Randomized Algorithms (Blum/Gupta) **Final Exam**
Date: X, 2011 **Due:** $X + 48$ hours, 2011

Groundrules

- There are 9 problems. Do any 7 of your choice.
- Work by yourself. You are free to use lecture notes, your own notes, or the textbook.

Problems

1. **(Mind the Integrality Gap.)** In class we showed that the naive LP for max-cut has an integrality gap of $1/2 + o(1)$. So let's consider the following LP:

$$\max \sum_{(i,j)\in E} z_{ij}$$
$$\sum_{(i,j)\in C} z_{ij} \leq |C| - 1 \qquad \forall \text{ odd cycles } C \subseteq E.$$

   This captures the fact that not all edges in an odd cycle can be cut by a solution. We will now show there exist graphs with an integrality gap of close to $1/2$, even for this LP: the optimal value of this LP will be close to m (the number of edges) whereas the optimal max-cut will be close to $m/2$. The idea will be to use graphs with only long cycles, so that we can set $z_{ij} \approx 1$ without violating the odd cycle inequalities.

   Consider the random graph $G(n, c/n)$ for some constant $c = 100/\varepsilon^2$, and $n \gg c/\varepsilon$.

   (a) Let $m = \frac{c(n-1)}{2}$. Show that $G(n, c/n)$ has at least $m(1-\varepsilon)$ edges whp.

   (b) In HW#5 we showed that removing a small number of edges (say n of them) from this graph gives us a graph H with only cycles of length at least $g = \frac{1}{2}\log_c n$. Show that setting $z_{ij} = \frac{g-1}{g}$ for each edge $\{i, j\} \in H$ gives us a feasible solution to the LP. Show that the value of this LP solution is at least $m(1 - O(\varepsilon))$ whp.

   (c) Show the probability that the number of edges in $G(n, c/n)$ crossing a fixed cut $(S, V \setminus S)$ exceeds $(1+\varepsilon)\frac{m}{2}$ is less than 2^{-n}.

   (d) Using the previous parts, show that the maxcut in H is $(1 + O(\varepsilon))\frac{m}{2}$ with high probability.

2. **(Linear Equalities.)** Given a set of m linear equalities over n variables mod p (e.g., if $p = 3$ we might have $2x_1 + x_2 + 2x_5 \equiv 1$ and $x_1 + x_3 \equiv 2$) it is easy to see that a random assignment to the variables will in expectation satisfy a $1/p$ fraction of them. (Here, p is prime and each equality has at least one variable in it.)

   Give a deterministic algorithm to find an assignment that satisfies at least a $1/p$ fraction of the equalities, and runs in poly(n, m) time.

3. **(A Matching Made in Heaven.)** Show that given an oracle for estimating the number of matchings in a bipartite graph (up to multiplicative $1 \pm 1/poly(n)$) one can produce a near-uniform-random generator of matchings in a bipartite graph (up to multiplicative $1 \pm 1/poly(n)$).

4. **(Dominating Sets.)** Given a graph $G = (V, E)$ with $|V| = n$, a *dominating set* for G is a subset $D \subseteq V$ such that each vertex $v \in V$ is either in D or has a neighbor in D.

   (a) Show that any graph with minimum degree δ has a dominating set of size at most $K = \frac{n \ln n}{\delta+1}$. (Hint: pick a random set of vertices.)

(b) Now improve the bound and show the existence of a dominating set of size at most $K' := \frac{n(1+\ln(1+\delta))}{1+\delta}$. (Hint: pick a smaller random set of vertices, and then add some more vertices as needed.)

5. **(Serious Discrepancies.)** Consider a collection of m sets $\mathcal{F} = \{S_1, S_2, \ldots, S_m\}$, each $S_i \subseteq U$ with $|U| = n$. Given an assignment $\chi : U \to \{1, -1\}$, the *discrepancy* of a set $S \subseteq U$ is

$$\mathsf{disc}_\chi(S) = |\sum_{x \in S} \chi(x)|.$$

(If you think of χ as coloring U by red or blue, then $\mathsf{disc}_\chi(S)$ is the the magnitude of the difference between the number of reds and blues in the set S.) The discrepancy of the entire collection $\mathcal{F}$ with respect to coloring χ is maximum discrepancy of any set in $\mathcal{F}$, i.e.,

$$\mathsf{disc}_\chi(\mathcal{F}) = \max_{S_i \in \mathcal{F}} \mathsf{disc}_\chi(S_i) = \max_{S_i \in \mathcal{F}} \Big| \sum_{x \in S_i} \chi(x) \Big|$$

(a) Show that, for any such set system $(U, \mathcal{F})$, a random coloring $\chi : U \to \{-1, 1\}$ gives a discrepancy of at most $\sqrt{3n \ln(2m/\delta)}$ with probability $1 - \delta$.

Using a different probabilistic argument, we can also give a lower bound: we show there exists a family $\mathcal{F}$ of n subsets of U such that *for every coloring* χ, the discrepancy $\mathsf{disc}_\chi(\mathcal{F})$ is at least $\sqrt{n}/c$ for some constant $c \geq 1$.

(b) Fix some assignment $\chi : U \to \{-1, 1\}$. Pick a random subset $A \subseteq U$ by including each element of U in A independently with prob $\frac{1}{2}$. Show that for some constant c and this fixed assignment χ:

$$\Pr_A \left[\mathsf{disc}_\chi(A) > \sqrt{n}/c \right] > 1/2.$$

(c) Let $\mathcal{F}$ consist of n sets picked independently as above. Infer that for any fixed assignment $\chi : U \to \{-1, 1\}$, $\mathsf{disc}_\chi(\mathcal{F})$ exceeds $\sqrt{n}/c$ with probability strictly greater than $1 - (1/2)^n$.

(d) Take a union bound over all 2^n assignments to show the existence of a family $\mathcal{F}$ with n sets for which *all assignments* $\chi : U \to \{-1, 1\}$ have discrepancy $> \sqrt{n}/c$.

6. **(Rewriting Expectations.)** Take n cards numbered $\{1, 2, \ldots, n\}$ and consider them in a perfectly random order, with each of the $n!$ orders being equally likely. Every time you see a card number larger than the previous ones, you mark the card. (n.b. you mark the first card.)

(a) If X denotes the number of marked cards, what is $E[X]$?

(b) Using, say, a Chernoff bound, show concentration: for $\varepsilon \in (0, 1)$, show

$$\Pr[X > (1 + \varepsilon)E[X]] \leq \frac{1}{n^{\Omega(\varepsilon^2)}}.$$

(BTW, be sure to show *independence* for the random variables if you use a Chernoff bound.)

7. **(Hooking up.)** Consider a random walk on a line of length n (at each step flip a coin and go left if heads and right if tails) where the rule at the endpoints is that if your coin flip tells you to go off the end of the line, then you stay where you are. Let us say that the walk begins at the leftmost point. The following is a "coupling" argument to show that this walk is rapidly mixing.

(a) Prove that the stationary distribution is uniform (equivalently that the uniform distribution is stationary).

Now, imagine that at the same time we are doing our random walk, a second "virtual particle", which begins at a *random* starting location is also doing a random walk, and using the *same* coin flips as we are (i.e., when we get a heads, it gets a heads, and when we get a tails, it gets a tails). Notice that if we and the virtual particle ever meet, then we will continue to follow exactly the same trajectory from then on (i.e., we will have "coupled").

(b) Prove that for sufficiently large constant c, after $t = cn^2 \log n$ coin flips, with probability $1 - 1/n$ we *will* have met the virtual particle.

(c) Argue that this implies the position at time step t is approximately uniform. In particular, given that with probability at least $1 - \epsilon$ we have coupled by time t, what can you say about $\sum_{i=1}^{n} |p_i - 1/n|$, where p_i is our probability of being at location i at time t?

(d) Suppose we were watching the above process and stopped it right when the two particles met. Are all points on the line equally likely to be the meeting point? Why or why not?

8. **(So Much in Common.)** A string s is a *subsequence* of string a if the letters of s occur in a in the same order (but perhaps not consecutively). Given two n-bit strings a and b, the *longest common subsequence* $\mathsf{lcs}(a, b)$ is the longest string s such that s is a subsequence of both a and b.

Suppose a and b are uniformly random n-bit strings drawn from $\{0,1\}^n$: it is easy to show that the length of $\mathsf{lcs}(a, b)$ is $\Theta(n)$ whp.; one can get pretty precise estimates, but we're not asking for this.

Show that the length of $\mathsf{lcs}(a, b)$ is tightly concentrated around its mean.

9. **(Hashing.)** Consider the setup of HW#5 problem 2 again: elements from $[D]$ stream by, x_p is the number of times element p was seen, and given a query $q \in [D]$, we want to return $\hat{x}_q$ such that $\hat{x}_q \in x_q \pm \varepsilon \|x\|_1$ with probability at least $1 - \delta$. (n.b.: $\|x\|_1 = \sum_i |x_i|$.)

The algorithm is this:

> Keep t hash functions $h_1, h_2, \ldots, h_t : [D] \to [d]$, and td counters C_{ij}. When you see element $e \in [D]$, increment the t counters $C_{1,h_1(e)}, C_{2,h_2(e)}, \ldots, C_{t\,h_t(e)}$. On query q, return
>
> $$\hat{x}_q := \min_i C_{i,h_i(q)}.$$

(a) Observe that $x_q \le \hat{x}_q$.

(b) Suppose we condition on $h_i(q) = j$, and consider counter C_{ij}. If each hash function $h_i()$ is pairwise independent, what is the probability that $h_i(e) = j$ for $e \neq q$? Show that the expected value of the "overshot" $C_{ij} - x_q$ is $\frac{1}{d} \sum_{e \neq q} x_e \le \frac{1}{d} \|x\|_1$.

(c) Suppose $d = 2/\varepsilon$, show that the probability that $C_{i,h_i(q)} > x_q + \varepsilon \|x\|_1$ is at most $1/2$.

(d) Suppose $t = \log_2(1/\delta)$ and the hash functions are all independent of each other, show that the probability that $\hat{x}_q > x_q + \varepsilon \|x\|_1$ is at most δ.

Hence, using $O(\varepsilon^{-1} \log \delta^{-1})$ space, we can estimate any one frequency value to within an additive error of $\varepsilon \|x\|_1$. Note the space usage is better compared to HW#5 problem 2 (number of counters is $O(\varepsilon^{-1} \log \delta^{-1})$ instead of $O(\varepsilon^{-2} \log \delta^{-1})$), but with a weaker ℓ_1 guarantee and not ℓ_2.