On the evolution of the expected gain of a greedy action in the bandit problem

Olivier Caelen and Gianluca Bontempi

Machine Learning Group, Département d'Informatique,
Université Libre de Bruxelles, Brussels, Belgium

Abstract. The K-armed bandit problem is a well-known formalization of the exploration versus exploitation dilemma. In a K-armed bandit problem, a player is confronted with a gambling machine with K arms where each arm is associated to an unknown gain distribution and the goal is to maximize the sum of the rewards. One of the simplest policies for this bandit problem is the greedy policy which keeps a gain estimation of the arms and at each round greedily chooses the arm which, on average, performed the best so far. This paper defines and gives an analytical definition of the *expected gain of a greedy action* μ_g and studies its evolution over the time. If the gambling machine has two arms, we derive an optimal exploration algorithm for the two arms case and we show that the evolution of μ_g under an exploitation greedy policy is upper-bounded. Otherwise if the player is confronted with a gambling machine with more than two arms we show experimentally that the μ_g evolution is much more complex and that the value of μ_g can decrease. Our results confirm analytically and experimentally that exploitation before acquiring enough knowledge on the arms is a bad practice.

In this paper, we study the well-known K-armed bandit problem, first introduced by Robbins [6]. In this learning problem, a casino player has to decide which arm of a K-slot machine to pull to maximize the total gain in a series of rounds. Each of the K arms of the slot machine returns a reward which is randomly distributed and unknown to the player. The player has to define a sequential selection policy on the basis of limited knowledge about the reward distributions which derive exclusively from previous results.

The K-armed bandit problem is a classical instance of situations involving a exploration/exploitation trade-off. An example of such a situation is the design of clinical trials to assess and compare a set of new medical treatments [4]. Here, the goal is to determine the best treatment while minimizing the inconveniences for the patients. Another example is the business problem of selecting the best supplier on the basis of incomplete information [2].

Several approaches have been proposed in literature to deal with the K-armed bandit problem. In this paper, we will focus on semi uniform algorithms [7]. A semi uniform algorithm is characterized by the alternation of two working modes, namely the exploration mode and the exploitation mode. The simplest semi uniform algorithm is the greedy policy in which the player keeps an updated estimation of the gains of the arms and at each round greedily chooses the arm which, on average, performed the best so far. If the player adopts the greedy policy, we say that he adopts a pure exploitation policy. This means that he uses current knowledge to select the seemingly best arm without reserving any time to explore what seems to be inferior arms. Exploitation is the right thing to do to maximize the short term gain of an action and exploration is the right thing to do to maximize gain in long term.

A possible alternative to pure exploitation is the ϵ-greedy policy [8] which preserves a fixed fraction of the time, e.g., quantified by a parameter ϵ, to perform random uniform selection. A variant of the ϵ-greedy is the ϵ-decreasing-greedy policy where the exploration rate is set initially to a high value and then gradually decreases. For specific conditions on the initial parameters of ϵ-decreasing-greedy, Auer et al. in [1] found an upper decreasing bound on the probability of selecting a sub-optimal arm. This means that after enough rounds, a player adopting the ϵ-decreasing-greedy policy has a high probability of always playing the best arm.

The probability that a greedy action selects the true best arm is well-known in simulation literature as the probability of correct selection (PCS) [5]. In that case the issue is to decide how many simulation trials should be conducted if we want to have a certain guarantee that the correct selection will be accomplished. In [3], we propose to use the PCS notion as a founding principle of a sequential strategy and as a measure of the effectiveness of an exploration step. In this paper, we extend existing work by (i) extending the notion of PCS to the notion of expected greedy reward and (ii) studying the behavior of the evolution in the time of this expected greedy reward.

This paper is structured as follows : Section 1 defines formally the bandit problem and the greedy action. An analytical definition of the expected greedy reward is given in Section 2 and the behavior of its evolution is studied in section 3.

1 The bandit problem

This section formally defines the K-armed bandit problem and introduces the notations used in this paper.

A K-armed bandit problem can be modeled by a set $\mathbf{z}_{\#} = \{\mathbf{z}_k\}$, $k = 1, \ldots, K$ of K random rewards[1] $\mathbf{z}_k$ with mean μ_k and standard deviation σ_k. Suppose that the goal of the player is to maximize the collected rewards. Once fixed the duration of the game to H rounds, at each round l, $l = 1, \ldots, H$ the player is expected to select an arm out of the K alternatives. Let $\mathbf{N}(l) = [\mathbf{n}_1(l), \ldots, \mathbf{n}_K(l)]$ be a *counting vector* whose kth term denotes the number of times that the kth arm was selected during the l first rounds and $\mathbf{Z}_k(l) = \left[\mathbf{z}_k^1,\ \mathbf{z}_k^2,\ \mathbf{z}_k^3,\ \ldots,\ \mathbf{z}_k^{\mathbf{n}_k(l)}\right]$ be the vector of identically and independently distributed observed rewards of the arm k up to time l.

[1] We use boldface symbols to denote random variables.

Based on the samples in the set $\{\mathbf{Z}_k(l)\}$, $k = 1, \dots, K$ and following a policy, the player selects iteratively one arm $\widehat{\mathbf{k}}$ at each round l. We define (i) the *observed state* $\widehat{\mathbf{s}}(l) \in \widehat{\mathcal{S}}$ of the game at the lth round as the whole set of observations $\{\mathbf{Z}_k(l)\}$, $k = 1, \dots, K$ and (ii) the *policy* as a function $\widehat{\pi} : \widehat{\mathcal{S}} \to \{1, \dots, K\}$ which returns for each state $\widehat{\mathbf{s}}$ the arm $\widehat{\mathbf{k}}$.

Two common adopted strategies are the pure random policy and the greedy policy. The pure random policy neglects any partial information about the state at time l and returns a selection

$$\widehat{\mathbf{k}}_r = \widehat{\pi}(\widehat{\mathbf{s}}(l)) \sim \text{Uni}(1/K, \dots, 1/K)$$

sampled according to an uniform distribution.

The greedy policy uses instead the information contained in $\widehat{\mathbf{s}}(l)$ and returns

$$\widehat{\mathbf{k}}_g = \arg\max \widehat{\boldsymbol{\mu}}_k(l)$$

where $\widehat{\boldsymbol{\mu}}_k(l)$ is the sampled average of the vector $\mathbf{Z}_k(l)$.

2 An analytical definition of the expected greedy reward

This section introduces and gives an analytical definition of the *expected gain of a greedy exploitation action* μ_g. Since $\mathbf{Z}_k(l)$ is a realization of a random vector, the state $\widehat{\mathbf{s}}$ and the output $\widehat{\mathbf{k}} = \widehat{\pi}(\widehat{\mathbf{s}})$ of the policy are random realizations too. At round l it is then possible to associate to a policy the *expected gain* $\mathbb{E}[\mu_{\widehat{\mathbf{k}}}]$ which quantifies the gain caused by the policy adoption at round l. Note that $\mu_{\widehat{\mathbf{k}}}$ is a random variable which denotes the mean of the arm $\widehat{\mathbf{k}}$ selected at the lth step.

Let

$$P_{\bar{k}}(l) = \text{Prob}\left\{\arg\max \widehat{\boldsymbol{\mu}}_k(l) = \bar{k}\right\} \tag{1}$$

be the probability that a greedy algorithm selects arm $\bar{k}$ at the lth step. The *expected gain of a greedy exploitation action* at round l is then defined as follows

$$\mathbb{E}[\mu_{\widehat{\mathbf{k}}}] = \sum_{k=1}^{K} P_k(l) \cdot \mu_k = \mu_g. \tag{2}$$

Note that $\mu_g < \mu_{k^*}$ where $k^* = \arg\max_k \mu_k$. Theorem 1 will show that an analytical expression of the probability $P_{\bar{k}}$ of selecting the $\bar{k}$ machine in a greedy strategy can be derived in case of a bandit problem with normally distributed arms.

Theorem 1.
Let $\mathbf{z}_{\#} = \{\mathbf{z}_1, \dots, \mathbf{z}_K\}$ be a set of $K > 1$ normal reward distributions $\mathbf{z}_k \sim \mathcal{N}[\mu_k, \sigma_k]$ with mean μ_k and standard deviation σ_k. If the selection policy is greedy then the probability of selecting $\mathbf{z}_{\bar{k}}$ is

$$P_{\bar{k}} = \mathit{Prob}\left\{\, \widehat{\mathbf{r}}_1 > 0\ ,\ \dots\ ,\ \widehat{\mathbf{r}}_{\bar{k}-1} > 0\ ,\ \widehat{\mathbf{r}}_{\bar{k}+1} > 0\ ,\ \dots\ ,\ \widehat{\mathbf{r}}_K > 0\ \right\}$$

where $\left(\widehat{\mathbf{r}}_1, \ldots, \widehat{\mathbf{r}}_{\bar{k}-1}, \widehat{\mathbf{r}}_{\bar{k}+1}, \ldots, \widehat{\mathbf{r}}_K\right)^T$ *follows a multivariate normal distribution*

$$\left(\widehat{\mathbf{r}}_1, \ldots, \widehat{\mathbf{r}}_{\bar{k}-1}, \widehat{\mathbf{r}}_{\bar{k}+1}, \ldots, \widehat{\mathbf{r}}_K\right)^T \sim \mathcal{N}\left[\mathit{\Gamma}, \mathit{\Sigma}\right]$$

with mean

$$\mathit{\Gamma} = \begin{pmatrix} \mu_{\bar{k}} - \mu_1 \\ \vdots \\ \mu_{\bar{k}} - \mu_{\bar{k}-1} \\ \mu_{\bar{k}} - \mu_{\bar{k}+1} \\ \vdots \\ \mu_{\bar{k}} - \mu_K \end{pmatrix},$$

and covariance matrix $\mathit{\Sigma}$

$$\mathit{\Sigma} = \begin{pmatrix} \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} + \frac{\sigma_1^2}{n_1} & \cdots & \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} & \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} & \cdots & \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} & \cdots & \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} + \frac{\sigma_{\bar{k}-1}^2}{n_{\bar{k}-1}} & \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} & \cdots & \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} \\ \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} & \cdots & \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} & \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} + \frac{\sigma_{\bar{k}+1}^2}{n_{\bar{k}+1}} & \cdots & \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} & \cdots & \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} & \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} & \cdots & \frac{\sigma_{\bar{k}}^2}{n_{\bar{k}}} + \frac{\sigma_K^2}{n_K} \end{pmatrix}$$

where n_k *is the number of observations of* $\mathbf{z}_k$*.*

Proof. According to (1), $P_{\bar{k}}$ is the probability that $\widehat{\boldsymbol{\mu}}_{\bar{k}}$ is the maximum in $\{\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_K\}$:

$$\begin{aligned} P_{\bar{k}} &= \text{Prob}\left\{\widehat{\mathbf{k}}_g = \bar{k}\right\} \\ &= \text{Prob}\left\{\bar{k} = \arg \max_{k \in [1 \ldots K]} \{\widehat{\boldsymbol{\mu}}_k\}\right\} \\ &= \text{Prob}\left\{\ \widehat{\boldsymbol{\mu}}_{\bar{k}} > \widehat{\boldsymbol{\mu}}_1\ ,\ \ldots,\ \widehat{\boldsymbol{\mu}}_{\bar{k}} > \widehat{\boldsymbol{\mu}}_{\bar{k}-1}\ ,\ \widehat{\boldsymbol{\mu}}_{\bar{k}} > \widehat{\boldsymbol{\mu}}_{\bar{k}+1}\ ,\ \ldots,\ \widehat{\boldsymbol{\mu}}_{\bar{k}} > \widehat{\boldsymbol{\mu}}_K\ \right\} \\ &= \text{Prob}\left\{\ \widehat{\mathbf{r}}_1 > 0\ ,\ \ldots\ ,\ \widehat{\mathbf{r}}_{\bar{k}-1} > 0\ ,\ \widehat{\mathbf{r}}_{\bar{k}+1} > 0\ ,\ \ldots\ ,\ \widehat{\mathbf{r}}_K > 0\ \right\}, \end{aligned}$$

where $\widehat{\mathbf{r}}_k = \widehat{\boldsymbol{\mu}}_{\bar{k}} - \widehat{\boldsymbol{\mu}}_k$. It follows that $P_{\bar{k}}$ denotes also the probability that all the components of the vector $\left(\widehat{\mathbf{r}}_1, \ldots, \widehat{\mathbf{r}}_{\bar{k}-1}, \widehat{\mathbf{r}}_{\bar{k}+1}, \ldots, \widehat{\mathbf{r}}_K\right)^T$ are positive. Under the assumption of Gaussianity, this vector is a multivariate normal random variable

$$\begin{pmatrix} \widehat{\mathbf{r}}_1 \\ \vdots \\ \widehat{\mathbf{r}}_{\bar{k}-1} \\ \widehat{\mathbf{r}}_{\bar{k}+1} \\ \vdots \\ \widehat{\mathbf{r}}_K \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\mu}}_{\bar{k}} - \widehat{\boldsymbol{\mu}}_1 \\ \vdots \\ \widehat{\boldsymbol{\mu}}_{\bar{k}} - \widehat{\boldsymbol{\mu}}_{\bar{k}-1} \\ \widehat{\boldsymbol{\mu}}_{\bar{k}} - \widehat{\boldsymbol{\mu}}_{\bar{k}+1} \\ \vdots \\ \widehat{\boldsymbol{\mu}}_{\bar{k}} - \widehat{\boldsymbol{\mu}}_K \end{pmatrix} \sim \mathcal{N}\left[\mathit{\Gamma}, \mathit{\Sigma}\right]$$

with mean vector

$$\mathit{\Gamma} = \begin{pmatrix} \mu_{\bar{k}} - \mu_1 \\ \vdots \\ \mu_{\bar{k}} - \mu_{\bar{k}-1} \\ \mu_{\bar{k}} - \mu_{\bar{k}+1} \\ \vdots \\ \mu_{\bar{k}} - \mu_K \end{pmatrix},$$

and covariance matrix

$$\mathit{\Sigma} = \begin{pmatrix} \sigma^2_{\widehat{\mathbf{r}}_1,\widehat{\mathbf{r}}_1} & \cdots & \sigma^2_{\widehat{\mathbf{r}}_1,\widehat{\mathbf{r}}_{\bar{k}-1}} & \sigma^2_{\widehat{\mathbf{r}}_1,\widehat{\mathbf{r}}_{\bar{k}+1}} & \cdots & \sigma^2_{\widehat{\mathbf{r}}_1,\widehat{\mathbf{r}}_K} \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ \sigma^2_{\widehat{\mathbf{r}}_{\bar{k}-1},\widehat{\mathbf{r}}_1} & \cdots & \sigma^2_{\widehat{\mathbf{r}}_{\bar{k}-1},\widehat{\mathbf{r}}_{\bar{k}-1}} & \sigma^2_{\widehat{\mathbf{r}}_{\bar{k}-1},\widehat{\mathbf{r}}_{\bar{k}+1}} & \cdots & \sigma^2_{\widehat{\mathbf{r}}_{\bar{k}-1},\widehat{\mathbf{r}}_K} \\ \sigma^2_{\widehat{\mathbf{r}}_{\bar{k}+1},\widehat{\mathbf{r}}_1} & \cdots & \sigma^2_{\widehat{\mathbf{r}}_{\bar{k}+1},\widehat{\mathbf{r}}_{\bar{k}-1}} & \sigma^2_{\widehat{\mathbf{r}}_{\bar{k}+1},\widehat{\mathbf{r}}_{\bar{k}+1}} & \cdots & \sigma^2_{\widehat{\mathbf{r}}_{\bar{k}+1},\widehat{\mathbf{r}}_K} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{\widehat{\mathbf{r}}_K,\widehat{\mathbf{r}}_1} & \cdots & \sigma^2_{\widehat{\mathbf{r}}_K,\widehat{\mathbf{r}}_{\bar{k}-1}} & \sigma^2_{\widehat{\mathbf{r}}_K,\widehat{\mathbf{r}}_{\bar{k}+1}} & \cdots & \sigma^2_{\widehat{\mathbf{r}}_K,\widehat{\mathbf{r}}_K} \end{pmatrix}$$

Now, since $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\boldsymbol{\mu}}_j$ are independent for $i \neq j$

$$\sigma^2_{\widehat{\mathbf{r}}_j,\widehat{\mathbf{r}}_j} = \mathrm{Var}\left(\widehat{\boldsymbol{\mu}}_{\bar{k}} - \widehat{\boldsymbol{\mu}}_j\right) = \frac{\sigma^2_{\bar{k}}}{n_{\bar{k}}} + \frac{\sigma^2_j}{n_j},$$

it follows that

$$\begin{aligned} \sigma^2_{\widehat{\mathbf{r}}_i,\widehat{\mathbf{r}}_j} = \sigma^2_{\widehat{\mathbf{r}}_j,\widehat{\mathbf{r}}_i} &= \mathrm{cov}\left[\widehat{\boldsymbol{\mu}}_{\bar{k}} - \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\mu}}_{\bar{k}} - \widehat{\boldsymbol{\mu}}_j\right] \\ &= \mathbb{E}\big[(\widehat{\boldsymbol{\mu}}_{\bar{k}} - \widehat{\boldsymbol{\mu}}_i - \mathbb{E}[\widehat{\boldsymbol{\mu}}_{\bar{k}} - \widehat{\boldsymbol{\mu}}_i]) \cdot \big(\widehat{\boldsymbol{\mu}}_{\bar{k}} - \widehat{\boldsymbol{\mu}}_j - \mathbb{E}\big[\widehat{\boldsymbol{\mu}}_{\bar{k}} - \widehat{\boldsymbol{\mu}}_j\big]\big)\big] \\ &= \mathbb{E}\big[(\widehat{\boldsymbol{\mu}}_{\bar{k}})^2\big] - \mu^2_{\bar{k}} \\ &= \frac{\sigma^2_{\bar{k}}}{n_{\bar{k}}}. \end{aligned}$$

□

3 The evolution of the expected greedy reward

A major difficulty of the bandit problem is the dynamic and multivariate nature of the terms involved in the definition of optimal policy. For this reason, it is important to study how the *expected reward of a greedy action* μ_g (2) changes with time.

Our analysis will be made in two parts. In the first one we assume that $\mathbf{z}_\#$ contains only two arms where k^* is the index of the best arm and $\bar{k}^*$ is the other index. This is the simplest case and some analytical results can be derived about the evolution in time of μ_g. The second part will analyse the most complex case, i.e. when $K > 2$.

3.1 The evolution of μ_g in the case where $K = 2$

First, we assume that the bandit machine has two arms. Theorem 2 shows that if $K = 2$ then testing $\mathbf{z}_1$ or $\mathbf{z}_2$ at round l will, in both cases, improve the value of μ_g at round $l + 1$. Theorem 3 proposes an optimal exploration policy which maximizes μ_g at round $l + 1$. Note that by definition, $\mu_g < \mu_{k^*}$. Theorem 4 shows that the evolution of μ_g of any strategy which always tests the same arm is upper-bounded by a value ub^{∞} which is smaller than μ_{k^*}. This is an argument in favor of exploration.

Theorem 2. *Let $K = 2$ and $\mu_g^{l+1}(k)$ be the next expected reward of a greedy action when arm k is tested at round l. Then*

$$\forall k \in \{1, 2\}, \ \mu_g^{l+1}(k) > \mu_g.$$

Proof. $\forall k \in \{1, 2\}, \mu_g^{l+1}(k) > \mu_g$ if and only if $\forall k \in \{1, 2\}, P_{k^*}^{l+1}(k) > P_{k^*}$ where $P_{k^*}^{l+1}(k)$ is the probability of selecting the best alternative through a pure greedy algorithm at round $l + 1$ if $\mathbf{z}_k$ is tested at round l.
By definition

$$P_{k^*} = \text{Prob}\left\{\widehat{\boldsymbol{\mu}}_{k^*} - \widehat{\boldsymbol{\mu}}_{\bar{k}^*} > 0\right\}$$

and, under the assumption of Gaussianity,

$$\left(\widehat{\boldsymbol{\mu}}_{k^*} - \widehat{\boldsymbol{\mu}}_{\bar{k}^*}\right) \sim \mathcal{N}\left(\mu_{k^*} - \mu_{\bar{k}^*}, \frac{\sigma_{k^*}^2}{n_{k^*}} + \frac{\sigma_{\bar{k}^*}^2}{n_{\bar{k}^*}}\right).$$

The mean of the Gaussian is positive. Reducing the variance of $(\widehat{\boldsymbol{\mu}}_{k^*} - \widehat{\boldsymbol{\mu}}_{\bar{k}^*})$ will improve the P_{k^*} at round $l + 1$ (see fig. 1). Testing $\mathbf{z}_{k^*}$ or $\mathbf{z}_{\bar{k}^*}$ will both reduce the variance of $(\widehat{\boldsymbol{\mu}}_{k^*} - \widehat{\boldsymbol{\mu}}_{\bar{k}^*})$, this proves the proposition. □

Theorem 3. *Let us consider a bandit problem with two normal arms $\mathbf{z}_{\#} = \{\mathbf{z}_{k_1}, \mathbf{z}_{k_2}\}$. The following policy*

$$\begin{cases} \textit{if } N_\Delta < 0 \textit{ then test } \mathbf{z}_{k_1} \\ \textit{if } N_\Delta > 0 \textit{ then test } \mathbf{z}_{k_2} \\ \textit{if } N_\Delta = 0 \textit{ then test either } \ \mathbf{z}_{k_1} \textit{ or } \mathbf{z}_{k_2} \end{cases}$$

will maximize μ_g at round $l + 1$ where

$$N_\Delta = n_{k_1} \cdot (n_{k_1} + 1) \cdot \left(\sigma_{k_2}^2 - \sigma_{k_1}^2\right) + \sigma_{k_1}^2 \cdot (n_{k_2} + n_{k_1} + 1) \cdot (n_{k_1} - n_{k_2}).$$

Proof. If $\left[\mu_g^{l+1}(k_1) - \mu_g^{l+1}(k_2)\right]$ is positive then $\mathbf{z}_1$ must be tested to maximize μ_g at round $l+1$ and if $\left[\mu_g^{l+1}(k_1) - \mu_g^{l+1}(k_2)\right]$ is negative then $\mathbf{z}_2$ must be tested.

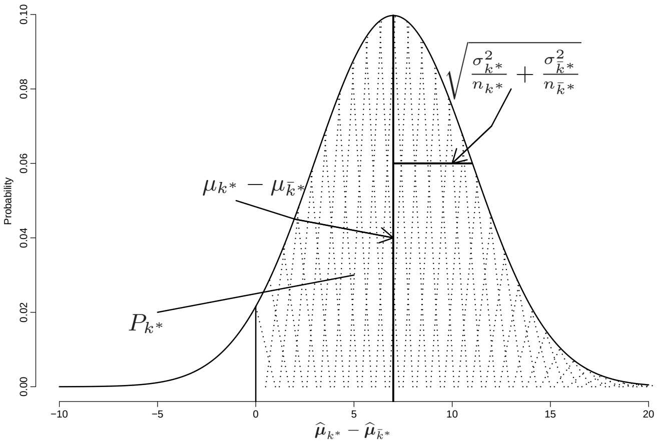

Fig. 1. If K=2, the probability of selecting the best random variable P_{k^*} is the surface under the Gaussian when the abscissa takes his value in the set $[0, +\infty]$. Note that the mean of the Gaussian is always positive.

$$\begin{aligned}
&\mu_g^{l+1}(k_1) - \mu_g^{l+1}(k_2) \\
&= P_{k^*}^{l+1}(k_1) \cdot \mu_{k^*} + P_{\bar{k}^*}^{l+1}(k_1) \cdot \mu_{\bar{k}^*} - P_{k^*}^{l+1}(k_2) \cdot \mu_{k^*} - P_{\bar{k}^*}^{l+1}(k_2) \cdot \mu_{\bar{k}^*} \\
&= \left[P_{k^*}^{l+1}(k_1) - P_{k^*}^{l+1}(k_2)\right] \cdot \mu_{k^*} + \left[P_{\bar{k}^*}^{l+1}(k_1) - P_{\bar{k}^*}^{l+1}(k_2)\right] \cdot \mu_{\bar{k}^*} \\
&= \left[P_{k^*}^{l+1}(k_1) - P_{k^*}^{l+1}(k_2)\right] \cdot \mu_{k^*} - \left[P_{k^*}^{l+1}(k_1) - P_{k^*}^{l+1}(k_2)\right] \cdot \mu_{\bar{k}^*} \\
&= \left[P_{k^*}^{l+1}(k_1) - P_{k^*}^{l+1}(k_2)\right] \cdot (\mu_{k^*} - \mu_{\bar{k}^*}) \\
&= \mathit{\Delta} P_{k^*} \cdot (\mu_{k^*} - \mu_{\bar{k}^*}) \,.
\end{aligned}$$

Since $(\mu_{k^*} - \mu_{\bar{k}^*}) > 0$, the sign of $\left[\mu_g^{l+1}(k_1) - \mu_g^{l+1}(k_2)\right]$ is the same as the sign of $\mathit{\Delta} P_{k^*}$.

Let $\mathrm{Var}^{k_1}\left(\widehat{\boldsymbol{\mu}}_{k^*}^{l+1} - \widehat{\boldsymbol{\mu}}_{\bar{k}^*}^{l+1}\right)$ and $\mathrm{Var}^{k_2}\left(\widehat{\boldsymbol{\mu}}_{k^*}^{l+1} - \widehat{\boldsymbol{\mu}}_{\bar{k}^*}^{l+1}\right)$ be respectively the variance of $\left(\widehat{\boldsymbol{\mu}}_{k^*}^{l+1} - \widehat{\boldsymbol{\mu}}_{\bar{k}^*}^{l+1}\right)$ at $l+1$ if either $\mathbf{z}_{k_1}$ or $\mathbf{z}_{k_2}$ is tested at l. $\mathit{\Delta}\mathrm{Var}$ is defined as the difference between the two variances at $l+1$. A reduction of the variance of $\left(\widehat{\boldsymbol{\mu}}_{k^*}^{l+1} - \widehat{\boldsymbol{\mu}}_{\bar{k}^*}^{l+1}\right)$ improves the probability of selecting the best random variable

$\mathbf{z}_{k^*}$. The sign of $\Delta\mathrm{Var}$ is thus the opposite of the sign of ΔP_{k^*}.

$$\begin{aligned}\Delta\mathrm{Var} &= \mathrm{Var}^{k_1}\left(\widehat{\boldsymbol{\mu}}^{l+1}_{k^*} - \widehat{\boldsymbol{\mu}}^{l+1}_{\bar{k}^*}\right) - \mathrm{Var}^{k_2}\left(\widehat{\boldsymbol{\mu}}^{l+1}_{k^*} - \widehat{\boldsymbol{\mu}}^{l+1}_{\bar{k}^*}\right)\\ &= \frac{\sigma^2_{k_1}}{n_{k_1}+1} + \frac{\sigma^2_{k_2}}{n_{k_2}} - \frac{\sigma^2_{k_1}}{n_{k_1}} - \frac{\sigma^2_{k_2}}{n_{k_2}+1}\\ &= \frac{N_\Delta}{D_\Delta}\end{aligned}$$

where

$$\begin{aligned}N_\Delta &= n_{k_2} n_{k_1}\left(n_{k_2}+1\right)\sigma^2_{k_1} + \left(n_{k_1}+1\right) n_{k_1}\left(n_{k_2}+1\right)\sigma^2_{k_2}\\ &\quad - \left(n_{k_1}+1\right) n_{k_2}\left(n_{k_2}+1\right)\sigma^2_{k_1} - \left(n_{k_1}+1\right) n_{k_2} n_{k_1}\sigma^2_{k_2}\\ &= n_{k_1}\left(\sigma^2_{k_2} - \sigma^2_{k_1}\right)\left(n_{k_1}+1\right) + \sigma^2_{k_1}\left(n_{k_1} - n_{k_2}\right)\left(n_{k_1}+n_{k_2}+1\right)\\ D_\Delta &= \left(n_{k_1}+1\right) n_{k_2} n_{k_1}\left(n_{k_2}+1\right)\end{aligned}$$

Since $D_\Delta > 0$, the following policy

$$\begin{cases}\text{if } N_\Delta < 0 \text{ then test } \mathbf{z}_{k_1}\\ \text{if } N_\Delta > 0 \text{ then test } \mathbf{z}_{k_2}\\ \text{if } N_\Delta = 0 \text{ then test either } \mathbf{z}_{k_1} \text{ or } \mathbf{z}_{k_2}\end{cases}$$

maximizes μ_g at round $l+1$.

□

Theorem 4. *Let $\mathbf{z}_\#$ be a set containing two normal random variables $\mathbf{z}_{k_1}$ and $\mathbf{z}_{k_2}$. Suppose that the player adopts a bandit strategy which always tests the same arm $\mathbf{z}_{k_1}$. $\mathbf{z}_{k_2}$ is the other arm which is never tested. If*

$$\mathbf{X} \sim \mathcal{N}\left(\mu_{k_1} - \mu_{k_2}, \frac{\sigma^2_{k_2}}{n_{k_2}}\right)$$

then

$$\forall l \geq 1\ , \quad \mu_g < ub^\infty < \mu_{k^*},$$

where

$$ub^\infty = \left(\mu_{k_1} - \mu_{k_2}\right) \cdot \mathit{Prob}\left\{\mathbf{X} > 0\right\} + \mu_{k_2}.$$

Proof. Let τ be a positive integer, we can find an upper-bound for μ_g (see Theorem 2) :

$$\begin{aligned}\mu_g &< \mu^{l+1}_g\left(k_1\right)\\ &\leq \mu^{l+\tau}_g\left(k_1\right).\end{aligned}$$

The future *expected greedy gain* when arm k_1 is tested $\tau \geq 1$ times is $\mu_g^{l+\tau}(k_1)$ and is analytically defined as follows

$$\begin{aligned}\mu_g^{l+\tau}(k_1) &= \mu_{k_1} \cdot \text{Prob}\left\{\left(\widehat{\boldsymbol{\mu}}_{k_1} - \widehat{\boldsymbol{\mu}}_{k_2}\right)_\tau > 0\right\} + \mu_{k_2} \cdot \left(1 - \text{Prob}\left\{\left(\widehat{\boldsymbol{\mu}}_{k_1} - \widehat{\boldsymbol{\mu}}_{k_2}\right)_\tau > 0\right\}\right) \\ &= (\mu_{k_1} - \mu_{k_2}) \cdot \text{Prob}\left\{\left(\widehat{\boldsymbol{\mu}}_{k_1} - \widehat{\boldsymbol{\mu}}_{k_2}\right)_\tau > 0\right\} + \mu_{k_2} \\ &= ub^\tau\end{aligned}$$

where

$$\left(\widehat{\boldsymbol{\mu}}_{k_1} - \widehat{\boldsymbol{\mu}}_{k_2}\right)_\tau \sim \mathcal{N}\left(\mu_{k_1} - \mu_{k_2}, Var_\tau\right)$$

and where

$$Var_\tau = \frac{\sigma_{k_1}^2}{n_{k_1} + \tau} + \frac{\sigma_{k_2}^2}{n_{k_2}}.$$

If the player tests infinitely the arm $\mathbf{z}_{k_1}$ without reserving any round to explore $\mathbf{z}_{k_2}$ then the variance of $\left(\widehat{\boldsymbol{\mu}}_{k_1} - \widehat{\boldsymbol{\mu}}_{k_2}\right)_\tau$ converges

$$\lim_{\tau \to \infty} Var_\tau = \lim_{\tau \to \infty} \left(\frac{\sigma_{k_1}^2}{n_{k_1} + \tau} + \frac{\sigma_{k_2}^2}{n_{k_2}}\right) = \frac{\sigma_{k_2}^2}{n_{k_2}}$$

and we define the random variable $\mathbf{X}$ as follows

$$\mathbf{X} = \left(\widehat{\boldsymbol{\mu}}_{k_1} - \widehat{\boldsymbol{\mu}}_{k_2}\right)_\infty \sim \mathcal{N}\left(\mu_{k_1} - \mu_{k_2}, \frac{\sigma_{k_2}^2}{n_{k_2}}\right).$$

The upper-bound of μ_g is defined as follows

$$ub^\infty = (\mu_{k_1} - \mu_{k_2}) \cdot \text{Prob}\left\{\mathbf{X} > 0\right\} + \mu_{k_2}$$

and we have $\mu_g < ub^\infty$. ub^∞ reaches μ_{k^*} only if $P_{k^*} = 1$ and P_{k^*} equals one only if $\frac{\sigma_{k_1}^2}{n_{k_1}} + \frac{\sigma_{k_2}^2}{n_{k_2}} = 0$ (see figure 1). In the case of a bandit strategy which never tests $\mathbf{z}_{k_2}$, the variance $\frac{\sigma_{k_1}^2}{n_{k_1}} + \frac{\sigma_{k_2}^2}{n_{k_2}}$ converges to $\frac{\sigma_{k_2}^2}{n_{k_2}}$ and we have thus $ub^\infty < \mu_{k^*}$. □

Theorem 2 shows that, whatever the arm selected, the value of μ_g is always increasing. Theorem 3 proposes, knowing the characteristics of the arms, a new exploration strategy and Theorem 4 shows that the evolution of μ_g of any bandit strategy which never explores the arms and which always tests the same arm, is upper-bounded by ub^∞ which is smaller than μ_{k^*}. We will now empirically test these properties on six synthetic problems (table 1) with horizon $H = 50$. Three algorithms are compared. At each round l, each of the three algorithms tests arm $\widetilde{k}$. The three algorithms are (i) an oracle greedy exploitation algorithm which always tests the best arm $\left[\widetilde{k} = k^*\right]$, (ii) a random exploration algorithm which periodically tests the two arms $\left[\widetilde{k} = (l \mod 2) + 1\right]$ and (iii) the optimal oracle exploration algorithm described in theorem 3 and called optiK2

	case 1		*case 2*		*case 3*		*case 4*		*case 5*		*case 6*	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
$\mathbf{z}_1$	0	2	0	1	0	3	0	2	0	0.01	0	0.01
$\mathbf{z}_2$	1	2	1	2	1	1	1	0.01	1	2	1	0.01

Table 1. Six synthetic problems to estimate the evolution of μ_g where $K = 2$ under (i) an oracle greedy algorithm, (ii) a random exploration algorithm and (iii) the optiK2 algorithm. For each case, the first and the second column contain respectively the mean and the standard deviation of the random variables.

$\left[\text{if } N_\Delta < 0 \text{ then } \widetilde{k} = 1 \text{ else } \widetilde{k} = 2\right]$. Initially, n_1 and n_2 are equal to one. At each round l and following one of the three algorithms, the value of $n_{\widetilde{k}}$ is increased by one and $\mu_g(l)$ (see eq. (2)) is computed. The evolution of μ_g in the time for the six synthetic problems are shown in figure 2.

In all the cases, the curve of $\mu_g(l)$ is increasing (see Theorem 2) and optiK2 is optimal (see Theorem 3). In case 1, σ_1 is equal to σ_2 such that $N_\Delta = \sigma_1^2(n_1 - n_2)(n_1 + n_2 + 1)$ and optiK2 is then equivalent to a random exploration strategy. In case 5, the standard deviation of the worst arm is very small such that the term $n_1(\sigma_2^2 - \sigma_1^2)(n_1 + 1)$ in N_Δ is very high and optiK2 is equivalent to an oracle greedy strategy which always tests $\mathbf{z}_2$. Note that, with a random exploration strategy, we have

$$\lim_{l \to \infty} \sqrt{\frac{\sigma_{k^*}^2}{n_{k^*}} + \frac{\sigma_{\bar{k}^*}^2}{n_{\bar{k}^*}}} = 0$$

and thus (see figure 1)

$$\lim_{l \to \infty} P_{k^*} = 1.$$

This means that playing a great number of random exploration actions guarantees that μ_g will always be improved and that, asymptotically, the value of μ_g will reach the value of μ_{k^*}. In the case of bandit strategies which always test the same arm (like the oracle greedy strategy), we are guaranteed that μ_g will always be improved but μ_g is upper-bounded by ub^∞ which is smaller than μ_{k^*} (see Theorem 4). Without exploration, the evolution of the gain of a greedy action is upper-bounded by ub^∞ and the only way to increase this upper-bound is to explore the other arm.

3.2 The evolution of μ_g in the case where $K > 2$

Consider now the case where $K > 2$ and let $\mu_g^{l+1}(k)$ be the next greedy expected reward when arm k is tested at round l. In this case, we will empirically show that the evolution of μ_g is more complex and that the monotone property $\mu_g^{l+1}(k) > \mu_g$ shown in Theorem 2 does not always hold. We will consider the evolution of μ_g under a pure exploitation and a pure exploration policy.

First, consider the evolution of μ_g under a pure oracle exploitation greedy algorithm which knows the index k^* and always tests this best arm. We will show that in some cases, mainly when the $\frac{\sigma}{\sqrt{n}}$ of the bad alternatives is high,

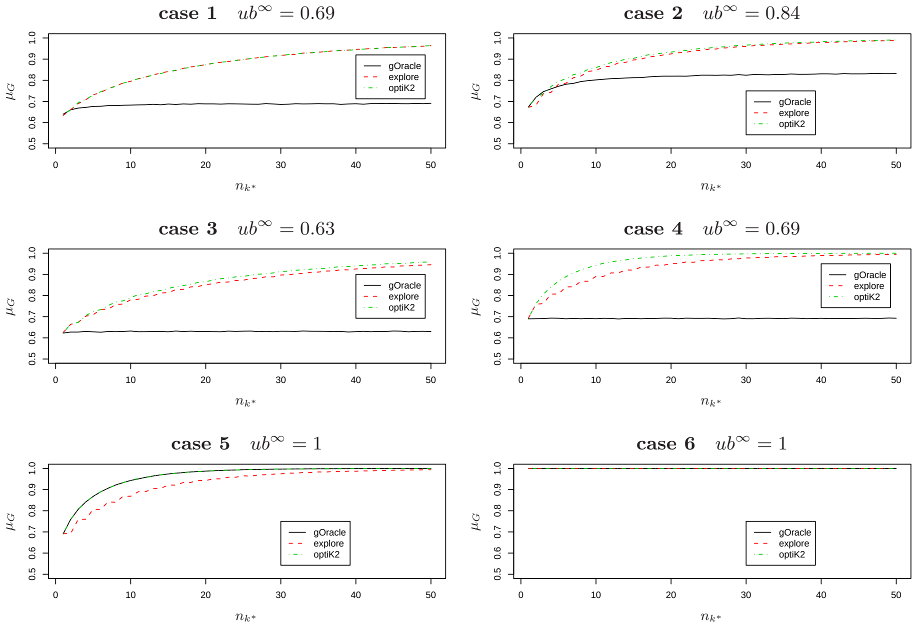

Fig. 2. Evolution of μ_g for the six synthetic problems under an oracle greedy algorithm, random exploration algorithm and the optiK2 algorithm. For each case, the value of the upper-bound ub^{∞} is given.

the value of $\mu_g^{l+1}(k) - \mu_g$ can be negative such that an oracle greedy action can reduce the value of μ_g at the next round $l+1$.

Six synthetic example cases (see table 2 and fig. 3) are tested in which n_{k^*} takes its value in $[1, 200]$ and where σ_{k^*} is always equal to 10. The first three cases concern problems where $K = 3$ and in the last three cases $K = 4$. In the cases 1, 2 and 5 the $\frac{\sigma}{\sqrt{n}}$ of the worst arms has the same value, a small value in case 1 and in case 5 and a higher value in case 2. The plots of the cases 1 and 5 show that increasing the n_{k^*} of the best alternative, increases μ_g if the $\frac{\sigma}{\sqrt{n}}$ of the bad arms is small. Case 2 is the opposite : if the $\frac{\sigma}{\sqrt{n}}$ of the bad arms is high, then increasing the number of tests made on the best alternative, decreases μ_g. In the cases 3, 4 and 6 the $\frac{\sigma}{\sqrt{n}}$ of the worst alternatives is not the same and the corresponding plots in figure 3 are more non-linear. At the beginning (i.e. when the $\frac{\sigma_{k^*}}{\sqrt{n_{k^*}}}$ is high), testing the best $\mathbf{z}_{k^*}$ will degrade μ_g and when $\frac{\sigma_{k^*}}{\sqrt{n_{k^*}}}$ is small enough, testing $\mathbf{z}_{k^*}$ will improve μ_g. Note that, like in case 4, this improvement of μ_g can be very small. The two last cases are example cases where the means of the bad alternatives are different.

Note that in the cases 1, 2, 3 and 4, we have $\mu_{k^*} = 1$ and the other means equal zero. In these four first cases, the expected gain of a greedy action μ_g equals thus the probability of correct selection P_{k^*}. In case 1, the P_{k^*} evolution,

Oracle exploitation greedy problem $K > 2$												
	case 1		*case 2*		*case 3*		*case 4*		*case 5*		*case 6*	
	μ	$\frac{\sigma}{\sqrt{n}}$	μ	$\frac{\sigma}{\sqrt{n}}$	μ	$\frac{\sigma}{\sqrt{n}}$	μ	$\frac{\sigma}{\sqrt{n}}$	μ	$\frac{\sigma}{\sqrt{n}}$	μ	$\frac{\sigma}{\sqrt{n}}$
$\mathbf{z}_1$	0	0.0001	0	5	0	1	0	0.0001	0	0.0001	0	0.0001
$\mathbf{z}_2$	0	0.0001	0	5	0	10	0	4	0.2	0.0001	0.2	4
$\mathbf{z}_3$	1	·	1	·	1	·	0	6	0.7	0.0001	0.7	6
$\mathbf{z}_4$							1	·	1	·	1	·

Table 2. Six synthetic problems to estimate the evolution of μ_g under an oracle exploitation greedy algorithm where $K > 2$. For each case, the first column contains the mean of the random variable and the second one contains the standard deviation of the sample average.

under an oracle greedy algorithm, is monotone increasing and in case 2 , P_{k^*} is monotone decreasing. In the cases 3 and 4, the P_{k^*} evolution is more non-linear. Given a set of K random variables $\{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$, where n_k is the number of observed rewards from $\mathbf{z}_k$ and where $\widehat{\boldsymbol{\mu}}_k$ is the corresponding sample average of $\mathbf{z}_k$. We have

$$P_{k^*} = \text{Prob}\left\{\ \widehat{\boldsymbol{\mu}}_{k^*} > \widehat{\boldsymbol{\mu}}_1\ ,\ \ldots,\ \ \widehat{\boldsymbol{\mu}}_{k^*} > \widehat{\boldsymbol{\mu}}_{k^*-1}\ ,\ \ \widehat{\boldsymbol{\mu}}_{k^*} > \widehat{\boldsymbol{\mu}}_{k^*+1}\ ,\ \ldots,\ \ \widehat{\boldsymbol{\mu}}_{k^*} > \widehat{\boldsymbol{\mu}}_K \right\} \tag{3}$$

$$= \text{Prob}\left\{\widehat{\boldsymbol{\mu}}_{k^*} > \mathbf{X}_{max}\right\} \tag{4}$$

where

$$\mathbf{X}_{max} = \max_{k \in [1K]/k^*} \widehat{\boldsymbol{\mu}}_k$$

In equation 3, the probability of correct selection is the probability that $\widehat{\boldsymbol{\mu}}_{k^*}$ is the highest in a set of K random variables $\{\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_K\}$. In equation 4, the bandit problem is transformed and the probability of correct selection is the probability that $\widehat{\boldsymbol{\mu}}_{k^*}$ is the highest in a set of only 2 random variables $\{\widehat{\boldsymbol{\mu}}_{k^*}, \mathbf{X}_{max}\}$. Figure 4 gives the probability density function of $\mathbf{X}_{max}$ in the four cases.

In case 1, the k^*th arm is confronted with $\mathbf{X}_{max}$ which has a mean which is almost equal to zero. In case 1, testing arm k^* reduces the standard deviation of $\widehat{\boldsymbol{\mu}}_{k^*}$ and increases thus the probability that arm k^* will be selected by a greedy action. In case 2, the special arm $\mathbf{X}_{max}$ has a higher mean than the mean of $\widehat{\boldsymbol{\mu}}_{k^*}$ which is one. $\widehat{\boldsymbol{\mu}}_{k^*}$ has more chance to win with a higher standard deviation. Testing arm k^* reduces the standard deviation of $\widehat{\boldsymbol{\mu}}_{k^*}$ and decreases the probability that arm k^* will be selected by a greedy action.

Note the non-regularity of the $\mathbf{X}_{max}$ probability density functions of cases 3 and 4.

We now consider the evolution of μ_g when $K > 2$ under a pure random exploration algorithm which periodically tests the two arms. Six synthetic example cases (see table 3 and fig. 5) are tested with horizon $H = 200$. We observe that, in all six cases, the value of μ_g is always increasing when a random exploration algorithm is applied. The rate of the increase is high at the beginning and lower at the end.

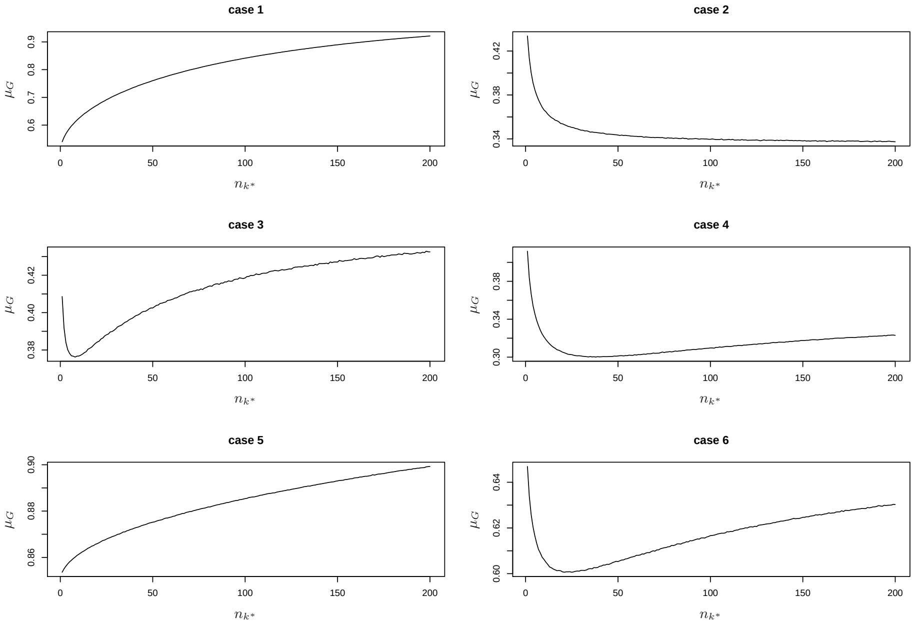

Fig. 3. Evolution of μ_g of the six synthetic problems under an oracle exploitation greedy algorithm. Note that in the cases 1, 2, 3 and 4, the expected gain of a greedy action μ_g equals the probability of correct selection P_{k^*}

As conclusion, applying a greedy policy when the $\frac{\sigma}{\sqrt{n}}$ of the bad arms is high decreases the value of the next μ_g. It is better to make exploration at the beginning to improve μ_g and when the $\frac{\sigma}{\sqrt{n}}$ of the bad arms is small enough the player must move to a more greedy exploitation mode.

4 Conclusion

In this paper, we give an analytical definition of the expected gain of the well-known greedy exploitation action and study its evolution in the context of the bandit problem. In the case where $K = 2$, we give an optimal exploration strategy and we show that testing an arm will always improve the next greedy expected gain μ_g. Results show also that, the μ_g evolution of strategies which always test the same arm are upper-bounded by a smaller value than μ_{k^*}. In the case where $K > 2$, the mode in which a semi-uniform bandit policy is working affects the evolution of μ_g. At the beginning of the problem or when the reward variances of the bandit machines are high, a greedy action can reduce the next μ_g and random exploration is thus a better working mode. When enough knowledge is collected, such that the $\frac{\sigma}{\sqrt{n}}$ values are small, the policy must converge to a more greedy exploitation mode. Future work will focus on the estimation

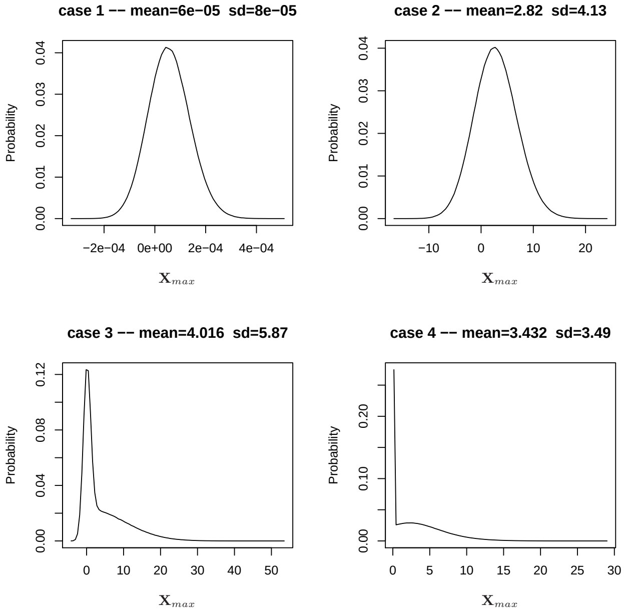

Fig. 4. Monte Carlo determination of $\mathbf{X}_{max}$ for the cases 1, 2, 3 and 4. The mean and the standard deviation are estimated in all the cases.

of μ_g based on the historical samples $\{\mathbf{Z}_k(l)\}$, $k = 1, \ldots, K$ and on the use of these μ_g estimators to propose a new bandit algorithm.

References

1. P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2/3):235–256, 2002.
2. R. Azoulay-Schwartz, S. Kraus, and J. Wilkenfeld. Exploitation vs. exploration: choosing a supplier in an environment of incomplete information. *Decision support systems*, 38(1):1–18, 2004.
3. O. Caelen and G. Bontempi. Improving the exploration strategy in bandit algorithms. In *the Proceedings of Learning and Intelligent OptimizatioN LION II*, 2007.
4. J. Hardwick and Q. Stout. Bandit strategies for ethical sequential allocation. *Computing Science and Statistics*, 23:421–424, 1991.
5. S. Kim and B. Nelson. *Handbooks in Operations Research and Management Science: Simulation*, chapter Selecting the Best System. Elsevier Science, 2006.
6. H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.
7. J. Vermorel and M. Mohri. Multi-armed bandit algorithms and empirical evaluation. In *16th European Conference on Machine Learning (ECML05)*, pages 437–448. ecml, 2005.

Random exploration problem $K > 2$												
	case 1		*case 2*		*case 3*		*case 4*		*case 5*		*case 6*	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
$\mathbf{z}_1$	0	1	0	3	0	3	0	3	0	6	0	6
$\mathbf{z}_2$	0	1	0	1	0	3	0	0.1	0.2	3	0.2	3
$\mathbf{z}_3$	1	0.1	1	0.1	1	0.1	1	3	0.7	0.1	0.7	0.1
$\mathbf{z}_4$									1	3	1	0.1

Table 3. Six synthetic problems to estimate the evolution of μ_g under a random exploration algorithm where $K > 2$. For each case, the first and the second column contain respectively the mean and the standard deviation of the random variables.

8. C. Watkins. *Learning From Delayed Rewards.* PhD thesis, Cambridge University, 1989.

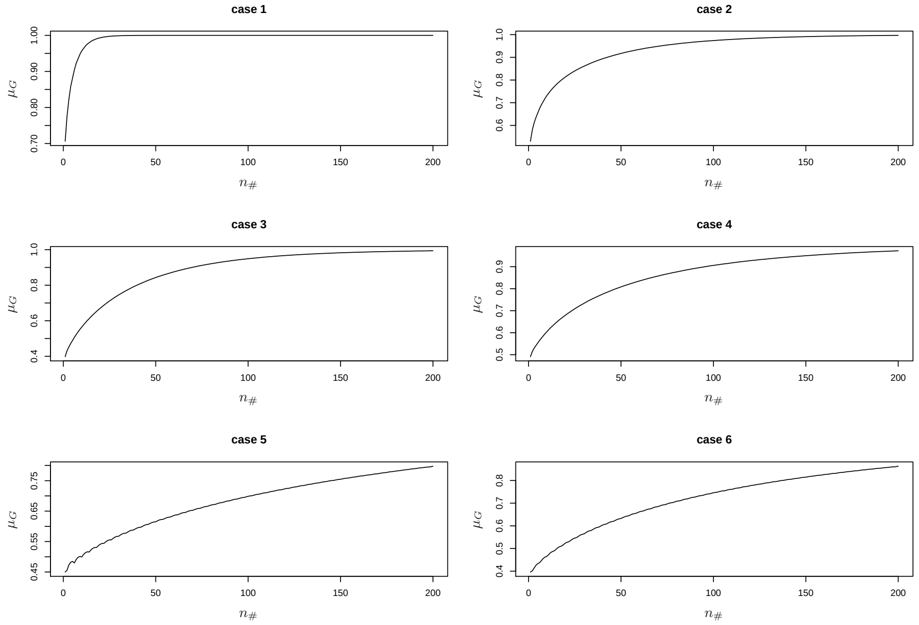

Fig. 5. Evolution of μ_g of the six synthetic problems under a random exploration algorithm.