

# Online Learning with Variable Stage Duration

Shie Mannor

Nahum Shimkin

*McGill University*

*Technion*

June 2006

# Outline

1. Model formulation
2. Best response envelope
3. Approachability and regret minimization
4. Calibrated play
5. Some challenges

## Part I: Model formulation

Online learning does not allow temporal variation between actions

Real world problems often have variation in duration (e.g., traversing different routes in a network, buying different options in a market).

Two players:

P1: reward maximizing (regret minimizing) and

P2: arbitrary.

## Model formulation II

$I$  and  $J$  denote the finite action sets of P1 and P2, respectively.

The stage game:

a reward function  $r : I \times J \rightarrow \mathbb{R}$ ,

and a **duration** function  $\tau : I \times J \rightarrow (0, \infty)$ .

At the beginning of each stage  $k$  (where  $k = 1, 2, \dots$ ):

P1 chooses  $i_k$  and P2 chooses  $j_k$ .

P1 obtains a reward  $r_k = r(i_k, j_k)$ , and the current stage proceeds for  $\tau_k = \tau(i_k, j_k)$ .

## Model formulation III

The average reward **per unit time** over the first  $n$  stages of play is thus given by

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

It may be easier to think about it through per-stage averages:

$$\hat{r}_n = \frac{1}{n} \sum_{k=1}^n r_k \quad \text{average stage reward}$$

$$\hat{\tau}_n = \frac{1}{n} \sum_{k=1}^n \tau_k \quad \text{average stage duration}$$

so that  $\rho_n = \hat{r}_n / \hat{\tau}_n$ .

We will consider the game from the viewpoint of P1, who seeks to maximize his long-term reward rate.

## Part II: Best response envelope

An element  $x \in \Delta(I)$  is a *mixed action* of P1, and similarly  $y \in \Delta(J)$  is a mixed action of P2. We shall use the bilinear extension of  $r$  and  $\tau$  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  $\tau$ .

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

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

Let  $\hat{y}_n(j) = \frac{1}{n} \sum_{k=1}^n \mathbf{1}\{j_k = j\}$  denote the empirical frequency until stage  $n$ .

## Best response envelope II

Objective: Attain the best response in the sense:

Best response is:

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

Attaining here means:

$$\liminf_{n \rightarrow \infty} (\rho_n - \rho^*(\hat{y}_n)) \geq 0 \quad \text{almost surely}$$

## Best response envelope III

A very simple example shows that the best-response is **not** attainable:

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

P2's strategy is to first play left action (for  $N$  stages) and the right action (for another  $N$  stages). We then show that:

1. Either  $\rho^*(y_N) - \rho_N > \epsilon$  or
2.  $\rho^*(y_{2N}) - \rho_{2N} > \epsilon$ .

We therefore set out to look for less ambitious goals.

## Part III: Approachability

It is possible to define an envelope via approachability a-la-Blackwell.

This envelope can be attained using a geometric policy in an appropriately defined space where a further convex relaxation is needed.

Reminiscent to our “regret minimization in stochastic games paper” (COLT 01’ and Mathematics of Operations Research 03’).

Envelope does better than worst-case (assuming P2 is adversarial) if possible.

For more details, read the paper ...

## Part IV: Calibrated play

Idea: if P1 can predict the (mixed) action of P2 accurately, P1 can play a best response against P2's predicted play.

**Definition:** A forecasting scheme is *calibrated* if for every (Borel measurable) set  $Q \subset \Delta(J)$  and every strategy of P2,

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

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

Intuition: Whenever  $q \in Q$  is predicted, either:

1. Prediction balances out on  $Q$ .
2. Forecasting  $Q$  does not happen too often (so the  $1/n$  kills it).

## Calibrated play II

Suggested policy:

Play the best response against the forecast.

If forecast is  $q_k$ , play  $i_k = \arg \max \rho(i, q_k)$ .

Side issue: Can we calibrate?

In principle: Yes.

Algorithmically: Maybe (Foster and Kakade). A pretty heavy algorithm.

In practice: Can calibrate if alphabet is binary or if the source is smooth in some sense. In that case the calibration can be done easily using weighted average of past observations (Mannor et al. MLJ, accepted).

## Calibrated play III

What does calibrated play get us?

We let  $I^*(y)$  denote the best response against  $y$  ( $\arg \max \rho(i, y)$ ).

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

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

**Theorem:** Calibrated play attains  $\rho^{\text{cal}}(\hat{y}_n)$ .

## Calibrated play IV

Some properties of  $\rho^{\text{cal}}$ :

1. It is never less than what approachability gets.
2. It is sometimes more than what approachability gets.
3. It equals  $\rho^*$  for pure actions.
4. It is less than or equal to  $\rho^*$ .
5. If P2 alone controls the duration then  $\rho^{\text{cal}} = \rho^*$  and is therefore optimal.
6. It is higher than an appropriately defined value of a zero-sum game: you never lose comparing to a very defensive (worst-case) strategy.

## Part V: Challenges

The model presented is simple and clean (much cleaner than stochastic games).

Still, the best response is not attainable.

Challenges:

1. What is the best envelope we can attain using *any* policy?
2. We required pretty heavy guns (calibration) to get  $\rho^{\text{cal}}$ . Can we get the same things with something simpler?
3. What would FPL/multilicative experts based algorithms do?