Algorithms for Big Data Fall 2020

Homework 3 Solutions

Problem 1: Tales of different norms

(a) Prove that the following two relationships hold for any vector $x \in \mathbb{R}^n$:

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \cdot \|x\|_\infty \qquad \text{and} \qquad \|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \cdot \|x\|_2.$$

Solution: To prove:

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \cdot \|x\|_\infty$$

Let X = $(x_1, x_2, ...x_n)$

$$\|x\|_\infty = \max_i |x_i|$$

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}$$

Consider LHS,

$$\max_i |x_i| \leq \sqrt{\sum_{i=1}^{n} x_i^2}$$

Consider RHS,

$$\sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x_1^2 + x_2^2 + ... + x_n^2} \leq \sqrt{n \cdot \max_i |x_i|^2} = \sqrt{n} \cdot \|x\|_\infty$$

To prove:

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \cdot \|x\|_2$$

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + ... + x_n^2} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \leq \sum_{i=1}^{n} \sqrt{x_i^2} = \sum_{i=1}^{n} |x_i| = \|x\|_1$$

Using the Cauchy-Schwarz inequality we get,

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \leq \sqrt{\sum_{i=1}^{n} |x_i|^2} \cdot \sqrt{\sum_{i=1}^{n} 1^2} = \sqrt{n} \cdot \|x\|_2$$

(b) Provide example vectors that satisfy each of the above four inequalities with an equality.

Solution: Let X = $(1, 0)$

$$\|x\|_1 = \|x\|_2 = \|x\|_\infty = 1$$

Let X = $(1, -1)$, $(n = 2)$

$$\|x\|_1 = 2$$

$$\|x\|_2 = \sqrt{2}$$

$$\|x\|_\infty = 1$$

$$\|x\|_2 \leq \sqrt{n} \cdot \|x\|_\infty \qquad \text{and} \qquad \|x\|_1 \leq \sqrt{n} \cdot \|x\|_2.$$

Problem 2: Streaming Sampling

Let $a_1, a_2, \ldots, a_n$ be a stream of n integers (not necessarily distinct) in the range $\{1, 2, \ldots, n\}$. The algorithm knows n up front. Each a_i will arrive one-by-one. The algorithm may compute something and update the storage, but then the value may not be accessed again (unless it is explicitly stored). The space is the maximum amount of memory used throughout. For each of these, you must prove that the algorithm works correctly, and provide a bound on the space.

(a) Assume that you know $A = \|\vec{a}\|_2^2 = \sum_{i=1}^n a_i^2$, the sum-of-squares of the values in the stream. Provide an algorithm using $O(\log n)$ space to sample an element a_i from the stream with probability exactly $p_i = \frac{a_i^2}{A}$.

Solution: Note that you can come up with an algorithm analogous to the Problem 3 of HW2 where at each step during the stream you choose to keep a_i with probability $\frac{a_i^2}{A_i}$ (A_i is the running sum up to a_i).
Here, we will discuss a different solution which requires you to know $\sum_{i=1}^n a_i^2$ beforehand. Before the stream starts choose a random number j in the range 1 to A with equal probability.

Initialize $running\_sum = 0$

For $i = 1, 2...n$

$running\_sum \leftarrow running\_sum + a_i^2$

If $running\_sum \geq j$:

Let $X \leftarrow a_i$

Output X

(b) Now, assume that you **do not know** A ahead of time. Provide an algorithm using $O\left(\log^2 n\right)$ space that samples a_i from the stream with probability approximately $p_i = \frac{a_i^2}{A}$. More precisely, you should sample a_i with probability $\widetilde{p}_i$ satisfying $\frac{p_i}{4} \leq \widetilde{p}_i \leq 4p_i$ for all $i \in [n]$.

Hint: Use many different samples like the ones from (a) depending on the true value of A, and in parallel, compute A exactly so that you know which sample to use for the output.

Solution: Run the above algorithm independently for $A' = n, 2n, 4n....$ Sample a total of $O(logn)$ elements from the stream.

Compute the value of A. Determine the specific A' for which the following condition is true

$$A' \leq A \leq 2A'$$

Choose a sample from A' with a probability $\frac{1}{2}$ and choose from $2A'$ with probability $\frac{1}{2}$.

Because we are off by a factor of at most 2 in the estimate for A, the probability will be off slightly as well. But it is an easy calculation to show the in the end element a_i is output with probability $\widetilde{p_i}$ satisfying $\frac{p_i}{4} \leq \widetilde{p_i} \leq 4p_i$ for all $i \in [n]$.

For lower bound, notice that a_i is output with probability at least $\frac{p_i}{2}$ in the case of using $2A'$. Since this case is chosen with probability $1/2$, we have that $\frac{p_i}{4} \leq \widetilde{p_i}$.

For the upper bound, it can be output in two ways: either with probability at most p_i for the $2A'$ case. Or with probability at most $2p_i$ for the A' case. Each happens with probability $1/2$ so we have that $\widetilde{p_i} \leq (1/2 + 2/2)p_i = (3/2)p_i$.

(c) Improve your algorithm from (b). Now, given ε in the range $0 < \varepsilon < 1$, your sampling probabilities should satisfy $(1-\varepsilon)p_i \leq \widetilde{p_i} \leq (1+\varepsilon)p_i$.

Solution: We sketch the high-level idea. As with the previous part, we use multiple estimators. But we want to use a finer granularity, so we look at $A' = (1+\varepsilon')^k n$ for some $\varepsilon' < \varepsilon$ and for k in the range $1 \leq k \leq \log_{1+\varepsilon'}(n^2)$. Since now A' does not have to an integer, instead of selecting a random number in the range of 1 to A', we sample j from a continuous uniform distribution defined on the interval $(0, A')$. Then, we consider A' or $(1+\varepsilon')A'$ for the estimates that are closest to A, satisfying $A' \leq A \leq (1+\varepsilon')A'$. We choose between the the outputs from these two cases with probabilities depending on ε'. Additionally, in the second case for $(1+\varepsilon')A'$, we only return an output if initial random sample j is less than or equal to A where $A = \sum_{i=1}^{j}({a_i}^2)$.

Since $A \geq A'$, the numbers for which the running sums are greater than A' have zero probability of being the output in the first case. Therefore, the output probabilities for such numbers only comes from the second case where the range $(1+\varepsilon')A'$ is greater than A. We select the output from A' with probability q and from $(1+\varepsilon')A'$ with probability $1-q$. The output probability for a number a_i whose running sum is greater than A' is given by:

$$\begin{aligned}
\tilde{p_i} &= (1-q)\Pr(j \leq A) \cdot \Pr(X = a_i | j \leq A) \\
&= (1-q)\frac{A}{(1+\varepsilon')A'}\frac{a_i^2}{A} \\
&= (1-q)\frac{A}{(1+\varepsilon')A'}p_i \\
&\geq \frac{(1-q)p_i}{1+\varepsilon'}
\end{aligned}$$

The output probabilities for all a_is will at least be $\frac{(1-q)p_i}{1+\varepsilon'}$. Therefore, lower bound on $\tilde{p_i}$ for all a_is is $\frac{(1-q)p_i}{1+\varepsilon'}$.

Now for the upper bound, the output probability for any number a_i can be bounded as,

$$\begin{aligned}
\tilde{p_i} &\leq q\frac{a_i^2}{A'} + (1-q)\frac{A}{(1+\varepsilon')A'}\frac{a_i^2}{A} \\
&\leq qp_i\frac{A}{A'} + (1-q)\frac{p_i}{(1+\varepsilon')}\frac{A}{A'}
\end{aligned}$$

Since, $\frac{A}{A'} \leq (1+\varepsilon')$, we get:

$$\begin{aligned}
\tilde{p_i} &\leq qp_i(1+\varepsilon') + (1-q)p_i \\
&\leq p(1+q\varepsilon')
\end{aligned}$$

Combining the two bounds on p_i, we get,

$$\frac{(1-q)p_i}{1+\varepsilon'} \leq \tilde{p_i} \leq p(1+q\varepsilon')$$

Observe that if we set $q = \varepsilon'^2$, we can write,

$$(1 - \varepsilon')p_i \leq \tilde{p}_i \leq p(1 + \varepsilon'^3) \leq p(1 + \varepsilon')$$

Since, $\varepsilon' < \varepsilon$,

$$(1 - \varepsilon)p_i \leq \tilde{p}_i \leq p(1 + \varepsilon)$$

Problem 3: Implementing a Sketching Algorithm

Implement and test one of the algorithms from the class, that is, choose one of the following options: (i) Morris for approximate counting, (ii) FM for distinct elements, or (iii) AMS for ℓ_2 estimation. Implement the algorithm and the + and ++ variants for the one you choose.

(a) Demonstrate/compare the performance of the three variants of the algorithm (normal, +, ++). Set the input size(s) to be large enough to see some difference in their performance.

(b) Provide results (in a table or plot, clearly labeled) for at least 2 different parameter settings (and list the parameters). Briefly discuss the results and any interesting observations.

(c) Provide the results of 10 repetitions for each of the two parameter settings (in a table or plot, clearly labeled), to demonstrate the probability of failure (and list the parameters). Briefly discuss the results and any interesting observations.

(d) Discuss how theory relates to practice, with quantitative results to back up your claims. For example, if the theory is pessimistic, then show that the results in practice are better with the same/improved parameters.

Solution: See other document.