Foundations of Large-Scale Sequential Experimentation

Aaditya Ramdas
aramdas@stat.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania

ABSTRACT

Large-scale sequential hypothesis testing (A/B-testing) is rampant in the tech industry, with internet companies running hundreds of thousands of tests per year. This experimentation is actually "doubly-sequential", since it consists of a sequence of sequential experiments. In this tutorial, the audience will learn about the various problems encountered in large-scale, asynchronous, doubly-sequential experimentation, both for the inner sequential process (a single sequential test) and for the outer sequential process (the sequence of tests), and learn about recently developed principles to tackle these problems. We will discuss error metrics both within and across experiments, and present state-of-the-art methods that provably control these errors, both with and without resorting to parametric or asymptotic assumptions. In particular, we will demonstrate how current common practices of peeking and marginal testing fail to control errors both within and across experiments, but how these can be alleviated using simple yet nuanced changes to the experimentation setup. We will also briefly discuss the role of multi-armed bandit methods for testing hypotheses (as opposed to minimizing regret), and the potential pitfalls due to selection bias introduced by adaptive sampling. This tutorial is timely because while almost every single internet company runs such tests, most practitioners in the tech industry focus mainly on how to run a single test correctly. However, ignoring the interplay with the outer sequential process could unknowingly inflate the number of false discoveries, as we will carefully explain in the second half of the tutorial.

CCS CONCEPTS

• **Mathematics of computing** → **Hypothesis testing and confidence interval computation**; • **General and reference** → *Experimentation*; • **Human-centered computing** → *HCI design and evaluation methods*; • **Social and professional topics** → Industry statistics;

KEYWORDS

large-scale A/B testing; false discovery rate; doubly sequential experimentation; peeking; always-valid p-values; confidence sequences; post-hoc statistical inference; selection bias; asynchronous

KDD '19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3332282

ACM Reference Format:
Aaditya Ramdas. 2019. Foundations of Large-Scale Sequential Experimentation. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'19), June 22–24, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332282

1 TUTORIAL OBJECTIVES

This **traditional** tutorial serves three objectives:

(1) To educate the community broadly about the pitfalls of the current methods of performing large scale sequential experimentation, specifically those of peeking at p-values and marginal testing ignoring multiplicity (covered in topics I.2 and II.1 in the outline).
(2) To summarize advances made over the last few years to provide a unified framework for doubly-sequential experimentation with end-to-end guarantees on error rates (covered in topics I.3, I.4, II.4, III.1, III.2 in the outline), since they are scattered over the ML and Statistics literatures.
(3) To encourage immediate adoption by practitioners of several concrete mature ideas, as well as suggest avenues for future investigation by researchers (topics IV, V in the outline).

2 TARGET AUDIENCE

We expect broad interest from academics (students, professors), researchers working with applied scientists, and industry participants (researchers and practitioners). There will be enough new concepts for all attendees, and connections to other areas of data mining and machine learning research will be pointed out as appropriate. The mathematical background required would be basic probability and statistics. Most topics will be introduced as needed at both an intuitive and mathematical level. No special expertise or prior knowledge is needed to understand this tutorial, and we expect that even beginning masters or PhD students will be able to grasp the material.

3 TUTOR BIOGRAPHY

Aaditya Ramdas (`aramdas@cmu.edu`) is an assistant professor in the Department of Statistics and Data Science, and affiliated with the Machine Learning Department at Carnegie Mellon University. Previously, he was a postdoctoral researcher in Statistics and EECS at UC Berkeley from 2015-18, hosted by Martin Wainwright and Michael Jordan. He finished his PhD at CMU in Statistics and Machine Learning with Aarti Singh and Larry Wasserman, winning the Umesh K. Gavaskar Memorial Thesis Award. His undergraduate degree was in Computer Science from IIT Bombay. A lot of his research focuses on modern aspects of reproducibility in science and technology, involving statistical testing and false discovery rate control in static and dynamic settings. He also works on some problems in sequential decision-making and online uncertainty quantification.

4 TUTORIAL OUTLINE

Introduction: introduction to doubly-sequential experimentation, motivations, and outline of tutorial.

(First half) I. The inner process: a single experiment.

(1) Hypothesis testing basics: null/alternative, p-value, type-1 error, power, duality with confidence intervals.
(2) The threat caused by the common practice of peeking: why repeatedly running batch tests on accumulating data invalidates the p-value and inflates type-1 error.
(3) How sequential tests avoid the peeking problem, Gaussian/binomial SPRT (sequential probability ratio tests), constructing always-valid p-values.
(4) Constructing confidence sequences (uniformly valid sequence of confidence intervals), nonparametric extensions to SPRT.
(5) The use of multi-armed bandits (MAB) as adaptive hypothesis tests, dealing with selection bias, constructing correct p-values from MAB experiments.

(Second half) II. The outer process: a sequence of experiments.

(1) The issue of multiplicity: why running thousands of experiments necessitates a correction, even if the experiments and hypotheses are all independent.
(2) Why familywise error rate is too conservative in the online setting, motivation for the false discovery rate (FDR) as a more natural error metric.
(3) Why Benjamini-Hochberg cannot be used online, motivation for alpha-investing.
(4) Online control of the FDR in the fully synchronized setting: algorithms beyond alpha-investing, like LORD, SAFFRON and ADDIS, and corresponding intuition.
(5) The LORD-CI algorithm for online control of the false coverage rate, and post-hoc analysis.

(Last stretch) III. The joint doubly-sequential process.

(1) Algorithms for fully asynchronous doubly-sequential experimentation, the principle of pessimism.
(2) End-to-end anytime post-hoc guarantees, both within and across experiments.

IV. An explicit example: sequential average treatment effect estimation in A/B tests (or equivalently randomized clinical trials) with adaptive randomization.

V. Issues, open directions for theory and practice: hierarchical error rates for large organizations, incentives to encourage groups to control both within and across group errors.

5 SOME RELATED WORK

(a) classical work on sequential testing, experiment design and confidence sequences [3, 4, 12, 16];
(b) modern work on uniform bounds for multi-armed bandits [7, 11, 19];
(c) offline and online control of the false discovery rate and false coverage rate [2, 5, 8, 13–15, 17];
(d) peeking, always-valid p-values, sequential nonparametric testing, multi-armed bandits for testing [1, 9, 10, 18];
(e) uniform, nonparametric confidence sequences for sequential estimation and testing [6];
(f) asynchronous multiple hypothesis testing for doubly sequential experimentation [20].

REFERENCES

[1] Akshay Balsubramani and Aaditya Ramdas. 2016. Sequential Nonparametric Testing with the Law of the Iterated Logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence (UAI'16)*. AUAI Press, Arlington, Virginia, 42–51.

[2] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)* 57, 1 (1995), 289–300.

[3] Herman Chernoff. 1959. Sequential design of experiments. *The Annals of Mathematical Statistics* 30, 3 (1959), 755–770.

[4] D. A. Darling and Herbert Robbins. 1967. Inequalities for the Sequence of Sample Means. *Proceedings of the National Academy of Sciences* 57, 6 (June 1967), 1577–1580.

[5] Dean Foster and Robert Stine. 2008. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 70, 2 (2008), 429–444.

[6] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. 2018. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240* (2018).

[7] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. 2014. lil' UCB: An Optimal Exploration Algorithm for Multi-Armed Bandits. In *Proceedings of The 27th Conference on Learning Theory (Proceedings of Machine Learning Research)*, Vol. 35. 423–439.

[8] Adel Javanmard and Andrea Montanari. 2018. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics* 46, 2 (2018), 526–554.

[9] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2017. Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1517–1525.

[10] Ramesh Johari, Leo Pekelis, and David J. Walsh. 2015. Always valid inference: Bringing sequential analysis to A/B testing. *arXiv preprint arXiv:1512.04922* (2015).

[11] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. 2016. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research* 17, 1 (2016), 1–42.

[12] Tze Leung Lai. 1976. On Confidence Sequences. *The Annals of Statistics* 4, 2 (March 1976), 265–280.

[13] Aaditya Ramdas, Rina Foygel Barber, Martin Wainwright, and Michael Jordan. 2019. A unified treatment of multiple testing with prior knowledge using the p-filter. *Annals of Statistics (accepted)* (2019).

[14] Aaditya Ramdas, Fanny Yang, Martin Wainwright, and Michael Jordan. 2017. Online control of the false discovery rate with decaying memory. In *Advances In Neural Information Processing Systems*. 5655–5664.

[15] Aaditya Ramdas, Tijana Zrnic, Martin Wainwright, and Michael Jordan. 2018. SAFFRON: an Adaptive Algorithm for Online Control of the False Discovery Rate. In *Proceedings of the 35th International Conference on Machine Learning*. 4286–4294.

[16] Abraham Wald. 1945. Sequential Tests of Statistical Hypotheses. *Annals of Mathematical Statistics* 16, 2 (1945), 117–186.

[17] Asaf Weinstein and Aaditya Ramdas. 2019. Online Control of the False Coverage Rate and False Sign Rate. *arXiv preprint arXiv:1905.01059* (2019).

[18] Fanny Yang, Aaditya Ramdas, Kevin G Jamieson, and Martin J Wainwright. 2017. A framework for Multi-A(rmed)/B(andit) Testing with Online FDR Control. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA, USA.

[19] Shengjia Zhao, Enze Zhou, Ashish Sabharwal, and Stefano Ermon. 2016. Adaptive Concentration Inequalities for Sequential Decision Problems. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*. Barcelona, Spain.

[20] Tijana Zrnic, Aaditya Ramdas, and Michael I Jordan. 2018. Asynchronous Online Testing of Multiple Hypotheses. *arXiv preprint arXiv:1812.05068* (2018).