# 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.