

Online Learning with Variable Stage Duration

Shie Mannor¹ and Nahum Shimkin²

¹ Department of Electrical and Computer Engineering
McGill University, Québec H3A-2A7
`shie@ece.mcgill.ca`

² Department of Electrical Engineering
Technion, Israel Institute of Technology
Haifa 32000, Israel
`shimkin@ee.technion.ac.il`

Abstract. We consider online learning in repeated decision problems, within the framework of a repeated game against an arbitrary opponent. For repeated matrix games, 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, in hindsight, by playing any fixed action against the observed actions of the opponent. In the present paper we consider the extended model 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 online learning in repeated games with variable stage duration 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.

1 Introduction

Consider a repeated game from the viewpoint of a specific 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.

This problem was considered by [12], in the context of repeated matrix games. Hannan introduced the Bayes envelope 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 .

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 [12] 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 ([3]), smooth fictitious play ([10]), calibrated forecasts ([6]), and multiplicative weights ([9]) among others. We refer the reader to [5] for a discussion and an extensive literature review.

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 start, for example, with board games like Chess (where the game duration can be taken as the number of moves or the actual time played), and continue with gambling (where different options can take a different time per round), investment options, and choosing between projects or treatments with different durations. 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.

Our purpose then is to examine decision strategies and performance goals that are suitable for adaptive play against an arbitrary opponent in repeated variable-duration games. While this model may be viewed as the simplest non-trivial extension of standard repeated games, it turns 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 attainable when the stage duration depends only on player 2's action. However, a 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.

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. Given this negative result, we look for strategies that offer some reasonable performance guarantees. In Section 4 we consider a strategy based on approachability. By applying a convexification procedure to the Bayes envelope, we exhibit a weaker performance goal, the convex Bayes envelope, which is indeed attainable. This strategy is reminiscent to our previously developed strategy in [15] for stochastic game and is provided here for reference; Section 4 can therefore be skipped by readers who are familiar with [15]. In Section 5 we introduce our main solution concept, calibrated play and its associated performance guarantees. Section 6 briefly offers directions for further study. Some of the proofs are omitted and appear in [16].

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 $r : I \times J \rightarrow \mathbb{R}$ and a strictly positive duration function $\tau : I \times J \rightarrow (0, \infty)$. 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. 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 $k = 1, 2, \dots$, 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

$$\rho_n = \frac{\sum_{k=1}^n r_k}{\sum_{k=1}^n \tau_k}. \quad (1)$$

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 seeks 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 $\sigma^1 : H \rightarrow \Delta(I)$, where H is the set of all possible history sequences of the form $h_k = (i_1, j_1, \dots, i_k, j_k)$, $k \geq 0$ (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 $\Gamma^\infty \equiv \Gamma^\infty(r, \tau)$.

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 $r(i, y) = \sum_j r(i, j)y_j$, and $r(x, y) = \sum_{i,j} x_i r(i, j)y_j$, and similarly for τ .

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

$$\rho(x, y) \triangleq \frac{r(x, y)}{\tau(x, y)} = \frac{\sum_{i,j} x_i r(i, j)y_j}{\sum_{i,j} x_i \tau(i, j)y_j}. \quad (2)$$

This function plays 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

$$\lim_{n \rightarrow \infty} \rho_n = \rho(x, y) \quad (a.s.) \quad (3)$$

$$\lim_{n \rightarrow \infty} \mathbb{E}(\rho_n) = \rho(x, y). \quad (4)$$

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) is *not* bilinear in its arguments. We next establish that this game has a value, which we denote by $v(r, \tau)$, as well as some additional properties of the reward-rate function ρ .

Lemma 1 (Basic properties of ρ).

- (i) $v(r, \tau) \triangleq \max_{x \in X} \min_{y \in Y} \rho(x, y) = \min_{y \in Y} \max_{x \in X} \rho(x, y)$.
- (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 $\rho^*(y) \triangleq \max_{x \in X} \rho(x, y)$ is Lipschitz continuous in y .

Proof. The stated results may be deduced from known ones for semi-Markov games; see [14]. For completeness, a proof can be found in [16]. \square

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.

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, $\rho^* : Y \rightarrow \mathbb{R}$, is defined by

$$\rho^*(y) \triangleq \max_{i \in I} \frac{r(i, y)}{\tau(i, y)} = \max_{i \in I} \rho(i, y). \quad (5)$$

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

$$\rho^*(y) = \max_{x \in X} \frac{r(x, y)}{\tau(x, y)} = \max_{x \in X} \rho(x, y), \quad (6)$$

as per Lemma 1(iii).

We consider the difference $\rho^*(\hat{y}_n) - \rho_n$ as P1's n -stage *regret*. 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 1 (No-regret strategies). *A strategy σ^1 of P1 is a no-regret strategy if, for every strategy of P2,*

$$\liminf_{n \rightarrow \infty} (\rho_n - \rho^*(\hat{y}_n)) \geq 0 \quad (a.s.). \quad (7)$$

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 2. *Suppose that P2 uses a fixed sequence of actions (j_1, \dots, 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.*

Proof. With $i_k \equiv i$ we obtain, by (1), $\rho_n = \frac{\sum_{k=1}^n r(i, j_k)}{\sum_{k=1}^n \tau(i, j_k)} = \frac{r(i, \hat{y}_n)}{\tau(i, \hat{y}_n)} = \rho(i, \hat{y}_n)$. The required conclusion follows by definition of ρ^* . \square

The last lemma indicates that ρ^* is indeed the natural extension of Hannan's best-response envelope. The Lemma 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.

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 the standard model with fixed stage durations.

Proposition 1. *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, $\rho^*(\hat{y}_n) = \max_i (r(i, \hat{y}_n) / \tau(\hat{y}_n))$. 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 $\liminf_{n \rightarrow \infty} (\hat{r}_n - \max_i r(i, \hat{y}_n)) \geq 0$. 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 known to exist. \square

The situation becomes more intricate when the stage durations do depend on P1's actions, as demonstrated in the following example.

Example 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 $(r(i, j), \tau(i, j))$.

Proposition 2. *The best-response envelope is not attainable by P1 in the game $\Gamma^\infty(r, \tau)$ defined by Example 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:

$$j_n = \begin{cases} 1 & \text{for } 1 \leq n \leq N, \\ 2 & \text{for } N + 1 \leq n \leq 2N. \end{cases} \quad (8)$$

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 $\zeta_1 = \sum_{i=1}^N 1\{i_k = 1\} / N$ denote the empirical distribution of P1's action 1 over the first N stages. It is easily seen that $\rho_N = (1 - \zeta_1) / (3 - 2\zeta_1)$, and $\rho^*(\hat{y}_N) = 1/3$ (which is obtained by action 2 of P1). Thus, to obtain $\rho_N \geq \rho^*(\hat{y}_N) - \epsilon_0$ we need $\zeta_1 \leq (9\epsilon_0) / (2 + 3\epsilon_0) = O(\epsilon_0)$. Next, at $k = 2N$ we have $y_{2N} = (0.5, 0.5)$ and $\rho^*(\hat{y}_{2N}) = \max\{5/2, 1/6\} = 5/2$, which is now obtained by action 1 of P1. To compute ρ_{2N} , let $\zeta_2 = \sum_{i=N+1}^{2N} 1\{i_k = 1\} / N$ denote the empirical distribution of P1's action 1 over the second N -stage period. Then maximizing ρ_{2N} over $\zeta_2 \in [0, 1]$ by $\zeta_2 = 1$ we get that $\rho_{2N} = (6 - \zeta_1) / (4 -$

$2\zeta_1$). A simple calculation now shows that to obtain $\rho_{2N} \geq \rho^*(\hat{y}_{2N}) - \epsilon_0$ we need $\zeta_1 \geq (2 - 2\epsilon_0)/(3 - 2\epsilon_0)$. It is evident that the requirements are contradictory for ϵ_0 small enough.

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. \square

We close this section with a sufficient condition for *non-existence* of no-regret strategies. This condition essentially follows 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. *Suppose there exist $y_1, y_2 \in Y$ and $\alpha \in (0, 1)$ such that:*

$$\rho^*(\alpha y_1 + (1 - \alpha)y_2) > \max_{x_1 \in X^*(y_1), x_2 \in X} \frac{\alpha r(x_1, y_1) + (1 - \alpha)r(x_2, y_2)}{\alpha \tau(x_1, y_1) + (1 - \alpha)\tau(x_2, y_2)}. \quad (9)$$

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

Proof. The proof of is similar to that of Proposition 2, 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 (9) 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 . \square

4 Approachability and Regret Minimization

The theory of approachability, introduced in [2], is one of the fundamental tools that have been used for obtaining no-regret strategies in repeated matrix games. The analysis in this section will 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. This strategy can be thought of as a first attempt at deriving an adaptive strategy which will be shown to be dominated by the strategy considered in Section 5. Much of the analysis here is similar to our paper [15] and most proofs are therefore deferred to [16].

4.1 The Temporal Best-Response Envelope

Following [3, 15], we attempt to construct a vector-valued payoff vector $\rho_n = (\rho_n, \hat{y}_n)$, so that attaining the best-response $\rho^*(y)$ is equivalent to approaching³ the set

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

However, two obstacles stand in the way of applying the approachability results. 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 the simple condition for convex sets in Blackwell's original result which is exploited in [3, 15] cannot be used.

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 $1/\hat{\tau}_n$. The set of feasible action rates is given by

$$\Pi = \left\{ \frac{y}{\tau} : y \in Y, \tau \in T(y) \right\}, \quad (10)$$

where $T(y)$ is the set of average stage durations τ which are feasible jointly with the empirical distribution y , that is, $T(y) = \left\{ \sum_j y(j) \tau(x^j, j) : x^j \in X \text{ for all } j \right\}$. 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 ([4]).

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 $\pi \in \Pi$,

$$\tilde{\rho}^*(\pi) \triangleq \rho^* \left(\frac{\pi}{|\pi|} \right) = \max_{i \in I} \frac{\sum_j r(i, j) \pi(j)}{\sum_j \tau(i, j) \pi(j)}, \quad (11)$$

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 $\tilde{\rho}^* : \Pi \rightarrow \mathbb{R}$ as the *temporal best-response envelope*.

Convexity of $\tilde{\rho}^*$ turns out to be a sufficient condition for existence of no-regret strategies. The proof is similar to [15] and is therefore omitted.

Theorem 1. *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 1), namely, a strategy that attains the best-response envelope $\rho^*(\hat{y})$.*

³ Formally, in approachability one has to define a vector-valued game and prove that the point-to-set distance between the average vector-valued reward and the target set goes to 0 almost surely. See [15] for an example of such analysis.

4.2 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 2 (Convex best-response envelope). *The convex best-response envelope $\tilde{\rho}^{\text{co}} : \Pi \rightarrow \mathbb{R}$ is defined as the lower convex hull of $\tilde{\rho}^*$ over its domain Π .*

We now have the following result. See [16] for a proof.

Theorem 2 ($\tilde{\rho}^{\text{co}}(\pi)$ is attainable). *The convex best-response envelope $\tilde{\rho}^{\text{co}}(\pi)$ is attainable by P1. Namely, there exists a strategy of P1 so that*

$$\liminf_{n \rightarrow \infty} (\rho_n - \tilde{\rho}^{\text{co}}(\pi_n)) \geq 0 \quad (\text{a.s.}) \quad (12)$$

for any strategy of P2.

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 to Y . For $\hat{y} \in Y$, define

$$\rho^{\text{co}}(\hat{y}) = \min\{\tilde{\rho}^{\text{co}}(\pi) : \pi \in \Pi, \frac{\pi}{|\pi|} = \hat{y}\}. \quad (13)$$

For simplicity we also refer to ρ^{co} as the convex best-response envelope (over Y). The following corollary to Theorem 2 is immediate.

Corollary 1 ($\rho^{\text{co}}(\hat{y})$ is attainable). *The convex best-response envelope $\rho^{\text{co}}(\hat{y})$ is attainable by P1. Namely, there exists a strategy of P1 so that*

$$\liminf_{n \rightarrow \infty} (\rho_n - \rho^{\text{co}}(\hat{y}_n)) \geq 0 \quad (\text{a.s.}) \quad (14)$$

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

Figure 1 illustrates the resulting convex best-response envelope for the game of Example 1. 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.

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

Lemma 3 (Properties of ρ^{co}). *The convex best-response envelope $\rho^{\text{co}}(y)$ satisfies the following properties. For each $y \in Y$,*

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

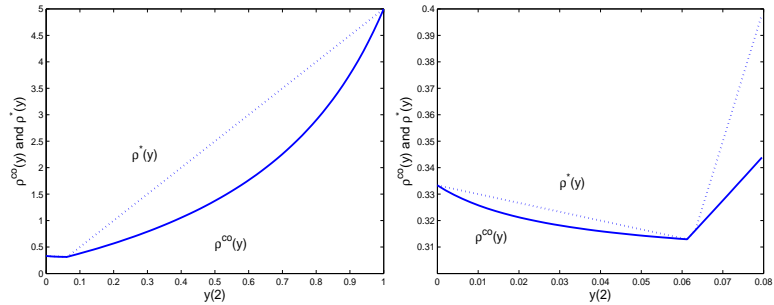


Fig. 1. $\rho^*(y)$ (dotted) and $\rho^{\text{co}}(y)$ (thick line) for the game of Example 1. The right figure zooms on the segment $[0, 0.08]$. Note that $v(r, \tau) = 5/16$ for that game.

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 $\rho^* \geq v(r, \tau)$, the same property is inherited by $\tilde{\rho}^*$, $\tilde{\rho}^{\text{co}}$ and ρ^{co} , again by their respective definitions.

(ii) We will show that $\rho^{\text{co}}(y) = v(r, \tau)$ implies that $\rho^*(y) = v(r, \tau)$. Suppose $\rho^{\text{co}}(y) = v(r, \tau)$. Then there exists some $\pi \in \Pi$ such that $\pi/|\pi| = y$ and $\tilde{\rho}^{\text{co}}(\pi) = v(r, \tau)$. By Caratheodory's Theorem there exist ℓ points π_1, \dots, π_ℓ in Π (where $\ell \leq 2 + |J|$) and coefficients $\alpha_1, \dots, \alpha_\ell > 0$ with $\sum_{m=1}^{\ell} \alpha_m = 1$ such that $\pi = \sum_{m=1}^{\ell} \alpha_m \pi_m$ and $v(r, \tau) = \tilde{\rho}^{\text{co}}(\pi) = \sum_{m=1}^{\ell} \alpha_m \rho^*(\pi_m)$. Since $\rho^*(\pi) \geq v(r, \tau)$, this implies that $\rho^*(\pi_m) = v(r, \tau)$ for all m . Recall now from Lemma 1(ii) that the set Y^* of mixed actions $y \in Y$ for which $\rho^*(y) = v(r, \tau)$ 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 ([4]). Noting that $\pi_m \in \Pi^*$ for all m (which follows from $\rho^*(\pi_m) = v(r, \tau)$) 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 $\rho^{\text{co}}(y) = v(r, \tau)$. \square

Both properties that were stated in the last lemma can be observed in Fig. 1.

5 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, [6] for an overview of the relation between regret minimization and 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 5.2 the *calibration envelope* $\rho^{\text{cal}}(\hat{y})$, and show that it is attained by calibrated play in the sense that $\rho_n \geq \rho^{\text{cal}}(\hat{y}_n)$ holds asymptotically. We then proceed to compare the calibration

envelope with the convex best-response envelope of the previous section, and show that $\rho^{\text{cal}} \geq \rho^{\text{co}}$.

5.1 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 $\mu_k : H_{k-1} \rightarrow \Delta(Y)$, $k \geq 1$, 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 .

We shall use the following definition of calibrated forecasts.

Definition 3 (Calibrated forecasts). *A forecasting scheme is calibrated if for every (Borel measurable) set $Q \subset Y$ and every strategy of P2,*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n 1\{q_k \in Q\} (e_{j_k} - q_k) = 0 \quad \text{a.s.}, \quad (15)$$

where e_j is a vector of zeros with 1 in the j th location.

This form of calibration property has been introduced into game theory by [7], and several algorithms have been devised to achieve it ([8, 11, 13]). These algorithms typically start with predictions that are restricted to a finite grid, gradually increasing the number of grid points (see [5] for such a construction).

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

$$I^*(y) = \arg \max_{i \in I} \frac{r(i, y)}{\tau(i, y)}, \quad y \in Y. \quad (16)$$

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

$$i_k = i^o(q_k) \quad (17)$$

where (q_k) is a calibrated forecast of P2's actions, and $i^o(y) \in I^*(y)$ 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 (5). 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.

5.2 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 $I_0(r, \tau)$. 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 $\hat{y} \in Y$:

$$\rho^{\text{cal}}(\hat{y}) = \inf \left\{ \frac{\int r(i(y), y)\mu(dy)}{\int \tau(i(y), y)\mu(dy)} : \mu \in \Delta_d(Y), m_\mu = \hat{y}, i(y) \in I^*(y) \right\}. \quad (18)$$

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 4.

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

$$\rho^{\text{cal}}(\hat{y}) = \min \left\{ \frac{\sum_{i \in I} \alpha_i r(i, y_i)}{\sum_{i \in I} \alpha_i \tau(i, y_i)} : \alpha \in \Delta(I), y_i \in co(Y_i^*), \sum_{i \in I} \alpha_i y_i = \hat{y} \right\}. \quad (19)$$

(ii) The infimum in (18) is attained by a measure μ of finite support.

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

Proof. (i) Note first that the minimum in (19) 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 (19). To show that $\rho^1 \leq \rho^{\text{cal}}$, 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 $\beta_{ij} \in \Delta(J)$. It follows that for each \hat{y} the argument of (19) can be written as the special case of the argument of (18), from which $\rho^1(\hat{y}) \leq \rho^{\text{cal}}(\hat{y})$ 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 (18) is thus reduced to the form of (19), which implies that $\rho^1(\hat{y}) \leq \rho^{\text{cal}}(\hat{y})$.

(ii) Follows immediately from the indicated reduction of the argument of (19) to that of (18).

(iii) Continuity follows since the minimized function in (19) is continuous in its arguments α and (y_i) , while the minimizing set is upper semi-continuous in y . \square

We next establish that calibrated play attains the calibration envelope.

Theorem 3 (ρ^{cal} is attainable). *Suppose P1 uses a calibrated strategy. Then, for any strategy of P2,*

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

Proof. It will be convenient to use for this proof the shorthand notations $a_n \stackrel{o(n)}{=} b_n$ for $\lim_{n \rightarrow \infty} (a_n - b_n) = 0$, and $a_n \stackrel{o(n)}{\geq} b_n$ for $\liminf_{n \rightarrow \infty} (a_n - b_n) \geq 0$. 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

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

The second ($o(n)$) equality follows from (15). Repeating the argument for τ we obtain

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

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

$$\rho_n \stackrel{o(n)}{=} \frac{\sum_{k=1}^n r(i^o(q_k), q_k)}{\sum_{k=1}^n \tau(i^o(q_k), q_k)}, \quad (20)$$

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

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

Thus, $\rho_n \stackrel{o(n)}{\geq} \rho^{\text{cal}}(\hat{q}_n)$. Note also that from (15), with $Q = Y$, we have $\hat{y}_n \stackrel{o(n)}{=} \hat{q}_n$. The required equality now follows by continuity for $\rho^{\text{cal}}(y)$ in y , as noted in Lemma 4. \square

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 2. *Suppose that $\rho^{\text{cal}}(y) = \rho^*(y)$ 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 4 (Properties of ρ^{cal}).

- (a) $v(r, \tau) \leq \rho^{\text{cal}}(\hat{y}) \leq \rho^*(\hat{y})$ for all $\hat{y} \in Y$.
- (b) $\rho^{\text{cal}}(\hat{y}) = \rho^*(\hat{y})$ at the extreme points of Y , which correspond to the pure action set I .
- (c) For each $\hat{y} \in Y$, $\rho^*(\hat{y}) > v(r, \tau)$ implies $\rho^{\text{cal}}(\hat{y}) > v(r, \tau)$.

Proof. The proof is technical and appears in [16].

5.3 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 4.2) 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 5 (ρ^{cal} dominates ρ^{co}).

- (i) $\rho^{\text{cal}}(\hat{y}) \geq \rho^{\text{co}}(\hat{y})$ 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 $\rho^{\text{cal}} = \rho^{\text{co}}$.

The proof is omitted; see [16].

Example 2. (ρ^{cal} strictly dominates ρ^{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 $(r(i, j), \tau(i, j))$. A plot of $\rho^{\text{cal}} = \rho^*$ and ρ^{co} for the last example is shown in Figure 2. A detailed account of the computation can be found in [16].

6 Directions for Future Work

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 an *optimal* performance envelope exists in general. 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 setup in repeated matrix games ([1]).

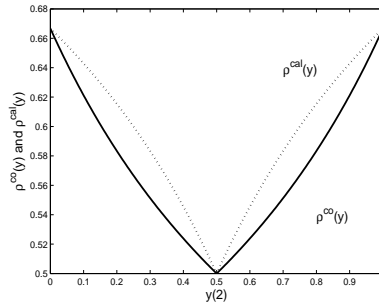


Fig. 2. $\rho^{\text{cal}}(y)$ (dotted line) and $\rho^{\text{co}}(y)$ (thick line) for the game of Example 2.

References

1. P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.
2. D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific J. Math.*, 6(1):1–8, 1956.
3. D. Blackwell. Controlled random walks. In *Proc. Int. Congress of Mathematicians 1954*, volume 3, pages 336–338. North Holland, Amsterdam, 1956.
4. S. Boyd and L. Vanderberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.
5. N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, 2006.
6. D. P. Foster and R. Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29:7–35, November 1999.
7. D. P. Foster and R. V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21:40–55, 1997.
8. D. P. Foster and R. V. Vohra. Asymptotic calibration. *Biometrika*, 85:379–390, 1998.
9. Y. Freund and R. E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29:79–103, 1999.
10. D. Fudenberg and D. Levine. Universal consistency and cautious fictitious play. *Journal of Economic Dynamic and Control*, 19:1065–1090, 1995.
11. D. Fudenberg and D. Levine. An easier way to calibrate. *Games and Economic Behavior*, 29:131–137, 1999.
12. J. Hannan. *Approximation to Bayes Risk in Repeated Play*, volume III of *Contribution to The Theory of Games*, pages 97–139. Princeton University Press, 1957.
13. S. Kakade and D. P. Foster. Deterministic calibration and Nash equilibrium. In *COLT*, volume 3120 of *Lecture Notes in Computer Science*, pages 33–48. Springer, 2004.
14. A. A. Lal and S. Sinha. Zero-sum two-person semi-Markov games. *J. Appl. Prob.*, 29:56–72–8, 1992.
15. S. Mannor and N. Shimkin. The empirical Bayes envelope and regret minimization in competitive Markov decision processes. *Mathematics of Operations Research*, 28(2):327–345, 2003.
16. S. Mannor and N. Shimkin. Regret minimization in repeated matrix games with variable stage duration. Technical Report EE-1524, Faculty of Electrical Engineering, Technion, February 2006.