FROM MARKOV CHAINS TO STOCHASTIC GAMES

ABRAHAM NEYMAN
Hebrew University of Jerusalem
Jerusalem, Israel

1. Introduction

Markov chains[1] and Markov decision processes (MDPs) are special cases of stochastic games. Markov chains describe the dynamics of the states of a stochastic game where each player has a single action in each state. Similarly, the dynamics of the states of a stochastic game form a Markov chain whenever the players' strategies are stationary. Markov decision processes are stochastic games with a single player. In addition, the decision problem faced by a player in a stochastic game when all other players choose a fixed profile of stationary strategies is equivalent to an MDP.

The present chapter states classical results on Markov chains and Markov decision processes. The proofs use methods that introduce the reader to proofs of more general analog results on stochastic games.

2. Finite State Markov Chains

A *transition matrix* is an $n \times n$ matrix P such that all entries $P_{i,j}$ of P are nonnegative and for every $1 \leq i \leq n$ we have $\sum_{j=1}^{n} P_{i,j} = 1$.

A finite state stationary Markov chain, or *Markov chain* for short, is a discrete stochastic process $z_1, \ldots, z_t, \ldots$ with values z_t in the finite set $S = \{1, \ldots, n\}$ and such that

$$\Pr(z_{t+1} = j \mid z_1, \ldots, z_t = i) = P_{i,j},$$

where P is an $n \times n$ transition matrix.

An $n \times n$ transition matrix P together with an (initial) distribution μ on $S = \{1, \ldots, n\}$ defines a discrete stochastic process $(z_1, \ldots, z_t, \ldots)$ with

[1]We use the term Markov chain for the more explicitly termed stationary Markov chain.

values z_t in S by the following formulas.

$$\Pr(z_1 = j) = \mu(j)$$

and

$$\Pr(z_{t+1} = j \mid z_1, \ldots, z_t = i) = P_{i,j}.$$

The interpretation is that the initial state $z_1 \in S$ is chosen according to the initial distribution μ and thereafter the state in stage $t+1$ depends stochastically on z_t. The probability of moving from state i to state j equals $P_{i,j}$.

It follows by induction on t that the probability of moving in $t \geq 0$ stages from state i to state j, i.e., $\Pr(z_{s+t} = j \mid z_1, \ldots, z_s = i)$, equals $(P^t)_{i,j}$. Indeed, it holds trivially for $t = 0, 1$. By the induction hypothesis it follows that for $t > 0$ we have $\Pr(z_{s+t-1} = k \mid z_1, \ldots, z_s = i) = (P^{t-1})_{i,k}$. Therefore,

$$\begin{aligned}
\Pr(z_{s+t} = j \mid \ldots, z_s = i) &= \sum_k \Pr(z_{s+t} = j \text{ and } z_{s+t-1} = k \mid \ldots, z_s = i) \\
&= \sum_k \Pr(z_{s+t-1} = k \mid \ldots, z_s = i) P_{k,j} \\
&= \sum_k (P^{t-1})_{i,k} P_{k,j} = (P^t)_{i,j}.
\end{aligned}$$

We proceed with a well-known and classical result.

Proposition 1 *Let P be an $n \times n$ transition matrix.*

(a) The sequence $\frac{I+P+\ldots+P^{k-1}}{k}$ converges as $k \to \infty$ to a transition matrix Q, and, moreover, the sequence $I + P + \ldots + P^{k-1} - kQ$ is bounded.

(b) rank $(I - P)$ + rank $Q = n$.

(c) For every $n \times 1$ column vector c, the system of equations

$$Px = x, \qquad Qx = Qc$$

has a unique solution.

(d) I-(P-Q) is nonsingular, and

$$H(\beta) = \sum_{t \geq 0} \beta^t (P^t - Q) \to_{\beta \to 1-} H = (I - P - Q)^{-1} - Q.$$

$$H(\beta)Q = QH(\beta) = HQ = QH = 0$$

and

$$(I - P)H = H(I - P) = I - Q.$$

3. Markov Decision Processes

A special subclass of stochastic games is the class of Markov decision processes, i.e., stochastic games with a single player. This section reexamines classical results on Markov decision processes.

A finite-state-and-action MDP consists of

- a finite set S, the set of states;
- for every $z \in S$ a finite action set $A(z)$;
- for every pair consisting of a state z in S and an action $a \in A(z)$ a reward $r(z,a)$;
- for every pair consisting of a state z in S and an action $a \in A(z)$ a probability distribution $p(z,a)$ on S;
- an initial distribution μ on S.

The interpretation is as follows. The set $A(z)$ is the set of feasible actions at state z. The initial distribution of the state z_1 is according to μ. If at stage t the state is z_t and action $a_t \in A(z_t)$ is played, the decision-maker gets a stage payoff of $r(z_t, a_t)$ at stage t and the conditional distribution of the next state z_{t+1} given all past states and actions $z_1, a_1, \ldots, z_t, a_t$ is given by $p(z_t, a_t)$. We use the common notational convention denoting the probability of $z_{t+1} = z'$ given $z_t = z$ and $a_t = a$, $p(z,a)[z']$, by $p(z' \mid z, a)$.

The quadruple $\langle S, A, r, p \rangle$ is called an MDP form.

3.1. STRATEGIES

A *pure strategy*[2] of the decision-maker in an MDP is a function σ that assigns to every finite string $h = (z_1, a_1, \ldots, z_t)$ an action $\sigma(h)$ in $A(z_t)$. The set of all pure strategies is denoted Σ. A *behavioral strategy* is a function σ that assigns to every finite string $h = (z_1, a_1, \ldots, z_t)$ a probability distribution $\sigma(h)$ on $A(z_t)$; $\sigma(h)[a]$ stands for the probability that the behavioral strategy σ chooses the action a (in $A(z_t)$) given the finite history $h = (z_1, a_1, \ldots, z_t)$. Obviously, when a pure strategy σ is seen as a map that assigns to the finite history h the Dirac measure concentrated on $\sigma(h)$, it is also a behavioral strategy. Note that the definition of a strategy in an MDP depends only on the state space S and the feasible action sets $A(z)$, $z \in S$.

Let H stand for the set of all finite histories $(z_1, a_1, \ldots, z_t)$, where t is a positive integer and for every $s < t$ the action a_s is in $A(z_s)$. Given $h = (z_1, a_1, \ldots, z_t) \in H$ we denote by $A(h)$ the set $A(z_t)$ of feasible actions at state z_t. A pure strategy σ is thus a point in the Cartesian product $\prod_{h \in H} A(h)$. This is a Cartesian product of countably many finite sets. Therefore, it is a metrizable compact space. A *mixed strategy* is a probability

[2]The classical literature on MDPs often refers to a strategy as a policy (or plan).

distribution on the space of pure strategies, i.e., an element of $\Delta(\Sigma)$ where $\Delta(B)$ stands for all probability distributions on B.

Let H_t be the set of all finite histories $(z_1, a_1, \ldots, z_t)$, where for every $s < t$ the action a_s is in $A(z_s)$. Then $H = \cup_{t \geq 1} H_t$. An infinite sequence $(z_1, a_1, \ldots, z_t, \ldots)$ such that $(z_1, a_1, \ldots, z_t) \in H_t$ for every t is called an *infinite play*. The space of all infinite plays is denoted H_∞. The algebra of subsets of H_∞ spanned by the coordinates $z_1, a_1, \ldots, z_t$ is denoted $\mathcal{H}_t$, and the σ-algebra of subsets spanned by $\cup_{t \geq 1} \mathcal{H}_t$ is denoted $\mathcal{H}_\infty$.

A probability measure P on the measurable space $(H_\infty, \mathcal{H}_\infty)$ induces a sequence of probability measures P_t on $(H_\infty, \mathcal{H}_t)$ by defining P_t to be the restriction of P to the algebra $\mathcal{H}_t$. Note that the restriction of P_t to $\mathcal{H}_s$ where $s \leq t$ is equal to P_s. Also, if P_t is a sequence of probability measures on $(H_\infty, \mathcal{H}_t)$ such that the restriction of P_t to the algebra of subsets $\mathcal{H}_s$ is equal to P_s, then there is a unique measure P on $(H_\infty, \mathcal{H}_\infty)$ whose restriction to $(H_\infty, \mathcal{H}_t)$ coincides with P_t. Therefore, a common way to define a probability on the space of infinite plays $(H_\infty, \mathcal{H}_\infty)$ is to define recursively a sequence of probability distributions P_t on $(H_\infty, \mathcal{H}_t)$ such that the restriction of P_t to $(H_\infty, \mathcal{H}_s)$ (where $s \leq t$) equals P_s. This last compatibility condition is achieved by defining the conditional probability of P_{t+1} given $\mathcal{H}_t$ and thus implicitly stipulating that the restriction of P_{t+1} to $\mathcal{H}_t$ coincides with P_t.

A pure or behavioral strategy σ together with the initial distribution μ induces a probability distribution P_σ^μ, or P_σ for short, on the space H_∞ as follows.

$$P_\sigma(z_1 = z) = \mu(z)$$

$$P_\sigma(a_t = a \mid z_1, a_1, \ldots, z_t) = \sigma(z_1, \ldots, z_t)[a]$$

$$P_\sigma(z_{t+1} = z \mid z_1, a_1, \ldots, z_t, a_t) = p\,(z \mid z_t, a_t).$$

Note that the right-hand side of the first and last equalities above is independent of σ.

A mixed strategy $\nu \in \Delta(\Sigma)$ is a mixture of pure strategies and therefore the probability P_ν that it induces on H_∞ is given by the following formula. Let X be a measurable subset of H_∞. Then

$$P_\nu(X) = E_\nu(P_\sigma(X)) = \int P_\sigma(X)\, d\nu(\sigma).$$

In particular,

$$P_\nu(z_1 = z) = \mu(z)$$

and

$$P_\nu(z_{t+1} = z \mid z_1, a_1, \ldots, z_t, a_t) = p\,(z \mid z_t, a_t).$$

In order to complete the definition of P_ν by means of conditional probabilities we have to derive the formula for $P_\nu(a_t = a \mid z_1, a_1, \ldots, z_t)$.

For every finite history $h = (z_1, a_1, \ldots, z_t)$, we denote by $\Sigma(h)$ the set of all pure strategies compatible with h, i.e., $\sigma \in \Sigma(h)$ if and only if for every $s < t$ we have $\sigma(z_1, a_1, \ldots, z_s) = a_s$. Then

$$P_\nu(a_t = a \mid z_1, a_1, \ldots, z_t) = \frac{\nu(\{\sigma \in \Sigma(h) \mid \sigma(h) = a\})}{\nu(\Sigma(h))}$$

whenever $\nu(\Sigma(h)) \neq 0$, where $h = (z_1, a_1, \ldots, z_t)$. The above conditional distribution when $\nu(\Sigma(h)) = 0$ is immaterial for the reconstruction of P_ν.

Given a mixed strategy $\nu \in \Delta(\Sigma)$, we define the following behavioral strategy τ. Let

$$\tau(z_1, a_1, \ldots, z_t)[a] = \frac{\nu(\{\sigma \in \Sigma(h) \mid \sigma(h) = a\})}{\nu(\Sigma(h))}$$

if $\nu(\Sigma(h)) \neq 0$ and $\tau(z_1, a_1, \ldots, z_t)$ is arbitrary if $\nu(\Sigma(h)) = 0$.

The formulas defining P_ν and P_τ by means of the conditional distributions are identical. Therefore, the probabilities induced on H_∞ by ν and by τ coincide. In addition, if τ is a behavioral strategy we can identify it with a point in the product $\prod_{h \in H} \Delta(A(h))$ and thus with a probability distribution ν on $\Sigma = \prod_{h \in H} A(h)$. The probability induced on H_∞ by the mixed strategy ν and the behavioral strategy τ coincide. Therefore, any distribution on H_∞ induced by a mixed strategy can be induced by a behavioral strategy and vice versa.

In particular, maximizing (respectively, taking the supremum of) the expectation of a bounded real-valued function defined on H_∞ by means of behavioral strategies or by means of mixed strategies (and therefore also by means of pure strategies) leads to the same maximal (respectively, supremum) value.

A special class of strategies is the class of stationary strategies. A behavioral strategy σ is *stationary* if $\sigma(z_1, \ldots, z_t)$ depends only on the state z_t. Thus, a stationary strategy is represented by a function $\sigma : S \to \cup_z \Delta(A(z))$ such that $\sigma(z) \in \Delta(A(z))$. Equivalently, a stationary strategy can be represented by a point $\sigma \in \prod_{z \in S} \Delta(A(z))$.

3.2. PAYOFFS

The objective of the decision-maker in an MDP is to maximize a specific evaluation of the stream $r(z_1, a_1), \ldots, r(z_t, a_t), \ldots$ of payoffs. In the present chapter we confine ourselves to two evaluations: the discounted evaluation and the limiting average evaluation.

In the β-discounted model, Γ_β, the payoff associated with the strategy σ and the initial state z is

$$\begin{aligned} v(z,\beta,\sigma) &:= E_\sigma^z\left((1-\beta)\sum_{t=1}^{\infty}\beta^{t-1}r(z_t,a_t)\right) \\ &= (1-\beta)\sum_{t=1}^{\infty}\beta^{t-1}E_\sigma^z\left(r(z_t,a_t)\right), \end{aligned}$$

where E_σ^z stands for the expectation with respect to the distribution induced by the initial state z and the strategy σ. The equality follows from the uniform convergence of $\sum_{t=1}^{T}\beta^{t-1}r(z_t,a_t)$, as $T\to\infty$, to the infinite sum $\sum_{t=1}^{\infty}\beta^{t-1}r(z_t,a_t)$.

For a pair consisting of an initial state z and a discount factor $\beta<1$ we set

$$v(z,\beta) := \max_\sigma v(z,\beta,\sigma).$$

The existence of the max follows from the fact that the space of pure strategies is a compact space and the function $\sigma\mapsto v(z,\beta,\sigma)$ is continuous in σ. Indeed, for any two pure strategies σ and τ that coincide on all finite histories of length $\le t$ we have $|v(z,\beta,\sigma)-v(z,\beta,\tau)|\le 2\beta^t\|r\|$, where $\|r\|=\max_{z,a}|r(z,a)|$. The existence of the max will follow also from the result in Section 3.3.

Note that for every state z, strategy σ, and $1>\beta>\gamma>0$ we have $|v(z,\beta,\sigma)-v(z,\gamma,\sigma)|\le\|r\|\sum_{t=0}^{\infty}|(1-\beta)\beta^t-(1-\gamma)\gamma^t|$. By the triangle inequality we have $|(1-\beta)\beta^t-(1-\gamma)\gamma^t|\le(\beta-\gamma)\beta^t+(1-\gamma)(\beta^t-\gamma^t)$. Therefore, $|v(z,\beta,\sigma)-v(z,\gamma,\sigma)|\le 2\|r\|(\beta-\gamma)/(1-\beta)$. Therefore, the functions $\gamma\mapsto v(z,\gamma,\sigma)$ and $\gamma\mapsto v(z,\gamma)$ are Lipschitz in the interval $[0,\beta)$.

In the limiting average model, Γ_∞, we wish to define the payoff associated with an initial state z and a strategy σ as the expectation of the limit of the average stage payoffs, $\frac{1}{T}\sum_{t=1}^{T}r(z_t,a_t)$. However, the limit need not exist. Therefore, optimality of a strategy in the limiting average model needs careful definition.

For a pair consisting of a state z and a strategy σ we set

$$\begin{aligned} v(z,\sigma) &:= E_\sigma^z\left(\liminf_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}r(z_t,a_t)\right) \\ u(z,\sigma) &:= \liminf_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}E_\sigma^z\left(r(z_t,a_t)\right). \end{aligned}$$

Note that (by Fatou's lemma) $u(z,\sigma)\ge v(z,\sigma)$ with the possible strict inequality. Consider for example the following MDP. The state space is

$S = \{1, 2\}$, there are two actions, T and B, in each state (i.e., $A(z) = \{T, B\}$), the payoff function is given by $r(z, T) = 1$ and $r(z, B) = 0$, and the transitions are described by $p(1 \mid \cdot, \cdot) = .5 = p(2 \mid \cdot, \cdot)$. Let σ be the pure strategy that at stage $t \geq 2$ plays T if either $z_2 = 2$ and $(2n)! \leq t < (2n+1)!$ or $z_2 = 1$ and $(2n+1)! \leq t < (2n+2)!$, and σ plays B otherwise. Then, $u(\cdot, \sigma) = 1/2$. However, $\liminf_{k\to\infty} \frac{1}{k} \sum_{t=1}^{k} r(z_t, a_t) = 0$ and therefore $v(z, \sigma) = 0$.

It will later be shown that in an MDP with finitely many states and actions there exists a pure stationary strategy σ which satisfies the following optimality conditions. There exists a constant C such that for every initial state z and every strategy τ we have $v(z, \sigma) \geq u(z, \tau) \geq v(z, \tau)$, and, moreover,

$$v(z, \sigma) = E_\sigma^z \left(\liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r(z_t, a_t) \right) \geq E_\tau^z \left(\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r(z_t, a_t) \right),$$

and

$$E_\sigma^z \left(\sum_{t=1}^{k} r(z_t, a_t) \right) \quad \geq \quad E_\tau^z \left(\sum_{t=1}^{k} r(z_t, a_t) \right) - C \quad \forall\, k.$$

3.3. THE DISCOUNTED MDP

In this section we prove the classical result that the MDP with finitely many states and actions and a fixed discount factor β has a pure stationary strategy that is optimal for every initial state.

Consider the map Ψ from $\mathbb{R}^S$ to itself defined by

$$(\Psi x)[z] = \max_{a \in A(z)} \left((1-\beta) r(z, a) + \beta \sum_{z' \in S} p(z' \mid z, a) x(z') \right).$$

Two immediate properties of this map follow. Monotonicity,

$$x \geq y \Rightarrow \Psi x \geq \Psi y,$$

where for $x, y \in \mathbb{R}^S$ we use the notation $x \geq y$ whenever $x(z) \geq y(z)$ for every coordinate z, and

$$\Psi(c\mathbf{1} + x) = \beta c\mathbf{1} + \Psi x,$$

where $\mathbf{1}$ stands for the vector with each of its coordinates equal to 1.

Therefore, since $x - \|x-y\|\mathbf{1} \le y \le x + \|x-y\|\mathbf{1}$, we have

$$\Psi y \le \Psi(x + \|x-y\|\mathbf{1}) = \beta\|x-y\|\mathbf{1} + \Psi x$$

and

$$\Psi y \ge \Psi(x - \|x-y\|\mathbf{1}) = -\beta\|x-y\|\mathbf{1} + \Psi x.$$

The two inequalities imply that

$$\|\Psi x - \Psi y\| \le \beta\|x-y\|.$$

Therefore, the map Ψ is a (uniformly) strict contraction of the (complete metric) space $\mathbb{R}^S$ and thus has a unique fixed point w. The (unique) fixed point w satisfies the following equalities.

$$w(z) = \max_{a\in A(z)} \left((1-\beta) r(z,a) + \beta \sum_{z'\in S} p\left(z' \mid z, \sigma(z)\right) w(z') \right). \tag{1}$$

Therefore, there is a pure stationary strategy σ, i.e., a function $\sigma : S \to \cup_z A(z)$ such that $\sigma(z) \in A(z)$, such that

$$w(z) = (1-\beta) r(z,\sigma(z)) + \beta \sum_{z'\in S} p\left(z' \mid z, \sigma(z)\right) w(z'). \tag{2}$$

It follows from equation (2) that

$$E_\sigma\left((1-\beta) r(z_t,a_t) + \beta w(z_{t+1}) \mid z_1, a_1, \ldots, z_t\right) = w(z_t) \tag{3}$$

and therefore by taking expectation of the conditional expectations in equation (3) and rearranging the terms we have

$$(1-\beta) E_\sigma\left(r(z_t,a_t) \mid z_1\right) = E_\sigma\left(w(z_t) \mid z_1\right) - \beta E_\sigma\left(w(z_{t+1}) \mid z_1\right). \tag{4}$$

Multiplying equation (4) by β^{t-1} and summing over $1 \le t < k$ we deduce that

$$(1-\beta)\sum_{t=1}^{k-1} \beta^{t-1} E_s\left(r(z_t,a_t) \mid z_1\right) = w(z_1) - \beta^k E_\sigma\left(w(z_k) \mid z_1\right)) \to_{k\to\infty} w(z_1)$$

and therefore

$$(1-\beta)\sum_{t=1}^{\infty} \beta^{t-1} E_\sigma\left(r(z_t,a_t) \mid z_1\right) = w(z_1).$$

Similarly, using equation (1) we have for every strategy τ that

$$E_\tau\left((1-\beta) r(z_t,a_t) + \beta w(z_{t+1}) \mid z_1, a_1, \ldots, z_t\right) \le w(z_t) \tag{5}$$

and therefore by taking expectation of the conditional expectations in equation (5) and rearranging the terms we have

$$(1-\beta)E_\tau\left(r(z_t,a_t)\mid z_1\right)\le E_\tau\left(w(z_t)\mid z_1\right)-\beta E_\tau\left(w(z_{t+1})\mid z_1\right). \quad (6)$$

Multiplying equation (6) by β^{t-1} and summing over $t\ge 1$ we deduce that

$$(1-\beta)\sum_{t=1}^{\infty}\beta^{t-1}E_\tau\left(r(z_t,a_t)\mid z_1\right)\le w(z_1).$$

We conclude that for every strategy τ and every initial state z we have $v(z,\beta,\sigma)=w(z)\ge v(z,\beta,\tau)$. This proves

Proposition 2 (Blackwell [3]) *For every MDP form $\langle S,A,r,p\rangle$ with finitely many states and actions and every discount factor $\beta<1$ there is a pure stationary strategy σ such that for every initial state z and every strategy τ we have*

$$v(z,\beta,\sigma)\ge v(z,\beta,\tau).$$

Moreover, the stationary strategy σ obeys, for every state z,

$$\begin{aligned} v(z,\beta) &= (1-\beta)r(z,\sigma(z))+\beta\sum_{z'\in S}p\left(z'\mid z,\sigma(z)\right)v(z',\beta)\\ &= \max_{a\in A(z)}\left((1-\beta)r(z,a)+\beta\sum_{z'\in S}p\left(z'\mid z,\sigma(z)\right)v(z',\beta)\right). \end{aligned}$$

The next result provides a formula for the payoff of a Markov stationary strategy (this result obviously applies to every stochastic game). For every (pure or) behavioral stationary strategy σ let P denote the $S\times S$ matrix where

$$P_{z,z'}=p\left(z'\mid z,\sigma(z)\right):=\sum_{a\in A(z)}(\sigma(z))[a]\,p\left(z'\mid z,a\right)$$

and set

$$r_\sigma(z)=r(z,\sigma(z)):=\sum_{a\in A(z)}(\sigma(z))[a]\,r(z,a).$$

Lemma 1 *a) The $S\times S$ matrix $I-\beta P$ is invertible and its inverse is given by*

$$(I-\beta P)^{-1}=\sum_{t=0}^{\infty}(\beta P)^t.$$

b) The payoff as a function of the initial state z, $v(z,\beta,\sigma)$, is given by

$$v(z,\beta,\sigma)=\sum_{z'\in S}(I-\beta P)^{-1}_{zz'}\,r_\sigma(z').$$

Proof. The series of finite sums $\sum_{t=0}^{n}\beta^t P^t$ converges, as $n\to\infty$, to the infinite sum $\sum_{t=0}^{\infty}\beta^t P^t$. The product $(I-\beta P)\sum_{t=0}^{n}\beta^t P^t$ equals $I-\beta^{n+1}P^{n+1}$, which converges, as $n\to\infty$, to the identity matrix I, and therefore $\sum_{t=0}^{\infty}\beta^t P^t$ is the inverse of the matrix $I-\beta P$, which proves (a).

Notice that $E_\sigma^z(\beta^{t-1}r(z_t,a_t))=\sum_{z'\in S}(\beta^{t-1}P^{t-1})_{z,z'}r_\sigma(z')$ and therefore $v(z,\beta,\sigma)=\sum_{t=1}^{\infty}\sum_{z'\in S}(\beta^{t-1}tP^{t-1})_{z,z'}r_\sigma(z')$, which proves (b). ∎

A corollary of the lemma is that, for every stationary strategy σ and every state z, the payoff $v(z,\beta,\sigma)$ of an MDP with finitely many states and actions is a rational function of the discount factor β, the stage payoffs $r(z,a)$, $z\in S$ and $a\in A(z)$, and the transition probabilities $p(z'\mid z,a)$, $z,z'\in S$ and $a\in A(z)$. In what follows, the symbol $\forall\tau$ (respectively, $\forall z$) means for every strategy τ (respectively, for every state z).

Proposition 3 *For every MDP form $\langle S,A,r,p\rangle$, there is a pure stationary strategy (policy) σ and a discount factor $0<\beta_0<1$ such that*
1) (Blackwell [3])

$$v(z,\beta,\sigma)\quad\geq\quad v(z,\beta,\tau)\quad\forall\beta_0\leq\beta<1\quad\forall z\quad\forall\tau.$$

2) (Blackwell [3]) *For every state z, the function $\beta\mapsto v(z,\beta)$ is a rational function on the interval $[\beta_0,1)$. In particular, the limit of $v(z,\beta)$ as $\beta\to 1-$ exists. Set $v(z):=\lim_{\beta\downarrow 1}v(z,\beta)$.*
3) There is a positive constant C such that for every initial state z_1, every strategy τ, and every $k>1$ we have

$$\begin{aligned}\frac{1}{k}\sum_{t=1}^{k}E_\tau^{z_1}\left(r(z_t,a_t)\right)&\leq v(z_1)+\frac{C}{k}\\&\leq\frac{1}{k}\sum_{t=1}^{k}E_\sigma^{z_1}\left(r(z_t,a_t)\right)+\frac{2C}{k};\end{aligned}$$

in particular, $u(z,\sigma)\geq u(z,\tau)$.
4) For every initial state z_1 and every strategy τ we have

$$E_\sigma^{z_1}\left(\liminf_{k\to\infty}\frac{1}{k}\sum_{t=1}^{k}r(z_t,a_t)\right)\geq E_\tau^{z_1}\left(\limsup_{k\to\infty}\frac{1}{k}\sum_{t=1}^{k}r(z_t,a_t)\right).$$

Proof. For every stationary strategy σ and every initial state z the function $\beta\to v(z,\beta,\sigma)$ is a rational function of β. For every two rational functions f and g defined on a left neighborhood of 1 there is $\gamma<1$ such that either $f(\beta)>g(\beta)$ for all $\gamma\leq\beta<1$, or $f(\beta)<g(\beta)$ for all $\gamma\leq\beta<1$, or $f(\beta)=g(\beta)$ for all $\gamma\leq\beta<1$. By Proposition 2, for every discount factor β there is a pure stationary strategy σ such that for every initial state z

and every pure (stationary) strategy τ we have $v(z,\beta,\sigma) \geq v(z,\beta,\tau)$. There are finitely many pure stationary strategies. Therefore, there is one pure stationary strategy σ which is optimal in the β_i discounted MDP, Γ_{β_i}, for a sequence of discount factors $\beta_i < 1$ that converges to 1. Therefore, there is a discount factor $\beta_0 < 1$ such that for every pure stationary strategy τ we have

$$v(\mu,\beta,\sigma) \geq v(\mu,\beta,\tau) \quad \forall \beta_0 \leq \beta < 1 \quad \forall \mu \in \Delta(S).$$

This completes the proof of 1).

In particular, for $\beta_0 \leq \beta < 1$, $v(z,\beta) = v(z,\beta,\sigma)$, and using Part 2) of Lemma 1 the function $\beta \mapsto v(z,\beta)$ is a rational function on $[\beta_0, 1)$. As $\beta \mapsto v(z,\beta)$ is a bounded rational function in a left neighborhood of 1, its limit as $\beta \downarrow 1$ exists. This completes the proof of 2).

We prove 3) by induction on k. The function $\beta \mapsto v(z,\beta)$ is a bounded rational function. Therefore, it is differentiable in a left neighborhood of 1 and its derivative there is bounded in absolute value, say by $C_1(z) \leq C_1$. Therefore, there is k_0 such that for every $1 - 1/k_0 \leq \beta < \gamma < 1$ and every state z we have $|v(z,\beta) - v(z,\gamma)| \leq C_1|\gamma - \beta|$. As the function $\beta \mapsto v(z,\beta)$ is Lipschitz in the interval $[0, 1 - 1/k_0)$ there is a positive constant C_2 such that for every $0 \leq \beta < \gamma \leq 1 - 1/k_0$ and every state z we have $|v(z,\beta) - v(z,\gamma)| \leq C_2|\gamma - \beta|$. Therefore, if $C \geq \max\{C_1, C_2\}$ we have

$$v(z, 1 - \frac{1}{k}) \leq v(z, 1 - \frac{1}{k+1}) + \frac{C}{k(k+1)} \quad \forall\, k \geq 1. \tag{7}$$

W.l.o.g. we assume that $C \geq 2\|r\|$. Define the function α by $\alpha(k) = \frac{C}{k}\sum_{n \leq k} n^{-2}$. Observe that for every k we have

$$\frac{k\alpha(k)}{k+1} + \frac{C}{(k+1)^2} = \alpha(k+1). \tag{8}$$

We prove by induction on k that for every $k \geq 1$ we have

$$E_\tau^{z_1}\left(\frac{\sum_{t=1}^{k} r(z_t,a_t)}{k}\right) \leq v(z_1, 1 - \frac{1}{k}) + \alpha(k) \quad \forall\, z_1 \ \ \forall\, \tau. \tag{9}$$

As $\alpha(1) = C \geq 2\|r\| \geq \|r\| + v(\cdot,\cdot)$, inequality (9) holds for $k = 1$. We will show that if inequality (9) holds for some fixed $k \geq 1$ then it also holds for $k+1$. As (9) holds for k, we have (using the equality $E_\tau^{z_1}(E_\tau^{z_1}(\cdot \mid \mathcal{H}_1)) = E_\tau^{z_1}(\cdot)$)

$$\frac{k}{k+1} E_\tau^{z_1}\left(\frac{\sum_{t=2}^{k+1} r(z_t,a_t)}{k}\right) \leq \frac{k}{k+1} E_\tau^{z_1}\left(v(z_2, 1 - \frac{1}{k})\right) + \frac{k\alpha(k)}{k+1}. \tag{10}$$

Recall that

$$E_\tau^{z_1}\left(\frac{r(z_1,a_1)}{k+1}+\frac{k}{k+1}v(z_2,1-\frac{1}{k+1})\right)\le v(z_1,1-\frac{1}{k+1}). \qquad (11)$$

Inequality (7) implies that

$$E_\tau^{z_1}\left(\frac{k}{k+1}v(z_2,1-\frac{1}{k})\right)\le E_\tau^{z_1}\left(\frac{k}{k+1}v(z_2,1-\frac{1}{k+1})\right)+\frac{C}{(k+1)^2}. \quad (12)$$

Summing inequalities (8), (10), (11) and (12) we deduce that

$$E_\tau^{z_1}\left(\frac{\sum_{t=1}^{k+1}r(z_t,a_t)}{k+1}\right)\le v(z_1,1-\frac{1}{k+1})+\alpha(k+1),$$

which proves that (9) holds for $k+1$, and thus (9) holds for every $k\ge 1$. As $v(z,1-\frac{1}{k}\le v(z)+C/k$ and $\alpha(k)\le 2C/k$, the first part of 3) follows.

Let (7*) (respectively, (9*), (10*), (11*) and (12*)) stand for (7) (respectively, (9), (10), (11) and (12)) where v is replaced by $w:=-v$, τ is replaced by σ and r is replaced by $g:=-r$. Then (7*), (11*) and (12*) hold for every k. Inequality (9*) holds for $k=1$. Therefore, if (9*) holds for some $k\ge 1$, then (10*) holds for k, and summing inequalities (8), (10*), (11*) and (12*) we deduce that

$$E_\sigma^{z_1}\left(\frac{\sum_{t=1}^{k+1}r(z_t,a_t)}{k+1}\right)\ge v(z_1,1-\frac{1}{k+1})-\alpha(k+1),$$

which proves that (9*) holds for $k+1$, and thus (9*) holds for every $k\ge 1$. As $v(z,1-\frac{1}{k})\ge v(z)-C/k$ and $\alpha(k)\le 2C/k$, the second part of 3) follows. This completes the proof of 3).

Fix $\beta_0<1$ sufficiently large such that for all $\beta_0\le\beta<1$

$$v(z,\beta) = (1-\beta)\,r(z,\sigma(z))+\sum_{z'}p\,(z'\mid z,\sigma(z))\beta v(z',\beta)\quad\forall z \quad (13)$$

$$= \max_{a\in A(z)}(1-\beta)\,r(z,a))+\sum_{z'}p\,(z'\mid z,a)\beta v(z',\beta)\quad\forall z. \quad (14)$$

Equation (13) implies that

$$E_\sigma((1-\beta)r(z_t,a_t)+\beta v(z_{t+1},\beta)\mid\mathcal{H}_t)=v(z_t,\beta)\quad\forall\beta_0\le\beta<1, \qquad (15)$$

and equation (14) implies that for every strategy τ we have

$$E_\tau((1-\beta)r(z_t,a_t)+\beta v(z_{t+1},\beta)\mid\mathcal{H}_t)\le v(z_t,\beta)\quad\forall\beta_0\le\beta<1. \qquad (16)$$

Equation (16) implies (by going to the limit as $\beta \to 1-$) that for every strategy τ we have

$$E_\tau\left(v(z_{t+1}) \mid \mathcal{H}_t\right) \leq v(z_t) = E_\sigma\left(v(z_{t+1}) \mid \mathcal{H}_t\right). \tag{17}$$

Therefore, the stochastic process $(v(z_t))$ is a bounded supermartingale (respectively, a martingale) w.r.t. the probability induced by τ (respectively, by σ) and thus has a limit v_τ (respectively, v_σ) almost everywhere w.r.t. the probability induced by τ (respectively, by σ), and

$$E_\tau^{z_1}(v_\tau) \leq E_\tau^{z_1}(v(z_t)) \leq v(z_1) = E_\sigma^{z_1}(v_\sigma) = E_\sigma^{z_1}(v(z_t)).$$

Let $\varepsilon > 0$. By Part 3) of Proposition 3 there is k sufficiently large so that for every m we have

$$E_\tau\left(\frac{1}{k}\sum_{t=m+1}^{m+k} r(z_t, a_t) \mid \mathcal{H}_m\right) \leq v(z_{m+1}) + \varepsilon. \tag{18}$$

For every integer $n \geq 1$ we set $Z_n := \frac{1}{k}\sum_{t=k(n-1)+1}^{kn} r(z_t, a_t)$, $Y_n := Z_n - E_\tau(Z_n \mid \mathcal{H}_{k(n-1)+1})$, and $u_n := v(z_{(n-1)k+1})$. Equation (18) implies that for every positive integer n we have

$$E_\tau\left(Z_n \mid \mathcal{H}_{(n-1)k+1}\right) \leq u_n + \varepsilon.$$

The stochastic process (Y_n) is a bounded sequence of martingale differences and therefore

$$\frac{Y_1 + \ldots + Y_n}{n} \to 0 \quad \text{a.e. w.r.t. } \tau.$$

As $(u_n)_n = (v(z_{(n-1)k+1}))_n$ is a subsequence of the sequence $(v(z_t))_t$ that converges a.e. w.r.t. τ to v_τ, and $E_\tau(v_\tau) \leq v(z_1)$ we have

$$\frac{u_1 + \ldots + u_n}{n} \to v_\tau \quad \text{a.e. w.r.t. } \tau.$$

As $Z_n \leq Y_n + u_n + \varepsilon$ we deduce that

$$\frac{\sum_{t\leq nk} r(z_t, a_t)}{nk} \leq \frac{Y_1 + \ldots, Y_n}{n} + \frac{u_1 + \ldots + u_n}{n} + \varepsilon \to v + \varepsilon \quad \text{a.e. w.r.t. } \tau.$$

We conclude that

$$E_\tau^{z_1}\left(\limsup_{k\to\infty} \frac{1}{k}\sum_{t=1}^{k} r(z_t, a_t)\right) \leq E_\tau^{z_1}(v_\tau) + \varepsilon \leq v(z_1) + \varepsilon.$$

As the last inequality holds for every $\varepsilon > 0$ we conclude that

$$E_\tau^{z_1}\left(\limsup_{k\to\infty}\frac{1}{k}\sum_{t=1}^{k} r(z_t,a_t)\right) \le v(z_1).$$

Similarly, setting $Y_n^\sigma := Z_n - E_\sigma(Z_n \mid \mathcal{H}_{k(n-1)+1})$ we deduce that the stochastic process (u_n) is a martingale w.r.t. the probability induced by σ and thus converges a.e. to a limit v_σ with $E_\sigma(v_\sigma) = v(z_1)$. By Part 3) it follows that for a sufficiently large k we have

$$E_\sigma\left(Z_n \mid \mathcal{H}_{(n-1)k+1}\right) \ge u_n - \varepsilon.$$

As $Z_n \ge Y_n^\sigma + u_n - \varepsilon$ we deduce that

$$E_\sigma^{z_1}\left(\liminf_{k\to\infty}\frac{r(z_1,a_1)+\ldots+r(z_k,a_k)}{k}\right) \ge E_\sigma^{z_1}(v_\sigma) - \varepsilon \ge v(z_1) - \varepsilon.$$

As the last inequality holds for every $\varepsilon > 0$ we conclude that

$$E_\sigma^{z_1}\left(\liminf_{k\to\infty}\frac{r(z_1,a_1)+\ldots+r(z_k,a_k)}{k}\right) \ge v(z_1).$$

∎

4. Remarks

In this section we discuss the extension of the above-mentioned results on MDPs to results on two-person zero-sum stochastic games with finitely many states and actions.

Remark 1 The definition of a finite-state-and-action stochastic game is a minor modification to the definition of an MDP. The set of players is a finite set I. The set $A(z)$ is the Cartesian product of finite sets $A^i(z)$, $i \in I$; $A^i(z)$ is the set of feasible actions of player $i \in I$ at state z. The reward $r(z,a)$ is the vector of rewards $\left(r^i(z,a)\right)_{i\in I}$; player i gets a stage payoff $r^i(z_t,a_t)$ at stage t. A vector of strategies $\sigma = (\sigma^i)_{i\in I}$ together with the initial distribution μ induces a probability distribution P_σ^μ on the space of infinite plays, exactly as in the setup of an MDP (with $\sigma(z_1,\ldots,z_t)[a] = \prod_i \sigma^i(z_1,\ldots,z_t)[a^i]$).

Remark 2 In a two-person zero-sum stochastic game we define the value of the β-discounted stochastic game by

$$v(z,\beta) = \max_{\sigma^1}\min_{\sigma^2} E_{\sigma^1,\sigma^2}^z\left((1-\beta)\sum_{t=1}^{\infty}\beta^{t-1}r^1(z_t,a_t)\right).$$

The map from $\mathbb{R}^S$ to itself, $v \mapsto \Psi v$, defined by

$$\Psi v[z] = \max_x \min_y \left((1-\beta) r^1(z,x,y) + \beta \sum_{z'} p(z' \mid z,x,y) v(z') \right),$$

where the max is over all $x \in \Delta(A^1(z))$, the min is over all $y \in \Delta(A^2(z))$, and $r^1(z,x,y)$ (respectively $p(z' \mid z,x,y)$) is the multilinear extension of $r^1(z,a^1,a^2)$ (respectively $p(z' \mid z,a^1,a^2)$), is a strict contraction [14], [17]. Its unique fixed point, $w(\cdot,\beta)$ (whose z-th coordinate is $w(z,\beta)$) is the value of the β-discounted stochastic game [14], [17].

Remark 3 In an MDP with finitely many states and actions there exist for every discount factor $\beta < 1$ a pure strategy that is optimal in the β-discounted MDP (Proposition 2), and a uniform optimal strategy, namely a strategy that is optimal in all β-discounted MDPs with the discount factor β sufficiently close to 1 (Proposition 3, Part 1)). These two results do not extend to results on two-person zero-sum stochastic games with finitely many states and actions. However, there are special classes of stochastic games where a pure optimal strategy and a uniform optimal strategy do exist (see, e.g., [13], [19], [20] and the references there).

Remark 4 Part 2) of Proposition 3 states that, for a fixed MDP with finitely many states and actions, the function $\beta \mapsto v(z,\beta)$ is a bounded rational function of β and thus, in particular, it can be expressed, in a sufficiently small left neighborhood of 1 ($\beta_0 < \beta < 1$), as a convergent series in powers of $1-\beta$. In a two-person zero-sum stochastic game with finitely many states and actions such an expression is no longer available. However, Bewley and Kohlberg [1] show that the value of the β-discounted stochastic game is expressed in a left neighborhood of 1 as a convergent series in fractional powers of $(1-\beta)$ (see also [9]).

Remark 5 Part 3) of Proposition 3 provides an approximation to the n-stage value of an MDP, $v(z,n) := \max_\sigma E^z_\sigma(\frac{1}{n}\sum_{t=1}^n r(z_t,a_t))$, by $v(z)$. The error term, $|v(z,n)-v(z)|$, is $O(1/n)$. As $|v(z)-v(z,1-1/n)| = O(1/n)$ we deduce that in an MDP with finitely many states and actions we have $|v(z,n)-v(z,1-1/n)| = O(1/n)$.

We now comment on the asymptotic properties of the values $v(z,n)$ ($v(z,\beta)$) of the n-stage (β-discounted) two-person zero-sum stochastic game with finitely many states and actions. The proof of Part 2) of Proposition 3 shows actually that if for $0 < \gamma < \beta < 1$ we have $|v(z,\beta)-v(z,\gamma)| \le C|\beta-\gamma|(1-\beta)^{-1/M}$, then $|v(z,n)-v(z,1-1/n)| = O(n^{-1/M})$. In particular, if the series in fractional powers of $(1-\beta)$, $\sum_{i=0}^\infty a_i(z)(1-\beta)^{i/M}$, converges in a left neighborhood of 1 to the value $v(z,\beta)$ of the β-discounted game, then there is a constant C such that $|v(z,n)-v(z,1-1/n)| \le Cn^{1/M-1}$.

In particular, it proves that the limit of $v(z,n)$ as $n \to \infty$ exists and equals $\lim_{\beta \to 1-} v(z,\beta)$ [1] (see also [11]).

However, other series in fractional powers of $1/n$ provide a better approximation of $v(z,n)$. There exists a series in fractional powers of $1/n$, $\sum_{i=0}^{\infty} b_i(z)(1/n)^{i/M}$ (where $b_i(z)$ are real numbers and M is a positive integer), that converges for sufficiently large n and such that

$$\left| v(z,n) - \sum_{i=0}^{\infty} a_i(z) n^{-i/M} \right| = O\left(\frac{\ln n}{n}\right)$$

[2]. It is impossible to improve on the error term [2].

Remark 6 It will be shown in a later chapter, [10], that the existence of a uniform optimal strategy σ in an MDP, Proposition 3, has the following counterpart of ε uniform optimality in two-person zero-sum stochastic games.

For every $\varepsilon > 0$ there are strategies σ^ε of player 1 and τ^ε of player 2, and a positive integer N and a discount factor $\beta_0 < 1$, such that for every strategy τ of player 2 and every strategy σ of player 1 we have

$$\varepsilon + E^z_{\sigma^\varepsilon,\tau}\left(\liminf_{n\to\infty} \bar{x}_n\right) \geq v(z) \geq E^z_{\sigma,\tau^\varepsilon}\left(\limsup_{n\to\infty} \bar{x}_n\right) - \varepsilon,$$

where $\bar{x}_n = \frac{1}{n}\sum_{t=1}^{n} r(z_t, a_t)$, and

$$\varepsilon + E^z_{\sigma^\varepsilon,\tau}(\bar{x}_n) \geq v(z) \geq E^z_{\sigma,\tau^\varepsilon}(\bar{x}_n) - \varepsilon \quad \forall n \geq N.$$

Remark 7 In this chapter we considered transition matrices with values in the field of real numbers. However, one could consider Markov chains with a transition matrix whose values are in any ordered field. A field that has proved especially useful in the study of stochastic games is the field of functions that have an expansion in a left neighborhood of 1 as a power series in a fraction of $1-\beta$ (here β does not necessarily refer to the discount factor, but may be simply a parameter). This construction allows one to study the sensitivity of various statistics of the Markov chain as one varies the parameter β in a left neighborhood of 1. For more details, see [15], [18], [16].

References

1. Bewley, T. and Kohlberg, E. (1976) The asymptotic theory of stochastic games, *Mathematics of Operations Research* **1**, 197–208.
2. Bewley, T. and Kohlberg, E. (1976) The asymptotic solution of a recursion equation occurring in stochastic games, *Mathematics of Operations Research* **1**, 321–336.

3. Blackwell, D. (1962) Discrete dynamic programming, *Annals of Mathematical Statistics* **33**, 719–726.
4. Denardo, E. V. (1982) *Dynamic Programming*, Prentice-Hall, Englewood Cliffs, NJ.
5. Derman, C. (1970) *Finite State Markov Decision Processes*, Academic Press, New York.
6. Filar, J.A. and Vrieze, O.J. (1996) *Competitive Markov Decision Processes*, Springer-Verlag, Berlin.
7. Hardy, G. H. and Littlewood, J. E. (1931) Notes on the theory of series (xiv): Two Tauberian theorems, *Journal of London Mathematical Society* **6**, 281–286.
8. Kemeny, J. G. and Snell, J. L. (1960) *Finite Markov Chains*, Van Nostrand Reinhold, New York.
9. Neyman, A. (2003) Real algebraic tools in stochastic games, in A. Neyman and S. Sorin (eds.), *Stochastic Games and Applications*, NATO Science Series C, Mathematical and Physical Sciences, Vol. 570, Kluwer Academic Publishers, Dordrecht, Chapter 6, pp. 57–75.
10. Neyman, A. (2003) Stochastic games: Existence of the minmax, in A. Neyman and S. Sorin (eds.), *Stochastic Games and Applications*, NATO Science Series C, Mathematical and Physical Sciences, Vol. 570, Kluwer Academic Publishers, Dordrecht, Chapter 11, pp. 173–193.
11. Neyman, A. (2003) Stochastic games and nonexpansive maps, in A. Neyman and S. Sorin (eds.), *Stochastic Games and Applications*, NATO Science Series C, Mathematical and Physical Sciences, Vol. 570, Kluwer Academic Publishers, Dordrecht, Chapter 26, pp. 397–415.
12. Puterman, M. L. (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley, New York.
13. Raghavan, T.E.S. (2003) Finite-step algorithms for single-controller and perfect information stochastic games, in A. Neyman and S. Sorin (eds.), *Stochastic Games and Applications*, NATO Science Series C, Mathematical and Physical Sciences, Vol. 570, Kluwer Academic Publishers, Dordrecht, Chapter 15, pp. 227–251.
14. Shapley, L.S. (1953) Stochastic games, *Proceedings of the National Academy of Sciences of the U.S.A.* **39**, 1095–1100 (Chapter 1 in this volume).
15. Solan, E. (2003) Perturbations of Markov chains with applications to stochastic games, in A. Neyman and S. Sorin (eds.), *Stochastic Games and Applications*, NATO Science Series C, Mathematical and Physical Sciences, Vol. 570, Kluwer Academic Publishers, Dordrecht, Chapter 17, pp. 265–280.
16. Solan, E. and Vieille, N. (2002), Correlated equilibrium in stochastic games, *Games and Economic Behavior* **38**, 362–399.
17. Sorin, S. (2003) Discounted stochastic games: The finite case, in A. Neyman and S. Sorin (eds.), *Stochastic Games and Applications*, NATO Science Series C, Mathematical and Physical Sciences, Vol. 570, Kluwer Academic Publishers, Dordrecht, Chapter 5, pp. 51–55.
18. Vieille, N. (2000) Small perturbations and stochastic games, *Israel Journal of Mathematics* **119**, 127–142.
19. Vrieze, O.J. (2003) Stochastic games and stationary strategies, in A. Neyman and S. Sorin (eds.), *Stochastic Games and Applications*, NATO Science Series C, Mathematical and Physical Sciences, Vol. 570, Kluwer Academic Publishers, Dordrecht, Chapter 4, pp. 37–50.
20. Vrieze, O.J. (2003) Stochastic games, practical motivation and the orderfield property for special classes, in A. Neyman and S. Sorin (eds.), *Stochastic Games and Applications*, NATO Science Series C, Mathematical and Physical Sciences, Vol. 570, Kluwer Academic Publishers, Dordrecht, Chapter 14, pp. 215–225.
21. White, D. J. (1993) *Markov Decision Processes*, John Wiley, Chichester.