Exercise 1 (Due: 15.11.2009)

Grading: Every question states its maximum grade. The sum of the question grades throughout the semester forms the basis on which the final course grade is calculated, so higher maximum grades imply higher weights.

Playing with vectors (3 points)

Given a probability distribution P over $\{1, \ldots, n\}$, such that p_i denotes the probability of the value i, let us denote the l_2 norm by $\|P\| = \sum_{i=1}^{n} p_i^2$. Show that if P is ϵ-far in the variation distance from the uniform distribution (see Assignment Zero for the definition of this distance), then $\|P\| \geq (1 + 4\epsilon^2)\frac{1}{n}$.

Comment: This result will be revisited in a future assignment.

Trying to ascertain uniformity (6 points)

Suppose we run an algorithm that can take independent samples from $\{1, \ldots, n\}$, each chosen according to the distribution P (which the algorithm does not know). The algorithm can compare which of the taken samples are equal to which, but you may assume that the algorithm cannot do any other operation on the values of the samples. The algorithm should say "yes" with probability $\frac{2}{3}$ if P is uniform, and should say "no" with probability $\frac{2}{3}$ if P is $\frac{1}{4}$-far from being uniform. Show that this cannot be done with $\frac{1}{10}\sqrt{n}$ samples.

Quicksort exposed (6 points)

A reminder: The quicksort algorithm for n elements $a_1, \ldots, a_n$ (assuming that all elements are distinct) first picks uniformly at random an index $1 \leq i \leq n$, compares a_i with all other elements, and then recursively runs the over the set of all elements smaller than a_i and the set of all elements larger than a_i. Show that the expected total number of comparisons is $2n(\ln(n) + \Theta(1))$.

Hint: Assuming that $s_1, \ldots, s_n$ are those elements in sorted order, define for $i < j$ the indicator variable $X_{i,j}$ for the event that a comparison took place between s_i and s_j.

Troubles moving forward (6 points)

We let $X_0 = 0$, and then for every $i > 0$ (going from 1 forward), with probability $\frac{1}{3}$ we set $X_i = X_{i-1} + 1$ and with probability $\frac{2}{3}$ we set $X_i = \max\{0, X_{i-1} - 1\}$, independently of previous choices. Let t_k be the expectation of the distance between the smallest i for which $X_i = k - 1$ and the smallest j for which $X_j = k$. For example, it is not hard to see that $t_1 = \sum_{r=1}^{\infty} \frac{1}{3}(\frac{2}{3})^{r-1} r = \sum_{r=1}^{\infty} \sum_{s=1}^{r} \frac{1}{3}(\frac{2}{3})^{r-1} = \sum_{s=1}^{\infty} \sum_{r=s}^{\infty} \frac{1}{3}(\frac{2}{3})^{r-1} = \sum_{s=1}^{\infty} (\frac{2}{3})^{s-1} = 3$ (yes, there may be a more elegant way to calculate this sum). Calculate t_k.

Exercise 2 (Due: 20.12.2009)

Three-sum-free (6 points)

Show the following: If A is a set of positive integers, then there exists a set $B \subseteq A$ with $|B| \geq \frac{1}{4}|A|$ so that B contains no three members (distinct or otherwise) whose sum is also in B. You can get most but not all of the points for this question by finding a set B with $|B| \geq (\frac{1}{4} - o(1))|A|$.

Note: There is in the lecture notes a similar proof for finding a set B with no two members whose sum is also in B. Its ideas are useful here.

Ascertaining uniformity (12 points)

Suppose again we are able to take independent samples from $\{1, \ldots, n\}$, again each chosen according to an unknown distribution P (see "Trying to ascentain uniformity" from Exercise 1). Now we do the following: For $k = \lceil 100\sqrt{n}/\epsilon^4 \rceil$, we take k samples, and denote their results by $a_1, \ldots, a_k$, and now count the number of $1 \leq i < j \leq k$ for which $a_i = a_j$. If this number is at most $(1 + 2\epsilon^2)\frac{1}{n} \cdot \binom{k}{2}$ then we say "yes", and otherwise we say "no". Show if P is uniform then the algorithm says "yes" with probability at least $\frac{2}{3}$, while if P is ϵ-far from uniform in variation distance then the algorithm says "no" with probability at least $\frac{2}{3}$.

Hint: Use (among other things) the result of "Playing with vectors" from Exercise 1.

Exercise 3 (Due: 17.1.2010)

It is a good idea to start thinking about this exercise well before the deadline. Both questions require more than just knowing the techniques.

Editing strings (12 points)

For a fixed k and a binary string $v = v_1 v_2 \dots v_n$ with size n, we define two "statistics" over the set of strings of size k.

- The "tile statistic": Counts the relative fraction of copies of $w \in \{0,1\}^k$ appearing as consecutive substrings of v starting at a multiple of k. Formally:

$$\mathcal{T}_v(w) = |\{0 \le i < \lfloor \frac{n}{k} \rfloor : v_{ki+1} = w_1, v_{ki+2} = w_2, \dots, v_{ki+k} = w_k\}| / \lfloor \frac{n}{k} \rfloor.$$

- The "shingle statistic": Counts the relative fraction of copies of $w \in \{0,1\}^k$ appearing as consecutive substrings of v starting at any index. Formally:

$$\mathcal{S}_v(w) = |\{0 \le i \le n-k : v_{i+1} = w_1, v_{i+2} = w_2, \dots, v_{i+k} = w_k\}| / (n-k+1).$$

Show that for any fixed k and ϵ there exists N, such that if $n > N$ then there is always a way to delete from v no more than ϵn bits and obtain a (not necessarily consecutive) substring v', so that $|\mathcal{T}_{v'}(w) - \mathcal{S}_v(w)| \le \epsilon$ for every $w \in \{0,1\}^k$ (in other words, the tile statistic of the new string v' should approximate the shingle statistic of the original v).

Satisfying all (9 points)

Assume that we have a CNF formula (a SAT instance), in which there are no clauses with fewer than two literals, and moreover any clause with between two and six literals contains no negated literals. Prove that if there is no clause which intersects (i.e. shares variables with) more than three other clauses, then there is an assignment to the variables satisfying all clauses at once.

Exercise 4 (Due: 21.2.2010)

Playing with FKG (6 points)

Let $f : \mathcal{P}(S) \to \mathbb{R}$ be non-negative and monotone non-decreasing, $h : \mathcal{P}(S) \to \mathbb{R}$ be non-negative and monotone non-increasing, and $\mu : \mathcal{P}(S) \to \mathbb{R}$ be non-negative log-super-modular. Show that

$$\left(\sum_{A \subseteq S} \mu(A) f(A) \right) \left(\sum_{A \subseteq S} \mu(A) h(A) \right) \geq \left(\sum_{A \subseteq S} \mu(A) f(A) h(A) \right) \left(\sum_{A \subseteq S} \mu(A) \right).$$

Just wandering (9 points)

Let G be a simple non-directed graph, s and t be two vertices, and u and v two others. To avoid multiple edges and loops in the following we also assume that u and v are not connected to each other and have no common neighbor. Suppose that there exists an automorphism σ of G (i.e. a permutation $\sigma : V \to V$ such that for all $x, y \in V$ we have $xy \in E$ if and only if $\sigma(x)\sigma(y) \in E$), that preserves s and t (i.e. $\sigma(s) = s$ and $\sigma(t) = t$) and satisfies $\sigma(u) = v$. Show that the commute time between s and t does not change if we "fuse" u and v, i.e. replace them with a single vertex whose set of neighbors is the union of the neighbors of u and v.

Independent in triples (6 points)

Show for any $k > 1$ a way to construct 2^k random variables, that depend only on $k+1$ uniformly and independently random bits, such that each of the constructed variables gets a uniformly random value from $\{0, 1\}$, and moreover any three distinct variables are (fully) independent of each other.