Book Announcement: "Introduction to Multi-Armed Bandits"

ALEKSANDRS SLIVKINS
Microsoft Research NYC

"Introduction to multi-armed bandits" is a broad and accessible textbook which emphasizes connections to economics and operations research.

Categories and Subject Descriptors: F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems; F.1.2 [**Computation by Abstract Devices**]: Modes of Computation—*Online computation*; J.4 [**Social and Behavioral Sciences**]: Economics

General Terms: Algorithms, Economics, Theory

I am pleased to announce "Introduction to multi-armed bandits" [Slivkins 2019], a broad and accessible textbook recently published in the "Foundations and Trends in Machine Learning" series. In addition to the machine learning issues, the book covers connections to economics and operations research. The book is geared towards a wide audience interested in machine learning, and may be of particular interest to the "economics and computation" community.

Multi-armed bandits is a simple but very powerful framework for algorithms that make decisions over time under uncertainty. In the basic version, an algorithm repeatedly chooses from a fixed set of actions (a.k.a., *arms*), and receives a reward for the chosen action. The reward comes from a fixed, action-specific distribution that is not known to the algorithm. Crucially, no "counterfactual" feedback is received: the algorithm does not know what would have happened had a different action been chosen. Hence, we have a tension between *exploring* different arms in order to acquire information and *exploiting* this information to maximize rewards, a.k.a. *exploration-exploitation tradeoff*.

This tradeoff is essential in many application scenarios. The term "multi-armed bandits" comes from a fictitious gambling scenario with multiple slot machines, a.k.a. one-armed bandits, which look identical to the gambler but may have different payout distributions. The original motivation comes from the design of medical trials: how to quickly phase out under-performing treatments without compromising statistical validity? Modern applications include web search (what are the best result for a given search query), content optimization (*e.g.,* what are the best news articles for a given user), online advertisement (what are the best ads to display for a given ad opportunity), and many others.

Aleksandrs Slivkins. "Introduction to multi-armed bandits".
Foundations and Trends in Machine Learning, vol. 12, pp. 1-286, November 2019.
At *Now Publishers*: `https://www.nowpublishers.com/article/Details/MAL-068`.
On Arxiv (in plain format): `https://arxiv.org/abs/1904.07272`.

Author's address: `slivkins@microsoft.com`,
`https://www.microsoft.com/en-us/research/people/slivkins/`.

I am particularly interested in how exploration-exploitation tradeoff plays out in "game-theoretic" environments, when the algorithm interacts with self-interested behavior. In retail, how to optimize prices and product assortments? In online advertisement, how to optimize parameters of an ad auction and how to allocate ads so as not to skew the advertisers' bidding behavior? In online labor markets, how to control quality and match workers to tasks, and which prices or contracts to offer? For online platforms that collect ratings and reviews, how to incentivize users to explore when they usually prefer to exploit? If two bandit algorithms play a repeated game against each other, does this game converge to an equilibrium, in which sense, and how fast? What if online platforms (*e.g.,* search engines) learn from interactions with users *and* compete with one another for the said users?

An enormous, multi-dimensional body of work has accumulated since the 1950-ies, with a big surge of interest in the past two decades. While the basic model described above captures something fundamental about the application scenarios, it can be extended – and made more realistic – in many different directions. To wit, what if the reward distributions can change over time? What if there is a huge number of actions, with some structure that can help us navigate the action space? What if more feedback is available to the algorithm, before and/or after each action is chosen? What if the algorithm consumes resources and operates under budget constraints thereon? What if the algorithm interacts with self-interested parties and needs to be compatible with their incentives? These questions, and their numerous refinements, prompted many distinct lines of work.

How to present all this work, let alone make it accessible? My approach is to present a broad picture, favoring fundamental ideas and elementary, teachable proofs over the strongest possible results. Each chapter handles one big direction in the problem space, covers the first-order concepts and results on a technical level, and provides a detailed literature review for further exploration. Most chapters conclude with exercises (which often introduce pertinent results that do not quite fit into the main technical narrative.) The book is teachable by design: each chapter corresponds to one week of my class. While some exposure to probability and statistics would help, a standard undergraduate course on algorithms should suffice for background.

I believe that multi-armed bandits are both deeply theoretical and deeply practical. Aside from all the math, I try to be careful and explicit about motivation. Our models might not capture the full complexity of application scenarios, and that's OK. Instead, the point is to capture some essential features present in many motivating examples. Practical aspects are discussed in considerable detail, based on a system for contextual bandits developed at Microsoft Research.[1]

For the "economics and computation" audience, the value proposition is as follows. Two chapters in my book (out of 11 total) cover connections to game theory and mechanism design, two more are directly motivated by dynamic pricing and similar problems, and the remaining chapters provide the necessary background. Economic aspects of multi-armed bandits is an increasingly popular subject, with ACM EC being the "conference home" for much of it.

[1] `https://www.microsoft.com/en-us/research/project/real-world-reinforcement-learning`. Also, see `https://www.microsoft.com/en-us/research/project/multi-world-testing-mwt`.

On a more personal note, this was some project. Once I sent my lecture notes to the publisher for the initial review, I thought it would take me a month or two to finish the book. As it happened, it took me *two and a half years*, with many interruptions in between ... I am planning a minor revision — some time after the pandemic is over, my son is back to school, and his parents are back to sanity — so, please send me bug reports and any other feedback.

P.S. Several books, published over the years, provide an in-depth treatment of various specific aspects of multi-armed bandits [Berry and Fristedt 1985; Cesa-Bianchi and Lugosi 2006; Gittins et al. 2011; Bubeck and Cesa-Bianchi 2012; Hazan 2015; Russo et al. 2018]. My book provides a more uniform, textbook-like treatment of the subject.

An upcoming book on bandits by Lattimore and Szepesvári [2020, preprint] provides a deeper treatment for a number of topics, and omits a few others. Evolving simultaneously and independently over the past 3 years, their book and mine reflect the authors' somewhat differing tastes and presentation styles, and, I believe, are complementary to one another.

REFERENCES

Berry, D. A. and Fristedt, B. 1985. *Bandit problems: sequential allocation of experiments*. Springer, Heidelberg, Germany.

Bubeck, S. and Cesa-Bianchi, N. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning 5,* 1, 1–122. Published with *Now Publishers* (Boston, MA, USA). Also available at `https://arxiv.org/abs/1204.5721`.

Cesa-Bianchi, N. and Lugosi, G. 2006. *Prediction, learning, and games*. Cambridge University Press, Cambridge, UK.

Gittins, J., Glazebrook, K., and Weber, R. 2011. *Multi-Armed Bandit Allocation Indices*, 2nd ed. John Wiley & Sons, Hoboken, NJ, USA. The first edition, single-authored by John Gittins, has been published in 1989.

Hazan, E. 2015. Introduction to Online Convex Optimization. *Foundations and Trends® in Optimization 2,* 3-4, 157–325. Published with *Now Publishers* (Boston, MA, USA). Also available at https://arxiv.org/abs/1909.05207.

Lattimore, T. and Szepesvári, C. 2020. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK. Preprint, to be published in 2020. Versions available at `https://banditalgs.com/` since 2018.

Russo, D., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. 2018. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning 11,* 1, 1–96. Published with *Now Publishers* (Boston, MA, USA). Also available at `https://arxiv.org/abs/1707.02038`.

Slivkins, A. 2019. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning 12,* 1-2 (Nov.), 1–286. Published with *Now Publishers* (Boston, MA, USA). Also available at `https://arxiv.org/abs/1904.07272`.