CS–174 Combinatorics & Discrete Probability, Spring 2023

Final Exam Solutions

11:30am–2:30pm, 10 May

Read these instructions carefully

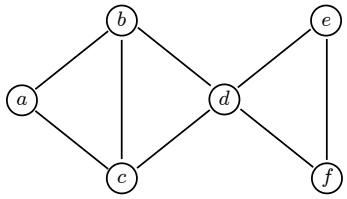
1. **Write your name and SID number on the front page, and your SID number on every page!**
2. This is a **closed book** exam, but you are allowed one two-sided cheat sheet and blank scratch paper. No phones, calculators or other electronic equipment.
3. The exam consists of 14 questions. The first 9 questions are multiple choice; the remaining 5 require written answers.
4. Approximate point totals for each question part are indicated in the margin. The maximum total number of points is 100.
5. **Multiple choice questions 1–6:** Answer these by **filling in** the **single** circle adjacent to the correct answer; there is **no** penalty for incorrect answers (unless you select more than one answer).
6. **Multiple choice questions 7–9:** Answer these by **filling in** the squares adjacent to **all** the answers that apply. Points will be allocated both for selecting correct answers and for not selecting incorrect answers, so you should select only those that apply.
7. **Other questions 10–14:** Write your answers to these in the spaces provided below them. None of these questions requires a very long answer, so you should have enough space—if not you are probably writing too much. **Always show your working for these questions!**
8. The questions vary in difficulty: if you get stuck on some part of a question, leave it and go on to the next one. Point totals and space provided are not necessarily an indication of difficulty.

Your First Name: **Your Last Name:**

Your SID Number:

[exam starts on next page]

1. Consider random walk on the following undirected graph, with vertices $\{a, b, c, d, e, f\}$.



(a) In equilibrium, the average proportion of time spent at vertex d is *3pts*

○ 0; ○ $\frac{1}{6}$; ● $\frac{1}{4}$; ○ $\frac{1}{3}$; ○ $\frac{1}{2}$

The stationary distribution π is proportional to the degrees of the vertices, so $\pi(d) = \frac{\deg(d)}{2|E|} = \frac{4}{16}$.

(b) If the walk starts at vertex d, the expected time until it is next at d is *3pts*

○ 1; ○ 2; ● 4; ○ 8; ○ 16

The expected return time to state d is $\frac{1}{\pi(d)}$.

(c) The expected hitting time from vertex e to vertex d is *3pts*

○ $\frac{1}{2}$; ○ 1; ○ $\frac{3}{2}$; ● 2; ○ 3

$h_{ed} = 1 + \frac{1}{2}h_{fd}$ and $h_{fd} = 1 + \frac{1}{2}h_{ed}$. Solving gives $h_{ed} = h_{fd} = 2$.

---

2. We are given a biased coin that comes up heads with probability p and tails with probability $1 - p$.

(a) We flip the coin until we get a heads. The expected number of flips, including the flip that comes up heads, is: *3pts*

○ $\frac{1}{p^2}$; ● $\frac{1}{p}$; ○ 2; ○ $\frac{1}{(1-p)^2}$; ○ $\frac{1}{1-p}$

This is just the expectation of a geometric r.v. with parameter p.

(b) We flip the coin until we get three heads in a row. The expected number of tails that we flip is: *3pts*

○ $\frac{1}{(1-p)^3}$; ● $\frac{1}{p^3} - 1$; ○ $\frac{1}{(1-p)^3} - 1$; ○ $\frac{1}{p^3(1-p)} - 1$; ○ $\frac{1}{p(1-p)^3}$

Think of a "trial" as an attempt to get 3 Heads. Thus a successful trial is a sequence of 3 Heads, and an unsuccessful trial is a sequence of 0, 1 or 2 Heads followed by a Tail. The number of trials until we get a success is geometric with parameter p^3. The number of Tails is equal to the number of *unsuccessful* trials, which has expectation $\frac{1}{p^3} - 1$.

[continued overleaf]

3. Consider the balls-and-bins model, in which n balls are thrown independently and u.a.r. into n bins.

(a) The probability that bin 1 is empty conditioned on bins $2, 3, \ldots, k+1$ being empty is *3pts*

○ $\exp(-1)$; ○ $\left(1-\frac{1}{n}\right)^n$; ○ $\left(1-\frac{k}{n}\right)^n$; ● $\left(1-\frac{1}{n-k}\right)^n$; ○ $\left(1-\frac{1}{n}\right)^k$

Conditioning effectively reduces the number of bins to $n-k$. All n balls must avoid bin 1.

(b) Under the Poisson approximation, the probability of the event in part (a) is *3pts*

● $\exp(-1)$; ○ $\left(1-\frac{1}{n}\right)^n$; ○ $\left(1-\frac{k}{n}\right)^n$; ○ $\left(1-\frac{1}{n-k}\right)^n$; ○ $\left(1-\frac{1}{n}\right)^k$

Under Poisson, the number of balls in bin 1 is Poisson with parameter $\lambda = 1$ and bin loads are independent; thus the conditioning has no effect, and the probability of bin 1 being empty is $\exp(-\lambda)$.

---

4. Consider the following two-round experiment. In round one, we flip n independent fair coins and count the number of heads X. In round two, we draw a sample Z that is Poisson with parameter X/n.

(a) The expected value of Z is *3pts*

○ $\frac{n}{2}$; ○ 1; ● $\frac{1}{2}$; ○ $\exp(-\frac{1}{2})$; ○ $\exp(-1)$

$\mathrm{E}[Z] = \mathrm{E}[\mathrm{E}[Z \mid X]] = \mathrm{E}[X]/n = (n/2)/n = 1/2$.

(b) The probability that $Z = 0$ is *3pts*

○ $\exp(-\frac{1}{2})$; ○ $\sum_{m=0}^{n/2} \frac{\binom{n/2}{m}\exp(-m/n)}{2^{n/2}}$; ○ $\frac{\binom{n}{n/2}\exp(-\frac{1}{2})}{2^n}$; ● ; $\sum_{m=0}^{n} \frac{\binom{n}{m}\exp(-m/n)}{2^n}$

$\Pr[Z=0] = \sum_{m=0}^{n} \Pr[Z=0 \mid X=m]\Pr[X=m]$.

---

5. Let $G \in \mathcal{G}_{n,p}$ be a random graph with edge probability p.

(a) The expected number of edges in G is *3pts*

○ np; ● $\binom{n}{2}p$; ○ $p^{\binom{n}{2}}$; ○ $\sum_{k=0}^{\binom{n}{2}} p^k(1-p)^{\binom{n}{2}-k}$; ○ $\binom{n}{2}p(1-p)$

By linearity of expectation: there are $\binom{n}{2}$ possible edges, each present with probability p.

(b) The expected number of vertices of degree exactly d in G is *3pts*

○ np^d; ○ $np^d(1-p)^{n-1-d}$; ● $n\binom{n-1}{d}p^d(1-p)^{n-1-d}$; ○ $n(1-p)^{n-1-d}$

Each of the n vertices has degree d with probability $\binom{n-1}{d}p^d(1-p)^{n-1-d}$.

---

6. Let $X_1, X_2, \ldots$ be a sequence of non-negative integer-valued random variables, such that $\mathrm{E}[X_n] \to \infty$ as $n \to \infty$. Which one of the following additional facts is sufficient to guarantee that $\Pr[X_n \geq 1] \to 1$ as $n \to \infty$? *3pts*

○ $\frac{\mathrm{Var}[X_n]}{\mathrm{E}[X_n]^2} \to \infty$; ○ $\frac{\mathrm{E}[X_n^2]}{\mathrm{E}[X_n]^2} \to \infty$; ○ $\frac{\mathrm{Var}[X_n]}{\mathrm{E}[X_n]^2} \to 1$; ● $\frac{\mathrm{E}[X_n^2]}{\mathrm{E}[X_n]^2} \to 1$; ○ none of these

By Chebyshev, $\Pr[X_n < 1] = \Pr[X_n = 0] \leq \Pr[|X_n - \mathrm{E}[X_n]| \geq \mathrm{E}[X_n]] \leq \frac{\mathrm{Var}[X_n]}{\mathrm{E}[X_n]^2} = \frac{\mathrm{E}[X_n^2]-\mathrm{E}[X_n]^2}{\mathrm{E}[X_n]^2} = \frac{\mathrm{E}[X_n^2]}{\mathrm{E}[X_n]^2} - 1$.

[continued overleaf]

Note: In Questions 7–9, shade the squares adjacent to all solutions that apply. Points will be awarded both for selecting the correct solutions and for not selecting the incorrect ones.

---

7. Let $X_1, ..., X_n$ be independent (not necessarily identically distributed) discrete random variables and let $Y = \sum_{i=1}^{n} X_i$. Suppose $\mathrm{E}[X_i] = 1$ and $\mathrm{Var}[X_i] = 2$ for all i. Which of the following statements **must** be true, where a is an arbitrary positive number? [Choose all that apply.] *3pts*

☐ $\Pr[Y \geq (1+a)n] \leq e^{-na^2/3}$; ■ $\mathrm{Var}[Y] = 2n$; ☐ $\Pr[X_1 = a] = \Pr[X_2 = a]$;

■ $\Pr[|X_1 - 1| \geq a] \leq \frac{2}{a^2}$; ■ $\mathrm{Cov}(Y, X_1) = 2$; ☐ $\Pr[X_1 \geq a] \leq \frac{1}{a}$

$\mathrm{Var}[Y] = \sum_i \mathrm{Var}[X_i] = 2n$ because the X_i are independent. $\Pr[|X_1 - 1| \geq a] \leq \frac{2}{a^2}$ is Chebyshev. $\mathrm{Cov}(Y, X_1) = \sum_i \mathrm{Cov}(X_i, X_1) = \mathrm{Cov}(X_1, X_1) = \mathrm{Var}[X_1] = 2$.

The Chernoff bound $\Pr[Y \geq (1+a)n] \leq e^{-na^2/3}$ assumes the X_i are 0-1 r.v.'s. $\Pr[X_1 = a] = \Pr[X_2 = a]$ assumes the X_i have the same distribution. $\Pr[X_1 \geq a] \leq \frac{1}{a}$ is Markov's inequality, which assumes X_1 is non-negative.

---

8. Let $X_0, X_1, \ldots$ be a sequence of non-negative iid random variables with mean $\mu > 0$ and variance σ^2. Which of the following sequences $Z_0, Z_1, \ldots$ is a martingale with respect to (X_t)? [Choose all that apply.] *3pts*

■ $Z_t := \sum_{i=0}^{t} X_i - \mu t$

☐ $Z_t := tX_0 - \mu t$

■ $Z_0 := 1$; $Z_t := \mu^{-t} \prod_{i=1}^{t} X_i$ for $t \geq 1$

■ $Z_t := (\sum_{i=0}^{t} X_i^2) - (\sigma^2 + \mu^2)t$

These can all be checked by straightforward calculations.

---

9. Bella visits a casino and repeatedly places \$1 bets on a fair game, winning or losing \$1 independently with probability $\frac{1}{2}$ each time. Let (X_t) denote her capital at time t, with $X_0 = \$1000$. Under which of the following stopping rules does the optional stopping theorem apply? [Choose all that apply.] *3pts*

☐ Bella stops when she has won \$200

■ Bella stops when she has either won \$200 or lost \$500

☐ Bella stops when she has spent all her money

☐ Bella stops the last time her capital reaches \$200

■ Bella stops after making 200 bets

In the first and third rules, $\mathrm{E}[T]$ is unbounded: in the first rule this is because Bella's capital hits 0 with some probability, after which she will never win \$200; in the third rule she is doing symmetric random walk on an unbounded interval. The fourth rule is not a valid stopping time. The second rule works because $\mathrm{E}[T]$ is finite (random walk on a bounded interval) and the jumps are bounded, while the fifth rule works because T itself is bounded.

[continued overleaf]

10. Strings and Codes [9pts total]

An (m, ℓ, ε)-*code* is a set of m strings of length ℓ over a large alphabet $\{1, 2, \ldots, a\}$ such that no pair of strings have the same symbols in more than $\varepsilon\ell$ positions. Codes of this kind have been useful in the theoretical study of the protein-folding problem. In this question we see how to construct an (m, ℓ, ε)-code using random strings.

(a) Suppose we pick two strings of length ℓ independently and u.a.r. from the set $\{1, 2, \ldots, a\}^\ell$. Use a Chernoff bound to give an upper bound in terms of a, ε, ℓ on the probability that the two strings agree at more than $\varepsilon\ell$ positions. You may assume that $\varepsilon > \frac{1}{a}$. [HINT: Recall the Chernoff bound of the form $\Pr[\sum_i X_i \geq \mu + \lambda] \leq \exp\{-\frac{2\lambda^2}{n}\}$, where $X_1, \ldots, X_n$ are independent 0-1 r.v.'s and $\mu = \sum_i \mathrm{E}[X_i]$.] *4pts*

Let the r.v. X denote the number of positions where the two strings agree. Then $X = \sum_{i=1}^{\ell} X_i$ for independent indicator r.v.'s X_i with $\Pr[X_i = 1] = \frac{1}{a}$, and the expectation of X is $\mu = \frac{\ell}{a}$. Using the Chernoff bound in the form given in the hint, with $\mu + \lambda = \varepsilon\ell$ and therefore $\lambda = \ell(\varepsilon - \frac{1}{a})$, we get

$$\Pr[X \geq \varepsilon\ell] \leq \exp\{-\tfrac{2\lambda^2}{\ell}\} = \exp\{-2\ell(\varepsilon - \tfrac{1}{a})^2\}.$$

(Note that the assumption $\varepsilon > \frac{1}{a}$ implies that $\lambda > 0$, which is required for the bound to hold.)

(b) Now suppose we choose m independent strings u.a.r. as in part (a), and let the length of the strings be $\ell = C \ln m$. Show how to choose the constant C (in terms of ε and a only) so that the probability that *any* pair of the strings agree at more than $\varepsilon\ell$ positions tends to zero as $m \to \infty$. *4pts*

We take a union bound over all $\binom{m}{2}$ pairs of strings, and use the upper bound of part (a) for each pair. This gives us the following upper bound on the probability that any pair agrees at more than $\varepsilon\ell$ positions:

$$\tbinom{m}{2} \exp\{-2\ell(\varepsilon - \tfrac{1}{a})^2\} \leq m^2 \exp\{-2C(\varepsilon - \tfrac{1}{a})^2 \ln m\},$$

where we have set $\ell = C \ln m$. Now we just need to choose C large enough so that the exponential kills the m^2 factor. Specifically, if we choose $C = (1 + \delta)(\varepsilon - \frac{1}{a})^{-2}$ for some $\delta > 0$, the above expression becomes $m^2 \exp\{-2(1 + \delta) \ln m\}$, which tends to 0 as $m \to \infty$.

(c) Why does part (b) guarantee the existence of (m, ℓ, ε)-codes with strings of length $\ell = O(\log m)$? *1pt*

Part (b) tells us that if we pick m random strings of length $\ell = C \ln m$ (for a specific C that depends on ε and a), then these strings will form a (m, ℓ, ε)-code with non-zero (actually, with high) probability. Thus, by the probabilistic method, there must exist (m, ℓ, ε)-codes with $\ell = C \ln m$.

[continued overleaf]

11. Unfair Games [12pts total]

Consider random walk on the integers with a *drift* downwards; i.e., if the process is currently at position z, it moves to $z+1$ with probability $p < \frac{1}{2}$ and to $z-1$ with probability $q = 1-p$. This process models the capital of a gambler repeatedly playing an *unfair* game, where she is more likely to lose than win, with a fixed stake of \$1 each time. We suppose that the process starts at 0 and ends at either of the endpoints $\pm m$ (i.e., when the gambler has either won or lost \$$m$ in total).

(a) Let X_t denote the position of the process at time t (so that $X_0 = 0$). Show that the process $Z_t := (q/p)^{X_t}$ is a martingale w.r.t. (X_t). *3pts*

Clearly Z_t is a function of X_t and $\mathrm{E}[|Z_t|]$ is finite for all t. Thus we just need to check that $\mathrm{E}[Z_t \mid X_0, \ldots, X_{t-1}] = Z_{t-1}$. We compute:

$$\begin{aligned} \mathrm{E}[Z_t \mid X_0, \ldots, X_{t-1}] &= p \times (q/p)^{(X_{t-1}+1)} + q \times (q/p)^{(X_{t-1}-1)} \\ &= p \times (q/p) \times Z_{t-1} + q \times (p/q) \times Z_{t-1} \\ &= (q+p) Z_{t-1} \\ &= Z_{t-1}. \end{aligned}$$

[A common error here was to write $\mathrm{E}[Z_t \mid X_0, \ldots, X_{t-1}] = (q/p)^{\mathrm{E}[X_t | X_0, \ldots, X_{t-1}]}$, which of course is just wrong and makes it impossible to obtain the desired property.]

(b) Use the Optional Stopping Theorem to show that the probability that the process ends at $-m$ (i.e., the gambler goes bankrupt) is given by *3pts*

$$p_{\text{bank}} = \frac{\rho^m - 1}{\rho^m - \rho^{-m}},$$

where $\rho > 1$ denotes the ratio $\frac{q}{p}$.

We use the OST with the stopping time $T =$ first time at which the process hits $\pm m$. The conditions for the OST are satisfied since $|Z_t| \leq \rho^m$ is bounded for all t, so we get $\mathrm{E}[Z_T] = Z_0 = 1$. Writing out $\mathrm{E}[Z_T]$ explicitly we therefore have

$$1 = \mathrm{E}[Z_T] = p_{\text{bank}} \times \rho^{-m} + (1 - p_{\text{bank}}) \times \rho^m,$$

and rearranging this gives us the desired expression for p_{bank}.

[Both here and in part (d), some people forgot to verify one of the conditions for the OST.]

[continued overleaf]

(c) Now show that the process $Y_t := X_t + (q-p)t$ is a martingale w.r.t. (X_t). *2pts*

Clearly Y_t is a function of X_t and $\mathrm{E}[|Y_t|]$ is finite for all t. We compute

$$\begin{aligned}\mathrm{E}[Y_t \mid X_0, \ldots, X_{t-1}] &= p \times (X_{t-1}+1) + q \times (X_{t-1}-1) + (q-p)t \\ &= (p+q)X_{t-1} + (p-q) + (q-p)t \\ &= Y_{t-1}.\end{aligned}$$

(d) Use the Optional Stopping Theorem again, along with part (b), to show that the expected number of steps until the process ends is given by *3pts*

$$\mathrm{E}[T] = \frac{m}{q-p}(2p_{\text{bank}} - 1).$$

We use the OST with the same stopping time as in part (b). Here the conditions of the OST are satisfied because $\mathrm{E}[T]$ is finite (as we have seen several times in class: T is the escape time of random walk on a bounded interval) and the expected jumps $\mathrm{E}[|Y_t - Y_{t-1}| \mid X_0, \ldots, X_{t-1}] \le 1 + (q-p)$ are bounded for all t. The OST tells us that $\mathrm{E}[Y_T] = Y_0 = 0$, so writing out $\mathrm{E}[Y_T]$ gives

$$\begin{aligned}0 = \mathrm{E}[Y_T] &= p_{\text{bank}} \times (-m + (q-p)\mathrm{E}[T]) + (1 - p_{\text{bank}}) \times (m + (q-p)\mathrm{E}[T]) \\ &= (q-p)\mathrm{E}[T] - m(2p_{\text{bank}} - 1).\end{aligned}$$

Rearranging gives us the claimed expression for $\mathrm{E}[T]$.

(e) Discuss briefly what the results of parts (b) and (d) tell you about unfair games. *1pt*

Fix the odds p, q (and therefore $\rho = \frac{q}{p}$). Then assuming m is reasonably large, the probability p_{bank} of going bankrupt in part (b) is close to $1 - \rho^{-m}$, so you will go bankrupt with all but exponentially small probability (as a function of your capital m). Part (d) tells us that the expected length of the game (the amount of time you have to enjoy the casino!) scales approximately linearly with your capital m. (Recall that this scaling is quadratic in the case of a fair game $p = q = \frac{1}{2}$; but in that case the above analysis fails and we instead use the analysis we did in class.)

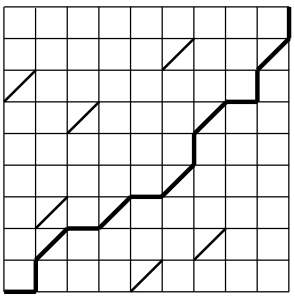
[We gave 1 point for any reasonable quantitative observation here.]

[continued overleaf]

12. Shortest Paths with Random Shortcuts [12 pts total]

Consider the $n \times n$ square grid whose points are $\{0, 1, \ldots, n\}^2$ and where there is an edge from each point to its (at most) four neighbors (N,S,E,W) in the grid. In addition, within each of the n^2 "squares" of the lattice, we add the SW-NE diagonal independently with probability p. We are interested in the length of a shortest path from the point $(0, 0)$ to the point (n, n). In the absence of the diagonals, this is of course exactly $2n$, but with the diagonals its expectation will be αn for some α in the range $[\sqrt{2}, 2]$ that depends on p (and which we will not calculate here).

Here is an example with $n = 9$; the unique shortest path is shown with bold lines.



(a) For $1 \le i, j \le n$, let X_{ij} be the Bernoulli r.v. for the event that the diagonal in the square with NE corner (i, j) is present. Denote by $f(X_{11}, \ldots, X_{nn})$ the shortest path from $(0, 0)$ to (n, n) for a particular setting of these r.v.'s. Show that f is $(2 - \sqrt{2})$-Lipschitz. *2pts*

Toggling the value of X_{ij} between 0 and 1 just corresponds to removing or adding the diagonal in the corresponding square. Any path that uses a given diagonal can be modified by replacing the diagonal by two sides of the square, thus increasing the distance by $2-\sqrt{2})$. Symmetrically, adding any diagonal can reduce the length of a path by at most the same amount. Hence f is $(2 - \sqrt{2})$-Lipschitz.

(b) Prove that *3pts*

$$\Pr[|f - \mathrm{E}[f]| \ge \lambda] \le 2 \exp\Big(-\frac{\lambda^2}{2(2-\sqrt{2})^2 n^2}\Big).$$

[NOTE: You should use Azuma's inequality in the form $\Pr[|Z_t - Z_0| \ge \lambda] \le 2\exp(-\lambda^2/2\sum_i c_i^2)$. Specify a martingale and justify your reasoning!]

We use the Doob martingale of f. Order the indices (i, j) in some arbitrary fashion, say left-to-right, top-to-bottom, starting with $(1, 1)$ and ending with (n, n). The Doob martingale of f is then $Z_{ij} := \mathrm{E}[f(X_{11}, \ldots, X_{nn}) \mid X_{11}, \ldots, X_{ij}]$, and $Z_0 := \mathrm{E}[f]$. Note that $Z_{nn} = f$. Since the X_{ij} are independent, and f is $(2 - \sqrt{2})$-Lipschitz by part (a), we may apply Azuma's inequality with bounded differences $c_i = 2 - \sqrt{2}$ for all i to get exactly the claimed bound on $\Pr[|f - \mathrm{E}[f]| \ge \lambda]$. (Note that the length of the martingale is n^2.)

[Many people forgot to mention the crucial fact that the X_{ij} are independent; as we saw in class and on HW, without this fact the Lipschitz condition cannot be used in Azuma's inequality.]

[continued overleaf]

(c) Now let $Y_i := (X_{1i}, \ldots, X_{ni})$ denote the collection of the above Bernoulli random variables in the ith column of the grid, and define the function $g(Y_1, \ldots, Y_n)$ to be the length of the shortest path from $(0, 0)$ to (n, n). (Thus g has the same value as f for any given realization of the diagonal edges, but is viewed as a function of a different set of variables.) Show that g is also $(2 - \sqrt{2})$-Lipschitz. *3pts*

The key observation here is that any shortest path from $(0, 0)$ to (n, n) can make use of at most one diagonal in any column (since once the path has crossed the column from left to right, it will never cross it again). Thus changing the value of any vector of values Y_i (which corresponds to toggling the presence/absence of any of the diagonals in the ith column) can increase/decrease the length of a shortest path by at most $2 - \sqrt{2}$, by the same reasoning as in part (a). Hence g is $(2 - \sqrt{2})$-Lipschitz.

(d) Show how to improve the bound of part (b) to the following: *2pts*

$$\Pr[|g - \mathrm{E}[g]| \geq \lambda] \leq 2 \exp\Big(-\frac{\lambda^2}{2(2 - \sqrt{2})^2 n}\Big).$$

Again, as in part (b), the Y_i are independent and $(2 - \sqrt{2})$-Lipschitz. However, the length of the martingale is now only n (since there are n columns). Thus Azuma's inequality gives us the claimed bound, which is the same as in part (b) with n^2 replaced by n in the denominator of the exponent.

(e) Why is the bound in part (d) much more useful than the one in part (b)? *2pts*

The bound in part (b) implies that f is concentrated in an interval of width $O(n)$ around its expectation, while the bound in part (d) reduces this width to $O(\sqrt{n})$. Since the mean shortest path length $\mu = \mathrm{E}[f] = \mathrm{E}[g] = \alpha n$, the bound in part (b) is essentially useless (the concentration width is of the same order as the mean itself), while the bound in part (d) implies good concentration within an interval of width $o(\mu)$.

[Many people failed to note the key point that the concentration interval in part (b) is of the same order as the mean itself, and therefore useless.]

[continued overleaf]

13. Another Construction of 2-Universal Hash Functions [10pts total]

Suppose we want to construct a family of 2-universal hash functions from the set $U = \{0,1\}^k$ to the set $T = \{0,1\}^\ell$, where $k > \ell$.

(a) Let $\mathcal{M}$ denote the set of all $k \times \ell$ 0,1 matrices. For $A \in \mathcal{M}$, define the function $h_A : U \to T$ by $h_A(u) = uA \bmod 2$, and define $\mathcal{H} = \{h_A : A \in \mathcal{M}\}$. Prove that $\mathcal{H}$ is a 2-universal family of hash functions from U to T. [NOTE: You do *not* need to show that the functions h_A map points in U uniformly into T, only that they are 2-universal.] *5pts*

First recall the definition of 2-universal hash families. To show that the family $\mathcal{H}$ is 2-universal, we need to prove that, for any $u, v \in \{0,1\}^k$ with $u \neq v$, $\Pr[h_A(u) = h_A(v)] \leq \frac{1}{2^\ell}$, where the probability is taken over the choice of the hash function $h_A \in \mathcal{H}$. Using the definition of h_A, this translates to the condition that $uA = vA \bmod 2$, where the probability is over the choice of a random matrix $A \in \mathcal{M}$.

Now note that $uA = vA \bmod 2$ is equivalent to $[(u-v)A]_i \bmod 2 = 0$ for $1 \leq i \leq \ell$. Now fix i, and note that $[(u-v)A]_i$ is the dot product of $(u-v)$ with the ith column of A. Since $u \neq v$, there is at least one coordinate, say j, for which $(u-v)_j \neq 0$. We can now use the Principle of Deferred Decisions to first randomly choose all elements of the ith column of A except for the jth element; given these choices, there is then a unique choice (0 or 1) for the remaining element A_{ji} that makes $(u-v)A \bmod 2 = 0$. And since A_{ji} is independent of the other entries in the column, this implies that $\Pr[[(u-v)A]_i \bmod 2 = 0] = \frac{1}{2}$.

Finally, we note that the events $[(u-v)A]_i \bmod 2 = 0$ for $1 \leq i \leq \ell$ are independent, so $\Pr[uA = vA \bmod 2] = \frac{1}{2^\ell}$, as desired.

[In both this part and part (b), many people either didn't mention the role of independence at all, or were not careful about it. This is particularly important in part (b), where the matrix entries are not all mutually independent.]

(b) Now consider the following alternative construction. Choose a random $k \times \ell$ 0,1 matrix by picking its first row and column independently and uniformly at random. Then, copy each of these elements along its corresponding diagonal. For example, if $k = 5$ and $\ell = 4$ and the first row and column are as shown in bold, then the resulting matrix is *4pts*

$$A = \begin{pmatrix} \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & 0 & 1 & 1 \\ \mathbf{1} & 1 & 0 & 1 \\ \mathbf{0} & 1 & 1 & 0 \\ \mathbf{1} & 0 & 1 & 1 \end{pmatrix}$$

Prove that this family of matrices also defines a 2-universal family of hash functions from U to T.

The argument here is essentially the same as in part (a), except that we have to be more careful about independence. Note that, with this new construction, it is still the case that the entries of each column of A are mutually independent, since each of them is derived from a different random "seed" bit inserted into the first row or column. Thus the argument in part (a) that $\Pr[[(u-v)A]_i \bmod 2 = 0] = \frac{1}{2}$ still holds, since the element A_{ji} on which the argument hinges is chosen independently of all the other elements in the ith column.

The second place in part (a) where we used independence was in the final step, where we argued that the events $[(u-v)A]_i \bmod 2 = 0$ for $1 \leq i \leq \ell$ are independent. But we can see that this also still holds with the new construction, since each of these events is determined by the entry A_{ji} for some fixed j as i varies, i.e., by the entries of the jth row of A. But, for the same reason as in the previous paragraph, the entries of each row of A are also mutually independent.

(c) Why is the family in part (b) preferable to the family in part (a)? *1pt*

Both families achieve pairwise independence. However, the construction in part (a) requires $k\ell$ random bits, while that in part (b) requires only $k + \ell - 1$ random bits.

[continued overleaf]

14. The Inverse Riffle Shuffle [12pts total]

Consider the following rather strange way of shuffling a deck of n cards. Write an independent, uniformly random binary digit (0 or 1) on the back of each card. Then pull out all the cards with a 0 and place them on top of all the cards with a 1, keeping the relative orderings of both sets of cards fixed. Repeat this process, each time appending a new random binary digit to the sequence of digits on the back of each card and pulling out the cards whose latest digit is 0. You may assume without proof that this process converges to the uniform distribution over all permutations of the deck. In this problem we analyze its mixing time.

(a) Let (X_t) and (Y_t) denote separate copies of the above process, starting from arbitrary initial permutations X_0 and Y_0. Propose a coupling between (X_t) and (Y_t) that attempts to minimize the time T until $X_T = Y_T$. *4pts*

Consider the following coupling: in both (X_t) and (Y_t), we give the same random digit to each card (regardless of its positions in the two copies of the deck). More specifically, we can let (X_t) choose independent random digits for each card, and then let (Y_t) simply copy the choices of (X_t) for each card. Clearly this is a valid coupling, since when we view each of (X_t) and (Y_t) separately, each card receives digit 0 or 1 independently and u.a.r.

[Some people designed a "coupling" in which the evolution of Y_t (and sometimes even of X_y) was not equivalent to the original shuffling process. This is of course not a valid coupling.]

(b) Let T be the first time at which the sequences of digits on the backs of the cards in (X_T) (interpreted as binary numbers with most significant digit written last) are all distinct. Show that $X_T = Y_T$. [NOTE: If the coupling you proposed in part (a) does not have this property, you should go back and revise your coupling.] *3pts*

The key observation here is that the shuffle performs a *sorting* operation on the binary numbers on the backs of the cards: specifically, if one card has a smaller binary number than another card, then the first card will be placed above the other in the deck. (Cards with the same number will remain in the same relative order as they started.) Thus, if we wait until the numbers on the backs of all cards are distinct, there is a unique ordering of the cards given by the ordering of those numbers. Since our coupling in part (a) ensures that each card has the same number in both (X_t) and (Y_t), the ordering of both decks must be the same, so $X_T = Y_T$.

[continued overleaf]

(c) Show that the random time in part (b) satisfies $\Pr[T \geq 2\log_2 n + \log_2(1/2\varepsilon)] \leq \varepsilon$. [HINT: You may use the fact from class that, in the birthday problem with n people and $m \geq \frac{c}{2}n^2$ possible birthdays, the probability that any two people have the same birthday is at most $\frac{1}{c}$.] *3pts*

Suppose we have run the process for t steps. The event that the numbers on the backs of the cards are all distinct is equivalent to the event that there are no common birthdays in a birthday problem with n people and $m = 2^t$ birthdays. Following the hint, if we want this probability to be at most $\frac{1}{c} = \varepsilon$, we need $m \geq \frac{c}{2}n^2$, or equivalently, $2^t \geq \frac{1}{2\varepsilon}n^2$. Taking logs gives us exactly the claimed condition for T.

(d) Deduce that the mixing time of this shuffle satisfies $\tau(\varepsilon) \leq 2\log_2 n + \log_2(1/2\varepsilon)$. *1pt*

This follows immediately from part (c) once we recall from the Coupling Lemma that the mixing time $\tau(\varepsilon)$ satisfies $\tau(\varepsilon) = \min_T\{\Pr[X_T \neq Y_T] \leq \varepsilon\}$, for any coupling.

(e) How does this mixing time bound compare with those that we derived in class for other shuffles (such as random-to-top and random transpositions)? Is this surprising? *1pt*

In class and homeworks we also discussed the random-to-top shuffle, whose mixing time is $\Theta(n \log n)$, and the random transposition shuffle, whose mixing time is $\Theta(n^2)$. The shuffle analyzed here has an exponentially smaller mixing time of $O(\log n)$. Intuitively, this is explained by the fact that this shuffle moves the entire deck of cards at each step, rather than just one or two cards as in the other shuffles.

NOTE: It turns out that the inverse riffle shuffle analyzed here has exactly the same mixing time as the riffle shuffle itself that we defined (but didn't analyze) in class! Indeed, this inverse version is by far the cleanest way to analyze the original riffle shuffle. The fact that they are equivalent follows from the fact that both are random walks on the group of permutations, and one is (in a group-theoretic sense) the inverse of the other. (This latter fact should be at least intuitively clear from the definitions of the two shuffles.)

[The End]