Randomness and Computation 2018/19
Solutions to Coursework 1 (formative)

1. We have a fair coin and want to generate a stream of (uniform) random bits. Instead of using the natural process for generating (say) the N needed bits (flip the coin N times), we have decided we will attempt to "reuse" the randomness and generate the N bits using only $n = \lceil 2\sqrt{N} \rceil$ random flips.

   To do this, we first generate $n = \lceil 2\sqrt{N} \rceil$ fair random bits $Z_1, \ldots, Z_{\lceil 2\sqrt{N} \rceil}$ with the fair coin. Next, we consider the index set of pairs $P = \{\{i, j\} : 1 \leq i < j \leq n\}$ and for every $p \in P$ we define Y_p as the exclusive-or $Z_i \oplus Z_j$ of the particular variables of the pair $p = \{i, j\}$.

   This will give rise to $|P|$ different Y_p random variables.

   We will also consider the "total" random variable $Y = \sum_{p=1}^{|P|} Y_p$.

   (a) We show that for every $p \in P$, Y_p is 0 with probability $1/2$ and 1 with probability $1/2$. *[4 marks]*

   To see this, remember $Y_p = Z_i \oplus Z_j$ where Z_i, Z_j are two different uniform random bits. Since Z_i and Z_j are independent, the joint distribution on $Z_i Z_j$ gives $00, 01, 10, 11$ with probability $1/4$ each. The two middle events will set Y_p to be 1 (with probability $1/4+1/4 = 1/2$) and the 00 and 11 events will give $Y_p = 0$ (with probability $1/2$).

   marking: Up to 4 marks for a well-argued justification which refers to the details of the joint distribution, and the independence of the two contributing bits.

   (b) We now count the number of different Y_p variables. *[4 marks]*

   We have one Y_p variable for every pair of indices $1 \leq i, j \leq n$ with $i \neq j$ and the number of such pairs is $\binom{n}{2}$, which is $\frac{n(n-1)}{2}$. Substituting in $n = \lceil 2\sqrt{N} \rceil$, this is

   $$\frac{\lceil 2\sqrt{N} \rceil (\lceil 2\sqrt{N} \rceil - 1)}{2} \;\geq\; \sqrt{N}(\lceil 2\sqrt{N} \rceil - 1).$$

   For any value $N \geq 1$, we have $2\sqrt{N} - 1 = \sqrt{N} + \sqrt{N} - 1$, and since $\sqrt{N} \geq 1$, this implies $2\sqrt{N} - 1 \geq \sqrt{N}$. Hence we certainly have $\lceil 2\sqrt{N} \rceil - 1 \geq \sqrt{N}$ for $N \geq 1$, and substituting into our expression above, the cardinality of P is at least $\sqrt{N} \cdot \sqrt{N} = N$, as required.

   marking: 2 of the 4 marks are going for the initial steps to get $\sqrt{N}(\lceil 2\sqrt{N} \rceil - 1)$, the other 2 are going for working with $(\lceil 2\sqrt{N} \rceil - 1)$ to show it is $\geq \sqrt{N}$. If they are careless with the $\lceil \cdot \rceil$ then they get *at most* 3.

   (c) We now show that every pair of the Y_p variables satisfy the definition of *pairwise independence*, and hence that that $E[Y_p Y_q] = E[Y_p]E[Y_q]$. *[4 marks]*

   We first note that for a 0/1 random variable X, that $E[X] = \Pr[X = 1]$. This is the case for all of Y_p, Y_q and $Y_p Y_q$; hence our aim is to show that $\Pr[Y_p Y_q = 1] = \Pr[Y_p = 1] \Pr[Y_q = 1]$ for p, q different pairs. There are two cases.

(a) Suppose the component variables Z_i, Z_j for Y_p do not overlap with the $Z_{i'}, Z_{j'}$ for Y_q. In this case the random variable $Y_P = Z_i \oplus Z_j$ is independent of $Y_q = Z_{i'} \oplus Z_{j'}$, hence $\Pr[Y_p Y_q = 1] = \Pr[Y_p = 1] \Pr[Y_q = 1]$.
(b) We might have two pairs p, q that overlap, say $Y_p = Z_i \oplus Z_j$, but $Z_q = Z_i \oplus Z_{j'}$, for $j' \neq j$ (we can't have both items in the pairs matching as then p would equal q). We already know $\Pr[Y_p = 1] = 1/2, \Pr[Y_q = 1] = 1/2$ from (a). To analyse, $\Pr[Y_p Y_q = 1]$ note that we can obtain $Y_p = 1$ and Y_q in just two cases of the variables $Z_i, Z_j, Z_{j'}$: (1, 0, 0) and (0, 1, 1). Each of these triplets has probability $1/8$ of being generated from the original three coin flips, giving $1/4$ overall for the $\Pr[Y_p Y_q = 1]$. Hence $\Pr[Y_p Y_q = 1] = \Pr[Y_p = 1] \cdot \Pr[Y_q = 1]$ in this case also.

marking: Up to 4 marks for a good argument.

(d) We now show that the collection of $\{Y_p : p \in P\}$ variable do *not* satisfy the definition of mutual independence. *[4 marks]*

We will consider any three Z_i, Z_j, Z_k of the original unbiased random bits, and focus on the induced distribution on $Y_{\{i,j\}}, Y_{\{i,k\}}, Y_{\{j,k\}}$. From our analysis in (c) we know that we will have $Y_{\{i,j\}} = 1, Y_{\{i,k\}} = 1$ if and only if Z_i, Z_j, Z_k is either (1,0,0) or (0,1,1). Note that in both these two situations, $Z_j = Z_k$ and hence $Y_{\{j,k\}} = 0$. Hence the probability that $Y_{\{i,j\}} Y_{\{i,k\}} Y_{\{j,k\}}$ is 111 in the joint distribution is 0, when it should instead be $1/8$. Hence we do not have mutual independence, even for triplets.

marking: They will need to give a detailed argument of where the mutual independence breaks. Just arguing that $2\sqrt{N}$ bits can't generate N truly independent bits won't cut it (1 mark only for this). They should give a concrete example of where the mutual independence breaks and detailed justification to get all 4 marks.

(e) What is the expected value $E[Y]$? *[4 marks]*

Using linearity of expectation (which does not require independence) the expected value is $|P|/2$, which is $\frac{n(n-1)}{4} = \frac{\lceil 2\sqrt{N} \rceil (\lceil 2\sqrt{N} \rceil - 1)}{4}$ (approximately N).

marking: 4 marks for a correct argument.

(f) We will now use the fact that the $\{Y_p : p \in P\}$ are pairwise independent to calculate $\mathrm{Var}[Y]$. *[4 marks]*

We will use the fact (because of pairwise independence) that $\mathrm{Var}[Y] = \mathrm{Var}[\sum_{p \in P} Y_p] = \sum_{p \in P} \mathrm{Var}[Y_p]$.

Now, for an arbitrary $p = \{i, j\}$, $\mathrm{Var}[Y_p] = E[(Y_p)^2] - E[Y_p]^2$, We already know $E[V_p]^2 = 1/4$. For Y_p^2, this can be either 0 or 1. Y_p^2 will be 1 only if Z_i and Z_j have differing values, which happens with probability $1/2$, otherwise it is 0. So $E[Y_p^2] = 1/2$ and $\mathrm{Var}[Y_p] = 1/4$.

Hence $\mathrm{Var}[Y] = \frac{|P|}{4} = \frac{n(n-1)}{8}$.

marking: 4 marks for a correct answer with some justification.

(g) We now use Chebyshev's inequality to prove an upper bound on $\Pr[|Y - E[Y]| \geq n]$. *[6 marks]*

Chebyshev's Inequality tells us that $\Pr[|Y - E[Y]| \geq n] \leq \frac{\mathrm{Var}[Y]}{n^2}$, and this bound is $\frac{n(n-1)}{8n^2}$ which is $< \frac{1}{8}$.

marking: 6 marks for a correct application of Chebyshev and the answer close to $1/8$.

2. We consider the coupon collector problem where the book only has space for $n/2$ players and we must fill these positions with $n/2$ different footballers. *[10 marks]*

   We can consider the purchases of cereal packets to be divided into a similar set of geometric random variables as in the standard coupon collector problem, with the random variable X_i for phase i being the number of trials to advance from $i-1$ cards to i cards. Each X_i will be distributed as a random variable with parameter $p_i = \frac{n-(i-1)}{n} = 1 - \frac{i-1}{n}$.

   To analyse the amount of time to get $n/2$ cards, we define $X = \sum_{i=1}^{n/2} X_i$, then we are interested in the expected value

$$E[X] \;=\; \sum_{i=1}^{n/2} E[X_i] \;=\; \sum_{i=1}^{n/2} \frac{n}{n-(i-1)} = \sum_{i=n}^{n/2+1} \frac{n}{i} = n\Big(\sum_{i=n/2+1}^{n} \frac{1}{i}\Big).$$

   We will rewrite $\sum_{i=n/2+1}^{n} \frac{1}{i}$ as $\sum_{i=1}^{n} \frac{1}{i} - \sum_{i=1}^{n/2} \frac{1}{i}$, to allow us to exploit the fact that $\sum_{i=1}^{k} \frac{1}{i}$ is a crude approximation to $\ln(k)$. In particular, we will use the fact that

$$\ln(k) \;<\; \sum_{i=1}^{k} \frac{1}{i} \;\leq\; \ln(k) + 1,$$

   using different sides of this bound for our two sums. We have

$$\begin{aligned} \sum_{i=n/2+1}^{n} \frac{1}{i} \;&=\; \sum_{i=1}^{n} \frac{1}{i} - \sum_{i=1}^{n/2} \frac{1}{i} \\ &\leq\; \ln(n) + 1 - \ln(n/2) \\ &=\; \ln(n) + 1 - \ln(n) + \ln(2) \end{aligned}$$

   which is < 2. The expected time to collect $n/2$ different cards is then less than $2n$.

   marking: They get 4 marks for going through the into stages and setting things up in relation to $\sum_{i=n/2+1}^{n} \frac{1}{i}$. then 3 marks for the re-writing as the difference of two sums, and 3 marks for finishing off with the bounds wrt $\ln(\cdot)$. If they don't use accurate bounds, but only the $\sim \ln(k)$, then take two off (at least, depends on other stuff too of course).

3. Consider a function $F : \{0, 1, \ldots, n-1\} \to \{0, 1, \ldots, m-1\}$ and suppose we know that for $0 \leq x, y \leq n-1$, $F((x+y) \bmod n) = (F(x)+F(y)) \bmod m$. The only way we know to evaluate $F(\cdot)$ is to examine the values in an array where the $F(\cdot)$ values have been stored (with entry i holding the value of $F(i)$). Unfortunately, a system failure has corrupted up to a $1/5$-fraction of the entries of the array, so we no longer have reliable values in all positions. We now design a simple randomized algorithm that, given an input $z \in \{0, \ldots, n-1\}$, outputs a value that equals $F(z)$ with probability at least $1/2$. *[10 marks]*

   solution: Our algorithm relies on the additive (modulo m) property and a simple sampling rule: regardless of the value z input, we will randomly draw a value y from $\{0, 1, \ldots, n-1\}$

uniformly at random. After having drawn y, we will define $x =_{\text{def}} (z - y) \bmod n$. Note that this ensures $z = (x + y) \bmod n$. We will then "lookup" the values of $F(x)$ and $F(y)$ from the table, add $F(x)$ and $F(y)$ together, and take the remainder with m as the final answer for $F(z)$.

The probability that this value returned is the true value $F(z)$ is exactly the probability that *both* $F(x)$ and $F(y)$ were *not corrupted*. Let's consider the fact that at most $\frac{n}{5}$ of the n entries of F have been corrupted; this means that in considering the n different (x, y) *pairs*, at most $\frac{2n}{5}$ of these might be somehow corrupted (either x corrupted, or y corrupted, or both). So $\frac{3n}{5}$ of the n (x, y) pairs for z are uncorrupted, and given that we drew y randomly from all possible n values, this means the probability we have an uncorrupted pair is $\geq \frac{3}{5} > \frac{1}{2}$. This result is independent of the particular z which was chosen.

marking: Give 4 marks for the algorithm above (with uniform sampling from $\{0, \ldots, n-1\}$ and up to 6 marks for a good detailed analysis. If they use more than 1 sample (from $\{0, 1, \ldots, n-1\}$) then I'd expect them to get at most 7marks overall and less again if it's getting complicated.

common error: Many solutions that were submitted assumed the errors were distributed uniformly in F; that makes it easier to obtain the result. This is not a legitimate assumption, the failures could be adversarially arranged; for example, maybe only the even-indexed values were affected. All we can rely on is that at most $\frac{1}{5}$ of the cells were damaged.

Now suppose we are allowed to repeat the initial algorithm times before you return a result. *[5 marks]*
When it says "repeat the algorithm" this implies a resampling of the y (and hence a new (x, y) pair) each time. The approach should be to view the three results returned for $F(z)$ and if ≥ 2 of those are the same, return that max value; otherwise, if all are different, randomly return any of them.

What is the probability of a correct answer? It is certainly at least as high as the probability that *at most one answer is corrupted*, which is $\frac{3}{5}^3 + 3\frac{2}{5}\frac{3}{5}^2$. If exactly two of the results are corrupted, which happens with probability $3\frac{3}{5}\frac{2}{5}^2$, there is still probability at most $\frac{1}{3}$ of the right answer being returned - unfortunately, this is not guaranteed, as we may be tricked if the two corrupted values match. So we can only say that the total probability of the right answer returned will be at least

$$\frac{9}{25}\frac{1}{5}(3 + 6) \;=\; \frac{81}{125},$$

which is almost .65, an improvement on the original result.

marking: Give 3 marks for explaining what to do, and the other 2 marks for the analysis. They may have decided to just work with the $\frac{1}{2}$ from the phrasing of the initial question, and shouldn't lost any marks for that (though they will only get the answer $\frac{1}{2}$ mirroring the analysis above). Delete at least one mark if they add extra probability of success for the "3 different values" scenario.

4. Our variation on the "Max-Cut" (or $\frac{|E|}{2}$-cut) algorithm will chose a random *balanced* cut of the graph (instead of a random cut of variable size, as we did in lecture 6). Specifically, we choose $S \subset V$ to be any random subset of size $\lfloor \frac{|V|}{2} \rfloor$ and take the cut $(S, V \setminus S)$ to be our cut. Let $n = |V|$. *[20 marks]*

We now analyse the expected size of this cut, in two cases.

n **even:** In this case, S will be any subset of size exactly $n/2$, and the same for $V \setminus S$. Let $e = (u, v) \in E$. Then the probability that u and v are on different sides of the cut is exactly

$$\frac{2\binom{n-2}{n/2-1}}{\binom{n}{n/2}} = \frac{2(n-2)!}{(n/2-1)!(n/2-1)!}\frac{(n/2)!(n/2)!}{n!} = \frac{2(n/2)^2}{n(n-1)} = \frac{n}{2n-2} > \frac{n}{2n-1}.$$

n **odd:** In this case, S will be any subset of size exactly $(n-1)/2$, with $V \setminus S$ of size $(n+1)/2$. Then for $e = (u, v) \in E$, the probability that u and v are on different sides of the cut is exactly

$$\frac{2\binom{n-2}{(n-3)/2}}{\binom{n}{(n-1)/2}} = \frac{2(n-2)!}{\frac{n-3}{2}!\frac{n-1}{2}!} \cdot \frac{\frac{n-1}{2}!\frac{n+1}{2}!}{n!} = \frac{2}{n(n-1)}\frac{n+1}{2}\frac{n-1}{2} = \frac{n+1}{2n}$$

which is also greater than $\frac{n}{2n-1}$, as required.

Now we consider the set E of all edges, and note that the number of edges in the random cut $(S, V\setminus S)$ is $E[X] = E[\sum_{\{u,v\}\in E} X_{u,v}]$, where $X_{u,v}$ is the indicator variable that is 1 when u and v end up on opposite sides of the cut. Applying linearity of expectation this is $\sum_{\{u,v\}\in E} E[X_{u,v}]$, which is $|E|E[X_{u,v}]$ for any $\{u, v\}$ (our analysis above did not depend on the particular edge we were interested in). This is at least $|E|\frac{n}{2n-1}$, as required.

marking: The students had to come up with the idea of the algorithm themselves. 6marks for the algorithm, with the idea of taking a balanced-size cut (and details of the value needed for $|S|$ in the odd and even cases). Then the other 14 marks should be split as 3 marks for applying linearity of expectation over the various $\{u, v\}$ (with explanation) and 11 marks for analysing $E[X_{u,v}]$ for a particular $\{u, v\}$. The 11 marks should be split between even and odd, with them getting 6 marks if they only do a (perfect) argument for one of the cases.

5. (a) Given the $X_1, ..., X_m$ of the random variable X supported in $[0, 1]$, we have the very simple algorithm which estimates $E[X]$ as $\widehat{X} = m^{-1}\sum_{i=1}^{m} X_i$ (adds up all the sample values, and then divides by the number of samples m). Our aim is to derive conditions on m to ensure that $|\widehat{X} - E[X]| \leq \epsilon$ with probability at least $1 - \delta$, for given $\delta, \epsilon \in (0, 1)$. *[10 marks]*

   If we consider $\mathbf{X} = \sum_{i=1}^{m} X_i$, then the condition above on $\widehat{X}$ is same as having $|\mathbf{X} - E[X]\cdot m| \leq \epsilon m$ with probability $\geq 1 - \delta$, ie, having $|\mathbf{X} - E[X] \cdot m| > \epsilon m$ with probability *at most* δ.

   Our only knowledge of X is that it takes values in the interval $[0, 1]$, therefore we are working with a less-understood r.v. that a Bernoulli r.v., say. So Chernoff Bounds are not at our disposal. However, we do have the option of using Chebyshev and are helped by the fact that in considering $\mathbf{X}$ the sum of all the X_i, we can use the independence of the different X_i trials to infer that $\mathrm{Var}[\mathbf{X}] = \sum_{i=1}^{m} \mathrm{Var}[X_i]$. Remember that for any individual i, that $\mathrm{Var}[X_i] = E[X_i^2] - E[X_i]^2$, and that because of the *support* of X, this value is at most $1 - E[X]^2$. Hence $\mathrm{Var}[\mathbf{X}] \leq m \cdot (1 - E[X]^2) \leq m$.

   Our goal is to choose m sufficient to show $\Pr[|\mathbf{X} - E[X] \cdot m| > \epsilon m] < \delta$. Chebyshev tells us that $\Pr[|\mathbf{X} - E[X] \cdot m| > \epsilon m] \leq \frac{\mathrm{Var}[\mathbf{X}]}{(\epsilon m)^2}$. Hence, given our knowledge of $\mathrm{Var}[\mathbf{X}]$, we know

$$\Pr[|\mathbf{X} - E[X] \cdot m| > \epsilon m] \leq \frac{m}{(\epsilon m)^2} = \frac{1}{\epsilon^2 m}.$$

Our aim is to achieve $\Pr[|\mathbf{X} - \mathrm{E}[X] \cdot m| > \epsilon m] \leq \delta$, hence we should take $m \geq \delta^{-1}\epsilon^{-2}$.

marking: 5marks for reformulating in terms of the summation r.v. and casting it in Chebyshev terms, then 5 for working down for the details of $\delta^{-1}\epsilon^{-2}$. If they used Chernoff (which is not applicable to this case) give some marks.

common errors: They might have thought they could use Chernoff, but that's not applicable when we have variables which can take on real values in the interval $[0, 1]$ (for Chernoff we need binomial values, or at worse (for more general forms of Chernoff/Hoeffding), variables which take on a small number of specific values).

(b) Given an undirected graph $G = (V, E)$, where $|V| = n$, we must assign labels to each vertex such that the sum of the labels of each vertex and its neighbours modulo $n + 1$ is nonzero. *[15 marks]*

Our procedure is to assign a label in the range $\{0, 1, 2, \ldots, n\}$ to each vertex independently and uniformly at random, check whether this satisfies the condition above, and if not, try again from the beginning. Note that the task of generating a single random labelling of the vertices, and checking the desired property at each vertex, will take time only $O(n^2)$ overall. Therefore our goal is to show we expect to need only polynomially-many trials before a correct labelling is generated.

Let us examine the situation for a single trial, focusing first on what happens to a single vertex v. Regardless of the structure of the graph, we are guaranteed that the probability of v (and its neighbours) violating the condition wrt $\mathrm{mod}(n + 1)$ is exactly $\frac{1}{n+1}$. This can be seen by using "deferred decisions" to first generate labels for all vertices in $\mathrm{Adj}(v)$, and randomly generate v's label last - note that after the adjacent vertices have value labels, that if we sum these modulo $(n+1)$, we get some value $k \in \{0, \ldots, n\}$. Then in generating v's label, the probability we generate $(n + 1 - k)$ is exactly $\frac{1}{n+1}$. These events (violating v's condition) are not necessarily independent for the various vertices v. However, just using the Union Bound (which does not require any independence), the probability that *some* vertex of the graph will violate the condition (on a single random trial generating labels for all vertices) is at most $\frac{n}{n+1}$, as there are only n vertices in the graph.

Now observe that the process of repeating the "generate labels for the whole graph" process until we succeed is actually a geometric random variable with success probability at least $p = \frac{1}{n+1}$. The results from lecture 5 of the course tell us that the expected number of trials will be at most $p^{-1} = n+1$. Then taking $(n+1) \cdot O(n^2)$, this is certainly polynomial.

marking: 5 marks for examining what happens at a single vertex and analysing the probability of a violation at v exactly. The next 4 marks for using/explaining the Union Bound and getting $\frac{n}{n+1}$ as an upper bound on the probability of a violation somewhere in the graph. 4 marks for the relation to Geometric random variables and inferring the $n+1$. And 2 marks for discussing the time for a single round of label-generation and checking.

Mary Cryan, 7th March 2019