No-Regret Learning and Game-Theoretic Equilibria

by
Casey Alvin Marks
Sc.B., Brown University, 2002
Sc.M., Brown University, 2005

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy in the
Department of Computer Science at Brown University

Providence, Rhode Island
May 2008

This dissertation by Casey Alvin Marks is accepted in its present form by the Department of Computer Science as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date ______________ ______________________________

Amy Greenwald, Director

Recommended to the Graduate Council

Date ______________ ______________________________

Odest Chadwicke Jenkins, Reader

Date ______________ ______________________________

Geoffrey J. Gordon, Reader
Carnegie Mellon University

Approved by the Graduate Council

Date ______________ ______________________________

Sheila Bonde
Dean of the Graduate School

Vita

Casey Alvin Marks was born in San Diego, California on 15 May 2008, the son of Marguerite Marks and Alvin Marks. After graduating first in his class from Francis Parker School, San Diego, California, he studied Applied Mathematics at Brown University, where he received a bachelor's degree in 2002. In 2003, he began the doctoral program in Computer Science at Brown University, receiving a master's degree in 2005 and a Ph.D. in 2008.

Acknowledgements

I am deeply indebted to all those who worked with me and supported me over the course of my graduate studies. To Sara Hillenmeyer, Michelle Hasday, my parents, my collaborators (Zheng Li, Warren Schudy, and especially Geoffrey Gordon), and above all to Amy Greenwald—thank you, from the bottom of my heart.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This work is most generally in the area of *multiagent learning.* The basic object of inquiry is the behavior that emerges when autonomous agents, capable of learning, interact with each other.

More specifically, we are interested in agents learning as they repeatedly play a game and how the resulting behavior relates to the game-theoretic equilibria of the game. It turns out that a particular measure of performance in this setting, called *regret*, is closely related to equilibrium concepts. The study of learning algorithms for which we can provide certain regret guarantees and the study of game-theoretic equilibria are mutually illuminating.

The no-regret approach to learning to play games is a particularly powerful one. Its strength is that it allows for algorithms that do well in either competitive or cooperative situations. That is, the algorithms take advantage of opportunities to cooperate with other players without making themselves vulnerable to exploitation. Further, no-regret results can be obtained without any model of the behavior of the other players.

The fundamental notion in this work is a *game*, in which some (finite) number of players each choose an *action.* Depending on the actions chosen by the players, each player receives some *reward.* Rewards are represented by real numbers and may be interpreted as monetary value or utility (in the economic sense), or in some other domain-appropriate way. The essential properties of rewards are that higher numbers are better (more desirable) and that taking probabilistic expectations is appropriate.

Given a game, we are interested in the equilibria of that game. Roughly, an equilibrium of the game is a configuration of strategies for each player such that no player has an incentive to deviate from that configuration. The classic equilibrium concept is due to Nash [1950], but Aumann's *correlated equilibrium* (1974) is arguably a more useful approach. (Aumann [1987] argues this point persuasively.)

Chapter 2 establishes a formalism for games of this sort (specifically, real-valued, one-shot games) and presents a framework for defining classes of equilibria of such games. Equilibrium concepts correspond to sets of transformations on the players' action sets. One of the original goals of this line of inquiry (particularly in [Greenwald et al., 2008]) was to discover a more powerful equilibrium concept. However, a consequence of Observation 2.4 and Proposition 2.6 is that correlated equilibrium is the strongest equilibrium concept achievable. Thus, the value of the general framework is called into question. Its worth does not become evident until we turn our attention to convex games (Chapter 5) and extensive-form games (Chapter 6).

Chapter 2 also considers the repeated game setting, in which players repeatedly play a one-shot game with each other and thus have an opportunity to learn. We define a class of *no-regret* properties of learning algorithms for this setting and show that algorithms with these properties converge to particular sets of equilibria in self-play. In this sense, we show how to "learn" equilibria.

In order to develop a comprehensive theory of no-regret learning algorithms, we first present, in Chapter 3, a theory of vector games, in which each player's rewards are vectors rather than real numbers. The work in this chapter is a development of Blackwell's seminal paper on approachability. We provide a variety of approachability results as well as two flavors of bounding results.

Once we have this sophisticated machinery for analyzing vector games, in Chapter 4 applies it to develop a class of regret-matching algorithms for matrix (finite-action) games, along with no-regret and regret bounding guarantees for these algorithms. We also demonstrate how to build a no-regret learning algorithm for the naïve setting (in which a player learns only its own reward on each trial) out of a learning algorithm for the informed setting (in which a player learns its opponents' actions). To our knowledge, this is the first general no-regret algorithm for the naïve setting.

Chapter 5 turns to convex games, in which each player's set of actions is a convex set, and rewards are multi-linear. We present a class of transformations for convex games, called σ transformations, which we show to be the strongest set of transformations in the framework. We also consider polyhedral games (convex games in which each player's action set has a finite number of corners) and show that we can consider only the corners of the action sets in these games without losing any power. These two results form the groundwork for Gordon et al. [2008], which presents a class of no-regret learning algorithms that are exponentially faster than previously known algorithms.

Finally, Chapter 6 considers extensive-form games, in which players take turns making choices and may or may not have knowledge of each other's choices. This framework has extremely broad applicability and is particularly appropriate to sequential interactions such as bargaining or poker. Extensive-form games can be represented as matrix games, but it is well known that standard equilibrium concepts, applied to the matrix-game representation, are deficient.

We develop sets of transformations that correspond to two types of extensive-form equilibrium (EFCE) concepts, permissive and reduced, and also present a much smaller set sufficient for reduced EFCE. Using these transformations with the algorithms of Chapter 4 yields the first class of algorithms that are theoretically capable of learning any EFCE. However, such an algorithm would require knowledge of the complete strategy of each opponent in each round. Such knowledge is generally not available. We solve this problem by making use of our class of general, no-regret naïve algorithms. The end result is a class of algorithms for extensive-form games that learn EFCE with minimal information requirements.

Chapter 2

Games and Equilibria

Here we establish a formalism for games of this sort (specifically, real-valued, one-shot games) and present a framework for defining classes of equilibria of such games. Equilibrium concepts correspond to sets of transformations on the players' action sets. One of the original goals of this line of inquiry (particularly in [Greenwald et al., 2008]) was to discover a more powerful equilibrium concept. However, a consequence of Observation 2.4 and Proposition 2.6 is that correlated equilibrium is the strongest equilibrium concept achievable. Thus, the value of the general framework is called into question. Its worth does not become evident until we turn our attention to convex games (Chapter 5) and extensive-form games (Chapter 6).

We also consider the repeated game setting, in which players repeatedly play a one-shot game with each other and thus have an opportunity to learn. We define a class of *no-regret* properties of learning algorithms for this setting and show that algorithms with these properties converge to particular sets of equilibria in self-play. In this sense, we show how to "learn" equilibria.

2.1 Games

The fundamental object of inquiry is a *game*, or more precisely a real-valued, one-shot game.

Definition 2.1 *A* **real-valued, one-shot game** *is a triple*

$$\left\langle N, \langle A_i \rangle_{i \in N}, \langle r_i \rangle_{i \in N} \right\rangle$$

where

- *N is finite set of players*
- *A_i is the set of actions available to player i, and*
- *$r_i : \prod_j A_j \to \mathbb{R}$ is the reward function for player i.*

	R	P	S
R	0,0	-1,1	1,-1
P	1,-1	0,0	-1,1
S	-1,1	1,-1	0,0

Table 2.1: Rock-paper-scissors reward matrix

The interpretation is that each player i independently selects an action from A_i and then receives a reward (or payoff, or utility) according to its reward function. That is, if each player j plays action $a_j \in A_j$, then player i obtains reward $r_i\left(\langle a_j \rangle_{j \in N}\right)$.

A game in which each A_i is a finite set is called a *matrix game*.

The set of possible configurations of actions for the players is called the *joint action space*, denoted $\vec{A}$ and defined as $\vec{A} = \prod_i A_i$. This gives rise to the simpler notation that if the players play joint action $a \in \vec{A}$, each player i obtains reward $r_i(a)$.

Sometimes we will factor a joint action a into two components: player i's action, a_i, and the joint action of all other players, denoted $a_{\neg i} \in A_{\neg i} = \prod_{j \neq i} A_j$. In this case, we may write $r_j(a_i, a_{\neg i})$, which is equivalent to $r(a)$.

For example, consider the game of rock-paper-scissors. It can be represented in this formalism as:

- $N = \{1, 2\}$
- $A_1 = A_2 = \{R, P, S\}$
- $r_1(R, S) = r_1(S, P) = r_1(P, R) = 1$,
  $r_1(S, R) = r_1(P, S) = r_1(R, P) = -1$
  $r_1(R, R) = r_1(P, P) = r_1(S, S) = 0$
- $r_2(a) = -r_1(a)$ for all a

A two-player matrix game, such as rock-paper-scissors, can be conveniently represented in a table, as in Figure 2.1.

The selection of an action for player 1 (the "row player") corresponds to a row of the table, and the selection of an action for player 2 (the "column player") corresponds to a column of the table. The cell in a particular row and column contains a pair indicating the rewards obtained by each player when they play the actions corresponding to the row and column.

Observe that in this particular game, $\sum_i r_i(a) = 0$ for all $a \in \vec{A}$. This property defines a *zero-sum* game.[1] Such games represent purely competitive games—one player's gain is exactly the other players' loss.

We will often be concerned with *bounded games*, in which the range of each r_i is bounded. Without loss of generality, we will assume each $r_i : A \to [0, 1]$ in this case. Observe that any matrix game is necessarily a bounded game.

[1] Equivalently, the rewards may sum to some constant, rather than zero.

2.1.1 Randomization

We consider players who may choose to play distributions over their action sets, rather than specific actions. These distributions are often called *mixed strategies*. In order to formally define these distributions, each action set A_i must have a σ-algebra associated with it. In this work we consider only two kinds of action sets, finite action sets and action sets which are subsets of $\mathbb{R}^d$. In the former case, we take the σ-algebra to be the power set of the action set; in the latter case, we use the Borel σ-algebra. Given appropriate σ-algebras, we denote the set of distributions over player i's action set by $\Delta(A_i)$.

We can also consider the set of joint distributions, $\Delta(\vec{A})$, using the product σ-algebra.[2] A joint distribution is *independent* if it can be expressed as the product of the marginal distributions for each player, or equivalently as an element of $\prod_i \Delta(A_i)$.

2.1.2 Transformations

The concept of an *action transformation* (or just "transformation") serves as the basis for our definitions of both equilibria and regret. An action transformation, denoted ϕ, is a measurable function from a set of actions to itself.[3] Measurability is defined with respect to the σ-algebra associated with the action set. Given player i's action set A_i, the set of all action transformations for A_i is called the set of *swap* transformations and is denoted $\Phi^{\text{SWAP}}(A_i)$. In the case of matrix games, all functions from the action set to itself are measurable.

There are several more limited sets of action transformations which we study, for example the set of all constant action transformations, which are usually called *external* transformations. Formally, we define the external transformation ϕ_α^{EXT}, for $\alpha \in A_i$, as

$$\phi_\alpha^{\text{EXT}}(x) = \alpha \quad \forall x \in A_i. \tag{2.1}$$

Given an action set A_i, the set of external transformations is denoted $\Phi^{\text{EXT}}(A_i)$.

In the study of matrix games, an important class of transformations is the *internal* transformations, which act as the identity except on a single action. Formally, for $\alpha, \beta \in A_i$,

$$\phi_{\alpha\to\beta}^{\text{INT}}(x) = \begin{cases} \beta & \text{if } x = \alpha \\ x & \text{otherwise} \end{cases} \tag{2.2}$$

and the set of internal transformations is denoted $\Phi^{\text{INT}}(A_i)$.

For convex and extensive-form games, there are other important classes of transformations; see Chapters 5 and 6.

[2] Given a nonempty set X, a σ-algebra for X is a nonempty collection of subsets of X that is closed under countable unions and finite complements.

[3] There are other formulations of transformations in the literature. There are discussed in Section 4.8.

2.2 Equilibria

An equilibrium of a game is a specification of how each player is to act, with the property that no player has an incentive to unilaterally deviate from that specification. That is, no player, assuming that all other players follow the specification, would be better off not following it.

The specification can be formalized in two ways: as a vector of mixed strategies, $\langle q_i \rangle \in \prod_i \Delta(A_i)$; or as a distribution over the joint action space, $q \in \Delta(\vec{A}) = \Delta(\prod_i A_i)$. The latter formulation is more general and the one we use. (A joint distribution can express any vector of mixed strategies; to obtain a joint distribution from a vector of mixed strategies, simply take their product.)

An equilibrium that can be represented as a vector of mixed strategies is called *independent*. The best-known equilibrium concept, Nash equilibrium [Nash, 1950] is an independent equilibrium. The formal definition of a Nash equilibrium in terms of joint distributions is:

Definition 2.2 *Given a game, a joint distribution $q \in \prod_i \Delta(A_i)$ is a* **Nash equilibrium** *if (a) it is independent, and (b) for all players i, for all actions $\alpha \in A_i$,*

$$\mathbb{E}\left[r_i(\alpha, a_{\neg i})\right] \leq \mathbb{E}\left[r_i(a)\right] \tag{2.3}$$

where the joint action $a = (a_i, a_{\neg i})$ is distributed according to q.

In general, the ways in which a player is allowed to deviate define an equilibrium concept. Formally, these allowable deviations are defined by a set of action transformations.

Definition 2.3 *Given a game $\left\langle N, \langle A_i \rangle_{i \in N}, \langle r_i \rangle_{i \in N} \right\rangle$ and a vector of sets of actions transformations $\langle \Phi_i \rangle_{i \in N}$, where each $\Phi_i \subseteq \Phi^{\text{SWAP}}(A_i)$, a joint distribution $q \in \Delta(\vec{A})$ is a $\langle \Phi_i \rangle$* **equilibrium** *if for all players i, for all action transformations $\phi \in \Phi_i$,*

$$\mathbb{E}\left[r_i(\phi(a_i), a_{\neg i}) - r_i(a)\right] \leq 0 \tag{2.4}$$

where the joint action $a = (a_i, a_{\neg i})$ is distributed according to q.

For convenience, we define the random variable $\rho_{i,\phi} = r_i(\phi(a_i), a_{\neg i}) - r_i(a)$. Then condition (2.4) becomes

$$\mathbb{E}\left[\rho_{i,\phi}\right] \leq 0 \tag{2.5}$$

If each Φ_i is of the same type, then we can refer to an equilibrium accordingly, e.g., a Φ^{SWAP} equilibrium, or a Φ^{EXT} equilibrium.

A common interpretation of an equilibrium is to add the presence of a *moderator* to the game setting. The moderator, who is trusted by the players, uses a joint distribution to generate a joint action. The moderator then privately gives each player a "suggestion" to play his component of that joint action. A joint distribution is an equilibrium if it has the property that no player can benefit by consistently deviating from the moderator's suggestion.

The larger the set of deviations according to which the players must have no incentive to deviate, the more restrictive the equilibrium concept.

Observation 2.4 *Given a game Γ and two vectors of sets of action transformations $\langle \Phi_i \rangle$ and $\langle \Phi'_i \rangle$, where for each i, $\Phi'_i \subseteq \Phi_i \subseteq \Phi^{\text{SWAP}}(A_i)$ a joint distribution $q \in \Delta(\vec{A})$ that is a $\langle \Phi_i \rangle$-equilibrium must also be a $\langle \Phi'_i \rangle$-equilibrium.*

Thus, the strongest equilibrium concept is a Φ^{SWAP} equilibrium. That is, every equilibrium is a Φ^{SWAP} equilibrium.

Each set of action transformations yields a convex set of equilibria.

Proposition 2.5 *Given a game Γ and a vector of sets of actions transformations $\langle \Phi_i \rangle$, where each $\Phi_i \subseteq \Phi^{\text{SWAP}}(A_i)$, the set of $\langle \Phi_i \rangle$-equilibria is convex.*

Proof Let q and q' be $\langle \Phi_i \rangle$-equilibria, and let $\lambda \in [0, 1]$. Define $q^* = \lambda q + (1 - \lambda) q'$. For arbitrary i and $\phi \in \Phi_i$,

$$\mathbb{E}_{q^*}\left[\rho_{i,\phi}\right] = \int \rho_{i,\phi} \, \mathrm{d}q^* \tag{2.6}$$

$$= \lambda \int \rho_{i,\phi} \, \mathrm{d}q + (1 - \lambda) \int \rho_{i,\phi} \, \mathrm{d}q' \tag{2.7}$$

$$= \lambda \mathbb{E}_q\left[\rho_{i,\phi}\right] + (1 - \lambda)\mathbb{E}_{q'}\left[\rho_{i,\phi}\right] \tag{2.8}$$

$$\leq 0 \tag{2.9}$$

Line (2.7) follows from the convexity of the integral with respect to the measure. Thus q^* is also a $\langle \Phi_i \rangle$-equilibrium. ■

2.2.1 Matrix Game Equilibria

The two most studied types of equilibria in matrix games are Nash equilibria and correlated equilibria [Aumann, 1974]. A related concept, coarse correlated equilibria [Moulin and Vial, 1978], is less well studied.

The moderator interpretation of a correlated equilibrium is that each player is given a suggested action by the moderator, who has sampled a joint action from a joint distribution. If the player has no incentive to deviate based on what he can infer about the suggestions received by the other players based on the suggestion he himself received, then the moderator's joint distribution is a correlated equilibrium. A coarse correlated equilibrium is characterized by a weaker property: each player must have no incentive to deviate without seeing his suggestion.

The following results are from Greenwald et al. [2008]:

Proposition 2.6 *Given a matrix game, the set of Φ^{INT}-equilibria of the game is identical to the set of correlated equilibria of the game.*

Proof A joint distribution q is a Φ^{INT}-equilibrium if and only if for all players i, for all $\alpha, \beta \in A_i$,

$$\mathbb{E}\left[r_i(\phi^{\text{INT}}_{\alpha\to\beta}(a_i), a_{\neg i}) - r_i(a)\right] \leq 0 \tag{2.10}$$

$$\sum_{a \in A} q(a)\left(r_i(\phi^{\text{INT}}_{\alpha\to\beta}(a_i), a_{\neg i}) - r_i(a)\right) \leq 0 \tag{2.11}$$

$$\sum_{a_{\neg i} \in A_{\neg i}} q(\alpha, a_{\neg i})\left(r_i(\beta, a_{\neg i}) - r_i(a)\right) \leq 0 \tag{2.12}$$

	L	M	R
T	0	0	−1
B	0	0	2

Table 2.2: r_1 for a Zero-Sum Game

Line (2.12) follows because $\phi^{\text{INT}}_{\alpha\to\beta}$ acts as the identity when $a_i \neq \alpha$, so $r_i(\phi^{\text{INT}}_{\alpha\to\beta}(a_i), a_{\neg i}) - r_i(a) = 0$ in those cases. Line (2.12) is the definition of correlated equilibrium [Aumann, 1974]. ▮

Proposition 2.7 *Given a matrix game, the set of Φ^{EXT}-equilibria of the game is identical to the set of coarse correlated equilibria of the game.*

Proof A joint distribution q is a Φ^{EXT}-equilibrium if and only if for all players i, for all $\alpha \in A_i$,

$$\begin{aligned} \mathbb{E}\left[r_i(\phi^{\text{EXT}}_{\alpha}(a_i), a_{\neg i}) - r_i(a)\right] &\leq 0 & (2.13)\\ \sum_{a\in A} q(a)\left(r_i(\phi^{\text{EXT}}_{\alpha}(a_i), a_{\neg i}) - r_i(a)\right) &\leq 0 & (2.14)\\ \sum_{a\in A} q(a)\left(r_i(\alpha, a_{\neg i}) - r_i(a)\right) &\leq 0 & (2.15)\end{aligned}$$

which is the definition of coarse correlated equilibrium [Moulin and Vial, 1978]. ▮

A coarse correlated equilibrium need not be a correlated equilibrium. This observation is intuitive for general-sum games, but perhaps less so for zero-sum games.

For example, in the two-player zero-sum game represented in Table 2.2, with player 1 the row player and player 2 the column player, the joint distribution with half its weight on (T,L) and the other half on (B,M) is a coarse correlated equilibrium, but not a correlated equilibrium. It is a coarse correlated equilibrium because player 1 has no incentive to deviate from its marginal distribution (half its weight on T and half on B), and player 2 has no incentive to deviate from its marginal distribution (half its weight on L and half on M). If player 2 were to deviate to R, it would expect to lose $\frac{1}{2}$ instead of 0. It is not, however, a correlated equilibrium: if player 2 is advised to play L, then it can deduce that player 1 is playing T, in which case player 2 actually prefers to play R, where it would win 1 instead of 0.

In the case of two-player, zero-sum games, we obtain the following result for coarse correlated equilibria (and consequently correlated equilibria), which is related to the result in Forges [1990]:

Proposition 2.8 *Given a two-player, zero-sum game with reward functions $r = r_1 = -r_2$ and value v, if q is a coarse correlated equilibrium, then (i) $r(q) = v$ and (ii) each player's marginal distribution is an optimal strategy (i.e., optimal for the maximizing player means: guarantees he wins at least v; optimal for the minimizing player means: guarantees he loses at most v).*

Proof Let q_1 and q_2 denote the marginal distributions of the maximizer and the minimizer in q, respectively. First, $r(q) \geq \max_{\alpha\in A_1} r(\alpha, q_2) \geq v$ since q is a coarse correlated equilibrium and v is the value of the game. Symmetrically, $r(q) \leq \max_{\beta\in A_2} r(q_1, \beta) \leq v$. Hence, $r(q) = v$.

Second, applying the definition of coarse correlated equilibrium again together with the above result, $v = r(q) \geq \max_{\alpha \in A_1} r(\alpha, q_2)$, so by playing q_2, player 2 loses at most v. Symmetrically, $v = r(q) \leq \max_{\beta \in A_2} r(q_1, \beta)$, so by playing q_1, player 1 wins at least v. ∎

2.2.2 Sufficiency of Φ^{INT}

From Observation 2.4, we know that Φ^{SWAP} yields the strongest equilibrium concept in this system; any such equilibrium is also a $\langle \Phi_i \rangle$ equilibrium for all choices of Φ_i. In the case of matrix games, the set of correlated equilibria is equivalent to the set of Φ^{SWAP} equilibria, and is thus the strongest equilibrium concept.

Proposition 2.9 *Given a matrix game, a joint distribution is a correlated equilibrium if and only if it is a Φ^{SWAP} equilibrium.*

Proof One direction follows directly from Observation 2.4. For the other direction, let q be a correlated (equivalently, Φ^{INT}) equilibrium. For arbitrary i and $\phi \in \Phi^{\text{SWAP}}(A_i)$,

$$\begin{aligned}
& \mathbb{E}\left[r_i(\phi(a_i), a_{\neg i}) - r_i(a)\right] && (2.16)\\
= \; & \sum_{a \in A} q(a)\left(r_i(\phi(a_i), a_{\neg i}) - r_i(a)\right) && (2.17)\\
= \; & \sum_{\alpha \in A_i} \sum_{a_{\neg i} \in A_{\neg i}} q(\alpha, a_{\neg i})\left(r_i(\phi(\alpha), a_{\neg i}) - r_i(\alpha, a_{\neg i})\right) && (2.18)\\
= \; & \sum_{\alpha \in A_i} \sum_{a_{\neg i} \in A_{\neg i}} q(\alpha, a_{\neg i})\left(r_i\left(\phi^{\text{INT}}_{\alpha \to \phi(\alpha)}(\alpha), a_{\neg i}\right) - r_i(\alpha, a_{\neg i})\right) && (2.19)\\
= \; & \sum_{\alpha \in A_i} \sum_{a_{\neg i} \in A_{\neg i}} \sum_{a_i \in A_i} q(a)\left(r_i\left(\phi^{\text{INT}}_{\alpha \to \phi(\alpha)}(a_i), a_{\neg i}\right) - r_i(a)\right) && (2.20)\\
= \; & \sum_{\alpha \in A_i} \sum_{a \in A} q(a)\left(r_i\left(\phi^{\text{INT}}_{\alpha \to \phi(\alpha)}(a_i), a_{\neg i}\right) - r_i(a)\right) && (2.21)\\
= \; & \sum_{\alpha \in A_i} \mathbb{E}\left[r_i\left(\phi^{\text{INT}}_{\alpha \to \phi(\alpha)}(a_i), a_{\neg i}\right) - r_i(a)\right] && (2.22)\\
\leq \; & 0 && (2.23)
\end{aligned}$$

The final line follows because q is a Φ^{INT} equilibrium, so each term in the summation in (2.22) is non-positive. ∎

Because an independent Φ^{SWAP} equilibrium is a Nash equilibrium, the existence result for Nash equilibria implies that the set of Φ equilibria is non-empty for any Φ.

2.3 Repeated Games

In this work we are interested in agents learning to play games by playing them repeatedly. In the *repeated game* set-up, we take a game Γ (sometimes called a "one-shot" game to distinguish it from a repeated game) and assign each agent to the role of one of the players in Γ. The agents then play Γ with one another on each of a potentially infinite sequence of trials. In the case of an infinitely-repeated version of the game Γ, we use the notation Γ^∞.

We mostly deal with the *informed* repeated game setting, which is generally characterized by each agent having knowledge of Γ and each trial t having the following steps:

(1) Each agent i selects an action distribution $q_i^{(t)} \in \Delta(A_i)$

(2) For each agent i, the action $a_i^{(t)}$ is sampled from $q_i^{(t)}$

(3) The agents all observe the joint action $a^{(t)} = \left\langle a_j^{(t)} \right\rangle_{j \in N}$

(4) Each agent i receives reward $r_i(a^{(t)})$

Step (3) and knowledge of Γ are the qualities that make the setting merit the label "informed." However, for the algorithms that we study here we will place more lenient requirements on the information available to the agents in the informed setting. Specifically, each agent needs only know its own action set, and step (3) is replaced by:

(3*) Each agent i is given the function $r_i^{(t)} : \alpha \mapsto r_i\left(\alpha, a_{\neg i}^{(t)}\right)$

That is, after each trial an agent is able to calculate the rewards it would have obtained for each of the actions available to it, given the actions of the other agents for that trial. In this formulation, an agent is ignorant of the actions played the other agents and the rewards attained by them. We refer to the function $r_i^{(t)}$ as player i's *marginal reward function.*

In Section 4.7 we consider the more challenging *naïve* setting in which step (3) is eliminated altogether, so that the agent learns only the reward that it attained on each trial.

In these repeated game settings, a learning algorithm for an agent provides an action distribution for the agent to play based upon the agent's behavior and observations during past trials. The concept of a *history* represents the accumulated behavior and observations of an agent. In our formulation of the informed setting, the information available to agent i corresponding to a single trial is an element of $\mathcal{I}_i = \Delta(A_i) \times A_i \times \{A_i \mapsto \mathbb{R}\}$. The information at trial t would be $\left(q_i^{(t)}, a_i^{(t)}, r_i^{(t)}\right)$: agent i's action distribution, action, and marginal reward function. The set of histories for agent i, denoted $\mathcal{H}_i$ is then the set of all finite sequences of elements of $\mathcal{I}_i$, along with a special null history, h_0, which corresponds to nothing having happened yet. Formally, $\mathcal{H}_i = \bigcup_{t \geq 0} \mathcal{I}_i^t$, where $\mathcal{I}_i^0 = \{h_0\}$.

We can now formally define a learning algorithm for the informed repeated game setting:

Definition 2.10 *Given a game and a player i for that game, a learning algorithm for player i for the informed repeated game is a mapping $L : \mathcal{H}_i \to \Delta(A_i)$.*

2.4 Regret

The fundamental notion for both the design and analysis of our learning algorithms is regret. Essentially, an agent calculates its regret by comparing the rewards it obtained over the course of the trials it has played to the rewards it would have obtained had it altered its own behavior during those trials in a particular way. These alterations are specified by action transformations. Given a repeated game, an agent i, and an action transformation for that agent, $\phi \in \Phi^{\text{SWAP}}(A_i)$, the *instantaneous ϕ regret* on trial t is

$$\rho_{i,\phi}^{(t)} = r_i \left(\phi \left(a_i^{(t)} \right), a_{\neg i}^{(t)} \right) - r_i \left(a^{(t)} \right), \tag{2.24}$$

or, in terms of the marginal reward function:

$$\rho_{i,\phi}^{(t)} = r_i^{(t)} \left(\phi \left(a_i^{(t)} \right) \right) - r_i^{(t)} \left(a_i^{(t)} \right). \tag{2.25}$$

The *cumulative ϕ regret* after trial T is simply the sum $\sum_{t=1}^{T} \rho_{i,\phi}^{(t)}$, and the *average ϕ regret* after trial T is

$$\bar{\rho}_{i,\phi}^{T} = \frac{1}{T} \sum_{t=1}^{T} \rho_{i,\phi}^{(t)} \tag{2.26}$$

Given a set of transformations $\Phi \subseteq \Phi^{\text{SWAP}}(A_i)$, the instantaneous Φ regret is the vector

$$\rho_{i,\Phi}^{(t)} = \left\langle \rho_{i,\phi}^{(t)} \right\rangle_{\phi \in \Phi} \tag{2.27}$$

and the cumulative and average Φ regret vectors are analogously defined.

Each $\rho_{i,\Phi}^{(t)}$ is an element of the regret-vector space $\mathbb{R}^{\Phi}$, which is formally a mapping from Φ to $\mathbb{R}$. In the case of finite Φ, $\mathbb{R}^{\Phi}$ is equivalent to the Euclidean space $\mathbb{R}^{|\Phi|}$. Regardless of the cardinality of Φ, we write elements of the regret-vector space as vectors indexed by elements of Φ.

The three most commonly studied forms of regret are external, internal, and swap. The notion of external regret is attributed to Hannan [1957]. Indeed, the no-external-regret property is often called "Hannan consistency," although it is also sometimes called "universal consistency" [Fudenberg and Levine, 1995]. Foster and Vohra [1999] introduced the notion of internal regret, and Blum and Mansour [2005] introduced the terminology "swap regret."

The frameworks of Lehrer [2003], Blum and Mansour [2005], and Fudenberg and Levine [1999], and our action-transformation framework, can all represent external, internal, and swap regret. The frameworks of Cesa-Bianchi and Lugosi [2003], Hart and Mas-Colell [2001], and Foster and Vohra [1999] can represent external and internal regret naturally.

Lehrer's (2003) framework is very general. In fact, it subsumes the frameworks studied in Cesa-Bianchi and Lugosi [2003], Herbster and Warmuth [1998], and Fudenberg and Levine [1999], as well as our action-transformation framework. However, it also does not allow for partially awake experts as in Blum and Mansour [2005].

2.4.1 No Regret

Intuitively, we want algorithms that minimize their regret. Formally, we define the *no-regret* property as a guarantee about an agent's average ϕ regret vector when playing an infinitely-repeated game.

Different researchers have given different characterizations of this guarantee. Some (e.g., Greenwald et al. [2008]) in terms of Blackwell approachability. Others formulate the property in terms of *almost surely* (a.s.) convergence, or equivalently (by the Hoeffding-Azuma lemma) in terms of an $o(t)$ bound on regret (e.g., Foster and Vohra [1999]). Blackwell no-regret is the stronger form; it implies a.s. no-regret. Regardless of the characterization, the idea is the same: the agent's average ϕ regrets are guaranteed to converge to the interval $(-\infty, 0]$.

An infinitely-repeated game, along with a learning algorithm for each agent, defines a probability space over the universe of infinite sequences of joint actions, A^∞. Each agent's instantaneous ϕ regret on trial t is then a random variable.

Definition 2.11 *Given an infinitely-repeated game $\langle N, \langle A_i\rangle_{i\in N}, \langle r_i\rangle_{i\in N}\rangle$, a player i, and a set of transformations $\Phi \subseteq \Phi^{SWAP}(A_i)$, a learning algorithm for player i is* almost surely no-Φregret *if regardless of the other agents' learning algorithms, for all $\phi \in \Phi$, agent i's average ϕ regret converges to the interval $(-\infty, 0]$ almost surely.*

The definition of Blackwell no-regret is deferred until Section 4.2.

2.4.2 Distribution Regret

The form of regret that we focus on in this work is sometimes called "action regret," because it is calculated with respect to the actions that the agent actually plays, i.e., $a_i^{(t)}$. However some authors consider a form of regret sometimes called "distribution regret." The distribution regret at time t is calculated with respect to the agent's mixed strategy $q_i^{(t)}$. It is essentially an expectation at time $t-1$ of the action regret at time t. We use the prefix δ to indicate the distribution version of a regret quantity. Thus we define the instantaneous distribution regret as

$$\delta\rho_{i,\phi}^{(t)} = \mathbb{E}_{t-1}\left[\rho_{i,\phi}^{(t)}\right] \quad (2.28)$$

$$= \mathbb{E}\left[\rho_{i,\phi}^{(t)} \mid a_i^{(t)} \sim q_i^{(t)}\right], \quad (2.29)$$

and the average distribution regret, $\delta\bar{\rho}_{i,\phi}^{(T)}$, accordingly.

2.5 Convergence to Equilibria

In this section, we establish a fundamental relationship between no-regret learning algorithms and game-theoretic equilibria. We prove that learning algorithms that satisfy no-$\vec{\Phi}$-regret converge to the set of $\vec{\Phi}$-equilibria. We derive as corollaries of this theorem the following two specific results: no-Φ^{EXT}-regret algorithms (i.e., no-external-regret algorithms) converge to the set of Φ^{EXT}-equilibria, which correspond to generalized minimax equilibria in zero-sum games; and no-Φ^{INT}-regret algorithms (i.e., no-internal-regret algorithms) converge to the set of Φ^{INT}-equilibria, which correspond to correlated equilibria in general-sum games. This latter result is well-known Hart and Mas-Colell [2000]. By Proposition 2.8, we arrive at another known result, namely, in two-player, zero-sum games, if each player plays using a no-external-regret learning algorithm, then each player's empirical distribution of joint play converges to his set of minimax strategies Hart and Mas-Colell [2001].

In addition to giving sufficient conditions for convergence to the set of $\vec{\Phi}$-equilibria, we also give *necessary* conditions. We show that multiagent learning converges to the set of $\vec{\Phi}$-equilibria only if the time-averaged Φ_i-regret experienced by each player i converges to the negative orthant.

Given an infinitely-repeated n-player game Γ_n^∞, a *run* of the game is a sequence of action vectors $\{\vec{a}_\tau\}_{\tau=1}^\infty$ with each $\vec{a}_\tau \in \vec{A}$. Given a run $\{\vec{a}_\tau\}_{\tau=1}^\infty$ of Γ_n^∞, the *empirical distribution of joint play through time* t, denoted z_t, is the element of $\Delta(\vec{A})$ given by:

$$z_t(\vec{b}) = \frac{1}{t}\sum_{\tau=1}^{t} \mathbf{1}_{\vec{a}_\tau = \vec{b}} \tag{2.30}$$

where $\mathbf{1}_{x=y}$ denotes the indicator function, which equals 1 whenever $x = y$, and 0 otherwise.

The results in this section rely on a technical lemma, the statement and proof of which appear in Appendix A. We apply this lemma via the following corollary, which relates the empirical distribution of joint play at equilibrium to the players' rewards at equilibrium.

Corollary 2.12 *Given an n-player game Γ_n and a vector of sets of action transformations $\vec{\Phi} = (\Phi_i)_{1\le i\le n}$ such that $\Phi_i \subseteq \Phi^{ALL}(A_i)$ for $1 \le i \le n$. If Z is the set of $\vec{\Phi}$-equilibria of Γ_n, then $d(z_t, Z) \to 0$ as $t \to \infty$ if and only if $\mathbb{E}_{a\sim z_t}\left[\rho^\phi(a)\right] \to \mathbb{R}_-$ as $t \to \infty$, for all players i and for all action transformations $\phi_i \in \Phi_i$.*

Proof For all players i and action transformations $\phi_i \in \Phi_i$, let $f_i^{\phi_i}(q) = \mathbb{E}_{a\sim q}\left[\rho^\phi(a)\right]$ and $Z_i^{\phi_i} = \{q \in \Delta(A_1 \times \ldots \times A_n) \mid f_i^{\phi_i}(q) \le 0\}$, for all $q \in \Delta(A_1 \times \ldots \times A_n)$. The set of $\vec{\Phi}$-equilibria is thus $Z = \cap_{1\le i\le n} \cap_{\phi_i\in\Phi_i} Z_i^{\phi_i}$. For each i and ϕ_i, apply Lemma A.1 to $f_i^{\phi_i}$ and $Z_i^{\phi_i}$ so that $d(z_t, Z_i^{\phi_i}) \to 0$ as $t \to \infty$ if and only if $\mathbb{E}_{a\sim z_t}\left[\rho^\phi(a)\right] \to \mathbb{R}_-$ as $t \to \infty$. ■

In words, Corollary 2.12 states that the empirical distribution of joint play converges to the set of $\vec{\Phi}$-equilibria if and only if the rewards each player i obtains exceed the rewards player i could have expected to obtain by playing according to any of the action transformations $\phi_i \in \Phi_i$ of component i of the empirical distribution of joint play.

Theorem 2.13 *Given an n-player game Γ_n and a vector of sets of action transformations $\vec{\Phi} = (\Phi_i)_{1\le i\le n}$ such that $\Phi_i \subseteq \Phi^{ALL}(A_i)$ is finite for $1 \le i \le n$. As $t \to \infty$, the average Φ_i-regret experienced by each player i through time t converges to the negative orthant if and only if the empirical distribution of joint play converges to the set of $\vec{\Phi}$-equilibria of Γ_n.*

Proof By Corollary 2.12, it suffices to show that, as $t \to \infty$, the average Φ_i-regret through time t experienced by each player i converges to the negative orthant if and only if for all players i and for all $\phi_i \in \Phi_i$, $\mathbb{E}_{a\sim z_t}\left[\rho^\phi(a)\right] \to \mathbb{R}_-$ as $t \to \infty$. But for arbitrary player i and for arbitrary $\phi_i \in \Phi_i$,

$$\mathbb{E}_{a\sim z_t}\left[\rho^\phi(a)\right] = \frac{1}{t}\sum_{\tau=1}^{t} \rho^\phi(a_\tau) \tag{2.31}$$

From this equivalence, the conclusion follows immediately. ■

By Theorem 2.13, if the time-averaged Φ_i-regret experienced by each player i converges to the negative orthant with probability 1, then empirical distribution of joint play converges to the set of $\vec{\Phi}$-equilibria with probability 1. But if each player i plays according to a no-Φ_i-regret learning algorithm, then the time-averaged Φ_i-regret experienced by each player i converges to the negative orthant with probability 1, regardless of the opposing algorithm:: i.e., on any run of the game. From this discussion, we draw the following general conclusion:

Theorem 2.14 *Given an n-player game Γ_n and a vector of sets of action transformations $\vec{\Phi} = (\Phi_i)_{1 \leq i \leq n}$ such that $\Phi_i \subseteq \Phi^{ALL}(A_i)$ is finite for $1 \leq i \leq n$. If all players i play no-Φ_i-regret learning algorithms, then the empirical distribution of joint play converges to the set of $\vec{\Phi}$-equilibria of Γ_n with probability 1.*

Thus, we see that if all players abide by no-internal-regret algorithms, then the distribution of play converges to the set of correlated equilibria. Moreover, in two-player, zero-sum games if all players abide by no-external-regret algorithms, then the distribution of play converges to the set of generalized minimax equilibria, that is, the set of minimax-valued joint distributions. Again, by Proposition 2.8, this latter result implies that each player's empirical distribution of joint play converges to his set of minimax strategies, under the stated assumptions.

Chapter 3

Vector Games

In this chapter we set aside our definition of a real-valued game (Definition 2.1), and develop a theory of vector games. The framework we develop will ultimately serve the purpose of allowing us to analyze the regret properties of repeated real-valued games.

In a *vector game* (also called a vector-valued game), we are only concerned with two players, called the protagonist and the opponent. Both players select actions, but only the protagonist obtains rewards, which are now vectors rather than real numbers. Formally:

Definition 3.1 *A* **vector game** *is a tuple $\langle A, A', V, \rho \rangle$, where*

- *A is the set of actions available to the protagonist,*
- *A' is the set of actions available to the opponent,*
- *V is a real Hilbert space, and*
- *$\rho : A \times A' \to V$ is the protagonist's reward function.*

A real Hilbert space is a vector space over $\mathbb{R}$ with an inner product, denoted $\langle \cdot, \cdot \rangle$; the space must be complete with respect to the norm $\|x\| = \sqrt{\langle x, x \rangle}$. The most accessible examples of a real Hilbert space, and the ones that we will predominantly work with, are the Euclidean spaces $\mathbb{R}^d$. In these cases, we will refer to the vector game as a *Euclidean game*.

As in the case of real-valued games, we assume that we have σ-algebras for A and A' so that we can define probability distributions (elements of $\Delta(A)$ and $\Delta(A')$) over them.

We say the game is a *bounded vector game* if the rewards attainable by the protagonist have bounded norm.

We consider the repeated play of a vector game just as we did with real-valued games in Chapter 2, and we similarly define the notions of a history and of a learning algorithm. In the repeated version of a vector game $\langle A, A', V, \rho \rangle$, each trial t has the following steps:

(1) The protagonist selects an action distribution $q_t \in \Delta(A)$ from which action $a^{(t)}$ is sampled

(2) The opponent selects an action distribution $q'_t \in \Delta(A')$ from which action $a'^{(t)}$ is sampled

(3) Both agents observe the joint action (a_t, a'_t)

(4) The protagonist receives reward $\rho\left(a_t, a'_t\right) \in V$

The set of histories, denoted $\mathcal{H}$, is the set of finite (possibly zero)-length sequences of pairs $(a, a') \in A \times A'$. A learning algorithm L for the protagonist is a function $\mathcal{H} \to \Delta(A)$, and a learning algorithm for the opponent is a function $H \to \Delta(A')$, each of which indicates the action distribution that the agent is to play on the next trial.

Following Blackwell, we define approachability, a notion of convergence for vector games. First, define the average reward vector at time T:

$$\bar{\rho}_T = \frac{1}{T} \sum_{t=1}^{T} \rho(a_t, a'_t) \tag{3.1}$$

3.1 Approachability

Definition 3.2 (Approachability) *Given an infinitely-repeated vector game $\langle A, A', V, \rho \rangle$, a set $U \subseteq V$, and a learning algorithm L for the protagonist, the set U is said to be* approachable *by L, if for all $\epsilon > 0$, there exists t_0 such that for any opposing learning algorithm L',*

$$P\left[\exists t \geq t_0 \quad s.t. \quad d\left(U, \bar{\rho}_T\right) \geq \epsilon\right] < \epsilon \tag{3.2}$$

where the distance function d is defined by V's inner product in the standard manner.[1]

Note that the definition of approachability requires equation (3.2) hold *for all opposing learning algorithms*. This stands in contrast to other characterizations (e.g., Greenwald and Jafari [2003], Blum and Mansour [2005]) which use the phrase *for all sequences of opposing actions*. The former characterizes for *adaptive* adversaries, while the latter only accounts for *oblivious* adversaries.

For example, consider the two-player zero-sum game rock-paper-scissors as a vector game with rewards in $\mathbb{R}$. The algorithm which always plays mixed strategy $\left\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right\rangle$ approaches the set in which it does at least as well as the minimax value of the game, 0, *for all opposing learning algorithms*. However, consider the alternating algorithm which plays mixed strategy $\left\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right\rangle$ on odd-numbered rounds, and on even rounds t plays the same action as it played in round $t - 1$. This algorithm will still approach 0 *for all sequences of opposing actions*, but is easily taken advantage of by an adaptive adversary.

3.1.1 Blackwell Bounds

We introduce three bounding concepts which we will employ in deriving approachability and bounding results for protagonist learning algorithms: *expectation Blackwell bounds*, *almost-surely Blackwell*

[1] $d(x, y) = \|x - y\| = \sqrt{\langle x - y, x - y \rangle}$

bounds, and *absolute Blackwell bounds*. An expectation Blackwell bound is a generalization of Blackwell's [Blackwell, 1956] sufficient condition for approachability where the set being approached is the negative orthant. An almost-surely Blackwell bound is a stronger version of an expectation Blackwell bound—it holds with probability one, rather than in expectation. An absolute Blackwell bound is a stronger version of an almost-surely Blackwell bound—it bounds the absolute value of the quantity.

Definition 3.3 *Let $\langle A, A', V, \rho \rangle$ be a vector game. Let L be a protagonist learning algorithm for the repeated game. Let g be a function, $g : V \to V$. If there exists a function $c : \mathbb{N} \to \mathbb{R}$ such that for any T, for any sequence $\{a_t, a'_t\}_{t=1}^{T-1} \in (A \times A')^{T-1}$, and for any $a' \in A'$,*

$$\mathbb{E}\left[\left\langle g\left(\sum_{t=1}^{T-1} \rho(a_t, a'_t)\right), \rho(a, a')\right\rangle\right] \leq c(T) \tag{3.3}$$

where the expectation is taken over the random variable a, which is distributed according to $L\left(\{a_t, a'_t\}_{t=1}^{T-1}\right)$, then c is said to be an **expectation Blackwell bound** *for L and g in the game.*

Definition 3.4 *Let $\langle A, A', V, \rho \rangle$ be a vector game. Let L be a protagonist learning algorithm for the repeated game. Let g be a function, $g : V \to V$. If there exists a function $c : \mathbb{N} \to \mathbb{R}$ such that for any T, for any sequence $\{a_t, a'_t\}_{t=1}^{T-1} \in (A \times A')^{T-1}$, and for any $a' \in A'$,*

$$\left\langle g\left(\sum_{t=1}^{T-1} \rho(a_t, a'_t)\right), \rho(a, a')\right\rangle \leq c(T) \tag{3.4}$$

almost surely, then c is said to be an **almost-surely Blackwell bound** *for L and g in the game.*

Definition 3.5 *Let $\langle A, A', V, \rho \rangle$ be a vector game. Let g be a function, $g : V \to V$. If there exists a function $C : \mathbb{N} \to \mathbb{R}$ such that for any T, for any sequence $\{a_t, a'_t\}_{t=1}^{T} \in (A \times A')^{T}$,*

$$\left|\left\langle g\left(\sum_{t=1}^{T-1} \rho(a_t, a'_t)\right), \rho(a_T, a'_T)\right\rangle\right| \leq C(t) \tag{3.5}$$

almost surely, then C is said to be an **absolute Blackwell bound** *for g in the game.*

3.2 Gordon Triples

Greenwald et al. [2006] introduced the concept of the Gordon triple and presented three specific Gordon triples.

Definition 3.6 (Gordon Triple) *Let V be an inner product space over $\mathbb{R}$. A triple of functions $\langle G, g, \gamma \rangle$, where $G : V \to \mathbb{R}$, $g : V \to V$, and $\gamma : V \to \mathbb{R}$, is a* Gordon triple *over V if*

$$G(x + y) \leq G(x) + \langle g(x), y \rangle + \gamma(y) \tag{3.6}$$

for all $x, y \in V$.

The function G can be thought of as a potential function, and the function g is often called a *link* function. If $V = \mathbb{R}^d$, G is smooth, and g is equal to the gradient of G, then $\gamma(y)$ is a bound on the higher order terms of the Taylor expansion of $G(x + y)$.

We now present three useful classes of Gordon triples over $\mathbb{R}^d$: the "large polynomial" triples, the "small polynomial" triples, and the exponential triples.

A large polynomial triple $\langle G, g, \gamma \rangle$ is parameterized by a real number $p > 2$ and defined by

$$G(x) = \|x^+\|_p^2 \tag{3.7}$$

$$g_i(x) = \begin{cases} 0 & \text{if } x_i \leq 0 \\ \frac{2x_i^{p-1}}{\|x^+\|_p^{p-2}} & \text{otherwise} \end{cases} \tag{3.8}$$

$$\gamma(x) = (p-1)\|x\|_p^2 \tag{3.9}$$

A small polynomial triple $\langle G, g, \gamma \rangle$ is parameterized by a real number $p \in [1, 2]$ and defined by $G(x) = \|x^+\|_p^p$, $g_i(x) = p(x_i^+)^{p-1}$, and $\gamma(x) = \|x\|_p^p$.

An exponential triple $\langle G, g, \gamma \rangle$ is parameterized by a real number $\eta > 0$ and defined by $G(x) = \frac{1}{\eta} \ln \left(\sum_i e^{\eta x_i}\right)$, $g_i(x) = \frac{e^{\eta x_i}}{\sum_j e^{\eta x_j}}$, and $\gamma(x) = \frac{\eta}{2}\|x\|_\infty^2$.

Proposition 3.7 *A large polynomial triple ($p > 2$) is a Gordon triple.*

Proposition 3.8 *A small polynomial triple ($p \in [1, 2]$) is a Gordon triple.*

Proposition 3.9 *An exponential triple is a Gordon triple.*

The proofs that these triples are Gordon triples appear in Appendix B.

3.3 Approachability Results

The following mathematical results allow the derivation of specific approachability results for repeated Euclidean games.

Proposition 3.10 *Let V be an inner product space over $\mathbb{R}$. Let $x_1, x_2, \ldots$ be a sequence of random vectors taking values in V. Define a sequence of random vectors $X_t \equiv \sum_{\tau=1}^t x_t$ for $t \geq 0$. Let $\langle G, g, \gamma \rangle$ be a Gordon triple over V. Further, let there be a constant $k \in \mathbb{R}$ and functions $c : \mathbb{N} \to \mathbb{R}$ and $C : \mathbb{N} \to \mathbb{R}$ such that for all $t \geq 1$*

$$\gamma(x_t) \quad \leq \quad k \quad a.s. \tag{3.10}$$

$$\mathbb{E}[\langle g(X_{t-1}), x_t \rangle] \quad \leq \quad c(t) \tag{3.11}$$

$$|\langle g(X_{t-1}), x_t \rangle| \quad \leq \quad C(t) \quad a.s.. \tag{3.12}$$

Then

$$P\left[G(X_t) \geq G(\vec{0}) + (c(t) + k)t + 2\epsilon t(C(t) + |c(t)| + k)\right] \leq e^{-\epsilon^2 t} \tag{3.13}$$

for any $\epsilon > 0$.

Proof Let $M_t = G(X_t) - (c(t)+k)t - G(\vec{0})$ for $t \geq 0$. We first show that M_t is a supermartingale. For $t \geq 1$,

$$\begin{aligned}
\mathbb{E}_{t-1}[M_t] &\leq \mathbb{E}_{t-1}[G(X_{t-1}) + \langle x_t, g(X_{t-1})\rangle + k] - (c(t)+k)t - G(\vec{0}) && (3.14)\\
&\leq G(X_{t-1}) + c(t) + k - (c(t)+k)t - G(\vec{0}) && (3.15)\\
&= M_{t-1} && (3.16)
\end{aligned}$$

We now show that $|M_t - M_{t-1}| \leq C(t) + |c(t)| + k$.

- $M_t - M_{t-1} \geq 0$.

$$\begin{aligned}
|M_t - M_{t-1}| &= G(X_t) - G(X_{t-1}) - (c+k) && (3.17)\\
&\leq \langle x_t, g(X_{t-1})\rangle + k - (c(t)+k) && (3.18)\\
&\leq C(t) - c(t) && (3.19)
\end{aligned}$$

- $M_t - M_{t-1} < 0$.

$$\begin{aligned}
|M_t - M_{t-1}| &= G(X_{t-1}) - G(X_t) + (c(t)+k) && (3.20)\\
&\leq - \cdot x_t g(X_{t-1}) + (c(t)+k) && (3.21)\\
&\leq C(t) + c(t) + k && (3.22)
\end{aligned}$$

Apply Lemma A.4 with $f(t) = C(t) + |c(t)| + k$, noting that $M_0 = 0$ a.s., to obtain the result. ▮

Theorem 3.11 *Let $\langle A, A', V, \rho\rangle$ be a vector game. Let $\langle G, g, \gamma\rangle$ be a Gordon triple over V. Let L be a protagonist learning algorithm for the repeated vector game. If γ is bounded on $\rho(A, A')$, C is an absolute Blackwell bound for g, and c is an expectation Blackwell bound for L and g, then*

$$P\left[G\left(\sum_{t=1}^{T} \rho\left(a_t, a_t'\right)\right) \geq G(\vec{0}) + (c(t)+k)t + 2\epsilon t(C(t) + |c(t)| + k)\right] \leq e^{-\epsilon^2 t} \qquad (3.23)$$

for any $\epsilon > 0$, where k is an upper bound on $\gamma(\rho(A, A'))$.

Proof Apply Proposition 3.10 with $x_t = \rho(a_t, a_t')$ ▮

Lemma 3.12 (Convergence Lemma) *Given a function $f : \mathbb{R} \to \mathbb{R}$ that maps the positive reals onto the positive reals and a stochastic process $(X_t : t \geq 0)$, if for all $\epsilon > 0$, there exists T such that for all $t \geq T$, $P[X_t \geq f(\epsilon)] \leq e^{-\epsilon t}$, then for all $\delta > 0$, there exists t_0 such that $P[\exists t \geq t_0 \text{ s.t. } X_t \geq \delta] < \delta$.*

Proof Given an arbitrary $\delta > 0$, choose $\epsilon \in f^{-1}(\delta)$. By assumption, there exists T such that for all $t > T$, $P[X_t \geq \delta] \leq e^{-\epsilon t}$. Now, for all $t' \geq T$,

$$P\left[\exists t \geq t' \text{ s.t. } X_t \geq \delta\right] = P\left[\bigcup_{t \geq t'} (X_t \geq \delta)\right] \qquad (3.24)$$

$$\begin{aligned}
&\leq \sum_{t \geq t'} P[X_t \geq \delta] && (3.25)\\
&\leq \sum_{t \geq t'} e^{-\epsilon t} && (3.26)\\
&= \frac{e^{-\epsilon t'}}{1 - e^{-\epsilon}} && (3.27)
\end{aligned}$$

Hence, for sufficiently large t_0, $P\left[\exists t \geq t_0 \text{ s.t. } X_t \geq \delta\right] < \delta$. $\blacksquare$

We can now derive approachability results. We do so first for large-polynomial link functions, then for small-polynomial link functions, and finally for exponential link functions.

Theorem 3.13 *For $p > 2$, define $g : \mathbb{R}^d \to \mathbb{R}^d_+$ to be*

$$g_i(x) = \begin{cases} 0 & \textit{if } x_i \leq 0 \\ \frac{2x_i^{p-1}}{\|x^+\|_p^{p-2}} & \textit{otherwise} \end{cases} \tag{3.28}$$

Given a bounded Euclidean game $\left\langle A, A', \mathbb{R}^d, \rho \right\rangle$, let L be a protagonist learning algorithm such that c is an expectation Blackwell bound for L and g. such that $\lim_{t\to\infty} \frac{c(t)}{t} = 0$. Then the set $\mathbb{R}^d_-$ is approachable by L.

Proof Let ρ_t denote $\rho(a_t, a'_t)$, let R_t denote $\sum_{\tau=1}^{t} \rho_\tau$, and let $\bar{\rho}_t$ denote $\frac{R_t}{t}$.

The game is bounded, so we can choose $b \in \mathbb{R}$ such that for all $a \in A$ and $a' \in A'$, $\|\rho(a, a')\|_\infty \leq b$.

If $\|R_{t-1}\|_\infty \leq 0$, then $|g(R_{t-1}) \cdot \rho_t| = 0$. Otherwise,

$$\begin{aligned}
|g(R_{t-1}) \cdot \rho| &= \frac{2}{\|R^+_{t-1}\|_p^{p-2}} \left| \sum_i ((R_{t-1})^+_i)^{p-1} \rho_i \right| && (3.29)\\
&\leq \frac{2}{\|R^+_{t-1}\|_p^{p-2}} \; b \sum_i ((R_{t-1})^+_i)^{p-1} && (3.30)\\
&= 2b \frac{\|R^+_{t-1}\|_{p-1}^{p-1}}{\|R^+_{t-1}\|_p^{p-2}} && (3.31)\\
&\leq 2b \frac{\|R^+_{t-1}\|_p^{p-1} ((2b)^d)^{\frac{1}{p(p-1)}}}{\|R^+_{t-1}\|_p^{p-2}} && (3.32)\\
&= (2b)^{\left(\frac{d}{p(p-1)}+1\right)} \|R^+_{t-1}\|_p && (3.33)\\
&= (2b)^{\left(\frac{d}{p(p-1)}+1\right)} \left(\sum_i \left((R_{t-1})^+_i\right)^p \right)^{\frac{1}{p}} && (3.34)\\
&\leq (2b)^{\left(\frac{d}{p(p-1)}+1\right)} \left(d(b(t-1))^p\right)^{\frac{1}{p}} && (3.35)\\
&= (2b)^{\left(\frac{d}{p(p-1)}+1\right)} \sqrt[p]{d}\, b(t-1) && (3.36)\\
&< (2b)^{(d+2)} dt && (3.37)
\end{aligned}$$

Line (3.32) follows from the theory of L^P spaces (see Proposition 6.12 in Folland [1999], for example).

Now apply Theorem 3.11 with the large polynomial Gordon triple and $C(t) = (2b)^{(d+2)}dt$, yielding

$$P\left[\|R_t^+\|_p^2 \geq (c(t)+k)t + 2\epsilon t(C(t) + |c(t)| + k)\right] \leq e^{-\epsilon^2 t} \tag{3.38}$$

for some k. Observe that $d\left(\bar{\rho}_t, \mathbb{R}_-^d\right) = d\left(\frac{R_t}{t}, \mathbb{R}_-^d\right) = \frac{\|R_t^+\|_2}{t}$. Thus,

$$P\left[d\left(\bar{\rho}_t, \mathbb{R}_-^d\right) \geq \sqrt{\frac{1}{t}(c(t)+k) + 2\epsilon(C(t) + |c(t)| + k)}\right] \leq e^{-\epsilon^2 t} \tag{3.39}$$

$$P\left[d\left(\bar{\rho}_t, \mathbb{R}_-^d\right) \geq \sqrt{\frac{c(t)}{t} + \frac{k}{t} + \frac{2\epsilon(|c(t)|+k)}{t} + \frac{2\epsilon C(t)}{t}}\right] \leq e^{-\epsilon^2 t} \tag{3.40}$$

Given an ϵ, for large enough t,

$$P\left[d\left(\bar{\rho}_t, \mathbb{R}_-^d\right) \geq \sqrt{\frac{3\epsilon C(t)}{t}}\right] \leq e^{-\epsilon^2 t} \tag{3.41}$$

$$P\left[d\left(\bar{\rho}_t, \mathbb{R}_-^d\right) \geq \sqrt{3\epsilon(2b)^{(d+2)}d}\right] \leq e^{-\epsilon^2 t} \tag{3.42}$$

And now apply the Convergence Lemma with $f(x) = \sqrt[4]{x}\sqrt{3(2b)^{(d+2)}d}$. ∎

Theorem 3.14 *For $p \in [1,2]$, let $g_i(x) = p(x_i^+)^{p-1}$. Given a bounded Euclidean game $\langle A, A', \mathbb{R}^d, \rho\rangle$, let L be a protagonist learning algorithm such that c is an expectation Blackwell bound for L and g such that $\lim_{t\to\infty}\frac{c(t)}{t} = 0$. Then the set $\mathbb{R}_-^d$ is approachable by L.*

Proof Let ρ_t denote $\rho(a_t, a_t')$, let R_t denote $\sum_{\tau=1}^{t}\rho_\tau$, and let $\bar{\rho}_t$ denote $\frac{R_t}{t}$.

The game is bounded, so we can choose $b \in \mathbb{R}$ such that for all $a \in A$ and $a' \in A'$, $\|\rho(a,a')\|_\infty \leq b$.

$$\begin{aligned} |g(R_{t-1})\cdot\rho| &= p\sum_i((R_{t-1})_i^+)^{(p-1)}(\rho_t)_i && (3.43)\\ &\leq pb\sum_i((R_{t-1})_i^+)^{(p-1)} && (3.44)\\ &\leq pbd(b(T-1))^{(p-1)} && (3.45)\\ &\leq pd(bT)^p && (3.46)\end{aligned}$$

Now apply Theorem 3.11 with the small polynomial Gordon triple and $C(t) = pd(bT)^p$, yielding

$$P\left[\left\|\frac{1}{t}R_t^+\right\|_p^p \geq \frac{1}{t^{(p-1)}}\left((c(t)+k) + 2\epsilon(C(t) + |c(t)| + k)\right)\right] \leq e^{-\epsilon^2 t} \tag{3.47}$$

for some k. Observe that $d\left(\bar{\rho}_t, \mathbb{R}_-^d\right) = \frac{\|R_t^+\|_2}{t} \leq \sqrt{d}\|\frac{1}{t}R_t^+\|_\infty \leq \sqrt{d}\|\frac{1}{t}R_t^+\|_p$, so

$$P\left[d\left(\bar{\rho}_t, \mathbb{R}_-^d\right) \geq \sqrt{d}\sqrt[p]{\frac{1}{t^{(p-1)}}\left((c(t)+k) + 2\epsilon(C(t) + |c(t)| + k)\right)}\right] \leq e^{-\epsilon^2 t} \tag{3.48}$$

Given an ϵ, for large enough t,

$$P\left[d\left(\bar{\rho}_t, \mathbb{R}_-^d\right) \geq \sqrt{d}\sqrt[p]{3\epsilon pbd)}\right] \leq e^{-\epsilon^2 t} \tag{3.49}$$

And now apply the Convergence Lemma with $f(x) = \sqrt{d}\sqrt[p]{3\sqrt{x}pbd)}$. ∎

Theorem 3.15 *Let g be the exponential function*

$$g_i(x) = \frac{e^{\eta x_i}}{\sum_j e^{\eta x_j}} \tag{3.50}$$

Given a bounded Euclidean game $\langle A, A', \mathbb{R}^d, \rho \rangle$, let L be a protagonist learning algorithm. If c is an expectation Blackwell bound for L and g such that for some t_0, $|c(t)| \leq q \ \forall t > t_0$, then the set $\{x \in \mathbb{R}^d \mid x_i \leq \frac{\eta}{2} b^2 + q \ \forall i\}$ is approachable by L, where b is an upper bound such that $\forall a, a', \|\rho(a, a')\|_\infty \leq b$.

Proof Let ρ_t denote $\rho(a_t, a'_t)$, let R_t denote $\sum_{\tau=1}^{t} \rho_\tau$, and let $\bar{\rho}_t$ denote $\frac{R_t}{t}$.

Observe that $\frac{\eta}{2}\|\rho_t\|_\infty^2 \leq \frac{\eta}{2} b^2$ almost surely for any t.

$$\begin{aligned} |g(R_{t-1}) \cdot \rho_t| &= \frac{1}{\sum_i e^{\eta R_i}} \sum_i e^{\eta R_i} \rho_i & (3.51) \\ &\leq \frac{1}{\sum_i e^{\eta R_i}} b \sum_i e^{\eta R_i} & (3.52) \\ &= b & (3.53) \end{aligned}$$

Now apply Theorem 3.11 with the Gordon triple from Lemma 3.9, $C(t) = b$, and $k = \frac{\eta}{2} b^2$, yielding

$$P\left[\frac{1}{\eta} \ln\left(\sum_i e^{(R_t)_i}\right) \geq G(\vec{0}) + (c(t) + \frac{\eta}{2} b^2)t + 2\epsilon t\left(b + |c(t)| + \frac{\eta}{2} b^2\right)\right] \leq e^{-\epsilon^2 t} \tag{3.54}$$

Equivalently,

$$P\left[\frac{1}{\eta t} \ln\left(\sum_i e^{(R_t)_i}\right) - \frac{\eta}{2} b^2 - c(t) \geq \frac{1}{\eta t} \ln d + 2\epsilon\left(b + |c(t)| + \frac{\eta}{2} b^2\right)\right] \leq e^{-\epsilon^2 t} \tag{3.55}$$

Which implies

$$P\left[\max_i \frac{1}{t}(R_t)_i - \frac{\eta}{2} b^2 - q \geq \frac{1}{\eta t} \ln d + 2\epsilon\left(b + q + \frac{\eta}{2} b^2\right)\right] \leq e^{-\epsilon^2 t} \tag{3.56}$$

For large enough t,

$$P\left[\max_i \frac{1}{t}(R_t)_i - \frac{\eta}{2} b^2 - q \geq \epsilon + 2\epsilon\left(b + q + \frac{\eta}{2} b^2\right)\right] \leq e^{-\epsilon^2 t} \tag{3.57}$$

Applying the Convergence Lemma, we get that for all δ, there exists T such that

$$P\left[\exists t \geq T \ \text{ s.t. } \ \max_i \frac{1}{t}(R_t)_i \geq \delta\right] < \delta. \tag{3.58}$$

The result follows. ▮

3.4 Bounding Results

3.4.1 Expectation Bounding Results

Our bounding theorem is a corollary of a modified version of Gordon's Gradient Descent Theorem (Gordon [2005]), which bounds the growth rate of any real-valued function on $\mathbb{R}^n$.

Theorem 3.16 *Let $\langle G, g, \gamma \rangle$ be a Gordon triple. Let $X_0 \in \mathbb{R}^n$, let $x_1, x_2, \ldots$ be a sequence of random vectors over $\mathbb{R}^n$, and define $X_t = X_{t-1} + x_t$ for all times $t \geq 1$.*

If there is a function $D : \mathbb{N} \to \mathbb{R}$ such that for all $t \geq 1$,

$$g(X_{t-1}) \cdot \mathbb{E}_{t-1}[x_t] + \mathbb{E}_{t-1}[\gamma(x_t)] \leq D(t) \quad \textit{a.s.} \tag{3.59}$$

then, for all $t \geq 0$,

$$\mathbb{E}[G(X_t)] \leq G(X_0) + \sum_{\tau=1}^{t} D(\tau) \tag{3.60}$$

Proof The proof is by induction on t. At time $t = 0$, $\mathbb{E}[G(X_t)] = G(X_0)$ and $\sum_{\tau=1}^{t} D(\tau) = 0$.

At time $t \geq 1$, since $X_t = X_{t-1} + x_t$ and $\langle G, g, \gamma \rangle$ is a Gordon triple,

$$\begin{aligned} G(X_t) &= G(X_{t-1} + x_t) && (3.61) \\ &\leq G(X_{t-1}) + g(X_{t-1}) \cdot x_t + \gamma(x_t) && (3.62) \end{aligned}$$

By Assumption 3.59, taking conditional expectations w.r.t. X_{t-1} yields

$$\begin{aligned} \mathbb{E}_{t-1}[G(X_t)] &\leq G(X_{t-1}) + g(X_{t-1}) \cdot \mathbb{E}_{t-1}[x_t] + \mathbb{E}_{t-1}[\gamma(x_t)] && (3.63) \\ &\leq G(X_{t-1}) + D(t) \quad \text{a.s.} && (3.64) \end{aligned}$$

Taking expectations and applying the law of iterated expectations yields

$$\begin{aligned} \mathbb{E}[G(X_t)] &= \mathbb{E}[\mathbb{E}_{t-1}[G(X_t)]] && (3.65) \\ &\leq \mathbb{E}[G(X_{t-1})] + D(t) && (3.66) \end{aligned}$$

Therefore, by the induction hypothesis,

$$\begin{aligned} \mathbb{E}[G(X_t)] &\leq \mathbb{E}[G(X_{t-1})] + D(t) && (3.67) \\ &\leq G(X_0) + \sum_{\tau=1}^{t-1} D(\tau) + D(t) && (3.68) \\ &= G(X_0) + \sum_{\tau=1}^{t} D(\tau) && (3.69) \end{aligned}$$

■

Applying Theorem 3.16 to the cumulative reward vector of a real-vector games gives the following result:

Theorem 3.17 *Let $\langle G, g, \gamma \rangle$ be a Gordon triple. Let $\left\langle A, A', \mathbb{R}^d, \rho \right\rangle$ be a real-vector game. Let L be a learning algorithm for the repeated game. Let c be an expectation Blackwell bound for L and g. Then playing according to L guarantees that for all T,*

$$\mathbb{E}\left[G\left(\sum_{t=1}^{T} \rho(a_t, a'_t)\right)\right] \leq G(0) + T \sup_{a,a'} \gamma\left(\rho(a, a')\right) + \sum_{t=1}^{T} c(t) \tag{3.70}$$

at all times $t \geq 0$ provided the supremum in question exists.

Proof Apply Theorem 3.16 with $x_t = \rho(a_t, a'_t)$. ■

Theorem 3.18 *For $d \in \mathbb{N}$ and $p > 2$, let $f : \mathbb{R}^d \to \mathbb{R}^d_+$ be*

$$f_i(x) = \begin{cases} 0 & \textit{if } x_i \leq 0 \\ \frac{2x_i^{p-1}}{\|x^+\|_p^{p-2}} & \textit{otherwise} \end{cases} \tag{3.71}$$

Let $\left\langle A, A', \mathbb{R}^d, \rho \right\rangle$ be a bounded real-vector game. Let L be a learning algorithm for the repeated game. Let $c(t) = 0$ be an expectation Blackwell bound for L and f. Then playing according to L guarantees that for all T,

$$\mathbb{E}\left[\frac{1}{T} \max_i \sum_{t=1}^{T} \rho_i(a_t, a'_t)\right] \leq \sqrt{\frac{1}{T}(p-1)} \sup_{a,a'} \|\rho(a, a')\|_p \tag{3.72}$$

at all times $T \geq 0$.

Proof Let R_T denote $\sum_{t=1}^{T} \rho(a_t, a'_t)$. Let $G(x) = \|x^+\|_p^2$.

$$\begin{aligned} \mathbb{E}\left[\frac{1}{T} \max_i \sum_{t=1}^{T} \rho_i(a_t, a'_t)\right]^2 &\leq \left(\mathbb{E}\left[\|R_T\|_\infty\right]\right)^2 && (3.73) \\ &\leq \mathbb{E}\left[\left\|(R_T)^+\right\|_p^2\right] && (3.74) \\ &= \mathbb{E}\left[G\left(R_T\right)\right] && (3.75) \\ &\leq G(0) + T \sup_{a,a'} \gamma\left(\rho(a, a')\right) && (3.76) \\ &= T(p-1)\left(\sup_{a,a'} \|\rho(a, a')\|_p^2\right) && (3.77) \end{aligned}$$

The second inequality follows from Lemma A.2, with $x = R_t^\Phi$, $q = 2$, and $p > 2$. The third inequality is an application of Corollary 3.17. ■

Theorem 3.19 *For $d \in \mathbb{N}$ and $p \in [1, 2]$, define $f : \mathbb{R}^d \to \mathbb{R}^d_+$ to be $f_i(x) = (x_i^+)^{p-1}$. Let $\langle G, g, \gamma \rangle$ be the small polynomial ($1 \leq p \leq 2$) Gordon triple. Let $\langle A, A', V, \rho \rangle$ be a bounded vector-valued game with $V \subset \mathbb{R}^d$. Let L be a learning algorithm for the repeated game. Let $c(t) = 0$ be an expectation Blackwell bound for L and g. Then playing according to L guarantees that for all T,*

$$\mathbb{E}\left[\frac{1}{T} \max_i \sum_{t=1}^{T} \rho_i(a_t, a'_t)\right] \leq T^{\left(\frac{1}{p}-1\right)} \sup_{a,a'} \|\rho(a, a')\|_p \tag{3.78}$$

at all times $t \geq 0$.

Proof Let R_T denote $\sum_{t=1}^{T} \rho(a_t, a'_t)$. Let $G(x) = \|x^+\|_p^p$.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\max_i \sum_{t=1}^{T} \rho_i(a_t, a'_t)\right]^p &\leq \left(\mathbb{E}\left[\|R_T\|_\infty\right]\right)^p && (3.79)\\
&\leq \mathbb{E}\left[\left\|(R_T)^+\right\|_p^p\right] && (3.80)\\
&= \mathbb{E}\left[G\left(R_T\right)\right] && (3.81)\\
&\leq G(0) + T\sup_{a,a'} \gamma\left(\rho(a,a')\right) && (3.82)\\
&\leq T\sup_{a,a'} \|\rho(a,a')\|_p^p && (3.83)
\end{aligned}$$

The second inequality follows from Lemma A.2, with $x = R_t^\Phi$, $q = p$, and $1 \leq p \leq 2$. The third inequality is an application of Corollary 3.17. ▮

Theorem 3.20 *Let $\langle G, g, \gamma\rangle$ be the exponential Gordon triple with parameter η. Let $\langle A, A', V, \rho\rangle$ be a bounded vector-valued game with $V \subset \mathbb{R}^d$. Let L be a learning algorithm for the repeated game. Let $c(t) = 0$ be an expectation Blackwell bound for L and g. Then playing according to L guarantees that for all T,*

$$\mathbb{E}\left[\frac{1}{T}\max_i \sum_{t=1}^{T} \rho_i(a_t, a'_t)\right] \leq \frac{\ln d}{\eta T} + \frac{\eta}{2}\sup_{a,a'} \|\rho(a,a')\|_\infty^2 \qquad (3.84)$$

at all times $T \geq 0$.

Proof Let R_T denote $\sum_{t=1}^{T} \rho(a_t, a'_t)$.

$$\begin{aligned}
\mathbb{E}\left[\eta \max_i (R_T)_i\right] &\leq \mathbb{E}\left[\ln \sum_i e^{\eta (R_T)_i}\right] && (3.85)\\
&= \eta\mathbb{E}\left[G\left(R_T\right)\right] && (3.86)\\
&\leq \eta\left(G(0) + T\sup_{a,a'} \gamma\left(\rho(a,a')\right)\right) && (3.87)\\
&= \eta\left(\frac{1}{\eta}\ln d + \frac{\eta T}{2}\sup_{a,a'} \|\rho(a,a')\|_\infty^2\right) && (3.88)\\
&= \ln d + \frac{\eta^2 T}{2}\sup_{a,a'} \|\rho(a,a')\|_\infty^2 && (3.89)
\end{aligned}$$

The first inequality follows from the following observation: for all $x \in \mathbb{R}^n$,

$$\max_i x_i = \max_i \ln e^{x_i} = \ln \max_i e^{x_i} \leq \ln \sum_i e^{x_i} \qquad (3.90)$$

The second inequality is an application of Corollary 3.17. ▮

3.4.2 Almost-Surely Bounding Results

We also present a bounding theorem for the case when inequality (3.59) holds without the expectations. This is essentially a simplified version of Theorem 3.16.

Theorem 3.21 *Let $\langle G, g, \gamma \rangle$ be a Gordon triple. Let $X_0 \in \mathbb{R}^n$, let $x_1, x_2, \ldots$ be a sequence of random vectors over $\mathbb{R}^n$, and define $X_t = X_{t-1} + x_t$ for all times $t \geq 1$.*

If there is a function $D : \mathbb{N} \to \mathbb{R}$ such that for all $t \geq 1$,

$$g(X_{t-1}) \cdot x_t + \gamma(x_t) \leq D(t) \qquad a.s. \tag{3.91}$$

then, for all $t \geq 0$,

$$G(X_t) \leq G(X_0) + \sum_{\tau=1}^{t} D(\tau) \tag{3.92}$$

Proof The proof is by induction on t. For $t = 0$, the result is immediate. For $t \geq 1$,

$$\begin{aligned} G(X_t) &= G(X_{t-1} + x_t) && (3.93) \\ &\leq G(X_{t-1}) + g(X_{t-1}) \cdot x_t + \gamma(x_t) && (3.94) \\ &\leq G(X_{t-1}) + D(t) && (3.95) \\ &\leq G(X_0) + \sum_{\tau=1}^{t-1} D(\tau) + D(t) && (3.96) \\ &= G(X_0) + \sum_{\tau=1}^{t} D(\tau) && (3.97) \end{aligned}$$

with all inequalities holding almost surely. ∎

Applying this result to real-vector games yields:

Theorem 3.22 *Let $\langle G, g, \gamma \rangle$ be a Gordon triple. Let $\left\langle A, A', \mathbb{R}^d, \rho \right\rangle$ be a real-vector game. Let L be a learning algorithm for the repeated game. Let c be an almost-surely Blackwell bound for L and g. Then playing according to L guarantees that for all T,*

$$G\left(\sum_{t=1}^{T} \rho(a_t, a'_t)\right) \leq G(0) + T \sup_{a,a'} \gamma\left(\rho(a, a')\right) + \sum_{t=1}^{T} c(t) \tag{3.98}$$

almost surely, at all times $t \geq 0$ provided the supremum in question exists.

Proof Apply Theorem 3.21 with $x_t = \rho(a_t, a'_t)$. ∎

We can now obtain analogues of Theorems 3.18, 3.19, and 3.20. (The proofs are identical except for the absence of expectations.)

Theorem 3.23 *For $d \in \mathbb{N}$ and $p > 2$, let $f : \mathbb{R}^d \to \mathbb{R}^d_+$ be*

$$f_i(x) = \begin{cases} 0 & \text{if } x_i \leq 0 \\ \frac{2x_i^{p-1}}{\|x^+\|_p^{p-2}} & \text{otherwise} \end{cases} \tag{3.99}$$

Let $\left\langle A, A', \mathbb{R}^d, \rho \right\rangle$ be a bounded real-vector game. Let L be a learning algorithm for the repeated game. Let $c(t) = 0$ be an almost-surely Blackwell bound for L and f. Then playing according to L guarantees that for all T,

$$\frac{1}{T} \max_i \sum_{t=1}^{T} \rho_i(a_t, a'_t) \leq \sqrt{\frac{1}{T}(p-1)} \sup_{a,a'} \|\rho(a, a')\|_p \tag{3.100}$$

at all times $T \geq 0$.

Theorem 3.24 *For $d \in \mathbb{N}$ and $p \in [1,2]$, define $f : \mathbb{R}^d \to \mathbb{R}^d_+$ to be $f_i(x) = (x_i^+)^{p-1}$. Let $\langle G, g, \gamma \rangle$ be the small polynomial ($1 \leq p \leq 2$) Gordon triple. Let $\langle A, A', V, \rho \rangle$ be a bounded vector-valued game with $V \subset \mathbb{R}^d$. Let L be a learning algorithm for the repeated game. Let $c(t) = 0$ be an almost-surely Blackwell bound for L and g. Then playing according to L guarantees that for all T,*

$$\frac{1}{T} \max_i \sum_{t=1}^{T} \rho_i(a_t, a'_t) \leq T^{\left(\frac{1}{p}-1\right)} \sup_{a,a'} \|\rho(a,a')\|_p \tag{3.101}$$

at all times $t \geq 0$.

Theorem 3.25 *Let $\langle G, g, \gamma \rangle$ be the exponential Gordon triple with parameter η. Let $\langle A, A', V, \rho \rangle$ be a bounded vector-valued game with $V \subset \mathbb{R}^d$. Let L be a learning algorithm for the repeated game. Let $c(t) = 0$ be an almost-surely Blackwell bound for L and g. Then playing according to L guarantees that for all T,*

$$\frac{1}{T} \max_i \sum_{t=1}^{T} \rho_i(a_t, a'_t) \leq \frac{\ln d}{\eta T} + \frac{\eta}{2} \sup_{a,a'} \|\rho(a,a')\|_\infty^2 \tag{3.102}$$

at all times $T \geq 0$.

Finally, we show how to obtain a convergence result from an almost-surely bounding result.

Theorem 3.26 *Let $\langle A, A', V, \rho \rangle$ be a bounded vector-valued game with $V \subset \mathbb{R}^d$. Let L be a learning algorithm for the repeated game. If there exists a convergent bounding function $B : \mathbb{N} \to \mathbb{R}$ such that for any sequence of opponent actions and for any t,*

$$\max_i \frac{1}{t} \sum_{\tau=1}^{t} \mathbb{E}_{\tau-1}\left[\rho_i(a_t, a'_t)\right] \leq B(t) \quad a.s. \tag{3.103}$$

and $\lim_{t\to\infty} B(t) = b$ then the set $S_b = \{x \in \mathbb{R}^n \mid \forall i \; x_i \leq b\}$ is approachable by $\mathcal{A}$.

Proof It is sufficient to show that the set is approachable in each individual dimension. Choose an arbitrary dimension i and let r_τ denote $\rho_i(a_\tau, a'_\tau)$. Let R_t denote $\sum_{\tau=1}^{t} r_\tau$.

The proof uses Lemma A.7 from Cesa-Bianchi and Lugosi [2006]. Define $V_t = r_t - \mathbb{E}_{t-1}[r_t]$ so that $V_1, V_2, \ldots$ is a martingale difference sequence. The game is bounded, so WLOG $r_t \in [0,1]$ and $V_t \in [-1,1]$. Applying the lemma, we get that for any t and any $\delta > 0$

$$P\left\{\frac{1}{t}\sum_{\tau=1}^{t} V_\tau > \sqrt{2\delta}\right\} \leq e^{-\delta t} \tag{3.104}$$

or equivalently

$$P\left\{\frac{1}{t}R_t > \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{E}_{\tau-1}\left[(\rho_\tau)_i\right] + \sqrt{2\delta}\right\} \leq e^{-\delta t} \tag{3.105}$$

Given an $\epsilon > 0$, choose t_0 s.t. for $t \geq t_0$, $B(t) < b + \frac{\epsilon}{2}$. Thus for $t \geq t_0$,

$$P\left\{\frac{1}{t}R_t > b + \frac{\epsilon}{2} + \sqrt{2\delta}\right\} \leq e^{-\delta t} \tag{3.106}$$

Setting $\delta = \frac{\epsilon^2}{8}$,

$$P\left\{\frac{1}{t}R_t > b + \epsilon\right\} \leq e^{-\delta t} \tag{3.107}$$

Finally, for $t' \geq t_0$,

$$\begin{aligned} P\left\{\exists t \geq t_0 \text{ s.t. } \frac{1}{t}R_t > b + \epsilon\right\} &\leq \sum_{t \geq t'} P\left\{\frac{1}{t}R_t > b + \epsilon\right\} && (3.108) \\ &\leq \sum_{t \geq t'} e^{-\delta t} && (3.109) \\ &= \frac{e^{-\delta t'}}{1 - e^{-\delta}} && (3.110) \end{aligned}$$

which, for large enough t', is smaller than ϵ. ▮

Chapter 4

Regret Matching

4.1 Regret Games

In order to analyze the regret properties of algorithms in the repeated game setting (from Section 2.3), we construct a vector game such that the "rewards" obtained in the vector game correspond to the regret experienced in the repeated game.

Definition 4.1 *Given a one-shot game* $\Gamma = \left\langle N, \langle A_i \rangle_{i \in N}, \langle r_i \rangle_{i \in N} \right\rangle$*, a player* i*, and a set of action transformations* $\Phi \subseteq \Phi^{SWAP}(A_i)$*, the* **$\Phi$-regret game** *for player* i *is the vector game*

$$\left\langle A_i, A_{\neg i}, \mathbb{R}^{\Phi}, \rho_i^{\Phi} \right\rangle \tag{4.1}$$

where $\rho_i^{\Phi} : A_i \times A_{\neg i} \to \mathbb{R}^{\Phi}$ *is defined as*

$$\rho_i^{\Phi}(a_i, a_{\neg i}) = \langle r_i\left(\phi\left(a_i\right), a_{\neg i}\right) - r_i(a_i, a_{\neg i}) \rangle_{\phi \in \Phi} \tag{4.2}$$

For example, let Γ be the game of rock-paper-scissors presented in Section 2.1. The Φ^{EXT} regret game for player 1 would be $\left\langle \{R, P, S\}, \{R, P, S\}, \mathbb{R}^{\Phi^{\text{EXT}}}, \rho \right\rangle$ where ρ is given by the matrix in Table 4.1. (The protagonist is the row player; the opponent is the column player. Reward vectors are written with respect to the basis $\langle \phi_R^{\text{EXT}}, \phi_P^{\text{EXT}}, \phi_S^{\text{EXT}} \rangle$.)

In the case of finite sets Φ we can treat $\mathbb{R}^{\Phi}$ as the Euclidean space $\mathbb{R}^{|\Phi|}$. For infinite sets of transformations, we must treat regret vectors as functions.[1] For this reason, the Blackwell no-regret

	R	P	S
R	$\langle 0, 1, -1 \rangle$	$\langle 0, 1, 2 \rangle$	$\langle 0, -2, -1 \rangle$
P	$\langle -1, 0, -2 \rangle$	$\langle -1, 0, 1 \rangle$	$\langle 2, 0, 1 \rangle$
S	$\langle 1, 2, 0 \rangle$	$\langle -2, -1, 0 \rangle$	$\langle 1, -1, 0 \rangle$

Table 4.1: Rock-paper-scissors Φ^{EXT} Regret Game

[1] If we were to equip Φ with a σ-algebra and a measure such that $\mathbb{R}^{\Phi}$ were an L^2 space then we would have a satisfactory Hilbert space.

concept has only been studied in the context of matrix games, where all sets of transformations are necessarily finite. However, the results we will be using from Chapter 3 are all in terms of Euclidean games, and so we restrict our attention here to the case of finite Φ.

4.2 Blackwell No-Regret

We can now define Blackwell no-Φ-regret.

Definition 4.2 *Given an infinitely-repeated game, a learning algorithm for agent i is* Blackwell no-Φ-regret *for a set of transformations $\Phi \subseteq \Phi^{SWAP}(A_i)$ if in player i's Φ-regret game, the set $\left\{x \in \mathbb{R}^\Phi \mid x_\phi \leq 0 \quad \forall \phi\right\}$ (the "negative orthant") is approachable by the algorithm.*

In the case of finite Φ, Blackwell no-regret is the strongest no-regret concept, implying a.s. no-regret.

We also define the more general property of Blackwell ϵ-no-regret, which requires only that a neighborhood of the negative orthant is approachable. For finite Φ, each component of the Φ-regret vector must approach the interval $(\infty, \epsilon]$.

Definition 4.3 *Given an infinitely-repeated game, a learning algorithm for agent i is* Blackwell ϵ-no-Φ-regret *for a set of transformations $\Phi \subseteq \Phi^{SWAP}(A_i)$ if in player i's Φ-regret game, the set $\left\{x \in \mathbb{R}^\Phi \mid x_\phi \in (\infty, \epsilon] \quad \forall \phi\right\}$ is approachable by the algorithm.*

4.3 Regret-Matching Algorithms

In this section, we define a general class of learning algorithms, called regret-matching algorithms,[2] for the repeated game setting. These algorithms are parameterized by a set of action transformations Φ and a link function f. We also prove the regret-matching theorem, which states that...

In the definition of regret-matching algorithms, we will consider action transformations on A_i as linear transformations on $\Delta(A_i)$. The linearization of a transformation ϕ is denote with square brackets, $[\phi]$, and formally defined as

$$[\phi](q)(s) = q\left(\phi^{-1}(s)\right) \tag{4.3}$$

for s a measurable subset of A_i.

Given a game, we can choose a finite set $\Phi \subseteq \Phi^{\text{SWAP}}(A_i)$ of action transformations and a link function $f : \mathbb{R}^\Phi \to \mathbb{R}^\Phi_+$. These parameters define the class of (Φ, f)*-regret-matching algorithms* for the game.

[2] We appropriate this terminology from Hart and Mas-Colell [2001], whose regret-matching algorithms based on Φ^{EXT} and the polynomial link functions are instances of this class.

Recall the definition of the cumulative regret vector from Section 2.4, which we will now denote $R^T_{i,\Phi}$, or just R^T where i and Φ are understood.

$$R^T_{i,\Phi} = \sum_{t=1}^{T} \left\langle \rho^{(t)}_{i,\phi} \right\rangle_{\phi \in \Phi} \tag{4.4}$$

If $R^T \in \mathbb{R}^\Phi_-$, so that each element of the regret vector is non-positive, then the agent does not regret its past actions. In this case, a (Φ, f)-regret-matching algorithm leaves the agent's next play unspecified. But if $R^T \notin \mathbb{R}^\Phi_-$, so that the agent "feels" regret in at least one dimension, we apply the link function f to this quantity, yielding a non-negative vector, call it $Y^T \in \mathbb{R}^\Phi_+$. Normalizing this vector, we compute the linear transformation M^T as a convex combination of the linear transformations $[\phi]$ as follows:

$$\begin{aligned} M^T &= \frac{\sum_{\phi\in\Phi} Y^{T-1}_\phi [\phi]}{\sum_{\phi\in\Phi} Y^{T-1}_\phi} && (4.5) \\ &= \frac{\sum_{\phi\in\Phi} f_\phi(R^{T-1})[\phi]}{\sum_{\phi\in\Phi} f_\phi(R^{T-1})} && (4.6) \end{aligned}$$

A (Φ, f)-regret-matching algorithm uses a fixed point of M^T as its mixed strategy at time T whenever $R^T_{i,\Phi} \notin \mathbb{R}^\Phi_-$.

Definition 4.4 (Regret-Matching Algorithm) *Given a game $\left\langle N, \langle A_i \rangle_{i\in N}, \langle r_i \rangle_{i \in N} \right\rangle$, a player i, a finite $\Phi \in \Phi^{\text{SWAP}}$, and a link function $f : \mathbb{R}^\Phi \to \mathbb{R}^\Phi_+$, an algorithm for agent i in playing the repeated game is a* (Φ, f)**-regret-matching algorithm** *if the mixed strategy it plays on trial T, $q^{(T)}_i$, is a fixed point of M^T (defined in Equation 4.5), whenever $R^T_{i,\Phi} \notin \mathbb{R}^\Phi_-$.*

If we take f to be $f_i(x) = (x^+_i)^{p-1}$, for $p > 1$, then we refer to a polynomial Φ-regret-matching algorithm. If we take f to be $f_i(x) = e^{\eta x_i}$, for $\eta > 0$, then we refer to an exponential Φ-regret-matching algorithm.

We prove an equivalence result for matching algorithms:

Lemma 4.5 *Given a game $\left\langle N, \langle A_i \rangle_{i\in N}, \langle r_i \rangle_{i \in N} \right\rangle$, a player i, a finite $\Phi \in \Phi^{\text{SWAP}}$, let f, f' be two link functions, both mapping $\mathbb{R}^\Phi$ to $\mathbb{R}^\Phi_+$. If there exists a strictly positive function $\psi : \mathbb{R}^\Phi \to \mathbb{R}$ such that $\psi(x)f(x) = f'(x)$ for all $x \in \mathbb{R}^\Phi$, then a (Φ, f)-regret-matching algorithm is also a (Φ, f')-regret-matching algorithm.*

Proof At an arbitrary trial T, let M^T and M'^T be defined according to Equation 4.5 for f and f', respectively. Since ψ is strictly positive, $M_t = M'_t$ so that a (Φ, f)-regret-matching algorithm plays a fixed point of M'_t, whenever $R^\Phi_{t-1}(h) \notin \mathbb{R}^\Phi_-$. ▮

Many well-known online learning algorithms arise as instances of (Φ, f)-regret matching, or the closely related class of (Φ, f)-distribution-regret matching. (Distribution regret matching is defined in Section 2.4.2.) The no-external-regret algorithm of Hart and Mas-Colell [2000] is the special case of (Φ, f)-regret matching in which $\Phi = \Phi^{\text{EXT}}$ and f is the polynomial link function with $p = 2$. The no-internal-regret algorithm of Foster and Vohra [1999] is equivalent to (Φ, f)-distribution-regret matching with $\Phi = \Phi^{\text{INT}}$ and the polynomial link function with $p = 2$. If f is the exponential link function, then (Φ, f)-distribution-regret reduces to Freund and Schapire's Hedge algorithm (1997) when $\Phi = \Phi^{\text{EXT}}$ and a variant of an algorithm discussed by Cesa-Bianchi and Lugosi (2003) when $\Phi = \Phi^{\text{INT}}$. For a more thorough comparison of these and related algorithms, see Section 4.8.

4.4 Implementation and Complexity

For finite A[3], the linear transformation M^T maps $\Delta(A)$, a nonempty compact convex set, into itself. Moreover, M^T is continuous, since all $[\phi]$ are linear functions in finite-dimensional Euclidean space, and hence continuous. Therefore, by Brouwer's fixed point theorem, M^T is guaranteed to have a fixed point. Therefore, for finite A, for all choices of f and Φ, a (Φ, f)-regret-matching algorithm exists.

Finite A also means that each linear transformation $[\phi]$ can be represented as an $|A| \times |A|$ stochastic matrix, as can M^T. Thus, a (Φ, f)-regret-matching algorithm can be easily implemented.

The pseudocode for the class of (Φ, f)-regret-matching algorithms is shown in Algorithm 1. The cumulative regret vector is initialized to zero. For all times $t = 1, \ldots, T$, the agent samples a pure action $a^{(t)}$ according to the distribution $q^{(t)}$, after which it observes its marginal reward function $r_i^{(t)}$. Given $a^{(t)}$ and $r_i(t)$, the agent computes its instantaneous regret with respect to each $\phi \in \Phi$, and updates the cumulative Φ-regret vector accordingly. A subroutine is then called to compute the mixed strategy that the agent learns to play at time $t + 1$. In this subroutine, the link function f is applied to the cumulative Φ-regret vector. (Recall that the co-domain of a link function is the positive orthant.) If this quantity is zero, then the subroutine returns an arbitrary mixed strategy. Otherwise, the subroutine returns a fixed point of the stochastic matrix M^T.

Algorithm 1 (Φ, f)-RegretMatchingAlgorithm()

```
 1: initialize cumulative regret vector $X_0 = 0$
 2: **for** $t = 1, 2, \ldots$ **do**
 3:    play mixed strategy $q^{(t)} = (\Phi, f)$-ComputeMixedStrategy($X_{t-1}$)
 4:    observe sampled pure action $a^{(t)} \sim q^{(t)}$
 5:    observe marginal reward function $r_i^{(t)}$
 6:    **for all** $\phi \in \Phi$ **do**
 7:       compute instantaneous regret $\rho_{i,\Phi}$
 8:       update cumulative regret vector $X_t = X_{t-1} + \rho_{i,\Phi}$
 9:    **end for**
10: **end for**
```

Each iteration of Algorithm 1 has time complexity $O(\max\{|\Phi||A|^2, |A|^3\})$, assuming the time complexity of f is linear in Φ (as it is for the polynomial and exponential link functions). Updating the cumulative regret vector in steps 5–8 of Algorithm 1 takes time $O(|\Phi||A|)$, since computing instantaneous regret for each $\phi \in \Phi$ (step 6) is an $O(|A|)$ operation. In Subroutine 2, step 1

[3] We use A to denote the action set of an arbitrary player.

Algorithm 2 (Φ, f)-ComputeMixedStrategy(regret vector X_t)

```
let $Y = f(X_t)$
if $Y = 0$ then
    return arbitrary $q \in \Delta(A)$
else
    let $M = \sum_{\phi \in \Phi} Y_\phi [\phi] / \sum_{\phi \in \Phi} Y_\phi$
    solve for a fixed point $q$ of $M$
    return $q$
end if
```

takes time $O(|\Phi|)$, by assumption. If we view the mixed strategy transformations $[\phi]$ as $|A| \times |A|$ stochastic matrices, then we can also view M_t as an $|A| \times |A|$ stochastic matrix, as it is a convex combination of the elements of Φ. Computing M_t in step 5 takes time $O(|\Phi||A|^2)$, since each matrix $[\phi]$ has dimensions $|A| \times |A|$. Finding the fixed point of an $n \times n$ stochastic matrix, which can be accomplished, for example, via Gaussian elimination, is an $O(n^3)$ operation.

If, however, $\Phi \subseteq \Phi^{\text{SWAP}}(A)$, then this time complexity reduces to $O(\max\{|\Phi||A|, |A|^3\})$, since in this special case, (i) computing instantaneous regret for each $\phi \in \Phi$ (step 6 of Algorithm 1) takes constant time so that updating the cumulative regret vector takes time $O(|\Phi|)$; and (ii) computing the stochastic matrix M_t in step 5 of Subroutine 2 is only an $O(|\Phi||A|)$ operation, since there are only $|A|$ nonzero entries in each $\phi \in \Phi$. In particular, if $\Phi = \Phi^{\text{INT}}(A)$, then the time complexity reduces to $O(|A|^3)$, because $|\Phi^{\text{INT}}(A)| = O(|A|^2)$. Moreover, if $\Phi = \Phi^{\text{EXT}}(A)$, then the time complexity reduces even further to $O(|A|)$, because matrix manipulation is not required in the special case of Φ^{EXT}-regret matching. The rows of M are constant: each is a copy of the (normalized) cumulative regret vector, which is precisely the fixed point of M. In particular, for all $q \in \Delta(A)$,

$$M_t(q) : a \quad \mapsto \quad \frac{\sum_{\phi \in \Phi^{\text{EXT}}} Y_t^\phi [\phi]}{\sum_{\phi \in \Phi^{\text{EXT}}} Y_t^\phi}(q) \tag{4.7}$$

$$= \quad \frac{\sum_{a' \in A} Y_t^{a'} \begin{cases} 1 & \text{if } a = a' \\ 0 & \text{otherwise} \end{cases}}{\sum_{a' \in A} Y_t^{a'}} \tag{4.8}$$

$$= \quad \frac{Y_t^a}{\sum_{a' \in A} Y_t^{a'}} \tag{4.9}$$

Observe that $M_t(q)$ is independent of q, so that

$$q : a \mapsto \frac{Y_t^a}{\sum_{a' \in A} Y_t^{a'}} \tag{4.10}$$

is the unique fixed point of M_t. Subroutine 3 computes the mixed strategy for an external regret-matching algorithm: at time $t+1$, play action a with probability proportional to Y_t^a.

The space complexity of Algorithm 1 (and Subroutine 2) is $O(|\Phi||A|^2) = O(\max\{|\Phi||A|^2, |A|^2\})$ because it is necessary to store the $|\Phi|$ matrices, each with dimensions $|A| \times |A|$, and computing the fixed point of an $|A| \times |A|$ stochastic matrix (via Gaussian elimination) requires $O(|A|^2)$ space. If, however, $\Phi \subseteq \Phi^{\text{SWAP}}(A)$, then the space complexity reduces to $O(\max\{|\Phi||A|, |A|^2\})$, since, in

Algorithm 3 (Φ^{EXT}, f)-ComputeMixedStrategy(regret vector X_t^A)

```
let $Y_t^A = f(X_t^A)$
**if** $Y_t^A = 0$ **then**
    set $q \in \Delta(A)$ arbitrarily
**else**
    **for all** $a \in A$ **do**
        set $q_a = Y_t^a / \sum_{a \in A} Y_t^a$
    **end for**
**end if**
**return** $q$
```

this case, there are only $|A|$ nonzero entries in each $\phi \in \Phi$. In particular, if $\Phi = \Phi^{\text{INT}}(A)$ then the space complexity reduces to $O(|A|^2)$, since it suffices to store cumulative regrets in a matrix of size $|A| \times |A|$. Similarly, if $\Phi = \Phi^{\text{EXT}}(A)$ (Subroutine 3), then the space complexity reduces to $O(|A|)$, since it suffices to store cumulative regrets in a vector of size $|A|$. Our discussion of the time and space complexity of Algorithm 1 and its subroutines is summarized in Table 4.2.

Table 4.2: Complexity of (Φ, f)-Regret Matching

	Time	Space
$\Phi \subseteq \Phi^{\text{ALL}}$	$O(\max\{|\Phi||A|^2, |A|^3\})$	$O(|\Phi||A|^2)$
$\Phi \subseteq \Phi^{\text{SWAP}}$	$O(\max\{|\Phi||A|, |A|^3\})$	$O(\max\{|\Phi||A|, |A|^2\})$
$\Phi = \Phi^{\text{INT}}$	$O(|A|^3)$	$O(|A|^2)$
$\Phi = \Phi^{\text{EXT}}$	$O(|A|)$	$O(|A|)$

4.5 Regret-Matching Theorem

We now prove the regret matching theorem, which states that regret-matching algorithms have an expectation Blackwell bound on their regret games.

Theorem 4.6 (Regret-Matching Theorem) *Given a game Γ, an agent i for the repeated game Γ^∞, a set of action transformations $\Phi \subseteq \Phi^{\text{SWAP}}(A_i)$, a link function $f : \mathbb{R}^\Phi \to \mathbb{R}_+^\Phi$, and a learning algorithm L for i, if L is a (Φ, f)-regret-matching algorithm, then the constant zero function is an expectation Blackwell bound for L and f in agent i's Φ-regret game.*

Proof We must show that for all T,

$$\mathbb{E}_{T-1}\left[\left\langle Y^{T-1}, \rho_{i,\Phi}^T \right\rangle\right] \leq 0 \tag{4.11}$$

When Y^{T-1} is the zero vector (i.e., $Y_\phi^{T-1} = 0$ for all ϕ), the bound is trivial. Otherwise, let q^* is a fixed point of M^T. Given some $a_{\neg i}$, let $r^* : A_i \to \mathbb{R}$ be defined as $r^*(x) = r_i(x, a_{\neg i})$ Let S denote $\sum_{\phi \in \Phi} Y_\phi^{T-1}$. The inner product can be written:

$$\mathbb{E}_{T-1}\left[\left\langle Y^{T-1}, \rho_{i,\Phi}^T \right\rangle\right] \tag{4.12}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left\langle Y^{T-1}, \langle r^*(\phi(a_i)) - r^*(a_i)\rangle\right\rangle_{\phi\in\Phi} \mid a_i \sim q^*\right] && (4.13)\\
&= \mathbb{E}\left[\sum_{\phi\in\Phi} Y_\phi^{T-1} r^*(\phi(a_i)) - r^* a \mid a_i \sim q^*\right] && (4.14)\\
&= \sum_{\phi\in\Phi} Y_\phi^{T-1}\left(\int r^*(\phi(\cdot)\ \mathrm{d})q^* - \int r^*\ \mathrm{d}q^*\right) && (4.15)\\
&= \sum_{\phi\in\Phi} Y_\phi^{T-1}\left(\int r^*\ \mathrm{d}[\phi](q^*) - \int r^*\ \mathrm{d}q^*\right) && (4.16)\\
&= \left(\int r^*\ \mathrm{d}\sum_{\phi\in\Phi} Y_\phi^{T-1}[\phi](q^*)\right) - S\int r^*\ \mathrm{d}q^* && (4.17)\\
&= S\int r^*\ \mathrm{d}M^T(q^*) - S\int r^*\ \mathrm{d}q^* && (4.18)\\
&= S\int r^*\ \mathrm{d}q^* - S\int r^*\ \mathrm{d}q^* && (4.19)\\
&= 0 && (4.20)
\end{aligned}
$$

Line (4.19) follows because q^* is the fixed point of M^T. ∎

We can now apply Theorems 3.13, 3.14, and 3.15 in the context of Φ-regret games to get approachability results for regret-matching algorithms.

Proposition 4.7 *Given a bounded repeated game, for finite Φ any polynomial Φ-regret-matching algorithm ($p > 1$) is no-Φ-regret.*

Proof For small polynomials ($p \in [1, 2]$), the result follows directly from Theorem 3.14 and the Regret Matching Theorem. For large polynomials ($p > 2$), the result follows from Lemma 4.5, Theorem 3.13, and the Regret Matching Theorem. ∎

Proposition 4.8 *Given a bounded repeated game, for finite Φ any exponential Φ-regret-matching algorithm for agent i, $\eta > 0$, is ϵ-no-Φ-regret, where $\epsilon = \frac{\eta}{2}$.*

Proof The result follows from Lemma 4.5, Theorem 3.15, the Regret Matching Theorem, and the observation that for rewards in $[0, 1]$, the maximal entry in an instantaneous regret vector is 1. ∎

4.6 Bounds

We can also apply the Regret Matching Theorem to get bounds on the average regret that an agent will experience at any time t. First, we define the maximal activation of a set of action transformations.

Given a finite set of action transformations $\Phi \subseteq \Phi^{\text{SWAP}}(A_i)$, the *maximal activation*, denoted $\mu(\Phi)$, is computed by maximizing, over all actions $a \in A_i$, the number of transformations ϕ that alter action a: i.e.,

$$\mu(\Phi) = \max_{a\in A_i} |\{\phi \in \Phi : \phi(a) \neq a\}| \tag{4.21}$$

Clearly, $\mu(\Phi) \leq |\Phi|$. In addition, observe that for finite A_i, $\mu(\Phi^{\text{EXT}}(A_i)) = \mu(\Phi^{\text{INT}}(A_i)) = |A_i| - 1$.

Lemma 4.9 *Given a bounded game and a player i, any finite set of action transformations for player i, $\Phi \subseteq \Phi^{SWAP}(A_i)$, has the property that*

$$\|\rho_{i,\Phi}(a)\|_p \leq (\mu(\Phi))^{1/p} \tag{4.22}$$

for any $a \in A$.

Proof Since rewards are bounded in $[0,1]$, regrets are bounded in $[-1,1]$, so that

$$\|\rho_{i,\Phi}(a)\|_p = \sqrt[p]{\sum_{\phi \in \Phi} (\rho_{i,\phi}(a))^p} \leq \sqrt[p]{\sum_{\phi \in \Phi} \mathbf{1}_{\phi(a) \neq a}} \leq \sqrt[p]{\mu(\Phi)} \tag{4.23}$$

■

Proposition 4.10 *If an agent plays a repeated bounded game according to a polynomial Φ-regret-matching algorithm with parameter $p > 2$, then the maximal entry in its average Φ regret vector is bounded as follows:*

$$\mathbb{E}\left[\max_{\phi \in \Phi} \bar{\rho}_{i,\phi}^{(T)}\right] \leq (\mu(\Phi))^{\frac{1}{p}} \sqrt{\frac{1}{T}(p-1)} \tag{4.24}$$

at all times $T \geq 0$.

Proof The result follows from Theorem 3.18, the Regret-Matching Theorem, and Lemma 4.9. ■

Proposition 4.11 *If an agent plays a repeated bounded game according to a polynomial Φ-regret-matching algorithm with parameter $p \in [1,2]$, then the maximal entry in its average Φ regret vector is bounded as follows:*

$$\mathbb{E}\left[\max_{\phi \in \Phi} \bar{\rho}_{i,\phi}^{(T)}\right] \leq (\mu(\Phi))^{\frac{1}{p}} T^{\left(\frac{1}{p}-1\right)} \tag{4.25}$$

at all times $T \geq 0$.

Proof The result follows from Theorem 3.19, the Regret-Matching Theorem, and Lemma 4.9. ■

Proposition 4.12 *If an agent plays a repeated bounded game according to an exponential Φ-regret-matching algorithm, for finite Φ, then the maximal entry in its average Φ regret vector is bounded as follows:*

$$\mathbb{E}\left[\max_{\phi \in \Phi} \bar{\rho}_{i,\phi}^{(T)}\right] \leq \frac{\ln |\Phi|}{\eta T} + \frac{\eta}{2} \tag{4.26}$$

at all times $T \geq 0$.

Proof The result follows directly from Theorem 3.20, the Regret-Matching Theorem, and the observation that for rewards in $[0,1]$,

$$\sup_{a \in A} \|\rho_{i,\Phi}(a)\|_\infty^2 \leq 1 \tag{4.27}$$

■

4.6.1 Distribution Regret

Here we consider regret-matching algorithms which use distribution regret instead of action regret. A distribution regret-matching algorithm is identical to the regret-matching algorithm of Definition 4.4, except that the cumulative distribution regret vector,

$$\delta R^{T}_{i,\Phi} = \sum_{t=1}^{T} \left\langle \delta\rho^{(t)}_{i,\phi} \right\rangle_{\phi\in\Phi} \tag{4.28}$$

is used to define the linear transformation M^T. That is,

$$M^{T} = \frac{\sum_{\phi\in\Phi} f_{\phi}(\delta R^{T-1})[\phi]}{\sum_{\phi\in\Phi} f_{\phi}(\delta R^{T-1})}. \tag{4.29}$$

So that we may reason about distribution regret-matching algorithms, we provide a version of the Regret-Matching Theorem.

Theorem 4.13 *Given a game Γ, an agent i for the repeated game Γ^{∞}, a set of action transformations $\Phi \subseteq \Phi^{SWAP}(A_i)$, a link function $f : \mathbb{R}^{\Phi} \to \mathbb{R}^{\Phi}_{+}$, and a learning algorithm L for i, if L is a (Φ, f)-distribution-regret-matching algorithm, then the constant zero function is an almost-surely Blackwell bound for L and f in agent i's Φ-distribution-regret game.*

The proof is a simplified version of the proof of the Regret-Matching Theorem.

We can now provide bounds on the average distribution regret of distribution-regret-matching algorithms, analogous to Propositions 4.10, 4.11, and 4.12.

Proposition 4.14 *If an agent plays a repeated bounded game according to a polynomial Φ-distribution-regret-matching algorithm with parameter $p > 2$, then the maximal entry in its Φ regret vector is bounded as follows:*

$$\max_{\phi\in\Phi} \delta\bar{\rho}^{(T)}_{i,\phi} \leq (\mu(\Phi))^{\frac{1}{p}} \sqrt{\frac{1}{T}(p-1)} \tag{4.30}$$

at all times $T \geq 0$.

Proof The result follows from Theorem 3.23, the Distribution-Regret-Matching Theorem, and Lemma 4.9. ∎

Proposition 4.15 *If an agent plays a repeated bounded game according to a polynomial Φ-distribution-regret-matching algorithm with parameter $p \in [1, 2]$, then the maximal entry in its average Φ regret vector is bounded as follows:*

$$\max_{\phi\in\Phi} \delta\bar{\rho}^{(T)}_{i,\phi} \leq (\mu(\Phi))^{\frac{1}{p}}\, T^{\left(\frac{1}{p}-1\right)} \tag{4.31}$$

at all times $T \geq 0$.

Proof The result follows from Theorem 3.24, the Distribution-Regret-Matching Theorem, and Lemma 4.9. ∎

Proposition 4.16 *If an agent plays a repeated bounded game according to an exponential Φ-distribution-regret-matching algorithm, for finite Φ, then the maximal entry in its average Φ distribution-regret vector is bounded as follows:*

$$\max_{\phi \in \Phi} \delta \bar{\rho}_{i,\phi}^{(T)} \leq \frac{\ln |\Phi|}{\eta T} + \frac{\eta}{2} \tag{4.32}$$

at all times $T \geq 0$.

Proof The result follows directly from Theorem 3.25, the Distribution-Regret-Matching Theorem, and the observation that for rewards in $[0, 1]$,

$$\sup_{a \in A} \|\rho_{i,\Phi}(a)\|_\infty^2 \leq 1 \tag{4.33}$$

▮

We can also apply Theorem 3.26 to obtain approachability results from distribution-regret bounds.

Proposition 4.17 *Given a bounded repeated game, a learning algorithm that guarantees*

$$\max_{\phi \in \Phi} \delta \bar{\rho}_{i,\phi}^{(T)} \leq B(T) \tag{4.34}$$

at all times $T \geq 0$, where $\lim_{T \to \infty} B(T) = \epsilon$, is ϵ-no-regret.

4.6.2 Summary of Bounds

Table 4.3 summarizes the bounds we derived on $\mathbb{E}\left[\max_{\phi \in \Phi} \frac{1}{t} R_t^\phi\right]$ (for action regret-matching) and $\max_{\phi \in \Phi} \frac{1}{t} \hat{R}_t^\phi$ (for distribution-regret-matching). Our two analyses of polynomial regret matching (Theorems 4.10 and 4.11) agree when $p = 2$. In general, our bounds on polynomial distribution-regret matching for $2 \leq p < \infty$ agree with those of Cesa-Bianchi and Lugosi [2003], although their bounds are computed in terms of the number of experts rather than the number of action transformations (see Section 4.8 for details). For external and internal polynomial distribution-regret matching, in particular, we improve upon the bounds that can be derived immediately from their results. Though the improvement is small for external regret (from a bound proportional to $\sqrt[p]{|A|}$ to a bound proportional to $\sqrt[p]{|A| - 1}$), it is more significant for internal regret (from $|A|^{2/p}$ to $(|A| - 1)^{1/p}$).

Table 4.3: Bounds for polynomial and exponential regret-matching algorithms.

$f_i(x)$	Condition	Bound for finite $\Phi \subseteq \Phi^{\text{ALL}}$	Bound for Φ^{EXT},Φ^{INT}
$(x_i^+)^{p-1}$	$2 < p < \infty$	$\sqrt{\frac{p-1}{t}} \sqrt[p]{\mu(\Phi)}$	$\sqrt{\frac{p-1}{t}} \sqrt[p]{\lvert A\rvert - 1}$
$(x_i^+)^{p-1}$	$1 \leq p \leq 2$	$t^{\left(\frac{1}{p}-1\right)} \sqrt[p]{\mu(\Phi)}$	$t^{\left(\frac{1}{p}-1\right)} \sqrt[p]{\lvert A\rvert - 1}$
$e^{\eta x_i}$	$\eta > 0$	$\frac{\ln \lvert\Phi\rvert}{\eta t} + \frac{\eta}{2}$	$\frac{\ln \lvert\Phi\rvert}{\eta t} + \frac{\eta}{2}$

Finally, observe that for $p > 1$ polynomial regret matching has the property that

$$\lim_{t\to\infty} \mathbb{E}\left[\max_{\phi\in\Phi} \frac{1}{t} R_t^{\phi}\right] \leq 0 \tag{4.35}$$

while exponential regret matching has the property that

$$\lim_{t\to\infty} \mathbb{E}\left[\max_{\phi\in\Phi} \frac{1}{t} R_t^{\phi}\right] \leq \frac{\eta}{2} \tag{4.36}$$

In particular, the bound on the time-averaged action-regret of any polynomial algorithm is eventually better than that of any exponential algorithm. An analogous result holds for distribution regret.

4.7 Naïve Algorithms

Here we construct a naïve learning algorithm for finite A_i with bounded distribution regret. The bound is derived from a bound on the distribution regret of an (informed) learning algorithm L which is run as a subroutine.

The algorithm L^*, parameterized by real number $\lambda \in (0, 1)$, proceeds as follows for each round t:

1. L generates a mixed strategy $q_i^{(t)}$
2. the mixed strategy $\widehat{q}_i^{(t)}$ is calculated as:

$$\widehat{q}_i^{(t)} = (1-\lambda) q_i^{(t)} + \lambda U \tag{4.37}$$

   where U is the uniform distribution over A_i

3. the action $a_i^{(t)}$ is sampled from $\widehat{q}_i^{(t)}$ and played
4. the algorithm observes the reward $r_i^{(t)}\left(a_i^{(t)}\right)$
5. the marginal reward function $\widehat{r}_i^{(t)}$ is calculated and reported to L

$$\widehat{r}_i^{(t)}(a) = \begin{cases} \lambda \frac{r_i^{(t)}\left(a_i^{(t)}\right)}{\widehat{q}_i^{(t)}\left(a_i^{(t)}\right)} & \text{if } \alpha = a_i^{(t)} \\ 0 & \text{otherwise} \end{cases} \tag{4.38}$$

Theorem 4.18 *Let $B : \mathbb{N} \to \mathbb{R}$ a bound on the time-averaged distribution regret of L when its rewards are bounded in $[0, 1]$, i.e., Then if L^* is faced with rewards in $[0, 1]$ it will guarantee for any ϕ,*

$$\frac{1}{T}\sum_{t=1}^{T} \hat{\rho}_{i,\phi}^{(t)} \leq \frac{1-\lambda}{\lambda} B(T) + \lambda \tag{4.39}$$

Proof First note that for any a_i, because $r_i^{(t)}(a_i) \in [0, 1]$ and $\widehat{q}_i^{(t)}(a_i) \geq \lambda$, we know that $\widehat{r}_i^{(t)}(a_i) \in [0, 1]$.

Observe that for any t and any $\alpha \in A_i$,

$$\frac{q_i^{(t)}(\alpha)}{\widehat{q}_i^{(t)}(\alpha)} < \frac{1}{1-\lambda} \tag{4.40}$$

For each t, let $a_*^{(t)}$ be a random variable distributed according to $q_i^{(t)}$.

We show that

$$\mathbb{E}\left[\widehat{r}_i^{(t)}\left(a_*^{(t)}\right) \mid q_i^{(t)}\right] = \lambda \mathbb{E}\left[r_i^{(t)}\left(a_*^{(t)}\right) \mid q_i^{(t)}\right]. \tag{4.41}$$

Derivation:

$$\begin{aligned}
\mathbb{E}\left[\widehat{r}_i^{(t)}\left(a_*^{(t)}\right) \mid q_i^{(t)}\right] &= \sum_{\alpha,\beta\in A_i} q_i^{(t)}(\alpha)\,\widehat{q}_i^{(t)}(\beta)\,\widehat{r}_i^{(t)}(\alpha)\Big|_{a_i^{(t)}=\beta} && (4.42)\\
&= \sum_{\alpha,\beta\in A_i} q_i^{(t)}(\alpha)\,\widehat{q}_i^{(t)}(\beta)\,\lambda\frac{r_i^{(t)}(\beta)}{\widehat{q}_i^{(t)}(\beta)}\mathbf{1}_{\alpha=\beta} && (4.43)\\
&= \sum_{\alpha\in A_i} q_i^{(t)}(\alpha)\,\lambda r_i^{(t)}(\alpha) && (4.44)\\
&= \lambda\mathbb{E}\left[r_i^{(t)}\left(a_*^{(t)}\right) \mid q_i^{(t)}\right] && (4.45)
\end{aligned}$$

Now we show that

$$\mathbb{E}\left[\widehat{r}_i^{(t)}\left(a_*^{(t)}\right) \mid q_i^{(t)}\right] < \frac{\lambda}{1-\lambda}\mathbb{E}\left[r_i^{(t)}\left(a_i^{(t)}\right) \mid q_i^{(t)}\right]. \tag{4.46}$$

Derivation:

$$\begin{aligned}
\mathbb{E}\left[\widehat{r}_i^{(t)}\left(a_*^{(t)}\right) \mid q_i^{(t)}\right] &= \sum_{\alpha\in A_i} q_i^{(t)}(\alpha)\,\lambda r_i^{(t)}(\alpha) && (4.47)\\
&= \sum_{\alpha\in A_i} \widehat{q}_i^{(t)}(\alpha)\,\frac{q_i^{(t)}(\alpha)}{\widehat{q}_i^{(t)}(\alpha)} r_i^{(t)}(\alpha) && (4.48)\\
&< \frac{\lambda}{1-\lambda}\sum_{\alpha\in A_i} \widehat{q}_i^{(t)}(\alpha)\,r_i^{(t)}(\alpha) && (4.49)\\
&= \frac{\lambda}{1-\lambda}\mathbb{E}\left[r_i^{(t)}\left(a_i^{(t)}\right) \mid q_i^{(t)}\right] && (4.50)
\end{aligned}$$

Line (4.47) is the same as Line (4.44). Line (4.48) is a multiplication of each term by 1. Line (4.49) follows because of Equation (4.40). Line (4.50) follows because $a_i^{(t)}$ is sampled from $\widehat{q}_i^{(t)}$.

Now we show that

$$\frac{\lambda}{1-\lambda}\mathbb{E}\left[r_i^{(t)}\left(\phi\left(a_i^{(t)}\right)\right) \mid q_i^{(t)}\right] - \frac{\lambda^2}{1-\lambda} \le \mathbb{E}\left[\widehat{r}_i^{(t)}\left(\phi\left(a_*^{(t)}\right)\right) \mid q_i^{(t)}\right]. \tag{4.51}$$

Derivation:

$$\begin{aligned}
&\mathbb{E}\left[r_i^{(t)}\left(\phi\left(a_i^{(t)}\right)\right) \mid q_i^{(t)}\right] && (4.52)\\
=\;& \sum_{\alpha\in A_i} \widehat{q}_i^{(t)}(\alpha)\,r_i^{(t)}\left(\phi\left(a_i^{(t)}\right)\right) && (4.53)
\end{aligned}$$

$$= \sum_{\alpha \in A_i} \left((1-\lambda) q_i^{(t)}(\alpha) + \frac{\lambda}{A_i} \right) r_i^{(t)} \left(\phi \left(a_i^{(t)} \right) \right) \tag{4.54}$$

$$= (1-\lambda) \mathbb{E} \left[r_i^{(t)} \left(\phi \left(a_*^{(t)} \right) \right) \mid q_i^{(t)} \right] + \frac{\lambda}{A_i} r_i^{(t)} \left(\phi \left(a_i^{(t)} \right) \right) \tag{4.55}$$

$$\leq (1-\lambda) \mathbb{E} \left[r_i^{(t)} \left(\phi \left(a_*^{(t)} \right) \right) \mid q_i^{(t)} \right] + \lambda \tag{4.56}$$

$$= \frac{1-\lambda}{\lambda} \mathbb{E} \left[\widehat{r}_i^{(t)} \left(\phi \left(a_*^{(t)} \right) \right) \mid q_i^{(t)} \right] + \lambda \tag{4.57}$$

Line (4.57) follows because of Equation (4.41).

From the distribution regret bound on L, we have

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[\widehat{r}_i^{(t)} \left(\phi \left(a_*^{(t)} \right) \right) - \widehat{r}_i^{(t)} \left(a_*^{(t)} \right) \mid q_i^{(t)} \right] \leq B(T) \tag{4.58}$$

Applying Equations (4.46) and (4.51) yields

$$\frac{\lambda}{1-\lambda} \left(\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[r_i^{(t)} \left(\phi \left(a_i^{(t)} \right) \right) - r_i^{(t)} \left(a_i^{(t)} \right) \mid q_i^{(t)} \right] - \lambda \right) < B(T) \tag{4.59}$$

Rearranging gives the desired result. ■

This naïve algorithm has the property that even if it uses a subroutine with $B(T) = 0$, the resulting distribution bound will be $\lambda > 0$, and therefore not provably no-regret. However, by using the standard doubling trick, which involves regularly restarting the algorithm with a smaller λ, we can obtain a distribution bound that goes to 0 as $T \to \infty$. Applying Proposition 4.17 shows that this is a no-regret algorithm. Thus we have presented the first no-regret algorithm for general Φ.

4.8 Related Work

The literature is rife with analyses of regret-minimization algorithms defined within a variety of frameworks. There are at least four dimensions on which these analyses vary. First, regret may be computed relative to *actions* or *distributions*. Second, different frameworks incorporate different kinds of *transformations* (e.g., Φ^{INT} and Φ^{EXT}), and consequently feature different kinds of regret. Third, there are two broad classes of results about online learning algorithms. *Bounding results*, derived primarily by computer scientists, provide functions that bound the time-averaged (or cumulative) regret vector at a particular time t. *Convergence results*, derived primarily by game theorists, establish guarantees on the behavior of the time-averaged regret vector as $t \to \infty$. Fourth is the *algorithm* itself. In the case of regret matching, this amounts to choosing, along with the variant of regret, a link function (or equivalently a potential function).

4.8.1 Convergence Results

Foster and Vohra [1999] focus their investigations on internal polynomial distribution-regret matching with $p = 2$ and derive an $o(t)$ bound on its cumulative internal distribution regret, which, by the Hoeffding-Azuma lemma, is sufficient for no internal regret.

Hart and Mas-Colell [2001] analyze external and internal action regret, the latter of which they call "conditional regret." They exhibit a class of no-regret algorithms parameterized by potential functions, which includes the polynomial action-regret-matching algorithm studied here.

Lehrer [2003] combines "replacing schemes," which are functions from $\mathcal{H} \times A$ to A, with "activeness functions" from $\mathcal{H} \times A$ to $\{0, 1\}$. Given a replacing scheme g and an activeness function I, Lehrer's framework compares the agent's rewards to the rewards that could have been obtained by playing action $g(h_t, a_t)$, but only if $I(h_t, a_t) = 1$, yielding a general form of action regret. Lehrer establishes the existence of (no-regret) algorithms whose action regret with respect to any countable set of pairs of replacing schemes and activeness functions, averaged over the number of times each pair is "active," approaches the negative orthant.

Fudenberg and Levine [1999] suppose the existence of a countable set of categories Ψ and consider "classification rules," functions from $\mathcal{H} \times A$ to Ψ. They then compare the agent's rewards to the rewards that could have been obtained under sequences of actions that are measurable with respect to each classification rule. Their framework, which is a special case of Lehrer's, also yields action regret. In it, they derive a variant of fictitious play, "categorical smooth fictitious play," with parameter ϵ that guarantees that the $\limsup$ of the maximal entry in the time-averaged action-regret vector converges to $(-\infty, \epsilon]$ a.s., a property they call "ϵ-universal conditional consistency."

Young [2004] presents an "incremental" conditional regret-matching algorithm, a variant of internal polynomial action-regret matching with $p = 2$. Rather than playing the fixed point of an $|A| \times |A|$ stochastic matrix, the computation of which is an $O(|A|^3)$ operation, Young incrementally updates the agent's mixed strategy based on the internal action-regret vector, whose maintenance is only an $O(|A|^2)$ operation. Young argues that his approach yields an algorithm that exhibits no internal regret. Presumably, this result can be generalized to yield an entire class of no-regret algorithms parameterized by action transformations Φ and link functions f.

4.8.2 Bounding Results

Most bounding results can be found in the computer science learning theory literature. Freund and Schapire [1997] introduce the Hedge algorithm, which in our framework arises as external exponential distribution-regret matching. Inspired by the method of Littlestone and Warmuth [1994], they derive a bound on its external distribution regret.

Herbster and Warmuth [1998] consider a finite set of "experts," which they define as functions from $\mathbb{N}$ to A. In the context of a prediction problem, they compare the agent's rewards to the rewards that could have been obtained had the agent played according to each such expert at each time t. They present an algorithm and bound its distribution regret with respect to alternatives constructed by dividing its history into finite-length segments and choosing the best expert for each segment. Bounds on external distribution regret can be obtained by specializing their framework.

Cesa-Bianchi and Lugosi [2003] develop a framework of "generalized" regret. They rely on the same notion of experts as Herbster and Warmuth [1998], but they pair experts $f_1, \ldots, f_N$ with activation functions $I_i : A \times \mathbb{N} \to \{0, 1\}$. At time t, for each i, if $I_i(a_t, t) = 1$, they compare the

Convergence	Action Regret	Distribution Regret
SWAP	Lehrer [2003] Greenwald et al. [2008] Fudenberg and Levine [1999]	
INT	Hart and Mas-Colell [2001] Young [2004]	Foster and Vohra [1999]
EXT		

Bounding	Action Regret	Distribution Regret
SWAP	Greenwald et al. [2006]	Greenwald et al. [2006] Blum and Mansour [2005]
INT		Cesa-Bianchi and Lugosi [2003]
EXT	Hannan [1957]	Freund and Schapire [1997] Herbster and Warmuth [1998]

Table 4.4: Related Work organized along three dimensions.

agent's rewards to the rewards the agent could have obtained by playing $f_i(t)$. This approach is more general than our action-transformation framework in that alternatives may depend on time. At the same time, it is more limited in that it does not naturally represent swap regret. Their calculations yield bounds on generalized distribution regret.

Blum and Mansour [2005]'s framework is similar to Lehrer's, but is applied to distribution rather than action regret. Their "modification rules" are the same as Lehrer's replacing schemes, but instead of activeness functions, they pair modification rules with "time selection functions," which are functions from $\mathbb{N}$ to the interval $[0, 1]$. The rewards an agent could have obtained under each modification rule are weighted according to how "awake" the rule is, as indicated by the corresponding time selection function. They present a method that, given a collection of algorithms whose external distribution regret is bounded above by $f(t)$ at time t, generates an algorithm whose swap distribution regret (and hence, internal distribution regret) is bounded above by $|A|f(t)$.

Table 4.4 summarizes related work. Note that SWAP results subsume INT results, which in turn subsume EXT results. However, many of these results are more general than their entry in this table suggests. For example, the framework of Herbster and Warmuth [1998] deals with time-varying experts as well as external regret. Also, an appropriate bound on distribution-regret can imply no-regret: i.e., convergence to zero of action regrets.

Chapter 5

Convex Games

Thus far we have focused on matrix games (in which each A_i is finite). Now we consider infinite action sets. In particular, we turn our attention to the case of convex games.

Definition 5.1 *A* **convex game** *is a (real-valued) game* $\left\langle N, \langle A_i \rangle_{i \in N}, \langle r_i \rangle_{i \in N} \right\rangle$ *such that*

- *each* A_i *is a convex, compact subset of Euclidean space, and*
- *each* r_i *is multi-linear.*

For such games we will use the Borel σ-algebras for each A_i to define $\Delta(A_i)$.

5.1 Equilibria

In the case of matrix games, Φ^{INT} was an important set of action transformations. We showed that the set of Φ^{INT} equilibria is equivalent to the set of correlated equilibria (Proposition 2.6), as well as the set of Φ^{SWAP} equilibria (Proposition 2.9). However, once we have infinite action sets, Φ^{INT} no longer yields an interesting equilibrium concept.

Consider, as a simple example, each $A_i = [0, 1]$. Suppose each player uses the uniform distribution as its mixed strategy. Then for any player i, for any $\phi = \phi^{\text{INT}}_{\alpha \to \beta} \in \Phi^{\text{INT}}(A_i)$, the random variable $\rho_{i,\phi}$ has value zero except (potentially) on α. However, under the uniform distribution every singleton set has measure 0, so $\mathbb{E}\left[\rho_{i,\phi}\right] = 0$. Thus, every player playing the uniform distribution (or any distribution that assigns singleton sets measure 0) is a Φ^{INT} equilibrium, regardless of the game's rewards. Clearly the internal transformation are not a useful concept in this realm.

Whereas in the case of matrix games we could take Φ^{INT} as defining the set of correlated equilibria, for the case of general (possibly non-finite) games, we will take the set of Φ^{SWAP} equilibria as the definition of correlated equilibria. This is consistent with our idea of what a correlated equilibrium means—given a suggestion from the moderator, there is *no* transformation of it from which the player would benefit.

Note that Φ^{EXT} is still applicable and has the same interpretation as it did in the finite case.

5.1.1 σ Transformations

However, we will find it useful to have a set of transformations that serve the role that internal transformations did for matrix games. That is, we want a set of transformations that is as powerful as Φ^{SWAP}, but significantly smaller. We introduce the σ transformations. Given a measurable set $S \subset A_i$ and an action $\alpha \in A_i$, we define the σ transformation $\phi^{\sigma}_{S\to\alpha}$ as

$$\phi^{\sigma}_{S\to\alpha}(x) = \begin{cases} \alpha & \text{if } x \in S \\ x & \text{otherwise} \end{cases} \tag{5.1}$$

Clearly $\Phi^{\text{INT}}(A_i) \subseteq \Phi^{\sigma}(A_i)$ for any A_i, as we can take S to be a singleton set.[1]

The set of σ transformations is indeed as powerful as Φ^{SWAP}. This follows from the measure theory result that any measurable non-negative function is the limit of a sequence of simple functions.

Proposition 5.2 *Given a convex game, for any joint distribution, player i, and transformation $\phi^* : A_i \to A_i$, if*

$$\sup_{\phi\in\Phi^{\sigma}} \mathbb{E}\left[\rho_{i,\phi}\right] \le 0, \tag{5.2}$$

then

$$\mathbb{E}\left[\rho_{i,\phi^*}\right] \le 0. \tag{5.3}$$

where the expectations are taken over the joint distribution as usual.

Proof Because ϕ^* is measurable, there is a sequence of "simple" transformations $\{\phi_n\}$ such that $\lim_{n\to\infty} \phi_n\ \phi^*$. A simple transformation is one whose range is finite. Given a set E, let χ_E be the indicator function of E. Each ϕ_n can be written:

$$\phi_n = \sum_{j=1}^{J_n} \alpha_{n,j}\chi_{E_{n,j}} \tag{5.4}$$

where J_n is a natural number, each $\alpha_{n,j} \in A_i$, and $E_{n,j}$ are mutually disjoint measurable subsets of A_i. Equivalently, each ϕ_n is the sum of σ transformations. Let $\phi_{n,j}$ be the σ transformation which maps $E_{n,j}$ to $\alpha_{n,j}$, and otherwise acts as the identity.

$$\phi_n = \sum_{j=1}^{J_n} \phi_{n,j} \tag{5.5}$$

First observe:

$$\begin{aligned}
& \mathbb{E}\left[r_i(\phi^*(a_i), a_{\neg i})\right] && (5.6)\\
= \quad & \mathbb{E}\left[r_i\left(\lim_{n\to\infty}\phi_n(a_i), a_{\neg i}\right)\right] && (5.7)\\
= \quad & \lim_{n\to\infty}\mathbb{E}\left[r_i(\phi_n(a_i), a_{\neg i})\right] && (5.8)\\
= \quad & \lim_{n\to\infty}\mathbb{E}\left[r_i\left(\sum_{j=1}^{J_n}\phi_{n,j}(a_i), a_{\neg i}\right)\right] && (5.9)
\end{aligned}$$

[1] Assuming the singleton sets are measurable, as they are in the Borel σ-algebra.

$$= \lim_{n\to\infty} \sum_{j=1}^{J_n} \mathbb{E}\left[r_i(\phi_{n,j}(a_i), a_{\neg i})\right] \tag{5.10}$$

$$= \lim_{n\to\infty} \sum_{j=1}^{J_n} \int_A r_i(\phi_{n,j}(a_i), a_{\neg i})\ dq \tag{5.11}$$

Then,

$$\begin{aligned}
& \mathbb{E}\left[r_i(\phi^*(a_i), a_{\neg i}) - r_i(a_i, a_{\neg i})\right] && (5.12)\\
= \;& \lim_{n\to\infty} \sum_{j=1}^{J_n} \int_A r_i(\phi_{n,j}(a_i), a_{\neg i}) - r_i(a_i, a_{\neg i})\ dq && (5.13)\\
\leq \;& \lim_{n\to\infty} J_n \sup_{1\leq j\leq J_n} \int_A r_i(\phi_{n,j}(a_i), a_{\neg i}) - r_i(a_i, a_{\neg i})\ dq && (5.14)\\
\leq \;& \lim_{n\to\infty} J_n \sup_{\phi\in\Phi^\sigma} \int_A r_i(\phi(a_i), a_{\neg i}) - r_i(a_i, a_{\neg i})\ dq && (5.15)\\
= \;& \lim_{n\to\infty} J_n \sup_{\phi\in\Phi^\sigma} \mathbb{E}\left[r_i(\phi(a_i), a_{\neg i}) - r_i(a_i, a_{\neg i})\right] && (5.16)\\
\leq \;& 0 && (5.17)
\end{aligned}$$

▮

From this result it follows that the set of correlated (Φ^{SWAP}) equilibria of the game is identical to the set of Φ^σ equilibria.

5.2 Corner and Polyhedral Games

The two properties of convex games (convex action sets and multi-linear rewards) allow us to treat them as if they were much simpler without losing any expressive power. We do this by only considering corners of the action sets. This approach is particularly effective in the case of polyhedral action sets, as it allows us to treat a convex game as a matrix game.

A *corner set* of a convex set is a minimal subset whose convex hull is the subset itself. When A_i is a polyhedron, it has a finite set of corners, denoted $\kappa(A_i)$. We refer to such a game as a *polyhedral game.*

Given a polyhedral game Γ, we can construct the corresponding *corner game*, denoted Γ_κ, in which the set of players N and the reward functions r_i are the same, but the action sets are replaced by their corner sets, $\kappa(A_i)$. We can think of this as a version of the game in which only corners may be played.[2]

Due to the multi-linearity of the reward functions, a convex game game and the corresponding corner game have equivalent sets of correlated equilibria—a correlated equilibrium of Γ_κ is also a

[2] We could in fact apply the corner restriction to only certain players, leaving the other players with their original action sets, and the analysis that follows would still hold. For simplicity, however, we assume that all players are limited to corner actions.

correlated equilibrium of Γ,[3] and for every correlated equilibrium of Γ, there is a payoff-equivalent correlated equilibrium of Γ_κ.[4]

Proposition 5.3 *Let Γ be a polyhedral game, and let q^* be a correlated equilibrium of the corner game Γ_κ. If $\bar{q}^*$ is an extension of q^* to the original joint action space A, then it is a correlated equilibrium of Γ.*

Proof Let $\vec{K}$ be the joint corner space (i.e., the joint action space of Γ_κ).[5] The extension $\bar{q}^*$ is defined, for measurable $B \subset \vec{A}$, as

$$\bar{q}^*(B) = q^*(B \cap \vec{K}). \tag{5.18}$$

Thus $\bar{q}^*(\vec{K}) = 1$; only corners have non-zero probability under $\bar{q}^*$. No player is better off transforming any set of its corners to any other corner. Because only corners are player, no player is better off transforming any measurable set of its actions to any corner. Because each players conditional rewards are linear, no corner being better implies that no interior point is better. Thus $\bar{q}^*$ is a Φ^σ equilibrium, and therefore a correlated equilibrium, of Γ. ▮

Proposition 5.4 *Let q^* be a correlated equilibrium of a polyhedral game Γ. There exists a correlated equilibrium for the corner game Γ_κ that has the same expected payoffs as q^* for each player.*

Proof For each A_i, let K_i denote the (finite) set of corners of A_i. Also, let $B^i : A_i \to \mathbb{R}^{n_i}$ be a barycentric coordinate mapping.[6] Each B^i has the property that

$$\forall a_i \in A_i \quad B^i(a) \cdot \langle j_1^i, \ldots, j_{n_i}^i \rangle = a_i, \tag{5.19}$$

where j_k^i is the corner of A_i corresponding to coordinate k of $\mathbb{R}^{n_i}$,

Let $K = \times_i K_i$ be the joint corner space. Define the joint barycentric coordinate mapping $B : \vec{A} \to \Delta(K)$ by

$$B(a_1, \ldots, a_N)(j_1, \ldots, j_N) = \prod_i B^i_{j_i}(a_i). \tag{5.20}$$

Define the mapping $\kappa : \Delta(A) \to \Delta(K)$, which will transform joint distributions for the convex game into joint distributions for the corner game, as

$$\kappa(q)(\vec{j}) = \int_A B(\cdot)(\vec{j})\ dq. \tag{5.21}$$

We make two claims: κ preserves rewards, and κ preserves equilibria. First, we show that for all joint distributions q, and for all players i, $\mathbb{E}_{\kappa(q)}[r_i] = \mathbb{E}_q[r_i]$. Let $\vec{j}$ denote a joint action in the

[3] Consequently, the existence result for matrix games implies an existence result for polyhedral games.

[4] Some of the results in this section could be extended to non-polyhedral convex games, but we see no benefit to doing so.

[5] Because we are use the Borel σ-algebra, $\vec{K}$ is a measurable subset of $\vec{A}$.

[6] See Gordon et al. [2008] for an explanation of barycentric coordinates.

corner game: $\left\langle j^i \right\rangle_{1\leq i\leq N}$. The result follows from the multi-linearity of r_i:

$$\begin{aligned}
\mathbb{E}\left[r_i \mid \kappa(q)\right] &= \sum_{\vec{j}\in K} r_i(\vec{j})\kappa(q)(\vec{j}) && (5.22)\\
&= \sum_{\vec{j}\in K} r_i(\vec{j}) \int_A B(\cdot)(\vec{j})\ dq && (5.23)\\
&= \int_A \sum_{\vec{j}\in K} B(\cdot)(\vec{j}) r_i(\vec{j})\ dq && (5.24)\\
&= \int_A \sum_{\vec{j}\in K} \prod_i B^i_{j^i}(a_i) r_i(\vec{j})\ dq(a) && (5.25)\\
&= \int_A \sum_{\vec{j}\in K} r_i\left(\left\langle B^i_{j^i}(a_i)\ j^i\right\rangle_{1\leq i\leq N}\right)\ dq(a) && (5.26)\\
&= \int_A \sum_{j^1\in K_1} \cdots \sum_{j^N\in K_N} r_i\left(\left\langle B^i_{j^i}(a_i)\ j^i\right\rangle_{1\leq i\leq N}\right)\ dq(a) && (5.27)\\
&= \int_A r_i\left(\left\langle \sum_{j^i\in K_i} B^i_{j^i}(a_i)\ j^i\right\rangle_{1\leq i\leq N}\right)\ dq(a) && (5.28)\\
&= \int_A r_i\left(\left\langle B^i(a)\cdot\langle j^i_1,\ldots,j^i_{n_i}\rangle\right\rangle_{1\leq i\leq N}\right)\ dq(a) && (5.29)\\
&= \int_A r_i\left(\langle a_i\rangle_{1\leq i\leq N}\right)\ dq(a) && (5.30)\\
&= \int_A r_i\ dq && (5.31)\\
&= \mathbb{E}_q\left[r_i\right] && (5.32)
\end{aligned}$$

Now, assume that q is a correlated equilibrium for the convex game. We prove that $\kappa(q)$ must also be a correlated equilibrium for the convex game by contradiction. Assume that $\kappa(q)$ is not a correlated equilibria. Then there exists a player i, a set $S\subseteq A_i$, and an action $\alpha\in A_i$ such that player i would benefit by deviating from $\kappa(q)$ according to $\phi^\sigma_{S\to\alpha}$. That is,

$$\mathbb{E}_{\kappa(q)}\left[r_i(\phi^\sigma_{S\to\alpha}(a_i),a_{\neg i})\right] > \mathbb{E}_{\kappa(q)}\left[r_i(a)\right] = \mathbb{E}_q\left[r_i(a)\right]. \tag{5.33}$$

We can rewrite the left-hand side using the same reasoning as the previous derivation:

$$\begin{aligned}
& \mathbb{E}_{\kappa(q)}\left[r_i(\phi^\sigma_{S\to\alpha}(a_i),a_{\neg i})\right] && (5.34)\\
=\ & \int_A \sum_{\vec{j}\in K} \prod_{1\leq k\leq N} B^k_{j^k}(a_k) r_i(\phi^\sigma_{S\to\alpha}(a_i),a_{\neg i})\ dq(a) && (5.35)\\
=\ & \int_A \sum_{j_i\in K_i} B^i_{j^i}(a_i) r_i(\phi^\sigma_{S\to\alpha}(j_i),a_{\neg i})\ dq(a) && (5.36)\\
=\ & \int_A r_i\left(\sum_{j_i\in K_i} B^i_{j^i}(a_i)\phi^\sigma_{S\to\alpha}(j_i), a_{\neg i}\right)\ dq(a) && (5.37)
\end{aligned}$$

Define a swap transformation $\phi^* : A_i \to A_i$ as $\phi^*(a_i) = \sum_{j_i \in K_i} B^i_{j^i}(a_i)\phi^\sigma_{S\to\alpha}(j_i)$. Thus, we have

$$\begin{aligned} \mathbb{E}_q\left[r_i(a)\right] &< \mathbb{E}_{\kappa(q)}\left[r_i(\phi^\sigma_{S\to\alpha}(a_i), a_{\neg i})\right] && (5.38)\\ &= \int_A r_i\left(\phi^*(a_i), a_{\neg i}\right)\ dq(a) && (5.39)\\ &= \mathbb{E}_q\left[r_i\left(\phi^*(a_i), a_{\neg i}\right)\right] && (5.40) \end{aligned}$$

But q is a correlated equilibrium, so no player prefers any swap transformation, so this is a contradiction.

We can think of Propositions 5.3 and 5.4 as soundness and completeness results, respectively, for the corner game reduction. Proposition 5.3 is particularly powerful. It provides the justification for the class of no-regret convex-game algorithms presented in Gordon et al. [2008], some of which are exponentially more efficient than previously known algorithms.

Chapter 6

Extensive-Form Games

Here we turn our attention to extensive-form games, in which players take turns making decisions and may or may not learn about the the decisions of others. This framework can represent not only a deterministic turn-taking game with complete (i.e., public) information, like tic-tac-toe, but also a game with incomplete (i.e., private) information and an element of chance, like poker. We present a formalism for describing extensive-form games. We discuss representing these games as matrix games, and we consider two classes of equilibria: permissive extensive-form equilibria and reduced extensive-form equilibria.

6.1 Formalism

To avoid confusion with the sequential properties of repeated games, we refer to each decision point in an extensive-form game as *turn*, and each possible decision at a turn as a *choice*.

As always, let N be a finite set of players. We allow for probabilistic events, represented as choices made by an additional player called the *chance player*. We represent the structure of the extensive-form game as a *game tree*. Each interior node of the tree represents a decision point and belongs to a player (possibly the chance player). Each non-chance player's interior nodes are partitioned into information sets, which represent collections of situations that are indistinguishable to the player. At each information set, a player has a finite number of choices available to it, and every node in that set has one child for each of those choices, with the edges labeled accordingly. The leaf nodes of the tree are the outcomes of the game; each outcome has a real-valued reward for each player.

Let $\mathcal{I}_i$ denote the set of all information sets for player $i \in N$. Given an information set h, let C_h denote the set of choices available at h. Thus, each node in h has $|C_h|$ children. Without loss of generality, each set of choices C_h is disjoint. We assume that each player i has the property of *perfect recall*: for any information set, $h \in \mathcal{I}_i$, the path from the root to each node in h defines the

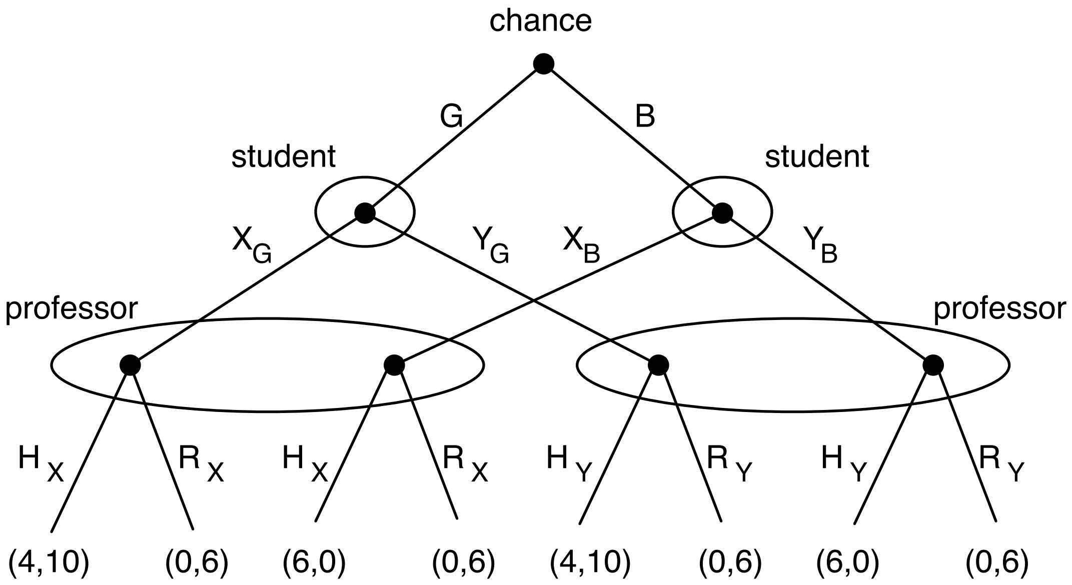

Figure 6.1: Hiring Game

same sequence of choices for i. We denote the set of all choices available to player i by

$$C_i^* = \bigcup_{h \in \mathcal{I}_i} C_h. \tag{6.1}$$

Observe that each play of an extensive-form game corresponds to a path through the game tree. An individual player's view of that path is a sequence of information sets (elements of $\mathcal{I}_i$). Every time a player makes a choice, he is at an information set which, by the perfect recall property, uniquely defines the sequence of information sets which he previously observed, as well as the sequence of choices he made at those information sets.

The example we give here is due to von Stengel and Forges [2006]. Figure 6.1 is a graphical representation of the Hiring Game, which represents a situation in which a student is applying for a research job with a professor. The players in this game are the chance player, the student, and a professor. First, the student receives either a good education (G) or a bad education (B) with equal probability (represented as a choice made by the chance player). Then, the student applies for a job with the professor by sending one of two signals, X or Y, represented by choices X_G and Y_G when chance chose G, and X_B and Y_B when chance chose B. Finally, the professor receives the signal sent by the student and decides to hire him (H_X or H_Y, depending on the signal) or reject him (R_X or R_Y). The professor knows only the signal he receives, not the education received by the student. Thus the professor has two information sets: one corresponds to receiving X and contains the children of X_G and X_B; the other corresponds to receiving Y and contains the children of Y_G and Y_B.

	H_XH_Y	H_XR_Y	R_XH_Y	R_XR_Y
X_GX_B	5,5	5,5	0,6	0,6
X_GY_B	5,5	2,8	3,3	0,6
Y_GX_B	5,5	3,3	2,8	0,6
Y_GY_B	5,5	0,6	5,5	0,6

Table 6.1: Hiring Game in strategic form

The rewards for the outcomes are assigned as follows. The best outcome for the student is to receive a bad education, which presumably requires less work, but be hired nevertheless. In this case his reward is 6. If he receives a good education and is hired, his reward is 4. If he is not hired, his reward is 0 regardless of his education. For the professor, the best outcome is to have hired a student who received a good education, in which case his reward is 10. If he does not hire the student, the professor can spend his grant money some other way and obtain reward 6. Hiring a bad student is the worst outcome for the professor—in this case he gets reward 0. Thus, with equal probability the players' preferences with respect to hiring or rejecting the student are aligned or in conflict.

We can represent an extensive-form game as a matrix game by converting it to *strategic form*. In this reduction, we consider each non-chance player's action set to be the set of strategies for playing the extensive-form game. A strategy for player i is an element of

$$\Sigma_i = \prod_{h \in \mathcal{I}_i} C_h \tag{6.2}$$

specifying which choice the player is to make at each of its information sets. $\vec{\Sigma}$ denotes the set of joint strategies, i.e.,

$$\vec{\Sigma} = \prod_{i \in N} \Sigma_i. \tag{6.3}$$

Rewards are calculated by taking an expectation over the choices of the chance player.

The matrix game representation of the Hiring Game is given in Table 6.1.

6.2 Permissive EFCE

Now that we can represent extensive-form games as matrix games, the framework developed in Chapter 2 can be applied to extensive-form games. In particular, the definitions of correlated and coarse correlated equilibria can be applied to yield equilibrium concepts for extensive-form games. However, these equilibria may not be appropriate to the setting of the game. Here, we consider alternative equilibrium concepts specific to extensive-form games.

Applying the moderator interpretation of correlated equilibria to an extensive-form game yields a story in which the moderator makes suggestions for strategies to the players before the game is played, and the players must decide whether to follow their suggestions before beginning game play. This set-up may result in a player getting "too much" information. In the Hiring Game, for example, if the student is told what signal to send if he gets a good education, then he will know that information even if he has the bad education, so he can always switch to a strategy in which he simulates having seen G even when he saw B.

von Stengel and Forges [2006] analyze the strategic form of the Hiring Game to show that in any correlated equilibrium, there can only be non-zero weights on joint strategies in which the professor plays $R_X R_Y$. Thus, the professor will never hire the student, and they always get reward vector $\langle 0, 6\rangle$. Observe that this outcome is not Pareto efficient; rewards of $\langle 2, 8\rangle$ are possible.

Forges and von Stengel [2002] introduced an alternative equilibrium concept for extensive-form games, called extensive-form correlated equilibrium (EFCE). The moderator interpretation for an EFCE is this: The moderator gives each player a collection of envelopes, one for each of the player's information sets. Inside each envelope is a suggestion of a choice to be made at the information set. The player is allowed to open an envelope and see the suggestion only when it reaches the corresponding information set, at which point she can choose to accept or reject the suggestion.

Whether or not a player is able to open an envelope after having chosen not to follow an earlier suggestion distinguishes between the two types of EFCE which we name here *permissive* EFCE (pEFCE), and *reduced* EFCE (rEFCE). In a pEFCE, the player continues to have access to envelopes even after choosing not to follow a suggestion; this is not the case in a rEFCE. Because the player has access to more information in a pEFCE and can therefore make finer-grained deviations, it is a tighter equilibrium concept. Though a game may have a rEFCE that is not a pEFCE, von Stengel and Forges [2006] show that pEFCE and rEFCE are payoff-equivalent (i.e., both concepts result in identical sets of achievable payoff vectors). In a game like the Hiring Game, in which each player makes only one decision, the two equilibrium concepts are equivalent, and we can refer to both simply as EFCE.

In the Hiring Game, a student with a bad education does not get to see the signal that the student with a good education is supposed to play. This prevents a student with a bad education from impersonating a student with a good education, the possibility of which prevents the professor from ever hiring in a correlated equilibrium. Both players can in fact do better with an EFCE than with a correlated equilibrium. For example, a distribution which puts weight of $\frac{1}{4}$ on each of the following joint strategies is an EFCE: $\langle X_G X_B, H_X R_Y\rangle$, $\langle X_G Y_B, H_X R_Y\rangle$, $\langle Y_G X_B, R_X H_Y\rangle$, and $\langle Y_G Y_B, R_X H_Y\rangle$. (Note that the student with a bad education is given the "correct" signal half the time; otherwise he would always do the opposite of his suggestion.) The expected reward vector for this equilibrium is $\left\langle 3\frac{1}{2}, 5\frac{1}{2}\right\rangle$—both players are better off than they would be at any correlated equilibria.

von Stengel and Forges [2006] give a formal definition of pEFCE in terms of an *extended game*, which, given a joint distribution $q \in \Delta(\vec{\Sigma})$, is constructed from the original extensive-form game. We add a chance player (the moderator) who first picks a joint strategy s^* according to q. The original game is then played, except that each player is informed before making a choice of a "recommended" choice, which is the corresponding component of s^*. A pEFCE is a q such that always following the recommendation is a Nash equilibrium of the extended game constructed using q.

Given a information set $h \in \mathcal{I}_i$, let $\mathcal{H}(h)$ denote the set of informations sets observed by player i up to and including h. In the extended game, each information set h is replaced with $\prod_{h' \in \mathcal{H}(h)} |C_{h'}|$ versions of it, one for each sequence of suggestions that the player could have received upon reaching h. Each of these versions has its own version of C_h as its choices.

Thus, the set of strategies Σ_i' for player i in the extended game is equivalent[1] to the set of vectors of functions:

$$\prod_{h \in \mathcal{I}_i} \left(\left(\prod_{h' \in \mathcal{H}(h)} C_{h'} \right) \mapsto C_h \right). \tag{6.4}$$

That is, a strategy for the extended game must specify for each information set h, for every sequence of choices corresponding to the information sets leading up to h, a choice in C_h.

One member of each player's strategy set corresponds to always choosing the recommended choice; we denote it $\textsc{obey}_i$, and define it by the property that $(\textsc{obey}_i)_h(\vec{c}) = \vec{c}_h$ for all $h \in \mathcal{I}_i$ and all $\vec{c} \in \prod_{h' \in \mathcal{H}(h)} C_{h'}$. If each player playing the pure strategy $\textsc{obey}_i$ is a Nash equilibrium of the strategic form of the extended game (i.e., with respect to expectation over the added chance player), then q is an pEFCE.

Definition 6.1 *Given an extensive-form game, a joint distribution q is a pEFCE if and only if for all players i, for all $s_i' \in \Sigma_i'$,*

$$r_i'(\text{OBEY}_i, \text{OBEY}_{\neg i}) \geq r_i'(s_i', \text{OBEY}_{\neg i}), \tag{6.5}$$

where r_i' is player i's reward in the (strategic form of the) extended game built using q.

We now give a set of action transformations corresponding to pEFCE. Each element of Σ_i' can be thought of as an action transformation for the strategic form of the original game. Given $s_i' \in \Sigma_i'$, we can define $\phi_{s_i'} \in \Phi^{\text{SWAP}}(\Sigma_i)$.

$$\phi_{s_i'}(s)(h) = s_i' \left(\langle s_{h'} \rangle_{h' \in \mathcal{H}(h)} \right) \tag{6.6}$$

Given an extensive-form game, we can then define a set of transformations for each player i, denoted Φ_i^{pEFCE}, to be the set of all $\phi_{s_i'}$.

Proposition 6.2 *Given an extensive-form game, the set of pEFCE is identical to the set of Φ^{pEFCE} equilibria.*

Proof Recall the definition of a pEFCE—for all players i, for all $s_i' \in \Sigma_i'$,

$$r_i'(\text{OBEY}_i, \text{OBEY}_{\neg i}) \geq r_i'(s_i', \text{OBEY}_{\neg i}). \tag{6.7}$$

Observe that

$$r_i'(\text{OBEY}_i, \text{OBEY}_{\neg i}) = \mathbb{E}_{s \sim q} \left[r_i(s) \right] \tag{6.8}$$

[1] Technically, we need each of the new information sets to have distinct sets of choices.

and for any $s'_i \in \Sigma'_i$,

$$r'_i(s'_i, \text{OBEY}_{\neg i}) = \mathbb{E}_{s\sim q}\left[r_i\left(\phi_{s'_i}(s_i), s_{\neg i}\right)\right]. \tag{6.9}$$

Thus, q is a pEFCE if and only if for all players i, for all $\phi \in \Phi_i^{\text{pEFCE}}$,

$$\mathbb{E}_{s\sim q}\left[r_i(s)\right] \geq \mathbb{E}_{s\sim q}\left[r_i\left(\phi_{s'_i}(s_i), s_{\neg i}\right)\right], \tag{6.10}$$

which is the definition of a Φ^{pEFCE} equilibrium. ∎

6.3 Reduced EFCE

The pEFCE concept has the undesirable property that the equilibrium is defined in terms of an extended game which is exponentially bigger than the original game. Consequently, Φ^{pEFCE} is an extremely large set. Consider an extensive-form game in which player i makes a sequence of binary decisions at n consecutive information sets and hence has 2^n strategies. In the extended game, the ith information set is replaced by 2^i information sets, so that player i has 2^{n+1} information sets and $2^{2^{n+1}}$ strategies in the extended game. Thus Φ^{pEFCE} for player i contains $2^{2^{n+1}}$ transformations.

In this section we present a more compact representation of game trees, reduced strategic form, and consider reduced EFCE, also due to von Stengel and Forges [2006].

6.3.1 Reduced Strategic Form

Thus far, we have overlooked the fact that the strategic-form representation of an extensive-form game can be grossly inefficient. This will be the case when a player may make a choice which renders some of his information sets unreachable, so that an element of Σ_i contains useless information. The Hiring Game does not have this structure, so to illustrate we consider the two-player game tree in Figure 6.2. Player 1 makes two decisions in each play of the game, first between choices A and B, and then between a pair of choices which depend on the choice it made previously and the choice made by player 2.

In the strategic-form representation of this game, Player 1 has $2^5 = 32$ strategies, because a strategy must indicate a choice for each Player 1's five information sets. However, many of these strategies are essentially identical. For example, the strategies $\langle A, a, c, e, g\rangle$ and $\langle A, a, c, e, h\rangle$ can be considered equivalent; they only disagree at an information set that neither can reach (because both specify choice A).

This inefficiency is the motivation for *reduced strategic form*, which collapses strategies that differ only at unreachable information sets. In this example, Player 1 only has eight strategies in reduced strategic form. Writing them without the superfluous choices, they are:

- $\langle A, a, c\rangle$
- $\langle A, b, c\rangle$
- $\langle A, a, d\rangle$

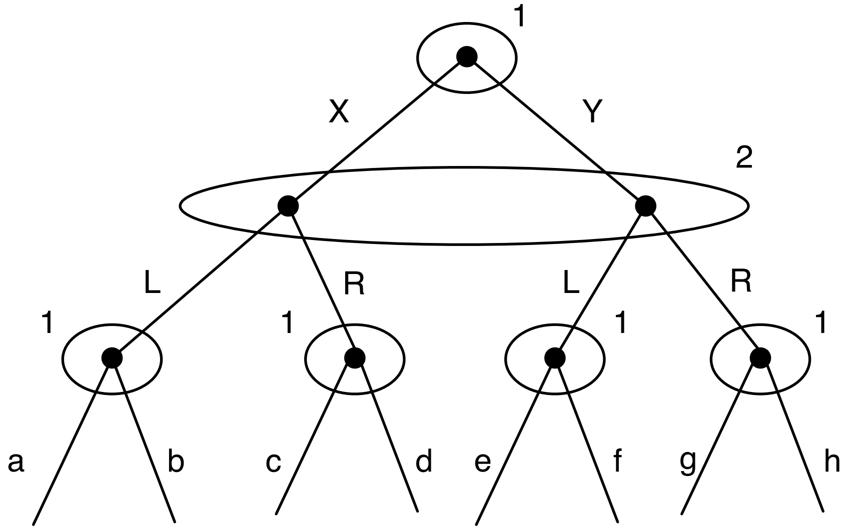

Figure 6.2: Another game tree

- $\langle A, b, d \rangle$
- $\langle B, e, g \rangle$
- $\langle B, f, g \rangle$
- $\langle B, e, h \rangle$
- $\langle B, f, h \rangle$

6.3.2 Reduced EFCE

The same reasoning that motivated reduced strategic form as an alternative to strategic form also motivates reduced EFCE (rEFCE) as an alternative to pEFCE. The moderator interpretation of an rEFCE is very similar to an that of a pEFCE, with the sole distinction being that once a player disregards the suggestion found in an envelope, the player is no longer allowed to open envelopes. We can thus think of the moderator as choosing reduced strategies for each player and putting the components of the reduced strategies in envelopes corresponding to the appropriate information set. However, information sets belonging to player i that are unreachable given player i's suggested strategy have empty envelopes.

Here we present a formalization of rEFCE in the Φ transformation framework by constructing a set of transformations, Φ^{rEFCE}, for the strategic form of an extensive-form game. We also present a subset of Φ^{rEFCE} that has equivalent power (i.e., the resulting sets of equilibria are identical).

In the moderator interpretation of coarse correlated equilibria, the player must choose whether to depart from his suggested move before he sees it. This limitation is implemented in Φ^{EXT} by requiring that the transformation give the same output regardless of its input (i.e., that it be a constant function). In order to correctly construct Φ^{rEFCE}, we must represent the inability of the player to know the contents of its unopened envelopes. We require that a transformation act the same on all parts of the strategy corresponding to the information sets that it has not reached at the point at which the agent departs from the moderator's suggestion.

We can think of a transformation in Φ^{rEFCE} as having three components. One component is the set of points at which the agent chooses to defect from the moderator's suggestion. These points are represented by information sets. Another component is the set of suggestions that will trigger a defection. These suggestions are choices in C_i^*, and because the sets of choices are disjoint, designating such a "trigger" choice also indicates the information set at which the defection occurs. The third component of a transformation in Φ^{rEFCE} is the player's behavior after defecting.

We denote the set of partial strategies providing choices sufficient to play the rest of the game starting in information set h by $\Sigma|_h$. A partial strategy σ must be *consistent.* That is, σ must not contain choices for information sets that are unreachable from h using the choices in σ. It must also be *complete.* That is, σ must contain choices for all information sets that are reachable from h using the choices in σ. Let $\Sigma_i^? = \prod_{h \in \mathcal{I}_i} \Sigma|_h$ be the set of all such partial strategies.

Given $c \in C_i^*$, let $h|_c \in \mathcal{I}_i$ be the information set at which choice c is available, i.e., h such that $c \in C_h$. Let $\Sigma|_c$ be shorthand for $\Sigma|_{h|_c}$. Given an information set $h \in \mathcal{I}_i$, let $\delta(h) \subset \mathcal{I}_i$ denote player i's information sets which contain nodes that are descendents of nodes in h. Similarly. let $\alpha(h)$ denote the set of ancestor information sets of h. Let YES represent not altering a suggestion.

We can thus identify the set $\Phi^{\text{rEFCE}}(\Sigma_i)$ with the set of functions

$$\psi : C_i^* \to \{\text{YES}\} \cup \Sigma_i^? \tag{6.11}$$

satisfying two conditions:

- for all c, $\psi(c) \in \{\text{YES}\} \cup \Sigma|_c$, and
- for all c, $\psi(c) \neq \text{YES}$ implies $\psi(c') = \text{YES}$ for all $c' \in C_{h'}$, where $h' \in \alpha(h|_c) \cup \delta(h|_c)$.

Once the player defects, he no longer gets to open envelopes, so a transformation can't be triggered more than once.

Just as the much smaller set Φ^{INT} proved to be as expressive as Φ^{SWAP} in the case of matrix games, we present the smaller set $\Phi^{\text{rEFCE-INT}}$, which is as expressive as Φ^{rEFCE}. Transformations in Φ^{INT} only change a single input, otherwise they act as the identity. Similarly, transformations in $\Phi^{\text{rEFCE-INT}}$ have only a single trigger choice; if the trigger choice is not in their input, they do not alter it.

The set $\Phi^{\text{rEFCE-INT}} \subset \Phi^{\text{rEFCE}}$ for player i can be identified with the set

$$\left\{(c, \sigma) \in C_i^* \times \Sigma_i^? \text{ s.t. } \sigma \in \Sigma|_c\right\}. \tag{6.12}$$

Each such element $(c, \sigma) \in C_i^* \times \Sigma_i^?$ corresponds to a $\phi_{c,\sigma}^{\text{rEFCE}} : \Sigma_i \to \Sigma_i$ defined, if h is reachable using

σ,

$$\left(\phi^{\text{rEFCE}}_{c,\sigma}(s)\right)(h) = \begin{cases} \sigma(h) & \text{if } s\left(h|_c\right) = c \\ s(h) & \text{otherwise} \end{cases} \tag{6.13}$$

and if h is not reachable using σ,

$$\left(\phi^{\text{rEFCE}}_{c,\sigma}(s)\right)(h) = s(h). \tag{6.14}$$

Proposition 6.3 *Given an extensive-form game in strategic form, a joint strategy distribution q is a Φ^{rEFCE} equilibrium if and only if it is a $\Phi^{rEFCE\text{-}INT}$ equilibrium.*

Proof One direction follows from Observation 2.4 and the fact that $\Phi^{\text{rEFCE-INT}} \subset \Phi^{\text{rEFCE}}$. For the other direction, assume that q is a $\Phi^{\text{rEFCE-INT}}$ equilibrium. Given a player i, let $\phi^* \in \Phi^{\text{rEFCE}}(\Sigma_i)$ be a transformation corresponding to a function $\psi : C_i^* \to \{\text{YES}\} \cup \Sigma_i^?$. Given a strategy $s_i \in \Sigma_i$ and a partial strategy $\sigma_i \in \Sigma_i^?$, let $s_i \setminus \sigma_i \in \Sigma_i$ denote the strategy which is equivalent to σ_i where σ_i is defined, and equivalent to s_i elsewhere.

Given a joint strategy $s \in \vec{\Sigma}$, let $c^*(s)$ be the first choice reached by s such that $\psi(c^*(s)) \neq i$. If there is such a c^*, let ϕ_s be $\phi^{\text{rEFCE}}_{c^*,\psi(c^*)}$. Otherwise, let ϕ_s be the identity.

Given a choice c, let $\Sigma|^c \subseteq \vec{\Sigma}$ be the set of joint strategies that lead to choice c being made.

Now we have

$$\begin{aligned}
\mathbb{E}\left[\rho_{i,\phi^*}\right] &= \sum_{s\in\vec{\Sigma}} q(s)\left(r_i\left(\phi^*(s_i), s_{\neg i}\right) - r_i\left(s\right)\right) &(6.15)\\
&= \sum_{s\in\vec{\Sigma}} q(s)\left(r_i\left(\phi_s(s_i), s_{\neg i}\right) - r_i\left(s\right)\right) &(6.16)\\
&= \sum_{s\in\vec{\Sigma}} q(s) \sum_{c:\psi(c)\neq i} \left(r_i\left(\phi^{\text{rEFCE}}_{c,\psi(c)}(s_i), s_{\neg i}\right) - r_i\left(s\right)\right) &(6.17)\\
&= \sum_{c:\psi(c)\neq i} \sum_{s\in\vec{\Sigma}} q(s) \left(r_i\left(\phi^{\text{rEFCE}}_{c,\psi(c)}(s_i), s_{\neg i}\right) - r_i\left(s\right)\right) &(6.18)\\
&= \sum_{c:\psi(c)\neq i} \mathbb{E}\left[\rho_{i,\phi^{\text{rEFCE}}_{c,\psi(c)}}\right] &(6.19)\\
&\leq 0 &(6.20)
\end{aligned}$$

Line (6.17) follows because the only c for which the difference in rewards will be non-zero is c^* (if such a c^* exists). Line (6.20) follows because q is a $\Phi^{\text{rEFCE-INT}}$ equilibrium. ▮

Because each player in the Hiring Game takes only one turn has only two information sets, the $\Phi^{\text{rEFCE-INT}}$ sets are quite small for this game. For the student, there are four possible triggers (X_G, Y_G, X_B, and Y_B) and two choices for each trigger. There are thus eight transformations for the student, but four of them are the identity, so $|\Phi^{\text{rEFCE-INT}}\left(\Sigma_{\text{student}}\right)| = 5$. Similar reasoning shows that $\Phi^{\text{rEFCE-INT}}\left(\Sigma_{\text{professor}}\right)$ also has cardinality 5.

6.4 Learning

Now that we have a representation of an extensive-form game as a matrix game, along with representations of both types of EFCE in terms of sets of transformations for that matrix game, Theorem 2.14 implies that any of the no-Φ-regret algorithms in Chapter 4 can learn either EFCE concept. However, while the informed setting often makes sense for ordinary matrix games, it generally is inappropriate for an extensive-form game. Unless the players reveal (honestly) their entire strategies to each other at the end of each trial, there is no way for a player to calculate his own regret. However, applying the class of naïve no-regret algorithms presented in Section 4.7 solves this problem.

One may also take a very different approach to learning EFCE. An extensive-form game can also be represented as a convex game using the *sequence form*. Gordon et al. [2008], building on the framework presented in Chapter 5, explain how to learn EFCE in this setting. It remains to be seen how the sequence form approach and the approach taken in this work compare in terms of efficiency.

In general, one could construct any number of equilibrium concepts for extensive-form games by conceiving of a moderator interpretation with particular rules. For example, a player could have some limited ability to preview the contents of envelopes. Given such a moderator, the extended game formalism used here to define pEFCE provides a method for formally defining the corresponding equilibrium concept. More specifically, given a joint distribution, q, construct an extended game that represents the rules of the moderator interpretation. If $\langle \textsc{obey}_i \rangle$ is a pure Nash equilibrium, then q is an equilibrium of this new type. One can then define a set of transformations corresponding to the ability of the players to deviate from the moderator's suggestions. Plugging this set of transformations into a regret-matching algorithm yields an algorithm which learns the equilibrium concept. Thus, the work here can be generalized to a variety of equilibrium concepts.

Appendix A

Technical Lemmas

A.1 Point-Set Topology Lemma

Lemma A.1 *Let (X, d_X) be a compact metric space and let (Y, d_Y) be a metric space. Let $\{x_t\}$ be an X-valued sequence, and let S be a nonempty, closed subset of Y. If $f : X \to Y$ is continuous and if $f^{-1}(S)$ is nonempty, then $d_X(x_t, f^{-1}(S)) \to 0$ as $t \to \infty$ if and only if $d_Y(f(x_t), S) \to 0$ as $t \to \infty$.*

Proof We write $d = d_X$ and $d = d_Y$, since the appropriate choice of distance metric is always clear from the context. To prove the forward implication, assume $d(x_t, f^{-1}(S))) \to 0$ as $t \to \infty$. Choose t_0 s.t. for all $t \geq t_0$, $d(x_t, f^{-1}(S)) < \frac{\delta}{2}$. Observe that for all x_t and for all $\gamma > 0$, there exists $q_t^{(\gamma)} \in f^{-1}(S)$ s.t. $d(x_t, q_t^{(\gamma)}) < d(x_t, f^{-1}(S)) + \gamma$. Now, since $d(x_t, q_t^{(\frac{\delta}{2})}) < \frac{\delta}{2} + \frac{\delta}{2} = \delta$, by the continuity of f, $d(f(x_t), f(q_t^{(\frac{\delta}{2})})) < \epsilon$, for all $\epsilon > 0$. Therefore, $d(f(x_t), S) < \epsilon$, since $f(q_t^{(\frac{\delta}{2})}) \in S$.

To prove the reverse implication, assume $d(f(x_t), S) \to 0$ as $t \to \infty$. We must show that for all $\epsilon > 0$, there exists a t_0 s.t. for all $t \geq t_0$, $d(x_t, f^{-1}(S)) < \epsilon$. Define $T = \{x \in X \mid d(x, f^{-1}(S)) \geq \epsilon\}$. If $T = \emptyset$, the claim holds. Otherwise, observe that T can be expressed as the complement of the union of open balls, so that T is closed and thus compact. Define $g : X \to \mathbb{R}$ as $g(x) = d(f(x), S)$. By assumption S is closed; hence, $g(x) > 0$, for all x. Because T is compact, g achieves some minimum value, say $L > 0$, on T. Choose t_0 s.t. $d(f(x_t), S) < L$ for all $t \geq t_0$. Thus, for all $t \geq t_0$, $g(x_t) < L \Rightarrow x_t \notin T \Rightarrow d(x_t, f^{-1}(S)) < \epsilon$. $\blacksquare$

A.2 Probability Lemmas

Lemma A.2 *If x is a random vector taking values in $\mathbb{R}^n$, then*

$$\left(\mathbb{E}\left[\max_i x_i\right]\right)^q \leq \mathbb{E}\left[\|x^+\|_p^q\right] \tag{A.1}$$

for all $p > 0$ and $q \geq 1$.

Proof Apply Jensen's inequality and the fact that $\|x\|_\infty \leq \|x^+\|_p$ for any $p > 0$. $\blacksquare$

Let $(\Omega, \mathcal{F}, P)$ be a probability space with a filtration $(\mathcal{F}_t : t \geq 0)$: that is, a sequence of σ-algebras with $\mathcal{F}_t \subseteq \mathcal{F}$ for all t and $\mathcal{F}_s \subseteq \mathcal{F}_t$ for all $s < t$. A stochastic process $(Z_t : t \geq 0)$ is said to be adapted to a filtration $(\mathcal{F}_t : t \geq 0)$ if Z_t is $\mathcal{F}_t$-measurable, for all times t: i.e., if the value of Z_t is determined by $\mathcal{F}_t$, the information available at time t.

We denote by $\mathbb{E}_t$ the conditional expectation with respect to $\mathcal{F}_t$: i.e., $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$.

Lemma A.3 (Product Lemma) *Assume the following:*

1. *$(Z_t : t \geq 0)$ is an adapted process such that $\forall t$, $0 \leq Z_t < k$ a.s., for some $k \in \mathbb{R}$;*

2. *$\mathbb{E}_{t-1}[Z_t] \leq c_t$ a.s., for $t \geq 1$, and $\mathbb{E}[Z_0] \leq c_0$, where $c_t \in \mathbb{R}$, for all t.*

For fixed T,

$$\mathbb{E}\left[\prod_{t=0}^{T} Z_t\right] \leq \prod_{t=0}^{T} c_t \tag{A.2}$$

Proof The proof is by induction. The claim holds for $T = 0$ by assumption. We assume it also holds for T and show it holds for $T + 1$:

$$\begin{aligned}
\mathbb{E}\left[\prod_{t=0}^{T+1} Z_t\right] &= \mathbb{E}\left[\mathbb{E}_T\left[\prod_{t=0}^{T+1} Z_t\right]\right] && \text{(A.3)}\\
&= \mathbb{E}\left[\mathbb{E}_T\left[\left(\prod_{t=0}^{T} Z_t\right) Z_{T+1}\right]\right] && \text{(A.4)}\\
&= \mathbb{E}\left[\left(\prod_{t=0}^{T} Z_t\right) \mathbb{E}_T\left[Z_{T+1}\right]\right] && \text{(A.5)}\\
&\leq \mathbb{E}\left[\left(\prod_{t=0}^{T} Z_t\right) c_{T+1}\right] && \text{(A.6)}\\
&= c_{T+1}\mathbb{E}\left[\prod_{t=0}^{T} Z_t\right] && \text{(A.7)}\\
&\leq \prod_{t=0}^{T+1} c_t && \text{(A.8)}
\end{aligned}$$

Line (A.3) follows from the tower property, also known as the law of iterated expectations: If a random variable X satisfies $\mathbb{E}\left[|X|\right] < \infty$ and $\mathcal{H}$ is a sub-σ-algebra of $\mathcal{G}$, which in turn is a sub-σ-algebra of $\mathcal{F}$, then $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathbb{E}[X \mid \mathcal{H}]$ almost surely [Williams, 1991]. Note that $\mathbb{E}\left[\prod_{t=0}^{T+1} Z_t\right] < \infty$. Line (A.5) follows because $\prod_{t=0}^{T} Z_t$ is $\mathcal{F}_T$-measurable and $\mathbb{E}\left[\prod_{t=0}^{T} Z_t\right] < \infty$. Line (A.6) follows by assumption, since Z_t, for $t \geq 0$, is nonnegative with probability 1. Line (A.8) follows from the induction hypothesis. ▮

Lemma A.4 (Supermartingale Lemma) *Assume the following:*

1. *$(M_t : t \geq 0)$ is a supermartingale, i.e. $(M_t : t \geq 0)$ is an adapted process s.t. for all t, $\mathbb{E}[|M_t|] < \infty$ and for $t \geq 1$, $\mathbb{E}_{t-1}[M_t] \leq M_{t-1}$ a.s.;*

2. f is a nondecreasing positive function s.t. for $t \geq 1$, $|M_t - M_{t-1}| \leq f(t)$ a.s..

If $M_0 = m \in \mathbb{R}$ a.s., then for fixed T,

$$P[M_T \geq 2\epsilon T f(T)] \leq e^{\epsilon m/f(0) - \epsilon^2 T}, \tag{A.9}$$

for all $\epsilon \in [0, 1]$.

Proof For $t \geq 0$, let

$$Y_t = \frac{M_t}{f(T)} \tag{A.10}$$

and for $t \geq 1$, let $X_t = Y_t - Y_{t-1}$, so that

$$Y_t = \sum_{\tau=0}^{t} X_\tau. \tag{A.11}$$

Note that $X_0 = Y_0 = m/f(0)$ a.s..

Because M_t is supermartingale and $f(t)$ is positive, for $t \geq 1$,

$$\mathbb{E}_{t-1}[X_t] = \mathbb{E}_{t-1}[Y_t] - Y_{t-1} = \frac{\mathbb{E}_{t-1}[M_t] - M_{t-1}}{f(T)} \leq 0 \quad \text{a.s.} \tag{A.12}$$

Because f is nondecreasing, $f(t) \leq f(T)$ for all t; hence, for $1 \leq t \leq T$,

$$|X_t| = |Y_t - Y_{t-1}| = \left| \frac{M_t}{f(T)} - \frac{M_{t-1}}{f(T)} \right| \leq \frac{|M_t - M_{t-1}|}{f(t)} \leq 1 \quad \text{a.s.} \tag{A.13}$$

Thus, for $t \geq 1$,

$$\mathbb{E}_{t-1}\left[e^{\epsilon X_t}\right] \leq 1 + \epsilon \mathbb{E}_{t-1}[X_t] + \epsilon^2 \mathbb{E}_{t-1}[X_t^2] \leq 1 + \epsilon^2 \quad \text{a.s.} \tag{A.14}$$

The first inequality follows from the fact that $e^y \leq 1 + y + y^2$ for $y \leq 1$, and $\epsilon X_t \leq 1$ a.s., since $\epsilon \in [0, 1]$ and $|X_t| \leq 1$ a.s. by Line (A.13). The second inequality follows from Line (A.12).

Therefore,

$$\begin{aligned}
P\left[M_T \geq 2\epsilon T f(T)\right] &= P\left[Y_T \geq 2\epsilon T\right] && \text{(A.15)}\\
&= P\left[e^{\epsilon Y_T} \geq e^{2\epsilon^2 T}\right] && \text{(A.16)}\\
&\leq \frac{\mathbb{E}\left[e^{\epsilon Y_T}\right]}{e^{2\epsilon^2 T}} && \text{(A.17)}\\
&= \frac{\mathbb{E}\left[e^{\epsilon \sum_{t=0}^{T} X_t}\right]}{e^{2\epsilon^2 T}} && \text{(A.18)}\\
&= \frac{\mathbb{E}\left[\prod_{t=0}^{T} e^{\epsilon X_t}\right]}{e^{2\epsilon^2 T}} && \text{(A.19)}\\
&\leq \frac{e^{\epsilon m/f(0)}(1+\epsilon^2)^T}{e^{2\epsilon^2 T}} && \text{(A.20)}\\
&\leq \frac{e^{\epsilon m/f(0)} e^{\epsilon^2 T}}{e^{2\epsilon^2 T}} && \text{(A.21)}\\
&= e^{\epsilon m/f(0) - \epsilon^2 T} && \text{(A.22)}
\end{aligned}$$

Line (A.17) follows from Markov's inequality. Line (A.20) follows from the Product Lemma (since $(X_t : t \geq 0)$ is an adapted process), Line (A.14), and the assumption that $M_0 = m$ a.s.. Line (A.21) follows from the fact that $(1 + x) \leq e^x$. ∎

Appendix B

Gordon Triple Proofs

In this appendix, we prove Lemmas 3.7, 3.8, and 3.9. Of particular interest is the proof of Lemma 3.7, because in this proof we identify a set of points (namely, the boundary of the negative orthant; see Lemma B.3) on which the polynomial link function for $p \geq 2$ is not differentiable.

Cesa-Bianchi and Lugosi [2003] apply Taylor's theorem to the polynomial (potential) function $G(x) = \|x^+\|_p^2$ for $p \geq 2$, whose gradient is the polynomial link function. In this application, the authors implicitly assume that G is twice differentiable everywhere; however, it is not.

B.1 Proof of Proposition 3.7

In Lemmas B.1, B.2, and B.3, we let i and j range over the set $\{1, \ldots, n\}$.

In addition, we rely on the following functions:

- for $p \geq 1$, $G : \mathbb{R}^n \to \mathbb{R}$, defined by:
$$G(x) = \|x^+\|_p^2 \tag{B.1}$$
- for $p \geq 2$, $g : \mathbb{R}^n \to \mathbb{R}^n$, where for all i,
$$g_i(x) = \begin{cases} 0 & \text{if } x_i \leq 0 \\ \frac{2x_i^{p-1}}{\|x^+\|_p^{p-2}} & \text{otherwise} \end{cases} \tag{B.2}$$
- for $p > 2$, $h : \mathbb{R}^n \to \mathbb{R}^{2n}$, where
$$h_{ii}(x) = \begin{cases} 0 & \text{if } x_i \leq 0 \\ 2(2-p)\left(\frac{x_i}{\|x^+\|_p}\right)^{2p-2} + 2(p-1)\left(\frac{x_i}{\|x^+\|_p}\right)^{p-2} & \text{otherwise} \end{cases} \tag{B.3}$$

  and for $i \neq j$
$$h_{ij}(x) = \begin{cases} 0 & \text{if } x_i \leq 0 \text{ or } x_j \leq 0 \\ 2(2-p)\left(\frac{x_i x_j}{\|x^+\|_p^2}\right)^{p-1} & \text{otherwise} \end{cases} \tag{B.4}$$

Lemma B.1 *For $p \geq 2$, G is C^1 (i.e., continuously differentiable) with gradient g.*

Proof We consider the negative orthant and its complement separately.

- G is differentiable with gradient g on the complement of the negative orthant.

  We (de)construct G as follows:

$$\begin{aligned} a_i(x) &= (x_i^+)^p && \text{(B.5)} \\ b(y) &= \sum_i a_i(y) && \text{(B.6)} \\ c(z) &= z^{\frac{2}{p}} && \text{(B.7)} \\ G &= c \circ b && \text{(B.8)} \end{aligned}$$

  Because the a_i are differentiable everywhere, b is C^1. Also, c is differentiable on $(0, \infty)$, so G is C^1 on the complement of the negative orthant. By straightforward calculus, for u in the complement of the negative orthant, $\left.\frac{\partial G}{\partial x_i}\right|_u = g_i(u)$.

- G is differentiable with gradient g on the negative orthant.

  Let $u \in \mathbb{R}^n_-$. We show the following:

$$\lim_{\delta \to 0} \frac{G(u + \delta e_i) - G(u)}{\delta} = g_i(u) \tag{B.9}$$

  Since $u_i \leq 0$, $g_i(u) = 0$ and $G(u) = 0$. If $u_i < 0$, then $G(u + \delta e_i) = 0$ for sufficiently small δ. If $u_i = 0$, then

$$\lim_{\delta \to 0} \frac{G(u + \delta e_i) - G(u)}{\delta} = \lim_{\delta \to 0} \frac{G(u + \delta e_i)}{\delta} = \lim_{\delta \to 0} \frac{(\delta^+)^2}{\delta} = 0 \tag{B.10}$$

  Hence, for u in the negative orthant, $\left.\frac{\partial G}{\partial x_i}\right|_u = g_i(u)$.

- g_i is continuous, for all i.

  Let $u \in \mathbb{R}^n$. If $u_i > 0$, then $g_i(x) = \frac{2(x_i)^{p-1}}{\|x^+\|_p^{p-2}}$ on a neighborhood of u, so g_i is continuous at u. Otherwise, for all $x \in \mathbb{R}^n$ such that $x_i > 0$, since, by assumption, $p > 2$, it follows that

$$0 < \frac{2(x_i)^{p-1}}{\|x^+\|_p^{p-2}} \leq \frac{2(x_i)^{p-1}}{|x_i|^{p-2}} = 2x_i, \tag{B.11}$$

  Hence, for all $x \in \mathbb{R}^n$, $0 \leq g_i(x) \leq 2x_i^+$. Now, if $\{x^{(\tau)}\}$ is a sequence such that $\lim x^{(\tau)} = u$, then $u_i \leq 0 \Rightarrow u_i^+ = 0 \Rightarrow \lim \left(x_i^{(\tau)}\right)^+ = 0 \Rightarrow \lim g_i\left(x^{(\tau)}\right) = 0 \Rightarrow \lim g_i\left(x^{(\tau)}\right) = g_i(u)$.

Lemma B.2 *For $p \geq 1$, G is smooth on the complement of the axes: i.e., on the set $\{x \in \mathbb{R}^n \mid x_i \neq 0$, for all $i\}$. Further, on the set where G is smooth, $\left.\frac{\partial G}{\partial x_i \partial x_j}\right|_u = h_{ij}(u)$, for all i, j.*

Proof For a point u not on an axis, we define an (everywhere) smooth function $\widetilde{G}_u$ by replacing the $+$ operator for each component of the argument of G with either the identity, if u is positive in that component, or zero, if it is not. Specifically, given $u \in \mathbb{R}^n$ such that $u_i \neq 0$ for all i, let

$$\widetilde{G}_u(x) = \left(\sum_{i:u_i>0} x_i^p \right)^{\frac{2}{p}} \tag{B.12}$$

Observe that $G = \widetilde{G}_u$ on a neighborhood of u, and for v in this neighborhood, $\left.\frac{\partial \widetilde{G}_u}{\partial x_i \partial x_j}\right|_v = h_{ij}(v)$. ∎

Lemma B.3 *Assume $p > 2$. All second-order partial derivatives of G exist and are continuous at a point u if and only if u is not on the boundary of the negative orthant. Furthermore, on the set where G is twice-differentiable, $\left.\frac{\partial G}{\partial x_i \partial x_j}\right|_u = h_{ij}(u)$, for all i, j.*

Proof By Lemma B.1, G is differentiable with gradient g on this set. By Lemma B.2, the result holds for u not on an axis. For u on an axis, we show that for all i, j,

$$\lim_{\delta \to 0} \frac{g_i(u + \delta \mathbf{e}_j) - g_i(u)}{\delta} = h_{ij}(u) \tag{B.13}$$

if and only if u is not on the boundary of the negative orthant.

- Case 1: $u_i = 0$.

  In this case, for all j, $h_{ij}(u) = 0$. For $i \neq j$, the limit in Equation B.13 is equal to 0. For $i = j$, the limit from the left:

  $$\lim_{\delta \to 0^-} \frac{g_i(u + \delta \mathbf{e}_i) - g_i(u)}{\delta} = \frac{0}{\delta} = 0 \tag{B.14}$$

  and from the right:

  $$\lim_{\delta \to 0^+} \frac{g_i(u + \delta \mathbf{e}_i) - g_i(u)}{\delta} \tag{B.15}$$

  $$= \lim_{\delta \to 0^+} \frac{2\delta^{p-1}}{\|(u + \delta \mathbf{e}_i)^+\|_p^{p-2}} \frac{1}{\delta} \tag{B.16}$$

  $$= \lim_{\delta \to 0^+} \frac{2\delta^{p-2}}{\|(u + \delta \mathbf{e}_i)^+\|_p^{p-2}} \tag{B.17}$$

  $$= \lim_{\delta \to 0^+} \frac{2\delta^{p-2}}{(C + \delta^p)^{1-\frac{2}{p}}} \tag{B.18}$$

  where $C = \sum_{k \neq i} (u_k^+)^p$. If u is not in the negative orthant (and thus not on the boundary of the negative orthant), then $C > 0$ and the limit in Equation B.18 is equal to 0.[1]

  If u is in the negative orthant (and thus on the boundary of the negative orthant), then $C = 0$ and the limit in Equation B.18 is not *equal to 0. On the contrary,*

  $$\lim_{\delta \to 0^+} \frac{2\delta^{p-2}}{(\delta^p)^{1-\frac{2}{p}}} = \lim_{\delta \to 0^+} \frac{2\delta^{p-2}}{\delta^{p-2}} = 2 \tag{B.19}$$

  In particular, g_i is not differentiable on the boundary of the negative orthant.

[1] Note that Equation B.18 simplifies to 2, if $p = 2$; hence, we assume that $p > 2$.

- Case 2: $u_i \neq 0, u_j = 0$.

  In this case, for all i, $h_{ij}(u) = 0$. Because $u_i \neq u_j$, it cannot be the case that $i = j$. If $u_i < 0$, then the limit in Equation B.13 is equal to 0. If $u_i > 0$, then $(u + \delta \mathbf{e}_j)^+ = u^+$ for $\delta < 0$ (since $u_j = 0$), so the left hand limit of Equation B.13 is also equal to 0. Now, for the right-hand limit:

  $$\lim_{\delta \to 0^+} \frac{1}{\delta} \left(\frac{2u_i^{p-1}}{\|(u + \delta \mathbf{e}_j)^+\|_p^{p-2}} - \frac{2u_i^{p-1}}{\|u^+\|_p^{p-2}} \right) \tag{B.20}$$

  $$= \lim_{\delta \to 0^+} \frac{2u_i^{p-1}}{\delta} \left(\frac{1}{(C + \delta^p)^{1-\frac{2}{p}}} - \frac{1}{C^{1-\frac{2}{p}}} \right) \tag{B.21}$$

  $$= \lim_{\delta \to 0^+} \frac{2u_i^{p-1}}{\delta} \frac{C^{1-\frac{2}{p}} - (C + \delta^p)^{1-\frac{2}{p}}}{C^{1-\frac{2}{p}} (C + \delta^p)^{1-\frac{2}{p}}} \tag{B.22}$$

  $$= \frac{2u_i^{p-1}}{C^{1-\frac{2}{p}}} \lim_{\delta \to 0^+} \frac{C^{1-\frac{2}{p}} - (C + \delta^p)^{1-\frac{2}{p}}}{\delta (C + \delta^p)^{1-\frac{2}{p}}} \tag{B.23}$$

  $$= \frac{2u_i^{p-1}}{C^{1-\frac{2}{p}}} \lim_{\delta \to 0^+} \frac{(p-2)\delta^{p-1} (C + \delta^p)^{\frac{2}{p}}}{(C + \delta^p)^{1-\frac{2}{p}} + (p-2)\delta^p (C + \delta^p)^{\frac{2}{p}}} \tag{B.24}$$

  $$= 0 \tag{B.25}$$

  where $C = \sum_{k \neq j} (u_k^+)^p > 0$ since $u_i > 0$. Line B.24 follows from L'Hôpital's rule.

- Case 3: $u_i \neq 0$, $u_j \neq 0$.

  To take partial derivatives in the ith and jth coordinates, we can project onto $\mathbb{R}^2$, holding all other coordinates constant. We then apply the technique of Lemma B.2 to this projection.

Finally, we must show that all h_{ij} are continuous on the complement of the boundary of the negative orthant. First, consider h_{ii}, for an arbitrary i. Clearly, h_{ii} is continuous on the (open) sets $\{x \in \mathbb{R}^n \mid x_i < 0\}$ and $\{x \in \mathbb{R}^n \mid x_i > 0\}$. It remains to show h_{ii} is continuous on the complement of these sets, namely $\{x \in \mathbb{R}^n \mid x_i = 0\}$, except where it intersects the boundary of the negative orthant. At each point u in the set $\{x \mid x_i = 0 \text{ and } \exists j\ x_j > 0\}$, $x_i = 0$, so $h_{ii}(u) = 0$. Moreover, as x approaches each such point u, $\frac{x_i}{\|x^+\|_p}$ approaches 0 so that $h_{ii}(x)$ does too. Thus, all h_{ii} are continuous on the complement of the boundary of the negative orthant.

Now consider h_{ij}, for arbitrary $i \neq j$. Clearly, h_{ij} is continuous on the (open) sets $\{x \in \mathbb{R}^n \mid x_i < 0 \text{ or } x_j < 0\}$ and $\{x \in \mathbb{R}^n \mid x_i > 0 \text{ and } x_j > 0\}$. It remains to show h_{ij} is continuous on the complement of these sets, except where it intersects the boundary of the negative orthant. At each point u in the relevant set, either $x_i = 0$ or $x_j = 0$, so $h_{ij}(u) = 0$. Moreover, as x approaches each such point u, $\frac{x_i x_j}{\|x^+\|_p^2}$ approaches 0 so that $h_{ij}(x)$ does too. Thus, all h_{ij} are continuous on the complement of the boundary of the negative orthant.

Lemma B.4 *If $p > 2$, then the functions*

$$G(x) = \|x^+\|_p^2 \tag{B.26}$$

$$g_i(x) = \begin{cases} 0 & \textit{if } x_i \le 0 \\ \frac{2x_i^{p-1}}{\|x^+\|_p^{p-2}} & \textit{otherwise} \end{cases} \tag{B.27}$$

$$\gamma(x) = (p-1)\|x\|_p^2 \tag{B.28}$$

satisfy the condition $G(x+y) \le G(x) + g(x) \cdot y + \gamma(y)$, for $x, y \in \mathbb{R}^n$ such that the boundary of the negative orthant does not intersect the line segment between them (inclusive).

Proof Let U be an open convex set containing x and $x+y$ but no points on the boundary of the negative orthant. By Lemma B.1, G is C^1 and the gradient of G is g. By Lemma B.3, G is C^2 on U,

$$\left.\frac{\partial^2 G}{\partial x_i^2}\right|_u = \begin{cases} 0 & \text{if } u_i \le 0 \\ 2(2-p)\left(\frac{u_i}{\|u^+\|_p}\right)^{2p-2} + 2(p-1)\left(\frac{u_i}{\|u^+\|_p}\right)^{p-2} & \text{otherwise} \end{cases} \tag{B.29}$$

and for $i \ne j$,

$$\left.\frac{\partial^2 G}{\partial x_i \partial x_j}\right|_u = \begin{cases} 0 & \text{if } u_i \le 0 \text{ or } u_j \le 0 \\ 2(2-p)\left(\frac{u_i u_j}{\|u^+\|_p^2}\right)^{p-1} & \text{otherwise} \end{cases} \tag{B.30}$$

By Taylor's theorem, for some $u \in U$,

$$G(x+y) = G(x) + g(x) \cdot y + \frac{1}{2}\sum_{i,j} \left.\frac{\partial^2 G}{\partial x_i \partial x_j}\right|_u y_i y_j \tag{B.31}$$

If $u \le 0$ so that $u^+ = 0$, then

$$\frac{1}{2}\sum_{i,j} \left.\frac{\partial^2 G}{\partial x_i \partial x_j}\right|_u y_i y_j = 0 \le (p-1)\|y\|_p^2 \tag{B.32}$$

Otherwise,

$$\frac{1}{2}\sum_{i,j} \left.\frac{\partial^2 G}{\partial x_i \partial x_j}\right|_u y_i y_j \tag{B.33}$$

$$= \sum_{i,j}(2-p)\left(\frac{u_i^+ u_j^+}{\|u^+\|_p^2}\right)^{p-1} y_i y_j + \sum_i (p-1)\left(\frac{u_i^+}{\|u^+\|_p}\right)^{p-2} y_i^2 \tag{B.34}$$

$$= (2-p)\|u^+\|_p^{2-2p}\left(\sum_i (u_i^+)^{p-1} y_i\right)^2 + (p-1)\|u^+\|_p^{2-p}\sum_i (u_i^+)^{p-2} y_i^2 \tag{B.35}$$

$$\le (p-1)\|u^+\|_p^{2-p}\sum_i (u_i^+)^{p-2} y_i^2 \tag{B.36}$$

$$\le (p-1)\|u^+\|_p^{2-p}\left(\sum_i \left((u_i^+)^{p-2}\right)^{\frac{p}{p-2}}\right)^{\frac{p-2}{p}}\left(\sum_i |y_i|^p\right)^{\frac{2}{p}} \tag{B.37}$$

$$= (p-1)\|y\|_p^2 \tag{B.38}$$

Line (B.36) follows from the fact that $p \ge 2$ (by assumption, $p > 2$). Line (B.37) is an application of Hölder's inequality. ∎

Proposition 3.7 *If $p > 2$, then the functions G, g, and γ defined in Lemma B.4 form a Gordon triple.*

Proof Define $z = x + y$. If the line segment between x and z (inclusive) does not intersect the boundary of the negative orthant. then apply Lemma B.4. Otherwise, three cases arise.

- Case 1: $x \in \mathbb{R}^n_-$

  If x is in the negative orthant, then $G(x) = 0$ and $g(x) = 0$. Hence, it suffices to show that

  $$G(x+y) = \|(x+y)^+\|_p^2 \leq \|y^+\|_p^2 \leq \|y\|_p^2 \leq (p-1)\|y\|_p^2 = \gamma(y) \tag{B.39}$$

- Case 2: $x \notin \mathbb{R}^n_-, z \in \mathbb{R}^n_-$

  If z is in the negative orthant, but x is not, then $G(z) = 0$. Hence, it suffices to show that

  $$||x^+||_p^2 + \left(\frac{2(x^+)^{p-1}}{||x^+||_p^{p-2}}\right) \cdot y + (p-1)||y||_p^2 \geq 0 \tag{B.40}$$

  which reduces to

  $$||x^+||_p^p + 2(x^+)^{p-1} \cdot y + (p-1)||y||_p^2||x||_p^{p-2} \geq 0 \tag{B.41}$$

  By Hölder's inequality,

  $$||y||_p^2||x||_p^{p-2} \geq \sum_i |y_i|^2 \cdot |x_i|^{p-2} \tag{B.42}$$

  So in fact, it suffices to show that for any $a, b \in \mathbb{R}$,

  $$(a^+)^p + 2(a^+)^{p-1}b + (p-1)b^2|a|^{p-2} \geq 0 \tag{B.43}$$

  Indeed,

  $$\begin{aligned} & (a^+)^p + 2(a^+)^{p-1}b + (p-1)b^2|a|^{p-2} && \text{(B.44)} \\ \geq\ & (a^+)^p + 2(a^+)^{p-1}b + (p-1)b^2(a^+)^{p-2} && \text{(B.45)} \\ =\ & (a^+)^{p-2}((a^+)^2 + 2a^+b + (p-1)b^2) && \text{(B.46)} \\ \geq\ & (a^+)^{p-2}((a^+)^2 + 2a^+b + b^2) && \text{(B.47)} \\ =\ & (a^+)^{p-2}(a^+ + b)^2 && \text{(B.48)} \\ \geq\ & 0 && \text{(B.49)} \end{aligned}$$

  Line (B.46) follows from the fact that $p \geq 2$.

- Case 3: $x \notin \mathbb{R}^n_-, z \notin \mathbb{R}^n_-$

  Neither x nor z is in the negative orthant, but, by assumption, the line segment between x and z intersects the boundary of the negative orthant. We claim the line segment between x^+ and z does not intersect this boundary. We prove this claim by contradiction.

  Assume, for the sake of contradiction, that there exists $\lambda \in (0, 1)$ such that $\lambda x^+ + (1-\lambda)z \in \mathbb{R}^n_-$. Define $I = \{i \mid z_i > 0\}$. Since $z \notin \mathbb{R}^n_-$, the set I is non-empty. Moreover, for all $i \in I$, $x_i < 0$.

For all $i \in I$, $z_i > 0$ and $x_i^+ = 0$, so $\lambda = 1$ (because $\lambda x^+ + (1-\lambda)z \in \mathbb{R}^n_-$). Consequently, it must be the case that $x^+ \in \mathbb{R}^n_-$. This is a contradiction, since $x \notin \mathbb{R}^n_-$, by assumption.

Therefore, we can apply Lemma B.4 to x^+ and $z - x^+$, which yields

$$G(x+y) = G(z) \leq G(x^+) + g(x^+) \cdot (z - x^+) + \gamma(z - x^+). \tag{B.50}$$

First, observe that $G(x^+) = G(x)$. Second, for all $i = 1, \ldots, n$, $g_i(x^+) = g_i(x) \geq 0$ and $y_i = z_i - x_i \geq z_i - x_i^+$. Hence, $g(x^+) \cdot (z - x^+) \leq g(x) \cdot (z - x) = g(x) \cdot y$. Third, for all $i = 1, \ldots, n$, $|z_i - x_i^+| \leq |z_i - x_i|$ Thus, $\|z - x^+\|_p^2 \leq \|z - x\|_p^2 = \|y\|_p^2$. Therefore,

$$G(x^+) + g(x^+) \cdot (z - x^+) + \gamma(z - x^+) \leq G(x) + g(x) \cdot y + \gamma(y) \tag{B.51}$$

Together, Equations B.50 and B.51 imply the desired conclusion.

B.2 Proof of Proposition 3.8

Proposition 3.8 *If $1 \leq p \leq 2$, then the following is a Gordon triple: $G(x) = \|x^+\|_p^p$, $g_i(x) = p(x_i^+)^{p-1}$, and $\gamma(x) = \|x\|_p^p$.*

Proof Because $\|x^+\|_p^p = \sum_i \left(x_i^+\right)^p$, it suffices to show that for any $a, b \in \mathbb{R}$,

$$((a+b)^+)^p \leq (a^+)^p + p(a^+)^{p-1}b + |b|^p \tag{B.52}$$

after which the result follows from a component-wise proof.

If $p = 1$, then Line (B.52) follows immediately because $(a+b)^+ \leq a^+ + b^+$. Otherwise, we use the basic inequality $x^\alpha + y^\alpha \geq (x+y)^\alpha$, for all $x, y \geq 0$ and $\alpha \in [0,1]$. Two cases arise.

- Case 1: $b \geq 0$. Define the function $h_c(z) = z^p + p(c^+)^{p-1}z - ((c+z)^+)^p$ for $c \in \mathbb{R}$ for $z \geq 0$. Taking its derivative yields $h_c'(z) = p\left(z^{p-1} + (c^+)^{p-1} - ((c+z)^+)^{p-1}\right)$. Applying the basic inequality with $x = z$, $y = c^+$, and $\alpha = p - 1$ yields

$$\begin{aligned} z^{p-1} + (c^+)^{p-1} &\geq (z + c^+)^{p-1} & \text{(B.53)} \\ &\geq ((c+z)^+)^{p-1} & \text{(B.54)} \end{aligned}$$

Thus, $h_c'(z) \geq 0$, i.e., h_c is non-decreasing on $[0, \infty)$. In particular,

$$\begin{aligned} h_a(0) &\leq h_a(b) & \text{(B.55)} \\ -(a^+)^p &\leq b^p + p(a^+)^{p-1}b - ((a+b)^+)^p & \text{(B.56)} \\ ((a+b)^+)^p &\leq (a^+)^p + p(a^+)^{p-1}b + b^p & \text{(B.57)} \\ &\leq (a^+)^p + p(a^+)^{p-1}b + |b|^p & \text{(B.58)} \end{aligned}$$

- Case 2: $b < 0$. Define the function $h_c(z) = z^p - p(c^+)^{p-1}z - ((c-z)^+)^p$ for $c \in \mathbb{R}$ for $z \geq 0$. Taking its derivative yields $h'_c(z) = p\left(z^{p-1} - (c^+)^{p-1} + ((c-z)^+)^{p-1}\right)$. Applying the basic inequality with $x = z$, $y = (c-z)^+$, and $\alpha = p-1$ yields

$$\begin{aligned} z^{p-1} + ((c-z)^+)^{p-1} &\geq (z + (c-z)^+)^{p-1} && \text{(B.59)} \\ &\geq (c^+)^{p-1} && \text{(B.60)} \end{aligned}$$

  Thus, $h'_c(z) \geq 0$, i.e., h_c is non-decreasing on $[0, \infty)$. In particular,

$$\begin{aligned} h_a(0) &\leq h_a(-b) && \text{(B.61)} \\ -(a^+)^p &\leq (-b)^p + p(a^+)^{p-1}b - ((a+b)^+)^p && \text{(B.62)} \\ ((a+b)^+)^p &\leq (a^+)^p + p(a^+)^{p-1}b + (-b)^p && \text{(B.63)} \\ &\leq (a^+)^p + p(a^+)^{p-1}b + |b|^p && \text{(B.64)} \end{aligned}$$

B.3 Proof of Proposition 3.9

Proposition 3.9 *If $\delta > 0$, then the following is a Gordon triple:*

$$G(x) = \frac{1}{\delta} \ln\left(\sum_i e^{\delta x_i}\right) \tag{B.65}$$

$$g_i(x) = \frac{e^{\delta x_i}}{\sum_j e^{\delta x_j}} \tag{B.66}$$

$$\gamma(x) = \frac{\delta}{2}\|x\|_\infty^2 \tag{B.67}$$

Proof We use the same technique as in the proof of Lemma B.4.

Observe that G is smooth and that the gradient of G is g. Moreover, for $i \neq j$,

$$\left.\frac{\partial^2 G}{\partial x_i \partial x_j}\right|_u = -\frac{\delta e^{\delta u_i} e^{\delta u_j}}{\left(\sum_i e^{\delta u_i}\right)^2} \tag{B.68}$$

and otherwise,

$$\left.\frac{\partial^2 G}{\partial x_i^2}\right|_u = -\frac{\delta e^{\delta u_i} e^{\delta u_i}}{\left(\sum_i e^{\delta u_i}\right)^2} + \frac{\delta e^{\delta u_i}}{\sum_i e^{\delta u_i}} \tag{B.69}$$

By Taylor's theorem, for some $u \in U$,

$$G(x+y) = G(x) + g(x) \cdot y + \frac{1}{2}\sum_{ij} \left.\frac{\partial^2 G}{\partial x_i \partial x_j}\right|_u y_i y_j \tag{B.70}$$

Finally,

$$\sum_{ij} \frac{\partial^2 G}{\partial x_i \partial x_j}\bigg|_u y_i y_j = \sum_{ij} -\frac{\delta e^{\delta u_i} e^{\delta u_j}}{\left(\sum_i e^{\delta u_i}\right)^2} y_i y_j + \sum_i \frac{\delta e^{\delta u_i}}{\sum_i e^{\delta u_i}} y_i^2 \tag{B.71}$$

$$= -\delta \left(\frac{\sum_i e^{\delta u_i} y_i}{\sum_i e^{\delta u_i}} \right)^2 + \sum_i \frac{\delta e^{\delta u_i}}{\sum_i e^{\delta u_i}} y_i^2 \tag{B.72}$$

$$\leq \sum_i \frac{\delta e^{\delta u_i}}{\sum_i e^{\delta u_i}} y_i^2 \tag{B.73}$$

$$\leq \sum_i \frac{\delta e^{\delta u_i}}{\sum_i e^{\delta u_i}} \|y\|_\infty^2 \tag{B.74}$$

$$= \delta \|y\|_\infty^2 \tag{B.75}$$

$\blacksquare$

Bibliography

Robert Aumann. Correlated equilibrium as an expression of bayesian rationality. *Econometrica*, 55 (1):1–18, 1987.

Robert Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1:67–96, 1974.

David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

Avrim Blum and Yishay Mansour. From external to internal regret. In *Proceedings of the 2005 Computational Learning Theory Conferences*, pages 621–636, June 2005.

Nicolo Cesa-Bianchi and Gabor Lugosi. Potential-based algorithms in on-line prediction and game theory. *Machine Learning*, 51(3):239–261, 2003.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, 2006.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications.* John Wiley & Sons, 1999.

Françoise Forges. Correlated equilibrium in two-person zero-sum games. *Econometrica*, 58(2):515, March 1990.

Françoise Forges and Bernhard von Stengel. Computationally efficient coordination in game trees. Technical Report LSE-CDAM-2002-02, London School of Economics, Department of Mathematics, 2002.

Dean P. Foster and Rakesh Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29:7–35, 1999.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

Drew Fudenberg and David K. Levine. Consistency and cautious fictitious play. *Journal of Economic Dyanmics and Control*, 19:1065–1090, 1995.

Drew Fudenberg and David K. Levine. Conditional universal consistency. *Games and Economic Behavior*, 29:104–130, 1999.

Geoffrey J. Gordon. No-regret algorithms for structured prediction problems. Technical Report 112, Carnegie Mellon University, Center for Automated Learning and Discovery, 2005.

Geoffrey J. Gordon, Amy Greenwald, and Casey Marks. No-regret learning in convex games. In *Proceedings of the 25th International Conference on Machine Learning*, July 2008.

Amy Greenwald and Amir Jafari. A general class of no-regret algorithms and game-theoretic equilibria. In *Proceedings of the 2003 Computational Learning Theory Conference*, pages 1–11, August 2003.

Amy Greenwald, Zheng Li, and Casey Marks. Bounds for regret-matching algorithms. In *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics*, 2006.

Amy Greenwald, Amir Jafari, and Casey Marks. A general class of no-regret algorithms and game-theoretic equilibria. Invited chapter in a forthcoming book to be published by the Indian Logic Association, 2008.

James Hannan. Approximation to Bayes risk in repeated plays. In M. Dresher, A.W. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games*, volume 3, pages 97–139. Princeton University Press, 1957.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150, 2000.

Sergiu Hart and Andreu Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98(1):26–54, 2001.

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2): 151–178, 1998.

Ehud Lehrer. A wide range no-regret theorem. *Games and Economic Behavior*, 42(1):101–115, 2003.

Nick Littlestone and Manfred Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212 – 261, 1994.

Hervé Moulin and Jean-Phillipe Vial. Strategically zero-sum games. *International Journal of Game Theory*, 7:201–221, 1978.

John Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

Bernhard von Stengel and Françoise Forges. Extensive form correlated equilibrium: Definition and computational complexity. Technical Report LSE-CDAM-2006-04, London School of Economics, Department of Mathematics, 2006.

David Williams. *Probability with Martingales.* Cambridge University Press, 1991.

Peyton Young. *Strategic Learning and its Limits.* Oxford University Press, Oxford, 2004.