In this lecture we define stochastic games and Markov perfect equilibrium.

1 Stochastic Games

A (discounted) stochastic game with N players consists of the following elements.

1. A *state space* $\mathcal{X}$ (which we assume to be finite for the moment).
2. For each player i and state x, a set $A_i(x)$ of actions available to player i in state x. (We also assume each $A_i(x)$ is finite for the moment.)
3. For each player i, state x, and action vector $\boldsymbol{a} \in \prod_i A_i(x)$, a stage payoff $Q_i(\boldsymbol{a}; x)$.
4. For each state x and action vector $\boldsymbol{a} \in \prod_i A_i(x)$, a *transition probability* $\mathbb{P}(x'|x, \boldsymbol{a})$ that is a distribution on the state space $\mathcal{X}$.
5. A discount factor δ, $0 < \delta < 1$.
6. An initial state x^0.

Play proceeds as follows. The game starts in state x^0. At each stage t, all players simultaneously choose (possibly mixed) actions a_i^t, with possible pure actions given by the set $A_i(x^t)$. The stage payoffs $Q_i(\boldsymbol{a}^t; x^t)$ are realized, and the next state is chosen according to $\mathbb{P}(\cdot|x^t, \boldsymbol{a}^t)$. All players observe the entire past history of play before choosing their actions at stage t. (This is the simplest assumption; versions with partial monitoring have also been studied.)

As usual, let s_i denote a strategy for player i in this dynamic game; it is a mapping from histories (including states and actions) to actions. (After any history leading to state x, player i's strategy must choose an action in $A_i(x)$.) Given strategies $s_1, \ldots, s_N$, the expected discounted payoff of player i starting from state x^0 is:

$$\Pi_i(s_1, \ldots, s_N; x^0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \delta^t Q_i(s_1(x^t), \ldots, s_N(x^t); x^t)\right].$$

Here the expectation is over both randomization in state transitions, and randomization in players' choice of actions after any history.

2 Markov perfect equilibrium

The overwhelming focus in stochastic games is on *Markov perfect equilibrium*. This refers to a (subgame) perfect equilibrium of the dynamic game where players' strategies depend only on the

current state. When s_i is a strategy that depends only on the state, by some abuse of notation we will let $s_i(x)$ denote the action that player i would choose in state x. Such a strategy is called a *Markov strategy*.

It is straightforward to check that if all players other than i are playing Markov strategies $\boldsymbol{s}_{-i}$, then player i has a best response that is a Markov strategy. The basic intuition is that if there exists a best response where player i plays a_i after a history h leading to state x, and plays a_i' after another history h' that also leads to state x, then both a_i and a_i' must yield the same expected payoff to player i. Thus, for each state x, there exists a value $V_i(x; \boldsymbol{s}_{-i})$ that is the highest possible payoff player i can achieve starting from state x, given that all other players play the Markov strategies $\boldsymbol{s}_{-i}$. It then follows that:

$$V_i(x; \boldsymbol{s}_{-i}) = \max_{a_i \in A_i(x)} \mathbb{E}\left[Q_i(a_i, \boldsymbol{s}_{-i}(x); x) + \delta \sum_{x' \in \mathcal{X}} \mathbb{P}(x' | a_i, \boldsymbol{s}_{-i}(x), x) V_i(x'; \boldsymbol{s}_{-i}) \right].$$

(This is the standard Bellman equation of dynamic programming.) A Markov best response is then identified by finding, for each state x, the action $a_i \in A_i(x)$ that maximizes the right hand side of the above equation.

3 Existence of MPE

It is straightforward to see that an MPE exists in any N-player game with a finite state space and finite action spaces. We use a reduction to a standard finite game.

For each player i and state x, we create a new player with action space $\tilde{A}(i, x) = A_i(x)$; we refer to any player in this new game as an *agent*. When the actions chosen by all agents are $\boldsymbol{a} = (a(i, x), x \in \mathcal{X}, i = 1, \ldots, N)$, the payoff to player (i, x) is:

$$R_{i,x}(\boldsymbol{a}) = \mathbb{E}\left[\sum_{t \geq 0} \delta^t Q_i(a(1, x^t), \ldots, a(N, x^t); x^t) \Big| x^0 = x \right].$$

Note that this game has finitely many players, and each player has finitely many actions. Thus the agent game has a (possibly mixed) Nash equilbrium, where agent (i, x) plays the (possibly mixed) action $a(i, x)$.

Define a strategy for player i where $s_i(x) = a(i, x)$. We claim that $\boldsymbol{s}$ is a MPE. Clearly each player's strategy depends only on the current state. Further, observe that by construction, the strategy of player i maximizes his payoff among all Markov strategies, given $\boldsymbol{s}_{-i}$. Since we saw in the previous section that each player i has a best response that is a Markov strategy when all opponents play Markov strategies, we conclude that $\boldsymbol{s}$ must be a MPE.

4 Two-Player, Zero-Sum Stochastic Games

In this section, following the development of Shapley [1], we consider two-player zero-sum stochastic games. We begin by reviewing the *minimax theorem* of Von Neumann for static two-player

zero-sum games. Let P be a $m \times n$ matrix of payoffs; player 1 has m actions (given by the set A_1), and player 2 has n actions (given by the set A_2). The entry P_{ij} is the payoff to player 1 when (i, j) is played; since the game is zero-sum, the payoff to player 2 is $-P_{ij}$ in this case. The minimax theorem states that:

$$\max_{s_1 \in \Delta(A_1)} \min_{s_2 \in \Delta(A_2)} s_1^\top \boldsymbol{P} s_2 = \min_{s_2 \in \Delta(A_2)} \max_{s_1 \in \Delta(A_1)} s_1^\top \boldsymbol{P} s_2. \tag{1}$$

Here s_i is a mixed strategy for player i, represented as a vector with entries indexed by the actions in A_i; $\Delta(A_i)$ is the set of all mixed strategies for player i; and $s_1^\top \boldsymbol{P} s_2 = \sum_{i,j} s_1(i) s_2(j) P_{ij}$ is the expected payoff to player 1 when (s_1, s_2) is played. There are many proofs, some using linear programming, and others using fixed point theorems. The standard game theoretic proof is that all finite games have Nash equilibria, and any Nash equilibrium of the minimax theorem yields (1). Indeed, at any Nash equilibrium, the expected payoff to player 1 is the value given in (1):

$$\mathrm{val}(\boldsymbol{P}) = \max_{s_1 \in \Delta(A_1)} \min_{s_2 \in \Delta(A_2)} s_1^\top \boldsymbol{P} s_2 = \min_{s_2 \in \Delta(A_2)} \max_{s_1 \in \Delta(A_1)} s_1^\top \boldsymbol{P} s_2. \tag{2}$$

The scalar $\mathrm{val}(\boldsymbol{P})$ is called the *value* of the matrix game defined by $\boldsymbol{P}$.

Shapley showed that the minimax theorem extends to two-player zero-sum stochastic games; all such games also have a value. The proof is via a technique that is very standard in dynamic programming, called *value iteration*. In dynamic programming, value iteration is used to find both optimal policies and the optimal cost or profit of a stochastic control problem.

We will need the following lemma.

Lemma 1 *For any two $m \times n$ matrices $\boldsymbol{B}, \boldsymbol{C}$, there holds:*

$$|\operatorname{val} \boldsymbol{B} - \operatorname{val} \boldsymbol{C}| \leq \max_{i,j} |B_{ij} - C_{ij}|.$$

Proof of Lemma. Let (s_1, s_2) be a NE of the zero-sum matrix game defined by $\boldsymbol{B}$, and let $(\overline{s}_1, \overline{s}_2)$ be a NE of the zero-sum matrix game defined by $\boldsymbol{C}$. Then $s_1^\top \boldsymbol{B} \overline{s}_2 \geq s_1^\top \boldsymbol{B} s_2$, and $\overline{s}_1^\top \boldsymbol{C} \overline{s}_2 \geq s_1^\top \boldsymbol{C} \overline{s}_2$, by the NE optimality conditions for each player. Thus:

$$s_1^\top \boldsymbol{B} s_2 - \overline{s}_1^\top \boldsymbol{C} \overline{s}_2 \leq s_1^\top \boldsymbol{B} \overline{s}_2 - s_1 \boldsymbol{C} \overline{s}_2 \leq \max_{i,j} |B_{ij} - C_{ij}|.$$

By symmetry the same claim holds with $\boldsymbol{B}$ and $\boldsymbol{C}$ reversed, proving the result. □

Now consider a two person zero-sum stochastic game; since the game is zero-sum, we drop the index i on the stage payoff Q.

Value iteration proceeds as follows. First, we pick an arbitrary function $\alpha : \mathcal{X} \to \mathbb{R}$, called a *value function*. For each $x \in \mathcal{X}$, define a matrix $\boldsymbol{R}_x(\alpha)$ as follows:

$$R_x(\alpha)(a_1, a_2) = Q(a_1, a_2; x) + \delta \sum_{x' \in \mathcal{X}} \mathbb{P}(x' | a_1, a_2, x) \alpha(x), \;\; a_1 \in A_1(x), \; a_2 \in A_2(x).$$

Value iteration initializes with a value function α_0. The value function α_k is defined by $\alpha_k(x) = \mathrm{val}(\boldsymbol{R}_x(\alpha_{k-1}))$. It is convenient to define the shorthand operator notation $(T\alpha)(x) = \mathrm{val}(\boldsymbol{R}_x(\alpha))$;

with this notation, $\alpha_k = T\alpha_{k-1}$. (Note that this is analogous to the dynamic programming operator; indeed, if player 2 only had one action available, this would be exactly the dynamic programming operator.)

To interpret $\alpha_k(x)$ consider a two-player zero-sum game with k stages, where stages are counted down from k to 1, and the game starts in state x at stage k. At any stage with a positive index, payoffs accrue according to the stage game payoff Q, and then play proceeds to the next state. At the terminal state payoffs are given by α_0. Since this game is zero-sum with finitely many (pure) strategies for each player, it has a value. This value is exactly $\alpha_k(x)$, which is easily shown by induction: if $\alpha_{k-1}(x')$ is the value of the $k-1$ stage game terminating with α_0, then player 1 can guarantee himself a payoff of $\alpha_k(x)$ in the k stage game starting from x, while player 2 can guarantee herself a payoff $-\alpha_k(x)$.

For any real vector $x \in \mathbb{R}^J$, let $\|x\|_\infty = \sup_j |x_j|$ (this is called the sup norm). Observe that for any two functions α, α' we have:

$$\begin{aligned}
\|T\alpha - T\alpha'\|_\infty &= \max_{x\in\mathcal{X}} |\operatorname{val}(\boldsymbol{R}_x(\alpha) - \operatorname{val}(\boldsymbol{R}_x(\alpha'))| \\
&\leq \delta \max_{x\in\mathcal{X}} \max_{a_1\in A_1(x), a_2\in A_2(x)} \left| \sum_{x'\in\mathcal{X}} \mathbb{P}(x'|a_1,a_2,x)(\alpha(x') - \alpha'(x')) \right| \\
&\leq \delta \max_{x'\in\mathcal{X}} |\alpha(x') - \alpha'(x')| \\
&= \delta \|\alpha - \alpha'\|_\infty.
\end{aligned}$$

(The first inequality follows from our lemma.)

Since $\delta \in (0,1)$, this argument establishes that T is a *contraction*; and thus, regardless of the initial value function α_0, the sequence α_k converges to a unique limit α^* that satisfies $\alpha^* = T\alpha^*$.

We now have two propositions: one that establishes that two-player zero-sum stochastic games have a value, and the second that finds optimal strategies for the players.

Theorem 2 *Given a two-player zero-sum stochastic game, define α^* as the unique solution to $\alpha^* = T\alpha^*$. A pair of strategies (s_1, s_2) is a subgame perfect equilibrium if and only if after any history leading to the state x, the expected discounted payoff to player 1 is exactly $\alpha^*(x)$.*

Proof. Suppose the game is in state x. Suppose that for the next k periods, player 1 plays an optimal strategy from the k stage game, with terminal payoffs $\alpha_0(x') = 0$ for all $x' \in \mathcal{X}$; after the first k periods, player 1 can play any strategy. Regardless of player 2's strategy, this approach guarantees player 1 an expected discounted payoff in the infinite game of at worst:

$$\alpha_k(x) - \frac{\delta^k}{1-\delta} M, \tag{3}$$

where:

$$M = \max_{x'\in X} \max_{a_1\in A_1(x'), a_2\in A_2(x')} |Q(a_1,a_2;x)|.$$

The guarantee follows by the fact that α_k is the value of the k-stage game with terminal payoffs α_0.

As $k \to \infty$ in (3), we conclude that player 1 can lower bound his payoff by $\alpha^*(x)$. A similar argument shows player 2 can lower bound her payoff by $-\alpha^*(x)$. This establishes the claim of the theorem. □

An *optimal strategy* for player 1 (resp., player 2) is a strategy for player 1 that guarantees player 1 (resp., player 2) a payoff of at least $\alpha^*(x)$ (resp., $-\alpha^*(x)$). Note that although the preceding proof establishes that all stochastic games have a value, it does not provide optimal stationary strategies for each player. We provide these in the following proposition.

Proposition 3 *Let $s_1(x), s_2(x)$ be optimal (possibly mixed) strategies for players 1 and 2 in zero-sum game defined by the matrix $\boldsymbol{R}_x(\alpha^*)$. Then s_1, s_2 are optimal strategies in the stochastic game for both players; in particular, (s_1, s_2) is an MPE.*

Proof. Fix a (possibly history dependent) strategy $\hat{s}_2$ for player 2. We first consider a k stage game, where terminal payoffs are given by α^*. In this game, it follows that player 1 can guarantee a payoff of at least $\alpha^*(x)$ by playing the strategy s_1 given in the proposition, regardless of the strategy of player 2. Thus we have:

$$\mathbb{E}\left[\sum_{t=0}^{k-1} \delta^t Q(s_1(x^t), \hat{s}_2(x^t); x^t) + \delta^k \alpha^*(x^k) \middle| x^0 = x\right] \geq \alpha^*(x).$$

The preceding implies:

$$\mathbb{E}\left[\sum_{t=0}^{k-1} \delta^t Q(s_1(x^t), \hat{s}_2(x^t); x^t) \middle| x^0 = x\right] \geq \alpha^*(x) - \delta^k \|\alpha^*\|_\infty.$$

This in turn implies:

$$\Pi(s_1, \hat{s}_2; x) \geq \alpha^*(x) - \delta^k \|\alpha^*\|_\infty - \frac{\delta^k}{1-\delta} M.$$

As $k \to \infty$, the right hand side approaches $\alpha^*(x)$, as required; the proof for player 2 is symmetric. □

Note that although such an approach *guarantees* that player 1 will do no worse than α^*, in practice "mistakes" by player 2 may allow player 1 to develop strategies that perform better than α^*. The study of learning in zero-sum stochastic games is devoted to exactly this problem: achieving α^* against an adversary, but also exploiting possibilities for further gain when the opponent is not adversarial.

References

[1] L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39:1095–1100, 1953.