COS 597J Streaming and Sketching Algorithms (Fall 2021)
Problem Set

Things to note before you start:

- Problems will be added to this pset as the course progresses. You are expected to work on $(\frac{1}{2} \pm \frac{1}{4})$-fraction of the problems in total. You will submit all your solutions in one single PDF, by **23:59 November 30, 2021 EST**. Please make it clear which problem each solution corresponds to.
- We will prove one single result in each problem, with the main steps outlined as subproblems. Sometimes the subproblems are stated *informally on purpose*. You will have to figure out the exact statement that you can prove and is useful towards the final goal (the final goal will always be precisely stated). However, if you have a different proof that does not follow the predesigned route, feel free to use it.
- The solutions to most problems are likely to exist somewhere on the Internet. Looking up for the solutions is strongly discouraged, but allowed. In any case, you must phrase your solutions in your own words. If you read any articles (e.g., paper, blog post, stackexchange, ...) that eventually *help* you in solving the problems, please list them in the reference. Please also acknowledge anyone with whom you had discussions.

This version contains all problems!

Problem 1. In this problem, we will analyze the Morris counter mentioned in the lecture, and prove that it solves approximate counting using $O(\log\log N + \log(1/\varepsilon) + \log\log(1/\delta))$ bits of space.

Recall that the approximate counting problem asks us to maintain a counter n up to N, supporting

- `inc()`: $n \leftarrow n + 1$;
- `query()`: output an estimate $\tilde{n}$ such that $\Pr[|\tilde{n} - n| > \varepsilon n] < \delta$.

A Morris counter has a parameter $\alpha > 0$. It maintains a variable X, initialized to 0. Each time `inc()` is called, X is incremented to $X+1$ with probability $(1+\alpha)^{-X}$. To answer `query()`, it returns $((1+\alpha)^X - 1)/\alpha$.

1. Let Y_k be the random variable denoting the number of `inc()` calls needed to increment X from $k-1$ to k. Derive the probability distribution of each Y_k.
2. We say a random variable Z is *subgamma* with parameters σ, B, if
$$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq e^{\lambda^2\sigma^2/2},$$
for all $|\lambda| < B$. Show that each Y_k is subgamma for some σ_k, B_k.
3. Show that the sum of independent subgamma random variables is also subgamma.
4. Show that if Z is subgamma with parameters σ, B, then for any $t > 0$, both $\Pr[Z - \mathbb{E}[Z] > t]$ and $\Pr[Z - \mathbb{E}[Z] < -t]$ are at most $\max\left\{e^{-\frac{t^2}{2\sigma^2}}, e^{-\frac{tB}{2}}\right\}$.
5. Fix n greater than some threshold T. Show that the probability that after n `inc()` calls, X is still too small is low, and that the probability that after n `inc()` calls, X is already too large is low.
6. Choose the right parameters T and α for N, ε, δ. Conclude that by also maintaining the exact count up to $n = T$, the Morris counter solves approximate counting using $O(\log\log N + \log(1/\varepsilon) + \log\log(1/\delta))$ bits.

Problem 2. In this problem, we will show how to improve the space usage of the FM sketch to $O(\varepsilon^{-2}\log(1/\delta)+\log\log n)$ bits, still assuming free access to random functions.

Recall that one FM sketch uses a random $h : [n] \to \mathbb{N}$, where $\Pr[h(x) = i] = 2^{-(i+1)}$ for all $x \in [n]$ and $i \geq 0$, and maintains R, the maximum h-value it sees. We maintain a total of $t = O(\varepsilon^{-2}\log(1/\delta))$ independent FM sketches, and apply median-of-means at the end of the stream. Let R_i be R in the i-th sketch ($i = 1, \ldots, t$). In the following, we will show that there is a space-efficient way to store all R_i.

1. At any given point, let F be the number of distinct elements in the stream so far. Let $\Delta_i := R_i - \lfloor \log F \rfloor$. Show that with probability at least $1 - 2^{-t}/\log n$, $\Delta_1, \ldots, \Delta_t$ can be stored using $c \cdot (t + \log\log n)$ bits for some absolute constant $c > 0$:
   (a) Upper bound the tail probabilities $\Pr[\log|\Delta_i| \geq D]$ for $D \geq 1$ (shown in the lecture).
   (b) Let $S > 0$. Show that if $\sum_{i=1}^{t} \log|\Delta_i| > S$, then there must exist some $k \geq 1$ such that at least $2^{-k}t$ Δ_i have $\log|\Delta_i| \geq S \cdot 2^{k/2}/(4t)$.
   (c) Upper bound the probability that $\sum_{i=1}^{t} \log|\Delta_i| \geq S$.
2. Show that suppose we know the value of $\lfloor \log F \rfloor$ at every point of the stream, then all R_i can be maintained with probability $\geq 1 - \delta$ using space $O(t + \log\log n)$.
3. Show that even if we don't know $\lfloor \log F \rfloor$, all R_i can still be maintained with probability $\geq 1 - \delta$ using space $O(t + \log\log n)$, and conclude.

Problem 3. In this problem, we will show that any algorithm that can approximate the number of distinct elements F up to $1 \pm \varepsilon$ relative error with constant probability must use $\Omega(\varepsilon^{-2})$ bits of space.

To prove this space lower bound, we will make a reduction from a (one-way) communication problem, called *Gap-Hamming*. Recall the Hamming distance between two binary strings, $\Delta(x, y) := |\{i : x_i \neq y_i\}|$, is the number of bits where they differ. In this communication problem, two players Alice and Bob receive, as inputs, d-bit binary strings x and y respectively. It is promised that either $\Delta(x, y) \leq d/2 - \sqrt{d}$, or $\Delta(x, y) \geq d/2 + \sqrt{d}$. Then Alice sends one message M to Bob, and Bob needs to decide which case they are in. We will show later that the (one-way) randomized communication complexity with public randomness of Gap-Hamming is at least $\Omega(d)$.

1. Assuming this Gap-Hamming communication lower bound, prove the claimed distinct elements lower bound.

2. Let $a \in \{-1, 1\}^{d'}$ be fixed, and $u \in \{-1, 1\}^{d'}$ be uniformly random, for some *odd* $d' \geq 1$. Also fix $i \in [d']$, let e_i be the unit vector $(0, \ldots, 1, \ldots, 0)$ with the only 1 in coordinate i. Compute the probability that $\langle a, u \rangle$ and $\langle e_i, u \rangle$ have the same *sign*. Note that since d' is odd, the inner products must be nonzero.

3. (Probabilistically) reduce the INDEX problem on d' bits to Gap-Hamming on d bits for $d = O(d')$, assuming public randomness.

4. Conclude by applying the INDEX lower bound.

In the next **three** problems, we will prove that ℓ_p-norm estimation for *constant* p requires $\Omega(n^{1-3/p})$ bits of space, even in the insertion-only setting. Note that in class, we mentioned that the best known lower bound is $\Omega(n^{1-2/p})$. The proof uses *information theory*.

- In Problem 4, we will prove basic facts in information theory;
- In Problem 5 and 6, we will use them to prove the space lower bound.

If you are familiar with information theory, you may also skip this problem. **Solving Problem 5 and 6 automatically gives you credit on Problem 4.**

Problem 4. All logarithms in this problem are with respect to base 2.

1. (Entropy) Let X be a random variable taking values in $[n]$. Then the entropy of X is

$$H(X) := \sum_{X \in [n]} \Pr[X = x] \log(1/\Pr[X = x]).$$

   It measures "how random X is".

   (a) Show that $0 \le H(X) \le \log n$. When do we have the equalities?

   (b) Let $(X, X) \in [n]^2$, how does $H(X, X)$ compare to $H(X)$?

   (c) Let $(X, Y) \in [n]^2$ be jointly distributed, show that $H(X, Y) \le H(X) + H(Y)$.

2. (Conditional entropy) Let X be a random variable, and let W be an *event*. The entropy of X conditioned on W is

$$H(X \mid W) := \sum_{x} \Pr[X = x \mid W] \log(1/\Pr[X = x \mid W]).$$

   Let Y be another *random variable*. The entropy of X conditioned on Y is

$$H(X \mid Y) := \sum_{y} \Pr[Y = y] \cdot H(X \mid Y = y).$$

   (Note that $Y = y$ is an event.)

   (a) Show that $H(X \mid Y) \le H(X)$. When do we have the equality? Is $H(X \mid W)$ also always at most $H(X)$?

   (b) Show that $H(X \mid Y) = H(X, Y) - H(Y)$.

3. (Mutual information) Let X, Y be two random variables. Their mutual information is

$$I(X; Y) := H(X) - H(X \mid Y).$$

   It measures "how much information Y reveals about X", i.e., "how much less random X becomes after seeing Y." Similarly, let Z be a random variable, the mutual information between X and Y conditioned on Z is

$$I(X; Y \mid Z) := H(X \mid Z) - H(X \mid Y, Z).$$

   (a) Show that $I(X; Y) = I(Y; X)$, and $I(X; Y) \le H(Y)$.

   (b) When is $I(X; Y)$ zero?

4. (Chain rule) Let $X, Y_1, Y_2, \ldots, Y_n$ be random variables. Show that

$$I(X; Y_1, Y_2, \ldots, Y_n) = I(X; Y_1) + I(X; Y_2 \mid Y_1) + \cdots + I(X; Y_n \mid Y_1, \ldots, Y_{n-1}).$$

   When $Y_1, Y_2, \ldots, Y_n$ are independent, show that

$$I(X; Y_1, Y_2, \ldots, Y_n) \ge I(X; Y_1) + I(X; Y_2) + \cdots + I(X; Y_n).$$

5. (KL divergence) Let P, Q be two distributions over $[n]$. Then the KL divergence from Q to P is

$$\mathbf{D}_{\mathrm{KL}}(P \,\|\, Q) := \sum_{x \in [n]} P(x) \log \frac{P(x)}{Q(x)}.$$

When Q is the prior distribution of some random variable X and P is the posterior distribution after an observation, the KL divergence can measure how "surprised" you are about the observation (in terms of X), e.g., you are infinitely surprised if P has non-zero mass outside the support of Q, you are not surprised if $P = Q$.

(a) Show that $\mathbf{D}_{\mathrm{KL}}(P \,\|\, Q) \geq 0$.

(b) Show that $I(X;Y) = \sum_y \Pr[Y = y] \cdot \mathbf{D}_{\mathrm{KL}}(\mathrm{dist}(X \mid Y = y) \,\|\, \mathrm{dist}(X))$, where $\mathrm{dist}(X)$ denotes the distribution of X, and $\mathrm{dist}(X \mid Y = y)$ denotes the distribution of X conditioned on $Y = y$.

(c) (Pinsker's inequality) Show that

$$\mathbf{D}_{\mathrm{TV}}(P, Q) \leq \sqrt{2\mathbf{D}_{\mathrm{KL}}(P \,\|\, Q)},$$

where $\mathbf{D}_{\mathrm{TV}}(P, Q) := \sum_x |P(x) - Q(x)|$ is the total variation distance.

The proof is a reduction from a communication problem, which we call RandomVsOne. RandomVsOne has t players, $P_1, \ldots, P_t$. Each player P_i receives one bit $X_i \in \{0,1\}$. Moreover, it is guaranteed that $(X_1, \ldots, X_t)$ are jointly sampled from either

- $\mathcal{D}_{\text{RAND}}$: $\Pr[X_i = 1] = t^{-1}$ independently, or
- $\mathcal{D}_{\text{ONE}}$: $X_1 = X_2 = \cdots = X_t = 1$ (not random).

In the RandomVsOne problem, P_1 first sends a message M_1 to P_2, then P_2 sends a message M_2 to P_3, and so forth. The goal of the players is to decide which case they are in. Finally, P_t will output M_t, which must be either RAND or ONE. If the inputs are sampled from $\mathcal{D}_{\text{RAND}}$ [resp. $\mathcal{D}_{\text{ONE}}$], P_t must output RAND [resp. ONE] with probability at least 0.9. We allow each player to be random, but will assume that there is *no public randomness*.

Instead of considering the number of bits in $M_1, \ldots, M_t$, we will study its *information cost* on $\mathcal{D}_{\text{RAND}}$. Fix such a (randomized) protocol, and consider the joint distribution $(X_1, \ldots, X_t, M_1, \ldots, M_t)$ where $(X_1, \ldots, X_t) \sim \mathcal{D}_{\text{RAND}}$. The information cost of the protocol is defined as

$$\text{IC} := I(X_1, \ldots, X_t; M_1, \ldots, M_t),$$

i.e., the amount information the messages reveal about the inputs. The messages may be arbitrarily long, and we only consider if they have low mutual information with the inputs.

The streaming lower bound proof will consist of two steps:

(I) prove that if there is a streaming algorithm for ℓ_p-norm estimation with $\varepsilon \cdot n^{1-3/p}$ bits of space, then there is a protocol for RandomVsOne with *information cost* $O(\varepsilon/t^2)$;

(II) prove that any protocol for RandomVsOne must have information cost $\Omega(1/t^2)$.

The two steps together imply the desired $\Omega(n^{1-3/p})$ lower bound.

Problem 5. In this problem, we will complete step (I) above. To this end, consider the n-fold RandomVsOne problem, which we denote by RandomVsOne^n. For $i = 1, \ldots, t$, player P_i gets $X_i^{(1)}, \ldots, X_i^{(n)} \in \{0,1\}$. For each $j = 1, \ldots, n$, let $X^{(j)} = (X_1^{(j)}, \ldots, X_t^{(j)})$. It is guaranteed that $(X^{(1)}, \ldots, X^{(n)})$ is sampled from either

- $\mathcal{D}_{\text{RAND}}^n$: all $X^{(j)} \sim \mathcal{D}_{\text{RAND}}$ independently, or
- $\mathcal{D}_{\text{ONE}}^n$: for *one uniformly random* $j^* \in \{1, \ldots, n\}$, $X^{(j^*)} \sim \mathcal{D}_{\text{ONE}}$, and all other $X^{(j)} \sim \mathcal{D}_{\text{RAND}}$ independently conditioned on j^*.

Similarly to RandomVsOne, the goal is to decide which distribution their inputs are sampled from, by sending one message to the next player in order.

1. Prove that if there is a streaming algorithm that approximates the ℓ_p-norm of the frequency vector x within a factor of 2 with probability 0.95 using S bits of space, then there is a protocol that solves RandomVsOne^n with probability 0.9 where each player sends a message of at most S bits (for appropriate value of t).

2. Fix a protocol for RandomVsOne^n, where each player P_i sends a message M_i of at most S bits to the next player (and M_t is the output). When $(X^{(1)}, \ldots, X^{(n)}) \sim \mathcal{D}_{\text{RAND}}^n$, show that

$$\mathop{\mathbb{E}}_{j} \left[I(X_1^{(j)}, \ldots, X_t^{(j)}; M_1, \ldots, M_t) \right] \le St/n,$$

   where j is a uniformly random index in $\{1, \ldots, n\}$.

3. Show that given a protocol for RandomVsOne^n where each player sends few bits, one can design a protocol for RandomVsOne with low *information cost*.

4. Conclude step (I) and state what you prove as a lemma.

Problem 6. In this problem, we will complete step (II) above. Fix a protocol for RandomVsOne. Consider the joint distribution of $(X_1, \ldots, X_t, M_1, \ldots, M_t)$, where $(X_1, \ldots, X_t) \sim \mathcal{D}_{\text{RAND}}$ and $(M_1, \ldots, M_t)$ is generated by the protocol given $(X_1, \ldots, X_t)$.

1. Denote $(M_1, \ldots, M_{i-1})$ by $M_{<i}$. Show that the information cost is equal to

$$\text{IC} = \sum_{i=1}^{t} I(X_i; M_i \mid M_{<i}).$$

2. For $i = 1, \ldots, t$, let $\epsilon_i = I(X_i; M_i \mid M_{<i})$. Let $P(M_{<i})$ be the distribution of M_i conditioned on $M_{<i}$, and $Q(M_{<i})$ be the distribution of M_i conditioned on $M_{<i}$ and $X_i = 1$. Show that

$$\mathop{\mathbb{E}}_{M_{<i}} [\mathbf{D}_{\text{TV}}(P(M_{<i}), Q(M_{<i}))] \leq O(\sqrt{t\epsilon_i}),$$

   where $\mathbf{D}_{\text{TV}}(\cdot, \cdot)$ is the total variation distance.

3. Show that if $\text{IC} < \varepsilon/t^2$ for a sufficiently small constant $\varepsilon > 0$, then the protocol cannot distinguish between $\mathcal{D}_{\text{RAND}}$ and $\mathcal{D}_{\text{ONE}}$ with probability 0.9. (If you also solved Problem 5, combine the lemmas you proved and conclude.)

Problem 7. In this problem, we will prove that the number of distinct elements can be approximated within a factor of $1 \pm \varepsilon$ using $O(\varepsilon^{-2} \log^2 n)$ bits with 0.9 probability in turnstile streaming, assuming the coordinates of the final vector x are at most M for $M \leq \operatorname{poly} n$. Recall that in turnstile streaming, the number of distinct elements is the number of nonzero coordinates in x, which is the frequency vector.

1. Let P be a degree n polynomial over $\mathbb{F}_p$ for prime $p \gg n$. Upper bound the probability that $P(y) = 0$ for a uniformly random $y \in \mathbb{F}_p$.

2. Show that there is a linear sketch using space $O(\log n)$ that can test if the final vector $x \in \mathbb{Z}^n$ is $\mathbf{0}$ with error probability $1/\operatorname{poly} n$, assuming the coordinates of x are at most M.

3. Show that there is an algorithm for counting distinct elements in turnstile streaming with the claimed space usage.

Problem 8. In this problem, we prove a lower bound for the SUPPORTFINDING problem. We will show that any streaming algorithm that finds an index $i \in [n]$ with $x_i \neq 0$ with 0.9 probability in turnstile streaming must use $\Omega(\log^2 n)$ bits.

Consider the following communication problem with two players, called universal relation ($\text{UR}^{\subset}$). Alice gets a set $S \subseteq [n]$, and Bob gets $T \subsetneq S$. Assume that they have shared public randomness. Alice sends one message M to Bob, and the goal is for Bob to output some element in $S \setminus T$ with probability 0.9. We will show that the length of M must be at least $\Omega(\log^2 n)$ bits.

1. Assuming the above communication lower bound for $\text{UR}^{\subset}$, prove the claimed streaming space lower bound.

2. Show that for $\text{UR}^{\subset}$, we may assume without loss of generality, that conditioned on the output $a \in S \setminus T$, a is a uniformly random element in $S \setminus T$.

3. Imagine that $S = S_1 \cup S_2 \cup \cdots$ such that $|S_i| = \Theta(\sqrt{n} \cdot 8^{-i})$, and all S_i are disjoint. We further assume that T is the union of $S_1 \cup S_2 \cup \cdots \cup S_i$ for some i that Alice does not know. Analyze the behavior of a correct protocol on such instances, and reduce $\text{UR}^{\subset}$ from AUGMENTED-INDEX.
(Recall that in AUGMENTED-INDEX, Alice gets a vector $x \in C^d$ for some set C, and Bob gets $t \in [d]$, and $x_1, \ldots, x_{t-1}$. The problem asks Alice to send one message to Bob so that Bob can recover x_t with constant probability.)

4. Conclude by applying the AUGMENTED-INDEX lower bound mentioned in Lecture 7 (you don't need to (re)prove it).

Problem 9. In this problem, we show that ℓ_2-norm estimation can be solved in the *sliding window* streaming model. In sliding window streaming, we receive a stream of elements, and maintain some function of the substream of the *latest T elements*. For ℓ_2-norm estimation, the stream consists of $a_1, a_2, \ldots, a_m \in [n]$. For every $k \geq T$, let $x^{(k)} \in \mathbb{Z}^n$ be the frequency vector of the substream $a_{k-T+1}, \ldots, a_k$. We will show that there is an algorithm using $\mathrm{poly}(\varepsilon^{-1}, \log n)$ bits such that with probability $1 - 1/n^{\Omega(1)}$, it outputs an $(1 \pm \varepsilon)$-approximation to $\|x^{(k)}\|_2^2$ after seeing a_k for all $k \geq T$.

In this problem, we assume that T is known to the algorithm from the beginning, and $m \leq \mathrm{poly}\, n$.

1. Let $x, y, z \in \mathbb{Z}^n$ such that $y_i \geq x_i \geq 0$ and $z_i \geq 0$ for all $i \in [n]$. Show that for $\delta \in (0, 1/2)$, if $\|y\|_2^2 \leq (1+\delta)\|x\|_2^2$, then
$$\|y + z\|_2^2 \leq (1 + O(\sqrt{\delta}))\|x + z\|_2^2.$$

2. Let k be the length of stream we have processed so far. Show that we can maintain a list $k = k_0 > k_1 > k_2 > \cdots$ and approximations to ℓ_2-norms of the substreams $a_{k_i}, \ldots, a_k$ for every i, such that $\|x^{(k)}\|_2^2$ can be approximated from the list, and conclude.

3. (Optional) Show that if the stream consists of turnstile updates (i, Δ), then ℓ_2-norm estimation in sliding window cannot be solved using $o(n)$ space.

Problem 10. In this problem, we are going to prove that one can *track* the ℓ_2-norm of an **insertion-only** stream using $O(\log n \log\log n)$ bits with multiplicative error 1 ± 0.1 and probability 0.9. That is, the algorithm must output the current ℓ_2-norm after seeing *every* element with multiplicative error 1 ± 0.1 (all outputs must be simultaneously accurate with 0.9 probability). Recall that the AMS sketch solves ℓ_2-estimation with error $1 \pm \varepsilon$ and probability $1 - \delta$ using space $O(\varepsilon^{-2} \log n \log(1/\delta))$. Thus, by setting $\delta = 1/\operatorname{poly} n$, $\varepsilon = 0.1$ and applying a union bound, the ℓ_2-norm of the stream can be tracked using $O(\log^2 n)$ bits. We will show that we only need an extra factor of $\log\log n$.[1]

In this problem, we assume that the length of the stream m is at most $\operatorname{poly} n$. For $t = 0, \ldots, m$, let $x^{(t)}$ be the frequency vector at time t. For simplicity, we are going to assume that the algorithm has free access to random bits.

1. Let $\sigma_1, \ldots, \sigma_n \in \{-1, 1\}$ be independently random signs, $\sigma = (\sigma_1, \ldots, \sigma_n)$, and let $x^{(0)}, x^{(1)}, x^{(2)}, \ldots, x^{(t)} \in \mathbb{R}^n$ be fixed vectors such that $0 = x^{(0)} \le x^{(1)} \le x^{(2)} \le \cdots \le x^{(t)}$ *coordinate-wisely*. Show that there exists a constant $C > 0$ such that

$$\Pr_{\sigma}\left[\sup_{1\le j\le t}\left|\left\langle \sigma, x^{(j)}\right\rangle\right| > C\cdot\sqrt{\log(1/\delta)}\|x^{(t)}\|_2\right] < \delta,$$

for all $\delta \in (0, 1/2)$, as follows.

   (a) Fix an integer constant $c \le t$. Show that there exist integers $t_0 < t_1 < t_2 < \cdots < t_c$ such that $t_0 = 0, t_c = t$ and $\|x^{(t_i-1)} - x^{(t_{i-1})}\|_2^2 \le \frac{1}{c}\cdot\|x^{(t)}\|_2^2$ for $i = 1, \ldots, c$. Show that

$$\Pr_{\sigma}\left[\sup_{j\in\{t_0,\ldots,t_c\}}\left|\left\langle \sigma, x^{(j)}\right\rangle\right| > C/2\cdot\sqrt{\log(1/\delta)}\|x^{(t)}\|_2\right] < \delta/2,$$

   by Hoeffding's inequality and union bound.

   (b) Fix an integer c', which depends on c. For $i = 0, \ldots, c-1$, suppose $c' \le t_{i+1} - t_i$, show that there exist integers $t_{i,0} < \cdots < t_{i,c'}$ such that $t_{i,0} = t_i, t_{i,c'} = t_{i+1}$ and $\|x^{(t_{i,i'}-1)} - x^{(t_{i,i'-1})}\|_2^2 \le \frac{1}{c'}\|x^{(t_{i+1}-1)} - x^{(t_i)}\|_2^2$ for $i' = 1, \ldots, c'$. Show that

$$\Pr_{\sigma}\left[\exists i\in[c], \sup_{j\in\{t_{i,0},\ldots,t_{i,c'}\}}\left|\left\langle \sigma, x^{(j)} - x^{(t_i)}\right\rangle\right| > C/4\cdot\sqrt{\log(1/\delta)}\|x^{(t)}\|_2\right] < \delta/4.$$

   (c) Prove the claim by repeatedly refining the intervals.

2. Let m_i be the first time such that $\|x^{(m_i)}\|_2^2 \ge (1+\varepsilon)^i$ for $\varepsilon = \Theta(1/\log n)$. Show that the AMS sketch with space $O(\log n \log\log n)$ reports accurate estimates at all times m_i with high probability. Use the first step to show that it is also accurate between all m_i and m_{i+1} with high probability.

[1] In fact, there is an algorithm that does not even lose the $\log\log n$ factor.

Problem 11. In this problem, we show that the CountSketch sketching matrix gives a version of distributional Johnson-Lindenstrauss with a (very fast) multiplication time of $O(\|x\|_0)$ and output dimension $O(1/\varepsilon^2\delta)$, where $\|x\|_0$ is the number of nonzero entries of x. It also gives an oblivious subspace embedding with multiplication time $O(\|x\|_0)$ and output dimension $O(d^2/\varepsilon^2)$. Recall that CountSketch solves ℓ_2-point query and heavy hitters. It samples a random hash function $h : [n] \to [k]$, and a random sign function $\sigma : [n] \to \{-1, 1\}$, and maintains k counters:

$$S_j = \sum_{i \in [n] : h(i) = j} \sigma(i) \cdot x_i.$$

Denote $(S_1, \ldots, S_k) \in \mathbb{R}^k$ by S.

1. Given h and σ, we have $S = \Pi x$ for some matrix $\Pi \in \mathbb{R}^{k \times n}$. Express Π in terms of h and σ.

2. Show that assuming random access to h and σ, given a sparse x encoded as a list of nonzero entries, Πx can be computed in $O(\|x\|_0)$ time, encoded also as a list of nonzero entries.

3. Given $x \in \mathbb{R}^n$, show that $\mathbb{E}[\|\Pi x\|_2^2] = \|x\|_2^2$, and bound $\mathrm{Var}[\|\Pi x\|_2^2]$.

4. Show that for any $x \in \mathbb{R}^n$, we have

$$\Pr_{h,\sigma}\left[\|\Pi x\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2\right] > 1 - \delta,$$

   by setting $k = c/\varepsilon^2\delta$ for some constant $c > 0$. Show that it suffices to have constant-wise independent h and σ.

5. Let $E \subset \mathbb{R}^n$ be a linear subspace of dimension d. Show that by setting $k = c \cdot d^2/\varepsilon^2$ for some constant $c > 0$, Π is a subspace embedding of E with probability 0.9 (constant-wise h and σ suffice), as follows.

   (a) Let $x, y \in \mathbb{R}^n$ such that $\langle x, y\rangle = 0$ and $\|x\|_2 = \|y\|_2 = 1$. Show that $\mathbb{E}\left[\langle \Pi x, \Pi y\rangle^2\right] \le O(1/k)$.

   (b) Fix an orthonormal basis $\{u_1, \ldots, u_d\}$ for E, and let $U \in \mathbb{R}^{n \times d}$ be the matrix with columns $u_1, \ldots, u_d$. Show that

$$\mathbb{E}\left[\mathrm{tr}\big((I_d - (\Pi U)^\top(\Pi U))^2\big)\right] \le O(d^2/k),$$

   where I_d is the $d \times d$ identity matrix.

   (c) Show that $\Pr[\|I_d - (\Pi U)^\top(\Pi U)\|_{\mathrm{op}} > \varepsilon] < 0.1$, and conclude.

Problem 12. In this problem, we show that given a matrix $A \in \mathbb{R}^{n\times d}$ encoded as a list of nonzero entries, there is an algorithm that runs in time $O(\mathrm{nnz}(A)\log n + \mathrm{poly}\, d)$ that approximates the leverage scores of the column space of A within a factor of $1 \pm \varepsilon$ for constant $\varepsilon > 0$, where $\mathrm{nnz}(A)$ is the number of nonzero entries of A (which we assume is at least n). Denote the column vectors of A by $a_1, \ldots, a_d$. For simplicity, *we assume that $a_1, \ldots, a_d$ are linearly independent.*

Let $u_1, \ldots, u_d$ be an orthonormal basis for the column space of A, let $U = (u_1, \ldots, u_d) \in \mathbb{R}^{n\times d}$. Recall that the i-th leverage score $p_i = \|e_i^\top U\|_2^2$, where $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ is the i-th coordinate (column) vector.

By Problem 11, there is an oblivious subspace embedding $\Pi \in \mathbb{R}^{n\times m}$ for $m = O(d^2/\varepsilon^2)$ such that given $x \in \mathbb{R}^n$, Πx can be computed in $O(\|x\|_0)$ time, where $\|x\|_0$ is the number of nonzero entries of x.[2]

The main idea is to first apply the oblivious subspace embedding on $a_1, \ldots, a_d$ to reduce the dimension to $m = O(d^2/\varepsilon^2)$, then run Gram-Schmidt orthogonalization on $\Pi a_1, \ldots, \Pi a_d$ to obtain an orthonormal basis. We then show that the *same process* of orthogonalization on $a_1, \ldots, a_d$ gives a set of vectors that is almost orthonormal, and projecting $e_1, \ldots, e_n$ to this set gives approximations of the leverage scores.

1. The algorithm first computes $Q \in \mathbb{R}^{m\times d}, R \in \mathbb{R}^{d\times d}$ such that

   - $\Pi A = QR$;
   - Q has orthonormal columns;
   - R is upper-triangular (and invertible since A has rank d).

   Hence, $\Pi A R^{-1} = Q$. Bound the computational time.

2. Show that $\|AR^{-1}x\|_2^2 = (1 \pm O(\varepsilon))\|x\|_2^2$ for all $x \in \mathbb{R}^d$ (columns of AR^{-1} are almost orthonormal).

3. Since U and AR^{-1} have the same columns space, let $T \in \mathbb{R}^{d\times d}$ be the matrix such that $UT = AR^{-1}$. Show that $\|Tx\|_2^2 = (1 \pm O(\varepsilon))\|x\|_2^2$.

4. Show that $\|e_i^\top AR^{-1}\|_2^2 = (1 \pm O(\varepsilon))p_i$ (recall that T and $T^\top$ have the same set of singular values).

5. Since AR^{-1} cannot be computed efficiently, the algorithm applies a distributional Johnson-Lindenstrauss on the row vectors of AR^{-1} to reduce its dimension. Let $G \in \mathbb{R}^{m\times d}$ such that $G_{i,j} \sim \frac{1}{\sqrt{m}}\{-1, 1\}$. Show that $\|e_i^\top AR^{-1}G^\top\|_2^2 = (1 \pm O(\varepsilon))p_i$ for all $i \in [n]$ (for a suitable value of m). Bound the computational time.

[2] See also subproblem 11.1 and 11.2.