

Online Learning with Constraints

Shie Mannor¹ and John N. Tsitsiklis²

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

² Laboratory for Information and Decision Systems
Massachusetts Institute of Technology, Cambridge, MA 02139
jnt@mit.edu

Abstract. We study online learning where the objective of the decision maker is to maximize her average long-term reward given that some average constraints are satisfied along the sample path. 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. We further provide an explicit strategy that attains this convex hull using a calibrated forecasting rule.

1 Introduction

We consider a repeated game from the viewpoint of a specific 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, regarding the opponent's choice of actions. This setting was considered by Hannan [1], in the context of repeated matrix games. Hannan introduced the Bayes utility against the current empirical distribution of the opponent's actions, as a performance goal for adaptive play. This quantity is the highest average reward that player P1 could achieve, 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 [1] 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.

In this paper we consider regret minimization under sample-path constraints. That is, in addition to maximizing the 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 weather, etc.). The decision maker may be interested in the rate of successful transmissions, while minimizing 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 [2] and references therein for a discussion on constrained average cost stochastic games and to [3] for constrained Markov decision problems.

The paper is organized as follows. In Section 2, we present formally the basic model, and provide a result that relates attainability and 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. Finally, in Section 6, we provide a simple strategy, based on calibrated forecasting, that attains the convex hull.

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 stage game is a game with two players: P1 (the decision maker of interest) and P2 (who represents Nature and is assumed arbitrary). In this context, 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 $A = \{1, 2, \dots, |A|\}$.
2. B is the set of actions of P2; we will assume $B = \{1, 2, \dots, |B|\}$.
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 accumulated 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 shall assume that T is convex and closed. Since C is 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

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

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

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

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 keeping the average penalty vector in T , pathwise:

$$\text{for every } \epsilon > 0, \quad \Pr(\text{dist}(\hat{c}_t, T) > \epsilon \text{ infinitely often}) = 0, \quad (2.3)$$

where $\text{dist}(\cdot)$ is the point-to-set Euclidean distance, i.e., $\text{dist}(x, T) = \inf_{y \in T} \|y - x\|_2$, 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 of $T = \{c \in \mathbb{R}^d : c \leq c_0\}$. 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:

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

where the inequality is interpreted componentwise.

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,

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

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 program³

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

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).

In 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 componentwise).

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) $\liminf_{t \rightarrow \infty} (\hat{r}_t - r(\hat{q}_t)) \geq 0$, a.s., and
- (ii) $\limsup_{t \rightarrow \infty} \text{dist}(\hat{c}_t, T) \rightarrow 0$, 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

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

2.3 The value of the game

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

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

³ If T is a polyhedron (specified by finitely many linear inequalities), then the optimization problem is a linear program.

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:

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

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 is outside T .

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

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

The existence of a strategy for P1 that attains the value was proven in [4] in the broader context of stochastic games.

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

- (i) *P1 has a strategy that guarantees that the constant function $r(q) \equiv v$ is attained with respect to T .*
- (ii) *For every number $v' > v$ there exists $\delta > 0$ such that P2 has a strategy that guarantees that either $\liminf_{t \rightarrow \infty} \hat{r}_t < v'$ or $\limsup_{t \rightarrow \infty} \text{dist}(\hat{c}_t, T) > \delta$, almost surely. (In particular, the constant function v' is not attainable.)*

Proof. The proof relies on Blackwell's approachability theory (see [5]). We construct a nested sequence of convex sets in \mathbb{R}^{d+1} denoted 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 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 [5] that every S_α is either approachable or excludable. If S_α is approachable, then S_β is approachable for every $\beta < \alpha$. We define $v_0 = \sup\{\beta \mid S_\beta \text{ is approachable}\}$. It follows that S_{v_0} is approachable (as the limit of approachable sets; see [6]). 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. \square

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 $v = \inf_q r^*(q)$.

Remark 1. We note in order to attain the value of the game, P1 may have to use a non-stationary strategy. This is in contrast to standard (non-constrained) games, in which P1 always has an optimal stationary strategy that attains the value of the game.

Remark 2. 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.

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 $0 \leq q \leq 1/2$, in which case $r^*(q) = 1$. For $1/2 \leq q \leq 1$, 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 while maximizing the reward. If we let α denote the frequency of choosing the first row, we see that the reward and penalty are:

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

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 $c(\alpha) \leq 0$, which is $\alpha(q) = 1/2q$, and the optimal reward is $1/2 + 1/2q - q$. We conclude that the reward-in-hindsight is:

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

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

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

Proposition 2. *If $c_0 = 0$, then there exists a strategy for P2 such that $r^*(q)$ cannot be attained.*

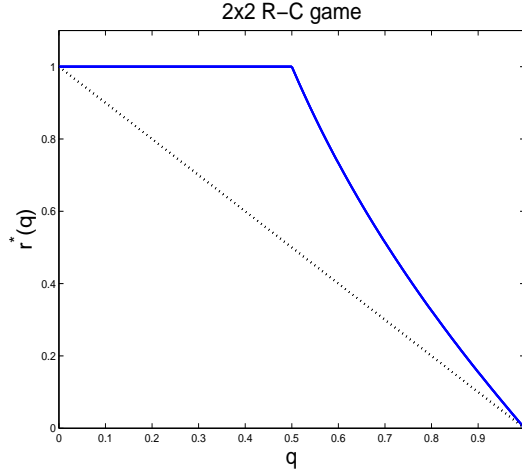


Fig. 1. The reward-in-hindsight of the constrained game. Here, $r^*(q)$ is the bold thick line, and the dotted line connects the two extreme values, for $q = 0$ and $q = 1$.

Proof. (Outline) Suppose that P2 starts by playing the second column for some long time τ . At time τ , P2's empirical frequency of choosing the second column is $\hat{q}_\tau = 1$. As computed before, $r^*(\hat{q}_\tau) = 0$. Since P1 tries to satisfy $\hat{c}_\tau \leq 0$, and also have the average reward by time τ as high as $r^*(\hat{q}_\tau)$, P1 must choose both rows with equal probability and obtain a reward of $\hat{r}_\tau = 0$, which equals $r^*(\hat{q}_\tau)$. This is essentially the best that can be achieved (neglecting negligible effects of order $1/\tau$). In the next τ time stages, P2 plays the first column. The empirical frequency of P2 at time 2τ is $\hat{q}_{2\tau} = 1/2$. During these last τ periods, P1 can choose the first row and achieve a reward of 1 (which is the best possible), and also satisfy the constraint. In that case, $\hat{r}_{2\tau} \leq 1/2$, while $r^*(\hat{q}_{2\tau}) = 1$. Player P2 can then repeat the same strategy, but replacing τ with some τ' which is much bigger than τ (so that the first 2τ stages are negligible). \square

Using the strategy that was described above, P2 essentially forces P1 to traverse the dotted line in Fig. 1. It so happens that $r^*(q)$ is not convex, and the dotted line is below $r^*(q)$ which 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 look for a more modest objective. More specifically, we look for functions $f : \Delta(B) \rightarrow \mathbb{R}$ 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 with respect to the constraint set. 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 $f : X \mapsto \mathbb{R}$, 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. 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.) Furthermore, for every q in the interior of $\Delta(B)$, we have

$$\begin{aligned} r^c(q) &= \inf_{q_1, q_2, \dots, q_k \in \Delta(B), \alpha_1, \dots, \alpha_k} \sum_{i=1}^k \alpha_i r^*(q_i) & (4.8) \\ \text{s.t. } & \sum_{i=1}^k \alpha_i q_i(b) = q(b), \quad b \in B, \\ & \alpha_i \geq 0, \quad i = 1, 2, \dots, k, \\ & \sum_{i=1}^k \alpha_i = 1, \end{aligned}$$

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., [7, 8]), and is based on the convexity of a target set that resides in an appropriately defined space.

Theorem 1. *Let Assumption 1 hold with respect to some 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 the b th location. It follows that the average vector-valued reward at time t , which we denote by $\hat{m}_t = \frac{1}{t} \sum_{\tau=1}^t m_\tau$, 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:

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

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 $m : \Delta(A) \times \Delta(B) \rightarrow \mathcal{M}$ describe the expected payoff in a one shot game,

when P1 and P2 choose actions p and q , respectively. That is,

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

Using the sufficient condition for approachability of convex sets ([5]), 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 $\liminf (\hat{r}_t - r^c(\hat{q}_t)) \geq 0$. \square

Remark 3. Convergence rate results also follow from general approachability theory, and are generally of the order of $t^{-1/3}$; see [9]. It may be possible, perhaps, to improve upon this rate (and obtain $t^{-1/2}$ as in the non-constrained case), but this is beyond the scope of this paper.

Remark 4. For every $q \in \Delta(B)$, we have $r^*(q) \geq v$, which implies that $r^c(q) \geq v$. 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, 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 observe 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 2. *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 $\tilde{r}(q) : \Delta(B) \mapsto \mathbb{R}$ be an attainable continuous function with respect to the scalar c_0 . Then, $r^c(q) \geq \tilde{r}(q)$ for all $q \in \Delta(B)$.*

Proof. The proof is constructive, as it provides a concrete strategy for P2, which 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. (4.8), it follows that there exist $q^1, \dots, q^k \in \Delta(B)$ and $\alpha_1, \dots, \alpha_k$ (with $k \leq |B| + 2$) such that

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

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

Let τ be a large number (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 $\alpha_i \tau$ time steps, in an appropriate order, then either P1 does not satisfy the constraint along the way or $\hat{r}_\tau \leq \tilde{r}(\hat{q}_\tau) - \epsilon/2$.

For $i = 1, \dots, k$, we define a function $f_i : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{-\infty\}$, by letting $f_i(c)$ be the maximum of

$$\sum_{a,b} p(a) q^i(b) R(a,b),$$

subject to

$$p \in \Delta(A), \quad \text{and} \quad \sum_{a,b} p(a) q^i(b) C(a,b) \leq c,$$

where the maximum over an empty set is defined to equal $-\infty$. We note that $f_i(c)$ is piecewise linear, concave, and nondecreasing in c . Furthermore, $f_i(c_0) = r^*(q^i)$. Let f_i^+ be the right directional derivative of f_i at $c = c_0$. From now on, we assume that the q^i have been ordered so that the sequence f_i^+ is non-increasing.

Suppose that P1 knows the sequence q^1, \dots, q^k (ordered as above) in advance, and that P2 will be following the strategy described earlier. We assume that τ is large enough so that we can ignore the effects of dealing with a finite sample, or of $\alpha_i \tau$ not being an integer. We allow P1 to choose any sequence of p^1, \dots, p^k , and introduce the constraints

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

These constraints are required 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:

$$\begin{aligned} & \max_{p^1, p^2, \dots, p^k} \sum_{i=1}^k \alpha_i \sum_{a,b} p^i(a) q^i(b) R(a, b) \\ & \text{s.t.} \sum_{i=1}^{\ell} \alpha_i \sum_{a,b} p^i(a) q^i(b) C(a, b) \leq c_0 \sum_{i=1}^{\ell} \alpha_i, \quad \ell = 1, 2, \dots, k, \quad (5.9) \\ & p^\ell \in \Delta(A), \quad \ell = 1, 2, \dots, k. \end{aligned}$$

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

$$\begin{aligned} & \max_{c_1, \dots, c_k} \sum \alpha_i f_i(c_i) \\ & \text{s.t.} \sum_{i=1}^{\ell} \alpha_i c_i \leq c_0 \sum_{i=1}^{\ell} \alpha_i, \quad \ell = 1, 2, \dots, k. \quad (5.10) \end{aligned}$$

We claim that letting $c_i = c_0$, for all i , is an optimal solution to the problem (5.10). This will then imply that the optimal value of the objective function for the problem (5.9) 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, \dots, c_0) . Suppose that $(\bar{c}_1, \dots, \bar{c}_k) \neq (c_0, \dots, c_0)$ is an optimal solution of the problem (5.10). If $\bar{c}_i \leq c_0$ for all i , the monotonicity of the f_i implies that (c_0, \dots, c_0) is also an optimal solution. Let us therefore assume that there exists some j for which $\bar{c}_j > c_0$. In order for the constraint (5.10) to be satisfied, there must exist some index $s < j$ such that $\bar{c}_s < c_0$. Let us perturb this solution by setting $\delta = \min\{\alpha_s(c_0 - \bar{c}_s), \alpha_j(\bar{c}_j - c_0)\}$, increasing \bar{c}_s to $\bar{c}_s = \bar{c}_s + \delta/\alpha_s$, and decreasing \bar{c}_j to $\bar{c}_j = \bar{c}_j - \delta/\alpha_j$. This new solution is clearly feasible. Let $f_s^- = \lim_{\epsilon \downarrow 0} (f_s(c_0) - f_s(c_0 - \epsilon))$, which is the left derivative of f_s at c_0 . Using concavity, and the earlier introduced ordering, we have $f_s^- \geq f_s^+ \geq f_j^+$, from which it follows easily (the detailed argument is omitted) that $f_s(\bar{c}_s) + f_j(\bar{c}_j) \geq f_s(\bar{c}_s) + f_j(\bar{c}_j)$. 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, \dots, c_0) is optimal. \square

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 standard repeated decision problems.

6 Attaining the Convex Hull Using Calibrated Forecasts

In this section we consider a specific strategy that attains the convex hull, thus strengthening Theorem 1. 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 forecast; it is well known that using *calibrated* forecasts leads to no-regret strategies in standard repeated matrix games. See [10, 11] 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 2 (Calibrated forecasts). *A forecasting scheme is calibrated if for every (Borel measurable) set $Q \subset \Delta(B)$ and every strategy of P2,*

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=1}^t 1\{q_\tau \in Q\} (e(b_\tau) - q_\tau) = 0 \quad a.s., \quad (6.11)$$

where $e(b)$ is a vector of zeroes, except for a 1 in the b th location.

Calibrated forecasts, as defined above, have been introduced into game theory in [10], and several algorithms have been devised to achieve them (see [11] 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 one-shot game). Formally, we let:

$$\begin{aligned} p^*(q) &= \arg \max_{p \in \Delta(A)} \sum_{a,b} p(a)q(b)R(a,b) & (6.12) \\ \text{s.t.} & \sum_{a,b} p(a)q(b)C(a,b) \in T, \end{aligned}$$

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 actions⁴. We call such a strategy a *calibrated strategy*.

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

Theorem 3. *Let Assumption 1 hold, and suppose that P1 uses a calibrated strategy. Then r^c is attainable with respect to T .*

Proof. (Outline) Fix $\epsilon > 0$. We need to show that by playing the calibrated strategy, P1 obtains $\liminf \hat{r}_t - r^c(\hat{q}_t) \geq -\epsilon$ and $\limsup \text{dist}(\hat{c}_t, T) \leq \epsilon$ almost surely. Due to lack of space, we only provide an outline of the proof.

Consider a partition of the simplex $\Delta(B)$ to finitely many measurable sets Q_1, Q_2, \dots, Q_ℓ such that $q, q' \in Q_i$ implies that $\|q - q'\| \leq \epsilon/K$ and $\|p^*(q) - p^*(q')\| \leq \epsilon/K$, where K is a large constant. (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. (6.11) 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 will have $\Gamma_t(i) \geq \epsilon t/K$. (If for some i , and t this condition is violated, the contribution of such an i in the expressions that follow will be $O(\epsilon)$.) In the sequel the approximate equality sign " \approx " will indicate the presence of an approximation error term, e_t , that satisfies $\limsup_{t \rightarrow \infty} e_t \leq L\epsilon$, almost surely, where L is a constant.

We have

$$\begin{aligned} \hat{c}_t &\approx \frac{1}{t} \sum_{\tau=1}^t C(a_\tau, b_\tau) \\ &= \sum_i \frac{\Gamma_t(i)}{t} \sum_{a,b} C(a,b) \frac{1}{\Gamma_t(i)} \sum_{\tau=1}^t 1\{q_\tau \in Q_i\} 1\{a_\tau = a\} 1\{b_\tau = b\} \\ &\approx \sum_i \frac{\Gamma_t(i)}{t} \sum_{a,b} C(a,b) p^i(a) \frac{1}{\Gamma_t(i)} \sum_{\tau=1}^t 1\{q_\tau \in Q_i\} 1\{b_\tau = b\} \\ &\approx \sum_i \frac{\Gamma_t(i)}{t} \sum_{a,b} C(a,b) p^i(a) q^i(b). \end{aligned} \tag{6.13}$$

The first approximate equality follows from laws of large numbers. The second approximate equality holds because whenever $q_\tau \in Q_i$, p_τ is approximately equal to $p^*(q_i) = p^i$, and by laws of large numbers, the frequency with which a will

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

be selected will be approximately $p^i(a)$. The last approximate equality holds by virtue of the calibration property (6.11) with $Q = Q_i$, and the fact that whenever $q_\tau \in Q_i$, we have $q_\tau \approx q^i$.

Note that the right-hand side expression (6.13) is a convex combination (because the $\Gamma_t(i)/t$ 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 ϵ . Note that in this argument, whenever $\Gamma_t(i)/t < \epsilon/K$, the summand corresponding to i is indeed of order $O(\epsilon)$ and can be safely ignored, as stated earlier.

Regarding the average reward, a similar argument yields

$$\begin{aligned} \hat{r}_t &\approx \sum_i \frac{\Gamma_t(i)}{t} \sum_{a,b} R(a,b) p^i(a) q^i(b) \\ &= \sum_i \frac{\Gamma_t(i)}{t} r^*(q^i) \\ &\geq r^c \left(\sum_i \frac{\Gamma_t(i)}{t} q^i \right) \\ &\approx r^c(\hat{q}_t). \end{aligned}$$

The first approximate equality is obtained similar to (6.13), with $C(a,b)$ replaced by $R(a,b)$. The equality that follows is a consequence of the definition of p^i . The inequality that follows is obtained because of the definition of r^c as the closed convex hull of r^* . The last approximate equality relies on the continuity of r^c , and the fact

$$\hat{q}_t \approx \frac{1}{t} \sum_{\tau=1}^t q_\tau \approx \sum_i \frac{\Gamma_t(i)}{t} q^i.$$

To justify the latter fact, the first approximate equality follows from the calibration property (6.11), with $Q = \Delta(B)$, and the second because q_t is approximately equal to q^i for a fraction $\Gamma_t(i)/t$ of the time.

The above outlined 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 $r^c - \epsilon$ is attainable with respect to the set T^ϵ defined by $T^\epsilon = \{x \mid \text{dist}(x, T) \leq \epsilon\}$. Since this is true for every $\epsilon > 0$, we conclude that the calibrated strategy attains r^c as claimed. \square

Acknowledgements

This research was partially supported by the National Science Foundation under contract ECS-0312921, by the Natural Sciences and Engineering Research Council of Canada, and by the Canada Research Chairs Program.

References

1. 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.

2. S. Mannor and N. Shimkin. A geometric approach to multi-criterion reinforcement learning. *Journal of Machine Learning Research*, 5:325–360, 2004.
3. E. Altman. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.
4. 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.
5. D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific J. Math.*, 6(1):1–8, 1956.
6. X. Spinat. A necessary and sufficient condition for approachability. *Mathematics of Operations Research*, 27(1):31–44, 2002.
7. D. Blackwell. Controlled random walks. In *Proc. Int. Congress of Mathematicians 1954*, volume 3, pages 336–338. North Holland, Amsterdam, 1956.
8. 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.
9. 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.
10. D. P. Foster and R. V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21:40–55, 1997.
11. N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, 2006.