Problem Set 1 Solutions

Point totals are in the margin; problem 2 was not graded. Max total number of points was 15.

1. Testing commutativity in groups

(a) Note first that some generator $g_i$ must lie outside $H$, since $H$ is a proper subgroup. So let $i$ be the minimum index such that $g_i \notin H$. Then we can write the random product as $h = a \circ g_i^b \circ c$, where $a \in H$. Now I claim that, for any choice of $a$ and $c$, we cannot have $h \in H$ for both $b_i = 0$ and $b_i = 1$. For if $c \in H$ then $h = a \circ g_i \circ c \notin H$, and if $c \notin H$ then $h = a \circ c \notin H$. Hence $\Pr[h \in H] \leq \frac{1}{2}$.

Common Issues: Many people claimed that if $g_i \notin H$, then any term with $g_i^1$ must not be in $H$. This is clearly false, since for any subgroup $H$, $1 \in H$, and $g \circ g^{-1} = 1 \in H$, even if $g \notin H$.

(b) When $G$ is not abelian, its center $C$ is a proper subgroup. Hence by part (a), $\Pr[h \notin C] \geq \frac{1}{2}$. Assuming $h \notin C$, we have that $Z(h)$ is a proper subgroup, so again by part (a) $\Pr[h' \notin Z(h)] \geq \frac{1}{2}$. Putting these together we see that $\Pr[h \circ h' \neq h' \circ h] = \Pr[h' \notin Z(h)|h \notin C] \Pr[h \notin C] \geq \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

Common Issues: Some people claimed that $\Pr[h \notin C] \Pr[h' \notin Z(h)] \geq 1/4$. This is false, since if $h \in C$, then $\Pr[h' \notin Z(h)] = 0$, since in that case $Z(h)$ is not a proper subgroup. One must instead consider the conditional probability $\Pr[h' \notin Z(h)|h \notin C]$.

2. More on pattern matching

Throughout, we will interpret an $m \times m$ binary matrix $Z$ as an $m^2$-bit binary number by writing out the matrix in row-major order (i.e., writing out the first row, then the second row, and so on), and we use as a fingerprint $F_p(Z) = Z \mod p$, where $p$ is a prime chosen at random from a suitable range. Writing $X[i, j]$ for the $m \times m$ submatrix of $X$ with top left corner at position $(i, j)$, we can apply the usual Karp-Rabin scheme as follows:

for $j = 1$ to $n - m + 1$ do  
    for $i = 1$ to $n - m + 1$ do 
        if $F_p(Y) = F_p(X[i, j])$ then return “match”  
        return “no match”

Given $F_p(X[i, j])$ the next fingerprint $F_p(X[i + 1, j])$ can be computed quickly using the rule

$$F_p(X[i + 1, j]) = \left(2^m \left(F_p(X[i, j]) - 2^{m(m - 1)}F_p(x[i, j])\right) + F_p(x(i + m, j))\right) \mod p,$$

where $x[i, j]$ denotes the $m$-bit number obtained from the matrix elements $x_{i,j}$ through $x_{i,j+m-1}$. If we have precomputed $F_p(x[i, j])$ for all $i, j$, this update takes time only $O(1)$, assuming that multiplication and addition mod $p$ can be done in constant time. Each time the column is shifted right, however, the initial fingerprint $F_p(X[1, j])$ takes time $O(m)$ to compute, since each of the $m$ rows must be handled separately. Therefore the total running time is $O(n^2 + mn) = O(n^2)$, plus the time to precompute the $F_p(x[i, j])$. This precomputation can also be handled column by column. The first column ($j = 1$) takes time $O(mn)$ since there are $n$ rows to deal with, each one being an $m$-bit string. Subsequent columns can be handled more efficiently as follows:

$$F_p(x[i, j + 1]) = \left(2F_p(x[i, j]) - 2^m x_{i,j} + x_{i,j+m}\right) \mod p.$$
This computation takes only constant time, so that the remaining $E_p(x[i,j])$ (after the first column) can be computed in $O(n^2)$ time, giving a total precomputation time of $O(mn + n^2) = O(n^2)$ and a total running time of $O(n^2)$. For comparison, note that the naive algorithm, which explicitly compares $Y$ against all submatrices $X[i, j]$, takes time $O(n^2m^2)$.

Just as in the one-dimensional case, if $Y$ is contained in $X$ then this algorithm is always correct. Otherwise, an error can occur only when $p$ divides $\prod_{i,j} |Y - X[i,j]|$, which is an $(m^2n^2)$-bit number. Thus choosing $p$ randomly from the primes in $\{2, \ldots, T\}$, where $T = cn^2$ for a suitable modest constant $c$, yields a small probability of error. This means that $p$ will have only $O(\log n)$ bits, so our assumption that arithmetic mod $p$ can be done in constant time is justified.

### 3. Probabilistic method for dominating sets

Pick a random subset $S$ of vertices by including each vertex of $V$ in $S$ independently with probability $p$ (a value to be chosen later). Let $T \subseteq U$ be the set of vertices in $U$ that are neither in $S$ nor adjacent to a vertex of $S$. Then clearly $S \cup T$ is a dominating set for $U$.

Note that a vertex $u \in U$ belongs to $T$ iff neither it nor any of its (at least $d$) neighbors belongs to $S$, which happens with probability at most $(1 - p)^{d+1}$. Hence the expected size of the dominating set is

$$E(|S \cup T|) = E(|S|) + E(|T|) \leq np + n(1 - p)^{d+1} \leq n(p + e^{-p(d+1)}).$$

Now we choose $p$ so as to minimize this expression. Differentiating the term in the parentheses, we see that we want to take $e^{-p(d+1)} = \frac{1}{d+1}$, or $p = \frac{\ln(d+1)}{d+1}$. Plugging this into the above expression gives $E(|S \cup T|) \leq n \cdot \frac{\ln(d+1)+1}{d+1}$, so we deduce that there must exist a dominating set of at most this size. Indeed, since the size of a dominating set must be an integer, we know there exists a dominating set of size $\lceil n \cdot \frac{\ln(d+1)+1}{d+1} \rceil$.

### 4. A threshold for isolated vertices

(a) For $i = 1, \ldots, n$, let $X_i$ be the indicator random variable for the event that vertex $i$ is isolated, and let $X = \sum_i X_i$ be the number of isolated vertices. To show that $p = \frac{\ln n}{n}$ is a threshold, we need to show the following two facts:

1. if $p \gg \frac{\ln n}{n}$, then $\Pr[X > 0] \to 0$ as $n \to \infty$.
2. if $p \ll \frac{\ln n}{n}$, then $\Pr[X > 0] \to 1$ as $n \to \infty$.

[Here, as in Lecture 6, for functions $f(n), g(n)$, $f \gg g$ means that $f(n)/g(n) \to \infty$ as $n \to \infty$, and $f \ll g$ means that $f(n)/g(n) \to 0$ as $n \to \infty$.]

Clearly $EX_i = n(1 - p)^{n-1} = \mu$. Note that

$$\ln \mu = \ln n + (n - 1) \ln(1 - p) \sim \ln n - (n - 1)p.$$  \hspace{1cm} (1)

[Here $\sim$ means that $f(n)/g(n) \to 1$ as $n \to \infty$.] Thus, as $n \to \infty$, we have $\mu \to 0$ if $p \gg \frac{\ln n}{n}$, and $\mu \to \infty$ if $p \ll \frac{\ln n}{n}$.

For $p \gg \frac{\ln n}{n}$ we immediately have by Markov’s inequality that

$$\Pr[X > 0] = \Pr[X \geq 1] \leq \frac{\mu}{n} \to 0 \quad \text{as } n \to \infty,$$

which establishes Fact 1 above.

To establish Fact 2, for $p \ll \frac{\ln n}{n}$ we need to use the second moment method, as follows:

$$\Pr[X = 0] \leq \Pr[|X - \mu| \geq \mu] \leq \frac{\Var X}{\mu^2}.$$

Thus it is sufficient to show that $\frac{\Var X}{\mu^2} \to 0$. To do this, note that

$$EX^2 = \sum_i EX_i^2 + \sum_{i \neq j} E[X_iX_j] = \mu + n(n - 1)E[X_iX_j].$$
Also, for any \( i \neq j \) we have \( \mathbb{E}[X_i X_j] = (1 - p)^{2n - 3} \), and thus \( n(n - 1)\mathbb{E}[X_i X_j] \leq \mu^2/(1 - p) \). Hence

\[
\frac{\text{Var}X}{\mu^2} \leq \frac{\mu + \mu^2/(1 - p) - \mu^2}{\mu^2} = \frac{1}{\mu} + \frac{p}{1 - p} \to 0,
\]

(2) since \( \mu \to 0 \) and \( p \to 0 \). This establishes Fact 2.

(b) If we set \( p = \frac{c \ln n}{n} \) then the above calculations carry over essentially unchanged. In particular, from (1) above we see that \( \mu \to 0 \) if \( c > 1 \) and \( \mu \to \infty \) if \( c < 1 \). Thus by Markov’s inequality we still get that \( \Pr[X > 0] \to 0 \) in the former case. And in the latter case the calculation in (2) still holds as well, so we get \( \Pr[X = 0] \to 0 \) in this case. This shows that the threshold is sharp; the Note in the problem set tells you that the width of the scaling window is actually \( \frac{1}{n} \).

5. Planted cliques and cryptography

(a) Consider the set of pairs \((C, G)\) where \( C \) is a set of \( k \) vertices, and \( G \) is a graph. Each pair uniquely corresponds to a specific derivation of a graph in \( G'_{n,1/2} \), where \( G \) is the graph chosen according to \( G_{n,1/2} \), and \( C \) is the set of \( k \) vertices chosen to become a clique. Clearly each pair is equally likely, arising with probability \( \frac{1}{\binom{n}{k} 2^\binom{k}{2}} \), where the first term is the probability of selecting \( C \), and the second is the probability of choosing \( G \), since all graphs are equally likely in \( G_{n,1/2} \). For a given graph \( G' \), the number of pairs \((C, G)\) which would give rise to it is \( f(G')2^\binom{k}{2} \), since \( C \) must be one of the cliques of \( G' \), and given the choice of \( C, G \) can have any subset of the \( \binom{k}{2} \) edges between vertices in \( C \). Thus

\[
\Pr'[G'] = \frac{f(G')2^\binom{k}{2}}{\binom{n}{k} 2^\binom{k}{2}} = \frac{f(G') \Pr[G']}{\mu}.
\]

**Common Issues**: Many people simply claimed that the number of graphs \( G \) that could have given rise to a graph \( G' \) (in \( G'_{n,1/2} \)) is \( f(G')2^\binom{k}{2} \). This is not true; to see a silly example, let \( G' \) be a graph with two disjoint \( k \)-cliques; I claim that the number of graphs that could have given rise to \( G' \) is \( 2 \cdot 2^\binom{k}{2} - 1 \). To see why the \(-1\) is there, note that for either choice of the clique to add, there \( 2^\binom{k}{2} \) possible graphs that could have yielded \( G' \), however, we are double-counting the graph \( G' \), since that graph could have had given rise to \( G' \) given either choice of which clique to add (because no new edges are needed). When counting, one must be very careful to avoid double-counting....

(b) By Chebyshev’s inequality, we have

\[
\Pr[G \text{ is } \alpha \text{-bad}] \leq \Pr[|f(G) - \mu| > (n^\alpha - 1)\mu] \leq \frac{\mathbb{E}(f^2)}{\mu^2(n^\alpha - 1)^2} = \frac{O(n^c \log n)}{(n^\alpha - 1)^2},
\]

where in the last step we used the given upper bound on \( \frac{\mathbb{E}(f^2)}{\mu^2} \). For large enough \( n \), this last expression is bounded above by \( n^{-2\alpha+c+\epsilon} \) for any \( \epsilon > 0 \).
(c) For $B_j$ defined as in the hint, we have

$$\Pr'[G \text{ is } \alpha\text{-bad}] = \sum_{j=2}^{\infty} \Pr'[G \in B_j]$$

$$\leq \sum_{j=2}^{\infty} n^{(j+1)\alpha/2} \Pr[G \in B_j]$$

$$\leq \sum_{j=2}^{\infty} n^{(j+1)\alpha/2} \Pr[f(G) > n^{j\alpha/2}]$$

$$\leq \sum_{j=2}^{\infty} n^{(j+1)\alpha/2} n^{-j\alpha+c+\epsilon}$$

$$= n^{\alpha/2+c+\epsilon} \sum_{j=2}^{\infty} n^{-j\alpha/2}$$

$$= n^{\alpha/2+c+\epsilon} \times O(n^{-\alpha}) = O(n^{-\alpha/2+c+\epsilon}).$$

The second line here follows from part (a), and the fourth line from part (b).

(d) We assume that $\Pr'[G \text{ has } \mathcal{P}] = \Omega(n^{-r})$. Now note that

$$\Pr[G \text{ has } \mathcal{P}] \geq \Pr[G \text{ has } \mathcal{P} \text{ and } G \text{ is not } \alpha\text{-bad}]$$

$$\geq n^{-\alpha} \Pr'[G \text{ has } \mathcal{P} \text{ and } G \text{ is not } \alpha\text{-bad}]$$

$$\geq n^{-\alpha} \left( \Pr'[G \text{ has } \mathcal{P}] - \Pr'[G \text{ is } \alpha\text{-bad}] \right)$$

$$= n^{-\alpha} \left( \Omega(n^{-r}) - O(n^{-\alpha/2+c+\epsilon}) \right).$$

(The second line here uses part (a), and the last line part (c).) Finally, we are free to choose $\alpha$ so that $\alpha/2 - c - \epsilon > r$, which implies that the negative term in the above final expression is $o(n^{-r})$. We conclude that $\Pr[G \text{ has } \mathcal{P}] = \Omega(n^{-r-\alpha}) = \Omega(n^{-r})$, as desired.

6. More on unbalancing lights

(a) Since $z \geq 0$ the required inequality, for any fixed $\alpha > 0$, is equivalent to

$$z - \frac{z^3}{\alpha} \leq \frac{2\sqrt{\alpha}}{3\sqrt{3}}.$$

Differentiating the left-hand side and setting to zero, we get a turning point at $z = \sqrt{\frac{2\alpha}{3}}$, and since the second derivative is negative this must be a maximum. There are no other turning points in $z \geq 0$, so the global maximum value is $\frac{2\sqrt{\alpha}}{3\sqrt{3}}$, as desired.

For any random variable $Z$, applying the above inequality to $|Z|$ and taking expectations we get

$$E(|Z|) \geq \frac{3\sqrt{3}}{2\sqrt{\alpha}} \left( E(Z^2) - \frac{E(Z^4)}{\alpha} \right).$$

This holds for any $\alpha > 0$, so we may choose $\alpha$ to maximize the right-hand side. (We may also assume that $E(Z^2) \neq 0$; otherwise $Z = 0$ with probability 1.) Differentiating, we see that the optimal $\alpha$ is $\alpha = \frac{3E(Z^4)}{E(Z^2)}$. Plugging this in and tidying up gives the claimed bound.

(b) Setting $Z = S_n$ in part (a), we need to evaluate $E(S_n^2)$ and $E(S_n^4)$. First we have

$$E(S_n^2) = \sum_i E(X_i^2) + \sum_{i \neq j} E(X_i X_j) = n.$$
To see this, note that $E(X_i^2) = 1$ since $X_i$ is $\pm 1$-valued; and $E(X_i X_j) = E(X_i) E(X_j) = 0$ since the $X_i$ are $4$-wise independent (and hence certainly pairwise independent). By similar considerations we have

$$E(S_n^4) = \sum_i E(X_i^4) + \frac{1}{2} \binom{4}{2} \sum_{i \neq j} E(X_i^2 X_j^2) = n + 3n(n - 1) = 3n^2 - 2n.$$  

Note that all other terms have expectation zero because of 4-wise independence and the fact that $E(X_i) = 0$. (The $\frac{1}{2} \binom{4}{2} = 3$ arises as the number of ways of partitioning the two copies of $X_i$ and $X_j$ among the four factors.) Plugging these two values into the bound from part (a) gives

$$E(|S_n|) \geq \frac{n^{3/2}}{(3n^2 - 2n)^{1/2}} \geq \sqrt{\frac{n}{3}}.$$  

(c) First, suppose we use the randomized algorithm of Lecture 5, except that now the $n$ coin tosses used to determine the column switch settings are only 4-wise independent. The analysis in Lecture 5 relied only on the fact that $E(|Z_i|)$ is asymptotically at least $\sqrt{\frac{2m}{\pi}}$, where $Z_i = \sum_{j=1}^{n} X_{ij}$ is the sum of light values for row $i$. Since the $X_{ij}$ are also 4-wise independent, we may substitute the result of part (b) to deduce that $E(|Z_i|)$ is asymptotically at least $\sqrt{\frac{2}{\pi}}$, and hence the asymptotic expected excess is $\frac{1}{\sqrt{3}} n^{3/2}$. (Note that $\frac{1}{\sqrt{3}} \approx 0.58$ while $\sqrt{\frac{2}{\pi}} \approx 0.80$.)

Now recall from Lecture 9 that we can construct a family of $r d$-wise independent fair coin flips using random polynomials of degree $d-1$ over a field of size $q = O(r)$. (Say $q = 2^m$, where $m = \lceil \log_2 r \rceil$.) In our present application, $r = n$ and $d = 4$. The number of points in this sample space (i.e., the number of such polynomials) is only $q^d = O(n^4)$, so instead of running the randomized algorithm we may exhaustively try all sample points in polynomial time. We are guaranteed to find at least one sample point for which the excess is at least the above expected value.

(d) We use the construction in the hint. First, let’s check that $E(|S_n|) = 1$. Note that if $v = (0, \ldots, 0, 0)$ then $Y_i = 0$ for all $i$, so $S_n = n$. Similarly, if $v = (0, \ldots, 0, 1)$ then $Y_i = 1$ for all $i$, so $S_n = -n$. Now consider any other value of $v$, and let $v_j$ be its leftmost non-zero entry (so $j \leq k$). For each $i \in \{0, 1, \ldots, n - 1\}$, let $i'$ denote the integer whose binary expansion differs from that of $i$ only in position $j$. Then it is clear that $Y_i = 1 - Y_{i'}$, and hence $X_i = -X_{i'}$. Thus the contributions to $S_n$ cancel out in pairs, and $S_n = 0$. Putting all this together we get

$$E(|S_n|) = \frac{1}{2^k+1} (n + n) = 1.$$  

To see that the $Y_i$ (and hence the $X_i$) are 3-wise independent, note first that, by the same argument as above, $\Pr[Y_i = 0] = \Pr[Y_i = 1] = \frac{1}{2}$ for all $i$. Now fix any three distinct integers $i_1, i_2, i_3 \in \{0, 1, \ldots, n - 1\}$; we need to show that

$$\Pr[Y_{i_1} = a_1 \land Y_{i_2} = a_2 \land Y_{i_3} = a_3] = \frac{1}{8}$$  

for any triple of values $a_j \in \{0, 1\}$. The sample points (vectors $v$) corresponding to this triple are solutions to the equations $Bv^T = a^T$, where $a = (a_1, a_2, a_3)$, and $B$ is the $3 \times (k + 1)$ matrix whose rows are $b_{i_1}, b_{i_2}$ and $b_{i_3}$. We claim that the rows of $B$ are linearly independent over $GF[2]$; for if not then some combination of them would have to sum to 0, and since all have 1 in the last position this means that some pair must sum to 0, which is not possible since all three rows are distinct. Linear independence means that the system of equations has full rank, so the dimension of the set of solutions is $k + 1 - 3 = k - 2$, i.e., the number of solutions is $2^{k-2}$. Since this is independent of the triple $a$, all eight triples must have the same probability, $\frac{1}{8}$.

Common Issues: There were a number of “fuzzy” arguments: to show that the variables are 3-wise independent, you must argue that all 8 assignments to any 3 variables have equal probabilities (1/8); it is not sufficient to just argue that no pair of variables completely determines the value of a third variable.