A Tutorial on Thompson Sampling

Other titles in Foundations and Trends® in Machine Learning

Non-convex Optimization for Machine Learningy
Prateek Jain and Purushottam Ka
ISBN: 978-1-68083-368-3

Kernel Mean Embedding of Distributions: A Review and Beyond
Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur and Bernhard Scholkopf
ISBN: 978-1-68083-288-4

Tensor Networks for Dimensionality Reduction and Large-scale Optimization: Part 1 Low-Rank Tensor Decompositions
Andrzej Cichocki, Anh-Huy Phan, Qibin Zhao, Namgil Lee, Ivan Oseledets, Masashi Sugiyama and Danilo P. Mandic
ISBN: 978-1-68083-222-8

Tensor Networks for Dimensionality Reduction and Large-scale Optimization: Part 2 Applications and Future Perspectives
Andrzej Cichocki, Anh-Huy Phan, Qibin Zhao, Namgil Lee, Ivan Oseledets, Masashi Sugiyama and Danilo P. Mandic
ISBN: 978-1-68083-276-1

Patterns of Scalable Bayesian Inference
Elaine Angelino, Matthew James Johnson and Ryan P. Adams
ISBN: 978-1-68083-218-1

Generalized Low Rank Models
Madeleine Udell, Corinne Horn, Reza Zadeh and Stephen Boyd
ISBN: 978-1-68083-140-5

A Tutorial on Thompson Sampling

Daniel J. Russo
Columbia University

Benjamin Van Roy
Stanford University

Abbas Kazerouni
Stanford University

Ian Osband
Google DeepMind

Zheng Wen
Adobe Research

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

D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband and Z. Wen. *A Tutorial on Thompson Sampling*. Foundations and Trends® in Machine Learning, vol. 11, no. 1, pp. 1–96, 2018.

ISBN: 978-1-68083-471-0

Foundations and Trends® in Machine Learning

Volume 11, Issue 1, 2018

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes 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, 2018, Volume 11, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Contents

A Tutorial on Thompson Sampling

Daniel J. Russo[1], Benjamin Van Roy[2], Abbas Kazerouni[2], Ian Osband[3] and Zheng Wen[4]

[1] *Columbia University*
[2] *Stanford University*
[3] *Google DeepMind*
[4] *Adobe Research*

ABSTRACT

Thompson sampling is an algorithm for online decision problems where actions are taken sequentially in a manner that must balance between exploiting what is known to maximize immediate performance and investing to accumulate new information that may improve future performance. The algorithm addresses a broad range of problems in a computationally efficient manner and is therefore enjoying wide use. This tutorial covers the algorithm and its application, illustrating concepts through a range of examples, including Bernoulli bandit problems, shortest path problems, product recommendation, assortment, active learning with neural networks, and reinforcement learning in Markov decision processes. Most of these problems involve complex information structures, where information revealed by taking an action informs beliefs about other actions. We will also discuss when and why Thompson sampling is or is not effective and relations to alternative algorithms.

Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband and Zheng Wen (2018), "A Tutorial on Thompson Sampling", Foundations and Trends® in Machine Learning: Vol. 11, No. 1, pp 1–96. DOI: 10.1561/2200000070.

In memory of Arthur F. Veinott, Jr.

1

Introduction

The multi-armed bandit problem has been the subject of decades of intense study in statistics, operations research, electrical engineering, computer science, and economics. A "one-armed bandit" is a somewhat antiquated term for a slot machine, which tends to "rob" players of their money. The colorful name for our problem comes from a motivating story in which a gambler enters a casino and sits down at a slot machine with multiple levers, or arms, that can be pulled. When pulled, an arm produces a random payout drawn independently of the past. Because the distribution of payouts corresponding to each arm is not listed, the player can learn it only by experimenting. As the gambler learns about the arms' payouts, she faces a dilemma: in the immediate future she expects to earn more by *exploiting* arms that yielded high payouts in the past, but by continuing to *explore* alternative arms she may learn how to earn higher payouts in the future. Can she develop a sequential strategy for pulling arms that balances this tradeoff and maximizes the cumulative payout earned? The following Bernoulli bandit problem is a canonical example.

Example 1.1. (*Bernoulli Bandit*) Suppose there are K actions, and when played, any action yields either a success or a failure. Action

$k \in \{1, ..., K\}$ produces a success with probability $\theta_k \in [0, 1]$. The success probabilities $(\theta_1, .., \theta_K)$ are unknown to the agent, but are fixed over time, and therefore can be learned by experimentation. The objective, roughly speaking, is to maximize the cumulative number of successes over T periods, where T is relatively large compared to the number of arms K.

The "arms" in this problem might represent different banner ads that can be displayed on a website. Users arriving at the site are shown versions of the website with different banner ads. A success is associated either with a click on the ad, or with a conversion (a sale of the item being advertised). The parameters θ_k represent either the click-through-rate or conversion-rate among the population of users who frequent the site. The website hopes to balance exploration and exploitation in order to maximize the total number of successes.

A naive approach to this problem involves allocating some fixed fraction of time periods to exploration and in each such period sampling an arm uniformly at random, while aiming to select successful actions in other time periods. We will observe that such an approach can be quite wasteful even for the simple Bernoulli bandit problem described above and can fail completely for more complicated problems.

Problems like the Bernoulli bandit described above have been studied in the decision sciences since the second world war, as they crystallize the fundamental trade-off between exploration and exploitation in sequential decision making. But the information revolution has created significant new opportunities and challenges, which have spurred a particularly intense interest in this problem in recent years. To understand this, let us contrast the Internet advertising example given above with the problem of choosing a banner ad to display on a highway. A physical banner ad might be changed only once every few months, and once posted will be seen by every individual who drives on the road. There is value to experimentation, but data is limited, and the cost of of trying a potentially ineffective ad is enormous. Online, a different banner ad can be shown to each individual out of a large pool of users, and data from each such interaction is stored. Small-scale experiments are now a core tool at most leading Internet companies.

Our interest in this problem is motivated by this broad phenomenon. Machine learning is increasingly used to make rapid data-driven decisions. While standard algorithms in supervised machine learning learn passively from historical data, these systems often drive the generation of their own training data through interacting with users. An online recommendation system, for example, uses historical data to optimize current recommendations, but the outcomes of these recommendations are then fed back into the system and used to improve future recommendations. As a result, there is enormous potential benefit in the design of algorithms that not only learn from past data, but also explore systemically to generate useful data that improves future performance. There are significant challenges in extending algorithms designed to address Example 1.1 to treat more realistic and complicated decision problems. To understand some of these challenges, consider the problem of learning by experimentation to solve a shortest path problem.

Example 1.2. (Online Shortest Path) An agent commutes from home to work every morning. She would like to commute along the path that requires the least average travel time, but she is uncertain of the travel time along different routes. How can she learn efficiently and minimize the total travel time over a large number of trips?

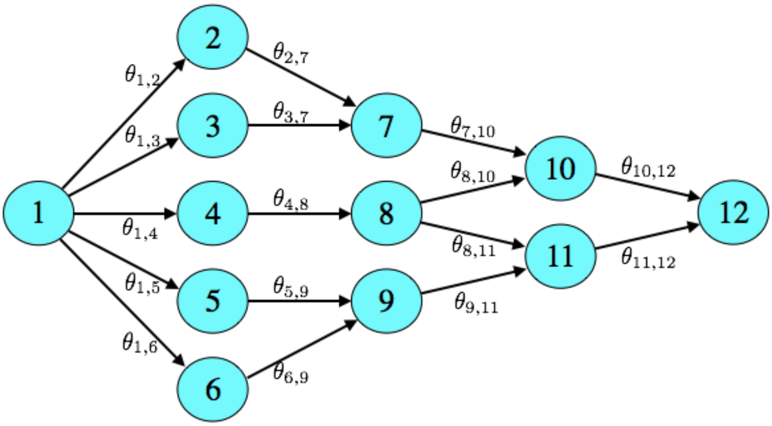

Figure 1.1: Shortest path problem.

We can formalize this as a shortest path problem on a graph $G = (V, E)$ with vertices $V = \{1, ..., N\}$ and edges E. An example is illustrated in Figure 1.1. Vertex 1 is the source (home) and vertex N is the destination (work). Each vertex can be thought of as an intersection, and for two vertices $i, j \in V$, an edge $(i, j) \in E$ is present if there is a direct road connecting the two intersections. Suppose that traveling along an edge $e \in E$ requires time θ_e on average. If these parameters were known, the agent would select a path $(e_1, .., e_n)$, consisting of a sequence of adjacent edges connecting vertices 1 and N, such that the expected total time $\theta_{e_1} + ... + \theta_{e_n}$ is minimized. Instead, she chooses paths in a sequence of periods. In period t, the realized time $y_{t,e}$ to traverse edge e is drawn independently from a distribution with mean θ_e. The agent sequentially chooses a path x_t, observes the realized travel time $(y_{t,e})_{e \in x_t}$ along each edge in the path, and incurs cost $c_t = \sum_{e \in x_t} y_{t,e}$ equal to the total travel time. By exploring intelligently, she hopes to minimize cumulative travel time $\sum_{t=1}^{T} c_t$ over a large number of periods T.

This problem is conceptually similar to the Bernoulli bandit in Example 1.1, but here the number of actions is the number of paths in the graph, which generally scales exponentially in the number of edges. This raises substantial challenges. For moderate sized graphs, trying each possible path would require a prohibitive number of samples, and algorithms that require enumerating and searching through the set of all paths to reach a decision will be computationally intractable. An efficient approach therefore needs to leverage the statistical and computational structure of problem.

In this model, the agent observes the travel time along each edge traversed in a given period. Other feedback models are also natural: the agent might start a timer as she leaves home and checks it once she arrives, effectively only tracking the total travel time of the chosen path. This is closer to the Bernoulli bandit model, where only the realized reward (or cost) of the chosen arm was observed. We have also taken the random edge-delays $y_{t,e}$ to be independent, conditioned on θ_e. A more realistic model might treat these as correlated random variables, reflecting that neighboring roads are likely to be congested at the same time. Rather than design a specialized algorithm for each possible statistical model, we seek a general approach to exploration that accommodates flexible modeling and works for a broad array of problems. We will see that Thompson sampling accommodates such flexible modeling, and offers an elegant and efficient approach to exploration in a wide range of structured decision problems, including the shortest path problem described here.

Thompson sampling – also known as *posterior sampling* and *probability matching* – was first proposed in 1933 (Thompson, 1933; Thompson, 1935) for allocating experimental effort in two-armed bandit problems arising in clinical trials. The algorithm was largely ignored in the academic literature until recently, although it was independently rediscovered several times in the interim (Wyatt, 1997; Strens, 2000) as an effective heuristic. Now, more than eight decades after it was introduced, Thompson sampling has seen a surge of interest among industry practitioners and academics. This was spurred partly by two influential articles that displayed the algorithm's strong empirical performance (Chapelle and Li, 2011; Scott, 2010). In the subsequent five years, the literature on Thompson sampling has grown rapidly. Adaptations of Thompson sampling have now been successfully applied in a wide variety of domains, including revenue management (Ferreira *et al.*, 2015), marketing (Schwartz *et al.*, 2017), web site optimization (Hill *et al.*, 2017), Monte Carlo tree search (Bai *et al.*, 2013), A/B testing (Graepel *et al.*, 2010), Internet advertising (Graepel *et al.*, 2010; Agarwal, 2013; Agarwal *et al.*, 2014), recommendation systems (Kawale *et al.*, 2015), hyperparameter tuning (Kandasamy *et al.*, 2018), and arcade games (Osband *et al.*, 2016a); and have been used at several companies, including Adobe, Amazon (Hill *et al.*, 2017), Facebook, Google (Scott, 2010; Scott, 2015), LinkedIn (Agarwal, 2013; Agarwal *et al.*, 2014), Microsoft (Graepel *et al.*, 2010), Netflix, and Twitter.

The objective of this tutorial is to explain when, why, and how to apply Thompson sampling. A range of examples are used to demonstrate how the algorithm can be used to solve a variety of problems and provide clear insight into why it works and when it offers substantial benefit over naive alternatives. The tutorial also provides guidance on approximations to Thompson sampling that can simplify computation as well as practical considerations like prior distribution specification, safety constraints and nonstationarity. Accompanying this tutorial we also release a Python package[1] that reproduces all experiments and figures presented. This resource is valuable not only for reproducible research, but also as a reference implementation that may help practioners build intuition for how to practically implement some of the ideas and algorithms we discuss in this tutorial. A concluding section discusses theoretical results that aim to develop an understanding of why Thompson sampling works, highlights settings where Thompson sampling performs poorly, and discusses alternative approaches studied in recent literature. As a baseline and backdrop for our discussion of Thompson sampling, we begin with an alternative approach that does not actively explore.

[1]Python code and documentation is available at https://github.com/iosband/ts_tutorial.

References

Abbasi-Yadkori, Y., D. Pál, and C. Szepesvári. 2011. "Improved algorithms for linear stochastic bandits". In: *Advances in Neural Information Processing Systems 24*. 2312–2320.

Abeille, M. and A. Lazaric. 2017. "Linear Thompson sampling revisited". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. 176–184.

Agarwal, D. 2013. "Computational advertising: the LinkedIn way". In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. ACM. 1585–1586.

Agarwal, D., B. Long, J. Traupman, D. Xin, and L. Zhang. 2014. "Laser: a scalable response prediction platform for online advertising". In: *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM. 173–182.

Agrawal, S., V. Avadhanula, V. Goyal, and A. Zeevi. 2017. "Thompson sampling for the MNL-bandit". In: *Proceedings of the 30th Annual Conference on Learning Theory*. 76–78.

Agrawal, S. and N. Goyal. 2012. "Analysis of Thompson sampling for the multi-armed bandit problem". In: *Proceedings of the 25th Annual Conference on Learning Theory*. 39.1–39.26.

Agrawal, S. and N. Goyal. 2013a. "Further optimal regret bounds for Thompson sampling". In: *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*. 99–107.

Agrawal, S. and N. Goyal. 2013b. “Thompson sampling for contextual bandits with linear payoffs”. In: *Proceedings of The 30th International Conference on Machine Learning*. 127–135.

Auer, P., N. Cesa-Bianchi, and P. Fischer. 2002. “Finite-time analysis of the multiarmed bandit problem”. *Machine Learning*. 47(2): 235–256.

Bai, A., F. Wu, and X. Chen. 2013. “Bayesian mixture modelling and inference based Thompson sampling in Monte-Carlo tree search”. In: *Advances in Neural Information Processing Systems 26*. 1646–1654.

Bastani, H., M. Bayati, and K. Khosravi. 2018. “Exploiting the natural exploration in contextual bandits”. *arXiv preprint arXiv:1704.09011*.

Besbes, O., Y. Gur, and A. Zeevi. 2014. “Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards”. In: *Advances in Neural Information Processing Systems 27*. 199–207.

Bubeck, S., R. Munos, G. Stoltz, and C. Szepesvári. 2011. “X-armed bandits”. *Journal of Machine Learning Research*. 12: 1655–1695.

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

Bubeck, S. and R. Eldan. 2016. “Multi-scale exploration of convex functions and bandit convex optimization”. In: *Proccedings of 29th Annual Conference on Learning Theory*. 583–589.

Bubeck, S., R. Eldan, and J. Lehec. 2018. “Sampling from a log-concave distribution with projected Langevin Monte Carlo”. *Discrete & Computational Geometry*.

Cappé, O., A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. 2013. “Kullback-Leibler upper confidence bounds for optimal sequential allocation”. *Annals of Statistics*. 41(3): 1516–1541.

Casella, G. and E. I. George. 1992. “Explaining the Gibbs sampler”. *The American Statistician*. 46(3): 167–174.

Chapelle, O. and L. Li. 2011. “An empirical evaluation of Thompson sampling”. In: *Advances in Neural Information Processing Systems 24*. 2249–2257.

Cheng, X. and P. Bartlett. 2018. “Convergence of Langevin MCMC in KL-divergence”. In: *Proceedings of the 29th International Conference on Algorithmic Learning Theory*. 186–211.

Craswell, N., O. Zoeter, M. Taylor, and B. Ramsey. 2008. "An experimental comparison of click position-bias models". In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM. 87–94.

Dani, V., T. Hayes, and S. Kakade. 2008. "Stochastic linear optimization under bandit feedback". In: *Proceedings of the 21st Annual Conference on Learning Theory*. 355–366.

Dimakopoulou, M. and B. Van Roy. 2018. "Coordinated exploration in concurrent reinforcement learning". *arXiv preprint arXiv:1802.01282*.

Durmus, A. and E. Moulines. 2016. "Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm". *arXiv preprint arXiv:1605.01559*.

Eckles, D. and M. Kaptein. 2014. "Thompson sampling with the online bootstrap". *arXiv preprint arXiv:1410.4009*.

Ferreira, K. J., D. Simchi-Levi, and H. Wang. 2015. "Online network revenue management using Thompson sampling". *Working Paper*.

Francetich, A. and D. M. Kreps. 2017a. "Choosing a Good Toolkit: Bayes-Rule Based Heuristics". *preprint*.

Francetich, A. and D. M. Kreps. 2017b. "Choosing a Good Toolkit: Reinforcement Learning". *preprint*.

Frazier, P., W. Powell, and S. Dayanik. 2009. "The knowledge-gradient policy for correlated normal beliefs". *INFORMS Journal on Computing*. 21(4): 599–613.

Frazier, P., W. Powell, and S. Dayanik. 2008. "A knowledge-gradient policy for sequential information collection". *SIAM Journal on Control and Optimization*. 47(5): 2410–2439.

Ghavamzadeh, M., S. Mannor, J. Pineau, and A. Tamar. 2015. "Bayesian reinforcement learning: A survey". *Foundations and Trends in Machine Learning*. 8(5-6): 359–483.

Gittins, J. and D. Jones. 1979. "A dynamic allocation index for the discounted multiarmed bandit problem". *Biometrika*. 66(3): 561–565.

Gittins, J., K. Glazebrook, and R. Weber. 2011. *Multi-armed bandit allocation indices*. John Wiley & Sons.

Gómez-Uribe, C. A. 2016. "Online algorithms for parameter mean and variance estimation in dynamic regression". *arXiv preprint arXiv:1605.05697v1*.

Gopalan, A., S. Mannor, and Y. Mansour. 2014. "Thompson sampling for complex online problems". In: *Proceedings of the 31st International Conference on Machine Learning*. 100–108.

Gopalan, A. and S. Mannor. 2015. "Thompson sampling for learning parameterized Markov decision processes". In: *Proceedings of the 24th Annual Conference on Learning Theory*. 861–898.

Graepel, T., J. Candela, T. Borchert, and R. Herbrich. 2010. "Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine". In: *Proceedings of the 27th International Conference on Machine Learning*. 13–20.

Hill, D. N., H. Nassif, Y. Liu, A. Iyer, and S. V. N. Vishwanathan. 2017. "An efficient bandit algorithm for realtime multivariate optimization". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1813–1821.

Honda, J. and A. Takemura. 2014. "Optimality of Thompson sampling for Gaussian bandits depends on priors". In: *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*. 375–383.

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

Kandasamy, K., A. Krishnamurthy, J. Schneider, and B. Poczos. 2018. "Parallelised Bayesian optimisation via Thompson sampling". In: *To appear in proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*.

Katehakis, M. N. and A. F. Veinott Jr. 1987. "The multi-armed bandit problem: decomposition and computation". *Mathematics of Operations Research*. 12(2): 262–268.

Kauffmann, E., N. Korda, and R. Munos. 2012. "Thompson sampling: an asymptotically optimal finite time analysis". In: *Proceedings of the 24th International Conference on Algorithmic Learning Theory*. 199–213.

Kaufmann, E., O. Cappé, and A. Garivier. 2012. “On Bayesian upper confidence bounds for bandit problems”. In: *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*. 592–600.

Kawale, J., H. H. Bui, B. Kveton, L. Tran-Thanh, and S. Chawla. 2015. “Efficient Thompson sampling for online matrix-factorization recommendation”. In: *Advances in Neural Information Processing Systems 28*. 1297–1305.

Kim, M. J. 2017. “Thompson sampling for stochastic control: the finite parameter case”. *IEEE Transactions on Automatic Control*. 62(12): 6415–6422.

Kleinberg, R., A. Slivkins, and E. Upfal. 2008. “Multi-armed bandits in metric spaces”. In: *Proceedings of the 40th ACM Symposium on Theory of Computing*. 681–690.

Kveton, B., C. Szepesvari, Z. Wen, and A. Ashkan. 2015. “Cascading bandits: learning to rank in the cascade model”. In: *Proceedings of the 32nd International Conference on Machine Learning*. 767–776.

Lai, T. and H. Robbins. 1985. “Asymptotically efficient adaptive allocation rules”. *Advances in applied mathematics*. 6(1): 4–22.

Li, L., W. Chu, J. Langford, and R. E. Schapire. 2010. “A Contextual-bandit approach to personalized news article recommendation”. In: *Proceedings of the 19th International Conference on World Wide Web*. 661–670.

Littman, M. L. 2015. “Reinforcement learning improves behaviour from evaluative feedback”. *Nature*. 521(7553): 445–451.

Liu, F., S. Buccapatnam, and N. Shroff. 2017. “Information directed sampling for stochastic bandits with graph feedback”. *arXiv preprint arXiv:1711.03198*.

Lu, X. and B. Van Roy. 2017. “Ensemble Sampling”. *Advances in Neural Information Processing Systems 30*: 3258–3266.

Mattingly, J. C., A. M. Stuart, and D. J. Higham. 2002. “Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise”. *Stochastic processes and their applications*. 101(2): 185–232.

Osband, I., D. Russo, and B. Van Roy. 2013. "(More) Efficient reinforcement learning via posterior sampling". In: *Advances in Neural Information Processing Systems 26*. 3003–3011.

Osband, I., C. Blundell, A. Pritzel, and B. Van Roy. 2016a. "Deep exploration via bootstrapped DQN". In: *Advances in Neural Information Processing Systems 29*. 4026–4034.

Osband, I., D. Russo, Z. Wen, and B. Van Roy. 2017. "Deep exploration via randomized value functions". *arXiv preprint arXiv:1703.07608*.

Osband, I. and B. Van Roy. 2014a. "Model-based reinforcement learning and the eluder dimension". In: *Advances in Neural Information Processing Systems 27*. 1466–1474.

Osband, I. and B. Van Roy. 2014b. "Near-optimal reinforcement learning in factored MDPs". In: *Advances in Neural Information Processing Systems 27*. 604–612.

Osband, I. and B. Van Roy. 2017a. "On optimistic versus randomized exploration in reinforcement learning". In: *Proceedings of The Multi-disciplinary Conference on Reinforcement Learning and Decision Making*.

Osband, I. and B. Van Roy. 2017b. "Why is posterior sampling better than optimism for reinforcement learning?" In: *Proceedings of the 34th International Conference on Machine Learning*. 2701–2710.

Osband, I., B. Van Roy, and Z. Wen. 2016b. "Generalization and exploration via randomized value functions". In: *Proceedings of The 33rd International Conference on Machine Learning*. 2377–2386.

Ouyang, Y., M. Gagrani, A. Nayyar, and R. Jain. 2017. "Learning unknown Markov decision processes: A Thompson sampling approach". In: *Advances in Neural Information Processing Systems 30*. 1333–1342.

Roberts, G. O. and J. S. Rosenthal. 1998. "Optimal scaling of discrete approximations to Langevin diffusions". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 60(1): 255–268.

Roberts, G. O. and R. L. Tweedie. 1996. "Exponential convergence of Langevin distributions and their discrete approximations". *Bernoulli*: 341–363.

Rusmevichientong, P. and J. Tsitsiklis. 2010. "Linearly parameterized bandits". *Mathematics of Operations Research*. 35(2): 395–411.

Russo, D. and B. Van Roy. 2013. "Eluder Dimension and the Sample Complexity of Optimistic Exploration". In: *Advances in Neural Information Processing Systems 26*. 2256–2264.

Russo, D. and B. Van Roy. 2014a. "Learning to optimize via information-directed sampling". In: *Advances in Neural Information Processing Systems 27*. 1583–1591.

Russo, D. and B. Van Roy. 2014b. "Learning to optimize via posterior sampling". *Mathematics of Operations Research*. 39(4): 1221–1243.

Russo, D. and B. Van Roy. 2016. "An Information-Theoretic analysis of Thompson sampling". *Journal of Machine Learning Research*. 17(68): 1–30.

Russo, D. 2016. "Simple bayesian algorithms for best arm identification". In: *Conference on Learning Theory*. 1417–1418.

Russo, D. and B. Van Roy. 2018a. "Learning to optimize via information-directed sampling". *Operations Research*. 66(1): 230–252.

Russo, D. and B. Van Roy. 2018b. "Satisficing in time-sensitive bandit learning". *arXiv preprint arXiv:1803.02855*.

Schwartz, E. M., E. T. Bradlow, and P. S. Fader. 2017. "Customer acquisition via display advertising using multi-armed bandit experiments". *Marketing Science*. 36(4): 500–522.

Scott, S. 2010. "A modern Bayesian look at the multi-armed bandit". *Applied Stochastic Models in Business and Industry*. 26(6): 639–658.

Scott, S. L. 2015. "Multi-armed bandit experiments in the online service economy". *Applied Stochastic Models in Business and Industry*. 31(1): 37–45.

Srinivas, N., A. Krause, S. Kakade, and M. Seeger. 2012. "Information-Theoretic regret bounds for Gaussian process optimization in the bandit setting". *IEEE Transactions on Information Theory*. 58(5): 3250–3265.

Strens, M. 2000. "A Bayesian framework for reinforcement learning". In: *Proceedings of the 17th International Conference on Machine Learning*. 943–950.

Sutton, R. S. and A. G. Barto. 1998. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge.

Teh, Y. W., A. H. Thiery, and S. J. Vollmer. 2016. "Consistency and fluctuations for stochastic gradient Langevin dynamics". *Journal of Machine Learning Research*. 17(7): 1–33.

Thompson, W. R. 1935. "On the theory of apportionment". *American Journal of Mathematics*. 57(2): 450–456.

Thompson, W. 1933. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". *Biometrika*. 25(3/4): 285–294.

Welling, M. and Y. W. Teh. 2011. "Bayesian learning via stochastic gradient Langevin dynamics". In: *Proceedings of the 28th International Conference on Machine Learning*. 681–688.

Wyatt, J. 1997. "Exploration and inference in learning from reinforcement". *PhD thesis*. University of Edinburgh. College of Science and Engineering. School of Informatics.