STAT 548 PhD Qualifying Papers Guide and Expectations

Benjamin Bloem-Reddy

September 15, 2023

Choosing a paper

At the end of this document is a list of papers and project ideas that I am interested in supervising as Qualifying Papers (QPs). I am generally interested in theoretical and methodological aspects of statistics and machine learning, and various problems in applied probability. My recent research efforts have focused on symmetry and invariance in various statistics and machine learning contexts, and on causal inference/discovery.

Expectations

If you are interested in doing a QP with me, the first step is to email me to schedule a one-on-one meeting. Please use the words "Qualifying Paper" in the subject line of your email (I have a filter set up so that I don't miss these). Please come to this meeting prepared to discuss the following:

- Your background.
- Your long-term research interests (it's okay if these are not yet well-defined).
- Why you are interested in the particular paper/project.
- When you will submit your report (typically four-six weeks after we meet).
- The details of the QP project and report.

It's not a terrible idea to have spent some time with the paper, at least enough time to know what the basic ideas are.

There are **two** milestones as part of your QP:

1. A one-on-one meeting 2-3 weeks after our preparatory meeting. In this meeting, you will give an informal 20-minute presentation on the paper (slides or whiteboard; whichever you feel is more effective), including relevant background/context and your technical/methodological summary and analysis. Essentially, you'll present the content that makes up parts 1-2 of your report (see below), with an emphasis on part 1. We'll have some discussion/Q&A, and then you will present 1-2 ideas/proposals for projects related to the QP. We'll talk through these and possibly other options, and agree on a small QP-specific project.
2. Your QP report, which includes the results of your project; details are below.

Ideally, the timeline for your QP should be: 2-3 weeks reading, understanding, presenting, and starting to write parts 1-3 of your report (see below); 3-4 weeks working on your project and finishing your report.

Report[1]

The report should have the following structure:

1. **Technical/methodological summary and analysis (five-page limit; you will be graded for clarity and conciseness).**
    - *Theory papers:* Give a precise summary of the main technical results, and how they are proved. What assumptions are made? Why? What existing results are relied on? What are the technical challenges? Are there any technical innovations? What is the overall structure of the argument? Can it be generalized? Are there different proof techniques?
    - *Methodological papers:* Give a detailed explanation of the proposed methodology, along with any theoretical properties. Under what conditions should the methodology be expected to work well? To fail? For performance be ambiguous? What about computational properties? (E.g., hardness of an optimization problem, computational complexity, etc.)
2. **Conceptual summary (maximum of three pages).** This section of the report should provide a high-level summary of the paper and the problem(s) it addresses, including previous work, the contributions/impact of the work in question (WIQ) (e.g., novel techniques, problem formulations, or synthesis; important results; subsequent work), *why* the results and contributions are important, and any limitations or shortcomings (e.g., restrictive assumptions or flawed methodology). The aim of this section is for you to synthesize the findings of a body of work and clearly present the important points.
3. **Mini-proposals for research projects (maximum of two pages per proposal, at least two proposals—no limit on number of proposals).** Each proposal should describe a research project that applies, extends, generalizes, adapts, or addresses shortcomings of the WIQ. Seemingly unrelated ideas inspired by the WIQ are also fine. A proposal should concisely describe: the primary problem to be addressed; an approach (or multiple approaches) for addressing the problem; any technical or conceptual sub-problems; the potential impact of the project. You are not expected to pursue any of these projects (though we can talk more if you would like to). The aim of this section is to get you thinking creatively about research, and to begin developing the skills necessary for writing research proposals.
4. **Paper-specific project results (8-page limit).** After writing a reasonably complete draft of sections 1-3 above, we will discuss a short paper-specific project. We can meet again (as many times as necessary) before the report due date. Your grade *will not* be affected by how good the results are, whether your approach improves on past work, or whether you achieve the initial goal of the project. I will use this project to evaluate your research potential, which includes:
    - clearly formulating a research question;
    - setting up a useful mathematical framework for the problem;
    - thinking creatively and independently to develop a solution;
    - relating the problem to existing work, in other fields if necessary;
    - being resourceful and asking questions when necessary;
    - learning from and moving past the inevitable setbacks;
    - reformulating the research problem when necessary;
    - implementing new methods in code (when applicable);
    - choosing appropriate experiments and metrics;
    - communicating and reflecting on progress, setbacks, and results;
    - thinking of future research directions.

[1] With thanks to Trevor Campbell, from whom I am borrowing liberally.

The report should be submitted as a PDF using the format in the GitHub repository downloaded from the template here: https://github.com/ben-br/qp-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.)

Any experimental/numerical results should be reproducible. Any code should be reusable, clearly commented/documented, and exist in a GitHub repository to which I have access as a collaborator. I prefer Julia or Python or R (in that order), but we can talk about using other languages if necessary.

Resources

- Some resources on technical/mathematical writing:
  - Nancy Heckman's page: http://ugrad.stat.ubc.ca/~nancy/writing/
  - Harry Joe on mathematical writing and typesetting in LaTeX: https://www.stat.ubc.ca/~harry/papers/
  - 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.

Candidate papers (last updated: September 15, 2023)

1. Learning theory for problems with geometric stability

Paper: A. Bietti, L. Venturi, and J. Bruna. *On the Sample Complexity of Learning with Geometric Stability*. 2021. arXiv: 2106.07148 [stat.ML]

Categories: theory, invariance, machine learning, kernel methods

2. Hypothesis testing for "maximal" group symmetry

Paper: L. G. Christie and J. A. D. Aston. "Estimating Maximal Symmetries of Regression Functions via Subgroup Lattices" (Mar. 2023). eprint: 2303.13616. URL: https://arxiv.org/pdf/2303.13616.pdf

Categories: theory, invariance, hypothesis testing

You may also find [CB23] relevant.

3. Hypothesis testing for specified group symmetry

Paper: K. Chiu and B. Bloem-Reddy. "Non-parametric Hypothesis Tests for Distributional Group Symmetry" (July 2023). eprint: 2307.15834. URL: https://arxiv.org/pdf/2307.15834.pdf

Categories: theory, invariance, hypothesis testing

You may also find [CA23] relevant.

4. Statistical aspects of triangular flows

Paper: N. J. Irons et al. "Triangular Flows for Generative Modeling: Statistical Consistency, Smoothness Classes, and Fast Rates" (Dec. 2021). eprint: 2112.15595. URL: https://arxiv.org/pdf/2112.15595.pdf

Categories: theory, machine learning, triangular maps

5. Approximation theory of measure-transport

Paper: R. Baptista et al. "An Approximation Theory Framework for Measure-Transport Sampling Algorithms" (Feb. 2023). eprint: 2302.13965. URL: https://arxiv.org/pdf/2302.13965.pdf

Categories: theory, machine learning, optimal transport

References

[Bap+23] R. Baptista et al. "An Approximation Theory Framework for Measure-Transport Sampling Algorithms" (Feb. 2023). eprint: 2302.13965. URL: https://arxiv.org/pdf/2302.13965.pdf.

[BVB21] A. Bietti, L. Venturi, and J. Bruna. *On the Sample Complexity of Learning with Geometric Stability*. 2021. arXiv: 2106.07148 [stat.ML].

[CB23] K. Chiu and B. Bloem-Reddy. "Non-parametric Hypothesis Tests for Distributional Group Symmetry" (July 2023). eprint: 2307.15834. URL: https://arxiv.org/pdf/2307.15834.pdf.

[CA23] L. G. Christie and J. A. D. Aston. "Estimating Maximal Symmetries of Regression Functions via Subgroup Lattices" (Mar. 2023). eprint: 2303.13616. URL: https://arxiv.org/pdf/2303.13616.pdf.

[Iro+21] N. J. Irons et al. "Triangular Flows for Generative Modeling: Statistical Consistency, Smoothness Classes, and Fast Rates" (Dec. 2021). eprint: 2112.15595. URL: https://arxiv.org/pdf/2112.15595.pdf.