Regret Minimization in Repeated Matrix Games with Variable Stage Duration

Shie Mannor
Department of Electrical and Computer Engineering, McGill University
3480 University Street, Montreal
Quebec, Canada
shie.mannor@mcgill.ca

Nahum Shimkin*
Department of Electrical Engineering
Technion, Israel Institute of Technology
Haifa 32000, Israel
shimkin@ee.technion.ac.il

June 18, 2007

Abstract

We consider a player who faces an arbitrary opponent (or environment), in the sense that actions of the latter are not predictable. Repeated games offer an opportunity for adaptive play against such an opponent, in the sense that the minimax payoff may be improved upon if the opponent deviates from his worst-case strategy. For repeated matrix games, in particular, well known results establish the existence of no-regret strategies; such strategies secure a long-term average payoff that comes close to the maximal payoff that could be obtained by playing a fixed action that is best, in hindsight, against the observed action sequence of the opponent. This paper considers the extension of these ideas to repeated games with variable stage duration, where the duration of each stage of the game may depend on the actions of both players, while the performance measure of interest is the average payoff per unit time. We start the analysis of this model by showing that no-regret strategies, in the above sense, do not exist in general. Consequently, we consider two classes of adaptive strategies, one based on Blackwell's approachability theorem and the other on calibrated forecasts, and examine their performance guarantees. In either case we show that the long-term average payoff is higher than a certain function of the empirical distribution of the opponent's actions, and in particular is strictly higher than the minimax value of the repeated game whenever that empirical distribution deviates from a minimax strategy in the stage game. Along the way, we provide sufficient conditions for existence of no-regret strategies in our model.

JEL Classification: C73; C44.

Keywords: no-regret strategies, regret minimization, Hannan consistency, best-response envelope, repeated matrix games, variable duration games, approachability, calibrated play.

*Corresponding author

1 Introduction

Consider a repeated game from the viewpoint of a certain player, say player 1, who faces an arbitrary opponent, say player 2. The opponent is arbitrary in the sense that player 1 has no prediction, statistical or strategic, regarding the opponent's choice of actions. Such an opponent can represent the combined effect of several other players, as well as arbitrary-varying elements of Nature's state. The questions that arise naturally are how should player 1 act in this situation, and what performance guarantees can he secure against an arbitrary opponent.

In a single-stage game, the most obvious option for player 1 is to play his maximin strategy (with respect to his own payoff function), thereby securing for himself the value of the corresponding zero-sum game. More is possible in a multi-stage game, provided that the previous actions of the other player (or at least some signals that depend on these actions) are observed by player 1. An *adaptive strategy* of player 1 thus seeks to improve upon the zero-sum value by exploiting observed deviations of the arbitrary player from a worst-case strategy. The point is that player 2 need not be antagonistic, and his observed action history may indeed reveal that he is not utilizing a worst-case strategy. Still, since past actions of an arbitrary player do not reveal anything about his subsequent choices, it is far from obvious how to capitalize on this information.

An elegant solution was provided by Hannan (1957), in the context of repeated matrix games. Hannan introduced the Bayes utility against the current (n-stage) empirical distribution of the opponent's actions as a performance goal for adaptive play. This quantity coincides with the highest average payoff that player 1 could achieve, in hindsight, by playing some fixed action against the observed action sequence of player 2. Player 1's *regret* can now be defined as the difference between the above Bayes utility and the actual n-stage average payoff obtained by player 1. Hannan established the existence of *no-regret strategies* for player 1, that guarantee non-positive regret in the long run. More precisely, an explicit strategy was presented for which the n-stage regret is (almost surely) bounded by an $O(n^{-1/2})$ term, without requiring any prior knowledge on player 2's strategy or the number of stages n. Consequently, when the empirical distribution of the opponent's actions deviates from his worst-case strategy in the matrix game, the long-term average payoff for player 1 exceeds the minimax value of the game.

Hannan's seminal work was continued in various directions. No-regret strategies in the above sense have been termed regret minimizing, Hannan consistent, and universally consistent. The original strategy proposed in Hannan (1957) is essentially perturbed fictitious play, namely playing best-response to the current empirical distribution of player 2, to which a random perturbation is added. Subsequent works developed no-regret strategies that rely on Blackwell's approachability theory (Blackwell, 1956b; Hart and Mas-Colell, 2000, 2001), smooth fictitious play (Fudenberg and Levine, 1995, 1999), calibrated forecasts (Foster and Vohra, 1997, 1999), multiplicative weights (Freund and Schapire, 1999), and online gradient ascent (Zinkevich, 2003). Extensions have considered monitoring of rewards only (Auer et al., 2002), general signal monitoring (Rustichini, 1999), and wider definitions of regret (Fudenberg and Levine, 1999a; Lehrer, 2003). For an overview see Foster and Vohra (1999), Hart (2005), Cesa-Bianchi and Lugosi (2006).

The model we consider in this paper extends the standard repeated matrix game model by associating with each stage of the game a temporal duration, which may depend on the actions chosen by both players at the beginning of that stage. Moreover, the performance measure of interest to player

1 is the average reward *per unit time* (rather than the per-stage average). We refer to this model as a *repeated variable-duration game*. The interest in this model is quite natural, as many basic games and related decision problems do have variable length: One can number in this group sequential gambling, board games, investment choices, medical treatment selection, and many others. The proposed model is then the relevant one provided that the player's interest is indeed in the average reward per unit time, rather than the average reward per stage.

Let us expand briefly on two (somewhat academic) examples to illustrate how the variable-duration model arises in specific problems. Consider the game of chess first. Each game instance (ending with win, loose or draw) can be considered as a matrix game, with a finite albeit impractically large number of actions. However, our player may well consider choosing at the start of each game between a much simpler set of decisions, such as the choice of an opening, or adopting an "aggressive", "mild" or "defensive" play (with similar choices for the opponent, augmented possibly by the opponent's strength in case of a random Internet opponent). Each of these options may lead to a different play time for each game (on the average). Now, if our player is interested in maximizing the percentage of games won, then the game duration is of no consequence to her and the standard repeated game model is appropriate. But if, alternatively, our player is interested in maximizing the number of points won *per unit time* (as might be the case if winning is each game is associate with monetary gain, or if our player wishes to increase her Internet player rating as quickly as possible), then the appropriate model is that of a variable-duration repeated game.

As a second illustration consider the problem of sequential investment in an arbitrarily varying market. This problem has been well studied, and overviews may be found in Cover and Thomas (2006), Cesa-Bianchi and Lugosi (2006). Consider the variant where the player needs to choose between several available investment vehicles, say a couple of non-redeemable fixed-term bonds with different maturity dates. In the standard problem formulation the player can change his choice at every time unit, which leads to a repeated game formulation. However, for non-redeemable bonds, this is not the case, and the simplest model that accommodates this restriction is that of a variable-duration repeated game.

Our purpose then is to examine decision strategies and performance goals that are suitable for adaptive play against a arbitrary opponent in repeated variable-duration games. While this model may be viewed as the simplest non-trivial extension of standard repeated games, it will quickly turn out that a direct extension of Hannan's no-regret framework is impossible in general. We start by formulating a natural extension of Hannan's empirical Bayes utility to the present model, to which we refer as the empirical best-response envelope. This average payoff level is easily seen to be attainable when the stage duration depends only on player 2's action. However, a relatively simple counter-example shows that it cannot be attained in general. Hence, in the rest of the paper we turn our attention to weaker performance goals that are attainable. This will be done using two of the basic tools that have previously been used for regret minimization in repeated matrix games, namely Blackwell's approachability theorem and calibrated play.

We note that the repeated variable-duration game model that we consider here is closely related to certain stochastic game models (in the sense of Shapley, 1953). This relationship and its consequences will be further discussed in Section 7.

The paper is organized as follows. Our repeated game model is presented in section 2, together with some preliminary properties. Section 3 defines the empirical Bayes envelope for this model, gives an example for a game in which this envelope is not attainable, and presents some more general conditions

under which the same conclusion holds. Section 4 discusses briefly certain *desiderata* for adaptive play against an arbitrary opponent. These properties will provide a yardstick for measuring the performance of the strategies we subsequently consider. In Section 5 we apply approachability theory to our model. By applying a convexification procedure to the Bayes envelope, we exhibit a weaker performance goal, the convex Bayes envelope, which is indeed attainable. In Section 6 we examine calibrated play and its associated performance guarantees. The final Section 7 considers some additional options of interest for adaptive strategies, discusses the relation to stochastic games, and closes with some directions for further study.

2 Model Formulation

We consider two players, player 1 (P1) and player 2 (P2), who repeatedly play a *variable-duration matrix game*. Let I and J denote the finite action sets of P1 and P2, respectively. The stage game is specified by a reward function $r : I \times J \to \mathbb{R}$ and a strictly positive duration function $\tau : I \times J \to (0, \infty)$. Thus, $r(i, j)$ denotes the reward corresponding to the action pair (i, j), and $\tau(i, j) > 0$ is the duration of the stage game[1]. Let $\Gamma(r, \tau)$ denote this single-stage game model. We note that the reward function r is associated with P1 alone, while P2 is considered an arbitrary player whose utility and goals need not be specified.

The repeated game proceeds as follows. At the beginning of each stage k, where $k = 1, 2, \ldots$, P1 chooses an action i_k and P2 simultaneously chooses an action j_k. Consequently P1 obtains a reward $r_k = r(i_k, j_k)$, and the current stage proceeds for $\tau_k = \tau(i_k, j_k)$ time units, after which the next stage begins. The average reward *per unit time* over the first n stages of play is thus given by

$$\rho_n = \frac{\sum_{k=1}^{n} r_k}{\sum_{k=1}^{n} \tau_k} \,. \tag{2.1}$$

We shall refer to ρ_n as the *(n-stage) reward-rate*. It will also be convenient to define the following per-stage averages:

$$\hat{r}_n = \frac{1}{n} \sum_{k=1}^{n} r_k \,, \quad \hat{\tau}_n = \frac{1}{n} \sum_{k=1}^{n} \tau_k$$

so that $\rho_n = \hat{r}_n / \hat{\tau}_n$. The beginning of stage k will be called the k-th decision epoch or k-th decision point.

We will consider the game from the viewpoint of P1, who wishes to maximize his long-term reward rate. P2 is an *arbitrary player* whose goals are not specified, and whose strategy is not a-priori known to P1. We assume that both players can observe and recall all past actions, and that the game parameters (r and τ) are known to P1. Thus, a strategy σ^1 of P1 is a mapping $\sigma^1 : H \to \Delta(I)$, where H is the set of all possible history sequences of the from $h_k = (i_1, j_1, \ldots, i_k, j_k)$, $k \geq 0$ (with h_0 the empty sequence), and $\Delta(I)$ denotes the set of probability measures over I. P1's action i_k is thus chosen randomly according to the probability measure $x_k = \sigma(h_{k-1})$. A strategy of P1 is *stationary* if $\sigma^1 \equiv x \in \Delta(I)$, and is then denoted by $(x)^\infty$. A strategy σ^2 of P2 is similarly defined as a mapping from H to $\Delta(J)$. We denote this repeated game model by $\Gamma^\infty \equiv \Gamma^\infty(r, \tau)$.

[1] In some applications it may be more natural to specify the model in terms of a reward-rate function $\rho(i, j)$, in which case $r(i, j) = \rho(i, j)\tau(i, j)$. The two representations are of course equivalent.

We next establish some additional notations and terminology. It will be convenient to denote $\Delta(I)$ by X and $\Delta(J)$ by Y. An element $x \in X$ is a *mixed action* of P1, and similarly $y \in Y$ is a mixed action of P2. We shall use the bilinear extension of r and τ to mixed actions, namely

$$r(i,y) = \sum_j r(i,j) y_j \,,$$
$$r(x,y) = \sum_{i,j} x_i r(i,j) y_j \,,$$

and similarly for τ.

The *reward-rate* function $\rho : X \times Y \to \mathbb{R}$ is defined as

$$\rho(x,y) \stackrel{\triangle}{=} \frac{r(x,y)}{\tau(x,y)} = \frac{\sum_{i,j} x_i r(i,j) y_j}{\sum_{i,j} x_i \tau(i,j) y_j} \,. \tag{2.2}$$

This function will play a central role in the following. It is easily seen (using the strong law of large numbers and the renewal theorem) that for any pair of stationary strategies $\sigma^1 = (x)^\infty$ and $\sigma^2 = (y)^\infty$ we have

$$\lim_{n\to\infty} \rho_n = \rho(x,y) \quad \textit{(a.s.)} \tag{2.3}$$
$$\lim_{n\to\infty} \mathbb{E}(\rho_n) = \rho(x,y) \,. \tag{2.4}$$

As usual, the *a.s.* qualifier indicates that the respective event holds with probability one under the probability measure induced by the players' respective strategies.

We further define an auxiliary (single-stage) game $\Gamma_0(r,\tau)$ as the zero-sum game with actions sets X for P1 and Y for P2, and payoff function $\rho(x,y)$ for P1. Note that ρ as defined by (2.2) is *not* bilinear in its arguments. We next establish that this game has a value, which we denote by $\mathrm{v}(r,\tau)$, as well as some additional properties of the reward-rate function ρ.

Lemma 2.1 (Basic properties of ρ)

(i) $\mathrm{v}(r,\tau) \stackrel{\triangle}{=} \max_{x\in X} \min_{y\in Y} \rho(x,y) = \min_{y\in Y} \max_{x\in X} \rho(x,y)$.

(ii) Let X^* denote the set of optimal mixed actions for P1 in $\Gamma_0(r,\tau)$, namely the maximizing set in the max-min expression above, and similarly let Y^* be the minimizing set in the min-max expression. Then X^* and Y^* are closed convex sets.

(iii) For every fixed y, $\rho(\cdot,y)$ is maximized in pure actions, namely

$$\max_{x\in X} \rho(x,y) = \max_{i\in I} \rho(i,y) \,.$$

(iv) The best-response payoff function $\rho^*(y) \stackrel{\triangle}{=} \max_{x\in X} \rho(x,y)$ is Lipschitz continuous in y.

Proof: As we note below, the stated results may be deduced from known ones for semi-Markov games. For completeness we outline a direct proof. Let $\underline{v}$ and $\overline{v}$ denote the max-min and min-max values in (i), respectively. Obviously $\underline{v} \le \overline{v}$, so we need to show that $\underline{v} \ge \overline{v}$. Let v_0 satisfy the equation

$$\mathrm{val}\{r(i,j) - v_0 \tau(i,j)\} = 0,$$

where val$\{m(i,j)\}$ is the value of the zero-sum matrix game with payoffs $m(i,j)$. Existence of such v_0 easily follows by continuity, as the left-hand side is clearly positive for v_0 small enough, and negative for v_0 large enough. It may easily be verified that $\underline{v} \geq v_0$ and $v_0 \geq \overline{v}$, thus establishing (i). Part (ii) now follows by verifying that the optimal action sets of P1 and P2 in the matrix game just described coincide with X^* and Y^*, respectively. Part (iii) follows similarly by noting that $\alpha = \max_{x \in X} \rho(x,y)$ is equivalent to $\max_{x \in X}(r(x,y) - \alpha\tau(x,y)) = 0$, where the last function is linear in x, hence attains its maximum at extreme points. Finally, (iv) follows since τ is bounded away from 0, so that ρ^* is the maximum of a finite number of functions (from (iii)) which are Lipschitz continuous over Y. □

Part (i) of this lemma together with (2.4) imply that $\mathrm{v}(r,\tau)$ is the min-max value of the *repeated* game $\Gamma^\infty(r,\tau)$ when both players are restricted to stationary strategies. We note that this is also the value for general strategies, namely

$$\mathrm{v}(r,\tau) = \inf_{\pi_2} \sup_{\pi_1} \limsup_{n\to\infty} \mathbb{E}(\rho_n) = \sup_{\pi_1} \inf_{\pi_2} \liminf_{n\to\infty} \mathbb{E}(\rho_n)\,.$$

where the infimum and supremum are taken over all the strategy sets of the respective players, and the expectation is according to the measure induced by the strategy pair in effect. This follows from the results of Lal and Sinha (1992), as the repeated game Γ^∞ that is considered here is a special case of Semi-Markov Games which are treated in that paper.

Remark: We assume for simplicity that the rewards and durations in the stage game are deterministic quantities. This model may be extended to accommodate random rewards and durations, with $r(i,j)$ and $\tau(i,j)$ now representing their expected values. All main results of this paper can be extended to this case under appropriate technical conditions — for example, bounded second moments for the reward and duration random variables, and stage durations which are bounded away from zero.

3 No-Regret Strategies and the Best-Response Envelope

In this section we define the empirical best-response envelope as a natural extension of the corresponding concept for fixed duration games. P1's regret is defined as the difference between this envelope and the actual reward-rate, and no-regret strategies must ensure that this difference becomes small (or negative) in the long run. We first observe that no-regret strategies indeed exist when the duration of the stage game depends only on P2's action (but not on P1's). However, the main result of this section is a negative one – namely that no-regret strategies need not exist in general. This is first shown in a specific example, and then shown to hold more generally under certain conditions on the game parameters.

We note that a counter example similar in spirit to Example 3.1 below has been given in Mannor and Shimkin (2003) in the context of regret minimization for stochastic games. As these two examples rely on the specifics of the relevant models, neither implies the other.

Let $\hat{y}_n \in Y$ denote the empirical distribution of P2's actions up to stage n. That is,

$$\hat{y}_n(j) = \frac{1}{n}\sum_{k=1}^{n} 1\{j_k = j\},$$

where $1\{C\}$ denotes the indicator function for a condition C. Clearly $\hat{y}_n \in Y$. The *best-response*

envelope (or Bayes envelope) of P1, $\rho^* : Y \to \mathbb{R}$, is defined by

$$\rho^*(y) \stackrel{\triangle}{=} \max_{i\in I} \frac{r(i,y)}{\tau(i,y)} = \max_{i\in I} \rho(i,y)\,. \tag{3.1}$$

Observe that $\rho^*(y)$ maximizes $\rho(x,y)$ over mixed actions as well, namely

$$\rho^*(y) = \max_{x\in X} \frac{r(x,y)}{\tau(x,y)} = \max_{x\in X} \rho(x,y)\,, \tag{3.2}$$

as per Lemma 2.1(iii).

We consider the difference $\rho^*(\hat{y}_n) - \rho_n$ as P1's n-stage *regret*. As elaborated below, this may be interpreted as P1's payoff loss for not playing his best action against $\hat{y}_n$ over the first n stages. This leads us to the following definition.

Definition 3.1 (No-regret strategies) *A strategy σ^1 of P1 is a* no-regret strategy *if, for every strategy of P2,*

$$\liminf_{n\to\infty} (\rho_n - \rho^*(\hat{y}_n)) \geq 0 \quad \textit{(a.s.)}\,. \tag{3.3}$$

A no-regret strategy of P1 is said to *attain* the best-response envelope. If such a strategy exists we say that the best-response envelope ρ^* is *attainable* by P1.

The following observations provide the motivation for our regret definitions.

Lemma 3.1

(i) Suppose that P2 uses a fixed sequence of actions $(j_1, \ldots, j_n)$, with corresponding empirical distribution $\hat{y}_n$. Then $\rho^*(\hat{y}_n)$ is the maximal reward-rate ρ_n that P1 could obtain by playing any fixed *action $i \in I$ over the first n stages.*

(ii) Let P1 play a mixed stationary strategy $(x)^\infty$. Then, for any strategy of P2,

$$\lim_{n\to\infty} (\rho_n - \rho(x,\hat{y}_n)) = 0 \quad \textit{(a.s.)}\,,$$

and consequently

$$\liminf_{n\to\infty} (\rho^*(\hat{y}_n) - \rho_n) \geq 0 \quad \textit{(a.s.)}\,.$$

Proof: (i) With $i_k \equiv i$ we obtain, by (2.1),

$$\rho_n = \frac{\sum_{k=1}^n r(i,j_k)}{\sum_{k=1}^n \tau(i,j_k)} = \frac{r(i,\hat{y}_n)}{\tau(i,\hat{y}_n)} = \rho(i,\hat{y}_n)\,. \tag{3.4}$$

The required conclusion follows by definition of ρ^*.

(ii) Using the strong low of large numbers for the Martingale difference sequence $D_n = \sum_{k=1}^n (r(i_k,j_k) - r(x,j_k))$, it follows that with probability 1

$$\lim_{n\to\infty} (\hat{r}_n - r(x,\hat{y}_n)) \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n (r(i_k,j_k) - r(x,j_k)) = 0\,.$$

The same holds with r replaced by τ. The first claim now follows since $\rho_n = \hat{r}_n/\hat{\tau}_n$ and $\rho(x,\hat{y}_n) = r(x,\hat{y}_n)/\tau(x,\hat{y}_n)$, with τ bounded away from 0. The second claim then follows by (3.2). □

The last lemma indicates that ρ^* is indeed the natural extension of Hannan's best-response envelope. Part (i) implies that $\rho^*(\hat{y}_n)$ is the best reward-rate that P1 could achieve by using any fixed action

given the empirical distribution $\hat{y}_n$ of P2's actions. Thus, the difference $\rho^*(\hat{y}_n) - \rho_n$ can be interpreted as P1's *regret* for not using that action throughout. Part (ii) implies that $\rho^*(\hat{y}_n)$ is also the best that P1 could achieve by any fixed *mixed* action, at least in the long run.

A particular case where the best-response envelope is attainable is when P1's actions do not affect the duration of the stage game. This includes of course the standard model with fixed stage durations.

Proposition 3.2 *Suppose that the stage duration depends on P2's actions only, namely $\tau(i,j) = \tau(j)$ for every action pair. Then the best-response envelope is attainable by P1.*

Proof: Since $\tau(i_k, j_k) = \tau(j_k)$, we obtain

$$\rho_n = \frac{\sum_{k=1}^n r(i_k, j_k)}{\sum_{k=1}^n \tau(j_k)} = \frac{\hat{r}_n}{\tau(\hat{y}_n)},$$

where $\tau(\hat{y}_n) = \frac{1}{n}\sum_{k=1}^n \tau(j_k)$. Similarly,

$$\rho^*(\hat{y}_n) = \max_i \frac{r(i, \hat{y}_n)}{\tau(\hat{y}_n)}\,.$$

By cancelling out the corresponding denominators it follows that the required inequality in the definition of a no-regret strategy reduces in this case to

$$\liminf_{n\to\infty} \Big(\hat{r}_n - \max_i r(i, \hat{y}_n)\Big) \geq 0\,.$$

This is just the standard definition for a repeated matrix game with fixed stage duration and reward function r, for which no-regret strategies are well known to exist. □

The situation becomes more intricate when the stage durations do depend on P1's actions. This is demonstrated in the following example, which serves as a starting point for the main part of this paper.

Example 3.1 (A game with unattainable best-response envelope). *Consider the variable duration matrix game $\Gamma(r,\tau)$ defined by the following matrix:*

$$\begin{pmatrix} (0,1) & (5,1) \\ (1,3) & (0,3) \end{pmatrix},$$

where P1 is the row player, P2 the column player, and the ij-th entry is $(r(i,j), \tau(i,j))$.

Figure 3 depicts the best-response envelope $\rho^*(y)$ for this example, which is just the maximum of $\rho(i,y)$, $i = 1, 2$. Note that both $\rho(1,y)$ and $\rho(2,y)$ are linear functions of y in this example, which is the case since $\tau(i,j)$ depends on P1's actions only. As a consequence $\rho^*(y)$ turns out to be a convex function.

Proposition 3.3 *The best-response envelop is not attainable by P1 in the game $\Gamma^\infty(r,\tau)$ defined by Example 3.1.*

Proof: We will specify a strategy of P2 against which $\rho^*(y)$ cannot be attained by P1. Let N be some pre-specified integer. Consider first the following strategy for P2 over the first $2N$ stages:

$$j_n = \begin{cases} 1 & \text{for } 1 \leq n \leq N, \\ 2 & \text{for } N+1 \leq n \leq 2N. \end{cases} \tag{3.5}$$

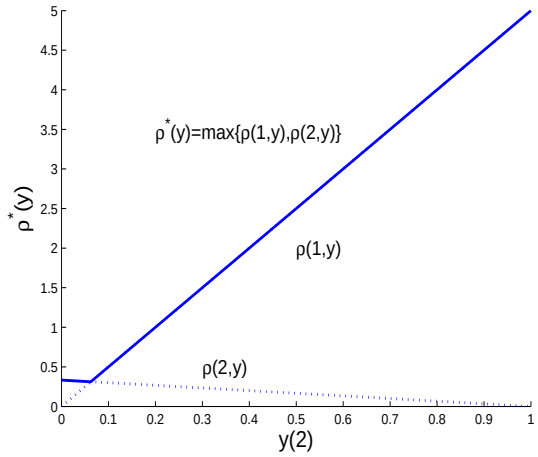

Figure 1: $\rho(1,y)$ and $\rho(2,y)$ as a function of y for the game of Example 3.1. $\rho^*(y)$ is the maximum of these two linear functions. Note that y is represented by its second coordinate $y(2)$.

We claim that for some $\epsilon_0 > 0$ and any strategy of P1, $\rho_k < \rho^*(\hat{y}_k) - \epsilon_0$ must hold either at $k = N$ or at $k = 2N$. To see that, let $\zeta_1 = \sum_1^N 1\{i_k = 1\}/N$ denote the empirical distribution of P1's action 1 over the first N stages. It is easily seen that

$$\rho_N = \frac{\zeta_1 \cdot 0 + (1-\zeta_1)\cdot 1}{\zeta_1 \cdot 1 + (1-\zeta_1)\cdot 3} = \frac{1-\zeta_1}{3-2\zeta_1}\,,$$

and

$$\rho^*(\hat{y}_N) = \max\left\{\frac{0}{1}, \frac{1}{3}\right\} = \frac{1}{3}$$

(which is obtained by action 2 of P1). Thus, to obtain $\rho_N \geq \rho^*(\hat{y}_N) - \epsilon_0$ we need

$$\zeta_1 \leq \frac{9\epsilon_0}{2+3\epsilon_0} = O(\epsilon_0)\,. \tag{3.6}$$

Next, at $k = 2N$ we have $y_{2N} = (0.5, 0.5)$ and

$$\rho^*(\hat{y}_{2N}) = \max\left\{\frac{0+5}{1+1}, \frac{1+0}{3+3}\right\} = \max\left\{\frac{5}{2}, \frac{1}{6}\right\} = \frac{5}{2}\,,$$

which is now obtained by action 1 of P1. To compute ρ_{2N}, let $\zeta_2 = \sum_{N+1}^{2N} 1\{i_k = 1\}/N$ denote the empirical distribution of P1's action 1 over the second N-stage period. Then

$$\rho_{2N} = \frac{(1-\zeta_1)N + 5\zeta_2 N}{(3-2\zeta_1)N + \zeta_2 N + 3(1-\zeta_2)N}\,,$$

which is maximized over $\zeta_2 \in [0,1]$ by $\zeta_2 = 1$, to give

$$\rho_{2N} = \frac{1-\zeta_1+5}{3-2\zeta_1+1} = \frac{6-\zeta_1}{4-2\zeta_1}\,.$$

A simple calculation now shows that to obtain $\rho_{2N} \geq \rho^*(\hat{y}_{2N}) - \epsilon_0$ we need

$$\zeta_1 \geq \frac{2-2\epsilon_0}{3-2\epsilon_0}\,. \tag{3.7}$$

It is evident that the requirements in (3.6) and (3.7) are contradictory for ϵ_0 small enough (say $\epsilon_0 = 0.1$).

To recapitulate, the essence of the above argument is: to obtain ρ_N close to $\rho^*(\hat{y}_N)$ P1 must use action 1 during most of the first N stages. But then the most he can get for ρ_{2N} is about 3/2, which falls short of $\rho^*(\hat{y}_{2N}) = 5/2$.

We conclude that P2's stated strategy forces P1 to have positive regret at the end of stage N or at the end of stage $2N$. P2 can repeat the same strategy with a new N' much larger than N, so that the first N stages have a negligible effect. This can be done repeatedly, so that P1 has non-zero regret (larger than, say, $\epsilon_0/2$) infinitely often. □

A few additional points should be noted regarding Example 3.1 and its consequences.

1. The proof of Proposition 3.3 uses a fixed strategy of P2, which may be disclosed to P1 *beforehand* without changing the conclusion. Thus, the inability to attain the best-response envelope is not a consequence of the unknown strategy of an arbitrary opponent. Rather, it may attributed to the different rates at which $\hat{y}_n$ and ρ_n can change due to the action-dependent time normalization of the latter.
2. To emphasize the last point, observe that even if P1 plays his best-response to P2's action at each stage, he still falls short of attaining $\rho^*(\hat{y}_n)$. Indeed, in our example, suppose that P1 reacts to the strategy (3.5) of P2 by playing $i_k = 1$ for the first N stages (a best response to $j_k = 1$) and $i_k = 2$ for the next N stages (his best response to $j_k = 2$). Then at $n = 2N$ he obtains $\rho_n = 3/2$, while $\hat{y}_n = (0.5, 0.5)$ and $\rho^*(\hat{y}_n) = 5/2$.
3. The stage durations $\tau(i, j)$ in our example depend only on i, the action of P1. Combined with Proposition 3.2, this implies that the inability to attain the best-response envelope in repeated variable-duration games can be fully attributed to the dependence on the stage durations on P1's actions.
4. As already noted, ρ^* is a convex function in this example (see Figure 3). This implies that convexity of ρ^* has no direct implication on its attainability. In the next section we will see that some related convexity conditions do provide a sufficient condition for ρ^* to be attainable.

We close this section with certain sufficient conditions for *non-existence* of no-regret strategies. These conditions essentially follow by similar reasoning to that of the last counterexample. We use $X^*(y)$ to denote the set of best response strategies against y. That is:

$$X^*(y) = \arg\max_{x \in X} \rho(x, y).$$

Proposition 3.4

(i) Suppose there exist $y_1, y_2 \in Y$ and $\alpha \in (0, 1)$ such that:

$$\rho^*\left(\alpha y_1 + (1-\alpha) y_2\right) > \max_{x_1 \in X^*(y_1),\, x_2 \in X} \frac{\alpha r(x_1, y_1) + (1-\alpha) r(x_2, y_2)}{\alpha \tau(x_1, y_1) + (1-\alpha) \tau(x_2, y_2)}. \tag{3.8}$$

Then the best-response envelope is not attainable by P1.

(ii) More generally, suppose there exist $y_1, y_2, \ldots, y_M \in Y$ and $\alpha_1, \ldots, \alpha_M > 0$ with $\sum_{m=1}^{M} \alpha_m = 1$ such that the following system of inequalities (in $x_1, x_2, \ldots, x_M \in X$):

$$\rho^*\left(\frac{\sum_{m=1}^{\ell} \alpha_m y_m}{\sum_{m=1}^{\ell} \alpha_m}\right) \leq \frac{\sum_{m=1}^{\ell} \alpha_m r(x_m, y_m)}{\sum_{m=1}^{\ell} \alpha_m \tau(x_m, y_m)}, \quad \ell = 1, 2, \ldots, M \tag{3.9}$$

does not have a solution. Then the best response is not attainable by P1.

Proof: The proof of (i) is very similar to that of Proposition 3.3, and we only provide a brief outline. The strategy used by P2 over the first N stages (with N a large pre-specified number) is to play y_1 for αN stages (taking the integer part thereof) and play y_2 for the remaining $(1-\alpha)N$ stages. We take N to be large enough so that stochastic fluctuations (due to possibly mixed actions) from the expected averages become insignificant. The empirical distribution of P1's actions at the end of the first period must then be close to some $x_1 \in X^*(y_1)$ to guarantee that ρ_n is close to $\rho^*(\hat{y}_n) \approx \rho^*(y_1)$ at $n = \alpha N$. However, equation (3.8) implies then that at the end of stage N the reward rate ρ_N falls short of the best response $\rho^*(\hat{y}_N)$, no matter what actions P1 uses against y_2.

The claim in (ii) is just a multi-period extension of (i). The strategy used by P2 is to play y_1 for $\alpha_1 N$ stages, then play y_2 for $\alpha_2 N$ and so forth. Again we can ignore stochastic effects by taking N large enough. Since there is no sequence (x_m) that satisfies (3.9), it follows that $\rho_n < \rho^*(\hat{y}_n)$ must hold at the end of one of these M periods. Furthermore, we claim that this inequality is satisfied with some uniform margin, namely that there exists $\epsilon_0 > 0$ (which depends only on $y_1, \ldots, y_M$) so that $\rho_n \leq \rho^*(\hat{y}_n) - \epsilon_0$. This follows by the compactness of X and continuity of the right-hand side of (3.9), which imply that at least one of the opposite inequalities to those in (3.9) is satisfied with some uniform margin $\epsilon_0 > 0$, independent of the choice of $x_1, \ldots, x_M$. As in the proof of Proposition 3.3, we can now extend P2's strategy to the entire time horizon by repeating it with increasingly larger N, so that P1 has non-zero regret (larger than, say, $\epsilon_0/2$) infinitely often. □

We note that the maximum over x and x^* in Equation (3.8) is in fact obtained in pure actions (see Lemma 2.1(iii)), and (3.8) can be simplified accordingly.

Remark: Corollary 6.3 in Section 6 provides a sufficient condition for attainability of the best-response envelope, which can be written in the following way. Suppose that for every $M > 1$, $y_1, y_2, \ldots, y_M \in Y$ and $\alpha_1, \ldots, \alpha_M > 0$ with $\sum_{i=1}^k \alpha_i = 1$ we have

$$\rho^*\left(\sum_{m=1}^{M} \alpha_m y_m\right) \leq \min_{x_m^* \in X^*(y_m)} \left\{ \frac{\sum_{m=1}^{M} \alpha_m r(x_m^*, y_m)}{\sum_{m=1}^{M} \alpha_m \tau(x_m^*, y_m)} \right\}.$$

Then the best response is attainable. This sufficient condition may be viewed as a partial converse to the necessary condition of (3.9).

4 Desiderata for Adaptive Play

Given the negative results of the previous section regarding the non-existence of no-regret strategies, it follows that in general we will need to settle for less ambitious goals. It will thus be useful to consider at this point some desired properties for adaptive play, against which the performance of specific strategies can be compared. Following Fudenberg and Levine (1995), in part, we consider the following *desiderata* for adaptive play of P1 against an arbitrary opponent.

(1) *Safety:* P1's long-term payoff should be at least his min-max payoff in the repeated game.

(2) *Adaptivity:* When the opponent's play deviates from a worst-case strategy (according to some pre-specified criterion), the long-term payoff for P1 should be strictly higher than his min-max payoff.

(3) *Best-response to stationary strategies:* If the opponent's strategy is stationary, then P1's long-term payoff should be as high as his best-response payoff to the opponent's strategy.

Property (2) requires a measure for deviation of P2 from a worst-case strategy. The standard choice for repeated games is the deviation of P2's empirical action distribution ($\hat{y}_n$) from his optimal adversarial mixed action in the stage game. This is the one we adopt in this paper as well. Furthermore, it is important to quantify the expected gain (over the min-max payoff) as a function of this deviation: without such a quantitative estimate, one could hardly justify the effort involved in implementing an adaptive strategies. Such estimates will indeed be provided below in the form of performance envelopes.

Property (3) is motivated by the observation that in a stationary environment, achieving the best-response payoff is easy. The point of course is that stationarity is not assumed here a-priori, and the required best-response property should be achieved together with properties (1) and (2). A no-regret strategy, as per Definition 3.1, is easily seen to satisfy all three properties. As we have seen, however, such a strategy need not exist in general. The approachability-related strategies proposed in the next section will be seen to satisfy properties (1) and (2), while the calibration-related strategies of Section 6 satisfy all three.

5 Approachability and Regret Minimization

The theory of approachability, introduced in Blackwell (1956a), is one of the fundamental tools that have been used for obtaining no-regret strategies in repeated matrix games. In this section we apply the approachability framework to our model of repeated variable-duration matrix games. The analysis will yield a sufficient condition on the model parameters for existence of no-regret strategies. It will also allow us to specify a relaxed goal for adaptive play, the convex best-response envelope, which is always attainable, and provides some useful performance guarantees.

The results of Section 6 show that the performance guarantees available for calibrated play dominate those obtained for approachability-related strategies. The latter are however easier to implement, and provide additional theoretical insight.

5.1 Approachability for Repeated Variable-Duration Games

We start by adapting the required definitions and results of approachability theory to our repeated variable-duration model. We augment the model of Section 2 by replacing the reward function with a vector-valued reward $\vec{r} : I \times J \to \mathbb{R}^L$, where $L \geq 1$. Thus $\vec{r}(i,j) = (r^1(i,j), \ldots, r^L(i,j))$. Consider the corresponding n-stage reward rate vector

$$\vec{\rho}_n = \frac{\sum_{k=1}^n \vec{r}(i_k, j_k)}{\sum_{k=1}^n \tau(i_k, j_k)} .$$

As in the scalar game we let $\vec{r}(x,y)$ denote the bilinear extension of $\vec{r}$ to mixed action (note that τ remains a scalar in the vector-valued game). The average vector-valued reward is

$$\vec{\rho}(x,y) = \frac{\vec{r}(x,y)}{\tau(x,y)} .$$

Let $\Gamma^\infty(\vec{r}, \tau)$ denote the corresponding repeated variable-duration matrix game.

Definition 5.1 *A set $B \subset \mathbb{R}^L$ is* approachable[2] *(by P1) if there exists a strategy σ^1 of P1 so that, for any strategy of P2,*

$$\lim_{n\to\infty} d(\vec{\rho}_n, B) = 0 \quad \textit{(a.s.)}$$

where $d(\vec{\rho}, B) = \inf_{b\in B} d(\vec{\rho}, b)$ denotes the Euclidean point-to-set distance in $\mathbb{R}^L$.

A strategy of P1 that satisfies this property is called an *approaching strategy* for the set B. The following theorem extends Blackwell's characterization of approachability in Blackwell (1956a) to the variable duration model. The proof follows in essence Blackwell's original argument, with some modifications that are required to handle the different time normalization of the average reward vector. For completeness we provide an outline in Appendix A.

Theorem 5.1 *Let B be a closed set in $\mathbb{R}^L$.*

(i) B is approachable if for every point $a \notin B$ there exists a mixed action $x \in X$ so that

$$\langle c_a - a, c_a - \vec{\rho}(x, y)\rangle \le 0 \quad \textit{for every } y \in Y, \tag{5.1}$$

where c_a is a closest point in B to a and $\langle a, b\rangle$ is the standard inner product in $\mathbb{R}^L$. An approaching strategy for P1 is then to play an arbitrary x_n if $\vec{\rho}_{n-1} \in B$, and otherwise play any x_n that satisfies the separation condition (5.1) with $a = \vec{\rho}_{n-1}$.

(ii) Assume that B is a convex set. Then the last condition is both necessary and sufficient for B to be approachable. Furthermore, it is equivalent to either one of the following conditions:

(a) For every unit vector $u \in \mathbb{R}^L$ there exists $x \in X$ so that

$$\langle u, \vec{\rho}(x, y)\rangle \ge \inf_{b\in B} \langle u, b\rangle \quad \textit{for every } y \in Y\,.$$

(b) For every $y \in Y$ there exists $x \in X$ so that $\vec{\rho}(x, y) \in B$.

An approaching strategy in this case is to play, whenever $a = \vec{\rho}_{n-1} \notin B$, a mixed action $x_n \in \arg\max_x \min_y \langle u, \vec{\rho}(x, y)\rangle$, where $u = (c_a - a)$.

5.2 The Temporal Best-Response Envelope

We now return to our original model with a scalar reward function r. We aim to formulate the no-regret requirement of Definition 3.1 (or a relaxed one) as an approachability condition, and apply the conditions of Theorem 5.1. Following Blackwell (1956b), our first attempt will be to define the payoff vector $\vec{\rho}_n = (\rho_n, \hat{y}_n)$, so that attaining the best-response $\rho^*(y)$ is equivalent to approaching the set

$$B_0 = \{(\rho, y) \in \mathbb{R} \times Y \,:\, \rho \ge \rho^*(y)\}\,.$$

However, two obstacles stand in the way of applying the approachability result of Theorem 5.1. First and foremost, ρ_n and $\hat{y}_n$ are normalized by different temporal factors. Second, B_0 need not be a convex set as the best-response envelope $\rho^*(y)$ is not convex in general, so that condition (b) in that theorem may not be applicable.

[2] Blackwell's definition of approachability requires also a uniform rate of convergence (independent of P2's strategies). We note that the proof of Theorem 5.1 indeed provides explicitly such a uniform rate. However, in the present paper we omit this requirement from the definition as it is not required for our results.

To address the first difficulty, we reformulate the approachability problem. Let π_n denote the vector of P2's *action rates*, namely

$$\pi_n = \frac{1}{\hat{\tau}_n}\hat{y}_n\,.$$

Note that $\pi_n(j)$ gives the temporal rate, in actions per unit time, in which action j was chosen over the first n stages. Obviously π_n is not a probability vector, as the sum of its elements is $\hat{\tau}_n$. The set of feasible action rates is given by

$$\Pi = \left\{\frac{y}{\tau} \,:\, y \in Y,\, \tau \in T(y)\right\}, \tag{5.2}$$

where $T(y)$ is the set of average stage durations τ which are feasible jointly with the empirical distribution y:

$$\begin{aligned} T(y) &= \left\{\sum_{i,j}\alpha_{ij}\tau(i,j) \,:\, \alpha \in \Delta(I\times J),\, \sum_i \alpha_{ij} = y(j) \text{ for all } j\right\} \\ &= \left\{\sum_j y(j)\tau(x^j,j) \,:\, x^j \in X \text{ for all } j\right\}. \end{aligned} \tag{5.3}$$

Note that Π is a convex set; indeed, it is the image of the convex set $\{y,\tau : y \in Y,\, \tau \in T(y)\}$ under a linear-fractional function (Boyd and Vanderberghe, 2004).

We proceed to formulate the set to be approached in terms of π instead of $\hat{y}$. Note first that the action rate vector π_n uniquely determines the empirical distribution vector $\hat{y}_n$ via $\hat{y}_n = \pi_n/|\pi_n|$, where $|\pi|$ is the sum of elements of π. Given P2's action-rate vector $\pi \in \Pi$, we define the best-response payoff for P1 as its best-response payoff against the empirical distribution $\hat{y} = \pi/|\pi|$ induced by π. That is, for $\pi \in \Pi$,

$$\tilde{\rho}^*(\pi) \stackrel{\triangle}{=} \rho^*\left(\frac{\pi}{|\pi|}\right) = \max_{i\in I}\frac{\sum_j r(i,j)\pi(j)}{\sum_j \tau(i,j)\pi(j)}\,, \tag{5.4}$$

where $|\pi|$ was cancelled out from the last expression. Thus, although defined on a different set, $\tilde{\rho}^*$ turns out to be identical in its functional form to ρ^*. We refer to $\tilde{\rho}^* : \Pi \to \mathbb{R}$ as the *temporal* best-response envelope.

Convexity of $\tilde{\rho}^*$ turns out to be a sufficient condition for existence of no-regret strategies.

Theorem 5.2 *Suppose the temporal best-response envelope $\tilde{\rho}^*(\pi)$ is convex over its domain Π. Then P1 has a no-regret strategy (in the sense of Definition 3.1), namely, a strategy that attains the best-response envelope $\rho^*(\hat{y})$.*

Proof: Suppose that we have a vector-valued game where the immediate reward vector at stage n is (r_n, e_{j_n}), where e_k is a vector of zeros except for the k-th position which is one. This vector-valued reward has as a first coordinate the (per-stage) reward and a vector indicating which action was chosen by P2 in the rest of the $|J|$ coordinates. Using our definitions for the vector-valued game, we obtain that the average reward vector is $\vec{\rho}_n = (\rho_n, \pi_n)$. We will now show that the following set is approachable by P1 with this reward vector:

$$B_1 = \{(\rho,\pi) \in \mathbb{R}\times\Pi \,:\, \rho \geq \tilde{\rho}^*(\pi)\}\,.$$

Indeed, B_1 is convex as $\tilde{\rho}^*$ is a convex function (by assumption) over a convex domain Π, and B_1 is its epigraph. We next verify condition (b) in Theorem 5.1. Note that

$$\vec{\rho}(x,y) = \frac{\vec{r}(x,y)}{\tau(x,y)} = (\rho(x,y), \frac{y}{\tau(x,y)}),$$

so that $\vec{\rho}(x,y) \in B_1$ is equivalent to

$$\rho(x,y) \geq \tilde{\rho}^* \left(\frac{y}{\tau(x,y)} \right) \equiv \rho^*(y). \tag{5.5}$$

For each y, we choose an x that maximizes $\rho(x,y)$, namely $\rho(x,y) = \rho^*(y)$. Thus the last inequality is satisfied with equality, and $\vec{\rho}(x,y) \in B_1$. Thus condition (b) is satisfied and B_1 is approachable.
Recall next that $\rho^*(y)$ is Lipschitz continuous by Lemma 2.1(iv). Since τ is bounded away from zero it follows that $\tilde{\rho}^*(\pi)$ is also Lipschitz continuous. It is therefore easily verified that $d(\vec{\rho}_n, B_1) \to 0$ implies that $\liminf_{n\to\infty}(\rho_n - \tilde{\rho}^*(\pi_n)) \geq 0$, and since $\tilde{\rho}^*(\pi_n) = \rho^*(\hat{y}_n)$ we obtain the required inequality in (3.3). Thus, any approaching strategy for B_1 is a no-regret strategy of P1. □

Note that an approaching strategy, as specified in Theorem 5.1, requires P1 to keep track only of π_n and ρ_n, or equivalently of $\hat{y}_n$ and $\hat{\tau}_n$.

The convexity condition in Theorem 5.2 is clearly satisfied for the standard model of fixed-duration matrix games, where $\tau \equiv 1$, $y = \pi$, and $\tilde{\rho}^*(\pi) = \rho^*(y) = \max_i r(i,y)$ is a convex function (as the maximum of linear functions). This assumption is also satisfied when P2 alone controls the game duration, namely $\tau(i,j) = \tau_0(j)$. We then obtain that $\Pi = \{y/\tau_0(y) : y \in Y\}$, so that $\sum_j \tau(j)\pi(j) = 1$ and by (5.4), $\tilde{\rho}^*(\pi) = \max_i \sum_j r(i,j)\pi(j)$. This is again convex (as the maximum of linear functions), and we thus recover the conclusion of Proposition 3.2. However, $\tilde{\rho}^*$ must be non-convex whenever the best-response envelope $\rho^*(\hat{y})$ is not attainable, as in the game of Example 3.1.

5.3 The Convex Best-Response Envelope

When $\tilde{\rho}^*$ is not convex, the preceding analysis provides no performance guarantees for P1. To proceed, we will need to relax the goal of attaining the best-response.

Definition 5.2 (Convex best-response envelope) *The convex best-response envelope $\tilde{\rho}^{co} : \Pi \to \mathbb{R}$ is defined as the lower convex hull of $\tilde{\rho}^*$ over its domain Π.*

We now have the following result.

Theorem 5.3 ($\tilde{\rho}^{\mathbf{co}}(\pi)$ is attainable) *The convex best-response envelope $\tilde{\rho}^{co}(\pi)$ is attainable by P1. Namely, there exists a strategy of P1 so that*

$$\liminf_{n\to\infty} (\rho_n - \tilde{\rho}^{co}(\pi_n)) \geq 0 \quad \textit{(a.s.)} \tag{5.6}$$

for any strategy of P2.

Proof: The proof is identical to that of Theorem 5.2, with $\tilde{\rho}^*$ replaced by $\tilde{\rho}^{\mathrm{co}}$. Clearly $\tilde{\rho}^{\mathrm{co}}$ satisfies the convexity requirement by its definition. Since $\rho^*(\pi)$ is Lipschitz continuous, it follows that its lower convex hull is Lipschitz continuous. Condition (b) of Theorem 5.1 is satisfied since $\tilde{\rho}^{\mathrm{co}} \leq \tilde{\rho}^*$, again by its definition, so that any point $\vec{\rho}(x,y)$ that belongs to the set B_1 also belongs to the relaxed set that corresponds to $\tilde{\rho}^{\mathrm{co}}$. □

It will be useful to formulate the performance guarantee of the last proposition in terms of the empirical distribution $\hat{y}_n$ rather than the action rates π_n. This is easily done by projecting $\tilde{\rho}^{\text{co}}$ from Π back Y. For $\hat{y} \in Y$, define

$$\rho^{\text{co}}(\hat{y}) = \min\{\tilde{\rho}^{\text{co}}(\pi) \,:\, \pi \in \Pi, \frac{\pi}{|\pi|} = \hat{y}\}\,. \tag{5.7}$$

For simplicity we also refer to ρ^{co} as the convex best-response envelope (over Y). The following corollary to Theorem 5.3 is immediate.

Corollary 5.4 ($\rho^{\text{co}}(\hat{y})$ is attainable) *The convex best-response envelope $\rho^{co}(\hat{y})$ is attainable by P1. Namely, there exists a strategy of P1 so that*

$$\liminf_{n\to\infty} (\rho_n - \rho^{co}(\hat{y}_n)) \geq 0 \quad \textit{(a.s.)}. \tag{5.8}$$

In fact, any strategy of P1 that attains $\tilde{\rho}^{co}(\pi)$ also attains $\rho^{co}(\hat{y})$.

Figure 5.3 illustrates the resulting convex best-response envelope for the game of Example 3.1. We note that ρ^{co} was computed analytically, but the computation is technical and is omitted here. As ρ^* is not attainable in this example, it is clear that ρ^{co} must be strictly smaller than ρ^* for some values of y, as is indeed the case.

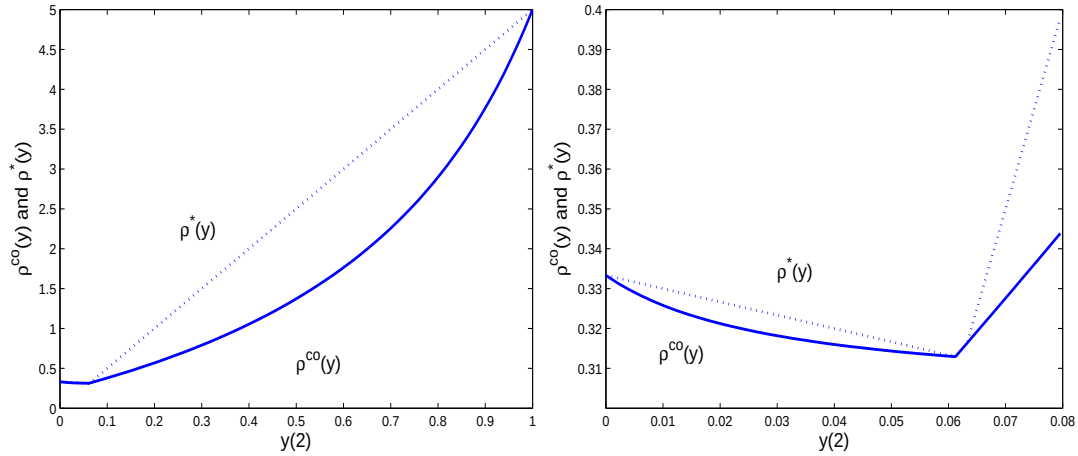

Figure 2: $\rho^*(y)$ (dotted) and $\rho^{\text{co}}(y)$ (thick line) for the game of Example 3.1. The right figure zooms on the segment $[0, 0.08]$.

The next lemma presents some general properties of ρ^{co} that will be related to its performance guarantees.

Lemma 5.5 (Properties of ρ^{co}) *The convex best-response envelope $\rho^{co}(y)$ satisfies the following properties. For each $y \in Y$,*

(i) $\mathrm{v}(r,\tau) \leq \rho^{co}(y) \leq \rho^*(y)$.

(ii) If $\rho^*(y) > \mathrm{v}(r,\tau)$, then $\rho^{co}(y) > \mathrm{v}(r,\tau)$.

Proof: (i) Fix y, and take any $\pi \in \Pi$ with $\pi/|\pi| = y$. Then

$$\rho^{\text{co}}(y) \leq \tilde{\rho}^{\text{co}}(\pi) \leq \tilde{\rho}^*(\pi) = \rho^*(y)\,,$$

where all inequalities follow directly from the definitions of the respective envelopes. Also, since $\rho^* \geq \mathrm{v}(r,\tau)$, the same property is inherited by $\tilde{\rho}^*$, $\tilde{\rho}^{\mathrm{co}}$ and ρ^{co}, again by their respective definitions.

(ii) We will show that $\rho^{\mathrm{co}}(y) = \mathrm{v}(r,\tau)$ implies that $\rho^{\mathrm{co}}(y) = \mathrm{v}(r,\tau)$. Suppose $\rho^{\mathrm{co}}(y) = \mathrm{v}(r,\tau)$. Then there exists some $\pi \in \Pi$ such that $\pi/|\pi| = y$ and $\tilde{\rho}^{\mathrm{co}}(\pi) = \mathrm{v}(r,\tau)$. By Caratheodory's Theorem there exist ℓ points $\pi_1, \ldots, \pi_\ell$ in Π (where $\ell \leq 2 + |J|$) and coefficients $\alpha_1, \ldots, \alpha_\ell > 0$ with $\sum_{m=1}^{\ell} \alpha_m = 1$ such that $\pi = \sum_{m=1}^{\ell} \alpha_m \pi_m$ and $\mathrm{v}(r,\tau) = \tilde{\rho}^{\mathrm{co}}(\pi) = \sum_{m=1}^{\ell} \alpha_m \rho^*(\pi_m)$. Since $\rho^*(\pi) \geq \mathrm{v}(r,\tau)$, this implies that $\rho^*(\pi_m) = \mathrm{v}(r,\tau)$ for all m. Recall now from Lemma 2.1(ii) that the set Y^* of mixed actions $y \in Y$ for which $\rho^*(y) = \mathrm{v}(r,\tau)$ is convex. The set $\Pi^* = \{\pi' \in \Pi : \pi'/|\pi'| \in Y^*\}$ is thus an image of a convex set under a linear-fractional transformation, and is therefore convex (Boyd and Vanderberghe, 2004). Noting that $\pi_m \in \Pi^*$ for all m (which follows from $\rho^*(\pi_m) = \mathrm{v}(r,\tau)$) and π is their convex combination, it follows that $\pi \in \Pi^*$ and in particular that $y = \pi/|\pi| \in Y^*$, which is equivalent to $\rho^{\mathrm{co}}(y) = \mathrm{v}(r,\tau)$. □

Both properties that were stated in the last lemma can be observed in Figure 5.3.

We can now examine the performance guarantees that P1 secures by attaining ρ^{co}. Referring to Section 4, *safety* is clearly implied since $\rho^{\mathrm{co}} \geq \mathrm{v}(r,\tau)$. More interestingly, *adaptivity* is also satisfied, as implied by Lemma 5.5(ii). Thus, P1's long-term reward rate will be strictly higher than the min-max value of the game whenever the empirical distribution of P2's actions deviates from as optimal strategy in the stage game. We note however that property (3), best-response to stationary strategies, need not hold in general. Specifically, if P2 uses a stationary strategy $(y)^\infty$ for which $\rho^{\mathrm{co}}(y) < \rho^*(y)$, then ρ_n may well fall short of $\rho^*(y)$, although it will still be strictly higher than $\mathrm{v}(r,\tau)$.

Remark: Our use of approachability for regret minimization follows the formulation of Blackwell (1956b), where the set to be approached is defined in the joint space of the average reward and empirical distribution of P2's actions. More recent applications of approachability to this problem have taken an alternative view, whereby the set to be approached is defined as the negative quadrant in the $|I|$-dimensional space where each coordinate corresponds to the regret with respect to some action $i \in I$ (see Hart and Mas-Colell, 2000). In terms of the present model, we could similarly define

$$L_n(i) \triangleq \rho(i,\hat{y}_n) - \rho_n = \frac{\sum_{k=1}^{n} r(i,j_k)}{\sum_{k=1}^{n} \tau(i,j_k)} - \frac{\sum_{k=1}^{n} r_k}{\sum_{k=1}^{n} \tau_k},$$

so that the no-regret requirement is equivalent to $\limsup_n L_n(i) \leq 0$, $i \in I$. Unfortunately, the approachability results in Theorem 5.1 cannot be applied here, as the two terms on the right hand side of the last equation have different temporal factors in their denominators.

6 Calibrated Play

In calibrated play, P1 uses at each stage a best-response to his forecasts of the other player's action at that stage. The quality of the resulting strategy depends of course on the quality of the forecast; it is well known that using *calibrated* forecasts leads to no-regret strategies in repeated matrix games. See, for example, Foster and Vohra (1997) for an overview of the regret concept and its relations to calibration. In this section we consider the consequences of calibrated play for repeated games with variable stage duration.

We start with a formal definition of calibrated forecasts and calibrated play in the next subsection. We then introduce in Subsection 6.2 the *calibration envelope* $\rho^{\mathrm{cal}}(\hat{y})$, and show that it is attained

by calibrated play in the sense that $\rho_n \geq \rho^{\text{cal}}(\hat{y}_n)$ holds asymptotically. As $\rho^{\text{cal}}(\hat{y}) \geq \text{v}(r,\tau)$ and $\rho^{\text{cal}}(\hat{y}) > \text{v}(r,\tau)$ whenever $\rho^*(\hat{y}) > \text{v}(r,\tau)$ (Lemma 6.4), it follows that calibrated play satisfies the safety and adaptivity properties from Section 4. We then proceed to compare the calibration envelope with the convex best-response envelope of the previous section, and show that $\rho^{\text{cal}} \geq \rho^{\text{co}}$. We identify certain classes of games where equality holds, but show that the inequality may be strict in general. Thus, the performance guarantees provided by the calibrated envelope are superior to those implied by the convex best-response envelope of the previous section. In Subsection 6.4 we show that calibrated play achieves the best-response to stationary strategies, namely attains the best-response payoff $\rho^*(\hat{y}_n)$ when P2 is stationary. It thus follows that calibrated play achieves all three desiderata of adaptive play that were proposed in Section 4.

6.1 Calibrated Forecasts and Calibrated Play

A forecasting scheme specifies at each decision point k a probabilistic forecast $q_k \in Y$ of P2's action j_k. More specifically, a (randomized) forecasting scheme is a sequence of maps $\mu_k : H_{k-1} \to \Delta(Y)$, $k \geq 1$, which associates with each possible history h_{k-1} a probability measure μ_k over Y. The forecast $q_k \in Y$ is selected at random according to the distribution μ_k. Note that the realized value q_k is included in the history sequence h_k.

We shall use the following definition of calibrated forecasts.

Definition 6.1 (Calibrated forecasts) *A forecasting scheme is* calibrated *if for every (Borel measurable) set $Q \subset Y$ and every strategy of P2,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 1\{q_k \in Q\}(e_{j_k} - q_k) = 0\,. \tag{6.1}$$

This form of calibration property has been introduced into game theory by Foster and Vohra (1997), and several algorithms have been devised to achieve it (Foster and Vohra, 1998; Foster, 1999; Fudenberg and Levine, 1999b; Kakade and Foster, 2004). These algorithms typically start with predictions that are restricted to a finite grid, and the requirement in the last definition may be achieved by gradually increasing the number of grid points. A notable deviation from grid-based procedures is presented in Mannor et al. (2007), where a computationally efficient calibration scheme is devised. However, except for some special cases, the scheme of Mannor et al. (2007) is calibrated only against a restricted class of opponents. Stronger notions of calibration were considered in Kalai et al. (1999), Sandroni et al. (2003).

In calibrated play, the active player (P1) essentially chooses a best-response action to his forecast of the other player's actions. That is: $i_k \in I^*(q_k)$, where

$$I^*(y) = \arg\max_{i\in I} \frac{r(i,y)}{\tau(i,y)}\,, \quad y \in Y\,. \tag{6.2}$$

To be more specific, we shall assume some fixed tie-breaking rule when $I^*(y)$ is not a singleton. Thus, we have the following definition.

Definition 6.2 (Calibrated Play) *A* calibrated strategy *for P1 in the variable-duration repeated game $\Gamma^\infty(r,\tau)$ is given by*

$$i_k = i^o(q_k) \tag{6.3}$$

where (q_k) is a calibrated forecast of P2's actions, and $i^0(y) \in I^*(y)$ for each $y \in Y$.

The choice of i_k as a best response to q_k in the game $\Gamma_0(r,\tau)$ with payoff $\rho(x,y)$ is motivated by the definition of the best-response envelope in (3.1). Note that the chosen action does *not* maximize expected one-stage reward rate, namely $\sum q_k(j)\frac{r(i,j)}{\tau(i,j)}$, which cannot be easily related to the repeated game payoff. In the final Section 7 we shall mention another reasonable option for choosing i_k in response to q_k.

6.2 The Calibration Envelope

Let

$$Y_i^* = \{y \in Y : i \in I^*(y)\}$$

denote the (closed) set of mixed actions to which $i \in I$ is a best response in $\Gamma_0(r,\tau)$. We shall assume that each Y_i^* is non-empty; actions i for which Y_i^* is empty will never be used and can be deleted from the game model.

Let $\Delta_d(Y)$ denote the set of discrete probability measures on Y, and let $m_\mu = \int y\mu(dy)$ denote the barycenter of $\mu \in \Delta_d(Y)$. The *calibration envelope* ρ^{cal} is defined as follows, for $\hat{y} \in Y$:

$$\rho^{\mathrm{cal}}(\hat{y}) = \inf\left\{\frac{\int r(i(y),y)\mu(dy)}{\int \tau(i(y),y)\mu(dy)} \;:\; \mu \in \Delta_d(Y),\, m_\mu = \hat{y},\, i(y) \in I^*(y)\right\}. \tag{6.4}$$

The restriction to discrete measures is for technical convenience only and is of no consequence, as the infimum is already attained by a measure of finite support. This follows from the next lemma which also provides an alternative expression for ρ^{cal}, alongside a useful continuity property.

Lemma 6.1

(i) Let $co(Y_i^*)$ denote the convex hull[3] of Y_i^*. Then

$$\rho^{cal}(\hat{y}) = \min\left\{\frac{\sum_{i\in I}\alpha_i r(i,y_i)}{\sum_{i\in I}\alpha_i \tau(i,y_i)} \;:\; \alpha \in \Delta(I),\, y_i \in co(Y_i^*), \sum_{i\in I}\alpha_i y_i = \hat{y}\right\}. \tag{6.5}$$

(ii) The infimum in (6.4) is attained by a measure μ of finite support.

(iii) $\rho^{cal}(\hat{y})$ is continuous in $\hat{y} \in Y$.

Proof: (i) Note first that the minimum in (6.5) is indeed attained, as we minimize a continuous function over a compact set ($co(Y_i^*)$ is closed since Y_i^* is closed). Let $\rho^1(\hat{y})$ denote the right-hand side of (6.5). To show that $\rho^1 \le \rho^{\mathrm{cal}}$, note that by Caratheodory's Theorem each $y_i \in co(Y_i^*)$ can be written as $y_i = \sum_{j\in J}\beta_{ij}y_{ij}$, with $y_{ij} \in Y_i^*$ and $\beta_i \in \Delta(J)$. It follows that for each $\hat{y}$ the argument of (6.5) can be written as the special case of the argument of (6.4), from which $\rho^1(\hat{y}) \le \rho^{\mathrm{cal}}(\hat{y})$ follows. Conversely, given $\mu \in \Delta_d(\hat{y})$ and the selection function $i(y) \in I^*(y)$, define $\alpha_i = \int_{y:i(y)=i}\mu(dy)$, and $y_i = \int_{y:i(y)=i} y\mu(dy)/\alpha_i$ (with y_i arbitrary if $\alpha_i = 0$). Note that $y_i \in co(Y_i^*)$, since $i(y) \in I^*(y)$ implies $y \in Y_i^*$, and y_i is defined as a convex combination of such y's. The argument of (6.4) is thus reduced to the form of (6.5), which implies that $\rho^1(\hat{y}) \le \rho^{\mathrm{cal}}(\hat{y})$.

[3] Y_i^* need not be convex, as the functions $\rho(i,y)$ are generally not linear in y. For concreteness, consider a two action game with $\rho(1,y)$ concave and $\rho(2,y)$ linear.

(ii) Follows immediately from the indicated reduction of the argument of (6.5) to that of (6.4).

(iii) Continuity follows since the minimized function in (6.5) is continuous in its arguments α and (y_i), while the minimizing set is upper semi-continuous in y. □

We next establish that calibrated play attains the calibration envelope.

Theorem 6.2 ($\rho^{\mathbf{cal}}$ is attainable) *Suppose P1 uses a calibrated strategy. Then, for any strategy of P2,*

$$\liminf_{n\to\infty}(\rho_n - \rho^{cal}(\hat{y}_n)) \geq 0 \quad (a.s.)\,.$$

Proof: It will be convenient to use for this proof the shorthand notations $a_n \overset{o(n)}{=} b_n$ for $\lim_{n\to\infty}(a_n - b_n) = 0$, and $a_n \overset{o(n)}{\geq} b_n$ for $\liminf_{n\to\infty}(a_n - b_n) \geq 0$. All relations between random variables are assumed by default to hold with probability 1. Let $Y_i = \{y \in Y : i^o(y) = i\}$, so that $q_k \in Y_i$ implies $i_k = i$; note that $Y_i \subset Y_i^*$. We thus have

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n} r(i_k, j_k) &= \frac{1}{n}\sum_{i\in I}\sum_{k=1}^{n} \mathbf{1}\{q_k \in Y_i\} r(i, j_k) \\
&\overset{o(n)}{=} \frac{1}{n}\sum_{i\in I}\sum_{k=1}^{n} \mathbf{1}\{q_k \in Y_i\} r(i, q_k) \\
&= \frac{1}{n}\sum_{i\in I}\sum_{k=1}^{n} \mathbf{1}\{q_k \in Y_i\} r(i^o(q_k), q_k) \\
&= \frac{1}{n}\sum_{k=1}^{n} r(i^o(q_k), q_k)\,.
\end{aligned}$$

The second ($o(n)$) equality follows from (6.1). Repeating the argument for τ we obtain

$$\frac{1}{n}\sum_{k=1}^{n} \tau(i_k, j_k) \overset{o(n)}{=} \frac{1}{n}\sum_{k=1}^{n} \tau(i^o(q_k), q_k)\,.$$

Since $\tau(i,j)$ is bounded away from zero, it follows that

$$\rho_n \overset{o(n)}{=} \frac{\sum_{k=1}^{n} r(i^o(q_k), q_k)}{\sum_{k=1}^{n} \tau(i^o(q_k), q_k)}\,, \tag{6.6}$$

while the latter expression satisfied the following inequality by definition of ρ^{cal}:

$$\frac{\sum_{k=1}^{n} r(i^o(q_k), q_k)}{\sum_{k=1}^{n} \tau(i^o(q_k), q_k)} \geq \rho^{\mathrm{cal}}(\hat{q}_n)\,, \quad \text{where } \hat{q}_n = \frac{1}{n}\sum_{k=1}^{n} q_k\,.$$

Thus,

$$\rho_n \overset{o(n)}{\geq} \rho^{\mathrm{cal}}(\hat{q}_n)\,.$$

Note also that from (6.1), with $Q = Y$, we have $\hat{y}_n \overset{o(n)}{=} \hat{q}_n$. The required equality now follows by continuity for $\rho^{\mathrm{cal}}(y)$ in y, as noted in Lemma 6.1. □

The following immediate consequence provides a sufficient condition for the best-response envelope ρ^* to be attainable, namely for the existence of no-regret strategies.

Corollary 6.3 *Suppose that $\rho^{cal}(y) = \rho^*(y)$ for all $y \in Y$. Then ρ^* is attainable by P1.*

The condition of the last corollary is satisfied in standard (fixed-duration) repeated matrix games. In general, however, ρ^{cal} can be strictly smaller than ρ^*. In particular, this must be the case when ρ^* is not attainable.

We proceed to establish some basic bounds on ρ^{cal}, that highlight the performance guarantees of calibrated play.

Proposition 6.4 (Properties of ρ^{cal})

(a) $\mathrm{v}(r,\tau) \le \rho^{cal}(\hat{y}) \le \rho^*(\hat{y})$ *for all* $\hat{y} \in Y$.

(b) $\rho^{cal}(\hat{y}) = \rho^*(\hat{y})$ *at the extreme points of* Y, *which correspond to the pure action set* I.

(c) For each $\hat{y} \in Y$, $\rho^*(\hat{y}) > \mathrm{v}(r,\tau)$ *implies* $\rho^{cal}(\hat{y}) > \mathrm{v}(r,\tau)$.

Proof: (a) $\rho^{\text{cal}}(\hat{y}) \le \rho^*(\hat{y})$ follows since the argument of (6.4) equals $\rho^*(\hat{y})$ when μ is concentrated entirely on $\hat{y}$. The inequality $\rho^{\text{cal}}(\hat{y}) \ge \mathrm{v}(r,\tau)$ is established in part (c) below.

(b) The stated equality follows since μ in (6.4) must be entirely concentrated on $\hat{y}$ when $\hat{y}$ is an extreme point.

(c) We first note that $\rho^{\text{cal}}(\hat{y}) \ge \mathrm{v}(r,\tau)$, which follows from (6.4) since for every y and $i(y) \in I^*(y)$ we have

$$\frac{r(i(y),y)}{\tau(i(y),y)} = \max_{i\in I} \frac{r(i,y)}{\tau(i,y)} \ge \mathrm{v}(r,\tau) \tag{6.7}$$

where the last equality holds by Lemma 2.1(iii). Recall next that $Y^* = \{y \in Y : \rho^*(y) = \mathrm{v}(r,\tau)\}$ is a closed convex set by Lemma 2.1(ii). Consider some $\hat{y}$ with $\rho^*(\hat{y}) > \mathrm{v}(r,\tau)$, namely $\hat{y} \notin Y^*$, and let μ and $\{i(y), y \in Y\}$ attain the minimum in (6.5). Since $\sum_{i\in I} \alpha_i y_i = \hat{y} \notin Y^*$ and Y^* is convex, there is at least one y_0 in the support of μ (namely, $\mu(\{y_0\}) > 0$) so that $y_0 \notin Y^*$. Since $i(y_0) \in I^*(y_0)$ we obtain

$$\frac{r(i(y_0),y_0)}{\tau(i(y_0),y_0)} = \rho^*(y_0) > \mathrm{v}(r,\tau).$$

Noting that (6.7) holds for every y, it follows from (6.4) that $\rho^{\text{co}}(\hat{y}) > \mathrm{v}(r,\tau)$ as claimed. □

We may now examine the performance guarantees of calibrated play, as per the desired properties in Section 4. Similar to our discussion in the previous section, it is seen that *safety* is ensured by part (a) of the last lemma, while *adaptivity* follows for part (c).

Our third desideratum, best-response to stationary strategies, cannot be deduced from the calibration envelope alone, as ρ^{cal} can be strictly smaller that ρ^* (and indeed must be so when ρ^* is not attainable). Still, we will show in Subsection 6.4 that this property does hold in general. However, before tending to that we proceed to show that the calibration envelope dominates the convex best-response envelope from the previous section.

6.3 Comparison with the Convex Best-Response Envelope

The results obtained so far establish that both the convex best-response envelope ρ^{co} (defined in Section 5.3) and the calibration envelope ρ^{cal} are attainable, using different strategies. Here we compare these two performance envelopes, and show that the calibration envelope dominates ρ^{co}. We first show that ρ^{cal} is at least as large as ρ^{co}, and identify certain class of variable-duration games for which equality holds. We then provide an example where ρ^{cal} is strictly larger than ρ^{co}.

Proposition 6.5 (ρ^{cal} dominates ρ^{co})

(i) $\rho^{cal}(\hat{y}) \geq \rho^{co}(\hat{y})$ *for all* $\hat{y} \in Y$.

(ii) If the stage durations depend on P2's actions only, namely $\tau(i,j) = \tau_0(j)$*, then* $\rho^{cal} = \rho^{co}$.

Proof: Let us rewrite $\rho^{\rm co}$ in more explicit form. Recall that $\rho^{\rm co}$ is defined in terms of $\tilde{\rho}^{\rm co}$ in (5.7), while $\tilde{\rho}^{\rm co}$ is the lower convex hull of $\tilde{\rho}^*$. Thus, by Caratheodory's theorem,

$$\tilde{\rho}^{\rm co}(\pi) = \min\left\{\sum_{l=1}^{L}\gamma_l\tilde{\rho}^*(\pi_l) \,:\, L\geq 1,\, \gamma\in\Delta_L,\, \pi_l\in\Pi,\, \sum_l \gamma_l\pi_l = \pi\right\},$$

where Δ_L is the unit simplex in $\mathbb{R}^L$. Therefore,

$$\begin{aligned}\rho^{\rm co}(\hat{y}) &= \min\left\{\tilde{\rho}^{\rm co}(\pi) \,:\, \pi\in\Pi, \frac{\pi}{|\pi|} = \hat{y}\right\}\\ &= \min\left\{\sum_{l=1}^{L}\gamma_l\tilde{\rho}^*(\pi_l) \,:\, L\geq 1,\, \gamma\in\Delta_L,\, \pi_l\in\Pi,\, \frac{\sum_l\gamma_l\pi_l}{|\sum_l\gamma_l\pi_l|} = \hat{y}\right\}.\end{aligned}$$

Now, by the definition of Π in (5.2), $\pi_l\in\Pi$ implies that $\pi_l = y_l/\tau_l$ for some $y_l\in Y$ and $\tau_l\in T(y_l)$ (where $T(y_l)$ is defined in (5.3)). Also note that $\tilde{\rho}^*(y_l/\tau_l) = \rho^*(y_l)$ by definition of $\tilde{\rho}^*$. Therefore

$$\rho^{\rm co}(\hat{y}) = \min\left\{\sum_{l=1}^{L}\gamma_l\rho^*(y_l) \,:\, L\geq 1,\, \gamma\in\Delta_L,\, y_l\in Y,\, \tau_l\in T(y_l),\, \frac{\sum_l(\gamma_l/\tau_l)y_l}{\sum_l\gamma_l/\tau_l} = \hat{y}\right\}, \tag{6.8}$$

where we have used the fact that $|y_l| = 1$. We now restrict the range of variables in the argument of the last expression by choosing, for each given y_l,

$$\tau_l\in\tilde{T}(y_l) \stackrel{\triangle}{=} \{\tau(i_l,y_l) : i_l\in I^*(y_l)\}\subseteq T(y_l)\,.$$

Note that for $i_l\in I^*(y_l)$ we have

$$\rho^*(y_l) = \frac{r(i_l,y_l)}{\tau(i_l,y_l)}\,.$$

Thus,

$$\begin{aligned}\rho^{\rm co}(\hat{y}) \leq \min\Bigg\{\sum_{l=1}^{L}\gamma_l\frac{r(i_l,y_l)}{\tau(i_l,y_l)} \,:\, L\geq 1,\, \gamma\in\Delta_L, y_l\in Y,\, i_l\in I^*(y_l),\,\dots\\ \tau_l = \tau(i_l,y_l),\, \frac{\sum_l(\gamma_l/\tau_l)y_l}{\sum_l\gamma_l/\tau_l} = \hat{y}\Bigg\}.\end{aligned} \tag{6.9}$$

We next parameterize $\gamma\in\Delta_L$ by $\alpha\in\Delta_L$ in the form

$$\gamma_l = \frac{\tau_l\alpha_l}{\sum_{l=1}^{L}\tau_l\alpha_l}\,.$$

This finally gives, after cancelling out $\tau_l = \tau(i_l,y_l)$ and some rearranging,

$$\begin{aligned}\rho^{\rm co}(\hat{y}) &\leq \min\left\{\frac{\sum_{l=1}^{L}\alpha_l r(i_l,y_l)}{\sum_{l=1}^{L}\alpha_l\tau(i_l,y_l)} \,:\, L\geq 1,\, \alpha\in\Delta_L,\, y_l\in Y,\, i_l\in I^*(y_l),\, \sum_l\alpha_l y_l = \hat{y}\right\}\\ &= \rho^{\rm cal}(\hat{y})\,,\end{aligned}$$

where the last equality is evident by the definition of $\rho^{\rm cal}$ in (6.4). This establishes part (i) of the proposition. Part (ii) follows after noting that under the stated condition $\tilde{T}(y_l) = \{\tau_0(y_l)\} = T(y_l)$, so that equality holds in (6.9). □

Part (ii) of this proposition implies, in particular, that $\rho^{\rm cal} = \rho^{\rm co}$ for the game of Example 3.1. The next example shows that $\rho^{\rm cal}$ and $\rho^{\rm co}$ need not coincide in general.

Example 6.1 ($\rho^{\rm cal}$ strictly dominates $\rho^{\rm co}$). Consider the variable duration matrix game $\Gamma(r,\tau)$ defined by the following matrix:

$$\begin{pmatrix} (0,1) & (2,3) \\ (2,3) & (0,1) \end{pmatrix}.$$

As before, P1 is the row player, P2 the column player, and the ij-th entry is $(r(i,j),\tau(i,j))$. As y is two-dimensional it is uniquely determined by its second coordinate $y(2)$, and we shall henceforth identify y with the scalar $y(2)$. It follow that

$$\rho^*(y) = \max\left\{\frac{2y}{1+2y}, \frac{2-2y}{3-2y}\right\}.$$

Note that $I^*(y) = \{2\}$ for $y < 0.5$, and $I^*(y) = \{1\}$ for $y > 0.5$. Similarly, $Y_1^* = [0.5, 1]$ and $Y_2^* = [0, 0.5]$, where both are convex sets. Further note that $\rho^*(y)$ is strictly decreasing in y for $y \in [0, 0.5)$, and strictly increasing for $y \in (0.5, 1]$.

We first claim that $\rho^{\rm cal} = \rho^*$. To see that, fix y and let $\alpha \in \Delta(I)$ and $y_i \in co(Y_i^*) = Y_i^*$, $i = 1, 2$, attain the minimum in (6.5). Assume first that $y \in Y_1^*$ (i.e., $y \geq 0.5$). If $\alpha_2 = 0$ then $y = y_1$ and $\rho^{\rm cal}(y) = r(1,y_1)/r(1,y_1) = \rho(1,y_1) = \rho^*(y)$. Suppose then that $\alpha_2 > 0$. It follows by symmetry of the game parameters for $r(1,y) = r(2,1-y)$ and similarly for τ. As a result we have

$$\begin{aligned}\rho^{\rm cal}(y) &= \frac{\alpha_1 r(1,y_1) + \alpha_2 r(2,y_2)}{\alpha_1 \tau(1,y_1) + \alpha_2 \tau(2,y_1)} = \frac{\alpha_1 r(1,y_1) + \alpha_2 r(1,1-y_2)}{\alpha_1 \tau(1,y_1) + \alpha_2 \tau(1,1-y_2)} \\ &\geq \rho^*(\alpha_1 y_1 + \alpha_2(1-y_2)) = \rho^*(y + \alpha_2(1-2y_2)) \;\geq\; \rho^*(y),\end{aligned}$$

where the last inequality follows since $\rho^*(y)$ is strictly increasing for $y \geq 1/2$ (as noted above), and this inequality is in fact strict if $y_2 \neq 0.5$. It follows that $\rho^{\rm cal}(y) = \rho^*(y)$ for $y \in Y_1^*$. The remaining case $y \in Y_2^*$ is symmetric and the required equality follows similarly.

We next estimate $\rho^{\rm co}(y)$ at a specific point. Let $y_1 = (0,1)$ and $y_2 = (1/2, 1/2)$. The following observations are immediate from the definitions in the previous sections:

1. $T(y_1) = [1,3]$, $T(y_2) = [1,3]$.
2. $\rho^*(y_1) = \frac{2}{3}$ and $\rho^*(y_2) = \frac{1}{2}$.
3. Let $\pi_1 = (0, \frac{1}{2}) \in \Pi$ and $\pi_2 = (\frac{1}{4}, \frac{1}{4}) \in \Pi$. Note that $\pi_1/|\pi_1| = y_1$ and $\pi_2/|\pi_2| = y_2$.
4. $\tilde{\rho}^*(\pi_1) = \rho^*(y_1) = \frac{2}{3}$ and $\tilde{\rho}^*(\pi_2) = \rho^*(y_2) = \frac{1}{2}$.
5. Let $\pi_3 = \frac{1}{2}\pi_1 + \frac{1}{2}\pi_2 = (\frac{1}{8}, \frac{3}{8})$, and let $y_3 = \pi_3/|\pi_3| = (\frac{1}{4}, \frac{3}{4})$.
6. By convexity, $\tilde{\rho}^{\rm co}(\pi_3) = \tilde{\rho}^{\rm co}(\frac{1}{2}\pi_1 + \frac{1}{2}\pi_2) \leq \frac{1}{2}(\frac{2}{3} + \frac{1}{2}) = \frac{7}{12}$, so that $\rho^{\rm co}(y_3) \leq \frac{7}{12}$.

But from our previous calculation we know that $\rho^{\rm cal}(y_3) = \rho^*(y_3) = \frac{3}{5}$. Therefore, $\rho^{\rm cal}(y_3) > \frac{7}{12} \geq \rho^{\rm co}(y_3)$. □

A plot of $\rho^{\rm cal} = \rho^*$ and $\rho^{\rm co}$ for the last example is shown in Figure 3. We note that $\rho^{\rm co}$ was computed analytically, but details are omitted.

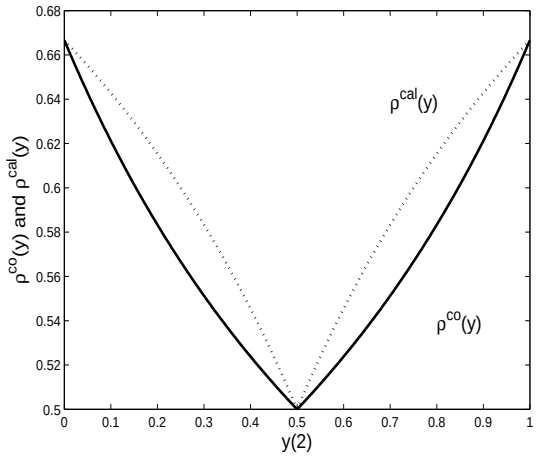

Figure 3: $\rho^{\mathrm{cal}}(y)$ (dotted line) and $\rho^{\mathrm{co}}(y)$ (thick line) for the game of Example 6.1.

6.4 Best-response to Stationary Strategies

We next show that the *best-response* payoff $\rho^*(\hat{y}_n)$ is attained by calibrated play when P2 follows a stationary strategy (or even close to that in a certain sense). We note that achieving the best response payoff when the opponent is restricted to stationary policies is easily obtainable by much simpler policies (such as fictitious play). The point here is, then, that calibrated play achieves that while simultaneously securing the above-presented performance guarantees against an arbitrary opponent.

We start the analysis with the following lemma.

Lemma 6.6 *Suppose the sequence $(y_k)_{k\geq 1}$ of P2's mixed actions Cesaro-converges to a stationary strategy $y_0 \in Y$, in the sense that*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} |y_k - y_0| = 0 \quad \textit{(a.s.)}\,. \tag{6.10}$$

Then any calibrated forecast (q_k) Cesaro-converges y_0 in the same sense; equivalently, for any $\epsilon > 0$ we have that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\{|q_k - y_0| > \epsilon\} = 0 \quad \textit{(a.s.)}\,. \tag{6.11}$$

Proof: Let P1 use a calibrated forecasting scheme, and let P2 use an arbitrary strategy. For the rest of this proof all probabilistic relations hold by default with probability 1, and we omit the *a.s.* quantifier. Observe first that the calibration requirement (6.1) implies the following *merging* property (see Kalai et al., 1999)

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 1\{q_k \in Q\}(y_k - q_k) = 0\,. \tag{6.12}$$

This follows by the strong law of large numbers, applied to the Martingale difference sequence $D_k = 1\{q_k \in Q\}(e_{j_k} - y_k)$. Combined with (6.10), this clearly implies that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 1\{q_k \in Q\}(y_0 - q_k) = 0\,. \tag{6.13}$$

Consider a set $Q \subset Y$ so that $\overline{co}(Q)$, the closed convex hull of Q, does not contain y_0. We claim that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\{q_k \in Q\} = 0\,. \tag{6.14}$$

To see that, denote

$$z_n = \frac{\sum_{k=1}^{n} \mathbf{1}\{q_k \in Q\} q_k}{\sum_{k=1}^{n} \mathbf{1}\{q_k \in Q\}}$$

(z_n may be chosen arbitrarily in Q when the denominator is 0), and observe that $z_n \in \overline{co}(Q)$ by its definition. Expanding on (6.13), we obtain

$$\begin{aligned} 0 &= \lim_{n\to\infty} \frac{1}{n} \left| \sum_{k=1}^{n} 1\{q_k \in Q\}(y_0 - q_k) \right| \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 1\{q_k \in Q\} |y_0 - z_n| \\ &\geq \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 1\{q_k \in Q\}\, d(y_0, \overline{co}(Q))\,. \end{aligned}$$

Recalling that $y_0 \notin \overline{co}(Q)$ by choice of Q, we have that $d(y_0, \overline{co}(Q)) > 0$, and (6.14) follows. To conclude, note that for any $\epsilon > 0$ the set $\{q \in Y : |q - y_0| > \epsilon\}$ may be written as a finite union $\bigcup_{m=1}^{M} Q_m$, where the sets Q_m are mutually exclusive and with $y_0 \notin \overline{co}(Q_m)$. Hence (6.14) holds for each Q_m, and since

$$\mathbf{1}\{|q_k - y_0| > \epsilon\} = \sum_{m=1}^{M} \mathbf{1}\{q_k \in Q_m\}$$

we obtain (6.11). The Cesaro convergence of q_k to y_0 follows from (6.11) since the sequence (q_k) is bounded. □

Proposition 6.7 *Let P1 use a the calibrated strategy (6.3), and suppose the sequence (y_k) of P2's mixed actions Cesaro-converges to a stationary strategy y_0, in the sense of (6.10). Then*

$$\lim_{n\to\infty} \rho_n = \lim_{n\to\infty} \rho^*(\hat{y}_n) = \rho^*(y_0) \quad \textit{(a.s.)}\,.$$

Proof: We first note that the assumption on (y_k) implies that $\lim_{n\to\infty} \hat{y}_n = y_0$, so that the second equality follows by continuity of ρ^* (see Lemma 2.1). To establish the first equality, observe from Lemma 6.6 that (6.11) holds for any $\epsilon > 0$, , so that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\{|q_k - y_0| \leq \epsilon\} = 1 \quad \textit{(a.s.)}\,. \tag{6.15}$$

Now, if $|q_k - y_0| \leq \epsilon$ it follows by the definition of $i^o(q)$ and the Lipschitz continuity of ρ^* (see *ibid*) that

$$\left| \frac{r(i^o(q_k), q_k)}{\tau(i^o(q_k), q_k)} - \rho^*(y_0) \right| = |\rho^*(q_k) - \rho^*(y_0)| \leq A\epsilon\,,$$

where $A > 0$ depends only on the game parameters (r and τ). Since (6.6) holds under myopic calibrated play, we obtain by combining the last two equations that

$$\limsup_{n\to\infty} |\rho_n - \rho^*(y_0)| = \limsup_{n\to\infty} \left| \frac{\sum_{k=1}^{n} r(i^o(q_k), q_k)}{\sum_{k=1}^{n} \tau(i^o(q_k), q_k)} - \rho^*(y_0) \right| \leq \epsilon A\,.$$

As this holds for any $\epsilon > 0$, the conclusion follows. □

The following remarks concern Proposition 6.7.

1. The conclusion of Proposition 6.7 can be seen to hold even when the stationary strategy to which P2 converges depends on the sample path. More precisely, on a set of probability 1, if (6.10) holds for some $y_0 = y_0(\omega)$, then $\lim_{n\to\infty} \rho_n = \rho^*(y_0)$.

2. Proposition 6.7 may also be generalized by considering the case where P2's mixed actions converge to any convex set $Y_0 \subset Y$, rather than a single point, in the sense that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} d(y_k, Y_0) = 0\,.$$

   In that case, a similar argument to the above shows that, for any $\epsilon > 0$,

$$\liminf_{n\to\infty} (\rho_n - \rho^{\mathrm{cal}}(\hat{y}_n, Y_0^\epsilon)) \geq 0 \ \textit{(a.s.)}\,,$$

   where Y_0^ϵ is an ϵ blowup of Y_0, and $\rho^{\mathrm{cal}}(\hat{y}, Y_0^\epsilon)$ is defined similarly to (6.4) except that Y_0^ϵ replaced Y as the support of the measure μ.

3. It should noted that convergence of the empirical distribution $\hat{y}_n$ to some value y_0 is not sufficient to obtain $\rho_n \geq \rho^*(y_0)$ asymptotically. For example, if P2 plays periodically the actions $a, b, a, b, ...$, then a best-response to a perfect prediction of these actions will yield the long-term reward rate $\rho_n \simeq \frac{r(i^o(a),a))+r(i^o(b),b)}{\tau(i^o(a),a))+\tau(i^o(b),b)}$. This may be smaller (or larger) than $\rho^*(\frac{1}{2}e_a + \frac{1}{2}e_b)$.

We have thus established that calibrated play satisfies all three desiderata that we posed for adaptive play. We close this section with some remarks that concern that complexity of calibration.

1. As noted before, devising a computationally feasible algorithm for calibration is a non-trivial task. With the exception of Foster and Vohra (1997) for binary sequences and Mannor et al. (2007) for some other restricted cases, currently available calibration algorithms are inefficient. Settling for ϵ-calibration, where the magnitude of the error in (6.1) is allowed to grow up to some $\epsilon > 0$, may lead to more feasible algorithms (e.g., Kakade and Foster, 2004). When used for calibrated play this would lead to a proportional reduction in the guaranteed performance envelope ρ^{cal}.

2. To attain ρ^{cal}, we actually need the calibration property in (6.1) to hold only for $Q = Y$ and $Q = Y_i^*$, $i \in I$ (as these are the properties used in the proof of Theorem 6.2). This should be simpler than general calibration. However, the resulting scheme does not guarantee best-response to stationary strategies.

7 Concluding Remarks

In this final section we consider briefly some possible alternatives and extensions to the adaptive strategies that were suggested above. We further discuss some relations between the present model and stochastic games, and finally conclude the paper.

7.1 Alternative Strategies

A number of variants and alternatives of interest exist for the adaptive strategies that were proposed in this paper. We mention the following two:

1. *Modified calibrated play:* According to the calibration-related strategy of Section 6, the adaptive player (P1) maximizes his single-stage payoff $\rho(x, q_k)$ against his calibrated forecast q_k of the opponent's action. A reasonable alternative for P1 would be to choose his action so as to maximize the expected average reward up to and including the next stage, namely

$$x_k \in \arg\max_{x \in X} \frac{r_1 + \ldots + r_{k-1} + r(x, q_k)}{\tau_1 + \ldots + \tau_{k-1} + \tau(x, q_k)} .$$

or even

$$x_k \in \arg\max_{x \in X} \sum_{i,j} x(i) q_k(j) \frac{r_1 + \ldots + r_{k-1} + r(i,j)}{\tau_1 + \ldots + \tau_{k-1} + \tau(i,j)} .$$

Recalling that P1 wishes to maximize his long-term average payoff, these schemes may be viewed as natural *greedy* choices with respect to the current average. However, no performance guarantees are currently available for these modified schemes when playing an arbitrary opponent. We note that all three options coincide for standard (fixed duration) matrix games.

2. *Fictitious play:* In fictitious play, P1 selects the best response against the current empirical distribution of the opponent's actions. No-regret play is obtained for repeated matrix games by slightly perturbing (or smoothing) this choice; see Fudenberg and Levine (1999) and references therein. Fictitious play and it variants may be defined in our model using

$$x_k \in \arg\max_{x} \frac{r(x, \hat{y}_{k-1})}{\tau(x, \hat{y}_{k-1})}$$

as a starting point. It should be easy to verify that such schemes nullify the regret when the opponent is stationary, or, more generally, when the empirical distribution of his actions converges. However, no performance guarantees against an arbitrary opponent are currently available.

A somewhat different approach to the definition of regret in variable duration games is the *super-game approach* considered in Mannor (2002). The idea is to aggregate a variable number of consecutive stages of the basic game into a larger "super-game", so that the overall length of this game is approximately constant (in relative terms), to within a required accuracy. We then consider no-regret strategies for the repeated super-game, where the opponents actions are now his strategies in the stage super-game. While these super-actions are not fully observable, we can still apply existing results that rely on the observed payoff only (e.g., Auer et al., 2002) to obtain (approximate) no-regret strategies in this super game. The downside of this approach is that the performance guarantees are not defined in the natural space of $\hat{y}_n$ (the opponent's empirical distribution in the actual stage game), and that the complexity and convergence time of the resulting strategy are exponential in the number of stages in the super-game, which quickly becomes prohibitive.

7.2 Regret minimization in stochastic games

There exist some interesting inter-relations between our model and certain classes of stochastic games which are worth pointing out. First, repeated variable-duration games can be represented as a special case of stochastic games. Specifically, when the stage durations $\tau(i,j)$ are integer (or, by scaling,

rational) numbers, the variable duration game can be represented as a finite-state stochastic game, where the decision points coincide with the arrival times to a fixed state. The long-term average reward in this stochastic game coincides with the long-term reward-rate in the repeated variable-duration game. It follows that the *negative* results derived here on the non-existence of no-regret strategies immediately apply to the more general class of stochastic games, even under the assumption of a fixed recurrent state.

We considered regret minimization for stochastic games in a previous work (Mannor and Shimkin, 2003). A counterexample was given there to existence of no-regret strategies, where regret is defined with respect to best response payoff against the stationary strategy of the opponent that corresponds to the conditional empirical distribution of the opponent's actions at each state. The example given in the present paper offers additional insight, as it pinpoints the problem to the variable duration of time between arrivals to the recurrent state, rather than to the state dynamics. In that work we also considered approachability-related adaptive strategies for recurrent stochastic games, and introduced the concept of the convex Bayes-envelope which parallels the convex best-response envelope of the present paper. However, the performance guarantees obtained here are stronger.

Regret minimization in stochastic games was considered also in Even-Dar et al. (2004). This model is restricted to the case where P1 alone affects the state transitions, it assumes that the potential reward in *all* states is revealed to P1 at every stage, and it considers the *expected* average reward criterion (rather than the sample-path average).

¿From the opposite direction, it is worth pointing out certain classes of stochastic games may in principle be reduced to repeated variable-duration games. Assuming the existence of a fixed state that is recurrent under all strategies (as in Mannor and Shimkin (2003) for example), we can consider segment of the game between successive visits to the recurrent state as a normal-form super-game, in which the player's actions are their pure strategies in such segment. This raises the possibility of translating regret minimizing schemes for repeated variable duration games to this class of stochastic games in the appropriate super-action space. The caveat is that the other player's super-actions are now not fully observed by P1, a problem which was not treated in the present paper.

7.3 Conclusion

In this paper we considered the extension of the regret minimization concept by allowing the stage game to be of variable duration. It was shown that a natural extension of the no-regret concept to this case cannot be supported in general (although no-regret strategies do exist in the special case when P1 does not affect the stage game duration). Motivated by this inherent limitation, we studied two classes of strategies for adaptive play that provide somewhat less favorable, but still significant, performance guarantees. The first one is based on approachability, and its performance guarantee is defined in terms of a certain function of $\hat{y}_n$, the empirical distribution of the opponent's actions, which was termed the convex best-response envelope. In particular, it was shown that this strategy is indeed *adaptive*, in the sense that it achieves more than the minimax value of the game when the opponent's empirical distribution is non-adversarial. The second strategy is based on calibrated play, namely playing the best response action in a related single-shot game against a calibrated forecast of the opponent's actions. This strategy was shown to attain the *calibration envelope*, which is at least as high as the convex best-response envelope of the previous strategy, and moreover attains the best

response (hence no regret) if the opponent happens to use a stationary strategy.

Several directions and issues remain for future work. First, the calibration-based scheme is quite demanding, and it should be of interest to obtain similar performance using simpler strategies. Second, a challenging question is to determine whether the performance guarantees of the calibration envelope can be improved upon, and indeed whether a sense of *optimal* performance envelope exists in general. Third, each of the alternative strategy options outlined above has some conceptual appeal, and a study of their properties is called for. Of particular interest here would be to develop fictitious play-like strategies with provable performance guarantees. Finally, it would be of interest to study adaptive strategies for the variable-duration model under incomplete observation of the opponent's action, similar to the bandit problem in repeated matrix games (Auer et al., 2002) or the general signalling model of Rustichini (1999), and make the connection mentioned above with stochastic games.

References

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, *32*(1), 48–77.

Blackwell, D. (1956a). An analog of the minimax theorem for vector payoffs. *Pacific J. Math.*, *6*(1), 1–8.

Blackwell, D. (1956b). Controlled random walks. In *Proc. Int. Congress of Mathematicians 1954,* Vol. 3, 335–338. Amsterdam: North Holland.

Boyd, S., and Vanderberghe, L. (2004). *Convex Optimization.* Cambridge, UK: Cambridge University Press.

Cesa-Bianchi, N., and Lugosi, G. (2006). *Prediction, Learning and Games.* New-York, NY: Wiley-Interscience.

Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Even-Dar, E., Kakade, S. M., and Mansour, Y. (2004). Experts in a markov decision process. In L. K. Saul, Y. Weiss, and L. Bottou (Eds.), *NIPS 17: Advances in Neural Information Processing Systems* (pp. 401–408). Cambridge, MA: MIT Press.

Foster, D. P. (1999). A proof of calibration via Blackwell's approachability theorem. *Games Econ. Behav.*, *29*(1-2), 73–78.

Foster, D. P., and Vohra, R. (1999). Regret in the on-line decision problem. *Games Econ. Behav.*, *29*, 7–35.

Foster, D. P., and Vohra, R. V. (1997). Calibrated learning and correlated equilibrium. *Games Econ. Behav.*, *21*, 40–55.

Foster, D. P., and Vohra, R. V. (1998). Asymptotic calibration. *Biometrika*, *85*, 379–390.

Freund, Y., and Schapire, R. E. (1999). Adaptive game playing using multiplicative weights. *Games Econ. Behav.*, *29*, 79–103.

Fudenberg, D., and Levine, D. (1995). Universal consistency and cautious fictitious play. *J. Econ. Dyn. Control*, *19*, 1065–1990.

Fudenberg, D., and Levine, D. (1999a). Conditional universal consistency. *Games Econ. Behav.*, *29*, 104–130.

Fudenberg, D., and Levine, D. (1999b). An easier way to calibrate. *Games Econ. Behav.*, *29*, 131–137.

Fudenberg, D., and Levine, D. K. (1999). *The Theory of Learning in Games.* Cambridge, Massachusetts: MIT Press.

Hannan, J. (1957). Approximation to Bayes risk in repeated play. In M. Dresher, A. W. Tucker, and P. Wolfe (Eds.), *Contributions to the Theory of Games,* Vol. III, Annals of Mathematical Studies, Vol. 39, pp. 97–193. Princton, NJ: Princeton University Press.

Hart, S. (2005). Adaptive heuristics. *Econometrica*, *73*(5), 1401–1430.

Hart, S., and Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, *68*, 1127-1150.

Hart, S., and Mas-Colell, A. (2001). A general class of adaptive strategies. *J. Econ. Theory*, *98*, 26–54.

Kakade, S., and Foster, D. P. (2004). Deterministic calibration and Nash equilibrium. In J. Shawe-Taylor and Y. Singer (Eds.), *COLT 2004: Proc. 17th Annual Conference on Learning Theory.* Lecture Notes in Computer Science, 3120, pp. 33–48. Springer.

Kalai, E., Lehrer, E., and Smorodinsky, R. (1999). Calibrated forecasting and merging. *Games Econ. Behav.*, *29*(1-2), 151-169.

Lal, A. A., and Sinha, S. (1992). Zero-sum two-person semi-Markov games. *J. Appl. Prob.*, *29*, 56–72.

Lehrer, E. (2003). A wide range no-regret theorem. *Games Econ. Behav.*, *42*, 101–115.

Mannor, S. (2002). *Reinforcement learning and adaptation in competitive environments.* Doctoral Dissertation, Faculty of Electrical Enginieering, Technion.

Mannor, S., Shamma, J., and Arslan, G. (2007). Online calibrated forecasts: Efficiency versus universality for learning in games. *Machine Learning*, *67*(2), 77–115.

Mannor, S., and Shimkin, N. (2003). The empirical Bayes envelope and regret minimization in competitive Markov decision processes. *Math. Oper. Res.*, *28*(2), 327–345.

Rustichini, A. (1999). Minimizing regret: the general case. *Games Econ. Behav.*, *29*, 224–243.

Sandroni, A., Smorodinsky, R., and Vohra, R. V. (2003). Calibration with many checking rules. *Math. Oper. Res.*, *28*(1), 141–153.

Shapley, L. (1953). Stochastic games. *Proc. Natl. Acad. Sci. USA*, *39*, 1095–1100.

Shimkin, N., and Shwartz, A. (1993). Guaranteed performance regions in markovian systems with competing decision makers. *IEEE Trans. Automat. Contr.*, *38*(1), 84–95.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proc. 20th Internat. Conf. on Mach. Learning (ICML 2003), Washington, DC,* (pp. 928–936). AAAI Press.

A Appendix

A.1 Proof of Theorem 5.1

We provide here an outline of the proof of the approachability result in Theorem 5.1. The proof of part (i) is essentially a simplified version of the proof of Theorem 3.1 in Shimkin and Shwartz (1993), which deals with approachability for stochastic games with a fixed recurrent state.

We will require the following Martingale-related convergence result (see, e.g., Shimkin and Shwartz, 1993, Proposition 4.1).

Proposition A.1 *Let $(X_n, \mathcal{F}_n)_{n\geq 0}$ be a stochastic sequence on some probability space, namely $(\mathcal{F}_n)$ is an increasing sequence of sigma-algebras and X_n is measurable on $\mathcal{F}_n$. Let $X_0 = 0$, and suppose there exists a constant Q such that, for each $n \geq 0$,*

$$\mathbb{E}(X_{n+1}^2|\mathcal{F}_n) \leq X_n^2 + Q \quad \textit{(a.s.)}\,.$$

Then $X_n/n \to 0$ (almost surely). Furthermore, the rate of convergence depends only on Q, in the sense that

$$P\left\{\sup_{n\geq N} \frac{|X_n|}{n} \geq \epsilon\right\} \leq \delta \quad \textit{for} \;\; N \geq \frac{6Q}{\delta\epsilon^2}\,.$$

Assume that the condition in Theorem 5.1(i) is satisfied for a given set B. Suppose that P1 uses the specified strategy, while P2 uses an arbitrary strategy. Let $\mathcal{F}_n$ be the sigma-algebra generated by the history sequence $h_n = (i_1, j_1, \ldots, i_n, j_n)$. Further denote

$$\begin{aligned}
&T_n = \textstyle\sum_{k=1}^n \tau_k\,, \text{ where } \tau_k = \tau(i_k, j_k),\\
&r_k = \vec{r}(i_k, j_k),\\
&C_n = c_{\rho_n},\ \text{the closest point in } B \text{ to } \rho_n, d_n \qquad\qquad = d(\rho_n, B) \equiv \|\rho_n - C_n\|_2,
\end{aligned}$$

where we use the standard Euclidean norm. We proceed to show that $E(T_{n+1}^2 d_{n+1}^2|\mathcal{F}_n) \leq T_n^2 d_n^2 + Q$ for some finite constant Q. If $d_n > 0$ (namely $\rho_n \notin B$) then, from (5.1) and the specified policy of P1 we have

$$\big\langle C_n - \rho_n,\, C_n - \vec{\rho}(x_{n+1}, y_{n+1})\big\rangle \leq 0\,. \tag{A.1}$$

If $d_n = 0$ this holds trivially since $C_n - \rho_n$. Observe that

$$\vec{\rho}(x_{n+1}, y_{n+1}) = \frac{\vec{r}(x_{n+1}, y_{n+1})}{\tau(x_{n+1}, y_{n+1})} = \frac{E(r_{n+1}|\mathcal{F}_n)}{E(\tau_{n+1}|\mathcal{F}_n)},$$

so that the inequality (A.1) can be written as

$$E\left(\big\langle C_n - \rho_n\,,\, C_n\tau_{n+1} - r_{n+1}\big\rangle|\mathcal{F}_n\right) \leq 0\,. \tag{A.2}$$

Next, by definition of C_{n+1} we have

$$d_{n+1} = \|\rho_{n+1} - C_{n+1}\| \leq \|\rho_{n+1} - C_n\| = \left\|\frac{T_n\rho_n + r_{n+1}}{T_{n+1}} - C_n\right\|.$$

Therefore,

$$T_{n+1}^2 d_{n+1}^2 \leq \|T_n\rho_n + r_{n+1} - C_n T_{n+1}\|^2 = \|T_n(\rho_n - C_n) + (r_{n+1} - C_n\tau_{n+1})\|^2\,.$$

Expanding the last squares, using (A.2) and recalling that $d_k = \|\rho_n - C_n\|$ gives

$$E(T_{n+1}^2 d_{n+1}^2 | \mathcal{F}_n) \leq T_n^2 d_n^2 + E(\|r_{n+1} - C_n \tau_{n+1}\|^2 | \mathcal{F}_n) \leq T_n^2 d_n^2 + Q\,,$$

where $Q = (r_{\max} + \tau_{\max} C_{\max})^2$ is a (finite) upper bound on the last term. Applying Proposition A.1 we can deduce that

$$\lim_{n\to\infty} \frac{1}{n} T_n d_n = 0\,.$$

But since $T_n/n \leq \tau_{\max}$, it follows that $\lim_{n\to\infty} d_n = 0$, which establishes part (i) of the Theorem. We note that a uniform convergence rate applies to all strategies of P2, give by

$$P\left\{\sup_{n\geq N} d_n \geq \epsilon\right\} \leq \delta \quad \text{for} \quad N \geq \frac{6Q}{\delta(\epsilon\tau_{\min})^2}\,.$$

Part (ii) of the theorem follows by using the minimax argument of Blackwell (1956a), after noting the minimax equality of Lemma 2.1(i) which applies to the projected payoff functions $\rho_u = \langle u, \vec{\rho}\rangle$. $\square$