JMLR: Workshop and Conference Proceedings vol (2013) 1–14

Opportunistic Strategies for Generalized No-Regret Problems

Andrey Bernstein ANDREYB@TX.TECHNION.AC.IL
Shie Mannor SHIE@EE.TECHNION.AC.IL
Nahum Shimkin SHIMKIN@EE.TECHNION.AC.IL
Faculty of Electrical Engineering
Technion – Israel Institute of Technology

Abstract

This paper considers a *generalized no-regret problem* with *vector-valued* rewards, defined in terms of a *desired reward set* of the agent. For each mixed action q of the opponent, the agent has a set $R^*(q)$ where the average reward should reside. In addition, the agent has a *response* mixed action p which brings the expected reward under these two actions, $r(p,q)$, to $R^*(q)$. If a strategy of the agent ensures that the average reward converges to $R^*(\bar{q}_n)$, where $\bar{q}_n$ is the empirical distribution of the opponent's actions, for any strategy of the opponent, we say that it is a *no-regret* strategy with respect to $R^*(q)$. When the multifunction $q \mapsto R^*(q)$ is convex, as is the case in the standard no-regret problem, no-regret strategies can be devised. Our main interest in this paper is in cases where this convexity property does not hold. The best that can be guaranteed in general then is the convergence of the average reward to $R^c(\bar{q}_n)$, the *convex hull* of $R^*(\bar{q}_n)$. However, as the game unfolds, it may turn out that the opponent's choices of actions are limited in some way. If these restrictions were known in advance, the agent could possibly ensure convergence of the average reward to some desired subset of $R^c(\bar{q}_n)$, or even approach $R^*(\bar{q}_n)$ itself. We formulate appropriate goals for *opportunistic* no-regret strategies, in the sense that they may exploit such limitations on the opponent's action sequence in an on-line manner, without knowing them beforehand. As the main technical tool, we propose a class of *approachability* algorithms that rely on a *calibrated forecast* of the opponent's actions, which are opportunistic in the above mentioned sense. As an application, we consider the online no-regret problem with average cost constraints, introduced in Mannor, Tsitsiklis, and Yu (2009). We show, in particular, that our algorithm does attain the best-response-in-hindsight for this problem if the opponent's play happens to be stationary, or close to stationary in a certain sense.

Keywords: No-regret algorithms, Blackwell's approachability, Calibrated play

1. Introduction

The notion of no-regret strategies, introduced by Hannan (1957) in the context of repeated matrix games, has played a central role in on-line learning and prediction problems; see, e.g., Cesa-Bianchi and Lugosi (2006) for an overview. The present paper is motivated by a *generalized no-regret problem*, defined as follows. Consider a repeated matrix game between two players, the agent and the opponent. For each pair of simultaneous actions a and z in the one-stage game, a *vector-valued* reward $r(a,z) \in \mathbb{R}^K$ is obtained by the agent. Let $r(p,q) \triangleq \sum_{a,z} p(a)q(z)r(a,z)$ denote the expected reward vector for a pair of mixed actions p

and q. Suppose that for each mixed action q of the opponent, the agent has a *desired reward set* $R^*(q) \subset \mathbb{R}^K$, and at least one mixed action $p = p^*(q)$ that satisfies $r(p,q) \in R^*(q)$. A generalized no-regret strategy for this problem may be defined as strategy (or on-line algorithm) for the agent in the repeated game that ensures $\lim_{n\to\infty} d(\bar{R}_n, R^*(\bar{q}_n)) = 0$ (almost surely) for any strategy of the opponent. Here $\bar{R}_n = n^{-1}\sum_{k=1}^n r(a_k, z_k)$ is the average reward, $\bar{q}_n = n^{-1}\sum_{k=1}^n 1(z_k)$ is the empirical distribution of the opponent's actions, and d is the Euclidean distance. For short, we say that $\bar{R}_n$ *approaches* $R^*(\bar{q}_n)$.

The standard no-regret problem is obtained as a special case for scalar rewards $r(a,z)$ and $R^*(q) = \{r \in \mathbb{R} : r \geq r^*(q)\}$, where $r^*(q) \triangleq \max_p r(p,q)$ is the maximal reward that the agent could obtain against a mixed action q of the opponent. The no-regret strategy proposed in Hannan (1957) essentially relied on perturbed best-response to the empirical frequencies of the opponent's actions. Subsequently, Blackwell (1954) showed that the problem can be formulated and solved as a particular case of his *theory of approachability* (Blackwell, 1956). The proof relies on the convexity of the multifunction $q \mapsto R^*(q)$, which in turn follows from the convexity of $r^*(q)$ in q.

A similar line of reasoning may be pursued for the generalized no-regret problem described above. Our main interest in this paper is in cases where the convexity property of $q \mapsto R^*(q)$ does not hold, as in the following problem of *constrained no-regret* (which is treated in detail in Section 5).

Example 1 Suppose the agent wishes to maximize the average $\bar{U}_n$ of a scalar reward $u(a,z)$, subject to a long-term average cost constraint of the form $\bar{C}_n \leq \gamma + o(1)$, where $\bar{C}_n$ is the n-step average of a (scalar or vector-valued) cost function $c(a,z)$. Let $u^*_\gamma(q) = \max_{p\in\Delta(\mathcal{A})}\{u(p,q) : c(p,q) \leq \gamma\}$ denote the maximal expected reward that the agent can secure against a mixed action $q \in \Delta(\mathcal{Z})$ of the opponent while keeping the constraints below γ. The desired set of the pairs (reward, cost) for this problem is given by $R^*(q) = \{r = (u,c) : u \geq u^*_\gamma(q), c \leq \gamma\}$. Observe that with a non-trivial side constraint, $q \mapsto R^*(q)$ is non-convex due to the non-convexity of $u^*_\gamma(q)$, and hence $R^*(\bar{q}_n)$ cannot be approached in general, as was shown in Mannor et al. (2009). □

Several other generalized no-regret problems can be formulated in the above form, including 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).

When $q \mapsto R^*(q)$ is non-convex, its convex hull $R^c(\cdot)$ is defined as follows.

Definition 1 (Convex Hull of a Multifunction) *The convex hull of a multifunction (or set-valued function) $q \mapsto R(q)$ is the smallest convex multifunction that contains it, namely R^c so that $R(q) \subset R^c(q)$ for each q, and $\alpha R^c(q_1)+(1-\alpha)R^c(q_2) \subset R^c(\alpha q_1+(1-\alpha)q_2)$ for any q_1, q_2 and $\alpha \in (0,1)$, where the first plus sign stands for the Minkowski sum. A smallest one exists by closedness to intersections.*

$R^c(\cdot)$ is still provably approachable, in the sense that there is a strategy for the agent that ensures $\lim_{n\to\infty} d(\bar{R}_n, R^c(\bar{q}_n)) = 0$ *a.s.* This appears to be the best that can be guaranteed in general.

But more can be achieved if the opponent happens to play "regularly" in some sense. For example, suppose that the opponent adopts a stationary strategy, namely repeats the same mixed action $q_n = q_0$ at each stage. In that case we can hope to refine our online strategy so that, by capitalizing on this regularity, it will still approach the original $R^*(q_0)$

(rather than the larger $R^c(q_0)$). Note that we still require that $\bar{R}_n$ approaches $R^c(\bar{q}_n)$ for any strategy of the opponent, and that any regularities in the opponnent's play are not imposed beforehands but rather need to be detected online. We refer to such strategies of the agent as *opportunistic*. A precise definition is given in Section 3.

We next illustrate these ideas, and the key role played by convexity.

Example 2 Consider a scalar reward matrix given by $r(a^1, z^1) = 2$, $r(a^2, z^2) = -2$, and $r(a^1, z^2) = r(a^2, z^1) = 0$, where a^1, a^2 are the actions available to the agent, and z^1, z^2 those of the opponent. Suppose the agent's goal is to have its long-term average reward larger or equal to 1 *in absolute value*, namely $|\bar{R}_n| \geq 1 - o(1)$. The desired set of rewards is hence $R^*(q) = (-\infty, -1] \cup [1, \infty)$ for all q. Now, it is easily seen that for any mixed action $q = (q(z^1), q(z^2))$ of the opponent, the agent has a response p so that $r(p, q) \in R^*(q)$. Thus, if the opponent is restricted a-priori to stationary strategies, the agent can easily devise a (possibly adaptive) strategy that approaches $R^*(\bar{q}_n)$. However, this is clearly not the case in general: for example, the opponent can ensure $\bar{R}_n \to 0$ by playing z^1 whenever $\bar{R}_{n-1} < 0$, and z^2 otherwise. We see that the agent cannot approach $R^*(\bar{q}_n)$ in general, but can hope to do so if the opponent happens to play a stationary strategy. □

Below is the outline of the paper. In Section 2, we review Blackwell's approachability theory, which is our main tool to devise opportunistic strategies. We then turn to treat a general (abstract) approachability problem for *non-convex sets*: (i) In Section 3, we introduce the concept of *opportunistic approachability*, which relies on an accompanying notions of *statistically and empirically restricted opponent*. (ii) Section 4 provides a background on *calibrated forecasts* (Dawid, 1985), presents our calibration-based approachability algorithm, and establishes its opportunistic properties. For the case of empirically restricted opponent, we require the calibrated forecast to be *slowly time-varying* in an appropriate sense, which we establish for a specific forecasting algorithm. Finally, in Section 5, we specialize to the extended no-regret problem, and in particular to constrained no regret. Specifically, we formulate the no-regret problem as approachability of the set $S = \{(r, q) : r \in R^*(q)\}$, which is non-convex and non-approachable in general.

Since Blackwell's original construction, several approachability algorithms and related results have been proposed in the literature (Hart and Mas-Colell, 2001; Spinat, 2002; Shimkin and Shwartz, 1993; Milman, 2006; Lehrer, 2002; Abernethy et al., 2011). The approachability policies discussed in these papers are based on Blackwell's *primal* condition, which is a geometric separation condition with respect to the fixed target set. Therefore, the existing algorithms are *not opportunistic* in the sense we advocate in this paper. The idea of best-response to calibrated forecasts was first introduced in Foster and Vohra (1997) in the context of attaining correlated equilibrium, and was subsequently used in Mannor and Shimkin (2008) and Mannor et al. (2009) in the context of regret minimization. An approachability strategy that is based on calibrated forecasts was first proposed in Perchet (2009); however, the discussion there is limited only to *convex* sets, and hence the opportunistic properties of the algorithm are not analyzed. We note that our calibration-based algorithm, while conceptually simple, is computationally challenging due to the computational complexity of obtaining calibrated forecasts. Given the result of Hazan and Kakade (2012), it is unlikely that there exists a polynomial-time algorithm to compute an exact calibrated forecast when the number of actions available to the opponent is large. The only polynomial algorithms known in the literature are for the case of *binary*

sequences (Mannor et al., 2007). We emphasize that the main goal in this paper is in formulating the concept of opportunistic strategies and in showing that there exist algorithms that fit this concept. Thus, the computational issues and the analysis of the convergence rates are left for future work.

2. Review of the Approachability Problem

In this Section, we present the main tool that is used to devise opportunistic strategies, namely the approachability theory as well as Blackwell's approachability algorithm.

Consider a repeated two-person game 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 $\mathcal{Z}$. At each time instance $n = 1, 2, ...$, the agent selects its action $a_n \in \mathcal{A}$, observes the action $z_n \in \mathcal{Z}$ chosen by the opponent, and obtains a *vector* reward $V_n = v(a_n, z_n) \in \mathbb{R}^\ell$, $\ell \geq 1$, where $v : \mathcal{A} \times \mathcal{Z} \to \mathbb{R}^\ell$ is a given function. The average reward up to time n is $\bar{V}_n = n^{-1} \sum_{k=1}^n V_k$. A *mixed* action of the agent is the probability distribution $p \in \Delta(\mathcal{A})$, where $p(a)$ specifies the probability of choosing action $a \in \mathcal{A}$. Similarly, $q \in \Delta(\mathcal{Z})$ denotes a mixed action of the opponent. Let $\bar{q}_n \in \Delta(\mathcal{Z})$ denote the empirical distribution of the opponent's actions at time n, with $\bar{q}_n(z) \triangleq n^{-1} \sum_{k=1}^n \mathbb{I}\{z_n = z\}$. We slightly abuse notation and let $v(p, q) \triangleq \sum_{a \in \mathcal{A}, z \in \mathcal{Z}} p(a)q(z)v(a, z)$ denote the expected reward under mixed actions $p \in \Delta(\mathcal{A})$ and $q \in \Delta(\mathcal{Z})$; the distinction between $v(a, z)$ and $v(p, q)$ should be clear by their arguments. The notation $v(p, z)$ and $v(a, q)$ is interpreted similarly.

Let $h_{n-1} \triangleq \{a_1, z_1, ..., a_{n-1}, z_{n-1}\} \in (\mathcal{A} \times \mathcal{Z})^{n-1}$ denote the history of the game up to time n. A *strategy* π of the agent is a collection of the decision rules $\pi_n : (\mathcal{A} \times \mathcal{Z})^{n-1} \to \Delta(\mathcal{A})$, $n \geq 1$, where each mapping π_n specifies the mixed action for the agent at time n, based on the observed history: $p_n = \pi_n(h_{n-1})$. The pure action a_n taken by the agent is then selected randomly according to p_n. Similarly, the opponent's strategy is denoted by $\sigma = \{\sigma_n\}_{n=1}^\infty$, with $\sigma_n : (\mathcal{A} \times \mathcal{Z})^{n-1} \to \Delta(\mathcal{Z})$.

Below is the classical definition of an approachable set from Blackwell (1956).

Definition 2 (Approachable Set) *A closed set $S \subseteq \mathbb{R}^\ell$ is approachable by the agent's strategy π if the average reward $\bar{V}_n$ converges to S in the point-to-set Euclidean distance, almost surely for every strategy σ of the opponent*[1]*. The set S is approachable if there exists such a strategy for the agent.*

In what follows, we find it convenient to state all our results in terms of the *expected* average reward, where the expected value is only with respect to the agent's mixed actions: $\bar{v}_n \triangleq n^{-1} \sum_{k=1}^n v(p_k, z_k)$. The stated convergence results are shown to hold *pathwise*, for any possible sequence of the opponent's actions. The corresponding almost sure results for the actual average reward can be easily deduced using martingale convergence theory, by noting that $d\left(\bar{V}_n, S\right) \leq \left\|\bar{V}_n - \bar{v}_n\right\| + d\left(\bar{v}_n, S\right)$. Now, the first term is the norm of the mean of the martingale difference sequence $D_k = v(a_k, z_k) - v(p_k, z_k)$ and can readily be shown to converge to zero at a uniform rate of $O\left(1/\sqrt{n}\right)$; see, e.g., Shiryaev (1995).

Next, we present a formulation of Blackwell's Theorem (Blackwell, 1956) which provides us with conditions for approachability of general and convex sets. To this end, for any

1. Blackwell's original definition requires almost sure convergence at a uniform rate over opponent's strategies. Our algorithms satisfy this definition provided that the convergence of the employed calibrated forecasts is uniform. However, we will not assume it here.

$x \notin S$, let $c(x) \in S$ denote a closest point in S to x. Also, for any $p \in \Delta(\mathcal{A})$ let $T(p) \triangleq \{v(p,q) : \ q \in \Delta(\mathcal{Z})\}$, which equals the convex hull of the points $\{v(p,z)\}_{z\in\mathcal{Z}}$.

Theorem 3

(i) ***Primal Condition and Algorithm.*** *A closed set $S \subseteq \mathbb{R}^\ell$ is called a B-set (where B stands for Blackwell) if for every $x \notin S$ there exists a mixed action $p^* = p^*(x) \in \Delta(\mathcal{A})$ such that the hyperplane through $y = c(x)$ perpendicular to the line segment xy, separates x from $T(p^*)$. Every B-set is approachable, by using at time n the mixed action $p^*(\bar{v}_{n-1})$ whenever $\bar{v}_{n-1} \notin S$. If $\bar{v}_{n-1} \in S$, an arbitrary action can be used.*

(ii) ***Dual Condition.*** *A closed set $S \subseteq \mathbb{R}^\ell$ is called a D-set (where D stands for Dual) if for every $q \in \Delta(\mathcal{Z})$ there exists $p \in \Delta(\mathcal{A})$ such that $v(p,q) \in S$. If a closed set S is approachable then it is a D-set.*

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

Corollary 4 *The convex hull of a D-set is approachable (and is also a B-set).*

Proof The convex hull of a D-set is a convex D-set. The claim follows by part (iii) of Theorem 3. ■

3. Opportunistic Approachability

In this Section, we define the desiderata for an opportunistic approachability algorithm. To that end, we first define appropriate notions of a *statistically* and an *empirically* restricted play of the opponent, as well as the response and goal function for the given target set. The proofs of the results of this and other sections are omitted due to space constraints, and can be found in Bernstein et al. (2013).

We start with the following assumption on the target set S.

Assumption 3.1 *The set S is a D-set. Namely, for every $q \in \Delta(\mathcal{Z})$ there exists a response $p \in \Delta(\mathcal{A})$ such that $v(p,q) \in S$.*

Observe that we do *not* assume that S is a convex set. Consequently, while $\mathrm{conv}(S)$ is approachable by Corollary 4, S itself need not be approachable.

Before making formal definitions, we state the idea of our approach. We propose algorithms that simultaneously achieve the following goals, for any D-set S: (i) The *convex hull* of S is approached, for any strategy of the opponent; (ii) If the *empirical frequencies* of the opponent's pure actions are restricted to a subset of its mixed actions space (in the sense of Definition 6 below), then the algorithm approaches a corresponding strict subset of $\mathrm{conv}(S)$. In particular, if the opponent is stationary, the set S itself is approached.

3.1. Restricted Opponent Play

We first consider the notion of a *statistically restricted play* of the opponent, in the sense that the sequence of its *mixed* actions $\{q_n\}$ is asymptotically restricted to some set $Q \subseteq \Delta(\mathcal{Z})$.

Definition 5 (Statistically Q-Restricted Play) *We say that the play of the opponent is statistically Q-restricted, if there exists a convex subset $Q \subseteq \Delta(\mathcal{Z})$ so that the sequence $\{q_n\}$ of the mixed actions of the opponent satisfies, for the given sample path, $\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} d(q_k, Q) = 0$. Here, $d(q, Q)$ is Euclidean point-to-set distance.*

It should be emphasized that Definition 5 is stated in terms of the *sample path* of the opponent's play, and not in terms of its overall policy.

A weakness of Definition 5 is that the mixed actions of the opponent are not generally revealed (when its strategy is not known), or may be meaningless (e.g., when the opponent is Nature). We therefore define a *weaker* notion of an *empirically restricted play* of the opponent, in terms of the *empirical frequencies* of the opponent's pure actions. To this end, we need to refer to a partition of the time axis into episodes on which these frequencies are computed. We let τ_m denote the length of episode $m = 1, 2, ...$ and $n_M = \sum_{m=1}^{M} \tau_m$ denote the time at the end of episode M. Finally, $\hat{q}_m \in \Delta(\mathcal{Z})$ denotes the empirical distribution of the opponent's actions at episode m, namely $\hat{q}_m(z) = \tau_m^{-1} \sum_{k=n_{m-1}+1}^{n_m} \mathbb{I}\{z_k = z\}$.

Definition 6 (Empirically Q-Restricted Play) *We say that the play of the opponent is empirically Q-restricted with respect to a partition $\{\tau_m\}$, if there exists a convex subset $Q \subseteq \Delta(\mathcal{Z})$ so that, for the given sample path, $\lim_{M\to\infty} n_M^{-1} \sum_{m=1}^{M} \tau_m d(\hat{q}_m, Q) = 0$.*

Our definition of empirically Q-restricted play involves a *general* partition $\{\tau_m\}$ rather than a partition with fixed lengths $\tau_m \equiv \tau$. The main reason behind this general definition is the fact that we would like to cover the case of *statistically* stationary sequences.

Lemma 7 *Suppose that, almost surely, the play of the opponent is statistically Q-restricted in the sense of Definition 5. Then the requirement of Definition 6 is satisfied with respect to any partition $\{\tau_m\}$ with super-logarithmically increasing episode lengths, namely with $\lim_{m\to\infty}(\log_a m/\tau_m) = 0$ for all $a > 0$.*

A given sequence of actions may satisfy Definition 6 under different partitions, as the following example demonstrates.

Example 3 Consider binary sequences of actions, and let $Q = \{(0.5, 0.5)\}$, a singleton. The sequence 0101... is empirically Q-restricted with respect to *any* partition with fixed *even* episode lengths, or with any strictly increasing episodes lengths. On the other hand, consider the sequence 01001100001111.... The empirical frequencies of this sequence does *not* converge to Q, but it is empirically Q-restricted with respect to a partition with *exponentially* increasing lengths $\tau_m = 2^m$. However, if we choose any partition with sub-exponentially increasing lengths, Definition 6 will not be satisfied. □

The next lemma shows that Definition 6 requires more than just convergence of $\bar{q}_n$ to Q. This requirement is motivated by the fact that we are interested in sequences for which Definition 6 can be satisfied on *sub-exponentially* increasing partitions.

Lemma 8 *If Definition 6 is satisfied with respect to a partition with sub-exponentially increasing episode lengths (namely with $\lim_{m\to\infty}(\tau_m/a^m) = 0$ for all $a > 0$), then $\bar{q}_n$ converges to Q. However, the converse is not true. Namely, there exist a sequence of actions so that $\bar{q}_n$ converges to Q, but there is no partition with sub-exponentially increasing block lengths so that Definition 6 is satisfied with respect to it.*

3.2. Response and Goal Functions

By definition of a D-set, one can define the following.

Definition 9 (Response Function) *A function $p^* : \Delta(\mathcal{Z}) \to \Delta(\mathcal{A})$ is a response function relative to the target set S if for each $q \in \Delta(\mathcal{Z})$, $v\left(p^*\left(q\right), q\right) \in S$.*

Below we demonstrate that p^* cannot be continuous in general. A *piecewise continuous* selection is always possible, as shown in Bernstein et al. (2013), Appendix A. We note that the smoother is p^*, the tighter would be the approachability guarantees (see Definition 11 below, and the discussion following it). However, we do not impose any additional assumptions on p^* in the following.

Example 4 (Example 2 continued) Recall the approachability problem with the target D-set $S = (-\infty, -1] \cup [1, \infty)$ presented in Example 2. Let $p \triangleq p(a^1)$ and $q \triangleq q(z^1)$, respectively, denote the probability of choosing action a^1 by the agent and z^1 by the opponent. Observe that for $q < 0.5$ and $p \leq 0.5 - q$, we have that $v(p, q) \leq -1$ and therefore $v(p, q) \in S$. Similarly, for $q > 0.5$ and $p \geq 1.5 - q$, we have that $v(p, q) \geq 1$, implying that $v(p, q) \in S$. We thus can define a response function as follows:

$$p^*(q) = \begin{cases} 0, & \text{for } q \leq 0.5 \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

While other selections possible, all of them will have a discontinuity at $q = 0.5$. □

The actual choice of p^* is problem dependent. In Section 5, we will see an example where p^* is naturally defined as a best-response map. In general, we make the following.

Assumption 3.2 *Let p^* be a response function relative to the given target set S, which we fix in the following.*

We note that Assumption 3.2 implies Assumption 3.1 by the definition of p^*. Hence, throughout, we suppose that Assumption 3.2 holds, and we do not refer to the target set S explicitly.

The specified response function p^* leads naturally to our next definition.

Definition 10 (Goal Function) *The goal function $v^* : \Delta(\mathcal{Z}) \to S$ is defined as $v^*(q) = v\left(p^*\left(q\right), q\right)$ for any $q \in \Delta(\mathcal{Z})$.*

3.3. Opportunistic Strategies

When the play of the opponent is Q-restricted, we essentially require the average reward to converge to $V(Q) = \text{conv}\{v^*(q) : q \in Q\}$, the convex hull of the image of Q under v^* (see Figure 1). Due to possible discontinuities in v^*, we need to slightly expand that definition.

Definition 11 (Closed Convex Image) *The closed convex image of a set $Q \subseteq \Delta(\mathcal{Z})$ under the goal function v^* is defined as $V^+(Q) \triangleq \bigcap_{\epsilon > 0} \text{conv}\left\{v^*(q) : \ d(q, Q) \leq \epsilon\right\}$.*

The set $V^+(Q)$ contains the convex hull of $v^*(q)$, $q \in Q$, *together with possible jumps in v^* on the boundary of Q*. Note that $V^+(Q) \subset \text{conv}(S)$, as $v^*(q) \in S$ by its definition.

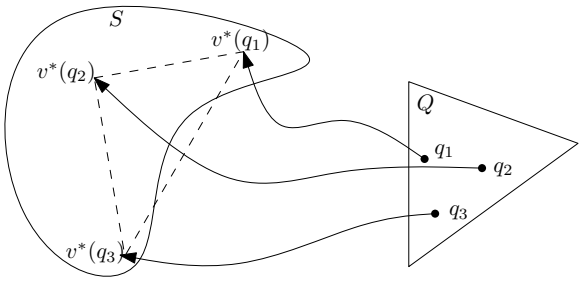

Figure 1: An illustration of the restriction set Q and $V(Q)$.

Example 5 (Example 4 continued) Consider the response function defined in (1). The corresponding goal function is given by $v^*(q) = -2(1-q)$, $q \leq 0.5$, and $v^*(q) = 2q$ otherwise. The closed convex image of singeltons is then $V^+(\{q\}) = v^*(q)$ for $q \neq 0.5$, while $V^+(\{q\}) = \text{conv}(\{-1,1\}) = [-1,1]$ for $q = 0.5$. Observe that the discontinuity of v^* at $q = 0.5$ is expressed by the fact that $V^+(\{q\})$ is the "jump interval" $[-1,1]$. □

We are now ready to define opportunistic approachability strategies.

Definition 12 (Statistically Opportunistic Approachability) *A strategy π is statistically opportunistic for a given target D-set S if it holds that $\lim_{n\to\infty} d(\bar{v}_n, V^+(Q)) = 0$ whenever the play of the opponent is statistically Q-restricted (Definition 5) for some set $Q \subseteq \Delta(\mathcal{Z})$.*

Definition 13 (Empirically Opportunistic Approachability) *A strategy π is empirically opportunistic for a given target D-set S w.r.t. a partition $\{\tau_m\}$ if $\lim_{n\to\infty} d(\bar{v}_n, V^+(Q)) = 0$ whenever the play of the opponent is empirically Q-restricted w.r.t. $\{\tau_m\}$ (Definition 6) for some set $Q \subseteq \Delta(\mathcal{Z})$.*

It should be emphasized that the definitions of opportunistic approachability strategies are based on the *sample path* properties of the opponent's play, and the related convergence results are required to hold *without knowing the restriction set* Q beforehand. Also, note that these Definitions include the standard definition of approachability as a special case, by setting $Q = \Delta(\mathcal{Z})$. Finally, observe that if a strategy is empirically opportunistic with respect to some partition with super-logarithmically increasing lengths, it is also statistically opportunistic (as follows from Lemma 7). But the converse is not necessarily true.

4. Calibration-Based Approachability

In this Section, we present the calibration-based algorithm that is the subject of this paper, and show that it is an opportunistic approachability algorithm in the sense of Definitions 12 and 13.

4.1. Calibrated Forecasts

A *forecaster* specifies at each time step n a probabilistic forecast $y_n \in \Delta(\mathcal{Z})$ of the opponent's action z_n, based on the history of observed actions and previous forecasts. The

forecaster's policy may be *randomized*, i.e. it specifies a probability measure η_n over $\Delta(\mathcal{Z})$. In this case, the forecast $y_n \in \Delta(\mathcal{Z})$ is drawn at random according to η_n. The following is a standard definition of a calibrated forecaster (see, e.g., Foster and Vohra (1997)).

Definition 14 (Calibrated Forecaster) *A forecaster is calibrated if for every Borel measurable set $Q \subseteq \Delta(\mathcal{Z})$ and every strategy of the opponent, it holds that*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}\{y_k \in Q\} (1(z_k) - y_k) = 0, \quad a.s., \tag{2}$$

where $1(z)$ is the probability vector in $\Delta(\mathcal{Z})$ concentrated on z.

A deterministic forecaster can not be calibrated for all possible sequences (Dawid, 1985). However, if the forecaster is allowed to randomize, calibration is possible (see the overview in Cesa-Bianchi and Lugosi (2006), as well as Mannor et al. (2007) and Foster et al. (2011)). The common approach is to use a finite ϵ-grid over $\Delta(\mathcal{Z})$ which is gradually refined in order to fulfil the requirement of Definition 14. To achieve ϵ-calibration, the algorithms usually process the entire grid for each prediction. The only computationally efficient algorithms known in the literature are for the case of *binary* sequences (Mannor et al., 2007). Moreover, it was shown in Hazan and Kakade (2012) that the existence of a general computationally efficient calibrated forecaster would imply the existence of an efficient algorithm for computing approximate Nash equilibria, thus implying the unlikely conclusion that every problem in PPAD (the class of problems that are polynomial time reducible to the problem of computing Nash equilibrium in a two player game) is solvable in polynomial time.

4.2. The Calibrated Approachability Algorithm and Main Result

Recall that p^* denotes a response function as per Definition 9. Our algorithm is *conceptually* very simple – at each time n use the mixed action p_n which is specified by

$$p_n = p^*(y_n), \tag{3}$$

where y_n is the calibrated forecast at time n.

The following Theorem summarizes the opportunistic approachability properties of this algorithm. The detailed analysis of the algorithm that establishes these properties appears in Bernstein et al. (2013), Appendix B.

Theorem 15 *Suppose that Assumption 3.2 holds. Then:*

(i) The Calibrated Approachability Algorithm specified by (3) is statistically opportunistic in the sense of Definition 12.

(ii) Suppose that the probability distribution η_n of the employed calibrated forecast is changing slowly. Namely, there exists $n_0 < \infty$ such that for all $n \geq n_0$,

$$\|\eta_n - \eta_{n-1}\|_\infty \leq \frac{C}{n^\xi}, \tag{4}$$

for some $\xi > 0$ and $C < \infty$. Then, the Calibrated Approachability Algorithm is empirically opportunistic in the sense of Definition 13, under either

(1) Bounded episodes lengths $\tau_m \leq \bar{\tau} < \infty$*, or*

(2) Growing episodes lengths $\tau_m = O(m^\nu)$ *with* $\nu > 0$*, under the condition that* $\xi > \nu/(\nu+1)$.

Proof Idea. The proof of general approachability is based on showing that the average reward converges to $n^{-1}\sum_{k=1}^{n} v^*(y_k)$ which is in $\mathrm{conv}(S)$. As a consequence, part (i) follows by showing that whenever the play of the opponent is statistically Q-restricted, so does the sequence of calibrated forecasts. Thus, $n^{-1}\sum_{k=1}^{n} v^*(y_k) \to V^+(Q)$. (See Bernstein et al. (2013), Appendix B.2 for the full proof.)

The empirical opportunistic approachability result (part (ii)) is obtained by showing that the calibration property (2) implies a similar property in terms of the *empirical frequencies* of the actions provided that the distributions of calibrated forecasts are changing slowly in the sense of (4). (See Bernstein et al. (2013), Appendix B.3 for the full proof.) ■

We illustrate the importance of requirement (4) using the setting of Example 4, where the goal is to approach the non-convex set $S = (-\infty, -1] \cup [1, \infty)$. Suppose that the opponent's actions are $0, 0, 1, 0, 0, 1, ...$, implying that $\bar{q}_n \to q_0 = (2/3, 1/3)$. An opportunistic approachability algorithm should ideally converge in this case to $V^+(\{q_0\})$ (see Definition 13). Indeed, the fixed forecaster $y_n = (2/3, 1/3)$ is calibrated, and the Calibrated Approachability Algorithm that uses this forecaster will approach $V^+(\{q_0\}) = v(p^*(q_0), q_0) = v((1,0),(2/3,1/3)) = 4/3$, where the first equality follows since q_0 is a continuity point of the response function p^* defined in (1). Now since $4/3 \in S$, the algorithm will approach S. However, consider a *perfect* forecaster that predicts $y_n = 1(z_n)$. If the Calibrated Approachability Algorithm uses this forecaster, it approaches $\frac{2}{3}v^*((1,0)) + \frac{1}{3}v^*((0,1)) = \frac{2}{3}v((1,0),(1,0)) + \frac{1}{3}v((0,1),(0,1)) = 2/3$, which is *not* in S. Hence, in this case, only convergence to $\mathrm{conv}(S)$ is guaranteed.

This example illustrates the fact that a perfect forecaster is bad for the purpose of empirically opportunistic approachability. In fact, we would prefer a fixed forecaster, or more generally, a slowly time-varying forecaster, as captured by condition (4). In Bernstein et al. (2013), Appendix C, we show that there exists a specific calibration algorithm that satisfies this property (see Corollary 35 there). We also conjecture that *all* calibrated forecasters possess a certain smoothness property. Indeed, since a calibrated forecaster should be calibrated for *all* possible sequences, it is reasonable to assume that it will do some kind of averaging based on the history of the observed actions, and thus will be smooth in some sense (but maybe not in the *uniform* sense of (4)).

5. Generalized No-Regret and Constrained Regret Minimization

Recall the generalized no-regret problem presented in the Introduction. The repeated game model is the same as above, except that the vector-valued reward function v is replaced by r. In addition, suppose that for each mixed action q of the opponent, the agent has a desired reward set $R^*(q) \subset \mathbb{R}^K$, and at least one mixed action $p = p^*(q)$ that satisfies $r(p,q) \in R^*(q)$. If a strategy of the agent ensures that $\lim_{n\to\infty} d(\bar{R}_n, R^*(\bar{q}_n)) = 0$ (a.s.) for any strategy of the opponent, we say that it is a *no-regret* or *regret minimizing* strategy with respect to $R^*(q)$. Observe that the existence of no-regret strategies is equivalent to

the approachability of the set $S = \{v = (r, q) : r \in R^*(q)\}$. As was mentioned, our main interest is in non-convex maps $q \mapsto R^*(q)$. Hence, the target set S is a non-convex D-set by construction, and the opportunistic approachability theory developed in this paper applies.

We next specialize to the problem of regret minimization subject to average cost constraints (Mannor et al., 2009). Consider the same repeated game model, except that we are given two functions: (i) a scalar reward (or utility) function $u : \mathcal{A} \times \mathcal{Z} \to \mathbb{R}$, and (ii) a vector-valued cost function $c : \mathcal{A} \times \mathcal{Z} \to \mathbb{R}^s$. In addition, we are given a convex set $\Gamma \subseteq \mathbb{R}^s$, the constraint set, defining the allowed values for the long-term average cost (see below). The typical case is that of linear constraints, that is $\Gamma = \{c \in \mathbb{R}^s : \ c_i \leq \gamma_i, \ i = 1, ..., s\}$ for some vector $\gamma \in \mathbb{R}^s$. The constraint set is assumed to be *not excludable*, in the sense that for every $q \in \Delta(\mathcal{Z})$, there exists $p \in \Delta(\mathcal{A})$, such that $c(p, q) \in \Gamma$.

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, the on-line problem of merging the output of multiple binary classifiers was considered, with the goal of maximizing the true-positive rate, while keeping the false-positive rate under a given threshold $0 < \gamma < 1$. As shown there, this problem may be formulated as a constrained regret minimization problem.

Suppose the agent knew in advance that the empirical distribution $\bar{q}_n$ equals to q. He could then maximize its expected average reward subject to the constraints by always choosing the mixed action p that solved the following program:

$$u^*_\Gamma(q) \triangleq \max_{p\in\Delta(\mathcal{A})} \{u(p, q) \ : \ c(p, q) \in \Gamma\}. \tag{5}$$

We consider $u^*_\Gamma(q)$ as the the *best-reward-in-hindsight* for the constrained problem. The goal of the agent would be then to attain u^*_Γ in the following sense.

Definition 16 (Constrained no-regret) *A strategy of the agent π is a constrained no-regret strategy with respect to a function u^*_Γ if: (i) $\limsup_{n\to\infty} (u^*_\Gamma(\bar{q}_n) - \bar{u}_n) \leq 0$; and (ii) $\lim_{n\to\infty} d(\bar{c}_n, \Gamma) = 0$ both hold almost surely, for every strategy of the opponent. If such a strategy exists, we say that $u^*_\Gamma(\cdot)$ is attainable.*

We can formulate the constrained regret minimization problem as a particular case of the generalized no-regret problem discussed in this paper. Indeed, let $r(a, z) = (u(a, z), c(a, z)) \in \mathbb{R}^{1+s}$ denote the vector-valued reward associated with this problem. The desired reward set for a given mixed action q of the opponent is then given by $R^*(q) = \{r = (u, c) \in \mathbb{R}^{1+s} : u \geq u^*_\Gamma(q), c \in \Gamma\}$. The goal of the agent would be to approach $R^*(\bar{q}_n)$. This is equivalent to the approachability of the set $S = \{v = (r, q) \in \mathbb{R}^{1+s} \times \Delta(\mathcal{Z}) : r \in R^*(q)\}$ with the vector-valued reward function $v(a, z) = (r(a, z), \mathbf{1}(z))$.

However, the function $u^*_\Gamma(q)$ is *not* convex in general, which implies that the set S is not convex. Therefore, one cannot invoke the dual condition to infer approachability of S, but only of its convex hull. Indeed, it was shown in Mannor et al. (2009) that S is not approachable in general.

A feasible (approachable) target set is then $\mathrm{conv}(S) = \left\{(r, q) \in \mathbb{R}^{s+1} \times \Delta(\mathcal{Z}) : \ r \in R^c(q)\right\}$, where $R^c(q) = \{r = (u, c) \in \mathbb{R}^{1+s} : u \geq \mathrm{conv}\,(u^*_\Gamma)\,(q), c \in \Gamma\}$ and the function $\mathrm{conv}\,(u^*_\Gamma)$ is the lower convex hull of $u^*_\Gamma(\cdot)$.

Two algorithms were proposed in Mannor et al. (2009) for attaining the relaxed goal function $\text{conv}(u_\Gamma^*)$. The first is a standard approachability algorithm for $\text{conv}(S)$, which requires the demanding calculation of projection directions to the convex hull of S. Further, this algorithm is not opportunistic, in the sense described below. The second algorithm relies on computing calibrated forecasts of the opponent's actions, and as we show below is equivalent to our calibration-based scheme when used for this special case.

In order to apply our algorithm, a response function p^* (Definition 9) is required. It is easily seen that any choice of $p^*(q) \in \text{argmax}_{p\in\Delta(\mathcal{A})} \{u(p,q) : \ c(p,q) \in \Gamma\}$ yields a response function. The goal function in this case is then

$$v^*(q) = (u_\Gamma^*(q), c(p^*(q),q), q)\,. \tag{6}$$

Thus, our Calibrated Approachability Algorithm can be applied, and the results of this section imply that the algorithm approaches $\text{conv}(S)$, hence attains the relaxed goal function $\text{conv}(u_\Gamma^*)$. In particular, S itself is approached whenever the opponent is either statistically or empirically restricted to a singleton $Q = \{q_0\}$ that is a continuity point of $p^*(q)$. Interestingly, in the present case the last continuity requirement can be removed.

Lemma 17 *For the model of the present section, $V^+(\{q\}) \subseteq S$ (rather than $\text{conv}(S)$) for every $q \in \Delta(\mathcal{Z})$.*

Proof Observe that the first component of v^* (defined in (6)) is continuous in q (see Mannor et al. 2009). Also, note that the jumps of $c(p^*(q),q)$, the second component of v^* , lie entirely in S. This is true since, for the fixed first component, the induced set is convex due to convexity of Γ. Consequently, the jumps of $v^*(q)$ around a given $q \in \Delta(\mathcal{Z})$ lie in S, which implies that $V^+(\{q\}) \subseteq S$ by its definition. ■

This brings us back to our requirement for a constrained no-regret algorithm, in Definition 16. While this requirement cannot be attained for any strategy of the opponent, it is attained whenever the opponent is asymptotically stationary, in the sense that its actions are (statistically or empirically) converging to a singleton.

Corollary 18 *For the model of the present section, whenever the play of the opponent is either empirically or statistically $\{q_0\}$-restricted, the Calibrated Approachability Algorithm attains $u_\Gamma^*(q_0)$ (rather than a relaxed goal) while satisfying the average cost constraints.*

6. Conclusion and Future Work

In this paper, our central goal was to formulate the concept of opportunistic strategies for the generalized no-regret problem. Our technical tool was Blackwell's approachability theory, which we extended to the opportunistic framework. We have also devised a class of calibration-based approachability algorithms and shown that they are opportunistic in the sense advocated here. The presented algorithms are computationally challenging in that they require the computation of calibrated forecasts. In addition, a procedure for the computation of the response function p^* is required, the complexity of which is problem dependent. However, given these two components, the computational burden is much lighter than standard approachability that requires computing the projection to the target set and

solving a zero-sum game in every stage. Specifically, it is sometimes difficult to compute the projection to the convex hull of a non-convex set; a step which our approach avoids.

We have applied our opportunistic framework to the problem of constrained regret minimization, and shown that the best-reward-in-hindsight (rather than its convex relaxation) is attained when the opponent turns out to be stationary in our sense.

It should be of interest to devise alternative algorithms that are computationally efficient and have optimal convergence rates. Specifically, we are currently considering a new class of algorithms that is based on online convex optimization methods.

Acknowledgments

This research was partially supported by the Israel Science Foundation under grant no. 920/12 and by the European Research Counsel under the European Union's Seventh Framework Program (FP7/2007-2013) / ERC Grant Agreement no. 306638.

References

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

A. Bernstein, S. Mannor, and N. Shimkin. Online classification with specificity constraints. In *NIPS*, 2010.

A. Bernstein, S. Mannor, and N. Shimkin. Opportunistic strategies for generalized no-regret problems: Full version. Technical report, Technion, Israel, 2013. http://tx.technion.ac.il/~andreyb/OppNoRegretColt13Full.pdf.

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, USA, 2006.

A. P. Dawid. The impossibility of inductive inference. *Journal of the American Statistical Association*, 80:340–341, 1985.

E. Even-Dar, R. Kleinberg, S. Mannor, and Y. Mansour. Online learning with global cost functions. 2009.

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

D. P. Foster, A. Rakhlin, K. Sridharan, and A. Tewari. Complexity-based approach to calibration with checking rules. In *Proceedings of the 24th*

Annual Conference on Learning Theory (COLT '11), pages 293–314, 2011. http://jmlr.csail.mit.edu/proceedings/papers/v19/foster11a/foster11a.pdf.

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 general class of adaptive strategies. *Journal of Economic Theory*, 98:26–54, 2001.

E. Hazan and S. Kakade. (weak) Calibration is computationally hard. *CoRR*, abs/1202.4478, 2012. http://arxiv.org/abs/1202.4478.

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

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

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

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

E. Milman. Approachable sets of vector payoffs in stochastic games. *Games and Economic Behavior*, 56(1):135–147, July 2006.

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

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.

A. N. Shiryaev. *Probability*. Springer, 1995.

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