A TWO ARMED BANDIT TYPE PROBLEM REVISITED

GILLES PAGÈS

Abstract. In Benaim and Ben Arous (2003) is solved a multi-armed bandit problem arising in the theory of learning in games. We propose a short and elementary proof of this result based on a variant of the Kronecker lemma.

Mathematics Subject Classification. 91A20, 91A12, 60F99.

Received December 10, 2004. Revised April 29, 2005.

In [2] a multi-armed bandit problem is addressed and investigated by Benaim and Ben Arous. Let \( f_0, \ldots, f_d \) denote \( d + 1 \) real-valued continuous functions defined on \([0, 1]^{d+1}\). Given a sequence \( x = (x_n)_{n \geq 1} \in \{0, \ldots, d\}^N \) (the strategy), set for every \( n \geq 1 \)
\[
\bar{x}_n := (\bar{x}_0^n, \bar{x}_1^n, \ldots, \bar{x}_d^n) \quad \text{with} \quad \bar{x}_i^n := \frac{1}{n} \sum_{k=1}^{n} 1_{\{x_k=i\}}, \quad i = 0, \ldots, d,
\]
and
\[
Q(x) = \liminf_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n-1} f_{x_k}(\bar{x}_k).
\]

In [2] an answer (see Th. 1 below) is provided to the following question

What are the good strategies (for the group)?

The authors rely on some recent tools developed in stochastic approximation theory (see e.g. [1]). The aim of this note is to provide an elementary and shorter proof based on a slight improvement of the Kronecker lemma. As an illustration, we emphasize that in such a game a greedy strategy is usually not optimal, even for the “individual winner”.

Keywords and phrases. Two-armed bandit problem, Kronecker lemma, learning theory, stochastic fictitious play.

1 Laboratoire de Probabilités et Modèles Aléatoires, UMR 7599, Université Paris 6, case 188, 4, place Jussieu, 75252 Paris Cedex 5, France; gpa@ccr.jussieu.fr

© EDP Sciences, SMAI 2005

Article published by EDP Sciences and available at http://www.edpsciences.org/ps or http://dx.doi.org/10.1051/ps:2005017
Let $S_d := \{ v = (v_1, \ldots, v_d) \in [0, 1]^d, \sum_{i=1}^d v_i \leq 1 \}$ and $P_{d+1} := \{ u = (u_0, u_1, \ldots, u_d) \in [0, 1]^{d+1}, \sum_{i=1}^{d+1} u_i = 1 \}$. Furthermore, for notational convenience, set

$$\forall v \in S_d, \; \bar{v} := \left(1 - \sum_{i=1}^d v_i, v_1, \ldots, v_d\right) \in P_{d+1}.$$

$$\forall u \in P_{d+1}, \; \sigma u := (u_1, \ldots, u_d) \in S_d. \tag{1}$$

The canonical inner product on $\mathbb{R}^d$ will be denoted by $(v|w) = \sum_{i=1}^d v_i w_i$. The interior of a subset $A$ of $\mathbb{R}^d$ will be denoted by $\overset{\circ}{A}$. For a sequence $u = (u_n)_{n \geq 0}, \Delta u_n := u_n - u_{n-1}, n \geq 1$.

The main result is the following theorem (first established in [2]).

**Theorem 1.** Assume there is a continuous function $\Phi : S_d \to \mathbb{R}$, continuously differentiable on $\overset{\circ}{S}_d$, having a continuous extension of its gradient $\nabla \Phi$ on $\overset{\circ}{S}_d$ and satisfying:

$$\forall v \in S_d, \quad \nabla \Phi(v) = (f_i(\bar{v}) - f_0(\bar{v}))_{1 \leq i \leq d}. \tag{2}$$

Set for every $u \in P_{d+1},$

$$q(u) := \sum_{i=0}^{d+1} u^i f_i(u)$$

and $Q^* := \max \{q(u), \; u \in P_{d+1}\}$. Then, for every strategy $x \in \{0, 1, \ldots, d\}^\mathbb{N}^*$,

$$Q(x) \leq Q^*.$$

Furthermore, for any strategy $x$ such that $\bar{x}_n \to \bar{x}_\infty$,

$$\frac{1}{n} \sum_{k=0}^{n-1} f_{x_{k+1}}(\bar{x}_k) \to q(\bar{x}_\infty) \quad as \quad n \to \infty \quad (so \ that \ Q(x) = q(\bar{x}_\infty)).$$

In particular there is no better strategy than choosing the player at random according to an i.i.d. “Bernouilli strategy” with parameter $\bar{x}_\infty \in \text{argmax}_{x} q$.

The key of the proof is the following slight extension of the Kronecker lemma.

**Lemma 1** (“à la Kronecker” lemma). Let $(b_n)_{n \geq 1}$ be a nondecreasing sequence of positive real numbers converging to $+\infty$ and let $(a_n)_{n \geq 1}$ be a sequence of real numbers. Then

$$\liminf_{n \to +\infty} \sum_{k=1}^{n} \frac{a_k}{b_k} \in \mathbb{R} \quad \Rightarrow \quad \liminf_{n \to +\infty} \frac{1}{b_n} \sum_{k=1}^{n} a_k \leq 0.$$

**Proof.** Set $C_n = \sum_{k=1}^{n} a_k / b_k$, $n \geq 1$, and $C_0 = 0$ so that $a_n = b_n \Delta C_n$. As a consequence, an Abel transform yields

$$\frac{1}{b_n} \sum_{k=1}^{n} a_k = \frac{1}{b_n} \sum_{k=1}^{n} b_k \Delta C_k = \frac{1}{b_n} \left( b_n C_n - \sum_{k=1}^{n} C_{k-1} \Delta b_k \right)$$

$$= C_n - \frac{1}{b_n} \sum_{k=1}^{n} C_{k-1} \Delta b_k.$$
Now, \( \liminf_{n \to +\infty} C_n \) being finite, for every \( \varepsilon > 0 \), there is an integer \( n_\varepsilon \) such that for every \( k \geq n_\varepsilon \), \( C_k \geq \liminf_{n \to +\infty} C_n - \varepsilon \). Hence
\[
\frac{1}{b_n} \sum_{k=1}^{n} C_{k-1} \Delta b_k \geq \frac{1}{b_n} \sum_{k=1}^{n_\varepsilon} C_{k-1} \Delta b_k + \frac{b_n - b_{n_\varepsilon}}{b_n} \left( \liminf_{k \to +\infty} C_k - \varepsilon \right).
\]
Consequently, \( \liminf_{n \to +\infty} C_n \) being finite, one concludes that for every \( \varepsilon > 0 \),
\[
\liminf_{n \to +\infty} \frac{1}{b_n} \sum_{k=1}^{n} a_k \leq \liminf_{n \to +\infty} C_n - 1 \times \left( \liminf_{k \to +\infty} C_k - \varepsilon \right) = \varepsilon.
\]

**Proof of Theorem 1.** First note that for every \( u = (u^0, u^1, \ldots, u^d) \in \mathcal{P}_{d+1} \),
\[
q(u) := \sum_{i=0}^{d+1} u^i f_i(u) = f_0(u) + \sum_{i=1}^{d} u^i (f_i(u) - f_0(u))
\]
so that
\[
Q^* = \sup_{v \in \mathcal{S}_d} \left\{ f_0(\bar{v}) + \sum_{i=1}^{d} v^i (f_i(\bar{v}) - f_0(\bar{v})) \right\} = \sup_{v \in \mathcal{S}_d} \{ f_0(\bar{v}) + (v | \nabla \Phi(v)) \}.
\]

Now, for every \( k \geq 0 \),
\[
f_{x_{k+1}}(\bar{x}_k) - q(\bar{x}_k) = \sum_{i=0}^{d} (f_i(\bar{x}_k) 1_{\{x_{k+1} = i\}} - \bar{x}_k^i f_i(\bar{x}_k)) = \sum_{i=1}^{d} f_i(\bar{x}_k) (1_{\{x_{k+1} = i\}} - \bar{x}_k^i)
\]
\[
= \sum_{i=0}^{d} f_i(\bar{x}_k) (k+1) \Delta \bar{x}_k^i
\]
\[
= (k+1) \sum_{i=1}^{d} (f_i(\bar{x}_k) - f_0(\bar{x}_k)) \Delta \bar{x}_k^i.
\]
The last equality reads using Assumption (2) and notation (1),
\[
f_{x_{k+1}}(\bar{x}_k) - q(\bar{x}_k) = (k+1)(\nabla \Phi(\bar{x}_k) | \Delta \bar{x}_{k+1}).
\]
Consequently, by the fundamental formula of calculus applied to \( \Phi \) on \((\sigma \bar{x}_k, \sigma \bar{x}_{k+1}) \subset \tilde{\mathcal{S}}_d \),
\[
\frac{1}{n} \sum_{k=0}^{n-1} f_{x_{k+1}}(\bar{x}_k) - q(\bar{x}_k) = \frac{1}{n} \sum_{k=0}^{n-1} (k+1) (\Phi(\sigma \bar{x}_{k+1}) - \Phi(\sigma \bar{x}_k)) - R_n
\]
with
\[
R_n := \frac{1}{n} \sum_{k=0}^{n-1} (\nabla \Phi(\xi_k) - \nabla \Phi(\sigma \bar{x}_k) | (k+1) \Delta \sigma \bar{x}_{k+1})
\]
and \( \xi_k \in (\sigma \bar{x}_k, \sigma \bar{x}_{k+1}) \), \( k = 0, \ldots, n-1 \). The fact that \( |(k+1) \Delta \sigma \bar{x}_{k+1}| \leq 1 \) implies
\[
|R_n| \leq \frac{1}{n} \sum_{k=0}^{n-1} w(\nabla \Phi | \Delta \sigma \bar{x}_{k+1})
\]
where \( w(g, \delta) \) denotes the uniform continuity \( \delta \)-modulus of a function \( g \). One derives from the uniform continuity of \( \nabla \Phi \) on the compact set \( S_d \) that 

\[
R_n \to 0 \quad \text{as} \quad n \to +\infty.
\]

Finally, the continuous function \( \Phi \) being bounded on the compact set \( S_d \), the partial sums

\[
\sum_{k=0}^{n-1} \Phi^\kappa(\sigma_{\bar{x}_k+1}) - \Phi^\kappa(\sigma_{\bar{x}_0}) = \Phi^\kappa(\sigma_{\bar{x}_{n+1}}) - \Phi^\kappa(\sigma_{\bar{x}_0})
\]

remain bounded as \( n \) goes to infinity. Lemma 1 then implies that

\[
\liminf_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n-1} (k+1) (\Phi^\kappa(\sigma_{\bar{x}_k+1}) - \Phi^\kappa(\sigma_{\bar{x}_k})) \leq 0.
\]

One concludes by noting that on one hand

\[
\limsup_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n-1} q(\bar{x}_k) \leq Q^* = \sup_{P_{d+1}} q
\]

and that, on the other hand, the function \( q \) being continuous,

\[
\lim_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n-1} q(\bar{x}_k) = q(x^*) \quad \text{as soon as} \quad \bar{x}_n \to x^*.
\]

\[\square\]

**Corollary 1.** When \( d+1 = 2 \) (two players), Assumption (2) is satisfied as soon as \( f_0 \) and \( f_1 \) are continuous on \( P_2 \) and then the conclusions of Theorem 1 hold true.

**Proof.** This follows from the obvious fact that the continuous function \( u^i \mapsto f_1(1-u^i, u^i) - f_0(1-u^i, u^i) \) on \([0,1]\) has an antiderivative. \(\square\)

**Further comments:**

- If one considers a slightly more general game in which some weighted strategies are allowed, the final result is not modified in any way provided the weight sequence satisfies a very light assumption. Namely, assume that at time \( n \) the reward is

\[
\Delta_{n+1} f_{x_{n+1}}(\bar{x}_n)
\]

instead of \( f_{x_{n+1}}(\bar{x}_n) \)

where the weight sequence \( \Delta = (\Delta_n)_{n \geq 1} \) satisfies

\[
\Delta_n \geq 0, \quad n \geq 1, \quad S_n = \sum_{k=1}^{n} \Delta_k \to +\infty, \quad \frac{\Delta_n}{S_n} \to 0 \text{ as } n \to \infty
\]

then the quantities \( \bar{x}_n^\Delta \in P_{d+1}, \quad \bar{x}_n^\Delta := (\bar{x}_n^\Delta, \ldots, \bar{x}_n^\Delta) \) with \( \bar{x}_n^\Delta = \frac{1}{S_n} \sum_{k=1}^{n} \Delta_k 1_{\{x_k = i\}}, \quad i = 0, \ldots, d, \quad n \geq 1, \)

and \( Q^\Delta(x) = \liminf_{n \to +\infty} \frac{1}{S_n} \sum_{k=0}^{n-1} \Delta_{k+1} f_{x_{k+1}}(\bar{x}_k^\Delta) \) satisfy all the conclusions of Theorem 1 \emph{mutatis mutandis}.

- Several applications of Theorem 1 to the theory of learning in games and to stochastic fictitious play are extensively investigated in [2] which we refer to for all these aspects. As far as we are concerned we will simply make a remark about some “natural” strategies which illustrates the theorem in an elementary way.

In the reward function at time \( k \), i.e. \( f_{x_k}(\bar{x}_{k-1}), \quad x_k \) represents the competitive term (“who will play?”) and \( \bar{x}_{k-1} \) represents a cooperative term (everybody’s past behaviour has influence on everybody’s reward).
This cooperative/competitive antagonism induces that in such a game a greedy competitive strategy is usually not optimal (when the players do not play a symmetric role). Let us be more specific. Assume for the sake of simplicity that $d + 1 = 2$ (two players). Then one may consider without loss of generality that $\bar{x}_n = \arg\max_{x} q_n$ i.e. that $\bar{x}_n$ is a $[0,1]$-valued real number. A greedy competitive strategy is defined by

$$f_{\bar{x}_n}(\bar{x}_{n-1}) = \max(f_0(\bar{x}_{n-1}), f_1(\bar{x}_{n-1}))$$

and it is clear that

$$f_{\bar{x}_n}(\bar{x}_{n-1}) - q(\bar{x}_{n-1}) = \max(f_0(\bar{x}_{n-1}), f_1(\bar{x}_{n-1})) - q(\bar{x}_{n-1}) =: \varphi(\bar{x}_{n-1}) \geq 0.$$ 

On the other hand, the proof of Theorem 1 implies that

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi(\bar{x}_k) \leq 0.$$ 

Hence, there is at least one weak limiting distribution $\mu_{\infty}$ of the sequence of empirical measures $\mu_n := \frac{1}{n} \sum_{0 \leq k \leq n-1} d_{\bar{x}_k}$ on the compact interval $[0,1]$ which is supported by the closed set $\{\varphi = 0\} \subset \{0,1\} \cup \{f_0 = f_1\}$; on the other hand supp($\mu_{\infty}$) is contained in the set $\bar{X}_{\infty}$ of the limiting values of the sequence $(\bar{x}_n)$ itself (in fact $\bar{X}_{\infty}$ is an interval since $(\bar{x}_n)_n$ is bounded and $\bar{x}_{n+1} - \bar{x}_n \to 0$). Hence $\bar{X}_{\infty} \cap (\{0,1\} \cup \{f_0 = f_1\}) \neq \emptyset$.

If the greedy strategy $(\bar{x}_n)_n$ is optimal then $\text{dist}(\bar{x}_n, \arg\max q) \to 0$ as $n \to \infty$ i.e. $\bar{X}_{\infty} \subset \arg\max q$. Consequently if

$$\arg\max q \cap (\{0,1\} \cup \{f_0 = f_1\}) = \emptyset$$

then the purely competitive strategy is never optimal for the group of two players.

Let us be more specific on the following example: set for two positive parameters $a \neq b$

$$f_0(x) := ax \quad \text{and} \quad f_1(x) := b(1-x), \quad x \in [0,1].$$

Then one checks that

$$\arg\max q = \{1/2\} \quad \text{and} \quad f_0(1/2) \neq f_1(1/2).$$

One first shows that the greedy strategy $x = (x_n)_{n \geq 1}$ defined by (3) satisfies

$$\bar{x}_n \to \frac{b}{a+b} \quad \text{and} \quad Q(x) = \frac{ab}{a+b} \quad \text{as} \quad n \to \infty.$$ 

On the other hand, any optimal (cooperative) strategy (like the i.i.d. Bernoulli$(1/2)$ one) yields an asymptotic (relative) global payoff rate

$$Q^* = \max_{[0,1]} q = \frac{a+b}{4}.$$ 

Note that $Q^* > \frac{ab}{(a+b)^2}$ since $a \neq b$. (When $a = b$ the greedy strategy becomes optimal.)

Now, if one looks at the individual performances (i.e. $\lim_n \frac{1}{n} \sum_{0 \leq k \leq n-1} f_i(\bar{x}_k)\mathbf{1}_{\{x_{k+1} = i\}}, i \in \{0,1\}$) of both players when the greedy strategy is played, one checks that:

- the “winner” of the game is player 1 if $b > a$ and player 0 if $a > b$,
- the asymptotic (relative) payoff rate of the winner is equal to $\frac{ab \max(a,b)}{(a+b)^2}$ (and $\frac{ab \min(a,b)}{(a+b)^2}$ for the “looser”).
If an optimal cooperative strategy is adopted by the players the “winner” remains the same but with an asymptotic payoff rate equal to \( \frac{\max(a,b)}{4} \) (the “looser” gets \( \frac{\min(a,b)}{4} \)). Consequently (when \( a \neq b \)), an optimal cooperative strategy always yields to the winner a strictly higher asymptotic payoff rate than the greedy one. This is also true for the looser.

- A more abstract version of Theorem 1 can be established using the same approach. The finite set \( \{0, 1, \ldots, d\} \) is replaced by a compact metric set \( K \), \( \mathcal{P}_{d+1} \) is replaced by the convex set \( \mathcal{P}_K \) of probability distributions on \( K \) equipped with the weak topology and the continuous function \( f : K \times \mathcal{P}_K \to \mathbb{R} \) is still supposed to derive from a potential function in some sense.

REFERENCES
