Exam 1

CSCI 4030 / CSCI 6220: Randomized Algorithms

Fall 2026

Rensselaer Polytechnic Institute

Name: _____ **RIN:** _____

Instructions. Answer all questions. You have 100 minutes. You may use one 8.5×11 inch double-sided crib sheet. No electronic devices are permitted. Show your reasoning clearly; unsupported answers may receive little or no credit. You may use any of the inequalities below without proof.

Useful inequalities. For any nonnegative random variable X and $t > 0$,

$$\Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t} \qquad \text{(Markov)}.$$

For any random variable X with finite variance and $t > 0$,

$$\Pr\big(|X - \mathbb{E}[X]| \geq t\big) \leq \frac{\mathrm{Var}(X)}{t^2} \qquad \text{(Chebyshev)}.$$

For events $A_1, \ldots, A_m$,

$$\Pr\left(\bigcup_{i=1}^{m} A_i\right) \leq \sum_{i=1}^{m} \Pr(A_i) \qquad \text{(union bound)}.$$

If $X = \sum_i X_i$ is a sum of independent Bernoulli random variables and $\mu = \mathbb{E}[X]$, then for $0 < \delta \leq 1$,

$$\Pr\big(X \geq (1+\delta)\mu\big) \leq \exp\left(-\frac{\delta^2 \mu}{3}\right), \qquad \Pr\big(X \leq (1-\delta)\mu\big) \leq \exp\left(-\frac{\delta^2 \mu}{2}\right).$$

Sample Solution

[12 points] Problem 1 (How much randomness do we really need?).

A randomized-algorithms library was originally implemented using fully independent random choices. Because random bits and stored random seeds are resources too, an engineer wants to replace full independence by weaker randomness whenever the analysis still goes through.

For parts (a)–(c), identify the *weakest* assumption among

no independence, pairwise independence, 3-wise independence, full independence

that is sufficient for the stated argument. Give a brief justification.

(a) (2 points) The running time is $T = \sum_{i=1}^{n} X_i$, and the analysis only computes $\mathbb{E}[T]$ from the individual expectations $\mathbb{E}[X_i]$.

no independence : linearity of expectation does not require it

(b) (3 points) The analysis uses $\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$.

$\mathrm{Var}(\sum X_i) = \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$ and pairwise independence says the covariance terms vanish

(c) (3 points) Three specified distinct keys x, y, z are hashed into m buckets, and the analysis needs

$$\Pr\big(h(x) = h(y) = h(z)\big) = \frac{1}{m^2}.$$

3-wise independence gives this equality via a conditioning argument similar to that used to show 2-universality

(d) (4 points) A Las Vegas algorithm always returns the correct answer and has expected running time at most T. We convert it to a Monte Carlo algorithm by stopping it after $10T$ steps and returning FAIL if it has not finished. Give an upper bound on the probability that the new algorithm returns FAIL, and justify your answer.

$\mathbb{P}(\text{MC algo returns FAIL}) = \mathbb{P}(\text{LV algo runs longer than } 10T)$

$\leq \dfrac{\mathbb{E}[\text{LV running time}]}{10T}$ by Markov's inequality

$\leq \dfrac{1}{10}$

[12 points] Problem 2 (How often do we update the current best?).

A system has n candidate configurations with distinct performance scores. The configurations are tested in a uniformly random order.

The first configuration tested is deployed. After that, whenever a newly tested configuration has a higher score than *all configurations tested before it*, the system replaces the currently deployed configuration with the new one.

Let R be the total number of deployments, including the initial deployment.

(a) (7 points) For $i = 1, \ldots, n$, let

$$X_i = \begin{cases} 1, & \text{if the } i\text{th configuration tested causes a deployment,} \\ 0, & \text{otherwise.} \end{cases}$$

Express R in terms of $X_1, \ldots, X_n$, and compute $\Pr(X_i = 1)$.

Hint: The scores of the first i configurations tested are distinct and are in uniformly random order.

$R = \sum_{i=1}^{n} X_i$ counts the total number of deployments over the n configurations

$\mathbb{P}(X_i = 1) = \mathbb{P}($the score of configuration i is the best of i distinct, randomly ordered scores$)$

$= \frac{1}{i}$

(b) (5 points) Compute $\mathbb{E}[R]$ exactly as a sum, and give its asymptotic order in n.

$$\mathbb{E}[R] = \sum_{i=1}^{n} \mathbb{P}(X_i = 1) = \sum_{i=1}^{n} \frac{1}{i} = H_n$$

so $\mathbb{E}[R] = \Theta(\ln n)$

[28 points] Problem 3 (Keeping a server farm balanced).

A service has n identical servers and receives n jobs. As a very low-overhead routing rule, each arriving job chooses one server independently and uniformly at random. Let L_j be the final load of server j.

You may use, without proof, the standard estimate

$$\binom{n}{k} \le \left(\frac{en}{k}\right)^k, \qquad 1 \le k \le n.$$

(a) (6 points) Fix a server j. Show that for every integer $1 \le k \le n$,

$$\Pr(L_j \ge k) \le \binom{n}{k}\left(\frac{1}{n}\right)^k \le \left(\frac{e}{k}\right)^k.$$

For the first inequality, use the following observation: if $L_j \ge k$, then there is some set of k jobs that all chose server j.

$\mathbb{P}(L_j \ge k) = \mathbb{P}(\text{a set of at least } k \text{ jobs is assigned to server } j)$
$= \mathbb{P}(\text{some set of exactly } k \text{ jobs is assigned to server } j)$
$= \sum_{\substack{S \subseteq [n] \\ |S|=k}} \mathbb{P}(\text{exactly } k \text{ jobs assigned to server } j)$
$= \binom{n}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{n-k} \le \binom{n}{k}\left(\frac{1}{n}\right)^k \le \left(\frac{en}{k}\right)^k\left(\frac{1}{n}\right)^k = \left(\frac{e}{k}\right)^k$

(b) (5 points) Suppose k is chosen so that

$$\left(\frac{e}{k}\right)^k \le \frac{1}{n^2}.$$

Deduce from part (a) that

$$\Pr\left(\max_{1\le j\le n} L_j \ge k\right) \le \frac{1}{n}.$$

$\mathbb{P}(\max_j L_j \ge k) = \mathbb{P}(\exists j : L_j \ge k) \le n\,\mathbb{P}(L_1 \ge k) \le n\left(\frac{e}{k}\right)^k$
$\le n \cdot \frac{1}{n^2} = \frac{1}{n}$

Remark. For all sufficiently large n, $k = \lceil 4 \log n / \log\log n \rceil$ satisfies the displayed condition. One way to verify this is to use $\log k - 1 \ge \frac{1}{2}\log\log n$. Thus one random choice gives maximum load $O(\log n / \log\log n)$ with high probability.

(c) (3 points) Now change the routing rule: each arriving job samples $d \geq 2$ servers independently and uniformly at random and joins a least-loaded sampled server. After t jobs have been placed, suppose at most a β_i fraction of the servers have load at least i. What is the probability that all d servers sampled by job $t+1$ have load at least i?

$\mathbb{P}(\text{all } d \text{ servers have load} \geq i) = \mathbb{P}(\text{a server has load} \geq i)^d$
$\leq \beta_i^d$

(d) (6 points) For $i \geq 0$, let $B_i(t)$ be the number of servers with load at least i after t jobs have been placed, and write $B_i = B_i(n)$ for the final value. First show deterministically that

$$B_4 \leq \frac{n}{4}.$$

Set $\beta_4 = 1/4$ and, for $i \geq 4$, define

$$\beta_{i+1} = 2\beta_i^d.$$

Suppose that throughout the process $B_i(t) \leq \beta_i n$. Using part (c), explain why each arriving job has conditional probability at most β_i^d of being placed at load level at least $i+1$. Why does this make $\beta_{i+1} n = 2\beta_i^d n$ a natural threshold for B_{i+1}?

i) Assume otherwise, that $B_4 > \frac{n}{4}$. Then there are more than $\frac{n}{4}$ bins with at least 4 balls, so there are more than n balls in total. This is a contradiction, so $B_4 \leq \frac{n}{4}$.

ii) $\mathbb{P}(\text{~~all d~~ job placed at load level} \geq i+1 \mid B_i \leq \beta_i n)$
$= \mathbb{P}(\text{all } d \text{ servers it considers has load} \geq i \mid B_i \leq \beta_i n)$
$\leq \beta_i^d$

We choose that threshold for B_{i+1} because by concentration (of a non-independent) Bernoulli sum, the number of jobs placed at load level $\geq i+1$ is less than $2\beta_i^d n$ with high probability conditioned on $B_i \leq \beta_i n$

(e) (8 points) Define

$$E_i = \{B_i \le \beta_i n\}.$$

You may assume the concentration step from the lecture analysis: for every relevant level $i \ge 4$,

$$\Pr\left(E_{i+1}^c \mid E_4 \cap E_5 \cap \cdots \cap E_i\right) \le \frac{1}{n^3}.$$

Confirm that part (d) above implies that after

$$r = O\left(\frac{\log\log n}{\log d}\right)$$

additional levels, the recurrence above reaches a level r for which $\beta_r n < 1/2$.

Show that

$$\Pr(E_4 \cap E_5 \cap \cdots \cap E_r) \ge 1 - \frac{r-4}{n^3},$$

and use $\beta_r n < 1/2$ to deduce a high-probability upper bound on the maximum server load. State the resulting asymptotic dependence on n and d when we place a job on the least loaded of d uniformly randomly selected servers.

i) Part d takes $\beta_4 \le \frac{1}{4}$ and $\beta_{i+1} = 2\beta_i^d$. Plugging in gives

$\beta_{4+1} \le 2\beta_4^d$

$\beta_{4+2} \le 2(2\beta_4^d)^d = 2^{d+1}\beta_4^{d^2}$

$\beta_{4+3} \le 2(2^{d+1}\beta_4^{d^2})^d \le 2^{d^2+d+1}\beta_4^{d^3}$

and we see

$\beta_{4+r} \le 2^{d^{r-1}+\cdots+1}\beta_4^{d^r}$. Note $d^{r-1}+\cdots+1 = \Theta(d^r)$

so $\beta_{4+r} = \Theta\left(2^{d^r}\left(\frac{1}{4}\right)^{d^r}\right) = \Theta\left(\left(\frac{1}{2}\right)^{d^r}\right)$

Note that when $r = \Omega\left(\frac{\ln\ln n}{\ln d}\right)$, $d^r = \Omega(\ln n)$, so $\left(\frac{1}{2}\right)^{d^r} = O\left(\frac{1}{n}\right)$

so $\beta_r n < \frac{1}{2}$ for $r = \Omega\left(\frac{\ln\ln n}{\ln d}\right)$

ii) $\mathbb{P}(E_4 \cap \cdots \cap E_r) = 1 - \mathbb{P}$(there is an i for which E_i does not hold, but E_j holds for $j<i$)

$$\ge 1 - \sum_{i=5}^{r} \mathbb{P}\left(E_i^c \mid \bigcap_{j=4}^{i-1} E_j\right) \ge 1 - \sum_{i=5}^{r} \frac{1}{n^3} = 1 - \frac{r-4}{n^3}$$

This means in particular that when $\beta_r n \le \frac{1}{2}$,

$\mathbb{P}(E_r) = \mathbb{P}(B_r \le \beta_r n) = \mathbb{P}(B_r \le \frac{1}{2}) = \mathbb{P}(B_r = 0) \ge 1 - \frac{r-4}{n^3}$

so all bins have loading $O\left(\frac{\ln\ln n}{\ln d}\right)$ with high probability

[22 points] Problem 4 (Reliable estimates from cheap random seeds).

A streaming algorithm maintains a small randomized sketch of a massive data stream in order to estimate F_0, the number of distinct items in the stream. One copy of the sketch returns an estimate within a factor $1 \pm \varepsilon$ of F_0 with probability at least $3/4$.

To increase confidence, we run k copies of the sketch and return the median of their estimates. If more than half of the copies are successful, then the median is also a $(1 \pm \varepsilon)$-approximation. Generating k fully independent hash seeds is costly, so an engineer proposes using seeds whose resulting success indicators are only *pairwise independent*.

Let

$$X_i = \mathbf{1}\{\text{copy } i \text{ is successful}\}, \qquad X = \sum_{i=1}^{k} X_i.$$

You may assume k is even, and count $X = k/2$ as a failure.

(a) (4 points) Compute $\mathbb{E}[X]$. by linearity of expectation,

$\mathbb{E}[X] = k\mathbb{E}[X_1] = k\,\mathbb{P}(\text{copy 1 is successful}) \geq \frac{3k}{4}$

(b) (6 points) Compute $\mathrm{Var}(X)$. Be explicit about where pairwise independence is used.

$\mathrm{Var}(X) = \sum_{i=1}^{k} \mathrm{Var}(X_i) + \underbrace{\sum_{i \neq j} \mathrm{Cov}(X_i, X_j)}_{=0 \text{ by pairwise independence}}$

and $\mathrm{Var}(X_i) = p(1-p)$ where p = prob of a successful copy $\geq 3/4$

$\leq \frac{3}{4} \cdot \frac{1}{4} = \frac{3}{16} \Rightarrow \boxed{\mathrm{Var}(X) \leq 3k/16}$

(c) (6 points) Use Chebyshev's inequality to show that the median estimate fails with probability at most $3/k$.

$\mathbb{P}(\text{median estimate fails}) = \mathbb{P}(X \leq k/2) = \mathbb{P}(X - \mathbb{E}X \leq \frac{k}{2} - \mathbb{E}X)$

$\leq \mathbb{P}(X - \mathbb{E}X \leq \frac{k}{2} - \frac{3k}{4}) = \mathbb{P}(X - \mathbb{E}X \leq -\frac{k}{4})$

$\leq \mathbb{P}(|X - \mathbb{E}X| \geq \frac{k}{4}) \leq \frac{\mathrm{Var}(X)}{(k/4)^2} \leq \frac{3k/16}{k^2/16} = \frac{3}{k}$

(d) (4 points) If the copies were fully independent, use the supplied Chernoff bound to show that the median estimate fails with probability at most $\exp(-k/24)$.

$\mathbb{P}(X \le k/2) = \mathbb{P}(X \le (1-\delta)\frac{3k}{4})$ where we choose δ s.t. $(1-\delta)\frac{3k}{4} = \frac{k}{2} \iff \delta = \frac{1}{3}$

$\le \mathbb{P}(X \le (1-\delta)\mathbb{E}X)$

$\le \exp\left(\frac{-\delta^2 \mathbb{E}X}{2}\right) \le \exp\left(\frac{-\delta^2 3k/4}{2}\right) = \exp\left(\frac{-k}{24}\right)$

(e) (2 points) Suppose the target failure probability is at most $1/n$. Determine, up to constant factors, the number of copies k sufficient under pairwise independence and under full independence. Which assumption reaches the target with asymptotically fewer copies?

Under pairwise independence

$\mathbb{P}(\text{failure}) \le \frac{3}{k} \le \frac{1}{n} \iff k = \Omega(n)$

Under full independence

$\mathbb{P}(\text{failure}) \le \exp\left(\frac{-k}{24}\right) \le \frac{1}{n} \iff k = \Omega(\ln n)$

full independence requires exponentially fewer copies

[26 points] Problem 5 (A hash table that looks random locally but can fail globally).

A dictionary stores p distinct keys, where $p \geq 3$ is prime, in p buckets using separate chaining. Storing a truly random hash function is impractical, so we want a compact randomized hash family.

Part I: What does a universal-family guarantee buy us?

Suppose first that h is drawn from a universal family, so that for every pair of distinct keys x, y,

$$\Pr\big(h(x) = h(y)\big) \leq \frac{1}{p}.$$

For a fixed stored key x, let C_x be the number of stored keys in the same chain as x, including x itself.

(a) (4 points) Express C_x as 1 plus a sum of collision indicators.

$$C_x = 1 + \sum_{y \neq x} \mathbb{1}_{\{h(x) = h(y)\}}$$

(b) (4 points) Show that $\mathbb{E}[C_x] < 2$.

$$\mathbb{E}[C_x] = 1 + \sum_{y \neq x} \mathbb{P}(h(x) = h(y)) \leq 1 + (p-1) \cdot \frac{1}{p} < 2$$

(c) (3 points) Use Markov's inequality to give an explicit upper bound on $\Pr(C_x \geq 6)$.

$$\mathbb{P}[C_x \geq 6] \leq \frac{\mathbb{E}[C_x]}{6} < \frac{1}{3}$$

(d) (4 points) An engineer concludes: "Part (c) shows that every chain is shorter than 6 with high probability." Determine the upper bound that parts (a)–(c) actually give on

$$\Pr\left(\max_x C_x \geq 6\right).$$

Explain why this bound does *not* establish the engineer's claim.

$\mathbb{P}[\max_x C_x \geq 6] \leq p\,\mathbb{P}(C_1 \geq 6)$ by a union bound

$< \frac{p}{3}$

This is not even a probability when $p > 3$, so it is certainly not a high probability guarantee (that would usually be a bound of the form $\mathbb{P}(\max_x C_x \geq 6) \leq \frac{1}{\text{poly}(p)}$)

Problem 5 continued.

Part II: Pairwise-random outputs can hide global structure.

Now identify both the keys and buckets with the finite field $\mathbb{F}_p$ and choose a, b independently and uniformly from $\mathbb{F}_p$. Define

$$h_{a,b}(x) = ax + b \pmod{p}.$$

(e) (5 points) Let $x \neq y$ and fix arbitrary buckets $u, v \in \mathbb{F}_p$. Show that there is exactly one pair $(a, b) \in \mathbb{F}_p^2$ satisfying

$$ax + b = u, \qquad ay + b = v.$$

Deduce that $h(x)$ and $h(y)$ are independent and uniform over $\mathbb{F}_p$.

$ax+b=u$, $ay+b=v$ $\iff \begin{bmatrix} x & 1 \\ y & 1 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix}$ and $x \neq y$ means $\begin{bmatrix} x & 1 \\ y & 1 \end{bmatrix}$ is rank 2 so invertible

$\iff$ there is one unique pair (a, b) solving these equations

This implies

$$\mathbb{P}_{a,b}(h_{a,b}(x)=u, h_{a,b}(y)=v) = \frac{\#\text{seeds } (a,b) \text{ for which } h_{a,b}(x)=u \text{ \& } h_{a,b}(y)=v}{\#\text{seeds in } \mathbb{F}_p \times \mathbb{F}_p} = \frac{1}{p^2}$$

Since $(h(x), h(y))$ is uniformly distributed over $\mathbb{F}_p \times \mathbb{F}_p$, $h(x) \perp h(y)$ and both are uniform over $\mathbb{F}_p$

(f) (3 points) What happens to the hash table when $a = 0$? Determine the probability of this event and use it to give a lower bound on

$$\Pr(\text{maximum chain length} = p).$$

If $a=0$ then all p inputs hash to $h_{a,b}(x) = 0 \cdot x + b = b$, so

$$\mathbb{P}(\text{maximum chain length} = p) \geq \mathbb{P}(a=0) = \frac{1}{p}$$

because a is uniformly sampled on $\mathbb{F}_p$

(g) (3 points) For a fixed bucket j, write

$$L_j = \sum_{x \in \mathbb{F}_p} \mathbf{1}\{h(x) = j\}.$$

The summands are pairwise independent. Identify the exact step in the standard MGF proof of a Chernoff bound at which replacing full independence by pairwise independence is not sufficient, and explain why.

To prove that L_j concentrates, standard Chernoff bounds use the fact that

$$M_{L_j}(\lambda) = \prod_{x \in \mathbb{F}_p} M_{\varphi_x}(\lambda)$$

where $L_j = \sum_{x \in \mathbb{F}_p} \varphi_x$ (so $\varphi_x = \mathbb{1}_{\{h(x)=j\}}$).

This convenient factorization of the MGF of L_j requires full independence of the summands φ_x.

Thus pairwise independence is not sufficient for a standard Chernoff bound to apply.