

# Online Learning with Constraints

Shie Mannor

John N. Tsitsiklis

*McGill University*

*MIT*

June 2006

# Outline

1. Model formulation
2. Reward-in-hindsight
3. The convex relaxation
4. Calibrated play
5. Conclusion

## Part I: Model formulation

Online learning models ignore constraints

Real world problems often call for considering constraints:

Example: An antenna sending information over a wireless link. The higher the power the more probable a successful transmission (but it depends on other factors such as the humidity, other antennas, etc.).

**Primary objective:** Maximize transmission rate.

**Secondary objective:** Have the average power consumption below some FCC threshold.

Two players:

P1: reward maximizing (regret minimizing) that tries to keep the constraints satisfied

P2: arbitrary.

## Model formulation II

$A$  and  $B$  denote the finite action sets of P1 and P2, respectively.

The stage game:

a reward function  $r : A \times B \rightarrow \mathbb{R}$ ,

and a penalty function (secondary objective)  $c : A \times B \rightarrow \mathbb{R}^d$ .

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

P1 chooses  $a_t$  and P2 chooses  $b_t$ .

P1 obtains a reward  $r_t = r(a_t, b_t)$ , and an immediate penalty of  $c_t = c(a_t, b_t)$ .

## Model formulation III

We define P1's average reward by time  $t$  to be

$$\widehat{r}_t = \frac{1}{t} \sum_{\tau=1}^t r_{\tau},$$

and P1's average penalty vector by time  $t$  to be

$$\widehat{c}_t = \frac{1}{t} \sum_{\tau=1}^t c_{\tau}.$$

P1 also has a desired convex target set  $T \subseteq \mathbb{R}^d$ . The average penalty should (hopefully and asymptotically) be in  $T$ .

## Part II: Reward-in-hindsight

Suppose P1 knew in advance P2's empirical play. Let the empirical frequency of P2's actions be:

$$\hat{q}_t(b) = \frac{1}{t} \sum_{\tau=1}^t \mathbf{1}_{\{b_\tau=b\}}, \quad b \in B.$$

Or equivalently, suppose P2 plays a stationary strategy  $q$ .

Then P1 would solve the convex program

$$\begin{aligned} r^*(q) &\triangleq \max_{p \in \Delta(A)} \sum_{a,b} p(a)q(b)R(a,b), && \text{Maximize reward} \\ &\text{s.t. } \sum_{a,b} p(a)q(b)C(a,b) \in T. && \text{Satisfy constraints} \end{aligned}$$

We call  $r^*$  the **reward-in-hindsight** w.r.t.  $T$ .

## Reward-in-hindsight II

Objective: Attain the same performance without knowing  $q$  in advance in the online sense while not violating the constraints infinitely often.

More formally:

A function  $r : \Delta(B) \mapsto \mathbb{R}$  is **attainable** by P1 in a constrained game **with respect to a set  $T$**  if there exists a strategy  $\sigma$  of P1 such that for every strategy  $\rho$  of P2:

$$(i) \liminf_{t \rightarrow \infty} (\hat{r}_t - r(\hat{q}_t)) \geq 0, \quad \text{a.s., and}$$

$$(ii) \limsup_{t \rightarrow \infty} \text{dist}(\hat{c}_t, T) \rightarrow 0, \quad \text{a.s.,}$$

dist here is the Euclidean distance, but any metric is OK since  $d$  is finite.

## Reward-in-hindsight III

**Assumption:** For every mixed action  $q \in \Delta(B)$  of P2, there exists a mixed action  $p \in \Delta(A)$  of P1, such that:

$$\sum_{a,b} p(a)q(b)C(a,b) \in T.$$

Without this assumption P2 can guarantee that the constraint is not satisfied by repeatedly playing the  $q$  violating the assumption.

## Some game theory connections

Consider the infinitely repeated constrained game between two players, P1 and P2 (with the same reward and penalty function structure).

P1's goal is to maximize his average reward while keeping the average penalty in  $T$ .

P2's goal is to keep the average penalty away from  $T$  (preferred) and minimize the reward (if cannot take the average penalty out of  $T$ ).

Define

$$v = \inf_{q \in \Delta(B)} \sup_{p \in \Delta(A), \sum_{a,b} p(a)q(b)C(a,b) \in T} \sum_{a,b} p(a)q(b)R(a,b).$$

This is the **value** of the constrained game.

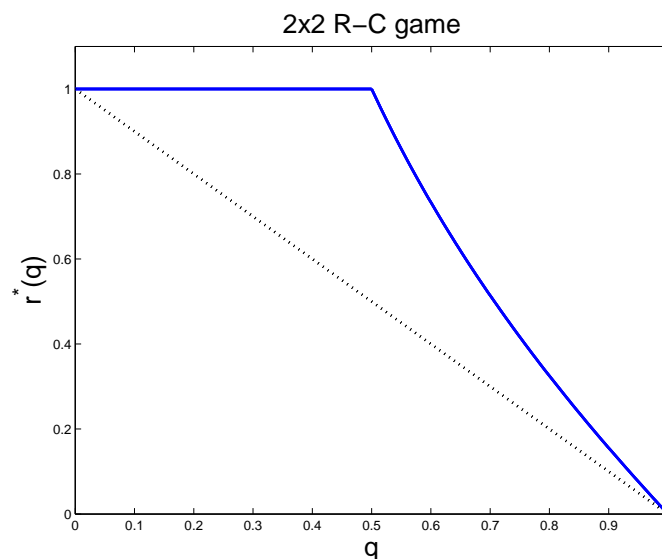
**Theorem:**  $v$  is the highest average reward P1 can guarantee while keeping the average penalty in  $T$  against any opponent.

## Reward-in-hindsight is too much to ask

Consider a  $2 \times 2$  constrained game ( $d = 1$ ) specified by with  $T = [-\infty, 0]$ : (each entry is a pair  $(r, c)$ )

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

It turns out that:



## Reward-in-hindsight is too much to ask II

P2 has a strategy that forces P1 to have an average reward less than  $r^*(q)$  or average penalty above 0.

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

1. In the first  $N$  stages P1 must choose the lower and upper actions with probability  $1/2$ , getting an average reward of 0 and an average penalty of 0.

2. In the second  $N$  stages P2 can choose the point he likes on the dotted curve, but the best thing for him is to play the upper action, leading to  $\hat{r}_{2N} = 1/2$ ,  $\hat{c}_{2N} = 1/2$ .

It follows that  $\hat{q}_{2N} = 1/2$ , but  $r^*(1/2) = 1$ .

## Part III: The convex relaxation

We consider a less ambitious goal - the convex relaxation. Namely:

$$\begin{aligned} r^c(q) &= \inf_{q_1, q_2, \dots, q_k \in \Delta(B), \alpha_1, \dots, \alpha_k} \sum_{i=1}^k \alpha_i r^*(q_i) \\ \text{s.t. } &\sum_{i=1}^k \alpha_i q_i(b) = q(b), \quad b \in B, \\ &\alpha_i \geq 0, \quad i = 1, 2, \dots, k, \\ &\sum_{i=1}^k \alpha_i = 1, \end{aligned}$$

**Theorem:** If the Assumption holds  $r^c(q)$  is attainable w.r.t.  $T$ .

**Proof:** A standard application of approachability and continuity arguments.

## The convex relaxation II

It follows that  $r^c(q) \geq v$  so that attaining the convex relaxation is always better (not worse) than playing under the assumption that P2 is adversarial.

Degenerate cases:

If either player affects the penalty function alone:

1.  $C(a, b) = C(a, b')$  for every  $a, b, b'$  or
2.  $C(a, b) = C(a', b)$  for every  $a, a', b$

Then  $r^* = r^c$ .

## The convex relaxation III

For the case  $d = 1$  and  $T = \{c \mid c \leq c_0\}$  we have a tightness result. That is:

**Theorem** Suppose that  $d = 1$ ,  $T$  is of the form  $T = \{c \mid c \leq c_0\}$ , where  $c_0$  is a given scalar, and that the Assumption is satisfied. Let  $\tilde{r}(q) : \Delta(B) \mapsto \mathbb{R}$  be an attainable continuous function with respect to  $T$ . Then,  $r^c(q) \geq \tilde{r}(q)$  for all  $q \in \Delta(B)$ .

The proof is based on geometric arguments tailored for 2 dimensions (reward + penalty).

This is the first tightness result of an envelope (other than the Bayes envelope) we are aware of.

## 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 that keeps the constraint satisfied.

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

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

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.
2. Forecasting  $Q$  does not happen often enough (so the  $1/n$  kills it).

## Calibrated play II

Suggested policy:

Play the best response against the forecast that keeps the constraint satisfied. If forecast is  $q_t$ , play  $p_t = p^*(q_t)$  where:

$$\begin{aligned} p^*(q) &= \arg \max_{p \in \Delta(A)} \sum_{a,b} p(a)q(b)R(a,b) \\ &\text{s.t. } \sum_{a,b} p(a)q(b)C(a,b) \in T. \end{aligned}$$

**Theorem:** Under the Assumption calibrated play attains  $r^c$  w.r.t.  $T$ .

## Part V: Conclusion

The model presented is a variation of “standard” online learning models. It attempts to capture side constraints.

The best response is not attainable and instead a relaxed goal is needed.

Challenges:

1. What is the best envelope we can attain using *any* policy for  $d > 1$ ?
2. Is there a “simple” way to attain  $r^c$  (without calibration)?
3. What would FPL/multiplicative experts based algorithm do? How to modify them to take the constraints under consideration?