This lecture introduces our study of learning in games. We first give a conceptual overview of the possible approaches to studying learning in repeated games; in particular, we distinguish between approaches that use a Bayesian model of the opponents, vs. nonparametric or "model-free" approaches to playing against the opponents. Our investigation will focus almost entirely on the second class of models, where the main results are closely tied to the study of *regret minimization* in online regret. We introduce two notions of regret minimization, and also consider the corresponding equilibrium notions. (Note that we focus attention on repeated games primarily because learning results in stochastic games are significantly weaker.)

Throughout the lecture we consider a finite N-player game, where each player i has a finite pure action set A_i; let $A = \prod_i A_i$, and let $A_{-i} = \prod_{j \neq i} A_j$. We let a_i denote a pure action for player i, and let $s_i \in \Delta(A_i)$ denote a mixed action for player i. We will typically view s_i as a vector in $\mathbb{R}^{A_i}$, with $s_i(a_i)$ equal to the probability that player i places on a_i. We let $\Pi_i(\boldsymbol{a})$ denote the payoff to player i when the composite pure action vector is $\boldsymbol{a}$, and by an abuse of notation also let $\Pi_i(\boldsymbol{s})$ denote the expected payoff to player i when the composite mixed action vector is $\boldsymbol{s}$. More generally, if $\boldsymbol{q}$ is a joint probability distribution on the set A, we let $\Pi_i(\boldsymbol{q}) = \sum_{\boldsymbol{a} \in A} q(\boldsymbol{a}) \Pi(\boldsymbol{a})$. We let $BR_i(\boldsymbol{s}_{-i})$ denote the best response mapping of player i; here $\boldsymbol{s}_{-i}$ is the composite mixed action vector of players other than i.

The game is played repeatedly by the players. We let $h^T = (\boldsymbol{a}^0, \ldots, \boldsymbol{a}^{T-1})$ denote the history up to time T. We let $p_i^T \in \Delta(A_i)$ denote the marginal empirical distribution of player i's play up to time T:

$$p_i^T(a_i) = \frac{1}{T} \sum_{t=0}^{T-1} \mathcal{I}\{a_i^t = a_i\}.$$

Similarly, we let q^T denote the joint empirical distribution of play up to time T:

$$q^T(\boldsymbol{a}) = \frac{1}{T} \sum_{t=0}^{T-1} \mathcal{I}\{\boldsymbol{a}^t = \boldsymbol{a}\}.$$

1 Introduction

We consider a setup where the same game is repeatedly played by a finite set of players. Game theory is concerned with predicting the behavior of players in such a situation; however, the traditional notions of equilibrium (e.g., sequential equilibrium as studied in Lecture 1) typically impose very strong coordination requirements on the players. In particular, despite the fact that sequential equilibrium is an equilibrium concept for *dynamic games*, it requires players to play best responses to the complete contingent plans of their opponents. Effectively, it is as if the players had access to their opponents' strategies before play began, reducing equilibrium analysis to a *static* problem. This problem is further complicated by the fact that games typically exhibit many equilibria: how do players coordinate on the "right" equilibrium?

In practice, players rarely interact with prior understanding of the complete (history-dependent) strategies their opponents will play. Instead, coordination occurs *through* the process of play. This perspective is referred to as *learning in games*. (Note that fictitious play and adaptive learning are examples.) It is worth noting that learning in games is distinguished from other areas of learning by the attention to the interaction of multiple decision makers; in particular, in learning in games we are typically concerned with arriving at a characterization of the limiting behavior of the players, and whether such behavior corresponds to any natural notion of equilibrium.

At a high level, learning in games requires attention to the following basic issues:

- How does a player judge her performance? In particular, what objective is being maximized (or minimized), and over what set of strategies?
- How does a player model her opponents? In particular, how does the player model the (history-dependent) strategy that will be played by an opponent?

Some examples:

1. In a *sequential equilibrium*, each player maximizes expected discounted payoff (or average payoff), over the class of all history-dependent strategies. This optimization is carried out given the exact knowledge of the history-dependent strategies of the opponents.
2. *Markov perfect equilibrium* has the same characteristics as a sequential equilibrium, except that a player maximizes only over the class of Markov strategies.
3. In *Bayesian learning*, each player again acts to maximize discounted payoffs or average payoffs; however, it is assumed that players are not certain about the strategies of their opponents. Players begin with a prior distribution over a class of models for their opponents, and update their beliefs on the basis of observed play. (Note the class of models need not include the actual strategies played by the opponents.)
4. In *fictitious play*, each player maximizes only their one period payoff, with respect to available pure actions; that is, they act myopically. Further, each player assumes the opponents are stationary, and plays according to the product of the marginal empirical distributions of past play.
5. In *best response dynamics*, each player acts as in fictitious play, but now it is assumed that opponents will play exactly as they did in the last time period.

As the preceding examples illustrate, the two elements—judging performance, and modeling one's opponent—interact strongly with each other: in order for a player to make "optimal" decisions, the player must also have a forecast or belief about the behavior of her opponents.

It is clear that best response dynamics and sequential equilibrium are at opposite extremes in the study of repeated games. Sequential equilibrium assumes players are rational over the infinite horizon of the game, and that they also optimize given the strategies of their opponents. By contrast, best response dynamics assumes that individuals are highly myopic, and further, they are very naive in modeling their opponents.

Our goal is to allow some more sophistication into the discussion, but without requiring that players have full access to their opponents' strategies before playing the game. There are two main approaches we could take. The first, discussed above, is Bayesian (or parametric) learning; such a model requires that a player have prior beliefs about his opponents' possible behavior. Unfortunately, general results in this settings are somewhat limited; typically, under assumptions that the prior contains information about the opponents' true behavior, it is possible to show that play approaches Nash equilibrium. However, the assumptions are difficult to verify in practice.

Instead, we will focus our attention on a model-free or nonparametric approach to learning in games, where a player has no prior assumptions about his opponents. In this model, a player evaluates play retrospectively; *a priori* optimal decisions are not well defined, since a player does not have any assumptions or model about how the opponents will behave. For this reason, nonparametric learning in games is closely connected to the theory of *regret minimization* in online optimization.

2 Regret Minimization: An Example

Recall the coordination game studied in Lecture 7:

		Player 2 a	Player 2 b
Player 1	A	(1,1)	(0,0)
	B	(0,0)	(1,1)

This is a coordination game with a unique fully mixed Nash equilibrium, where both players put probability $1/2$ on each action. Recall that discrete time fictitious play, where we start with $p_1^0(A) = 1 - p_1^0(B) = 3/4$, and $p_2^0(a) = 1 - p_2^0(b) = 1/4$, has the following trajectory:

t	p_1^t	p_2^t	a_1^t	a_2^t
0	(3/4, 1/4)	(1/4, 3/4)	B	a
1	(3/4, 5/4)	(5/4, 3/4)	A	b
2	(7/4, 5/4)	(5/4, 7/4)	B	a
3	(7/4, 9/4)	(9/4, 7/4)	A	b
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Clearly, each player will receive zero payoff every time period. Each player would have obviously been better off using a strategy that uniformly randomized each time period, rather than playing using the prescribed action of DTFP. Alternatively, with hindsight, each player would have been better with a constant strategy (i.e., a strategy that always plays the same action regardless of history), instead of the prescription of DTFP. An even more sophisticated view would be to

observe that switching actions (i.e., playing A whenever B was played, and playing B whenever A was played) would have yielded perfect coordination along the history of play.

This example embodies the idea of *regret*. Each player may have begun with no model of their opponent; however, after play has progressed, each player can look into the past and ask whether they could have done better. The difference in payoff between an alternate strategy and the actual strategy pursued is called the regret. Regret minimization algorithms aim to ensure that long term average regret is small. We define two notions of regret minimization in this lecture, that vary depending on the space of strategies against which regret is computed. The first is *external regret minimization*, also known as Hannan consistency or universal consistency; and the second is *internal regret minimization*.

3 External Regret Minimization and the Hannan Set

In external regret minimization, a player compares his average payoff to the payoff that would have been received against the best constant action. We start by defining regret against the (possibly mixed) action s_i, after history h^T:

$$ER_i(h^T; s_i) = \sum_{t=0}^{T-1} \Pi_i(s_i, \boldsymbol{a}_{-i}^t) - \Pi_i(a_i^t, \boldsymbol{a}_{-i}^t).$$

We now let $ER_i(h^T)$ be the external regret of player i after history h^T:

$$ER_i(h^T) = \max_{a_i \in A_i} ER_i(h^T; a_i).$$

(Note the maximum is only over pure actions; since player i maximizes expected payoff, this is equivalent to the maximum over mixed strategies.)

We pause to note an important feature of regret minimization that is particularly problematic in the context of games. Note that although player i considers the payoff that might have been achieved using the best constant action, he does *not* assume that his opponents would have reacted differently if this constant action had been played—even though, in general, history-dependent strategies of the opponents would have led to a different realized path of play. This is a critical aspect of the model-free approach to learning. Since there is no strategic representation of the opponents from player i's point of view, there is no way for player i to determine how his opponents would have reacted had player i played differently. Of course, this means that regret may ultimately have been an illusory objective: even the regret minimizing action may not have performed well if actually played. This same issue remains even under more sophisticated definitions of regret.

The goal of external regret minimizing algorithms is to ensure that the time average regret approaches zero, regardless of the opponents' strategies. We say that a (history-dependent) strategy for player i is *external regret minimizing*, or *Hannan consistent*, if for all strategies of players other than i, there holds:

$$\limsup_{T \to \infty} \frac{1}{T} ER_i(h^T) \leq 0, \quad \text{almost surely.}$$

Note that "almost surely" in the preceding statement is with respect to any randomization employed by the players. Importantly, players other than i are not allowed to observe the randomization employed by player i; this is implicit in the fact that the strategies of the other players are dependent only on the history of pure actions played.

We will later establish the remarkable fact that Hannan consistent strategies exist (including many variants of stochastic fictitious play). For now, we investigate the consequences of Hannan consistency for the joint empirical distribution of play. In particular, we ask: if all players use Hannan consistent strategies, then what is their limiting behavior? Does it correspond to a natural equilibrium notion?

3.1 Zero-Sum Games

We start by studying the implications of Hannan consistency in two player zero-sum games, i.e., games where $N = 2$, and $\Pi_1(a_1, a_2) = -\Pi_2(a_1, a_2)$ for all (a_1, a_2).

Proposition 1 *Suppose player 1 uses a Hannan consistent strategy. Then regardless of the strategy of player 2,*

$$\liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \Pi_1(a_1^t, a_2^t) \geq \mathrm{val}(\Pi_1), \quad \textit{almost surely}.$$

(Here $\mathrm{val}(\Pi_1)$ is the value of the zero-sum game with payoff matrix Π_1.)

Proof. Since player 1 is using a Hannan consistent strategy, it suffices to show:

$$\liminf_{T\to\infty} \frac{1}{T} \max_{a_1\in A_1} \sum_{t=0}^{T} \Pi_1(a_1, a_2^t) \geq \mathrm{val}(\Pi_1), \quad \text{almost surely}.$$

Observe that:

$$\begin{aligned}
\frac{1}{T} \max_{a_1\in A_1} \sum_{t=0}^{T} \Pi_1(a_1, a_2^t) &= \frac{1}{T} \max_{s_1\in\Delta(A_1)} \sum_{t=0}^{T} \Pi_1(s_1, a_2^t) \\
&= \max_{s_1\in\Delta(A_1)} \Pi_1(s_1, p_2^T) \\
&\geq \min_{s_2\in\Delta(A_2)} \max_{s_1\in\Delta(A_1)} \Pi_1(s_1, s_2) = \mathrm{val}(\Pi_1).
\end{aligned}$$

(Here p_2^T is the empirical distribution of player 2's play up to time T.) This establishes the desired result. □

We emphasize that the preceding proposition would hold regardless of whether the game were a zero-sum game or not; in other words, in any game, a player using a Hannan consistent strategy achieves a long run average payoff that is at least his maximin value (i.e., the best payoff that could be achieved if it were known in advance that other players would play adversarially).

The preceding proposition also establishes that if both players use Hannan consistent strategies, then:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \Pi_1(a_1^t, a_2^t) = \text{val}(\Pi_1).$$

Thus asymptotically, player 1's average payoff is equal to the value of the game. Of course, using a Hannan consistent strategy means that if player 2 were not adversarial, then player 1 could potentially earn a significantly higher average payoff than his maximin value.

We conclude by noting that the same argument as used in the proposition can also yield a learning-theoretic proof of the minimax theorem for zero-sum games. To see this, observe that if player 1 uses a Hannan consistent strategy, then the proof of the proposition implies:

$$\liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \Pi_1(a_1^t, a_2^t) \geq \min_{s_2\in\Delta(A_2)} \max_{s_1\in\Delta(A_1)} \Pi_1(s_1, s_2).$$

Similarly, if player 2 uses a Hannan consistent strategy, then we have:

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \Pi_1(a_1^t, a_2^t) \leq \max_{s_1\in\Delta(A_1)} \min_{s_2\in\Delta(A_2)} \Pi_1(s_1, s_2).$$

Finally, the following inequality holds by an elementary argument:

$$\min_{s_2\in\Delta(A_2)} \max_{s_1\in\Delta(A_1)} \Pi_1(s_1, s_2) \leq \max_{s_1\in\Delta(A_1)} \min_{s_2\in\Delta(A_2)} \Pi_1(s_1, s_2).$$

Combining these results shows that:

$$\min_{s_2\in\Delta(A_2)} \max_{s_1\in\Delta(A_1)} \Pi_1(s_1, s_2) = \max_{s_1\in\Delta(A_1)} \min_{s_2\in\Delta(A_2)} \Pi_1(s_1, s_2),$$

which is exactly the minimax theorem. (In fact, Hannan consistent strategies can also be used to show the existence of optimal strategies in two-player zero-sum games.)

3.2 The Hannan Set

We now turn our attention to characterizing limiting play in general games, when all players use Hannan consistent strategies. Note that Hannan consistency for player i implies that for any $a_i' \in A_i$, the following limit holds almost surely:

$$\limsup_{T\to\infty} \sum_{\boldsymbol{a}\in A} q^T(\boldsymbol{a}) \left(\Pi_i(a_i', \boldsymbol{a}_{-i}) - \Pi_i(\boldsymbol{a})\right) \leq 0.$$

From the preceding expression, it is not difficult to show that if all players use Hannan consistent strategies, then the joint frequency of play converges (almost surely) to the following set:

$$\mathcal{H} = \{q \in \Delta(A) : \sum_{\boldsymbol{a}\in A} q(\boldsymbol{a})\Pi_i(a_i', \boldsymbol{a}) \leq \sum_{\boldsymbol{a}\in A} q(\boldsymbol{a})\Pi_i(\boldsymbol{a}), \text{ for all } i, a_i'\}.$$

(In other words, $\lim_{T\to\infty} d(q^T, \mathcal{H}) = 0$ almost surely, where $d(q, S) = \inf_{q'\in S} \|q - q'\|$ is the distance from q to the set S.)

The set $\mathcal{H}$ of joint distributions is called the *Hannan set* of the game. We interpret it as follows. Suppose that each player i sees a private "recommendation" of play a_i from an oracle, and that the joint vector of recommendations $\boldsymbol{a}$ is sampled according to $q(\boldsymbol{a})$. If $q \in \mathcal{H}$, then a unilateral deviation by player i to a_i'—regardless of the recommendation seen by i—is not profitable, assuming that other players play according to their recommendations. We can thus conclude that external regret minimization leads to a notion of equilibrium defined by the Hannan set.

The Hannan set provides a somewhat strange equilibrium concept. In general, the recommendation a_i will be correlated with the recommendations of the other players, and thus provide information about the expected payoff to player i under action a_i'. Thus, in particular, it may be that for a distribution q in the Hannan set, player i may wish to deviate *after* seeing his recommendation a_i, despite the fact that *before* seeing the recommendation no profitable deviation existed (in expectation). Robustness to this stronger notion of deviation leads us to the definition of *correlated equilibrium*, and the study of regret minimization algorithms that lead to correlated equilibrium.

4 Internal Regret Minimization and Correlated Equilibrium

Correlated equilibrium is an equilibrium concept that strengthens the deviation property allowed in the definition of the Hannan set. In particular, we now allow a player to deviate *after* having observed his recommendation. Thus player i should now compute his expected payoff *conditional* on having observed the recommendation a_i, and assuming all other players play according to their recommendations. Formally, the set of correlated equilibria is:

$$\mathcal{CE} = \{q \in \Delta(A) : \sum_{\boldsymbol{a}_{-i}\in A_{-i}} q(\boldsymbol{a}_{-i}|a_i)\Pi_i(a_i', \boldsymbol{a}) \leq \sum_{\boldsymbol{a}_{-i}\in A_{-i}} q(\boldsymbol{a}_{-i}|a_i)\Pi_i(\boldsymbol{a}), \text{ for all } i, a_i, a_i'\}.$$

Clearly, $\mathcal{CE} \subset \mathcal{H}$. Further, it is clear that any Nash equilibrium is a correlated equilibrium, since it is a product distribution. Finally, we note that $\mathcal{CE}$ (as well as $\mathcal{H}$) is a closed, convex subset of the set of joint distributions on A. To see this, rewrite the definition of $\mathcal{CE}$ as follows:

$$\mathcal{CE} = \{q \in \Delta(A) : \sum_{\boldsymbol{a}_{-i}\in A_{-i}} q(a_i, \boldsymbol{a}_{-i})\Pi_i(a_i', \boldsymbol{a}) \leq \sum_{\boldsymbol{a}_{-i}\in A_{-i}} q(a_i, \boldsymbol{a}_{-i})\Pi_i(\boldsymbol{a}), \text{ for all } i, a_i, a_i'\}.$$

Thus $\mathcal{CE}$ is defined by a collection of linear inequalities, and hence convex.

Typically, the study of correlated equilibrium assumes the existence of a correlated randomization device that can provide private recommendations to each player. However, in dynamic play, that correlated randomization occurs *implicitly* through the past history of play; indeed, this is why external regret minimization algorithms lead to joint empirical distributions in the Hannan set, despite the lack of an explicit correlation device. We now ask: do there exist analogous algorithms that lead to correlated equilibrium?

To address this issue, we introduce *internal regret minimization*. In internal regret minimization, player i considers switching any time she played a_i to a_i'. We start by defining regret of a_i

against a_i', after history h^T:

$$IR_i(h^T; a_i, a_i') = \sum_{t=0}^{T-1} \mathcal{I}\{a_i^t = a_i\} \left(\Pi_i(a_i', \boldsymbol{a}_{-i}^t) - \Pi_i(a_i, \boldsymbol{a}_{-i}^t)\right).$$

Thus the preceding expression compares the payoff player i would have achieved if she had switched to playing a_i' at every time in the past that she played a_i. The *interal regret* of player i after history h^T is then:

$$IR_i(h^T) = \max_{a_i, a_i' \in A_i} IR_i(h^T; a_i, a_i').$$

We say that a (history-dependent) strategy for player i is *internal regret minimizing* if for all strategies of players other than i, there holds:

$$\limsup_{T\to\infty} \frac{1}{T} IR_i(h^T) = 0, \quad \text{almost surely.}$$

As before, "almost surely" here refers to any randomization employed by player i and/or the other players. We will show later that internal regret minimizing strategies exist; in fact, it is possible to construct such strategies using external regret minimizing strategies.

Note that we can rewrite internal regret minimization as follows: for all players i, and for all pairs $a_i, a_i' \in A_i$, there holds (almost surely):

$$\limsup_{T\to\infty} \sum_{\boldsymbol{a}_{-i} \in A_{-i}} q^T(a_i, \boldsymbol{a}_{-i}) \left(\Pi_i(a_i', \boldsymbol{a}_{-i}) - \Pi_i(a_i, \boldsymbol{a}_{-i})\right) \leq 0.$$

(Note that equality in the definition of internal regret minimization follows by considering $a_i' = a_i$.) Using the previous expression, and an argument analogous to that used for external regret minimization, we can easily conclude that the joint empirical distribution approaches the set of correlated equilibria (almost surely) if all players use internal regret minimizing strategies.