1 10 pts

Estimate

As an estimate, return $x$ which has rank $k \cdot \frac{t}{m}$ in the subsample.

Note: You do not need to worry about rounding when $k \cdot \frac{t}{m}$ is not an integer.

Approach

For $\text{rank}(x)$ to lie in $[(1 - \epsilon)k, (1 + \epsilon)k]$, it suffices to show that two events happen simultaneously:

1. The number of sampled items with $\text{rank} \leq k(1 + \epsilon)$ is $\geq k \cdot \frac{t}{m}$.
2. The number of sampled items with $\text{rank} \leq k(1 - \epsilon)$ is $\leq k \cdot \frac{t}{m}$.

If we set $t = O\left(\frac{m}{k} \cdot \frac{1}{1+\epsilon} \log(1/\delta)\right)$ then we can show that both of each of these events occur with probability $(1 - \delta/2)$ using a Chernoff bound.

Applying Chernoff

Specifically, the expected number of subsampled items with rank $\leq k(1 + \epsilon)$ is $\frac{k(1+\epsilon)}{m} \cdot t$. For Event 1, we want to bound the probability that this expectation is exceed by a $\frac{1}{1+\epsilon}$ factor. For sufficiently small $\epsilon$, $\frac{1}{1+\epsilon} = (1 - c\epsilon)$ for a constant $c$. We apply the following Chernoff bound:

$$\Pr(X < (1 - c\epsilon)\mu) \leq e^{-\mu(\epsilon c)^2/2}$$

This quantity is less than $\delta/2$ as long as our expected value $\frac{k(1+\epsilon)}{m} \cdot t$ is $O\left(\frac{1}{\epsilon^2} \log 1/\delta\right)$ which requires setting $t = O\left(\frac{m}{k} \cdot \frac{1}{\epsilon^2} \log(1/\delta)\right)$.

Similarly, the expected number of subsampled items with rank $\leq k(1 - \epsilon)$ is $\frac{k(1-\epsilon)}{m} \cdot t$. For Event 2 we want to bound the probability that this expectation is undercut by a $\frac{1}{1-\epsilon}$ factor.
For sufficiently small $\epsilon$, $\frac{1}{1-\epsilon} = (1 - c\epsilon)$ for a constant $c$. We apply the following Chernoff bound:

$$Pr(X > (1 + c\epsilon)\mu) \leq e^{-\mu(\epsilon c)^2/(2+c\epsilon)}$$

This quantity is less than $\delta/2$ as long as our expected value $\frac{k(1-\epsilon) \cdot t}{m}$ is $O(\frac{1}{\epsilon^2} \log 1/\delta)$ which requires setting $t = O(\frac{m}{k} \cdot \frac{1}{\epsilon^2} \log(1/\delta))$.

So Event 1 occurs with probability $(1 - \delta/2)$ and Event 2 occurs with probability $(1 - \delta/2)$ as long as we set $t = O(\frac{m}{k} \cdot \frac{1}{\epsilon^2} \log(1/\delta))$. By a union bound we conclude that both events occur simultaneously with probability $(1 - \delta)$ as desired.
2 10 pts

(a)
For a completely random hash, \(Pr_{h}[s = 0]\) is equal to the probability that none of the non-zero entries in \(x\) hash into the range \([0, \frac{1}{2T}]\). This is only true because we required that, at the end of our stream, each \(x_i \geq 0\).

Each \(x_i\) hashes into that range with probability \((1 - \frac{1}{2T})\) and there are \(\|x\|_0\) non-zero entries in the vector so \(Pr_{h}[s = 0]\) can be computed as:

\[
Pr_{h}[s = 0] = \left(1 - \frac{1}{2T}\right)^{\|x\|_0} = \left(\left(1 - \frac{1}{2T}\right)^{2T}\right)^{\|x\|_0/2T}
\]

To prove part (i) we just notice that \((1 - \frac{1}{2T})^{2T} \leq \frac{1}{e}\) and when \(T \leq \frac{1}{2}\|x\|_0\), \(\|x\|_0/2T \geq 1\).

Accordingly, \(\left((1 - \frac{1}{2T})^{2T}\right)^{\|x\|_0/2T} \leq \frac{1}{e}\) as desired.

To prove part (ii) we just notice that \((1 - \frac{1}{2T})^{2T} \geq \frac{1}{3}\) when \(T \geq 1\). We won’t need any lower values of \(T\). When \(T \geq 2\|x\|_0\), \(\|x\|_0/2T \leq \frac{1}{4}\). Accordingly, \(\left((1 - \frac{1}{2T})^{2T}\right)^{\|x\|_0/2T} \geq .7 > .5\) as desired.

(b)

Our procedure will be as follows. Count the number of events where \(s_i = 0\) and the number where \(s_i > 1\). If the ratio of events with \(s_i = 0\) is less than .45 output LOW. Otherwise output high. To prove that this works when we collect \(O(\log(1/\delta))\) individual counters, we just apply a Chernoff bound:

If \(Pr_{h}[s = 0] \leq .37\) then the expected ratio of events where \(s_i = 0\) is \(< .37\). The probability that this ratio exceeds its expectation by a constant factor, \(.(45 - .37)\), is \(\delta\) by a Chernoff bound as long as we average over \(O(\log(1/\delta))\) events.

If \(Pr_{h}[s = 0] > .5\) then the expected ratio of events where \(s_i = 0\) is \(> .5\). The probability that this ratio is below its expectation by a constant factor, \(.(5 - .45)\), is \(\delta\) by a Chernoff bound as long as we average over \(O(\log(1/\delta))\) events.

Note: There is nothing special about .45. You could have used any value in place of .45 that is between .37 and .5, but it had to be at least some constant distance away from both.

(c)

Run the above algorithm with \(T = 1, 2, 4, 8, \ldots, m\) where \(m\) is the first power of 2 above \(N\). This equates to \(O(\log N)\) runs. Any value of \(\delta = 1/N^c\) will ensure that all copies of our
testing procedure succeed with high probability, simultaneously, although you didn’t need to say so to receive full credit.

Return the lowest value of $T$ that returns HIGH as our estimate for $F$. If our procedure above returns HIGH for a value of $T$, we can at least guarantee that $T \geq \frac{1}{2} \|x\|_0$. If the next lower tested value (which equal $T/2$) returned LOW then we similarly know that $T/2 \leq 2\|x\|_0$. Combining these inequalities ensures that:

$$\frac{1}{2} \|x_0\|_2 \leq F \leq 4 \|x_0\|_2$$

which satisfies the requirements of the problem.

There is one corner case to worry about: if you never see an output of LOW, you can’t guess $F = 1$ because this won’t be a good estimate if in fact $\|x\|_0 = 0$. There are a number of ways to handle this case. For example, we can just maintain a counter that does no subsampling of the stream. If the counter comes out as 0 at the end of the stream then it must be that $\|x\|_0 = 0$ (since again we guarantee that each $x_i \leq 0$. Otherwise $\|x\|_0 > 0$ and we can safely output $F = 1$.

**Note:** You did not need to handle the edge case to receive full credit.
3 10 pts

(a) Suppose we have one entry $x_i$ with value $\epsilon n$ and every other entry equal to 1. With Count-Min we obtain a 2-factor approximation to $x_i$. If we have error depending on $\|x\|_2$ however, our estimate will lie in the interval $[\epsilon n - \epsilon \sqrt{n}, \epsilon n + \epsilon \sqrt{n}]$, which is much tighter for large $n$.

(b) Let $I_j$ be an indicator variable for the event that $h(x_j) = h(x_i)$. In other words, $I_j = 1$ when this event occurs and 0 otherwise. By linearity of expectation

$$E[y_i] = E \left[ g(i) \sum_{j=1}^{n} I_j g(j) x_j \right]$$

$$= \sum_{j=1}^{n} E \left[ I_j g(i) g(j) x_j \right]$$

$$= x_i + \sum_{j \neq i} E \left[ I_j g(i) g(j) x_j \right]$$

$$= x_i + \sum_{j \neq i} E[I_j] E \left[ g(i) g(j) \right] x_j$$

The second to last step follows from the fact that $g(i) g(i) x_i I_i$ simply equals $x_i$. The last step follows from the independence of hash functions $h$ and $g$. Finally, we obtain the result by noting that, when $j \neq i$, $E \left[ g(i) g(j) \right] = 0$, by independence of $g$’s. I.e. $g(i) g(j)$ is just a random sign when $i \neq g$, so it has expectation 0.

(c) Let’s start by computing $E[y_i^2]$.

$$E[y_i^2] = E \left[ \left( g(i) \sum_{j=1}^{n} I_j g(j) x_j \right)^2 \right]$$

$$= E \left[ \sum_{j=1}^{n} \sum_{k=1}^{n} g(j) g(k) x_j x_k I_j I_k \right]$$

$$= E \left[ \sum_{j=1}^{n} g(j) g(j) x_j x_j I_j I_j \right] + E \left[ \sum_{j=1}^{n} \sum_{k>j}^{n} g(j) g(k) x_j x_k I_j I_k \right]$$

$$= \sum_{j=1}^{n} x_j^2 E \left[ I_j^2 \right] + 0$$
Again the last step follows from noting that $\mathbb{E}[g(i)g(j)] = 0$ when $j \neq i$ and 1 otherwise. Now, for all $j \neq i$, for a fully random $h(\cdot)$, $\mathbb{E}[I_j^2]$ simply equals $\frac{1}{k}$ since there is a $\frac{1}{k}$ chance that any given $x_j$ collides with $x_i$. At the same time, $\mathbb{E}[I_i^2] = 1$. Accordingly,

$$\mathbb{E}[y_i^2] = \sum_{j=1}^{n} x_j^2 \mathbb{E}[I_j^2]$$

$$= x_i^2 \mathbb{E}[I_i^2] + \sum_{j \neq i} x_j^2 \mathbb{E}[I_j^2]$$

$$= x_i^2 + \sum_{j \neq i} x_j^2 \frac{1}{q}$$

$$\leq x_i^2 + \frac{1}{q} \|x\|_2^2$$

The last equation is an inequality because the sum is missing the contribution from $x_i$.

Finally, as shown in part (b), $\mathbb{E}[y_i]^2 = x_i^2$ so we conclude that $Var[y_i] = \mathbb{E}[y_i^2] - \mathbb{E}[y_i]^2 \leq \frac{\|x\|_2^2}{q}$

**(d)**

Chebyshev’s inequality says that:

$$Pr \left[ |x_i - y_i| \geq c \sqrt{Var[y_i]} \right] \leq 1/c^2$$

$$Pr \left[ |x_i - y_i| \geq c \|x\|_2 / \sqrt{q} \right] \leq 1/c^2$$

Setting $c = \epsilon \sqrt{q}$ gives the result.

**(e)**

We need to set $q = O(\frac{1}{\epsilon^2})$. This compares to a total of $O(\frac{1}{\epsilon})$ for Count-Min when we run with a comparable failure probability. **Note:** Count-min only has a log $n$ dependence when you want it to succeed with high probability, $1 - \frac{1}{n^c}$. 

6
4 10 pts

(a)
To prove that this statement is true, we just need to prove that for a set of points $x_1, \ldots, x_m$ with mean $\mu$,

$$\sum_{i=1}^{m} \|x_i - \mu\|^2 = \frac{1}{m} \sum_{i,j<i} \|x_i - x_j\|^2$$

(1)

$$\sum_{i=1}^{m} \|x_i - \mu\|^2 = \sum_{i=1}^{m} \|x_i - \frac{1}{m} \sum_{j=1}^{m} x_j\|^2$$

$$\quad = \sum_{i=1}^{m} \left[ \|x_i\|^2 + \frac{1}{m} \sum_{j=1}^{m} x_j^2 - 2 \frac{1}{m} \sum_{j=1}^{m} x_j \right]$$

$$\quad = \sum_{i=1}^{m} \langle x_i, x_i \rangle + \frac{1}{m} \sum_{j=1}^{m} \langle x_j, x_j \rangle - 2 \frac{1}{m} \sum_{j=1}^{m} \langle x_j, \frac{1}{m} \sum_{j=1}^{m} x_j \rangle$$

$$\quad = \sum_{i=1}^{m} \langle x_i, x_i \rangle - \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{m} \langle x_i, x_j \rangle$$

Additionally, for the righthand side of (1):

$$\frac{1}{m} \sum_{i,j<i} \|x_i - x_j\|^2 = \frac{1}{2m} \sum_{j=1}^{m} \sum_{i=1}^{m} \|x_i - x_j\|^2$$

$$\quad = \frac{1}{2m} \sum_{j=1}^{m} \sum_{i=1}^{m} \left[ \langle x_i, x_i \rangle + \langle x_j, x_j \rangle - 2 \langle x_i, x_j \rangle \right]$$

$$\quad = \frac{1}{2m} \left[ 2m^2 \sum_{i=1}^{m} \langle x_i, x_i \rangle - 2 \sum_{j=1}^{m} \sum_{i=1}^{m} \langle x_i, x_j \rangle \right]$$

Multiplying through by $2m^2$ we conclude that (1) holds. We just apply the equality to every cluster individually to get the result.
(b)

Since every term of the sum
\[
\sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{j,\ell \in C_i, j < \ell} \|\tilde{x}_j - \tilde{x}_\ell\|_2^2 = \sum_{i=1}^{k} \sum_{j,\ell \in C_i, j < \ell} \frac{1}{|C_i|} \|\tilde{x}_j - \tilde{x}_\ell\|_2^2
\]
is positive and is bounded in the interval
\[
(1 - \epsilon) \frac{1}{|C_i|} \|x_j - x_\ell\|_2^2, (1 + \epsilon) \frac{1}{|C_i|} \|x_j - x_\ell\|_2^2
\]
by JL, it must be that the entire sum is bounded in the interval
\[
\left[ (1 - \epsilon) \sum_{i=1}^{k} \sum_{j,\ell \in C_i, j < \ell} \frac{1}{|C_i|} \|x_j - x_\ell\|_2^2, (1 + \epsilon) \sum_{i=1}^{k} \sum_{j,\ell \in C_i, j < \ell} \frac{1}{|C_i|} \|x_j - x_\ell\|_2^2 \right].
\]
The fact holds for all clusters simultaneously because, once we condition on the success of our JL for all distance pairs \(\|x_j - x_\ell\|_2\) (remember there are only \(O(n^2)\) of them), the terms being bounded is just a deterministic statement.

(c)

Using part (b):
\[
Cost(x_1, \ldots, x_n, \bar{C}_1, \ldots, \bar{C}_k) \leq \frac{1}{1 - \epsilon} Cost(\bar{x}_1, \ldots, \bar{x}_n, \bar{C}_1, \ldots, \bar{C}_k)
\]
\[
\leq \frac{1}{1 - \epsilon} \gamma \cdot Cost(\bar{x}_1, \ldots, \bar{x}_n, \bar{C}_1^*, \ldots, \bar{C}_k^*)
\]
\[
\leq \frac{1}{1 - \epsilon} \gamma \cdot Cost(\bar{x}_1, \ldots, \bar{x}_n, C_1^*, \ldots, C_k^*)
\]
\[
\leq \frac{1 + \epsilon}{1 - \epsilon} \gamma \cdot Cost(x_1, \ldots, x_n, C_1^*, \ldots, C_k^*)
\]
For small \(\epsilon\), \(\frac{1 + \epsilon}{1 - \epsilon} = (1 + O(\epsilon))\), giving the result.

(d)

The proof for \((9 + \epsilon)\) appears in [http://arxiv.org/pdf/1410.6801v3.pdf](http://arxiv.org/pdf/1410.6801v3.pdf) if you are interested in checking it out. The TA’s would love to know if it could be improved.