PhD course: average case complexity and statistical inference

This course will introduce some theory of, and cover some recent results in computational complexity of statistical inference problems - detecting/recovering signals in noisy data.

First lecture: Thursday 26th Jan 15:15-17.00 in room 64119.

Details: The course is 5hp and will run throughout the spring term to June. Assessment will take the form of 2 exercise sheets and 1 longer piece to focus on tractability/hardness/impossibility of a particular problem not covered in the lectures or possibly a novel application of our methods assessed via a written report or talk.

Lecturer: Fiona Skerman, `fiona.skerman@math.uu.se`. Feel free to contact me with any questions about the course.

Description:

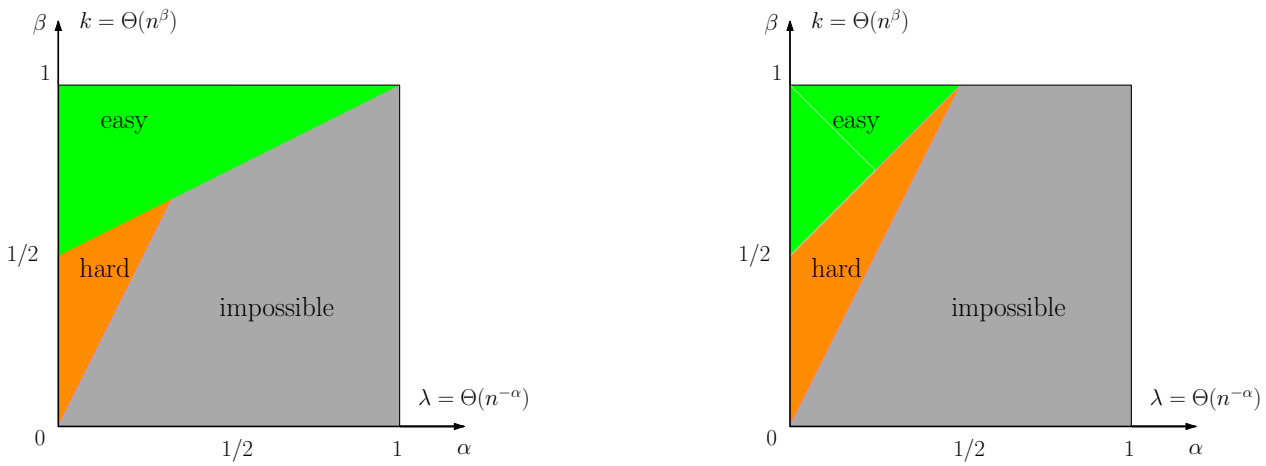
Statistical inference to us is detecting or recovering a signal in noisy data. An example is finding a submatrix with entries normally distributed with mean λ and variance one in a matrix with all other entries normally distributed with mean zero and variance one. For some signal strengths λ and some sizes of submatrix fast algorithms are known which succeed with probability near 1; for some regions of parameters it is information theoretically impossible to succeed with probability near 1.

We are interested also in the third region - parameter values for which there are brute-force algorithms which succeed but no fast algorithms are known. In particular there is a regime of parameter values where finding the submatrix is conjectured hard: and we are interested in rigorous results which give evidence of hardness i.e. showing failure of restricted classes of algorithms and showing average-case reductions to problems we believe are hard.

The course will use two running example problems, finding / detecting a submatrix with elevated mean in a large random matrix, and finding / detecting a dense subgraph within a large random graph; and will establish the phase transition diagrams for these problems (see below). These exhibit a detection-recovery gap, it is easier to detect the presence of the planted substructure than to recover it even approximately. We will also cover definitions, ideas and techniques necessary to establish phase-transitions of hardness for statistical inference problems: including analysis of algorithms on random structures: spectral techniques, SDPs & brute-force, bounding chi-square divergence, low-degree polynomial method, average-case reductions. Note the probabilistic aspect means one has to be careful what a reduction is (it is allowed to fail on some instances for example) and the proofs have different techniques.

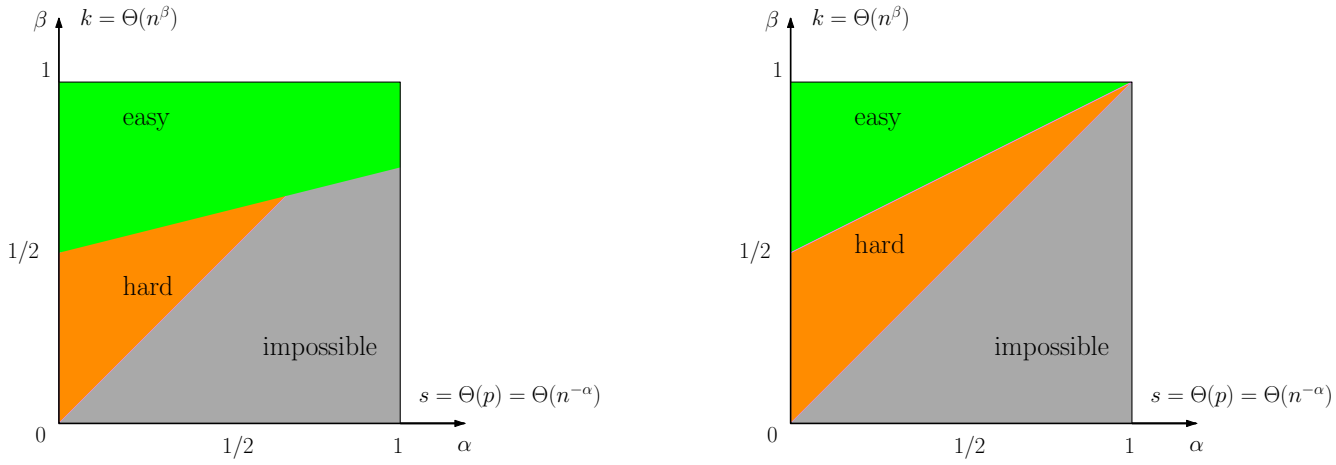
The following phase transition diagrams show parameter regimes where the example problems are **easy** (fast algorithms succeed with probability near 1), **hard** (brute-force/slow algorithms succeed with probability near 1 and evidence no such fast algorithms exists) and **impossible** (information theoretically impossible for any algorithm to succeed with probability near 1). In each diagram the size $\sim k(n)$ of the planted structure increases along the y-axis and the strength of the signal of the planted substructure decreases along the x-axis. The **detection problem** is, given a (random) sample from either H_0, a distribution with no planted substructure or H_1, a distribution with a planted substructure to determine which distribution it likely came from. The **recovery problem**

is, given a (random) sample from H_1 a distribution with a planted substructure, output (exactly or approximately) the location of the planted substructure.



(a) detection (b) recovery

Figure 1: **Spiked Matrix Model** (planted submatrix with elevated mean).
H_0: random $n \times n$ matrix with each entry independent with distribution $N(0,1)$.
H_1: $n \times n$ matrix with each index in set S independently with probability k/n. Each entry independent with distribution $N(\lambda,1)$ if $i,j \in S$ and with distribution $N(0,1)$ otherwise.



(a) detection (b) recovery

Figure 2: **Planted dense subgraph**.
H_0: $G(n,q)$ random graph on n vertices where each edge is present independently with probability q.
H_1: $G(n,k,q,s)$ with $s > 0$, random graph on n vertices where each vertex is part of 'community' S independently with probability k/n. Each edge ij is present independently either with probability $q+s$ if $i,j \in S$ or with probability q otherwise.

Further particulars The course will comprise ∼15 lectures and ∼5 problems sessions. The assessment, all of which can be done in small groups (up to 2-3), will be exercise sheets (2×25%) and 1 longer project (50%). The first exercise sheet will be out Friday 3rd and due Monday 21st February, the second will be out Friday 24th March and due 17th April.

For the longer project is to understand the proof of tractability, hardness or impossibility of a particular problem. List of suggestions will be provided (by 21st April) including some reductions in total variation from a paper by Brennan and Breser, spectral method to achieve the threshold in stochastic block from a paper by Lelarge, Bordenave and Massoulié as well as some candidate lemmas which together will prove some new results (probably a new testing problem where both H_0 and H_1 consist of different planted structures instead of planted and null: with lemmas to prove low-deg hardness, find fast algorithms, info-theoretic thresholds). Hand in either ∼5-10 pages give or 25 minutes talk each person end of May / early June.

Dates (provisional) Lectures and problem sessions all in 64119 unless otherwise indicated, and will start 15min past the hour.

L1 Thu 26th Jan 3-5pm
L2 Wed 1st Feb 3-5pm
L3 Thur 9th Feb 3-5pm
L4 Wed 15th Feb 3-5pm
L5 Wed 22nd Feb 3-5pm
L6 Wed 1st Mar 3-5pm
L7 Wed 8th Mar 3-5pm