CS 395T: Sublinear Algorithms, Fall 2020 September 17, 2020

Lecture 7: More Quantile Estimation

Prof. Eric Price *Scribe: Tongzheng Ren, Shuo Yang*

NOTE: THESE NOTES HAVE NOT BEEN EDITED OR CHECKED FOR CORRECTNESS

1 Skill Drills

For the following X, what is

- mean
- "typical" variation (i.e. 90% region)
- Prob $1-\delta$ region

1. $X = \frac{1}{n}\sum_{i=1}^n X_i$, $X_i \in [-n, n]$, independent with mean μ_i.
2. $X = \frac{1}{n}\sum_{i=1}^n X_i$, $X_i \in [-a_i, a_i]$, independent with mean 0.
3. $X = \frac{1}{n}\sum_{i=1}^n X_i$, $X_i \in [-a_i, a_i]$, mean 0 but pairwise independent.
4. $X = \frac{1}{n}\sum_{i=1}^n X_i$, $X_i \in [-a_i, a_i]$, mean 0, but 6-wise independent.

1.1 Case 1

We have mean $\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^n \mu_i$, $\mathrm{Var}(X) \le n$, with Hoeffding's inequality we have

$$\mathbb{P}\left[\left|X - \frac{1}{n}\sum_{i=1}^n \mu_i\right| \ge t\right] \le 2\exp\left(-\frac{t^2}{2n}\right)$$

Here we need $t = \Theta(\sqrt{n})$ to get a non-trivial bound. So we have the prob $1-\delta$ region with width $\Theta(\sqrt{n\log\frac{1}{\delta}})$

1.2 Case 2

We have mean $\mathbb{E}[X] = 0$, $\mathrm{Var}(X) \le \frac{1}{n^2}\sum_{i=1}^n a_i^2$, and with Hoeffding's inequality

$$\mathbb{P}\left[|X| \ge t\right] \le 2\exp\left(-\frac{n^2t^2}{2\sum_{i=1}^n a_i^2}\right)$$

The Prob $1-\delta$ region depends on a_i, for example, if $a_i = i$, we need $t = \Theta(\sqrt{n})$ to get a non-trivial bound, as $\sum_{i=1}^n i^2 = \Theta(n^3)$. So we have the prob $1-\delta$ region with width $\Theta(\frac{\|a_i\|_2}{n}\sqrt{\log\frac{1}{\delta}})$

1.3 Case 3

As we just have pairwise independent, we now can only use Chebyshev's inequality:

$$\mathbb{P}\left[|X| \geq t\right] \leq \frac{\sum_{i=1}^{n} a_i^2}{n^2 t^2}$$

So we have the prob $1-\delta$ region with width $\Theta(\frac{\|a_i\|_2}{n}\delta^{-0.5})$

1.4 Case 4

Here we can use the following bound:

$$\mathbb{P}\left[|X| \geq t\right] \leq \frac{\mathbb{E}[X^6]}{t^6} \leq \frac{\sum_{i=1}^{n} a_i^6}{n^6 t^6}$$

So we have the prob $1-\delta$ region with width $\Theta(\frac{\|a_i\|_6}{n}\delta^{-\frac{1}{6}})$

2 More Quantile Estimation

2.1 A deterministic algorithm

We want to find rank(x): $X_r \leq x < X_{r+1}$. Our goal is rank$(x) \pm \epsilon n$. Last class we have $O(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon\delta})$ results. Now we want $\frac{1}{\epsilon}\log n$ rather than $\frac{1}{\epsilon^2}$.

Intuitively, we just need $X_{\epsilon n}, X_{2\epsilon n}, \cdots, X_{\frac{1}{\epsilon}\epsilon n}$ to achieve the goal. Consider "compress" any input into a "smaller" input with ϵ-close answers, which can be formally described as $\forall X$, $\exists X'$ with $\frac{1}{\epsilon}$ distinct values, s.t.

$$\text{rank}_X(y) = \text{rank}_{X'}(y) \pm \epsilon n,$$

where rank$_X(y)$ is the prediction of rank(y) given the information of X. We call X' a "coreset" for quantiles. In fact, $\exists$ coresets for many problems, including graph sparsifies, k-means, regression, etc.

The estimator we consider in the following analysis is

$$\widehat{\text{rank}}_{X'}(y) = \text{quantile}_{X'}(y) \cdot n.$$

We now solve in a streaming fashion. Assume each time we receive a new batch of data with size $\frac{1}{\epsilon}$. We maintain the compression level by level and stored in a binary tree. At each level, keep 1 compression, recompress & put at next level when you get second compression. We output the compression at the highest level finally, and in total we have $\log(\epsilon n)$ levels.

Lemma 1. *If S' is ϵ-close to S (quantile$_{S'}(x)$ = quantile$_S(x) \pm \epsilon$), T' is ϵ-close to T, $|S'| = |T'|$, $|S| = |T|$, then quantile$_{S' \cup T'}(x)$ = quantile$_{S\cup T}(x) \pm \epsilon$*

Proof.

$$\begin{aligned}
\text{quantile}_{S'\cup T'}(x) =& \frac{1}{|S'|+|T'|}(\text{rank}_{S'}(x)+\text{rank}_{T'}(x))\\
=& \frac{|S'|}{|S'|+|T'|}\text{quantile}_{S'}(x)+\frac{|T'|}{|S'|+|T'|}\text{quantile}_{T'}(x)\\
=& \frac{|S|}{|S|+|T|}\text{quantile}_{S}(x)+\frac{|T|}{|S|+|T|}\text{quantile}_{T}(x)\pm\epsilon\\
=& \text{quantile}_{S\cup T}(x)\pm\epsilon
\end{aligned}$$

□

Lemma **??** shows that union two compression will not increase the error on quantile, and notice that compression can introduce at most ϵ additional error on quantile, so each level will have at most ϵ quantile error, and the top level will have at most $\epsilon\log(\epsilon n)$ quantile error, and equivalently ϵn *log*(ϵn) rank error. Space we used is $O(\frac{1}{\epsilon}\log(\epsilon n))$, as each compression needs $O(\frac{1}{\epsilon})$ and we have $\log(\epsilon n)$ levels. Run this algorithm with $\epsilon'=\frac{\epsilon}{\log(\epsilon n)}$, we will get accuracy ϵ with space $O(\frac{1}{\epsilon}\log^2(\epsilon n))$

2.2 Improve the result with randomization

Now we further improve the analysis based on randomization. If we compress with random offset (e.g. for ordered sequence $X_1,X_2,\cdots,X_{2k}$ we uniformly choose $X_1,X_3,\cdots,X_{2k-1}$ or $X_2,X_4,\cdots,X_{2k}$ as the compression), with simple calculation, we know

$$\text{rank}_{X'}(x_i)=\text{rank}_X(x_i)+\eta_i,\quad\text{where}\quad\eta_i=0, i\text{ even};\quad\eta_i=\pm1, i\text{ odd}.$$

In our algorithm: in compression at level i,

$$\begin{aligned}
&\text{rank}_{X^{(i+1)}}(x_j)=\text{rank}_{X^{(i)}}(x_j)+\eta_j^{(i)},\\
&\text{where}\quad\eta_j^{(i)}=0,\text{if the } i\text{-th significant bit of } j \text{ is } 0;\quad\eta_j^{(i)}=\pm2^i,\text{if the } i\text{-th significant bit of } j \text{ is } 1.
\end{aligned}$$

Final error for $\text{rank}(x_j)$ is $\widehat{\text{rank}}(x_j)-\text{rank}(x_j)=\sum_{i=1}^{\log(\epsilon n)}\eta_j^{(i)}$ where η_j^i have mean zero and $|\eta_j^{(i)}|\le2^i$.

$$\text{Var}(\widehat{\text{rank}}(x_j)-\text{rank}(x_j))=\sum_{i=1}^{\log(\epsilon n)}\text{Var}(\eta_j(i))\le\sum_{i=1}^{\log n}2^{2i}\le2\cdot2^{2\log(\epsilon n)}=2(\epsilon n)^2.$$

Thus $|\widehat{\text{rank}}(x_j)-\text{rank}(x_j)|\le O(\epsilon n)$ with 0.9 probability using Chebyshev's inequality. In fact with Hoeffding we can get $|\widehat{\text{rank}}(x_j)-\text{rank}(x_j)|\le\epsilon n\sqrt{\log\frac{1}{\delta}}$ w.p. $1-\delta$. Set $\epsilon'=\frac{\epsilon}{\sqrt{\log\frac{1}{\delta}}}$, then we have the space complexity $O(\frac{\sqrt{\log\frac{1}{\delta}}}{\epsilon}\log(\epsilon n))$, i.e. an improvement of $\log n$ over deterministic variant.