Project Guidelines

B9145: Topics in Trustworthy AI

Hongseok Namkoong

The course project is perhaps the most important and interesting component of this class. You may do this in pairs, or alone. Please use the course staff, as well as your peers, for continual feedback on the project. I *highly encourage you* to schedule short meetings with me early on in the semester.

You are required to submit a final report for course projects by **May 4, 11:59pm** to Canvas. This should be limited to at most 10 pages (excluding the bibliography for citations), single-spaced with 11 point font, with at least 1 inch margins. For technical projects, the final report should be typeset in Latex.

Description: The main goal of the course project is to formulate an interesting and novel research question related to the themes of the class. You should schedule individual meetings or drop by office hours to discuss potential directions; I am happy to discuss ideas and suggest potential topics. I encourage you to continually discuss progress throughout the semester.

You should begin choosing a project topic (and a partner, if you'd like to work in pairs) in late February. You should write a short description of the proposed topic ($\leq$3 paragraphs), and upload this to Canvas by **Mar 14**. Before submitting your proposal, I expect you to have discussed the topic with me at least once.

If making progress on a research project proves difficult, you may do a pedagogical project instead. See below for a detailed instruction on this modality; you should choose this option *only after* having pushed on your chosen research project topic for the majority of the semester.

Research Project: Your research project can exist anywhere in the applied-to-theoretical spectrum of possible topics. Please view the project as an opportunity to develop a critical perspective on a particular literature, ideate a high-impact research project, and make progress throughout the semester (and ideally beyond). While the expectation is not to produce publishable research (this would be very difficult to do in a semester!), you should view this as an opportunity to start a project that can *potentially* result in a high-quality publication.

If you are already working on a problem in learning and optimization—broadly interpreted—or an application area (e.g. healthcare, digital marketing, finance etc), you may use it as a course project as long as you read at least three papers related to the themes of this class and discuss potential interfaces in your final report. Ideally, there would be a natural and interesting interface between your research and the themes of this class; I'm willing to let this be broadly interpreted.

The following is a brief description of what I expect in a good final report.

1. Motivation: You should clearly describe why the problem you are studying is interesting. This can be applied (e.g. "this is a practically important problem not considered by previous works", "previously proposed algorithm is difficult to implement"), or technical (e.g., "analysis of previous work was loose in some aspects", "you want to relax an unrealistic assumption made by a previous work"). Usually, you will have a combination of both: a problem bears

practical importance—you will need to carefully argue this—and is also technically interesting because existing works do not address it. You may find it helpful to illustrate your research motivation using an example; this can be very concrete or somewhat abstract, depending on the topic.

2. Related work: You should have a thorough discussion of the literature surrounding the problem. In addition to papers related to the high-level motivation of the problem, you should delineate related works in terms of solution approaches. Your discussion should demonstrate a detailed technical level of understanding of the literature, and make clear where your problem is situated within it. A good literature review provides a critical perspective of related works, tailored to the particular problem at hand.

3. Problem formulation: The final report should provide a precise description of the problem you are studying. You should clearly formulate the problem of interest, and articulate and justify your modeling assumptions. For example, this could be the notion of optimality you are studying or the solution concept such as the notion of equilibrium for strategic agents. This could also be benchmarks you are comparing your solution against both theoretically and empirically.

4. Contributions: You should clearly describe your overall approach to the problem. If you have made tangible progress, you should write up your results carefully, alongside a discussion of its implications and limitations. If you have tried a few approaches without apparent success, you may also describe your efforts and discuss why these approaches did not work. You are also encouraged to discuss potential approaches you may undertake in the future.

Pedagogical Project: Since research is an uncertain process, the pedagogical project option intends to eliminate the cost of a research project not working out. This is a less preferred option, and should be viewed as a last resort. In the case that making progress on your research project prove difficult, you may perform a critical review of at least two papers relevant to the course material. This can consist of a combination of the following:

1. A peer-review style discussion on the strengths and weaknesses of the paper, alongside potential proposals to improve upon authors' results. Are the motivation of the paper well-justified? Are the assumptions realistic? Does the solution concept provide a natural goalpost? Is the proposed procedure practical? Does the analysis provide interesting insights, either from the results or proof techniques? Does the empirical evaluation provide a good description of the strength and weaknesses of the authors' proposal? What is notably missing in the discussion?

2. A critical reproduction of empirical results. Is the authors' approach practically relevant? Are the results reproducible? Are the results robust against hyperparameter choices, shifts in data distribution, realistic violations to the modeling assumption etc?

3. A couple of exercises from key technical results. Develop a couple of homework-style problems of similar difficulty as problem sets in class. The goal here is to provide a succinct description of the main technical insights inside a result. A good exercise gives you insight on the overall technical picture, without covering all the details.