15-496/859X: Computer Science Theory for the Information Age — Spring 2012
Carnegie Mellon University — V. Guruswami & R. Kannan

# Take Home Final

Due by 6pm, Friday, May 4

---

## Instructions

- You should work on the problems and the write-up **entirely by yourself.** *No collaboration allowed.*
- Reference to any external material besides the course text, problem set solutions, and material covered in lecture is **not allowed.** In particular, you are not allowed to search for answers or hints on the web. You are encouraged to contact the instructors or the TA if you feel stuck or are otherwise unsure about the meaning of a problem statement.
- Solutions typeset in LaTeX are preferred.
- You are urged to start work on the final early; no extensions will be granted on the final.
- Each problem is worth 10 points unless indicated otherwise.

---

1. (15 points)

   (a) True or False: For every matrix $A$, there exists a matrix $B$ of rank at most $k$ such that $\|A-B\|_F \leq \frac{\|A\|_F}{\sqrt{k}}$ ? Justify your answer.

   (b) True or False: For every matrix $A$, there exists a matrix $B$ of rank at most $k$ such that $\|A-B\|_2 \leq \frac{\|A\|_F}{\sqrt{k}}$ ? Justify your answer.

   (c) Suppose an $n \times d$ matrix $A$ is given and you are allowed to pre-process it. You are then given $m$ vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m} \in \mathbb{R}^d$ and your task is to compute the vector $A\mathbf{x_i}$ approximately for each $i$, in the sense that you must find a vector $\mathbf{u_i} \in \mathbb{R}^n$ such that $\|\mathbf{u_i} - A\mathbf{x_i}\|_2 \leq \epsilon \|A\|_F \|\mathbf{x_i}\|_2$ for some error bound $\epsilon > 0$. Describe a method to accomplish this in time $O\left(\frac{d+n}{\epsilon^2}\right)$ per $\mathbf{x_i}$ not counting the preprocessing time.

2. Consider the clique $K_n$ with vertex set $\{1, 2, \ldots, n\}$.

   (a) What is the hitting time $h_{1,2}$?

   (b) Suppose each edge of the clique is removed independently with probability $p$. What is the probability that the nodes 1 and 2 have distance 3 or more (i.e., every path connecting them has at least 3 edges)?

3. (5 points) For $X \sim N(0,1)$, what is $\mathbb{E}[|X|]$?

4. In lecture, we proved that for a set system with VC dimension $d$, picking a random sample of $m = \Omega\left(\frac{d}{\epsilon}\log\frac{d}{\epsilon}\right)$ examples implies that the probability of missing any set $S$ with probability mass at least $\epsilon$ is atmost $\exp(-\Omega(\epsilon m))$. Prove that in fact a linear in $d$ number of examples suffices, specifically that for

$$m \geq \Omega\left(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{d}{\epsilon}\log\frac{1}{\epsilon}\right) ,$$

   the probability that there exists a set $S$ with probability mass at least $\epsilon$ that is missed by the sample is at most $\delta$.

5. Suppose someone hides a clique (a subset of nodes which are all pairwise adjacent) of size $k \geq 10\sqrt{n \ln n}$ in a $G(n, 1/2)$ random graph. That is, someone chooses an arbitrary subset $S$ of $k$ vertices and puts in the missing edges to make $S$ a clique. You are given this modified graph and asked to identify $S$. Given an algorithm running in $O(n^2)$ time for this task that correctly finds $S$ with high probability (over the choice of the $G(n, 1/2)$ random graph).

   Hint: How do the degrees of nodes in $S$ compare to the degrees of the rest of the nodes?

6. Suppose you plan to invest in a stock based on predictions of $n$ websites, each of which makes a Boolean up/down prediction on the stock for the next day. Your bottomline depends on correctly guessing whether the market will go up/down the next day. Naturally, therefore, you would like to minimize the number of days you guess wrongly by devising a judicious strategy based on the history of predictions and actual up/down truth revealed to you each day.

   (a) First, assume that at least one of the $n$ websites (you don't know which one!) is always right. Give a strategy that guarantees making incorrect guesses on at most $O(\log n)$ days.

   (b) Now suppose you invest for $D$ days, and the best of the $n$ websites is wrong on at most $m$ days (again, you do *not* which website has this guarantee). Give a strategy that is guaranteed to make at most $O(m + \log n)$ incorrect guesses. Thus it is possible to do not much worse than the best website in hindsight.

   Hint: Weight the predictions of the websites and use a multiplicative update rule to penalize mistakes in each round.