# STAT 547C Final Project

Benjamin Bloem-Reddy

October 25, 2019

## Logistics

The final project consists of a report and a written peer review of a classmate's report. The report is due on **December 11 at 5:00 pm**. The peer review is due one week later, **December 18 at 5:00 pm**. Your project outline is due on **November 4 at 3:00 pm**.

## Choosing a project

Broadly speaking, choose a topic that you're interested in from a research perspective. If you're already involved in research, use this project to learn more about/develop your understanding of/take steps towards new results using probability in your research area. 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; feel free to discuss with me in more detail before choosing. Alternatively, you may propose your own topic; please meet with me individually to give me a high-level idea of what you have in mind—I want to make sure the project is properly 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 a combination of research paper and lecture notes. It's helpful to have an audience in mind when you're writing—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? 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.

Out of respect for your reviewers' (i.e., your classmate and me) time, there is a strict **12-page limit** (including figures, tables, etc.). You may include an appendix of unbounded length but your reviewers are not obligated to read them, 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).

## Peer review

Good peer review is an essential part of academic research. Giving a careful and constructive review is something you should take seriously. (If you've ever had the experience of getting a negative but unthoughtful review that offers no feedback on how to improve your paper, this should resonate with you.) Serving as a reviewer is one way to be a good academic citizen. Some guidelines:

- Set aside sufficient time. A hurried review is typically unhelpful.
- Be constructive: start with the mindset that the paper is *good*, and your job is to offer feedback on how to make the paper *better*.
- Focus on content.

Some resources:

- Wiley's top tips for reviewers: `https://authorservices.wiley.com/Reviewers/journal-reviewers/how-to-perform-a-peer-review/top-tips-for-peer-reviewers.html`.
- Wiley's detailed reviewer guidelines: `https://authorservices.wiley.com/Reviewers/journal-reviewers/how-to-perform-a-peer-review/step-by-step-guide-to-reviewing-a-manuscript.html`.
- The NeurIPS reviewer guidelines: `https://nips.cc/Conferences/2019/PaperInformation/ReviewerGuidelines`. See, in particular, the sections "Reviewer best practices", "Review content", and "Examples of Review Content".

Details on format are forthcoming.

## 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. Computational and statistical uses of exchangeability

**Reference(s):** [Ber96; OR15]

A classic article by Kingman [Kin78] and a more recent follow-up by Aldous [Ald10] survey probabilistic uses of exchangeability. Give a detailed treatment of its computational and statistical uses, including a survey of methods that rely on exchangeability. (Some overlap with the next project.)

### 2. Conditional independence in computational statistics and machine learning

**Reference(s):** [Daw79; Daw80]

The classic paper by Dawid [Daw79] presents core ideas of statistics in terms of conditional independence. Update the review, with a focus on computationally intensive methods in statistics and machine learning. (Some overlap with the previous project.)

### 3. Probabilistic symmetry in deep learning

**Reference(s):** [KT18; BT18a; BT18b] (see me for an updated version of the latter)

Learn about probabilistic (and deterministic) symmetry in deep learning, and take steps towards a probabilistic version of the main result in [KT18].

### 4. Stein's method in machine learning

**Reference(s):** [Ros11] for fundamentals, [GM15; LW16; Zhu+18] for some potential applications.

Learn about the fundamentals of Stein's method and give detailed treatment of some applications in machine learning.

### 5. 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.

### 6. Statistical learning theory and generalization bounds

**Reference(s):** [BBL03] is a nice overview of classical results, https://bcourses.berkeley.edu/courses/1409209 is the course webpage from one of the leaders in the field. [MA13; Gue19] for an overview of PAC-Bayes theory, [Cat07] for a rigorous but not-very-accessible presentation of PAC-Bayes theory.

Give a detailed treatment of some aspect of statistical learning theory.

### 7. 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]).

**8. Branching processes**

**Reference(s):** [AN72; Dur15]

Give a detailed treatment of some aspect of branching processes, ideally with an application.

**9. Preferential attachment graphs**

**Reference(s):** [Hof16; BO17]

Learn about preferential attachment graphs.

## References


[Ald10] D. J. Aldous. "More uses of exchangeability: representations of complex random structures". In: *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman*. Ed. by N. H. Bingham and C. M. Goldie. London Mathematical Society Lecture Note Series. Cambridge University Press, 2010, pp. 35–63.

[AN72] K. B. Athreya and P. E. Ney. *Branching Processes*. Springer, 1972.

[Ber96] J. M. Bernardo. "The concept of exchangeability and its applications". In: *Far East Journal of Mathematical Sciences* 4 (1996), pp. 111–122.

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

[BT18a] B. Bloem-Reddy and Y. W. Teh. "Neural network models of exchangeable sequences". In: *NeurIPS Workshop on Bayesian Deep Learning*. 2018.

[BT18b] B. Bloem-Reddy and Y. W. Teh. *Probabilistic symmetry and invariant neural networks*. 2018. arXiv: `1901.06082 [stat.ML]`.

[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.

[BBL03] O. Bousquet, S. Boucheron, and G. Lugosi. "Introduction to statistical learning theory". In: *Summer School on Machine Learning*. 2003, pp. 169–207. URL: `http://www.econ.upf.edu/~lugosi/mlss_slt.pdf`.

[Cat07] O. Catoni. "PAC-Bayesian supervised classification: the thermodynamics of statistical learning". In: *Institute of Mathematical Statistics* (2007).

[Daw79] A. P. Dawid. "Conditional Independence in Statistical Theory". In: *J. Royal Stat. Soc. Ser. B* 41.1 (1979), pp. 1–31.

[Daw80] A. P. Dawid. "Conditional Independence for Statistical Operations". In: *The Annals of Statistics* 8.3 (1980), pp. 598–617.

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

[Dur15] R. Durrett. "Branching Process Models of Cancer". In: *Branching Process Models of Cancer*. Mathematical Biosciences Institute Lecture Series. Springer International Publishing, 2015, pp. 1–63. URL: `http://dx.doi.org/10.1007/978-3-319-16065-8_1`.

[GM15] J. Gorham and L. Mackey. "Measuring Sample Quality with Stein's Method". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. 2015, pp. 226–234. URL: `http://papers.nips.cc/paper/5768-measuring-sample-quality-with-steins-method.pdf`.

[Gue19] B. Guedj. *A Primer on PAC-Bayesian Learning*. 2019. arXiv: `1901.05353 [stat.ML]`.

[Hof16] R. v. d. Hofstad. *Random Graphs and Complex Networks*. Vol. 1. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Nov. 2016.

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

[Kin78] J. F. C. Kingman. "Uses of Exchangeability". In: *The Annals of Probability* 6.2 (1978), pp. 183–197. URL: http://dx.doi.org/10.1214/aop/1176995566.

[KT18] R. Kondor and S. Trivedi. "On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups". In: *Proc. ICML 35*. Vol. 80. PMLR. 2018, pp. 2747–2755.

[LW16] Q. Liu and D. Wang. "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. 2016, pp. 2378–2386. URL: http://papers.nips.cc/paper/6338-stein-variational-gradient-descent-a-general-purpose-bayesian-inference-algorithm.pdf.

[MA13] D. McAllester and T. Akinbiyi. "PAC-Bayesian Theory". In: *Empirical Inference*. Springer Berlin Heidelberg, 2013. Chap. 10, pp. 95–103. URL: http://dx.doi.org/10.1007/978-3-642-41136-6_10.

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

[OR15] P. Orbanz and D. M. Roy. "Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2 (2015), pp. 437–461.

[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.

[Ros11] N. Ross. "Fundamentals of Stein's method". In: *Probability Surveys* 8 (2011), pp. 210–293. URL: http://dx.doi.org/10.1214/11-PS182.

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

[Zhu+18] J. Zhuo et al. "Message Passing Stein Variational Gradient Descent". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. PMLR, 2018, pp. 6018–6027. URL: http://proceedings.mlr.press/v80/zhuo18a.html.