CS 225: Pseudorandomness **Prof. Salil Vadhan**

Problem Set 2

Assigned: Feb. 20, 2007 Due: Mar. 7, 2007 (1 PM)

- Recall that your problem set solutions must be typed. You can email your solutions to `cs225-hw@eecs.harvard.edu`, or turn in it to Carol Harlow in MD 343. You may write formulas or diagrams by hand. Aim for clarity and conciseness in your solutions, emphasizing the main ideas over low-level details.
- If you use LaTeX, please submit both the source (`.tex`) and the compiled file (`.ps` or `.pdf`). Name your files `PS2-yourlastname`.
- Starred problems are extra credit.

Problem 1. (Spectral Graph Theory) Let M be the random-walk matrix for a d-regular undirected graph $G = (V, E)$ on n vertices. We allow G to have self-loops and multiple edges. Recall that the uniform distribution (or all-ones vector) is an eigenvector of M of eigenvalue $\lambda_1 = 1$. Prove the following statements. (Hint: for intuition, it may help to think about what the statements mean for the behavior of the random walk on G.)

(a). All eigenvalues of M have absolute value at most 1.

(b). G is disconnected $\iff$ 1 is an eigenvalue of multiplicity at least 2.

(c). Suppose G is connected. Then G is bipartite $\iff$ -1 is an eigenvalue of M.

(d). G connected $\Rightarrow$ all eigenvalues of M other than λ_1 are $\leq 1 - 1/\mathrm{poly}(n, d)$. To do this, it may help to first show that the second largest eigenvalue of M (not necessarily in absolute value) equals

$$\max_x \langle Ax, x \rangle = 1 - \frac{1}{d} \cdot \min_x \sum_{(i,j) \in E} (x_i - x_j)^2,$$

where the maximum/minimum is taken over all vectors x of length 1 such that $\sum_i x_i = 0$, and $\langle x, y \rangle = \sum_i x_i y_i$ is the standard inner product. For intuition, consider restricting the above maximum/minimum to $x \in \{+\alpha, -\beta\}^n$ for $\alpha, \beta > 0$.

(e). G connected and nonbipartite $\Rightarrow$ all eigenvalues of M (other than 1) have absolute value at most $1 - 1/\mathrm{poly}(n, d)$.

(f*) Extra credit: Establish the (tight) bound $1 - \Omega(1/d \cdot D \cdot n)$ in Part (d), where D is the diameter of the graph, and show that a simple graph satisfies $D \leq O(n/d)$. (The $1 - \Omega(1/d \cdot D \cdot n)$ bound also holds for Part (e), even though I'm not asking you to prove it.)

Problem 2. (Derandomizing RP versus BPP) Show that $\mathbf{prRP} = \mathbf{prP} \Rightarrow \mathbf{prBPP} = \mathbf{prP}$ (and in particular, $\mathbf{BPP} = \mathbf{P}$). (Hints: Look at the proof that $\mathbf{NP} = \mathbf{P} \Rightarrow \mathbf{BPP} = \mathbf{P}$.)

Problem 3. (Designs) Designs (also known as packings) are collections of sets which are nearly disjoint. Later in the course, we will see how they are useful in the construction of pseudorandom generators. Formally, a collection of sets $S_1, S_2, \ldots, S_m \subseteq [d]$[1] is called an (ℓ, a)-*design* if

- For all i, $|S_i| = \ell$.
- For all $i \neq j$, $|S_i \cap S_j| < a$.

For given ℓ, we'd like m to be large, a to be small, and d to be small. That is, we'd like to pack many sets into a small universe with small intersections.

(a). Prove that if $m < \binom{d}{a} / \binom{\ell}{a}^2$, then there exists an (ℓ, a)-design $S_1, \ldots, S_m \subseteq [d]$. Guideline: Use the Probabilistic Method. Specifically, show that if the sets are chosen randomly, then

$$\mathop{\mathrm{E}}_{S_i}\left[\#\{j < i : |S_i \cap S_j| \geq a\}\right] < 1.$$

(b). Conclude that for every $\epsilon > 0$, there is a constant c_ϵ such that for all ℓ, there is a design with $a \leq \epsilon\ell$, $m \geq 2^{\epsilon\ell}$, and $d \leq c_\epsilon \ell$. That is, in a universe of size $O(\ell)$, we can fit *exponentially many* sets of size ℓ whose intersections are an arbitrarily small constant fraction of ℓ.

(c). Using the Method of Conditional Expectations, show how to construct a design as in Part (a) *deterministically* in time $\mathrm{poly}(m, d)$.

Problem 4. (Frequency Moments of Data Streams) Given one pass through a huge 'stream' of data items $(a_1, a_2, \ldots, a_k)$, where each $a_i \in \{0,1\}^n$, we want to compute statistics on the distribution of items occurring in the stream while using small space (not enough to store all the items or maintain a histogram). In this problem, you will see how to compute the *2nd frequency moment* $f_2 = \sum_a m_a^2$, where $m_a = \#\{i : a_i = a\}$.

The algorithm works as follows: Before receiving any items, it chooses t random *4-wise* independent hash function $h_1, \ldots, h_t : \{0,1\}^n \to \{+1, -1\}$, and sets counters $X_1 = X_2 = \cdots = X_t = 0$. Upon receiving the i'th item a_i, it adds $h_j(a_i)$ to counter X_j. At the end of the stream, it outputs $Y = (X_1^2 + \cdots + X_t^2)/t$.

Notice that the algorithm only needs space $O(t \cdot n)$ to store the hash functions h_j and space $O(t \cdot \log k)$ to maintain the counters X_j (compared to space $k \cdot n$ to store the entire stream, and space $2^n \cdot \log k$ to maintain a histogram).

(a). Show that for every data stream $(a_1, \ldots, a_k)$ and each j, we have $\mathrm{E}[X_j^2] = f_2$, where the expectation is over the choice of the hash function h_j.

(b). Show that $\mathrm{Var}[X_j^2] \leq 2f_2^2$.

(c). Conclude that for a sufficiently large constant t (independent of n and k), the output Y is within 1% of f_2 with probability at least .99.

[1] The notation $[n]$ is shorthand for the set $\{1, 2, \ldots, n\}$.