A two armed bandit type problem revisited
Gilles Pagès

To cite this version:
Gilles Pagès. A two armed bandit type problem revisited. ESAIM: Probability and Statistics, 2005, 9, pp.277-282. 10.1051/ps:2005017. hal-00004198

HAL Id: hal-00004198
https://hal.science/hal-00004198
Submitted on 9 Feb 2005

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A two armed bandit type problem revisited

Gilles Pagès *

February 9, 2005

Abstract

In [2] M. Benaïm and G. Ben Arous solve a multi-armed bandit problem arising in the theory of learning in games. We propose an short elementary proof of this result based on a variant of the Kronecker Lemma.

Key words: Two-armed bandit problem, Kronecker Lemma, learning theory, stochastic fictitious play.

In [2] a multi-armed bandit problem is addressed and investigated by M. Benaïm and G. Ben Arous. Let $f_0, \ldots, f_d$ denote $d + 1$ real-valued continuous functions defined on $[0, 1]^{d+1}$. Given a sequence $x = (x_n)_{n \geq 1} \in \{0, \ldots, d\}^{\mathbb{N}^*}$ (the strategy), set for every $n \geq 1$

$$\bar{x}_n := (\bar{x}_0, \bar{x}_1, \ldots, \bar{x}_d) \quad \text{with} \quad \bar{x}_i := \frac{1}{n} \sum_{k=1}^{n} 1_{\{x_k = i\}}, \quad i = 0, \ldots, d,$$

and

$$Q(x) = \liminf_{n \to +\infty} \frac{1}{n} \sum_{k=1}^{n-1} f_{x_k+1}(\bar{x}_k).$$

$(\bar{x}_0 := (\bar{x}_0^0, \bar{x}_0^1, \ldots, \bar{x}_0^d) \in [0, 1]^{d+1}, \bar{x}_0^0 + \cdots + \bar{x}_0^d = 1$ is a starting distribution). Imagine $d + 1$ players enrolled in a cooperative/competitive game with the following simple rules: if player $i \in \{0, \ldots, d\}$ plays at time $n$ he is rewarded by $f_i(\bar{x}_n)$, otherwise he gets nothing; only one player can play at the same time. Then the sequence $x$ is a playing strategy for the group of players and $Q(x)$ is the global cumulative worst payoff rate of the strategy $x$ for the whole community of players (regardless of the cumulative payoff rate of each player).

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

**What are the good strategies (for the group) ?**

---

*Laboratoire de Probabilités et Modélisation aléatoire, UMR 7599, Université Paris 6, case 188, 4, pl. Jussieu, F-75252 Paris Cedex 5. E-mail: gpa@ccr.jussieu.fr*
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.

Let \( S_d := \{ v \in [0,1]^d, \sum_{i=1}^d v_i \leq 1 \} \) and \( P_{d+1} := \{ u \in [0,1]^{d+1}, \sum_{i=1}^{d+1} u_i = 1 \} \). Furthermore, for notational convenience, set

\[
\forall v = (v_1, \ldots, v_d) \in S_d, \quad \tilde{v} := (1 - \sum_{i=1}^d v_i, v_1, \ldots, v_d) \in P_{d+1},
\]

\[
\forall u = (u_0, u_1, \ldots, u_d) \in P_{d+1}, \quad \hat{u} := (u_1, \ldots, u_d) \in S_d.
\]

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

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

**Theorem 1** Assume there is a function \( \Phi : S_d \to \mathbb{R} \), continuously differentiable on \( S_d^\circ \) having a continuous extension \( \nabla \Phi \) on \( S_d \) and satisfying:

\[
\forall v \in S_d, \quad \nabla \Phi(v) = (f_i(\tilde{v}) - f_0(\tilde{v}))_{1 \leq i \leq d}.
\]

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 \( \bar{x}_n \to \bar{x}_\infty \),

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

In particular there is no better strategy than choosing the player at random according to an i.i.d. strategy with distribution \( \bar{x}^* \in \arg \max q \).

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

\[
\liminf_{n \to +\infty} \frac{1}{n} \sum_{k=1}^n \frac{a_k}{b_k} \in \mathbb{R} \quad \implies \quad \liminf_{n \to +\infty} \frac{1}{b_n} \sum_{k=1}^n a_k \leq 0.
\]
Proof. Set $C_n = \sum_{k=1}^{n} \frac{a_k}{b_k}$, $n \geq 1$ 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

$$\frac{1}{b_n} \sum_{k=1}^{n} C_{k-1} \Delta b_k \geq \frac{1}{b_n} \sum_{k=1}^{n} C_{k-1} \Delta b_k + \frac{b_n - b_{n_\varepsilon}}{b_n} \left( \liminf_{k \to +\infty} C_k - \varepsilon \right).$$

Consequently, $\liminf_{n \to +\infty} C_n$ being finite, one concludes that

$$\liminf_{n \to +\infty} \frac{1}{b_n} \sum_{k=1}^{n} a_k \leq \liminf_{n \to +\infty} C_n - 0 - 1 \times \left( \liminf_{k \to +\infty} C_k - \varepsilon \right) = \varepsilon. \quad \diamond$$

Proof of Theorem 1. First note that for every $u = (u_0, \ldots, u_d) \in \mathcal{P}_{d+1}$,

$$q(u) := \sum_{i=0}^{d+1} u_i f_i(u) = f_0(u) + \sum_{i=1}^{d} u_i (f_i(u) - f_0(u))$$

so that

$$Q^* = \sup_{v \in \mathcal{S}_d} \left\{ f_0(\tilde{v}) + \sum_{i=1}^{d} v_i (f_i(\tilde{v}) - f_0(\tilde{v})) \right\} = \sup_{v \in \mathcal{S}_d} \left\{ f_0(\tilde{v}) + (v|\nabla \Phi(v)) \right\}.$$

Now, for every $k \geq 0$

$$f_{x_{k+1}}(\tilde{x}_k) - q(\tilde{x}_k) = \sum_{i=0}^{d} (f_i(\tilde{x}_k) \mathbf{1}_{\{x_{k+1}=i\}} - \tilde{x}_k^i f_i(\tilde{x}_k)) = \sum_{i=0}^{d} f_i(\tilde{x}_k) (1_{\{x_{k+1}=i\}} - \tilde{x}_k^i)$$

$$= \sum_{i=0}^{d} f_i(\tilde{x}_k) (k+1) \Delta \tilde{x}_{k+1}^i$$

$$= (k+1) \sum_{i=1}^{d} (f_i(\tilde{x}_k) - f_0(\tilde{x}_k)) \Delta \tilde{x}_{k+1}^i.$$

The last equality reads using Assumption (1),

$$f_{x_{k+1}}(\tilde{x}_k) - q(\tilde{x}_k) = (k+1)(\nabla \Phi(\tilde{x}_k)|\Delta \tilde{x}_{k+1})$$

3
Consequently, by the fundamental formula of calculus applied to $\Phi$ on $(\hat{x}_k, \hat{x}_{k+1}) \subset \hat{S}_d$,

$$\frac{1}{n} \sum_{k=0}^{n-1} f_{x_{k+1}}(\bar{x}_k) - q(\bar{x}_k) = \frac{1}{n} \sum_{k=0}^{n-1} (k + 1) (\Phi(\hat{x}_{k+1}) - \Phi(\hat{x}_k)) - R_n$$

with

$$R_n := \frac{1}{n} \sum_{k=0}^{n-1} \left( \nabla \Phi(\hat{x}_k) - \nabla \Phi(\hat{x}_k) \right) (k + 1) \Delta \hat{x}_{k+1}$$

and $\hat{x}_k \in (\hat{x}_k, \hat{x}_{k+1})$, $k = 1, \ldots n$. The fact that $|(k + 1)\Delta \hat{x}_{k+1}| \leq 1$ implies

$$|R_n| \leq \frac{1}{n} \sum_{k=0}^{n-1} w(\nabla \Phi, |\Delta \hat{x}_{k+1}|)$$

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 $\mathcal{S}_d$ that

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

Finally, the continuous function $\Phi$ being bounded on the compact set $\mathcal{S}_d$, the partial sums

$$\sum_{k=0}^{n-1} \Phi(\hat{x}_{k+1}) - \Phi(\hat{x}_k) = \Phi(\hat{x}_{n+1}) - \Phi(\hat{x}_0)$$

remain bounded as $n$ goes to infinity. Lemma 1 then implies that

$$\lim \inf_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n-1} (k + 1) (\Phi(\hat{x}_{k+1}) - \Phi(\hat{x}_k)) \leq 0.$$ 

One concludes by noting that on one hand

$$\limsup_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n-1} q(\bar{x}_k) \leq Q^* = \sup_{\mathcal{P}_{d+1}} q$$

and that, on the other hand, the function $q$ being continuous,

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

**Corollary 1** When $d + 1 = 2$ (two players), Assumption (1) is satisfied as soon as $f_0$ and $f_1$ are continuous on $\mathcal{P}_2$ and then the conclusions of Theorem 1 hold true.

**Proof:** This follows from the obvious fact that the continuous function $u_1 \mapsto f_1(1 - u_1, u_1) - f_0(1 - u_1, u_1)$ on $[0, 1]$ has an antiderivative. ♦

**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
\[
\Delta_{n+1} f_{x_{n+1}}(\bar{x}_n) \quad \text{instead of} \quad f_{x_{n+1}}(\bar{x}_n)
\]
where the weight sequence \( \Delta = (\Delta_n)_{n \geq 1} \) satisfies
\[
\Delta_n \geq 0, \quad n \geq 1, \quad S_n = \sum_{k=1}^{n} \Delta_k \to +\infty, \quad \frac{\Delta_n}{S_n} \to 0 \quad \text{as} \quad n \to \infty
\]
then the quantities \( \bar{x}^\Delta = (\bar{x}^\Delta_0, \ldots, \bar{x}^\Delta_d) \) with \( \bar{x}^\Delta_i = \frac{1}{S_n} \sum_{k=1}^{n-1} \Delta_k 1\{x_k = i\}, \ i = 0, \ldots, d, \ n \geq 1, \) and \( Q^\Delta(x) = \liminf_{n \to +\infty} \frac{1}{S_n} \sum_{k=0}^{n-1} \Delta_{k+1} f_{x_{k+1}}(\bar{x}^\Delta_k) \) satisfy all the conclusions of Theorem 1 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. \( f_{x_k}(\bar{x}_{k-1}) \), \( x_k \) 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 rôle). 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 \( \bar{x}_n = \hat{\bar{x}}_n \) i.e. that \( \bar{x}_n \) is a \([0, 1]\)-valued real number. A greedy competitive strategy is defined by
\[
\text{player 1 plays at time } n \ (i.e. \ x_n = 1) \text{ iff } f_1(\bar{x}_{n-1}) \geq f_0(\bar{x}_{n-1}) (2)
\]
i.e. the player with the highest reward is nominated to play. Note that such a strategy is anticipative from a probabilistic viewpoint. Then, for every \( n \geq 1, \)
\[
f_{x_n}(\bar{x}_{n-1}) = \max(f_0(\bar{x}_{n-1}), f_1(\bar{x}_{n-1}))
\]
and it is clear that
\[
f_{x_n}(\bar{x}_{n-1}) - q(\bar{x}_n) = \max(f_0(\bar{x}_{n-1}), f_1(\bar{x}_{n-1})) - q(\bar{x}_n) =: \varphi(\bar{x}_n) \geq 0.
\]
On the other hand, the proof of Theorem 1 implies that
\[
\liminf_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi(\bar{x}_n) \leq 0.
\]
Hence, there is at least one weak limiting distribution \( \bar{\mu}_\infty \) of the sequence of empirical measures \( \bar{\mu}_n := \frac{1}{n} \sum_{0 \leq k \leq n-1} \delta_{\bar{x}_k} \) which is supported by the closed set \( \{\varphi = 0\} \subset \{0, 1\} \cup \{f_0 = f_1\} \); on the other \( \text{supp}(\mu_\infty) \) 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 \( \text{dist}(\bar{x}_n, \text{argmax } q) \to 0 \) as \( n \to \infty \) i.e. \( \bar{X}_\infty \subset \text{argmax } q \). Consequently if

\[
\text{argmax } q \cap (\{0, 1\} \cup \{f_0 = f_1\}) = \emptyset \tag{3}
\]

then the purely competitive strategy is never optimal.

So is the case if

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

for some positive parameters \( a \neq b \), then

\[
\text{argmax } q = \{1/2\} \quad \text{and} \quad f_0(1/2) \neq f_1(1/2).
\]

In fact, one shows that the greedy strategy \( x = (x_n)_{n \geq 1} \) defined by (2) satisfies

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

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

- 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 \( K \), \( P_{d+1} \) is replaced by the convex set \( P_K \) of probability distributions on \( K \) equipped with the weak topology and the continuous function \( f : K \times P_K \to \mathbb{R} \) still derives from a potential function in some sense.

References

[1] M. Benaïm (1999). Dynamics of stochastic algorithms, in J. Azéma et al. eds, Séminaire de probabilités XXXIII, L.N. in Math. 1708, 1-68, Springer Verlag, Berlin.

[2] M. Benaïm, G. Ben Arous (2003). A two armed bandit type problem, Game Theory, 32(3), 3-16.