Stochastic Shortest Path Games: Theory and Algorithms

by

Stephen David Patek

B.S., The University of Tennessee, Knoxville (1991)
S.M.E.E., Massachusetts Institute of Technology (1994)

Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1997

Author .....
Department of Electrical Engineering and Computer Science
September 5, 1997

Certified by .....
Dimitri Bertsekas
Professor
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Stochastic Shortest Path Games: Theory and Algorithms

by

Stephen David Patek

Submitted to the
Department of Electrical Engineering and Computer Science
on September 5, 1997,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

This thesis is devoted to stochastic shortest path games. These are two-player, zero-sum stochastic games where one player seeks to drive an underlying finite-state system to a terminal state along a least expected-cost path. The other player acts to prevent termination and otherwise seeks to make termination as expensive as possible. At every stage of the game, the players implement control actions selected from compact constraint sets. They make their decisions simultaneously with full knowledge of the current state of the system but without knowledge of what the other player will do. We impose relatively mild assumptions about the transition probabilities and cost functions. One special case of this formulation is that of mixed strategies over finite action sets. We employ an undiscounted, additive cost structure which generalizes many of the stochastic games previously considered. Policies for the minimizer which permit the maximizer to prolong the game indefinitely are allowable as long as the resulting cost is infinite from some initial state. We do not assume nonnegativity of cost; we make alternative assumptions which guarantee that the terminal state is reached with probability one, at least under equilibrium policies.

Our main results relate to proving the existence of equilibrium solutions, characterizing the equilibria, and establishing the convergence of algorithms. We show that both players have stationary equilibrium policies. We also show that the equilibrium cost-to-go function is characterized as the unique solution to a game-theoretic generalization of Bellman's equation. We prove that value iteration and policy iteration converge to the equilibrium. We also consider several interesting variations of policy iteration, some of which converge in theory and others only in practice. We end the thesis by relating stochastic shortest path games to a broad class of average cost games. This allows us to derive new results for the latter concerning the existence of solutions and the convergence of algorithms.

Thesis Supervisor: Dimitri Bertsekas
Title: Professor

Acknowledgments

My thanks go principally to my advisor, Dimitri Bertsekas, a great innovator, scholar, and educator. He has faithfully provided the means for me to finish this thesis even though I haven't always been on the shortest path to completion.

I gratefully acknowledge John Tsitsiklis and Dimitris Bertsimas (the other members of my thesis committee), who have provided much insight and inspiration. Thanks also go to my other mentors at MIT: Michael Athans, Munther Dahleh, Alvin Drake, Robert Gallager, Sanjoy Mitter, and James Munkres.

Next, I would like to thank my wife, Dana, for her constant love and encouragement throughout my graduate career. My thanks also go to my parents, brother, and sister for their total confidence and willingness to brag. I would like to thank our close personal friends, John and Robin (Koh) Krienke, and their children, Melissa and Lydia, for companionship and fun over the years. My thanks also go to our friend Wesley McDermott: colleague, wind sportsman, and fellow savior of the universe.

I thank all of my friends and colleagues at LIDS, MIT, and Alphatech: Jinane Abounadi, Frank Aguirre, Randy Berry, Constantinos Boussios, Alan Chao, Mike Daniel, Joel Douglas, Maruska and Nicola Elia, Marcos Escobar, Austin Frakt, Seema Jaggi, Patrick Kreidl, David Logan, Peter Marbach ("Snow Tiger"), Gina Mourtzinou, Dimitris Papaioannou, Ioannis Paschalidis, Georgia Perakis, Lazaros Polymenakos, Nils Sandell, Sekhar Tatikonda, Sean Warnick, Benjamin Van Roy, John Wissinger, and Cynara Wu. I have learned a lot from all of these guys!

This research was conducted at the MIT Laboratory for Information and Decision Systems and was supported by an Office of Naval Research Fellowship, US Army Research Office grants ARO-ASSERT DAAH04-93-G-0169 and ARO DAAH04-95-1-0103, and a National Science Foundation grant NSF 9206379-NCR.

Contents

List of Figures

Chapter 1

Introduction

Game theory, the theory of multiagent optimization, has intrigued researchers in the mathematical sciences for years. This interest is well justified since game-theoretic models emerge naturally in many fields including economics, finance, biology, engineering systems, and operations research.

This thesis is devoted to the theory of a new class of games, called stochastic shortest path games. These are two-player, zero-sum stochastic games where one of the players (the minimizer) seeks to drive an underlying finite-state system to a terminal state along a least expected-cost path. The other player (the maximizer) acts to prevent termination and otherwise seeks to make termination as expensive as possible. These games generally fit into the framework of Markov decision processes, however, instead of having a single controller, we now have two players who act independently. These "competitive Markov decision processes" are often called stochastic games, and we refer the reader to a new book by Filar and Vrieze [20] for a general introduction. The main distinguishing feature of stochastic shortest path games is that they employ a more general cost structure than earlier stochastic games. Specifically, by assuming the existence of a terminal state which is reached with probability one under equilibrium policies, we allow stochastic shortest path games to evolve on an infinite time horizon without cost discounting or averaging. A second distinguishing feature of stochastic shortest path games is that they allow the players to choose actions from more general constraint sets than previously consid-

ered. The traditional approach in stochastic games is to restrict attention to the case where each player has only finitely many control options at each state. To achieve an equilibrium in such games it is usually necessary for the players to employ "mixed actions," which are probability distributions over the underlying finite action sets. In stochastic shortest path games, however, the players are allowed to choose from more general compact constraint sets. Our main results extend the state of the art in several areas, particularly in establishing the existence of equilibrium solutions, characterizing the equilibria, and establishing the convergence of algorithms.

Stochastic shortest path games are useful in modeling conflict situations where there is a notion of the game ending (terminating) at some random future time and where one of the players wishes to minimize the cost of reaching termination while the other wishes to maximize this cost. We list the following conceptual examples.

1. **Pursuit and Evasion:** minimize the time required to capture an intelligent, evasive opponent. This is a very common model in logistics, e.g. aircraft avoiding other aircraft or missiles, submarines avoiding destroyers, etc. The "vehicles" in these games do not have to be physical devices. An example we will consider in the sequel is an inspection game involving a manufacturer who must dump waste and an inspector who wants to catch the manufacturer "in the act" two nights in a row.

2. **Minimax Resource Allocation:** apply resources to mitigate the effect of a worst-case opponent (perhaps as a model for nature). As an example from warfare, consider the problem of theater missile defense where it is necessary to defend against incoming missile attacks with a limited number of interceptors. The game ends when the attacker runs out of missiles, and the objective of the defender is to minimize the expected worst-case damage.

3. **Parlor Games:** move pieces on a playing-board to achieve a certain goal. In particular, consider two-player games where there is one clear winner such as backgammon or chess.

1.1 Outline

In Chapter 2, we provide a rigorous formulation of stochastic shortest path games. We indicate how previously studied stochastic games, such as the terminating games of Shapley, can be viewed as special cases of the stochastic shortest path model. We also describe two concrete examples of stochastic shortest path games. These will be revisited in subsequent chapters.

In Chapter 3, we establish the existence of a long term equilibrium cost function (equilibrium value function) for stochastic shortest path games. We show that the equilibrium can be achieved by stationary policies of the opposing players and is characterized as the unique solution of a game-theoretic generalization of Bellman's equation.

In Chapter 4, we consider various dynamic programming algorithms for solving stochastic shortest path games. Specifically, we prove the convergence of game-theoretic versions of value iteration and policy iteration. The policy iteration we discuss generates a sequence of policies for only one of the players. (Policy evaluation amounts to computing the worst-case response of the opposing player.) We introduce and prove the convergence of an asynchronous form of policy iteration, indicating that policy iteration is relatively robust. To complete this notion, we prove an error-bound result which indicates that, for some stochastic shortest path games, if the policy evaluations and updates are accurate enough, then eventually the worst-case costs of the policies generated by the algorithm approach the equilibrium infinitely often. Next, we consider an alternative form of policy iteration, called naive policy iteration, where a sequence of pairs of policies is generated (one for each player) and where policy evaluation requires only the computation of expected costs in a Markov reward process. It is well-known that naive policy iteration is not a globally convergent algorithm, even in the benign case of discounted cost games. We are led to examine a modification of the algorithm due to Filar and Tolwinski [19] which relies on an interpretation of naive policy iteration as Newton's method. Next, we show how some of the general dynamic programming algorithms specialize to the case of

symmetric sequential games. It turns out that the resulting symmetric value iteration is globally convergent, whereas symmetric policy iteration, being equivalent to naive policy iteration, is not. Thus, symmetry (by itself) is not enough to provide a mechanism for convergence of naive policy iteration.

In Chapter 5, we explore the relationship between stochastic shortest path games and stochastic games with an average cost objective function. Games of the latter type have been the subject of intense study almost since the time of Shapley. We first describe an alternative proof for a result from Chapter 3 which uses well-known results from the literature on average cost games. We then show how our more general theory about stochastic shortest path games provides insight into a broad class of average cost games, leading to new results.

Chapter 6 finishes the thesis with a summary of results and a statement of open problems. Two appendices are included. Appendix A states and proves a number of technical results which are used throughout Chapters 3, 4, and 5. Appendix B lists a number of alternative algorithms for stochastic games which have been suggested since the time of Shapley. A number of these algorithms can be applied to stochastic shortest path games, but their convergence properties are unclear at present.

The introductory section of each chapter contains a review of the pertinent literature.

1.2 Contributions

Aside from the specific technical results we develop, several broad themes have emerged in this work which represent the true contributions of this thesis.

1. Our stochastic shortest path assumption gives rise to a more general cost structure than has previously been considered in stochastic games.

2. Our regularity assumptions are also more general than those previously considered, especially in the context in average cost games. Our assumptions allow both players to choose actions from arbitrary compact subsets of metric spaces.

In the case of mixed strategies, the players are effectively required to choose from simplicial constraint sets. Our results show that this type of special structure is unnecessary in proving the existence of solutions to games and in establishing the convergence of algorithms.

3. Stochastic shortest path games can be viewed as a direct extension of the stochastic shortest path problems originally studied by Bertsekas and Tsitsiklis in [8]. The results contained within this thesis serve to show the extent to which their analytical techniques and constructions apply. Similarly, our results about recurrent-state average cost games (in Chapter 5) illustrate the power and flexibility of some new analytical devices introduced by Bertsekas in [3].

4. As a whole, our analytical techniques represent an alternative to the so-called "limit discount equation approach" which is the traditional way in which undiscounted stochastic games are studied and requires the players to optimize with respect to mixed strategies over finite sets of actions.

1.3 Scope

While we are somewhat concerned with approximations, this thesis will not continue the analysis of Whitt [65] or Tidball and Altman [58] which is directed toward games whose state spaces are infinite, where solutions are obtained (or at least proven to exist) through state aggregation and the solution of a sequence of simpler finite-state games. In a similar vein, we do not consider two recent algorithms of Harmon and others [23, 24, 22] (advantage updating and advantage learning), which are geared toward solving discrete approximations of differential games.

We omit from consideration a number of algorithms which take advantage of special structure in games. In particular, we leave out linear programming-based methods that are specialized to games with a single controller, switching controller, and/or games with separable rewards or additive rewards and transitions. For a survey of these topics, we refer the reader to Raghavan et al. [43] or Filar and Vrieze [20].

Chapter 2

Formulation

This chapter serves to define stochastic shortest path games. Generally speaking, stochastic shortest path games are two-player, zero-sum games where one of the players (the minimizer) seeks to drive an underlying finite-state system to a special terminal state along a least expected cost path. The terminal state is absorbing and cost-free. The other player (the maximizer) seeks to interfere with the first player's progress so as to maximize the expected total cost. In actual play, the players choose actions from compact constraint sets at each stage. They have full knowledge of the state of the system but are unaware of what the other player will do. We employ an undiscounted additive cost structure, and we admit 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. We make alternative assumptions which guarantee that, at least under optimal (equilibrium) policies, the terminal state is reached with probability one. The formal assumptions for stochastic shortest path games generalize (to the case of two-players) those for stochastic shortest path problems of Bertsekas and Tsitsiklis [8]. In particular, our games are characterized by either inevitable termination (under all policies) or an incentive for the minimizer to drive the system to termination in a finite expected number of stages.

Dynamic games with stochastic payoff and transition sequences have been studied for some time. The field was initiated by Shapley in his famous paper [52]. In Shapley's games, two players are successively faced with matrix games (in mixed

strategies) where the costs to the minimizer and the transitions to new matrix games are influenced by the decisions of the players at each stage. In this formulation, the state of the system is exactly the matrix game currently being played. It is assumed that this set of states is finite and that there is a nonzero minimal probability that at any stage the game will transition to a terminal state, ending the sequence of rewards and payoffs.

In the time since Shapley's paper, game theorists have actively studied extensions to his terminating-game model. Kushner and Chamberlain in [28] studied undiscounted, pursuit-evasion games where, in a state space of $n + 1$ elements, there is a terminal state corresponding to the evader being "caught." After making some regularity assumptions on the transition probability and cost functions, they consider pure strategies over compact action spaces. In addition, they assume that either of the following are true.

1. There exists $\epsilon > 0$ such that for all pairs of stationary policies the probability of terminating within n stages is at least ϵ.

2. The transition costs are uniformly bounded below by $\delta > 0$ and there exists a stationary policy for the pursuer and an $\epsilon > 0$ such that for every stationary policy of the evader the probability of terminating within n stages is at least ϵ.

In [60], van der Wal considered a special case of Kushner and Chamberlain's games, where the pursuer is endowed with more power to drive the system to termination. In [20], Filar and Vrieze described "transient" stochastic games, where there is no cost discounting but the expected number of stages to termination is always finite. As will be shown in this and the following chapter, the class of stochastic shortest path games includes all of these earlier games as special cases.

Other extensions of Shapley's model are possible. Maitra and Parthasarathy [32, 33], Parthasarathy [38], and Kumar and Shiau [27] have considered stochastic games with general state spaces and action sets. These models lie beyond the scope of the present work. Game theorists have also studied stochastic games with average cost objectives. We shall consider games of this type in Chapter 5.

2.1 Mathematical Formulation

Let $S = \{1, 2, \ldots, n\} \cup \{\Omega\}$ denote a finite state space. For each $i \in S$, let $U(i)$ and $V(i)$ denote 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 $c_i(u, v)$. The state Ω, called the terminal state, has special significance to us; it is absorbing and has zero-cost. That is, $p_{\Omega\Omega}(u, v) = 1$ and $c_\Omega(u, v) = 0$ for all $u \in U(\Omega)$ and $v \in V(\Omega)$.

Let M and N denote the sets of allowable one-stage policies for the minimizer and maximizer, respectively, defined by

$$\begin{aligned} M &= \left\{ \mu : S \mapsto \bigcup_{i \in S} U(i) \quad | \quad \mu(i) \in U(i), \quad \forall\, i \in S \right\}, \\ N &= \left\{ \nu : S \mapsto \bigcup_{i \in S} V(i) \quad | \quad \nu(i) \in V(i), \quad \forall\, i \in S \right\}. \end{aligned}$$

The players are allowed to use nonstationary policies to optimize additive cost. Let $\bar{M}$ and $\bar{N}$ denote the sets of allowable nonstationary policies, defined by

$$\begin{aligned} \bar{M} &= \{\pi_M = \{\mu_0, \mu_1, \ldots\} \quad | \quad \mu_k \in M, \quad \forall\, k\}, \\ \bar{N} &= \{\pi_N = \{\nu_0, \nu_1, \ldots\} \quad | \quad \nu_k \in N, \quad \forall\, k\}. \end{aligned}$$

Given $\mu \in M$ and $\nu \in N$,

$$P(\mu, \nu) = \begin{bmatrix} p_{11}(\mu(1), \nu(1)) & \cdots & p_{1n}(\mu(1), \nu(1)) \\ \vdots & \vdots & \vdots \\ p_{n1}(\mu(n), \nu(n)) & \cdots & p_{nn}(\mu(n), \nu(n)) \end{bmatrix}, \quad c(\mu, \nu) = \begin{pmatrix} c_1(\mu(1), \nu(1)) \\ \vdots \\ c_n(\mu(n), \nu(n)) \end{pmatrix}$$

are the corresponding transition probability matrix and cost vector, respectively. (Note that the row-sums of $P(\mu, \nu)$ are less than or equal to one.)

Given two policies, $\pi_M = \{\mu_0, \mu_1, \ldots\} \in \bar{M}$ and $\pi_N = \{\nu_0, \nu_1, \ldots\} \in \bar{N}$,

$$J_{\pi_M,\pi_N}(i) = \liminf_{t\to\infty} \; h^t_{\pi_M,\pi_N}(i), \qquad i = 1, \ldots, n, \tag{2.1}$$

is the objective function which the two players seek to minimize and maximize, respectively, where

$$h^t_{\pi_M,\pi_N}(i) \triangleq \left[c(\mu_0,\nu_0) + \sum_{k=1}^{t}[P(\mu_0,\nu_0)P(\mu_1,\nu_1)\cdots P(\mu_{k-1},\nu_{k-1})]c(\mu_k,\nu_k)\right]_i . \tag{2.2}$$

Note that $h^t_{\pi_M,\pi_N}(i)$ can be interpreted as the expected $(t+1)$-stage cost from i under π_M and π_N. Similarly, $J_{\pi_M,\pi_N}(i)$ can be (loosely) interpreted as the expected total cost from i. Because the terminal state $\Omega \in S$ is absorbing and cost free, the expected total cost from Ω is zero under all pairs of policies. It is often convenient to think of the cost functions J_{π_M,π_N} as vectors in $\Re^n$ with components $J_{\pi_M,\pi_N}(1), \ldots, J_{\pi_M,\pi_N}(n)$.

We now state a few useful definitions. We say that a policy $\pi_M = \{\mu_0, \mu_1, \ldots\} \in \bar{M}$ 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. Now consider an arbitrary pair of policies, $\pi_M = \{\mu_0, \mu_1, \ldots\} \in \bar{M}$ and $\pi_N = \{\nu_0, \nu_1, \ldots\} \in \bar{N}$. We say that the corresponding Markov chain terminates with probability one if the following limit satisfies

$$\lim_{t\to\infty} P(\mu_0,\nu_0)P(\mu_1,\nu_1)\cdots P(\mu_t,\nu_t) = 0_{n\times n},$$

where $0_{n\times n}$ is the $n \times n$ matrix whose elements are all zero. 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. If the pair (π_M, π_N) is not terminating with probability one, then 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 π_N for the maximizer such that under (μ, π_N) there

is a positive probability that the game will never terminate from some initial state.] The designation of proper (or improper) applies only to stationary policies of the minimizer.

Let $\mathcal{J}$ denote the set of all functions J that map $\{1,\ldots,n\}$ to $\Re$. Again, it is often useful to view the elements of $\mathcal{J}$ as vectors in $\Re^n$. Let $\mathbf{0} \in \mathcal{J}$ be the vector whose components are all zeros, and similarly let $\mathbf{1} \in \mathcal{J}$ be the vector whose components are all ones. Given $J \in \mathcal{J}$ and a scalar γ, then $\gamma J \in \mathcal{J}$ is defined so that $(\gamma J)(i) = \gamma J(i)$. Also, given $J, \bar{J} \in \mathcal{J}$, 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 apply on $\mathcal{J}$:

$$T_{\mu\nu}J = c(\mu,\nu) + P(\mu,\nu)J, \qquad \mu \in M,\ \nu \in N; \tag{2.3}$$

$$T_\mu J = \sup_{\nu\in N}\left[c(\mu,\nu) + P(\mu,\nu)J\right], \qquad \mu \in M; \tag{2.4}$$

$$TJ = \inf_{\mu\in M}\sup_{\nu\in N}\left[c(\mu,\nu) + P(\mu,\nu)J\right]; \tag{2.5}$$

$$\tilde{T}_\nu J = \inf_{\mu\in M}\left[c(\mu,\nu) + P(\mu,\nu)J\right], \qquad \nu \in N; \tag{2.6}$$

$$\tilde{T}J = \sup_{\nu\in N}\inf_{\mu\in M}\left[c(\mu,\nu) + P(\mu,\nu)J\right]. \tag{2.7}$$

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).

Stochastic shortest path games are now formally defined by the following two assumptions. (They are in effect from here through Chapter 4, unless specifically stated otherwise.)

Assumption 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 $p_{ij}(u,v)$ are continuous with respect to $(u,v) \in U(i) \times V(i)$. 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)

3. *The outer extrema in the operators T and $\tilde{T}$ are achieved by elements of M and N, respectively. (That is, for every $J \in \mathcal{J}$, there exists $\mu \in M$ and $\nu \in N$ such that $TJ = T_\mu J \in \mathcal{J}$ and $\tilde{T}J = \tilde{T}_\nu J \in \mathcal{J}$.)*

4. *For every $J \in \mathcal{J}$, we have $TJ = \tilde{T}J$.*

Assumption SSP *(Stochastic Shortest Path) The following are true.*

1. *There exists a 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 such that* $\lim_{t\to\infty} h^t_{\pi_M,\pi_N}(i) = \infty$.

Note that part 1 of Assumption R implies that the sets M and N are compact. Moreover, parts 1 and 2 of Assumption R, along with the extreme value theorem (a.k.a. the Weierstrass Theorem), imply that the operators T_μ and $\tilde{T}_\nu$ are achieved by elements of N and M, respectively. (That is, for every $J \in \mathcal{J}$, there exists $\nu \in N$ such that $T_\mu J = T_{\mu\nu} J \in \mathcal{J}$, and similarly there exists $\mu \in M$ such that $\tilde{T}_\nu J = T_{\mu\nu} J \in \mathcal{J}$.) Part 4 of Assumption R is satisfied under conditions for which a minimax theorem can be used to interchange "inf" and "sup". Some examples of games which satisfy Assumption R are given in Section 2.3.

We remark that the regularity assumptions of Kushner and Chamberlain in [28] differ from ours only by requiring the functions c_i to be continuous.

The following lemmas characterize three important properties of the dynamic programming operators. These results are generally well known and their proofs appear in Appendix A.

Lemma A.1 [Monotonicity] *Given $J, \bar{J} \in \mathcal{J}$, if $J \leq \bar{J}$, then*

$$TJ \leq T\bar{J}.$$

The same is true of the other dynamic programming operators.

Lemma A.2 *Given $J \in \mathcal{J}$ and a positive scalar r,*

$$T(J + r\mathbf{1}) \leq TJ + r\mathbf{1}.$$

The same inequality holds for the other dynamic programming operators. The inequalities are reversed if $r < 0$.

Lemma A.3 [Continuity] *Given $J, \bar{J} \in \mathcal{J}$, then*

$$\|T(J - \bar{J})\|_\infty \leq \|J - \bar{J}\|_\infty.$$

Thus, T is nonexpansive on $\mathcal{J}$ and therefore continuous. The same is true of the other dynamic programming operators.

2.2 The Generality of Assumption SSP

Here we collect some remarks which indicate the generality of the stochastic shortest path cost structure.

2.2.1 Testing for Proper Policies

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 $\nu \in N$. Furthermore, if $\mu \in M$ is improper, then we can always find a stationary policy $\nu \in N$ which is

prolonging when paired with μ. This is clear from the following lemma whose proof appears in Appendix A.

Lemma A.4 *If $\mu \in M$ is such that the pair (μ, ν) is terminating with probability one for all $\nu \in N$, then μ is proper.*

2.2.2 Relation to Discounted Cost Games

Discounted cost stochastic games incorporate a discount factor $\alpha \in [0, 1)$ which attenuates the present value of future costs. Specifically, α^t is the present value of a unit of cost incurred t stages in the future. As a result, the minimizer's expected $(t+1)$-stage discounted cost from state $i \in \{1, \ldots, n\}$, would be defined as

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

and the corresponding operator T would be defined as

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

We note that every discounted cost game can be transformed into an equivalent stochastic shortest path game. Indeed, if $S = \{1, \ldots, n\}$ is the state space of a discounted cost game, then we may augment S with an extra state Ω which is absorbing and cost-free. The effect of the discount factor can be obtained by defining transition costs $\bar{c}_i(u, v)$ in the new game to be the same as in the original game and the transition probabilities $\bar{p}_{ij}(u, v)$ as $\bar{p}_{ij}(u, v) = \alpha p_{ij}(u, v)$ and $\bar{p}_{i\Omega}(u, v) = 1 - \alpha$ for all $i, j \in \{1, \ldots, n\}$ and $u \in U(i)$, $v \in V(i)$.

2.2.3 Relation to Terminating Games

Shapley's Terminating Games The terminating games of Shapley in [52] are defined by the assumption that for each state i (in a finite state space) there is a minimal probability $t_i > 0$ of terminating in the next transition. In this case, it is

clear that all stationary policies of the minimizer are proper and Assumption SSP is satisfied. In Chapter 3, we will show that Kushner and Chamberlain's pursuit-evasion games [28] satisfy Assumption SSP.

Transient Games The "transient" games of Filar and Vrieze [20] are defined by the assumption that the expected number of stages to termination is finite for all pairs of policies; that is

$$\sum_{t=0}^{\infty}[P(\mu_0,\nu_0)\cdots P(\mu_{t-1},\nu_{t-1})]\cdot\mathbf{1}<\infty$$

for all $\pi_M=\{\mu_0,\mu_1,\ldots\}\in\bar{M}$ and $\pi_N=\{\nu_0,\nu_1,\ldots\}\in\bar{N}$. Consequently, the Markov chain associated with each pair (π_M,π_N) is terminating with probability one, and each such game satisfies Assumption SSP.

2.3 The Generality of Assumption R

Assumption R is stated in such a way that it is easy to prove mathematical results about stochastic shortest path games. It is, in fact, a very general assumption which includes a large number of interesting special cases.

2.3.1 Games with Finite Constraint Sets

Games with finite constraint sets are allowable within the context of Assumption R. Let W denote a generic set of actions available to one of the players at some state. W can be viewed as a finite subset of a metric space (e.g. a finite set of real numbers). As a finite set, W is compact. With the subspace topology on W, every map from W to $\Re$ is continuous, so parts 1 and 2 of Assumption R are satisfied. Moreover, since "inf" and "sup" are always achieved on finite sets, part 3 of Assumption R is also satisfied. All that remains to be verified is part 4. This will be true whenever the one-stage game defined by the operators T and $\tilde{T}$ has equilibrium solutions in *pure* strategies for all terminal costs $J\in\mathcal{J}$. Unfortunately, this is not true in general,

and it may be necessary to allow the players to use randomized policies (i.e. mixed strategies), as discussed in Section 2.3.4.

2.3.2 Sequential Games

Various types of sequential games satisfy Assumption R. Generically, what makes a game "sequential" is the fact that, at any given stage, only one player has a decision to make. To capture this in our framework, we specify that, at each state, at least one of the players must choose from a singleton constraint set. We now consider two types of sequential games; they are distinguished by the order in which decisions affect state transitions.

The first type of sequential game is "truly sequential" in that state transitions occur immediately after each nontrivial decision. As an example, consider a board-game where players move pieces in turn (sequentially). Each movement here corresponds to a change of state in a finite state space. At every stage, depending on the state of the game, one of the players must choose from a set of actions containing only a single element, while the other player chooses from an arbitrary compact subset of a metric space. Given that part 2 of Assumption R is satisfied, parts 3 and 4 hold trivially.

The second type of sequential game is one where the conflict situation is naturally modeled as follows: at any given stage a "leading player" announces his decision first; the "following player", knowing the decision of the leading player, chooses an action in response; and finally, after both players have made their decisions, a state transition occurs within a finite state space. We shall refer to these as *leader/follower* games. (As an example, consider an attack/defense game where the attacking player always leads with a profile of attack resources against various targets, and the defending player then responds with a profile of defense resources.) As long as the leading player always has only finitely many actions from which to choose, leader/follower games can be transformed into sequential games of the first type. This can be done by augmenting the original state space S with a set of new states $\bar{S}$ which consists of all pairs (i, w) with $i \in S$ and $w \in W(i)$, where $W(i) = U(i)$ if the leading player at

state i is the minimizer, and $W(i) = V(i)$ otherwise. Then, $S \bigcup \bar{S}$ is the state space of a sequential stochastic shortest path game of the first type. (In the transformed version of the game, the system transitions from $i \in S$ to $(i, w) \in \bar{S}$ when the leading player chooses w (knowing i); the system then probabilistically transitions back to a state in S once the following player chooses an action. Since the leading player always has only finitely many options, the augmented state space $S \bigcup \bar{S}$ is finite.)

2.3.3 Symmetric Sequential Games

Many interesting games are symmetric in the sense that, if the system is in a particular state, an action that would be rational for one of the players would be rational for the other player if the "tables were turned". Many popular board games, such as chess and backgammon, fit this description. In this section, we formalize the notion of symmetric sequential games, and later we will show how the general results for stochastic shortest path games apply.

Let the minimizing player be identified as player 1, and let the maximizing player be identified as player 2. Let $\tilde{S}$ denote a finite set of positions, with elements labeled $i = 1, \ldots, \tilde{n}$. The state space for the symmetric sequential game is $S = (\{1, 2\} \times \tilde{S}) \cup \{\Omega\}$, where Ω denotes a terminal state. As in Section 2.3.2, there is, at every state in the game, a notion of "whose turn it is". In this particular framework, when the state of the game is $(z, i) \in S$, player $z \in \{1, 2\}$ gets to choose from a set of actions $W(i)$ (which is a compact subset of a metric space), while the other player chooses from a singleton. Let $s : \{1, 2\} \mapsto \{1, 2\}$ be a "switching function" such that $s(1) = 2$ and $s(2) = 1$. The probability of transitioning from $(z, i) \in S$ to $(z, j) \in S$ under $a \in W(i)$ is denoted $\bar{p}_{ij}(a)$. The probability of transitioning from $(z, i) \in S$ to $(s(z), j) \in S$ under $a \in W(i)$ is denoted $\bar{r}_{ij}(a)$. Clearly, $1 - \sum_{j \in \tilde{S}} \bar{p}_{ij}(a) - \sum_{j \in \tilde{S}} \bar{r}_{ij}(a)$ is the probability of transitioning from (z, i) to Ω under the action $a \in W(i)$. The expected cost to player 1 of transitioning from $(1, i) \in S$ under $a \in W(i)$ is denoted $\bar{c}_i(a)$. The expected cost to player 1 of transitioning from $(2, i) \in S$ under $a \in W(i)$ is $-\bar{c}_i(a)$. We assume that, under all pairs of policies, the terminal state is reached with probability one.

At this point it is possible to introduce sets of stationary policies, M and N, and nonstationary policies, $\bar{M}$ and $\bar{N}$, for the opposing players, subject to the sequential constraints of the game. The space of cost functions may be expressed as

$$\mathcal{J} = \left\{ J = \begin{pmatrix} X \\ Y \end{pmatrix} \quad \middle| \quad X : \tilde{S} \mapsto \Re, \quad Y : \tilde{S} \mapsto \Re \right\}.$$

J, X, and Y may be interpreted as vectors in Euclidean spaces with appropriate dimensions. The transition probability matrix under $\mu \in M$ and $\nu \in N$ may be expressed as

$$P(\mu, \nu) = \begin{bmatrix} \bar{P}(\mu) & \bar{R}(\mu) \\ \bar{R}(\nu) & \bar{P}(\nu) \end{bmatrix}$$

where

1. $\bar{P}(\mu)$ is the matrix whose (i,j)-th element is $\bar{p}_{ij}(\mu(1,i))$,
2. $\bar{P}(\nu)$ is the matrix whose (i,j)-th element is $\bar{p}_{ij}(\nu(2,i))$,
3. $\bar{R}(\mu)$ is the matrix whose (i,j)-th element is $\bar{r}_{ij}(\mu(1,i))$, and
4. $\bar{R}(\nu)$ is the matrix whose (i,j)-th element is $\bar{r}_{ij}(\nu(2,i))$.

The expected transition cost vector under $\mu \in M$ and $\nu \in N$ may be expressed as

$$c(\mu, \nu) = \begin{pmatrix} \bar{c}(\mu) \\ -\bar{c}(\nu) \end{pmatrix}$$

where

1. $\bar{c}(\mu)$ is the vector whose i-th component is $\bar{c}_i(\mu(1,i))$ and
2. $\bar{c}(\nu)$ is the vector whose i-th component is $\bar{c}_i(\nu(2,i))$.

2.3.4 Games in Mixed Strategies Over Finite Action Sets

In many games where the control constraint sets $U(i)$ and $V(i)$ are finite, the equality of TJ and $\tilde{T}J$ for all $J \in \mathcal{J}$ cannot be guaranteed. One way to get around this is to modify the game so that the new constraint sets are actually the sets of probability distributions over the underlying pure actions. That is, in formulating their policies, the players choose probability distributions ("mixed actions") rather than specific pure actions. In the literature on games, these randomized policies are often called mixed strategies. For convenience, whenever we refer to "mixed strategies", we are referring to probability distributions over *finite* underlying sets of actions. The case of mixed strategies over arbitrary action sets is beyond the scope of this thesis. Also, we do not consider "behavioral strategies," where the mixed actions selected at each stage can depend on the entire past history of play.

At this point it is useful to define a special notation for the case of mixed strategies. Let $A(i)$ and $B(i)$ denote the finite sets of underlying actions available to the minimizer and maximizer (respectively) at state i. The players' control constraint sets for the game are

$$\begin{aligned} U(i) &= \left\{ u \in \Re^{|A(i)|} \;\middle|\; \sum_{a \in A(i)} u_a = 1; \quad u_a \geq 0 \right\}, \\ V(i) &= \left\{ v \in \Re^{|B(i)|} \;\middle|\; \sum_{b \in B(i)} v_b = 1; \quad v_b \geq 0 \right\}. \end{aligned}$$

The functions $p_{ij}(u,v)$ and $c_i(u,v)$ are respectively of the form

$$\begin{aligned} p_{ij}(u,v) &= \sum_{a \in A(i)} \sum_{b \in B(i)} \underline{p}_{ij}(a,b) u_a v_b, \\ c_i(u,v) &= \sum_{j \in S} \sum_{a \in A(i)} \sum_{b \in B(i)} \underline{g}_{ij}(a,b) \underline{p}_{ij}(a,b) u_a v_b. \end{aligned}$$

where the functions $\underline{p}_{ij}$ and $\underline{g}_{ij}$ are the transition probabilities and costs associated with the pure actions. The Minimax Theorem of von Neumann [61] implies that Assumption R is satisfied.

2.3.5 Games Satisfying a Generalized Minimax Theorem

Assumption R is satisfied if:

1. the sets $U(i)$ and $V(i)$ are nonempty, convex, 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 matrix with dimensions commensurate with $U(i)$ and $V(i)$, and

3. the functions $c_i(u, v)$ are

   (a) convex and lower semi-continuous as functions of $u \in U(i)$ with v fixed, and

   (b) concave and upper semi-continuous as functions of $v \in V(i)$ with u fixed.

This follows from the Sion-Kakutani theorem (see [55], p.232 or [46], p. 397), which can be viewed as a generalization of the Minimax Theorem of von Neumann [61].

By restricting the domain of the dynamic programming operators to

$$\mathcal{J}_+ \stackrel{\Delta}{=} \{ \text{ functions } : \{1, \ldots, n\} \mapsto \Re_+ \},$$

the Sion-Kakutani theorem also implies that part 4 of Assumption R is satisfied if statements 1 and 3 above are true and statement 2 is weakened so that the functions $p_{ij}(u, v)$ are

1. convex and lower semi-continuous as functions of $u \in U(i)$ with v fixed, and

2. concave and upper semi-continuous as functions of $v \in V(i)$ with u fixed.

However, to ensure that T and $\tilde{T}$ map $\mathcal{J}_+$ to $\mathcal{J}_+$ we would also have to assume that the transition costs $c_i(u, v)$ are nonnegative.

2.4 Examples

2.4.1 A Tabletop Pursuit-Evasion Game

Consider a game which is played around a table with four corners. One player, the pursuer, is attempting to "catch" the other player, the evader, in minimum time. The game evolves in stages. At every stage, the 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, 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 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 $p \in (0,1)$. (The evader will fail to move with probability $1-p$.) Thus, the pursuer can ultimately catch the evader as long as 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.

We now describe the game in the notation for stochastic shortest path games. There are three states: evader caught (state Ω), players adjacent to one another (state 1), and players across from one another (state 2). Thus, $S = \{1,2\} \cup \{\Omega\}$. In state one, 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, $A(1) = B(1) = \{1,2,3\}$. From our earlier description, it's not hard to see that

$$p_{1\Omega}(u,v) = u_1[(v_1+v_3)(1-p)+v_2] + u_2 v_1 p,$$

$$
\begin{aligned}
p_{11}(u,v) &= (u_1+u_3)(v_1+v_3)p + u_2[(v_1+v_3)(1-p)+v_2], \\
p_{12}(u,v) &= u_2v_3p + u_3[(v_1+v_3)(1-p)+v_2].
\end{aligned}
$$

In state two (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, $A(2) = B(2) = \{1,2,3\}$. It's not hard to see that

$$
\begin{aligned}
p_{2\Omega}(u,v) &= u_1v_3p + u_3v_1p, \\
p_{21}(u,v) &= (u_1+u_3)[(v_1+v_3)(1-p)+v_2] + u_2(v_1+v_3)p, \\
p_{22}(u,v) &= u_1v_1p + u_2[(v_1+v_3)(1-p)+v_2] + u_3v_3p.
\end{aligned}
$$

The transition costs reflect time away from termination, so $c_1(u,v) = 1$ for all $u \in U(1)$, and $v \in V(1)$ and $c_2(u,v) = 1$ for all $u \in U(2)$ and $v \in V(2)$.

2.4.2 An Industrial-waste Inspection Game

In this section we describe a more practical game of pursuit and evasion. The players in this game are

1. a manufacturer who produces industrial waste which has to be dumped (illegally) every night and

2. an inspector who seeks to detect the manufacturer dumping two nights in a row in an effort to put the manufacturer out of business.

The game we describe is similar to one Filar and Vrieze used as an example in [20]. It is different in that our cost structure reflects a true stochastic shortest path game and the dynamics of the game are augmented so that state transitions are not governed by a single player.

Let us suppose there is a finite number of geographically disparate sites where industrial waste can be dumped. The manufacturer must dump waste at one of these

sites every night while avoiding detection by the inspector. To detect dumping activity, the inspector must arrange to be at the same site as a dump and even then there is a nonzero probability of failure to detect. The probability of detection (conditional on the manufacturer and the inspector being at the same location) depends upon the following considerations.

1. The closer the current dump site is to the preceding dump site, the greater the probability of detection. (This is due to the environment's limited ability to absorb the waste.)

2. The closer the current inspection site is to the preceding inspection site, the greater the probability of detection. (The more time the inspector spends traveling and setting up equipment, the less time there is to look for dumping activity.)

If the inspector manages to detect the manufacturer two nights in a row, then, according to the state's environmental protection laws, the inspector can put the manufacturer out of business. The inspector's objective is to minimize the number of days to shut-down, while the manufacturer seeks to maximize it's time in business. In deciding where to go each day, the manufacturer and the inspector are both aware of where the last night's dump and inspection occurred. Moreover, they are both aware of whether the manufacturer got caught last night.

To give a mathematical description, let $L = \{s_1, \ldots, s_N\} \in \Re^2$ represent the sites where dumping may occur. Let $d(s, \bar{s})$ denote the Euclidean distance between the sites s and $\bar{s}$, and let $\bar{d} = \max_{s,\bar{s}\in L} d(s, \bar{s})$ be the maximum distance between any two sites. Let x_t denote the site where the inspector searched during stage $(t-1)$, let y_t be the site where dumping occurred in stage $(t-1)$, and let z_t be a boolean variable which is TRUE if the manufacturer was caught dumping during stage $(t-1)$. The triple (x_t, y_t, z_t) describes the state of the system at stage t. (There are $2N^2$ nonterminal states.) Suppose that the inspector chooses to search at site $a_t \in L$ and the manufacturer chooses (independently) to dump at site $b_t \in L$. Given that the manufacturer has not yet been shut down, the probability that the manufacturer will

be detected at stage t is defined to be

$$p(x_t, y_t, a_t, b_t) = \begin{cases} p_1 + \frac{p_2 - p_1}{d(k_1+k2)}[k_1 d(a_t, x_t) + k_2 d(b_t, y_t)] & \text{if } a_t = b_t, \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < p_2 < p_1 < 1$ are worst-case and ideal probabilities of detection and k_1 and k_2 are positive weighting factors. If z_t = FALSE, then the system transitions to state (a_t, b_t, TRUE) with probability $p(x_t, y_t, a_t, b_t)$; otherwise the system transitions to (a_t, b_t, FALSE). If z_t = TRUE, then the game terminates with probability $p(x_t, y_t, a_t, b_t)$; otherwise the system transitions to (a_t, b_t, FALSE). The cost of transitions from all nonterminal states is one, regardless of the controls applied. By allowing the players to randomize their actions we obtain a stochastic shortest path game of the type described in Section 2.3.4. Note that any policy for the inspector which is pure (i.e. not random) is improper because, knowing this policy, the manufacturer can always arrange to avoid being detected. On the other hand, the random policy which selects sites for the inspector uniformly will eventually result in the manufacturer getting caught, so there exists a proper policy. The essence of the inspector's problem is to choose a set of probability distributions to minimize the worst-case expected pursuit time.

2.5 Chapter Summary

The purpose of this chapter was to formulate stochastic shortest path games. We observed that, thanks to Assumption SSP, stochastic shortest path games generalize the terminating games originally considered by Shapley. Next, we showed that the regularity conditions encompassed by Assumption R are consistent with a variety of important dynamic games, including stochastic (dynamic) games in mixed strategies, sequential games, and symmetric sequential games. Following these general remarks, we described two concrete examples of stochastic shortest path games.

Chapter 3

Existence and Characterization of Equilibria

A fundamental question about stochastic shortest path games is whether they have "value." That is, given a stochastic shortest path game, is it true that

$$\inf_{\pi_M \in \bar{M}} \sup_{\pi_N \in \bar{N}} J_{\pi_M,\pi_N}(i) = \sup_{\pi_N \in \bar{N}} \inf_{\pi_M \in \bar{M}} J_{\pi_M,\pi_N}(i), \quad i = 1, \ldots, n?$$

If equality holds for all i, then the game is said to have value in nonstationary policies and the common quantity for each i is called the value of the game from state i. If equality prevails when $\bar{M}$ and $\bar{N}$ are replaced by M and N respectively, then the game is said to have value in stationary policies. The existence of value is a fundamental question in zero-sum games. A priori, we can only be sure that

$$\inf_{\pi_M \in \bar{M}} \sup_{\pi_N \in \bar{N}} J_{\pi_M,\pi_N}(i) \geq \sup_{\pi_N \in \bar{N}} \inf_{\pi_M \in \bar{M}} J_{\pi_M,\pi_N}(i), \quad i = 1, \ldots, n,$$

and strict equality is crucially dependent on $J_{\cdot,\cdot}(i)$ and the sets $\bar{M}$ and $\bar{N}$.

Now suppose for the moment that a stochastic shortest path game has value in nonstationary policies. This by itself is not enough to assure the existence of policies

$\pi_M^* \in \bar{M}$ and $\pi_N^* \in \bar{N}$ such that

$$\begin{aligned} \inf_{\pi_M \in \bar{M}} \sup_{\pi_N \in \bar{N}} J_{\pi_M,\pi_N}(i) &= \sup_{\pi_N \in \bar{N}} J_{\pi_M^*,\pi_N}(i), \quad i = 1,\ldots,n, \\ \sup_{\pi_N \in \bar{N}} \inf_{\pi_M \in \bar{M}} J_{\pi_M,\pi_N}(i) &= \inf_{\pi_M \in \bar{M}} J_{\pi_M,\pi_N^*}(i), \quad i = 1,\ldots,n. \end{aligned} \tag{3.1}$$

A pair of such policies (π_M^*, π_N^*) is called an *equilibrium solution* of the game, and π_M^* (respectively π_N^*) is called an equilibrium solution for the minimizer (respectively maximizer). If an equilibrium solution exists, then the game is said to have a *equilibrium* (also called a Nash equilibrium), and $J_{\pi_M^*,\pi_N^*}$ is called the *equilibrium value function* of the game. Note that by definition, an equilibrium solution (π_M^*, π_N^*) satisfies

$$\begin{aligned} J_{\pi_M,\pi_N^*} &\geq J_{\pi_M^*,\pi_N^*}, \qquad \forall\ \pi_M \in \bar{M}, \\ J_{\pi_M^*,\pi_N} &\leq J_{\pi_M^*,\pi_N^*} J, \qquad \forall\ \pi_N \in \bar{N}, \end{aligned}$$

where, as usual, the inequalities are taken componentwise. Notice that if a game has a value but does not have an equilibrium solution, then the players can only choose policies which are arbitrarily close to achieving the value of the game, and the game is said to have an *ϵ-equilibrium*. A policy (or pair of policies) which comes to within ϵ of the value of the game at each state is called an *ϵ-equilibrium solution*.

The main result of this chapter is that all stochastic shortest path games have stationary equilibrium solutions (and therefore have a value in stationary policies). Another main result is that the equilibrium value function of a stochastic shortest path game can be characterized as the unique solution to a functional equation, known as Bellman's equation.

The results of this chapter generalize Shapley's earlier results for terminating stochastic games. Shapley's analysis is relatively straightforward because the corresponding dynamic programming operator is a contraction mapping. In the more general class of stochastic shortest path games, the corresponding dynamic programming operator loses this contraction property, and a much more sophisticated analysis is required. The results of the present chapter also generalize the findings of Kushner

and Chamberlain [28] and Filar and Vrieze [20] (regarding transient games).

In general, questions about the existence of value and the characterization of equilibria are difficult to answer. Maitra and Parthasarathy [32, 33], Parthasarathy [38], and Kumar and Shiau [27] analyzed stochastic games with mild assumptions about the state space and control constraint sets. In the most recent of these papers, Kumar and Shiau established the existence of an extended-real value in games with nonnegative cost when the players are allowed to use behavioral strategies (where the mixed actions chosen at each stage can depend on all of the past states and controls, as well as the current state). They showed that the minimizing player can always achieve this value using a stationary mixed strategy. They also showed that if the state space is finite, then the maximizing player has ϵ-equilibrium solutions in the set of stationary mixed strategies. Because of our assumptions relating to termination, the results of this chapter are stronger than the conclusions of Kumar and Shiau [27] for games whose state-spaces and underlying action sets are finite.

Assumptions R and SSP are standing assumptions throughout this chapter; they hold unless specifically stated otherwise.

3.1 Preliminaries

We collect here a few lemmas which will be useful in establishing the main results of the chapter. Most of the proofs of these lemmas are deferred to Appendix A.

3.1.1 A Contraction Property

Given a vector $w \in \Re^n$ whose elements are positive, the corresponding weighted sup-norm, denoted $\|\cdot\|_\infty^w$, is defined for all $J \in \mathcal{J}$ by

$$\|J\|_\infty^w = \max_{i=1,\ldots,n} |J(i)/w_i| .$$

The following lemma states that when all stationary policies for the minimizer are proper, the operator T is a contraction with respect to a weighted sup-norm.

Lemma 3.1 *Assume that all stationary policies for the minimizer are proper. Then, there exists a positive vector* $w \in \Re^n$ *and a scalar* $\beta < 1$ *such that* $\|TJ - T\bar{J}\|_\infty^w \leq \beta\|J - \bar{J}\|_\infty^w$ *for all* J, $\bar{J} \in \mathcal{J}$. *Moreover, for all* $\mu \in M$ *(all of which are proper) and* $\nu \in N$, *the operators* T_μ, $\tilde{T}_\nu$, *and* $T_{\mu,\nu}$ *are contraction mappings with respect to the same weighted sup-norm, and the same contraction modulus* β *applies. The vector of weights* w *can be scaled by a positive factor without affecting the contraction modulus.*

Proof: First we will establish the result for the operator T. Our strategy is to identify a vector of weights w and to show that this set of weights is such that T is a contraction with respect to $\|\cdot\|_\infty^w$.

We begin by defining a new single-player stochastic shortest path problem (of the type considered by Bertsekas and Tsitsiklis in [8][1]). This problem is defined by setting all of the transition costs from $i = 1, \ldots, n$ to -1 and allowing the two players (in the original game) to work together to minimize cost. That is, we set $c_i(u, v) = -1$ for all $i \in S$ and $(u, v) \in U(i) \times V(i)$, and we make $\bar{M} \times \bar{N}$ the decision space for the single player of the new problem. We leave all of the transition probabilities unchanged. The result is a stochastic shortest path problem where all stationary policies are proper. Using the results of [8], there is an optimal cost vector $\tilde{J} \in \mathcal{J}$ 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 nonterminal states are set to -1, we have $\tilde{J} \leq -\mathbf{1}$. Moreover, from Bellman's equation we have

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

Also, for all $\mu \in M$ and $\nu \in N$,

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

[1] In [8], Bertsekas and Tsitsiklis studied single-player stochastic shortest path problems. They imposed two main assumptions which are equivalent to our Assumptions R and SSP (when restricted to a single player). They proved the existence of optimal stationary policies and characterized the optimal value function as the unique solution to Bellman's equation. They also proved the convergence of value iteration and policy iteration.

Thus, for all $\mu \in M$, $\nu \in N$, and for all $i = 1, \ldots, n$,

$$\begin{aligned} \sum_{j=1}^{n} -p_{ij}(\mu(i), \nu(i))\tilde{J}(j) &\leq -\tilde{J}(i) - 1 \\ &\leq -\tilde{J}(i)\beta, \end{aligned} \tag{3.2}$$

where

$$\beta \stackrel{\triangle}{=} \max_{i=1,\ldots,n} (\tilde{J}(i) + 1)/\tilde{J}(i).$$

Notice that $\beta \in [0, 1)$ since $\tilde{J} \leq -\mathbf{1}$. Also, $w \stackrel{\triangle}{=} -\tilde{J}$ is strictly positive.

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

$$\begin{aligned} T\bar{J} - TJ &= T\bar{J} - T_\mu J \\ &\leq T_\mu \bar{J} - T_\mu J \\ &= T_{\mu\nu}\bar{J} - T_\mu J \\ &\leq T_{\mu\nu}\bar{J} - T_{\mu\nu}J. \end{aligned}$$

Thus, $T\bar{J} - TJ \leq P(\mu, \nu) \cdot (\bar{J} - J)$ and for all $i = 1, \ldots, n$

$$\begin{aligned} \frac{(T\bar{J})(i) - (TJ)(i)}{cw(i)} &\leq \frac{1}{cw_i} \sum_{j=1}^{n} p_{ij}(\mu(i), \nu(i))(\bar{J}(j) - J(j)) \\ &\leq \frac{1}{w(i)} \sum_{j=1}^{n} p_{ij}(\mu(i), \nu(i))w(j) \\ &= \frac{1}{-\tilde{J}(i)} \sum_{j=1}^{n} p_{ij}(\mu(i), \nu(i))(-\tilde{J}(j)) \\ &\leq \frac{1}{-\tilde{J}(i)}(-\tilde{J}(i))\beta = \beta. \end{aligned}$$

(The last inequality follows from (3.2).) Thus, we get

$$\frac{(T\bar{J})(i) - (TJ)(i)}{w(i)} \leq c\beta, \quad \forall\, i = 1, \ldots, n.$$

Similarly, we can show that

$$\frac{(T\bar{J})(i) - (TJ)(i)}{w(i)} \geq -c\beta, \quad \forall\, i = 1, \ldots, n.$$

Combining these inequalities, we see that $\|TJ - T\bar{J}\|_\infty^w \leq c\beta$. This proves the result for the operator T.

The reasoning of the preceding paragraph also applies to the other "T operators." For example, suppose μ is a stationary policy for the minimizer. Given J and $\bar{J}$ such that $\|J - \bar{J}\|_\infty^w = c$, choose $\nu \in M$ so that $T_\mu \bar{J} = T_{\mu\nu}\bar{J}$. Then, $T_\mu \bar{J} - T_\mu J \leq T_{\mu\nu}\bar{J} - T_{\mu\nu}J$, and it follows that

$$\frac{(T_\mu\bar{J})(i) - (T_\mu J)(i)}{w(i)} \leq c\beta, \quad \forall\, i = 1, \ldots, n.$$

Similar arguments show that

$$\frac{(T_\mu\bar{J})(i) - (T_\mu J)(i)}{w(i)} \geq -c\beta, \quad \forall\, i = 1, \ldots, n,$$

which proves the result for T_μ. (This does not affect the contraction modulus.)

The last statement of the proposition follows from $\|\cdot\|_\infty^{\alpha w} = (1/\alpha)\|\cdot\|_\infty^w$, with $\alpha > 0$. **Q.E.D.**

Remark 1: This proof uses a style of argument which is common throughout this thesis. In general, we shall make heavy use of the theory that already exists for the single-player case.

Remark 2: The modulus of contraction β is equal to $1 - 1/N_{max}$, where N_{max} is the largest expected time to termination given that both players work together to prolong the game.

Corollary 3.1 *If μ is proper, then T_μ is a contraction with respect to a weighted sup-norm.*

Proof: Apply Lemma 3.1 to a new stochastic shortest path game where for each i the control constraint set $U(i)$ is set equal to $\{\mu(i)\}$. **Q.E.D.**

Remark: In [28], Kushner and Chamberlain describe a class of terminating games more general than Shapley's in [52]. In particular, they assumed that either of the following are true.

1. The n-stage probability transition matrix $[P(\mu,\nu)]^n$ is a "uniform contraction." (That is, for some $\epsilon > 0$ and all pairs (μ,ν), the matrix $[P(\mu,\nu)]^n$ has row-sums less than $1-\epsilon$.)

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 with respect to $\nu \in N$.

In light of Lemma 3.1, the first alternative would be true in the special case of stochastic shortest path games where all stationary policies of the minimizer are proper. The second alternative above is very close to our statement of Assumption SSP but is slightly stronger in that it requires that the costs to the minimizer be strictly positive and bounded away from zero. (Under Assumption SSP, $c_i(u,v)$ can take on negative values for some i, u, and v.)

3.1.2 On Fixing a Policy for One of the Players

Given a policy π_M for the minimizer, we use J_{π_M} to denote the worst-case cost of π_M, defined by

$$J_{\pi_M}(i) = \liminf_{t\to\infty} \max_{\pi_N \in \bar{N}} h^t_{\pi_M,\pi_N}(i). \tag{3.3}$$

Appendix A shows that thanks to Assumption R the maximum in (3.3) is attained for every t (see Lemma A.6). The cost of a stationary policy μ for the minimizer is denoted J_μ and is computed according to (3.3) where $\pi_M = \{\mu,\mu,\ldots\}$.

Similarly, given a policy π_N for the maximizer, we use J_{π_N} to denote the worst-case cost of π_N, defined by

$$J_{\pi_N}(i) = \liminf_{t\to\infty} \min_{\pi_N \in \bar{M}} h^t_{\pi_M,\pi_N}(i). \tag{3.4}$$

As before, Assumption R assures that the minimum above is attained for every t. The cost of a stationary policy ν for the maximizer is denoted J_ν and is computed according to (3.4) where $\pi_N = \{\nu, \nu, \ldots\}$.

Lemma A.9 *Given a proper policy μ, the following are true.*

1. *J_μ is the unique fixed point of T_μ within $\mathcal{J}$.*
2. *$J_\mu = \sup_{\pi_N \in \bar{N}} J_{\mu,\pi_N}$.*
3. *$T^t_\mu J \to J_\mu$ for all $J \in \mathcal{J}$, with linear convergence.*

Lemma A.10 *For any $\nu \in N$, the following are true.*

1. *J_ν is the unique fixed point of $\tilde{T}_\nu$ within $\mathcal{J}$.*
2. *$J_\nu = \inf_{\pi_M \in \bar{M}} J_{\pi_M,\nu}$.*
3. *$\tilde{T}^t_\nu J \to J_\nu$ for all $J \in \mathcal{J}$. If for all $\mu \in M$, the pair (μ, ν) terminates with probability one, then the convergence is linear.*

3.1.3 Testing for Proper Policies

The following lemma provides a useful test for determining whether a stationary policy is proper.

Lemma A.11 *Given $\mu \in M$, if there exists $J \in \mathcal{J}$ such that $J \geq T_\mu J$, then μ is proper.*

3.2 Main Results

In this section we establish the existence of a unique equilibrium value for stochastic shortest path games. Central to the proof is the fact that the operator T has a unique fixed point in $\mathcal{J}$. This fact is obvious for the case that all stationary policies for the minimizer are proper (cf. Lemma 3.1 and the contraction mapping theorem). When there exists an improper policy a more sophisticated argument is required.

Proposition 3.1 *The operator T has a unique fixed point J^* in $\mathcal{J}$.*

Proof: We begin by showing that T has at most one fixed point in $\mathcal{J}$. Suppose J and $\bar{J}$ are both fixed points of T in $\mathcal{J}$. By Assumption R, we can select $\mu \in M$ and $\bar{\mu} \in M$ such that $J = TJ = T_\mu J$ and $\bar{J} = T\bar{J} = T_{\bar{\mu}}\bar{J}$. By Lemma A.11, we have that μ and $\bar{\mu}$ are proper. Lemma A.9 implies that $J = J_\mu$ and $\bar{J} = J_{\bar{\mu}}$. From the definition of T, $J = TJ \leq T_{\bar{\mu}}J$. By induction, using the monotonicity of $T_{\bar{\mu}}$, we obtain $J = T^t J \leq T^t_{\bar{\mu}}J$ for all $t > 0$. Thus, by Lemma A.9, we have that $J \leq \lim_{t\to\infty} T^t_{\bar{\mu}}J = J_{\bar{\mu}} = \bar{J}$. Similar arguments show that $\bar{J} \leq J$. Thus, $J = \bar{J}$ and T has at most one fixed point in $\mathcal{J}$.

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 A.9, we have that $J_\mu = T_\mu J_\mu$. Thus, $J_\mu \geq TJ_\mu$. Similarly, by fixing a stationary policy $\nu \in N$ for the maximizer, we obtain from Lemma A.10 that $J_\nu = \tilde{T}_\nu J_\nu$. Thus, $J_\nu \leq \tilde{T}J_\nu = TJ_\nu$. We now claim that $J_\nu \leq J_\mu$. To see this, note that, for every $\pi_M \in \bar{M}$, $\pi_N \in \bar{N}$, and $t > 0$,

$$h^t_{\pi_M,\pi_N} \leq \max_{\tilde{\pi}_N \in \bar{N}} h^t_{\pi_M,\tilde{\pi}_N},$$

where $h^t_{\pi_M,\pi_N}$ is the vector whose components are defined in (2.2). We also have that

$$h^t_{\pi_M,\pi_N} \geq \min_{\tilde{\pi}_M \in \bar{M}} h^t_{\tilde{\pi}_M,\pi_N}.$$

Thus, for any $\pi_N \in \bar{N}$ and for any $\pi_M \in \bar{M}$

$$\min_{\tilde{\pi}_M \in \bar{M}} h^t_{\tilde{\pi}_M, \pi_N} \leq \max_{\tilde{\pi}_N \in \bar{N}} h^t_{\pi_M, \tilde{\pi}_N}.$$

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

Using the monotonicity of T we have that

$$J_\nu \leq TJ_\nu \leq TJ_\mu \leq J_\mu.$$

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

$$J_\nu \leq T^{t-1}J_\nu \leq T^t J_\nu \leq J_\mu.$$

Thus, the sequence $\{T^t J_\nu\}$ converges to a vector $J^\infty \in \mathcal{J}$. From the continuity of T, we have that $J^\infty = TJ^\infty$. Thus, T has a fixed point in $\mathcal{J}$. **Q.E.D.**

Remark: Now that we have established the existence of a unique fixed point of T, it is relatively easy to prove that the fixed point is the equilibrium value function of the game.

Proposition 3.2 *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 $J \in \mathcal{J}$, $\mu \in M$, 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. $J_{\mu,\pi_N} \leq J_{\mu,\nu}, \quad \forall\, \pi_N \in \bar{N}$.

Proof: That there exists a unique fixed point $J^* = TJ^*$ follows from the Proposition 3.1. Let $\mu^* \in M$ be such that $J^* = T_{\mu^*}J^*$, and let $\nu^* \in N$ be such that

$J^* = \tilde{T}_{\nu^*} J^*$. (Such policies exist thanks to Assumption R.) By Lemma A.11, we have that μ^* is proper. Thus, by Lemma A.9, we have that $J^* = J_{\mu^*} = \sup_{\pi_N \in \bar{N}} J_{\mu^*,\pi_N}$. Similarly, by Lemma A.10, we have that $J^* = \tilde{J}_{\nu^*} = \inf_{\pi_M \in \bar{M}} J_{\pi_M,\nu^*}$. Combining these results we obtain

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

Since in general we have

$$\inf_{\pi_M \in \bar{M}} \sup_{\pi_N \in \bar{N}} J_{\pi_M,\pi_N} \geq \sup_{\pi_N \in \bar{N}} \inf_{\pi_M \in \bar{M}} J_{\pi_M,\pi_N}$$

(a statement of the minimax inequality), we obtain the desired result:

$$\inf_{\pi_M \in \bar{M}} \sup_{\pi_N \in \bar{N}} J_{\pi_M,\pi_N} = J^* = \sup_{\pi_N \in \bar{N}} \inf_{\pi_M \in \bar{M}} J_{\pi_M,\pi_N}$$

Q.E.D.

3.3 Example: Tabletop Pursuit-Evasion

Recall the tabletop pursuit-evasion game from the preceding chapter. We will show that the equilibrium value function of this game is

$$J^* = \left(\frac{1}{1-p}, \frac{2-p}{1-p} \right)'. \tag{3.5}$$

Moreover, we will show that the mixed strategies μ^* and ν^* such that

$$\begin{aligned} \mu^*(1) &= (1,0,0)', & \mu^*(2) &= (u_1,0,u_3)', \\ \nu^*(1) &= (v_1,0,v_3)', & \nu^*(2) &= (0,1,0)', \end{aligned} \tag{3.6}$$

form an equilibrium solution to the game. Thus, in state one (where the players are adjacent),

1. any mixed decision for the evader is optimal, as long he never remains at his

current location, and

2. the pursuer's best option to move toward the evader with probability one.

On the other hand, in state two (where the players are across from one another),

1. any mixed decision for the pursuer is optimal as long as he never remains at his current location, and

2. the evader's best option is to stay at his current location with probability one.

When both players use this equilibrium solution, the game will transition from state 2 to 1 in a single stage. Note that this is reflected in the equilibrium cost function: $J^*(2) = \frac{2-p}{1-p} = 1 + J^*(1)$.

To verify that we have found an equilibrium solution, we will show that $J^* = TJ^* = T_{\mu^*}J^* = \tilde{T}_{\nu^*}J^*$. (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 Proposition 3.2.)

Let us first consider the case where the two players are adjacent (i.e. state 1). Let J denote a generic estimate of the equilibrium value function. (We shall soon consider the case where $J = J^*$, as suggested by (3.5).) To evaluate $(TJ)(1)$, we must compute

$$\min_{u \in U(1)} \max_{v \in V(1)} u' G_1(J) v,$$

where the matrix $G_1(J)$ is computed as

$$G_1(J) = \begin{bmatrix} 1 + pJ(1) & 1 & 1 + pJ(1) \\ 1 + (1-p)J(1) & 1 + J(1) & 1 + (1-p)J(1) + pJ(2) \\ 1 + pJ(1) + (1-p)J(2) & 1 + J(2) & 1 + pJ(1) + (1-p)J(2) \end{bmatrix}.$$

In other words, $(TJ)(1)$ is equal to the value of the matrix game $G_1(J)$. It is well known that matrix games have solutions via linear programming [62]. Thus,

$$\frac{1}{(TJ)(1)} = \begin{array}{l} \min \mathbf{1}'\breve{v} \\ \text{subject to } G_1(J)\breve{v} \geq \mathbf{1},\ \breve{v} \geq 0, \end{array}$$

$$\frac{v^*}{(TJ)(1)} \in \begin{array}{l} \arg\min \mathbf{1}'\breve{v} \\ \text{subject to } G_1(J)\breve{v} \geq \mathbf{1},\ \breve{v} \geq 0, \end{array}$$

where 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 corresponding dual problem characterizes equilibrium strategies u^* for the minimizer:

$$\frac{u^*}{(TJ)(1)} \in \begin{array}{l} \arg\max \mathbf{1}'\breve{u} \\ \text{subject to } G_1(J)'\breve{u} \leq \mathbf{1},\ \breve{u} \geq 0. \end{array}$$

Now consider $G_1(J^*)$. Using (3.6), define

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

It is straightforward to verify that $\breve{v}^*$ is feasible for the primal problem and that $\breve{u}^*$ is feasible for the dual problem. Moreover, the primal cost corresponding to $\breve{v}^*$ is exactly $1-p$, just as the dual value of $\breve{u}^*$ is also exactly $1-p$. Thus, we have found a primal/dual feasible pair for which the primal cost equals the dual value. According to the duality theorem of linear programming, $\breve{v}^*$ and $\breve{u}^*$ are primal/dual optimal, and the optimal values of the primal and dual problems equal $1-p$ which is exactly $\frac{1}{J^*(1)}$. This verifies that $J^*(1) = (TJ^*)(1)$ and that $\mu^*(1)$ and $\nu^*(1)$ form an equilibrium solution at state 1.

Let us now consider the case where the two players are across from one another (i.e. state 2). To evaluate $(TJ)(2)$ for general $J \in \mathcal{J}$, we must compute

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

where $G_2(J)$ is a matrix computed as

$$G_2(J) = \begin{bmatrix} 1+(1-p)J(1)+pJ(2) & 1+J(1) & 1+(1-p)J(1) \\ 1+pJ(1)+(1-p)J(2) & 1+J(2) & 1+pJ(1)+(1-p)J(2) \\ 1+(1-p)J(1) & 1+J(1) & 1+(1-p)J(1)+pJ(2) \end{bmatrix}.$$

Thus, $(TJ)(2)$ is equal to the value of the matrix game $G_2(J)$, which can be found via linear programming:

$$\begin{aligned} &\min \mathbf{1}'\breve{v} \\ &\text{subject to } G_2(J)\breve{v} \geq \mathbf{1},\ \breve{v} \geq 0, \end{aligned}$$

$$\begin{aligned} &\max \mathbf{1}'\breve{u} \\ &\text{subject to } G_2(J)'\breve{u} \leq \mathbf{1},\ \breve{u} \geq 0. \end{aligned}$$

Now consider $G_2(J^*)$. Using (3.6), define

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

Again, it is straightforward to verify that $\breve{v}$ and $\breve{u}$ form a feasible primal/dual pair, where the primal cost of $\breve{v}$ equals the dual value of $\breve{u}$. Thus, by the duality theorem, $\breve{v}$ and $\breve{u}$ are primal/dual optimal. This time the optimal cost works out to be $\frac{1-p}{2-p}$ which is exactly $\frac{1}{J^*(2)}$. This verifies that $J^*(2) = (TJ^*)(2)$ and that $\mu^*(2)$ and $\nu^*(2)$ form an equilibrium solution at state 2.

3.4 Chapter Summary

We have shown that stochastic shortest path games have value in stationary policies and that they have stationary equilibrium solutions. The equilibrium value J^* is the unique fixed point of Bellman's equation $TJ = J$.

Chapter 4

Dynamic Programming Algorithms

In this chapter we are concerned with dynamic programming algorithms for stochastic shortest path games. The most basic algorithm we consider is value iteration. We establish the convergence of this method despite the fact that the corresponding dynamic programming operator is not a contraction. This result generalizes the earlier convergence result of Shapley for terminating stochastic games. Another basic algorithm we consider is policy iteration, where a sequence of policies for the minimizer is generated based on worst-case evaluations of cost and corresponding policy improvements. Extending the earlier convergence result of Rao et al. [44], we use the existence of a unique fixed point of the dynamic programming operator to establish the convergence of this method. We also consider several variations on policy iteration including asynchronous policy iteration (which generalizes van der Wal's algorithm [59]), naive policy iteration (usually called the algorithm of Pollatschek and Avi-Itzhak [40]), and the modified Newton's method of Filar and Tolwinski [19]. After this long discussion about general algorithms, we show how several of the dynamic programming algorithms specialize to the case of symmetric sequential games. The chapter ends with an application of the general algorithms to the inspection game of Section 2.4.2.

As in Chapter 3, Assumptions R and SSP are standing assumptions for all of the results of this chapter.

4.1 General Algorithms

In this section we consider algorithms for the general class of games described in Chapter 2. Later, we will consider how some of these algorithms can be adapted to the special case of symmetric games.

4.1.1 Value Iteration

In [52], Shapley proposed an iterative algorithm for solving discounted cost stochastic games. The same algorithm may be applied in our present context and is called value iteration.

Algorithm 4.1.1 *(Value Iteration)*

1. *Choose an initial cost function* $J_0 \in \mathcal{J}$.

2. *Given* $J_{k-1} \in \mathcal{J}$, *compute* $J_k \in \mathcal{J}$ *as*

$$J_k = TJ_{k-1}.$$

In games with a discount factor and in games where termination is inevitable under all policies, the dynamic programming operator T is a contraction mapping, and this fact can be used (as Shapley did in [52]) to prove the convergence of value iteration. Unfortunately, in general stochastic shortest path games (where not all stationary policies of the minimizer are proper), the dynamic programming operator loses this contraction property. Despite this, it is possible to prove the convergence of value iteration, as in the following proposition.

Proposition 4.1 *For every* $J \in \mathcal{J}$, *there holds*

$$\lim_{k\to\infty} T^k J = J^*,$$

where J^* *is the unique equilibrium cost vector.*

Proof: The existence and uniqueness of a fixed point for T was established in Proposition 3.1. Let J^* be the unique fixed point, and let $\mu^* \in M$ (proper) be such that $TJ^* = T_{\mu^*}J^*$. Our objective is to show that $T^k J \to J^*$ for all $J \in \mathcal{J}$. Let δ be some positive scalar. Let J^δ be the unique vector in $\mathcal{J}$ satisfying $T_{\mu^*}J^\delta = J^\delta - \delta\mathbf{1}$. (Such a vector exists because μ^* is proper, making the operator $T_{\mu^*}(\cdot) + \delta\mathbf{1}$ a contraction.) Note that

$$\begin{aligned} J^\delta &= T_{\mu^*}J^\delta + \delta\mathbf{1} \\ &= \max_{\nu \in N}[c(\mu^*, \nu) + P(\mu^*, \nu)J^\delta] + \delta\mathbf{1} \\ &= \max_{\nu \in N}[c(\mu^*, \nu) + \delta\mathbf{1} + P(\mu^*, \nu)J^\delta]. \end{aligned}$$

Thus, J^δ is the worst-case cost of the fixed policy μ^* with the immediate transition cost vector $c(\mu^*, \cdot)$ replaced with $c(\mu^*, \cdot) + \delta\mathbf{1}$. We have that

$$J^\delta = T_{\mu^*}J^\delta + \delta\mathbf{1} \geq T_{\mu^*}J^\delta.$$

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

$$T^k_{\mu^*}J^\delta \leq J^\delta.$$

By taking the limit as $k \to \infty$, we see that $J_{\mu^*} \leq J^\delta$. (This is also implied by our interpretation of J^δ above.)

Now using the monotonicity of T and the fact that $J^* = J_{\mu^*}$, we get

$$J^* = TJ^* \leq TJ^\delta \leq T_{\mu^*}J^\delta = J^\delta - \delta\mathbf{1} \leq J^\delta,$$

Proceeding inductively, we get

$$J^* \leq T^k J^\delta \leq T^{k-1}J^\delta \leq J^\delta.$$

Hence, $\{T^k J^\delta\}$ is a monotonically decreasing sequence which is bounded below and

therefore converges to some $J^\infty \in \mathcal{J}$. By continuity of the operator T, we must have that $J^\infty = TJ^\infty$. By the uniqueness of the fixed point of T, we have that $J^\infty = J^*$.

We now examine the convergence of the operator T^k applied to $J^* - \delta\mathbf{1}$. Note that,

$$J^* - \delta\mathbf{1} = TJ^* - \delta\mathbf{1} \leq T(J^* - \delta\mathbf{1}) \leq TJ^* = J^*,$$

where the first inequality follows from the fact that $P(\mu,\nu)\mathbf{1} \leq \mathbf{1}$ for all $\mu \in M$ and $\nu \in N$ (see also Lemma A.2). Once again, the monotonicity of T prevails, implying that $T^k(J^* - \delta\mathbf{1})$ is monotonically increasing and bounded above. From the continuity of T we have that $\lim_{k\to\infty} T^k(J^* - \delta\mathbf{1}) = J^*$.

We saw earlier that $J^\delta = T_{\mu^*}J^\delta + \delta\mathbf{1}$ and that $J^\delta \geq J^*$. Then,

$$J^\delta = T_{\mu^*}J^\delta + \delta\mathbf{1} \geq T_{\mu^*}J^* + \delta\mathbf{1} = J^* + \delta\mathbf{1}.$$

Thus, for any $J \in \mathcal{J}$ we can find $\delta > 0$ such that $J^* - \delta\mathbf{1} \leq J \leq J^\delta$. By the monotonicity of T, we then have

$$T^k(J^* - \delta\mathbf{1}) \leq T^kJ \leq T^kJ^\delta, \quad \forall\, k \geq 1.$$

Taking limits, we see that $\lim_{k\to\infty} T^kJ = J^*$. **Q.E.D.**

Remark: In addition to generalizing Shapley's result [52], Proposition 4.1 also generalizes Kushner and Chamberlain's convergence result [28] since Assumption SSP does not require the costs $c(\mu,\nu)$ to be strictly positive and bounded away from $\mathbf{0}$.

4.1.2 Policy Iteration

In [25], Hoffman and Karp proposed an iteration in policy-space for solving average-cost stochastic games. A modification to their algorithm may be applied to stochastic shortest path games. We will refer to this algorithm as policy iteration.

Algorithm 4.1.2 *(Policy Iteration)*

1. *Choose an initial proper policy* $\mu_0 \in M$.

2. *Given* $\mu_{k-1} \in M$*:*

   (a) *(Policy Evaluation) Compute the unique fixed point* $J_{\mu_{k-1}} \in \mathcal{J}$ *of the* $T_{\mu_{k-1}}$ *operator.*

   (b) *(Policy Improvement) Compute* $\mu_k \in M$ *such that* $TJ_{\mu_{k-1}} = T_{\mu_k} J_{\mu_{k-1}}$.

It was shown in [44] by Rao et al. that policy iteration converges for discounted cost games in mixed strategies. We provide a similar result below for stochastic shortest path games.

Proposition 4.2 *Given a proper policy* $\mu_0 \in M$*, we have that*

$$J_{\mu_k} \to J^*$$

where J^* *is the unique equilibrium cost vector and* $\{\mu_k\}$ *is a sequence of policies generated by policy iteration.*

Proof: Choose $\mu_1 \in M$ such that $T_{\mu_1} J_{\mu_0} = TJ_{\mu_0}$. (Assumption SSP implies that an initial proper policy μ_0 exists.) We have $T_{\mu_1} J_{\mu_0} = TJ_{\mu_0} \leq T_{\mu_0} J_{\mu_0} = J_{\mu_0}$. By Lemma A.11, μ_1 is proper. By the monotonicity of T_{μ_1} and Lemma A.9, we have that for all t

$$J_{\mu_0} \geq TJ_{\mu_0} \geq T_{\mu_1}^{t-1} J_{\mu_0} \geq T_{\mu_1}^t J_{\mu_0}.$$

Thus,

$$J_{\mu_0} \geq TJ_{\mu_0} \geq \lim_{t\to\infty} T_{\mu_1}^t J_{\mu_0} = J_{\mu_1}.$$

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

$$J_{\mu_k} \geq TJ_{\mu_k} \geq J_{\mu_{k+1}} \geq J^*, \quad \forall\, k = 0, 1, \ldots . \tag{4.1}$$

Since $\{J_{\mu_k}\}$ is monotonically decreasing and bounded below by J^*, we have that the entire sequence converges to some vector J^∞. From (4.1) and the continuity of T, we have that $J^\infty = TJ^\infty$. Since J^* is the unique fixed point of T on $\mathcal{J}$, we have that $J_{\mu_k} \to J^*$. **Q.E.D.**

Remarks: Each step of policy iteration requires the maximizer to solve a dynamic programming problem where termination is inevitable under all policies. This is to be contrasted with conventional, one-player policy iteration, where evaluating a policy μ_k is equivalent to solving a system of linear equations. As a result, each policy evaluation step in policy iteration (for games) involves a significant amount of computation. Note that the dual form of policy iteration, where a sequence of policies for the maximizer is generated, also converges by the same argument.

4.1.3 Asynchronous Policy Iteration

In [66], Williams and Baird introduced an algorithm for single-player Markov decision problems which was revisited by Bertsekas and Tsitsiklis in [9]. The algorithm is called asynchronous policy iteration, and it has a direct extension to stochastic shortest path games, as presented below.

Algorithm 4.1.3 *(Asynchronous Policy Iteration)*

1. *Start with an initial estimate for the equilibrium cost function, $J_0 \in \mathcal{J}$, and an initial policy $\mu_0 \in M$.*

2. *Given (J_{k-1}, μ_{k-1}), select a subset S_k of the states, and compute (J_k, μ_k) by either*

   (a) updating the cost function estimate on S_k according to

$$J_k(i) = \begin{cases} (T_{\mu_{k-1}} J_{k-1})(i), & \text{if } i \in S_k, \\ J_{k-1}(i), & \text{otherwise,} \end{cases} \tag{4.2}$$

   while leaving the policy unchanged by setting $\mu_k = \mu_{k-1}$, or

(b) updating the policy on S_k according to

$$\mu_k(i) \in \begin{cases} \arg\min_{u \in U(i)} h_i(u), & \text{if } i \in S_k, \\ \{\mu_{k-1}(i)\}, & \text{otherwise,} \end{cases} \tag{4.3}$$

where

$$h_i(u) = \max_{v \in V(i)} \left(c_i(u,v) + \sum_{j \in S} p_{ij}(i,u,v,j) J_{k-1}(j) \right),$$

while leaving the cost function estimate unchanged by setting $J_k = J_{k-1}$.

Note that the value update rule [cf. (4.2)] corresponds to a single step of value iteration for the maximizer in computing the worst-case cost of the minimizer's policy. Also, note that in the policy update rule [cf. (4.3)] we used the symbol "$\in$" in place of an equals sign to indicate that there may not exist a unique minimum. It was shown in [66, 9] (for the case of a single player) that if J_0 is such that $T_{\mu_0} J_0 \leq J_0$, then the sequence of cost function estimates J_k converges to the optimal cost-to-go function J^*. The same result is also true for stochastic shortest path games, as shown in the following proposition.

Proposition 4.3 *Let (J_k, μ_k) be the sequence generated by asynchronous policy iteration. Assuming that*

1. *the updates in (4.2) and (4.3) are executed infinitely often for all states, and*
2. *the initial conditions (J_0, μ_0) are such that $T_{\mu_0} J_0 \leq J_0$;*

the cost functions J_k converge to J^* and all policies μ_k generated by the algorithm are proper.

Proof: The proof is essentially identical to the proof in [9]. The difference, of course, is that here the operator T involves a minimax operation (instead of just a minimization) and T_μ involves a maximization (instead of nothing at all). Since T and T_μ are monotonic and continuous, the same proof holds. We spell out the proof below for completeness.

First we claim that for all k, if $T_{\mu_k} J_k \leq J_k$, then

$$T_{\mu_{k+1}} J_{k+1} \leq J_{k+1} \leq J_k. \tag{4.4}$$

To see this, suppose that at iteration k we have $T_{\mu_k} J_k \leq J_k$. Consider the following two possibilities.

1. *The value update (4.2) is executed next.* Then we have

$$J_{k+1}(i) = (T_{\mu_k} J_k)(i) \leq J_k(i), \quad \text{if } i \in S_k, \tag{4.5}$$

and

$$J_{k+1}(i) = J_k(i), \quad \text{if } i \notin S_k, \tag{4.6}$$

so that $J_{k+1} \leq J_k$. From the monotonicity of T_{μ_k} and the fact that $\mu_{k+1} = \mu_k$, we have that

$$T_{\mu_{k+1}} J_{k+1} = T_{\mu_k} J_{k+1} \leq T_{\mu_k} J_k. \tag{4.7}$$

From (4.5) we have

$$(T_{\mu_k} J_k)(i) = J_{k+1}(i), \quad \text{if } i \in S_k,$$

while from (4.6) and the hypothesis that $T_{\mu_k} J_k \leq J_k$, we have that

$$(T_{\mu_k} J_k)(i) \leq J_k(i) = J_{k+1}(i), \quad \text{if } i \notin S_k.$$

These two relations imply that $T_{\mu_k} J_k \leq J_{k+1}$, which when coupled with (4.7), shows that $T_{\mu_{k+1}} J_{k+1} \leq J_{k+1}$, completing the proof of (4.4) for the case of a value update.

2. *The policy update (4.3) is executed next.* In this case, we have that $J_{k+1} = J_k$,

and using the hypothesis that $T_{\mu_k} J_k \leq J_k$, we obtain

$$\begin{aligned}(T_{\mu_{k+1}} J_{k+1})(i) = (T_{\mu_{k+1}} J_k)(i) = (T J_k)(i) \leq (T_{\mu_k} J_k)(i) \\ \leq J_k(i) = J_{k+1}(i), \quad \text{if } i \in S_k,\end{aligned} \tag{4.8}$$

and

$$\begin{aligned}(T_{\mu_{k+1}} J_{k+1})(i) = (T_{\mu_{k+1}} J_k)(i) = (T_{\mu_k} J_k)(i) \\ \leq J_k(i) = J_{k+1}(i), \quad \text{if } i \notin S_k,\end{aligned} \tag{4.9}$$

so that $T_{\mu_{k+1}} J_{k+1} \leq J_{k+1}$, and (4.4) is shown for the case of a policy update.

Equation (4.4) and the hypothesis that $T_{\mu_0} J_0 \leq J_0$ imply that

$$J_{k+1} \leq J_k, \quad T J_k \leq T_{\mu_k} J_k \leq J_k, \quad \forall\, k. \tag{4.10}$$

From this, Lemma A.11 implies that all policies μ_k generated by the algorithm are proper. Moreover, from the monotonicity of T, we also have $T^m J_k \leq J_k$ for all m, and by taking the limit as $m \to \infty$, we obtain $J^* \leq J_k$ for all k. From this equation and (4.10), we see that J_k converges to some limit $\bar{J} \in \mathcal{J}$ satisfying

$$T \bar{J} \leq \bar{J} \leq J_k, \quad \forall\, k. \tag{4.11}$$

Furthermore, from (4.7)-(4.9), we have that

$$T_{\mu_{k+1}} J_{k+1} \leq T_{\mu_k} J_k, \quad \forall\, k. \tag{4.12}$$

To form a contradiction, suppose there is a state i such that $(T\bar{J})(i) < \bar{J}(i)$. From the continuity of T, there exists an integer $\bar{k}$ such that for all $k \geq \bar{k}$ we have $(T J_k)(i) < \bar{J}(i)$. Let $k \geq \bar{k}$ be an iteration index such that the policy update (4.3) is executed for state i. Let k' be the first iteration index with $k' > k$ such that the value update (4.2) is executed for state i. Then

$$J_{k'+1}(i) \;\; = \;\; (T_{\mu_{k'}} J_{k'})(i)$$

$$
\begin{aligned}
&\leq (T_{\mu_{k+1}} J_{k+1})(i) \\
&\leq (T_{\mu_{k+1}} J_k)(i) \\
&= (T J_k)(i) \\
&< \bar{J}(i), \qquad (4.13)
\end{aligned}
$$

where the first equality follows from the value update (4.2), the first inequality follows from (4.12), the second inequality follows from the relation $J_{k+1} \leq J_k$ and the monotonicity of $T_{\mu_{k+1}}$, and the second equality follows from the policy update (4.3). The relation (4.13) contradicts (4.11). Thus we must have that $(T\bar{J})(i) = \bar{J}(i)$ for all i, which implies that $\bar{J} = J^*$ since J^* is the unique fixed point of T. **Q.E.D.**

Remark: In [59] van der Wal proposed the following algorithm for discounted cost stochastic games.

Algorithm 4.1.4 *(Generalized Policy Iteration [59])*[1]

1. *Choose an initial cost function $J_0 \in \mathcal{J}$ with the property that $TJ_0 \leq J_0$. Also, choose $\epsilon > 0$ and a positive integer m.*

2. *Given $J_{k-1} \in \mathcal{J}$,*

   (a) (Policy Improvement) Compute $\mu_k \in M$ such that $TJ_{k-1} = T_{\mu_k} J_{k-1}$.

   (b) (Policy evaluation) Compute $J_k = T_{\mu_k}^m J_{k-1}$.

3. *Stop when*

$$\phi_{k-1} - \psi_{k-1} \leq \epsilon(1-\alpha)/\alpha,$$

   where α is the discount factor and

$$\phi_{k-1} = \max_{i \in S}[(TJ_{k-1})(i) - J_{k-1}(i)] \qquad \psi_{k-1} = \min_{i \in S}[(TJ_{k-1})(i) - J_{k-1}(i)].$$

[1]For single-player Markov decision processes, this algorithm would be recognized as modified policy iteration (see [42, 4]).

We observe that generalized policy iteration is actually a special case of asynchronous policy iteration. In [59] it was shown that (for discounted cost games in mixed strategies) generalized policy iteration will terminate after a finite number of stages. Moreover it is shown that, given the terminal value function approximation $\bar{J} \in \mathcal{J}$, the policies $\bar{\mu} \in M$ and $\bar{\nu} \in N$ such that $T\bar{J} = T_{\bar{\mu}}\bar{J}$ and $\tilde{T}_{\bar{\nu}}\bar{J} = \tilde{T}\bar{J}$ form an ϵ-equilibrium policy pair.

4.1.4 Approximate Policy Iteration

We now consider what happens when approximations are used in the policy iteration algorithm.

Algorithm 4.1.5 *(Approximate Policy Iteration with Function Approximation)*

1. *Choose an initial proper policy* $\mu_0 \in M$.
2. *Given the policy* $\mu_{k-1} \in M$*:*
   (a) *(Approximate Policy Evaluation) Compute an approximation* $J(\cdot, r^k)$ *of* $J_{\mu_{k-1}}$.
   (b) *(Approximate Policy Improvement) Compute a proper policy* μ_k *such that* $TJ(\cdot, r^k) \approx T_{\mu_k}J(\cdot, r^k)$.

For single-player Markov decision problems there is a recent result due to Bertsekas and Tsitsiklis ([9], Proposition 6.3 on page 279) which states that

1. if the cost function approximations for the successive policies are accurate (in a specific mathematical sense) and
2. if the policy updates are computed accurately with respect to the approximations,

then the resulting policies will ultimately yield costs which are close to optimal. The bounds on limiting performance are proportional to the accuracy of the approximations. This type result guarantees that there are no hidden bugs in the methodology

which can lead a priori to poor performance. The following proposition provides an analogous result for stochastic shortest path games.

Proposition 4.4 *Let M_p denote the set of proper policies for the minimizer. Define*

$$\rho = \max_{i=1,\ldots,n} \sup_{\mu \in M_p} \sup_{\nu \in N} Prob(i_n \neq \Omega \mid i_0 = i, \mu, \nu).$$

Assume that $\rho < 1$. In addition, assume that in implementing approximate policy iteration the following are true

$$\|J(\cdot, r^{k+1}) - J_{\mu_k}\|_\infty \leq \epsilon, \quad k = 0, 1, \ldots, \tag{4.14}$$

$$\|(T_{\mu_{k+1}} J(\cdot, r^{k+1})) - (TJ(\cdot, r^{k+1}))\|_\infty \leq \delta, \quad k = 0, 1, \ldots, \tag{4.15}$$

where ϵ and δ are fixed positive scalars. If each policy μ_k generated by approximate policy iteration is proper, then

$$\limsup_{k \to \infty} \|J_{\mu_k} - J^*\|_\infty \leq \frac{n(1 - \rho + n)(\delta + 2\epsilon)}{(1 - \rho)^2}. \tag{4.16}$$

Remark: The assumption that $\rho < 1$ is satisfied whenever the set of proper policies M_p is compact. This is true, for example, in sequential games with finite constraint sets and in games where all stationary policies for the minimizer are proper. In the special case of stochastic shortest path games which reflect a discount factor $\alpha < 1$, the bound in (4.16) can be improved to $(\delta + 2\alpha\epsilon)/(1 - \alpha)^2$.

The following lemma will be helpful in proving Proposition 4.4.

Lemma 4.1 *Let $P = P(\mu, \nu)$, where $\mu \in M$ is proper and $\nu \in N$, and let c be a nonnegative scalar. Assume that ρ (as defined in the statement of Proposition 4.4) satisfies $\rho < 1$.*

1. *If a vector x satisfies $x \leq Px + c\mathbf{1}$, then*

$$x(i) \leq \frac{nc}{1 - \rho}, \quad \forall\, i = 1, \ldots, n.$$

2. If a sequence of vectors x_k satisfies $x_{k+1} \leq Px_k + c\mathbf{1}$ for all k, then

$$\limsup_{k\to\infty} x_k(i) \leq \frac{nc}{1-\rho}, \quad \forall\, i = 1, \ldots, n.$$

Proof: (a) Let $y(i) = \max\{0, x(i)\}$, $i = 1, \ldots, n$. Then, $x \leq Px + c\mathbf{1} \leq Py + c\mathbf{1}$, which together with the relation $0 \leq Py + c\mathbf{1}$, implies $y \leq Py + c\mathbf{1}$. We then have

$$y \leq P(Py + c\mathbf{1}) + c\mathbf{1} \leq P^2 y + 2c\mathbf{1}.$$

By repeating this process $n-1$ times, we have

$$y \leq P^n y + nc\mathbf{1}.$$

By the definition of ρ, we have

$$P^n y \leq \rho \left(\max_i y(i)\right) \mathbf{1},$$

and it follows that

$$\max_i y(i) \leq \rho \max_i y(i) + nc.$$

Hence, $x(i) \leq \max_i y(i) \leq nc/(1-\rho)$, as desired.

(b) Proceeding as in part (a), we obtain

$$\max_i y_{k+n}(i) \leq \rho \max_i y_k(i) + nc, \quad \forall\, k,$$

where $y_k(i) = \max\{0, x_k(i)\}$. Hence,

$$\limsup_{k\to\infty} \left(\max_i y_{k+n}(i)\right) \leq \rho \limsup_{k\to\infty} \left(\max_i y_k(i)\right) + nc,$$

and the result follows. **Q.E.D.**

Proof of Proposition 4.4: To simplify the notation, let us define J_k to be equal to $J(\cdot, r^{k+1})$ which is the approximation of the worst-case cost J_{μ_k} of the minimizer using the policy μ_k. From (4.14) and (4.15) we have for all k,

$$T_{\mu_{k+1}} J_{\mu_k} - \epsilon \mathbf{1} \leq T_{\mu_{k+1}} J_k \leq T J_k + \delta \mathbf{1}.$$

From (4.14), we have for all k,

$$T J_k \leq T J_{\mu_k} + \epsilon \mathbf{1}.$$

Combining these relations, we obtain for all k,

$$T_{\mu_{k+1}} J_{\mu_k} \leq T J_{\mu_k} + (\delta + 2\epsilon)\mathbf{1} \leq T_{\mu_k} J_{\mu_k} + (\delta + 2\epsilon)\mathbf{1}. \tag{4.17}$$

From (4.17) and the fact that $T_{\mu_k} J_{\mu_k} = J_{\mu_k}$, we have

$$T_{\mu_{k+1}} J_{\mu_k} \leq J_{\mu_k} + (\delta + 2\epsilon)\mathbf{1}.$$

By subtracting from this relation the equation $T_{\mu_{k+1}} J_{\mu_{k+1}} = J_{\mu_{k+1}}$, we obtain

$$T_{\mu_{k+1}} J_{\mu_k} - T_{\mu_{k+1}} J_{\mu_{k+1}} \leq J_{\mu_k} - J_{\mu_{k+1}} + (\delta + 2\epsilon)\mathbf{1}.$$

Let $\bar{\nu}$ achieve the maximum in $T_{\mu_{k+1}} J_{\mu_{k+1}}$. We obtain

$$J_{\mu_{k+1}} - J_{\mu_k} \leq P(\mu_{k+1}, \bar{\nu})(J_{\mu_{k+1}} - J_{\mu_k}) + (\delta + 2\epsilon)\mathbf{1}. \tag{4.18}$$

Define $\xi_k = \|J_{\mu_{k+1}} - J_{\mu_k}\|_\infty$. Then, from (4.18) and Lemma 4.1(a), we obtain

$$\xi_k \leq \frac{n(\delta + 2\epsilon)}{1 - \rho}. \tag{4.19}$$

Let μ^* be an equilibrium policy for the minimizer. Note that μ^* must be proper.

From (4.17), we have

$$\begin{aligned} T_{\mu_{k+1}} J_{\mu_k} &\leq T J_{\mu_k} + (\delta + 2\epsilon)\mathbf{1} \\ &\leq T_{\mu^*} J_{\mu_k} + (\delta + 2\epsilon)\mathbf{1} \\ &= T_{\mu^*} J_{\mu_k} - T_{\mu^*} J_{\mu^*} + J^* + (\delta + 2\epsilon)\mathbf{1} \\ &\leq P(\mu^*, \tilde{\nu})(J_{\mu_k} - J^*) + J^* + (\delta + 2\epsilon)\mathbf{1}, \end{aligned}$$

where $\tilde{\nu}$ achieves the maximum in $T_{\mu^*} J_{\mu_k}$. (We have used the fact that $T_{\mu^*} J_{\mu^*} = J_{\mu^*} = J^*$.) We also have that

$$T_{\mu_{k+1}} J_{\mu_k} = J_{\mu_{k+1}} + T_{\mu_{k+1}} J_{\mu_k} - T_{\mu_{k+1}} J_{\mu_{k+1}} \geq J_{\mu_{k+1}} + P(\mu_{k+1}, \breve{\nu})(J_{\mu_k} - J_{\mu_{k+1}}),$$

where $\breve{\nu}$ achieves the maximum in $T_{\mu_{k+1}} J_{\mu_{k+1}}$. Combining these relations, we obtain

$$J_{\mu_{k+1}} - J^* \leq P(\mu^*, \tilde{\nu})(J_{\mu_k} - J^*) + P(\mu_{k+1}, \breve{\nu})(J_{\mu_{k+1}} - J_{\mu_k}) + (\delta + 2\epsilon)\mathbf{1}.$$

From the definition of ξ_k, we get

$$\begin{aligned} J_{\mu_{k+1}} - J^* &\leq P(\mu^*, \tilde{\nu})(J_{\mu_k} - J^*) + \xi_k P(\mu_{k+1}, \breve{\nu})\mathbf{1} + (\delta + 2\epsilon)\mathbf{1} \\ &\leq P(\mu^*, \tilde{\nu})(J_{\mu_k} - J^*) + \xi_k \mathbf{1} + (\delta + 2\epsilon)\mathbf{1}. \end{aligned}$$

With (4.19), this implies

$$J_{\mu_{k+1}} - J^* \leq P(\mu^*, \tilde{\nu})(J_{\mu_k} - J^*) + \frac{(1 - \rho + n)(\delta + 2\epsilon)}{1 - \rho}\mathbf{1}.$$

From Lemma 4.1 we obtain the desired result:

$$\limsup_{k \to \infty} (J_{\mu_k}(i) - J^*(i)) \leq \frac{n(1 - \rho + n)(\delta + 2\epsilon)}{(1 - \rho)^2}, \quad \forall\, i \in S.$$

Q.E.D.

4.1.5 Naive Policy Iteration

In the preceding chapter we described the policy iteration method (Algorithm 4.1.2) for computing the equilibria of stochastic shortest path games. In policy iteration, proper policies for the minimizer are generated via a sequence of policy evaluations and policy improvements. In evaluating the policy μ_k it is necessary to compute the worst-case cost $J_{\mu_k} = \sup_{\pi_N \in \tilde{N}} J_{\mu_k,\pi_N}$ which is equivalent to solving a dynamic programming problem for the maximizer (with the minimizer's policy fixed). Given the "complexity" of this method, we are led to examine other related algorithms which are at least conceptually easier to implement. In this and the following section we will consider two such algorithms: naive policy iteration and modified Newton's method

To introduce naive policy iteration, we cite the work of Pollatschek and Avi-Itzhak [40]. They proposed an algorithm which is similar to policy iteration in that it proceeds as a sequence of policy evaluations and improvements. The key difference is that instead of generating a sequence of policies just for the minimizer, a sequence of *pairs* of policies is generated (one for each player). As a result, policy evaluation only requires computing the expected additive cost for a Markov reward process.

Algorithm 4.1.6 *(Naive Policy Iteration [40])* [2]

1. *Choose an initial proper policy $\mu_0 \in M$ and choose an initial stationary policy for the maximizer $\nu_0 \in N$. (Alternatively, start with an initial guess $J \in \mathcal{J}$ for the equilibrium cost function and skip to step 2.(b), setting $J_{\mu_{-1},\nu_{-1}}$ equal to J.)*

2. *Given $\mu_{k-1} \in M$ and $\nu_{k-1} \in N$:*

   (a) *(Policy Evaluation) Compute $J_{\mu_{k-1},\nu_{k-1}}$.*

   (b) *(Policy Improvement)*

      i. *Compute $\mu_k \in M$ such that $TJ_{\mu_{k-1},\nu_{k-1}} = T_{\mu_k} J_{\mu_{k-1},\nu_{k-1}}$.*

      ii. *Compute $\nu_k \in N$ such that $\tilde{T}J_{\mu_{k-1},\nu_{k-1}} = \tilde{T}_{\nu_k} J_{\mu_{k-1},\nu_{k-1}}$.*

[2] This algorithm is often called the "Algorithm of Pollatschek and Avi-Itzhak."

Pollatschek and Avi-Itzhak derived sufficient conditions [40] which guarantee that the algorithm converges for any initial pair of stationary policies in discounted cost games. Unfortunately, these conditions are very restrictive, and it is not clear how they can be improved. In their numerical examples (none of which satisfied the sufficient conditions), the algorithm converged very quickly to the equilibrium. This led many researchers to conjecture (and even publish incorrect proofs, as in Rao et al. [44]) that the algorithm is globally convergent. However, van der Wal in [59] presented an example which shows that this is untrue.

Naive policy iteration has an extra failure-mode when applied to stochastic shortest path games: it may generate at some point a pair of stationary policies (μ, ν) which is not terminating with probability one. When this happens, the cost to the minimizer $J_{\mu,\nu}(i)$ is infinite for at least one initial state $i \in S$, and there is no basis for the algorithm to continue.

Interpretation as Newton's Method

In their original analysis [40], Pollatschek and Avi-Itzhak used a geometric interpretation of their algorithm as Newton's method. This interpretation, reproduced here, is useful for two reasons:

1. It provides insight into why naive policy iteration sometimes fails. (In particular, we'll see how van der Wal's example works.)

2. It suggests a technique for improving naive policy iteration, resulting in "modified Newton's method," to be discussed in the following subsection.

Let us formally define the Bellman error function $\Psi : \mathcal{J} \mapsto \mathcal{J}$ as

$$\Psi(J) = TJ - J.$$

Pure Newton's method for finding the roots of $\Psi(J)$ is given by the recursion:

$$J := J - [\nabla \Psi(J)']^{-1} \Psi(J). \tag{4.20}$$

Suppose $\mu \in M$ and $\nu \in N$ are such that $TJ = T_\mu J$ and $\tilde{T}J = \tilde{T}_\nu J$. Then,

$$\Psi(J) = c(\mu, \nu) + (P(\mu, \nu) - I_n)\, J,$$

where I_n is the $n \times n$ identity matrix. Wherever the minimax solution (μ, ν) in the evaluation of TJ is unique, Ψ is differentiable and the gradient may be computed as

$$\nabla\Psi(J) = [P(\mu, \nu) - I_n]'.$$

Using the matrix inversion lemma, it can be shown [40] that the cost $J_{\mu,\nu}$ associated with the pair (μ, ν) equals the right hand side of (4.20).

Van der Wal's Counter-example

We describe here van der Wal's example which shows that naive policy iteration is not a globally convergent algorithm. Consider the game shown in Figure 4-1. This is a discounted cost game with two states. While in state 1, each player has two control options. In order for there to be a nonzero probability of transitioning to state 2, the minimizer must implement action 2, and the maximizer must implement action 1. Assumption SSP is satisfied in this game because of the discount factor $\alpha = 3/4$. The transition costs $g(u, v)$ are shown in the figure.

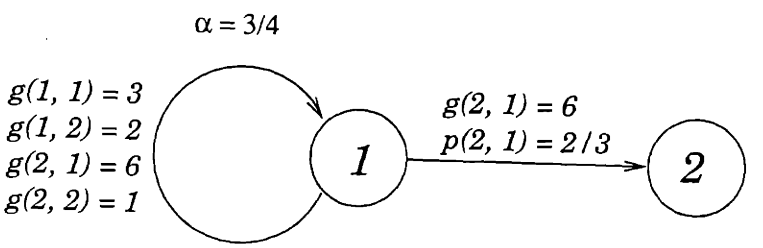

Figure 4-1: The example of van der Wal.

It turns out that this game has an equilibrium in pure strategies. Clearly, the equilibrium cost from state 2 is zero, so the only unknown is the equilibrium cost from the state 1. As shown in Figure 4-2, the policies which achieve T and $\tilde{T}$ are also pure for all estimates of the equilibrium cost-to-go. In Figure 4-3, we graph the

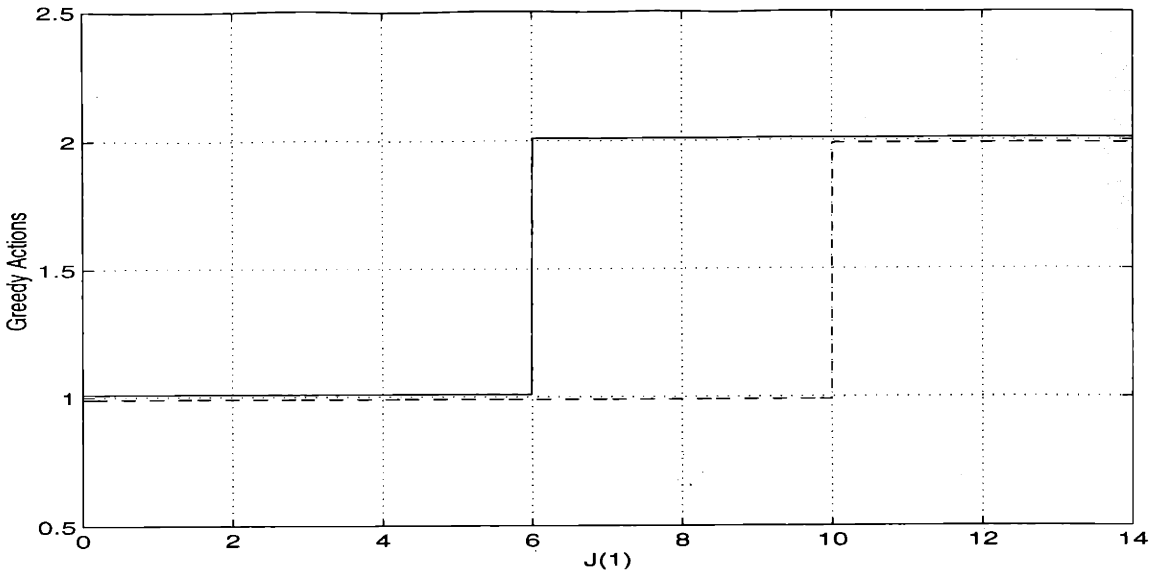

Figure 4-2: Greedy policies for van der Wal's example plotted as a function of estimates $J(1)$ of the equilibrium cost-to-go from state 1. The solid line represents the greedy action for the minimizer, and the dashed line represents the greedy action for the maximizer.

Bellman error Ψ evaluated at state 1 as a function of estimates $J(1)$ of the equilibrium cost-to-go from 1. We note that there two points where Ψ is not differentiable. These are points for which either pure strategy is optimal for one of the players. In the discussion which follows, we will constrain ourselves to $J(2) = 0$. (Naive policy iteration will only produce estimates of the equilibrium cost within this subspace.) Since Ψ here is effectively a function of a single variable, Newton's method can easily be described geometrically. Given an estimate of the equilibrium cost $J(1)$, draw the line tangent to Ψ at $J(1)$. The point where the tangent line intersects the $J(1)$-axis is the corresponding Newton step. Thus, the Figure 4-3 tells us everything we need to know about van der Wal's example. For initial estimates $J(1)$ in the set $(-\infty, 6) \cup (10, \infty)$, succeeding iterates of naive policy iteration will oscillate between 4 and 12. For initial estimates in $(6, 10)$, the method converges to the equilibrium in a single step. What results when $J(1) \in \{6, 10\}$ depends on which pair of policies is selected.

Apparently, it is not the fact that Ψ is nondifferentiable which causes naive policy iteration to fail. Rather, the problem stems from the fact that Ψ has linear segments

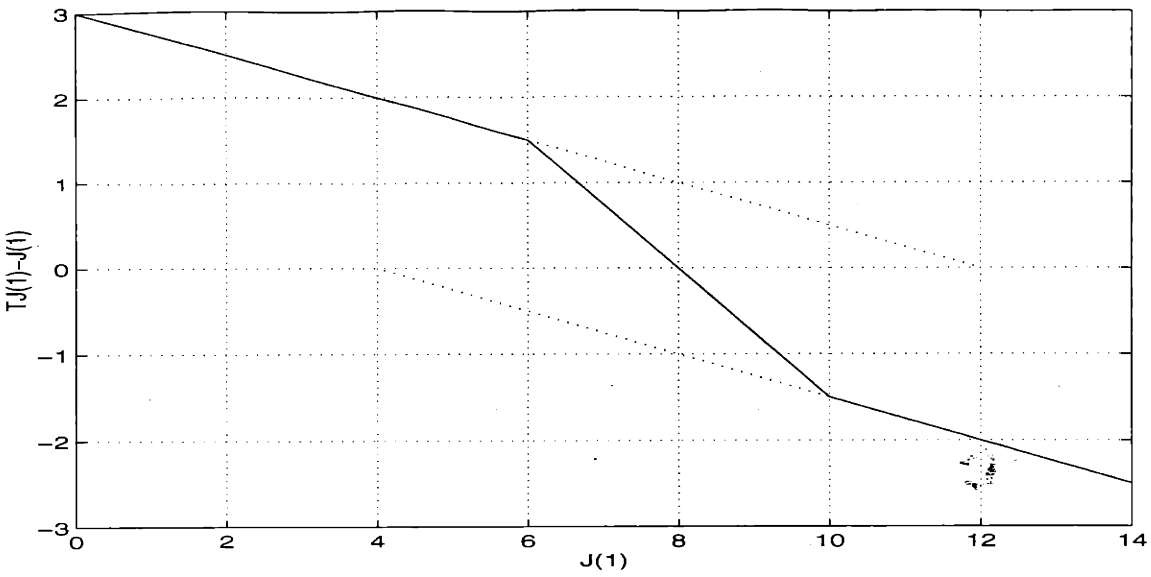

Figure 4-3: Bellman error $\Psi(J)(1)$ for van der Wal's example plotted as a function of estimates $J(1)$ of the equilibrium cost-to-go from 1. The solid line is the graph of Ψ. The dashed traces are the lines tangent to Ψ extended to the points where they intercept the $J(1)$-axis.

which are aligned so as to produce an oscillation in Newton's method. We would obtain this oscillatory behavior even if the nondifferentiable points were somehow smoothed out.

4.1.6 Modified Newton's Method

In [19], Filar and Tolwinski reconsidered the interpretation of naive policy iteration as Newton's method. They proposed a modification of the algorithm in which an Armijo-like stepsize rule is used to prevent the Newton updates from being too ambitious. To be more precise, after selecting an update direction according to (4.20), their algorithm employs Armijo's rule to guarantee a certain minimal relative improvement to

$$G(J) \triangleq \frac{1}{2}\Psi(J)'\Psi(J).$$

We give a formal statement of their algorithm, called modified Newton's method, below.

Algorithm 4.1.7 *(Modified Newton's Method [19])*

1. *Select two parameter values:* $\beta_1 \in (0,1)$ *and* $\beta_2 \in [.5,.8]$. *Also, choose an initial estimate for the equilibrium cost function,* $J_0 \in \mathcal{J}$.

2. *Given* $J_{k-1} \in \mathcal{J}$,

   (a) *Compute* $\mu_{k-1} \in M$ *and* $\nu_{k-1} \in N$ *such that* $TJ_k = T_{\mu_{k-1}} J_k$ *and* $\tilde{T}J_k = \tilde{T}_{\nu_{k-1}} J_k$.

   (b) *Compute* $D_{k-1} \in \mathcal{J}$ *as*

   $$D_{k-1} = J_{\mu_{k-1},\nu_{k-1}} - J_{k-1},$$

   where $J_{\mu_{k-1},\nu_{k-1}}$ *is the cost function associated with* (μ_{k-1}, ν_{k-1}).

   (c) *Set* $\gamma_k = 1$.

   (d) *Test the inequality*

   $$G(J_{k-1} + \gamma_k D_k) - G(J_{k-1}) \leq \beta_1 \gamma_k \nabla G(J_{k-1})' D_k.$$

   If the inequality is satisfied, then set $J_k = J_{k-1} + \gamma_k D_k$. *Otherwise, set* $\gamma_k := \beta_2 \gamma_k$ *and re-test the inequality.*

If Ψ is differentiable at J, then the gradient of G can be computed as

$$\nabla G(J) = [\nabla \Psi(J)] \Psi(J).$$

Moreover, as we remarked earlier, if Ψ is differentiable at J_{k-1}, then

$$D_k = -[\nabla \Psi(J_{k-1})']^{-1} \Psi(J_{k-1})$$

and we have that

$$\nabla G(J_{k-1})' D_k = -\Psi(J_{k-1})' \nabla \Psi(J)' [\nabla \Psi(J_{k-1})']^{-1} \Psi(J_{k-1}) = -\|\Psi(J_{k-1})\|^2 \leq 0.$$

Thus, the update directions for modified Newton's method are directions of nonincrease for the objective function G. Unfortunately, this argument breaks down at points where Ψ is no longer differentiable. Indeed, the algorithm itself is not well defined since the gradient of G may fail to exist at some J_k. If this is the case, then it is easy to let the algorithm proceed by replacing $\nabla G(J_k)$ with a directional derivative. However, it is possible that the algorithm can get stuck at a point where Armijo's rule cannot find a stepsize which yields a decrease in G.

In [19], Filar and Tolwinski argued (erroneously) that, with initial value function approximations $J_0 \in \mathcal{J}$ in some nonempty bounded set, modified Newton method will converge to the unique equilibrium value function for discounted cost games. Their argument uses classical results from the theory of nonlinear programming for *continuously differentiable* objective functions. In particular, they attempt to use a well-known result which states that the limit points of a gradient-related recursion under Armijo's rule (see [35, 5]) are stationary. This type of argument cannot be employed because here because the notion of a gradient-related sequence is not defined for the nondifferentiable objective function G.

Despite these difficulties, modified Newton's method seems to work well on academic examples. In particular, the method converges for van der Wal's example (as can be seen from Figure 4-3). Modified Newton's method also converges for the inspection game of Section 2.4.2, although it is quite slow and computationally intense.

4.2 Algorithms for Symmetric Sequential Games

In this section, we specialize our earlier results to the case of symmetric sequential games (the type described in Section 2.3.3). Recall that the state space of these games consists of pairs (z, i), where $z \in \{1, 2\}$ indicates whose turn it is, and $i \in \tilde{S} = \{1, \ldots, \tilde{n}\}$ indicates the relative positions of the two players. Also, for each i, the set of controls available to player z is denoted $W(i)$ and is assumed to be a compact subset of a metric space. Finally, we assume that termination is inevitable under all pairs of policies. Thus, these games fit into the framework for stochastic shortest

path games, and Propositions 3.1 and 3.2 apply. In particular, each such game has a unique equilibrium cost function J^* which can be found as the limit of value iteration. We note that J^* has a special structure due to the symmetric nature of the game. In particular,

$$J^* = \begin{pmatrix} X^* \\ -X^* \end{pmatrix}. \tag{4.21}$$

To see this, consider the value iteration algorithm applied to an initial cost function estimate J_0 of the form

$$J_0 = \begin{pmatrix} X_0 \\ -X_0 \end{pmatrix}.$$

Due to the symmetry of the game, each iterate J_k in value iteration has the same (symmetric) form. Since the iterates converge, they must converge to a vector of the form in (4.21).

We are led to the definition of two new operators that map $\mathcal{X}$, the space of real valued functions from $\{1, \ldots, \tilde{n}\}$, to itself. Given $\mu \in M$, define $Z_\mu : \mathcal{X} \mapsto \mathcal{X}$ as

$$Z_\mu X = \bar{c}(\mu) + \bar{P}(\mu)X - \bar{R}(\mu)X.$$

Now in analogy with T, define $Z : \mathcal{X} \mapsto \mathcal{X}$ as

$$ZX = \inf_{\mu \in M} \left[\bar{c}(\mu) + \bar{P}(\mu)X - \bar{R}(\mu)X \right].$$

Given a stationary policy μ for the minimizer, define the μ-symmetric policy ν_μ for the maximizer such that $\nu_\mu(2, i) = \mu(1, i)$ for all $i \in \tilde{S}$. Note that due to the symmetry of the game we have

$$T_{\mu\nu_\mu} \begin{pmatrix} X \\ -X \end{pmatrix} = \begin{pmatrix} Z_\mu X \\ -Z_\mu X \end{pmatrix},$$

$$T \begin{pmatrix} X \\ -X \end{pmatrix} = \begin{pmatrix} ZX \\ -ZX \end{pmatrix}.$$

Thus, the operators Z and Z_μ capture the essence of the computations that underly T and $T_{\mu\nu_\mu}$ when applied to symmetric estimates of the equilibrium cost function.

4.2.1 Symmetric Value Iteration

From the preceding discussion, we see that Z and Z_μ have unique fixed points which can be computed as the limits of the recursions $X := ZX$ and $X := Z_\mu X$, respectively. (We'll refer to this recursion as symmetric value iteration.) A closer analysis reveals that Z and Z_μ are contraction mappings. (This is obvious if the transition probabilities reflect a discount factor. Otherwise, given that termination is inevitable under all policies, the operators are still contractions with respect to a weighted sup-norm.)

4.2.2 Symmetric Policy Iteration

When policy iteration is applied to a symmetric game, two symmetric estimates of equilibrium cost-to-go are maintained on the respective halves of the state space. In this section we introduce a new algorithm which can be described as policy iteration for only half of the state space.

Algorithm 4.2.1 *(Symmetric Policy Iteration)*

1. *Choose an initial stationary policy μ_0 for the minimizer. (The maximizer will implicitly play the μ_0-symmetric policy ν_{μ_0}.)*
2. *Given $\mu_{k-1} \in M$:*
   (a) *(Symmetric Policy Evaluation) Compute the unique fixed point $X_{\mu_{k-1}} \in \mathcal{X}$ of the $Z_{\mu_{k-1}}$ operator.*
   (b) *(Symmetric Policy Improvement) Compute $\mu_k \in M$ such that*

$$ZX_{\mu_{k-1}} = Z_{\mu_k} X_{\mu_{k-1}}.$$

Because the operators Z and Z_μ do not have the monotonicity properties that are enjoyed by T and T_μ, the convergence of this algorithm is not immediately clear. Since

symmetric policy iteration can be viewed as a special case of naive policy iteration, there is good reason to suspect that the method is not globally convergent. Still, without doing any analysis, one may hope that the special structure of symmetric sequential games provides an extra mechanism for convergence. Unfortunately, a simple example shows that this is not the case. Consider the game illustrated in Figure 4-4. This is a discounted cost game where play strictly alternates between the two players. For every state of the system, one player gets to choose one out of two possible actions. Thus, each player has exactly eight pure policies. (There is no need to consider the mixed extension of this game since only one player acts at a time.) Figure 4-5 lists the costs to the minimizer of each of the eight policies. (In a fashion consistent with symmetric policy iteration, it is assumed that whenever the minimizer uses a policy μ the maximizer uses the μ-symmetric policy ν_μ.) Also listed are the corresponding unique policies which achieve the minimum in the Z operator applied to the cost evaluations. (We call these policies "greedy".) Note that policy 0 is the only policy which is greedy with respect to its own cost evaluation; thus it is the unique equilibrium policy of the game. Policy 0 is also greedy with respect to the evaluation of policy 7. In other words, if we initialized symmetric policy iteration with policy 7, it would converge to the equilibrium solution in one step. To see that this game actually represents a counter-example, we must examine the remaining policies. Note that

1. policy 2 is greedy with respect to the evaluation of policy 1,
2. policy 4 is greedy with respect to the evaluation policy 2, and
3. policy 1 is greedy with respect to the evaluation of policy 4.

Thus, policies 1, 2, and 4 form a cycle for symmetric policy iteration. Moreover, this cycle attracts all of the remaining policies: 3, 5, and 6. Obviously, if we were to initialize symmetric policy iteration with any of these policies, then the algorithm would fail to converge. The game is rigged in such a way that the cost-evaluations of policies 1, 2, and 4 are "misleading" to the policy improvement step.

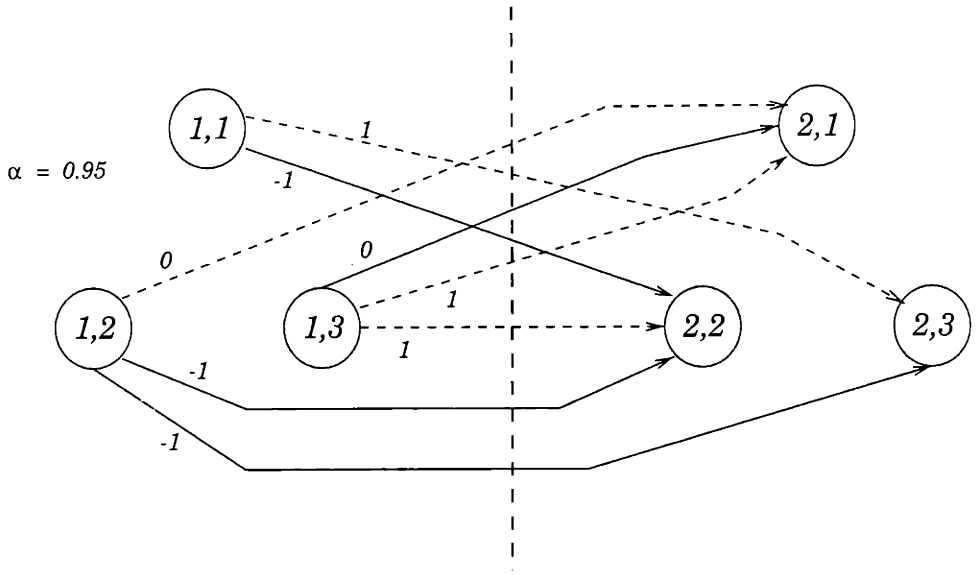

Figure 4-4: A symmetric game for which symmetric policy iteration is not globally convergent. Solid lines indicate possible transitions under action $a = 1$, while dashed lines represent possible transitions under $a = 2$. Whenever there is more than one possible transition under a particular action, the probabilities are assigned uniformly. The transition costs (to the minimizer) are indicated in the figure on the corresponding arcs; the corresponding transitions from the states $(2, i)$ are the negatives of the ones shown in the figure.

Policy	$\mu(1)$	$\mu(2)$	$\mu(3)$	$X_\mu(1)$	$X_\mu(2)$	$X_\mu(3)$	Greedy Policy
0	1	1	1	-0.275784	-0.762336	0.261992	0
1	2	1	1	10.2564	2.4598	-9.74359	2
2	1	2	1	-10.2564	9.74359	9.74359	4
3	2	2	1	10.2564	-9.74359	-9.74359	2
4	1	1	2	0.10376	-1.16185	1.50259	1
5	2	1	2	-1.03314	-1.35716	2.14014	1
6	1	2	2	-10.2564	9.74359	1.24359	4
7	2	2	2	0.0511542	-0.485965	0.998785	0

Figure 4-5: Table listing the eight policies of the game in Figure 4-4, along with their expected costs and corresponding greedy policies.

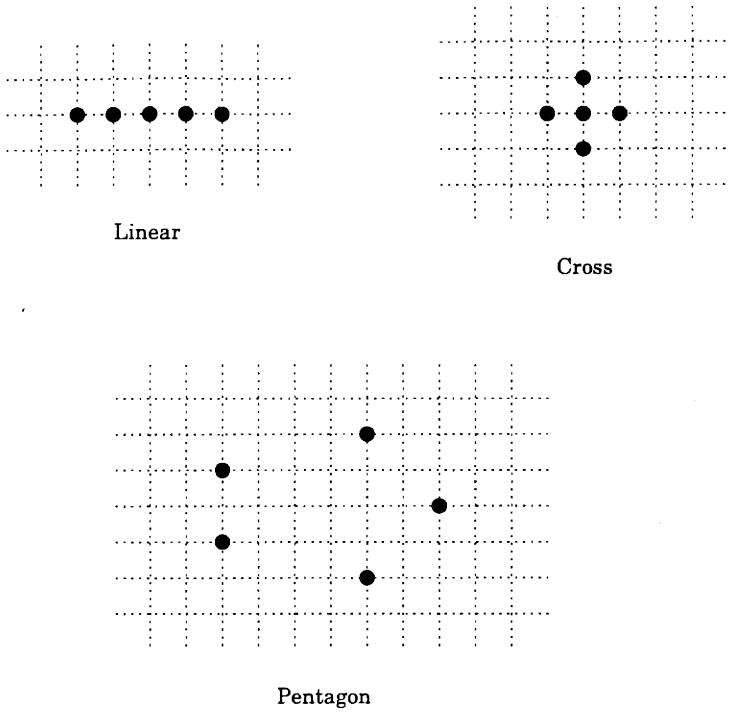

Figure 4-6: Three configurations for the industrial waste inspection game: linear, cross, pentagon.

4.3 Example: Industrial-waste Inspection

We return now to the industrial-waste dumping and inspection game of Section 2.4.2. Our purpose is to report on a small numerical study where we applied several dynamic programming algorithms to games with 5 dump sites (50 states). We consider 3 different configurations of the dump sites: linear, cross, and pentagon; as shown in Figure 4-6. The game parameters were chosen as follows: $p_1 = .95$, $p_2 = .15$, $k_1 = 2$, and $k_2 = 1$. Thus, the worst case probability of detecting a dump is .15, the best case probability is .95, and the effect of the inspector changing sites is twice as strong as the manufacturer changing sites. The dynamic programming algorithms we applied include value iteration, policy iteration, asynchronous policy iteration, naive policy iteration, and modified Newton's method. We list details specific to each algorithm below.

Value Iteration We initialized this algorithm with an initial cost function of all zeroes.

Policy Iteration We initialized this algorithm with the "uniform" policy where each site is chosen with equal probability. (This policy is proper.)

Asynchronous Policy Iteration Let S_D denote the subset of states where the manufacturer was detected last night. Each iteration of this algorithm consisted of four steps in sequence: (1) a cost update on $S - S_D$, (2) a policy update on S_D, (3) a cost update on S_D, and (4) a policy update on $S - S_D$. We initialized this algorithm with the uniform policy and an initial cost function of 2500 times the function of all ones. (We chose the initial cost estimate so that $T_{\mu_0} J_0 \leq J_0$, and we chose the initial policy to be proper.)

Naive Policy Iteration We initialized this algorithm with the cost function that results when the minimizer and maximizer both use the uniform policy.

Modified Newton's Method We set $\beta_1 = .1$ and $\beta_2 = .8$, and chose an initial cost function equal to the cost that results when the minimizer and maximizer both use the uniform policy. (Using an initial cost function of all zeros caused the algorithm to fail. It resulted in a target cost which was uniformly greater than the initial cost, so that the Armijo loop would never terminate.)

Results: The results of this numerical study are summarized in the table of Figure 4-7. The table lists the number of iterations that were required for the sup-norm difference between successive estimates of equilibrium cost to be less than 10^{-4}. In addition to this, various other statistics are listed which indicate the efficiency of the respective algorithms, including the total number of times the operator T was applied. (To evaluate $(TJ)(i)$, a linear program with five variables and five constraints must be solved. To evaluate TJ, this has to be done fifty times.) In describing policy iteration, we list the number of 1-player dynamic programming problems which had

Method:	Configuration: Notes:	Linear	Cross	Pentagon
Value Iteration	Iterations	1062	1649	1757
	T applications	1062	1649	1757
Policy Iteration	Iterations	11	10	12
	T applications	11	10	12
	Worst Case Evals.	11	10	12
Asynchronous Policy Iteration	Iterations	1526	2352	2497
	T applications	1526	2352	2497
	T_{μ_k} applications	1526	2352	2497
Naive Policy Iteration	Iterations	4	3	4
	T applications	4	3	4
	Pair Evaluations	4	3	14
Modified Newton's Method	Iterations	212	234	233
	Armijo's/Iteration	15	15	15
	T applications	3180	3510	3495
	Pair Evaluations	3180	3510	3495

Figure 4-7: Table summarizing the computational difficulty of the inspection game.

to be solved before termination of the iteration. In describing asynchronous policy iteration, we list the total number of times the T and T_{μ_k} operators were applied. (Note: we counted one T and one T_{μ_k} per global step of the method.) In describing naive policy iteration, we list the total number of times an exact evaluation of a pair of policies was done. (Each evaluation involves inverting a 50×50 matrix.) In describing modified Newton's method, we list the average number of Armijo steps which were required for each iteration. (Each Armijo step involves an application of the T operator along with the exact evaluation of a pair of policies.)

For each configuration of sites, the algorithms produced estimates of equilibrium cost-to-go that agree to at least four significant digits. As measured by compute-time, the naive policy iteration and policy iteration algorithms were the fastest. (Both naive policy iteration and modified Newton's method converged even though there are no theoretical guarantees that they should.) Modified Newton's method was the slowest of the algorithms, with asynchronous policy iteration being a close second.

Comparing Naive Policy Iteration and Modified Newton's Method Figures 4-8 through 4-10 show the evolution of modified Newton's method as applied to each of the three configurations of dump sites. Each global step of the method typically required 14 to 15 Armijo steps, ultimately resulting in stepsizes between $5.5 \cdot 10^{-2}$ and $4.4 \cdot 10^{-2}$. Consequently, each global update in modified Newton's method resulted in a very small adjustment to the estimate of the equilibrium cost-to-go function. This is to be contrasted with naive policy iteration which would always employ a stepsize of one in this framework. As shown in the figures, it is not the case that modified Newton's method approaches the pure Newton's method (i.e. naive policy iteration) in the limit; a large number of Armijo steps is required even as the method approaches a solution. In terms of computational effort, each Armijo step in modified Newton's method is roughly equivalent to a single step of naive policy iteration. As a result, modified Newton's method took at least $14 \cdot 212/4$ times longer than naive policy iteration to produce an accurate estimate of the equilibrium cost function.

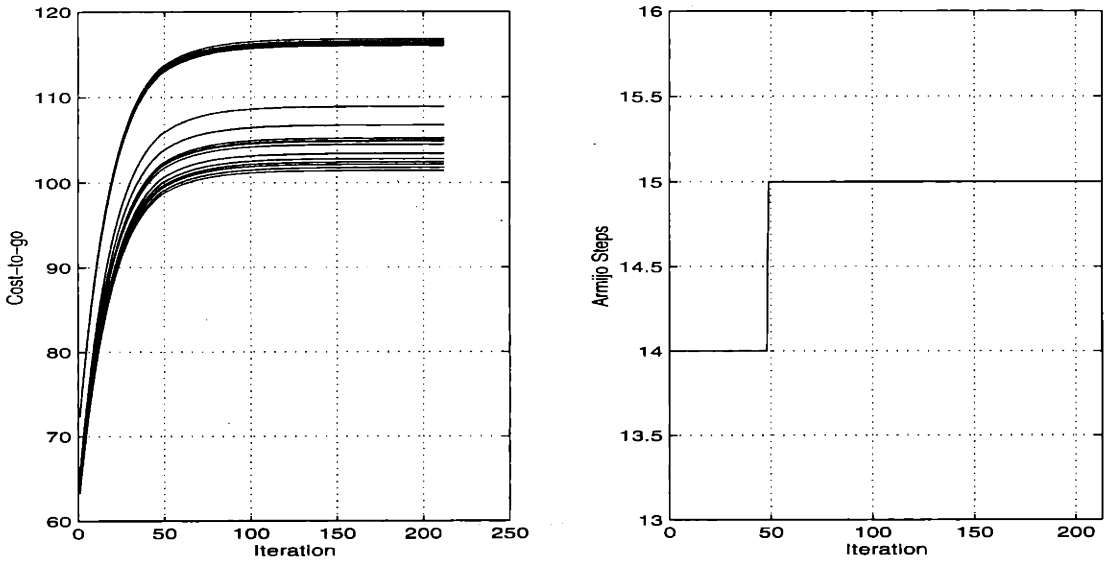

Figure 4-8: Modified Newton's Method applied to the linear configuration.

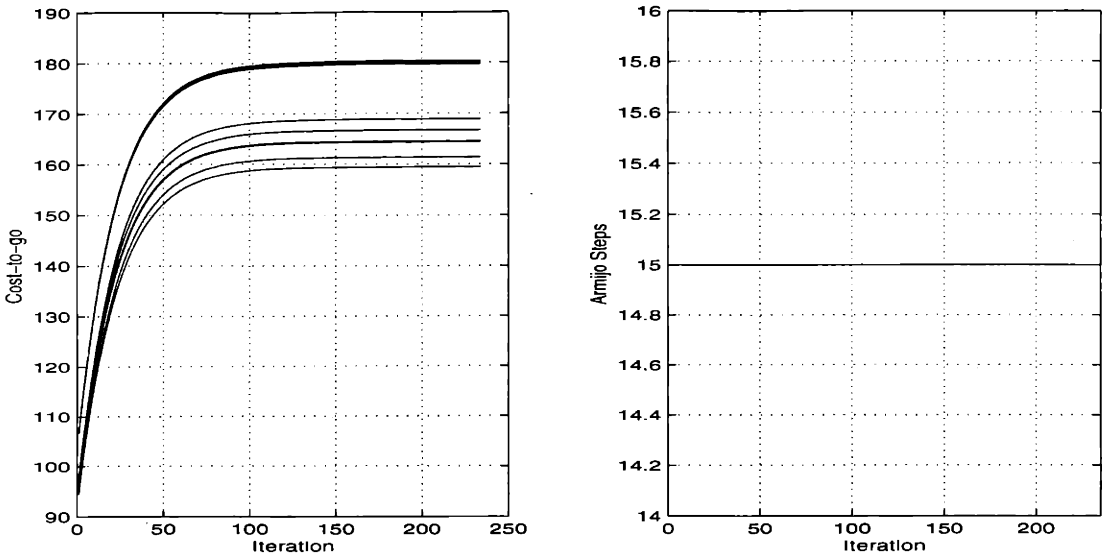

Figure 4-9: Modified Newton's Method applied to the cross configuration.

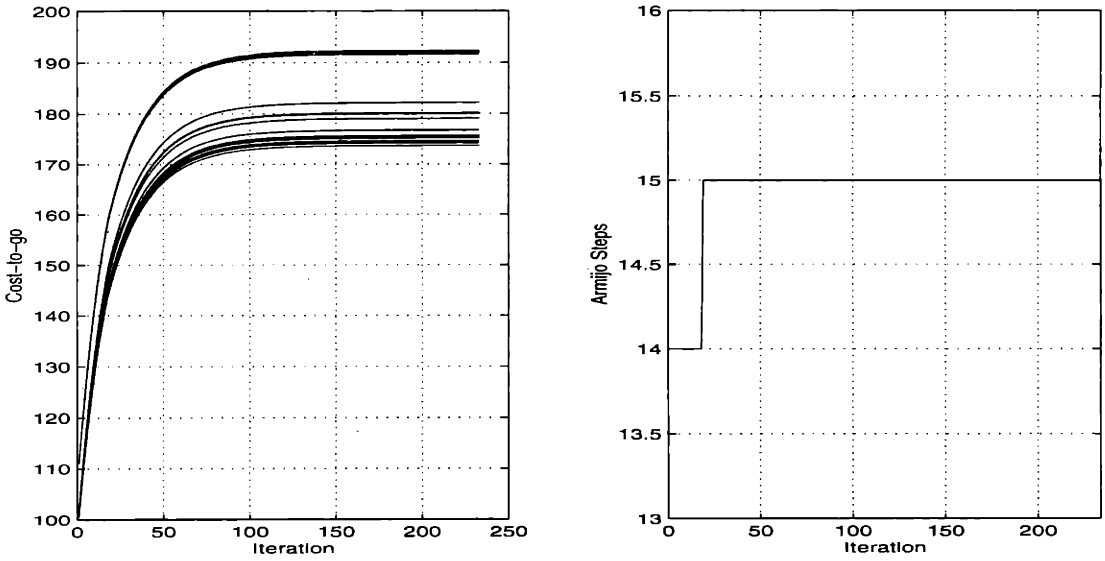

Figure 4-10: Modified Newton's Method applied to the pentagonal configuration.

4.4 Chapter Summary

In this chapter we analyzed dynamic programming algorithms for stochastic shortest path games. Regarding general algorithms, we first showed that value iteration converges to J^* for every initial estimate of the equilibrium cost. This is nontrivial given that the operator T isn't always a contraction. Next, we considered a policy iteration (of the same type as considered by Rao et al. [44]), where a sequence of policies for the minimizer is generated based on worst-case evaluations of cost. We proved that this form of policy iteration yields a sequence of policies whose costs converge monotonically to the equilibrium value of the game. We showed that a corresponding form of asynchronous policy iteration also converges to the equilibrium, implying a certain robustness for policy iteration. We next proved an error-bound result for approximate policy iteration. For completeness, we described naive policy iteration (due to Pollatschek and Avi-Itzhak [40]) which is a variation on the earlier policy iteration method. While this method is easier to implement than policy iteration, it is not globally convergent, as illustrated by an example of van der Wal [59]. Filar and Tolwinski proposed a related algorithm called modified Newton's method-so named for a geometric interpretation originally due to Pollatschek and Avi-Itzhak. We discussed this method in brief detail only to show that Filar and Tolwinski's proof of convergence is incorrect, and it is unclear at present whether convergence actually prevails. At this point, we switched gears to discuss specialized algorithms for symmetric sequential games. After defining two new "dynamic programming operators", we showed that the corresponding symmetric value iteration method is globally convergent, while the corresponding symmetric policy iteration is not. The last section of the chapter was devoted to a computational study where we applied the general algorithms of Section 4.1 to the inspection game of Section 2.4.2. All of the algorithms we tested converged to an equilibrium solution of the game, including naive policy iteration and modified Newton's method (for which convergence results do not exist). It is interesting to note that, while naive policy iteration was the fastest to produce and equilibrium solution, modified Newton's method was the slowest. This

indicates that the introduction of an Armijo's rule to naive policy iteration significantly impacts the evolution of the algorithm. In fact, modified Newton's method never implemented a pure Newton step.

Chapter 5

The Average Cost Connection

In this chapter we explore relationships between stochastic shortest path games and games with an average cost objective. We first show that the existing literature on average cost games can be used to obtain a subset of the results from Chapter 3 about stochastic shortest path games. It turns out that this line of reasoning is quite limited because the analysis of average cost games (to date) has been restricted to the case where players optimize with respect to mixed strategies over finite underlying sets of actions (as in Section 2.3.4). After making this point, we turn the tables and use the general theory of Chapters 3 and 4 to obtain *new* results for average cost games. In particular, we establish the existence of stationary equilibrium solutions for a broad class of average cost games where players choose actions from arbitrary compact constraint sets. We characterize the equilibrium value function of these games as the effectively-unique solution to a form of Bellman's equation, and establish the convergence of several dynamic programming algorithms. In the remainder of this introductory section we define basic terminology and review the pertinent literature on average cost games.

As in earlier chapters, we restrict attention to finite state games. However, because we will make use of an average cost objective, we do not require the existence of an extra terminal state Ω. Given a pair of policies $\pi_M \in \bar{M}$ and $\pi_N \in \bar{N}$, the average

cost to the minimizer from state i is defined to be

$$\bar{J}_{\pi_M,\pi_N}(i) = \liminf_{t\to\infty} \frac{1}{t+1} h^t_{\pi_M,\pi_N}(i), \tag{5.1}$$

where $h^t_{\pi_M,\pi_N}(i)$ is the expected $(t+1)$-stage cost from i under (π_M, π_N) which was defined in (2.2). Unlike the existing literature for average cost games, we assume that the control constraint sets for both players are arbitrary compact subsets of metric spaces.

Games with average cost objectives have been studied for a long time, starting with Gillette [21] in 1957. Gillette's original formulation included finite underlying sets of actions with the players allowed to use mixed strategies. In a famous example, the big match [21], Gillette showed that average-cost games do not generally have value in stationary policies. Additionally, he attempted to show that some special classes of games (with enough special structure) not only have value but also have stationary equilibrium solutions. The special classes he considered are

1. sequential games with finite constraint sets (cf. Section 2.3.2, otherwise known as games of perfect information) and

2. irreducible games where, under all pairs of pure policies, the associated Markov chain is irreducible. (Irreducible games are also sometimes called recurrent.)

Unfortunately, Gillette's analysis of these games relied upon an incorrect generalization of the Hardy-Littlewood theorem which relates discounted summations to Cesaro-averaged summations. Gillette's results were eventually proven to be correct. In [29], 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 [25], 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 [16] and van der Wal [60] 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 [54] 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. [We will refer to these as recurrent-state games. These are to be distinguished from irreducible (recurrent) games.] Equilibria in Stern's games are characterized (but not uniquely) by solutions to a generalized form of Bellman's equation. In [53], 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. (We will refer to these as unichain games. Notice that it is not required that the same recurrent class prevail for all profiles of policies.) In [47], using a different technique, Rogers obtained similar results.

Generally, as demonstrated by Gillette, average cost games do not have equilibrium (or even ϵ-equilibrium) solutions in stationary policies. This left researchers wondering whether average cost games have value in more general classes of policies. In [11], Blackwell and Ferguson showed that the big match does, in fact, have an equilibrium value. They showed that one player has a stationary equilibrium solution, while the other player has an ϵ-equilibrium solution in the form of a behavioral strategy (which specifies mixed actions based on the entire history of play). The question remained as to whether all average cost games have ϵ-equilibria in behavioral strategies. This question was eventually answered in the affirmative. Around 1980, Mertens and Neyman [36] and (independently) Monash [37] proved that all average cost stochastic games have ϵ-equilibria in behavioral strategies. Both sets of researchers employed the earlier results of Bewley and Kohlberg in [10].

Remark 1: All of the results discussed above make use of Gillette's original assumption that the players are optimizing with respect to mixed strategies over finite sets of actions. (We are unaware of any literature on average cost games where this assumption is not made.) Thus, one purpose of this chapter 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.

Remark 2: The literature on *single-player*, finite-state, average cost games with compact constraint sets is fairly well developed.[1] In the following, we briefly review (in chronological order) some of the highlights of this literature. Martin-Löf [34] established the existence of optimal stationary policies in irreducible average cost problems. Bather [2] established the existence of a solution to a single optimality equation given that an "accessibility" condition holds where, for any pair of states i and j, there is a stationary policy such that after a finite number of transitions there is a nonzero probability of reaching j from i. In this way, Bather also established the existence of optimal stationary policies. (The single optimality equation Bather considered is analogous to the Bellman's equation we develop in the sequel.) Fainberg [14] established the existence of stationary optimal policies under either the unichain assumption or the assumption that for each state the set of transition probabilities contains a finite set of extreme points. Fainberg [15] established the existence of ϵ-optimal policies in general (multichain) average cost problems (without the conditions he required in his earlier paper.) Federgruen and Tijms [18] considered semi-Markov decision processes with countable state spaces. They gave three "recurrency" conditions (one of which is equivalent to our recurrent-state assumption when restricted to problems with finite state spaces) which guarantee the existence of a bounded solution to a single optimality equation (analogous to the Bellman's equation we develop in the sequel). Moreover, by allowing instantaneous transitions, they are able to prove the convergence of a policy iteration algorithm. Federgruen, Schweitzer, and Tijms [17] proved the existence of a solution to a pair of optimality equations for multichain average problems where an "accessibility" assumption

[1] We refer the reader to [1] and [41] for surveys of the literature on single-player average cost problems.

holds. Schweitzer [48] showed that if the optimal average cost (within the class of stationary policies) is identical for all states then a single optimality equation (Bellman's equation) has a solution. Schweitzer [49] showed that a solution exists to a pair of optimality equations for multichain average cost problems if and only if at least one nonrandomized maximal-gain policy exists and the bias-vectors of all such maximum-gain policies are uniformly bounded above. Schweitzer [50] generalized Bather's results [2] by using Brouwer's fixed point theorem to prove the existence of a solution to a single optimality equation (Bellman's equation) in average cost problems satisfying an "accessibility" condition. Hordijk and Puterman [26] examined policy iteration as applied to unichain Markov (not semi-Markov) decision processes. They proved that the costs associated with the policies generated by the method converge to the optimal average cost and if there is a unique minimizer in the dynamic programming operator then the associated bias-vectors also converge to a solution of a single optimality equation (Bellman's equation). They establish these results by using a Newton's method interpretation of policy iteration which does not generalize to the two-player case. (Their analysis does not require differentiability of the objective function in the Newton's method interpretation.) Dekker [13] gave counter examples which show that (1), in the unichain case, Hordijk and Puterman's [26] policy iteration method does not converge finitely and (2) without the unichain assumption the method may not converge to the optimal average cost.

5.1 An Alternative Proof for Proposition 3.1

Proposition 3.1 states that there exists a unique solution to Bellman's equation in additive cost games satisfying Assumptions R and SSP. It turns out that there are several ways of proving this important result. The proof we gave in Section 3.2 is quite different from an earlier proof (cf. [39]), where we showed that policy iteration converges to a unique limit which solves $J = TJ$. This section is devoted to a third proof, suggested by an anonymous SIAM reviewer, which applies to the case of mixed strategies over finite underlying sets of actions. We outline the argument below.

1. Consider the average reward version of the stochastic shortest path game. (That is, simply impose the average cost objective function of (5.1), keeping the underlying transition probabilities and cost functions the same.)

2. Fix a proper policy $\mu \in M$ for the minimizer. (One exists thanks to Assumption SSP.) By Lemma A.9, we have that $-\infty < J_\mu(i) < \infty$ for all states $i \in S$. As a result, we claim that

$$\bar{J}_{\mu,\pi_N} = \mathbf{0}$$

for all $\pi_N \in \bar{N}$, where $\mathbf{0}$ denotes the zero function in $\mathcal{J}$. This implies that

$$\inf_{\pi_M \in \bar{M}} \sup_{\pi_N \in \bar{N}} \bar{J}_{\pi_M,\pi_N} \leq \mathbf{0}.$$

3. Now fix a stationary policy $\nu \in N$ for the maximizer. From Lemma A.10, we know that $-\infty < J_\nu(i) < \infty$ for all states $i \in S$. As before, this implies that $\bar{J}_{\pi_M,\nu} = \mathbf{0}$ for all $\pi_M \in \bar{M}$. As a result,

$$\sup_{\pi_N \in \bar{N}} \inf_{\pi_M \in \bar{M}} \bar{J}_{\pi_M,\pi_N} \geq \mathbf{0}.$$

4. Combining these inequalities, we see that the average cost version of the game has a stationary equilibrium solution and the equilibrium cost from all initial states $i \in S$ is zero. (Any pair of stationary policies for the two players will achieve this value.) By Lemma 8.1.3 in [63], we have that there exists a fixed point of T.

Remark 1: In studying this argument, it is clear that Lemma 8.1.3 from [63] is crucial. The proof of this lemma, as given in [63], begins with one of the players fixing a stationary equilibrium policy (for the average cost version of the game). This gives rise to a Markov decision problem faced by the remaining player. Using the fact that the underlying sets of actions for this problem are finite, Theorem 3.1 from [51][2]

[2]See also Theorem 9.1.4 in [41] or Proposition 2.4 in Chapter 4 of [4].

assures that there exists a solution to the associated Bellman's equation. This fact is used to complete the proof. As a result, the preceding argument only applies to stochastic shortest path games of the type considered in Section 2.3.4. [Before leaving this point, we note that Federgruen was the first to prove the result of Lemma 8.1.3 (cf. Corollary 7.3.5 in [16]). While Federgruen's basic approach was quite different from Vrieze's, it is also crucially dependent on the assumption that the underlying sets of actions are finite.]

Remark 2: The results of [8] are necessary in order for the alternative argument to hold. In particular, in the third step of the argument where we fix a stationary policy $\nu \in N$ for the maximizer, we need the earlier results to be sure that the minimizer's best response results in bounded cost.

5.2 Recurrent-state Average Cost Games

In this section we return to the recurrent-state games considered by Stern in [54]. That is, 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 maintain the assumption that there is a finite state-space and that the control constraint sets are compact subsets of metric spaces.) 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 will use a line of reasoning which appeared originally in [3] for the case of a single player. Our results, while restricted to the zero-sum case, generalize the results of Stern since the control constraint sets are general compact subsets of metric spaces. Consequently, we also generalize previously known results about so-called irreducible games [21, 29] (see also [20]).

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 and $\bar{M}$ 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 the maximizer. Given $\mu \in M$ and $\nu \in N$, let $P(\mu, \nu)$ and $c(\mu, \nu)$ be the corresponding transition probability matrix and expected transition cost vector, respectively. Let all of the usual dynamic programming operators be defined as in (2.3)-(2.7). (Eventually, we will interpret $\mathcal{J}$ as the space of all *differential* cost functions for the average cost game.) We make the following regularity assumptions which are slightly more restrictive than Assumption R.

Assumption $\bar{\mathrm{R}}$ *(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 $(u, v) \in U(i) \times V(i)$. (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 $H \in \mathcal{J}$, there exists $\mu \in M$ and $\nu \in N$ such that $TH = T_\mu H \in \mathcal{J}$ and $\tilde{T}H = \tilde{T}_\nu H \in \mathcal{J}$.)*

3. *For every $H \in \mathcal{J}$, we have $TH = \tilde{T}H$.*

The following assumption will play a central role in the results of this section.

Assumption RS *(Recurrent State) The Markov chain associated with each pair of stationary policies (μ, ν) is unichain. Moreover, the state $n \in S$ is recurrent under every pair of stationary policies.*

Given Assumption RS, we can view the recurrent state n as a terminal state which is inevitably reached in an infinite sequence of conventional stochastic shortest path games. The results of Section 3.2 then help to establish the existence of an equilibrium and the convergence of dynamic programming algorithms. To make a formal definition, consider an average cost game satisfying Assumptions $\bar{\mathrm{R}}$ and RS, along with an estimate λ of the equilibrium average cost. The associated stochastic

shortest path game (λ-SSPG), with transition probabilities $\bar{p}_{ij}(u,v)$ and costs $\bar{c}_i(u,v)$, is obtained by

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.)

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 3.1). In fact, there is a positive vector $w \in \mathcal{J}$ 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. Throughout the rest of this chapter, 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 $H \in \mathcal{J}$ 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 $\bar{\mathcal{J}} = \{H \in \mathcal{J} \mid H(n) = 0\}$. 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 5.1 *The following are statements are true for average cost games satisfying Assumptions $\bar{R}$ and RS.*

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

$$J_{\lambda,\mu,\nu}(i) = J_{\lambda',\mu,\nu}(i) + (\lambda' - \lambda) N_{\mu,\nu}(i), \quad i = 1, \ldots, n. \tag{5.2}$$

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

$$\begin{aligned} &J_{\lambda',\mu}(i) + N_{min}(\lambda' - \lambda) \leq J_{\lambda,\mu}(i) \leq J_{\lambda',\mu}(i) + N_{max}(\lambda' - \lambda), \quad \text{if } \lambda' \geq \lambda, \\ &J_{\lambda',\mu}(i) + N_{max}(\lambda' - \lambda) \leq J_{\lambda,\mu}(i) \leq J_{\lambda',\mu}(i) + N_{min}(\lambda' - \lambda), \quad \text{if } \lambda' \leq \lambda, \end{aligned} \tag{5.3}$$

for all $i = 1, \ldots, n$.

3. *The functions $J_\lambda(i)$ are continuous and decreasing as functions of λ and satisfy*

$$\begin{aligned} J_{\lambda'}(i) + N_{min}(\lambda' - \lambda) \leq J_\lambda(i) \leq J_{\lambda'}(i) + N_{max}(\lambda' - \lambda), \quad \text{if } \lambda' \geq \lambda, \\ J_{\lambda'}(i) + N_{max}(\lambda' - \lambda) \leq J_\lambda(i) \leq J_{\lambda'}(i) + N_{min}(\lambda' - \lambda), \quad \text{if } \lambda' \leq \lambda, \end{aligned} \tag{5.4}$$

for all $i = 1, \ldots, n$.

Proof: To prove statement 1, note that the second term on the right hand side of (5.2) is the expected differential cost associated with λ' in the associated stochastic shortest path game.

To prove statement 2, note that the continuity of the functions $J_{\lambda,\mu}(i)$ follows from Proposition 7.32 in [7] 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

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

Finally, to see (5.3), let $\lambda' \geq \lambda$ be given; then, for all $\nu \in N$ we have $J_{\lambda,\mu,\nu}(i) = J_{\lambda',\mu,\nu}(i) + (\lambda' - \lambda) N_{\mu,\nu}(i) \geq J_{\lambda',\mu,\nu}(i) + (\lambda' - \lambda) N_{min}$. The right-most expression is maximized by some $\bar{\nu} \in N$. Thus,

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

The remaining inequalities of (5.3) follow similarly.

To prove statement 3, note that the continuity of $J_\lambda(i)$ follows from Proposition

7.32 in [7] 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

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

Finally, we obtain (5.4) from (5.3) and similar arguments. **Q.E.D.**

Remark: 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.

5.2.1 Existence and Characterization of Equilibria

In this subsection, we establish the existence of stationary equilibrium solutions in average cost games satisfying Assumptions $\bar{\text{R}}$ and RS. We characterize the equilibrium value function as the effectively unique solution to a form of Bellman's equation. The results of this subsection are analogous to the propositions of Section 3.2, and can be viewed as a generalization of the results in [54] (cf. chapter 2, restricted to the zero-sum case). Our techniques depend upon the analysis of [3].

Proposition 5.1 *The following statements are true for average cost games satisfying Assumptions $\bar{R}$ and RS.*

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 $H^* \in \mathcal{J}$ which, along with λ^*, satisfies Bellman's equation*

$$\lambda^* \mathbf{1} + H^* = TH^*. \tag{5.5}$$

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 (5.5), there is a unique solution for which $H(n) = 0$.

2. *If a scalar λ and a function $H \in \mathcal{J}$ satisfy (5.5), 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 $H_\mu \in \mathcal{J}$ 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 $H_\nu \in \mathcal{J}$ 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 $H_{\mu\nu} \in \mathcal{J}$ 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 $\bar{\text{R}}$ 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

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

Thus,

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

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 [3]. 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 [7], 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

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

Observe that for all $\mu \in M$

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

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 $H^* \in \mathcal{J}$ (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 $\bar{\text{R}}$,

$$H^*(i) + \tilde{\lambda} = \max_{v \in V(i)} \min_{u \in U(i)} \left[c_i(u,v) + \sum_{j=1}^{n} p_{ij}(u,v) H^*(j) \right], \quad i \in \{1, \ldots, n\}. \tag{5.6}$$

Because we have found an equilibrium of the associated stochastic shortest path game a policy $\nu^* \in N$ which achieves the maximum in (5.6) 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 5.1 *Under Assumptions $\bar{R}$ and RS, the following are true.*

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.*

Remark: 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. In particular, it is possible (as in [4]) to use the existence of Blackwell optimal policies[3] to show that if every policy which is optimal

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

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 [4] cannot be generalized easily to prove the existence of a solution to Bellman's equation in unichain average cost problems with compact constraint sets.[4] Similarly, the analysis of [4] cannot be generalized easily to prove the existence of solutions to (5.5) in unichain games satisfying Assumption $\bar{\text{R}}$.

5.2.2 Dynamic Programming Algorithms

In this subsection we state and discuss the convergence properties of several dynamic programming algorithms.

Value Iteration

The first algorithm we consider is the value iteration algorithm of Chapter 3. It turns out that given any terminal cost function $J \in \mathcal{J}$, the k-horizon equilibrium cost divided by k approaches the equilibrium average cost of the game.

Proposition 5.2 *Under Assumptions $\bar{R}$ and RS we have that*

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

for every $J \in \mathcal{J}$, where λ^* is the equilibrium average cost of the game.

Proof: The proof we give is nearly identical to an argument in [4] (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 without any modifications.

[4] The existence of solutions to Bellman's equation under the unichain assumption was established in [50] by other methods.

By Proposition 5.1, there exist λ^* and H^* which solve Bellman's equation. Set $J_0^* = H^*$ and recursively define $J_{k+1}^* = TJ_k^*$. Since $\lambda^*\mathbf{1} + H^* = TH^*$, we can show by induction that

$$J_k^* = k\lambda^*\mathbf{1} + H^*$$

for all $k \leq 0$. Moreover, since T is nonexpansive, we have that

$$|T^k J(i) - J_k^*(i)| \leq \|J - H^*\|_\infty$$

for all $k \geq 0$ and $i \in S$. Combining these inequalities, we obtain for all $i \in S$

$$|T^k J(i) - k\lambda^*| \leq \|J - H^*\|_\infty + \|H^*\|_\infty.$$

Thus, $[T^k J(i)]/k$ converges to λ^*. **Q.E.D.**

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 5.2.1 *(Relative Value Iteration)*

1. *Choose $\tau \in (0, 1]$, $t \in S$, and an initial $H_0 \in \mathcal{J}$.*

2. *Given H_k, compute*

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

Remark 1: If this algorithm converges, say to $\bar{H}$, then the limit satisfies

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

As a result, $T(\tau H_k)(t)$ converges to the equilibrium average cost λ^*, and $\bar{H} = (1/\tau)H^*$, where H^* is the unique solution to $TH = H + \lambda^*\mathbf{1}$ with $H^*(t) = 0$.

Remark 2: 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. Unfortunately, this does not seem to be of much help in establishing the convergence of the method. To see this, let $\beta < 1$ be the weighted sup-norm contraction modulus associated with the dynamic programming operators of the associated stochastic shortest path game, and recall that $\|\cdot\|$ denotes the corresponding weighted sup-norm whose weights are scaled so that the weight on state n is one. Let λ^* be the equilibrium average cost of the game and let H^* be the unique solution to $TH = H + \lambda^*\mathbf{1}$ with $H(n) = 0$. Define $\bar{H} = (1/\tau)H^*$, and let λ_k denote $T(\tau H_k)(n)$. Using the fact that $H_k(n) = H_{k+1}(n) = \bar{H}(n) = H^*(n) = 0$, we have for all $k \geq 1$

$$\begin{aligned} |\lambda_k - \lambda^*| &= |T(\tau H_k)(n) - (TH^*)(n)| \\ &\leq \|T(\tau H_k) - TH^*\| \\ &\leq \beta\|\tau H_k - H^*\| \\ &= \tau\beta\|H_k - \bar{H}\|. \end{aligned}$$

Thus,

$$\|(\lambda_k - \lambda^*)\mathbf{1}\| = |\lambda_k - \lambda^*| \cdot \|\mathbf{1}\| \leq \tau\beta\|\mathbf{1}\| \cdot \|H_k - \bar{H}\|.$$

From the definition of H_k and after some algebraic manipulation we get

$$H_k - \bar{H} = (1 - \tau)[H_{k-1} - \bar{H}] + [T(\tau H_{k-1}) - T(\tau\bar{H})] + (\lambda_{k-1} - \lambda^*)\mathbf{1}.$$

From the triangle inequality we obtain

$$\|H_k - \bar{H}\| \leq [(1 - \tau) + \tau\beta(1 + \|\mathbf{1}\|)] \cdot \|H_{k-1} - \bar{H}\|.$$

Thus, the sequence of iterates H_k converges to $\bar{H}$ if

$$[(1-\tau)+\tau\beta(1+\|\mathbf{1}\|)]<1 \quad \Leftrightarrow \quad \beta(1+\|\mathbf{1}\|)<1.$$

From the proof of Lemma 3.1, we have

$$\beta(1+\|\mathbf{1}\|)=(1-1/N_{max})(1+N_{max}/N_{max}(n)),$$

where $N_{max}(n)=\max_{\mu\in M}\max_{\nu\in N}N_{\mu,\nu}(n)$. Since $N_{max}\geq N_{max}(n)\geq 2$, it is impossible to have $\beta(1+\|\mathbf{1}\|)<1$, and the sufficient condition can never be satisfied.

Proposition 5.3 *In addition to Assumptions $\bar{R}$ and RS, assume that there exists a positive integer m such that for every pair of admissible policies $\pi_M=\{\mu_0,\mu_1,\ldots\}$ and $\pi_N=\{\nu_0,\nu_1,\ldots\}$, there exists an $\epsilon>0$ such that*

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

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.)

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

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

Defining $q_k=H_{k+1}-H_k$, we obtain

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

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

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,

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

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

$$q_k \geq [P(\mu_k, \underline{\nu}_k) \dots P(\mu_{k-m+1}, \underline{\nu}_{k-m+1})] q_{k-1} + (\lambda_{k-m} - \lambda_k)\mathbf{1} \tag{5.7}$$

$$q_k \leq [P(\mu_{k-1}, \bar{\nu}_k) \dots P(\mu_{k-m}, \bar{\nu}_{k-m+1})] q_{k-1} + (\lambda_{k-m} - \lambda_k)\mathbf{1}. \tag{5.8}$$

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

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

From (5.8), we obtain

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

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 (5.7), 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 [4]. **Q.E.D.**

Remark: As described in [4], 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. We note that our results about relative value iteration generalize some of the earlier results of Federgruen [16] and van der Wal [60].

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 [6] and is motivated by the connection with stochastic shortest path games.

Algorithm 5.2.2 *(Contracting Value Iteration)*

1. *Start with an initial estimate* (λ_0, H_0) *of a solution to Bellman's equation (5.5).*
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 5.4 *Under Assumptions $\bar{R}$ and RS, 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 (5.5) with $H^*(n) = 0$.

Proof: The proof uses Lemma 5.1 and Corollary 5.1 and closely follows the proof of Proposition 1 in [6]. 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.

Policy Iteration

We now examine the policy iteration algorithm of Hoffman and Karp [25].

Algorithm 5.2.3 *(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*

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

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

Remark: This algorithm is known to converge [25] 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 $\bar{\mathrm{R}}$ and RS. Unfortunately, it falls short of actually proving convergence to a solution of Bellman's equation.

Proposition 5.5 *Under Assumptions $\bar{R}$ and RS, in the policy iteration algorithm, for each k we either have*

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

or else we have

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

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 cost function associated with the minimizer's policy μ_{k+1} when the terminal cost function is H_{μ_k}. Thanks to Proposition 5.2 we have that for every $i \in S$

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

By Proposition 5.1 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 5.1 implies that both μ_k and μ_{k+1} are equilibrium policies for the minimizer. **Q.E.D.**

Corollary 5.2 *If the minimizer has only finitely many policies, then policy iteration converges in a finite number of iterations.*

Remark: Convergence of policy iteration in the more general case (where $U(i)$ and $V(i)$ are arbitrary compact subsets of metric spaces) is not clear. To begin the analysis, we note that the pairs $(\lambda_{\mu_k}, H_{\mu_k})$ produced by the algorithm are all contained within a compact subset of $\Re^{n+1}$. To see this, recall that, as a function of (μ, ν), the unique solution $(\lambda_{\mu,\nu}, H_{\mu,\nu})$ to the equations

$$\begin{aligned} \lambda \mathbf{1} + H &= T_{\mu,\nu} H, \\ H(n) &= 0 \end{aligned}$$

is continuous over the compact product space $M \times N$. Thus, the space of all possible average and differential cost pairs is compact. Since for each μ_k there exists a stationary policy ν_k such that $(\lambda_{\mu_k}, H_{\mu_k}) = (\lambda_{\mu_k,\nu_k}, H_{\mu_k,\nu_k})$, the assertion is true. As a result, there is a convergent subsequence $\{(\lambda_{\mu_k}, H_{\mu_k})\}_{k \in \mathcal{K}_1}$ with limit $(\bar{\lambda}, \bar{H})$, where $\bar{H}(n) = 0$. Consider the corresponding subsequence of policies $\{\mu_{k+1}\}_{k \in \mathcal{K}_1}$. Since M is compact, there is a convergent subsequence of policies $\{\mu_{k+1}\}_{k \in \mathcal{K}_2}$ (where $\mathcal{K}_2 \subset \mathcal{K}_1$), with limit $\bar{\mu}$. It is not difficult to show that $T\bar{H} = T_{\bar{\mu}}\bar{H}$. Since $\lambda_{\mu_k}\mathbf{1} + H_{\mu_k} = T_{\mu_k} H_{\mu_k} \geq T H_{\mu_k}$ for all k, we have (by taking the limit of the subsequence and the continuity of T) that $\bar{\lambda}\mathbf{1} + \bar{H} \geq T\bar{H} = T_{\bar{\mu}}\bar{H}$. This implies $\lambda_{\bar{\mu}} \leq \bar{\lambda}$ by an induction argument. On the other hand, since λ_{μ_k} is monotonically decreasing, $\bar{\lambda} \leq \lambda_{\mu_{k+1}}$ for all k. Thus, from the continuous dependence of λ_μ on μ, we have $\bar{\lambda} \leq \lambda_{\bar{\mu}}$. It follows that $\bar{\lambda} = \lambda_{\bar{\mu}}$, and therefore $H_{\bar{\mu}} \leq \bar{H}$. If equality prevails in the latter, then we are done. However, if the inequality is strict for some state i, then there is nothing else to say. As a result, 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.[5]

[5] Unfortunately, we cannot pursue the type of analysis of Hordijk and Puterman in [26] which relies upon a Newton's method interpretation of single-player policy iteration which does not generalize to the two-player case.

ϵ-Policy Iteration

In this subsection, we describe a variation of policy iteration which yields policies that are arbitrarily close to equilibrium.

Algorithm 5.2.4 *(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*

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

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

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

   (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

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

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

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

Remark: 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 5.1, 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 5.6 *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 $J_{\lambda_{\mu_k},\tilde{\mu}_j}(n) < -\epsilon$. Since $J_{\lambda,\tilde{\mu}_j}(n)$ is strictly decreasing as a function of λ, it is true that $\lambda_{\tilde{\mu}_j} < \lambda_{\mu_k}$. Moreover, from (5.3), associating λ' with λ_{μ_k}, λ with $\lambda_{\tilde{\mu}_j}$, and μ with $\tilde{\mu}_j$, we have that

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

   which implies that $\lambda_{\mu_k} - \lambda_{\tilde{\mu}_j} \geq \epsilon/N_{max}$. 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 (5.4), 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.**

Naive Policy Iteration

As with stochastic shortest path games, it is possible to define a naive policy iteration for recurrent-state average cost games. Pollatschek and Avi-Itzhak [40] studied such an algorithm and were unable to prove it's convergence. Generally speaking, the average-cost version of naive policy iteration is "well-known" to not work, and we do not pursue it further here.

5.3 Chapter Summary

The purpose of this chapter was 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 that the existing theory for average cost games can be used to prove (easily) a subset of the results established in Chapter 3. Unfortunately, this line of reasoning does not apply to the general case where $U(i)$ and $V(i)$ are arbitrary compact subsets of metric spaces. On the other hand, the results of Chapter 3 and 4 can be used to extend the theory of recurrent-state average cost games. We established 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 examine several dynamic programming algorithms for recurrent-state average cost games. One important conclusion to be drawn from this chapter 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.

Chapter 6

Conclusion

6.1 Summary

We have seen that stochastic shortest path games represent a very general class of stochastic games, extending to two players the stochastic shortest path problems of Bertsekas and Tsitsiklis [8]. Our formulation includes the terminating games of Shapley [52] (which includes the class of discounted cost games), the pursuit-evasion games of Kushner and Chamberlain [28], and the transient games of Filar and Vrieze [20]. Our stochastic shortest path assumption (Assumption SSP) represents the main point of deviation from the existing literature on stochastic games. It stipulates the existence of a proper policy for the minimizer which forces termination regardless of the maximizer's policy. Improper policies, which permit the maximizer to prolong the game indefinitely, are allowed as long as the resulting cost to the minimizer is infinite for at least one initial state. Our regularity assumption (Assumption R), which agrees with Kushner and Chamberlain's formulation, allows both players to choose actions at each state from arbitrary compact subsets of metric spaces. Because of this, it is necessary to impose certain continuity properties on the transition probability and cost functions. We must also make the extra assumption that the minimization and maximization in the dynamic programming operator be interchangeable. That is,

given any estimate J of the equilibrium cost-to-go function,

$$TJ \triangleq \min_{\mu \in M} \max_{\nu \in N} c(\mu, \nu) + P(\mu, \nu)J = \max_{\nu \in N} \min_{\mu \in M} c(\mu, \nu) + P(\mu, \nu)J \triangleq \tilde{T}J.$$

Since we do not require the constraint sets of the two players to be simplicial and we do not require the transition probability and cost functions to be bilinear with respect to the players action's, our regularity assumptions generalize the conventional "mixed-strategy" formulation for stochastic games.

Despite the generality of our formulation, we are able to obtain all of the standard results for stochastic games:

1. the existence of stationary equilibrium solutions,

2. a characterization of the equilibrium value function J^* as the unique solution of Bellman's equation, and

3. a characterization of equilibrium solutions for both players in terms of the policies which achieve extrema in $\min_{\mu \in M} T_\mu J^*$ and $\max_\nu \tilde{T} J^*$.

Moreover, we obtain these results despite the fact that the dynamic programming operator T is not generally a contraction mapping. Our analysis makes use of only the most basic properties of T. In particular, the monotonicity and continuity of T are used repeatedly in this thesis, and without these properties there would be no hope of establishing our main results. We are aided by the theory previously developed by Bertsekas and Tsitsiklis in [8] for single-player stochastic shortest path problems. That theory laid the groundwork for our present extension to two player, zero-sum games.

Another main result of this thesis is that the classical methods of dynamic programming are effective in solving stochastic shortest path games. We have determined that value iteration and policy iteration both converge to the equilibrium value function. The value iteration algorithm dates back to Shapley [52] who used the fact that the dynamic programming operator is a contraction (in his terminating games) to prove convergence. The policy iteration algorithm dates back to a related algorithm

by Hoffman and Karp [25] for average cost games. Convergence of the discounted cost version of policy iteration was established by Rao et al. [44]. (The earlier results of Shapley, Hoffman and Karp, and Rao et al. were derived for the case of mixed strategies over finite sets of actions.)

We note that the policy iteration algorithm generates a sequence of stationary policies for one of the players based on evaluations of worst-case cost and corresponding policy improvements. We have shown that if the policies are for the minimizing player, then they are all proper and the costs of the respective policies converge monotonically to the equilibrium value function. Since a corresponding asynchronous policy iteration also converges, we see that policy iteration exhibits a fair degree of robustness. Given some extra assumptions, we obtain an error-bound for approximate policy iteration.

Naive policy iteration (due to Pollatschek and Avi-Itzhak [40]) generates a sequence of stationary policies for both players and is easier to implement than the official form of policy iteration (at least conceptually). Unfortunately, naive policy iteration is known not to converge, even in the presence of a discount factor [59]. A well-known fix to naive policy iteration, called modified Newton's method, is due to Filar and Tolwinski [19]. It uses Armijo's rule to prevent the policy updates from being "too greedy" with respect to the 2-norm of the Bellman error function. However, due to the nonsmooth nature of the Bellman error function, Filar and Tolwinski's proof of convergence is incorrect, even for discounted cost games. We note that both naive policy iteration and modified Newton's method have an extra failure mode when applied to stochastic shortest path games: they may generate at some point a pair of policies for which the corresponding Markov chain does not terminate with probability one.

In the last chapter of the thesis we examined connections between average cost games and stochastic shortest path games. It turns out that each type of game has something to say about the other. When restricted to the case of mixed strategies, standard results from the theory of average cost games can be used to prove the existence of a solution to Bellman's equation in stochastic shortest path games. The

restriction to the case of mixed strategies is typical (if not universally observed) in the literature on average cost games. Thus, average cost games cannot tell us everything we want to know about stochastic shortest path games. On the other hand, by proceeding in the reverse direction, we can derive new results for a broad class of average cost games. Specifically, by transforming to associated stochastic shortest path games, we can analyze recurrent-state average cost games, where the associated Markov chain is unichain for all pairs of stationary policies and there exists a state which is common to each recurrent class. This type of analysis was pioneered by Bertsekas in [4]. Because of the generality of our regularity assumptions, we are able to prove the existence of equilibrium solutions in recurrent-state games when both players choose actions from arbitrary compact subsets of metric spaces. We are also able to establish the convergence of several dynamic programming algorithms. All of this can be done without resorting to the "limit discount equation approach" [20] which is dominant in recent the literature on average cost games.

6.2 Future Work

The theory of finite-state, zero-sum stochastic games is more or less complete, thanks to the efforts of many researchers over the years. However, there remains at least one important open question:

What is there to say about naive policy iteration and its variations?

As we have discussed, it is easy to show that this algorithm does not always converge, but these examples are highly contrived. In practice, naive policy iteration seems to work quite well, and when it converges, it does so very quickly (as in Newton's method). A convincing way of rectifying these observations has eluded researchers for many years, and this thesis provides no new insight. Modified Newton's method does seem to improve the convergence properties of naive policy iteration. (For example, it converges in games where naive policy iteration fails.) On the other hand, as shown in

our computational example (the inspection game), it is often quite slow and produces a sequence of iterates which look nothing like that produced by naive policy iteration.

Other open questions remain about algorithms for stochastic shortest path games. In this thesis we have focused primarily on conventional dynamic programming algorithms: value iteration and policy iteration, and their closest variants. However, a number of other algorithms are possible, some of which have been proven to converge in the case of discounted cost games. Appendix B contains a list of some alternative algorithms. The convergence properties of these algorithms as applied to stochastic shortest path games have yet to be determined.

Many open questions remain regarding extensions of the stochastic shortest path model. For example, N-player, nonzero-sum, and infinite state versions of stochastic shortest path games are possible and largely unexplored. Establishing the existence of equilibrium solutions, characterizing these solutions, and developing appropriate algorithms should be interesting topics for future research in this area.

Appendix A

Proofs of Lemmas

For all of the results of this appendix we assume that Assumption R holds. We do not require Assumption SSP (unless specifically stated otherwise). We remind the reader of the following notation from Chapter 2 [cf. equation (2.2)]:

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

We will use $h^t_{\pi_M,\pi_N}$ to denote the vector in $\mathcal{J}$ whose components are $h^t_{\pi_M,\pi_N}(i)$ for $i = 1,\ldots,n$.

Lemma A.1 (Monotonicity) *Given $J, \bar{J} \in \mathcal{J}$, if $J \leq \bar{J}$, then*

$$TJ \leq T\bar{J}.$$

The same is true of the other dynamic programming operators.

Proof: Suppose $J \leq \bar{J} \in \mathcal{J}$. If we are given $\mu \in M$ and $\nu \in N$, then

$$\begin{aligned} T_{\mu\nu}J &= c(\mu,\nu) + P(\mu,\nu)J \\ &\leq c(\mu,\nu) + P(\mu,\nu)\bar{J} \\ &= T_{\mu\nu}\bar{J}. \end{aligned}$$

Given $\mu \in M$, Assumption R implies that there exists $\nu \in N$ such that $T_\mu J = T_{\mu\nu} J$. Thus,

$$\begin{aligned} T_\mu J = T_{\mu\nu} J &\leq T_{\mu\nu} \bar{J} \\ &\leq T_\mu \bar{J}. \end{aligned}$$

Similarly, Assumption R implies that there exists $\mu \in M$ such that $T\bar{J} = T_\mu \bar{J}$. Thus,

$$\begin{aligned} TJ &\leq T_\mu J \\ &\leq T_\mu \bar{J} = T\bar{J}. \end{aligned}$$

Similar arguments apply for the operators $\tilde{T}_\nu$ (given $\nu \in N$) and $\tilde{T}$, showing that $\tilde{T}_\nu J \leq \tilde{T}_\nu \bar{J}$ and $\tilde{T}J \leq \tilde{T}\bar{J}$. **Q.E.D.**

Lemma A.2 *Given $J \in \mathcal{J}$ and a positive scalar r,*

$$T(J + r\mathbf{1}) \leq TJ + r\mathbf{1}.$$

The same inequality holds for the other dynamic programming operators. The inequalities are reversed if $r < 0$.

Proof: Suppose $J \in \mathcal{J}$ and $r > 0$. For every $\mu \in M$ and $\nu \in N$, we have

$$T_{\mu\nu}(J + r\mathbf{1}) = c(\mu,\nu) + P(\mu,\nu) \cdot (J + r\mathbf{1}) \leq (c(\mu,\nu) + P(\mu,\nu)J) + r\mathbf{1} = T_{\mu\nu}J + r\mathbf{1}.$$

(The inequality holds because the row sums of $P(\mu,\nu)$ are less than or equal to one.) Consequently, given $\mu \in M$,

$$T_\mu(J + r\mathbf{1}) = \max_{\nu \in N} T_{\mu\nu}(J + r\mathbf{1}) \leq \max_{\nu \in N} T_{\mu\nu}J + r\mathbf{1} = T_\mu J + r\mathbf{1}.$$

Therefore,

$$T(J + r\mathbf{1}) = \min_{\mu \in M} T_\mu(J + r\mathbf{1}) \leq \min_{\mu \in M} T_\mu J + r\mathbf{1} = TJ + r\mathbf{1}.$$

Similar arguments hold for the operators $\tilde{T}_\nu$ and $\tilde{T}$, showing that $\tilde{T}_\nu(J+r\mathbf{1}) \leq \tilde{T}_\nu J+r\mathbf{1}$ and $\tilde{T}(J + r\mathbf{1}) \leq \tilde{T}J + r\mathbf{1}$.

The case that $r < 0$ is handled analogously. **Q.E.D.**

Lemma A.3 (Continuity) *Given $J, \bar{J} \in \mathcal{J}$, then*

$$\|T(J - \bar{J})\|_\infty \leq \|J - \bar{J}\|_\infty.$$

Thus, T is nonexpansive on $\mathcal{J}$ and therefore continuous. The same is true of the other dynamic programming operators.

Proof: Let J and $\bar{J}$ be any two elements of $\mathcal{J}$, and let $r = \|J - \bar{J}\|_\infty$, where $\|\cdot\|_\infty$ denotes the usual sup-norm on $\mathcal{J}$. Then,

$$J - r\mathbf{1} \leq \bar{J} \leq J + r\mathbf{1},$$

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

$$TJ - r\mathbf{1} \leq T\bar{J} \leq TJ + r\mathbf{1}.$$

Thus,

$$\|TJ - T\bar{J}\|_\infty \leq \|J - \bar{J}\|_\infty,$$

which shows that T is nonexpansive on $\mathcal{J}$ and therefore continuous. Similar arguments hold for the other dynamic programming operators. **Q.E.D.**

Lemma A.4 *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 [8]. 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 [8]. (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 [8] 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.**

Lemma A.5 *For any $(n \times n)$ matrix of nonnegative elements $\bar{P}$ and any $J \in \mathcal{J}$,*

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

Proof: It is sufficient to show that the first equation holds. The remaining equations follow as a corollary by redefining the control constraint sets for the respective players as $U(i) = \{\mu(i)\}$ and $V(i) = \{\nu(i)\}$.

The i-th component of $\bar{P}\left[c(\mu,\nu) + P(\mu,\nu)J\right]$ 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 $g_s(u,v) = c_s(u,v) + \sum_{j=1}^{n} p_{sj}(u,v)J(j)$ 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

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

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

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

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

Lemma A.6 *For every $J \in \mathcal{J}$,*

$$\begin{aligned}\min_{\pi_M=\{\mu_0,\ldots,\mu_t\}}\max_{\pi_N=\{\nu_0,\ldots,\nu_t\}}\left[h^t_{\pi_M,\pi_N}+P(\mu_0,\nu_0)\cdots P(\mu_t,\nu_t)J\right] &= T^{t+1}J,\\ \max_{\pi_N=\{\nu_0,\ldots,\nu_t\}}\left[h^t_{\mu,\pi_N}+P(\mu,\nu_0)\cdots P(\mu,\nu_t)J\right] &= T^{t+1}_{\mu}J,\\ \min_{\pi_N=\{\mu_0,\ldots,\mu_t\}}\left[h^t_{\pi_M,\nu}+P(\mu_0,\nu)\cdots P(\mu_t,\nu)J\right] &= \tilde{T}^{t+1}_{\nu}J,\end{aligned}$$

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 a corollary by redefining the control constraint sets for the respective players.

Notice that

$$\begin{aligned}&\min_{\pi_M=\{\mu_0,\ldots,\mu_t\}}\max_{\pi_N=\{\nu_0,\ldots,\nu_t\}}\left[h^t_{\pi_M,\pi_N}+P(\mu_0,\nu_0)\cdots P(\mu_t,\nu_t)J\right]\\ &= \min_{\pi_M=\{\mu_0,\ldots,\mu_t\}}\max_{\pi_N=\{\nu_0,\ldots,\nu_t\}}\left\{h^{t-1}_{\pi_M,\pi_N}+\bar{P}\left[c(\mu_t,\nu_t)+P(\mu_t,\nu_t)J\right]\right\}\end{aligned}$$

$$
\begin{aligned}
&= \min_{\pi_M=\{\mu_0,\ldots,\mu_t\}} \max_{\pi_N=\{\nu_0,\ldots,\nu_{t-1}\}} \left\{ h^{t-1}_{\pi_M,\pi_N} + \max_{\nu_t} \bar{P}\left[c(\mu_t,\nu_t) + P(\mu_t,\nu_t)J\right]\right\} \\
&= \min_{\pi_M=\{\mu_0,\ldots,\mu_{t-1}\}} \min_{\mu_t} \max_{\pi_N\{\nu_0,\ldots,\nu_{t-1}\}} \left\{ h^{t-1}_{\pi_M,\pi_N} + \max_{\nu_t} \bar{P}\left[c(\mu_t,\nu_t) + P(\mu_t,\nu_t)J\right]\right\} \\
&\geq \min_{\pi_M=\{\mu_0,\ldots,\mu_{t-1}\}} \max_{\pi_N=\{\nu_0,\ldots,\nu_{t-1}\}} \min_{\mu_t} \left\{ h^{t-1}_{\pi_M,\pi_N} + \max_{\nu_t} \bar{P}\left[c(\mu_t,\nu_t) + P(\mu_t,\nu_t)J\right]\right\} \\
&= \min_{\pi_M=\{\mu_0,\ldots,\mu_{t-1}\}} \max_{\pi_N\{\nu_0,\ldots,\nu_{t-1}\}} \left\{ h^{t-1}_{\pi_M,\pi_N} + \min_{\mu_t}\max_{\nu_t} \bar{P}\left[c(\mu_t,\nu_t) + P(\mu_t,\nu_t)J\right]\right\},
\end{aligned}
$$

where $\bar{P} = P(\mu_0,\nu_0)\cdots P(\mu_{t-1},\nu_{t-1})$. (The inequality follows from the minimax inequality.)

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

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

To see this, notice that

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

where the first and last equalities follows from the preceding lemma and $\bar{\mu}$ is the minimax solution to $\min_{\mu_t\in M}\max_{\nu_t\in N}\left[c(\mu_t,\nu_t) + P(\mu_t,\nu_t)J\right]$. (The existence of $\bar{\mu}$ follows from Assumption R.) Thus,

$$
\begin{aligned}
&\min_{\pi_M=\{\mu_0,\ldots,\mu_{t-1}\}} \min_{\mu_t} \max_{\pi_N=\{\nu_0,\ldots,\nu_{t-1}\}} \left\{ h^{t-1}_{\pi_M,\pi_N} + \max_{\nu_t} \bar{P}\left[c(\mu_t,\nu_t) + P(\mu_t,\nu_t)J\right]\right\} \\
&= \min_{\pi_M=\{\mu_0,\ldots,\mu_{t-1}\}} \max_{\pi_N=\{\nu_0,\ldots,\nu_{t-1}\}} \left\{ h^{t-1}_{\pi_M,\pi_N} + \min_{\mu_t}\max_{\nu_t} \bar{P}\left[c(\mu_t,\nu_t) + P(\mu_t,\nu_t)J\right]\right\} \\
&\leq \min_{\pi_M=\{\mu_0,\ldots,\mu_{t-1}\}} \max_{\pi_N=\{\nu_0,\ldots,\nu_{t-1}\}} \left\{ h^{t-1}_{\pi_M,\pi_N} + \max_{\nu_t} \bar{P}\left[c(\bar{\mu},\nu_t) + P(\bar{\mu},\nu_t)J\right]\right\} \\
&= \min_{\pi_M=\{\mu_0,\ldots,\mu_{t-1}\}} \max_{\pi_N=\{\nu_0,\ldots,\nu_{t-1}\}} \left\{ V_t + \min_{\mu_t}\max_{\nu_t} \bar{P}\left[c(\mu_t,\nu_t) + P(\mu_t,\nu_t)J\right]\right\}.
\end{aligned}
$$

Combining these inequalities, we obtain

$$\begin{aligned}
&\min_{\pi_M=\{\mu_0,\ldots,\mu_t\}} \max_{\pi_N=\{\nu_0,\ldots,\nu_t\}} \left[h^t_{\pi_M,\pi_N} + P(\mu_0,\nu_0)\cdots P(\mu_t,\nu_t)J\right] \\
&\quad = \min_{\pi_M=\{\mu_0,\ldots,\mu_{t-1}\}} \max_{\pi_N=\{\nu_0,\ldots,\nu_{t-1}\}} \left\{h^{t-1}_{\pi_M,\pi_N} + \min_{\mu_t}\max_{\nu_t} \bar{P}\left[c(\mu_t,\nu_t) + P(\mu_t,\nu_t)J\right]\right\} \\
&\quad = \min_{\pi_M=\{\mu_0,\ldots,\mu_{t-1}\}} \max_{\pi_N=\{\nu_0,\ldots,\nu_{t-1}\}} \left\{h^{t-1}_{\pi_M,\pi_N} + \bar{P}\min_{\mu_t}\max_{\nu_t} \left[c(\mu_t,\nu_t) + P(\mu_t,\nu_t)J\right]\right\} \\
&\quad = \min_{\pi_M=\{\mu_0,\ldots,\mu_{t-1}\}} \max_{\pi_N=\{\nu_0,\ldots,\nu_{t-1}\}} \left[h^{t-1}_{\pi_M,\pi_N} + \bar{P}TJ\right] \\
&\quad = \min_{\pi_M=\{\mu_0,\ldots,\mu_{t-1}\}} \max_{\pi_N=\{\nu_0,\ldots,\nu_{t-1}\}} \left[h^{t-1}_{\pi_M,\pi_N} + P(\mu_0,\nu_0)\cdots P(\mu_{t-1},\nu_{t-1})TJ\right].
\end{aligned}$$

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

Lemma A.7 *For any $(n \times n)$ matrix with nonnegative elements $\bar{P}$ and any $J \in \mathcal{J}$*

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

Proof: The proof of this is exactly analogous to that given for Lemma A.5. The interchange of the max and min has no bearing on the logical flow of the argument. **Q.E.D.**

Lemma A.8 *For any $J \in \mathcal{J}$,*

$$\max_{\pi_N=\{\nu_0,\ldots,\nu_t\}} \min_{\pi_M=\{\mu_0,\ldots,\mu_t\}} \left[h^t_{\pi_M,\pi_N} + P(\mu_0,\nu_0)\cdots P(\mu_t,\nu_t)J\right] \quad = \quad \tilde{T}^{t+1}J,$$

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.6. **Q.E.D.**

Lemma A.9 *Given a proper policy μ, the following are true.*

1. $J_\mu \stackrel{\triangle}{=} \liminf_{t\to\infty} \max_{\pi_N\in\bar{N}} h^t_{\mu,\pi_N}$ *is the unique fixed point of* T_μ *within* $\mathcal{J}$.

2. $J_\mu = \sup_{\pi_N\in\bar{N}} J_{\mu,\pi_N}$.

3. *We have* $T^t_\mu J \to J_\mu$ *for all* $J \in \mathcal{J}$, *with geometric convergence.*

Proof: Lemma A.6 implies that

$$T^{t+1}_\mu \mathbf{0} = \max_{\pi_N=\{\nu_0,\ldots,\nu_t\}} h^t_{\mu,\pi_N},$$

where $\mathbf{0}$ is the zero vector in $\mathcal{J}$. Thus, from Corollary 3.1 and the definition of J_μ, it is clear that

$$J_\mu = \lim_{t\to\infty} T^{t+1}_\mu \mathbf{0},$$

and J_μ is the unique fixed point of the contraction mapping T_μ within $\mathcal{J}$.

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

$$\sup_{\pi_N\in\bar{N}} \liminf_{t\to\infty} \left\{ c(\mu,\nu_0) + \sum_{k=1}^{t} [P(\mu,\nu_0)\cdots P(\mu,\nu_{k-1})]c(\mu,\nu_k) \right\}.$$

This problem is covered by the theory developed in [8] 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 $\sup_{\pi_N\in\bar{N}} J_{\mu,\pi_N}$, and (according to the theory of [8]) it is equal to the limit of value iteration applied to this problem, which is $\lim_{t\to\infty} T^{t+1}_\mu \mathbf{0}$. Thus, $\sup_{\pi_N\in\bar{N}} J_{\mu,\pi_N}$ is equal to the unique fixed point of T_μ.

Finally, the linear convergence of $T^{t+1}_\mu J$ follows from the fact that T_μ is a contraction. **Q.E.D.**

Lemma A.10 *In addition to Assumption R, let Assumption SSP hold. Then, for any $\nu \in N$, the following are true.*

1. $J_\nu \stackrel{\triangle}{=} \liminf_{t\to\infty} \min_{\pi_M \in \bar{M}} h^t_{\mu,\pi_N}$ *is the unique fixed point of $\tilde{T}_\nu$ within $\mathcal{J}$.*
2. $J_\nu = \inf_{\pi_M \in \bar{M}} J_{\pi_M,\nu}$.
3. *We have $\tilde{T}^t_\nu J \to J_\nu$ for all $J \in \mathcal{J}$. 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 (single-player) stochastic shortest path problems. [8] **Q.E.D.**

Lemma A.11 *In addition to Assumption R let Assumption SSP hold. Given $\mu \in M$, if there exists $J \in \mathcal{J}$ such that $J \geq T_\mu J$, then μ is proper.*

Proof: To reach a contradiction, suppose μ is improper. According to Assumption SSP and Lemma A.4, there exists a stationary policy $\bar{\nu}$ for the maximizer such that $(\mu,\bar{\nu})$ is prolonging and results in unbounded expected cost from some initial state when played against μ. Thus, some subsequence of

$$\left\{\sum_{k=0}^{t} [P(\mu,\bar{\nu})]^k\, c(\mu,\bar{\nu})\right\}_{t=1}^{\infty} \tag{A.1}$$

must have a coordinate that tends to infinity.

On the other hand, there exists (by hypothesis) $J \in \mathcal{J}$ such that $J \geq T_\mu J$. Applying T_μ to J, we have that

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

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

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

where the last inequality follows again from the definition of T_μ. Proceeding inductively (using the same steps) we obtain

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

Since the matrices $P(\mu,\bar{\nu})^{t+1}$ are stochastic, the term involving J remains bounded. Thus, we have contradicted our earlier observation about (A.1), and μ must be proper. **Q.E.D.**

Appendix B

Other Algorithms

In this appendix we collect a few alternative algorithms for stochastic shortest path games. To be consistent with the framework in which these algorithms were originally developed, we include a discount factor $\alpha \in (0, 1]$. Generally, questions about convergence remain for the full generality of Assumptions R and SSP (with $\alpha = 1$). In some cases, these algorithms are meaningful only in the context of mixed strategies over finite sets of actions.

B.1 Q-learning

In [30], Littman proposed Minimax-Q, a simulation-based algorithm for games in mixed strategies over finite sets of actions (based on Q-learning). In our statement of the algorithm below, let $A(i)$ and $B(i)$ denote the finite sets of underlying actions available to the minimizer and maximizer (respectively) at state i, with $U(i)$ and $V(i)$ the corresponding sets of probability distributions.

Algorithm B.1.1 *(Q-learning [30])*

1. *Let $Q_0(i, a, b)$ be an initial estimate of optimal long-term cost of the minimizer applying $a \in A(i)$ and the maximizer applying $b \in B(i)$ at state $i \in S$. Let γ_t be a positive stepsize factor for which $\sum_t \gamma_t = \infty$ and $\sum_t (\gamma_t)^2 < \infty$.*

2. *Given Q_{k-1},*

(a) Pick an element $i_k \in S$, and pick actions $a_k \in A(i_k)$ and $b_k \in B(i_k)$. (Make sure that each triple is visited infinitely often.)

(b) Realize a successor state $\bar{i}$ and transition cost $\bar{c}$ from i_k under the actions a_k and b_k.

(c) Compute a new Q-factor estimate Q_k as:

$$Q_k(i,a,b) = \begin{cases} & \text{if } i = i_k, \\ (1-\gamma_t)Q^{k-1}(i_k,a_k,b_k) + \gamma_t(\bar{c} + \alpha J_{k-1}(\bar{i})) & a = a_k, \\ & b = b_k \\ \\ Q_{k-1}(i,a,b) & \text{otherwise} \end{cases}$$

where,

$$J_{k-1}(i) = \min_{u \in U(i)} \max_{v \in V(i)} \sum_{a \in A(i)} \sum_{b \in B(i)} Q_{k-1}(i,a,b) u_a v_b. \tag{B.1}$$

This algorithm is interesting in that it may be used in situations the game parameters are not available explicitly but are built into a simulator which can be used to generate sample data for the game. It was shown in [31] and [9] that J_k converges with probability one to J^* in discounted cost games. Note that the finiteness of the sets $A(i)$ and $B(i)$ is essential. An extension to the general case of compact constraint sets does not seem possible.

Note that it is easy to evaluate J_{k-1} in sequential games since at least one of the control constraint sets $A(i)$ or $B(i)$ is a singleton. Thus, it is no longer necessary to extend the matrix game $Q_{k-1}(i,\cdot,\cdot)$ to mixed strategies, and (B.1) may be replaced with

$$J_{k-1}(i) = \min_{a \in A(i)} \max_{b \in B(i)} Q_{k-1}(i,a,b).$$

where at least one of the suprema is degenerate.

In *symmetric* sequential games (cf. Section 2.3.3), a further simplification may be employed: we only need to maintain Q-factor estimates on $\tilde{S}$.

Algorithm B.1.2 *(Symmetric Q-learning [31])*

1. *Let $Q_0(i,a)$ be an initial estimate of optimal long-term cost of player 1 implementing action $a \in W(i)$ at position $i \in \tilde{S}$. Let γ_t be a positive stepsize factor for which $\sum_t \gamma_t = \infty$ and $\sum_t (\gamma_t)^2 < \infty$.*

2. *Given Q_{k-1},*

   (a) *Pick an element $i_k \in \tilde{S}$ and an action $a_k \in W(i_k)$. (Make sure that each double is visited infinitely often.)*

   (b) *Realize a successor state $(\bar{z}, \bar{i})$ and transition cost $\bar{c}$ from $(1, i_k)$ under the action a_k.*

   (c) *Compute a new Q-factor estimate Q_k as:*

$$Q_k(i,a) = \begin{cases} (1-\gamma_t)Q_{k-1}(i_k,a_k) + \gamma_t\left[\bar{c} + \alpha(-1)^{\bar{z}-1}J_{k-1}(\bar{i})\right] & i = i_k,\ a = a_k \\ Q_{k-1}(i,a) & \text{otherwise} \end{cases}$$

   where,

$$J_{k-1}(i) = \min_{a \in W(i)} Q_{k-1}(i,a).$$

B.2 Fictitious Play and Sequential Improvement

In [12], Brown proposed an iterative algorithm (fictitious play) for computing the value of matrix games in mixed strategies. The algorithm proceeds as an infinite sequence of fictitious realizations of the game, where the players make decisions which are optimal with respect to running estimates of the equilibrium value and the other player's best action. The convergence of this algorithm (for matrix games) was established by Robinson in [45]. Later, Vrieze and Tijs [64] studied an extension (also called fictitious play) for discounted cost stochastic games.

Algorithm B.2.1 *(Fictitious Play [64])*

1. *Let $R_0(i,a)$ be an initial estimate of the minimizer's "worst case" cost of applying the pure action $a \in A(i)$. Let $S_0(i,v)$ be an initial estimate of the maxi-*

mizer's "worst case" long-term cost of applying the pure action $b \in B(i)$. Make sure that

$$\begin{aligned} \min_{a \in A(i)} R_0(i,a) &= \max_{b \in B(i)} S_0(i,b), \\ \min_{b \in B(i)} S_0(i,b) &\geq J^*(i). \end{aligned}$$

for every state i. Define $J_0 \in \mathcal{J}$ such that $J_0(i) = \max_{b \in B(i)} S_0(i,b)$. Let $\gamma_k = 1/k$ be a stepsize rule.

2. *Given J_{k-1}, R_{k-1}, and S_{k-1}, compute new iterates as follows:*

$$\begin{aligned} a_k(i) &\in \arg\min_{a \in A(i)} R_{k-1}(i,a) \\ b_k(i) &\in \arg\max_{b \in B(i)} S_{k-1}(i,b) \\ J_k(i) &= \min\{ \max_{b \in B(i)} S_{k-1}(i,b),\ J_{k-1}(i) \}, \end{aligned}$$

$$R_k(i,a) = (1-\gamma_k) R_{k-1}(i,a) + \gamma_k \left[c_i(a, b_k(i)) + \alpha \sum_{j \in 1}^{n} p_{ij}(a, b_k(i)) J_k(j) \right],$$

$$S_k(i,b) = (1-\gamma_k) S_{k-1}(i,b) + \gamma_k \left[c_i(a_k(i), b) + \alpha \sum_{j \in 1}^{n} p_{ij}(a_k(i), b) J_k(j) \right].$$

In studying this algorithm, Vrieze and Tijs showed that the matrix-game version of fictitious play also applies to a convergent sequence of matrix games (as would be defined by an evolving estimate of the equilibrium cost-to-go in a stochastic game). In the end, they showed that J_k converges with probability one to J^* in discounted cost games.

Before we were aware of the work of Vrieze and Tijs [64], we independently proposed a closely related algorithm, called sequential improvement. In stating this algorithm we make use of two new operators $ArgT_\mu$ and $Arg\tilde{T}_\nu$. Given a proper policy $\mu \in M$ and a cost function estimate $J \in \mathcal{J}$, we define $(ArgT_\mu J)(i)$ to be a uniquely-determined degenerate probability distribution which achieves the maximum in $(T_\mu J)(i)$. We define $(Arg\tilde{T}_\nu J)(i)$ similarly.

Algorithm B.2.2 *(Sequential Improvement)*

1. *Start with an initial estimate of the equilibrium cost function J_0, an initial proper policy for the minimizer $\mu_0 \in M$, and an initial stationary policy for the maximizer $\nu_0 \in N$. Let $\gamma_k > 0$ be a decreasing stepsize rule such that $\sum_k \gamma_k = \infty$ and $\sum_k (\gamma_k)^2 < \infty$.*

2. *Given $(J_{k-1}, \mu_{k-1}, \nu_{k-1})$,*

   (a) *Compute an intermediate estimate of the equilibrium cost function $\bar{J}_{k-1} \in \mathcal{J}$ according to*

   $$\bar{J}_{k-1} = T_{\mu_{k-1}} J_{k-1}.$$

   (b) *Update the maximizer's policy according to*

   $$\nu_k(i) = (1-\gamma_k)\nu_{k-1}(i) + \gamma_k \left(ArgT_{\mu_{k-1}} J_{k-1}\right)(i), \quad i = 1, \ldots, n.$$

   (c) *Compute the new equilibrium cost function estimate J_k according to*

   $$J_k = \tilde{T}_{\nu_k} \bar{J}_{k-1}. \tag{B.2}$$

   (d) *Update the minimizer's policy according to*

   $$\mu_k(i) = (1-\gamma_k)\mu_{k-1}(i) + \gamma_k \left(Arg\tilde{T}_{\nu_k} \bar{J}_{k-1}\right)(i), \quad i = 1, \ldots, n.$$

What motivates this algorithm is the interpretation of J_k as an approximation of $T^2 J_{k-1}$ for each k. Note that a number of variations of this algorithm are possible. One possibility is to use ν_{k-1} in place of ν_k in (B.2). Another possibility is to reverse the order in which the maximizer's and minimizer's policies are updated. There are many other possibilities, including that of "asynchronizing" the various types of updates, as in Asynchronous Policy Iteration.

B.3 Approximate Naive Policy Iteration

In Section 4.1.4, we described an approximate version of the policy iteration algorithm. Here we present an approximate version of naive policy iteration. (Caveat emptor! Recall from Section 4.1.5 that (exact) naive policy iteration fails to converge in some examples.) The interest in approximate naive policy iteration generally comes from the reinforcement learning and artificial intelligence research communities. Indeed, the algorithm we present below is already well-known within the context of Neuro-Dynamic Programming (NDP) [9].

Algorithm B.3.1 *(Naive Policy Iteration with Function Approximation [9])*

1. *Choose an initial proper policy μ_0 for the minimizer, an initial stationary policy ν_0 for the maximizer, and an initial parameter vector r^0. (Alternatively, start with just an initial parameter vector r^0 and skip to step 2(b).)*

2. *Given r^{k-1} and the policies $\mu_{k-1} \in M$, $\nu_{k-1} \in N$:*

   (a) *(Approximate Policy Evaluation)*

      i. *Use a simulation to generate sample state/ cost-to-go data $\mathcal{D}_k$ under the policies $\mu_{k-1} \in M$, $\nu_{k-1} \in N$.*

      ii. *Let r^k be the end result of the training algorithm:*

$$r^k = TrainingAlgorithm(\mathcal{D}_k, r^{k-1}).$$

      The parameter vector r^k should be such that either $\tilde{J}(\cdot, r^k) \approx J_{\mu_{k-1},\nu_{k-1}}(\cdot)$ or $(r^k - r^{k-1})$ is a step toward obtaining such an approximation.

   (b) *(Approximate Policy Improvement)*

      i. *Compute $\mu_k \in M$ such that $T\tilde{J}(\cdot, r^k) \approx T_{\mu_k}\tilde{J}(\cdot, r^k)$.*

      ii. *Compute $\nu_k \in N$ such that $\tilde{T}\tilde{J}(\cdot, r^k) \approx \tilde{T}_{\nu_k}\tilde{J}(\cdot, r^k)$.*

In general, $J(\cdot, r^k)$ represents an approximation of $J_{\mu_{k-1},\nu_{k-1}}(\cdot)$. Note that it is not necessary for the policies μ_k and ν_k to be explicitly computed and stored (in appropriate data structures); all that is needed is the ability to do on-line Monte Carlo

simulation of the various control options (at each stage) to determine the actions which are best with respect to $J(\cdot, r^k)$. There are two basic modes of operation for this algorithm:

1. *approximate*, where a large amount of sample data is generated for each pair of policies and the training algorithm is rigorous enough to yield an accurate approximation of $J_{\mu_{k-1},\nu_{k-1}}$, and

2. *optimistic*, where very little training data is generated and r^k usually represents a small change to r^{k-1}.

While Algorithm B.3.1 has been discussed in the literature [9], it has not (as far as we know) been used in practice. On the other hand, a variation on this algorithm has seen considerable use. Specifically, Tesauro's TD-Gammon [56, 57] (a computer backgammon playing program) taught itself to play at a world competition level using an (optimistic) approximate version of symmetric policy iteration. We state this algorithm below.

Algorithm B.3.2 *(Symmetric Policy Iteration with Function Approximation)*

1. *Choose an initial proper policy $\mu_0 \in M$ for the minimizer. (The maximizer will implicitly "play" the μ-symmetric policy ν_μ.) (Alternatively, start with an initial parameter vector r^0 and skip to step 2(b).)*

2. *Given r^{k-1} and the policy $\mu_{k-1} \in M$:*

   (a) (Approximate Symmetric Policy Evaluation)

      i. Generate sample data $\mathcal{D}_k$ under the policy $\mu_{k-1} \in M$.

      ii. Let r^k be the end result of the training algorithm:

$$r^k = TrainingAlgorithm(\mathcal{D}_k, r^{k-1}).$$

   (b) (Approximate Symmetric Policy Improvement) Compute $\mu_k \in M$ such that

$$Z\tilde{X}(\cdot, r^k) \approx Z_{\mu_k}\tilde{X}(\cdot, r^k).$$

B.4 Approximate Q-learning

In [9], Bertsekas proposed the following approximate form of Q-learning.

Algorithm B.4.1 *(Minimax-Q with Function Approximation [9])*

1. *Let $\tilde{Q}(i,a,b,r^0)$ be an initial approximation of the equilibrium long-term cost of the minimizer applying $a \in A(i)$ and the maximizer applying $b \in B(i)$ at state $i \in S$. Let γ_t be a positive stepsize factor for which $\sum_t \gamma_t = \infty$ and $\sum_t (\gamma_t)^2 < \infty$.*

2. *Given r^{k-1},*

   (a) *Pick an element $i_k \in S$, and pick actions $a_k \in A(i_k)$, and $b_k \in B(i_k)$.*

   (b) *Realize a successor state $\bar{i}$ and transition cost $\bar{c}$ from i_k under the actions a_k and b_k.*

   (c) *Compute a new parameter vector r^k for the equilibrium Q-factor approximation as:*

$$r^k = r^{k-1} - \gamma_k \nabla_r E(i_k, a_k, b_k, \bar{i}, \bar{c}, r^{k-1}),$$

   where

$$E(i,a,b,j,c,r) = \frac{1}{2}\left[c + \alpha \tilde{J}(j,r) - \tilde{Q}(i,a,b,r)\right]^2$$

   and

$$\tilde{J}(i,r) = \min_{u \in U(i)} \max_{v \in V(i)} \sum_{a \in A(i)} \sum_{b \in B(i)} \tilde{Q}(i,a,b,r) u_a v_b.$$

One problem with this algorithm (as stated) is that E is not generally differentiable for all values of the parameter vector r. In particular, E fails to be differentiable wherever the minimax solution in the evaluation of $\tilde{J}$ is not unique. The usual approach in NDP practice is to ignore this possibility and use a subgradient in place of the gradient and hope for the best. As with the exact version of this algorithm, there are important simplifications for the cases of sequential and symmetric-sequential games.

Bibliography

[1] A. Arapostathis, V. S. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh, and S. I. Marcus. Discrete-time Controlled Markov Processes with Average Cost Criterion: a Survey. *SIAM Journal on Control and Optimization*, 31(2):282–344, 1993.

[2] J. Bather. Optimal Decision Procedures for Finite Markov Chains. Part II: Communicating Systems. *Advances in Applied Probability*, 5:521–540, 1973.

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

[4] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, Belmont, MA, 1995.

[5] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 1995.

[6] D. P. Bertsekas. A New Value Iteration Method for the Average Cost Dynamic Programming Problem. Preprint: to appear in SIAM Journal on Optimization. (An older version is available as LIDS-P-2307, Laboratory for Information and Decision Systems, MIT, Cambridge MA), October 1996.

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

[8] D. P. Bertsekas and J. N. Tsitsiklis. Analysis of Stochastic Shortest Path Problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

[9] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, Belmont, MA, 1996.

[10] T. Bewley and E. Kohlberg. The Asymptotic Theory of Stochastic Games. *Mathematics of Operations Research*, 1(3):197–208, August 1976.

[11] D. Blackwell and T. S. Ferguson. The Big Match. *The Annals of Mathematical Statistics.*, 39(1):159–163, 1968.

[12] G. W. Brown. Some Notes on Computation of Games Solutions. Technical Report P-78, The RAND Corporation, Santa Monica, CA, April 1949.

[13] R. Dekker. Counter Examples for Compact Action Markov Decision Chains with Average Reward Criteria. *Communications in Statistics: Stochastic Models*, 3(3):357–368, 1987.

[14] E. A. Fainberg. On Controlled Finite State Markov Processes with Compact Control Sets. *Theory of Probability and its Applications*, 20(4):856–862, 1975. (SIAM translation of this journal has same title, published in 1976.).

[15] E. A. Fainberg. The Existence of a Stationary ϵ- Optimal Policy for a Finite Markov Chain. *Theory of Probability and its Applications*, 23(2):297–313, 1978. (SIAM translation of this journal has same title, published in 1979.).

[16] A. Federgruen. *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).

[17] A. Federgruen, P. J. Schweitzer, and H. C. Tijms. Denumerable Undiscounted Semi-Markov Decision Processes with Unbounded Rewards. *Mathematics of Operations Research*, 8(2):298–313, 1983.

[18] A. Federgruen and H. C. Tijms. The Optimality Equation in Average Cost Denumerable State Semi-Markov Decision Problems, Recurrency Conditions and Algorithms. *Journal of Applied Probability*, 15:356–373, 1978.

[19] J. A. Filar and B. Tolwinski. On the Algorithm of Pollatschek and Avi-Itzhak. In T. E. S. Raghavan et al., editors, *Stochastic Games and Related Topics, In Honor of Professor L. S. Shapley*, pages 59–70. Kluwer, Dordrecht, 1991.

[20] J. A. Filar and K. Vrieze. *Competitive Markov Decision Processes.* Springer, New York, 1997.

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

[22] M. E. Harmon and L. C. Baird. Multi-agent residual advantage learning with general function approximation. Technical Report WL-TR-96-1065, Wright-Patterson Air Force Base Ohio: Wright Laboratory, 1996.

[23] M. E. Harmon, L. C. Baird, and A. H. Klopf. Advantage Updating Applied to a Differential Game. In G. J. Tesauro et al., editors, *Advances in Neural Information Processing Systems*, volume 7, pages 353–360. MIT Press, Cambridge, MA, 1995.

[24] M. E. Harmon, L. C. Baird, and A. H. Klopf. Reinforcement Learning Applied to a Differential Game. *Adaptive Behavior*, 4(1):3–28, 1996.

[25] A. K. Hoffman and R. M. Karp. On Nonterminating Stochastic Games. *Management Science*, 12(5):359–370, January 1966.

[26] 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(1):163–176, February 1969.

[27] P. R. Kumar and T. H. Shaiu. Existence of Value and Randomized Strategies in Zero-Sum Discrete-Time Stochastic Dynamic Games. *SIAM Journal on Control and Optimization*, 19(5):617–634, September 1981.

[28] H. J. Kushner and S. G. Chamberlain. Finite State Stochastic Games: Existence Theorems and Computational Procedures. *IEEE Transactions on Automatic Control*, AC-14(3):248–255, 1969.

[29] T. M. Liggett and S. A. Lippman. Stochastic Games with Perfect Information and Time Average Payoff. *SIAM Review*, 11(4):604–607, October 1969.

[30] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In Morgan Kaufmann, editor, *Proceedings of the Eleventh International Conference on Machine Learning* , pages 157–163, San Fransisco, CA, 1994.

[31] M. L. Littman. *Algorithms for Sequential Decision Making.* PhD thesis, Brown University, May 1996.

[32] A. Maitra and T. Parthasarathy. On Stochastic Games. *Journal on Optimization Theory and Applications*, 5:289–300, 1970.

[33] A. Maitra and T. Parthasarathy. On Stochastic Games II. *Journal on Optimization Theory and Applications*, 8:154–160, 1971.

[34] A. Martin-Löf. Existence of a Stationary Control for the Markov Chain Maximizing the Average Reward. *Operations Research*, 15:866–871, 1967.

[35] G. P. McCormick. *Nonlinear Programming.* Wiley, New York, 1983.

[36] J. F. Mertens and A. Neyman. Stochastic Games. *Inernational Journal of Game Theory*, 10(2):53–66, 1980.

[37] C. A. Monash. *Stochastic Games: The minimax theorem.* PhD thesis, Department of Mathematics, Harvard University, Cambridge, MA, 1979.

[38] T. Parthasarathy. Discounted, Positive, and Noncooperative Stochastic Games. *International Journal of Game Theory*, 2(1), 1973.

[39] S. D. Patek and D. P. Bertsekas. Stochastic Shortest Path Games. Technical Report LIDS-R-2319, LIDS, Massachusetts Institute of Technology, 1996. Also accepted by SIAM Journal on Control and Optimization.

[40] M. A. Pollatschek and B. Avi-Itzhak. Algorithms for Stochastic Games with Geometrical Interpretation. *Management Science*, 10(7):399–415, 1969.

[41] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., New York, 1994.

[42] M. L. Puterman and M. C. Shin. Modified Policy Iteration Algorithms for Discounted Markov Decision Problems. *Management Science*, 24:1127–1137, 1978.

[43] T. E. S. Raghavan and J. A. Filar. Algorithms for Stochastic Games - A Survey. *Methods and Models of Operations Research (Zeitschrift für O. R.)*, 35:437–472, 1991.

[44] S. S. Rao, R. Chandrasekaran, and K. P. K. Nair. Algorithms for Discounted Stochastic Games. *Journal of Optimization Theory and Applications*, 11:627–637, 1973.

[45] J. Robinson. An Iterative Method of Solving a Game. *Annals of Mathematics*, 54(2):296–301, 1951.

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

[47] 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.).

[48] P. J. Schweitzer. On the Solvability of Bellman's Functional Equations for Markov Renewal Programming. *Journal of Mathematical Analysis and Applications*, 96:13–23, 1983.

[49] P. J. Schweitzer. On Undiscounted Markovian Decision Processes with Compact Action Spaces. *RAIRO Recherche opérationnelle*, 19(1):71–86, 1985.

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

[51] P. J. Schweitzer and A. Federgruen. The Functional Equations of Undiscounted Markov Renewal Programming. *Mathematics of Operations Research*, 3(4):308–321, November 1978.

[52] L. S. Shapley. Stochastic Games. *Proceedings of the National Academy of Sciences, Mathematics*, 39:1095–1100, 1953.

[53] M. J. Sobel. Noncooperative Stochastic Games. *The Annals of Mathematical Statistics*, 42(6):1930–1935, 1971.

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

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

[56] G. J. Tesauro. TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play. *Neural Computation*, 6:321–323, 1994.

[57] G. J. Tesauro. Temporal Differences Learning and TD-Gammon. *Communications of the ACM*, 38:58–68, 1995.

[58] M. M. Tidball and E. Altman. Approximations in Dynamic Zero-sum Games. Technical report, INRIA, Centre Sophia-Antipolis, 1993.

[59] J. van der Wal. Discounted Markov Games: Generalized Policy Iteration Method. *Journal of Optimization Theory and Applications*, 25(1):125–138, 1978.

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

[61] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, NJ, 1944.

[62] N. N. Vorob'ev. *Game Theory, Lectures for Economists and Systems Scientists.* Springer-Verlag, New York, 1977.

[63] 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.

[64] O. J. Vrieze and S. H. Tijs. Fictitious Play Applied to Sequences of Games and Discounted Stochastic Games. *International Journal of Game Theory*, 11(2):71–85, 1982.

[65] W. Whitt. Representation and Approximation of Noncooperative Sequential Games. *SIAM Journal of Control and Optimization*, 18(1):33–48, 1980.

[66] R. J. Williams and L. C. Baird. Analysis of Some Incremental Variants of Policy Iteration: First Steps Toward Understanding Actor-Critic Learning Systems. Report NU-CCS-93-11, College of Computer Science, Northeastern University, Boston, MA, 1993.