Journal of Machine Learning Research 16 (2015) ?-? Submitted 10/13; Revised 7/14; Published 4/15

Response-Based Approachability with Applications to Generalized No-Regret Problems

Andrey Bernstein ANDREY.BERNSTEIN@EPFL.CH
School of Computer and Communication Sciences
EPFL—École Polytechnique Fédérale de Lausanne
Lausanne CH-1015, Switzerland

Nahum Shimkin SHIMKIN@EE.TECHNION.AC.IL
Department of Electrical Engineering
Technion—Israel Institute of Technology
Haifa 32000, Israel

Editor: Alexander Rakhlin

Abstract

Blackwell's theory of approachability provides fundamental results for repeated games with vector-valued payoffs, which have been usefully applied in the theory of learning in games, and in devising online learning algorithms in the adversarial setup. A target set S is approachable by a player (the agent) in such a game if he can ensure that the average payoff vector converges to S, no matter what the opponent does. Blackwell provided two equivalent conditions for a convex set to be approachable. Standard approachability algorithms rely on the primal condition, which is a geometric separation condition, and essentially require to compute at each stage a projection direction from a certain point to S. Here we introduce an approachability algorithm that relies on Blackwell's *dual* condition, which requires the agent to have a feasible *response* to each mixed action of the opponent, namely a mixed action such that the expected payoff vector belongs to S. Thus, rather than projections, the proposed algorithm relies on computing the response to a certain action of the opponent at each stage. We demonstrate the utility of the proposed approach by applying it to certain generalizations of the classical regret minimization problem, which incorporate side constraints, reward-to-cost criteria, and so-called global cost functions. In these extensions, computation of the projection is generally complex while the response is readily obtainable.

Keywords: approachability, no-regret algorithms

1. Introduction

Consider a repeated matrix game with *vector-valued* rewards that is played by two players, the *agent* and the *opponent*, where the latter may stand for an arbitrarily-varying learning environment. For each pair of simultaneous actions a and b of the agent and the opponent in the one-stage game, a reward vector $r(a, b) \in \mathbb{R}^\ell$, $\ell \geq 1$, is obtained. In the approachability problem formulated in (Blackwell, 1956), the agent's goal is to ensure that the long-term average reward vector *approaches* a given target set S, namely converges to S almost surely in the point-to-set distance. If that convergence can be guaranteed irrespectively of the opponent's strategy, the set S is said to be *approachable*, and the strategy of the agent that satisfies this property is an approachability strategy (or algorithm) for S. Refinements and extensions of Blackwell's results have been considered, among others, in Vieille (1992); Shimkin and Shwartz (1993); Hart and Mas-Colell (2001); Spinat (2002); Lehrer (2002); Lehrer and Solan (2009); Abernethy et al. (2011).

Blackwell's approachability results have been broadly used in the theoretical work on learning in games, encompassing equilibrium analysis in repeated games with incomplete information (Aumann and Maschler, 1995), calibrated forecasting (Foster, 1999), and convergence to correlated equilibria (Hart and Mas-Colell, 2000). An application of approachability to multi-criteria reinforcement learning was considered in Mannor and Shimkin (2004). The earliest application, however, concerned the notion of *no-regret strategies*, that was introduced in Hannan (1957). Even before Hannan's paper appeared in print, it was shown in Blackwell (1954) that regret minimization can be formulated as a particular approachability problem, leading to an elegant no-regret strategy. More recently, approachability was used in Rustichini (1999) to establish a no-regret result for games with imperfect monitoring, and Hart and Mas-Colell (2001) proposed an alternative approachability formulation of the no-regret problem (see Section 5 for more details). An overview of approachability and no-regret in the context of learning in games can be found in Fudenberg and Levine (1998) and Young (2004), while Cesa-Bianchi and Lugosi (2006) highlights the connection with the modern theory of on-line learning and prediction algorithms. The recent article Perchet (2014) reviews the inter-relations between approachability, regret minimization and calibration.

Standard approachability algorithms require, at each stage of the game, the computation of the direction from the current average reward vector to a closest point in the target set S. This is implied by Blackwell's *primal* geometric separation condition, which is a sufficient condition for approachability of a target set. For *convex* sets, this step is equivalent to computing the *projection direction* of the average reward onto S. In this paper, we introduce an approachability algorithm that avoids this projection computation step. Instead, the algorithm relies on the availability of a *response map*, that assigns to each mixed action q of the opponent a mixed action p of the agent so that $r(p, q)$, the expected reward vector under these two mixed actions, is in S. Existence of such a map is based on the Blackwell's *dual* condition, which is also a necessary and sufficient condition for approachability of a convex target set.

The idea of defining an approachable set in terms of a general response map appears in Lehrer and Solan (2007), in the context of internal no-regret strategies. An explicit approachability algorithm which is based on computing the response to *calibrated forecasts* of the opponent's actions has been proposed in Perchet (2009), and further analyzed in Bernstein et al. (2014). However, the algorithms in these papers are essentially based on computing calibrated forecasts of the opponent's actions, a task which is computationally hard (Hazan and Kakade, 2012). In contrast, the algorithms proposed in the present paper retain the dimensionality of the single-stage game, similarly to Blackwell's original algorithm. An approachability algorithm that combines the response map with no-regret learning was proposed in Bernstein (2013). The algorithm accommodates some additional adaptive properties, but its temporal convergence rate is $O(n^{-1/4})$ rather than $O(n^{-1/2})$. A similar algorithm was employed in Mannor et al. (2014) to elegantly establish approachability results for unknown games.

Our motivation for the proposed algorithms is mainly derived from certain generalizations of the basic no-regret problem, where the set to be approached is geometrically complicated so that computing the projection direction may be hard, while the response map is explicit by construction. These generalizations include the constrained regret minimization problem (Mannor et al., 2009), regret minimization with global cost functions (Even-Dar et al., 2009), regret minimization in variable duration repeated games (Mannor and Shimkin, 2008), and regret minimization in stochastic game models (Mannor and Shimkin, 2003). In these cases, the computation of a response reduces to computing a *best-response* in the underlying regret minimization problem, and hence can be carried out efficiently. The application of our algorithm to some of these problems is discussed in Section 5 of this paper.

The paper proceeds as follows. In Section 2 we review the approachability framework along with available approachability algorithms. Section 3 presents our basic algorithm and establishes its approachability properties. In Section 4, we provide an interpretation of the proposed algorithm, and examine some variants and extensions. Section 5 presents the application to generalized no-regret problems. We conclude the paper in Section 6.

2. Review of Approachability Theory

Let us start with a brief review of the approachability problem. Consider a repeated two-person matrix game, played between an agent and an arbitrary opponent. The agent chooses its actions from a finite set $\mathcal{A}$, while the opponent chooses its actions from a finite set $\mathcal{B}$. At each step $n = 1, 2, ...$, the agent selects its action $a_n \in \mathcal{A}$, observes the action $b_n \in \mathcal{B}$ chosen by the opponent, and obtains a *vector-valued* reward $R_n = r(a_n, b_n) \in \mathbb{R}^\ell$, where $\ell \geq 1$, and $r : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^\ell$ is a given reward function. The average reward vector obtained by the agent up to time n is then $\bar{R}_n = n^{-1} \sum_{k=1}^{n} R_k$. A *mixed* action of the agent is a probability vector $p \in \Delta(\mathcal{A})$, where $p(a)$ specifies the probability of choosing action $a \in \mathcal{A}$, and $\Delta(\mathcal{A})$ denotes the set of probability vectors over $\mathcal{A}$. Similarly, $q \in \Delta(\mathcal{B})$ denotes a mixed action of the opponent. Let $\bar{q}_n \in \Delta(\mathcal{B})$ denote the empirical distribution of the opponent's actions at time n, namely

$$\bar{q}_n(b) \triangleq \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}\left\{b_n = b\right\}, \quad b \in \mathcal{B},$$

where $\mathbb{I}$ denotes the indicator function. Further define the Euclidean span of the reward vector as

$$\rho \triangleq \max_{a,b,a',b'} \left\| r(a,b) - r(a',b') \right\|, \tag{1}$$

where $\|\cdot\|$ is the Euclidean norm. The inner product between two vectors $v \in \mathbb{R}^\ell$ and $w \in \mathbb{R}^\ell$ is denoted by $v \cdot w$.

In what follows, we use the shorthand notation

$$r(p,q) \triangleq \sum_{a \in \mathcal{A}, b \in \mathcal{B}} p(a) q(b) r(a,b)$$

for the expected reward under mixed actions $p \in \Delta(\mathcal{A})$ and $q \in \Delta(\mathcal{B})$; the distinction between $r(a,b)$ and $r(p,q)$ should be clear from their arguments. We similarly denote

$r(p,b) = \sum_{a\in\mathcal{A}} p(a)r(a,b)$ for the expected reward under mixed action $p \in \Delta(\mathcal{A})$ and pure action $b \in \mathcal{B}$.

Let $h_n \triangleq \{a_1, b_1, ..., a_n, b_n\} \in (\mathcal{A}\times\mathcal{B})^n$ denote the history of the game up to stage n. A *strategy* $\pi = (\pi_n)$ of the agent is a collection of decision rules $\pi_n : (\mathcal{A}\times\mathcal{B})^{n-1} \to \Delta(\mathcal{A})$, $n \geq 1$, where each mapping π_n specifies a mixed action $p_n = \pi_n(h_{n-1})$ for the agent at time n. The agent's pure action a_n is sampled from p_n. Similarly, the opponent's strategy is denoted by $\sigma = (\sigma_n)$, with $\sigma_n : (\mathcal{A}\times\mathcal{B})^{n-1} \to \Delta(\mathcal{B})$. Let $\mathbb{P}^{\pi,\sigma}$ denote the probability measure on $(\mathcal{A}\times\mathcal{B})^\infty$ induced by the strategy pair (π,σ).

Let S be a given target set in the reward space. We may assume that S is closed as approachability of a set and its closure are equivalent.

Definition 1 (Approachable Set) *A closed set $S \subseteq \mathbb{R}^\ell$ is* approachable *by the agent if there exists a strategy π of the agent such that $\bar{R}_n = n^{-1}\sum_{k=1}^n R_k$ converges to S in the Euclidean point-to-set distance $d(\cdot, S)$, almost surely for every strategy σ of the opponent, at a uniform rate over the opponent's strategies. That is, for every $\epsilon > 0$ there exists an integer N such that*

$$\mathbb{P}^{\pi,\sigma}\{\sup_{n\geq N} d(\bar{R}_n, S) \geq \epsilon\} \leq \epsilon$$

for any strategy σ of the opponent.

In the sequel, we will find it convenient to state most of our results in terms of the time averaged *expected* rewards, where expectation is applied only to the agent's mixed actions:

$$\bar{r}_n = \frac{1}{n}\sum_{k=1}^n r_k, \quad \text{where} \quad r_k = r(p_k, b_k).$$

With these smoothed rewards, the stated convergence results and bounds can be shown to hold *pathwise*, for any possible sequence of the opponent's actions. See, e.g., Theorem 4, which states that $d(\bar{r}_n, S) \leq \frac{\rho}{\sqrt{n}}$ for all n. The corresponding almost sure convergence for the actual average reward $\bar{R}_n$ readily follows using martingale convergence theory. Indeed, observe that

$$d\left(\bar{R}_n, S\right) \leq \left\|\bar{R}_n - \bar{r}_n\right\| + d\left(\bar{r}_n, S\right),$$

where the first normed term is the time average of the vector-valued and uniformly bounded martingale difference sequence $D_k = r(a_k, b_k) - r(p_k, b_k)$. By standard martingale results, this average converges to zero at a uniform rate of $O(n^{-1/2})$.

We proceed to present a formulation of Blackwell's results, which provide a sufficient condition for approachability of general sets, and two sets of necessary and sufficient conditions for approachability of *convex* sets. For any $x \notin S$, let $c(x) \in S$ denote a closest point in S to x. Also, for any $p \in \Delta(\mathcal{A})$, let $T(p) = \{r(p,q) : q \in \Delta(\mathcal{B})\}$ denote the set of mean reward vectors that are achievable by the opponent. This evidently coincides with the convex hull of the vectors $\{r(p,b)\}_{b\in\mathcal{B}}$.

Definition 2 (Approachability Conditions)

(i) ***B-sets:*** *A closed set $S \subseteq \mathbb{R}^\ell$ will be called a B-set if for every $x \notin S$ there exists a mixed action $p^* = p^*(x) \in \Delta(\mathcal{A})$ and a closest point $c(x) \in S$ such that the hyperplane through $c(x)$ perpendicular to the line segment x-$c(x)$, separates x from $T(p^*)$.*

(ii) ***D-sets:*** *A closed set $S \subseteq \mathbb{R}^\ell$ will be called a D-set if for every $q \in \Delta(\mathcal{B})$ there exists a mixed action $p \in \Delta(\mathcal{A})$ so that $r(p, q) \in S$. We shall refer to such p as a* response *(or S-response) of the agent to q.*

Theorem 3 (Blackwell, 1956)

(i) ***Primal Condition and Algorithm.*** *A B-set is approachable, by using at stage n the mixed action $p^*(\bar{r}_{n-1})$ whenever $\bar{r}_{n-1} \notin S$. If $\bar{r}_{n-1} \in S$, an arbitrary action can be used.*

(ii) ***Dual Condition.*** *A closed set S is approachable only if it is a D-set.*

(iii) ***Convex Sets.*** *Let S be a closed convex set. Then, the following statements are equivalent: (a) S is approachable, (b) S is a B-set, (c) S is a D-set.*

We note that the approachability algorithm in Theorem 3(i) remains valid if $\bar{r}_{n-1}$ in the primal condition is replaced by $\bar{R}_{n-1}$. Blackwell's algorithm was generalized in Hart and Mas-Colell (2001) to a class of approachability algorithms, where the required steering directions are generated as gradients of a suitable potential function (rather than Euclidean projections). An alternative construction was recently proposed in Abernethy et al. (2011), where the steering directions are generated through a no-regret algorithm. Finally, as already mentioned, calibration-based approachability algorithms were considered in Perchet (2009) and Bernstein et al. (2014).

3. Response-Based Approachability

In this section we present our basic response-based algorithm, and establish its convergence properties. In the remainder of the paper, we shall assume that the target set S satisfies the following assumption.

Assumption 1 *The set S is a closed, convex and approachable set.*

It follows by Theorem 3 that S is a D-set, so that for all $q \in \Delta(\mathcal{B})$ there exists an S-response $p \in \Delta(\mathcal{A})$ such that $r(p, q) \in S$. It is further assumed that the agent can compute a response to any q.

We note that in some cases of interest, including those discussed in Section 5, the target S may itself be defined through an appropriate response map. Suppose that for each $q \in \Delta(\mathcal{B})$, we are given a mixed action $p^*(q) \in \Delta(\mathcal{A})$, devised so that $r(p^*(q), q)$ satisfies some desired properties. Then the convex hull $S = \text{conv}\{r(p^*(q), q),\, q \in \Delta(\mathcal{B})\}$ is a convex D-set by construction, hence approachable.

The proposed approachability strategy is presented in Algorithm 1. The general idea is as follows. At each stage n of the algorithm, a *steering vector* $\lambda_{n-1} = \bar{r}^*_{n-1} - \bar{r}_{n-1}$ is computed as the difference between the current average reward and the average of a certain sequence of *target points* r^*_k in S. The target point r^*_n is computed as $r(p^*_n, q^*_n)$, where p^*_n is chosen as an S-response to a certain fictitious action q^*_n of the opponent. Both p_n (the actual mixed action of the agent) and q^*_n are computed in step 3 of the algorithm, as the optimal strategies in the scalar game obtained by projecting the payoff vectors in the direction of

```
**Algorithm 1** Response-Based Approachability
**Initialization:** At time step $n = 1$, use arbitrary mixed action $p_1$ and set an arbitrary target point $r_1^* \in S$.
**At time step** $n = 2, 3, ...$:
    1. Set an approachability direction
        $$\lambda_{n-1} = \bar{r}_{n-1}^* - \bar{r}_{n-1},$$
       where
        $$\bar{r}_{n-1} = \frac{1}{n-1} \sum_{k=1}^{n-1} r(p_k, b_k), \qquad \bar{r}_{n-1}^* = \frac{1}{n-1} \sum_{k=1}^{n-1} r_k^*$$
       are, respectively, the average (smoothed) reward vector and the average target point.
    2. Solve the zero-sum matrix game with payoff matrix defined by $r(a, b)$ projected in the direction $\lambda_{n-1}$. Namely, find the equilibrium strategies $p_n$ and $q_n^*$ that satisfy
        $$p_n \in \operatorname{argmax}_{p \in \Delta(\mathcal{A})} \min_{q \in \Delta(\mathcal{B})} \lambda_{n-1} \cdot r(p, q), \qquad (2)$$
        $$q_n^* \in \operatorname{argmin}_{q \in \Delta(\mathcal{B})} \max_{p \in \Delta(\mathcal{A})} \lambda_{n-1} \cdot r(p, q), \qquad (3)$$
    3. Choose action $a_n$ according to $p_n$.
    4. Pick $p_n^*$ so that $r(p_n^*, q_n^*) \in S$, and set the target point $r_n^* = r(p_n^*, q_n^*)$.
```

λ_{n-1}. As shown in the proof, and further elaborated in Subsection 4.1, this choice implies the convergence of the difference $\lambda_n = \bar{r}_n^* - \bar{r}_n$ to 0. Since $\bar{r}_n^* \in S$ by construction, this in turn implies convergence of $\bar{r}_n$ to S.

We may now present our main convergence result and its proof, followed by some additional comments on the algorithm. Recall that ρ is reward span as defined in (1).

Theorem 4 *Let Assumption 1 hold, and suppose that the agent follows the strategy specified in Algorithm 1. Then*

$$d\left(\bar{r}_n, S\right) \leq \|\lambda_n\| \leq \frac{\rho}{\sqrt{n}}, \quad n \geq 1, \tag{4}$$

for any strategy of the opponent.

The proof follows from the next result, which also provides more general conditions on the required properties of (p_n, q_n^*, p_n^*).

Proposition 5 *(i) Suppose that at each time step $n \geq 1$, the agent chooses the triple (p_n, q_n^*, p_n^*) so that*

$$\lambda_{n-1} \cdot \left(r(p_n, b) - r(p_n^*, q_n^*)\right) \geq 0, \quad \forall b \in \mathcal{B}, \tag{5}$$

and sets $r_n^* = r(p_n^*, q_n^*)$. Then $\|\lambda_n\| \leq \frac{\rho}{\sqrt{n}}$ for $n \geq 1$.

(ii) If, in addition, p_n^* is chosen as an S-response to q_n^*, so that $r_n^* = r(p_n^*, q_n^*) \in S$, then

$$d\left(\bar{r}_n, S\right) \leq \|\lambda_n\| \leq \frac{\rho}{\sqrt{n}}, \quad n \geq 1, \tag{6}$$

Proof We first observe that

$$n^2 \|\lambda_n\|^2 \leq (n-1)^2 \|\lambda_{n-1}\|^2 + 2(n-1)\lambda_{n-1} \cdot (r_n^* - r_n) + \rho^2, \tag{7}$$

for any $n \geq 1$. Indeed,

$$\begin{aligned} \|\bar{r}_n^* - \bar{r}_n\|^2 &= \left\| \frac{n-1}{n}\left(\bar{r}_{n-1}^* - \bar{r}_{n-1}\right) + \frac{1}{n}\left(r_n^* - r_n\right) \right\|^2 \\ &= \left(\frac{n-1}{n}\right)^2 \|\lambda_{n-1}\|^2 + \frac{1}{n^2}\|r_n^* - r_n\|^2 + 2\frac{n-1}{n^2}\lambda_{n-1} \cdot (r_n^* - r_n) \\ &\leq \left(\frac{n-1}{n}\right)^2 \|\lambda_{n-1}\|^2 + \frac{\rho^2}{n^2} + 2\frac{n-1}{n^2}\lambda_{n-1} \cdot (r_n^* - r_n)\,. \end{aligned}$$

Now, under condition (5),

$$\lambda_{n-1} \cdot (r_n^* - r_n) = \lambda_{n-1} \cdot (r(p_n^*, q_n^*) - r(p_n, b_n)) \leq 0.$$

Hence, by (7),

$$n^2 \|\lambda_n\|^2 \leq (n-1)^2 \|\lambda_{n-1}\|^2 + \rho^2, \quad n \geq 1.$$

Applying this inequality recursively, we obtain that $n^2 \|\lambda_n\|^2 \leq n\rho^2$, or $\|\lambda_n\|^2 \leq \rho^2/n$, as claimed in part (i). Part (ii) now follows since $r_n^* \in S$ implies that $\bar{r}_n^* \in S$ (by convexity of S), hence

$$d\left(\bar{r}_n, S\right) \leq \|\bar{r}_n - \bar{r}_n^*\| = \|\lambda_n\|\ .$$

■

Proof [Theorem 4] It only remains to show that the choice of (p_n, q_n^*) in equations (2)-(3) implies the required inequality in (5). Indeed, under (2) and (3) we have that

$$\begin{aligned} \lambda_{n-1} \cdot r(p_n, b_n) &\geq \max_{p \in \Delta(\mathcal{A})} \min_{q \in \Delta(\mathcal{B})} \lambda_{n-1} \cdot r(p,q) \\ &= \min_{q \in \Delta(\mathcal{B})} \max_{p \in \Delta(\mathcal{A})} \lambda_{n-1} \cdot r(p,q) \\ &\triangleq \max_{p \in \Delta(\mathcal{A})} \lambda_{n-1} \cdot r(p, q_n^*), \end{aligned}$$

where the equality follows by the minimax theorem for matrix games. Therefore, condition (5) is satisfied for any p_n^*, and in particular for the one satisfying $r(p_n^*, q_n^*) \in S$. This concludes the proof of Theorem 4. ■

Additional Comments:

1. Observe that the projection directions in Blackwell's algorithm are replaced, in a sense, by the steering vectors λ_n. These vectors are computed based on the agent's S-responses to a fictitious sequence (q_n^*) of the opponent's mixed actions, which is computed as part of the algorithm.

2. Theorem 4 clearly implies that the set S is approachable with the specified strategy, and provides an explicit rate of convergence. In fact, the result is somewhat stronger as it implies convergence of the average reward vector to $\bar{r}_n^* \in S$. This property will be found useful in Proposition 13 below, where certain properties that do not follow from approachability alone are established for the reward-to-cost maximization problem.

3. A stated in Proposition 5, the condition in (5) on the triplets (p_n, q_n^*, p_n^*) is sufficient to ensure the convergence $\lambda_n \to 0$. Equations (2)-(3) specify a specific choice of (p_n, q_n^*) which satisfies these conditions. This choice is useful as it implies (5) for *any* choice of p_n^*.

4. The computational requirements of Algorithm 1 are as follows. At each time step n, two major computations are needed:

   a. Computing (p_n, q_n^*)—the equilibrium strategies in the zero-sum matrix game with the reward function $\lambda_{n-1} \cdot r(p,q)$. This boils down to the solution of the related primal and dual linear programs, and hence can be done efficiently. Note that, given the vector λ_{n-1}, this computation does not involve the target set S.

   b. Computing the S-response p_n^* to q_n^* and the target point $r_n^* = r(p_n^*, q_n^*)$, which is problem dependent. Specific examples are discussed in Section 5.

4. Interpretation and Extensions

We open this section with an illuminating interpretation of the proposed algorithm in terms of a certain approachability problem in an auxiliary game. We then proceed to present three variants and extensions to the basic algorithm; we note that these are not essential for the remainder of the paper and can be skipped at first reading. While each of these variants is presented separately, they may also be combined when appropriate.

4.1 An Auxiliary Game Interpretation

A central part of Algorithm 1 is the choice of the pair (p_n, q_n^*) so that $\bar{r}_n$ tracks $\bar{r}_n^*$, namely $\lambda_n = \bar{r}_n^* - \bar{r}_n \to 0$ (see Equations (2)-(3) and Proposition 5). If fact, the choice of (p_n, q_n^*) in (2)-(3) can be interpreted as Blackwell's strategy for a specific approachability problem in an auxiliary game, which we define next.

Suppose that at stage n, the agent chooses a *pair* of actions $(a, b^*) \in \mathcal{A} \times \mathcal{B}$ and the opponent chooses a pair of actions $(a^*, b) \in \mathcal{A} \times \mathcal{B}$. The vector payoff function, now denoted by v, is given by

$$v((a, b^*), (a^*, b)) = r(a^*, b^*) - r(a, b),$$

so that

$$V_n = r(a_n^*, b_n^*) - R_n.$$

Consider the single-point target set $S_0 = \{0\} \subset \mathbb{R}^\ell$. This set is clearly convex, and we next show that it is a D-set in the auxiliary game. We need to show that for any $\eta \in \Delta(\mathcal{A} \times \mathcal{B})$ there exists $\mu \in \Delta(\mathcal{A} \times \mathcal{B})$ so that $v(\mu, \eta) \in S_0$, namely $v(\mu, \eta) = 0$. That that end, observe that

$$v(\mu, \eta) = r(p^*, q^*) - r(p, q)$$

where p and q^* are the marginal distributions of μ on $\mathcal{A}$ and $\mathcal{B}$, respectively, while p^* and q are the respective marginal distributions of η. Therefore we obtain $v(\mu, \eta) = 0$ by choosing μ with the same marginals as η, for example $\{\mu(a, b) = p(a)q^*(b)\}$ with $p = p^*$ and $q^* = q$. Thus, by Theorem 3, S_0 is approachable.

We may now apply Blackwell's approachability strategy to this auxiliary game. Since S_0 is the origin, the direction from S_0 to the average reward $\bar{v}_{n-1}$ is just the average reward vector itself. Therefore, the primal (geometric separation) condition here is equivalent to

$$\bar{v}_{n-1} \cdot v(\mu, \eta) \leq 0, \quad \forall\, \eta \in \Delta(\mathcal{A} \times \mathcal{B})$$

or

$$\bar{v}_{n-1} \cdot (r(p^*, q^*) - r(p, q)) \leq 0, \quad \forall\, p^* \in \Delta(\mathcal{A}), q \in \Delta(\mathcal{B}).$$

Now, a pair (p, q^*) that satisfies this inequality is any pair of equilibrium strategies in the zero-sum game with reward v projected in the direction of $\bar{v}_{n-1}$. That is, for

$$p \in \operatorname*{argmax}_{p \in \Delta(\mathcal{A})} \min_{q \in \Delta(\mathcal{B})} \bar{v}_{n-1} \cdot r(p, q), \tag{8}$$

$$q^* \in \operatorname*{argmin}_{q \in \Delta(\mathcal{B})} \max_{p \in \Delta(\mathcal{A})} \bar{v}_{n-1} \cdot r(p, q), \tag{9}$$

it is easily verified that

$$\bar{v}_{n-1} \cdot r(p^*, q^*) \geq \bar{v}_{n-1} \cdot r(p, q), \quad \forall\, p^* \in \Delta(\mathcal{A}), q \in \Delta(\mathcal{B})$$

as required.

The choice of (p_n, q_n^*) in Equations (2)-(3) follows (8)-(9), with λ_{n-1} replacing $\bar{v}_{n-1}$. We note that the two are not identical, as $\bar{v}_n$ is the temporal average of $V_n = r(a_n^*, b_n^*) - r(a_n, b_n)$ while λ_n is the average of the expected difference $r(p_n^*, q_n^*) - r(p_n, b_n)$; however this does not change the approachability result above, and in fact either can be used. More generally, any approachability algorithm in the auxiliary game can be used to choose the pair (p_n, q_n^*) in Algorithm 1.

We note that in our original problem, the mixed action p_n^* is not chosen by an "opponent" but rather specified as part of Algorithm 1. But since the approachability result above holds for an arbitrary choice of p_n^*, it also holds for this particular one.

We proceed to present some additional variants of our algorithm.

4.2 Idling when S is Reached

Recall that in the original approachability algorithm of Blackwell, an *arbitrary* action a_n can be chosen by the agent whenever $\bar{r}_{n-1} \in S$. This may alleviate the computational burden of the algorithm, and adds another degree of freedom that may be used to optimize other criteria.

Such an arbitrary choice of a_n (or p_n) when the average reward is in S is also possible in our algorithm. However, some care is required in setting the average target point $\bar{r}_n^*$ at these time instances, as otherwise the two terms of the difference $\lambda_n = \bar{r}_n^* - \bar{r}_n$ may drift apart. As it turns out, $\bar{r}_n^*$ should be reset at these times to $\bar{r}_n$, which leads to the following recursion. Set $\bar{r}_0^* = 0$, and let

$$\bar{r}_n^* = \begin{cases} \frac{n-1}{n}\bar{r}_{n-1}^* + \frac{1}{n}r_n^* & \text{if } \bar{r}_n \notin S \\ \bar{r}_n & \text{if } \bar{r}_n \in S \end{cases} \tag{10}$$

for $n \geq 1$. The definition of λ_n as $\bar{r}_n^* - \bar{r}_n$ is retained, so that it satisfies the modified recursion:

$$\lambda_n = \begin{cases} \frac{n-1}{n}\lambda_{n-1} + \frac{1}{n}(r_n^* - r_n), & \text{if } \bar{r}_n \notin S \\ 0, & \text{if } \bar{r}_n \in S, \end{cases} \tag{11}$$

with $\lambda_0 = 0$. Thus, the steering vector λ_n is reset to 0 whenever the average reward $\bar{r}_n$ is in S. With this modified definition, the convergence properties of the algorithm are retained (with the same rates). The proof can be found in Bernstein and Shimkin (2013).

4.3 Directionally Unbounded Target Sets

In some applications of interest, the target set S may be unbounded in certain directions. It is often natural to define the agent's goal in this way even if the reward function is bounded, as it reflects clearly the agent's desire of obtaining a reward which is as large as possible in these directions.[1] Indeed, this is the case in the approachability formulations of the no-regret problem, where the goal is essentially to make the (scalar) average reward as large as possible in hindsight.

In such cases, the requirement that $\lambda_n = \bar{r}_n^* - \bar{r}_n \to 0$, which is a property of our basic algorithm, may be too strong, and may even be counter-productive. For example, suppose that our goal is to increase the first coordinate of the average reward vector $\bar{r}_n$ as much as possible. In that case, allowing negative values of λ_n in that component makes sense (rather than steering it to 0 by reducing $\bar{r}_n$). We propose here a modification of our algorithm that addresses this issue

Given the (closed and convex) target set $S \subset \mathbb{R}^\ell$, let D_S be the set of vectors $d \in \mathbb{R}^\ell$ such that $d + S \subset S$. It may be seen that D_S is a closed and convex cone, which trivially equals $\{0\}$ if (and only if) S is bounded. We refer to the unit vectors in D_S as directions in which S is unbounded.

Referring to the auxiliary game interpretation of our algorithm in Section 4.1, we may now relax the requirement that λ_n approaches $\{0\}$ to the requirement that λ_n approaches $-D_S$. Indeed, if we maintain $\bar{r}_n^* \in S$ as before, then $\lambda_n \in -D_S$ suffices to verify that $\bar{r}_n = \bar{r}_n^* - \lambda_n \in S$.

We may now apply Blackwell's approachability strategy to the cone D_S in place of the origin. The required modification to the algorithm is simple: replace the steering direction λ_n in (2)-(3) or (5) with the direction from the closest point in $-D_S$ to λ_n:

$$\tilde{\lambda}_n = \lambda_n - \mathrm{Proj}_{-D_S}(\lambda_n)$$

That projection is particularly simple in case S is unbounded along primary coordinates, so that the cone D_S is a quadrant, generated by a collection $e_j, j \in J$ of orthogonal unit

[1] Clearly, it is always possible to intersect S with the bounded set of feasible reward vectors without changing its approachability properties. We find it useful here to retain S in its unbounded form.

vectors. In that case, clearly,

$$\mathrm{Proj}_{-D_S}(\lambda) = -\sum_{j\in J}(e_j \cdot \lambda)^- \,.$$

Thus, the negative components of λ_n in directions (e_j) are nullified.

The modified algorithm admits analogous bounds to those of the basic algorithm, with (4) or (6) replaced by

$$d\left(\bar{r}_n, S\right) \leq d(\lambda_n, -D_S) \leq \frac{\rho}{\sqrt{n}}, \quad n \geq 1.$$

The proof is identical, and is obtained by replacing λ_n with $\tilde{\lambda}_n = \lambda_n - \mathrm{Proj}_{-D_S}(\lambda_n)$ in all the relations. See Bernstein and Shimkin (2013) for details.

4.4 Using the Actual Rewards

In the basic algorithm of Section 3, the definition of the steering direction λ_n employs the expected rewards $r(p_k, b_k)$ rather than the actual rewards $R_k = r(a_k, b_k)$. We consider here the variant of the algorithm which employs the latter. This is essential in case that the opponent's action b_k is not observed, so that $r(p_k, b_k)$ cannot be computed, while the reward vector R_k is observed directly. It also makes some sense in general since the quantity we are actually interested in is the average reward $\bar{R}_n$, and not its expected version $\bar{r}_n$.

Thus, we replace λ_{n-1} with

$$\tilde{\lambda}_{n-1} = \bar{r}^*_{n-1} - \bar{R}_{n-1}.$$

The rest of the algorithm remains the same as Algorithm 1. We have the following result for this variant.

Theorem 6 *Let Assumption 1 holds. If the agent uses Algorithm 1, with λ_{n-1} replaced by*

$$\tilde{\lambda}_{n-1} = \bar{r}^*_{n-1} - \bar{R}_{n-1},$$

it holds that

$$\lim_{n\to\infty} \|\tilde{\lambda}_n\| = 0,$$

almost surely, for any strategy of the opponent, at a uniform rate of $O(1/\sqrt{n})$ over all strategies of the opponent. More precisely, for every $\epsilon > 0$,

$$\mathbb{P}\left\{\sup_{k\geq n} \|\tilde{\lambda}_k\| \geq \epsilon\right\} \leq \frac{2\rho^2}{n\epsilon^2}. \tag{12}$$

Proof First observe that Lemma 7 still holds if $r_n = r(p_n, b_n)$ is replaced with $R_n = r(a_n, b_n)$ throughout. Namely,

$$n^2\|\tilde{\lambda}_n\|^2 \leq (n-1)^2\|\tilde{\lambda}_{n-1}\|^2 + 2(n-1)\tilde{\lambda}_{n-1} \cdot (r^*_n - r(a_n, b_n)) + \rho^2, \quad n \geq 1.$$

Let $\{\mathcal{F}_n\}$ denote the filtration induced by the history. We have that

$$\begin{aligned}\mathbb{E}\left[n^2\|\tilde{\lambda}_n\|^2 \;\middle|\; \mathcal{F}_{n-1}\right] &\leq (n-1)^2\|\tilde{\lambda}_{n-1}\|^2 + 2(n-1)\tilde{\lambda}_{n-1}\cdot\mathbb{E}\left[(r_n^* - r(a_n,b_n)) \mid \mathcal{F}_{n-1}\right] + \rho^2 \\ &= (n-1)^2\|\tilde{\lambda}_{n-1}\|^2 + 2(n-1)\tilde{\lambda}_{n-1}\cdot(r_n^* - \mathbb{E}\left[r(a_n,b_n) \mid \mathcal{F}_{n-1}\right]) + \rho^2 \\ &\leq (n-1)^2\|\tilde{\lambda}_{n-1}\|^2 + \rho^2, \qquad (13)\end{aligned}$$

where the equality follows since q_n^* and p_n^* are determined by the history up to time $n-1$ and hence so does $r_n^* = r(p_n^*, q_n^*)$, and the last inequality holds since

$$\tilde{\lambda}_{n-1}\cdot(r_n^* - \mathbb{E}\left[r(a_n,b_n) \mid \mathcal{F}_{n-1}\right]) = \tilde{\lambda}_{n-1}\cdot(r_n^* - r(p_n,b_n)) \leq 0,$$

similarly to the proof of Theorem 4.

From (13) we may deduce the almost sure convergence $\|\tilde{\lambda}_n\|$ to zero, at a rate the depends on ρ only. The argument may follow the original proof of Blackwell's theorem (Blackwell (1956), Theorem 1), or its adaptation in Shimkin and Shwartz (1993, Proposition 4.1) or Mertens et al. (1994, p. 125) which rely on Doob's maximal inequality for supermartingales. In particular, following the latter reference, we obtain the bound stated in (12). ■

5. Applications to Generalized No-Regret Problems

Our response-based approachability algorithm can be usefully applied to several generalized regret minimization problems, for which computation of a projection onto the target set is involved, but a response is readily obtainable. In the next Subsection, we briefly review the basic no-regret problem and its two standard formulations as an approachability problem. In Subsection 5.2 we first outline a generic *generalized no-regret* problem, using a general set-valued goal function, and then specialize the discussion to some specific problems that have been considered in the recent literature, namely constrained regret minimization, reward-to-cost maximization, and the so-called global cost function problem. In each case, we specify the performance obtainable by a suitable approachability algorithm, along with the corresponding response map that is needed in our algorithm. For the reward-to-cost problem, we also derive some performance guarantees that rely on specific properties of the proposed approachability algorithm.

We do not specify convergence rates in this section, but rather focus on asymptotic convergence results. Convergence rates can be derived by referring to our bounds in the previous sections, namely (4) or (12).

5.1 Approachability-Based No-Regret Algorithms

Let us start by reviewing the basic no-regret problem for repeated matrix games, along with its two alternative formulations as an approachability problem by Blackwell (1954) and Hart and Mas-Colell (2001). Consider, as before, an agent that faces an arbitrarily varying environment (the opponent). The repeated game model is the same as above, except that the vector reward function r is replaced by a scalar reward (or utility) function

$u : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$. Let $\bar{U}_n \triangleq n^{-1} \sum_{k=1}^{n} U_k$ denote the average reward by time n, and let

$$u^*(\bar{q}_n) \triangleq \max_{a \in \mathcal{A}} u(a, \bar{q}_n) = \frac{1}{n} \max_{a \in \mathcal{A}} \sum_{k=1}^{n} u(a, b_k) \tag{14}$$

denote the *best reward-in-hindsight* of the agent after observing $b_1, ..., b_n$, which is a *convex* function u^* of the empirical distribution $\bar{q}_n$. Hannan (1957) introduced the following notion of a no-regret strategy:

Definition 7 (No-Regret Algorithm) *A strategy of the agent is termed a no-regret algorithm (or Hannan Consistent) if*

$$\limsup_{n \to \infty} \left(u^*(\bar{q}_n) - \bar{U}_n \right) \leq 0$$

with probability 1, for any strategy of the opponent.

a. Blackwell's No-Regret Algorithm. Following Hannan's seminal paper, Blackwell (1954) used his approachability theorem to elegantly show the existence of regret minimizing strategies. Define the vector-valued rewards $R_n \triangleq (U_n, \mathbf{1}(b_n)) \in \mathbb{R} \times \Delta(\mathcal{B})$, where $\mathbf{1}(b)$ is the probability vector in $\Delta(\mathcal{B})$ supported on b. The corresponding average reward is $\bar{R}_n \triangleq n^{-1} \sum_{k=1}^{n} R_k = \left(\bar{U}_n, \bar{q}_n\right)$. Finally, define the target set

$$S = \left\{ (u, q) \in \mathbb{R} \times \Delta(\mathcal{B}) : \ u \geq u^*(q) \right\}.$$

This set is a D-set by construction: An S-response to q is given by any $p^* \in \Delta(\mathcal{A})$ that maximizes $u(p, q)$, as $u(p^*, q) = u^*(q)$ implies that $r(p^*, q) = (u(p^*, q), q) \in S$. Also, S is a convex set by the convexity of $u^*(q)$ in q. Hence, by Theorem 3, S is approachable, and by the continuity of $u^*(q)$, an algorithm that approaches S also minimizes the regret in the sense of Definition 7. Application of Blackwell's approachability strategy to the set S therefore results in a no-regret algorithm. We note that the required projection of the average reward vector onto S cannot be defined explicitly in this formulation. However, the computation of the S-response is explicit and straightforward: We just need to solve the original optimization problem $\max_{p \in \Delta(\mathcal{A})} u(p, q)$, which clearly admits a solution in pure actions.

b. Regret Matching. An alternative formulation due to Hart and Mas-Colell (2001) leads to a simple and explicit no-regret algorithm. Let

$$L_n(a') \triangleq \frac{1}{n} \sum_{k=1}^{n} \left(u(a', b_k) - u(a_k, b_k) \right) \tag{15}$$

denote the regret accrued due to not using action a' exclusively up to time n. The no-regret requirement in Definition 7 is now equivalent to $\limsup_{n \to \infty} L_n(a) \leq 0, a \in \mathcal{A}$, a.s. for any strategy of the opponent. This property, in turn, is equivalent to the approachability of the negative orthant $S = (\mathbb{R}_-)^{\mathcal{A}}$ in the game with vector payoff $r = (r_{a'}) \in \mathbb{R}^{\mathcal{A}}$, defined as $r_{a'}(a, b) = u(a', b) - u(a, b)$.

To verify the dual condition, observe that $r_{a'}(p, q) = u(a', q) - u(p, q)$. Choosing $p \in \operatorname{argmax}_p u(p, q)$ clearly ensures $r(p, q) \in S$, hence is an S-response to q (in the sense of

Definition 2(ii)), and S is a D-set. Note that the response here can always be taken as a pure action.

It was shown in Hart and Mas-Colell (2001) that the application of Blackwell's approachability strategy (or some generalizations thereof) to this formulation is simple and leads to explicit no-regret algorithms, namely the so-called *regret matching* algorithm and its variants.

5.2 Generalized No-Regret

Consider a repeated matrix game as before, except that the vector-valued reward $r(a, b)$ is now denoted by $v(a, b) \in \mathbb{R}^K$. Suppose that for each mixed action q of the opponent, the agent defines a *target set* $V^*(q) \subset \mathbb{R}^K$ which is non-empty and closed. Let $V^* : \Delta(\mathcal{B}) \rightrightarrows \mathbb{R}^K$ denote the corresponding set-valued map, which assigns to each q the subset $V^*(q)$. We refer to V^* as the agent's *goal function.* Denote[2] $v_n = v(a_n, b_n)$, $\bar{v}_n = \frac{1}{n}\sum_{k=1}^n v_k$.

Definition 8 (Attainability) *A strategy of the agent is said to be no-regret strategy with respect to the set-valued goal function V^* if*

$$\lim_{n\to\infty} d(\bar{v}_n, V^*(\bar{q}_n)) = 0 \quad (a.s),$$

for any strategy of the opponent. If such a strategy exists we say that V^* is attainable by the agent.

The classical no-regret problem is obtained as a special case, with scalar rewards $v(a, b)$ and target set $V^*(q) = \{u \in \mathbb{R} : u \geq v^*(q)\}$, where $v^*(q) \triangleq \max_p u(p, q)$.

Attainability is closely related to approachability of the graph of V^*. Recall that the graph of a set-valued map $V : \Delta(\mathcal{B}) \rightrightarrows \mathbb{R}^K$ is defined as

$$\mathrm{Graph}(V) \triangleq \left\{(v, q) \in \mathbb{R}^K \times \Delta(\mathcal{B}) : \; v \in V(q)\right\}.$$

(For this and other properties of set-valued maps see, e.g., Aubin and Frankowska, 1990 or Rockafellar and Wets, 1997, Chapter 5.) It is easily seen that attainability of V^* implies approachability of Graph(V), in the game with augmented vector rewards $r(p, q) = (v(p, q), q)$. The converse is also true under a continuity requirement.

Lemma 9 *Let $V : q \mapsto V^*(q) \cap \mathcal{V}_0$ denote the restriction of V^* to the compact set $\mathcal{V}_0 =$ conv$\{v(a, b)\}$ of feasible reward vectors. Suppose that V is continuous in the Hausdorff metric. If* Graph(V^*) *is approachable in the repeated game with reward vector $r(p, q) = (v(p, q), q)$, then V^* is attainable. Specifically, any approachability strategy for* Graph(V^*) *is a no-regret strategy for V^*.*

Proof Clearly, since $\bar{v}_n \in \mathcal{V}_0$, if Graph($V^*$) is approachable then so is Graph(V), and we may restrict attention to the latter. Recall that the Hausdorff distance $d_{\mathcal{H}}$ between sets X and Y, defined by

$$d_{\mathcal{H}}(X, Y) = \max\{\sup_{x\in X} d(x, Y), \sup_{y\in Y} d(y, X)\},$$

[2] For notational convenience, we will not use here the capitalized notation $V_n = v(a_n, b_n)$ to distinguish the latter from $v(p_n, b_n)$, as was done above for r. In fact, v_n can stand for either in the following, depending on whether Algorithm 1 or its variant in Subsection 4.4 is used.

is a metric on the space of non-empty compact subsets of $\mathbb{R}^K$. Now, V may be viewed as a map from the compact set $\Delta(\mathcal{B})$ to the metric space of non-empty compact subsets of $\mathbb{R}^K$ with the Hausdorff metric, and is continuous in that metric by assumption. Hence, by the Heine-Cantor Theorem, V is *uniformly* continuous.

Now, since $S = \text{Graph}(V)$ is approachable, we have (w.p. 1) that $d\left((\bar{v}_n, \bar{q}_n), S\right) \to 0$, implying that

$$\|\bar{v}_n - v_n^*\| \to 0, \quad \|\bar{q}_n - q_n^*\| \to 0,$$

for some sequences $v_n^* \in V(q_n^*)$, $q_n^* \in \Delta(\mathcal{B})$. The uniform continuity of V in the Hausdorff distance $d_{\mathcal{H}}$ then implies that $d_{\mathcal{H}}\left(V(\bar{q}_n), V(q_n^*)\right) \to 0$, hence

$$d(\bar{v}_n, V(\bar{q}_n)) \leq \|\bar{v}_n - v_n^*\| + d_{\mathcal{H}}\left(V(\bar{q}_n), V(q_n^*)\right) \to 0,$$

so that V is attainable by Definition 8. Attainability of V^* now follows since $V(\bar{q}_n) \subseteq V^*(\bar{q}_n)$. ■

We may now formulate a sufficient condition for attainability of a goal function by employing the *dual* condition for approachability of convex sets. Recall that a set-valued map $V : \Delta(\mathcal{B}) \rightrightarrows \mathbb{R}^K$ is called *convex* if its graph Graph(V) is a convex set. The convex hull conv(V) of V is the unique set-valued map whose graph is conv(Graph(V), the convex hull of Graph(V). Similarly, the closed convex hull $\overline{\text{co}}(V)$ of V is the unique set-valued map whose graph is the closure of conv(Graph(V)).

Proposition 10 *Suppose that the set-valued goal function V^* is feasible, in the following sense:*

- *For each mixed action $q \in \Delta(\mathcal{B})$ of the opponent, there exists some mixed action $p = p^*(q)$ of the agent so that $v(p, q) \in V^*(q)$. We refer to $p^*(q)$ as the agent's response to q.*

Denote $V^c = \overline{\text{co}}(V^*)$. Then

(i) The set Graph(V^c) *is approachable by the agent.*

(ii) The set-valued goal function V^c is attainable by the agent (in the sense of Definition 8), and any approachability strategy for Graph(V^c) *is a no-regret strategy for V^c.*

Proof Let us first redefine V^* as its restriction to the compact set $\mathcal{V}_0$, as in Lemma 9. It is clear that this restricted V^* still satisfies the feasibility requirement of the Proposition, and that establishing the claimed attainability property for the restricted version implies the same for the original one.

Let $V^c = \overline{\text{co}}(V^*)$. We first claim that Graph(V^c) is approachable. By the assumed feasibility of V^*, for any q there exists p such that $(v(p, q), q) \in S \triangleq \text{Graph}(V^*)$. Therefore $\overline{\text{co}}(S)$ is a convex D-set, which is approachable by Theorem 3. Now, observe that $\overline{\text{co}}(S) = \overline{\text{co}}(\text{Graph}(V^*)) = \text{Graph}(V^c)$ by definition of V^c.

To conclude that V^c is attainable, it remains to verify that it satisfies the continuity requirement in Lemma 9. Observe that $V^c : \Delta(\mathcal{B}) \rightrightarrows \mathcal{V}_0$ is a convex, compact-valued multifunction whose domain is a polytope. By Mac̆kowiak (2006, Corrolary 2), V^c is lower semi-continuous.[3] Furthermore, since the graph of $V^c = \overline{\text{co}}(V^*)$ is closed by its definition, V^c is upper-semi-continuous (Rockafellar and Wets, 1997, Theorem 5.7). It follows that V^c is a continuous map. Finally, since standard (Kuratowski) continuity and Hausdorff-continuity are equivalent for compact-valued map (*Ibid.*, 4.40(a)), the required continuity property of V_c follows. This concludes the proof. ■

Proposition 10 implies that a feasible and continuous goal function V^* that is *convex* is attainable. When V^* is not convex, as is often the case in the following examples, we need to resort to its convex relaxation $V^c = \overline{\text{co}}(V^*)$. The suitability of V^c as a goal function needs to be examined for each specific problem.

Proposition 10 asserts also that V^c can be attained by any approachability algorithm applied to the convex set $S = \text{Graph}(V^c)$. However, a projection onto that set as required in the standard approachability algorithms may be hard to compute. This is especially true when V^* itself is non-convex, so that V^c is not explicitly specified. In such cases, the response-based approachability algorithm proposed in this paper offers a convenient alternative, as it only requires to compute at each stage a response $p^*(q)$ of the agent to a mixed action q of the opponent, which is inherent in the definition of V^*.

The resulting generalized no-regret algorithm is presented in Algorithm 2. It is merely an application of Algorithm 1 to the problem of approaching $S = \text{Graph}(V^c)$, with augmented reward vectors $r(p, q) = (u(p, q), q)$.

We next specialize the discussion to certain concrete problems of interest.

5.2.1 Constrained Regret Minimization

The following constrained regret minimization problem was introduced in Mannor et al. (2009). Consider the repeated game model as before, where we are given a scalar reward (or utility) function $u : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ and a vector-valued cost function $c : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^s$. We are also given a closed and convex set $\Gamma \subseteq \mathbb{R}^s$, the constraint set, which specifies the allowed values for the long-term average cost. A specific case is that of upper bounds on each cost component, that is $\Gamma = \{c \in \mathbb{R}^s : \ c_i \leq \gamma_i, \ i = 1, ..., s\}$ for some given vector $\gamma \in \mathbb{R}^s$. The constraint set is assumed to be *feasible* (or *non-excludable*), in the sense that for every $q \in \Delta(\mathcal{B})$, there exists $p \in \Delta(\mathcal{A})$ such that $c(p, q) \in \Gamma$.

Let $\bar{U}_n \triangleq n^{-1} \sum_{k=1}^{n} u_k$ and $\bar{C}_n \triangleq n^{-1} \sum_{k=1}^{n} c_k$ denote, respectively, the average reward and cost by stage n. The agent is required to satisfy the cost constraints, in the sense that $\lim_{n \to \infty} d(\bar{C}_n, \Gamma) = 0$ must hold, irrespectively of the opponent's play. Subject to these constraints, the agent wishes to maximize its average reward $\bar{U}_n$.

We note that a concrete learning application for the constrained regret minimization problem was proposed in Bernstein et al. (2010). There, we considered the on-line problem of merging the output of multiple binary classifiers, with the goal of maximizing the true-positive rate, while keeping the false-positive rate under a given threshold $0 < \gamma < 1$. As shown in that paper, this may be formulated as a constrained regret minimization problem.

[3] This is a generalization of the Gale-Klee-Rockfellar Theorem from convex analysis to set-valued maps. The point is of course continuity at the boundary points.

```
Algorithm 2 Generalized No-Regret Algorithm
Input: The reward function $v : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^K$; a set-valued goal function $V^* : \Delta(\mathcal{B}) \rightrightarrows \mathbb{R}^K$;
and for each $q \in \Delta(\mathcal{B})$, a mixed action (or actions) $p \in \Delta(\mathcal{A})$ such that $v(p, q) \in V^*(q)$.
Initialization: At step $n = 1$, apply an arbitrary mixed action $p_1$, and choose arbitrary
values $v_1^* \in \mathbb{R}^K$, $q_1^* \in \Delta(\mathcal{B})$.
At step $n = 2, 3, ...$:
    1. Set
           $\lambda_{n-1}^v = \bar{v}_{n-1}^* - \bar{v}_{n-1}, \quad \lambda_{n-1}^q = \bar{q}_{n-1}^* - \bar{q}_{n-1},$
       where
           $(\bar{v}_m^*, \bar{v}_m) = \frac{1}{m} \sum_{k=1}^{m} (v_k^*, v_k), \quad \bar{q}_m^* = \frac{1}{m} \sum_{k=1}^{m} q_k^*, \quad \bar{q}_m = \frac{1}{m} \sum_{k=1}^{m} \mathbb{I}_{\{b_k = \cdot\}},$
       and $v_k = v(p_k, b_k)$ or $v(a_k, b_k)$.
    2. Solve the following zero-sum matrix game:
           $p_n \in \operatorname*{argmax}_{p \in \Delta(\mathcal{A})} \min_{q \in \Delta(\mathcal{B})} \left( \lambda_{n-1}^v \cdot v(p, q) + \lambda_{n-1}^q \cdot q \right),$
           $q_n^* \in \operatorname*{argmin}_{q \in \Delta(\mathcal{B})} \max_{p \in \Delta(\mathcal{A})} \left( \lambda_{n-1}^v \cdot v(p, q) + \lambda_{n-1}^q \cdot q \right).$
    3. Draw an action $a_n$ randomly from $p_n$.
    4. Pick $p_n^* \in \Delta(\mathcal{A})$ such that $v\left(p_n^*, q_n^*\right) \in V^*(q_n^*)$, and set $v_n^* = v(p_n^*, q_n^*)$.
```

A natural extension of the best-reward-in-hindsight $u^*(q)$ in (14) to the constrained setting is given by

$$u_\Gamma^*(q) \triangleq \max_{p \in \Delta(\mathcal{A})} \left\{ u(p, q) \ : \ c(p, q) \in \Gamma \right\}. \tag{16}$$

We can now define the target set of the pairs $v = (u, c) \in \mathbb{R}^{1+s}$ in terms of $u_\Gamma^*(q)$ and Γ:

$$V^*(q) \triangleq \left\{ v = (u, c) \in \mathbb{R}^{1+s} : u \geq u_\Gamma^*(q), c \in \Gamma \right\}.$$

Note that $u_\Gamma^*(q)$ is *not* convex in general, and consequently $V^*(q)$ is not convex as well. Indeed, it was shown in Mannor et al. (2009) that $V^*(q)$ is not attainable in general. The closed convex hull of $V^*(q)$ may be written as

$$V^c(q) = \left\{ (u, c) \in \mathbb{R}^{s+1} : \ u \geq \overline{\text{conv}} \left(u_\Gamma^* \right)(q), \ c \in \Gamma \right\}, \tag{17}$$

where the real-valued function $\overline{\text{conv}}\left(u_\Gamma^*\right)$ is the closure of the lower convex hull of u_Γ^* over $\Delta(\mathcal{A})$.

Two algorithms were proposed in Mannor et al. (2009) for attaining $V^c(q)$. The first is a standard (Blackwell) approachability algorithm for $S = \{(v, q) \ : \ v \in V^c(q)\}$, which requires the demanding computation of S and the projection directions to S. The second algorithm employs a best-response to calibrated forecasts of the opponent's mixed actions. As mentioned in the introduction, obtaining these forecasts is computationally hard. In contrast, our algorithm mainly requires the computation of the response $p^*(q)$ by solving the maximization problem in (16), which is a convex program. This further reduces to a linear program when the constraints are linear.

Specializing Proposition 10 to this case, we obtain the following result.

Corollary 11 *Consider Algorithm 2 applied to the present model. Thus, the response p_n^* to q_n^* is chosen as any maximizing action in (16) with $q = q_n^*$, and the target point is set to $v_n^* = (u(p_n^*, q_n^*), c(p_n^*, q_n^*))$. Then the goal function V^c is attainable in the sense of Definition 8, which implies that*

$$\liminf_{n\to\infty} \left(\bar{U}_n - \overline{\text{conv}}\,(u_\Gamma^*)\,(\bar{q}_n)\right) \geq 0, \quad \textit{and} \quad \lim_{n\to\infty} d\left(\bar{C}_n, \Gamma\right) = 0 \quad (a.s.)$$

for any strategy of the opponent.

We further note that $V^c(q)$ is unbounded in the direction of its first coordinate u, so that the variant of the algorithm presented in Subsection 4.3 can be applied. In this case, the first coordinate of the steering direction λ_n can be set to zero in $\tilde{\lambda}_n$ whenever it is negative. This corresponds to $\bar{u}_{n-1} \geq \bar{u}_{n-1}^*$, thereby avoiding an unnecessary reduction in $\bar{u}_{n-1}$. Similarly, for a component-wise constraint set of the form $\{c_i \leq \gamma_i\}$, the c_i-coordinate of λ_n may be nullified whenever $[\bar{c}_{n-1}]_i \leq [\bar{c}_{n-1}^*]_i$. The results of Corollary 11 are maintained of course.

5.2.2 Reward-to-Cost Maximization

Consider the repeated game model as before, where the goal of the agent is to maximize the ratio $\bar{U}_n/\bar{C}_n$. Here, $\bar{U}_n$ is, as before, the average of a scalar reward function $u(a,b)$ and $\bar{C}_n$ is the average of a *scalar* and *positive* cost function $c(a,b)$. This problem is mathematically equivalent to regret minimization in repeated games with variable stage duration considered in Mannor and Shimkin (2008) (MS08 for short; in that paper, the cost was specifically taken as the stage duration). Moreover, it can be seen that this problem is a particular case of the global cost function model presented below. However, a direct application of Proposition 10 does not yield a meaningful result in this specific case. We therefore resort to specific analysis which relies on additional properties of our response-based approachability algorithm. This yields a similar bound to that of Proposition 14(ii) below, but without the requirement that G be convex.

Similar bounds to the ones established below were obtained in MS08. The algorithm there was based on playing a best-response to calibrated forecasts of the opponent's mixed actions. The present formulation therefore offers an alternative algorithm which is considerably less demanding computationally.

Denote

$$\rho(a,q) \triangleq \frac{u(a,q)}{c(a,q)}, \quad \rho(p,q) \triangleq \frac{u(p,q)}{c(p,q)}.$$

and let

$$\text{val}(\rho) \triangleq \max_{p\in\Delta(\mathcal{A})} \min_{q\in\Delta(\mathcal{B})} \rho(p,q) = \min_{q\in\Delta(\mathcal{B})} \max_{p\in\Delta(\mathcal{A})} \rho(p,q)$$

(the last equality is proved in MS08; note that $\rho(p,q)$ is not generally concave-convex). As further shown in MS08, $\mathrm{val}(\rho)$ is the value of the zero-sum repeated game with payoffs $\bar{U}_n/\bar{C}_n$, hence serves as a security level for the agent. A natural goal for the agent would be to improve on $\mathrm{val}(\rho)$ whenever the opponent's actions deviate (in terms of their empirical mean) from the minimax optimal strategy.

We next propose an attainable goal function that satisfies this requirement. To that end, let

$$\rho^*(q) \triangleq \max_{p\in\Delta(A)} \rho(p,q)$$

denote the *best ratio-in-hindsight.* Let us apply Algorithm 2, with $v = (u,c)$, and the vector-valued goal function

$$V^*(q) = \left\{ v = (u,c) : \ \frac{u}{c} \geq \rho^*(q) \right\} \tag{18}$$

(observe that $\rho^*(q)$ and $V^*(q)$ are non-convex functions in general). The agent's response is given by any mixed action

$$p^*(q) \in P^*(q) \triangleq \operatorname*{argmax}_{p\in\Delta(\mathcal{A})} \rho(p,q).$$

It is readily verified that the maximum can always be obtained here in pure actions (MS08; see also the proof of Prop. 13 below). Hence, computing the response is trivial in this case.

Denote

$$A^*(q) \triangleq \operatorname*{argmax}_{a\in\mathcal{A}} \rho(a,q),$$

and define the following relaxation of $\rho^*(q)$:

$$\begin{aligned}\rho_1(q) \triangleq \inf &\left\{ \frac{\sum_{j=1}^J u(a_j,q_j)}{\sum_{j=1}^J c(a_j,q_j)} \ : \ J \geq 1,\, q_j \in \Delta(\mathcal{B}),\, \frac{1}{J}\sum_{j=1}^J q_j = q,\, a_j \in A^*(q_j) \right\} \\ &\leq \rho^*(q).\end{aligned} \tag{19}$$

We will show below that ρ_1 is attainable by applying Algorithm 2 to this problem. First, however, we show that ρ_1 never falls below the security level $\mathrm{val}(\rho)$, and is strictly better in typical cases.

Lemma 12

(i) $\rho_1(q) \geq \mathrm{val}(\rho)$ *for all* $q \in \Delta(\mathcal{B})$.

(ii) $\rho_1(q) > \mathrm{val}(\rho)$ *whenever* $\rho^*(q) > \mathrm{val}(\rho)$.

(iii) $\rho_1(q) = \rho^*(q)$ *for the* q*'s that represent pure actions.*

(iv) $\rho_1(q)$ *is a continuous function of* q.

Proof To prove this Lemma, we first derive a more convenient expression for $\rho_1(q)$. For $a \in \mathcal{A}$, let

$$Q_a \triangleq \{q \in \Delta(\mathcal{B}) \,:\, a \in A^*(q)\}$$

denote the (closed) set of mixed actions to which a is a best-response action. Observe that for given J, $q_1, ..., q_J$ and $a_j \in A^*(q_j)$, we have

$$\frac{\sum_{j=1}^{J} u(a_j, q_j)}{\sum_{j=1}^{J} c(a_j, q_j)} = \frac{\sum_{a \in \mathcal{A}} N_a u(a, \bar{q}_a)}{\sum_{a \in \mathcal{A}} N_a c(a, \bar{q}_a)},$$

where

$$N_a = \sum_{j=1}^{J} \mathbb{I}\{a_j = a\}, \quad \bar{q}_a = \frac{1}{N_a} \sum_{j=1}^{J} \mathbb{I}\{a_j = a\} q_j \,.$$

Note that $\bar{q}_a \in \text{conv}(Q_a)$ as it is a convex combination of $q_j \in Q_a$. Therefore, the definition in (19) is equivalent to

$$\rho_1(q) = \min \left\{ \frac{\sum_{a \in \mathcal{A}} \alpha_a u(a, q_a)}{\sum_{a \in \mathcal{A}} \alpha_a c(a, q_a)} \,:\, \alpha \in \Delta(\mathcal{A}), q_a \in \text{conv}(Q_a), \sum_{a \in \mathcal{A}} \alpha_a q_a = q \right\}. \tag{20}$$

Now, this is exactly the definition of the so-called *calibration envelope* in Mannor and Shimkin (2008), and the claims of the lemma follow by Lemma 6.1 and Proposition 6.4 there. ■

It may be seen that $\rho_1(q)$ does not fall below the security level val(q), and is strictly above it when q is not a minimax action with respect to $\rho(p, q)$. Furthermore, at the vertices vertices of $\Delta(\mathcal{B})$, it actually coincides with the best ratio-in-hindsight $\rho^*(q)$.

We proceed to the following result that proves the attainability of $\rho_1(q)$.

Proposition 13 *Consider Algorithm 2 applied to the present model, with the goal function V^* defined in* (18)*. Thus, the agent's response q_n^* is chosen as any action $p_n^* \in P^*(q_n^*)$, and the target point is set to $v_n^* = (u(p_n^*, q_n^*), c(p_n^*, q_n^*))$. Then,*

$$\liminf_{n \to \infty} \left(\frac{\bar{U}_n}{\bar{C}_n} - \rho_1(\bar{q}_n) \right) \geq 0 \quad (a.s.)$$

for any strategy of the opponent.

Proof Algorithm 2 guarantees that, with probability 1,

$$\|\bar{q}_n - \bar{q}_n^*\| \to 0, \tag{21}$$

$$\left| \bar{U}_n - \frac{1}{n} \sum_{k=1}^{n} u(p_k^*, q_k^*) \right| \to 0, \quad \left| \bar{C}_n - \frac{1}{n} \sum_{k=1}^{n} c(p_k^*, q_k^*) \right| \to 0; \tag{22}$$

see Theorem 4 and recall the asymptotic equivalence of expected and actual averages. Noting that the cost c is positive and bounded away from zero, (22) implies that

$$\lim_{n \to \infty} \left| \frac{\bar{U}_n}{\bar{C}_n} - \frac{\sum_{k=1}^{n} r(p_k^*, q_k^*)}{\sum_{k=1}^{n} c(p_k^*, q_k^*)} \right| = 0. \tag{23}$$

Let

$$\rho_2(q) \triangleq \inf\left\{\frac{\sum_{j=1}^J u(p_j, q_j)}{\sum_{j=1}^J c(p_j, q_j)} \,:\, J \geq 1,\, q_j \in \Delta(\mathcal{B}),\, \frac{1}{J}\sum_{j=1}^J q_j = q,\, p_j \in P^*(q_j)\right\}. \tag{24}$$

Clearly,

$$\frac{\sum_{k=1}^n r(p_k^*, q_k^*)}{\sum_{k=1}^n c(p_k^*, q_k^*)} \geq \rho_2(\bar{q}_n^*). \tag{25}$$

Furthermore, we verify below that the infimum in (24) is obtained in pure actions $a_j \in A^*(q_j)$, implying that

$$\rho_2(q) = \rho_1(q). \tag{26}$$

Indeed, note that the inequality

$$\frac{\sum_{j=1}^J u(p_j, q_j)}{\sum_{j=1}^J c(p_j, q_j)} \leq K$$

is equivalent to

$$\sum_{j=1}^J u(p_j, q_j) - K \sum_{j=1}^J c(p_j, q_j) \leq 0.$$

Now, consider minimizing the left-hand-side over $p_j \in P^*(q_j)$. Due to the linearity in p_j and the fact that $P^*(q_j)$ is just the mixture of actions in $A^*(q_j)$, the optimal actions are pure (that is, in $A^*(q_j)$).

Combining (23), (25), and (26), we obtain that

$$\liminf_{n\to\infty}\left(\frac{\bar{U}_n}{\bar{C}_n} - \rho_1(\bar{q}_n^*)\right) \geq 0.$$

The proof is concluded by applying (21) and the continuity (hence, uniform continuity) of ρ_1 (see Lemma 12). ■

We finally note that the algorithm variant from Subsection 4.3 can be applied here as well. Specifically, observe that the goal function V^* in (18) is unbounded in the u coordinate, and negatively unbounded in the c coordinate. Therefore, the u-coordinate of λ_n can be set to zero whenever $\bar{u}_{n-1} \geq \bar{u}_{n-1}^*$, while the c-coordinate of λ_n may be nullified whenever $\bar{c}_{n-1} \leq \bar{c}_{n-1}^*$.

5.2.3 Global Cost Functions

The following problem of regret minimization with global cost functions was introduced in Even-Dar et al. (2009). (A similar problem was recently addressed in Azar et al. (2014), using a relaxed regret criterion over sub-intervals.) Suppose that the goal of the agent is to minimize a general (i.e., non-linear) function of the average reward vector $\bar{v}_n$. In particular, we are given a *continuous* function $G : \mathbb{R}^K \to \mathbb{R}$, and the goal is to minimize $G(\bar{v}_n)$. For example, G may be some *norm* of $\bar{v}_n$. We define the best-cost-in-hindsight, given a mixed action q of the opponent, as

$$G^*(q) \triangleq \min_{p \in \Delta(\mathcal{A})} G(v(p,q)), \tag{27}$$

so that the target set may be defined as

$$V^*(q) = \{v \in \mathcal{V}_0 : G(v) \leq G^*(q)\}, \tag{28}$$

where $\mathcal{V}_0 = \mathrm{conv}\{v(a,b) : a \in \mathcal{A}, b \in \mathcal{B}\} \subset \mathbb{R}^K$ is the set of feasible reward vectors. Clearly, the agent's response to q is any mixed action that minimizes $G(v(p,q))$, namely

$$p^*(q) \in \operatorname*{argmin}_{p \in \Delta(\mathcal{A})} G(v(p,q)). \tag{29}$$

By Proposition 10, the closed convex hull $V^c = \overline{\mathrm{co}}(V^*)$ is attainable by the agent, and Algorithm 2 can be used to attain it. Observe that, in addition to solving a zero-sum matrix game, the algorithm requires solving the optimization problem (29). The computational complexity of the latter depends on the cost function G. For example, if G is convex, then (29) is a convex optimization problem. For specific instances, see Even-Dar et al. (2009) and Example 1 below.

The relation between V^c and V^* depends on the convexity properties of G and G^*. In particular, we have the following result (a slight extension of Even-Dar et al. (2009)).

Proposition 14 *For $q \in \Delta(\mathcal{B})$,*

$$V^c(q) \subset \tilde{V}(q) \triangleq \{v \in \mathcal{V}_0 \,:\, \mathrm{conv}(G)(v) \leq \mathrm{conc}(G^*)(q)\}, \tag{30}$$

where $\mathrm{conv}(G)$ is the lower convex hull of G, and $\mathrm{conc}(G^*)$ is the upper concave hull of G^*. Consequently, any no-regret strategy with respect to $V^c = \overline{\mathrm{co}}(V^*)$ guaranties that, for any strategy of the opponent,

$$\limsup_{n\to\infty} \left(\mathrm{conv}(G)(\bar{v}_n) - \mathrm{conc}(G^*)(\bar{q}_n)\right) \leq 0 \quad \textit{(a. s.)}. \tag{31}$$

In particular, if G is a convex function G^* a concave function, then $V^c = V^*$ and V^* itself is attained, namely

$$\limsup_{n\to\infty} \left(G(\bar{v}_n) - G^*(\bar{q}_n)\right) \leq 0 \quad \textit{(a. s.)}.$$

Proof To show (30), recall that the graph of $V^c = \overline{\mathrm{co}}(V^*)$, by its definition, is given by

$$\mathrm{Graph}(V^c) = \overline{\mathrm{co}}(\mathrm{Graph}(V^*)),$$

and, by (28),

$$\mathrm{Graph}(V^*) = \{(v,q) \in \mathcal{V}_0 \times \Delta(\mathcal{B}) \,:\, G(v) \leq G^*(q)\}.$$

Also, for $\tilde{V}$ as defined in (30),

$$\mathrm{Graph}(\tilde{V}) = \{(v,q) \in \mathcal{V}_0 \times \Delta(\mathcal{B}) \,:\, \mathrm{conv}(G)(v) \leq \mathrm{conc}(G^*)(q)\}.$$

It is clear from these expressions that $\mathrm{Graph}(\tilde{V})$ is a convex set that contains $\mathrm{Graph}(V^*)$, hence $\mathrm{conv}(\mathrm{Graph}(V^*)) \subset \mathrm{Graph}(\tilde{V})$. Furthermore, since G is a continuous function by assumption, the G and G^* are continuous functions on compact sets, so that $\mathrm{conv}(G)$ and $\mathrm{conc}(G^*)$ are continuous functions, which implies that $\mathrm{Graph}(\tilde{V})$ is a closed set. Therefore $\overline{\mathrm{co}}(\mathrm{Graph}(V^*)) \subset \mathrm{Graph}(\tilde{V})$, and (30) follows. The other claims in the Proposition now follow directly from Proposition 10. ■

Clearly, if G^* is not concave, the attainable goal function is weaker than the original one. Still, this relaxed goal is meaningful, at least when G is convex. Noting the definition of G^* in (27), if follows that $G^*(q) \leq \max_{q'} \min_p G(v(p, q'))$ for all q, so that

$$\mathrm{conc}(G^*)(q) \leq \max_{q' \in \Delta(\mathcal{B})} \min_{p \in \Delta(\mathcal{A})} G(v(p, q')) \leq \min_{p \in \Delta(\mathcal{A})} \max_{q' \in \Delta(\mathcal{B})} G(v(p, q')) \,. \tag{32}$$

The latter min-max bound is just the security level of the agent in the repeated game, namely the minimal value of $G(\bar{v}_n)$ that can be secured (as $n \to \infty$) by playing a *fixed* (non-adaptive) mixed action q'. Note that the second inequality in Equation (32) will be strict except for special cases where the min-max theorem holds for $G(v(p, q))$ (which is hardly expected if $G^*(q)$ is non-concave).

Convexity of $G(v)$ depends on its definition, and will hold for cases of interest such as norm functions. Concavity of $G^*(q)$, on the other hand, is more demanding and will hold only in special cases. In Section 5.2.2 we already considered a specific instance of this model where $G(v) = -u/c$ is not convex and $G^*(q) = -\max_p\{u(p, q)/c(p, q)\}$ is not concave, hence specific analysis was required to obtain meaningful bounds. Another concrete model was considered in Even-Dar et al. (2009), motivated by load balancing and job scheduling problems. Under appropriate conditions, it was shown there that G is convex, while G^* can be seen to be concave, and the agent's response was computed in closed form. The details can be found in that reference and will not be elaborated here. These properties allow an easy application of Algorithm 2 above to attain V^* itself.

We close this section with a simple example, in which G is convex while G^* is not necessarily concave.

Example 1 (Absolute Value) Let $v : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ be a scalar reward function, and suppose that we wish to minimize the deviation of the average reward $\bar{v}_n$ from a certain preset value, say 0. Define then $G(v) = |v|$, and note that G is a convex function. Now,

$$G^*(q) \triangleq \min_{p \in \Delta(\mathcal{A})} |v(p, q)| = \begin{cases} \min_{a \in \mathcal{A}} v(a, q), & \text{if } \forall a \in \mathcal{A},\ v(a, q) > 0 \\ \min_{a \in \mathcal{A}} (-v(a, q)), & \text{if } \forall a \in \mathcal{A},\ v(a, q) < 0 \\ 0, & \text{otherwise.} \end{cases}$$

The response $p^*(q)$ of the agent is obvious from these relations. We can observe two special cases in this example:

(*i*) The problem reduces to the classical no-regret problem if the rewards $v(a, b)$ all have the same sign (positive or negative), as the absolute value can be removed. Indeed, in this case $G^*(q)$ is concave, as a minimum of linear functions.

(ii) If the set $\{v(a,q), a \in \mathcal{A}\}$ includes elements of opposite signs (0 included) for each q, then $G^* = 0$, and the point $v = 0$ becomes attainable.

In general, however, $|v(p,q)|$ may be a *strictly* convex function of q for a fixed p, and the minimization above need not lead to a concave function. In that case, Proposition 14 implies only the attainability of $\mathrm{conc}(G^*)(q)$.

We note that the computation of $\mathrm{conc}(G^*)$ may be fairly complicated in general, which implies the same for computing the projection onto the associated goal set $S = \{(v,q) : |v| \leq \mathrm{conc}(G^*)(q)\}$. However, these computations are not needed in the response-based approachability algorithm, where the required computation of the agent's response $p^*(q)$ is straightforward.

6. Conclusion

We have introduced in this paper an approachability algorithm that is based on Blackwell's dual, rather than primal, approachability condition. The proposed algorithm and its variants rely directly on the availability of a response function, rather than projection onto the goal set (or related geometric quantities), and are therefore convenient in problems where the latter may be hard to compute. At the same time, the additional computational requirements are generally comparable to those of the standard Blackwell algorithm and its variants.

The proposed algorithms were applied to a class of generalized no-regret problems, that includes as specific cases the constrained no-regret problem and reward-to-cost maximization. The resulting algorithms are apparently the first computationally efficient algorithms in this generalized setting.

In this paper we have focused on a repeated *matrix* game model, where the action sets of the agent and the opponent in the stage game are both finite. It is worth pointing out that the essential results of this paper should apply directly to models with *convex* action sets, say X and Y, provided that the reward vector $r(x,y)$ is bilinear in its arguments. In that case the (observed) actions x and y simply take the place of the mixed actions p and q, leading to similar algorithms and convergence results. Such a continuous-action model is relevant to linear classification and regression problems.

Other extensions of possible interest for the approach of this paper may include stochastic game models, problems of partial monitoring, and nonlinear (concave-convex) reward functions. These are left for future work.

Acknowledgements

Most of this work was done while the first author was a PhD student at the Department of Electrical Engineering, Technion. This research was supported by the Israel Science Foundation grant No. 1319/11. We wish to thank Shie Mannor for useful discussions, and for pointing out the application to regret minimization with global cost functions. We also thank two anonymous reviewers for their useful comments on this manuscript that helped improve the presentation as well as certain technical aspects.

References

J. Abernethy, P. L. Bartlett, and E. Hazan. Blackwell approachability and low-regret learning are equivalent. In *Proceedings of the 24th Conference on Learning Theory (COLT'11)*, pages 27–46, Budapest, Hungary, June 2011.

J.-P. Aubin and H. Frankowska. *Set-Valued Analysis.* Birkhauser, Boston, MA, 1990.

R.J. Aumann and M. Maschler. *Repeated Games with Incomplete Information.* MIT Press, Boston, MA, 1995.

Y. Azar, U. Felge, M. Feldman, and M. Tennenholtz. Sequential decision making with vector outcomes. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science (ITCS'14)*, pages 195–206, January 2014.

A. Bernstein. *Approachability in Dynamic Games: Algorithms, Refinements, and Applications to No-Regret Problems.* PhD thesis, Technion, Haifa, Israel, October 2013.

A. Bernstein and N. Shimkin. Response-based approachability with applications to generalized no-regret problems. October 2013. Preprint, http://arxiv.org/abs/1312.7658.

A. Bernstein, S. Mannor, and N. Shimkin. Online classification with specificity constraints. In *Proceedings of the 23rd Conference on Neural Information Processing Systems (NIPS'10)*, pages 190–198, Vancouver, Canada, December 2010.

A. Bernstein, S. Mannor, and N. Shimkin. Opportunistic approachability and generalized no-regret problems. *Mathematics of Operations Research*, 39(4):1057–1083, 2014. Also in *Proc. COLT 2013.*

D. Blackwell. Controlled random walks. In *Proceedings of the International Congress of Mathematicians*, volume III, pages 335–338, 1954.

D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, New York, NY, 2006.

E. Even-Dar, R. Kleinberg, S. Mannor, and Y. Mansour. Online learning with global cost functions. In *Proceedings of the 22nd Conference on Learning Theory (COLT'09)*, 2009.

D. Foster. A proof of calibration via Blackwell's approachability theorem. *Games and Economic Behavior*, 29:73–78, 1999.

D. Fudenberg and D. K. Levine. *The Theory of Learning in Games.* MIT Press, Boston, MA, 1998.

J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150, 2000.

S. Hart and A. Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98:26–54, 2001.

E. Hazan and S. Kakade. (weak) Calibration is computationally hard. In *Proceeding of the 25th Conference on Learning Theory (COLT'12)*, pages 3.1–3.10, Edinburgh, Scotland, June 2012.

E. Lehrer. Approachability in infinite dimensional spaces. *International Journal of Game Theory*, 31:253–268, 2002.

E. Lehrer and E. Solan. Learning to play partially-specified equilibrium. Manuscript, available online: `http://www.math.tau.ac.il/~lehrer/Papers/LearningPSCE-web.pdf`, 2007.

E. Lehrer and E. Solan. Approachability with bounded memory. *Games and Economic Behavior*, 66(2):995–1004, 2009.

P. Maćkowiak. Some remarks on lower hemicontinuity of convex multivalued mappings. *Economic Theory*, 28(1):227–233, 2006.

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 and N. Shimkin. Regret minimization in repeated matrix games with variable stage duration. *Games and Economic Behavior*, 63(1):227–258, 2008.

S. Mannor, J. N. Tsitsiklis, and J. Y. Yu. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10:569–590, 2009.

S. Mannor, V. Perchet, and G. Stoltz. Approachability in unknown games: Online learning meets multi-objective optimization. In *Proceeding of the 27th Conference on Learning Theory (COLT'14)*, pages 339–355, Barcelona, Spain, May 2014.

J.F. Mertens, S. Sorin, and S. Zamir. *Repeated Games.* CORE Discussion Papers 9420-9422, Universitè Catholique de Louvain, 1994.

V. Perchet. Calibration and internal no-regret with partial monitoring. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT '09)*, Porto, Portugal, October 2009.

V. Perchet. Approachability, regret and calibration: Implications and equivalences. *Journal of Dynamics and Games*, 1:181–254, 2014.

R.T. Rockafellar and R. Wets. *Variational Analysis.* Springer-Verlag, 1997.

A. Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29: 224–243, 1999.

N. Shimkin and A. Shwartz. Guaranteed performance regions in Markovian systems with competing decision makers. *IEEE Transactions on Automatic Control*, 38(1):84–95, 1993.

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

N. Vieille. Weak approachability. *Mathematics of Operations Research*, 17(4):781–791, 1992.

H. P. Young. *Strategic Learning and Its Limits.* Oxford University Press, 2004.