STAT 547C Final Project

Benjamin Bloem-Reddy

October 18, 2022

Logistics

The final project of a project outline (due **November 4 at 11:59 pm**) and a report (due **December 14 at 11:59 pm**).

Templates for the outline and for the report are included in the GitHub repository template (https://github.com/ben-br/stat547c-project-template).

Choosing a project

Broadly speaking, choose a topic that you're interested in from a research perspective. Or, a topic that you're genuinely curious about. Hopefully your topic meets both criteria.

If you're already involved in research, use this project to:

- formulate your problem(s) in rigorous probabilistic terms;
- learn more about your research area;
- develop a deeper understanding;
- take steps towards new results using probability/measure theory.

If you're not already involved in research, or you'd like to start something new, I have included a list of potential topics at the end of this document; **please discuss with me in more detail before choosing**. Alternatively, you may propose your own topic; please meet with me individually to give me an idea of what you have in mind—I want to make sure the project is appropriately scoped. It should go without saying that whatever you do should have a heavy dose of probability.

You have substantial freedom to develop the project as you see fit. You might focus on any of the following (this is not an exhaustive list):

- The proofs of key results from a particular area.
- An update of a classical survey of some aspect of probability.
- An open research problem that you are already working on, or that you have been thinking about working on.

Report

The report should read like lecture notes: expository and informal for clarity, with plenty of precision and rigor. In general, it's helpful to have an audience in mind when you're writing. For this report, write for your classmates. The report should have the following high-level structure:

1. **Background.** What is the topic? What is known? What is not known? What are the major results? Any faults or shortcomings? Use this section to set notation, provide the basic ideas and definitions.

2. **Body.** The actual structure of the body will vary depending on the subject and type of report, but this is core of the report. Technical developments, proofs, intermediate results, simulation results, etc., should be included here.
3. **Open problems/research directions.** Your final section, which also serves as the conclusion, should propose at least one open problem or research direction, along with your ideas of how to approach the problem—the more precise, the better. The problem should be clearly motivated by the previous parts of the report, and your ideas of how to approach the problem should include what you view to be the most important obstacles, and any relevant related work.
4. **Appendix: Exercises/problems.** Demonstrate your understanding of your subject by creating at least two exercises or assignment problems (with solutions) suitable for graduate-level probability students. Simplified versions of results from the literature are acceptable, with proper attribution, as long as the pedagogical intention is clear.

There is a strict **12-page limit** (including figures, tables, etc.). You may include an appendix of unbounded length but I may not read it, with the exception of the exercises/problems section, which I will read.

Submitting your report

The report should be submitted as a GitHub repository based on the template found at: `https://github.com/ben-br/stat547c-project-template`. The template includes a LaTeX style file that should be used for the report. (Detailed instructions for usage can be found in the repository's README file.)

Add me as a collaborator to your project repo, and when you're ready to submit your project, `git commit` with the message `final project submission`.

Any experimental/numerical results should be reproducible. All code should be reusable, clearly commented/documented, and exist in a GitHub repository to which I have access as a collaborator. (My GitHub ID is ben-br).

Resources

- Some resources on technical/mathematical writing:
  - Trevor Campbell's "How to Explain Things" talk
  - Knuth, Larrabee, and Roberts on mathematical writing: `http://www.jmlr.org/reviewing-papers/knuth_mathematical_writing.pdf`
  - Halmos on writing mathematics: `https://www.math.uh.edu/~tomforde/Books/Halmos-How-To-Write.pdf` (a transcribed, searchable PDF with some typos: `https://entropiesschool.sciencesconf.org/data/How_to_Write_Mathematics.pdf`)
- Getting started with Git: chapters 1 and 2 of `https://git-scm.com/book/en/v2` should be all you need for this report. You may also find `https://uoftcoders.github.io/studyGroup/lessons/git/collaboration/lesson/` helpful.

Potential topics

The following are some potential topics. Feel free to propose a variation or your own topic.

1. Probabilistic foundations of causal inference

Reference(s): [PJS17] as background, then any number of directions, e.g., [CS19; Daw21]

Causal (probability) models are different from non-causal (i.e., observational) probability models in a number of ways; probably the biggest difference is that causal models consider what happens to a distribution when it is intervened upon. Give a detailed account of causal probability models, with a focus on what happens to the underlying probability space under various kinds of interventions.

2. Symmetry in statistics and machine learning

Reference(s): ML: [Ele21]; Statistics: [Hel04; HB66; HB67; Kie57; ES99; TL13]

Learn about an application of symmetry to statistics or machine learning, and take steps towards new results (discuss with me for ideas). Note that many of the classical statistical ideas have largely unexplored connections to modern problems in machine learning.

3. Probabilistic programs

Reference(s): In order of increasing technical difficulty: [Mee+18], [Rai17, Ch. 4–7], [Roy11, Ch. II-III]

Give a treatment of the probabilistic foundations of probabilistic programming.

4. Neutral stochastic processes

Reference(s): [Dok74; Jam06; BO17; Blo+18]

Learn about neutral-to-the-right stochastic processes, and take steps towards a theory of neutral-to-the-left stochastic processes defined on arbitrary spaces (in analogy to [Jam06]).

References

[BO17] B. Bloem-Reddy and P. Orbanz. *Preferential Attachment and Vertex Arrival Times.* 2017. arXiv: 1710.02159 [math.PR].

[Blo+18] B. Bloem-Reddy et al. “Sampling and inference for Beta Neutral-to-the-Left models of sparse networks”. In: *Proc. 34th Intl. Conf. Uncertainty in Artificial Intelligence.* 2018.

[CS19] I. Cabreros and J. D. Storey. “Causal models on probability spaces”. In: (July 2019). eprint: 1907.01672. URL: https://arxiv.org/abs/1907.01672.

[Daw21] P. Dawid. “Decision-theoretic foundations for statistical causality”. In: *Journal of Causal Inference* 9.1 (Jan. 2021), pp. 39–77. DOI: 10.1515/jci-2020-0008. URL: https://doi.org/10.1515%2Fjci-2020-0008.

[Dok74] K. Doksum. “Tailfree and Neutral Random Probabilities and Their Posterior Distributions”. In: *Ann. Probab.* 2.2 (Apr. 1974), pp. 183–201.

[ES99] M. L. Eaton and W. D. Sudderth. “Consistency and Strong Inconsistency of Group-Invariant Predictive Inferences”. In: *Bernoulli* 5.5 (1999), pp. 833–854.

[Ele21] B. Elesedy. “Provably Strict Generalisation Benefit for Invariance in Kernel Methods”. In: *Advances in Neural Information Processing Systems.* Vol. 34. 2021, pp. 17273–17283. URL: https://proceedings.neurips.cc/paper/2021/file/8fe04df45a22b63156ebabbb064fcd5e-Paper.pdf.

[Hel04] I. S. Helland. “Statistical Inference under Symmetry”. In: *International Statistical Review / Revue Internationale de Statistique* 72.3 (2004), pp. 409–422.

[HB66] R. B. Hora and R. J. Buehler. “Fiducial Theory and Invariant Estimation”. In: *The Annals of Mathematical Statistics* 37.3 (1966), pp. 643–656.

[HB67] R. B. Hora and R. J. Buehler. “Fiducial Theory and Invariant Prediction”. In: *The Annals of Mathematical Statistics* 38.3 (1967), pp. 795–801.

[Jam06] L. F. James. “Poisson calculus for spatial neutral to the right processes”. In: *The Annals of Statistics* 34.1 (2006), pp. 416–440.

[Kie57] J. Kiefer. “Invariance, Minimax Sequential Estimation, and Continuous Time Processes”. In: *The Annals of Mathematical Statistics* 28.3 (1957), pp. 573–601.

[Mee+18] J.-W. van de Meent et al. *An Introduction to Probabilistic Programming.* 2018. arXiv: 1809.10756 [stat.ML].

[PJS17] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms.* MIT Press, 2017. URL: https://library.oapen.org/handle/20.500.12657/26040.

[Rai17] T. Rainforth. “Automating Inference, Learning, and Design using Probabilistic Programming”. PhD thesis. University of Oxford, 2017. URL: http://www.robots.ox.ac.uk/~twgr/assets/pdf/rainforth2017thesis.pdf.

[Roy11] D. M. Roy. “Computability, inference and modeling in probabilistic programming”. PhD thesis. Massachusetts Institute of Technology, 2011.

[TL13] G. Taraldsen and B. H. Lindqvist. “Fiducial theory and optimal inference”. In: *The Annals of Statistics* 41.1 (2013).