Blackwell's Approachability and No-Regret Learning Algorithms

Nahum Shimkin

Department of Electrical Engineering, Technion

VALUETOOLS 2014: Dec. 9-11, Bratislava, Slovakia

Outline:

- Blackwell's Approachability
- No-Regret Learning
- Approachability and No-Regret
- *Generalized* No-Regret and Approachability
- Response-Based Approachability
- Calibrated Approachability

Blackwell's Theory of Approachability for Repeated Games with Vector Payoffs

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

Matrix Games

- We consider a repeated matrix game between Player 1 (the *agent*) and Player 2 (the *adversary*, which may be *Nature*).
- Recall that a **matrix game** Γ is defined by:
  - Finite action sets I and J for the two Players
  - A reward function $r : I \times J \to \mathbb{R}$, with $r(i, j)$ the payoff to Player 1
  - Mixed (randomized) actions $p \in \Delta(I)$, $q \in \Delta(J)$
- Denote $r(p,q) = \sum_{i \in I, j \in J} p(i)q(j)r(i,j)$
- Von Neumann's minimax theorem:

$$\max_p \min_q r(p,q) = \min_q \max_p r(p,q) := \mathrm{val}(\Gamma)$$

$\mathrm{val}(\Gamma)$ provides a *security level* for Player 1 against an arbitrary opponent.

Repeated Matrix Games

- In a **repeated** matrix game, Γ_∞, the same matrix game Γ is played sequentially at stages $k = 1, 2, \ldots$
- At each stage k, actions i_k and j_k are taken, and a reward $r_k = r(i_k, j_k)$ is obtained.
- The players may choose their mixed action p_k and q_k at stage k as functions of the observed game history $H_k = (i_t, j_t, r_t)_{1 \le t \le k-1}$.
- Let $\overline{r}_n$ denote the average n-stage reward:

$$\overline{r}_n = \frac{1}{n} \sum_{k=1}^{n} r_k$$

Blackwell's Approachability Framework

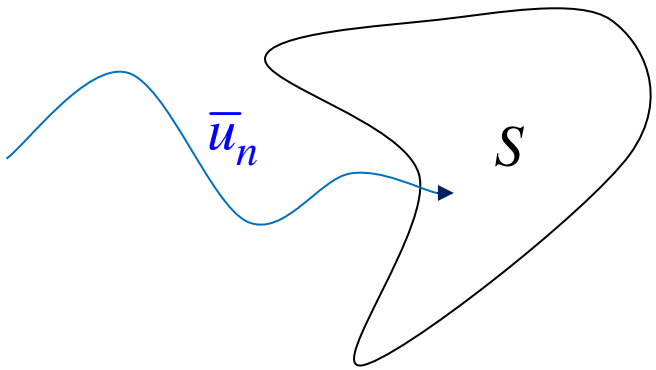
- Consider the repeated matrix game model above, but with a **vector-valued** payoff function: $u(i,j) \in \mathbb{R}^{\ell}$. Denote

$$\bar{u}_n = \frac{1}{n}\sum_{k=1}^{n} u(i_k, j_k) \in \mathbb{R}^{\ell}$$

- A set $S \subset \mathbb{R}^{\ell}$ is **approachable** by Player 1, if she has a strategy σ_1 so that, for any strategy σ_2 of Player 2,

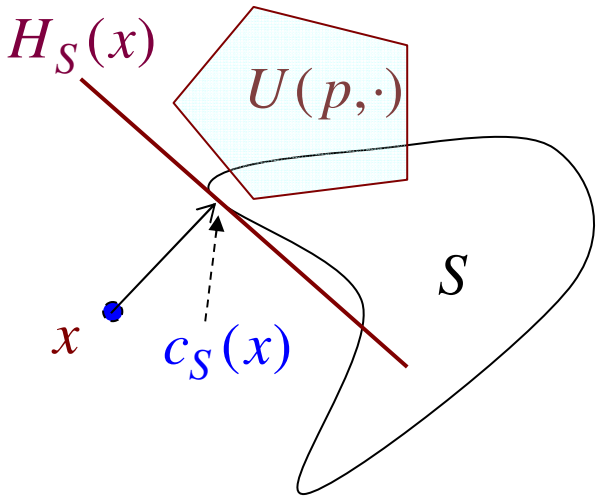
$$\lim_{n\to\infty} d(\bar{u}_n, S) = 0 \quad \textit{(a.s.)}$$

  - We actually require the a.s. convergence rate to be uniform over the opponent's strategies.



Blackwell's Sufficient (Primal) Condition

- *Notation:*
  - Let $S \subset \mathbb{R}^{\ell}$ be a closed ***target set***.
  - For $x \notin S$, $c_S(x)$ denotes a closest point to x in S
  - $H_S(x)$ denotes the hyperplane through $c_S(x)$ perpendicular to the line segment $x - c_S(x)$.
  - $U(p,\cdot) = \{u(p,q) : q \in \Delta(J)\}$ for $p \in \Delta(I)$.

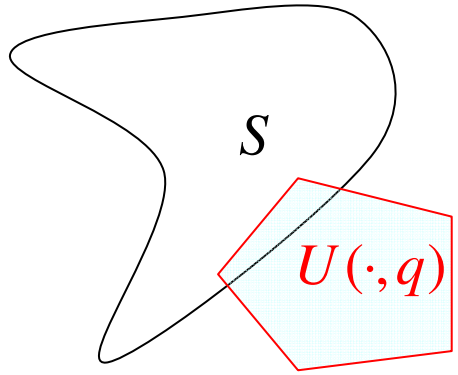


- **Definition:** A closed set $S \subset \mathbb{R}^{\ell}$ is a ***B-set*** if, for any point $x \notin S$, there exists $p = p^*(x) \in \Delta(I)$ s.t. $H_S(x)$ separates x from $U(p,\cdot)$
- **Theorem** [Blackwell 1956]:
  - A B-set is approachable, with convergence rate $O(n^{-\frac{1}{2}})$
  - Blackwell's approachability strategy: $p_{k+1} = p^*(\bar{u}_k)$.

Blackwell's Necessary (Dual) Condition

- $S \subset \mathbb{R}^{\ell}$ satisfies **Blackwell's dual condition** if, for any $q \in \Delta(J)$ there exists $p \in \Delta(I)$ so that $u(p,q) \in S$.

  Equivalently, $U(\cdot,q)$ must intersect S for every $q \in \Delta(J)$.

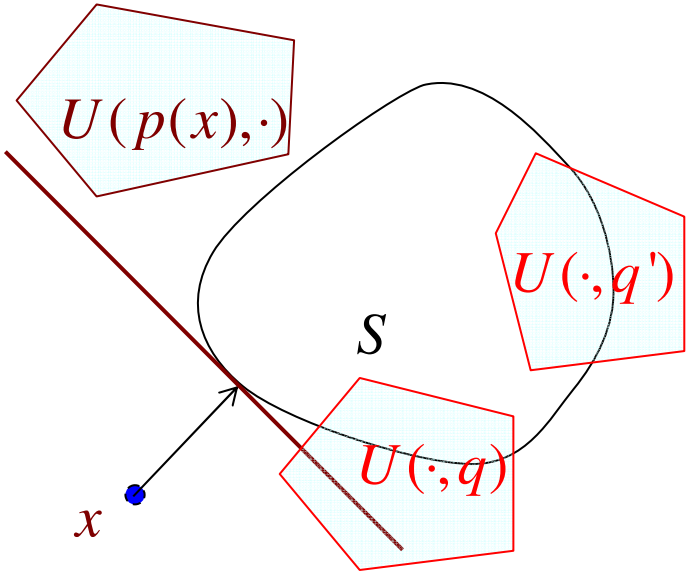


- The dual condition is clearly ***necessary*** for approachability.

Convex Target Sets

Suppose the target set S is convex. Recall:

- The primal condition:
  $\forall x \notin S \ \exists p$ s.t. $H_S(x)$ separates x from $U(p,\cdot)$
- The dual Condition:
  $\forall q \ \exists p$ s.t. $u(p,q) \in S$



Theorem [Blackwell: "An analog of the minimax theorem for vector payoffs"]:

- If S is **convex**, then the primal and dual conditions are equivalent.
- Hence either one is both necessary and sufficient for approachability.

Approachability Algorithms (for convex target sets)

- Existing approachability algorithms include:
  - Blackwell's strategy: Steering the reward vector in the projection direction.
  - HART & MAS-COLLEL (2001): Potential-induced steering directions.
  - ABERNETHY, BARTLETT & HAZAN (2012): Steering direction generated by a no-regret strategy, related to the *support function* of the set S.
- These algorithms rely on Blackwell's primal condition – existence of a separating hyperplane – and essentially require computing a projection direction onto S.
- We propose below approachability algorithms that ***rely on Blackwell's dual condition***: existence of a response map $q \mapsto p^*(q)$ such that $u(p^*(q), q) \in S$.
- Such algorithms are useful in certain problems (to be discussed) where:
  - The target set S is complicated, so that computing projections is hard,
  - but the response map is readily available.

No-Regret Learning

Origins

- No-regret strategies were introduced by Hannan in the context of repeated matrix games:
  - J. Hannan, Approximation to Bayes risk in repeated play, 1957.

- Regret is now the de-facto standard for measuring the learning-loss performance of on-line learning algorithms.
- We will focus here on the original repeated matrix game framework. However, most of the discussion is relevant for more general models, such as regression and classification, that involve continuous action spaces.

Regret (1)

- We consider again a repeated matrix game, Γ_∞, with a **scalar** reward function $r(i, j)$. Consider the cumulative expected reward:

$$R_n = E(\sum_{k=1}^{n} r_k), \quad r_k = r(i_k, j_k)$$

- If Player 1 uses repeatedly her maximin strategy p^*, she can evidently secure

$$R_n \geq n \cdot val(\Gamma), \quad val(\Gamma) = \max_p \min_q r(p, q)$$

- Can she do better – against an arbitrary opponent – by observing the past actions of the opponent ?

Regret (2)

- Define the (cumulative, n-stage) **regret**:

$$L_n = \max_{i \in I} \sum_{k=1}^{n} r(i, j_k) \; - \; \sum_{k=1}^{n} r(i_k, j_k)$$

- The first term is what we *could achieve by any fixed action*, with the benefit of *hindsight*. The second term is what was actually obtained.

  * Regret: "What we feel when we realize that we could have done better."

- A strategy of Player 1 is called a ***no-regret strategy*** or ***Hannan consistent*** if

$$\limsup_{n \to \infty} \left(\frac{1}{n} L_n\right) \le 0 \qquad (a.s.)$$

  for any (causal) strategy of Player 2. Here $\frac{1}{n} L_n$ is the *average* regret of Player 1.

- We write the above succinctly as $L_n \le o(n)$, or $\frac{1}{n} L_n \le o(1)$.

Regret: Bayes Envelope Formulation

- Recall that $\overline{r}_n = \frac{1}{n}\sum_{k=1}^{n} r(i_k, j_k)$ denotes the average n-stage reward, and let $\overline{q}_n$ denote the *empirical mean* of Player 2's actions:

$$\overline{q}_n = \frac{1}{n}\sum_{k=1}^{n} e(j_k)$$

- The average regret can now be written as

$$\frac{1}{n}L_n = \max_{i\in I} r(i, \overline{q}_n) - \overline{r}_n = \underline{r^*(\overline{q}_n) - \overline{r}_n}$$

where $r^*(q)$ is the *Bayes envelope* of the game:

$$r^*(q) = \max_{i\in I} r(i,q) \equiv \max_{p\in\Delta(I)} r(p,q)$$

- The no-regret property is therefore equivalent to

$$\overline{r}_n \geq r^*(\overline{q}_n) - o(1)$$

Example 1: Universal Prediction

- Suppose we wish to forecast the occurrence of *rain* on the following day.
- Let $j_k \in \{rain, sun\}$ (Nature's "choice"), $i_k \in \{rain, sun\}$ (our rain prediction), and $r(i, j) = \mathbf{I}\{i = j\}$ (unity reward for correct prediction).
- Let us identify $\overline{q}_n$ with the relative occurrence of *rain*: $\overline{\xi}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{I}\{j_k = rain\} \in [0,1]$.
- Observe that $r^*(\xi) = \max_{i \in \{rain, sun\}} r(i, \xi) = \max\{1-\xi, \xi\}$.

  For instance:
  - If *rain* proportion is 60%, we want to be correct 60% of the time.
  - If *rain* proportion is 30%, we wish to be correct 70% of the time.
- Note that a no-regret strategy should achieve this (asymptotically) even if the occurrence of rain is completely unpredictable, and without knowing anything about weather forecasting…

Example 2: Prediction with Expert Advice

- Suppose now we have access to M *experts* (e.g., weather prediction algorithms) , each providing a prediction for tomorrow's rain.
- Our goal is to **do as well as the best expert** (in terms of proportion of the successful predictions).
- This may be seen to be a no-regret problem, where
  - Player 1's action is the next choice of expert: $I = \{1, 2, \ldots, M\}$
  - Player 2's action aggregates the experts' predictions and the actual outcome.

Example 3 : Online Convex Programming (a modern extension)

- Let $C = \{c : X \to \mathbb{R}\}$ be a set of convex functions over a convex set X.
- Suppose at each stage k, Player 1 (the learning algorithm) chooses a point $x_k \in X$, nature (Player 2) chooses a function $c_k \in C$, and a cost $c_k(x_k)$ is incurred.
- A no-regret algorithm must achieve

$$\sum_{k=1}^{n} c_k(x_k) \le \min_{x \in X} \sum_{k=1}^{n} c_k(x) + o(n)$$

Specific problems:

- Linear regression: $c_k(x) = \| z_k - \varphi_k^T x \|^2$
- No-regret Routing: the best-path in hindsight in a network with arbitrarily varying link costs.

No-Regret Algorithms

- Follow the Leader / Fictitious Play:

$$i_{k+1} \in \arg\max_{i \in I} \sum_{t=1}^{k} r(i, j_t)$$

  - Unfortunately this fails, as does any deterministic strategy.

- Follow the perturbed leader [Hannan '57]:

$$i_{k+1} \in \arg\max_{i \in I} \{\sum_{t=1}^{k} r(i, j_t) + \zeta_k(i)\}$$

  where $\{\zeta_k(i)\}$ are independent RVs of diminishing size.

- Multiplicative updates [Littlestone '88]:

$$p_k[i] = \frac{w_k[i]}{\| w_k \|_1}, \quad w_{k+1}[i] := w_k[i] \exp(\eta_k r(i, j_k))$$

- Gradient ascent [Zinkevic '03]:
  - $p_{k+1} = proj_{\Delta(I)}(p_k + \eta_k r(\cdot, j_k))$, $\eta_k = O(\frac{1}{\sqrt{k}})$

- Convergence rates are generally $O(\sqrt{n})$.

No-Regret Algorithms (Partial List)

A variety of no-regret algorithms for repeated matrix games have been proposed and analyzed over the years, including:

- Hannan ('57)– perturbed fictitious play (original formulation and proof)
- Blackwell ('56) – approachability theory
- Littlestone & Warmuth ('94) – weighed majority alg. (experts problem)
- Fudenberg and Levine ('95) – smooth fictitious play
- Foster and Vohra ('97) – calibrated play
- Auer, Cesa-Bianchi, Freund & Schapire ('02) – multiplicative weights (bandit formulation)
- Hart & Mas-Colell ('00) – regret matching
- Hart & Mas-Colell ('01), Cesa-Bianchi & Lugosi ('03) – potential-based
- Zinkevich ('03) – gradient ascent

No-regret algorithms are closely related to several other problems of interest, such as universal prediction, calibrated predictions, and equilibrium dynamics in games.

Approachability and No-Regret:

Formulating the no-regret goal as an approachability problem

Blackwell's Formulation

- Recall the agent's (Player 1) no-regret goal:

$$\frac{1}{n}\sum_{k=1}^{n} r(i_k, j_k) \geq \max_i r(i, \bar{q}_n) - o(1),$$

or

$$\bar{r}_n \geq r^*(\bar{q}_n) - o(1)$$

- Define the following payoff vector:

$$u(i, j) = (r(i, j), e(j)) \in \mathbb{R} \times \Delta(J)$$

so that

$$u(p, q) = (r(p, q), q), \qquad \bar{u}_n = (\bar{r}_n, \bar{q}_n)$$

- No-regret is then equivalent to approaching the following target set:

$$S = \{(r, q) \in \mathbb{R} \times \Delta(J) \ : \ r \geq r^*(q)\}$$

Blackwell's Formulation (2)

- We next verify that the set $S = \{u = (r,q) : r \geq r^*(q)\}$ is approachable. Indeed,
  - The function $q \mapsto r^*(q) = \max_i r(i,q)$ is convex, hence S is a convex set.
  - The dual condition is satisfied by construction: For every $q \in \Delta(J)$,

    $\exists p$ s.t. $u(p,q) \in S$ is satisfied by $p(q) \in \arg\max_p r(p,q)$

- Blackwell's strategy requires, at each stage $k+1$, to
  - Compute the projection direction λ_k from $\bar{u}_k = (\bar{r}_k, \hat{q}_k)$ onto S
  - Compute a maximin strategy in the game with payoff function $\lambda_k \cdot u(i,j)$, and apply it as p_{k+1}

Regret Matching

Hart & Mas-Collel (2001) proposed a different approachability-based formulation of the no-regret problem, leading to a more explicit algorithm.

- The no-regret property $\overline{r}_n \geq \max_i r(i, \hat{q}_n) - o(1)$ may be written as
$$\overline{r}_n \geq r(i', \hat{q}_n) - o(1), \quad i' \in I$$
- Define the payoff vector $u(i, j) = (r(i', j) - r(i, j))_{i' \in I}$, so that
$$\overline{u}_n = (r(i', \hat{q}_n) - \overline{r}_n)_{i' \in I}$$
- No-regret is now equivalent to approaching the negative quadrant:
$$S = \left\{ u \in R^{|I|} : u \leq 0 \right\}$$
- Blackwell's strategy is then explicitly given:
$$p_{k+1}[i] = \frac{[r(i, \hat{q}_k) - \overline{r}_n]_+}{\| [r(\cdot, \hat{q}_k) - \overline{r}_n]_+ \|_1}$$

Some Generalized No-Regret Problems

* That lead to complex / nonconvex target sets

Constrained No-Regret

- Consider a repeated game model, with a scalar reward function $r(i, j)$, and with an additional cost function $c(i, j)$ (possibly vector-valued).

- As before, we are interested in maximizing the average reward

$$\overline{r}_n = \frac{1}{n}\sum_{k=1}^{n} r(i_k, j_k),$$

but subject to the **average cost constraints**:

$$\overline{c}_n := \frac{1}{n}\sum_{k=1}^{n} c(i_k, j_k) \le \gamma + o(1)$$

Constrained No-Regret (2)

No regret formulation (MANNOR, TSITSIKLIS & YU, 2009):

- Let $$r_\gamma(q) = \max_{p \in \Delta(I)} \{ r(p,q) : c(p,q) \le \gamma \} \qquad (*)$$
- A ***constrained no-regret algorithm*** should satisfy, for any strategy of the opponent:

$$\overline{r}_n \ge r_\gamma(\hat{q}_n) - o(1),$$
$$\overline{c}_n \le \gamma + o(1)$$

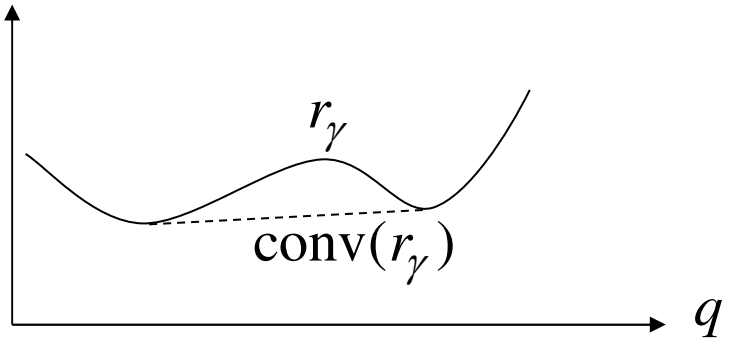
- Approachability formulation: Define

$$u = (r, c, \delta_j), \qquad S = \{ (r,c,q) : r \ge r_\gamma(q),\ c \le \gamma \}$$

- The response function here is, by construction, the optimal action $p^*_\gamma(q)$ in (*).

Constrained No-Regret (3)

- Alas, the constrained Bayes envelope $r_\gamma(q)$ need not be a convex function, hence S is not a convex set.
- Indeed, constrained no-regret ***cannot*** be achieved in general. We should therefore settle with a little less.
- Let $\mathrm{co}(S)$ denote the (closed) *convex hull* of S. Then $\mathrm{co}(S)$ is a convex set that satisfied Blackwell's dual condition, hence approachable.



- It may be seen that
$$\mathrm{co}(S) = \{(r,c,q) \ : \ r \geq \mathrm{conv}(r_\gamma)(q),\ c \leq \gamma\}$$
where $\mathrm{conv}(r_\gamma)$ is the **lower convex hull** of $q \mapsto r_\gamma(q)$.
- Therefore, approaching $\mathrm{co}(S)$ will guarantee
$$\overline{r}_n \geq \mathrm{conv}(r_\gamma)(\hat{q}_n) - o(1), \quad \text{and} \quad \overline{c}_n \leq \gamma + o(1).$$
- *This is still more than the maximin value of the constrained game*.

Constrained No-Regret (4)

- Since the set $\mathrm{co}(S)$ is generally complex, it is hard to apply here Blackwell's direct algorithm (which requires projection onto this set).
- On the other hand, the response function is readily calculated by solving the linear program

$$p^*(q) = \underset{p \in \Delta(I)}{\arg\max} \{ r(p,q) \,:\, c(p,q) \le \gamma \}$$

- A response-based approachability algorithm is therefore useful here.

Additional Generalized No-Regret Problems

Other generalized no-regret problems of similar nature, where the Response-Based Approachability algorithm may be efficiently applied:

- Maximizing a reward-to-cost ratio (MANNOR & S., 2008):

$$\frac{\overline{r}_n}{\overline{c}_n} \geq \max_{p \in \Delta(I)} \frac{r(p, \hat{q}_n)}{c(p, \hat{q}_n)}$$

- Regret minimization in **stochastic games**, i.e., Markovian models with arbitrarily-varying rewards and transitions (MANNOR & S., 2003).

Response-Based Approachability

Response-based Approachability: Basic Ideas

- Let S be a convex target set, and suppose we are given a ***response map*** $q \mapsto p^*(q)$ such that $u(p^*(q), q) \in S$.
- *If* we knew the adversary's next mixed action q_k, we could choose $p_k = p^*(q_k)$, so that $u_k = u(p_k, q_k) \in S$.
- Since S is convex, this obtains $\overline{u}_n = \frac{1}{n}\sum_{k=1}^{n} u_k \in S$.

However, as q_k is not known is advance, we must resort to substitutes:

- *Calibrated approachability* (BERNSTEIN, MANNOR & S., 2013) uses calibrated forecasts of q_k in place of q_k itself. The algorithm has some additional interesting "opportunistic" properties, but is computationally hard.
- *Response-based approachability* (BERNSTEIN & S., 2014) replaces q_k by some fictituous signal q_k^*, as describes next.

Algorithm Template

- At each stage k, in addition to the actual action p_k, we choose fictitious mixed actions p_k^*, q_k^*. Let

$$\bar{u}_n = \frac{1}{n}\sum_{k=1}^{n} u(p_k, j_k), \quad \bar{u}_n^* = \frac{1}{n}\sum_{k=1}^{n} u(p_k^*, q_k^*)$$

  (i) The pair (p_k, q_k^*) is chosen so that, for *any* (adversarial) choice of (p_k^*, q_k), we obtain

$$\| \bar{u}_n - \bar{u}_n^* \| \le \frac{c}{\sqrt{n}}$$

  (ii) p_k^* is chosen as an *S*-response to q_k^*: $p_k^* = p^*(q_k^*)$.

- As a consequence of (i), we have $\bar{u}_n^* \in S$. Therefore, by (ii), $d(\bar{u}_n, S) \le \frac{c}{\sqrt{n}} \to 0$

Achieving property (i)

- The required convergence $\| \bar{u}_n - \bar{u}_n^* \| \to 0$ in part (i) of the algorithm is achieved by choosing (p_k, q_k^*) as follows:

$$p_k \in \arg\max_{p \in \Delta(I)} \min_{q \in \Delta(J)} \rho_k(p,q)$$

$$q_k^* \in \arg\min_{q \in \Delta(J)} \max_{p \in \Delta(I)} \rho_k(p,q)$$

where

$$\rho_k(p,q) = (\bar{u}_{k-1}^* - \bar{u}_{k-1}) \cdot u(p,q)$$

That is: (p_k, q_k^*) are the saddle-point strategies in the game with payoff $\rho_k(p,q)$, which is the vector payoff $u(p,q)$ projected unto $(\bar{u}_{k-1}^* - \bar{u}_{k-1})$.

- That choice can be interpreted as an approachability strategy for an auxiliary game with respective actions (p_k, q_k^*), (p_k^*, q_k), and target point $S = \{0\}$

To Summarize:

Response-Based Approachability Algorithm:

For $k = 1, 2, \ldots$,

- $\lambda_k := \bar{u}^*_{k-1} - \bar{u}_{k-1} = \frac{1}{k-1}\sum_{t=1}^{k-1} u(p^*_t, q^*_t) - \frac{1}{k-1}\sum_{t=1}^{k-1} u(p_t, j_t)$.
- $p_k \in \arg\max_{p \in \Delta(I)} \min_{q \in \Delta(J)} \lambda_k \cdot u(p,q)$
- $q^*_k \in \arg\min_{q \in \Delta(J)} \max_{p \in \Delta(I)} \lambda_k \cdot u(p,q)$
- $p^*_k = p^*(q^*_k)$

Theorem (BERNSTEIN & S., 2014)

$d(\bar{u}_n, S) \le \frac{c}{\sqrt{n}}$ for all $n \ge 1$, with $c = \max \| u(p,q) - u(p',q') \|$.

Calibration-Based Approachability

and its Opportunistic Properties

Calibrated Forecasts

- Recall that we wish to replace q_k in $p_k = p^*(q_k)$ by some forecast of the opponent's actions.
- A sequence $(y_k \in \Delta(J))$ is a **calibrated forecast** of $(j_k \in J)$ if, for any measurable set $B \subset \Delta(J)$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 1\{y_k \in B\}(e_{j_k} - y_k) = 0$$

- Algorithms for generating calibrated forecasts of arbitrary sequences exist, but the problem is computationally hard (HAZAN & KAKADE, 2012).

Calibrated Approachability

- **The algorithm**:

$$p_k = p^*(y_k),$$

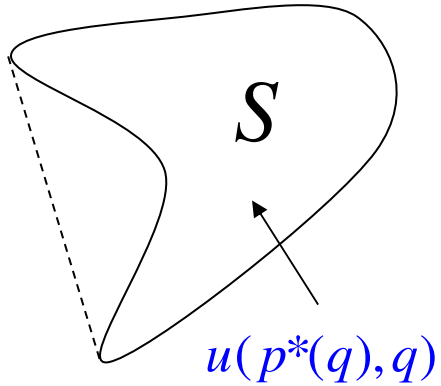
where (y_k) is a calibrated forecast of (j_k), and $p^*(y)$ an *S*-response to y such that $u(p^*(y), y) \in S$.

The following is easy to show:

- **Proposition**. The calibration-based algorithm is an approachability strategy for $conv(S)$, the convex hull of S.

Opportunistic Properties

- Let $S \subset \mathbb{R}^{\ell}$ be a **non-convex** set that satisfies Blackwell's dual condition, namely, for every q there exists an *S*-response $p^*(q)$ such that $u(p^*(q),q) \in S$.
- As we know, $conv(S)$ is then approachable. However, S itself need not be so.
- Observe that if the opponent is known to be stationary, we can easily obtain $d(\bar{u}_n, S) \to 0$ (a.s.), simply by playing an *S*-response to the sample mean $\bar{q}_n$ of (j_k).



- It may be shown that the calibration-based algorithm provides similar properties without known in advance that the opponent is stationary. Namely:
  - $conv(S)$ is always approached.
  - If the opponent's actions happens to be stationary in some sense, then S itself is approached, namely $d(\bar{u}_n, S) \to 0$.

Opportunistic Properties

Theorem (BERNSTEIN, MANNOR & S., 2013)

- Suppose the opponent's mixed actions (q_k) are restricted to a convex set $Q \subset \Delta(J)$, in the sense that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} d(q_k, Q) = 0$$

Then

$$\lim_{n\to\infty} d(\overline{u}_n, p^*(Q)) = 0,$$

where

$$p^*(Q) = \mathrm{conv}\{ p^*(q),\ q \in Q\} \subset \mathrm{conv}(S).$$

- In particular, if $Q = \{q_0\}$ is a singleton, then $\overline{r}_n \to p^*(q_0) \in S$.

A Brief Summary

- We have outlined Blackwell's approachability theory and its application to no-regret algorithms.
- A couple of approachability algorithms that are based on Blackwell's dual condition (existence of a response function) were introduced.

Future work:

- An approachability algorithm with opportunistic properties that avoids the computationally-hard calibrated forecasting.
- Extensions to arbitrarily-varying Markov Decision Problems.

Thanks for your attention

Enjoy the conference