Online learning with sample path constraints

Journal of Machine Learning Research 0 (2008) 0-0 Submitted 07/08; Published 00/00

Online Learning with Sample Path Constraints

Shie Mannor SHIE.MANNOR@MCGILL.CA
Department of Electrical and Computer Engineering
McGill University, Québec H3A-2A7

John N. Tsitsiklis JNT@MIT.EDU
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology, Cambridge, MA 02139

Jia Yuan Yu JIA.YU@MCGILL.CA
Department of Electrical and Computer Engineering
McGill University, Québec H3A-2A7

Editor: Gábor Lugosi

Abstract

We study online learning where a decision maker interacts with Nature with the objective of maximizing her long-term average reward subject to some sample path average constraints. We define the reward-in-hindsight as the highest reward the decision maker could have achieved, while satisfying the constraints, had she known Nature's choices in advance. We show that in general the reward-in-hindsight is *not* attainable. The convex hull of the reward-in-hindsight function is, however, attainable. For the important case of a single constraint, the convex hull turns out to be the highest attainable function. Using a calibrated forecasting rule, we provide an explicit strategy that attains this convex hull. We also measure the performance of heuristic methods based on non-calibrated forecasters in experiments involving a CPU power management problem.

1. Introduction

We consider a repeated game from the viewpoint of a decision maker (player P1) who plays against Nature (player P2). The opponent (Nature) is "arbitrary" in the sense that player P1 has no prediction, statistical or strategic, of the opponent's choice of actions. This setting was considered by Hannan (1957), in the context of repeated matrix games. Hannan introduced the Bayes utility with respect to the current empirical distribution of the opponent's actions, as a performance goal for adaptive play. This quantity, defined as the highest average reward that player P1 could have achieved, in hindsight, by playing some fixed action against the observed action sequence of player P2. Player P1's *regret* is defined as the difference between the highest average reward-in-hindsight that player P1 could have hypothetically achieved, and the actual average reward obtained by player P1. It was established in Hannan (1957) that there exist strategies whose regret converges to zero as the number of stages increases, even in the absence of any prior knowledge on the strategy of player P2. For recent advances on online learning, see Cesa-Bianchi and Lugosi (2006).

In this paper we consider regret minimization under sample-path constraints. That is, in addition to maximizing the average reward, or more precisely, minimizing the regret, the decision maker has some side constraints that need to be satisfied on the average. In particular, for every joint action of the players, there is an additional penalty vector that is accumulated by the decision maker. The decision maker has a predefined set in the space of penalty vectors, which represents the acceptable tradeoffs between the different components of the penalty vector. An important special case arises when the decision maker wishes to keep some constrained resource below a certain threshold. Consider, for example, a wireless communication system where the decision maker can adjust the transmission power to improve the probability that a message is received successfully. Of course, the decision maker does not know a priori how much power will be needed (this depends on the behavior of other users, the channel conditions, etc.). Still, a decision maker is usually interested in both the rate of successful transmissions, and in the average power consumption. In an often considered variation of this problem, the decision maker wishes to maximize the transmission rate, while keeping the average power consumption below some predefined threshold. We refer the reader to Mannor and Shimkin (2004) and references therein for a discussion of constrained average cost stochastic games and to Altman (1999) for constrained Markov decision problems. We note that the reward and the penalty are not treated the same; otherwise they could have been combined into a single scalar value, resulting in a much simpler problem.

The paper is organized as follows. In Section 2, we present formally the basic model, and provide a result that relates attainability with the value of the game. In Section 3, we provide an example where the reward-in-hindsight cannot be attained. In light of this negative result, in Section 4 we define the closed convex hull of the reward-in-hindsight, and show that it is attainable. Furthermore, in Section 5, we show that when there is a single constraint, this is the maximal attainable objective. In Section 6, we provide a simple strategy, based on calibrated forecasting, that attains the closed convex hull. Section 7 presents heuristic algorithms derived from an online forecaster, while incorporating strictly enforced constraints. The application of the algorithms of Section 7 to a power management domain is presented in Section 8. We finally conclude in Section 9 with some open questions and directions for future research.

2. Problem definition

We consider a repeated game against Nature, in which a decision maker tries to maximize her reward, while satisfying some constraints on certain time-averages. The underlying stage game is a game with two players: P1 (the decision maker of interest) and P2 (who represents Nature and is assumed arbitrary). For our purposes, we only need to define rewards and constraints for P1.

A constrained game with respect to a set T is defined by a tuple (A, B, R, C, T) where:

1. A is the set of actions of P1; we will assume $A = \{1, 2, \ldots, |A|\}$.
2. B is the set of actions of P2; we will assume $B = \{1, 2, \ldots, |B|\}$.
3. R is an $|A| \times |B|$ matrix where the entry $R(a, b)$ denotes the expected reward obtained by P1, when P1 plays action $a \in A$ and P2 action $b \in B$. The actual rewards obtained

at each play of actions a and b are assumed to be IID random variables, with finite second moments, distributed according to a probability law $\Pr_R(\cdot \,|\, a, b)$. Furthermore, the reward streams for different pairs (a, b) are statistically independent.

4. C is an $|A| \times |B|$ matrix, where the entry $C(a, b)$ denotes the expected d-dimensional penalty vector incurred by P1, when P1 plays action $a \in A$ and P2 action $b \in B$. The actual penalty vectors obtained at each play of actions a and b are assumed to be IID random variables, with finite second moments, distributed according to a probability law $\Pr_C(\cdot \,|\, a, b)$. Furthermore, the penalty vector streams for different pairs (a, b) are statistically independent.

5. T is a set in $\mathbb{R}^d$ within which we wish the average of the penalty vectors to lie. We assume that T is convex and closed. Since the entries of C are bounded, we will also assume, without loss of generality, that T is bounded.

The game is played in stages. At each stage t, P1 and P2 simultaneously choose actions $a_t \in A$ and $b_t \in B$, respectively. Player P1 obtains a reward r_t, distributed according to $\Pr_R(\cdot \,|\, a_t, b_t)$, and a penalty c_t, distributed according to $\Pr_C(\cdot \,|\, a_t, b_t)$. We define P1's average reward by time t to be

$$\hat{r}_t = \frac{1}{t} \sum_{\tau=1}^{t} r_\tau, \tag{2.1}$$

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

$$\hat{c}_t = \frac{1}{t} \sum_{\tau=1}^{t} c_\tau. \tag{2.2}$$

A *strategy* for P1 (resp. P2) is a mapping from the set of all possible past histories to the set of mixed actions on A (resp. B), which prescribes the (mixed) action of that player at each time t, as a function of the history in the first $t-1$ stages. Loosely, P1's goal is to maximize the average reward while having the average penalty vector converge to T, pathwise:

$$\limsup_{t \to \infty} \operatorname{dist}(\hat{c}_t, T) \to 0, \qquad \text{a.s.}, \tag{2.3}$$

where $\operatorname{dist}(\cdot)$ is the point-to-set Euclidean distance, i.e., $\operatorname{dist}(x, T) = \inf_{y \in T} \|y - x\|_2$, and the probability measure is the one induced by the policy of P1, the policy of P2, and the randomness in the rewards and penalties.

We will often consider the important special case where $T = \{c \in \mathbb{R}^d : c \leq c_0\}$, for some given $c_0 \in \mathbb{R}^d$, with the inequality interpreted component-wise. We simply call such a game a constrained game with respect to (a vector) c_0. For that special case, the requirement (2.3) is equivalent to:

$$\limsup_{t \to \infty} \hat{c}_t \leq c_0, \quad \text{a.s.}.$$

For a set D, we will use the notation $\Delta(D)$ to denote the set of all probability measures on D. If D is finite, we will identify $\Delta(D)$ with the set of probability vectors of the same size as D. If D is a subset of Euclidean space, we will assume that it is endowed with the Borel σ-field.

2.1 Reward-in-hindsight

We define $\hat{q}_t \in \Delta(B)$ as the empirical distribution of P2's actions by time t, that is,

$$\hat{q}_t(b) = \frac{1}{t}\sum_{\tau=1}^{t} 1_{\{b_t=b\}}, \quad b \in B. \tag{2.4}$$

If P1 knew in advance that $\hat{q}_t$ will equal q, and if P1 were restricted to using a fixed action, then P1 would pick an optimal response (generally a mixed action) to the mixed action q, subject to the constraints specified by T. In particular, P1 would solve the convex program[1]

$$\begin{aligned} \max_{p\in\Delta(A)} \quad & \sum_{a,b} p(a)q(b)R(a,b), \\ \text{s.t.} \quad & \sum_{a,b} p(a)q(b)C(a,b) \in T. \end{aligned} \tag{2.5}$$

By playing a p that solves this convex program, P1 would meet the constraints (up to small fluctuations that are a result of the randomness and the finiteness of t), and would obtain the maximal average reward. We are thus led to define P1's reward-in-hindsight, which we denote by $r^* : \Delta(B) \mapsto \mathbb{R}$, as the optimal objective value in the program (2.5), as a function of q. The function r^* is often referred to as the *Bayes envelope*.

For the special case of a constrained game with respect to a vector c_0, the convex constraint $\sum_{a,b} p(a)q(b)C(a,b) \in T$ is replaced by $\sum_{a,b} p(a)q(b)C(a,b) \leq c_0$ (the inequality is to be interpreted component-wise).

The following examples show some of the properties of the Bayes envelope. Consider a 2×2 constrained game with respect to a scalar c_0 specified by:

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

where each entry (pair) corresponds to $(R(a,b), C(a,b))$ for a pair of actions a and b. (Here a and b correspond to a choice of row and column, respectively.) Suppose first that $c_0 = 1$. In that case the constraint does not play a part in the problem, and we are dealing with a version of the matching pennies game. So, if we identify q with the frequency of the first action, we have that $r^*(q) = \max(q, 1-q)$. Suppose now that $c_0 = 1/2$. In this case, it is not difficult to show that $r^*(q) = 1/2$, since P1 cannot take advantage of any deviation from $q = 1/2$ while satisfying the constraint.

The next example involves a game where P2's action does not affect the constraints; such games are further discussed in Section 4.1. Consider a 2×2 constrained game with respect to a scalar c_0, specified by:

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

1. If T is a polyhedron (specified by finitely many linear inequalities), then the optimization problem is a linear program.

where each entry (pair) corresponds to $(R(a,b), C(a,b))$ for a pair of actions a and b. We identify q with the frequency of the second action of P2 as before. Suppose first that $c_0 = 1$. As before, the constraint has no effect and $r^*(q) = \max(q, 1-q)$. Suppose now that $c_0 = 1/2$. It is not hard to show that in this case $r^*(q) = \max(q, 1/2)$. Finally, if $c_0 = 0$, P1 is forced to choose the second action; in this case, $r^*(q) = q$. The monotonicity of $r^*(q)$ in c_0 is to be expected since the lower c_0 is, the more stringent the constraint in Eq. (2.5).

2.2 The Objective

Formally, our goal is to attain a function r in the sense of the following definition. Naturally, the higher the function r, the better.

Definition 1 *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 σ of P1 such that for every strategy ρ of P2:*

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

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

where the almost sure convergence is with respect to the probability measure induced by σ and ρ.

In constrained games with respect to a vector c_0 we can replace (ii) in the definition with

$$\limsup_{t\to\infty} \hat{c}_t \leq c_0, \quad \text{a.s.}$$

2.3 The value of the game

In this section, we consider the attainability of a constant function $r : \Delta(B) \mapsto \mathbb{R}$, i.e., $r(q) = \alpha$, for all q. We will establish that attainability is equivalent to having $\alpha \leq \mathrm{v}$, where v is a naturally defined "value of the constrained game."

We first introduce the assumption that P1 is always able to satisfy the constraint.

Assumption 1 *For every mixed action $q \in \Delta(B)$ of P2, there exists a mixed action $p \in \Delta(A)$ of P1, such that:*

$$\sum_{a,b} p(a)q(b)C(a,b) \in T. \tag{2.6}$$

For constrained games with respect to a vector c_0, the condition (2.6) reduces to the inequality $\sum_{a,b} p(a)q(b)C(a,b) \leq c_0$.

If Assumption 1 is not satisfied, then P2 can choose a q such that for every (mixed) action of P1, the constraint is violated in expectation. By repeatedly playing this q, P1's average penalty vector will be outside T, and the objectives of P1 will be impossible to meet.

The following result deals with the attainability of the value, v, of an average reward repeated constrained game, defined by

$$\mathrm{v} = \inf_{q\in\Delta(B)} \sup_{p\in\Delta(A):\sum_{a,b} p(a)q(b)C(a,b)\in T} \sum_{a,b} p(a)q(b)R(a,b). \tag{2.7}$$

The existence of a strategy for P1 that attains the value was proven in (Shimkin, 1994) in the broader context of stochastic games.

Proposition 2 *Suppose that Assumption 1 holds. Then,*

(i) P1 has a strategy that guarantees that the constant function $r(q) \equiv \mathrm{v}$ is attained with respect to T.

(ii) For every number $\mathrm{v}' > \mathrm{v}$ there exists $\delta > 0$ such that P2 has a strategy that guarantees that either $\liminf_{t\to\infty} \hat{r}_t < \mathrm{v}' - \delta$ or $\limsup_{t\to\infty} \mathrm{dist}(\hat{c}_t, T) > \delta$, almost surely. (In particular, the constant function v' is not attainable.)

Proof The proof relies on Blackwell's approachability theory (Blackwell, 1956a). We construct a nested family of convex sets in $\mathbb{R}^{d+1}$ defined by $S_\alpha = \{(r, c) \in \mathbb{R}\times\mathbb{R}^d : r \geq \alpha, c \in T\}$. Obviously, $S_\alpha \subset S_\beta$ for $\alpha > \beta$. Consider the vector-valued game in $\mathbb{R}^{d+1}$ associated with the constrained game. In this game, P1's vector-valued payoff at time t is the $d+1$ dimensional vector $m_t = (r_t, c_t)$ and P1's average vector-valued payoff is $\hat{m}_t = (\hat{r}_t, \hat{c}_t)$. Since S_α is convex, it follows from approachability theory for convex sets (Blackwell, 1956a) that each S_α is either approachable[2] or excludable[3]. If S_α is approachable, then S_β is approachable for every $\beta < \alpha$. We define $\mathrm{v}_0 = \sup\{\beta \mid S_\beta \text{ is approachable}\}$. It follows that S_{v_0} is approachable (as the limit of approachable sets; see Spinat (2002)). By Blackwell's theorem, for every $q \in \Delta(B)$, an approachable convex set must intersect the set of feasible payoff vectors when P2 plays q. Using this fact, it is easily shown that v_0 equals v, as defined by Eq. (2.7), and part (i) follows. Part (ii) follows because a convex set which is not approachable is excludable. ■

Note that part (ii) of the proposition implies that, essentially, v is the highest average reward P1 can attain while satisfying the constraints, if P2 plays an adversarial strategy. By comparing Eq. (2.7) with Eq. (2.5), we see that $\mathrm{v} = \inf_q r^*(q)$. On the other hand, if P2 does not play adversarially, P1 may be able to do better, perhaps attaining $r^*(q)$. Our subsequent results address the question whether this is indeed the case.

Remark 3 *In general, the infimum and supremum in (2.7)* cannot *be interchanged. This is because the set of feasible p in the inner maximization depends on the value of q. Moreover, it can be shown that the set of (p, q) pairs that satisfy the constraint $\sum_{a,b} p(a)q(b)C(a, b) \in T$ is not necessarily convex.*

2. A set X is approachable if there exists a strategy for the agent such that for every $\epsilon > 0$, there exists an integer N such that, for every opponent strategy:

$$\Pr\left(\mathrm{dist}\left(\frac{1}{n}\sum_{i=1}^{n} m_t, X\right) \geq \epsilon \text{ for some } n \geq N\right) < \epsilon.$$

3. A set X is excludable if there exists a strategy for the opponent such that there exists $\delta > 0$ such that for every $\epsilon > 0$, there exists an integer N such that, for every agent strategy:

$$\Pr\left(\mathrm{dist}\left(\frac{1}{n}\sum_{i=1}^{n} m_t, X\right) \geq \delta \text{ for all } n \geq N\right) > 1 - \epsilon.$$

2.4 Related works

Notwithstanding the apparent similarity, the problem that we consider is not an instance of online convex optimization (Zinkevich, 2003; Hazan and Megiddo, 2007). In the latter setting, there is a convex feasible domain $\mathcal{F} \subset \mathbb{R}^n$, and an arbitrary sequence of convex functions $f_t : \mathcal{F} \to \mathbb{R}$. At every step t, the decision maker picks $x_t \in \mathcal{F}$ based on the past history, without knowledge of the future functions f_t, and with the objective of minimizing the regret

$$\sum_{t=1}^{T} f_t(x_t) - \min_{y \in \mathcal{F}} \sum_{t=1}^{T} f_t(y). \tag{2.8}$$

An analogy with our setting might be possible, by identifying x_t and f_t with a_t and b_t, respectively, and by somehow relating the feasibility constraints described by $\mathcal{F}$ to our constraints. However, this attempt seems to run into some fundamental obstacles. In particular, in our setting, feasibility is affected by the opponent's actions, whereas in online convex optimization, the feasible domain $\mathcal{F}$ is fixed for all time steps. For this reason, we do not see a way to reduce the problem of online learning with constraints to an online convex optimization problem, and given the results below, it is unlikely that such a reduction is possible.

3. Reward-in-Hindsight Is Not Attainable

As it turns out, the reward-in-hindsight cannot be attained in general. This is demonstrated by the following simple 2×2 matrix game, with just a single constraint.

Consider a 2×2 constrained game specified by:

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

where each entry (pair) corresponds to $(R(a,b), C(a,b))$ for a pair of actions a and b. At a typical stage, P1 chooses a row, and P2 chooses a column. We set $c_0 = 0$. Let q denote the frequency with which P2 chooses the second column. The reward of the first row dominates the reward of the second one, so if the constraint can be satisfied, P1 would prefer to choose the first row. This can be done as long as $0 \le q \le 1/2$, in which case $r^*(q) = 1$. For $1/2 \le q \le 1$, player P1 needs to optimize the reward subject to the constraint. Given a specific q, P1 will try to choose a mixed action that satisfies the constraint (on the average) while maximizing the reward. If we let α denote the frequency of choosing the first row, we see that the reward and penalty are:

$$r(\alpha, q) = \alpha - (1-\alpha)q\,, \quad c(\alpha, q) = 2\alpha q - 1,$$

respectively. We observe that for every q, $r(\alpha)$ and $c(\alpha)$ are monotonically increasing functions of α. As a result, P1 will choose the maximal α that satisfies $c(\alpha) \le 0$, which is $\alpha(q) = 1/2q$, and the optimal reward is $1/2 + 1/2q - q$. We conclude that the reward-in-

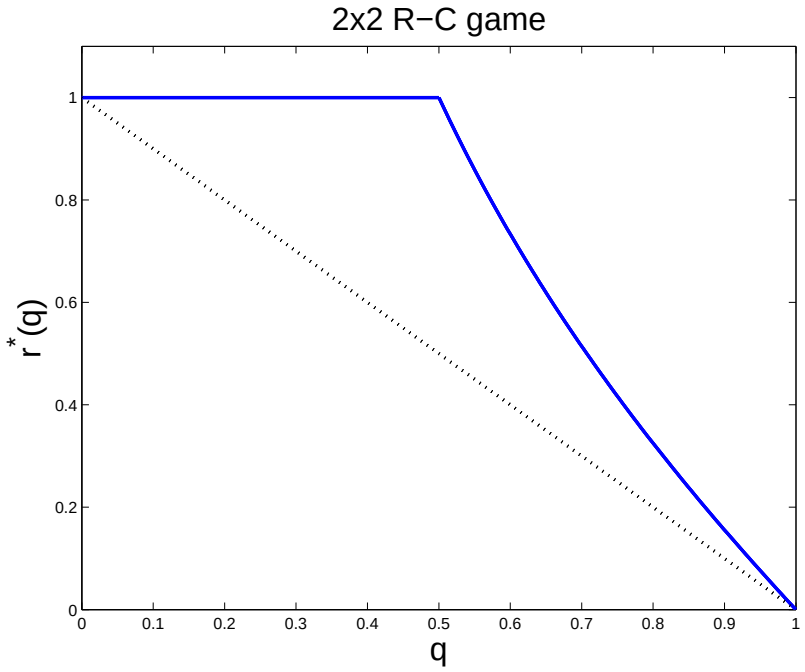


Figure 1: The reward-in-hindsight of the constrained game. Here, $r^*(q)$ is the solid line, and the dotted line connects the two extreme values, for $q = 0$ and $q = 1$.

hindsight is:

$$r^*(q) = \begin{cases} 1, & \text{if } 0 \le q \le 1/2, \\ \frac{1}{2} + \frac{1}{2q} - q, & \text{if } 1/2 \le q \le 1. \end{cases}$$

The graph of $r^*(q)$ is the solid line in Figure 1.

We now claim that P2 can make sure that P1 does not attain r^*.

Proposition 4 *If $c_0 = 0$, then there exists a strategy for P2 such that r^* cannot be attained.*

Proof Suppose that the opponent, P2, plays according to the following strategy. Initialize a counter $k = 1$. Let $\hat{\alpha}_t$ be the empirical frequency with which P1 chooses the *first* row during the first t time steps. Similarly, let $\hat{q}_t$ be the empirical frequency with which P2 chooses the *second* column during the first t time steps.

1. While $k = 1$ or $\hat{\alpha}_{t-1} > 3/4$, P2 chooses the second column, and k is incremented by 1.

2. For the next k times, P2 chooses the first column. Then, reset the counter k to , and go back to Step 1.

We now show that if

$$\limsup_{t\to\infty} \hat{c}_t \le 0, \quad \text{a.s.}, \tag{3.9}$$

then a strict inequality holds for the regret:

$$\liminf_{t\to\infty}(\hat{r}_t - r^*(\hat{q}_t)) < 0, \quad \text{a.s.}$$

Suppose that Step 2 is entered only a finite number of times. Then, after some finite time, P2 keeps choosing the second column, and $\hat{q}_t$ converges to 1. For P1 to satisfy the constraint $\limsup_{t\to\infty} \hat{c}_t \leq 0$, we must have $\limsup \hat{\alpha}_t \leq 1/2$. But then, the condition $\hat{\alpha}_{t-1} > 3/4$ will be eventually violated. This shows that Step 2 is entered an infinite number of times. In particular, there exist infinite sequences t_i and t'_i such that $t_i < t'_i < t_{i+1}$ and (i) if $t_i < t \leq t'_i$, P2 chooses the second column (Step 1); (ii) if $t'_i < t \leq t_{i+1}$, P2 chooses the first column (Step 2).

Note that Steps 1 and 2 last for an equal number of time steps. Thus, we have $\hat{q}_{t_i} = 1/2$, and $r^*(\hat{q}_{t_i}) = 1$, for all i. Furthermore, $t_{i+1} - t'_i \leq t'_i$, or $t'_i \geq t_{i+1}/2$. Note that $\hat{\alpha}_{t'_i} \leq 3/4$, because otherwise P2 would still be in Step 1 at time $t'_i + 1$. Thus, during the first t_{i+1} time steps, P1 has played the first row at most

$$3t'_i/4 + (t_{i+1} - t'_i) = t_{i+1} - t'_i/4 \leq 7t_{i+1}/8$$

times. Due to the values of the reward matrix, we have $\limsup_{t\to\infty} \hat{r}_t < \limsup_{i\to\infty} \hat{r}_{t_i}$. In particular, we have $\hat{r}_{t_{i+1}} \leq 7/8$, and $\liminf_{t\to\infty}(\hat{r}_t - r^*(\hat{q}_t)) \leq 7/8 - 1 < 0$. ■

Intuitively, the strategy that was described above allows P2 to force P1 to move, back and forth, between the extreme points ($q = 0$ and $q = 1$) that are linked by the dotted line in Figure 1. Since $r^*(q)$ is not convex, and since the dotted line is strictly below $r^*(q)$ for $q = 1/2$, this strategy precludes P1 from attaining $r^*(q)$. We note that the choice of c_0 is critical in this example. With other choices of c_0 (for example, $c_0 = -1$), the reward-in-hindsight may be attainable.

4. Attainability of the Convex Hull

Since the reward-in-hindsight is not attainable in general, we have to settle for a more modest objective. More specifically, we are interested in functions $f : \Delta(B) \to \mathbb{R}$ that are attainable with respect to a given constraint set T. As a target we suggest the closed convex hull of the reward-in-hindsight, r^*. After defining it, we prove that it is indeed attainable. In the next section, we will also show that it is the highest possible attainable function, when there is a single constraint.

Given a function $f : X \mapsto \mathbb{R}$, over a convex domain X, its *closed convex hull* is the function whose epigraph is

$$\overline{\text{conv}}(\{(x, r) : r \geq f(x)\}),$$

where $\text{conv}(D)$ is the convex hull, and $\overline{D}$ is the closure of a set D. We denote the closed convex hull of r^* by r^c.

We will make use of the following facts. Forming the convex hull and then the closure results in a larger epigraph, hence a smaller function. In particular, $r^c(q) \leq r^*(q)$, for all q. Furthermore, the closed convex hull is guaranteed to be continuous on $\Delta(B)$. (This would

not be true if we had considered the convex hull, without forming its closure.) Finally, for every q in the interior of $\Delta(B)$, we have:

$$\begin{aligned} r^c(q) \quad = \quad & \inf_{q_1,q_2,\ldots,q_k \in \Delta(B), \alpha_1,\ldots,\alpha_k} \sum_{i=1}^{k} \alpha_i r^*(q_i) \\ & \text{s.t.} \quad \sum_{i=1}^{k} \alpha_i q_i(b) = q(b), \quad \forall b \in B, \\ & \qquad \alpha_i \geq 0, \quad i = 1, 2, \ldots, k, \\ & \qquad \sum_{i=1}^{k} \alpha_i = 1, \end{aligned} \tag{4.10}$$

where k can be taken equal to $|B| + 2$ by Caratheodory's Theorem.

The following result is proved using Blackwell's approachability theory. The technique is similar to that used in other no-regret proofs (e.g., Blackwell (1956b); Mannor and Shimkin (2003)), and is based on the convexity of a target set in an appropriately defined space.

Theorem 5 *Let Assumption 1 hold for a given convex set $T \subset \mathbb{R}^d$. Then r^c is attainable with respect to T.*

Proof Define the following game with vector-valued payoffs, where the payoffs belong to $\mathbb{R} \times \mathbb{R}^d \times \Delta(B)$ (a $|B| + d + 1$ dimensional space, which we denote by $\mathcal{M}$). Suppose that P1 plays a_t, P2 plays b_t, P1 obtains an immediate reward of r_t and an immediate penalty vector of c_t. Then, the vector-valued payoff obtained by P1 is

$$m_t = (r_t, c_t, e(b_t)),$$

where $e(b)$ is a vector of zeroes, except for a 1 in its bth component. It follows that the average vector-valued reward at time t, which we define as $\hat{m}_t = \frac{1}{t}\sum_{\tau=1}^{t} m_\tau$, satisfies: $\hat{m}_t = (\hat{r}_t, \hat{c}_t, \hat{q}_t)$, where $\hat{r}_t$, $\hat{c}_t$, and $\hat{q}_t$ were defined in Eqs. (2.1), (2.2), and (2.4), respectively. Consider the sets:

$$\mathcal{B}_1 = \{(r, c, q) \in \mathcal{M} : \, r \geq r^c(q)\}, \qquad \mathcal{B}_2 = \{(r, c, q) \in \mathcal{M} : \, c \in T\},$$

and let $\mathcal{B} = \mathcal{B}_1 \cap \mathcal{B}_2$. Note that $\mathcal{B}$ is a convex set. We claim that $\mathcal{B}$ is approachable. Let $m : \Delta(A) \times \Delta(B) \to \mathcal{M}$ describe the expected payoff in a single stage game, when P1 and P2 choose actions p and q, respectively. That is,

$$m(p, q) = \Big(\sum_{a,b} p(a)q(b)R(a, b), \; \sum_{a,b} p(a)q(b)C(a, b), \; q \Big).$$

Using the sufficient condition for approachability of convex sets (Blackwell, 1956a), it suffices to show that for every q there exists a p such that $m(p, q) \in \mathcal{B}$. Fix $q \in \Delta(B)$. By Assumption 1, the constraint $\sum_{a,b} p(a)q(b)C(a, b) \in T$ is feasible, which implies that the program (2.5) has an optimal solution p^*. It follows that $m(p^*, q) \in \mathcal{B}$. We now claim that a strategy that approaches $\mathcal{B}$ also attains r^c in the sense of Definition 1. Indeed, since $\mathcal{B} \subseteq \mathcal{B}_2$ we have that $\Pr(d(c_t, T) > \epsilon \text{ infinitely often}) = 0$ for every $\epsilon > 0$. Since $\mathcal{B} \subseteq \mathcal{B}_1$ and using

the continuity of r^c, we obtain $\liminf (\hat{r}_t - r^c(\hat{q}_t)) \geq 0$. ■

We note that Theorem 5 is not constructive. Indeed, a strategy that approaches $\mathcal{B}$, based on a naive implementation Blackwell's approachability theory, requires an efficient procedure for computing the closest point in $\mathcal{B}$,and therefore a computationally efficient description of $\mathcal{B}$, which may not be available (we do not know whether $\mathcal{B}$ can be described efficiently). This motivates the development of the calibration based scheme in Section 6.

Remark 6 *Convergence rate results also follow from general approachability theory, and are generally of the order of $t^{-1/3}$; see (Mertens et al., 1994). It may be possible, perhaps, to improve upon this rate and obtain $t^{-1/2}$, which is the best possible convergence rate for the unconstrained case.*

Remark 7 *For every $q \in \Delta(B)$, we have $r^*(q) \geq \mathrm{v}$, which implies that $r^c(q) \geq \mathrm{v}$. Thus, attaining r^c guarantees an average reward at least as high as the value of the game.*

4.1 Degenerate Cases

In this section, we consider the degenerate cases where the penalty vector is affected by only one of the players. We start with the case where P1 alone affects the penalty vector, and then discuss the case where P2 alone affects the penalty vector.

If P1 alone affects the penalty vector, that is, if $C(a,b) = C(a,b')$ for all $a \in A$ and $b, b' \in B$, then $r^*(q)$ is convex. Indeed, in this case, Eq. (2.5) becomes (writing $C(a)$ for $C(a,b)$)

$$r^*(q) = \max_{p \in \Delta(A): \sum_a p(a) C(a) \in T} \sum_{a,b} p(a) q(b) R(a,b),$$

which is the maximum of a collection of linear functions of q (one function for each feasible p), and is therefore convex.

If P2 alone affects the penalty vector, that is, if $c(a,b) = c(a',b)$ for all $b \in B$ and $a, a' \in A$, then Assumption 1 implies that the constraint is always satisfied. Therefore,

$$r^*(q) = \max_{p \in \Delta(A)} \sum_{a,b} p(a) q(b) R(a,b),$$

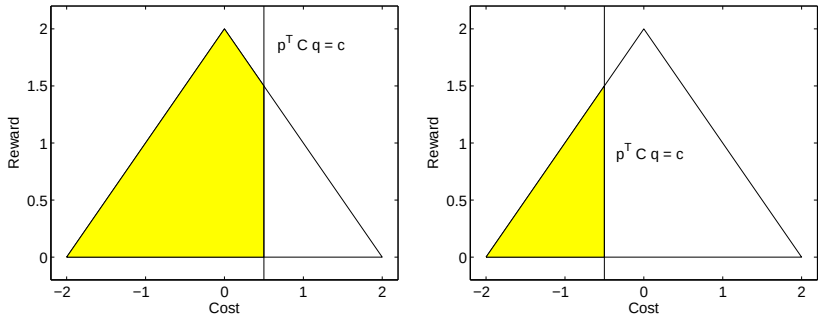
which is again a maximum of linear functions, hence convex.

We conclude that in both degenerate cases, if Assumption 1 holds, then the reward-in-hindsight is attainable.

5. Tightness of the Convex Hull

We now show that r^c is the maximal attainable function, for the case of a single constraint.

Theorem 8 *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 Assumption 1 is satisfied. Let $\tilde{r} : \Delta(B) \mapsto \mathbb{R}$ be a continuous attainable function with respect to the scalar c_0. Then, $r^c(q) \geq \tilde{r}(q)$ for all $q \in \Delta(B)$.*



(a) (b)

Figure 2: In either part (a) or (b) of the figure, we fix some $q \in \Delta(B)$. The triangle is the set of possible reward-cost pairs, as we vary p over the set $\Delta(A)$. Then, for a given value c in the upper bound on the cost (cf. (5.12)), the shaded region is the set of reward-cost pairs that also satisfy the cost constraint.

Proof The proof is constructive, as it provides a concrete strategy for P2 that prevents P1 from attaining $\tilde{r}$, unless $r^c(q) \geq \tilde{r}(q)$ for every q. Assume, in order to derive a contradiction, that there exists some $\tilde{r}$ that violates the theorem. Since $\tilde{r}$ and r^c are continuous, there exists some $q^0 \in \Delta(B)$ and some $\epsilon > 0$ such that $\tilde{r}(q) > r^c(q) + \epsilon$ for all q in an open neighborhood of q^0. In particular, q^0 can be taken to lie in the interior of $\Delta(B)$. Using Eq. (4.10), it follows that there exist $q^1, \ldots, q^k \in \Delta(B)$ and $\alpha_1, \ldots, \alpha_k$ (with $k \leq |B| + 2$, due to Caratheodory's Theorem) such that

$$\sum_{i=1}^{k} \alpha_i r^*(q^i) \leq r^c(q^0) + \frac{\epsilon}{2} < \tilde{r}(q^0) - \frac{\epsilon}{2};$$

$$\sum_{i=1}^{k} \alpha_i q^i(b) = q^0(b), \quad \forall\, b \in B; \qquad \sum_{i=1}^{k} \alpha_i = 1; \qquad \alpha_i \geq 0,\ \forall\, i.$$

Let τ be a large positive integer (τ is to be chosen large enough to ensure that the events of interest occur with high probability, etc.). We will show that if P2 plays each q^i for $\lceil \alpha_i \tau \rceil$ time steps, in an appropriate order, then either P1 does not satisfy the constraint along the way or $\hat{r}_\tau \leq \tilde{r}(\hat{q}_\tau) - \epsilon/2$.

We let q^i, $i = 1, \ldots, k$, be fixed, as above, and define a function $f_i : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$ as:

$$f_i(c) = \max_{p \in \Delta(A)} \sum_{a,b} p(a) q^i(b) R(a,b), \tag{5.11}$$

$$\text{subject to} \qquad \sum_{a,b} p(a) q^i(b) C(a,b) \leq c, \tag{5.12}$$

where the maximum over an empty set is defined to equal $-\infty$. Observe that the feasible set (and hence, optimal value) of the above linear program depends on c. Figure 2 illustrates how the feasible sets to (5.12) may depend on the value of c. By viewing Eqs. (5.11)-(5.12)

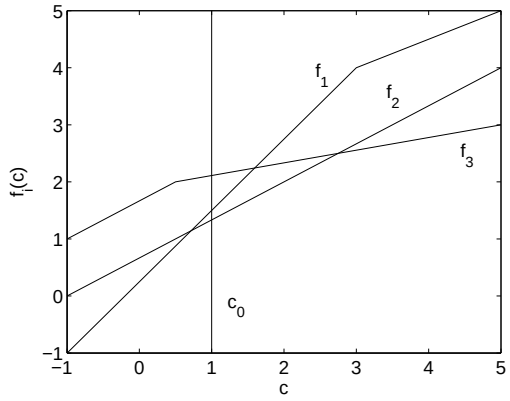


Figure 3: An example of functions f_i ordered according to ∂f_i^+.

as a parametric linear program, with a varying right-hand side parameter c, we see that $f_i(c)$ is piecewise linear, concave, and nondecreasing in c (Bertsimas and Tsitsiklis, 1997). Furthermore, $f_i(c_0) = r^*(q^i)$. Let ∂f_i^+ be the right directional derivative of f_i at $c = c_0$, and note that $\partial f_i^+ \geq 0$. From now on, we assume that the q^i have been ordered so that the sequence ∂f_i^+ is nonincreasing (e.g., as in Figure 3). To visualize the ordering that we have introduced, consider the set of possible pairs (r, c), given a fixed q. That is, consider the set $M(q^i) = \{(r, c) : \exists p \in \Delta(A) \text{ s.t. } r = \sum_{a,b} p(a)q^i(b)R(a, b),\ c = \sum_{a,b} p(a)q^i(b)C(a, b)\}$. The set $M(q^i)$ is the image of the simplex under a linear transformation, and is therefore a polytope, as illustrated by the triangular areas in Figure 2. The strategy of P2 is to first play q^i such that the p that maximizes the reward (Eq. (5.11)) satisfies Eq. (5.12) with equality. (Such a q^i results in a set $M(q^i)$ like the one shown in Figure 2(b).) After all these q^i are played, P2 plays those q^i for which the p that maximizes the reward (Eq. (5.11)) satisfies Eq. (5.12) with strict inequality, and $\partial f_i^+ = 0$. (Such a q^i results in a set $M(q^i)$ like the one shown in Figure 2(a).)

Suppose that P1 knows the sequence $q^1, \ldots, q^k$ (ordered as above) in advance, and that P2 follows the strategy described earlier. We assume that τ is large enough so that we can ignore the effects of dealing with a finite sample. Let p^i be the average of the mixed actions chosen by P1 while player P2 plays q^i. We introduce the constraints

$$\sum_{i=1}^{\ell} \alpha_i \sum_{a,b} p^i(a)q^i(b)C(a, b) \leq c_0 \sum_{i=1}^{\ell} \alpha_i, \quad \ell = 1, 2, \ldots, k.$$

These constraints must be satisfied in order to guarantee that $\hat{c}_t$ has negligible probability of substantially exceeding c_0, at the "switching" times from one mixed action to another. If P1 exploits the knowledge of P2's strategy to maximize her average reward at time τ, the resulting expected average reward at time τ will be the optimal value of the objective

function in the following linear programming problem:

$$\begin{aligned}
&\max_{p^1,p^2,\ldots,p^k} \sum_{i=1}^{k} \alpha_i \sum_{a,b} p^i(a) q^i(b) R(a,b) \\
&\quad \text{s.t.} \sum_{i=1}^{\ell} \alpha_i \sum_{a,b} p^i(a) q^i(b) C(a,b) \leq c_0 \sum_{i=1}^{\ell} \alpha_i, \quad \ell = 1,2,\ldots,k, \\
&\quad p^\ell \in \Delta(A), \quad \ell = 1,2,\ldots,k.
\end{aligned} \tag{5.13}$$

Of course, given the value of $\sum_{a,b} p^i(a) q^i(b) C(a,b)$, to be denoted by c_i, player P1 should choose a p^i that maximizes rewards, resulting in $\sum_{a,b} p^i(a) q^i(b) R(a,b) = f_i(c_i)$. Thus, the above problem can be rewritten as

$$\begin{aligned}
&\max_{c_1,\ldots,c_k} \sum \alpha_i f_i(c_i) \\
&\quad \text{s.t.} \sum_{i=1}^{\ell} \alpha_i c_i \leq c_0 \sum_{i=1}^{\ell} \alpha_i, \quad \ell = 1,2,\ldots,k.
\end{aligned} \tag{5.14}$$

We claim that letting $c_i = c_0$, for all i, is an optimal solution to the problem (5.14). This will then imply that the optimal value of the objective function for the problem (5.13) is $\sum_{i=1}^{k} \alpha_i f_i(c_0)$, which equals $\sum_{i=1}^{k} \alpha_i r^*(q^i)$, which in turn, is bounded above by $\tilde{r}(q^0) - \epsilon/2$. Thus, $\hat{r}_\tau < \tilde{r}(q^0) - \epsilon/2 + \delta(\tau)$, where the term $\delta(\tau)$ incorporates the effects due to the randomness in the process. By repeating this argument with ever increasing values of τ (so that the stochastic term $\delta(\tau)$ is averaged out and becomes negligible), we obtain that the event $\hat{r}_t < \tilde{r}(q^0) - \epsilon/2$ will occur infinitely often, and therefore $\tilde{r}$ is not attainable.

It remains to establish the claimed optimality of $(c_0,\ldots,c_0)$. Suppose that $(\overline{c}_1,\ldots,\overline{c}_k) \neq (c_0,\ldots,c_0)$ is an optimal solution of the problem (5.14). If $\overline{c}_i \leq c_0$ for all i, the monotonicity of the f_i implies that $(c_0,\ldots,c_0)$ is also an optimal solution. Otherwise, let j be the smallest index for which $\overline{c}_j > c_0$. If $\partial f_j^+ = 0$ (as in the case shown in Figure 2(b)) we have that $f_i(c)$ is maximized at c_0 for all $i \geq j$ and $(c_0,\ldots,c_0)$ is optimal. Suppose that $\partial f_j^+ > 0$. In order for the constraint (5.14) to be satisfied, there must exist some index $s < j$ such that $\overline{c}_s < c_0$. Let us perturb this solution by setting $\delta = \min\{\alpha_s(c_0 - \overline{c}_s), \alpha_j(\overline{c}_j - c_0)\}$, increasing $\overline{c}_s$ to $\tilde{c}_s = \overline{c}_s + \delta/\alpha_s$, and decreasing $\overline{c}_j$ to $\tilde{c}_j = \overline{c}_j - \delta/\alpha_j$. This new solution is clearly feasible. Let $\partial f_s^- = \lim_{\epsilon \downarrow 0}(f_s(c_0) - f_s(c_0 - \epsilon))/\epsilon$, which is the left derivative of f_s at c_0. Using the concavity of f_s, and the earlier introduced ordering, we have $\partial f_s^- \geq \partial f_s^+ \geq \partial f_j^+$. Observe that

$$\begin{aligned}
f_s(\tilde{c}_s) &= f_s(\overline{c}_s) + \partial f_s^- \delta/\alpha_s, \\
f_j(\tilde{c}_j) &= f_j(\overline{c}_j) - \partial f_j^+ \delta/\alpha_j,
\end{aligned}$$

so that $\alpha_s f_s(\tilde{c}_s) + \alpha_j f_j(\tilde{c}_j) \geq \alpha_s f_s(\overline{c}_s) + \alpha_j f_j(\overline{c}_j)$. Therefore, the new solution must also be optimal, but has fewer components that differ from c_0. By repeating this process, we eventually conclude that $(c_0,\ldots,c_0)$ is an optimal solution of (5.14). ■

To the best of our knowledge, this is the first tightness result for a performance envelope (the reward-in-hindsight) different than the Bayes envelope, for repeated games. On the other hand, we note that our proof relies crucially on the assumption of a single constraint ($d = 1$), which allows us to order the ∂f_i^+.

6. Attaining the Convex Hull Using Calibrated Forecasts

In this section, we consider a specific strategy that attains the convex hull, thus providing a constructive proof for Theorem 5. The strategy is based on forecasting P2's action, and playing a best response (in the sense of Eq. (2.5)) against the forecast. The quality of the resulting strategy depends, of course, on the quality of the forecasts; it is well known that *calibrated* forecasts lead to no-regret strategies in standard repeated matrix games. See (Foster and Vohra, 1997; Cesa-Bianchi and Lugosi, 2006) for a discussion of calibration and its implications in learning in games. In this section we consider the consequences of calibrated play for repeated games with constraints.

We start with a formal definition of calibrated forecasts and calibrated play, and then show that calibrated play attains r^c in the sense of Definition 1.

A forecasting scheme specifies at each stage k a probabilistic forecast $q_k \in \Delta(B)$ of P2's action b_k. More precisely a (randomized) forecasting scheme is a sequence of maps that associate with each possible history h_{k-1} during the first $k-1$ stages a probability measure μ_k over $\Delta(B)$. The forecast $q_k \in \Delta(B)$ is then selected at random according to the distribution μ_k. Let us clarify that for the purposes of this section, the history is defined to include the realized past forecasts.

We shall use the following definition of calibrated forecasts.

Definition 9 (Calibrated forecasts) *A forecasting scheme is* calibrated *if for every (Borel measurable) set $Q \subset \Delta(B)$ and every strategy of P1 and P2*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} 1\{q_\tau \in Q\}(e(b_\tau) - q_\tau) = 0, \quad a.s., \tag{6.15}$$

where $e(b)$ is a vector of zeroes, except for a 1 in its bth component.

Calibrated forecasts, as defined above, have been introduced into game theory in (Foster and Vohra, 1997), and several algorithms have been devised to achieve them (see (Cesa-Bianchi and Lugosi, 2006) and references therein). These algorithms typically start with predictions that are restricted to a finite grid, and gradually increase the number of grid points.

The proposed strategy is to let P1 play a best response against P2's forecasted play while still satisfying the constraints (in expectation, for the single stage game). Formally, we let:

$$\begin{aligned} p^*(q) \quad = \quad & \operatorname*{argmax}_{p\in\Delta(A)} \sum_{a,b} p(a)q(b)R(a,b) \\ & \text{s.t.} \ \sum_{a,b} p(a)q(b)C(a,b) \in T, \end{aligned} \tag{6.16}$$

where in the case of a non-unique maximum we assume that $p^*(q)$ is uniquely determined by some tie-breaking rule; this is easily done, while keeping $p^*(\cdot)$ a measurable function. The strategy is to play $p_t = p^*(q_t)$, where q_t is a calibrated forecast of P2's actions[4]. We call such a strategy a *calibrated strategy*.

The following theorem states that a calibrated strategy attains the convex hull.

Theorem 10 *Let Assumption 1 hold, and suppose that P1 uses a calibrated strategy. Then, r^c is attained with respect to T.*

Proof Fix $\epsilon > 0$. We need to show that by playing the calibrated strategy, P1 obtains $\liminf_{t\to\infty}(\hat{r}_t - r^c(\hat{q}_t)) \geq 0$ and $\limsup_{t\to\infty} \mathrm{dist}(\hat{c}_t, T) \leq 0$, almost surely.

Fix some $\epsilon > 0$. Consider a partition of the simplex $\Delta(B)$ to finitely many measurable sets $Q_1, Q_2, \ldots, Q_\ell$ such that $q, q' \in Q_i$ implies that $\|q - q'\| \leq \epsilon$ and $\|p^*(q) - p^*(q')\| \leq \epsilon$. (Such a partition exists by the compactness of $\Delta(B)$ and $\Delta(A)$. The measurability of the sets Q_i can be guaranteed because the mapping $p^*(\cdot)$ is measurable.) For each i, let us fix a representative element $q^i \in Q_i$, and let $p^i = p^*(q^i)$.

Since we have a calibrated forecast, Eq. (6.15) holds for every Q_i, $1 \leq i \leq \ell$. Define $\Gamma_t(i) = \sum_{\tau=1}^t 1\{q_\tau \in Q_i\}$ and assume without loss of generality that $\Gamma_t(i) > 0$ for large t (otherwise, eliminate those i for which $\Gamma_t(i) = 0$ for all t, and renumber the Q_i). To simplify the presentation, we assume that for every i, and for large enough t, we have $\Gamma_t(i) \geq \epsilon t$. (If for some i, and t this condition is violated, the contribution of such an i in the expressions that follow will be $O(\epsilon)$.)

By a law of large numbers for martingales, we have

$$\lim_{t\to\infty} \left(\hat{c}_t - \frac{1}{t} \sum_{\tau=1}^t C(a_\tau, b_\tau) \right) = 0, \qquad \text{a.s.} \tag{6.17}$$

By definition, we have

$$\frac{1}{t} \sum_{\tau=1}^t C(a_\tau, b_\tau) = \sum_i \frac{\Gamma_t(i)}{t} \sum_{a,b} C(a,b) \frac{1}{\Gamma_t(i)} \sum_{\tau=1}^t 1\{q_\tau \in Q_i\} 1\{a_\tau = a\} 1\{b_\tau = b\}.$$

Observe that whenever $q_\tau \in Q_i$, we have $\left\| p_\tau - p^i \right\| \leq \epsilon$, where $p_\tau = p^*(q_\tau)$ and $p^i = p^*(q^i)$ because of the way the sets Q_i were constructed. By martingale convergence, the frequency with which a will be selected whenever $q_\tau \in Q_i$ and $b_\tau = b$, will be approximately $p^i(a)$. Hence, for all b,

$$\limsup_{t\to\infty} \left| \frac{1}{\Gamma_t(i)} \sum_{\tau=1}^t 1\{q_\tau \in Q_i\} 1\{a_\tau = a\} 1\{b_\tau = b\} - p^i(a) \frac{1}{\Gamma_t(i)} \sum_{\tau=1}^t 1\{q_\tau \in Q_i\} 1\{b_\tau = b\} \right| \leq \epsilon,$$

almost surely. By the calibration property (6.15) for $Q = Q_i$, and the fact that whenever $q, q' \in Q_i$, we have $\|q - q'\| \leq \epsilon$, we obtain

$$\limsup_{t\to\infty} \left| \frac{1}{\Gamma_t(i)} \sum_{\tau=1}^t 1\{q_\tau \in Q_i\} 1\{b_\tau = b\} - q^i(b) \right| \leq \epsilon, \qquad \text{a.s.}$$

4. When the forecast μ_t is mixed, q_t is the realization of the mixed rule.

By combining the above bounds, we obtain

$$\lim_{t\to\infty}\left|\hat{c}_t - \sum_i \frac{\Gamma_t(i)}{t}\sum_{a,b} C(a,b)p^i(a)q^i(b)\right| \le 2\epsilon, \qquad \text{a.s.} \tag{6.18}$$

Note that the sum over index i in Eq. (6.18) is a convex combination (because the coefficients $\Gamma_t(i)/t$ sum to 1) of elements of T (because of the definition of p^i), and is therefore an element of T (because T is convex). This establishes that the constraint is asymptotically satisfied within $O(\epsilon)$. Note that in this argument, whenever $\Gamma_t(i)/t < \epsilon$, the summand corresponding to i is indeed of order $O(\epsilon)$ and can be safely ignored, as stated earlier.

Regarding the average reward, an argument similar to the above yields

$$\liminf_{t\to\infty}\hat{r}_t \ge \liminf_{t\to\infty}\sum_i \frac{\Gamma_t(i)}{t}\sum_{a,b} R(a,b)p^i(a)q^i(b) - 2\epsilon, \qquad \text{a.s.}$$

Next, observe that

$$\sum_i \frac{\Gamma_t(i)}{t}\sum_{a,b} R(a,b)p^i(a)q^i(b) = \sum_i \frac{\Gamma_t(i)}{t} r^*(q^i) \ge r^c\Big(\sum_i \frac{\Gamma_t(i)}{t} q^i\Big),$$

where the equality is a consequence of the definition of p^i, and the inequality follows by the definition of r^c as the closed convex hull of r^*. Observe also that the calibration property (6.15), with $Q = \Delta(B)$, implies that

$$\lim_{t\to\infty}\left\|\hat{q}_t - \frac{1}{t}\sum_{\tau=1}^{t} q_\tau\right\| = 0, \qquad \text{a.s.}$$

In turn, since $\left\|q_\tau - q^i\right\| \le \epsilon$ for a fraction $\Gamma_t(i)/t$ of the time,

$$\limsup_{t\to\infty}\left\|\hat{q}_t - \sum_i \frac{\Gamma_t(i)}{t} q^i\right\| = \limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{\tau=1}^{t} q_\tau - \sum_i \frac{\Gamma_t(i)}{t} q^i\right\| \le \epsilon, \qquad \text{a.s.}$$

Recall that the function r^c is continuous, hence uniformly continuous. Thus, there exists some function g, with $\lim_{\epsilon\downarrow 0} g(\epsilon) = 0$, such that when the argument of r^c changes by at most ϵ, the value of r^c changes by at most $g(\epsilon)$. By combining the preceding results, we obtain

$$\liminf_{t\to\infty}\hat{r}_t \ge r^c(\hat{q}_t) - 2\epsilon - g(\epsilon), \qquad \text{a.s.}$$

The above argument involves a fixed ϵ, and a fixed number ℓ of sets Q_i, and lets t increase to infinity. As such, it establishes that for any $\epsilon > 0$ the function $r^c - 2\epsilon - g(\epsilon)$ is attainable with respect to the set T^ϵ defined by $T^\epsilon = \{x \mid \mathrm{dist}(x,T) \le 2\epsilon\}$. Since this is true for every $\epsilon > 0$, we conclude that the calibrated strategy attains r^c as claimed. ■

7. Algorithms

The results in the previous section motivate us to develop algorithms for online learning with constraints, perhaps based on calibrated forecasts. For practical reasons, we are interested in computationally efficient methods, but there are no known computationally efficient calibrated forecasting algorithms. For this reason, we will consider related heuristics that are similar in spirit, even if they do not have all the desired guarantees.

We first consider a method based on the weighted average predictor. The algorithm in Table 1 keeps track of the performance of the different actions in the set A, updating a corresponding set of weights accordingly at each step. The main idea is to quantify "performance" by a linear combination of the total reward and the magnitude of the constraint violation. The parameter $\lambda > 0$ of the algorithm, which acts similar to a Lagrange multiplier, determines the tradeoff between these two objectives. When the average penalty is higher than c_0 (i.e., there is a violation), the weight of the cost term increases. When the average penalty is lower than c_0, the weight of the cost term decreases. The parameters $\overline{M}$ and $\underline{M}$ are used to bound the magnitude of the weight of the cost term; in the experiments reported in Section 8, they were set to 1000 and 0.001, respectively.

1. Set λ, w_0, $\overline{M}$, and $\underline{M}$.
2. For $t = 1, 2, \ldots$:
   (a) Sample an independent random variable a_t distributed so that
   $$a_t = a, \quad \text{with probability } \frac{w_t(a)}{\sum_{a \in A} w_t(a)} \text{ for } a \in A. \tag{7.19}$$
   (b) Compute:
   $$w_t(a) = w_{t-1}(a) \, \exp\left(\eta \left(R(a, b_t) - \lambda C(a, b_t)\right)\right), \quad a \in A. \tag{7.20}$$
   (c) For $t = 1, 2, \ldots$, update λ:
   $$\lambda := \begin{cases} \min(2\lambda, \overline{M}), & \text{if } \hat{c}_t > c_0, \\ \max(\lambda/2, \underline{M}), & \text{otherwise.} \end{cases}$$

Table 1: Exponentially weighted average predictor.

The second algorithm uses the tracking forecaster (Mannor et al., 2007) as the forecasting method. This forecaster predicts that the distribution of the next action as a weighted average of previous actions, weighing recent actions more than less recent ones others. For the special case of only two actions, it is calibrated, but *not* calibrated in general. There are, however, some special cases where it is calibrated, in particular if the sequence it tries to calibrate comes from a source with some specific properties; see Mannor et al. (2007) for details. The algorithm is presented in Table 2. If there is a current violation, it selects an action that minimizes the immediate forecasted cost. If the current average penalty does

not violate the constraint, it selects a best response to the forecasted action of P2, while satisfying the constraints.

1. Set $\rho \in (0,1)$, c_0, and $f_0 = (1/|B|)\vec{1}$.
2. For $t = 1, 2, \ldots$:
   (a) If $t = 1$ or $\hat{c}_t > c_0$, choose an action that minimizes the worst-case cost:
   $$a_t \in \operatorname*{argmin}_{a \in A} \left(C(a,b) f_{t-1}(b) \right),$$
   (b) Otherwise (if $\hat{c}_t \leq c_0$ and $t > 1$), solve
   $$\begin{aligned} \max_{p \in \Delta(A)} \quad & \sum_{a,b} p(a) R(a,b) f_{t-1}(b), \\ \text{subject to} \quad & \sum_{a,b} p(a) C(a,b) f_{t-1}(b) \leq c_0. \end{aligned}$$
   and choose a random action distributed according to the solution to the above linear program.
   (c) After observing b_t, update the forecast f_t on the probability distribution of the next opponent action b_{t+1}:
   $$f_t = f_{t-1} + (1/t)^{\rho} \left(e_{b_t} - f_{t-1} \right),$$
   where e_b is a unit vector in $\mathbb{R}^{|B|}$ with the element 1 in the component corresponding to $b \in B$.

Table 2: Tracking forecaster.

8. Experimental setup

Our experiment addresses the problem of minimizing power consumption in a computer with a human user. The agent is a low-level software controller that decides when to put the central processor (CPU) into a low-power state, thereby reducing power expenditures during periods when the user is idle. The system is driven by a human user, as well as different hardware processes, and can be realistically assumed to be non-stationary. The actions of the system correspond to hardware interrupts (most interrupts are generated by hardware controllers on the motherboard such as direct memory access, hard disk interrupts and networking interrupts) and the ongoing running processes. In the particular application at hand, there is a software interrupt (generated by the Windows operating system) every 16 milliseconds. The times of these interrupts are the decision epochs, at which the software controller can decide if and when to put the CPU to sleep before the next scheduled periodic interrupt.

However, saving energy by putting the processor in the low-power state comes at a cost. In the low-power state, a delay is incurred each time that the processor moves back into the high-power state in response to user-generated interrupts. We wish to limit the delay perceived by the human user. For this purpose, we assign a cost to the event that an interrupt arrives while the processor is in the low-power state, and impose a constraint on the time average of these costs. A similar model was used in Kveton et al. (2008), and we refer the reader to that work for further details.

We formulate the problem as follows. We divide a typical 16 millisecond interval into ten intervals. We let P1's action set be $A = \{0, 0.1, 0.2, \ldots, 1\}$, where action a corresponds to turning off the CPU after $16a$ milliseconds (the action $a = 1$ means the CPU is not turned off during the interval while the action $a = 0$ means it is turned off for the whole interval). Similarly, the action set of P2 is $B = \{0, 0.1, 0.2, \ldots, 0.9\}$, where action b corresponds to an interrupt after $16b$ milliseconds. (Note that the action $b = 0$ means there is no interrupt and that there is no point in including an action $b = 1$ in B since it would coincide with the known periodic interrupt.) The assumption is that an interrupt is handled instantaneously so if the CPU chooses a slightly larger than b it maximizes the power savings while incurring no penalty for observed delay (it is assumed for the sake of discussion that only a single interrupt is possible in each 16 millisecond interval). We define the reward at each stage as follows:

$$R(a,b) = \begin{cases} 1-a, & \text{if } b = 0 \text{ or } a > b, \quad \text{i.e., if no interrupt occurs or an interrupt occurs} \\ & \text{before the CPU turns off,} \\ b-a, & \text{if } b > 0 \text{ and } a \leq b, \quad \text{i.e., if there is an interrupt} \\ & \text{after the CPU is turned off.} \end{cases}$$

The cost is:

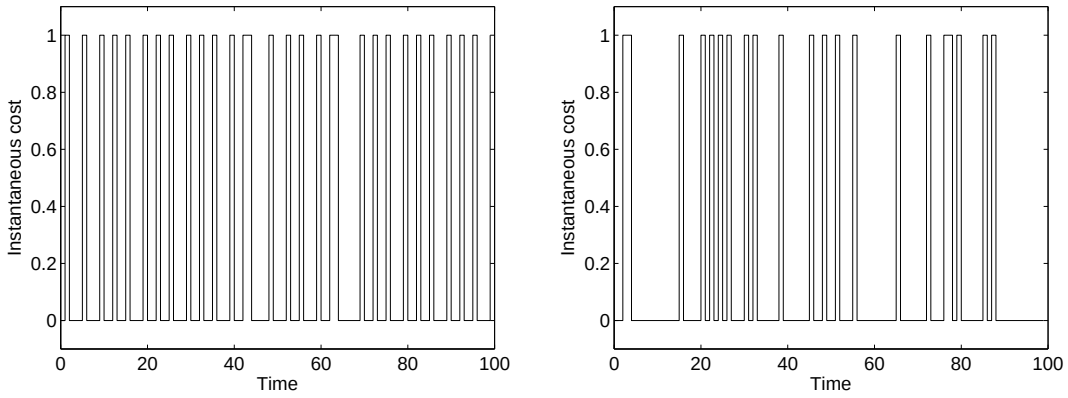
$$C(a,b) = \begin{cases} 1, & \text{if } a \leq b \text{ and } b > 0, \\ 0, & \text{otherwise.} \end{cases}$$

In "normal" operation where the CPU is powered throughout, the action is $a = 1$ and in that case there is no reward (no power saving) and no cost (no perceived delay). When $a = 0$ the CPU is turned off immediately and in this case the reward will be proportional to the amount of time until an interrupt (or until the next decision). The cost in the case $a = 0$ is 0 only is there is no interrupt ($b = 0$).

We used the real data trace obtained from what is known as MobileMark 2005 (MM05), a performance benchmark that simulates the activity of an average Microsoft Windows user. This CPU activity trace is 90 minutes long and contains more than 500,000 interrupts, including the periodic scheduled interrupts mentioned earlier. The exponentially weighted algorithm (Table 1) and the tracking forecaster (Table 2) were run on this data set. Figure 4 shows the performance of the two algorithms. The straight line shows the tradeoff between constraint violation and average reward by picking a fixed action over the entire time horizon. The different points for the exponential weighted predictor (Table 1) or the tracking forecaster (Table 2) correspond to different values of c_0. We observe that for the same average cost, the tracking forecast performs better (i.e., gets higher reward).

We selected $c_0 = 0.3$ and used both algorithms for the MM05 trace. Figures 5(a) and 5(b) show the instantaneous cost incurred by the tracking forecaster and the weighted

Figure 4: Plot of average reward against constraint violation frequency from experiments in power management for the MM05 data.



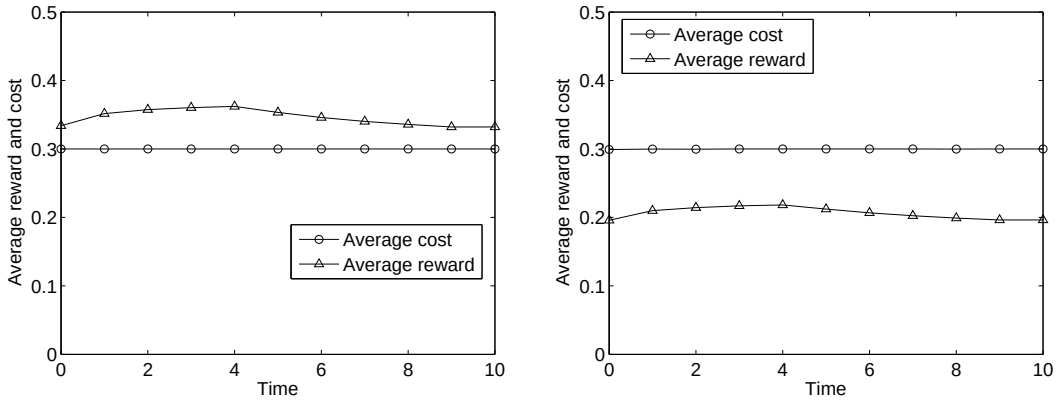
(a) Tracking forecaster (b) Weighted average predictor

Figure 5: Instantaneous cost incurred by the tracking forecaster and weighted average predictor with target constraint $c_0 = 0.3$ for the MM05 data.

average forecaster over the same short period. It should be observed that the cost of the algorithms is different, reflecting the fact that different policies are employed. Figures 6(a) and 6(b) show the time evolution of the average reward and average cost for the same experiment. In spite of not being calibrated, the tracking forecast based algorithm outperforms the exponentially weighted based algorithm.

9. Conclusions

There are several open problems and directions for future research that are worth mentioning. First, the issue of convergence rate is yet to be settled. We noted that there exists an algorithm based on approachability that converges at the rate of $t^{-1/3}$, and that the usual lower bound of $t^{-1/2}$ holds. The other algorithm based on calibration suffers from potentially even worse convergence rate, as we are not aware of any approximate calibration algorithm with comparable convergence rates. Second, the complexity of these two online learning algorithms leaves much to be desired. The complexity of a policy based on approachability theory is left undetermined because we do not have a specific procedure for computing P1's action at each stage. The per stage complexity is unknown for calibrated forecasts, but is exponential for approximately calibrated schemes (Cesa-Bianchi and Lugosi, 2006). Moreover, it is not clear whether online learning with constraints is as hard computationally as finding a calibrated forecast. Third, we only established the tightness of the lower convex hull of the Bayes envelope for the case of a one-dimensional penalty function. This is a remarkable result because it establishes the tightness of an envelope other than the Bayes envelope, and we are not aware of any such results for similar settings.



(a) Tracking forecaster (b) Weighted average predictor

Figure 6: Time evolution of average reward and average cost for the tracking forecaster and weighted average forecaster with $c_0 = 0.3$ for the MM05 data.

However, it is not clear whether such a result also holds for two-dimensional penalties. In particular, the proof technique of the tightness result does not seem to extend to higher dimensions.

Our formulation of the learning problem (learning with pathwise constraints) is only a first step in considering multi-objective problems in online learning. In particular, other formulations, e.g., that consider the number of time-windows where the constraints are violated, are of interest; see Kveton et al. (2008).

Acknowledgements

This research was partially supported by the National Science Foundation under contracts ECS-0312921 and ECCS-0701623, by the Natural Sciences and Engineering Research Council of Canada, and by the Canada Research Chairs Program. We are grateful to Georgios Theocharous and Branislav Kveton for supplying us with the CPU power management data and explaining this problem domain to us. We thank two anonymous reviewers for the comments.

References

E. Altman. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.

D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

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

D. Blackwell. Controlled random walks. In *Proc. Int. Congress of Mathematicians 1954*, volume 3, pages 336–338. North Holland, Amsterdam, 1956b.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

D. P. Foster and R. V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21:40–55, 1997.

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.

E. Hazan and N. Megiddo. Online learning with prior information. 2007. Proceedings of 20th Annual Conference on Learning Theory.

B. Kveton, J.Y. Yu, G. Theocharous, and S. Mannor. Online learning with expert advice and finite-horizon constraints. 2008. AAAI 2008, in press.

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.

S. Mannor and N. Shimkin. A geometric approach to multi-criterion reinforcement learning. *Journal of Machine Learning Research*, 5:325–360, 2004.

S. Mannor, J. S. Shamma, and G. Arslan. Online calibrated forecasts: Memory efficiency versus universality for learning in games. *Machine Learning*, 67(1–2):77–115, 2007.

J. F. Mertens, S. Sorin, and S. Zamir. Repeated games. CORE Reprint Dps 9420, 9421 and 9422, Center for Operation Research and Econometrics, Universite Catholique De Louvain, Belgium, 1994.

N. Shimkin. Stochastic games with average cost constraints. In T. Basar and A. Haurie, editors, *Advances in Dynamic Games and Applications*, pages 219–230. Birkhauser, 1994.

X. Spinat. A necessary and sufficient condition for approachability. *Mathematics of Operations Research*, 27(1):31–44, 2002.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of ICML*, 2003.