CMSC 858S: Mid-term, due on March 16th (Fri.) at the start of class

Notes: Please work on this individually (and not as a part of your group). You can consult your own notes or the handed-out lecture notes for the class; consulting other sources (including other people, papers, books, the Web) is not allowed. Write your solutions neatly; if you are able to make partial progress by making some additional assumptions, then state these assumptions clearly and submit your partial solution.

There are five questions total.

1. We consider tail bounds here.

(a) For any two random variables $Y$ and $Z$ such that $Y, Z \in [0, 1]$, show that $\text{coVar}[Y, Z] \leq E[YZ] \leq (E[Y] + E[Z])/2$. (Here, $\text{coVar}[Y, Z]$ denotes the covariance of $Y$ and $Z$, i.e., $E[YZ] - E[Y] \cdot E[Z]$.) \hfill (2 points)

(b) Prove that there is a constant $c > 0$ such that the following holds.

Suppose $X_1, X_2, \ldots, X_n$ are random variables, each lying in $[0, 1]$, such that each $X_i$ is dependent on at most $d$ of the other $X_j$. (For instance, if $d = 3$, then $X_1$ may depend on $X_2, X_4, X_7$ and be independent of the others, $X_2$ may depend on $X_1, X_8$ and be independent of the others, etc.) Let $X = \sum_i X_i$ with $\mu = E[X]$; assume $d \geq 1$. Then, for any $\delta \in (0, 1)$,
\[
\text{Pr}[|X - \mu| \geq \mu\delta] \leq c \cdot \frac{d}{\mu\delta^2}.
\]

(4 points)

2. All matrices in this problem have all entries lying in the set $\{0, 1\}$. Given a matrix with $k$ columns, call the matrix “$k$-nice” if each possible $k$-bit vector occurs as some row of the matrix (note that there are $2^k$ many $k$-bit vectors). Next, given a matrix $A$ with $n \geq k$ columns and some number $m$ of rows, call $A$ “$(n,k)$-good” if the $m \times k$ submatrix of $A$ obtained by taking all $m$ rows and any $k$ columns of $A$, is $k$-nice.

Prove that there is a constant $c > 0$ such that for any given integers $k$ and $n \geq k$ (also, $n \geq 2$), there is a matrix $A$ with $n$ columns and at most $ck2^k \log n$ rows, such that $A$ is $(n,k)$-good. \hfill (8 points)

3. This problem is about “continuous estimates” of the value $F_2$ of a data stream. Recall that a data stream is a sequence of arrivals $a = a_1, a_2, \ldots, a_m$, where each $a_i$ lies in the set $\{1, 2, \ldots, n\}$. For any $t$, $1 \leq t \leq m$, let:

- $\ell_{i,t}$ denote the number of occurrences of $i$ among the first $t$ elements of the stream; i.e., $\ell_{i,t}$ is the number of occurrences of $i$ among the elements $a_1, a_2, \ldots, a_t$.
- $F_{2,t}$ denote $\sum_{i=1}^n \ell_{i,t}^2$.

1
Thus, in particular, $F_{2,m}$ is the value $F_2$ that we discussed in class, and $F_{2,1} = 1$.

What we aim for is a data-stream algorithm with small storage, which will output an estimate $X_{2,t}$ for $F_{2,t}$ soon after $a_t$ arrives, for all $t$ ($1 \leq t \leq m$). That is, after each new arrival of a data element, we want to output an estimate of the $F_2$ value of the elements that have arrived so far. For any given parameter $\epsilon > 0$, design a data-stream algorithm for this “continuous estimation” task, which has space complexity that is bounded by a polynomial of $\log n$, $\log m$ and $1/\epsilon$, and which has the property

$$\Pr[\forall t \in \{1, 2, \ldots, m\}, |X_{2,t} - F_{2,t}| \leq \epsilon \cdot F_{2,t}] \geq 0.99.$$  

Note that all of the estimates $X_{2,t}$ should be simultaneously good, with probability at least 0.99. You can assume that $n, m$ and (of course) $\epsilon$ are known when you design the algorithm. (5 points)

4. Given a matrix $A$, any nonzero subset $R$ of the rows of $A$, and any nonzero subset $C$ of the columns of $A$, we can naturally define $A_{R,C}$, the “submatrix of $A$ induced by $R$ and $C$”: $A_{R,C}$ is the matrix of all entries of $A$ that simultaneously lie in some row in $R$, and in some column in $C$. Thus, if $R$ is a set of $r$ rows and $C$ is a set of $c$ columns, then $A_{R,C}$ is a matrix with $r$ rows and $c$ columns.

Given a matrix $B$ with entries lying in $\{-1, +1\}$, define $D(B)$ to be the absolute value of the sum of all entries in $B$: $D(B) = |\sum_{i,j} B_{i,j}|$. Thus, if $B$ is an $m \times n$ matrix with entries in $\{-1, +1\}$, then $D(B)$ is an integer in the range $\{0, 1, \ldots, mn\}$.

Prove that there is a constant $c > 0$ such that the following holds: for all $n$, there exists an $n \times n$ matrix $A$ with entries lying in $\{-1, +1\}$, such that for any nonzero subset $R$ of the rows of $A$ and any nonzero subset $C$ of the columns of $A$, $D(A_{R,C}) \leq cn^{3/2}$. (10 points)

5. Let $A_1, A_2, \ldots, A_n$ be random variables each taking values in $\{0, 1\}$, and not necessarily independent. Prove that for any positive integers $t$ and $a$ with $t \leq a$,

$$\Pr[A_1 + A_2 + \cdots + A_n \geq a] \leq \left( \sum_{1 \leq i_1 < i_2 < \cdots < i_t \leq n} \Pr[A_{i_1} = A_{i_2} = \cdots = A_{i_t} = 1] \right) \frac{1}{\binom{n}{t}}.$$ 

Note that there are $\binom{n}{t}$ terms in the sum above. (6 points)