Statistics 311/Electrical Engineering 377
Statistics and Information Theory
Course information and syllabus

Instructor: John Duchi (jduchi@stanford.edu)
Lectures: Tuesday/Thursday, 13:30 – 14:50, 200-205
TA: Hongseok Namkoong (hnank@stanford.edu)
Office hours: John: Tuesday/Thursday 15:00 - 16:00 in 126 Sequoia Hall, Hong: Wednesday 13:15 - 15:00 Packard 106.

Description: Information theory was developed to solve fundamental problems in the theory of communications, but its connections to statistical estimation and inference date nearly to the birth of the field. With their focus on fundamental limits, information theoretic techniques have provided deep insights into optimal procedures for a variety of inferential tasks. In addition, the basic quantities of information theory—entropy and relative entropy and their generalizations to other divergence measures such as $f$-divergences—are central in many areas of mathematical statistics and probability.

The application of these tools are numerous. In mathematical statistics, for example, they allow characterization of optimal error probabilities in hypothesis testing, determination of minimax rates of convergence for estimation problems, demonstration of equivalence between (ostensibly) different estimation problems, and lead to penalized estimators and the minimum description length principle. In probability, they provide insights into the central limit theorem, large deviations theory (via Sanov’s theorem and other results), and appear in empirical process theory and concentration of measure. Information theoretic techniques also arise in game playing, gambling, stochastic optimization and approximation, among other areas.

In this course, we will study information theoretic quantities, and their connections to estimation and statistics, in some depth, showing applications to many of the areas above. Except to provide background, we will not cover standard information-theoretic topics such as source-coding or channel-coding, focusing on the probabilistic and statistical consequences of information theory.

Prerequisites: One of EE376A (or equivalent) or Statistics 300A (or equivalent).

Grading and Evaluation: There will be a problem set approximately every two weeks; the problem sets will count for $\frac{2}{3}$ of the grade.

Students taking the course for a grade will complete a final project for the course; this is not required for students taking the class pass/fail (though students taking the class pass/fail may do a project if they desire). More information about the course project will be posted in the first few weeks of class, but roughly the project should correspond to one of the following:

1. Research project: original research related to the content of the course, and at the end of the course the student(s) will present a brief writeup to the course staff detailing their work. Projects here will likely turn into publishable work.

2. Pedagogical project: develop a set of problems 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 extracting exercises from a few papers around topics in the class.
Course outline (subject to change)

I. Basic information theory review (one week): divergence measures, connections with testing, basic coding, notions of entropy.

II. Minimax rates and estimation (two to three weeks): reductions from estimation to testing, the Le Cam, Fano, and Assouad methods for lower bounding minimax risk. Examples in regression and high-dimensional statistics, convex optimization, and non-parametric estimation. Optimal hypothesis testing and lower bounds, Chernoff-Stein theorem.

III. Concentration inequalities and large deviations (two weeks): concentration inequalities, sub-Gaussian random variables, the entropy method for concentration, applications in coding and statistical learning theory.

IV. Experiments and exploration (two weeks): multi-armed bandit problems, game playing, reinforcement learning, exploration/exploitation tradeoffs, information-directed sampling, randomized experiments, adaptive testing.

V. Surrogate risks and \( f \)-divergences (two weeks): connections between minimization of convex and non-convex losses, quantization and distributed detection, bounds and distance measures between distributions.

VI. Universal prediction and minimum description length (if time remains): universal coding and prediction, minimax redundancy and regret, penalization methods and model selection.

Textbooks  There will be no textbook for the course. The following, however, may contain useful reference material for different parts of the course. Starred items are available online on the Stanford network.