From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning

Rémi Munos
INRIA Lille – Nord Europe
remi.munos@inria.fr

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

R. Munos. *From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning*. Foundations and Trends® in Machine Learning, vol. 7, no. 1, pp. 1–129, 2014.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-60198-767-9

Foundations and Trends® in Machine Learning
Volume 7, Issue 1, 2014

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles on the theory, algorithms and applications of machine learning, including the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive, and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2014, Volume 7, 4 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Foundations and Trends® in Machine Learning
Vol. 7, No. 1 (2014) 1–129

DOI: 10.1561/2200000038

now
the essence of knowledge

From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning

Rémi Munos
INRIA Lille – Nord Europe
remi.munos@inria.fr

Contents

Abstract

This work covers several aspects of the *optimism in the face of uncertainty* principle applied to large scale optimization problems under finite numerical budget. The initial motivation for the research reported here originated from the empirical success of the so-called *Monte-Carlo Tree Search* method popularized in Computer Go and further extended to many other games as well as optimization and planning problems. Our objective is to contribute to the development of theoretical foundations of the field by characterizing the complexity of the underlying optimization problems and designing efficient algorithms with performance guarantees.

The main idea presented here is that it is possible to decompose a complex decision making problem (such as an optimization problem in a large search space) into a sequence of elementary decisions, where each decision of the sequence is solved using a *(stochastic) multi-armed bandit* (simple mathematical model for decision making in stochastic environments). This so-called *hierarchical bandit* approach (where the reward observed by a bandit in the hierarchy is itself the return of another bandit at a deeper level) possesses the nice feature of starting the exploration by a quasi-uniform sampling of the space and then focusing progressively on the most promising area, at different scales, according to the evaluations observed so far, until eventually performing a local search around the global optima of the function. The performance of the method is assessed in terms of the optimality of the returned solution as a function of the number of function evaluations.

Our main contribution to the field of function optimization is a class of hierarchical optimistic algorithms designed for general search spaces (such as metric spaces, trees, graphs, Euclidean spaces) with different algorithmic instantiations depending on whether the evaluations are noisy or noiseless and whether some measure of the "smoothness" of the function is known or unknown. The performance of the algorithms depends on the "local" behavior of the function around its global optima expressed in terms of the quantity of near-optimal states measured with some metric. If this local smoothness of the function is known then one can design very efficient optimization algorithms (with

convergence rate independent of the space dimension). When this information is unknown, one can build adaptive techniques which, in some cases, perform almost as well as when it is known.

In order to be self-contained, we start with a brief introduction to the stochastic multi-armed bandit problem in Chapter 1 and describe the UCB (Upper Confidence Bound) strategy and several extensions. In Chapter 2 we present the Monte-Carlo Tree Search method applied to Computer Go and show the limitations of previous algorithms such as UCT (UCB applied to Trees). This provides motivation for designing theoretically well-founded optimistic optimization algorithms. The main contributions on hierarchical optimistic optimization are described in Chapters 3 and 4 where the general setting of a semi-metric space is introduced and algorithms designed for optimizing a function assumed to be locally smooth (around its maxima) with respect to a semi-metric are presented and analyzed. Chapter 3 considers the case when the semi-metric is known and can be used by the algorithm, whereas Chapter 4 considers the case when it is not known and describes an adaptive technique that does almost as well as when it is known. Finally in Chapter 5 we describe optimistic strategies for a specific structured problem, namely the planning problem in Markov decision processes with infinite horizon discounted rewards.

R. Munos. *From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning*. Foundations and Trends® in Machine Learning, vol. 7, no. 1, pp. 1–129, 2014.

DOI: 10.1561/2200000038.

About optimism...

Optimists and pessimists inhabit different worlds, reacting to the same circumstances in completely different ways.

Learning to Hope, Daisaku Ikeda.

Habits of thinking need not be forever. One of the most significant findings in psychology in the last twenty years is that individuals can choose the way they think.

Learned Optimism, Martin Seligman.

Humans do not hold a positivity bias on account of having read too many self-help books. Rather, optimism may be so essential to our survival that it is hardwired into our most complex organ, the brain.

The Optimism Bias: A Tour of the Irrationally Positive Brain, Tali Sharot.

1

The stochastic multi-armed bandit problem

We start with a brief introduction to the stochastic multi-armed bandit setting. This is a simple mathematical model for sequential decision making in unknown random environments that illustrates the so-called *exploration-exploitation trade-off*. Initial motivation in the context of clinical trials dates back to the works of Thompson [1933, 1935] and Robbins [1952]. In this chapter we consider the *optimism in the face of uncertainty* principle, which recommends following the optimal policy in the most favorable environment among all possible environments that are reasonably compatible with the observations. In a multi-armed bandit the set of "compatible environments" is the set of possible distributions of the arms that are likely to have generated the observed rewards. More precisely we investigate a specific strategy, called UCB (where UCB stands for upper confidence bound) introduced by Auer, Cesa-Bianchi, and Fischer in [Auer et al., 2002], that uses simple high-probability confidence intervals (one for each arm) for the set of possible "compatible environments". The strategy consists of selecting the arm with highest upper-confidence-bound (the optimal strategy for the most favorable environment).

We introduce the setting of the multi-armed bandit problem in Sec-

tion 1.1.1, then present the UCB algorithm in Section 1.1.2 and existing lower bounds in Section 1.1.3. In Section 1.2 we describe extensions of the optimistic approach to the case of an infinite set of arms, either when the set is denumerable (in which case a stochastic assumption is made) or where it is continuous but the reward function has a known structure (e.g. linear, Lipschitz).

1.1 The K-armed bandit

1.1.1 Setting

Consider K arms (actions, choices) defined by distributions $(\nu_k)_{1\leq k\leq K}$ with bounded support (here we will assume that the support lies in $[0,1]$) that are initially unknown to the player. At each round $t = 1,\dots,n$, the player selects an arm $I_t \in \{1,\dots,K\}$ and obtains a reward $X_t \sim \nu_{I_t}$, which is a random sample drawn from the distribution ν_{I_t} corresponding to the selected arm I_t, and is assumed to be independent of previous rewards. The goal of the player is to maximize the sum of obtained rewards in expectation.

Define $\mu_k = \mathbb{E}_{X\sim\nu_k}[X]$ as the mean values of each arm, and $\mu^* = \max_k \mu_k = \mu_{k^*}$ as the mean value of one best arm k^* (there may exist several).

If the arm distributions were known, the agent would select the arm with the highest mean at each round and obtain an expected cumulative reward of $n\mu^*$. However, since the distributions of the arms are initially unknown, he needs to pull each arm several times in order to acquire information about the arms (this is called *exploration*) and while his knowledge about the arms improves, he should pull increasingly often the apparently best ones (this is called *exploitation*). This illustrates the so-called *exploration-exploitation trade-off*.

In order to assess the performance of any strategy, we compare its performance to an oracle strategy that would know the distributions in advance (and would thus play the optimal arm). For that purpose we define the notion of *cumulative regret*: at round n,

$$R_n \stackrel{\text{def}}{=} n\mu^* - \sum_{t=1}^{n} X_t. \tag{1.1}$$

This defines the loss, in terms of cumulative rewards, resulting from not knowing from the beginning the reward distributions. We are thus interested in designing strategies that have a low cumulative regret. Notice that using the tower rule, the expected regret can be written:

$$\mathbb{E}R_n = n\mu^* - \mathbb{E}\Big[\sum_{t=1}^{n} \mu_{I_t}\Big] = \mathbb{E}\Big[\sum_{k=1}^{K} T_k(n)(\mu^* - \mu_k)\Big] = \sum_{k=1}^{K} \mathbb{E}[T_k(n)]\Delta_k, \tag{1.2}$$

where $\Delta_k \stackrel{\text{def}}{=} \mu^* - \mu_k$ is the *gap* in terms of expected rewards, between the optimal arm and arm k, and $T_k(n) \stackrel{\text{def}}{=} \sum_{t=1}^{n} \mathbf{1}\{I_t = k\}$ is the number of pulls of arm k up to time n.

Thus a good algorithm should not pull sub-optimal arms too often. Of course, in order to acquire information about the arms, one needs to explore all the arms and thus pull sub-optimal arms. The regret measures how fast one can *learn* relevant quantities about one's unknown environment while simultaneously *optimizing* some criterion. This combined learning-optimizing objective is central to the exploration-exploitation trade-off.

Proposed solutions: Since initially formulated by Robbins [1952], several approaches have addressed this exploration-exploitation problem, including:

- *Bayesian exploration*: A prior is assigned to the arm distributions and an arm is selected as a function of the posterior (such as Thompson sampling [Thompson, 1933, 1935] which has been analyzed recently in [Agrawal and Goyal, 2012, Kauffmann et al., 2012, Agrawal and Goyal, 2013, Kaufmann et al., 2013], the Gittins indexes, see [Gittins., 1979, Gittins et al., 1989], and optimistic Bayesian algorithms such as in [Srinivas et al., 2010, Kauffman et al., 2012]).

- ϵ-*greedy exploration*: The empirical best arm is played with probability $1 - \epsilon$ and a random arm is chosen with probability ϵ (see e.g. Auer et al. [2002] for an analysis),

- *Soft-max exploration*: An arm is selected with a probability that depends on the (estimated) performance of this arm given previous reward samples (such as the EXP3 algorithm introduced in Auer et al. [2003], see also the *learning-from-expert* setting of Cesa-Bianchi and Lugosi [2006]).

- *Follow the perturbed leader*: The empirical mean reward of each arm is perturbed by a random quantity and the best perturbed arm is selected (see e.g. Kalai and Vempala [2005], Kujala and Elomaa [2007]).

- *Optimistic exploration*: Select the arm with the largest high-probability upper-confidence-bound (initiated by Lai and Robbins [1985], Agrawal [1995b], Burnetas and Katehakis [1996a]), an example of which is the UCB algorithm [Auer et al., 2002] described in the next section.

1.1.2 The Upper Confidence Bounds (UCB) algorithm

The Upper Confidence Bounds (UCB) strategy by Auer et al. [2002] consists of selecting at each time step t an arm with largest B-values:

$$I_t \in \arg\max_{k\in\{1,\dots,K\}} B_{t,T_k(t-1)}(k),$$

where the B-value of an arm k is defined as:

$$B_{t,s}(k) \stackrel{\text{def}}{=} \hat{\mu}_{k,s} + \sqrt{\frac{3\log t}{2s}}, \tag{1.3}$$

where $\hat{\mu}_{k,s} \stackrel{\text{def}}{=} \frac{1}{s}\sum_{i=1}^{s} X_{k,i}$ is the empirical mean of the s first rewards received from arm k, and $X_{k,i}$ denotes the reward received when pulling arms k for the i-th time (i.e., by defining the random time $\tau_{k,i}$ to be the instant when we pull arm k for the i-th time, we have $X_{k,i} = X_{\tau_{k,i}}$). We described here a slightly modified version where the constant defining the confidence interval is $3/2$ instead of 2 for the original version UCB1 described in [Auer et al., 2002].

This strategy follows the so-called *optimism in the face of uncertainty* principle since it selects the optimal arm in the most favorable environments that are (in high probability) compatible with the observations. Indeed the B-values $B_{t,s}(k)$ are high-probability upper-confidence-bounds on the mean-value of the arms μ_k. More precisely for any $1 \leq s \leq t$, we have $\mathbb{P}(B_{t,s}(k) \geq \mu_k) \leq 1 - t^{-3}$. This bound comes from the Chernoff-Hoeffding inequality which is described below. Let $Y_i \in [0,1]$ be independent copies of a random variable of mean μ. Then

$$\mathbb{P}\Big(\frac{1}{s}\sum_{i=1}^{s} Y_i - \mu \geq \epsilon\Big) \leq e^{-2s\epsilon^2} \quad \text{and} \quad \mathbb{P}\Big(\frac{1}{s}\sum_{i=1}^{s} Y_i - \mu \leq -\epsilon\Big) \leq e^{-2s\epsilon^2}. \tag{1.4}$$

Thus for any fixed $1 \leq s \leq t$,

$$\mathbb{P}\Big(\hat{\mu}_{k,s} + \sqrt{\frac{3\log t}{2s}} \leq \mu_k\Big) \leq e^{-3\log(t)} = t^{-3}, \tag{1.5}$$

and

$$\mathbb{P}\Big(\hat{\mu}_{k,s} - \sqrt{\frac{3\log t}{2s}} \geq \mu_k\Big) \leq e^{-3\log(t)} = t^{-3}. \tag{1.6}$$

We now deduce a bound on the expected number of plays of sub-optimal arms by noticing that with high probability, the sub-optimal arms are not played whenever their UCB is below μ^*.

Proposition 1.1. Each sub-optimal arm k is played in expectation at most

$$\mathbb{E}T_k(n) \leq 6\frac{\log n}{\Delta_k^2} + \frac{\pi^2}{3} + 1$$

time. Thus the cumulative regret of UCB is bounded as

$$\mathbb{E}R_n = \sum_k \Delta_k \mathbb{E}T_k(n) \leq 6 \sum_{k:\Delta_k>0} \frac{\log n}{\Delta_k} + K\Big(\frac{\pi^2}{3} + 1\Big).$$

First notice that the dependence in n is logarithmic. This says that out of n pulls, the sub-optimal arms are played only $O(\log n)$ times, and thus the optimal arm (assuming there is only one) is played $n - O(\log n)$ times. Now, the constant factor in the logarithmic term is $6\sum_{k:\Delta_k>0}\frac{1}{\Delta_k}$ which deteriorates when some sub-optimal arms are very close to the

optimal one (i.e., when Δ_k is small). This may seem counter-intuitive, in the sense that for any fixed value of n, if all the arms have a very small Δ_k, then the regret should be small as well (and this is indeed true since the regret is trivially bounded by $n \max_k \Delta_k$ whatever the algorithm). So this result should be understood (and is meaningful) for a fixed problem (i.e., fixed Δ_k) and for n sufficiently large (i.e., $n > \min_k 1/\Delta_k^2$).

Proof. Assume that a sub-optimal arm k is pulled at time t. This means that its B-value is larger than the B-values of the other arms, in particular that of the optimal arm k^*:

$$\hat{\mu}_{k,T_k(t-1)} + \sqrt{\frac{3\log t}{2T_k(t-1)}} \geq \hat{\mu}_{k^*,T_{k^*}(t-1)} + \sqrt{\frac{3\log t}{2T_{k^*}(t-1)}}. \tag{1.7}$$

Now, either one of the two following inequalities hold:

- The empirical mean of the optimal arm is not within its confidence interval:

$$\hat{\mu}_{k^*,T_{k^*}(t-1)} + \sqrt{\frac{3\log t}{2T_{k^*}(t-1)}} < \mu^*, \tag{1.8}$$

- The empirical mean of the arm k is not within its confidence interval:

$$\mu_{k,T_k(t-1)} > \mu_k + \sqrt{\frac{3\log t}{2T_k(t-1)}}, \tag{1.9}$$

or (when both previous inequalities (1.8) and (1.9) do not hold), then we deduce from (1.7) that

$$\mu_k + 2\sqrt{\frac{3\log t}{2T_k(t-1)}} \geq \mu^*,$$

which implies $T_k(t-1) \leq \frac{6\log t}{\Delta_k^2}$.

This says that whenever $T_k(t-1) \geq \frac{6\log t}{\Delta_k^2} + 1$, either arm k is not pulled at time t, or one of the two small probability events (1.8) or

(1.9) holds. Thus writing $u \stackrel{\text{def}}{=} \frac{6\log t}{\Delta_k^2} + 1$, we have:

$$\begin{aligned} T_k(n) &\leq u + \sum_{t=u+1}^{n} \mathbf{1}\{I_t = k; T_k(t) > u\} \\ &\leq u + \sum_{t=u+1}^{n} \mathbf{1}\{(1.8) \text{ or } (1.9) \text{ holds}\}. \end{aligned} \tag{1.10}$$

Now, the probability that (1.8) holds is bounded by

$$\mathbb{P}\Big(\exists 1 \leq s \leq t, \hat{\mu}_{k^*,s} + \sqrt{\frac{3\log t}{2s}} < \mu^*\Big) \leq \sum_{s=1}^{t} \frac{1}{t^3} = \frac{1}{t^2},$$

using Chernoff-Hoeffding inequality (1.5). Similarly the probability that (1.9) holds is bounded by $1/t^2$, thus by taking the expectation in (1.10) we deduce that

$$\begin{aligned} \mathbb{E}[T_k(n)] &\leq \frac{6\log(n)}{\Delta_k^2} + 1 + 2\sum_{t=u+1}^{n} \frac{1}{t^2} \\ &\leq \frac{6\log(n)}{\Delta_k^2} + \frac{\pi^2}{3} + 1 \end{aligned} \tag{1.11}$$

□

The previous bound depends on some properties of the distributions: the gaps Δ_k. The next result states a problem-independent bound.

Corollary 1.1. The expected regret of UCB is bounded as:

$$\mathbb{E}R_n \leq \sqrt{Kn(6\log n + \frac{\pi^2}{3} + 1)} \tag{1.12}$$

Proof. Using Cauchy-Schwarz inequality,

$$\begin{aligned} \mathbb{E}R_n &= \sum_k \Delta_k \sqrt{\mathbb{E}T_k(n)}\sqrt{\mathbb{E}T_k(n)} \\ &\leq \sqrt{\sum_k \Delta_k^2 \mathbb{E}T_k(n) \sum_k \mathbb{E}T_k(n)}. \end{aligned}$$

The result follows from (1.11) and that $\sum_k \mathbb{E}T_k(n) = n$ □

1.1.3 Lower bounds

There are two types of lower bounds: (1) The problem-dependent bounds [Lai and Robbins, 1985, Burnetas and Katehakis, 1996b] say that for any problem in a given class, an "admissible" algorithm will suffer -asymptotically- a logarithmic regret with a constant factor that depends on the arm distributions, (2) The problem-independent bounds [Cesa-Bianchi and Lugosi, 2006, Bubeck, 2010] states that for any algorithm and any time-horizon n, there exists an environment on which this algorithm will suffer a regret lower-bounded by some quantity.

Problem-dependent lower bounds: Lai and Robbins [1985] considered a class of one-dimensional parametric distributions and showed that any admissible strategy (i.e. such that the algorithm pulls each sub-optimal arm k a sub-polynomial number of times: $\forall \alpha > 0$, $\mathbb{E}T_k(n) = o(n^\alpha)$) will asymptotically pull in expectation any sub-optimal arm k a number of times such that:

$$\liminf_{n\to\infty} \frac{\mathbb{E}T_k(n)}{\log n} \geq \frac{1}{\mathcal{K}(\nu_k, \nu_{k^*})} \tag{1.13}$$

(which, from (1.2), enables the deduction of a lower bound on the regret), where $\mathcal{K}(\nu_k, \nu_{k^*})$ is the Kullback-Leibler (KL) divergence between ν_k and ν_{k^*} (i.e., $\mathcal{K}(\nu, \kappa) \stackrel{\text{def}}{=} \int_0^1 \frac{d\nu}{d\kappa} \log \frac{d\nu}{d\kappa} d\kappa$ if ν is dominated by κ, and $+\infty$ otherwise).

Burnetas and Katehakis [1996b] extended this result to several classes $\mathcal{P}$ of multi-dimensional parametric distributions. By writing

$$\mathcal{K}_{\inf}(\nu, \mu) \stackrel{\text{def}}{=} \inf_{\kappa \in \mathcal{P}: E(\kappa) > \mu} \mathcal{K}(\nu, \kappa),$$

(where μ is a real number such that $E(\nu) < \mu$), they showed the improved lower bound on the number of pulls of sub-optimal arms:

$$\liminf_{n\to\infty} \frac{\mathbb{E}T_k(n)}{\log n} \geq \frac{1}{\mathcal{K}_{\inf}(\nu_k, \mu^*)}. \tag{1.14}$$

Those bounds consider a fixed problem and show that any algorithm that is reasonably good on a class of problems (i.e. what we called an admissible strategy) cannot be extremely good on any specific instance,

and thus needs to suffer some incompressible regret. Note also that these problem-independent lower-bounds are of an asymptotic nature and do not say anything about the regret at any finite time n.

A problem independent lower-bound: In contrast to the previous bounds, we can also derive finite-time bounds that do not depend on the arm distributions: For any algorithm and any time horizon n, there exists an environment (arm distributions) such that this algorithm will suffer some incompressible regret on this environment [Cesa-Bianchi and Lugosi, 2006, Bubeck, 2010]:

$$\inf \sup \mathbb{E}R_n \geq \frac{1}{20}\sqrt{nK},$$

where the inf is taken over all possible algorithms and the sup over all possible (bounded) reward distributions of the arms.

1.1.4 Recent improvements

Notice that in the problem-dependent lower-bounds (1.13) and (1.14), the rate is logarithmic, like for the upper bound of UCB, however the constant factor is not the same. In the lower bound it uses KL divergences whereas in the upper bounds the constant is expressed in terms of the difference between the means. From Pinsker's inequality (see e.g. [Cesa-Bianchi and Lugosi, 2006]) we have: $\mathcal{K}(\nu,\kappa) \geq (E[\nu]-E[\kappa])^2$ and the discrepancy between $\mathcal{K}(\nu,\kappa)$ and $(E[\nu]-E[\kappa])^2$ can be very large (e.g. for Bernoulli distributions with parameters close to 0 or 1). It follows that there is a potentially large gap between the lower and upper bounds, which motivated several recent attempts to reduce this gap. The main line of research consisted in tightening the concentration inequalities defining the upper confidence bounds.

A first improvement was made by Audibert et al. [2009] who introduced UCB-V (UCB with variance estimate) that uses a variant of Bernstein's inequality to take into account the empirical variance of the rewards (in addition to their empirical mean) to define tighter UCB on the mean reward of the arms:

$$B_{t,s}(k) \stackrel{\text{def}}{=} \hat{\mu}_{k,s} + \sqrt{2\frac{V_{k,s}\log(1.2t)}{s}} + \frac{3\log(1.2t)}{s}, \tag{1.15}$$

where $V_{k,s}$ is the empirical variance of the rewards received from arm k. They proved that the regret is bounded as follows:

$$\mathbb{E}R_n \leq 10\Big(\sum_{k:\Delta_k>0} \frac{\sigma_k^2}{\Delta_k} + 2\Big)\log(n),$$

which scales with the actual variance σ_k^2 of the arms.

Then Honda and Takemura [2010, 2011] proposed the DMED (Deterministic Minimum Empirical Divergence) algorithm and proved an asymptotic bound that achieves the asymptotic lower-bound of Burnetas and Katehakis [1996b]. Notice that Lai and Robbins [1985] and Burnetas and Katehakis [1996b] also provided an algorithm with asymptotic guarantees (under more restrictive conditions). It is only in [Garivier and Cappé, 2011, Maillard et al., 2011, Cappé et al., 2013] that a finite-time analysis was derived for KL-based UCB algorithms, KL-UCB and $\mathcal{K}_{\text{inf}}$-UCB, that achieve the asymptotic lower bounds of [Lai and Robbins, 1985] and [Burnetas and Katehakis, 1996b] respectively. Those algorithms make use of KL divergences in the definition of the UCBs and use the full empirical reward distribution (and not only the two first moments). In addition to their improved analysis in comparison to regular UCB algorithms, several experimental studies showed their improved numerical performance.

Finally let us also mention that the logarithmic gap between the upper and lower problem-independent bounds (see (1.12) and (1.14)) has also been closed (up to a constant factor) by the MOSS algorithm of Audibert and Bubeck [2009], which achieves a minimax regret bound of order $\sqrt{Kn}$.

1.2 Extensions to many arms

The principle of optimism in the face of uncertainty has been successfully extended to several variants of the multi-armed stochastic bandit problem, notably when the number of arms is large (possibly infinite) compared to the number of rounds. In those situations one cannot even pull each arm once and thus in order to achieve meaningful results we need to make some assumptions about the unobserved arms. There are two possible situations:

- When the previously observed arms do not give us any information about unobserved arms. This is the case when there is no structure in the rewards. In those situations, we may rely on a probabilistic assumption on the mean value of any unobserved arm.

- When the previously observed arms can give us some information about unobserved arms: this is the case of structured rewards, for example when the mean reward function is a linear, convex, or Lipschitz function of the arm position, or also when the rewards depend on some tree, graph, or combinatorial structure.

1.2.1 Unstructured rewards

The so-called *many-armed bandit problem* considers a countably infinite number of arms where there is no structure among arms. Thus at any round t the rewards obtained by pulling previously observed arms do not give us information about the value of the unobserved arms.

To illustrate, think of the problem of selecting a restaurant for dinner in a big city like Paris. Each day you go to a restaurant and receive a reward indicating how much you enjoyed the food you were served. You may decide to go back to one of the restaurants you have already visited either because the food there was good (exploitation) or because you have not been there many times and want to try another dish (exploration). However you may also want to try a new restaurant (discovery) chosen randomly (maybe according to some prior information). Of course there are many other applications of this exploration-exploitation-discovery trade-off, such as in marketing (e.g. you want to send catalogs to good customers, uncertain customers, or random people), in mining for valuable resources (such as gold or oil) where you want to exploit good wells, explore unknown wells, or start digging at a new location.

A strong probabilistic assumption that has been made by Banks and Sundaram [1992], Berry et al. [1997] to model such situations is that the mean-value of any unobserved arm is a random variable that follows some known distribution. More recently this assumption

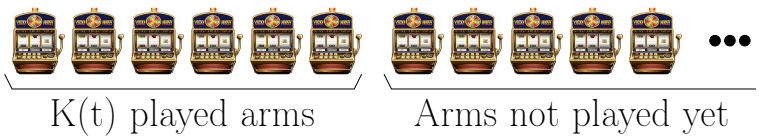

Figure 1.1: The UCB-AIR strategy: UCB-V algorithm is played on an increasing number $K(t)$ or arms

has been weakened by Wang et al. [2008] with an assumption focusing on this distribution upper tail only. More precisely, they assume that there exists $\beta > 0$ such that the probability that the mean-reward μ of a new randomly chosen arm is ϵ-optimal, is of order ϵ^β:

$$\mathbb{P}(\mu(\text{new arm}) > \mu^* - \epsilon) = \Theta(\epsilon^\beta), \tag{1.16}$$

[1]

where $\mu^* = \sup_{k \geq 1} \mu_k$ is the supremum of the mean-reward of the arms.

Thus the parameter β characterizes the probability of selecting a near-optimal arm. A large value of β indicates that there is a small chance that a new random arm will be good, thus an algorithm trying to achieve a low regret (defined like in (1.1) with respect to μ^*) would have to pull many new arms. Conversely, if β is small, then there is a reasonably large probability that a very good arm will be obtained by pulling a small number of new arms.

The UCB-AIR (UCB with Arm Increasing Rule) strategy introduced in Wang et al. [2008] consists of playing a UCB-V strategy [Audibert et al., 2009] (see (1.15)) on a set of current arms, whose number is increasing with time. At each round, either an arm already played is chosen according to the UCB-V strategy, or a new random arm is selected. Theorem 4 of [Wang et al., 2008] states that by selecting at each round t a number of active arms defined by

$$K(t) = \begin{cases} \lfloor t^{\frac{\beta}{2}} \rfloor & \text{if } \beta < 1 \text{ and } \mu^* < 1 \\ \lfloor t^{\frac{\beta}{\beta+1}} \rfloor & \text{if } \beta \geq 1 \text{ or } \mu^* = 1 \end{cases}$$

then the expected regret of UCB-AIR is upper-bounded as:

[1] We write $f(\epsilon) = \Theta(g(\epsilon))$ if $\exists c_1, c_2, \epsilon_0, \forall \epsilon \leq \epsilon_0, c_1 g(\epsilon) \leq f(\epsilon) \leq c_2 g(\epsilon)$.

$$\mathbb{E}R_n \leq \begin{cases} C(\log n)^2\sqrt{n} & \text{if } \beta < 1 \text{ and } \mu^* < 1 \\ C(\log n)^2 n^{\frac{\beta}{1+\beta}} & \text{if } \mu^* = 1 \text{ or } \beta \geq 1 \end{cases},$$

where C is a (numerical) constant.

This setting illustrates the *exploration-exploitation-discovery trade-off* where exploitation means pulling an apparently good arm (based on previous observations), exploration means pulling an uncertain arm (already pulled), and discovery means trying a new (unknown) arm.

An important aspect of this model is that the coefficient β characterizes the probability of choosing randomly a near-optimal arm (thus the proportion of near-optimal arms), and the UCB-AIR algorithm requires the knowledge of this coefficient (since β is used for the choice of $K(t)$). An open question is whether it is possible to design an *adaptive strategy* that could show similar performance when β is initially unknown.

Here we see an important characteristic of the performance of the optimistic strategy in a stochastic bandit setting, that will appear several times in different settings in the next chapters: The performance of a sequential decision making problem in a stochastic environment depends on a measure of the **quantity of near-optimal solutions,** as well as on **our knowledge** about this quantity.

1.2.2 Structured bandit problems

In structured bandit problems we assume that the mean-reward of an arm is a function of some arm parameters, where the function belongs to some known class. This includes situations where "arms" denote paths in a tree or a graph (and the reward of a path being the sum of rewards obtained along the edges), or points in some metric space where the mean-reward function possesses a specific structure.

A well-studied case is the *linear bandit* problem where the set of arms $\mathcal{X}$ lies in a Euclidean space $\mathbb{R}^d$ and the mean-reward function is linear with respect to (w.r.t.) the arm position $x \in \mathcal{X}$: at time t, one selects an arm $x_t \in \mathcal{X}$ and receives a reward $r_t \overset{\text{def}}{=} \mu(x_t) + \epsilon_t$, where the mean-reward is $\mu(x) \overset{\text{def}}{=} x \cdot \theta$ with $\theta \in \mathbb{R}^d$ is some (unknown) parameter, and ϵ_t is a (centered, independent) observation noise. The cumulative

regret is defined w.r.t. the best possible arm $x^* \stackrel{\text{def}}{=} \arg\max_{x\in\mathcal{X}} \mu(x)$:

$$R_n \stackrel{\text{def}}{=} n\mu(x^*) - \sum_{t=1}^{n} \mu(x_t).$$

Several optimistic algorithms have been introduced and analyzed, such as the *confidence ball* algorithms in [Dani et al., 2008], as well as refined variants in [Rusmevichientong and Tsitsiklis, 2010, Abbasi-Yadkori et al., 2011]. See also [Auer, 2003] for a pioneering work on this topic. The main bounds on the regret are either problem-dependent, of the order $O\Big(\frac{\log n}{\Delta}\Big)$ (where Δ is the mean-reward difference between the best and second best extremal points), or problem-independent of the order[2] $\widetilde{O}(d\sqrt{n})$. Several extensions to the linear setting have been considered, such as *Generalized Linear models* [Filippi et al., 2010] and *sparse linear bandits* [Carpentier and Munos, 2012, Abbasi-Yadkori et al., 2012].

Another popular setting is when the mean-reward function $x \mapsto \mu(x)$ is convex [Flaxman et al., 2005, Agarwal et al., 2011] in which case regret bounds of order $O(\text{poly}(d)\sqrt{n})$ can be achieved[3]. Other weaker assumptions on the mean-reward function have been considered, such as Lipschitz condition [Kleinberg, 2004, Agrawal, 1995a, Auer et al., 2007, Kleinberg et al., 2008b] or even weaker local assumptions in [Bubeck et al., 2011a, Valko et al., 2013]. This setting of bandits in metric spaces as well as more general spaces will be further investigated in Chapters 3 and 4.

1.3 Conclusions

It is worth mentioning that there have been a huge development of the field of Bandit Theory over the last few years which have produced emerging fields such as *contextual bandits* (where the rewards depend on some observed contextual information), *adversarial bandits* (where the rewards are chosen by an adversary instead of being stochastic), and has drawn strong links with other fields such as *online-learning*

[2] where $\widetilde{O}$ stands for a O notation up to a polylogarithmic factor
[3] where $\text{poly}(d)$ refers to a polynomial in d

(where a statistical learning task is performed online given limited feedback) and *learning from experts* (where one uses a set of recommendations given by experts). The interested reader may find additional references and developments in the following books and PhD theses [Cesa-Bianchi and Lugosi, 2006, Bubeck, 2010, Maillard, 2011, Bubeck and Cesa-Bianchi, 2012].

This chapter presented a brief overview of the multi-armed bandit problem which can be seen as a tool for rapidly selecting the best action among a set of possible ones, under the assumption that each reward sample provides information about the value (mean-reward) of the selected action. In the next chapters we will use this tool as a building block for solving more complicated problems where the action space is structured (for example when it is a sequence of actions, or a path in a tree) with a particular interest for *combining bandits in a hierarchy*. The next chapter introduces the historical motivation for our interest in this problem while the later chapters provide algorithmic and theoretical contributions.

References

Y. Abbasi-Yadkori, D. Pal, and Cs. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011.

Y. Abbasi-Yadkori, D. Pal, and Cs. Szepesvári. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, 2012.

Bruce Abramson. Expected-outcome: A general model of static evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:182–193, 1990.

A. Agarwal, D. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, 2011.

Alekh Agarwal, Peter Bartlett, Pradeep Ravikumar, and Martin Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transcations on Information Theory*, 58:5, 2012.

R. Agrawal. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33:1926–1951, 1995a.

R. Agrawal. Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995b.

S Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, 2012.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for Thompson sampling. In *Sixteenth International Conference on Artificial Intelligence and Statistics*, 2013.

John Asmuth and Michael L. Littman. Learning is planning: near Bayes-optimal reinforcement learning via Monte-Carlo tree search. In *Uncertainty in Artificial Intelligence*, 2011.

J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.

J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, 2010. URL `.files/ABM10.pdf`.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In Sanjot Dasgupta and Adam Klivans, editors, *Proceedings of the 22nd annual Conference On Learning Theory*, COLT '09, Montreal,Quebec, Canada, jun 2009.

P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. *20th Conference on Learning Theory*, pages 454–468, 2007.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, March 2003. ISSN 1532-4435.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32:48–77, January 2003. ISSN 0097-5397.

A. Auger and N. Hansen. *Theory of Randomized Search Heuristics: Foundations and Recent Developments*, chapter Theory of Evolution Strategies: A New Perspective, pages 289–325. World Scientific Publishing, 2011.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19:357–367, 1967.

J. S. Banks and R. Sundaram. Denumerable-armed bandits. *Econometrica*, 60:1071–1096, 1992.

Nicole Bäuerle and Ulrich Rieder. *Markov Decision Processes with Applications to Finance*. 2011.

D. A. Berry, R. W. Chen, A. Zame, D. C. Heath, and L. A. Shepp. Bandit problems with infinitely many arms. *Annals of Statistics*, (25):2103–2116, 1997.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Amine Bourki, Guillaume Chaslot, Matthieu Coulm, Vincent Danjean, Hassen Doghmen, Jean-Baptiste Hoock, Thomas Hérault, Arpad Rimmel, Fabien Teytaud, Olivier Teytaud, Paul Vayssière, and Ziqin Yu. Scalability and parallelization of monte-carlo tree search. In *International Conference on Computers and Games*, 2012.

B. Bouzy and B. Helmstetter. Monte-Carlo go developments. In Hiroyuki Iida Ernst A. Heinz H. Jaap van den Herik, editor, *Advances in Computer Games*, page 159âĂŞ174. Kluwer Academic Publishers, 2003.

Bruno Bouzy and Tristan Cazenave. Computer Go: an AI oriented survey. *Artif. Intell.*, 132(1):39–103, October 2001. ISSN 0004-3702. . URL `http://dx.doi.org/10.1016/S0004-3702(01)00127-8`.

Cameron Browne, Edward Powley, Daniel Whitehouse, Simon Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AIin Games*, 4(1), March 2012.

B. Brügmann. Monte Carlo Go. Technical report, Syracuse University, NY, USA, 1993.

S. Bubeck and R. Munos. Open loop optimistic planning. In *Conference on Learning Theory*, 2010.

S. Bubeck, R. Munos, G. Stoltz, and Cs. Szepesvári. Online optimization of X-armed bandits. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 22, pages 201–208. MIT Press, 2008. URL `http://hal.inria.fr/inria-00329797/en/`.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Proc. of the 20th International Conference on Algorithmic Learning Theory*, 2009.

S. Bubeck, R. Munos, G. Stoltz, and Cs. Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, 2011a. URL `http://arxiv.org/abs/1001.4475`.

S. Bubeck, G. Stoltz, and J. Y. Yu. Lipschitz bandits without the Lipschitz constant. In *International Conference on Algorithmic Learning Theory*, 2011b.

Sébastien Bubeck. *Bandits Games and Clustering Foundations*. PhD thesis, Université de Lille 1, 2010.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

L. Buşoniu, R. Munos, and R. Babuska. Optimistic planning in Markov decision processes. In Frank Lewis and Derong Liu, editors, *In Reinforcement Learning and Adaptive Dynamic Programming for feedback control*. Wiley, 2011a.

Lucian Buşoniu and Rémi Munos. Optimistic planning for markov decision processes. In *Proceedings 15th International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, page 182âĂŞ189, 2012.

Lucian Buşoniu, Robert Babuška, Bart De Schutter, and Damien Ernst. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. Automation and Control Engineering. Taylor & Francis CRC Press, 2010.

Lucian Buşoniu, Rémi Munos, Bart De Schutter, and Robert Babuška. Optimistic planning for sparsely stochastic systems. In *2011 IEEE International Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL-11)*, Paris, France, 11–15 April 2011b. Submitted to special session on *Active Reinforcement Learning*.

Adam D. Bull. Adaptive-treed bandits. Technical report, arXiv:1302.2489v2, 2013.

A.N. Burnetas and M.N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17:122–142, 1996a.

Apostolos N. Burnetas and Michaël N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2): 122–142, 1996b.

E. F. Camacho and C. Bordons. *Model Predictive Control*. Springer-Verlag, 2004.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

Alexandra Carpentier and Rémi Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *International Conference on Artificial Intelligence and Statistics*, 2012.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Hyeong Soo Chang, Michael C. Fu, Jiaqiao Hu, and Steven I. Marcus. *Simulation-based Algorithms for Markov Decision Processes*. Springer, London, 2007.

Guillaume Chaslot. *Monte-Carlo Tree Search*. PhD thesis, Maastricht University, 2010.

P.-A. Coquelin and R. Munos. Bandit algorithms for tree search. In *Uncertainty in Artificial Intelligence*, 2007.

Rémi Coulom. Efficient selectivity and backup operators in Monte-Carlo Tree Search. In LNCS, editor, *Computer Games*, volume 4630, pages 72–83, 2006.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In Rocco A. Servedio and Tong Zhang, editors, *Proceedings of the 21st annual Conference On Learning Theory*, volume 80 of *COLT '08*, pages 355–366, Helsinki, Finland, jul 2008. Omnipress.

Boris Defourny, Damien Ernst, and Louis Wehenkel. Lazy planning under uncertainties by optimizing decisions on an ensemble of incomplete disturbance trees. In S. Girgin, M. Loth, R. Munos, P. Preux, and D. Ryabko, editors, *Recent Advances in Reinforcement Learning*, volume 5323 of *Lecture Notes in Computer Science*, pages 1–14. Springer, 2008.

Michael Duff. *Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, Department of Computer Science, University of Massachusetts, Amherst, 2002.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvari. Parametric bandits: The generalized linear case. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 586–594. 2010.

D. E. Finkel and C. T. Kelley. Convergence analysis of the direct algorithm. Technical report, North Carolina State University, Center for, 2004.

Abraham D. Flaxman, Adam Tauman Kalai, and Hugh Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the 16th annual ACM-SIAM Symposium On Discrete Algorithms*, SODA '05, pages 385–394. SIAM, 2005.

C.A. Floudas. *Deterministic Global Optimization: Theory, Algorithms and Applications*. Kluwer Academic Publishers, Dordrecht / Boston / London, 1999.

R. Fonteneau, L. Busoniu, and R. Munos. Optimistic planning for belief-augmented Markov decision processes. In *IEEE International Symposium on Adaptive Dynamic Programming and reinforcement Learning*, 2013.

J. M. X. Gablonsky. *Modifications of the Direct algorithm*. PhD thesis, 2001.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, COLT '11, 2011.

S. Gelly, Y. Wang, R. Munos, and O. Teytaud. Modification of UCT with patterns in monte-carlo go. Technical report, INRIA RR-6062, 2006. URL `http://hal.inria.fr/inria-00117266_v3/`.

Sylvain Gelly and David Silver. Combining online and offline knowledge in UCT. In Zoubin Ghahramani, editor, *International Conference on Machine Learning*, volume 227 of *ICML '07, ACM International Conference Proceeding Series*, pages 273–280, Corvalis, Oregon, USA, jun 2007. ACM. ISBN 978-1-59593-793-3.

Sylvain Gelly and David Silver. Monte-carlo tree search and rapid action value estimation in computer Go. *Artificial Intelligence*, 175:1856–1875, 2011.

J.C. Gittins. Bandit processes and dynamic allocation indices. *In Journal of the Royal Statistical Society Series B*, 41(2):148–177, 1979.

John C. Gittins, Richard Weber, and Kevin Glazebrook. *Multi-armed Bandit Allocation Indices*. Wiley, 1989.

Arthur Guez, David Silver, and Peter Dayan. Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Advances in Neural Information Processing Systems*, 2012.

E.R. Hansen. *Global Optimization Using Interval Analysis*. Marcel Dekker, New York, 1992.

Eric A. Hansen and Shlomo Zilberstein. A heuristic search algorithm for Markov decision problems. In *Proceedings Bar-Ilan Symposium on the Foundation of Artificial Intelligence*, Ramat Gan, Israel, 23–25 June 1999. URL `http://rbr.cs.umass.edu/shlomo/papers/HZbisfai99.html`.

J. Honda and A. Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85: 361–391, 2011.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In Adam Tauman Kalai and Mehryar Mohri, editors, *Proceedings of the 23rd annual Conference On Learning Theory*, pages 67–79. Omnipress, June 2010. ISBN 978-0-9822529-2-5.

R. Horst and H. Tuy. *Global Optimization ? Deterministic Approaches*. Springer, Berlin / Heidelberg / New York, 3rd edition, 1996.

J-F. Hren and R. Munos. Optimistic planning of deterministic systems. In European Workshop on Reinforcement Learning Springer LNAI 5323, editor, *Recent Advances in Reinforcement Learning*, pages 151–164, 2008.

D. R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1):157–181, 1993.

Leslie P. Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71:291–307, 2005.

Emilie Kauffman, Olivier Cappé, and Aurélien Garivier. On Bayesian upper confidence bounds for bandit problems. In *International Conference on Artificial Intelligence and Statistics*, 2012.

Emilie Kauffmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite time analysis. In *International Conference on Algorithmic Learning Theory*, 2012.

Emilie Kaufmann, Nathan Korda, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Neural Information Processing Systems*, 2013.

R. B. Kearfott. *Rigorous Global Search: Continuous Problems*. Kluwer Academic Publishers, Dordrecht / Boston / London, 1996.

M. Kearns, Y. Mansour, and A.Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markovian decision processes. In *Machine Learning*, volume 49, pages 193–208, 2002a.

Michael J. Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning*, 49(2-3):193–208, 2002b.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th ACM Symposium on Theory of Computing*, 2008a.

Robert D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Proceedings of the 18th conference on advances in Neural Information Processing Systems*, NIPS '04, Vancouver, British Columbia, Canada, dec 2004. MIT Press.

Robert D. Kleinberg, Alexander Slivkins, and Eli Upfal. Multi-armed bandit problems in metric spaces. In *Proceedings of the 40th ACM symposium on Theory Of Computing*, TOC '08, pages 681–690, 2008b.

L. Kocsis and Cs. Szepesvári. Bandit based Monte-Carlo planning. In *Proceedings of the 17th European Conference on Machine Learning (ECML-2006)*, pages 282–293. 2006.

Jussi Kujala and Tapio Elomaa. Following the perturbed leader to gamble at multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, 2007.

Steven M. La Valle. *Planning Algorithms*. Cambridge University Press, 2006.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Chang-Shing Lee, Mei-Hui Wang, Guillaume Chaslot, Jean-Baptiste Hoock, Arpad Rimmel, Olivier Teytaud, Shang-Rong Tsai, Shun-Chin Hsu, and Tzung-Pei Hong. The computational intelligence of MoGo revealed in Taiwan's computer Go tournaments. *IEEE Trans. Comput. Intellig. and AI in Games*, 1(1):73–89, 2009.

J. M. Maciejowski. *Predictive Control with Constraints*. Prentice Hall, 2002.

Odalric Ambrym Maillard. *Apprentissage séquentiel: Bandits, Statistique et Renforcement*. PhD thesis, Université des Sciences et des Technologies de Lille 1, 2011.

Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of the 24th annual Conference On Learning Theory*, COLT '11, 2011.

R. Munos. Optimistic optimization of deterministic functions without the knowledge of its smoothness. In *Advances in Neural Information Processing Systems*, 2011.

A.S. Nemirovsky and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons Ltd, 1983.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

Neumaier. *Interval Methods for Systems of Equations*. Cambridge University Press, 1990.

N.J. Nilsson. *Principles of Artificial Intelligence*. Tioga Publishing, 1980.

L. Péret and F. Garcia. On-line search for solving large Markov decision processes. In *Proceedings of the 16th European Conference on Artificial Intelligence*, 2004.

Joelle Pineau, Geoffrey J. Gordon, and Sebastian Thrun. Anytime point-based approximations for large POMDPs. *Journal of Artificial Intelligence Research (JAIR)*, 27:335–380, 2006.

J.D. Pintér. *Global Optimization in Action (Continuous and Lipschitz Optimization: Algorithms, Implementations and Applications)*. Kluwer Academic Publishers, 1996.

M.L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994.

Arpad Rimmel, Fabien Teytaud, and Olivier Teytaud. Biasing Monte-Carlo simulations through RAVE values. In *International Conference on Computers and Games*, 2010.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.

S. Ross, J. Pineau, S. Paquet, and B. Chaib-Draa. Online planning algorithms for POMDPs. *Journal of Artificial Intelligence Research*, 32:663–704, 2008.

Stéphane Ross, Jeoelle Pineau, Brahim Chaib-draa, and Pierre Kreitmann. A Bayesian approach for learning and planning in partially observable Markov decision processes. *Journal of Machine Learning Research*, 12:1655–1696, 2011.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Math. Oper. Res.*, 35:395–411, May 2010.

Olivier Sigaud and Olivier Buffet, editors. *Markov Decision Processes in Artificial Intelligence*. Wiley, 2010.

David Silver. *Reinforcement Learning and Simulation-Based Search in Computer Go*. PhD thesis, University of Alberta, 2009.

David Silver and Joel Veness. Monte-carlo planning in large POMDPs. In *Advances in Neural Information Processing Systems*, 2012.

Aleksandrs Slivkins. Contextual bandits with similarity information. In *Proceedings of the 24th annual Conference On Learning Theory*, COLT '11, 2011.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, pages 1015–1022, 2010.

R.G. Strongin and Ya.D. Sergeyev. *Global Optimization with Non-Convex Constraints: Sequential and Parallel Algorithms*. Kluwer Academic Publishers, Dordrecht / Boston / London, 2000.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Morgan & Claypool Publishers, 2010.

William R. Thompson. On the theory of apportionment. *American Journal of Mathematics*, 57:450–456, 1935.

W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

J. F. Traub, G. W. Wasilkowski, and H. Wozniakowski. *Information-based Complexity*. Academic Press, New York, 1988.

Michal Valko, Alexandra Carpentier, and Rémi Munos. Stochastic simultaneous optimistic optimization. In *International Conference on Machine Learning*, 2013.

Nikos Vlassis, Mohammad Ghavamzadeh, Shie Mannor, and Pascal Poupart. *Reinforcement Learning: State of the Art*, chapter Bayesian Reinforcement Learning. Springer Verlag, 2012.

Tao Wang, Daniel Lizotte, Michael Bowling, and Dale Schuurmans. Bayesian sparse sampling for on-line reward optimization. In *International Conference on Machine Learning*, 2005.

Yizao Wang and Sylvain Gelly. Modifications of UCT and sequence-like simulations for Monte-Carlo Go. In *IEEE Symposium on Computational Intelligence and Games,*, pages 175–182, 2007.

Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for infinitely many-armed bandits. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *Proceedings of the 22nd conference on advances in Neural Information Processing Systems*, NIPS '08, pages 1729–1736, Vancouver, British Columbia, Canada, dec 2008. MIT Press.