CS 574: Randomized Algorithms, Fall 2026
Homework 4

Released: Friday, November 6, 2026

Due: Friday, November 20, 2026 at 11:59pm

Instructions & Course Policies:

- **Collaboration:** You may work in groups of up to three, but each student must write their solutions independently, in their own words. List all collaborators and external resources on the first page of your submission.
- **AI and other resources:** You are encouraged to use AI agents to help format your solutions, and may also consult them about the mathematics. Declare all textbooks, papers, websites, and AI tools consulted, and explain how you used them, in a footnote or bibliography.
- **Submission:** Submit a single typeset LaTeX PDF on Gradescope. Start each problem you submit on a new page.
- **Clarity and rigor:** Define your random variables and justify each use of independence, conditioning, and a tail inequality. You may use a preceding subpart's stated conclusion even if you have not proved it. Hints are suggestions, not additional requirements.
- **Review parts:** A part marked **Review** revisits an algorithm, result, or proof already developed in the lecture notes, possibly with minor variations. Its purpose is to check understanding: reconstruct the argument in your own words and justify the requested details; a new approach is not expected. You may consult and cite the notes, but citing a result alone does not replace a requested proof. Review parts carry the stated points and are required within the problem you choose. A **Review and extension** label identifies a part that also asks for an additional application or example.
- **Grading:** There are four problems, each worth 100 points. **You have to solve only one problem of your choice**, for a total of 100 points. Think about all four: they are fair game for the midterm.

Conventions. Write $[r] = \{1, \dots, r\}$ and $H_r = \sum_{j=1}^{r} 1/j$. Logarithms are natural unless a base is shown. All streams, point sets, and metrics are fixed independently of the algorithm's random choices. Different repetitions use independent randomness.

Lecture coverage. This assignment continues after Homework 3 and covers the teaching lectures from October 21 through November 6, inclusive (lectures 17 and 19–22): distinct-element estimation; finite metrics, random partitions, and embeddings; dimension reduction and the Johnson–Lindenstrauss lemma; entropy, randomness extraction, Huffman coding, and lossless compression. See the corresponding notes on the course lecture schedule. General background is available in [MR95, MU17].

Problem 1: Counting Distinct Elements

[100 points]

Let $x_1, \dots, x_m \in [U]$ be a nonempty insertion-only stream, where $U \geq 2$. Let S be its set of distinct keys and $D = |S|$. Assume $0 < \varepsilon, \delta < 1/2$. In (A)–(B) only, use an ideal random hash function whose values on distinct keys are independent uniform numbers in $(0, 1)$. Repeated occurrences of a key have the same hash value. Count stored real numbers and hash oracles in this ideal model; do not claim a finite-bit implementation for it.

(A) **[15 points] Review: The smallest hash.** Let $Y = \min_{v \in S} h(v)$. Derive its survival function, expectation, and variance, and prove

$$\mathbb{E}[Y] = \frac{1}{D+1}, \qquad \mathbb{V}[Y] = \frac{D}{(D+1)^2(D+2)}.$$

Give its streaming update rule. Is $1/Y - 1$ an unbiased estimator of D? Justify your answer, including whether its expectation is finite.

Hint. Use $\mathbb{P}[Y > y]$ for the moments of Y, and examine the density near zero for the reciprocal.

(B) **[20 points] Review: Invert after amplification.** Use independent copies of Y to obtain a number Z such that

$$\mathbb{P}[(1 - \varepsilon/4)/(D+1) \leq Z \leq (1 + \varepsilon/4)/(D+1)] \geq 1 - \delta.$$

Specify group sizes and the number of groups up to absolute constants. Prove that $\widehat{D} = 1/Z - 1$ is a $(1 \pm \varepsilon)$ approximation to D on this event, and give the space bound in the ideal model. Explain why (A)'s infinite expectation does not contradict this guarantee.

(C) **[20 points] Review: Sample keys with limited independence.** For a fixed $p \in (0, 1]$, suppose the indicators B_v, $v \in [U]$, are pairwise independent and satisfy $\mathbb{P}[B_v = 1] = p$. Let $R = \{v \in S : B_v = 1\}$ and $X = |R|$. Prove

$$\mathbb{E}[X] = pD, \qquad \mathbb{V}[X] = p(1-p)D,$$
$$\mathbb{P}[|X/p - D| > \varepsilon D] \leq \frac{1}{\varepsilon^2 pD}.$$

Describe how to maintain R with a cap of K stored keys: if a $(K+1)$st distinct sampled key arrives, discard the sample and mark this copy as overflowed permanently. Explain why overflow is equivalent to the final, uncapped count exceeding K.

(D) **[30 points] Choose a scale without assuming independence.** Set $L = \lceil \log_2 U \rceil$, $p_j = 2^{-j}$ for $0 \le j \le L$, and $K = \lceil 64\varepsilon^{-2} \rceil$. Run (C) at every scale, using pairwise independent indicators within each scale. You may assume each scale's indicators have an $O(\log U)$-bit description and can be evaluated in constant time. No independence between scales is required. At the end, choose the smallest nonoverflowed index J and return X_J/p_J; return zero if every scale overflowed.

Prove success probability at least $7/8$, with $O(\varepsilon^{-2} \log U)$ words in every execution. Then obtain failure probability at most δ using independent repetitions and state the resulting space and update-time bounds. One word holds a key or an $O(\log U)$-bit hash description; set operations may be implemented with balanced search trees.

Hint. If $D \le K/2$, scale zero is exact. Otherwise let j_* be the first index with $\mu_{j_*} = p_{j_*} D \le K/2$. Union bound the relative-error events only for $j \le j_*$. The sum $\sum_{j=0}^{j_*} 1/\mu_j$ is geometric. Do not apply a fixed-scale bound conditionally on the selected index.

(E) **[15 points] Merge summaries, not answers.** Two machines process streams with possibly overlapping key sets. Explain how to merge their summaries for both (B) and (D), so that the merged state is exactly what processing the concatenated streams would produce. Specify the shared randomness and the handling of overflow. Give a counterexample to adding the two final estimates, even if each local distinct count is known exactly.

Class notes: Approximating the Number of Distinct Elements in a Stream.

Problem 2: Random Partitions and Metric Embeddings

[100 points]

Let (X, d) be an n-point metric, $n \geq 2$, and let $0 < \delta < 1/2$. Use closed balls $B(x, r) = \{y : d(x, y) \leq r\}$. For a scale $\Delta > 0$, sample a uniform permutation of X and, independently, a uniform radius $R \in [\Delta/4, \Delta/2]$. Assign each point to the first center in the permutation within distance R; discard empty clusters. Write $P(x)$ for the cluster containing x.

(A) **[25 points] Review: A ball survives a partition.** Prove that every cluster has diameter at most Δ, and that for every $x \in X$ and $r > 0$,

$$\mathbb{P}[B(x, r) \not\subseteq P(x)] \leq \min\left\{1, \frac{8rH_n}{\Delta}\right\}.$$

Deduce a bound on the probability that a fixed pair x, y is separated.

Hint. Put $S = B(x, r)$. For each center z, define

$$a_z = \min_{w \in S} d(z, w), \qquad b_z = \max_{w \in S} d(z, w).$$

Order centers by a_z, breaking ties deterministically. The first ball to touch S can split it only if $R \in [a_z, b_z)$. Why is $b_z - a_z \leq 2r$, and why does the kth center have probability at most $1/k$ of being the first one to touch S, for a fixed $R \geq a_z$?

(B) **[15 points] Review and extension: A tree can stretch a cycle.** A hierarchically well-separated tree (HST) has leaves X, positive labels on internal nodes, zero labels on leaves, and each child's label at most half its parent's. Define $d_T(x, y)$ for distinct leaves as the label of their least common ancestor, and set $d_T(x, x) = 0$. Prove the strong triangle inequality

$$d_T(x, z) \leq \max\{d_T(x, y), d_T(y, z)\}.$$

For the shortest-path metric of the unweighted cycle C_n, $n \geq 3$, prove that every such d_T satisfying $d_T \geq d$ stretches some cycle edge by at least $\lfloor n/2 \rfloor$. Why does this not rule out a distribution of dominating trees with $\mathbb{E}[d_T(x, y)] \leq O(\log n) d(x, y)$ for every fixed pair?

(C) **[20 points] Review: Transfer a clustering solution.** You may use the lecture theorem that a random dominating HST satisfies $\mathbb{E}[d_T(x, y)] \leq \alpha d(x, y)$ for every pair, where $\alpha = C \log n$ for an absolute constant C. You may also assume an exact algorithm for k-median on this HST. Centers must be leaves in X, and $1 \leq k < n$.

For a k-element center set A, define $\mathrm{cost}_d(A) = \sum_{x \in X} \min_{a \in A} d(x, a)$. Prove that solving k-median on a sampled tree and evaluating its centers in d gives expected cost at most $\alpha\,\mathrm{OPT}_d$. Obtain cost at most $2\alpha\,\mathrm{OPT}_d$ with probability at least $1 - \delta$ using $O(\log(1/\delta))$ independent trees. Explain why the analysis fixes an optimal assignment in the original metric before taking expectations.

(D) **[20 points] Review: One coordinate from a partition.** Let $A \subseteq X$ be the union of clusters of P chosen independently with probability $1/2$. Define

$$f(z) = \min\{\Delta, d(z, A)\}, \qquad d(z, A) = \min_{a \in A} d(z, a),$$

with $f(z) = \Delta$ when $A = \varnothing$. Prove $|f(x) - f(y)| \le d(x, y)$ for every pair. If $d(x, y) > \Delta$, show that

$$\mathbb{P}[|f(x) - f(y)| \ge \Delta/(16H_n)] \ge 1/8.$$

Hint. Apply (A) with $r = \Delta/(16H_n)$. On the padding event, exclude $P(x)$ from A and include $P(y)$.

(E) **[20 points] Review and extension: Combine the scales.** Rescale d so its minimum nonzero distance is one, and let its diameter be $\Phi \ge 1$. At each scale $\Delta_i = 2^i$, for $-1 \le i \le \lceil \log_2 \Phi \rceil$, independently construct $K = \lceil 64 \ln(n^2/\delta) \rceil$ coordinates as in (D), where $0 < \delta < 1/2$. Let $F(x)$ be the vector of all coordinates and $\ell = \lceil \log_2 \Phi \rceil + 2$ the number of scales. Prove that, with probability at least $1 - \delta$, simultaneously for every pair,

$$\frac{\sqrt{K}}{256 H_n} d(x, y) \le \|F(x) - F(y)\|_2 \le \sqrt{\ell K}\, d(x, y).$$

State the dimension and the resulting distortion (the ratio of the upper and lower multiplicative constants). No dimension-reduction theorem may be used.

Hint. For each fixed pair choose a scale with $d(x, y)/4 \le \Delta_i \le d(x, y)/2$. At that scale, show that at least $K/16$ coordinates separate the pair as in (D), except with probability at most $e^{-K/64}$. Then union bound over pairs.

Class notes: Finite Metric Spaces and Partitions.

Problem 3: Dimension Reduction: What Does One Map Preserve?

[100 points]

Let $0 < \varepsilon, \delta < 1/2$ and let A be a $k \times d$ matrix whose entries are independent $N(0, 1/k)$ random variables. Vectors are columns, so $Ax \in \mathbb{R}^k$ for $x \in \mathbb{R}^d$. You may use the normalized standard Gaussian density $(2\pi)^{-1/2} e^{-t^2/2}$ and elementary Gaussian integration.

(A) **[15 points] Review: A fixed vector sees a Gaussian.** Prove that the standard Gaussian distribution in $\mathbb{R}^d$ is invariant under orthogonal transformations. Deduce that, for every fixed nonzero v,

$$\frac{\|Av\|_2^2}{\|v\|_2^2} \overset{\mathrm{d}}{=} \frac{1}{k} \sum_{i=1}^{k} Z_i^2, \qquad Z_i \text{ independent } N(0,1).$$

Compute the expectation and variance of this ratio. Explain both the factor $1/\sqrt{k}$ in the map and where independence is used.

(B) **[25 points] Review and extension: Both tails, with constants.** For $Z \sim N(0,1)$, prove $\mathbb{E}\left[e^{tZ^2}\right] = (1-2t)^{-1/2}$ for $t < 1/2$. Use exponential Markov bounds, with a positive parameter for the upper tail and a negative parameter for the lower tail, to prove

$$\mathbb{P}\left[\left|\frac{1}{k}\sum_{i=1}^{k} Z_i^2 - 1\right| > \varepsilon\right] \le 2e^{-k\varepsilon^2/8}.$$

Do not obtain the lower tail by asserting symmetry of Z^2.

Hint. Optimize the parameter separately for each tail. For $0 \le u \le 1$, show $u - \ln(1+u) \ge u^2/4$; for $0 \le u < 1$, show $-u - \ln(1-u) \ge u^2/2$. Integration of the derivatives suffices.

(C) **[20 points] Review and extension: One map for a finite metric.** Let $P \subset \mathbb{R}^d$ be a fixed set of $n \ge 2$ points. Prove that

$$k \ge \lceil 8\varepsilon^{-2} \ln(n^2/\delta) \rceil$$

suffices to preserve all squared pairwise distances within factors $1 \pm \varepsilon$, with probability at least $1 - \delta$. Translate this to a guarantee for ordinary distances.

Next suppose a randomized construction first embeds an n-point metric into some Euclidean space with distortion at most β, except with probability $\delta/2$. Apply an independent Gaussian map to these images. Show how to obtain dimension $O(\varepsilon^{-2} \log(n/\delta))$, failure probability at most δ, and distortion at most $\beta\sqrt{(1+\varepsilon)/(1-\varepsilon)}$. In particular, explain how this applies to Problem 2(E).

(D) **[25 points] The geometric route to the same dimension bound.** You may use this spherical concentration consequence of the Brunn–Minkowski argument from the supplementary geometry slides: there is an absolute $c > 0$ such that, for $d \geq 2$, uniform $U \in S^{d-1}$, a 1-Lipschitz function f on the sphere (with Euclidean distance), and a median M of $f(U)$,

$$\mathbb{P}[|f(U) - M| > t] \leq 2e^{-cdt^2} \qquad (t > 0).$$

For $1 \leq k \leq d$, let $f(u) = \|(u_1, \dots, u_k)\|_2$. Show that f is 1-Lipschitz, $\mathbb{E}[f(U)^2] = k/d$, and

$$|M - \sqrt{k/d}| \leq \sqrt{2/(cd)}.$$

Deduce, for $k \geq 8/(c\varepsilon^2)$, that

$$\mathbb{P}\left[|f(U) - \sqrt{k/d}| > \varepsilon\sqrt{k/d}\right] \leq 2e^{-c\varepsilon^2 k/4}.$$

Explain how a uniform random rotation, projection onto the first k coordinates, and scaling by $\sqrt{d/k}$ give another JL map. Why does the required dimension depend on $\log n$, rather than on d? If the sufficient bound on k exceeds d, use the identity map.

Hint. Integrate the concentration tail to bound $\mathbb{E}[(f(U) - M)^2]$, then apply the reverse triangle inequality for the L_2 norm of random variables. A fixed unit vector under a uniform random rotation is uniform on the sphere.

(E) **[15 points] A nearest neighbor, and the order of quantifiers.** Fix a data set P and a query q before sampling A. Suppose all squared distances in $P \cup \{q\}$ are preserved as in (C), and let $\hat{p}$ minimize $\|Aq - Ap\|_2$ over $p \in P$. Prove

$$\|q - \hat{p}\|_2 \leq \sqrt{\frac{1+\varepsilon}{1-\varepsilon}} \min_{p \in P} \|q - p\|_2.$$

Does the same fixed-set proof guarantee distance preservation for every query chosen after inspecting A? Give a counterexample when $k < d$, and state precisely which quantifier changes.

Class notes: Dimension Reduction; Geometry behind JL (supplementary slides).

Problem 4: Entropy and Compression

[100 points]

All logarithms here are base two, written lg. For a finite random variable X, write $p_x = \mathbb{P}[X = x]$ and $H(X) = \sum_{x:p_x>0} p_x \lg(1/p_x)$. Write $h(p) = -p \lg p - (1-p)\lg(1-p)$, with $0 \lg 0 = 0$. Let $S = (S_1, \ldots, S_n)$ be independent Bernoulli(p) bits, $0 < p < 1$, and $J = \sum_i S_i$. A deterministic extractor outputs a string Y of length L, uniform on $\{0,1\}^\ell$ conditional on $L = \ell$; empty output is allowed. Compression is lossless, with codebooks and block boundaries known to the decoder. Ignore encoding time and codebook storage.

(A) **[15 points] Review: Entropy counts possibilities.** Prove that independent finite random variables satisfy $H(X,Y) = H(X)+H(Y)$, and hence $H(S) = nh(p)$. Show $h(p) \le 1$, with equality exactly when $p = 1/2$. For $0 \le j \le n$, prove (including the endpoints)

$$\frac{2^{nh(j/n)}}{n+1} \le \binom{n}{j} \le 2^{nh(j/n)}.$$

Hint. For a binomial variable with parameter j/n, verify by adjacent-term ratios that j is a mode.

(B) **[15 points] Review: Fair bits from a uniform range.** For uniform $V \in \{0, \ldots, M-1\}$, write $M = \sum_i a_i 2^i$, where $a_i \in \{0,1\}$. Partition the range into consecutive blocks of these sizes. In a block of size 2^i, output the offset within that block in exactly i bits, including leading zeros. Prove conditional uniformity given the output length and

$$\mathbb{E}[L] = \frac{1}{M} \sum_i a_i 2^i i \ge \lfloor \lg M \rfloor - 1 > \lg M - 2.$$

Compute the expected output length when $M = 13$.

Hint. For $k = \lfloor \lg M \rfloor$, compare $\sum_{i<k} a_i 2^i (k-i-1)$ with 2^k, using $\sum_{r\ge1}(r-1)2^{-r} = 1$.

(C) **[20 points] Extract almost all the entropy without knowing p.** Conditional on $J = j$, rank S lexicographically among the $\binom{n}{j}$ strings of weight j, starting at rank zero, and apply (B) to this rank. Prove that this is an extractor for every $p \in (0,1)$, even though the algorithm does not know p. Show the quantitative bound

$$\mathbb{E}[L] \ge nh(p) - \lg(n+1) - 2.$$

Deduce that its expected number of output bits per input bit tends to $h(p)$ for every fixed p, using (D) for the upper bound.

Hint. First prove the exact identity $H(S) = H(J) + \mathbb{E}[\lg \binom{n}{J}]$ by grouping the strings by weight. Use $H(J) \le \lg(n+1)$; you may use the general fact that a variable on at most r outcomes has entropy at most $\lg r$. Uniformity must still hold after averaging over J.

(D) **[10 points] Review: Extraction cannot create entropy.** For any finite input X, prove $\mathbb{E}[L] \le H(X)$ for a deterministic extractor, allowing variable output length. Why is the absence of extra random bits necessary?

Hint. For $y = \mathrm{Ext}(x)$, compare $\mathbb{P}[X = x]$ and $\mathbb{P}[Y = y]$.

(E) **[20 points] Review: Why Huffman's greedy merge is safe.** A binary prefix-free code assigns a leaf of a binary tree to each symbol. For positive symbol probabilities $p_1, \dots, p_r$, $r \ge 2$, its expected length is $\sum_i p_i \ell_i$, where ℓ_i is leaf depth. Show that some optimal tree has two least-probable symbols as deepest siblings. Prove that contracting them reduces the optimum by their combined probability, and deduce Huffman's optimality. Run the rule for probabilities $(2/5, 3/10, 1/5, 1/10)$ and give an optimal code and its expected length.

Hint. Remove unary internal nodes. Then swap labels, keeping the tree fixed, before contracting the deepest sibling pair.

(F) **[20 points] Review and extension: The compression rate, even without prefixes.** You may use Kraft's converse: nonnegative integer lengths $\ell_1, \dots, \ell_r$ with $\sum_i 2^{-\ell_i} \le 1$ can be realized by a binary prefix-free code. For any finite source X, construct a prefix-free code with expected length less than $H(X) + 1$. Deduce that a Huffman code on entire blocks S has expected length less than $nh(p) + 1$.

For the converse, let C_n be *any* injective map from $\{0,1\}^n$ to finite binary strings; it need not be prefix-free. For fixed $p \in (0,1)$, prove

$$\liminf_{n\to\infty} \frac{\mathbb{E}[|C_n(S)|]}{n} \ge h(p).$$

Also show that no injective code can shorten every n-bit input.

Hint. Use lengths $\lceil \lg(1/p_x) \rceil$ for the upper bound. For the lower bound, fix $0 < \eta < h(p)/2$ and put $I(S) = -\lg \mathbb{P}[S]$. Show $\mathbb{P}[I(S) < n(h(p) - \eta)] = o(1)$. There are fewer than $2^{n(h(p)-2\eta)+1}$ binary strings of length below $n(h(p) - 2\eta)$, and an input with $I(S) \ge n(h(p) - \eta)$ has probability at most $2^{-n(h(p)-\eta)}$.

Class notes: Entropy; Entropy II (Huffman coding and compression).

References

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

[MU17] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis.* 2nd ed. Cambridge University Press, 2017.