Homework 7

CSCI 4030 / CSCI 6220: Randomized Algorithms — Fall 2026

Rensselaer Polytechnic Institute — 110 points

Justify every claim. Acknowledge collaborators and external sources, including AI tools, and explain their role in your submission. Throughout, log is the natural logarithm and $\log_2$ is the base-2 logarithm. Ignore integer rounding only when it does not affect the argument.

The first three problems concern how much independence a randomized algorithm actually needs. The last two use the Local Lemma: first to choose Wi-Fi channels, then to improve a randomized-rounding guarantee for network routing.

[10 points] Problem 1 (k-wise independent polynomial hashing).

Many machines need to compute the same hash value from a key. A fully random function $h : \mathbb{F}_P \to \mathbb{F}_P$ requires a table of P values; a short random seed is much easier to distribute. We will prove that a polynomial with k random coefficients behaves like a fully random function whenever we examine at most k distinct keys.

Let P be prime, let $1 \le k \le P$, and let $\mathbb{F}_P = \{0, 1, \dots, P-1\}$ with arithmetic modulo P. Choose $A = (A_0, \dots, A_{k-1})$ uniformly from $\mathbb{F}_P^k$ and define

$$p_A(x) = A_0 + A_1 x + \cdots + A_{k-1} x^{k-1} \qquad \text{in } \mathbb{F}_P.$$

Our goal is to prove that for distinct $\alpha_1, \dots, \alpha_k$ and arbitrary $y_1, \dots, y_k \in \mathbb{F}_P$,

$$\mathbb{P}[p_A(\alpha_1) = y_1, \dots, p_A(\alpha_k) = y_k] = P^{-k}.$$

You may use these facts over a field: a square matrix is invertible exactly when its nullspace is $\{0\}$, and a nonzero polynomial of degree at most d has at most d distinct roots.

(a) (4 points) For distinct $\alpha_1, \dots, \alpha_k$, consider

$$V = \begin{pmatrix} 1 & \alpha_1 & \alpha_1^2 & \cdots & \alpha_1^{k-1} \\ 1 & \alpha_2 & \alpha_2^2 & \cdots & \alpha_2^{k-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \alpha_k & \alpha_k^2 & \cdots & \alpha_k^{k-1} \end{pmatrix}.$$

For $c = (c_0, \dots, c_{k-1})^{\mathsf{T}}$, interpret Vc as evaluations of a polynomial. Use that interpretation to prove V is invertible.

(b) (4 points) Fix the α_i and the y_i. Show that exactly one seed A satisfies $p_A(\alpha_i) = y_i$ for every i, and deduce the displayed probability. If $k < P$, show that evaluations at $k+1$ distinct points are not independent: explain why the first k evaluations determine the last one.

(c) (2 points) A cluster of M machines needs the same hash function. Compare the number of bits needed to describe (i) a completely random function $\mathbb{F}_P \to \mathbb{F}_P$ and (ii) a polynomial hash function from this problem. If the description is sent separately to each machine, how many bits are transmitted altogether?

[20 points] Problem 2 (Distributing objects across cache servers).

A content delivery network assigns an object to a cache server by hashing its ID. Every front end can then find the object without consulting a directory. Assume for this problem that all objects require the same storage and generate requests at the same rate, so the number of assigned objects measures a server's load. We want to control the largest load.

There are n distinct objects and n servers. With independent uniform assignments, the maximum load is $\Theta(\log n / \log\log n)$ with high probability (the balls-into-bins experiment). Instead, we use the short-seed hash function

$$h(x) = \big(p_A(x) \bmod P\big) \bmod n, \qquad p_A(x) = \sum_{r=0}^{k-1} A_r x^r \quad \text{in } \mathbb{F}_P,$$

where the object IDs are distinct elements of $\mathbb{F}_P$. Assume $n \geq 2$ and $P \geq kn$.

(a) (5 points) Let $x_1, \ldots, x_k$ be distinct objects and $b_1, \ldots, b_k$ arbitrary servers. Prove

$$\mathbb{P}[h(x_1) = b_1, \ldots, h(x_k) = b_k] \leq \frac{e}{n^k}.$$

(b) (7 points) Show that

$$\mathbb{P}[\text{some server has at least } k \text{ objects}] \leq \frac{en}{k!}.$$

Deduce that a sufficiently large constant multiple of $\log n / \log\log n$ suffices for maximum load less than k with probability at least $1 - 1/n$. For $n = 10^6$, find the smallest integer k for which the displayed upper bound is at most 10^{-6} and state the resulting load guarantee. You may use $\binom{n}{k} \leq n^k/k!$.

(c) (6 points) Now set $k = 2$. Show that, with probability at least $1/2$, the maximum load is $O(\sqrt{n})$. Your argument should use only the independence and marginal-distribution properties established above.

(d) (2 points) A colleague argues that the weaker guarantee in (c) merely reflects a weaker analysis, and that pairwise-independent hashing may perform just as well in practice. Do your results support or contradict this claim? Explain.

Further context. Pătraşcu and Thorup (2012) proved that simple tabulation hashing, although only 3-wise independent, has strong load guarantees. Independence assumptions are sufficient conditions, not a complete description of a hash family's behavior.

[30 points] Problem 3 (Balanced A/B tests with a short seed).

A website is testing a redesigned checkout page. The N users are assigned either to treatment (the new page, $x_i = +1$) or control (the old page, $x_i = -1$). User i has fixed potential outcomes u_i and v_i under the two pages; only the assignment is random. We observe

$$Y_i = \begin{cases} u_i, & x_i = +1, \\ v_i, & x_i = -1, \end{cases} \qquad \tau = \frac{1}{N}\sum_{i=1}^{N}(u_i - v_i).$$

We estimate τ by $\hat{\tau} = (2/N)\sum_i x_i Y_i$. Part (a) investigates this estimator and relates its error to the composition of the two arms.

Users with similar characteristics may have systematically different purchasing behavior, regardless of which page they see. For example, returning customers may buy more frequently than first-time visitors. This suggests balancing not only the total number of users in each arm but also the composition of the arms across relevant groups. Let $A_1, \ldots, A_m \subseteq \{1, \ldots, N\}$ be these groups, called *segments*; take $A_1 = \{1, \ldots, N\}$. Examples include mobile users and returning users; segments may overlap. Define

$$S_j = \sum_{i \in A_j} x_i, \qquad \mathrm{imb}(x) = \max_{1 \le j \le m} |S_j|.$$

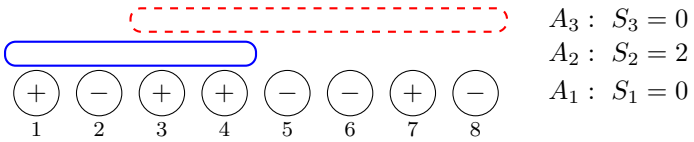


Figure 1: Eight assignments, with $A_2 = \{1, 2, 3, 4\}$ and $A_3 = \{3, 4, 5, 6, 7, 8\}$. The worst segment imbalance is 2.

A returning user must see the same page on later visits, even across servers. The servers therefore assign arms by hashing the user's ID with a shared seed. Assume the arriving users and their segments are fixed independently of that random seed.

(a) (6 points) Put $a_i = (u_i + v_i)/2$.

(i) Prove, for every assignment,

$$x_i Y_i = a_i x_i + \tfrac{1}{2}(u_i - v_i), \qquad \hat{\tau} - \tau = \frac{2}{N}\sum_i a_i x_i.$$

(ii) Show that $\hat{\tau}$ is unbiased if each x_i is a uniform sign; mutual independence is not needed.

(iii) Suppose typical outcomes are explained by segment memberships: $a_i = \sum_{j : i \in A_j} \beta_j$, where β_j is the additive effect of belonging to segment j. Prove

$$|\hat{\tau} - \tau| \le \frac{2\sum_j |\beta_j|}{N}\,\mathrm{imb}(x).$$

Explain why this inequality makes segment imbalance relevant to the accuracy of the experiment under the stated model.

(b) (4 points) First suppose the x_i are independent uniform signs. Prove $\mathbb{P}[|S_j| \geq \lambda] \leq 2e^{-\lambda^2/(2N)}$. Deduce that, with probability at least $1 - \delta$,

$$\mathrm{imb}(x) \leq \sqrt{2N \log(2m/\delta)}.$$

What estimation-error bound does this imply under the model in (a)(iii)?

(c) (6 points) User IDs are distinct elements of $\mathbb{F}_P$, where P is prime and $P \geq 2N$. Choose an even $t \leq N$ and use Problem 1's degree- $(t-1)$ polynomial. Set $x_i = +1$ when $p_A(\mathrm{ID}_i) \bmod 2 = 0$ and $x_i = -1$ otherwise.

(i) Show that $\mathbb{E}[x_i] = 1/P$ and that the signs are t-wise independent. Why are they not exactly unbiased?

(ii) Show that $\mathbb{E}[\hat{\tau}] - \tau = (2/(NP)) \sum_i a_i$. If $u_i, v_i \in [0, 1]$, bound the absolute bias by $2/P$.

(iii) Let $y_1, \ldots, y_N$ be fully independent signs with the same marginals as the x_i, and put $T_j = \sum_{i \in A_j} y_i$. Prove $\mathbb{E}[S_j^t] = \mathbb{E}[T_j^t]$. Explain why only t-wise independence is needed.

(d) (8 points) *Bounding a high moment.* For the independent biased signs y_i defined in (c), and every even $t \leq N \leq P$, prove

$$\mathbb{E}[T_j^t] \leq 2e \left(\frac{tN}{e} \right)^{t/2}.$$

You may use $\cosh \alpha \leq e^{\alpha^2/2}$ for real α. Your proof should justify how a bound on an exponential moment leads to a bound on the ordinary t-th moment.

(e) (6 points) *Choosing the independence.* Use (c) and (d) to derive an upper bound on $\mathrm{imb}(x)$ that holds with probability at least $1 - \delta$. Choose an even $t \leq N$ to make your bound as small as possible, and justify your choice. Assume N is large enough for the resulting choice. Compare your guarantee with (b), including the corresponding estimation-error bound under the segment model. How many random bits specify the assignment?

Further context. Large experimentation systems do assign users by hashing identifiers; see Tang et al. (2010) and Kohavi, Tang, and Xu (2020). Separate experiments use different salts or seeds. The underlying limited-independence concentration principle is developed more generally by Schmidt, Siegel, and Srinivasan (1995). The high-moment argument in (c)–(e) is a concrete instance of the limited-independence concentration principle: its proof only needs joint distributions of at most t signs, even though the sum may contain many more terms.

[25 points] Problem 4 (Wi-Fi channels: choosing the right bad events).

An apartment building has wireless access points (APs). Nearby APs using the same channel interfere. Represent the APs by a graph $G = (V, E)$, where $uv \in E$ means that u and v can interfere. Each AP v has a list $L(v)$ of exactly ℓ allowed channels. We seek a choice $X_v \in L(v)$ with $X_u \neq X_v$ on every edge.

Write Δ for the maximum degree. Greedy coloring succeeds when $\ell > \Delta$, but this ignores channel restrictions: many neighbors might have lists disjoint from $L(v)$. For $c \in L(v)$ define the number of neighbors that could compete for this particular channel,

$$d_c(v) = |\{u : uv \in E,\ c \in L(u)\}|, \qquad D = \max_v \max_{c \in L(v)} d_c(v).$$

Assume G has at least one edge. In parts (b)–(e), each AP independently chooses a channel uniformly from its list. We will compare two ways to describe conflicts under this random assignment.

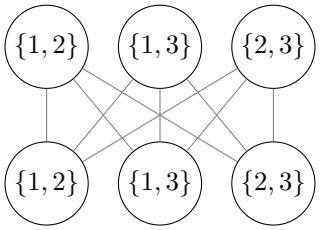


Figure 2: A $K_{3,3}$ with an allowed-channel list at each AP. Here $\ell = D = 2$.

(a) (4 points) Prove the instance in the figure has no valid channel assignment. Explain why this shows that $\ell \geq D$ is not sufficient to guarantee a valid assignment.

(b) (5 points) *Edge events.* For each edge uv, define $B_{uv} = \{X_u = X_v\}$. Compute its probability and identify the dependencies among these events. Use the Local Lemma to obtain a sufficient condition on ℓ and Δ for a valid assignment. Compare it with the greedy condition $\ell > \Delta$: does this approach improve on greedy coloring?

(c) (9 points) *Channel-specific events.* For each edge uv and each channel $c \in L(u) \cap L(v)$, define $A_{uv,c} = \{X_u = X_v = c\}$. Use the Local Lemma to derive an explicit sufficient condition for a valid channel assignment in terms of ℓ and D. Justify your dependency count carefully. Explain why the resulting condition may be useful even when $\ell \leq \Delta$.

(d) (3 points) Explain the difference between the two event descriptions. What information did the refined description exploit? When is Moser–Tardos resampling potentially useful even though greedy coloring cannot be guaranteed to succeed?

(e) (4 points) *Finding the assignment.* Assuming the condition from (c), describe a Moser–Tardos algorithm that finds a valid channel assignment. Show that the expected number of resampling steps is $O(|V|)$.

Further context. This is a list-coloring problem. The Local Lemma approach goes back to Reed (1999); Haxell (2001) proved that $\ell \geq 2D$ suffices, and Reed and Sudakov (2002) gave an asymptotic $(1 + o(1))D$ condition.

[25 points] Problem 5 (Randomized rounding for network routing).

A data-center network is a directed graph $G = (V, E)$ with m links. Connections $i = 1, \dots, N$ each need one path from s_i to t_i. For each connection we are given a finite, nonempty set $\mathcal{P}_i$ of candidate paths; assume at least one connection needs a nonempty path. The *congestion* is the maximum number of chosen paths using any link. Let $\mathrm{OPT} \geq 1$ be the smallest congestion achievable by choosing one path per connection.

An LP can first allow each connection to split fractionally over its paths:

$$\begin{aligned} \text{minimize} \quad & W \\ \text{subject to} \quad & \sum_{P \in \mathcal{P}_i} f_{i,P} = 1 && \text{for every connection } i, \\ & \sum_{i} \sum_{P \in \mathcal{P}_i : e \in P} f_{i,P} \leq W && \text{for every link } e, \\ & f_{i,P} \geq 0. \end{aligned}$$

Write $(f^\star, W^\star)$ for an optimal fractional solution. Fractions describe a benchmark, not a valid final choice: the rounded routing must select one entire path per connection.

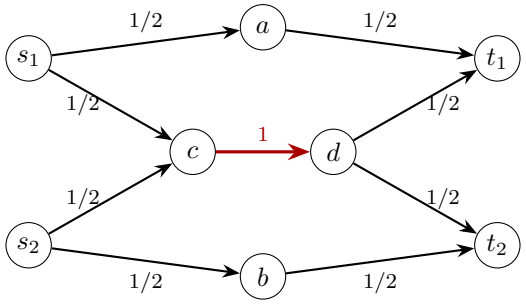


Figure 3: Each of two connections has a private path and a path using the shared link $c \to d$. Labels show link loads if each connection splits evenly; the shared link then carries one unit.

(a) (5 points) Determine the optimal fractional congestion $W^\star$ and the minimum integral congestion OPT for the network in the figure. Explain the difference. More generally, prove that $W^\star \leq \mathrm{OPT}$.

(b) (5 points) *Randomized rounding.* Independently for each connection i, choose one path $P \in \mathcal{P}_i$ with probability $f^\star_{i,P}$. For the network in the figure, determine the probability distribution of the resulting congestion when an optimal fractional solution is rounded. For a general link e, show that its load L_e is a sum of independent Bernoulli variables and $\mathbb{E}[L_e] \leq W^\star$.

(c) (6 points) *A global guarantee.* Put $\mu = \max\{W^\star, 1\} \leq \mathrm{OPT}$. You may use this Chernoff bound for a sum L of independent Bernoulli variables with $\mathbb{E}[L] \leq \mu$ and $\mu \geq 1$:

$$\mathbb{P}[L \geq c\mu] \leq (e/c)^c, \qquad c > 0.$$

Show that, for a suitable $c = O(\log m / \log\log m)$, the rounded routing has congestion less than $c\,\mathrm{OPT}$ with probability at least $1/2$ (for sufficiently large m).

(d) (7 points) *A local guarantee.* Suppose the union of all candidate paths for any one connection uses at most D links, and any link belongs to candidate paths of at most Δ connections. Use the Local Lemma to show that the congestion guarantee can depend on ΔD instead of m. In particular, for growing ΔD, show that a factor $c = O(\log(\Delta D)/\log\log(\Delta D))$ suffices; for bounded ΔD, a constant factor suffices. Specify the bad events and their dependencies in your proof. Then describe how Moser–Tardos finds such a routing and show that the expected number of resampling steps is $O(m)$.

(e) (2 points) Compare (d) with the elementary deterministic bounds that follow from the assumptions on candidate paths. Under what circumstances is the Local Lemma guarantee a substantial improvement?

Optional extension (no points). Construct an instance on which independent rounding achieves congestion $\Omega(\log m/\log\log m)$ with probability at least $1/2$ even though $W^\star = \mathrm{OPT} = 1$. You may use the balls-into-bins lower bound.

Further context. Randomized rounding for routing originates with Raghavan and Thompson (1987); Raghavan (1988) gives derandomization methods. A related Local Lemma application to packet scheduling appears in Leighton, Maggs, and Rao (1994).