Full text available at: http://dx.doi.org/10.1561/2200000024

Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems

Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems

Sébastien Bubeck

Princeton University
Princeton, NJ 08544
USA
sbubeck@princeton.edu

Nicolò Cesa-Bianchi

Università degli Studi di Milano
Milano 20135
Italy
nicolo.cesa-bianchi@unimi.it

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
USA
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 S. Bubeck and N. Cesa-Bianchi, Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems, Foundation and Trends® in Machine Learning, vol 5, no 1, pp 1–122, 2012.

ISBN: 978-1-60198-626-9

Foundations and Trends® in Machine Learning
Volume 5 Issue 1, 2012

Editorial Board

Editor-in-Chief:
Michael Jordan
Department of Electrical Engineering and Computer Science
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720-1776

Editors

Editorial Scope

Foundations and Trends® in Machine Learning will publish survey and tutorial articles in 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, 2012, Volume 5, 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. 5, No. 1 (2012) 1–122

DOI: 10.1561/2200000024

now
the essence of knowledge

Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems

Sébastien Bubeck[1] and Nicolò Cesa-Bianchi[2]

[1] *Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA, sbubeck@princeton.edu*
[2] *Dipartimento di Informatica, Università degli Studi di Milano, Milano 20135, Italy, nicolo.cesa-bianchi@unimi.it*

Abstract

Multi-armed bandit problems are the most basic examples of sequential decision problems with an exploration–exploitation trade-off. This is the balance between staying with the option that gave highest payoffs in the past and exploring new options that might give higher payoffs in the future. Although the study of bandit problems dates back to the 1930s, exploration–exploitation trade-offs arise in several modern applications, such as ad placement, website optimization, and packet routing. Mathematically, a multi-armed bandit is defined by the payoff process associated with each option. In this monograph, we focus on two extreme cases in which the analysis of regret is particularly simple and elegant: i.i.d. payoffs and adversarial payoffs. Besides the basic setting of finitely many actions, we also analyze some of the most important variants and extensions, such as the contextual bandit model.

Contents

1

Introduction

A multi-armed bandit problem (or, simply, a bandit problem) is a sequential allocation problem defined by a set of actions. At each time step, a unit resource is allocated to an action and some observable payoff is obtained. The goal is to maximize the total payoff obtained in a sequence of allocations. The name *bandit* refers to the colloquial term for a slot machine ("one-armed bandit" in American slang). In a casino, a sequential allocation problem is obtained when the player is facing many slot machines at once (a "multi-armed bandit") and must repeatedly choose where to insert the next coin.

Bandit problems are basic instances of sequential decision making with limited information and naturally address the fundamental trade-off between exploration and exploitation in sequential experiments. Indeed, the player must balance the exploitation of actions that did well in the past and the exploration of actions that might give higher payoffs in the future.

Although the original motivation of Thompson [162] for studying bandit problems came from clinical trials (when different treatments are available for a certain disease and one must decide which treatment to use on the next patient), modern technologies have created

many opportunities for new applications, and bandit problems now play an important role in several industrial domains. In particular, online services are natural targets for bandit algorithms, because there one can benefit from adapting the service to the individual sequence of requests. We now describe a few concrete examples in various domains.

Ad placement is the problem of deciding which advertisement to display on the web page delivered to the next visitor of a website. Similarly, website optimization deals with the problem of sequentially choosing design elements (font, images, layout) for the web page. Here the payoff is associated with visitor's actions, e.g., clickthroughs or other desired behaviors. Of course there are important differences with the basic bandit problem: in ad placement the pool of available ads (bandit arms) may change over time, and there might be a limit on the number of times each ad could be displayed.

In source routing a sequence of packets must be routed from a source host to a destination host in a given network, and the protocol allows to choose a specific source-destination path for each packet to be sent. The (negative) payoff is the time it takes to deliver a packet, and depends additively on the congestion of the edges in the chosen path.

In computer game-playing, each move is chosen by simulating and evaluating many possible game continuations after the move. Algorithms for bandits (more specifically, for a tree-based version of the bandit problem) can be used to explore more efficiently the huge tree of game continuations by focusing on the most promising subtrees. This idea has been successfully implemented in the MoGo player of Gelly et al. [85], which plays Go at world-class level. MoGo is based on the UCT strategy for hierarchical bandits of Kocsis and Szepesvári [123], which in turn is derived from the UCB bandit algorithm — see Section 2.

There are three fundamental formalizations of the bandit problem depending on the assumed nature of the reward process: stochastic, adversarial, and Markovian. Three distinct playing strategies have been shown to effectively address each specific bandit model: the UCB algorithm in the stochastic case, the Exp3 randomized algorithm in the adversarial case, and the so-called Gittins indices in the Markovian case. In this monograph, we focus on stochastic and adversarial bandits,

and refer the reader to the monograph by Mahajan and Teneketzis [130] or to the recent monograph by Gittins et al. [86] for an extensive analysis of Markovian bandits.

In order to analyze the behavior of a player or forecaster (i.e., the agent implementing a bandit strategy), we may compare its performance with that of an optimal strategy that, for any horizon of n time steps, consistently plays the arm that is best in the first n steps. In other terms, we may study the *regret* of the forecaster for not playing always optimally. More specifically, given $K \geq 2$ arms and sequences $X_{i,1}, X_{i,2}, \ldots$ of unknown rewards associated with each arm $i = 1, \ldots, K$, we study forecasters that at each time step $t = 1, 2, \ldots$ select an arm I_t and receive the associated reward $X_{I_t,t}$. The regret after n plays $I_1, \ldots, I_n$ is defined by

$$R_n = \max_{i=1,\ldots,K} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}. \tag{1.1}$$

If the time horizon is not known in advance we say that the forecaster is *anytime*.

In general, both rewards $X_{i,t}$ and forecaster's choices I_t might be stochastic. This allows to distinguish between the two following notions of averaged regret: the *expected regret*

$$\mathbb{E}\, R_n = \mathbb{E}\left[\max_{i=1,\ldots,K} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}\right] \tag{1.2}$$

and the *pseudo-regret*

$$\overline{R}_n = \max_{i=1,\ldots,K} \mathbb{E}\left[\sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}\right]. \tag{1.3}$$

In both definitions, the expectation is taken with respect to the random draw of both rewards and forecaster's actions. Note that pseudo-regret is a weaker notion of regret, since one competes against the action which is optimal only in expectation. The expected regret, instead, is the expectation of the regret with respect to the action which is optimal on the sequence of reward realizations. More formally one has $\overline{R}_n \leq \mathbb{E} R_n$.

In the original formalization of Robbins [146], which builds on the work of Wald [164] — see also Arrow et al. [16], each arm $i = 1, \ldots, K$

corresponds to an unknown probability distribution ν_i on $[0,1]$, and rewards $X_{i,t}$ are independent draws from the distribution ν_i corresponding to the selected arm.

The stochastic bandit problem

Known parameters: number of arms K and (possibly) number of rounds $n \geq K$.
Unknown parameters: K probability distributions $\nu_1, \ldots, \nu_K$ on $[0,1]$.

For each round $t = 1, 2, \ldots$

(1) the forecaster chooses $I_t \in \{1, \ldots, K\}$;
(2) given I_t, the environment draws the reward $X_{I_t,t} \sim \nu_{I_t}$ independently from the past and reveals it to the forecaster.

For $i = 1, \ldots, K$ we denote by μ_i the mean of ν_i (mean reward of arm i). Let

$$\mu^* = \max_{i=1,\ldots,K} \mu_i \quad \text{and} \quad i^* \in \operatorname*{argmax}_{i=1,\ldots,K} \mu_i.$$

In the stochastic setting, it is easy to see that the pseudo-regret can be written as

$$\overline{R}_n = n\mu^* - \sum_{t=1}^{n} \mathbb{E}[\mu_{I_t}]. \tag{1.4}$$

The analysis of the stochastic bandit model was pioneered in the seminal paper of Lai and Robbins [125], who introduced the technique of upper confidence bounds for the asymptotic analysis of regret. In Section 2 we describe this technique using the simpler formulation of Agrawal [9], which naturally lends itself to a finite-time analysis.

In parallel to the research on stochastic bandits, a game-theoretic formulation of the trade-off between exploration and exploitation has been independently investigated, although for quite some time this alternative formulation was not recognized as an instance of the multi-armed bandit problem. In order to motivate these game-theoretic bandits, consider again the initial example of gambling on slot machines. We now assume that we are in a rigged casino, where for each slot machine $i = 1, \ldots, K$ and time step $t \geq 1$ the owner sets the gain $X_{i,t}$ to some arbitrary (and possibly maliciously chosen) value

$g_{i,t} \in [0,1]$. Note that it is not in the interest of the owner to simply set all the gains to zero (otherwise, no gamblers would go to that casino). Now recall that a forecaster selects sequentially one arm $I_t \in \{1,\dots,K\}$ at each time step $t = 1,2,\dots$ and observes (and earns) the gain $g_{I_t,t}$. Is it still possible to minimize regret in such a setting?

Following a standard terminology, we call adversary, or opponent, the mechanism setting the sequence of gains for each arm. If this mechanism is independent of the forecaster's actions, then we call it an *oblivious* adversary. In general, however, the adversary may adapt to the forecaster's past behavior, in which case we speak of a *nonoblivious* adversary. For instance, in the rigged casino the owner may observe the way a gambler plays in order to design even more evil sequences of gains. Clearly, the distinction between oblivious and nonoblivious adversary is only meaningful when the player is randomized (if the player is deterministic, then the adversary can pick a bad sequence of gains right at the beginning of the game by simulating the player's future actions). Note, however, that in the presence of a nonoblivious adversary the interpretation of regret is ambiguous. Indeed, in this case the assignment of gains $g_{i,t}$ to arms $i = 1,\dots,K$ made by the adversary at each step t is allowed to depend on the player's past randomized actions $I_1,\dots,I_{t-1}$. In other words, $g_{i,t} = g_{i,t}(I_1,\dots,I_{t-1})$ for each i and t. Now, the regret compares the player's cumulative gain to that obtained by playing the single best arm for the first n rounds. However, had the player consistently chosen the same arm i in each round, namely $I_t = i$ for $t = 1,\dots,n$, the adversarial gains $g_{i,t}(I_1,\dots,I_{t-1})$ would have been possibly different than those actually experienced by the player.

The study of nonoblivious regret is mainly motivated by the connection between regret minimization and equilibria in games — see, e.g., [24, Section 9]. Here we just observe that game-theoretic equilibria are indeed defined similarly to regret: in equilibrium, the player has nSo incentive to behave differently, provided the opponent does not react to changes in the player's behavior. Interestingly, regret minimization has been also studied against *reactive opponents*, see for instance the works of Pucci de Farias and Megiddo [144] and Arora et al. [14].

The adversarial bandit problem

Known parameters: number of arms $K \geq 2$ and (possibly) number of rounds $n \geq K$.

For each round $t = 1, 2, \ldots$

(1) the forecaster chooses $I_t \in \{1, \ldots, K\}$, possibly with the help of external randomization,
(2) simultaneously, the adversary selects a gain vector $g_t = (g_{1,t}, \ldots, g_{K,t}) \in [0,1]^K$, possibly with the help of external randomization, and
(3) the forecaster receives (and observes) the reward $g_{I_t,t}$, while the gains of the other arms are not observed.

In this adversarial setting the goal is to obtain regret bounds in high probability or in expectation with respect to any possible randomization in the strategies used by the forecaster or the opponent, and irrespective of the opponent. In the case of a nonoblivious adversary this is not an easy task, and for this reason we usually start by bounding the pseudo-regret

$$\overline{R}_n = \max_{i=1,\ldots,K} \mathbb{E}\left[\sum_{t=1}^{n} g_{i,t} - \sum_{t=1}^{n} g_{I_t,t}\right].$$

Note that the randomization of the adversary is not very important here since we ask for bounds which hold for any opponent. On the other hand, it is fundamental to allow randomization for the forecaster — see Section 3 for details and basic results in the adversarial bandit model. This adversarial, or nonstochastic, version of the bandit problem was originally proposed as a way of playing an unknown game against an opponent. The problem of playing a game repeatedly, now a classical topic in game theory, was initiated by the groundbreaking work of James Hannan and David Blackwell. In Hannan's seminal paper Hannan [92], the game (i.e., the payoff matrix) is assumed to be known by the player, who also observes the opponent's moves in each play. Later, Baños [28] considered the problem of a repeated unknown game, where in each game round the player only observes its own payoff. This problem turns out to be exactly equivalent to the adversarial bandit

problem with a nonoblivious adversary. Simpler strategies for playing unknown games were more recently proposed by Foster and Vohra [81] and Hart and Mas-Colell [93, 94]. Approximately at the same time, the problem was re-discovered in computer science by Auer et al. [24]. It was them who made apparent the connection to stochastic bandits by coining the term nonstochastic multi-armed bandit problem.

The third fundamental model of multi-armed bandits assumes that the reward processes are neither i.i.d. (like in stochastic bandits) nor adversarial. More precisely, arms are associated with K Markov processes, each with its own state space. Each time an arm i is chosen in state s, a stochastic reward is drawn from a probability distribution $\nu_{i,s}$, and the state of the reward process for arm i changes in a Markovian fashion, based on an underlying stochastic transition matrix M_i. Both reward and new state are revealed to the player. On the other hand, the state of arms that are not chosen remains unchanged. Going back to our initial interpretation of bandits as sequential resource allocation processes, here we may think of K competing projects that are sequentially allocated a unit resource of work. However, unlike the previous bandit models, in this case the state of a project that gets the resource may change. Moreover, the underlying stochastic transition matrices M_i are typically assumed to be known, thus the optimal policy can be computed via dynamic programming and the problem is essentially of computational nature. The seminal result of Gittins [87] provides an optimal greedy policy which can be computed efficiently.

A notable special case of Markovian bandits is that of Bayesian bandits. These are parametric stochastic bandits, where the parameters of the reward distributions are assumed to be drawn from known priors, and the regret is computed by also averaging over the draw of parameters from the prior. The Markovian state change associated with the selection of an arm corresponds here to updating the posterior distribution of rewards for that arm after observing a new reward.

Markovian bandits are a standard model in the areas of Operations Research and Economics. However, the techniques used in their analysis are significantly different from those used to analyze stochastic and adversarial bandits. For this reason, in this monograph we do not cover Markovian bandits and their many variants.

References

[1] Y. Abbasi-Yadkori, D. Pal, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[2] N. Abe and P. M. Long, "Associative reinforcement learning using linear probabilistic concepts," in *Proceedings of the International Conference on Machine Learning (ICML)*, 1999.

[3] J. Abernethy, E. Hazan, and A. Rakhlin, "Competing in the dark: An efficient algorithm for bandit linear optimization," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2008.

[4] J. Abernethy and A. Rakhlin, "Beating the adaptive bandit with high probability," in *In Proceedings of the Annual Conference on Learning Theory (COLT)*, 2009.

[5] A. Agarwal, P. Bartlett, and M. Dama, "Optimal allocation strategies for the dark pool problem," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR Workshop and Conference Proceedings Volume 9, 2010.

[6] A. Agarwal, O. Dekel, and L. Xiao, "Optimal algorithms for online convex optimization with multi-point bandit feedback," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2010.

[7] A. Agarwal, J. Duchi, P. L. Bartlett, and C. Levrard, "Oracle inequalities for computationally budgeted model selection," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, JMLR Workshop and Conference Proceedings Volume 19, 2011.

[8] A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin, "Stochastic convex optimization with bandit feedback," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[9] R. Agrawal, "Sample mean based index policies with $\mathcal{O}(\log n)$ regret for the multi-armed bandit problem," *Advances in Applied Mathematics*, vol. 27, pp. 1054–1078, 1995.

[10] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, JMLR Workshop and Conference Proceedings Volume 23, 2012.

[11] C. Allenberg, P. Auer, L. Györfi, and G. Ottucsák, "Hannan consistency in on-line learning in case of unbounded losses under partial monitoring," in *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2006.

[12] K. Amin, M. Kearns, and U. Syed, "Bandits, query learning, and the haystack dimension," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, JMLR Workshop and Conference Proceedings Volume 19, 2011.

[13] A. Antos, V. Grover, and C. Szepesvári, "Active learning in multi-armed bandits," in *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2008.

[14] R. Arora, O. Dekel, and A. Tewari, "Online bandit learning against an adaptive adversary: From regret to policy regret," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

[15] S. Arora, E. Hazan, and S. Kale, "The multiplicative weights update method: A meta-algorithm and applications," *Theory of Computing*, vol. 8, pp. 121–164, 2012.

[16] K. J. Arrow, D. Blackwell, and M. A. Girshick, "Bayes and minimax solutions of sequential decision problems," *Econometrica*, pp. 213–244, 1949.

[17] J.-Y. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2009.

[18] J.-Y. Audibert and S. Bubeck, "Regret bounds and minimax policies under partial monitoring," *Journal of Machine Learning Research*, vol. 11, pp. 2635–2686, 2010.

[19] J.-Y. Audibert, S. Bubeck, and G. Lugosi, "Minimax policies for combinatorial prediction games," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, JMLR Workshop and Conference Proceedings Volume 19, 2011.

[20] J.-Y. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2010.

[21] J.-Y. Audibert, R. Munos, and C. Szepesvári, "Exploration-exploitation trade-off using variance estimates in multi-armed bandits," *Theoretical Computer Science*, vol. 410, pp. 1876–1902, 2009.

[22] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research*, vol. 3, pp. 397–422, 2002.

[23] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multi-armed bandit problem," *Machine Learning Journal*, vol. 47, no. 2–3, pp. 235–256, 2002.

[24] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire, "The non-stochastic multi-armed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[25] P. Auer and R. Ortner, "UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem," *Periodica Mathematica Hungarica*, vol. 61, pp. 55–65, 2010.

[26] P. Auer, R. Ortner, and C. Szepesvári, "Improved rates for the stochastic continuum-armed bandit problem," in *Proceedings of the International Conference on Learning Theory (COLT)*, 2007.

[27] B. Awerbuch and R. Kleinberg, "Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches," in *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, 2004.

[28] A. Baños, "On pseudo-games," *Annals of Mathematical Statistics*, vol. 39, pp. 1932–1945, 1968.

[29] M. Babaioff, R. D. Kleinberg, and A. Slivkins, "Truthful mechanisms with implicit payment computation," in *ACM Conference on Electronic Commerce 2010 (EC)*, 2010.

[30] M. Babaioff, Y. Sharma, and A. Slivkins, "Characterizing truthful multi-armed bandit mechanisms," in *ACM Conference on Electronic Commerce 2009 (EC)*, 2009.

[31] F. Bach and E. Moulines, "Non-asymptotic analysis of stochastic approximation algorithms for machine learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[32] K. Ball, "An elementary introduction to modern convex geometry," in *Flavors of Geometry*, (S. Levy, ed.), pp. 1–58, Cambridge University Press, 1997.

[33] P. Bartlett, V. Dani, T. Hayes, S. Kakade, A. Rakhlin, and A. Tewari, "High probability regret bounds for online optimization," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2008.

[34] G. Bartok, D. Pal, and C. Szepesvári, "Toward a classification of finite partial-monitoring games," in *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2010.

[35] G. Bartok, D. Pal, C. Szepesvári, and I. Szita, "Online learning," *Lecture Notes*, 2011.

[36] A. Beck and M. Teboulle, "Mirror Descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[37] A. Ben-Tal and A. Nemirovski, "The conjugate barrier Mirror Descent method for non-smooth convex optimization," Technical Report, MINERVA Optimization Center Report, Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa, 1999.

[38] D. A. Berry, R. W. Chen, D. C. Heath, A. Zame, and L. A. Shepp, "Bandit problems with infinitely many arms," *Annals of Statistics*, vol. 25, pp. 2103–2116, 1997.

[39] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire, "Contextual bandit algorithms with supervised learning guarantees," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR Workshop and Conference Proceedings Volume 15, 2011.

[40] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. E. Schapire, "Contextual bandit algorithms with supervised learning guarantees," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR Workshop and Conference Proceedings Volume 15, 2011.

[41] S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification: A survey of recent advances," *ESAIM: Probability and Statistics*, vol. 9, pp. 323–375, 2005.

[42] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[43] S. Bubeck, "Bandits games and clustering foundations," PhD thesis, Université Lille 1, 2010.

[44] S. Bubeck, "Introduction to online optimization," *Lecture Notes*, 2011.

[45] S. Bubeck, N. Cesa-Bianchi, and S. M. Kakade, "Towards minimax policies for online linear optimization with bandit feedback," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, JMLR Workshop and Conference Proceedings Volume 23, 2012.

[46] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi, "Bandits with heavy tail," Arxiv preprint arXiv:1209.1727, 2012.

[47] S. Bubeck, D. Ernst, and A. Garivier, "Optimal discovery with probabilistic expert advice," Arxiv preprint arXiv:1110.5447, 2011.

[48] S. Bubeck and R. Munos, "Open loop optimistic planning," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2010.

[49] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2009.

[50] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in finitely-armed and continuously-armed bandits," *Theoretical Computer Science*, vol. 412, pp. 1832–1852, 2011.

[51] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári, "Online optimization in $\mathcal{X}$-armed bandits," in *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[52] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári, "$\mathcal{X}$-armed bandits," *Journal of Machine Learning Research*, vol. 12, pp. 1587–1627, 2011.

[53] S. Bubeck and A. Slivkins, "The best of both worlds: Stochastic and adversarial bandits," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, JMLR Workshop and Conference Proceedings Volume 23, 2012.

[54] S. Bubeck, T. Wang, and N. Viswanathan, "Multiple identifications in multi-armed bandits," Arxiv preprint arXiv:1205.3181, 2012.

[55] L. Bui, R. Johari, and S. Mannor, "Committing bandits," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[56] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for Markov decision processes," *Mathematics of Operations Research*, pp. 222–255, 1997.

[57] R. Busa-Fekete and B. Kegl, "Fast boosting using adversarial bandits," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

[58] O. Cappé, A. Garivier, O. Maillard, R. Munos, and G. Stoltz, "Kullback-Leibler upper confidence bounds for optimal sequential allocation," Arxiv preprint arXiv:1210.1136, 2012.

[59] A. Carpentier, A. Lazaric, M. Ghavamzadeh, R. Munos, and P. Auer, "Upper confidence bounds algorithms for active learning in multi-armed bandits," in *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2011.

[60] A. Carpentier and R. Munos, "Finite time analysis of stratified sampling for Monte Carlo," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[61] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[62] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits," *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1404–1422, 2012.

[63] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz, "Minimizing regret with label efficient prediction," *IEEE Transactions on Information Theory*, vol. 51, pp. 2152–2162, 2005.

[64] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz, "Improved second-order bounds for prediction with expert advice," *Machine Learning*, vol. 66, pp. 321–352, 2007.

[65] O. Chapelle and L. Li, "An empirical evaluation of Thompson sampling," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[66] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR Workshop and Conference Proceedings Volume 15, 2011.

[67] A. Conn, K. Scheinberg, and L. Vicente, "Introduction to derivative-free optimization," in *Society for Industrial and Applied Mathematics (SIAM)*, 2009.

[68] E. W. Cope, "Regret and convergence bounds for a class of continuum-armed bandit problems," *IEEE Transactions on Automatic Control*, vol. 54, no. 6, pp. 1243–1253, 2009.

[69] P.-A. Coquelin and R. Munos, "Bandit algorithms for tree search," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

[70] K. Crammer and C. Gentile, "Multiclass classification with bandit feedback using adaptive regularization," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

[71] V. Dani, T. Hayes, and S. Kakade, "The price of bandit information for online optimization," in *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[72] V. Dani, T. Hayes, and S. Kakade, "Stochastic linear optimization under bandit feedback," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2008.

[73] N. Devanur and S. M. Kakade, "The price of truthfulness for pay-per-click auctions," in *ACM Conference on Electronic Commerce 2009 (EC)*, 2009.

[74] M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang, "Efficient optimal learning for contextual bandits," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2011.

[75] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and Markov decision processes," in *Proceedings of the Annual Conference on Computational Learning Theory (COLT)*, 2002.

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

[77] S. Filippi, O. Cappé, and A. Garivier, "Optimally sensing a single channel without prior information: The tiling algorithm and regret bounds," *Selected Topics in Signal Processing*, vol. 5, no. 1, pp. 68–76, 2011.

[78] S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári, "Parametric bandits: The generalized linear case," in *Neural Information Processing Systems (NIPS)*, 2010.

[79] A. Flaxman, A. Kalai, and B. McMahan, "Online convex optimization in the bandit setting: Gradient descent without a gradient," in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2005.

[80] D. Foster and A. Rakhlin, "No internal regret via neighborhood watch," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR Workshop and Conference Proceedings Volume 22, 2012.

[81] D. Foster and R. Vohra, "Asymptotic calibration," *Biometrika*, vol. 85, pp. 379–390, 1998.

[82] V. Gabillon, M. Ghavamzadeh, A. Lazaric, and S. Bubeck, "Multi-bandit best arm identification," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[83] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, JMLR Workshop and Conference Proceedings Volume 19, 2011.

[84] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2011.

[85] S. Gelly, Y. Wang, R. Munos, and O. Teytaud, "Modification of UCT with patterns in Monte-Carlo Go," Technical Report RR-6062, INRIA, 2006.

[86] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed Bandit Allocation Indices*. John Wiley and Sons, 2nd ed., 2011.

[87] J. C. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 148–177, 1979.

[88] A. Grove, N. Littlestone, and D. Schuurmans, "General convergence results for linear discriminant updates," *Machine Learning*, vol. 43, pp. 173–210, 2001.

[89] S. Grünewälder, J.-Y. Audibert, M. Opper, and J. Shawe-Taylor, "Regret bounds for Gaussian process bandit problems," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR Workshop and Conference Proceedings Volume 9, 2010.

[90] A. György, L. Kocsis, I. Szabó, and C. Szepesvári, "Continuous time associative bandit problems," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[91] A. György, T. Linder, G. Lugosi, and G. Ottucsák, "The on-line shortest path problem under partial monitoring," *Journal of Machine Learning Research*, vol. 8, pp. 2369–2403, 2007.

[92] J. Hannan, "Approximation to Bayes risk in repeated play," *Contributions to the Theory of Games*, vol. 3, pp. 97–139, 1957.

[93] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, pp. 1127–1150, 2000.

[94] S. Hart and A. Mas-Colell, "A general class of adaptive strategies," *Journal of Economic Theory*, vol. 98, pp. 26–54, 2001.

[95] E. Hazan, "The convex optimization approach to regret minimization," in *Optimization for Machine Learning*, (S. Sra, S. Nowozin, and S. Wright, eds.), pp. 287–303, MIT Press, 2011.

[96] E. Hazan and S. Kale, "Better algorithms for benign bandits," in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 38–47, 2009.

[97] E. Hazan and S. Kale, "NEWTRON: An efficient bandit algorithm for online multiclass prediction," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[98] E. Hazan, S. Kale, and M. Warmuth, "Learning rotations with little regret," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2010.

[99] D. Helmbold and S. Panizza, "Some label efficient learning results," in *Proceedings of the Annual Conference on Computational Learning Theory (COLT)*, 1997.

[100] D. P. Helmbold and M. Warmuth, "Learning permutations with exponential weights," *Journal of Machine Learning Research*, vol. 10, pp. 1705–1736, 2009.

[101] M. Herbster and M. Warmuth, "Tracking the best expert," *Machine Learning*, vol. 32, pp. 151–178, 1998.

[102] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*. Springer, 2001.

[103] J. Honda and A. Takemura, "An asymptotically optimal bandit algorithm for bounded support models," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2010.

[104] J.-B. Hoock and O. Teytaud, "Bandit-based genetic programming," in *Proceedings of the European Conference on Genetic Programming (EuroGP)*, 2010.

[105] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *Journal of Machine Learning Research*, vol. 11, pp. 1563–1600, 2010.

[106] A. Juditsky, A. Nazin, A. Tsybakov, and N. Vayatis, "Recursive aggregation of estimators by the Mirror Descent algorithm with averaging," *Problems of Information Transmission*, vol. 41, pp. 368–384, 2005.

[107] S. Kakade, S. Shalev-Shwartz, and A. Tewari, "Regularization techniques for learning with matrices," *Journal of Machine Learning Research*, vol. 13, pp. 1865–1890, 2012.

[108] S. M. Kakade, "On the sample complexity of reinforcement learning," PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[109] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari, "Efficient bandit algorithms for online multiclass prediction," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[110] A. Kalai and S. Vempala, "Efficient algorithms for online decision problems," *Journal of Computer and System Sciences*, vol. 71, pp. 291–307, 2005.

[111] S. Kale, L. Reyzin, and R. Schapire, "Non-stochastic bandit slate problems," in *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[112] V. Kanade, B. McMahan, and B. Bryan, "Sleeping experts and bandits with stochastic action availability and adversarial rewards," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR Workshop and Conference Proceedings Volume 5, 2009.

[113] V. Kanade and T. Steinke, "Learning hurdles for sleeping experts," in *Proceedings of the Innovations in Theoretical Computer Science Conference*, 2012.

[114] E. Kaufmann, O. Cappé, and A. Garivier, "On Bayesian upper confidence bounds for bandits problems," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR Workshop and Conference Proceedings Volume 22, 2012.

[115] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2012.

[116] M. Kearns, Y. Mansour, and A. Y. Ng, "A sparse sampling algorithm for near-optimal planning in large Markovian decision processes," *Machine Learning*, vol. 49, pp. 193–208, 2002.

[117] J. Kiefer, "Sequential minimax search for a maximum," *Proceedings of the American Mathematical Society*, vol. 4, no. 3, pp. 502–506, 1953.

[118] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Annals of Mathematical Statistics*, vol. 23, pp. 462–466, 1952.

[119] J. Kivinen and M. Warmuth, "Relative loss bounds for multidimensional regression problems," *Machine Learning*, vol. 45, pp. 301–329, 2001.

[120] R. Kleinberg, "Nearly tight bounds for the continuum-armed bandit problem," in *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[121] R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma, "Regret bounds for sleeping experts and bandits," *Machine Learning*, vol. 80, pp. 245–272, 2010.

[122] R. Kleinberg, A. Slivkins, and E. Upfal, "Multi-armed bandits in metric spaces," in *Proceedings of the ACM Symposium on Theory of Computing (STOC)*, 2008.

[123] L. Kocsis and C. Szepesvári, "Bandit based Monte-Carlo planning," in *Proceedings of the European Conference on Machine Learning (ECML)*, 2006.

[124] W. Koolen, M. Warmuth, and J. Kivinen, "Hedging structured concepts," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2010.

[125] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.

[126] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," in *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[127] P. Lezaud, "Chernoff-type bound for finite Markov chains," *Annals of Applied Probability*, vol. 8, pp. 849–867, 1998.

[128] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the International Conference on World Wide Web (WWW)*, 2010.

[129] K. Liu, Q. Zhao, and B. Krishnamachari, "Dynamic multichannel access with imperfect channel state detection," *IEEE Transactions on Signal Processing*, vol. 58, pp. 2795–2808, 2010.

[130] A. Mahajan and D. Teneketzis, "Multi-armed bandit problems," in *Foundations and Applications of Sensor Management*, pp. 121–151, Springer, 2008.

[131] O. Maillard and R. Munos, "Adaptive bandits: Towards the best history-dependent strategy," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR Workshop and Conference Proceedings Volume 15, 2011.

[132] O.-A. Maillard, R. Munos, and G. Stoltz, "A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, JMLR Workshop and Conference Proceedings Volume 19, 2011.

[133] S. Mannor and O. Shamir, "From bandits to experts: On the value of side-observations," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[134] S. Mannor and J. N. Tsitsiklis, "The sample complexity of exploration in the multi-armed bandit problem," *Journal of Machine Learning Research*, vol. 5, pp. 623–648, 2004.

[135] H. McMahan and A. Blum, "Online geometric optimization in the bandit setting against an adaptive adversary," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2004.

[136] H. B. McMahan and M. Streeter, "Tighter bounds for multi-armed bandits with expert advice," in *Proceedings of the International Conference on Learning Theory (COLT)*, 2009.

[137] A. Nemirovski, "Efficient methods for large-scale convex optimization problems," *Ekonomika i Matematicheskie Metody*, vol. 15, 1979. (In Russian).

[138] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, vol. 19, pp. 1574–1609, 2009.

[139] A. Nemirovski and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

[140] Y. Nesterov, "Random gradient-free minimization of convex functions," Core discussion papers, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2011.

[141] G. Neu, A. Gyorgy, C. Szepesvari, and A. Antos, "Online Markov decision processes under bandit feedback," in *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[142] R. Ortner, D. Ryabko, P. Auer, and R. Munos, "Regret bounds for restless Markov bandits," arXiv preprint arXiv:1209.2693, 2012.

[143] V. Perchet and P. Rigollet, "The multi-armed bandit problem with covariates," Arxiv preprint arXiv:1110.6084, 2011.

[144] D. Pucci de Farias and N. Megiddo, "Combining expert advice in reactive environments," *Journal of the ACM*, vol. 53, no. 5, pp. 762–799, 2006.

[145] A. Rakhlin, "Lecture notes on online learning," 2009.

[146] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematics Society*, vol. 58, pp. 527–535, 1952.

[147] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, pp. 400–407, 1951.

[148] P. Rusmevichientong and J. Tsitsiklis, "Linearly parameterized bandits," *Mathematics of Operations Research*, vol. 35, pp. 395–411, 2010.

[149] A. Salomon and J.-Y. Audibert, "Deviations of stochastic bandit regret," in *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2011.

[150] Y. Seldin, P. Auer, F. Laviolette, J. Shawe-Taylor, and R. Ortner, "PAC-Bayesian analysis of contextual bandits," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[151] S. Shalev-Shwartz, "Online learning: Theory, algorithms, and applications," PhD thesis, The Hebrew University of Jerusalem, 2007.

[152] A. Slivkins, "Contextual bandits with similarity information," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, JMLR Workshop and Conference Proceedings Volume 19, 2011.

[153] A. Slivkins and E. Upfal, "Adapting to a changing environment: The Brownian restless bandits," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2008.

[154] N. Srebro, K. Sridharan, and A. Tewari, "On the universality of online Mirror Descent," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[155] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[156] G. Stoltz, "Incomplete information and internal regret in prediction of individual sequences," PhD thesis, Université Paris-Sud, 2005.

[157] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[158] C. Szepesvári, "Algorithms for Reinforcement Learning," Morgan and Claypool, 2010.

[159] E. Takimoto and M. Warmuth, "Paths kernels and multiplicative updates," *Journal of Machine Learning Research*, vol. 4, pp. 773–818, 2003.
[160] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5588–5611, 2012.
[161] F. Teytaud and O. Teytaud, "Creating an upper-confidence-tree program for Havannah," in *Advances in Computer Games*, pp. 65–74, 2009.
[162] W. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Bulletin of the American Mathematics Society*, vol. 25, pp. 285–294, 1933.
[163] T. Uchiya, A. Nakamura, and M. Kudo, "Algorithms for adversarial bandit problems with multiple plays," in *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2010.
[164] A. Wald, *Sequential Analysis*. J. Wiley and Sons, 1947.
[165] C. C. Wang, S. R. Kulkarni, and H. V. Poor, "Arbitrary side observations in bandit problems," *Advances in Applied Mathematics*, vol. 34, no. 4, pp. 903–938, 2005.
[166] C. C. Wang, S. R. Kulkarni, and H. V. Poor, "Bandit problems with side observations," *IEEE Transactions on Automatic Control*, vol. 50, no. 3, pp. 338–355, 2005.
[167] Y. Wang, J.-Y. Audibert, and R. Munos, "Algorithms for infinitely many-armed bandits," in *Advances in Neural Information Processing Systems (NIPS)*, 2008.
[168] M. Warmuth, W. Koolen, and D. Helmbold, "Combining initial segments of lists," in *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2011.
[169] M. Warmuth and D. Kuzmin, "Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension," *Journal of Machine Learning Research*, vol. 9, pp. 2287–2320, 2008.
[170] C. A. Wilkens and B. Sivan, "Single-call mechanisms," in *ACM Conference on Electronic Commerce (EC)*, 2012.
[171] J. Y. Yu and S. Mannor, "Unimodal bandits," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.
[172] J. Y. Yu, S. Mannor, and N. Shimkin, "Markov decision processes with arbitrary reward processes," *Mathematics of Operations Research*, vol. 34, pp. 737–757, 2009.
[173] Y. Yue, J. Broder, R. Kleinberg, and T. Joachims, "The k-armed dueling bandits problem," in *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2009.
[174] Y. Yue and T. Joachims, "Beat the mean bandit," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.
[175] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient Ascent," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2003.