Online Learning Methods for Networking

Cem Tekin
University of California, Los Angeles
cmtkn@ucla.edu

Mingyan Liu
University of Michigan
mingyan@umich.edu

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Networking

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

C. Tekin and M. Liu. *Online Learning Methods for Networking*. Foundations and Trends® in Networking, vol. 8, no. 4, pp. 281–409, 2013.

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

ISBN: 978-1-60198-917-8

Foundations and Trends® in Networking
Volume 8, Issue 4, 2013

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Networking publishes survey and tutorial articles in the following topics:

- Modeling and analysis of:
  - Ad hoc wireless networks
  - Sensor networks
  - Optical networks
  - Local area networks
  - Satellite and hybrid networks
  - Cellular networks
  - Internet and web services
- Protocols and cross-layer design
- Network coding
- Energy-efficiency incentives/pricing/utility-based
- Games (co-operative or not)
- Security
- Scalability
- Topology
- Control/Graph-theoretic models
- Dynamics and asymptotic behavior of networks

Information for Librarians

Foundations and Trends® in Networking, 2013, Volume 8, 4 issues. ISSN paper version 1554-057X. ISSN online version 1554-0588. Also available as a combined paper and online subscription.

Foundations and Trends® in Networking
Vol. 8, No. 4 (2013) 281–409

DOI: 10.1561/1300000050

now
the essence of knowledge

Online Learning Methods for Networking

Cem Tekin
University of California, Los Angeles
cmtkn@ucla.edu

Mingyan Liu
University of Michigan
mingyan@umich.edu

Contents

Abstract

In this monograph we provide a tutorial on a family of sequential learning and decision problems known as the multi-armed bandit problems. We introduce a wide range of application scenarios for this learning framework, as well as its many different variants. The more detailed discussion is focused on the stochastic bandit problems, with rewards driven by either an IID or a Markov process, and when the environment consists of a single or multiple simultaneous users. We also present literature on the learning of MDPs, which captures coupling among the evolution of different options that a classical MAB problem does not.

C. Tekin and M. Liu. *Online Learning Methods for Networking*. Foundations and Trends® in Networking, vol. 8, no. 4, pp. 281–409, 2013.

DOI: 10.1561/1300000050.

1

Introduction

This monograph provides a tutorial on a family of sequential learning and decision problems of the following nature. Consider a system operating in discrete time, consisting of a single user (also referred to as a player or a learner) and N choices (also referred to as options). At each discrete time step, the user is to select (also referred to as play or activate) one of the choices and in return it obtains an award associated with that choice whose amount is unknown a priori. The user's objective is to maximize its total reward over a finite or infinite horizon. Since the rewards are unknown, the user's success depends on its play strategy (i.e., the sequence of moves it makes). The user is assumed to have perfect recall, and generally speaking the decision at a given time instance is a function of all its past decisions and past observations of the rewards.

In this context, a particular decision serves one or both of two purposes: *exploration* and *exploitation*. The former refers to making selections for the purpose of finding out the quality of an option, e.g., by choosing an option that has been rarely chosen; the latter refers to making selections for the purpose of obtaining large rewards, e.g., by choosing an option that has been observed in the past to generate large

rewards. A good decision process often involves carefully balancing exploration with exploitation, e.g., forgoing immediate rewards in order to find out the quality of an unknown option so as to obtain larger rewards in the future. This balancing act between exploration and exploitation is characteristic of this type of "learning-on-the-go" problem, where we have to instantaneously apply what we have learned so far even as we continue to learn.

Before getting into the technical details, it is always instructive to take a look at the range of practical problem scenarios to which the above abstraction applies. Below we introduce a number of interesting applications where this type of online learning formulations has been extensively used. Generally speaking, this type of work finds use in any scenario involving sequential decision with some type of resource constraint (otherwise one could simply sample all options at each time). Our primary focus is on networking applications, but to provide a broader view we also sample some applications from various other fields.

1.1 Applications

Opportunistic spectrum access In this application a user represents a radio transceiver that is capable of rapidly switching between operating frequencies and detecting/sensing the transmission quality of a channel. The user has access to a set of channels of time-varying and unknown conditions, as a result of random fading and/or other users' activities, and therefore must determine in a sequence of moves how to select which channel to switch to for use so as to maximize its long term reward (in the form of total transmission rate, etc.). This scenario often arises in a cognitive radio or software defined radio context [30], where the user is *secondary* to a *primary* user who holds the license for the channels; the primary user's activities contribute to the time-varying channel condition the secondary user sees and it is critical for the latter to make judicious channel switching decisions so as to take advantage of instantaneous channel availability. Specifically, at each time step, the user senses or probes a subset of the channels to find out their

condition, and is allowed to use the channels in a way consistent with their conditions. For instance, good channel conditions result in higher data rates or lower power for the user and so on. In some cases, channel conditions are simply characterized as being available and unavailable, and the user is allowed to use all channels sensed to be available.

Dynamic demand response in the emerging smart grid Electric loads participating in demand response programs provide a variety of benefits to electric power systems including increased power system reliability and power market efficiency [17]. In particular, when a demand aggregator deploy (activate) loads, an available load can potentially reduce its energy consumption in response to the curtailment signals. This is especially useful during peak loads in helping to shape the load on the grid. However, the demand aggregator is faced with a large number of loads, and their responses to curtailment signals can be highly uncertain [39] because load behavior is complex and influenced by a variety of stochastic factors including weather and human behavior. Generally, detailed load models are unavailable and we do not have full access to realtime information about load models, states, or disturbances due to limited communications. Subsequently, a load's ability to curtail is often only known after it has been told to curtail (i.e., deployed) and observed. For instance, the aggregator does not know a priori which load is available to be deployed (e.g., a refrigerator must be in working mode so it can be shut off), or how much savings might result from a particular load (e.g., this may be local temperature dependent). This results in the exploration and exploitation tradeoff illustrated earlier, i.e., pursuing potential gain from poorly characterized loads so as to improve our characterization which hopefully leads to future gains versus harvesting immediate benefits from well-characterized loads. A very interesting additional challenge here is that the aggregator generally does not get to observe the reward (response in this case) from individual loads, but only the aggregate.

Social networks and recommendation systems In this application an individual wishes to learn from others' recommendations to decide what

choices to make (e.g., restaurants, movies, parks, etc.) in a sequential fashion so as to maximize its own satisfaction over a period of time. This individual can obviously sample these choices all on its own which leads to a similar online learning problem as illustrated in the previous applications. An interesting question here is whether it could learn faster by taking others' recommendations into account; the associated challenge is that in many instances recommendations are reflections of subjective opinion or tastes, and thus one not only needs to learn the quality of the many choices on its own, but it must also learn how to judiciously weigh different recommendations so as to speed up her overall learning process.

Playing slot machine This is perhaps the original motivating (or illustrating) application of this type of learning problems. In this context a gambler plays a sequence of slot machines to maximize its gain over time. Each machine when played generates a reward/payoff that is unknown a priori, and the machines can potentially have different average payoff, which is also unknown to the gambler. The question then arises as to in what sequence should the slot machines be played as a function of the payoffs that the gambler has obtained in the past, so as to maximize its total payoff.

Clinical trial One of the earliest motivating applications of this learning framework is clinical trials [31]. The goal of a clinical trial is to find out the most effective drug (or dosage of a drug) from a set of given drugs. For this purpose, patients are sequentially treated by giving to each patient a drug from the set of available drugs and observing the outcome before deciding on which drug to give to the next patient. The sequential learning framework outlined earlier seeks to identify the most effective drug with a high confidence, and at the same time to also ensure that the number of patients who receive suboptimal drugs among the patients that participated in clinical trials is minimized.

Ad placement The revenues of most of the web search engines depend on advertisements. Usually the web search engines implement a

pay-per-click rule such that they obtain a fixed amount of payment whenever a user clicks on an advertisement. Therefore, it is of vital importance to learn which advertisements to display based on the search query of the user to maximize the number of clicks. Under this learning framework, this task can be accomplished by identifying the ads with the highest click-through rate (CTR), without losing too much revenue by exploring suboptimal ads in an effort to find out the best ad to show for a specific search query.

1.2 The Multi-Armed Bandit Problem Formalization

The sequential learning and decision problems arising from the array of applications discussed above are often also referred to as the family of *bandit* problems, where the set of options are collectively referred to as a multi-armed bandit (MAB), each option an *arm*. It should be noted that there is also an optimization version of the bandit problem first analyzed by Gittins [22], where the rewards are given by Markov chains whose statistics are perfectly known a priori. Therefore the problem is one of *optimization* rather than learning the unknowns to optimize: the goal is to determine offline an optimal policy of playing the arms so as to maximize the discounted reward over an infinite horizon. This was also referred to as the *deterministic* bandit problem by [31].

1.2.1 Learning algorithms and performance measures

The performance of a particular learning algorithm is typically measured by the notion of *regret*. This is defined as the difference between the expected reward that can be gained by an "infeasible" or ideal policy, i.e., a policy that requires either a priori knowledge of some or all statistics of the arms or hindsight information, and the expected reward of the user's algorithm.

The most commonly used infeasible policy is the *best single-action* policy, that is the best among all policies that continue to play the same arm. An ideal policy could play, for instance, the arm that has the highest expected reward (which requires statistical information but not hindsight). This type of regret is sometimes also referred to as the

weak regret. A stronger performance measure is accordingly and aptly referred to as the *strong regret*, where the infeasible policy is the best dynamic policy one may construct with a priori full information on the statistics of the underlying reward processes.

The regret of a learning algorithm is often expressed using asymptotic notation. In the literature there exists two types of regret bounds: a bound on the regret that is only proven asymptotically as the time horizon T goes to infinity and a bound on regret that holds uniformly over time. For instance, a bound of the form $\limsup_{T\to\infty} R(T)/\log T \leq C$, for some $C > 0$ is an asymptotic bound, while a bound of the form $R(T) \leq C \log T$ for all $T > 0$ for some $C > 0$ is a uniform bound. For the purpose of brevity, we use the standard Big-O notation $O(\cdot)$ for both types of bounds. For instance, both of the regret bounds illustrated above can be represented as $R(T) = O(\log T)$. However, we clearly make the distinction between a bound that holds uniformly and a bound that holds asymptotically when necessary. We also use $\tilde{O}(\cdot)$, which hides the constants in the bound similar to the standard Big-O notation, but also hides additional terms of logarithmic growth. For instance $R(T) = \tilde{O}(\sqrt{T})$ means that $R(T) = O(\sqrt{T}\log T)$.

1.2.2 Classification of multi-armed bandit problems

There are many variants of this basic version of the MAB problem, which we detail below.

Non-stochastic vs. stochastic When the rewards are driven by unknown and possibly arbitrary processes, the problem is referred to as *non-stochastic*. This is typically used to capture an adversarial setting, i.e., with the rewards being generated by an adversary playing against the user. In this case no probabilistic assumptions are made on the reward processes. By contrast, in the *stochastic* learning problem, the rewards are assumed to be driven by well-defined, though unknown, stochastic processes. Furthermore, they are often assumed to be of a certain structure with unknown components/parameters, e.g., finite-state Markov processes with unknown transition probability matrices, or IID processes with unknown distribution but finite support.

Rested vs. restless In the case of Markov reward processes (of which IID is a special case), we further make the distinction between a process whose state only evolves upon activation and one that continues to evolve regardless of the user's actions. The former is referred to as a rested bandit, the latter a restless bandit. The user is often assumed to only observe the state (or reward) of the arm it chooses to play. This leads to the following crucial difference. In the rested case, since a process' state remains frozen until it is played again, the user in effect has complete information on the states of all the arms. In the restless case, the user is facing a problem of *incomplete information* as it does not know the states of those arms except for the one it currently uses.

This distinction is inconsequential when the reward processes are IID, because knowing the current state does not alter the user's prediction of the next state when the rewards are given by an IID process; it however introduces substantial technical differences when the processes are Markov as we shall soon see.

Controlled vs. uncontrolled In the case of restless bandits, an arm is called uncontrolled if its state evolution upon activation follows the same probabilistic law as when it is not played. By contrast, a controlled arm follows different laws depending on the user's action. For instance, a controlled Markov reward process may be governed by two transition probability matrices, one for when the arm is not played and the other when the arm is played. This distinction is less important if one is only interested in weak regret because the reference, best single-action policy continuously plays a single arm, effectively invoking only one of the transition probability matrices. This distinction however makes a crucial difference when considering strong regret where a dynamic policy must take into consideration both matrices.

Other structural variations The user may not be restricted in selecting only one arm at a time. The problem is referred to as with *multiple plays* if the user can select up to M options at a time.

There is also an important *decentralized* version of the bandit problem, whereby multiple uncoordinated players each makes its own deci-

sion on which arm to play in each step. An interesting twist here is that if two or more players select the same arm simultaneously, a collision results and the reward they get is discounted, e.g., none of the players selecting that arm gets a reward, or the reward goes to only one of the players selecting that arm, or they each gets a share of the reward.

1.3 Organization

The remainder of this monograph is organized as follows. We start by discussing the simplest case, i.e., the single user IID MAB problem in Chapter 2. We review several learning algorithms that achieve the optimal tradeoff between exploration and exploitation. Single user Markov MAB problems, algorithms and regret analysis are discussed in Chapter 3. In Chapter 4 we review a different strand of tools that have been developed and used for online learning in Markov Decision Processes (MDP); we shall see that these tools bear striking resemblance to those developed for MAB problems. Chapter 5 is devoted to the study of multi-user MAB problems involving decentralized decision makers. Chapter 6 summarizes key intuitions and principles underlying the design and analysis of the algorithms, and concludes the monograph.

References

[1] S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. *arXiv preprint arXiv:1111.1797*, 2011.

[2] A. Anandkumar, N. Michael, A.K. Tang, and A. Swami. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications (JSAC)*, 29(4):731–745, 2011.

[3] V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: IID rewards. *IEEE Transactions on Automatic Control*, pages 968–975, November 1987.

[4] O. Atan, C. Tekin, and M. van der Schaar. Global bandits with Hölder continuity. *arXiv preprint arXiv:1410.7890*, 2014.

[5] J.Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proc. of the Annual Conference on Learning Theory (COLT)*, 2009.

[6] J.Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research*, 11:2785–2836, 2010.

[7] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 3:397–422, 2003.

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

[9] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proc. of the 36th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 322–331. IEEE, 1995.

[10] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2003.

[11] P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 89–96, 2009.

[12] J. Bartroff, T.L. Lai, and M.C. Shih. *Sequential Experimentation in Clinical Trials: Design and Analysis*, volume 298. Springer, 2012.

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

[14] N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

[15] Yiwei Chen and Vivek F Farias. Simple policies for dynamic pricing with imperfect forecasts. *Operations Research*, 61(3):612–624, 2013.

[16] E. Chlebus. An approximate formula for a partial sum of the divergent p-series. *Applied Mathematics Letters*, 22:732–737, 2009.

[17] DOE. Benefits of demand response in electricity markets and recommendations for achieving them. Technical report, Department of Energy Report to the US Congress, 2006.

[18] Y. Gai, B. Krishnamachari, and R. Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking (TON)*, 20(5):1466–1478, 2012.

[19] Y. Gai, B. Krishnamachari, and M. Liu. Online learning for combinatorial network optimization with restless Markovian rewards. In *Proc. of the 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, pages 28–36, 2012.

[20] A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proc. of the 24th Annual Conference on Learning Theory (COLT)*, 2011.

[21] S. Geirhofer, L. Tong, and B.M. Sadler. A measurement-based model for dynamic spectrum access in WLAN channels. In *Proc. of the IEEE Military Communications Conference (MILCOM)*, pages 1–7, 2006.

[22] J.C. Gittins and D.M. Jones. A dynamic allocation index for sequential design of experiments. *Progress in Statistics, Euro. Meet. Statis.*, 1:241–266, 1972.

[23] E. Hazan and S. Kale. Better algorithms for benign bandits. *The Journal of Machine Learning Research*, 12:1287–1311, 2011.

[24] W. Hsu, R.K. Taira, S. El-Saden, H. Kangarloo, and A.T. Bui. Context-based electronic health record: toward patient specific healthcare. *IEEE Transactions on Information Technology in Biomedicine*, 16(2):228–234, 2012.

[25] Senhua Huang, Xin Liu, and Zhi Ding. Opportunistic spectrum access in cognitive radio networks. In *Proc. of the 27th IEEE INFOCOM*, 2008.

[26] Yu-Gang Jiang, Subhabrata Bhattacharya, Shih-Fu Chang, and Mubarak Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2):73–101, 2013.

[27] L.P. Kaelbling, M.L. Littman, and A.W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, pages 237–285, 1996.

[28] E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *International Conference on Artificial Intelligence and Statistics*, pages 592–600, 2012.

[29] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pages 199–213. Springer, 2012.

[30] J. Kennedy and M. Sullivan. Direction Finding and Smart Antennas Using Software Radio Architectures. *IEEE Communications Magazine*, pages 62–68, May 1995.

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

[32] P. Lezaud. Chernoff-type bound for finite Markov chains. *Annals of Applied Probability*, pages 849–867, 1998.

[33] L. Li, W. Chu, J. Langford, and R.E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proc. of the 19th International Conference on World Wide Web*, pages 661–670, 2010.

[34] H. Liu, K. Liu, and Q. Zhao. Learning in a changing world: Non-Bayesian restless multi-armed bandit. *Techinal Report, UC Davis*, October 2010.

[35] H. Liu, K. Liu, and Q. Zhao. Learning and sharing in a changing world: Non-Bayesian restless bandit with multiple players. In *Information Theory and Applications Workshop (ITA)*, January 2011.

[36] Haoyang Liu, Keqin Liu, and Qing Zhao. Learning in a changing world: Restless multiarmed bandit with unknown dynamics. *IEEE Transactions on Information Theory*, 59(3):1902–1916, 2013.

[37] K. Liu and Q. Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, 58(11):5667–5681, 2010.

[38] O.L. Mangasarian, W.N. Street, and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, 1995.

[39] J.L. Mathieu, D.S. Callaway, and S. Kiliccote. Variability in automated responses of commercial buildings and industrial facilities to dynamic electricity prices. *Energy and Buildings*, 43:3322–3330, 2011.

[40] A.J. Mersereau, P. Rusmevichientong, and J.N. Tsitsiklis. A structured multiarmed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, 54(12):2787–2802, 2009.

[41] R. Ortner, D. Ryabko, P. Auer, and R. Munos. Regret bounds for restless Markov bandits. In *Algorithmic Learning Theory*, pages 214–228. Springer, 2012.

[42] C. Papadimitriou and J. Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, May 1999.

[43] M.L. Pinedo. *Scheduling: theory, algorithms, and systems*. Springer, 2012.

[44] M.L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*, volume 414. John Wiley & Sons, 2009.

[45] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proc. of the 25th International Conference on Machine Learning (ICML)*, pages 784–791. ACM, 2008.

[46] C. Stauffer and W.E. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[47] C. Tekin and M. Liu. Online algorithms for the multi-armed bandit problem with Markovian rewards. In *Proc. of the 48th Annual Allerton Conference on Communication, Control, and Computing*, pages 1675–1682, 2010.

[48] C. Tekin and M. Liu. Adaptive learning of uncontrolled restless bandits with logarithmic regret. In *Proc. of the 49th Annual Allerton Conference on Communication, Control, and Computing*, pages 983–990, September 2011.

[49] C. Tekin and M. Liu. Online learning in opportunistic spectrum access: A restless bandit approach. In *Proc. of the 30th Annual IEEE International Conference on Computer Communications (INFOCOM)*, pages 2462 – 2470, April 2011.

[50] C. Tekin and M. Liu. Performance and convergence of multi-user online learning. In *Proc. of the 2nd International Conference on Game Theory for Networks (GAMENETS)*, April 2011.

[51] C. Tekin and M. Liu. Online learning in decentralized multi-user spectrum access with synchronized explorations. In *Proc. of IEEE MILCOM*, 2012.

[52] C. Tekin and M. Liu. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611, 2012.

[53] C. Tekin, M. Liu, R. Southwell, J. Huang, and S.H.A. Ahmad. Atomic congestion games on graphs and their applications in networking. *IEEE/ACM Transactions on Networking*, 20(5):1541 –1552, October 2012.

[54] C. Tekin and M. van der Schaar. An experts learning approach to mobile service offloading. In *Proc. of the 52nd Annual Allerton Conference on Communication, Control, and Computing*, 2014.

[55] Cem Tekin and Mingyan Liu. Optimal adaptive learning in uncontrolled restless bandit problems. *arXiv preprint arXiv:1107.4042*, 2011.

[56] A. Tewari and P.L. Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. In *Proc. of the Conference on Advances in Neural Information Processing Systems (NIPS)*, volume 20, pages 1505–1512, 2008.

[57] S. Vakili, K. Liu, and Q. Zhao. Deterministic sequencing of exploration and exploitation for multi-armed bandit problems. *IEEE Journal on Selected Topics in Signal Processing (JSTSP)*, 7(5):759–767, October 2013.

[58] Hong Shen Wang and Pao-Chi Chang. On verifying the first-order markovian assumption for a rayleigh fading channel model. *Vehicular Technology, IEEE Transactions on*, 45(2):353–357, 1996.

[59] J.Y. Yu, S. Mannor, and N. Shimkin. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

[60] Q. Zhao and B. Sadler. A survey of dynamic spectrum access. *IEEE Signal Processing Magazine: Special Issue on Resource-Constrained Signal Processing, Communications, and Networking*, 24:79–89, May 2007.