COMS-4771 Spring 2008
Paper 4 Quiz Name: ______________________________

Dasgupta and Gupta. *An elementary proof of a theorem of Johnson and Lindenstrauss.*

You have 30 minutes to complete the questions. The quiz is worth 10 points, so you can choose either one of the two problems below.

Question 1 (10 points): The Johnson-Lindenstrauss Lemma says that for any $\epsilon > 0$, any set of n points in a Euclidean space can be embedded into a Euclidean space of dimension $k = c(\epsilon)\log n$ with distortion at most ϵ, where $c(\epsilon) \le O(1/\epsilon^2)$. Show that this is nearly tight, i.e., $c(\epsilon)$ must be at least $\Omega(1/\epsilon^2 \log(1/\epsilon))$.

You can use the following theorem:

Theorem (Alon): Let B be an $n \times n$ symmetric real-valued matrix with $b_{ii} = 1$ and $|b_{ij}| \le \epsilon$ for all $i \ne j$. If the rank of B is d and $\frac{1}{\sqrt{n}} \le \epsilon \le 1/2$, then

$$d \ge \Omega\left(\frac{1}{\epsilon^2 \log(1/\epsilon)} \log n\right).$$

Solution: If we have $n+1$ points in R^k, and the distance between any pair of distinct points in the set is between $1-\epsilon$ and $1+\epsilon$ (think of embedding the unit $(n+1)$-cube), we can put one of the points at the origin and shift the other n points by at most $O(\epsilon)$ making sure that their distance from this origin is exactly 1. By the triangle inequality, the distance between any pair of the shifted points is still $1 + O(\epsilon)$. Denote point i by v_i and consider an $n \times n$ matrix A with $a_{ij} = v_i \cdot v_j$. All diagonal entries are 1. All other entries are $1/2 + O(\epsilon)$ (since $v_i \cdot v_j \approx \|v_i\|\|v_j\|\cos(60°) = 1/2 + O(\epsilon)$). Since the points are in R^k, the rank is at most k. To apply Alon's theorem above, consider the matrix $B = 2A - I$, where I is the $n \times n$ matrix with every entry being 1 (so that B satisfies the conditions of the theorem). The rank of B is at most $k + 1 = O(\log n)$. According to the theorem, this rank is at least $\Omega\left(\frac{1}{\epsilon^2 \log(1/\epsilon)} \log n\right)$, implying the theorem.

Question 2 (10 points): For algorithmic applications of the Johnson-Lindenstrauss lemma, it is important to be able to generate and evaluate the random linear mapping fast. So the quest has been to create sparse mappings. Consider the following $k \times n$ sparse random matrix where the fraction of nonzero entries tends to 0:

The entries of are independent random variables and each of them attains value 0 with probability $1-q$, and a value drawn from the normal distribution with zero mean and variance $1/q$ with probability q, for $q = O(1/n)$.

What is the problem with using such a matrix directly?

Answer: Call the map defined by the matrix $f : \mathrm{R}^n \to \mathrm{R}^k$. Recall that we need

$$\mathbf{Pr}[(1-\epsilon)\|x\| \le \|f(x)\| \le (1+\epsilon)\|x\|] \ge 1 - \frac{1}{n^2}.$$

The problem is that the length of the image $\|f(x)\|$ is not sufficiently concentrated for some vectors, for example, for $x = (1, 0, \ldots, 0)$. The vector x has to be well-spread in the sense that $\max_j x_j$ is close to $1/\sqrt{n}$. See Ailon and Chazelle (Approximating nearest neighbors and the fast JL transform, STOC 2006) for a clever way of dealing with this obstacle.