7

Prediction and Playing Games

7.1 Games and Equilibria

The prediction problems studied in previous chapters have been often represented as repeated games between a forecaster and the environment. Our use of a game-theoretic formalism is not accidental: there exists an intimate connection between sequential prediction and some fundamental problems belonging to the theory of learning in games. We devote this chapter to the exploration of some of these connections.

Rather than giving an exhaustive account of the area of learning in games, we only focus on “regret-based” learning procedures (i.e., situations in which the players of the game base their strategies only on regrets they have suffered in the past) and our fundamental concern is whether such procedures lead to equilibria. We also limit our attention to finite strategic or normal form games.

In this introductory section we present the basic definitions of the games we consider, describe some notions of equilibria, and introduce the model of playing repeated games that we investigate in the subsequent sections of this chapter.

K-Person Normal Form Games

A (finite) K-person game given in its strategic (or normal) form is defined as follows. Player \( k = 1, \ldots, K \) has \( N_k \) possible actions (or pure strategies) to choose from, where \( N_k \) is a positive integer. If the action of each player \( k = 1, \ldots, K \) is \( i_k \in \{1, \ldots, N_k\} \) and we denote the \( K \)-tuple of all the players’ actions by \( i = (i_1, \ldots, i_K) \in \bigotimes_{k=1}^{K} \{1, \ldots, N_k\} \), then the loss suffered by player \( k \) is \( \ell_k(i) \), where \( \ell_k : \bigotimes_{k=1}^{K} \{1, \ldots, N_k\} \rightarrow [0, 1] \) for each \( k = 1, \ldots, K \) are given loss functions for all players. Note that, slightly deviating from the usual game-theoretic terminology, we consider losses as opposed to the more standard payoffs. The reader should keep in mind that the goal of each player is to minimize his loss, which is the same as maximizing payoffs if one defines payoffs as negative losses. We use this convention to harmonize notation with the rest of the book.

A mixed strategy for player \( k \) is a probability distribution \( \mathbf{p}^{(k)} = (p_1^{(k)}, \ldots, p_{N_k}^{(k)}) \) over the set \( \{1, \ldots, N_k\} \) of actions. When mixed strategies are used, players randomize, that is, choose an action according to the distribution specified by the mixed strategy. Denote the action played by player \( k \) by \( I^{(k)} \). Thus, \( I^{(k)} \) is a random variable taking values in the set \( \{1, \ldots, N_k\} \) and distributed according to \( \mathbf{p}^{(k)} \). Let \( \mathbf{I} = (I^{(1)}, \ldots, I^{(K)}) \) denote the \( K \)-tuple of actions played by all players. If the random variables \( I^{(1)}, \ldots, I^{(K)} \) are independent (i.e., the players randomize independently of each other), we denote their joint distribution by \( \pi \). That is, \( \pi \) is the joint distribution over the set \( \bigotimes_{k=1}^{K} \{1, \ldots, N_k\} \) of all possible \( K \)-tuples of actions.
obtained by the product of the mixed strategies \( p^{(1)}, \ldots, p^{(K)} \). The product distribution \( \pi \) is called a mixed strategy profile. Thus, for all \( i = (i_1, \ldots, i_K) \in \otimes_{k=1}^{K} \{1, \ldots, N_k\}, \)

\[
\pi(i) = \mathbb{P}[I = i] = p^{(1)}_{i_1} \times \cdots \times p^{(K)}_{i_K}.
\]

The expected loss of player \( k \) is

\[
\pi \ell^{(k)} \overset{\text{def}}{=} \mathbb{E} \ell^{(k)}(I) = \sum_{i \in \otimes_{k=1}^{K} \{1, \ldots, N_k\}} \pi(i) \ell^{(k)}(i) = \sum_{i_1=1}^{N_1} \cdots \sum_{i_K=1}^{N_K} p^{(1)}_{i_1} \times \cdots \times p^{(K)}_{i_K} \ell^{(k)}(i_1, \ldots, i_K).
\]

**Nash Equilibrium**

Perhaps the most important notion of game theory is that of a Nash equilibrium. A mixed strategy profile \( \pi = p^{(1)} \times \cdots \times p^{(K)} \) is called a Nash equilibrium if for all \( k = 1, \ldots, K \) and all mixed strategies \( q^{(k)}_k \), if \( \pi' = p^{(1)}_1 \times \cdots \times q^{(k)}_k \times \cdots \times p^{(K)}_K \) denotes the mixed strategy profile obtained by replacing \( p^{(k)}_k \) by \( q^{(k)}_k \) and leaving all other players’ mixed strategies unchanged, then

\[
\pi \ell^{(k)} \leq \pi' \ell^{(k)}.
\]

This means that if \( \pi \) is a Nash equilibrium, then no player has an incentive of changing his mixed strategy if all other players do not change theirs (i.e., every player is happy). A celebrated result of Nash [222] shows that every finite game has at least one Nash equilibrium. (The proof is typically based on fixed-point theorems.) However, a game may have multiple Nash equilibria, and the set \( N \) of all Nash equilibria can have a quite complex structure.

**Two-Person Zero-Sum Games**

A simple but important special class of games is the class of two-person zero-sum games. These games are played by two players (i.e., \( K = 2 \)) and the payoff functions are such that for each pair of actions \( i = (i_1, i_2) \), where \( i_1 \in \{1, \ldots, N_1\} \) and \( i_2 \in \{1, \ldots, N_2\} \), the losses of the two players satisfy

\[
\ell^{(1)}(i) = -\ell^{(2)}(i).
\]

Thus, in such games the objective of the second player (often called *column player*) is to maximize the loss of the first player (the *row player*). To simplify notation we will just write \( \ell \) for \( \ell^{(1)} \), replace \( N_1, N_2 \) by \( N \) and \( M \), and write \((i, j)\) instead of \((i_1, i_2)\). Mixed strategies of the row and column players will be denoted by \( p = (p_1, \ldots, p_N) \) and \( q = (q_1, \ldots, q_M) \).

It is immediate to see that the product distribution \( \pi = p \times q \) is a Nash equilibrium if and only if for all \( p' = (p'_1, \ldots, p'_N) \) and \( q' = (q'_1, \ldots, q'_M) \),

\[
\sum_{i=1}^{N} \sum_{j=1}^{M} p_i q'_j \ell(i, j) \leq \sum_{i=1}^{N} \sum_{j=1}^{M} p_i q_j \ell(i, j) \leq \sum_{i=1}^{N} \sum_{j=1}^{M} p'_i q_j \ell(i, j).
\]
Introducing the simplifying notation

\[ \ell(p, q) = \sum_{i=1}^{N} \sum_{j=1}^{M} p_i q_j \ell(i, j) \]

the above is equivalent to

\[ \max_{q'} \ell(p, q') = \ell(p, q) = \min_{p'} \ell(p', q). \]

This obviously implies that

\[ \max_{q'} \ell(p, q') \leq \max_{q'} \min_{p'} \ell(p', q'), \]

and therefore the existence of a Nash equilibrium \( p \times q \) implies that

\[ \min_{p'} \max_{q'} \ell(p', q') \leq \max_{q'} \min_{p'} \ell(p', q'). \]

On the other hand, clearly, for all \( p \) and \( q' \), \( \ell(p, q') \geq \max_{q'} \min_{p'} \ell(p', q') \) and therefore for all \( p, \max_{q'} \ell(p, q') \geq \max_{q'} \min_{p'} \ell(p', q') \), and in particular,

\[ \min_{p'} \max_{q'} \ell(p', q') \geq \max_{q'} \min_{p'} \ell(p', q'). \]

In summary, the existence of a Nash equilibrium implies that

\[ \min_{p'} \max_{q'} \ell(p', q') = \max_{q'} \min_{p'} \ell(p', q'). \]

The common value of the left-hand and right-hand sides is called the value of the game and will be denoted by \( V \). This equation, known as von Neumann’s minimax theorem, is one of the fundamental results of game theory. Here we derived it as a consequence of the existence of Nash equilibria (which, in turn, is based on fixed-point theorems), but significantly simpler proofs may be given. In Section 7.2 we offer an elementary “learning-theoretic” proof, based on the basic techniques introduced in Chapter 2, of a powerful minimax theorem that, in turn, implies von Neumann’s minimax theorem for two-person zero-sum games.

It is also clear from the argument that any Nash equilibrium \( p \times q \) achieves the value of the game in the sense that

\[ \ell(p, q) = V \]

and that any product distribution \( p \times q \) with \( \ell(p, q) = V \) is a Nash equilibrium.

**Correlated Equilibrium**

An important generalization of the notion of Nash equilibrium, introduced by Aumann [16], is the notion of correlated equilibrium. A probability distribution \( P \) over the set \( \otimes_{k=1}^{K} \{1, \ldots, N_k\} \) of all possible \( K \)-tuples of actions is called a correlated equilibrium if for all \( k = 1, \ldots, K \),

\[ \mathbb{E} \ell^{(k)}(I) \leq \mathbb{E} \ell^{(k)}(I^-, \tilde{I}^{(k)}), \]

where the random variable \( I = (I^{(1)}, \ldots, I^{(k)}) \) is distributed according to \( P \) and \( (I^-, \tilde{I}^{(k)}) = (I^{(1)}, \ldots, I^{(k-1)}, \tilde{I}^{(k)}, I^{(k+1)}, \ldots, I^{(K)}) \), where \( \tilde{I}^{(k)} \) is an arbitrary \( \{1, \ldots, N_k\} \)-valued random variable that is a function of \( I^{(k)} \).
The distinguishing feature of the notion is that, unlike in the definition of Nash equilibria, the random variables \( I^{(k)} \) do not need to be independent, or, in other words, \( P \) is not necessarily a product distribution (hence the name “correlated”). Indeed, if \( P \) is a product measure, a correlated equilibrium becomes a Nash equilibrium. The existence of a Nash equilibrium of any game thus assures that correlated equilibria always exist.

A correlated equilibrium may be interpreted as follows. Before taking an action, each player receives a recommendation \( I^{(k)} \) such that the \( I^{(k)} \) are drawn randomly according to the joint distribution of \( P \). The defining inequality expresses that, in an average sense, no player has an incentive to divert from the recommendation, provided that all other players follow theirs. Correlated equilibria model solutions of games in which the actions of players may be influenced by external signals.

A simple equivalent description of a correlated equilibrium is given by the following lemma whose proof is left as an exercise.

**Lemma 7.1.** A probability distribution \( P \) over the set of all \( K \)-tuples \( i = (i_1, \ldots, i_K) \) of actions is a correlated equilibrium if and only if, for every player \( k \in \{1, \ldots, K\} \) and actions \( j, j' \in \{1, \ldots, N_k\} \), we have

\[
\sum_{1: i_k = j} P(i) \left( \ell^{(k)}(\bar{\mathbf{i}}^+) - \ell^{(k)}(\bar{\mathbf{i}}^-) \right) \leq 0,
\]

where \( (\mathbf{i}^-, j') = (i_1, \ldots, i_{k-1}, j', i_{k+1}, \ldots, i_K) \).

Lemma 7.1 reveals that the set of all correlated equilibria is given by an intersection of closed halfspaces and therefore it is a closed and convex polyhedron. For example, any convex combination of Nash equilibria is a correlated equilibrium. (This can easily be seen by observing that the inequalities of Lemma 7.1 trivially hold if one takes \( P \) as any weighted mixture of Nash equilibria.) However, there may exist correlated equilibria outside of the convex hull of Nash equilibria (see Exercise 7.4). In general, the structure of the set of correlated equilibria is much simpler than that of Nash equilibria. In fact, the existence of correlated equilibria may be proven directly and without having to resort to fixed point theorems. We do this implicitly in Section 7.4. Also, as we show in Section 7.4, given a game, it is computationally very easy to find a correlated equilibrium. Computing a Nash equilibrium, on the other hand, appears to be a significantly harder problem.

**Playing Repeated Games**

The most natural application of the ideas of randomized prediction of Chapters 4 and 6 is in the theory of playing repeated games. In the model we investigate, a \( K \)-person game (the so-called one-shot game) is played repeatedly such that at each time instant \( t = 1, 2, \ldots \) player \( k \ (k = 1, \ldots, K) \) selects a mixed strategy \( p_t^{(k)} = (p_{1,t}^{(k)}, \ldots, p_{N_k,t}^{(k)}) \) over the set \( \{1, \ldots, N_k\} \) of his actions and draws an action \( I_t^{(k)} \) according to this distribution. We assume that the randomizations of the players are independent of each other and of past randomizations. After the actions are taken, player \( k \) suffers a loss \( \ell^{(k)}(\mathbf{I}_t) \), where \( \mathbf{I}_t = (I_t^{(1)}, \ldots, I_t^{(K)}) \). Formally, at time \( t \) player \( k \) has access to \( U_t^{(k)} \), where the \( U_t^{(k)} \) are independent random
variables uniformly distributed in $[0, 1]$, and chooses $I_t^{(k)}$ such that

$$I_t^{(k)} = i \quad \text{if and only if} \quad U_t^{(k)} \in \left[ \sum_{j=1}^{i-1} p_j^{(k)}, \sum_{j=1}^i p_j^{(k)} \right]$$

so that

$$\mathbb{P}\left[ I_t^{(k)} = i \mid \text{past plays} \right] = p_{i,t}^{(k)}, \quad i = 1, \ldots, N_k.$$  

In the basic setup we assume that after taking an action $I_t^{(k)}$ each player observes all other players’ actions, that is, the whole $K$-tuple $I_t = (I_t^{(1)}, \ldots, I_t^{(K)})$. This means that the mixed strategy $p_t^{(k)}$ played at time $t$ may depend on the sequence of random variables $I_1, \ldots, I_{t-1}$.

However, we will be concerned only with “uncoupled” ways of playing, that is, when each player knows his own loss (or payoff) function but not those of the other players. More specifically, we only consider regret-based procedures in which the mixed strategy chosen by every player depends, in some way, on his past losses. We focus our attention on whether such simple procedures can lead to some kind of equilibrium of the (one-shot) game. In Section 7.3 we consider the simplest situation, the case of two-person zero-sum games. We show that a simple application of the regret-minimization procedures of Section 4.2 leads to a solution of the game in a quite robust sense.

In Section 7.4 it is shown that if all players play to keep their internal regret small, then the joint empirical frequencies of play converge to the set of correlated equilibria. The same convergence may also be achieved if every player uses a well-calibrated forecasting strategy to predict the $K$-tuple of actions $I_t$, and chooses an action that is a best reply to the forecasted distribution. This is shown in Section 7.6. A model for learning in games even more restrictive than uncoupledness is the model of an “unknown game.” In this model the players cannot even observe the actions of the other players; the only information available to them is their own loss suffered after each round of the game. In Section 7.5 we point out that the forecasting techniques for the bandit problem discussed in Chapter 6 allow convergence of the empirical frequencies of play even in this setup of limited information.

In Sections 7.7 and 7.8 we sketch Blackwell’s approachability theory. Blackwell considered two-person zero-sum games with vector-valued losses and proved a powerful generalization of von Neumann’s minimax theorem, which is deeply connected with the regret-minimizing forecasting methods of Chapter 4.

Sections 7.9 and 7.10 discuss the possibility of reaching a Nash equilibrium in uncoupled repeated games. Simple learning dynamics, versions of a method called “regret testing,” are introduced that guarantee that the joint plays approach a Nash equilibrium in some sense. The case of unknown games is also investigated.

In Section 7.11 we address an important criticism of the notion of Hannan consistency. In fact, the basic notions of regret compare the loss of the forecaster with that of the best constant action, but without taking into account that the behavior of the opponent may depend on the actions of the forecaster. In Section 7.11 we show that asymptotic regret minimization is possible even in the presence of opponents that react, although we need to impose certain restrictions on the behavior of the opponents.
Fictitious Play

We close this introductory section by discussing perhaps the most natural strategy for playing repeated games: fictitious play. Player \( k \) is said to use fictitious play if, at every time instant \( t \), he chooses an action that is a best response to the empirical distribution of the opponents’ play up to time \( t - 1 \). In other words, the player chooses

\[
I_t^{(k)} = \arg\min_{i_k \in \{1, \ldots, N_k\}} \frac{1}{t-1} \sum_{s=1}^{t-1} \ell^{(k)}(I_s^-, i_k),
\]

where \((I_s^-, i_k) = (I_s^{(1)}, \ldots, i_k, \ldots, I_s^{(K)})\). The nontrivial behavior of this simple strategy is a good demonstration of the complexity of the problem of describing regret-based strategies that lead to equilibrium. As we already mentioned in Section 4.3, fictitious play is not Hannan consistent. This fact may invite one to conjecture that there is no hope to achieve equilibrium via fictitious play. It may come as a surprise that this is often not the case. First, Robinson [246] proved that if in repeated playing of a two-person zero-sum game at each step both players use fictitious play, then the product distribution formed by the frequencies of actions played by both players converges to the set of Nash equilibria. This result was extended by Miyasawa [218] to general two-person games in which each player has two actions (i.e., \( N_k = 2 \) for all \( k = 1, 2 \)); see also [220] for more special cases in which fictitious play leads to Nash equilibria in the same sense. However, Shapley [265] showed that the result cannot be extended even for two-person non-zero-sum games. In fact, the empirical frequencies of play may not even converge to the set of correlated equilibria (see Exercise 7.2 for Shapley’s game).

7.2 Minimax Theorems

As a first contact with game-theoretic applications of the prediction problems studied in earlier chapters, we derive a simple learning-style proof of a general minimax theorem that implies von Neumann’s minimax theorem cited in the introduction of this chapter. In particular, we prove the following.

**Theorem 7.1.** Let \( f(x, y) \) denote a bounded real-valued function defined on \( \mathcal{X} \times \mathcal{Y} \), where \( \mathcal{X} \) and \( \mathcal{Y} \) are convex sets and \( \mathcal{X} \) is compact. Suppose that \( f(\cdot, y) \) is convex and continuous for each fixed \( y \in \mathcal{Y} \) and \( f(x, \cdot) \) is concave for each fixed \( x \in \mathcal{X} \). Then

\[
\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y) = \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y).
\]

**Proof.** For any function \( f \), one obviously has

\[
\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y) \geq \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y).
\]

(To see this, just note that for all \( x' \in \mathcal{X} \) and \( y \in \mathcal{Y} \), \( f(x', y) \geq \inf_x f(x, y) \), so that for all \( x' \in \mathcal{X} \), \( \sup_y f(x', y) \geq \sup_y \inf_x f(x, y) \), which implies the statement.)

To prove the reverse inequality, without loss of generality, we may assume that \( f(x, y) \in [0, 1] \) for each \( (x, y) \in \mathcal{X} \times \mathcal{Y} \). Fix a small \( \varepsilon > 0 \) and a large positive integer \( n \). By the compactness of \( \mathcal{X} \) one may find a finite set of points \( \{x^{(1)}, \ldots, x^{(N)}\} \subset \mathcal{X} \) such that each
$x \in \mathcal{X}$ is within distance $\varepsilon$ of at least one of the $x^{(i)}$. We define the sequences $x_1, \ldots, x_n \in \mathcal{X}$ and $y_1, \ldots, y_n \in \mathcal{Y}$ recursively as follows. $y_0$ is chosen arbitrarily. For each $t = 1, \ldots, n$ let

$$x_t = \frac{\sum_{i=1}^{N} x^{(i)} e^{-\eta \sum_{s=0}^{t-1} f(x^{(s)}, y_s)}}{\sum_{j=1}^{N} e^{-\eta \sum_{s=0}^{t-1} f(x^{(s)}, y_s)},}$$

where $\eta = \sqrt{8 \ln N / n}$ and $y_t$ is such that $f(x_t, y_t) \geq \sup_{y \in \mathcal{Y}} f(x_t, y) - 1/n$. Then, by the convexity of $f$ in its first argument, we obtain (by Theorem 2.2) that

$$\frac{1}{n} \sum_{t=1}^{n} f(x_t, y_t) \leq \min_{i=1, \ldots, N} \frac{1}{n} \sum_{t=1}^{n} f(x^{(i)}, y_t) + \sqrt{\frac{\ln N}{2n}}. \quad (7.1)$$

Thus, we have

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y) \leq \sup_{y \in \mathcal{Y}} \left( \frac{1}{n} \sum_{t=1}^{n} x_t, y \right) \leq \sup_{y \in \mathcal{Y}} \frac{1}{n} \sum_{t=1}^{n} f(x_t, y) \quad \text{(by convexity of } f(\cdot, y))$$

$$\leq \frac{1}{n} \sum_{t=1}^{n} \sup_{y \in \mathcal{Y}} f(x_t, y) \leq \frac{1}{n} \sum_{t=1}^{n} f(x_t, y_t) + \frac{1}{n} \quad \text{(by definition of } y_t)$$

$$\leq \frac{1}{n} \sum_{t=1}^{n} f(x_t, y_t) + \frac{1}{n} \quad \text{(by inequality (7.1))}$$

$$\leq \min_{i=1, \ldots, N} \frac{1}{n} \sum_{t=1}^{n} f(x^{(i)}, y_t) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n} \quad \text{(by concavity of } f(x, \cdot))$$

$$\leq \min_{i=1, \ldots, N} f(x^{(i)}, y) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n}.$$

Thus, we have, for each $n$,

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y) \leq \sup_{y \in \mathcal{Y}} \min_{i=1, \ldots, N} f(x^{(i)}, y) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n},$$

so that, letting $n \to \infty$, we get

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y) \leq \sup_{y \in \mathcal{Y}} \min_{i=1, \ldots, N} f(x^{(i)}, y).$$

Letting $\varepsilon \to 0$ and using the continuity of $f$, we conclude that

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y) \leq \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y),$$

as desired. \[\square\]
Remark 7.1 (von Neumann’s minimax theorem). Observe that Theorem 7.1 implies von Neumann’s minimax theorem for two-person zero-sum games. To see this, just note that function $\ell$ is bounded and linear in both of its arguments and the simplex of all mixed strategies $p$ (and similarly $q$) is a compact set. In this special case the infima and suprema are achieved.

7.3 Repeated Two-Player Zero-Sum Games

We start our investigation of regret-based strategies for repeated game playing from the simple case of two-person zero-sum games. Recall that in our model, at each round $t$, based on the past plays of both players, the row player chooses an action $I_t \in \{1, \ldots, N\}$ according to the mixed strategy $p_t = (p_{1,t}, \ldots, p_{N,t})$ and the column player chooses an action $J_t \in \{1, \ldots, M\}$ according to the mixed strategy $q_t = (q_{1,t}, \ldots, q_{M,t})$. The distributions $p_t$ and $q_t$ may depend on the past plays of both. The row player’s loss at time $t$ is $\ell(I_t, J_t)$, and the column player’s loss is $-\ell(I_t, J_t)$. At each time instant, after making the play, the row player observes the losses $\ell(i, J_t)$ he would have suffered had he played strategy $i$, $i = 1, \ldots, N$.

In view of studying the convergence to equilibrium in such games, we consider the problem of minimizing the cumulative loss the row player. If the row player knew the column player’s actions $J_1, \ldots, J_n$ in advance, he would, at each time instant, choose his actions to satisfy $I_t = \arg\min_{i = 1, \ldots, N} \ell(i, J_t)$ invoking a total loss $\sum_{t=1}^{n} \min_{i = 1, \ldots, N} \ell(i, J_t)$. Achieving a cumulative loss close to this minimum without knowing the column player’s actions is, except for trivial cases, impossible (see Exercise 7.6), and so the row player has to put up with a less ambitious goal. A meaningful objective is to play almost as well as the best constant strategy. Thus, we consider the problem of minimizing the difference between the row player’s cumulative loss and the cumulative loss of the best constant strategy, that is,

$$\sum_{t=1}^{n} \ell(I_t, J_t) - \min_{i = 1, \ldots, N} \sum_{t=1}^{n} \ell(i, J_t).$$

By a simple application of regret-minimizing forecasters we show that simple strategies indeed exist such that this difference grows sublinearly (almost surely) no matter how the column player plays. This result is a simple consequence of the Hannan consistency results of Chapter 4.

It is natural to consider regret-minimizing strategies for both players. For example, we may assume that the players play according to Hannan consistent forecasting strategies. More precisely, assume that the row player chooses his actions $I_t$ such that, regardless of what the column player does,

$$\limsup_{n \to \infty} \left( \frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t) - \min_{i = 1, \ldots, N} \frac{1}{n} \sum_{t=1}^{n} \ell(i, J_t) \right) \leq 0 \quad \text{almost surely.}$$

Recall from Section 4 that several such Hannan-consistent procedures are available. For
example, this may be achieved by the exponentially weighted average mixed strategy

\[ p_{i,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell(i, J_s)\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \ell(k, J_s)\right)}, \quad i = 1, \ldots, N, \]

where \( \eta > 0 \). For this particular forecaster we have, with probability at least \( 1 - \delta \),

\[ \sum_{t=1}^{n} \ell(I_t, J_t) - \min_{i=1,\ldots,N} \sum_{t=1}^{n} \ell(i, J_t) \leq \frac{\ln N}{\eta} + \frac{n\eta}{8} + \sqrt{\frac{n}{2} \ln \frac{1}{\delta}} \]

(see Corollary 4.2).

**Remark 7.2 (Nonoblivious opponent).** It is important to point out that in the definition of regret, the cumulative loss \( \sum_{t=1}^{n} \ell(i, J_t) \) associated with the “constant” action \( i \) corresponds to the sequence of plays \( J_1, \ldots, J_n \) of the opponent. The plays of the opponent may depend on the forecaster’s actions, which, in this case, are \( I_1, \ldots, I_n \). Therefore, it is important to keep in mind that if the opponent is nonoblivious (recall the definition from Chapter 4), then \( \sum_{t=1}^{n} \ell(i, J_t) \) is not the same as the cumulative loss the forecaster would have suffered had he played action \( I_t = i \) for all \( t \). This issue is investigated in detail in Section 7.11.

**Remark 7.3 (Time-varying games).** The inequality above may be extended, in a straightforward way, to the case of time-varying games, that is, when the loss matrix is allowed to change with time as long as the entries \( \ell_t(i, j) \) stay uniformly bounded, say, between 0 and 1. The loss matrix \( \ell_t \) does not need to be known in advance by the row player. All we need to assume is that before making the play at time \( t \), the column \( \ell_{t-1}(. , J_{t-1}) \) of the loss matrix corresponding to the opponent’s play in the previous round is revealed to the row player. In a more difficult version of the problem, only the suffered loss \( \ell_{t-1}(I_{t-1}, J_{t-1}) \) is observed by the row player before time \( t \). These problems may be handled by the techniques of Section 6.7 (see also Section 7.5).

We now show the following remarkable fact: if the row player plays according to any Hannan consistent strategy, then his average loss cannot be much larger than the value of the game, regardless of the opponent’s strategy. (Note that by von Neumann’s minimax theorem this may also be achieved if the row player plays according to any minimax strategy; see Exercise 7.7.)

Recall that the value of the game characterized by the loss matrix \( \ell \) is defined by

\[ V = \max_{q} \min_{p} \overline{\ell}(p, q), \]

where the maximum is taken over all probability vectors \( q = (q_1, \ldots, q_M) \), the minimum is taken over all probability vectors \( p = (q_1, \ldots, p_N) \), and

\[ \overline{\ell}(p, q) = \sum_{i=1}^{N} \sum_{j=1}^{M} p_i q_j \ell(i, j). \]
7.3 Repeated Two-Player Zero-Sum Games

(Note that the maximum and the minimum are always achieved.) With some abuse of notation we also write

$$\bar{\ell}(p, j) = \sum_{i=1}^{N} p_i \ell(i, j) \quad \text{and} \quad \bar{\ell}(i, q) = \sum_{j=1}^{M} q_j \ell(i, j).$$

**Theorem 7.2.** Assume that in a two-person zero-sum game the row player plays according to a Hannan-consistent strategy. Then

$$\lim \sup_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t) \leq V \quad \text{almost surely.}$$

**Proof.** By the assumption of Hannan consistency, it suffices to show that

$$\min_{i=1, \ldots, N} \frac{1}{n} \sum_{t=1}^{n} \ell(i, J_t) \leq V.$$

This may be seen easily as follows. First,

$$\min_{i=1, \ldots, N} \frac{1}{n} \sum_{t=1}^{n} \ell(i, J_t) = \min_{p} \frac{1}{n} \sum_{t=1}^{n} \bar{\ell}(p, J_t)$$

since $\sum_{t=1}^{n} \bar{\ell}(p, J_t)$ is linear in $p$ and its minimum, over the simplex of probability vectors, is achieved in one of the corners. Then, letting $\hat{q}_{j,n} = \frac{1}{n} \sum_{t=1}^{n} 1_{[J_t=j]}$ be the empirical probability of the row player’s action being $j$,

$$\min_{p} \frac{1}{n} \sum_{t=1}^{n} \bar{\ell}(p, J_t) = \min_{p} \sum_{j=1}^{M} \hat{q}_{j,n} \bar{\ell}(p, j)$$

$$= \min_{p} \bar{\ell}(p, \hat{q}_n) \quad \text{(where $\hat{q}_n = (\hat{q}_{1,n}, \ldots, \hat{q}_{M,n})$)}$$

$$\leq \max_{q} \min_{p} \bar{\ell}(p, q) = V. \quad \blacksquare$$

Theorem 7.2 shows that, regardless of what the opponent plays, if the row player plays according to a Hannan-consistent strategy, then his cumulative loss is guaranteed to be asymptotically not more than the value $V$ of the game. It follows by symmetry that if both players use the same strategy, then the cumulative loss of the row player converges to $V$.

**Corollary 7.1.** Assume that in a two-person zero-sum game, both players play according to some Hannan consistent strategy. Then

$$\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t) = V \quad \text{almost surely.}$$

**Proof.** By Theorem 7.2

$$\lim \sup_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t) \leq V \quad \text{almost surely.}$$
The same theorem, applied to the column player, implies, using the fact that the column player’s loss that is the negative of the row player’s loss, that
\[
\liminf_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t) \geq \min_{\mathbf{p}} \max_{\mathbf{q}} \mathbb{E}(\mathbf{p}, \mathbf{q}) \quad \text{almost surely.}
\]

By von Neumann’s minimax theorem, the latter quantity equals \( V \).

**Remark 7.4 (Convergence to equilibria).** If both players follow some Hannan consistent strategy, then it is also easy to see that the product distribution \( \hat{\mathbf{p}}_n \times \hat{\mathbf{q}}_n \) formed by the (marginal) empirical distributions of play
\[
\hat{p}_{i,n} = \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}_{\{I_t = i\}} \quad \text{and} \quad \hat{q}_{j,n} = \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}_{\{J_t = j\}}
\]
of the two players converges, almost surely, to the set of Nash equilibria \( \pi = \mathbf{p} \times \mathbf{q} \) of the game (see Exercise 7.11). However, it is important to note that this does not mean that the players’ joint play is close to a Nash equilibrium in the long run. Indeed, one cannot conclude that the **joint empirical frequencies of play**
\[
\hat{P}_n(i, j) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}_{\{I_t = i, J_t = j\}}
\]
converge to the set of Nash equilibria. All one can say is that \( \hat{P}_n \) converges to the Hannan set of the game (defined later), which, even for zero-sum games, may include joint distributions that are not Nash equilibria (see Exercise 7.15).

### 7.4 Correlated Equilibrium and Internal Regret

In this section we consider repeated play of general \( K \)-person games. In the previous section we have shown the following fact: if both players of a two-person zero-sum game follow a Hannan-consistent strategy (i.e, play so that their external regret vanishes asymptotically), then in the long run equilibrium is achieved in the sense that the product of the marginal empirical frequencies of play converges to the set of Nash equilibria. It is natural to ask whether the joint empirical frequencies of play converge in any general \( K \)-person game. The answer is easily seen to be negative in general by the following argument. Assume that in a repeated \( K \)-person game each player follows a Hannan consistent strategy; that is, if the \( K \)-tuple of plays of all players at time \( t \) is \( \mathbf{I}_t = (I_t^{(1)}, \ldots, I_t^{(K)}) \), then for all \( k = 1, \ldots, K \) the cumulative loss of player \( k \) satisfies
\[
\limsup_{n \to \infty} \left( \frac{1}{n} \sum_{t=1}^{n} \ell^{(k)}(I_t) - \frac{1}{n} \min_{i_k=1,\ldots,N_k} \sum_{t=1}^{n} \ell^{(k)}(I_t^{(1)}, \ldots, i_k, \ldots, I_t^{(K)}) \right) \leq 0 \quad \text{almost surely.}
\]

Writing
\[
\hat{P}_n(i) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}_{\{I_t = i\}}, \quad i = (i_1, \ldots, i_K) \in \bigotimes_{k=1}^{K} \{1, \ldots, N_k\},
\]
for the empirical joint distribution of play, the property of Hannan consistency may be rewritten as follows: for all \( k = 1, \ldots, K \) and for all \( j = 1, \ldots, N_k \),
\[
\limsup_{n \to \infty} \left( \sum_i \hat{P}_n(i) \ell^{(k)}(i) - \sum_i \hat{P}_n(i) \ell^{(k)}(i^-, j) \right) \leq 0
\]
almost surely, where \((i^-, j) = (i_1, \ldots, j, \ldots, i_K)\) denotes the \( K \)-tuple obtained when the \( k \)th component \( i_k \) of \( i \) is replaced by \( j \). Writing the condition of Hannan consistency in this form reveals that if all players follow such a strategy, the empirical frequencies of play converge, almost surely, to the set \( \mathcal{H} \) of joint distributions \( P \) over \( \bigotimes_{k=1}^K \{1, \ldots, N_k\} \) defined by
\[
\mathcal{H} = \left\{ P : \forall k = 1, \ldots, K, \forall j = 1, \ldots, N_k, \sum_i P(i) \ell^{(k)}(i) \leq \sum_i P(i) \ell^{(k)}(i^-, j) \right\}
\]
(see Exercise 7.14). The set \( \mathcal{H} \) is called the Hannan set of the game. Since \( C \) is an intersection of \( \sum_{k=1}^K N_k \) halfspaces, it is a closed and convex subset of the simplex of all joint distributions. In other words, if all players play according to any strategy that asymptotically minimizes their external regret, the empirical frequencies of play converge to the set \( \mathcal{H} \). Unfortunately, distributions in the Hannan set do not correspond to any natural equilibrium concept. In fact, by comparing the definition of \( \mathcal{H} \) with the characterization of correlated equilibria given in Lemma 7.1, it is easy to see that \( \mathcal{H} \) always contains the set \( C \) of correlated equilibria. Even though for some special games (such as games in which all players have two actions; see Exercise 7.18) \( \mathcal{H} = C \), in typical cases \( C \) is a proper subset of \( \mathcal{H} \) (see Exercise 7.16). Thus, except for special cases, if players are merely required to play according to Hannan-consistent strategies (i.e., to minimize their external regret), there is no hope to achieve a correlated equilibrium, let alone a Nash equilibrium.

However, by requiring just a little bit more, convergence to correlated equilibria may be achieved. The main result of this section is that if each player plays according to a strategy minimizing internal regret as described in Section 4.4, the joint empirical frequencies of play converge to a correlated equilibrium.

More precisely, consider the model of playing repeated games described in Section 7.1. For each player \( k \) and pair of actions \( j, j' \in \{1, \ldots, N_K\} \) define the conditional instantaneous regret at time \( t \) by
\[
\hat{r}^{(k)}_{(j,j'),t} = \mathbb{I}_{(I_t^{(1)})=j} \left( \ell^{(k)}(I_t) - \ell^{(k)}(I_t^-, j') \right),
\]
where \((I_t^-, j') = (I_t^{(1)}, \ldots, I_t^{(k-1)}, j', I_t^{(k+1)}, \ldots, I_t^{(K)})\) is obtained by replacing the play of player \( k \) by action \( j' \). Note that the expected value of the conditional instantaneous regret \( \hat{r}^{(k)}_{(j,j'),t} \), calculated with respect to the distribution \( p_t^{(k)} \) of \( I_t^{(k)} \), is just the instantaneous internal regret of player \( k \) defined in Section 4.4. The conditional regret \( \sum_{t=1}^n \hat{r}^{(k)}_{(j,j'),t} \) expresses how much better player \( k \) could have done had he chosen action \( j' \) every time he played action \( j \).

The next lemma shows that if each player plays such that his conditional regret remains small, then the empirical distribution of plays will be close to a correlated equilibrium.

**Lemma 7.2.** Consider a \( K \)-person game and denote its set of correlated equilibria by \( C \). Assume that the game is played repeatedly so that for each player \( k = 1, \ldots, K \) and pair
of actions \( j, j' \in \{1, \ldots, N_k \} \) the conditional regrets satisfy
\[
\limsup_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \hat{r}_{(j,j'),t}^{(k)} \leq 0.
\]
Then the distance \( \inf_{P \in C} \sum_i |P(i) - \hat{P}_n(i)| \) between the empirical distribution of plays and the set of correlated equilibria converges to 0.

Proof. Observe that the assumption on the conditional regrets may be rewritten as
\[
\limsup_{n \to \infty} \sum_{i, i_i = j} \hat{P}_n(i) \left( \ell^{(k)}(i) - \ell^{(k)}(i^{-}, j') \right) \leq 0,
\]
where \( (i^{-}, j') \) is the same as \( i \) except that its \( k \)th component is replaced by \( j' \). Assume that the sequence \( \hat{P}_n \) does not converge to \( C \). Then by the compactness of the set of all probability distributions there is a distribution \( P^* \notin C \) and a subsequence \( \hat{P}_{n_k} \) of empirical distributions such that \( \lim_{k \to \infty} \hat{P}_{n_k} = P^* \). But since \( P^* \) is not a correlated equilibrium, by Lemma 7.1 there exists a player \( k \in \{1, \ldots, K \} \) and a pair of actions \( j, j' \in \{1, \ldots, N_k \} \) such that
\[
\sum_{i, i_i = j} P^*(i) \left( \ell^{(k)}(i) - \ell^{(k)}(i^{-}, j') \right) > 0,
\]
which contradicts the assumption. ■

Now it is easy to show that if all players play according to an internal-regret-minimizing strategy, such as that described in Section 4.4, a correlated equilibrium is reached asymptotically. More precisely, for each player \( k = 1, \ldots, K \) define the components of the internal instantaneous regret vector by
\[
r_{(j,j'),t}^{(k)} = p_{j,t}^{(k)} \left( \ell^{(k)}(I_t) - \ell^{(k)}(I_t^{-}, j') \right),
\]
where \( j, j' \in \{1, \ldots, N_k \} \) and \( (I_t^{-}, j') \) is as defined above. In Section 4.4 we saw that player \( k \) has a simple strategy \( p_t^{(k)} \) such that, regardless of the other players’ actions, it is guaranteed that the internal regret satisfies
\[
\max_{j, j'} \frac{1}{n} \sum_{t=1}^{n} r_{(j,j'),t}^{(k)} \leq c \sqrt{\ln N_k / n}
\]
for a universal constant \( c \). Now, clearly, \( r_{(j,j'),t}^{(k)} \) is the conditional expectation of \( \hat{r}_{(j,j'),t}^{(k)} \) given the past and the other players’ actions. Thus \( r_{(j,j'),t}^{(k)} - \hat{r}_{(j,j'),t}^{(k)} \) is a bounded martingale difference sequence for any fixed \( j, j' \). Therefore, by the Hoeffding–Azuma inequality (Lemma A.7) and the Borel–Cantelli lemma, we have, for each \( k \in \{1, \ldots, K \} \) and \( j, j' \in \{1, \ldots, N_k \} \),
\[
\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \left( r_{(j,j'),t}^{(k)} - \hat{r}_{(j,j'),t}^{(k)} \right) = 0 \quad \text{almost surely},
\]
which implies that for each \( k \)
\[
\limsup_{n \to \infty} \max_{j, j'} \frac{1}{n} \sum_{t=1}^{n} \hat{r}_{(j,j'),t}^{(k)} \leq 0 \quad \text{almost surely}.
\]

The following theorem summarizes what we have just proved.
7.4 Correlated Equilibrium and Internal Regret

Theorem 7.3. Consider a $K$-person game and denote its set of correlated equilibria by $C$. Assume that a game is played repeatedly such that each player $k = 1, \ldots, K$ plays according to an internal-regret-minimizing strategy (such as the ones described in Section 4.4). Then the distance $\inf_{P \in C} \sum_i |P(i) - \hat{P}_n(i)|$ between the empirical distribution of plays and the set of correlated equilibria converges to 0 almost surely.

Remark 7.5 (Existence of correlated equilibria). Observe that the theorem implicitly entails the existence of a correlated equilibrium of any game. Of course, this fact follows from the existence of Nash equilibria. However, unlike the proof of existence of Nash equilibria, the above argument avoids the use of fixed-point theorems.

Remark 7.6 (Computation of correlated equilibria). Given a $K$-person game, it is a complex and important problem to exhibit a computationally efficient procedure that finds a Nash equilibrium (or even better, the set of all Nash equilibria) of the game. As of today, no polynomial-time algorithm is known to approximate a Nash equilibrium. (The algorithm is required to be polynomial in the number of players and the number of actions of each player.) The difficulty of the problem may be understood by noting that even if every player in a $K$-person game has just two actions to choose from, there are $2^K$ possible action profiles $i = (i_1, \ldots, i_K)$, and so describing the payoff functions already takes exponential time. Interesting polynomial-time algorithms for computing Nash equilibria are available for important special classes of games that have a compact representation, such as symmetric games, graphical games, and so forth. We refer the interested reader to Papadimitriou [230, 231], Kearns and Mansour [179], Papadimitriou and Roughgarden [232]. Here we point out that the regret-based procedures described in this section may also be used to efficiently approximate correlated equilibria in a natural computational model. For any $\varepsilon > 0$, a probability distribution $P$ over the set of all $K$-tuples $i = (i_1, \ldots, i_K)$ of actions is called an $\varepsilon$-correlated equilibrium if, for every player $k \in \{1, \ldots, K\}$ and actions $j, j' \in \{1, \ldots, N_k\}$,

$$\sum_{i: i_k = j} P(i) (\ell(k)(i) - \ell(k)(i^-, j')) \leq \varepsilon.$$  

Clearly, $\varepsilon$-correlated equilibria approximate correlated equilibria in the sense that if $C_\varepsilon$ denotes the set of all $\varepsilon$-correlated equilibria, then $\bigcap_{\varepsilon > 0} C_\varepsilon = C$. Assume that an oracle is available that outputs the values of the loss functions $\ell(k)(i)$ for any action profile $i = (i_1, \ldots, i_K)$. Then it is easy to define an algorithm that calls the oracle polynomially many times and outputs an $\varepsilon$-correlated equilibrium. One may simply simulate as if all players played an internal-regret-minimizing procedure and calculate the joint distribution $\hat{P}_n$ of plays. By the results of Section 4.4, the Hoeffding–Azuma inequality and the union bound, with probability at least $1 - \delta$, for each $k = 1, \ldots, K$,

$$\sum_{i: i_k = j} \hat{P}_n(i) (\ell(k)(i) - \ell(k)(i^-, j')) \leq 2 \sqrt{\frac{\ln N_k}{n}} + \sqrt{\frac{\ln(K/\delta)}{2n}}.$$  

Thus, with probability at least $1 - \delta$, $\hat{P}_n$ is an $\varepsilon$-correlated equilibrium if

$$n \geq \max_{k=1, \ldots, K} \frac{16}{\varepsilon^2} \frac{N_k K}{\delta \ln N_k K}.$$
To compute all regrets necessary to run this algorithms, the oracle needs to be called not more than
\[
\max_{k=1,\ldots,K} \frac{16N_k K}{\varepsilon^2} \ln \frac{N_k K}{\delta}
\]
times to find an \( \varepsilon \)-correlated equilibrium, with probability at least \( 1 - \delta \). Thus, computation of an \( \varepsilon \)-correlated equilibrium is surprisingly fast: it takes time proportional to \( 1/\varepsilon^2 \) and is polynomial (in fact, barely superlinear) in terms of the number of actions and the number of players of the game.

### 7.5 Unknown Games: Game-Theoretic Bandits

In this section we mention the possibility of learning correlated equilibria in an even more restricted model than the uncoupled model studied so far in this chapter. We assume that the \( K \) players of a game play repeatedly, and at each time instance \( t \), after taking an action \( I_t^{(k)} \), the \( k \)th player observes his loss \( \ell_t^{(k)}(I_t) \) (where \( I_t = (I_t^{(1)}, \ldots, I_t^{(K)}) \) is the joint action profile played by the \( K \) players) but does not know his entire loss function \( \ell^{(k)} \) and cannot observe the other players’ actions \( I_t^{-} \). In fact, player \( k \) may not even know the number of players participating in the game, let alone the number of actions the other players can choose from.

It may come as a surprise that, with such limited information, the players have a way of playing that guarantees that the game reaches an equilibrium in some sense. Here we point out that there is a strategy such that if all players play according to it, then the joint empirical frequencies of play converge to a correlated equilibrium. The issue of convergence to a Nash equilibrium is addressed in Section 7.10.

The fact that convergence of the empirical frequencies of play to a correlated equilibrium may be achieved follows directly from Theorem 7.3 and the fact that each player can guarantee a small internal regret even if he cannot observe the actions of the other players. Indeed, in order to achieve the desired convergence, each player’s job now is to minimize his internal regret in the bandit problem described in Chapter 6. Exercise 6.15 shows that such internal-regret minimizing procedures exist and therefore, if all players follow such a strategy, the empirical frequencies of play will converge to the set of correlated equilibria of the game.

### 7.6 Calibration and Correlated Equilibrium

In Section 4.5 we described the connection between calibrated and internal-regret-minimizing forecasting strategies. According to the results of Section 7.4, internal-regret-minimizing forecasters may be used, in a straightforward way, to achieve correlated equilibrium in a certain sense. In the present section we close the circle and point out an interesting connection between calibration and correlated equilibria; that is, we show that if each player bases his decision on a calibrated forecaster in an appropriate way (where the calibrated forecasters used by the different players may be completely different), then the joint empirical frequencies of play converge to the set of correlated equilibria.
7.6 Calibration and Correlated Equilibrium

The strategy we assume the players follow is quite natural: on the basis of past experience, each player predicts the mixed strategy his opponents will use in the next round and selects an action that would minimize his loss if the opponents indeed played according to the predicted distribution. More precisely, each player forecasts the joint probability of each possible outcome of the next action of his opponents and then chooses a “best response” to the forecasted distribution. We saw in Section 4.5 that no matter how the opponents play, each player can construct a (randomized) well-calibrated forecast of the opponents’ sequence of play. The main result of this section shows that the mild requirement that all players use well calibrated forecasters guarantees that the joint empirical frequencies of play converge to the set $C$ of correlated equilibria.

To lighten the exposition, we present the results for $K = 2$ players, but the results trivially extend to the general case (left as exercise). So assume that the actions $I_t$ and $J_t$ selected by the two players at time $t$ are determined as follows. Depending on the past sequence of plays, for each $j = 1, \ldots, M$ the row player determines a probability forecast $\hat{q}_{j,t} \in [0, 1]$ of the next play $J_t$ of the column player where we require that $\sum_{j=1}^M \hat{q}_{j,t} = 1$. Denote the forecasted mixed strategy by $\hat{q}_t = (\hat{q}_{1,t}, \ldots, \hat{q}_{M,t})$ and write $J_t = (\mathbb{I}_{\{J_t = 1\}}, \ldots, \mathbb{I}_{\{J_t = M\}})$. We only require that the forecast be well calibrated. Recall from Section 4.5 that this means that for any $\varepsilon > 0$, the function

$$\rho_n(A) = \frac{\sum_{t=1}^n J_t \mathbb{I}_{\{\hat{q}_t \in A\}}}{\sum_{t=1}^n \mathbb{I}_{\{\hat{q}_t \in A\}}}$$

defined for all subsets $A$ of the probability simplex (defined to be 0 if $\sum_{t=1}^n \mathbb{I}_{\{\hat{q}_t \in A\}} = 0$) satisfies, for all $A$ and for all $\varepsilon > 0$,

$$\limsup_{n \to \infty} \left| \rho_n(A_\varepsilon) - \frac{\int_{A_\varepsilon} x \, dx}{\lambda(A_\varepsilon)} \right| < \varepsilon,$$

(where $A_\varepsilon = \{x : \exists y \in A \text{ such that } \|x - y\| < \varepsilon\}$ is the $\varepsilon$-blowup of $A$ and $\lambda$ stands for the uniform probability measure over the simplex.) Recall from Section 4.5 (more precisely Exercise 4.17) that regardless of what the sequence $J_1, J_2, \ldots$ is, the row player can determine a randomized, well-calibrated forecaster. We assume that the row player best responds to the forecasted probability distribution in the sense that

$$I_t = \arg\min_{i = 1, \ldots, N} \ell^{(1)}(i, \hat{q}_t) = \arg\min_{i = 1, \ldots, N} \sum_{j=1}^M \hat{q}_{j,t} \ell^{(1)}(i, j).$$

Ties may be broken by an arbitrary but constant rule. The tie-breaking rule is described by the sets

$$\hat{B}_i = \{q : \text{row player plays action } i \text{ if } \hat{q}_t = q\}, \quad i = 1, \ldots, N.$$  

Note that for each $i$, $\hat{B}_i$ is contained in the closed and convex set $B_i$ of $q$’s to which $i$ is a best reply defined by

$$B_i = \left\{ q : \ell^{(1)}(i, q) = \min_{i' = 1, \ldots, N} \ell^{(1)}(i', q) \right\}.$$
Prediction and Playing Games

Assume also that the column player proceeds similarly, that is,

\[ J_t = \arg\min_{j=1,\ldots,M} \ell^{(2)}(\hat{p}_t, j) = \arg\min_{j=1,\ldots,M} \sum_{t=1}^N \hat{p}_{i,t} \ell^{(2)}(i, j), \]

where for each \( i \), the sequence \( \hat{p}_{i,t} (t = 1, 2, \ldots) \) is a well calibrated forecaster of \( I_1, I_2, \ldots \).

**Theorem 7.4.** Assume that in a two-person game both players play by best responding to a calibrated forecast of the opponent’s sequence of plays, as described above. Then the joint empirical frequencies of play

\[ \hat{P}_n(i, j) = \frac{1}{n} \sum_{t=1}^n I_{\{I_t=i, J_t=j\}} \]

converge to the set \( C \) of correlated equilibria in the sense that

\[ \inf_{P \in C} \sum_{(i,j)} |P(i,j) - \hat{P}_n(i,j)| \to 0 \quad \text{almost surely.} \]

**Proof.** Consider the sequence of empirical distributions \( \hat{P}_n \). Since the simplex of all joint distributions over \([1, \ldots, N] \times [1, \ldots, M] \) is a compact set, every sequence has a convergent subsequence. Thus, it suffices to show that the limit of every convergent subsequence of \( \hat{P}_n \) is in \( C \). Let \( \hat{P}_{n_k} \) be such a convergent subsequence and denote its limit by \( P \). We need to show that \( P \) is a correlated equilibrium.

Note that by Lemma 7.1, \( P \) is a correlated equilibrium if and only if for each \( i \in \{1, \ldots, N\} \) the conditional distribution

\[ \hat{q}(. \mid i) = \left( q(1 \mid i), \ldots, q(M \mid i) \right) = \left( \frac{P(i, 1)}{\sum_{j'=1}^M P(i, j')}, \ldots, \frac{P(i, M)}{\sum_{j'=1}^M P(i, j')} \right) \]

is in the set \( B_i \) and the symmetric statement for the other player holds as well. Therefore, it suffices to show that, for each \( i \in \{1, \ldots, N\} \), the distribution

\[ \hat{q}_{n_k}(\cdot \mid i) = \left( \frac{\hat{P}_{n_k}(i, 1)}{\sum_{j'=1}^M \hat{P}_{n_k}(i, j')}, \ldots, \frac{\hat{P}_{n_k}(i, M)}{\sum_{j'=1}^M \hat{P}_{n_k}(i, j')} \right) \]

approaches the set \( B_i \). By the definition of \( I_t \), for each \( j = 1, \ldots, M \)

\[ \hat{P}_{n_k}(i, j) = \frac{1}{n_k} \sum_{s=1}^{k} \mathbb{I}_{\{q_{n_k} \in \hat{B}_i\}} \mathbb{I}_{\{J_{n_k}=j\}} \]

and therefore

\[ \hat{q}_{n_k}(\cdot \mid i) = \frac{\sum_{s=1}^{k} \mathbb{I}_{\{q_{n_k} \in \hat{B}_i\}}}{\sum_{s=1}^{k} \mathbb{I}_{\{q_{n_k} \in \hat{B}_i\}}} = \rho_{n_k}(\hat{B}_i). \]

Since \( \hat{P}_{n_k} \) is convergent, the limit

\[ \bar{x} = \lim_{k \to \infty} \rho_{n_k}(\hat{B}_i) \]
exists. We need to show that \( x \in B_i \). If \( (\hat{B}_i)_{\varepsilon} \) denotes the \( \varepsilon \)-blowup of \( \hat{B}_i \) then well-calibration of the forecaster \( \hat{q}_t \) implies that, for all \( \varepsilon > 0 \),

\[
\limsup_{k \to \infty} \left| \rho_{n_k}((\hat{B}_i)_{\varepsilon}) - \frac{\int_{(\hat{B}_i)_{\varepsilon}} x \, dx}{\lambda((\hat{B}_i)_{\varepsilon})} \right| < \varepsilon.
\]

Define

\[
x' \overset{\text{def}}{=} \lim_{\varepsilon \to 0} \int_{(\hat{B}_i)_{\varepsilon}} x \, dx / \lambda((\hat{B}_i)_{\varepsilon}).
\]

(The limit exists by the continuity of measure.) Because for all \( n_k \) we have \( \lim_{\varepsilon \to 0} \rho_{n_k}((\hat{B}_i)_{\varepsilon}) = \rho_{n_k}(\hat{B}_i) \), it is easy to see that \( x = x' \). But since \( \hat{B}_i \subset B_i \) and \( B_i \) is convex, the vector \( \int_{(\hat{B}_i)_{\varepsilon}} x \, dx / \lambda((\hat{B}_i)_{\varepsilon}) \) lies in the \( \varepsilon \)-blowup of \( B_i \). Therefore we indeed have \( x = x' \in B_i \), as desired.

### 7.7 Blackwell’s Approachability Theorem

We investigate a powerful generalization, introduced by Blackwell, of the problem of playing repeated two-player zero-sum games. Consider the situation of Section 7.3 with the only but essential difference that losses are vector valued. More precisely, the setup is described as follows. Just as before, at each time instance the row player selects an action \( i \in \{1, \ldots, N\} \) and the column player selects an action \( j \in \{1, \ldots, M\} \). However, the “loss” \( \ell(i, j) \) suffered by the row player is not a real number in [0, 1] but may take values in a bounded subset of \( \mathbb{R}^m \). (We use bold characters to emphasize that losses are vector valued.)

For the sake of concreteness we assume that all losses are in the euclidean unit ball, that is, the entries of the loss matrix \( \ell \) are such that \( \|\ell(i, j)\| \leq 1 \).

In the simpler case of scalar losses, the purpose of the row player is to minimize his average loss regardless of the actions of the column player. Then von Neumann’s minimax theorem (together with martingale convergence, e.g., the Hoeffding–Azuma inequality) asserts that no matter how the column player plays, the row player can always keep the normalized accumulated loss \( \frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t) \) in, or very close to, the set \( (-\infty, V] \), but not in the set \( (-\infty, V - \varepsilon] \) if \( \varepsilon > 0 \) (see Exercise 7.7). In the case of vector-valued losses the general question is to determine which subsets of \( \mathbb{R}^m \) can the row player keep his average loss close to. To this end, following Blackwell [28], we introduce the notion of approachability: a subset \( S \) of the unit ball of \( \mathbb{R}^m \) is approachable (by the row player) if the row player has a (randomized) strategy such that no matter how the column player plays,

\[
\lim_{n \to \infty} d \left( \frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t), S \right) = 0 \quad \text{almost surely},
\]

where \( d(u, S) = \inf_{v \in S} \|u - v\| \) denotes the euclidean distance of \( u \) from the set \( S \). Because any set is approachable if and only if its closure is approachable, it suffices to consider only closed sets.

Our purpose in this section is to characterize which convex sets are approachable. In the one-dimensional case the minimax theorem can be rephrased as follows: a closed interval
\((-\infty, c]\) is approachable if and only if \(c \geq V\). In other words, \((-\infty, c]\) is approachable if and only if the row player has a mixed strategy \(p\) such that \(\max_{j=1,\ldots,M} \ell(p, j) \leq c\).

In the general vector-valued case it is easy to characterize approachability of halfspaces. Consider a halfspace defined by \(H = \{u : a \cdot u \leq c\}\), where \(|a| = 1\). If we define an auxiliary game with scalar losses \(\ell(i, j) = a \cdot \ell(i, j)\), then clearly, \(H\) is approachable if and only if the set \((-\infty, c]\) is approachable in the auxiliary game. By the above-mentioned consequence of the minimax theorem, this happens if and only if \(\max_{j=1,\ldots,M} \ell(p, j) = \max_{j=1,\ldots,M} a \cdot \ell(p, j) \leq c\). (Note that, just as earlier, \(\ell(p, j) = \sum_j p_j \ell(i, j)\).) Thus, we have proved the following characterization of the approachability of closed halfspaces.

**Lemma 7.3.** A halfspace \(H = \{u : a \cdot u \leq c\}\) is approachable if and only if there exists a probability vector \(p = (p_1, \ldots, p_N)\) such that

\[
\max_{j=1,\ldots,M} a \cdot \ell(p, j) \leq c.
\]

The lemma states that the halfspace \(H\) is approachable by the row player in a repeated play of the game if and only if in the one-shot game the row player has a mixed strategy that keeps the expected loss in the halfspace \(H\). This is a simple and natural fact. What is interesting and much less obvious is that to approach any convex set \(S\), it suffices that the row player has a strategy for each hyperplane not intersecting \(S\) to keep the average loss on the same side of the hyperplane as the set \(S\). This is Blackwell’s Approachability Theorem, stated and proved next.

**Theorem 7.5 (Blackwell’s approachability theorem).** A closed convex set \(S\) is approachable if and only if every halfspace \(H\) containing \(S\) is approachable.

It is important to point out that for approachability of \(S\), every halfspace containing \(S\) must be approachable. Even if \(S\) can be written as an intersection of finitely many closed halfspaces, that is, \(S = \bigcap_i H_i\), the fact that all \(H_i\) are approachable does not imply that \(S\) is approachable: see Exercise 7.21.

The proof of Theorem 7.5 is constructive and surprisingly simple. At every time instance, if the average loss is not in the set \(S\), then the row player projects the average loss to \(S\) and uses the mixed strategy \(p\) corresponding to the halfspace containing \(S\), defined by the hyperplane passing through the projected loss vector, and perpendicular to the direction of the projection. In the next section we generalize this “approaching” algorithm to obtain a whole family of strategies that guarantee that the average loss approaches \(S\) almost surely.

Introduce the notation \(A_t = \frac{1}{t} \sum_{i=1}^t \ell(i, J_s)\) for the average loss vector at time \(t\) and denote by \(\pi_S(u) = \arg\min_{v \in S} \|u - v\|\) the projection of \(u \in \mathbb{R}^m\) onto \(S\). (Note that \(\pi_S(u)\) exists and is unique if \(S\) is closed and convex.)

**Proof.** To prove the statement, note first that \(S\) is clearly not approachable if there exists a halfspace \(H \supset S\) that is not approachable.

Thus, it remains to show that if all halfspaces \(H\) containing \(S\) are approachable, then \(S\) is approachable as well. To this end, assume that the row player’s mixed strategy \(p_t\) at time \(t = 1, 2, \ldots\) is arbitrary if \(A_{t-1} \in S\), and otherwise it is such that

\[
\max_{j=1,\ldots,M} a_{t-1} \cdot \ell(p_t, j) \leq c_{t-1},
\]
7.7 Blackwell's Approachability Theorem

Figure 7.1. Approachability: the expected loss $\ell(p_t, J_t)$ is forced to stay on the same side of the hyperplane \( \{ u : a_{t-1} \cdot u = c_{t-1} \} \) as the target set \( S \), opposite to the current average loss \( A_{t-1} \).

where

\[
a_{t-1} = \frac{A_{t-1} - \pi_S(A_{t-1})}{\| A_{t-1} - \pi_S(A_{t-1}) \|} \quad \text{and} \quad c_{t-1} = a_{t-1} \cdot \pi_S(A_{t-1}).
\]

(Define \( A_0 \) as the zero vector.) Observe that the hyperplane \( \{ u : a_{t-1} \cdot u = c_{t-1} \} \) contains \( \pi_S(A_{t-1}) \) and is perpendicular to the direction of projection of the average loss \( A_{t-1} \) to \( S \). Since the halfspace \( \{ u : a_{t-1} \cdot u \leq c_{t-1} \} \) contains \( S \), such a strategy \( p_t \) exists by assumption (see Figure 7.1). The defining inequality of \( p_t \) may be rewritten as

\[
\max_{j=1, \ldots, M} a_{t-1} \cdot (\ell(p_t, j) - \pi_S(A_{t-1})) \leq 0.
\]

Since \( A_t = \frac{t-1}{t} A_{t-1} + \frac{1}{t} \ell(I_t, J_t) \), we may write

\[
d(A_t, S)^2 = \| A_t - \pi_S(A_t) \|^2 \\
\leq \| A_t - \pi_S(A_{t-1}) \|^2 \\
= \left\| \frac{t-1}{t} A_{t-1} + \frac{\ell(I_t, J_t)}{t} - \pi_S(A_{t-1}) \right\|^2 \\
= \left( \frac{t-1}{t} \right)^2 \| A_{t-1} - \pi_S(A_{t-1}) \|^2 + \frac{1}{t^2} \| \ell(I_t, J_t) - \pi_S(A_{t-1}) \|^2 \\
+ 2 \left( \frac{t-1}{t} \right) \left( A_{t-1} - \pi_S(A_{t-1}) \right) \cdot \left( \ell(I_t, J_t) - \pi_S(A_{t-1}) \right).
\]

Using the assumption that all losses, as well as the set \( S \), are in the unit ball, and therefore \( \| \ell(I_t, J_t) - \pi_S(A_{t-1}) \| \leq 2 \), and rearranging the obtained inequality, we get

\[
t^2 \| A_t - \pi_S(A_t) \|^2 - (t-1)^2 \| A_{t-1} - \pi_S(A_{t-1}) \|^2 \\
\leq 4 + 2(t-1)(A_{t-1} - \pi_S(A_{t-1})) \cdot (\ell(I_t, J_t) - \pi_S(A_{t-1})).
\]
Summing both sides of the inequality for $t = 1, \ldots, n$, the left-hand side telescopes and becomes $n^2 \| A_n - \pi_S(A_n) \|^2$. Next, dividing both sides by $n^2$, and writing $K_{t-1} = \frac{1}{n} \| A_{t-1} - \pi_S(A_{t-1}) \|$, we have
\[
\| A_n - \pi_S(A_n) \|^2 \\
\leq \frac{4}{n} + \frac{2}{n} \sum_{t=1}^{n} K_{t-1} a_{t-1} \cdot (\ell(I_t, J_t) - \pi_S(A_{t-1})) \\
\leq \frac{4}{n} + \frac{2}{n} \sum_{t=1}^{n} K_{t-1} a_{t-1} \cdot (\ell(I_t, J_t) - \ell(p_t, J_t)),
\]
where at the last step we used the defining property of $p_t$. Since the random variable $K_{t-1}$ is bounded between 0 and 2, the second term on the right-hand side is an average of bounded zero-mean martingale differences, and therefore the Hoeffding–Azuma inequality (together with the Borel–Cantelli lemma) immediately implies that $\| A_n - \pi_S(A_n) \|^2 \to 0$ almost surely, which is precisely what we wanted to prove.

**Remark 7.7 (Rates of convergence).** The rate of convergence to 0 of the distance $\| A_n - \pi_S(A_n) \|$ that one immediately obtains from the proof is of order $n^{-1/4}$ (since the upper bound for the squared distance contains a sum of bounded martingale differences that, if bounded by the Hoeffding–Azuma inequality, gives a term that is $O_p(n^{-1/2})$). However, this bound can be improved substantially by a simple modification of the definition of the mixed strategy $p_t$ used in the proof. Indeed, if instead of the average loss vector $A_t = \frac{1}{t} \sum_{s=1}^{t} \ell(I_s, J_s)$ one uses the “expected” average loss $\overline{A}_t = \frac{1}{t} \sum_{s=1}^{t} \ell(p_s, J_s)$, one easily obtains a bound for $\| A_n - \pi_S(A_n) \|$ that is of order $n^{-1/2}$; see Exercise 7.23.

**Approachability and Regret Minimization**

To demonstrate the power of Theorem 7.5 we show how it can be used to show the existence of Hannan consistent forecaster. Recall the problem of randomized prediction described in Section 4.1. In this case the goal of forecaster is to determine, at each round of play, a distribution $p$ (i.e., a mixed strategy in the terminology of this chapter) so that regardless of what the outcomes $Y_t$ (the opponents’ play) are, the per-round regret
\[
\frac{1}{n} \left( \sum_{i=1}^{n} \ell(I_t, Y_t) - \min_{i=1, \ldots, N} \sum_{i=1}^{n} \ell(i, Y_t) \right)
\]
has a nonpositive limsup when each $I_t$ is drawn randomly according to the distribution $p_t$. Equivalently, the forecaster tries to keep the per-round regret vector $\frac{1}{n} R_{i,n} = \frac{1}{n} \sum_{t=1}^{n} (\ell(I_t, Y_t) - \ell(i, Y_t))$, close to the nonpositive orthant
\[
S = \{ u = (u_1, \ldots, u_N) : \forall i = 1, \ldots, N, u_i \leq 0 \}.
\]
Thus, the existence of a Hannan consistent forecaster is equivalent to the approachability of the orthant $S$ in a two-player game in which the vector-valued losses of the row player are defined by $\ell(i, j)$ whose $k$th component is
\[
\ell^{(k)}(i, j) = \ell(i, j) - \ell(k, j).
\]
(Not that with this choice the loss vectors fall in the ball, centered at the origin, of radius $2\sqrt{N}$. In the formulation of Theorem 7.5 we assumed that the loss vectors take their
The normal vector $\mathbf{a}$ of any hyperplane corresponding to halfspaces containing the nonpositive orthant has nonnegative components.

By Theorem 7.5 and Lemma 7.3, $S$ is approachable if for every halfspace $\{ \mathbf{u} : \mathbf{a} \cdot \mathbf{u} \leq c \}$ containing $S$ there exists a probability vector $\mathbf{p} = (p_1, \ldots, p_N)$ such that

$$\max_{j=1,\ldots,M} \mathbf{a} \cdot \overline{\ell}(\mathbf{p}, j) \leq c.$$  

Clearly, it suffices to consider halfspaces of the form

$$\max_{j=1,\ldots,M} \mathbf{a} \cdot \overline{\ell}(\mathbf{p}, j) \leq 0,$$

where the normal vector $\mathbf{a} = (a_1, \ldots, a_N)$ is such that all its components are nonnegative (see Figure 7.2). This condition is equivalent to requiring that, for all $j = 1, \ldots, M$,

$$\sum_{k=1}^{N} a_k \ell^{(k)}(\mathbf{p}, j) = \sum_{k=1}^{N} a_k (\overline{\ell}(\mathbf{p}, j) - \ell(k, j)) \leq 0,$$

Choosing

$$\mathbf{p} = \frac{\mathbf{a}}{\sum_{k=1}^{N} a_k}$$

the inequality clearly holds for all $j$ (with equality). Since $\mathbf{a}$ has only nonnegative components, $\mathbf{p}$ is indeed a valid mixed strategy. In summary, Blackwell’s approachability theorem indeed implies the existence of a Hannan consistent forecasting strategy. Note that the constructive proof of Theorem 7.5 defines such a strategy. Observe that this strategy is just the weighted average forecaster of Section 4.2 when the quadratic potential is used. In the next section we describe a whole family of strategies that, when specialized to the forecasting problem discussed here, reduce to weighted average forecasters with more general potential functions. The approachability theorem may be (and has been) used to handle more general problems. For example, it is easy to see that it implies the existence of internal-regret-minimizing strategies (see Exercise 7.22).
7.8 Potential-based Approachability

The constructive proof of Blackwell’s approachability theorem (Theorem 7.5) shows that if \( S \) is an approachable convex set, then for any strategy of the row player such that

\[
\max_{j=1,\ldots,M} (A_{t-1} - \pi_S(A_{t-1})) \cdot \ell(p_t, j) \leq \sup_{x \in S} (A_{t-1} - \pi_S(A_{t-1})) \cdot x
\]

the average loss \( A_n \) converges to the set \( S \) almost surely. The purpose of this section is to introduce, as proposed by Hart and Mas-Colell [146], a whole class of strategies for the row player that achieve the same goal. These strategies are generalizations of the strategy appearing in the proof of Theorem 7.5.

The setup is the same as in the previous section, that is, we consider vector-valued losses in the unit ball, and the row player’s goal is to guarantee that, no matter what the opponent does, the average loss \( A_n = \frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t) \) converges to a set \( S \) almost surely. Theorem 7.5 shows that if \( S \) is convex, this is possible if and only if all linear halfspaces containing \( S \) are approachable. Throughout this section we assume that \( S \) is closed, convex, and approachable, and we define strategies for the row player guaranteeing that \( d(A_n, S) \to 0 \) with probability 1.

To define such a strategy, we introduce a nonnegative potential function \( \Phi : \mathbb{R}^m \to \mathbb{R} \) whose role is to score the current situation of the average loss. A small value of \( \Phi(A_t) \) means that \( A_t \) is “close” to the target set \( S \). All we assume about \( \Phi \) is that it is convex, differentiable for all \( x \not\in S \), and \( \Phi(x) = 0 \) if and only if \( x \in S \). Note that such a potential always exists by convexity and closedness of \( S \). One example is the function \( \Phi(x) = \inf_{y \in S} \|x - y\|^2 \).

The row player uses the potential function to determine, at time \( t \), a mixed strategy \( p_t \) satisfying, whenever \( A_{t-1} \in S \),

\[
\max_{j=1,\ldots,M} a_{t-1} \cdot \ell(p_t, j) \leq c_{t-1},
\]

where

\[
a_{t-1} = \frac{\nabla \Phi(A_{t-1})}{\|\nabla \Phi(A_{t-1})\|} \quad \text{and} \quad c_{t-1} = \sup_{x \in S} a_{t-1} \cdot x.
\]

If \( A_{t-1} \in S \), the row player’s action can be arbitrary. Observe that the existence of such a \( p_t \) is implied by the fact that \( S \) is convex and approachable and by Theorem 7.5. Geometrically, the hyperplane tangent to the level curve of the potential function \( \Phi \) passing through \( A_{t-1} \) is shifted so that it intersects \( S \) but \( S \) falls entirely on one side of the shifted hyperplane. The distribution \( p_t \) is determined so that the expected loss \( \ell(p_t, j) \) is forced to stay on the same side of the hyperplane as \( S \) (see Figure 7.3). Note also that, in the special case when \( \Phi(x) = \inf_{y \in S} \|x - y\|^2 \), this strategy is identical to Blackwell’s strategy defined in the proof of Theorem 7.5, and therefore the potential-based strategy may be thought of as a generalization of Blackwell’s strategy.

**Remark 7.8 (Bregman projection).** For any \( x \not\in S \) we may define

\[
\pi_S(x) = \arg\max_{y \in S} \nabla \Phi(x) \cdot y.
\]
7.8 Potential-based Approachability

\[ \{ x : \Phi(x) = \text{const.} \} \]

**Figure 7.3.** Potential-based approachability. The hyperplane \( \{ u : a_{t-1} \cdot u = c_{t-1} \} \) is determined by shifting the hyperplane tangent to the level set \( \{ x : \Phi(x) = \text{const.} \} \) containing \( A_{t-1} \) to the Bregman projection \( \pi_S(A_{t-1}) \). The expected loss \( \bar{\ell}(p_s, J_s) \) is guaranteed to stay on the same side of the shifted hyperplane as \( S \).

It is easy to see that the conditions on \( \Phi \) imply that the maximum exists and is unique. In fact, since \( \Phi(y) = 0 \) for all \( y \in S \), \( \pi_S(x) \) may be rewritten as

\[
\pi_S(x) = \left( \arg\min_{y \in S} \Phi(y) - \Phi(x) - \nabla \Phi(x) \cdot (y - x) \right) = \arg\min_{y \in S} D_{\Phi}(y, x) .
\]

In other words, \( \pi_S(x) \) is just the projection, under the Bregman divergence, of \( x \) onto the set \( S \). (For the definition and basic properties of Bregman divergences and projections, see Section 11.2.)

This potential-based strategy has a very similar version with comparable properties. This version simply replaces \( A_{t-1} \) in the definition of the algorithm by \( \overline{A}_{t-1} \) where, for each \( t = 1, 2, \ldots, \overline{A}_t = \frac{1}{t} \sum_{s=1}^{t} \bar{\ell}(p_s, J_s) \) is the “expected” average loss, see also Exercise 7.23. Recall that

\[
\bar{\ell}(p_s, J_s) = \sum_{i=1}^{N} p_{i,s} \ell(i, J_s) = \mathbb{E} \left[ \ell(I_s, J_s) \bigg| I_s^{t-1}, J^s \right] .
\]

The main result of this section states that for any potential function, the row player’s average loss approaches the target set \( S \) no matter how the opponent plays. The next theorem states this fact for the modified strategy (based on \( \overline{A}_{t-1} \)). The analog statement for the strategy based on \( A_{t-1} \) may be proved similarly, and it is left as an exercise.
**Theorem 7.6.** Let $S$ be a closed, convex, and approachable subset of the unit ball in $\mathbb{R}^n$, and let $\Phi$ be a convex, twice differentiable potential function that vanishes in $S$ and is positive outside of $S$. Denote the Hessian matrix of $\Phi$ at $x \in \mathbb{R}^n$ by $H(x)$ and let $B = \sup_{x: \|x\| \leq 1} \|H(x)\| < \infty$ be the maximal norm of the Hessian in the unit ball. For each $t = 1, 2, \ldots$ let the potential-based strategy $p_t$ be any strategy satisfying

$$
\max_{j=1, \ldots, M} a_{t-1} \cdot \bar{\ell}(p_t, j) \leq c_{t-1} \quad \text{if } A_{t-1} \notin S \text{ if } \overline{A}_{t-1} \notin S,
$$

where

$$
a_{t-1} = \frac{\nabla \Phi(\overline{A}_{t-1})}{\|\nabla \Phi(\overline{A}_{t-1})\|} \quad \text{and} \quad c_{t-1} = \sup_{x \in S} a_{t-1} \cdot x = a_{t-1} \cdot \pi_S(\overline{A}_{t-1}).
$$

(If $\overline{A}_{t-1} \in S$, then the row player’s action can be arbitrary.) Then for any strategy of the column player, the average “expected” loss satisfies

$$
\Phi(\overline{A}_n) \leq \frac{2B(\ln n + 1)}{n}.
$$

Also, for the average loss $A_n = \frac{1}{n} \sum_{i=1}^n \ell(I_t, J_t)$, we have

$$
\lim_{n \to \infty} d(A_n, S) = 0 \quad \text{with probability 1}.
$$

**Proof.** First note that for any $x \notin S$ and $y \in S$,

$$
\nabla \Phi(x) \cdot (y - x) \leq -\Phi(x).
$$

This follows simply from the fact that by convexity of $\Phi$,

$$
0 = \Phi(y) \geq \Phi(x) + \nabla \Phi(x) \cdot (y - x).
$$

The rest of the proof is based on a simple Taylor expansion of the potential $\Phi(\overline{A}_t)$ around $\Phi(\overline{A}_{t-1})$. Since $\overline{A}_t = \overline{A}_{t-1} + \frac{1}{t} \left( \bar{\ell}(p_t, J_t) - \overline{A}_{t-1} \right)$, we have

$$
\Phi(\overline{A}_t) = \Phi(\overline{A}_{t-1}) + \frac{1}{t} \nabla \Phi(\overline{A}_{t-1}) \cdot \left( \bar{\ell}(p_t, J_t) - \overline{A}_{t-1} \right) + \frac{1}{2t^2} \left( \bar{\ell}(p_t, J_t) - \overline{A}_{t-1} \right)^\top H(\xi) \left( \bar{\ell}(p_t, J_t) - \overline{A}_{t-1} \right)
$$

(where $\xi$ is a vector between $\overline{A}_t$ and $\overline{A}_{t-1}$)

$$
\leq \Phi(\overline{A}_{t-1}) + \frac{1}{t} \nabla \Phi(\overline{A}_{t-1}) \cdot \left( \bar{\ell}(p_t, J_t) - \overline{A}_{t-1} \right) + \frac{1}{2t^2} \left\| \bar{\ell}(p_t, J_t) - \overline{A}_{t-1} \right\|^2 \sup_{\xi: \|\xi\| \leq 1} \|H(\xi)\|
$$

(by the Cauchy–Schwarz inequality)

$$
\leq \Phi(\overline{A}_{t-1}) + \frac{1}{t} \nabla \Phi(\overline{A}_{t-1}) \cdot \left( \pi_S(\overline{A}_{t-1}) - \overline{A}_{t-1} \right) + \frac{2B}{t^2}
$$

(by the defining property of $p_t$)

$$
\leq \Phi(\overline{A}_{t-1}) - \frac{1}{t} \Phi(\overline{A}_{t-1}) + \frac{2B}{t^2}
$$

(by the property noted at the beginning of the proof).
Multiplying both sides of the obtained inequality by $t$ we get

$$t \Phi(\overline{A}_t) \leq (t - 1) \Phi(\overline{A}_{t-1}) + \frac{2B}{t}$$

for all $t \geq 1$.

Summing this inequality for $t = 1, \ldots, n$, we have

$$n \Phi(\overline{A}_n) \leq \sum_{t=1}^{n} \frac{2B}{t},$$

which proves the first statement of the theorem. The almost sure convergence of $\overline{A}_n$ to $S$ may now be obtained easily using the Hoeffding–Azuma inequality, just as in the proof of Theorem 7.5. ■

### 7.9 Convergence to Nash Equilibria

We have seen in the previous sections that if all players follow simple strategies based on (internal) regret minimization, then the joint empirical frequencies of play approach the set of correlated equilibria at a remarkably fast rate. Here we investigate the considerably more difficult problem of achieving Nash equilibria.

Just as before, we only consider uncoupled strategies; that is, each player knows his own payoff function but not that of the rest of the players. In this section we assume standard monitoring; that is, after taking an action, all players observe the other players’ moves. In other words, the participants of the game know what the others do, but they do not know why they do it. In the next section we treat the more difficult case in which the players only observe their own losses but not the actions the other players take.

Our main concern here is to describe strategies of play that guarantee that no matter what the underlying game is, the players end up playing a mixed strategy profile close to a Nash equilibrium. The difficulty is that players cannot optimize their play because of the assumption of uncoupledness. The problem becomes more pronounced in the case of an “unknown game” (i.e., when players only observe their own payoffs but know nothing about what the other participants of the game do), treated in the next section.

As mentioned in the introduction, in many cases, including two-person zero-sum games and $K$-person games in which every player has two actions (i.e., $N_k = 2$ for all $k$), simple fictitious play is sufficient to achieve convergence to a Nash equilibrium (in a certain weak sense). However, the case of general games is considerably more difficult.

To describe the first simple idea, assume that the game has a pure action Nash equilibrium, that is, a Nash equilibrium $\pi$ concentrated on a single vector $i = (i_1, \ldots, i_K)$ of actions. In this case it is easy to construct a randomized uncoupled strategy of repeated play such that if all players follow such a strategy, the joint strategy profiles converge to a Nash equilibrium almost surely. Perhaps the simplest such procedure is described as follows.
AN UNCOUPLED STRATEGY TO FIND PURE
NASH EQUILIBRIA

Strategy for player $k$.

For $t = 1, 2, \ldots$

1. if $t$ is odd, choose an action $I_t^{(k)}$ randomly, according to the uniform distribution over $\{1, \ldots, N_k\}$;
2. if $t$ is even, let
   
   $$I_t^{(k)} = \begin{cases} 
   1 & \text{if } \ell^{(k)}(I_{t-1}) \leq \min_{i_k = 1, \ldots, N_k} \ell^{(k)}(I_{t-1}, i_k) \\
   2 & \text{otherwise.} 
   \end{cases}$$

3. If $t$ is even and all players have played action 1, then repeat forever the action played in the last odd period.

In words, at each odd period the players choose an action randomly. At the next period they check if their action was a best response to the other players’ actions and communicate it to the other players by playing action 1 or 2. If all players have best responded at some point (which they confirm by playing action 1 in the next round), then they repeat the same action forever. This strategy clearly realizes a random exhaustive search. Because the game is finite, eventually, almost surely, by pure chance, at some odd time period a pure action Nash equilibrium $i = (i_1, \ldots, i_K)$ will be played (whose existence is guaranteed by assumption). The expected time it takes to find such an equilibrium is at most $2 \prod_{k=1}^{K} N_k$, a quantity exponentially large in the number $K$ of players. By the definition of Nash equilibrium all players best respond at the time $i$ is played and therefore stay with this choice forever.

**Remark 7.9 (Exhaustive search).** The algorithm shown above is a simple form of exhaustive search. In fact, all strategies that we describe here and in the next section are variants of the same principle. Of course, this also means that convergence is extremely slow in the sense that the whole space (exponentially large as a function of the number of players) needs to be explored. This is in sharp contrast to the internal-regret-minimizing strategies of Section 7.4 that guarantee rapid convergence to the set of correlated equilibria.

**Remark 7.10 (Nonstationarity).** The learning strategy described above may not be very appealing because of its nonstationarity. Exercise 7.25 describes a stationary strategy proposed by Hart and Mas-Colell [149] that also achieves almost sure convergence to a pure action Nash equilibrium whenever such an equilibrium exists.

**Remark 7.11 (Multiple equilibria).** The procedure always converges to a pure action Nash equilibrium. Of course, the game may have several Nash equilibria: some pure, others mixed. Some equilibria may be “better” than others, sometimes in quite a strong sense. However, the procedure does not make any distinction between different equilibria, and the one it finds is selected randomly, according to the uniform distribution over all pure action Nash equilibria.
Remark 7.12 (Two players). It is important to note that the case of \( K = 2 \) players is significantly simpler. In a two-player game (under a certain genericity assumption) it suffices to use a strategy that repeats the previous play if it was a best response and selects an action randomly otherwise (see Exercise 7.26). However, if the game is not generic or in the case of at least three players, this procedure may enter in a cycle and never converge (Exercise 7.27).

Next we show how the ideas described above may be extended to general games that may not have a pure action Nash equilibrium. The simplest version of this extension does not quite achieve a Nash equilibrium but approximates it. To make the statement precise, we introduce the notion of an \( \varepsilon \)-Nash equilibrium. Let \( \varepsilon > 0 \). A mixed strategy profile \( \pi = p^{(1)} \times \cdots \times p^{(K)} \) is an \( \varepsilon \)-Nash equilibrium if for all \( k = 1, \ldots, K \) and all mixed strategies \( q^{(k)} \),

\[
\pi^{(k)} \leq \pi^{(k)} + \varepsilon,
\]

where \( \pi^{(k)} = p^{(1)} \times \cdots \times q^{(k)} \times \cdots \times p^{(K)} \) denotes the mixed strategy profile obtained by replacing \( p^{(k)} \) by \( q^{(k)} \) and leaving all other players’ mixed strategies unchanged. Thus, the definition of Nash equilibrium is modified simply by allowing some slack in the defining inequalities. Clearly, it suffices to check the inequalities for mixed strategies \( q^{(k)} \) concentrated on a single action \( i_k \). The set of \( \varepsilon \)-Nash equilibria of a game is denoted by \( N_\varepsilon \).

The idea in extending the procedure discussed earlier for general games is that each player first selects a mixed strategy randomly and then checks whether it is an approximate best response to the others’ mixed strategies. Clearly, this cannot be done in just one round of play because players cannot observe the mixed strategies of the others. But if the same mixed strategy is played during sufficiently many periods, then each player can simply test the hypothesis whether his choice is an approximate best response. This procedure is formalized as follows.

### A REGRET-TESTING PROCEDURE TO FIND NASH EQUILIBRIA

**Strategy for player \( k \).**

**Parameters:** Period length \( T \), confidence parameter \( \rho > 0 \).

**Initialization:** Choose a mixed strategy \( \pi_0^{(k)} \) randomly, according to the uniform distribution over the simplex \( D_k \) of probability distributions over \( \{1, \ldots, N_k\} \).

For \( t = 1, 2, \ldots, \)

1. if \( t = mT + s \) for integers \( m \geq 0 \) and \( 1 \leq s \leq T - 1 \), choose \( I_t^{(k)} \) randomly according to the mixed strategy \( \pi_m^{(k)} \);
2. if \( t = mT \) for an integer \( m \geq 1 \), then let

\[
I_t^{(k)} = \begin{cases} 
1 & \text{if } \frac{1}{T-1} \sum_{s=(m-1)T+1}^{mT-1} \ell^{(k)}(I_s) \leq \min_{i_1, \ldots, i_{N_k}} \frac{1}{T-1} \sum_{s=(m-1)T+1}^{mT-1} \ell^{(k)}(I_s, i_k) + \rho \\
2 & \text{otherwise};
\end{cases}
\]

3. If \( t = mT \) and all players have played action 1, then play the mixed strategy \( \pi_{m-1}^{(k)} \) forever; otherwise, choose \( \pi_m^{(k)} \) randomly, according to the uniform distribution over \( D_k \).
Prediction and Playing Games

In this procedure every player tests, after each period of length $T$, whether he has been approximately best responding and sends a signal to the rest of the players at the end of the period. Once every player accepts the hypothesis of almost best response, the same mixed strategy profile is repeated forever. The next simple result summarizes the performance of this procedure.

**Theorem 7.7.** Assume that every player of a game plays according to the regret-testing procedure described above that is run with parameters $T$ and $\rho \leq 1/2$ such that $(T - 1)\rho^2 \geq 2 \ln \sum_{k=1}^{K} N_k$. With probability 1, there is a mixed strategy profile $\pi = p^{(1)} \times \cdots \times p^{(K)}$ such that $\pi$ is played for all sufficiently large $m$. The probability that $\pi$ is not a $2\rho$-Nash equilibrium is at most

$$8T \rho^{2-d} \ln \left( T \rho^{2-d} \right) e^{-2(T-1)\rho^2},$$

where $d = \sum_{k=1}^{K} (N_k - 1)$ is the number of free parameters needed to specify any mixed strategy profile.

It is clear from the bound of the theorem that for an arbitrarily small value of $\rho$, if $T$ is sufficiently large, the probability of not ending up in a $2\rho$-Nash equilibrium can be made arbitrarily small. This probability decreases with $T$ at a remarkably fast rate. Just note that if $T \rho^2$ is significantly larger than $d \ln d$, the probability may be further bounded by $e^{-T\rho^2}$. Thus, the size of the game plays a minor role in the bound of the probability. However, the time the procedure takes to reach the near-equilibrium state depends heavily on the size of the game. The estimates derived in the proof reveal that this stopping time is typically of order $T \rho^{-d}$, exponentially large in the number of players. However, convergence to $\epsilon$-Nash equilibria cannot be achieved with this method for an arbitrarily small value of $\epsilon$, because no matter how $T$ and $\rho$ are chosen, there is always a positive, though tiny, probability that all players accept a mixed strategy profile that is not an $\epsilon$-Nash equilibrium. In the next section we describe a different procedure that may be converted into an almost surely convergent strategy.

**Proof of Theorem 7.7.** Let $M$ denote the (random) index for which every player plays 1 at time $MT$. We need to prove that $M$ is finite almost surely. Consider the process $\pi_0, \pi_1, \pi_2, \ldots$ of mixed strategy profiles found at times 0, $T$, $2T$, \ldots by the players using the procedure. For convenience, assume that players keep drawing a new random mixed strategy $\pi_m$ even if $m \geq M$ and the mixed strategy is not used anymore. In other words, $\pi_0, \pi_1, \pi_2, \ldots$ are independent random variables uniformly distributed over the $d = \sum_{k=1}^{K} (N_k - 1)$-dimensional set obtained by the product of the $K$ simplices of the mixed strategies of the $K$ players.

For any $\epsilon \in (0, 1)$ we may write

$$P[M \geq m_0] = \sum_{j=0}^{m_0-1} P[\pi_m \in N_\epsilon \ j \ times] \times P[\pi_m \ is \ rejected \ for \ every \ m < m_0 \ | \ \pi_m \in N_\epsilon \ j \ times].$$

We bound the terms on the right-hand side to derive the desired estimate for $P[M \geq m_0]$. 
First observe that since the game has at least one Nash equilibrium \( \bar{\pi} = \bar{\pi}^{(1)} \times \cdots \times \bar{\pi}^{(K)} \), any mixed strategy profile \( \bar{\pi} = \bar{\pi}^{(1)} \times \cdots \times \bar{\pi}^{(K)} \) such that

\[
\max_{k=1,\ldots,K} \max_{i_k=1,\ldots,N_k} |\bar{\pi}^{(k)}(i_k) - \bar{\pi}^{(k)}(i_k)| \leq \varepsilon
\]

is an \( \varepsilon \)-Nash equilibrium. Thus, for every \( m = 0, 1, \ldots \) we have

\[
P[\pi_m \in \mathcal{N}_\varepsilon] \geq \varepsilon^d.
\]

This implies that

\[
P[\pi_m \in \mathcal{N}_\varepsilon \ j \ times] = P[\pi_m \notin \mathcal{N}_\varepsilon \ m_0 - j \ times] \leq \binom{m_0}{j} (1 - \varepsilon^d)^{m_0 - j} \leq m_0^j e^{-(m_0 - j)\varepsilon^d}.
\]

To bound the other term in this expression of \( P[M \geq m_0] \), note that if \( \pi_m \in \mathcal{N}_\varepsilon \) exactly \( j \) times for \( m = 0, \ldots, m_0 - 1 \), then \( M \geq m_0 \) implies that at least one player observes a regret at least \( \rho \) at all of these \( j \) periods. Note that at any given period \( m \), if \( \varepsilon > \rho \), by using the fact that the regret estimates are sums of \( T - 1 \) independent, identically distributed random variables, Hoeffding’s inequality and the union-of-events bound imply that

\[
P[M \neq m | \pi_m \in \mathcal{N}_\varepsilon] \leq P[\exists k \leq K : I_m^{(k)} = 2 | \pi_m \in \mathcal{N}_\varepsilon] \leq \sum_{k=1}^{K} N_k e^{-2(T-1)(\varepsilon-\rho)^2}.
\]

Therefore,

\[
P[\pi_m \text{ is rejected for every } m < m_0 \mid \pi_m \in \mathcal{N}_\varepsilon \ j \ times] \leq \left( \sum_{k=1}^{K} N_k \right)^j e^{-2j(T-1)(\varepsilon-\rho)^2}.
\]

Putting everything together, letting \( \varepsilon = 2\rho \), using the assumption \( (T - 1)\rho^2 \geq 2 \ln \sum_{k=1}^{K} N_k \), and fixing a \( j_0 < m_0 \), we have

\[
P[M \geq m_0] \leq \sum_{j \leq j_0} m_0^j e^{-(m_0 - j)(2\rho)^d} + \sum_{j > j_0} \left( \sum_{k=1}^{K} N_k \right)^j e^{-j(T-1)\rho^2} \leq j_0 m_0^{j_0} e^{-(m_0 - j_0)(2\rho)^d} + m_0 e^{-j_0(T-1)\rho^2}/2.
\]

Since we are free to choose \( j_0 \), we take \( j_0 = \lfloor m_0 \rho^{d-2}/(T \ln m_0) \rfloor \). To guarantee that \( j_0 < m_0 \) we assume that \( m_0 \geq \rho^{-d} (2 \log \rho^{-d})^2 \). This implies that \( m_0 > \rho^{-d} (\ln m_0)^2 \), which in turn implies the desired condition \( j_0 < m_0 \). Thus, using this choice of \( m_0 \) and after some simplification, we get

\[
P[M \geq m_0] \leq 2e^{-m_0 \rho^d/(2 \ln m_0)}.
\]

But then the Borel–Cantelli lemma implies that \( M \) is finite almost surely.

Now let \( \pi \) denote the mixed strategy profile that the process ends up repeating forever. (This random variable is well defined with probability 1 according to the fact that \( M < \infty \) almost surely.) Then the probability that \( \pi \) is not a \( 2\rho \)-Nash equilibrium may be bounded
by writing
\[ P[\pi \not\in N_{2\rho}] = \sum_{m=1}^{\infty} P[M = m, \pi \not\in N_{2\rho}]. \]

For \( m \geq m_0 \) (where the value of \( m_0 \) is determined later) we simply use the bound \( P[M = m, \pi \not\in N_{2\rho}] \leq P[M = m] \). On the other hand, for any \( m \), by Hoeffding’s inequality,
\[ P[M = m, \pi \not\in N_{2\rho}] \leq P[M = m | \pi \not\in N_{2\rho}] \leq e^{-2(T-1)\rho^2}. \]

Thus, for any \( m_0 \), we get
\[ P[\pi \not\in N_{2\rho}] \leq m_0 e^{-2(T-1)\rho^2} + P[M \geq m_0]. \]

Here we may use the estimate (7.2), which is valid for all \( m_0 \geq \rho^{-d} \left(2 \log \rho^{-d}\right)^2 \). A convenient choice is to take \( m_0 \) to be the greatest integer such that \( m_0/(2 \ln m_0) \leq T \rho^{2-d} \).

Then \( m_0 \leq 4T \rho^{2-d} \ln \left(T \rho^{2-d}\right) \), and we obtain
\[ P[\pi \not\in N_{2\rho}] \leq 8T \rho^{2-d} \ln \left(T \rho^{2-d}\right) e^{-2(T-1)\rho^2} \]
as stated. 

7.10 Convergence in Unknown Games

The purpose of this section is to show that even in the much more restricted framework of “unknown” games described in Section 7.5, it is possible to achieve, asymptotically, an approximate Nash equilibrium. Recall that in this model, as the players play a game repeatedly, they not only do not know the other players’ loss function but they do not even know their own, and the only information they receive is their loss suffered at each round of the game, after taking an action.

The basic idea of the strategy we investigate in this section is somewhat reminiscent of the regret-testing procedure described in the previous section. On a quick inspection of the procedure it is clear that the assumption of standard monitoring (i.e., that players are able to observe their opponents’ actions) is used twice in the definition of the procedure. On the one hand, the players should be able to compute their estimated regret in each period of length \( T \). On the other hand, they need to communicate to the others whether their estimated regret is smaller than a certain threshold or not. It turns out that it is easy to find a solution to the first problem, because players may easily and reliably estimate their regret even if they do not observe others’ actions. This should not come as a surprise after having seen that regret minimization is possible in the bandit problem (see Sections 6.7, 6.8, and 7.5). In fact, the situation here is significantly simpler. However, the lack of ability of communication with the other players poses more serious problems because it is difficult to come up with a stopping rule as in the procedure shown in the previous section.

The solution is a procedure in which, just as before, time is divided into periods of length \( T \), all players keep testing their regret in each period, and they stay with their previously chosen mixed strategy if they have had a satisfactorily small estimated regret. Otherwise they choose a new mixed strategy randomly, just like above. To make the ideas more transparent, first we ignore the issue of estimating regrets in the model of unknown game
and assume that, after every period \((m-1)T + 1, \ldots, mT\), each player \(k = 1, \ldots, K\) can compute the regrets

\[
r^{(k)}_{m, i_k} = \frac{1}{T} \sum_{s=(m-1)T+1}^{mT} \ell^{(k)}(I_s) - \frac{1}{T} \sum_{s=(m-1)T+1}^{mT} \ell^{(k)}(I_s^{(k)}),
\]

for all \(i_k = 1, \ldots, N_k\). Now we may define a regret-testing procedure as follows.

### EXPERIMENTAL REGRET TESTING

**Strategy for player \(k\).**

**Parameters:** Period length \(T\), confidence parameter \(\rho > 0\), exploration parameter \(\lambda > 0\).

**Initialization:** Choose a mixed strategy \(\pi^{(k)}_0\) randomly, according to the uniform distribution over the simplex \(D_k\) of probability distributions over \(\{1, \ldots, N_k\}\);

For \(t = 1, 2, \ldots\)

1. if \(t = mT + s\) for integers \(m \geq 0\) and \(1 \leq s < T\), choose \(I^{(k)}_t\) randomly according to the mixed strategy \(\pi^{(k)}_m\);
2. if \(t = mT\) for an integer \(m \geq 1\), then
   - if \(\max_{i_k = 1, \ldots, N_k} r^{(k)}_{m, i_k} > \rho\) then choose \(\pi^{(k)}_m\) randomly according to the uniform distribution over \(D_k\);
   - otherwise, with probability \(1 - \lambda\) let \(\pi^{(k)}_m = \pi^{(k)}_{m-1}\), and with probability \(\lambda\) choose \(\pi^{(k)}_m\) randomly according to the uniform distribution over \(D_k\).

The parameters \(\rho\) and \(T\) play a similar role to that played in the previous section. The introduction of the exploration parameter \(\lambda > 0\) is technical; it is needed for the proofs given below but it is unclear whether the natural choice \(\lambda = 0\) would give similar results. With \(\lambda > 0\), even if a player has all regrets below the threshold \(\rho\), the player will reserve a positive probability of exploration. A strictly positive value of \(\lambda\) guarantees that the sequence \(\{\pi_m\}\) of mixed strategy profiles forms a rapidly mixing Markov process. In fact, the first basic lemma establishes this property. For convenience we denote the set \(\prod_{k=1}^K D_k\) of all mixed strategy profiles (i.e., product distributions) by \(\Sigma\). Also, we introduce the notation \(N = \sum_{k=1}^K N_k\).

**Lemma 7.4.** The stochastic process \(\{\pi_m\}, m = 0, 1, 2, \ldots\), defined by experimental regret testing with \(0 < \lambda < 1\), is a homogeneous, recurrent, and irreducible Markov chain satisfying Doeblin’s condition. In particular, for any measurable set \(A \subset \Sigma\),

\[
P(\pi \rightarrow A) \geq \lambda^N \mu(A)
\]

for every \(\pi \in \Sigma\), where \(P(\pi \rightarrow A) = P[\pi_{m+1} \in A \mid \pi_m = \pi]\) denotes the transition probabilities of the Markov chain, \(\mu\) denotes the uniform distribution on \(\Sigma\), and \(m\) is any nonnegative integer.
Proof. To see that the process is a Markov chain, note that at each \( m = 0, 1, 2, \ldots \), \( \pi_m \) depends only on \( \pi_{m-1} \) and the regrets \( r_{m,i}^{(k)} \) \((i_k = 1, \ldots, N_k, k = 1, \ldots, K)\). It is irreducible since at each 0, \( T, 2T, \ldots \), the probability of reaching some \( \pi'_m \in A \) for any open set \( A \subset \Sigma \) from any \( \pi_{m-1} \in \Sigma \) is strictly positive when \( \lambda > 0 \), and it is recurrent since

\[
E \left[ \sum_{m=0}^{\infty} 1_{\{\pi_m \in A\}} \mid \pi_0 \in A \right] = \infty \text{ for all } \pi_0 \in A.
\]

Doeblin’s condition follows simply from the presence of the exploration parameter \( \lambda \) in the definition of experimental regret testing. In particular, with probability \( \lambda^N \) every player chooses a mixed strategy randomly, and conditioned on this event the distribution of \( \pi_m \) is uniform. ■

The lemma implies that \( \{\pi_m\} \) is a rapidly mixing Markov chain. The behavior of such Markov processes is well understood (see, e.g., the monograph of Meyn and Tweedie [217] for an excellent coverage). The properties we use subsequently are summarized in the following corollary.

**Corollary 7.2.** For \( m = 0, 1, 2, \ldots \) let \( P_m \) denote the distribution of the mixed strategy profile \( \pi_m = (\pi_m^{(1)}, \ldots, \pi_m^{(K)}) \) chosen by the players at time \( mT \), that is, \( P_m(A) = \mathbb{P}[\pi_m \in A] \). Then there exists a unique probability distribution \( Q \) over \( \Sigma \) (the stationary distribution of the Markov process) such that

\[
\sup_A |P_m(A) - Q(A)| \leq (1 - \lambda^N)^m,
\]

where the supremum is taken over all measurable sets \( A \subset \Sigma \) (see [217, Theorem 16.2.4]). Also, the ergodic theorem for Markov chains implies that

\[
\lim_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} \pi_m = \int \Sigma \pi \, dQ(\pi) \quad \text{almost surely.}
\]

The main idea behind the regret-testing heuristics is that, after a not very long search period, by pure chance, the mixed strategy profile \( \pi_m \) will be an \( \varepsilon \)-Nash equilibrium, and then, since all players have a small expected regret, the process gets stuck with this value for a much longer time than the search period. The main technical result needed to justify such a statement is summarized in Lemma 7.5. This implies that if the parameters of the procedure are set appropriately, the length of the search period is negligible compared with the length of time the process spends in an \( \varepsilon \)-Nash equilibrium. The proof of Lemma 7.5 is quite technical and is beyond the scope of this book. See the bibliographic remarks for the appropriate pointers. In addition, the proof requires certain properties of the game that are not satisfied by all games. However, the necessary conditions hold for almost all games, in the sense that the Lebesgue measure of all those games that do not satisfy these conditions is 0. (Here we consider the representation of a game as the \( K \sum_{k=1}^{K} N_k \)-dimensional vector of all losses \( \ell^{(i)}(i,k) \). Let \( \bar{N}_\varepsilon = \Sigma \setminus N_\varepsilon \) denote the complement of the set of \( \varepsilon \)-Nash equilibria.

**Lemma 7.5.** For almost all \( K \)-person games there exist positive constants \( c_1, c_2 \) such that, for all sufficiently small \( \rho > 0 \), the \( K \)-step transition probabilities of experimental regret testing satisfy

\[
P^{(K)}(\bar{N}_\rho \to N_\rho) \geq c_1 \rho^{c_2},
\]
where we use the notation $P^{(K)}(A \rightarrow B) = \mathbb{P}[\pi_{m+K} \in B \mid \pi_m \in A]$ for the $K$-step transition probabilities.

On the basis of this lemma, we can now state one of the basic properties of the experimental regret-testing procedure. The result states that, in the long run, the played mixed strategy profile is not an approximate Nash equilibrium at a tiny fraction of time.

**Theorem 7.8.** Almost all games are such that there exists a positive number $\varepsilon_0$ and positive constants $c_1, \ldots, c_4$ such that for all $\varepsilon < \varepsilon_0$ if the experimental regret-testing procedure is used with parameters

$$\rho \in (\varepsilon, \varepsilon + \varepsilon^\delta), \quad \lambda \leq c_2 \varepsilon^\delta, \quad \text{and} \quad T \geq -\frac{1}{2(\rho - \varepsilon)^2} \log (c_4 \varepsilon^\delta),$$

then for all $M \geq \log(\varepsilon/2)/\log(1 - \lambda N)$,

$$P_M(\mathcal{N}_\varepsilon) = \mathbb{P}[\sigma MT \notin \mathcal{N}_\varepsilon] \leq \varepsilon.$$

**Proof.** First note that by Corollary 7.2,

$$P_M(\mathcal{N}_\varepsilon) \leq Q(\mathcal{N}_\varepsilon) + (1 - \lambda N)^M,$$

so that it suffices to bound the measure of $\mathcal{N}_\varepsilon$ under the stationary probability $Q$. To this end, first observe that, by the defining property of the stationary distribution,

$$Q(\mathcal{N}_\rho) = Q(\mathcal{N}_\rho) P^{(K)}(\mathcal{N}_\rho \rightarrow \mathcal{N}_\rho) + Q(\mathcal{N}_\rho) P^{(K)}(\mathcal{N}_\rho \rightarrow \mathcal{N}_\rho).$$

Solving for $Q(\mathcal{N}_\rho)$ gives

$$Q(\mathcal{N}_\rho) = \frac{P^{(K)}(\mathcal{N}_\rho \rightarrow \mathcal{N}_\rho)}{1 - P^{(K)}(\mathcal{N}_\rho \rightarrow \mathcal{N}_\rho) + P^{(K)}(\mathcal{N}_\rho \rightarrow \mathcal{N}_\rho)}. \quad (7.3)$$

To derive a lower bound for the expression on the right-hand side, we write the elementary inequality

$$P^{(K)}(\mathcal{N}_\rho \rightarrow \mathcal{N}_\rho) = \frac{Q(\mathcal{N}_\varepsilon) P^{(K)}(\mathcal{N}_\varepsilon \rightarrow \mathcal{N}_\rho)}{Q(\mathcal{N}_\rho)} + \frac{Q(\mathcal{N}_\rho \setminus \mathcal{N}_\varepsilon) P^{(K)}(\mathcal{N}_\rho \setminus \mathcal{N}_\varepsilon \rightarrow \mathcal{N}_\rho)}{Q(\mathcal{N}_\rho)} \geq \frac{Q(\mathcal{N}_\varepsilon) P^{(K)}(\mathcal{N}_\varepsilon \rightarrow \mathcal{N}_\rho)}{Q(\mathcal{N}_\rho)}. \quad (7.4)$$

To bound $P^{(K)}(\mathcal{N}_\varepsilon \rightarrow \mathcal{N}_\rho)$, note that if $\pi_m \in \mathcal{N}_\varepsilon$, then the expected regret of all players is at most $\varepsilon$. Since the regret estimates $r_{i,ik}^{(k)}$ are sums of $T$ independent random variables taking values between 0 and 1 with mean at most $\varepsilon$, Hoeffding’s inequality implies that

$$\mathbb{P}[r_{i,ik}^{(k)} \geq \rho] \leq e^{-2T(\rho - \varepsilon)^2}, \quad i_k = 1, \ldots, N_k, \quad k = 1, \ldots, K.$$

Then the probability that there is at least one player $k$ and a strategy $i_k \leq N_k$ such that $r_{i,ik}^{(k)} \geq \rho$ is bounded by $\sum_{k=1}^{K} N_k e^{-2T(\rho - \varepsilon)^2} = N e^{-2T(\rho - \varepsilon)^2}$. Thus, with probability at least
(1 − \lambda)^K \left( 1 - Ne^{-2T(\rho - \varepsilon)^2} \right), all players keep playing the same mixed strategy, and therefore
\[ P(N_\varepsilon \rightarrow N_\varepsilon) \geq (1 - \lambda)^K \left( 1 - Ne^{-2T(\rho - \varepsilon)^2} \right). \]
Consequently, since \rho > \varepsilon, we have \[ P^{(K)}(N_\varepsilon \rightarrow N_\rho) \geq P^{(K)}(N_\varepsilon \rightarrow N_\varepsilon) \] and hence
\[ P^{(K)}(N_\varepsilon \rightarrow N_\rho) \geq (1 - \lambda)^K \left( 1 - Ne^{-2T(\rho - \varepsilon)^2} \right)^K \geq 1 - K^2 \lambda - NK e^{-2T(\rho - \varepsilon)^2} \]
(where we assumed that \lambda \leq 1 and \(Ne^{-2T(\rho - \varepsilon)^2} \leq 1\)). Thus, using (7.4) and the obtained estimate, we have
\[ P^{(K)}(N_\rho \rightarrow N_\rho) \geq \frac{Q(N_\varepsilon)}{Q(N_\rho)} \left( 1 - K^2 \lambda - NK e^{-2T(\rho - \varepsilon)^2} \right). \]

Next we need to show that, for proper choice of the parameters, \( P^{(K)}(N_\rho \rightarrow N_\rho) \) is sufficiently large. For almost all of \( K \)-person games, this follows from Lemma 7.5, which asserts that
\[ P^{(K)}(N_\rho \rightarrow N_\rho) \geq C_1 \rho^{C_2} \]
for some positive constants \(C_1\) and \(C_2\) that depend on the game. Hence, from (7.3) we obtain
\[ Q(N_\rho) \geq \frac{C_1 \rho^{C_2}}{1 - \left( 1 - K^2 \lambda - NK e^{-2T(\rho - \varepsilon)^2} \right) \frac{Q(N_\varepsilon)}{Q(N_\rho)} + C_1 \rho^{C_2}}. \]
It remains to estimate the measure \(Q(N_\varepsilon)/Q(N_\rho)\). We need to show that the ratio is close to 1 whenever \(\rho - \varepsilon \ll \varepsilon\). It turns out that one can show that, in fact, for almost every game there exists a constant \(C_3\) such that
\[ \frac{Q(N_\varepsilon)}{Q(N_\rho)} \geq 1 - \frac{C_3 (\rho - \varepsilon)^{C_4}}{\rho^{C_5}}, \]
where \(C_3\) and \(C_4\) are positive constants depending on the game. This inequality is not surprising, but the rigorous proof of this statement is somewhat technical and is skipped here. It may be found in [126]. Summarizing,
\[ Q(N_\varepsilon) \geq Q(N_\rho) \left( 1 - \frac{C_3 (\rho - \varepsilon)^{C_4}}{\rho^{C_5}} \right) \]
\[ \geq \left( 1 - \frac{C_3 (\rho - \varepsilon)^{C_4}}{\rho^{C_5}} \right) \times \frac{C_1 \rho^{C_2}}{1 - \left( 1 - K^2 \lambda - NK e^{-2T(\rho - \varepsilon)^2} \right) \left( 1 - \frac{C_3 (\rho - \varepsilon)^{C_4}}{\rho^{C_5}} \right) + C_1 \rho^{C_2}} \]
for some positive constants \(C_1, \ldots, C_5\). Choosing the parameters \(\rho, \lambda, T\) with appropriate constants \(c_1, \ldots, c_4\), we have
\[ Q(N_\rho) \leq \varepsilon/2. \]
If \(M\) is so large that \((1 - \lambda^M)^M \leq \varepsilon/2\), we have \(P_M(N_\varepsilon) \leq \varepsilon\), as desired. ■
Theorem 7.8 states that if the parameters of the experimental regret-testing procedure are set in an appropriate way, the mixed strategy profiles will be in an approximate equilibrium most of the time. However, it is important to realize that the theorem does not claim convergence in any way. In fact, if the parameters $T, \rho, \text{ and } \lambda$ are kept fixed forever, the process will periodically abandon the set of $\varepsilon$-Nash equilibria and wander around for a long time before it gets stuck in a (possibly different) $\varepsilon$-Nash equilibrium. Then the process stays there for an even much longer time before leaving again. However, since the process $\{\pi_m\}$ forms an ergodic Markov chain, it is easy to deduce convergence of the empirical frequencies of play. Specifically, we show next that if all players play according to the experimental regret-testing procedure, then the joint empirical frequencies of play converge almost surely to a joint distribution $\overline{P}$ that is in the convex hull of $\varepsilon$-Nash equilibria. The precise statement is given in Theorem 7.9.

Recall that for each $t = 1, 2, \ldots$ we denote by $I^{(k)}_t$ the pure strategy played by the $k$th player. $I^{(k)}_t$ is drawn randomly according to the mixed strategy $\pi^{(k)}_m$ whenever $t \in \{mT + 1, \ldots, (m + 1)T\}$. Consider the joint empirical distribution of plays $\hat{P}_t$ defined by

$$\hat{P}_t(i) = \frac{1}{t} \sum_{s=1}^{t} \mathbb{I}_{[I_s = i]}, \quad i \in \prod_{k=1}^{K} \{1, \ldots, N_k\}.$$

Denote the convex hull of a set $A$ by $\text{co}(A)$.

**Remark 7.13.** Recall that Nash equilibria and $\varepsilon$-Nash equilibria $\pi$ are mixed strategy profiles, that is, product distributions, and have been considered, up to this point, as elements of the set $\Sigma$ of product distributions. However, a product distribution is a special joint distribution over the set $\prod_{k=1}^{K} \{1, \ldots, N_k\}$ of pure strategy profiles, and it is this “larger” space in which the convex hull of $\varepsilon$-Nash equilibria is defined. Thus, elements of the convex hull are typically not product distributions. (Recall that the convex hull of Nash equilibria is a subset of the set of correlated equilibria.)

**Theorem 7.9.** For almost every game and for every sufficiently small $\varepsilon > 0$, there exists a choice of the parameters $(T, \rho, \lambda)$ such that the following holds: there is a joint distribution $\overline{P}$ over the set of $K$-tuples $i = (i_1, \ldots, i_K)$ of actions in the convex hull $\text{co}(\mathcal{N}_\varepsilon)$ of the set of $\varepsilon$-Nash equilibria such that the joint empirical frequencies of play of experimental regret testing satisfy

$$\lim_{t \to \infty} \hat{P}_t \to \overline{P} \quad \text{almost surely.}$$

**Proof.** If $\pi = (\pi^{(1)} \times \cdots \times \pi^{(K)}) \in \Sigma$ is a product distribution, introduce notation

$$P(\pi, i) = \prod_{k=1}^{K} \pi^{(k)}(i_k),$$

where $i = (i_1, \ldots, i_K)$. In other words, $P(\pi, \cdot)$ is a joint distribution over the set of action profiles $i$, induced by $\pi$. 

---

7.10 Convergence in Unknown Games
First observe that, since at time $t$ the vector $I_t$ of actions is chosen according to the mixed strategy profile $\pi_{\lfloor t/T \rfloor}$, by martingale convergence, for every $i$,

$$\hat{P}_t(i) - \frac{1}{t} \sum_{s=1}^{t} P(\pi_{\lfloor s/T \rfloor}, i) \to 0 \quad \text{almost surely},$$

Therefore, it suffices to prove convergence of $\frac{1}{t} \sum_{s=1}^{t} P(\pi_{\lfloor s/T \rfloor}, i)$. Since $\pi_{\lfloor s/T \rfloor}$ is unchanged during periods of length $T$, we obviously have

$$\lim_{t \to \infty} \frac{1}{t} \sum_{s=1}^{t} P(\pi_{\lfloor s/T \rfloor}, i) = \lim_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} P(\pi_m, i).$$

By Corollary 7.2,

$$\lim_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} \pi_m = \bar{\pi} \quad \text{almost surely},$$

where $\bar{\pi} = \int_{\Sigma} \pi \, dQ(\pi)$. (Recall that $Q$ is the unique stationary distribution of the Markov process.) This, in turn, implies by continuity of $P(\pi, i)$ in $\pi$ that there exists a joint distribution $\overline{P}(i) = \int_{\Sigma} P(\pi, i) \, dQ(\pi)$ such that, for all $i$,

$$\lim_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} P(\pi_m, i) = \overline{P}(i) \quad \text{almost surely}.$$

It remains to show that $\overline{P} \in \text{co}(N_{\varepsilon})$.

Let $\varepsilon' < \varepsilon$ be a positive number such that the $\varepsilon'$ blowup of $\text{co}(N_{\varepsilon'})$ is contained in $\text{co}(N_{\varepsilon})$, that is,

$$\{ P \in \Sigma : \exists P' \in \text{co}(N_{\varepsilon'}) \text{ such that } \| P - P' \|_1 < \varepsilon' \} \subset \text{co}(N_{\varepsilon}).$$

Such an $\varepsilon'$ always exists for almost all games by Exercise 7.28. In fact, one may choose $\varepsilon' = \varepsilon/c_3$ for a sufficiently large positive constant $c_3$ (whose value depends on the game).

Now choose the parameters $(T, \rho, \lambda)$ such that $Q(N_{\varepsilon'}) < \varepsilon'$. Theorem 7.8 guarantees the existence of such a choice.

Clearly,

$$P(\overline{P}, i) = \int_{\Sigma} P(\pi, i) \, dQ(\pi) = \int_{N_{\varepsilon'}} P(\pi, i) \, dQ(\pi) + \int_{N_{\varepsilon'}} P(\pi, i) \, dQ(\pi).$$

Since $\int_{N_{\varepsilon'}} P(\pi) \, dQ(\pi) \in \text{co}(N_{\varepsilon'})$, we find that the $L_1$ distance of $\overline{P}$ and $\text{co}(N_{\varepsilon'})$ satisfies

$$d_1(\overline{P}, \text{co}(N_{\varepsilon'})) \leq \left\| \int_{N_{\varepsilon'}} P(\pi) \, dQ(\pi) \right\|_1 \leq \int_{N_{\varepsilon'}} dQ(\pi) = Q(N_{\varepsilon'}) < \varepsilon'.$$

By the choice of $\varepsilon'$ we indeed have $\overline{P} \in \text{co}(N_{\varepsilon}).$ \hfill $\blacksquare$

**Remark 7.14 (Convergence of the mixed strategy profiles).** We only mention briefly that the experimental regret testing procedure can be extended to obtain an uncoupled strategy such that the mixed strategy profiles converge, with probability 1, to the set of Nash equilibria for almost all games. Note that we claim convergence not only of the empirical frequencies of plays but also of the actual mixed strategy profiles. Moreover, we claim
convergence to $\mathcal{N}$ and to the convex hull $\text{co}(\mathcal{N}_\varepsilon)$ of all $\varepsilon$-Nash equilibria for a fixed $\varepsilon$. The basic idea is to “anneal” experimental regret testing such that first it is used with some parameters $(T_1, \rho_1, \lambda_1)$ for a number $M_1$ of periods of length $T_1$ and then the parameters are changed to $(T_2, \rho_2, \lambda_2)$ (by increasing $T$ and decreasing $\rho$ and $\lambda$ properly) and experimental regret testing is used for a number $M_2 \gg M_1$ of periods (of length $T_2$), and so on. However, this is not sufficient to guarantee almost sure convergence, because at each change of parameters the process is reinitialized and therefore there is an infinite set of indices $t$ such that $\sigma_t$ is far away from any Nash equilibrium. A possible solution is based on “localizing” the search after each change of parameters such that each player limits its choice to a small neighborhood of the mixed strategy played right before the change of parameters (unless a player experiences a large regret in which case the search is extended again to the whole simplex). Another challenge one must face in designing a genuinely uncoupled procedure is that the values of the parameters of the procedure (i.e., $T_\ell, \rho_\ell, \lambda_\ell, M_\ell, \ell = 1, 2, \ldots$) cannot depend on the parameters of the game, because by requiring uncoupledness we must assume that the players only know their payoff function but not those of the other players. We leave the details as an exercise.

Remark 7.15 (Nongeneric games). All results of this section up to this point hold for almost every game. The reason for this restriction is that our proofs require an assumption of genericity of the game. We do not know whether Theorems 7.8 and 7.9 extend to all games. However, by a simple trick one can modify experimental regret testing such that the results of these two theorems hold for all games. The idea is that before starting to play, each player slightly perturbs the values of his loss function and then plays as if his losses were the perturbed values. For example, define, for each player $k$ and pure strategy profile $i$,

$$\tilde{\ell}^{(k)}(i) = \ell^{(k)}(i) + Z_{i, s},$$

where the $Z_{i, s}$ are i.i.d. random variables uniformly distributed in the interval $[-\varepsilon, \varepsilon]$. Clearly, the perturbed game is generic with probability 1. Therefore, if all players play according to experimental regret testing but on the basis of the perturbed losses, then Theorems 7.8 and 7.9 are valid for this newly generated game. However, because for all $k$ and $i$ we have $|\tilde{\ell}^{(k)}(i) - \ell^{(k)}(i)| < \varepsilon$, every $\varepsilon$-Nash equilibrium of the perturbed game is a $2\varepsilon$-Nash equilibrium of the original game.

Finally, we show how experimental regret testing can be modified so that it can be played in the model of unknown games with similar performance guarantees. In order to adjust the procedure, recall that the only place in which the players look at the past is when they calculate the regrets

$$r_{m, i_k}^{(k)} = \frac{1}{T} \sum_{s=(m-1)T+1}^{mT} \ell^{(k)}(\mathbf{i}_s) - \frac{1}{T} \sum_{s=(m-1)T+1}^{mT} \ell^{(k)}(\mathbf{i}_s, i_k).$$

However, each player may estimate his regret in a simple way. Observe that the first term in the definition of $r_{m, i_k}^{(k)}$ is just the average loss player $k$ over the $m$th period, which is available to the player, and does not need to be estimated. However, the second term is the average loss suffered by the player if he had chosen to play action $i_k$ all the time during this period. This can be estimated by random sampling. The idea is that, at each time instant, player...
k flips a biased coin and, if the outcome is head (the probability of which is very small),
then instead of choosing an action according to the mixed strategy \( \pi \) the player chooses
one uniformly at random. At these time instants, the player collects sufficient information
to estimate the regret with respect to each fixed action \( i_k \).

To formalize this idea, consider a period between times \((m-1)T + 1\) and \(mT\). During
this period, player \( k \) draws \( n_k \) samples for each \( i_k = 1, \ldots, N_k \) actions, where \( n_k \ll T \)
is to be determined later. Formally, define the random variables \( U_{k,s} \in \{0, 1, \ldots, N_k\} \),
where, for \( s \) between \((m-1)T + 1\) and \( mT\), for each \( i_k = 1, \ldots, N_k \), there are exactly
\( n_k \) values of \( s \) such that \( U_{k,s} = i_k \), and all such configurations are equally probable; for
the remaining \( s \), \( U_{k,s} = 0 \). (In other words, for each \( i_k = 1, \ldots, N_k \), \( n_k \) values of \( s \) are
chosen randomly, without replacement, such that these values are disjoint for different
\( i_k \)'s.) Then, at time \( s \), player \( k \) draws an action \( i_s^{(k)} \) as follows: conditionally on the past up
to time \( s-1 \),

\[
I_s^{(k)} = \begin{cases} 
\text{is distributed as } \pi^{i_{m-1}} & \text{if } U_{k,s} = 0 \\
\text{equals } i_k & \text{if } U_{k,s} = i_k.
\end{cases}
\]

The regret \( r_{m,i_k}^{(k)} \) may be estimated by

\[
r_{m,i_k}^{(k)} = \frac{1}{T - N_k n_k} \sum_{s=(m-1)T+1}^{mT} \ell^{(k)}(I_s^{(k)})I_{[U_{k,s}=0]} - \frac{1}{n_k} \sum_{s=(m-1)T+1}^{mT} \ell^{(k)}(I_s^{(k)}, i_k)I_{[U_{k,s}=i_k]}
\]

\[k = 1, \ldots, N_k.\] The first term of the definition of \( r_{m,i_k}^{(k)} \) is just the average of the losses
of player \( k \) over those periods in which the player does not “experiment,” that is, when
\( U_{k,s} = 0 \). (Note that there are exactly \( T - N_k n_k \) such periods.) Since \( N_k n_k \ll T \), this
average should be close to the first term in the definition of the average regret \( r_{m,i_k}^{(k)} \).
The second term is the average over those time periods in which player \( k \) experiments, and
he plays action \( i_k \) (i.e., when \( U_{k,s} = i_k \)). This may be considered as an estimate, obtained
by sampling without replacement, of the second term in the definition of \( r_{m,i_k}^{(k)} \). Observe
that \( r_{m,i_k}^{(k)} \) only depends on the past payoffs experienced by player \( k \), and therefore these
estimates are feasible in the unknown game model.

In order to show that the estimated regrets work in this case, we only need to establish
that the probability that the estimated regret exceeds \( \rho \) is small if the expected regret is not
more than \( \varepsilon \) (whenever \( \varepsilon < \rho \)). This is done in the following lemma. It guarantees that if
the experimental regret-testing procedure is run using the regret estimates described above,
then results analogous to Theorems 7.8 and 7.9 may be obtained, in a straightforward way,
in the unknown-game model.

**Lemma 7.6.** Assume that in a certain period of length \( T \), the expected regret
\( \mathbb{E}[r_{m,i_k}^{(k)} | I_1, \ldots, I_{mT}] \) of player \( k \) is at most \( \varepsilon \). Then, for a sufficiently small \( \varepsilon \), with the choice of parameters of Theorem 7.8,

\[
\mathbb{P}[r_{m,i_k}^{(k)} \geq \rho] \leq cT^{-1/3} + \exp \left( -T^{1/3} (\rho - \varepsilon)^2 \right).
\]

**Proof.** We show that, with large probability, \( r_{m,i_k}^{(k)} \) is close to \( i_s^{(k)} \). To this end, first we
compare the first terms in the expression of both. Observe that at those periods \( s \) of time
when none of the players experiments (i.e., when \( U_{k,s} = 0 \) for all \( k = 1, \ldots, K \), the
corresponding terms of both estimates are equal. Thus, by a simple algebra it is easy to see that the first terms differ by at most \( \frac{2}{T} \sum_{k=1}^{K} N_k n_k \).

It remains to compare the second terms in the expressions of \( \hat{r}_{m,i_k}^{(k)} \) and \( r_{m,i_k}^{(k)} \). Observe that if there is no time instant \( s \) for which \( U_{k,s} = 1 \) and \( U_{k',s} = 1 \) for some \( k' \neq k \), then

\[
\frac{1}{n_k} \sum_{s=t+1}^{t+T} \ell(k) (I_s, i_k) \mathbb{I}_{[U_{k,s}=i_k]}
\]

is an unbiased estimate of

\[
\frac{1}{T} \sum_{s=t+1}^{t+T} \ell(k) (I_s, i_k)
\]

obtained by random sampling. The probability that no two players sample at the same time is at most

\[
TK^2 \max_{k,k' \leq k} \frac{N_k n_k}{T} \frac{N_k n_{k'}}{T},
\]

where we used the union-of-events bound over all pairs of players and all \( T \) time instants. By Hoeffding’s inequality for an average of a sample taken without replacement (see Lemma A.2), we have

\[
\mathbb{P} \left[ \left| \frac{1}{n_k} \sum_{s=t+1}^{t+T} \ell(k) (I_s, i_k) \mathbb{I}_{[U_{k,s}=i_k]} - \frac{1}{T} \sum_{s=t+1}^{t+T} \ell(k) (I_s, i_k) \right| > \alpha \right] \leq e^{-2n_k \alpha^2},
\]

where \( \mathbb{P} \) denotes the distribution induced by the random variables \( U_{k,s} \). Putting everything together,

\[
\mathbb{P} \left[ \hat{r}_{m,i_k}^{(k)} \geq \rho \right] \leq TK^2 \max_{k,k' \leq k} \frac{N_k n_k}{T} \frac{N_k n_{k'}}{T} + \exp \left( -2n_k \left( \rho - \varepsilon - 2 \sum_{k=1}^{K} \frac{N_k n_k}{T} \right)^2 \right).
\]

Choosing \( n_k \sim T^{1/3} \), the first term on the right-hand side is of order \( T^{-1/3} \) and \( \frac{1}{T} \sum_{k=1}^{K} N_k n_k = O(T^{-2/3}) \) becomes negligible compared with \( \rho - \varepsilon \). ■

### 7.11 Playing Against Opponents That React

Regret-minimizing strategies, such as those discussed in Sections 4.2 and 4.3, set up the goal of predicting as well as the best constant strategy in hindsight, assuming that the actions of the opponents would have been the same had the forecaster been following that constant strategy. However, when a forecasting strategy is used to play a repeated game, the actions prescribed by the forecasting strategy may have an effect on the behavior of the opponents, and so measuring regret as the difference of the suffered cumulative loss and that of the best constant action in hindsight may be very misleading.
To simplify the setup of the problem, we consider playing a two-player game such that, at time $t$, the row player takes an action $I_t \in \{1, \ldots, N\}$ and the column player takes action $J_t \in \{1, \ldots, M\}$. The loss suffered by the row player at time $t$ is $\ell(I_t, J_t)$. (The loss of the column player is immaterial in this section. Note also that since we are only concerned with the loss of the first player, there is no loss of generality in assuming that there are only two players, since otherwise $J_t$ can represent the joint play of all other players.) In the language of Chapter 4, we consider the case of a nonoblivious opponent; that is, the actions of the column player (the opponent) may depend on the history $I_1, \ldots, I_{t-1}$ of past moves of the row player.

To illustrate why regret-minimizing strategies may fail miserably in such a scenario, consider the repeated play of a prisoners’ dilemma, that is, a $2 \times 2$ game in which the loss matrix of the row player is given by

<table>
<thead>
<tr>
<th></th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>1/3</td>
<td>1</td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>2/3</td>
</tr>
</tbody>
</table>

In the usual definition of the prisoners’ dilemma, the column player has the same loss matrix as the row player. In this game both players can either cooperate (“c”) or defect (“d”). Regardless of what the column player does, the row player is better off defecting (and the same goes for the column player). However, it is better for the players if they both cooperate than if they both defect.

Now assume that the game is played repeatedly and the row player plays according to a Hannan-consistent strategy; that is, the normalized cumulative loss $\frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t)$ approaches $\min_{i=c,d} \frac{1}{n} \sum_{t=1}^{n} \ell(i, J_t)$. Clearly, the minimum is achieved by action “d” and therefore the row player will defect basically all the time. In a certain worst-case sense this may be the best one can hope for. However, in many realistic situations, depending on the behavior of the adversary, significantly smaller losses can be achieved. For example, the column player may be willing to try cooperation. Perhaps the simplest such strategy of the opponent is “tit for tat,” in which the opponent repeats the row player’s previous action. In such a case, by playing a Hannan consistent strategy, the row player’s performance is much worse than what he could have achieved by following the expert “c” (which is the worse action in the sense of the notions of regret we have used so far).

The purpose of this section is to introduce forecasting strategies that avoid falling in traps similar to the one described above under certain assumptions on the opponent’s behavior.

To this end, consider the scenario where, rather than requiring Hannan consistency, the goal of the forecaster is to achieve a cumulative loss (almost) as small as that of the best action, where the cumulative loss of each action is calculated by looking at what would have happened if that action had been followed throughout the whole repeated game.

It is obvious that a completely malicious adversary can make it impossible to estimate what would have happened if a certain action had been played all the time (unless that action is played all the time). But under certain natural assumptions on the behavior of the adversary, such an inference is possible. The assumptions under which our goal can be reached require a kind of “stationarity” and bounded memory of the opponent and are certainly satisfied for simple strategies such as tit for tat.
Remark 7.16 (Hannan consistent strategies are sometimes better). We have argued that in some cases it makes more sense to look for strategies that perform as well as the best action if that action had been played all the time rather than playing Hannan consistent strategies. The repeated prisoners’ dilemma with the adversary playing tit for tat is a clear example. However, in some other cases Hannan consistent strategies may perform much better than the best action in this new sense. The following example describes such a situation: assume that the row player has \( N = 2 \) actions, and let \( n \) be even such that \( n/2 \) is odd. Assume that in the first \( n/2 \) time periods the losses of both actions are 1 in each period. After \( n/2 \) periods the adversary decides to assign losses 00...00 (\( n/2 \) times) to the action that was played less times during the first \( n/2 \) rounds and 11...11 to the other action. Clearly, a Hannan consistent strategy has a cumulative loss of about \( n/2 \) during the \( n \) periods of the game. On the other hand, if any of the two actions is played constantly, its cumulative loss is \( n \).

The goal of this section is to design strategies that guarantee, under certain assumptions on the behavior of the column player, that the average loss \( \frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t) \) is not much larger than \( \min_{i=1,\ldots,N} \mu_{i,n} \), where \( \mu_{i,n} \) is the average loss of a hypothetical player who plays the same action \( I_t = i \) in each round of the game.

A key ingredient of the argument is a different way of measuring regret. The goal of the forecaster in this new setup is to achieve, during the \( n \) periods of play, an average loss almost as small as the average loss of the best action, where the average is computed over only those periods in which the action was chosen by the forecaster. To make the definition formal, denote by

\[
\hat{\mu}_t = \frac{1}{t} \sum_{s=1}^{t} \ell(I_s, J_s)
\]

the averaged cumulative loss of the forecaster at time \( t \) and by

\[
\mu_{i,t} = \frac{\sum_{s=1}^{t} \ell(I_s, J_s) \mathbb{I}_{[I_s=i]}}{\sum_{s=1}^{t} \mathbb{I}_{[I_s=i]}}
\]

the averaged cumulative loss of action \( i \), averaged over the time periods in which the action was played by the forecaster. If \( \sum_{s=1}^{t} \mathbb{I}_{[I_s=i]} = 0 \), let \( \mu_{i,t} \) take the maximal value 1. At this point it may not be entirely clear how the averaged losses \( \mu_{i,t} \) are related to the average loss of a player who plays the same action \( i \) all the time. However, shortly it will become clear that these quantities can be related under some assumptions of the behavior of the opponent and certain restrictions on the forecasting strategy.

The property that the forecaster needs to satisfy for our purposes is that, asymptotically, the average loss \( \hat{\mu}_n \) is not larger than the smallest asymptotic average loss \( \mu_{i,n} \). More precisely, we need to construct a forecaster that achieves

\[
\lim \sup_{n \to \infty} \hat{\mu}_n \leq \min_{i=1,\ldots,N} \lim \sup_{n \to \infty} \mu_{i,n}.
\]

Surprisingly, there exists a deterministic forecaster that satisfies this asymptotic inequality regardless of the opponent’s behavior. Here we describe such a strategy for the case of \( N = 2 \) actions. The simple extension to the general case of more than two actions is left as an exercise (Exercise 7.30). Consider the following simple deterministic forecaster.
DETERMINISTIC EXPLORATION–EXPLOITATION

For each round \( t = 1, 2, \ldots \)

(1) \((\text{Exploration})\) if \( t = k^2 \) for an integer \( k \), then set \( I_t = 1 \); if \( t = k^2 + 1 \) for an integer \( k \), then set \( I_t = 2 \);

(2) \((\text{Exploitation})\) otherwise, let \( I_t = \arg\min_{i=1,2} \mu_{i,t-1} \) (in case of a tie, break it, say, in favor of action 1).

This simple forecaster is a version of fictitious play, based on the averaged losses, in which the exploration step simply guarantees that every action is sampled infinitely often. There is nothing special about the time instances of the form \( t = k^2, k^2 + 1 \); any sparse infinite sequence would do the job. In fact, the original algorithm of de Farias and Megiddo [85] chooses the exploration steps randomly.

Observe, in passing, that this is a “bandit”-type predictor in the sense that it only needs to observe the losses of the played actions.

**Theorem 7.10.** Regardless of the sequence of outcomes \( J_1, J_2, \ldots \) the deterministic forecaster defined above satisfies

\[
\limsup_{n \to \infty} \hat{\mu}_n \leq \min_{i=1,2} \limsup_{n \to \infty} \mu_{i,n}.
\]

**Proof.** For each \( t = 1, 2, \ldots \), let \( i^*_t = \arg\min_{i=1,2} \mu_{i,t} \) and let \( t_1, t_2, \ldots \) be the time instances such that \( i^*_t \neq i^*_{t-1} \), that is, the “leader” is switched. If there is only a finite number of such \( t_k \)’s, then, obviously,

\[
\hat{\mu}_n - \min_{i=1,2} \mu_{i,n} \to 0,
\]

which implies the stated inequality. Thus, we may assume that there is an infinite number of switches and it suffices to show that whenever \( T = \max\{t_k : t_k \leq n\} \), then either

\[
\hat{\mu}_n - \min_{i=1,2} \mu_{i,T} \leq \frac{\text{const.}}{T^{1/4}}
\]

or

\[
\hat{\mu}_n - \min_{i=1,2} \mu_{i,n} \leq \frac{\text{const.}}{T^{1/4}},
\]

which implies the statement.

First observe that, due to the exploration step, for any \( t \geq 3 \) and \( i = 1, 2 \),

\[
\sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}} \geq \left\lfloor \sqrt{t-1} \right\rfloor \geq \sqrt{t}/2.
\]

But then

\[
|\mu_{1,T} - \mu_{2,T}| \leq \frac{2}{\sqrt{T}}.
\]

This inequality holds because by the boundedness of the loss, at time \( T \), the averaged loss of action \( i \) can change by at most \( 1/\sum_{s=1}^{T} \mathbb{1}_{\{I_s = i\}} \leq 2/\sqrt{T} \), and the definition of the switch is that the one that was larger in the previous step becomes smaller, which is only possible if the averaged losses of the two actions were already \( 2/\sqrt{T} \) close to each other. But then
the averaged loss of the forecaster at the time $T$ (the last switch before time $n$) may be bounded by
\[
\hat{\mu}_T = \frac{1}{T} \left( \sum_{s=1}^{T} \ell(I_s, J_s) \mathbb{1}_{[i_s=1]} + \sum_{s=1}^{T} \ell(I_s, J_s) \mathbb{1}_{[i_s=2]} \right)
\]
\[
= \frac{1}{T} \left( \mu_{1,T} \sum_{s=1}^{T} \mathbb{1}_{[i_s=1]} + \mu_{2,T} \sum_{s=1}^{T} \mathbb{1}_{[i_s=2]} \right)
\]
\[
\leq \min_{i=1,2} \mu_{i,T} + \frac{2}{\sqrt{T}}.
\]
Now assume that $T$ is so large that $n - T \leq T^{3/4}$. Then clearly, $|\hat{\mu}_n - \hat{\mu}_T| \leq T^{3/4}/n \leq T^{-1/4}$ and (7.5) holds.

Thus, in the rest of the proof we assume that $n - T \geq T^{3/4}$. It remains to show that $\hat{\mu}_n$ cannot be much larger than $\min_{i=1,2} \mu_{i,n}$. Introduce the notation
\[
\delta = \hat{\mu}_n - \min_{i=1,2} \mu_{i,T}.
\]
Since
\[
\hat{\mu}_n = \frac{1}{n} \left( \sum_{t=1}^{T} \ell(I_t, J_t) + \sum_{t=T+1}^{n} \ell(I_t, J_t) \right)
\]
\[
\leq \frac{1}{n} \left( T \min_{i=1,2} \mu_{i,T} + 2\sqrt{T} + \sum_{t=T+1}^{n} \ell(I_t, J_t) \right)
\]
we have
\[
\sum_{t=T+1}^{n} \ell(I_t, J_t) \geq (n - T) \min_{i=1,2} \mu_{i,T} + \delta n - 2\sqrt{T}.
\]
Since, apart from at most $\sqrt{n - T}$ exploration steps, the same action is played between times $T + 1$ and $n$, we have
\[
\min_{i=1,2} \mu_{i,n} \geq \frac{\mu_{i,T}}{n} \sum_{s=1}^{T} \mathbb{1}_{[i_s=i_s^*]} + \sum_{t=T+1}^{n} \ell(I_t, J_t) - \sqrt{n - T}
\]
\[
\geq \frac{\mu_{i,T}}{n} \sum_{s=1}^{T} \mathbb{1}_{[i_s=i_s^*]} + \sum_{t=T+1}^{n} \ell(I_t, J_t) - \sqrt{n - T}
\]
\[
\geq \frac{\mu_{i,T}}{n} \sum_{s=1}^{T} \mathbb{1}_{[i_s=i_s^*]} + \sum_{t=T+1}^{n} \ell(I_t, J_t) - \sqrt{n - T}
\]
\[
= \frac{\mu_{i,T}}{n} \left( \sum_{s=1}^{T} \mathbb{1}_{[i_s=i_s^*]} + \sum_{t=T+1}^{n} \mathbb{1}_{[i_t=i_t^*]} \right) + \delta n - 2\sqrt{T} - \sqrt{n - T}
\]
\[
\geq \mu_{i,T} + \delta - 2\sqrt{T} \frac{\sqrt{n - T}}{n - T} - \frac{1}{\sqrt{n - T}}
\]
\[
\geq \mu_{i,T} + \delta - 3T^{-1/4}
\]
\[
= \hat{\mu}_n - 3T^{-1/4},
\]
where at the last inequality we used $n - T \geq T^{3/4}$. Thus, (7.6) holds in this case. \[\square\]
There is one more ingredient we need in order to establish strategies of the desired behavior. As we have mentioned before, our aim is to design strategies that perform well if the behavior of the opponent is such that the row player can estimate, for each action, the average loss suffered by playing that action all the time. In order to do this, we modify the forecaster studied above such that whenever an action is chosen, it is played repeatedly sufficiently many times in a row so that the forecaster gets a good picture of the behavior of the opponent when that action is played. This modification is done trivially by simply repeating each action $\tau$ times, where the positive integer $\tau$ is a parameter of the strategy.

**REPEATED DETERMINISTIC EXPLORATION–EXPLOITATION**

**Parameter:** Number of repetitions $\tau$.

For each round $t = 1, 2, \ldots$

1. *(Exploration)* if $t = k^2 \tau + s$ for integers $k$ and $s = 0, 1, \ldots, \tau - 1$, then set $I_t = 1$;
   
   if $t = (k^2 + 1) \tau + s$ for integers $k$ and $s = 0, 1, \ldots, \tau - 1$, then set $I_t = 2$;

2. *(Exploitation)* otherwise, let $I_t = \arg\min_{i=1,2} \mu_i \cdot \tau \lfloor t/\tau \rfloor - 1$ (in case of a tie, break it, say, in favor of action 1).

Theorem 7.10 (as well as Exercise 7.30) trivially extends to this case and the strategy defined above obviously satisfies

$$
\lim_{n \to \infty} \tilde{\mu}_n \leq \min_{i=1,2} \lim_{n \to \infty} \mu_{i,n}
$$

regardless of the opponent’s actions and the parameter $\tau$.

Our main assumption on the opponent’s behavior is that, for every action $i$, there exists a number $\mu_i \in [0, 1]$ such that for any time instance $t$ and past plays $I_1, \ldots, I_t$,

$$
\frac{1}{\tau} \sum_{s=t+1}^{t+\tau} \ell(i, J_s) - \overline{\mu}_i \leq \varepsilon_{\tau},
$$

where $\varepsilon_{\tau}$ is a sequence of nonnegative numbers converging to 0 as $\tau \to \infty$. (Here the average loss is computed by assuming that the row player’s moves are $I_1, \ldots, I_t, i, i, \ldots, i$.) After de Farias and Megiddo [86], we call an opponent satisfying this condition flexible. Clearly, if the opponent is flexible, then for any action $i$ the average loss of playing the action forever is at most $\overline{\mu}_i$. Moreover, the performance bound for the repeated deterministic exploration–exploitation immediately implies the following.

**Corollary 7.3.** Assume that the row player plays according to the repeated deterministic exploration–exploitation strategy with parameter $\tau$ against a flexible opponent. Then the asymptotic average cumulative loss of the row player satisfies

$$
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \ell(I_i, J_i) \leq \min_{i=1,\ldots,N} \overline{\mu}_i + \varepsilon_{\tau}.
$$
The assumption of flexibility is satisfied in many cases when the opponent’s long-term behavior against any fixed action can be estimated by playing the action repeatedly for a stretch of time of length $\tau$. This is satisfied, for example, when the opponent is modeled by a finite automata. In the example of the opponent playing tit for tat in the prisoners’ dilemma described at the beginning of this section, the opponent is clearly flexible with $\varepsilon_\tau = 1/\tau$. Note that in these cases one actually has

$$
\left| \frac{1}{\tau} \sum_{s=t+1}^{t+\tau} \ell(i, J_s) - \bar{\mu}_i \right| \leq \varepsilon_\tau
$$

(with $\bar{\mu}_1 = 1/3$ and $\bar{\mu}_2 = 2/3$); that is, the estimated average losses are actually close to the asymptotic performance of the corresponding action. However, for Corollary 7.3 it suffices to require the one-sided inequality.

Corollary 7.3 states the existence of a strategy of playing repeated games such that, against any flexible opponent, the average loss is at most that of the best action (calculated by assuming that the action is played constantly) plus the quantity $\varepsilon_\tau$ that can be made arbitrarily small by choosing the parameter $\tau$ of the algorithm sufficiently large. However, sequence $\varepsilon_\tau$ depends on the opponent and may not be known to the forecaster. Thus, it is desirable to find a forecaster whose average loss actually achieves $\min_{i=1,\ldots,N} \bar{\mu}_i$ asymptotically. Such a method may now easily be constructed.

**Corollary 7.4.** There exists a forecaster such that whenever the opponent is flexible,

$$
\limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \ell(I_t, J_t) \leq \min_{i=1,\ldots,N} \bar{\mu}_i.
$$

We leave the details as a routine exercise (Exercise 7.32).

**Remark 7.17 (Randomized opponents).** In some cases it may be meaningful to consider strategies for the adversary that use randomization. In such cases our definition of flexibility, which poses a deterministic condition on the opponent, is not realistic. However, the definition may be easily modified to accommodate a possibly randomized behavior. In fact, the original definition of de Farias and Megiddo [86] involves a probabilistic assumption.

### 7.12 Bibliographic Remarks

Playing and learning in repeated games is an important branch of game theory with an extensive literature. In this chapter we addressed only a tiny corner of this immense subject. The interested reader may consult the monographs of Fudenberg and Levine [119], Sorin [276], and Young [316]. Hart [144] gives an excellent survey of regret-based uncoupled learning dynamics.

von Neumann’s minimax theorem is the classic result of game theory (see von Neumann and Morgenstern [296]), and most standard textbooks on game theory provide a proof. Various generalizations, including stronger versions of Theorem 7.1, are due to Fan [93] and Sion [271] (see also the references therein). The proof of Theorem 7.1 shown here is a
generalization of ideas of Freund and Schapire [114], who prove von Neumann’s minimax theorem using the strategy described in Exercise 7.9.

The notion and the proof of existence of Nash equilibria appears in the celebrated paper of Nash [222]. For the basic results on the Nash convergence of fictitious play, see Robinson [246], Miyasawa [218], Shapley [265], Monderer and Shapley [220]. Hofbauer and Sandholm [163] consider stochastic fictitious play, similar, in spirit, to the follow-the-perturbed-leader forecaster considered in Chapter 4, and prove its convergence for a class of games. See the references within [163] for various related results. Singh, Kearns, and Mansour [270] show that a simple dynamics based on gradient-descent yields average payoffs asymptotically equivalent of those of a Nash equilibrium in the special case of two-player games in which both players have two actions.

The notion of correlated equilibrium was first introduced by Aumann [16, 17]. A direct proof of the existence of correlated equilibria, using just von Neumann’s minimax theorem (as opposed to the fixed point theorem needed to prove the existence of Nash equilibria) was given by Hart and Schmeidler [150]. The existence of adaptive procedures leading to a correlated equilibrium was shown by Foster and Vohra [105]; see also Fudenberg and Levine [118, 121] and Hart and Mas-Colell [145, 146]. Stoltz and Lugosi [278] generalize this to games with an infinite, but compact, set of actions. The connection of calibration and correlated equilibria, described in Section 7.6, was pointed out by Foster and Vohra [105]. Kakade and Foster [171] take these ideas further and show that if all players play according to a best response to a certain common, “almost deterministic,” well-calibrated forecaster, then the joint empirical frequencies of play converge not only to the set of correlated equilibria but, in fact, to the convex hull of the set of Nash equilibria. Hart and Mas-Colell [145] introduce a strategy, the so-called regret matching, conceptually much simpler than the internal regret minimization procedures described in Section 4.4, which has the property that if all players follow this strategy, the joint empirical frequencies converge to the set of correlated equilibria; see also Cahn [44]. Kakade, Kearns, Langford, and Ortiz [172] consider efficient algorithms for computing correlated equilibria in graphical games. The result of Section 7.5 appears in Hart and Mas-Colell [147].

Blackwell’s approachability theory dates back to [28], where Theorem 7.5 is proved. It was also Blackwell [29] who pointed out that the approachability theorem may be used to construct Hannan-consistent forecasting strategies. Various generalizations of this theorem may be found in Vielle [295] and Lehrer [193]. Fabian and Hannan [92] studied rates of convergence in an extended setting in which payoffs may be random and not necessarily bounded. The potential-based strategies of Section 7.8 were introduced by Hart and Mas-Colell [146] and Theorem 7.6 is due to them. In [146] the result is stated under a weaker assumption than convexity of the potential function.

The problem of learning Nash equilibria by uncoupled strategies has been pursued by Foster and Young [108, 109]. They introduce the idea of regret testing, which the procedures studied in Section 7.10 are based on. Their procedures guarantee that, asymptotically, the mixed strategy profiles are within distance \( \epsilon \) of the set of Nash equilibria in a fraction of at least \( 1 - \epsilon \) of time. On the negative side, Hart and Mas-Colell [148, 149] show that it is impossible to achieve convergence to Nash equilibrium for all games if one is restricted to use stationary strategies that have bounded memory. By “bounded memory” they mean that there is a finite integer \( T \) such that each player bases his play only on the last \( T \) rounds of play. On the other hand, for every \( \epsilon > 0 \) they show a randomized bounded-memory stationary uncoupled procedure, different from those presented in Sections 7.9 and 7.10,
for which the joint empirical frequencies of play converge almost surely to an $\varepsilon$-Nash equilibrium. Germano and Lugosi [126] modify the regret testing procedure of Foster and Young to achieve almost sure convergence to the set of $\varepsilon$-Nash equilibria for all games. The analysis of Section 7.10 is based on [126]. In particular, the proof of Lemma 7.5 is found in [126], though the somewhat simpler case of two players is shown in Foster and Young [109].

A closely related branch of literature that is not discussed in this chapter is based on learning rules that are based on players updating their beliefs using Bayes’ rule. Kalai and Lehrer [176] show that if the priors “contain a grain of truth,” the play converges to a Nash equilibrium of the game. See also Jordan [169, 170], Dekel, Fudenberg, and Levine [87], Fudenberg and Levine [117, 119], and Nachbar [221].

Kalai, Lehrer, and Smorodinsky [177] show that this type of learning is closely related to stronger notions of calibration and merging. See also Lehrer, and Smorodinsky [196], Sandroni and Smorodinsky [258].

The material presented in Section 7.11 is based on the work of de Farias and Megiddo [85, 86], though the analysis shown here is different. In particular, the forecaster of de Farias and Megiddo is randomized and conceptually simpler than the deterministic predictor used here.

7.13 Exercises

7.1 Show that the set of all Nash equilibria of a two-person zero-sum game is closed and convex.

7.2 (Shapley’s game) Consider the two-person game described by the loss matrices of the two players (“$R$” and “$C$”), known as Shapley’s game:

\[
\begin{array}{ccc}
R \backslash C & 1 & 2 & 3 \\
1 & 0 & 1 & 1 \\
2 & 1 & 0 & 1 \\
3 & 1 & 1 & 0 \\
\end{array}
\]

Show that if both players use fictitious play, the empirical frequencies of play do not converge to the set of correlated equilibria (Foster and Vohra [105]).

7.3 Prove Lemma 7.1.

7.4 Consider the two-person game given by the losses

\[
\begin{array}{ccc}
R \backslash C & 1 & 2 \\
1 & 1 & 5 \\
2 & 0 & 7 \\
\end{array}
\]

Find all three Nash equilibria of the game. Show that the distribution given by $P(1, 1) = 1/3$, $P(1, 2) = 1/3$, $P(2, 1) = 1/3$, $P(2, 2) = 0$ is a correlated equilibrium that lies outside of the convex hull of the Nash equilibria. (Aumann [17]).

7.5 Show that a probability distribution $P$ over $\bigotimes_{k=1}^{K} \{ 1, \ldots, N_k \}$ is a correlated equilibrium if and only if for all $k = 1, \ldots, K$,

\[
\mathbb{E} \ell^k(I) \leq \mathbb{E} \ell^k(I^-),
\]

where $I = (I^{(1)}, \ldots, I^{(K)})$ is distributed according to $P$ and the random variable $\tilde{I}^k$ is any function of $I^{(k)}$ and of a random variable $U$ independent of $I$.

7.6 Consider the repeated time-varying game described in Remark 7.3, with $N = M = 2$. Assume that there exist positive numbers $\varepsilon, \delta$ such that, for every sufficiently large $n$, at least for $n\delta$ time
steps $t$ between time 1 and $n$,

$$\max_{j=1,2} |\ell_t(1, j) - \ell_t(2, j)| > \epsilon.$$ 

Show that then for any sequence of mixed strategies $p_1, p_2, \ldots$ of the row player, the column player can choose his mixed strategies $q_1, q_2, \ldots$ such that the row player’s cumulative loss satisfies

$$\sum_{t=1}^n \ell_t(I_t, J_t) - \sum_{t=1}^n \min_{i=1,2} \ell_t(i, J_t) > \gamma n$$

for all sufficiently large $n$ with probability 1, where $\gamma$ is positive.

7.7 Assume that in a two-person zero-sum game, for all $t$, the row player plays according to the constant mixed strategy $p_t = p$, where $p$ is any mixed strategy for which there exists a mixed strategy of the column player such that $\ell(p, q) = V$. Show that

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{t=1}^n \ell(I_t, J_t) \leq V.$$ 

Show also that, for any $\epsilon > 0$, the row player, regardless of how he plays, cannot guarantee that

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{t=1}^n \ell(I_t, J_t) \leq V - \epsilon.$$ 

7.8 Consider a two-person zero-sum game and assume that the row player plays according to the exponentially weighted average mixed strategy

$$p_{i,t} = \exp \left( -\eta \sum_{s=1}^{t-1} \ell(i, q_s) \right) \exp \left( -\eta \sum_{s=1}^{t-1} \ell(k, q_s) \right), \quad i = 1, \ldots, N.$$ 

Show that, with probability at least $1 - \delta$, the average loss of the row player satisfies

$$\frac{1}{n} \sum_{t=1}^n \ell(I_t, J_t) \leq V + \frac{\ln N}{n\eta} + \frac{\eta}{8} + \sqrt{\frac{2}{n} \ln \frac{2N}{\delta}}.$$ 

7.9 Freund and Schapire [113] investigate the weighted average forecaster in the simplified version of the setup of Section 7.3, in which the row player gets to see the distribution $q_{t-1}$ chosen by the column player before making the play at time $t$. Then the following version of the weighted average strategy for the row player is feasible:

$$p_{i,t} = \frac{\exp \left( -\eta \sum_{s=1}^{t-1} \ell(i, q_s) \right)}{\sum_{k=1}^N \exp \left( -\eta \sum_{s=1}^{t-1} \ell(k, q_s) \right)}, \quad i = 1, \ldots, N,$$

with $p_{i,1}$ set to $1/N$, where $\eta > 0$ is an appropriately chosen constant. Show that this strategy is an instance of the weighted average forecaster (see Section 4.2), which implies that

$$\sum_{t=1}^n \bar{\ell}(p, q) \leq \min_p \sum_{t=1}^n \bar{\ell}(p, q) + \frac{\ln N}{\eta} + \frac{n\eta}{8},$$

where

$$\bar{\ell}(p, q) = \sum_{i=1}^N \sum_{j=1}^M p_{i,t} q_{j,t} \ell(i, j).$$
Show that if \( I_t \) denotes the actual randomized play of the row player, then with an appropriately chosen \( \eta = \eta_t \),

\[
\lim_{n \to \infty} \frac{1}{n} \left( \sum_{t=1}^{n} \ell(I_t, q_t) - \min_{i=1, \ldots, N} \sum_{t=1}^{n} \ell(i, q_t) \right) = 0 \quad \text{almost surely}
\]

(Freund and Schapire [113]).

7.10 Improve the bound of the previous exercise to

\[
\sum_{t=1}^{n} \ell(p_t, q_t) \leq \min_{p} \left( \sum_{t=1}^{n} \ell(p, q_t) - \frac{H(p)}{\eta} + \frac{\ln N}{\eta} + \frac{n\eta}{8} \right),
\]

where \( H(p) = -\sum_{i=1}^{N} p_i \ln p_i \) denotes the entropy of the probability vector \( p = (p_1, \ldots, p_N) \). *Hint:* Improve the crude bound \( \frac{1}{\eta} \ln \left( \sum_{i} e^{R_i} \right) \geq \max_{1 \leq i \leq N} R_i \) to \( \frac{1}{\eta} \ln \left( \sum_{i} e^{R_i} \right) \geq \max_{1 \leq i \leq N} (R_i \cdot p + H(p)/\eta) \).

7.11 Consider repeated play in a two-person zero-sum game in which both players play such that

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t) = V \quad \text{almost surely.}
\]

Show that the product distribution \( \hat{p}_n \times \hat{q}_n \) with

\[
\hat{p}_{i,n} = \frac{1}{n} \sum_{t=1}^{n} I_{t,i} \quad \text{and} \quad \hat{q}_{j,n} = \frac{1}{n} \sum_{t=1}^{n} I_{t,j}
\]

converges, almost surely, to the set of Nash equilibria. *Hint:* Check the proof of Theorem 7.2.

7.12 Robinson [246] showed that if in repeated playing of a two-person zero-sum game both players play according to fictitious play (i.e., choose the best pure strategy against the average mixed strategy of their opponent), then the product of the marginal empirical frequencies of play converges to a solution of the game. Show, however, that fictitious play does not have the following robustness property similar to the exponentially weighted average strategy deduced in Theorem 7.2: if a player uses fictitious play but his opponent does not, then the player’s normalized cumulative loss may be significantly larger than the value of the game.

7.13 (Fictitious conditional regret minimization) Consider a two-person game in which the loss matrix of both players is given by

\[
\begin{array}{cc}
1 & 2 \\
1 & 0 \\
2 & 1 \\
\end{array}
\]

Show that if both players play according to fictitious play (breaking ties randomly if necessary), then Nash equilibrium is achieved in a strong sense.

Consider now the “conditional” (or “internal”) version of fictitious play in which both players \( k = 1, 2 \) select

\[
I^{(k)}_t = \arg\min_{i \in \{1, 2\}} \frac{1}{t-1} \sum_{s=1}^{t-1} \ell^{(k)}(I^{(k)}_s, i_k).
\]

Show that if the play starts with, say, \( (1, 2) \), then both players will have maximal loss in every round of the game.

7.14 Show that if in a repeated play of a \( K \)-person game all players play according to some Hannan consistent strategy, then the joint empirical frequencies of play converge to the Hannan set of the game.
7.15 Consider the two-person zero-sum game given by the loss matrix
\[
\begin{pmatrix}
0 & 0 & -1 \\
0 & 0 & 1 \\
1 & -1 & 0
\end{pmatrix}
\]
Show that the joint distribution
\[
\begin{pmatrix}
1/3 & 1/3 & 0 \\
1/3 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\]
is a correlated equilibrium of the game. This example shows that even in zero-sum games the set of correlated equilibria may be strictly larger than the set of Nash equilibria (Forges [101]).

7.16 Describe a game for which \( \mathcal{H} \setminus \mathcal{C} \neq \emptyset \), that is, the Hannan set contains some distributions that are not correlated equilibria.

7.17 Show that if \( P \in \mathcal{H} \) is a product measure, then \( P \in \mathcal{N} \). In other words, the product measures in the Hannan set are precisely the Nash equilibria.

7.18 Show that in a \( K \)-person game with \( N_k = 2 \), for all \( k = 1, \ldots, K \) (i.e., each player has two actions to choose from), \( \mathcal{H} = \mathcal{C} \).

7.19 Extend the procedure and the proof of Theorem 7.4 to the general case of \( K \)-person games.

7.20 Construct a game with two-dimensional vector-valued losses and a (nonconvex) set \( S \subset \mathbb{R}^2 \) such that all halfspaces containing \( S \) are approachable but \( S \) is not.

7.21 Construct a game with vector-valued losses and a closed and convex polytope such that if the polytope is written as a finite intersection of closed halfspaces, where the hyperplanes defining the halfspaces correspond to the faces of the polytope, then all these closed halfspaces are approachable but the polytope is not.

7.22 Use Theorem 7.5 to show that, in the setup of Section 7.4, each player has a strategy such that the limsup of the conditional regrets is nonpositive regardless of the other players’ actions.

7.23 This exercise presents a strategy that achieves a significantly faster rate of convergence in Blackwell’s approachability theorem than that obtained in the proof of the theorem in the text. Let \( S \) be a closed and convex set, and assume that all halfspaces \( H \) containing \( S \) are approachable. Define \( \bar{A}_0 = \mathbf{0} \) and \( \bar{A}_t = \frac{1}{t} \sum_{s=1}^{t} \bar{\ell}(p_s, J_s) \) for \( t \geq 1 \). Define the row player’s mixed strategy \( p_t \) at time \( t = 1, 2, \ldots \) as arbitrary if \( \bar{A}_{t-1} \in S \) and by
\[
\max_{j=1,\ldots,M} \bar{a}_{t-1} \cdot \bar{\ell}(p_t, j) \leq c_{t-1}
\]
otherwise, where
\[
\bar{a}_{t-1} = \frac{\bar{A}_{t-1} - \pi_S(\bar{A}_{t-1})}{\|\bar{A}_{t-1} - \pi_S(\bar{A}_{t-1})\|} \quad \text{and} \quad c_{t-1} = \bar{a}_{t-1} \cdot \pi_S(\bar{A}_{t-1}).
\]
Prove that there exists a universal constant \( C \) (independent of \( n \) and \( d \)) such that, with probability at least \( 1 - \delta \),
\[
\|A_n - \pi_S(A_n)\| \leq \frac{2}{\sqrt{n}} + C \sqrt{\frac{\ln(1/\delta)}{n}}.
\]
Hint: Proceed as in the proof of Theorem 7.5 to show that \( \|\bar{A}_n - \pi_S(\bar{A}_n)\| \leq 2/\sqrt{n} \). To obtain a dimension-free constant when bounding \( \|A_n - \bar{A}_n\| \), you will need an extension of the Hoeffding–Azuma inequality to vector-valued martingales; see, for example, Chen and White [58].
7.24 Consider the potential-based strategy, based on the average loss $A_{t-1}$, described at the beginning of Section 7.8. Show that, under the same conditions on $S$ and $\Phi$ as in Theorem 7.6, the average loss satisfies $\lim_{n \to \infty} d(A_n, S) = 0$ with probability 1. *Hint:* Mimic the proof of Theorem 7.6.

7.25 *(A stationary strategy to find pure Nash equilibria)* Assume that a $K$-person game has a pure action Nash equilibrium and consider the following strategy for player $k$: If $t = 1, 2$, choose $I_k(t)$ randomly. If $t > 2$, if all players have played the same action in the last two periods (i.e., $I_{t-1} = I_{t-2}$) and $I_k(t-1)$ was a best response to $I_{t-1}$, then repeat the same play, that is, define $I_k(t) = I_k(t-1)$. Otherwise, choose $I_k(t)$ uniformly at random.

Prove that if all players play according to this strategy, then a pure action Nash equilibrium is eventually achieved, almost surely. (Hart and Mas-Colell [149].)

7.26 *(Generic two-player game with a pure Nash equilibrium)* Consider a two-player game with a pure action Nash equilibrium. Assume also that the player is generic in the sense that the best reply is always unique. Suppose at time $t$ each player repeats the play of time $t-1$ if it was a best response and selects an action randomly otherwise. Prove that a pure action Nash equilibrium is eventually achieved, almost surely [149].

*Hint:* The process $I_1, I_2, \ldots$ is a Markov chain with state space $\{1, \ldots, N_1\} \times \{1, \ldots, N_2\}$. Show that given any state $i = (i_1, i_2)$, which is not a Nash equilibrium, the two-step transition probability satisfies

$$P[\text{I is a Nash equilibrium} \mid \text{I}_{t-2} = (i_1, i_2)] \geq c$$

for a constant $c > 0$.

7.27 *(A nongeneric game)* Consider a two-player game (played by “R” and “C”) whose loss matrices are given by

$$\begin{array}{ccc}
R \setminus C & 1 & 2 & 3 \\
1 & 0 & 1 & 0 \\
2 & 1 & 0 & 0 \\
3 & 1 & 1 & 0 \\
\end{array} \quad \begin{array}{ccc}
R \setminus C & 1 & 2 & 3 \\
1 & 1 & 0 & 1 \\
2 & 0 & 1 & 1 \\
3 & 0 & 0 & 0 \\
\end{array}$$

Suppose both players play according to the strategy described in Exercise 7.26. Show that there is a positive probability that the unique pure Nash equilibrium is never achieved. (This example appears in Hart and Mas-Colell [149].)

7.28 Show that almost all games (with respect to the Lebesgue measure) are such that there exist constants $c_1, c_2 > 0$ such that for all sufficiently small $\varepsilon > 0$, the set $N_\varepsilon$ of approximate Nash equilibria satisfies

$$D_\infty(N_\varepsilon, c_1 \varepsilon) \subset N_\varepsilon \subset D_\infty(N_\varepsilon, c_2 \varepsilon),$$

where $D_\infty(N, \varepsilon) = \{\pi \in \Sigma : \|\pi - \pi'\|_\infty \leq \varepsilon, \pi' \in N\}$ is the $L_\infty$ neighborhood of the set of Nash equilibria, of radius $\varepsilon$. (See, e.g., Germano and Lugosi [126].)

7.29 Use the procedure of experimental regret testing as a building block to design an uncoupled strategy such that if all players follow the strategy, the mixed strategy profiles converge almost surely to a Nash equilibrium of the game for almost all games. *Hint:* Follow the ideas described in Remark 7.14 and the Borel–Cantelli lemma. (Germano and Lugosi [126].)

7.30 Extend the forecasting strategy defined in Section 7.11 to the case of $N > 2$ actions such that, regardless of the sequence of outcomes,

$$\limsup_{n \to \infty} \hat{\mu}_n \leq \min_{i=1, \ldots, N} \limsup_{n \to \infty} \mu_{i,n}.$$  

*Hint:* Place the $N$ actions in the leaves of a rooted binary tree and use the original algorithm recursively in every internal node of the tree. The strategy assigned to the root is the desired forecasting strategy.
7.31 Assume that both players follow the deterministic exploration–exploitation strategy while playing the prisoners’ dilemma. Show that the players will end up cooperating. However, if their play is not synchronized (e.g., if the column player starts following the strategy at time $t = 3$), both players will defect most of the time.

7.32 Prove Corollary 7.4. Hint: Modify the repeated deterministic exploration–exploitation forecaster properly either by letting the parameter $\tau$ grow with time or by using an appropriate doubling trick.