Statistics 311/Electrical Engineering 377 Project

The final component of the course is to do a course project. The course project should (broadly) be related to some part of the content of the course. The project may be performed in teams of up to three students—though single and pair projects are fine as well.

Please use the course staff—John and Leighton—as well as your peers for feedback on the projects. Come talk to us during office hours, and talk early and often. We would love to give you feedback!

1 Types of projects

Here we list a few of the types of projects we expect to see—if you would like to deviate from this, please email the course staff to double check that this is OK. The final product of the research should be a four to six page report detailing your work.

1. Research project: original research related to the content of the course. This can be either theoretical work or applied, and ideally will be related to work you are already performing so as to help with your current research interests. These projects could include application of some of the techniques in class to new areas, so that the projects are very applied, or theoretical work.

2. Pedagogical project: develop a set of lecture notes and a few problems—of roughly the same difficulty as the problem sets in class—relating to some aspect of the course. A project in this vein may consist of replicating the empirical work in a paper and developing problems associated with it, or surveying an area of statistics and information theory and extracting exercises from a few papers around the area.

2 Logistics and dates

There are three main components for the projects: a proposal, a midterm report, and a final report and presentation. You should submit all parts of the project to the course staff at stats311submit@gmail.com. All page limits do not include references; we will not count those against the limit.

Part of the project is to give peer review: if you are in the class and doing a project, you will be randomly assigned as a reviewer to two other projects (so that each project will have several reviewers). You will give feedback on these projects throughout the quarter as detailed below.

Proposal: Due Thursday, 1/31, 1:30pm. The proposal is a (maximum) two-page document including the names of the student(s) working on the project and describing the planned project. The proposal should include a paragraph of background; a paragraph describing the approach that will be taken; and a paragraph about the hoped for (or desired) results as well as potential failure points of the project.

We will evaluate the proposals based on the following criteria.

i. Clarity: is what the project attempting to solve clear? Can anyone who has been in the class for the past few weeks be reasonably expected to understand the material, or at least the introduction?
ii. Relevance: is this project relevant to the course (information theory or statistics)?

iii. Writing style: does the writing have good mathematical notation? Is the grammar acceptable?

**Midterm report:** Due Tuesday, 2/26 at 5pm. The midterm progress report should be a maximum 4 page document that includes a clear problem statement for your project as well as preliminary results. By the time you write your midterm report, you should have a good idea of your approach.

As with the proposal, peer review is an essential part of the midterm report. By Friday, 2/22 at 6pm, you should email a draft of your midterm report to your feedback team. Peer feedback is due by Monday, 2/25 at 8am, and should cc stats311submit@gmail.com. In addition to feedback on clarity, relevance, and writing style, your feedback should begin with a one paragraph summary of the research.

Submit the midterm progress report, along with all received peer reviews, in a single stapled physical document, by Tuesday 2/26 at 5pm. You should also email the documents to stats311submit@gmail.com.

**Final report and presentation:** Monday, 3/18, 12:15pm – 3:15pm.

Your final report is a maximum six-page document that details your research. The final report should be similar to the midterm progress report, but detailing your approach. Your report should include a clear problem statement for your project, background on the area that can be understood by anyone who has taken the class, and then a description of your approach and results.

We do not require peer review for the final project, though you should feel free to use your peer feedback groups for comments before the final due date.

### 2.1 Important dates

**Tuesday, 1/31** Submit proposal

**Friday, 2/22** Draft midterm report to peer review teams

**Tuesday, 2/26** Midterm report

**Monday, 3/18** Submit final report, project presentation

Your proposal, midterm report, and final report *must* be written in \LaTeX, with appropriate and clear mathematical notation, formulae, etc. This means that mathematics is treated as part of sentences, so that equations end with appropriate punctuation.

### 3 Project ideas

Below we list a few different project ideas; this list is *by no means* exhaustive. Some of these projects may require reading ahead in the lecture notes or reading the lecture notes from last year’s offering of the course to be able to complete the project. The course staff can give more information and references on any of the projects if desired. Note that the length of a section does not necessarily correspond to anything we believe to be correlated with the quality of the resulting research; some simply took longer to write.
Recent work in theoretical computer science, among other areas, has studied adaptive inference, meaning that after performing some computation on a sample or dataset, we ask additional questions. As examples, we might consider high-dimensional inference (see, for example, the survey by Dezeure et al. [9], or papers [29, 18]). In such settings, we have a matrix $X \in \mathbb{R}^{n \times d}$, where $n \ll d$, and observe $Y = X\beta + \epsilon$, where $\epsilon$ is a mean-zero independent noise variable, and $\beta$ is sparse, meaning that $\|\beta\|_0 = k \ll d$. We may consider a two-phase procedure in which first we estimate non-zero elements of $\beta$, and then in a second (the adaptive phase) phase—after this selection—give confidence intervals or other information on $\beta$. As another example, we might collect a sample $X = \{X_1, \ldots, X_n\}$ and evaluate a few hypotheses on the sample, say that functions $\phi_1, \ldots, \phi_m$ are mean-zero. Based on the outputs, we select new functions $\tilde{\phi}_1, \ldots, \tilde{\phi}_m$ and wish to test properties of them using the sample. Of course, we have already “used” the sample; there is a natural question of what we can do to avoid over-fitting on the sample in this scenario, which has been studied by Dwork et al. [13, 14], Bassily et al. [2] and in an information-theoretic setting by Russo and Zou [28].

There are numerous questions in this area; we enumerate a few.

a. In classical statistics, asymptotic analyses (central limit theorems) give exact constants, allowing easier use in designing sampling strategies and performing tests and inference—we simply make a normal approximation. Is there a similar asymptotic theory for adaptive inference (i.e. correct constants)? (Obviously answering this in detail is hard; demonstrating interesting progress in this direction would be great.)

b. We know that adaptive algorithms may have bad worst-case behavior. What about more average-case or non-adversarial notions in interactive data analyses?

c. We discuss in the lecture notes how to use stability in KL-divergences to control first and second moments of “generalization” in adaptive analyses. On the other hand, ideas based on differential privacy (stability notions in infinite-divergence) yield stronger, high-probability guarantees [13, 14, 2]. Can we more easily get either (i) higher moment bounds or (ii) high-probability bounds using alternative divergences that are simpler, e.g., allowing composition more easily? For example, stability in different Rényi divergences may be useful.

Consider a multi-armed bandit problem, where we have $k$ different means $\mu_1, \ldots, \mu_k$, and we play a game where in each round $t = 1, 2, \ldots, T$, we play a particular arm $i_t$ and get feedback $Y_t = \mu_{i_t} + \epsilon_t$ for some noise $\epsilon_t$. Such settings occur, for example, in developing drug treatments and medical applications.

In some scenarios [23, 12], it makes sense to consider non-adaptive strategies, meaning that the strategy we use to choose arms to pull can change only infrequently. (In the two-armed case, roughly $\log \log T$ changes in strategy are sufficient to attain the minimax regret [23].) What effect does this have on the regret in these problems? How does this vary with $k$?

Developing fundamental limits in communication complexity beyond independent coordinate settings. In most results on communication complexity in estimation, there are strong assumptions about independence of “coordinates” to be estimated; see [32, 15, 5, 16]. This is natural: if I tell you ahead of time that each vector $X \in \mathbb{R}^d$ you observe has the form $X = vX_1$ for a known vector $v \in \{\pm 1\}^d$, then clearly all one must communicate is the first coordinate $X_1$. Can we build distributed estimation/learning protocols that use fewer than $O(d \log \frac{1}{\epsilon})$-bits of communication per machine while still achieving optimal (centralized) rates of convergence?
(4) Ranking and collaborative filtering from comparison models. In one version of the collaborative filtering problem (see, for example, the thesis of Oh [20]), we have \(n\) users and \(m\) items, where the matrix \(A \in \mathbb{R}^{n \times m}\) has entries \(A_{ij}\) corresponding to user \(i\)'s rating for item \(j\). When the matrix \(A\) is (nearly) low rank, it is possible to accurately reconstruct it from noisy measurements of its entries. In many situations—such as shopping data or web-search—it is easier to recover comparison information (i.e. user \(i\) preferred item \(j\) to item \(j'\)) than actual value information, so that it is of interest to recover matrices \(A\) under alternate models of observation. Investigate how low-rank recovery types of results may be extended to (noisy) comparison measurements of the matrix \(A\). See also [24, 25, 6, 19, 8].

(5) Ranking and collaborative filtering from more advanced comparison models. As in project (4), except that instead of just comparison information, sub-lists of compared items may be observed.

(6) Matrix reconstruction with different types of side information. In many ranking or collaborative filtering scenarios, we have access to side information relating objects. For example, in a music recommendation system, we have a matrix \(A \in \mathbb{R}^{n \times m}\) where \(A_{ij}\) has user \(i\)'s rating for song \(j\), while a matrix \(B \in \mathbb{R}^{m \times d}\) may consist of characteristics of the songs, so that \(B_{jk}\) is a measure of how much song \(j\) is associated with a genre \(k\). Medical and bio-informatics scenarios also suggest applications of such problems. Investigate matrix recovery under these types of models, either theoretically or by developing algorithmic schemes and applying them.

(7) In Bayesian statistics, one has a prior belief (represented by a prior distribution \(\pi\)) on a parameter \(\theta \in \Theta\), and observes a sample \(X\) drawn from a distribution \(P_\theta\) indexed by the (unknown) \(\theta\). Reference analysis (see the survey by Bernardo [4]) advocates choosing a prior \(\pi\) that “allows the data to speak for itself as much as possible,” meaning that the prior maximizes the mutual information between the sample \(X\) and the parameter \(\theta\),

\[
I_\pi(\theta; X) = \int \pi(\theta)p(x \mid \theta) \log \frac{p(x \mid \theta)}{\tilde{p}_\pi(x)},
\]

where \(\tilde{p}_\pi(x) = \int p(x \mid \theta)\pi(\theta)\).

This choice does not take into account the task at hand, however, including any loss function. Indeed, assume that \(L : \Theta \times \Theta \to \mathbb{R}_+\) is a loss function (measuring some performance of an estimate \(\hat{\theta}\) for \(\theta\)). It is possible to generalize information (as we see later in the class) by looking at the gap between the prior risk and posterior risk, that is, defining

\[
I_{L,\pi}(\theta; X) := \inf_{\hat{\theta}} \mathbb{E}_\pi[L(\hat{\theta}, \theta)] - \mathbb{E} \left[ \inf_{\hat{\theta}} \mathbb{E}[L(\hat{\theta}, \theta) \mid X] \right] \geq 0.
\]

Investigate the consequences of choosing a prior to maximize this loss-sensitive notion of information. What properties does it have? Are there interesting examples in which it is efficient to compute? Is it robust to prior mis-specification?

(8) Priors in online learning (including reinforcement learning and bandit) scenarios. A variety of recent works have studied Thompson-sampling and other Bayesian-based formulations for bandit learning [1]. Many of the convergence guarantees depend on the prior being correctly specified [26, 27]. Investigate the properties of prior mis-specification in such settings.
(9) Pedagogical project: explore Good-Turing estimation for different sequences. See, e.g., [21, 22].

(10) Learning problems in which the training distribution and testing distribution differ. Can we give optimality results or optimal procedures in this setting? What are good ways to do this? See Ben-David et al. [3], Duchi and Namkoong [10, 11].

(11) Representation learning and experimental “design.” In many machine learning problems, an important object is a function called the kernel, which measures similarity or difference between two objects (i.e. inputs $x, x'$). A theorem of Bochner is that any kernel that is only a function of the difference $x - x'$ can be written as the characteristic function of a random vector $W$ with (some) distribution $P$:

$$k(x, x') = \mathbb{E}_P[e^{iW^\top(x - x')}].$$

Given a sample, can we efficiently learn a distribution $P$ that generates a kernel with good performance?

(12) Distribution property testing and distribution functional estimation: in theoretical computer science and information theory, a recent body of work (see, among many other papers, Jiao et al. [17], Chan et al. [7], and Valiant and Valiant [30]) investigates estimation of quantities such as entropy and testing whether two distributions are close or far apart using a variety of tools. Give a general characterization of these testing (or other) results based on information theoretic quantities such as metric entropy (cf. [31]).

References


