, as long as the pursuer decides to not stay at his current location (when the two players are across from one another), any mixed decision for the pursuer (at state 3) is optimal. This time, it is the evader's strategy which is inflexible. His optimal action is to stay at his current location. Thus, when both players play optimally, the game will always transition from state $i = 3$ to $i = 2$ in one stage. Happily, the equilibrium cost reflects this: $x_3^* = \frac{2-p}{1-p} = 1 + x_2^*$.

To verify that these are indeed equilibrium policies, we will show that $x^* = T(x^*) = T_{\mu^*}(x^*) = \tilde{T}_{\nu^*}(x^*)$. (Notice that the policy μ^* corresponds to one where the pursuer always decides to move in the direction of the current location of the evader. This policy is clearly proper. The desired result follows from Corollary 4.6.)

Let us first consider the case where the two players are adjacent (i.e. state 2). Let a general cost-to-go vector be given as $x = (0, x_2, x_3)' \in X$. (Shortly, we shall

consider the case where $x = x^*$, as suggested above.) To evaluate the second element of $T(x)$, we must compute

$$\min_{u \in U(2)} \max_{v \in V(2)} u' G_2(x) v,$$

where the matrix $G_2(x)$ is computed as

$$G_2(x) = \begin{bmatrix} 1 + px_2 & 1 & 1 + px_2 \\ 1 + (1-p)x_2 & 1 + x_2 & 1 + (1-p)x_2 + px_3 \\ 1 + px_2 + (1-p)x_3 & 1 + x_3 & 1 + px_2 + (1-p)x_3 \end{bmatrix}.$$

In other words, the second element of $T(x)$ is evaluated as the value of the matrix game (in mixed strategies) defined by $G_2(x)$. It is well known that the equilibrium cost and equilibrium strategies for a matrix game are characterized as the optimal value and solutions to a particular linear program and it's dual [12]. In particular,

$$\begin{aligned} \frac{1}{[T(x)]_2} &= \min e' \check{v} \\ &\text{subject to } G_2(x) \check{v} \geq e, \check{v} \geq 0, \\ \frac{v^*}{[T(x)]_2} &\in \arg \min e' \check{v} \\ &\text{subject to } G_2(x) \check{v} \geq e, \check{v} \geq 0, \end{aligned}$$

where e is the vector of all ones in R^3 , and v^* is an equilibrium strategy for the maximizer in the matrix-game. We shall refer to the linear program above as the “primal” problem. The dual of the primal problem characterizes equilibrium strategies u^* for the minimizer of the matrix game:

$$\frac{u^*}{[T(x)]_2} \in \arg \max e' \check{u} \\ \text{subject to } G_2(x)' \check{u} \leq e, \check{u} \geq 0.$$

Now consider $G_2(x^*)$ and define

$$\begin{aligned} \check{u}^* &= \mu^*(2)/x_2^* = (1-p)(1, 0, 0)', \\ \check{v}^* &= \nu^*(2)/x_2^* = (1-p)(v_1, 0, v_3)'. \end{aligned}$$

It is straightforward to verify that \check{v}^* is feasible for the primal linear program and that \check{u}^* is feasible for the dual problem. Moreover, the primal cost corresponding to \check{v}^* is exactly $1-p$, just as the dual value of \check{u}^* is also exactly $1-p$. Thus, we have found a primal/dual feasible pair for which the primal cost equals the dual value. Then, according to the duality theorem of linear programming, \check{v}^* and \check{u}^* are primal/dual optimal, and the optimal values of the primal and dual problems equal $1-p$ which is exactly $\frac{1}{x_2^*}$. This verifies that $x_2^* = [T(x^*)]_2$ and that $\mu^*(2)$ and $\nu^*(2)$ form an equilibrium pair of mixed decisions at state $2 \in S$.

Let us now consider the case where the two players are across from one another (i.e. state 3). To evaluate the third element of $T(x)$ for general $x \in X$, we must compute

$$\min_{u \in U(3)} \max_{v \in V(3)} u' G_3(x) v,$$

where $G_3(x)$ is a matrix computed as

$$G_3(x) = \begin{bmatrix} 1 + (1-p)x_2 + px_3 & 1 + x_2 & 1 + (1-p)x_2 \\ 1 + px_2 + (1-p)x_3 & 1 + x_3 & 1 + px_2 + (1-p)x_3 \\ 1 + (1-p)x_2 & 1 + x_2 & 1 + (1-p)x_2 + px_3 \end{bmatrix}.$$

Thus the third element of $T(x)$ is evaluated as the value of the matrix game defined by $G_3(x)$. As before, the solution to this matrix game can be computed by solving a primal/dual pair of linear programs:

$$\begin{aligned} & \min e' \check{v} \\ & \text{subject to } G_3(x) \check{v} \geq e, \check{v} \geq 0, \end{aligned}$$

$$\begin{aligned} & \max e' \check{u} \\ & \text{subject to } G_3(x)' \check{u} \leq e, \check{u} \geq 0. \end{aligned}$$

Now consider the primal and dual problems given by $G_3(x^*)$. Define

$$\begin{aligned} \check{u}^* &= \mu^*(3)/x_3^* = \frac{1-p}{2-p} (u_1, 0, u_3)', \\ \check{v}^* &= \nu^*(3)/x_3^* = \frac{1-p}{2-p} (0, 1, 0)'. \end{aligned}$$

Again, it is straightforward to verify that \check{v} and \check{u} form a feasible primal/dual pair where the primal cost of \check{v} equals the dual value of \check{u} . Thus, by the duality theorem, \check{v} and \check{u} are primal/dual optimal. This time the optimal cost works out to be $\frac{1-p}{2-p}$ which is exactly $\frac{1}{x_3^*}$. This verifies that $x_3^* = [T(x^*)]_3$ and that $\mu^*(3)$ and $\nu^*(3)$ form an equilibrium pair of mixed decisions at state $3 \in S$.

Acknowledgments. Our proof of Lemma 4.1 was inspired by an argument by John Tsitsiklis for a similar result in one-player stochastic shortest path problems. We would like to thank our anonymous SIAM reviewers who, through their persistence, have helped us to find the shortest path to establishing our results.

REFERENCES

- [1] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
- [2] D. P. Bertsekas and J. N. Tsitsiklis, *Analysis of Stochastic Shortest Path Problems*, Mathematics of Operations Research, 16 (1991), pp. 580-595.
- [3] P. R. Kumar and T. H. Shiao, *Zero Sum Dynamic Games*, in Control and Dynamic Games, C. T. Leondes, ed., Academic Press, 1981, pp. 1345-1378.
- [4] H. J. Kushner and S. G. Chamberlain, *Finite State Stochastic Games: Existence Theorems and Computational Procedures*, IEEE Transactions on Automatic Control, Vol. AC-14, No. 3, 1969.
- [5] J. F. Mertens and A. Neyman, *Stochastic Games*, Int. Journal of Game Theory, 10 (1980), pp. 53-66.
- [6] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, 1970.
- [7] L. S. Shapley, *Stochastic Games*, Proceedings of the National Academy of Sciences, Mathematics, 39 (1953), pp. 1095-1100.
- [8] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions I*, Springer-Verlag, New York, 1970.
- [9] P. Tseng, *Solving H-horizon, Stationary Markov Decision Problems in Time Proportional to Log(H)*, Operations Research Letters, 9 (1990), pp. 287-297.
- [10] J. van der Wal, *Stochastic Dynamic Programming*, Mathematical Centre Tracts 139, Mathematisch Centrum, Amsterdam 1981.
- [11] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.
- [12] N. N. Vorob'ev, *Game Theory, Lectures for Economists and Systems Scientists*. Springer-Verlag, New York, 1977.
- [13] O. J. Vrieze, *Stochastic Games with Finite State and Action Spaces*. CWI Tract 33, Centrum voor Wiskunde en Informatica (Centre for Mathematics and Computer Science), 1009 AB Amsterdam, The Netherlands, 1987.

Appendix A. Proofs of Lemmas.

We collect here some useful but well-known results. We give proofs for completeness. We require Assumption R throughout.

The following lemmas summarize some important properties of the operators $T_{\mu\nu}$, T_μ , T , \tilde{T}_ν , and \tilde{T} .

LEMMA A.1 (Monotonicity). *Suppose $x \in R^n$ and $x' \in R^n$ (or $x \in X$ and $x' \in X$) are such that $x \leq x'$ componentwise. Then*

$$(A.1) \quad T_{\mu\nu}(x) \leq T_{\mu\nu}(x'), \quad \mu \in M, \nu \in N,$$

$$(A.2) \quad T_\mu(x) \leq T_\mu(x'), \quad \mu \in M,$$

$$(A.3) \quad T(x) \leq T(x'),$$

$$(A.4) \quad \tilde{T}_\nu(x) \leq \tilde{T}_\nu(x'), \quad \nu \in N,$$

$$(A.5) \quad \tilde{T}(x) \leq \tilde{T}(x').$$

Proof. This is straightforward using the definitions of various “ T ”-operators.

Q.E.D.

LEMMA A.2 (Cost-shifting). *For all $x \in R^n$, scalars $r \in R$, integers $t > 0$, and functions $\mu^k \in M$ and $\nu^k \in N$ for $k = 1, \dots, t$ we have*

$$(A.6) \quad (T_{\mu^1\nu^1}T_{\mu^2\nu^2}\cdots T_{\mu^t\nu^t})(x + r \cdot \mathbf{1}) = (T_{\mu^1\nu^1}T_{\mu^2\nu^2}\cdots T_{\mu^t\nu^t})(x) + r \cdot \mathbf{1},$$

$$(A.7) \quad (T_{\mu^1}T_{\mu^2}\cdots T_{\mu^t})(x + r \cdot \mathbf{1}) = (T_{\mu^1}T_{\mu^2}\cdots T_{\mu^t})(x) + r \cdot \mathbf{1},$$

$$(A.8) \quad (TT\cdots T)(x + r \cdot \mathbf{1}) = (TT\cdots T)(x) + r \cdot \mathbf{1},$$

where $\mathbf{1} = (1, \dots, 1)' \in R^n$. The same relationships hold for T_μ , $T_{\mu\nu}$, \tilde{T}_{nu} , and \tilde{T} .

Proof. This follows by induction and by the definition of $T_{\mu\nu}$, T_μ , and T . **Q.E.D.**

LEMMA A.3 (Continuity). *The mappings T , T_μ , $T_{\mu\nu}$, \tilde{T}_{nu} , and \tilde{T} are continuous over R^n .*

Proof. Let x and x' be any two elements of R^n , and let $r = \|x - x'\|_\infty$, where $\|\cdot\|_\infty$ denotes the usual sup-norm (l_∞ -norm) on X . Then we have

$$x - r \cdot \mathbf{1} \leq x' \leq x + r \cdot \mathbf{1}$$

where $\mathbf{1} = (1, \dots, 1)' \in X$. Lemmas A.1 and A.2 imply that

$$\begin{aligned} T(x) - r \cdot \mathbf{1} &\leq T(x') \leq T(x) + r \cdot \mathbf{1}, \\ T_\mu(x) - r \cdot \mathbf{1} &\leq T_\mu(x') \leq T_\mu(x) + r \cdot \mathbf{1}, \\ T_{\mu\nu}(x) - r \cdot \mathbf{1} &\leq T_{\mu\nu}(x') \leq T_{\mu\nu}(x) + r \cdot \mathbf{1}. \end{aligned}$$

Therefore,

$$\begin{aligned} \|T(x) - T(x')\|_\infty &\leq \|x - x'\|_\infty, \\ \|T_\mu(x) - T_\mu(x')\|_\infty &\leq \|x - x'\|_\infty, \\ \|T_{\mu\nu}(x) - T_{\mu\nu}(x')\|_\infty &\leq \|x - x'\|_\infty. \end{aligned}$$

Thus, T is continuous on R^n . Similar arguments hold for T_μ , $T_{\mu\nu}$, \tilde{T}_{nu} , and \tilde{T} . **Q.E.D.**

The remainder of this appendix examines finite-horizon dynamic games where at each stage the maximizer has access to the minimizer’s decision. We show that minimax and maximin versions of these games can be solved in a straightforward

manner through dynamic programming. In doing so, we prove several results relevant to the main body of this paper.

LEMMA A.4. *Let M , N , T , T_μ , and \tilde{T}_ν all be defined as in previous sections. Then, for any square matrix of nonnegative elements \bar{P} and any $x \in X$ we have*

$$\begin{aligned} \min_{\mu \in M} \max_{\nu \in N} \bar{P} [c(\mu, \nu) + P(\mu, \nu)x] &= \bar{P} \min_{\mu \in M} \max_{\nu \in N} [c(\mu, \nu) + P(\mu, \nu)x] = \bar{P}T(x), \\ \max_{\nu \in N} \bar{P} [c(\mu, \nu) + P(\mu, \nu)x] &= \bar{P} \max_{\nu \in N} [c(\mu, \nu) + P(\mu, \nu)x] = \bar{P}T_\mu(x), \\ \min_{\mu \in M} \bar{P} [c(\mu, \nu) + P(\mu, \nu)x] &= \bar{P} \min_{\mu \in M} [c(\mu, \nu) + P(\mu, \nu)x] = \bar{P}\tilde{T}_\nu(x). \end{aligned}$$

Proof. It is sufficient to show that the first equation holds. The remaining equations follow as corollaries by redefining the control constraint sets for the minimizing and maximizing players as $\tilde{U}(i) = \{\mu(i)\}$ and $\tilde{V}(i, u) = \{\nu(i)\}$, respectively.

The i -th component of $\bar{P} [c(\mu, \nu) + P(\mu, \nu)x]$ can be expressed as

$$\sum_{s=1}^n \bar{p}_{is} g_s(\mu(s), \nu(s)),$$

where \bar{p}_{is} is the $(i \times s)$ -th component of \bar{P} and $g_s(u, v) \triangleq c_s(u, v) + \sum_{j=1}^n p_{sj}(u, v)J(j)$ for $u \in U(s)$ and $v \in V(s)$.

Since the min and max are taken componentwise and since the elements of P are nonnegative, we have that

$$\begin{aligned} \min_{\mu \in M} \max_{\nu \in N} \sum_{s=1}^n \bar{p}_{is} g_s(\mu(s), \nu(s)) &= \min_{\mu \in M} \max_{v^1 \in V(1), \dots, v^n \in V(n)} \sum_{s=1}^n \bar{p}_{is} g_s(\mu(s), v^s) \\ &= \min_{\mu \in M} \sum_{s=1}^n \bar{p}_{is} \max_{v^s \in V(s)} g_s(\mu(s), v^s). \end{aligned}$$

Similarly, because the elements of \bar{P} are nonnegative,

$$\begin{aligned} \min_{\mu \in M} \sum_{s=1}^n \bar{p}_{is} \max_{v^s \in V(s)} g_s(\mu(s), v^s) &= \min_{u^1 \in U(1), \dots, u^n \in U(n)} \sum_{s=1}^n \bar{p}_{is} \max_{v^s \in V(s)} g_s(u^s, v^s) \\ &= \sum_{s=1}^n \bar{p}_{is} \min_{u^s \in U(s)} \max_{v^s \in V(s)} g_s(u^s, v^s) \\ &= \sum_{s=1}^n \bar{p}_{is} (TJ)(s). \end{aligned}$$

Since this same expression applies for all $i = 1, \dots, n$, the desired result holds. **Q.E.D.**

LEMMA A.5. *Let M , N , T , T_μ , and \tilde{T}_ν all be defined as in previous sections. Then, for any $x \in X$ we have*

$$\begin{aligned} \min_{\pi_M = \{\mu^0, \dots, \mu^t\}} \max_{\pi_N = \{\nu^0, \dots, \nu^t\}} [h_{\pi_M, \pi_N}^t + P(\mu^0, \nu^0) \cdots P(\mu^t, \nu^t)x] &= T^{t+1}(x), \\ \max_{\pi_N = \{\nu^0, \dots, \nu^t\}} [h_{\mu, \pi_N}^t + P(\mu, \nu^0) \cdots P(\mu, \nu^t)x] &= T_\mu^{t+1}(x), \\ \min_{\pi_M = \{\mu^0, \dots, \mu^t\}} [h_{\pi_M, \nu}^t + P(\mu^0, \nu) \cdots P(\mu^t, \nu)x] &= \tilde{T}_\nu^{t+1}(x), \end{aligned}$$

where μ and the μ^k are elements of M , and ν and the ν^k are elements of N .

Proof. It is sufficient to show that the first equation holds. The remaining equations follow as corollaries by redefining the control constraint sets for the minimizing and maximizing players as $\tilde{U}(i) = \{\mu(i)\}$ and $\tilde{V}(i) = \{\nu(i)\}$, respectively.

Notice that

$$\begin{aligned}
& \min_{\pi_M = \{\mu^0, \dots, \mu^t\}} \max_{\pi_N = \{\nu^0, \dots, \nu^t\}} [h_{\pi_M, \pi_N}^t + P(\mu^0, \nu^0) \cdots P(\mu^t, \nu^t)x] \\
&= \min_{\pi_M} \max_{\pi_N} \{h_{\pi_M, \pi_N}^{t-1} + \bar{P}(\pi_M, \pi_N) [c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x]\} \\
&= \min_{\pi_M} \max_{\pi_N} \left\{ h_{\pi_M, \pi_N}^{t-1} + \max_{\nu^t} \bar{P}(\pi_M, \pi_N) [c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x] \right\} \\
&= \min_{\pi_M} \min_{\mu^t} \max_{\pi_N} \left\{ h_{\pi_M, \pi_N}^{t-1} + \max_{\nu^t} \bar{P}(\pi_M, \pi_N) [c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x] \right\} \\
&\geq \min_{\pi_M} \max_{\pi_N} \min_{\mu^t} \left\{ h_{\pi_M, \pi_N}^{t-1} + \max_{\nu^t} \bar{P}(\pi_M, \pi_N) [c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x] \right\} \\
&= \min_{\pi_M} \max_{\pi_N} \left\{ h_{\pi_M, \pi_N}^{t-1} + \min_{\mu^t} \max_{\nu^t} \bar{P}(\pi_M, \pi_N) [c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x] \right\},
\end{aligned}$$

where

$$\bar{\pi}_M \triangleq \{\mu^0, \dots, \mu^{t-1}\}, \quad \bar{\pi}_N \triangleq \{\nu^0, \dots, \nu^{t-1}\},$$

and $\bar{P}(\pi_M, \pi_N) \triangleq P(\mu^0, \nu^0) \cdots P(\mu^{t-1}, \nu^{t-1})$. (The inequality follows from the minimax inequality.)

We now prove the reverse relationship. First, we claim there exists a policy $\bar{\mu} \in M$ such that

$$\min_{\mu^t \in M} \max_{\nu^t \in N} \bar{P}(\pi_M, \pi_N) [c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x] = \max_{\nu^t \in N} \bar{P}(\pi_M, \pi_N) [c(\bar{\mu}, \nu^t) + P(\bar{\mu}, \nu^t)x].$$

To see this, notice that

$$\begin{aligned}
& \min_{\mu_t \in M} \max_{\nu_t \in N} \bar{P}(\pi_M, \pi_N) [c(\mu_t, \nu_t) + P(\mu_t, \nu_t)x] \\
&= \bar{P}(\pi_M, \pi_N) \min_{\mu_t \in M} \max_{\nu_t \in N} (c(\mu_t, \nu_t) + P(\mu_t, \nu_t)x) \\
&= \bar{P}(\pi_M, \pi_N) \max_{\nu_t \in N} (c(\bar{\mu}, \nu_t) + P(\bar{\mu}, \nu_t)x) \\
&= \max_{\nu_t \in N} \bar{P}(\pi_M, \pi_N) [c(\bar{\mu}, \nu_t) + P(\bar{\mu}, \nu_t)x],
\end{aligned}$$

where the first and last equalities follows from the preceding lemma, and $\bar{\mu}$ is the minimax solution to $\min_{\mu_t \in M} \max_{\nu_t \in N} (c(\mu_t, \nu_t) + P(\mu_t, \nu_t)x)$. This completes the proof of our claim. Thus,

$$\begin{aligned}
& \min_{\pi_M} \min_{\mu^t} \max_{\pi_N} \left\{ h_{\pi_M, \pi_N}^{t-1} + \max_{\nu^t} \bar{P}(\pi_M, \pi_N) [c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x] \right\} \\
&= \min_{\pi_M} \max_{\pi_N} \left\{ h_{\pi_M, \pi_N}^{t-1} + \min_{\mu^t} \max_{\nu^t} \bar{P}(\pi_M, \pi_N) [c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x] \right\} \\
&\leq \min_{\pi_M} \max_{\pi_N} \left\{ h_{\pi_M, \pi_N}^{t-1} + \max_{\nu^t} \bar{P}(\pi_M, \pi_N) [c(\bar{\mu}, \nu^t) + P(\bar{\mu}, \nu^t)x] \right\} \\
&= \min_{\pi_M} \max_{\pi_N} \left\{ h_{\pi_M, \pi_N}^{t-1} + \min_{\mu^t} \max_{\nu^t} \bar{P}(\pi_M, \pi_N) [c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x] \right\}.
\end{aligned}$$

Combining the preceding inequalities, we see that

$$\begin{aligned}
& \min_{\pi_M} \max_{\pi_N} [h_{\pi_M, \pi_N}^t + P(\mu^0, \nu^0) \cdots P(\mu^t, \nu^t)x] \\
&= \min_{\bar{\pi}_M} \max_{\bar{\pi}_N} \left\{ h_{\bar{\pi}_M, \bar{\pi}_N}^{t-1} + \min_{\mu^t} \max_{\nu^t} \bar{P}(\bar{\pi}_M, \bar{\pi}_N) [c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x] \right\} \\
&= \min_{\bar{\pi}_M} \max_{\bar{\pi}_N} \left\{ h_{\bar{\pi}_M, \bar{\pi}_N}^{t-1} + \bar{P}(\bar{\pi}_M, \bar{\pi}_N) \min_{\mu^t} \max_{\nu^t} [c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x] \right\} \\
&= \min_{\bar{\pi}_M} \max_{\bar{\pi}_N} [h_{\bar{\pi}_M, \bar{\pi}_N}^{t-1} + P(\mu^0, \nu^0) \cdots P(\mu^{t-1}, \nu^{t-1})T(x)],
\end{aligned}$$

where the second inequality follows from Lemma A.4.

Mathematical induction, repeating the same argument above, gives the desired result. **Q.E.D**

LEMMA A.6. *Let M , N , and \tilde{T} all be defined as in previous sections. Then, for any square matrix with nonnegative elements \bar{P} and any $x \in X$*

$$\max_{\nu \in N} \min_{\mu \in M} \bar{P} [c(\mu, \nu) + P(\mu, \nu)x] = \bar{P} \max_{\nu \in N} \min_{\mu \in M} [c(\mu, \nu) + P(\mu, \nu)x] = \bar{P}\tilde{T}(x).$$

Proof. The proof of this is exactly analogous to that given for Lemma A.4. The interchange of the max and min has no bearing on the logical flow of the argument.

Q.E.D.

LEMMA A.7. *Let M , N , and \tilde{T} all be defined as in previous sections. Then, for any $x \in X$*

$$\max_{\pi_n = \{\nu^0, \dots, \nu^t\}} \min_{\pi_M = \{\mu^0, \dots, \mu^t\}} [h_{\pi_M, \pi_N}^t + P(\mu^0, \nu^0) \cdots P(\mu^t, \nu^t)x] = \tilde{T}^{t+1}(x),$$

where the μ^k are elements of M , and the ν^k are elements of N .

Proof. The proof of this is symmetrical to that given for Lemma A.5. **Q.E.D.**

Using the fact that $T(x) = \tilde{T}(x)$ (under Assumption R) we obtain

$$\begin{aligned}
& \min_{\pi_M = \{\mu^0, \dots, \mu^t\}} \max_{\pi_N = \{\nu^0, \dots, \nu^t\}} [h_{\pi_M, \pi_N}^t + P(\mu^0, \nu^0) \cdots P(\mu^t, \nu^t)x] \\
&= T^{t+1}(x) \\
&= \tilde{T}^{t+1}(x) \\
&= \max_{\pi_N = \{\nu^0, \dots, \nu^t\}} \min_{\pi_M = \{\mu^0, \dots, \mu^t\}} [h_{\pi_M, \pi_N}^t + P(\mu^0, \nu^0) \cdots P(\mu^t, \nu^t)x].
\end{aligned}$$

Thus, for finite-horizon games (with or without a terminal state), an equilibrium cost exists and can be found via dynamic programming iterations.