CS 395T: Sublinear Algorithms, Fall 2020 09/24/20

Lecture 9: On Estimation of Symmetric Random Variables

Prof. Eric Price *Scribe: Ajil Jalal*

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

1 Overview

In the last lecture we analyzed the FrequentElements and Count-Min Sketch algorithms.

In this lecture we will analyze symmetric random variables, and their concentrations, which will give us bounds for the CountSketch algorithm.

2 Estimate mean of symmetric random variables

Let x be a random variable over $\mathbb{R}$ that is symmetric about some unknown μ, with variance σ^2.

Given samples $x_1, x_2, \cdots, x_n$ of x, how do we estimate μ?

- The empirical mean requires $O\left(\frac{1}{\varepsilon^2\delta}\right)$ samples to generate $\widehat{\mu}$ satisfying $|\widehat{\mu} - \mu| \leq \varepsilon\sigma$ with probability $\geq 1 - \delta$.
- The median-of-means algorithm requires $O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ samples to guarantee $\widehat{\mu}$ satisfying $|\widehat{\mu} - \mu| \leq \varepsilon\sigma$ with probability $\geq 1 - \delta$.

However, the median-of-means algorithm requires us to decide on ε and δ in advance. Can we give an algorithm that works simultaneously for all ε?

The guarantee we want is:

$$\widehat{\mu} \text{ such that } \mathbb{P}[|\widehat{\mu} - \mu| \geq \varepsilon\sigma] \leq \exp\left(-\Omega\left(\varepsilon^2 m\right)\right) \forall\varepsilon \text{ simultaneously.}$$

In general, this cannot be done. However, when the variables are symmetric, then we can create an estimator that works simultaneously for all ε.

2.1 Warmup

As a warmup, let's consider a univariate Gaussian. For ε sufficiently small, we have

$$\mathbb{P}[|x_i - \mu| \geq \varepsilon\sigma] = 1 - \mathbb{P}[|X_i - \mu| \leq \varepsilon\sigma] \approx \frac{\varepsilon}{\sqrt{1+\varepsilon}} \leq \Omega(\varepsilon).$$

Define the indicator random variable $z_i = \mathbf{1}|x_i - \mu| \geq \varepsilon\sigma$.

From the previous inequality, we have

$$\mathbb{P}[z_i = 1] \leq O(\varepsilon).$$

This gives

$$\begin{aligned}\mathbb{P}[|\text{median}_i\ x_i - \mu| \geq \varepsilon\sigma] &= \mathbb{P}[\sum z_i \geq \frac{n}{2}], \\ &\leq e^{-\Omega(\varepsilon^2 n)}.\end{aligned}$$

This shows that the median of a univariate Gaussian is a good estimator of the mean, for all ε.

2.2 For general symmetric random variables

In the previous analysis, we only required

$$\mathbb{P}[|x - \mu| \leq \varepsilon\sigma] \gtrsim \varepsilon \forall \varepsilon < 1.$$

This is not true in general for all symmetric random variables.

This raises the following question: given $x_1, \cdots, x_n$, can we construct x' such that

$$\mathbb{P}[|x' - \mu| \leq \varepsilon\sigma] \gtrsim \varepsilon.$$

Using the following claim and the previous analysis, we can conclude that

$$\text{median}_{i\in[n/2]} \frac{x_{2i+1} + x_{2i+2}}{2},$$

will give a good estimate of the mean.

Note that is a simpler version of the Hodges-Lehmann estimator [**?**].

Claim 1. *If x_1, x_2 are i.i.d. and symmetric, then*

$$x' = \frac{x_1 + x_2}{2},$$

satisfies

$$\mathbb{P}[|x' - \mu| \leq \varepsilon\sigma] \gtrsim \varepsilon.$$

We now prove the claim.

Proof. Let

$$F_x(t) = \mathbb{E}_x[e^{i2\pi xt}]$$

denote the Fourier Transform of the random variable x.

Since x is symmetric, we have

$$F_x(t) = \mathbb{E}_x[\cos(2\pi xt)],$$

which is a real valued function.

By the definition of $x' = \frac{x_1+x_2}{2}$, we have

$$F_{x'}(t) = \frac{F_x^2(t)}{4} \geq 0,$$

which is non-negative everywhere because F_x is real-valued.

The following Lemma completes the proof:

Lemma 2 ([?]). *For any y such that $F_y(t) \geq 0$ symmetric about 0, $\operatorname{var}(y) = \sigma^2$, we have*

$$\forall \varepsilon < 1, \mathbb{P}[|y| \leq \varepsilon\sigma] \geq \Omega(\varepsilon).$$

□

Proof of Lemma **??**. Since y is symmetric about 0, we have

$$F_y(t) = \mathbb{E}_y[\cos(2\pi yt)] \geq \mathbb{E}[1 - \frac{(2\pi yt)^2}{2}] = 1 - 2\pi^2 t^2 \sigma^2.$$

Define the rectangular function

$$rect(y) = \mathbf{1}\left\{|y| \leq \varepsilon\sigma\right\},$$

and the corresponding triangular function

$$tri(y) = \mathbf{1}\left\{|y| \leq \varepsilon\sigma\right\}.$$

Note that the triangular function has a Fourier transform of

$$G(t) = \varepsilon\sigma sinc^2(\pi\sigma t) := \begin{cases} \varepsilon\sigma & t = 0, \\ \varepsilon\sigma \frac{\sin^2 \pi\sigma t}{(\pi\sigma t)^2} & \text{otherwise.} \end{cases}$$

We have

$$\begin{aligned} \mathbb{P}[|y| \leq \varepsilon\sigma] &= \int_y p(y) rect\,(y)\, dy, \\ &\geq \int_y p(y) tri(y) dy, \\ &= \int_t F_y(t) \varepsilon\sigma sinc^2(\pi\sigma t) dt, \\ &\geq \Omega(\frac{1}{\sigma}\varepsilon\sigma) = \Omega(\varepsilon), \end{aligned}$$

where the last bound follows since F_y is a parabola that is greater than constant for a width of $\Theta(\frac{1}{\sigma})$ and the Fourier transform of the triangle function has a value greater than $\Omega(\varepsilon\sigma)$ over a width of $\Theta(\frac{1}{\sigma})$.

□

References

[HL63] Hodges, J. L.; Lehmann, E. L. Estimates of Location Based on Rank Tests. Ann. Math. Statist. 34 (1963), no. 2, 598–611.

[MP14] Minton, Gregory T., and Eric Price. ”Improved concentration bounds for count-sketch.” Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2014.