An approach for handling risk and uncertainty in multiarmed bandit problems

Stefano Perabò*, Fabrice Clerot*

*France Télécom Division Recherche & Développement
2, avenue Pierre Marzin, 22307 Lannion Cedex
stefano.perabo@orange.fr, fabrice.clerot@orange-ftgroup.com

Abstract. An approach is presented to deal with risk in multiarmed bandit problems. Specifically, the well known exploration-exploitation dilemma is solved from the point of view of maximizing an utility function which measures the decision maker's attitude towards risk and uncertain outcomes. A link with the preference theory is thus established. Simulations results are provided for in order to support the main ideas and to compare the approach with existing methods, with emphasis on the short term (small sample size) behavior of the proposed method.

1 Introduction

A "multiarmed bandit problem" can be formulated as follows: given for $t = 1, 2 \dots T$ a sequence of K-dimensional random vectors $\mathrm{r}(t) = [\mathrm{r}_1(t) \dots \mathrm{r}_K(t)]$, called *rewards* and whose probability distribution is not known a priori, the objective is to determine *on line* a sequence of *actions* $\mathrm{a}(t)$ (also called *strategy* or *policy*) where each $\mathrm{a}(t)$ is a discrete random variable defined on the set $\{1, 2, \dots, K\}$, that maximizes the expectation of the *cumulative gain*, $G(T) = \mathbb{E}[\sum_{t=1}^{T} \mathrm{r}_{\mathrm{a}(t)}(t)]$, by observing for each t one (and only one) realization $r_{a(t)}(t)$ [1].
The main difficulty of the problem consists in the fact that the objective function is not known in advance. In fact, if the means $\mu_a(t) = \mathbb{E}[\mathrm{r}_a(t)]$ were available, the best strategy would be obviously to *play* the action $a^*(t) = \arg\max_a \mu_a(t)$. Hence, at each time instant t, the choice of an action is the result of a compromise trying to estimate (*learn*) the objective function (by *exploring* the actions whose mean rewards have not yet been determined with enough confidence) and, at the same time, to maximize it (by *exploiting* those which, based on the preceding observations, are estimated to provide for the best rewards).
This represents a prototype decision problem where the decision maker is faced to the so called *exploration/exploitation dilemma*: while pursuing the second objective (exploitation) by using, unavoidably, a suboptimal strategy, he might incur losses that could be avoided if better estimates of the rewards means were available; on the contrary, while pursuing the first objective (exploration) by using some other suboptimal strategy, he might renounce to play the *supposed*

1. Italic characters like r and a represent realizations of the corresponding random variables which are denoted by using roman characters like r and a.

best action found so far.

This is a problem that, along with many variants, can model basic decision tasks commonly encountered in resources allocation and marketing. A very popular example is borrowed from the internet advertising community: the decision maker must periodically display one ad on a web page (the action) by choosing it from a known set of ads, the objective being the maximization of the number of visitors clicks on the displayed ad (the rewards), which in turn translates into the maximization of the decision maker's revenues. The exploration-exploitation dilemma comes from the fact that users *interests* are not known in advance and thus, in order to judge which ad attracts the largest number of clicks, each ad must be displayed (tested) a certain number of times.
In more realistic scenarios, before taking a decision the decision maker could have access to some kind of side information (often called *context*) such as, in the above example, a profile of the user currently visiting the web page, or a list of keywords extracted from the web page. How this information could be used in order to better choose the ad to display represents an extension of the bandit problem which is usually called a "contextual bandit problem". It can be formalized by introducing an additional sequence of random quantities, the contexts $\mathrm{x}(t)$, and by assuming that a correlation exists between the contexts and the rewards $\mathrm{r}(t)$. The exploration issue consists thus in estimating this correlation in order to better predict which is the best action to perform *conditionally* on a given context.

In this paper an approach is proposed to deal with the exploration-exploitation dilemma in multiarmed bandit problems. Its extension to the more interesting contextual bandit problems is not pursued. However, it should be clear to the readers interested in such an extension that the same line of reasoning could be applied without changes also in the presence of a context, the main difficulties being more of technical rather than conceptual nature (these difficulties arise, in particular, when dealing with huge data sets requiring simultaneously the application of classification techniques).
Moreover, here the focus is on the *classic* framework, in which the rewards $\mathrm{r}(t)$ are modeled as independent and identically distributed (IID) random variables, independent from the decision maker strategy $\mathrm{a}(t)$. Besides these assumptions, only bounded rewards are considered, that is on the case $r_a(t) \in [0, 1]$ without loss of generality (this assumption is not really necessary for the method to work, but it is stated for the sake of comparing it with two other existing ones).

2 Relation to the state of the art

The literature on bandit problems since the seminal work of Robbins (1952) is abundant. The main recent contributions to the solution of multiarmed bandit problem with IID and bounded rewards, notably the strategies UCB1 (Auer et al., 2002) and UCB-V (Audibert et al., 2007), guarantee that the expected cumulative gain $\mathrm{G}(T)$ is lower bounded by a function of the form $G^*(T) - c \log T$, where c is some constant and $G^*(T)$ is the maximum value of the objective function. As shown in Lai and Robbins (1985), the logarithmic term $c \log T$ (called *regret*) cannot be improved further in the following sense: *any* possible strategy suffers a regret which is lower bounded by a function $O(\log T)$. From another point of view, the strategy SUCCESSIVEELIMINATION (Even-Dar et al., 2006) has been proved to find the optimal action

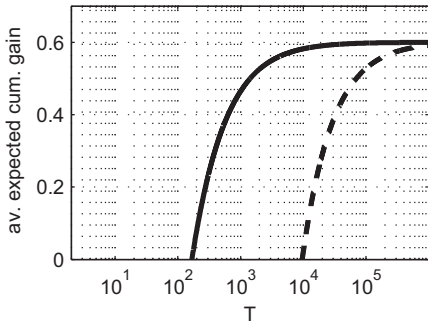

FIG. 1 – *An example of lower bounds on the average expected cumulative gain for the algorithms* UCB1 *(solid line) and* UCB-V *(dashed line). These bounds apply to the case where rewards are distributed as described in section §4 (the second example with* $K = 5$*, the rewards have support* $[0, 1]$ *and* $G^*(T)/T = 0.6$*).*

a^* with an arbitrary prespecified probability in a finite number of steps.
All these strategies have been derived by handling (with considerable mathematical insight) concentration inequalities for sums of random variables. More precisely, such inequalities state the following: if at time t the action a has been played $n_a(t)$ times, then its empirical mean, defined by $\hat{\mu}_a(t) = \sum_{\tau=1}^{t} 1(a(\tau) = a) r_a(\tau)/n_a(t)$, is within a distance ε from the true mean μ_a with a probability greater or equal than a certain threshold value $1-\alpha$. The functional form of ε and α, depends on the kind of inequality used; however it is generally true that if α is kept fixed, then ε decreases for increasing n_a, while if ε is kept fixed, then α increases for increasing n_a. Hence, by playing action a once more one is sure to increase the *confidence* with which the true mean μ_a is known. The *art* of strategy design consists thus in finding the *best* trade off between two competing objectives: sampling in order to maximize the gain and sampling in order to increase the confidence on rewards means.
The regret of the strategies mentioned above has been proved to be very close to the optimal for arbitrary rewards distributions, because they rely on concentration inequalities that are valid for arbitrary probability distributions. This is, however, also the main drawback of these approaches, because the confidence regions derived from such inequalities may be, in some practical case, very conservative, resulting in large constants in front of the $\log T$ factor of the regret term. As a consequence, the lower bounds on the gain become *useful* only in the long term, that is the quantity $G^*(T) - c \log T$ is positive only for a sufficiently large time horizon T. As an example, Figure 1 plots the *average* lower bounds (that is the quantity $(G^*(T) - c \log T)/T$) for the algorithms UCB1 and UCB-V in the particular case when $K = 5$ and the rewards are distributed as described in section §4, where the results of some numerical simulations will be presented. Clearly, for $T < 200$ nothing can be said.

The main criticism that can be addressed to the current approaches to solve bandit problems is to concentrate all the effort in upper bounding the regret of an algorithm as tightly as possible, which in fact leaves no room to account for user's preferences and constraints. Consider, in

particular, what happens during the first period following the start of a bandit algorithm, where exploration is the predominant activity. The question becomes when and how to switch from exploration to exploitation given the information provided by only a finite number of samples. The answer clearly depends on many factors, the most influential, maybe, being: the available time horizon (that is how many more samples can be drawn) and the decision maker's attitude towards uncertainty and risky choices.
The objective of this paper is thus to propose a strategy design procedure that allows such preferences and constraints to have an impact on algorithm behaviour and to be taken into account somehow automatically. In particular the exploration-exploitation dilemma will be reformulated as the problem of maximizing an *utility function* which *quantifies* the decision maker's preferences over a set of appropriate confidence intervals. More precisely, each alternative confidence interval is associated to an estimation of the future gain that could be obtained by playing a certain strategy. The role of the utility function is thus to *measure* the attitude towards risk and uncertainty, attitude that drives the choice of one confidence interval. A specific utility function which is designed to express aversion to risky strategies is provided for. The related algorithm will be called LCB1. The results of some numerical simulations are presented in order to compare the short term (small time horizon T) performance of this new algorithm with that of the algorithms UCB1 and UCB-V cited above.

3 Description of the approach

Pretend for a while that the rewards probability distributions are known. It is possible to view the problem of maximizing the gain $G(T)$ as a *finite horizon planning* problem in a degenerate markov decision process (MDP), that is the problem of determining the optimal strategy (often called *policy*) in a MDP consisting of only one deterministic state [2]. It is known (Puterman, 1987) that the optimal strategy can be found by a backward induction algorithm. If it is applied here, under the independence assumptions stated in the introduction, it simply amounts to find for each time instant t the functions $Q^*(a,t)$ (called *optimal action-value functions* in the MDP jargon) that solve

$$Q^*(a,t) = \mathbb{E}[\mathrm{r}_{\mathrm{a}(t)}(t) + \max_b Q^*(b,t+1) \parallel \mathrm{a}(t) = a] \qquad Q^*(a,T+1) = 0 \tag{1}$$

The optimal action-value function $Q^*(a,t)$ equals the maximum expected gain that can be obtained in the interval $[t,T]$, conditioned on the fact that at time t action a is played. The above formula can thus be read as follows: in order to maximize the *future* gain on the interval $[t,T]$, it is sufficient to play at time t the action maximizing the expected reward $\mathbb{E}[\mathrm{r}_a(t)]$ and then to follow the best strategy in the interval $[t+1,T]$. In other words, the best strategy is to play *with probability one* the action $a^*(t) = \arg\max_a Q^*(a,t)$, for all $t \in [1,T]$. In this case, the solution of the backward induction gives obviously $Q^*(a,t) = \mu_a + (T-t)\mu_*$, where $\mu_* = \max_a \mu_a$, and confirms the above interpretation.

2. In a MDP there is a stochastic process $\mathrm{x}(t)$ defined by the transition probabilities $\mathbb{P}[\mathrm{x}(t+1) = w \| \mathrm{x}(t) = z, \mathrm{a}(t) = a]$ (giving the probability of reaching one possible next state given the current state and action) and the vector of rewards depend on the current state through the conditional probability $\mathbb{P}[\mathrm{r}(t) = r \| \mathrm{x}(t) = z]$. The IID framework considered here is thus obtained when the only allowed transition is form one deterministic state $\mathrm{x}(t) \equiv z$ to itself, irrespective of which action is taken.

Now, assume that the decision maker is allowed to play a *randomized strategy*, that is the action $a(t)$ is drawn from a random variable taking values in the set $\{1, 2, \dots, K\}$. Define $p_a(t) = \mathbb{P}[\mathrm{a}(t) = a]$ to be the probability of taking action a at time t, and the quantity

$$V^*(t, p(t)) = \sum_a Q^*(a, t)\mathbb{P}[\mathrm{a}(t) = a] = \sum_a \mu_a p_a(t) + (T - t)\mu_* \tag{2}$$

where $p(t) = [p_1(t) \dots p_K(t)]$. Clearly $V^*(t, p(t))$ represents the maximum expected gain that can be obtained in the interval $[t, T]$ whenever a randomized strategy is played at time t. Moreover, $V^*(t, p(t)) \leq \max_a Q^*(a, t)$ for any choice of $p(t)$, the equality holding when $p_{a^*(t)}(t) = 1$ (case in which $V^*(t, p(t)) = Q^*(a^*(t), t)$, quantity that it is often called *optimal value function*). Hence, in a planning problem (that is when rewards distribution are known) there is no advantage in adopting a randomized strategy with respect to the optimal deterministic one obtained by the backward induction outlined above.

Consider then the case of no prior knowledge about the rewards distribution. Neither the functions $Q^*(a, t)$ nor $V^*(t, p(t))$ can be evaluated because the means μ_a are not known. However, suppose that for a given $\alpha > 0$, and for each t, a $100(1 - \alpha)\%$ *confidence interval* for $V^*(t, p(t))$, say

$$C(t, p(t)) = \left[V^*_{lo}(t, p(t)), V^*_{up}(t, p(t))\right] \tag{3}$$

can be computed based on the set of observation $\{r_{a(s)}(s)\}$ on the interval $[1, t - 1]$. Recall the meaning of a confidence interval: it is an interval whose extremes are functions of the empirical data and that contains with probability $(1 - \alpha)$ the true value. By varying $p(t)$, different confidence intervals are obtained. Hence it is reasonable to define the optimal strategy $p^*(t)$ as the strategy such that the confidence interval $C(t, p^*(t))$ is *preferred* over all the intervals $C(t, p(t))$ obtained for $p(t) \neq p^*(t)$. For example, an interval $[a + \Delta a, b + \Delta b]$ should be preferred over $[a, b]$ whenever $\Delta a, \Delta b > 0$ because the maximum expected gain of the corresponding strategy is likely to be larger. It is thus necessary to show how these intervals can be constructed in practice and to define a *preference relation* over confidence intervals.

Concerning the first issue, here an heuristic approach is adopted. It is known that, as $n_a(t) \to \infty$, the empirical means $\hat{\mu}_a(t) = \sum_{\tau=1}^{t} 1(a(\tau) = a) r_a(\tau)/n_a(t)$ tend to be distributed as a $\mathcal{N}(\mu_a, \sigma_a^2/n_a(t))$ random variable, that is a gaussian whose mean and variance are respectively the true mean μ_a and the true variance σ_a^2 divided by the number of samples. The true variance can also be replaced by the unbiased empirical variance $\hat{\sigma}_a^2(t) = \sum_{\tau=1}^{t} 1(a(\tau) = a)(r_a(\tau) - \hat{\mu}_a(t))^2/(n_a(t) - 1)$. Even though these are asymptotic results, pretend that they represent acceptable approximations also for a finite number of samples $n_a(t)$. Hence, $V^*(t, p(t))$ in (2) is approximately distributed as $\mathcal{N}(\hat{\mu}(t), \hat{\sigma}^2(t))$ with

$$\hat{\mu}(t) = \sum_a \hat{\mu}_a(t) p_a(t) + (T - t)\hat{\mu}_*(t) \qquad \hat{\sigma}^2(t) = \sum_a \frac{\hat{\sigma}_a^2(t)}{n_a(t)} p_a(t)^2 + (T - t)\frac{\hat{\sigma}_*^2(t)}{n_*(t)} \tag{4}$$

and where $\hat{\mu}_*(t)$ is the largest empirical mean, and $\hat{\sigma}_*^2(t)$ and $n_*(t)$ are the empirical variance and number of samples of the corresponding action. For a numerical example see Figure 2. The extremes of a $100(1 - \alpha)\%$ confidence interval are thus

$$V^*_{lo}(t, p(t)) = \hat{\mu}(t) - f\hat{\sigma}(t) \qquad V^*_{up}(t, p(t)) = \hat{\mu}(t) + f\hat{\sigma}(t) \tag{5}$$

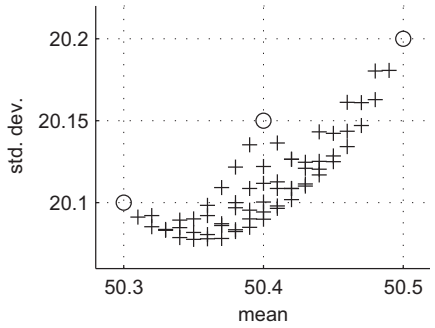

FIG. 2 – *In the plot, the crosses indicate the couples* $(\hat{\mu}(t)), \hat{\sigma}(t))$ *obtained by varying each action selection probability* $p_a(t)$ *in the range* $[0, 1]$ *in steps of* 0.1 *under the constraint* $\sum_a p_a(t) = 1$ *and by using the following data:* $K = 3$*,* $[\hat{\mu}_a(t)] = [0.5, 0.4, 0.3]$*,* $[\hat{\sigma}_a(t)/\sqrt{n_a(t)}] = [0.2, 0.15, 0.1]$ *and* $(T - t) = 100$*. The circles represent the values obtained when* $p_a(t) = 1$ *for some* $a \in \{1, 2, 3\}$*.*

where $\hat{\sigma}(t) = \sqrt{\hat{\sigma}^2(t)}$ and f is the $100(1 - \alpha)\%$ quantile of a standard gaussian.

Next, a preference relation is needed that is able to represent the decision maker's attitude towards the uncertainty represented by the fact that, for a given strategy, the corresponding gain can not be evaluated exactly. A preference relation is a binary relation that will be denoted by $\succ$, so that if interval D is preferred over interval C, one writes $D \succ C$. The classical result is that, under appropriate assumptions, there exists a quantitative representation of a preference relation in terms of a real *utility function* u defined on the set of intervals, say $\mathcal{I}$, such that

$$D \succ C \iff u(D) > u(C) \tag{6}$$

Besides the technical assumptions which are necessary to account for the fact that $\mathcal{I}$ is uncountable (see for example Fishburn (1999) for details), it can be proved that the above representation holds if and only if the relation $\succ$ on $\mathcal{I}$ is a *weak order*. This means that $\succ$ is transitive ($E \succ D$ and $D \succ C \;\Rightarrow\; E \succ C$), irreflexive ($C \succ C$ never holds) and that the relation $\sim$ (defined by $C \sim D$ if neither $C \succ D$ nor $D \succ C$) is also transitive.
In the so called *normative* approach to preference modelling (Tsoukiàs, 2008), that will be followed here, the existence of a preference relation with the above properties is postulated (a decision maker adhering to such a relation is termed *rational*). This amounts, in practice, to set the utility function to an appropriate form, depending on the application at hand, by choosing perhaps in a set of many possible good choices. In other words, the existence of an *optimal* utility function is not stated here: different forms might be reasonable and behave equally well in practice. In this paper, the following very simple utility function is proposed and tested by numerical simulation:

$$u(t, p(t)) = \exp\left(\frac{\hat{\mu}(t) - \hat{\sigma}(t)}{T - t + 1}\right) \tag{7}$$

a	α	β	μ_a	σ_a
1	0.333	0.222	0.600	0.393
2	1.200	0.800	0.600	0.283
3	17.000	12.000	0.586	0.090
4	0.222	0.333	0.400	0.393
5	2.000	18.000	0.100	0.066

TAB. 1 – *The parameters of the beta distributions that have been used to model the rewards outcomes in the numerical simulations, along with the corresponding expectations and standard deviations*

Note that it is a parametric function of the confidence intervals for $V^*(t, p(t))$ through $\hat{\mu}(t)$ and $\hat{\sigma}(t)$ which are the quantities used to define the interval itself.
The rationale behind this choice is the following. The tangent at a point on a level curve of u is equal to 1 everywhere in the plane $(\hat{\mu}, \hat{\sigma})$. Hence, the decision maker is indifferent between a confidence interval specified by the couple $(\hat{\mu}, \hat{\sigma})$ and another specified by $(\hat{\mu} + \Delta\hat{\mu}, \hat{\sigma} + \Delta\hat{\sigma})$, where $\Delta\hat{\mu}$ and $\Delta\hat{\sigma}$ are two increments related by $\Delta\hat{\sigma} = \Delta\hat{\mu}$. Since the lower bounds of the corresponding confidence intervals are equal, $\hat{\mu} - \hat{\sigma} = (\hat{\mu} + \Delta\hat{\mu}) - (\hat{\sigma} + \Delta\hat{\sigma})$, this means that the decision maker accepts to trade one strategy with another (in other words, they have the same utility) only if the confidence intervals share the same lower bound. Higher lower bounds are preferred while upper bounds are ignored, thus indicating aversion to risky strategies.

In conclusion, the proposed strategy, that will be called LCB1 (for "Lower Confidence Bound"), is the following:

1. *for* $1 \leq t \leq 2K$ *sample twice each action, so that the unbiased empirical variances can be set to their initial value;*
2. *for each time* $t > 2K$ *choose the action* a *with probability* $p_a^*(t)$*, where* $p^*(t) = \arg\max_p u(t, p)$*, and* $u(t, p)$ *is the utility function in* (7)*.*

4 Some numerical simulations

In this section the results of two numerical simulations are provided for. Five rewards distributions are considered, that belong to the family of beta distributions $\mathcal{B}e(\alpha, \beta)$. The parameters and plots of the corresponding probability densities are shown in Table 1 and Figure 3 respectively. The time horizon is $T = 200$ in both tests and $f = 1.96$ (which gives a 95% confidence interval for a standard gaussian random variable).
In the first test $K = 3$, corresponding to the rewards distributions 1 to 3. Figure 4 show plots of the probability distribution $F_{G(t)}(g) = \mathbb{P}[G(t) \leq g]$ of the gain obtained after $t = 50$, 100, 150 and 200 steps by the algorithms UCB1 (blue line), UCB-V (black line) and the proposed algorithm LCB1 (red line). These distributions have been obtained by running these algorithms over 1000 different realizations of the rewards sequences. There are no remarkable differences between the three algorithms. In fact reward means are comparable and the risk reduction that could be obtained by sampling action 3 more often is counterbalanced by a reduction of the

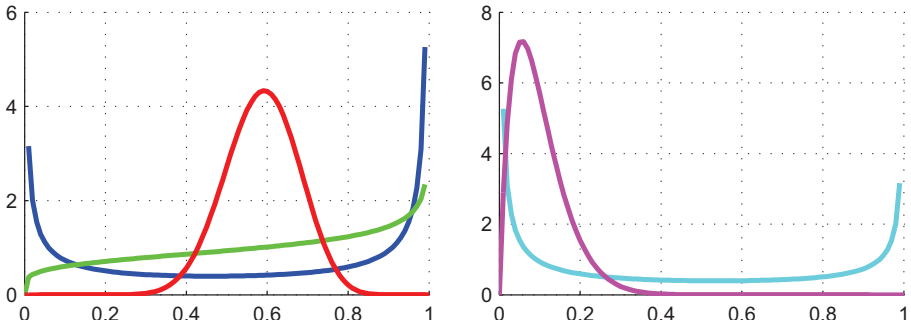

FIG. 3 – *Plots of the rewards probability densities, on the left for* $a = 1, 2, 3$ *(blue, green and red lines respectively) and on the right for* $a = 4, 5$ *(cyan and magenta lines respectively).*

expected gain.
In the second test $K = 5$ and actions 4 and 5 have been added to actions 1, 2 and 3 of the first test. The test has been conducted in the same way as described above. Plots of the probability distribution of the gain are shown in Figure 5. Contrary to the case considered in the first test, here the performance of LCB1 is superior. In fact, as it can be checked in Figure 6, LCB1 *gets rid* of the *inferior* actions 4 and 5 faster than the other two algorithms.

5 Discussion

An approach for the solution of the multiarmed bandit problem in the IID framework has been proposed. Contrary to well known existing methods, the approach is somewhat heuristic in that it is not based on rigorous *convergence proofs*. However, the numerical simulations show a very good performance on the short term, at least in the cases considered, thus motivating further analyses. In practical applications, the short term behavior is particular important because the assumption of identical distribution in time might be violated in the long term.
Two margins of improvements is worth indicating. On one hand, the way confidence intervals are computed: better established finite sample statistical techniques such as the bootstrap, for example, might provide for more correct and even asymmetric intervals. On the other, utility functions might exist that would represent different balances between risk avoiding and risk seeking behaviors, that is different ways of managing the exploration-exploitation dilemma.

The main contribution of this paper, however, is the link established with the theory of preferences which has been and currently is a large research subfield in economics. In fact, the approach presented here shares many similarities with very basic portfolio selection methods (Markowitz, 1952) and the possibility of introducing more advanced techniques (the so called *risk-sensitive* or *risk-averse*) in the context of bandit problems should be considered for future investigations.
It is also worth mentioning the possibility of extending the IID framework to allow *discounted* rewards, that is to solve an exploration-exploitation dilemma where, at each time t, a time de-

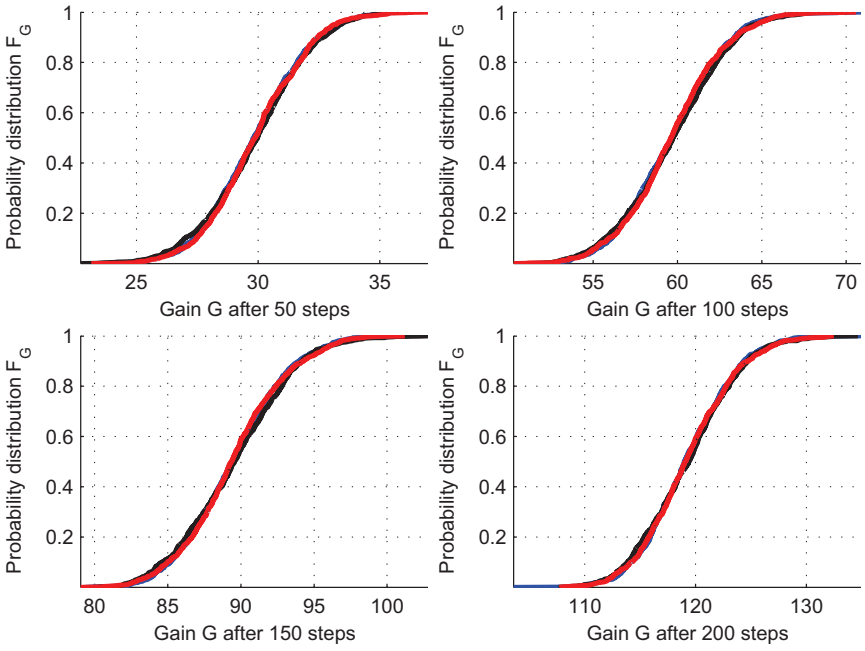

FIG. 4 – *The results of the first test ($K = 3$, rewards distributions 1 to 3): distribution of the gain obtained by the algorithm* UCB1 *(blue line),* UCB-V *(black line) and* LCB1 *(red line), after* 50, 100, 150 *and* 200 *time steps.*

pendent objective function of the form $G(t) = \mathbb{E}[\sum_{s=t}^{T(t)} w(t,s) \cdot \mathrm{r}_{\mathrm{a}(s)}(s)]$ has to be *maximized*, where $T(t) \geq t$ and the real numbers $w(t,s)$ are given weights (only the case $T(t) \equiv T$ and $w(t,s) \equiv 1$ has been discussed in this paper). In fact, decisions involve very often tradeoffs among uncertain costs and benefits occurring at different points in time under the constraint of different time horizons. In such cases, it is known (Frederick et al., 2002) that the choice of a weighting sequence is related to the decision maker's *time preferences*: different preferences demand different forms of weighting. It should be clear that, whenever these preferences are known (along with the utility function over confidence intervals), the presented approach can be applied quite straightforwardly.

References

Audibert, J.-Y., R. Munos, and C. Szepesvári (2007). Variance estimates and exploration function in multi-armed bandit. Technical Report 07-31, CERTIS, France.

Auer, P., N. Cesa-Bianchi, and P. Fischer (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning 47*, 235–256.

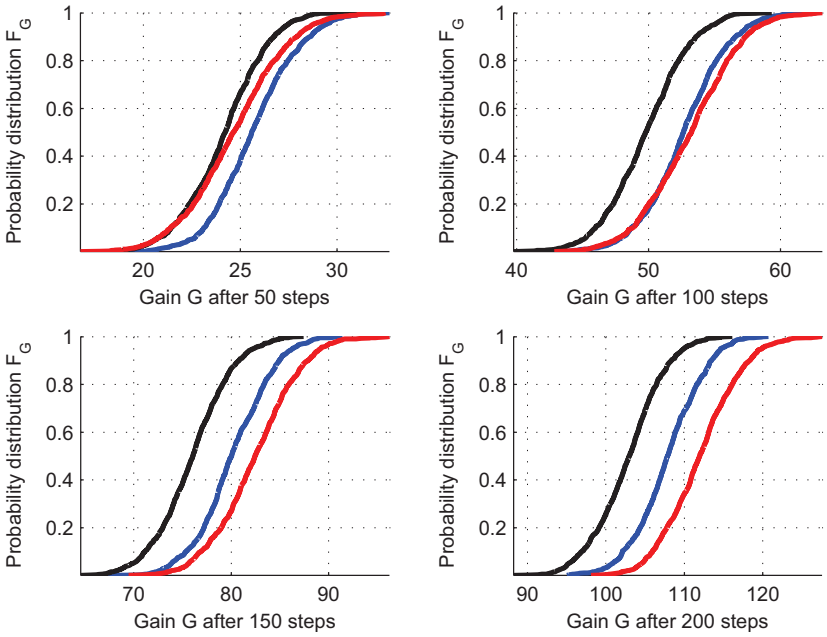

FIG. 5 – *The results of the second test ($K = 5$, rewards distributions 1 to 5): distribution of the gain obtained by the algorithm* UCB1 *(blue line),* UCB-V *(black line) and* LCB1 *(red line), after* 50, 100, 150 *and* 200 *time steps.*

Even-Dar, E., S. Mannor, and Y. Mansour (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research 7*, 1079–1105.

Fishburn, P. (1999). Preference structures and their numerical representation. *Theoretical Computer Science* (217), 359–383.

Frederick, S., G. Loewenstein, and T. O'donoghue (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature 40*(2), 351–401.

Lai, T. and H. Robbins (1985). Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math. 6*, 4–22.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance 7*(1), 77–91.

Puterman, M. (1987). Dynamic programming. In R. Meyers (Ed.), *Encyclopedia of Physical Science and Technology*, Volume 4, pp. 438–463. Academic Press.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc. 58*(5), 527–535.

Tsoukiàs, A. (2008). From decision theory to decision aiding methodology. *European Journal of Operational Research 187*, 138–161.

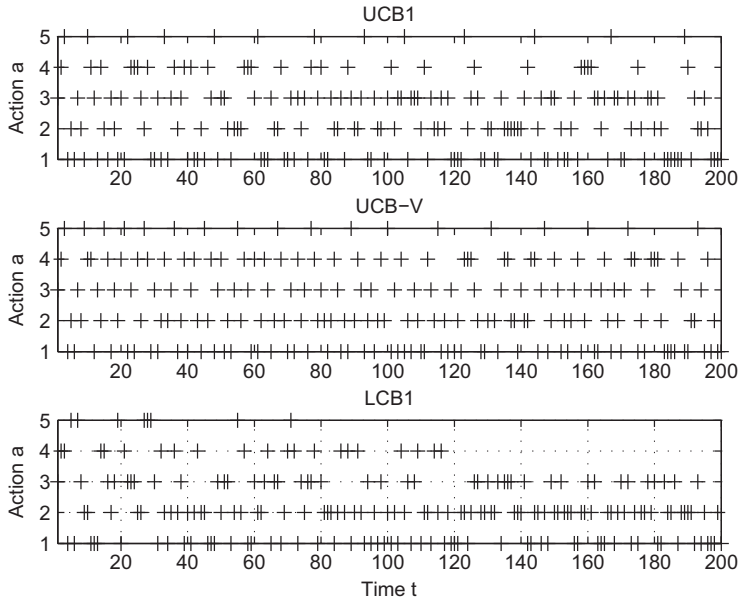

FIG. 6 – *Typical behaviours of algorithms* UCB1*,* UCB-V *and* LCB1 *(from top to bottom) obtained in the second test (*$K = 5$*, rewards distributions* 1 *to* 5*). A "+" symbol indicates that the corresponding action has been selected at time* t*.*

Résumé

Une approche est présentée pour la prise en considération du risque dans le problème du "bandit manchot". Plus précisément, le dilemme dit d'exploration-exploitation est reformulée comme un problème de maximisation d'une fonction d'utilité qui mesure l'attitude du décideur envers le risque et l'incertitude. Un lien avec la théorie de la préférence est donc établi. La méthode proposée est testée et comparée par simulation numérique avec d'autres bien connues dans la littérature, avec un intérêt particulier pour le comportement à court terme (petit nombre d'échantillons).