CS 395T: Sublinear Algorithms Fall 2016

Lecture 7 — Sept. 15, 2016

Prof. Eric Price *Scribe: Zhao Liu, Changyong Hu*

1 Overview

In today's lecture, we will discuss the following problems:

1. Count-Min Review [CM05]
2. Count-Sketch [CCF02]
3. Fourier analysis of Count-Sketch [MP14]

2 Count-Min Review

Sketch:

Consider the problem for finding heavy hitters of a histogram x of a data stream, i.e. the k most frequent items (x_k is large). Assume the items are turnstile in $[n]$. Last time, we use R different hashes $y^{(r)}$ $(r = 1, ..., R)$ with each $y^{(r)}$ served as a counter of size $B = O(k)$. This is equivalent to think about storing a "table" with R rows and B columns. Each row of the table stands for a counter $y^{(r)}$, and a counter $y_i^{(r)}$ for each cell (r, i) of the table. The following is a summary:

1. Choose R pair-wise independent hash functions $h_1, h_2, \cdots, h_R : [n] \to [B]$.
2. For each hash function h_r, the counters $y_i^{(r)} = \sum_{u, h_r(u)=i} x_u$, where $r \in [R], i \in [B]$.

This is a linear function of x, so it can be expressed as a matrix. Given y, because it's an overestimate, in order to let failure probability decay exponentially, our recovered estimate $\hat{x}$ of x is taken as the minimum of each hash.

Recovery Algorithm

1. In each row, estimate $\hat{x}_u^{(r)} = y_{h_r(u)}^{(r)}$.
2. Overall, estimate $\hat{x}_u = \min_r \hat{x}_u^{(r)}$.

The error and space for count-min is

$$\|\hat{x} - x\|_\infty \leq \frac{\|x - x_k\|_1}{k} \text{ with probability } 1 - \frac{1}{n}, \text{space of } O(k \log(n)).$$

The intuition of this algorithm is trying to separate large terms from small terms and making use of sparsity. We briefly review the analysis of error here:

Analysis:

Let $H = \{1, \cdots, k\}$ be the indices of k most frequent elements and $T = \{k+1, \cdots, n\}$ be the rest. For a particular hash function h_r and element i:

$$\begin{aligned} |\hat{x}_i^{(r)} - x_i| &= \sum_{j \in H, h_r(i)=h_r(j)} x_j + \sum_{j \in T, h_r(i)=h_r(j)} x_j \\ &\leq \underbrace{0}_{\text{with probability } 1-\frac{k}{B}} + \underbrace{\|x_T\|_1/B}_{\text{in expectation}} \\ &\leq \underbrace{0}_{\text{with probability } \frac{9}{10}} + \underbrace{\frac{\|x_T\|_1}{k}}_{\text{with probability } \frac{9}{10}} \end{aligned}$$

if we set $B = 10k$. Thus for each r and i, by a union bound we have

$$\hat{x}_i^{(r)} - x_i = |\hat{x}_i^{(r)} - x_i| \leq \frac{\|x_T\|_1}{k} \text{ with probability } \frac{8}{10}.$$

This implies that

$$\hat{x}_i = \min_r \hat{x}_i^{(r)} \leq x_i + \frac{\|x_T\|_1}{k} \text{ with probability } 1 - (\frac{1}{5})^R$$

Choose $R = O(\log n)$, then

$$BR = O(k \log n),\ 1 - (\frac{1}{5})^R = 1 - n^{-c},\ \text{where } c \text{ is a constant value.}$$

What if some coordinates are negative? For some error $\sigma = O(\|x_T\|_1/k)$, we still have $Pr[|\hat{x}_i^{(r)} - x_i| \leq \sigma] \geq \frac{4}{5}$. Then after R samples, with $1 - e^{-O(R)}$ probability we will have that at least $\frac{n}{2}$ of the $\hat{x}_i^{(r)}$ will land in $x_i \pm \sigma$. Their median then lands in that region. So "count-median" will work in this case.

3 Count-Sketch

3.1 Problem

Suppose x has one large element ($O(n)$) and $n-1$ small elements ($O(1)$). With count-min algorithm:

- $\hat{x_u} \approx \frac{n}{B}$ if x_u is not the big element.
- $\hat{x_u} \approx C + \frac{n}{B}$ if x_u is the big element.

Then the error will approximately be $\|\hat{x} - x\|_\infty \approx \frac{n}{B}$. We want to get a more precise estimation of error like, for example, $\sqrt{\frac{n}{B}}$.

3.2 Setup

The idea of count-sketch is to introduce random signs in the summation, so that the errors tend to cancel each other out. For example, think about the case when the $O(1)$ elements are ± 1 in the count-min algorithm. Then $\hat{x}_u$ has mean 0 (or c) and standard deviation of $O(\sqrt{\frac{n}{B}})$ (because $\hat{x}_u$ is sum of about $\frac{n}{B}$ x_i's.)

Sketch:

1. Choose R pair-wise independent hash functions $h_1, h_2, \cdots, h_R : [n] \to [B]$ and $s_1, \ldots, s_R : [n] \to \{\pm 1\}$. For each hash function h_r, we need B counters.
2. $\forall r \in [R], \forall i \in [B], y_i^{(r)} = \sum_{u,s.t.h_r(u)=i} s_r(u) x_u$.

It is a linear function of x, which can be expressed as a matrix (with random entries in $\{\pm 1\}$). Given y, we recover our estimate $\hat{x}$ of x by:

Recovery Algorithm

1. In each row, estimate $\hat{x}_u^{(r)} = s_r(u) y_{h_r(u)}^{(r)}$.
2. Overall, estimate $\hat{x}_u = \underset{r \in [R]}{\text{median}}\ \hat{x}_u^{(r)}$.

The only difference from Count-Min is the introduction of random signs s_r, and the use of the median for estimation. As follows, we estimate its error.

3.3 Analysis of error

First, let's bound the term $|\hat{x}_u^{(r)} - x_u|$ for any $r \in [R]$ by upper bounding its variance:

$$\begin{aligned}
\mathbb{E}[(\hat{x}_u^{(r)} - x_u)^2] &= \mathbb{E}[(\sum_{u' \neq u, h(u')=h(u)} x_{u'} \cdot s_r(u') \cdot s_r(u))^2] \\
&= \mathbb{E}[\sum_{u' \neq u, h(u')=h(u)} {x_{u'}}^2 \cdot 1_{h(u')=h(u)} + \sum_{u' \neq u'' \neq u, h(u')=h(u'')=h(u)} x_{u'} x_{u''} s_r(u') s_r(u'')] \\
&= \frac{1}{B} \cdot \sum_{u' \neq u, h(u')=h(u)} x_{u'}^2 \\
&\leq \frac{\|x\|_2^2}{B}
\end{aligned}$$

where the third equality follows from pairwise independence of s_r. Applying Chebyshev Inequality,

$$Pr\big[|\hat{x}_u^{(r)} - x_u| \leq \frac{2\|x\|_2}{\sqrt{B}}\big] \geq \frac{3}{4}$$

But the above upper bound is not optimal. It is reasonable to think that after applying median, as $R \to \infty$, the error will converge to 0.

We claim that usually, we have

$$|\hat{x}_u^{(r)} - x_u| \ll \frac{\|x\|_2}{\sqrt{B}}$$

To show this claim and estimate the error, we start with considering a simplified example: sampling from independent normal distribution. Let $z_1, z_2, \cdots, z_R \sim N(0,1)$, $z^* = \underset{r\in[R]}{\text{median}}\ z_r$. How will the median z^* decay with R? Let $V_r := 1_{\{\text{event } z_r \geq \varepsilon\}}$. Because z_r's are symmetric,

$$Pr[|z^*| > \varepsilon] \leq 2Pr[z^* > \varepsilon] \leq 2Pr[at\ least\ \frac{R}{2}\ of\ z_r\ are\ \geq \varepsilon]$$

$$\begin{aligned}\mathbb{E}[V_r] =& Pr[z_r \geq \epsilon] \\ =& \frac{1}{2} \cdot Pr[0 \leq z_r \leq \varepsilon] \\ =& \frac{1}{2} - \Omega(\varepsilon)\end{aligned}$$

where the last equality follows from $Pr[0 \leq z_r \leq \epsilon] = \int_0^\epsilon \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \approx O(\frac{\epsilon}{\sqrt{2\pi}})$. Therefore,

$$E[\sum V_r] \leq R(\frac{1}{2} - \Theta(\varepsilon))$$

From Chernoff bound,

$$Pr\big[\sum V_r - E[\sum V_r] \geq \Theta(\varepsilon R)\big] \leq e^{-\frac{\varepsilon^2 R^2}{2R}}$$

Thus we have

$$Pr[|z^*| \geq \varepsilon] \leq Pr\big[\sum V_r \geq \frac{R}{2}\big] \leq 2e^{-\Theta(\varepsilon^2 R)}$$

We can set $\varepsilon = \frac{1}{\sqrt{R}}$ for constant probability.

If the random variables z_r are not Gaussian, with some appropriate conditions (symmetric and concentration of probability around 0), following similar arguments as above, we have

Theorem 3.1. *Let z_i be independent variables symmetric around 0, with $E[{z_i}^2] = 1$ and $Pr[|z_i| \leq \varepsilon] \gtrsim \varepsilon$ for any ϵ less than some constant, then their median*

$$|z^*| \lesssim \frac{1}{\sqrt{R}}\ with\ \frac{3}{4}\ probability.$$

Back to our case, let $z^{(r)} = \hat{x}_u^{(r)} - x_u = \sum_{u'=u, h(u')=h(u)} x_{u'} \cdot s_r(u') \cdot s_r(u)$. These are symmetric random variables. To satisfy the conditions of above theorem, one can resort to Fourier Analysis described below.

4 Fourier Analysis of Count-Sketch[MP14]

You can actually give a tighter analysis of Count-Sketch, which shows that *most* coordinates are estimated to higher precision, if your hash functions are fully independent. As we described in an earlier class, the assumption of fully independent hash functions is unfortunate, but it can be justified using cryptographic hash functions and computational assumptions on the adversarial input, or assuming the input has high entropy.

Note that the details of this analysis also can be found in Eric Price's presentation slide of SODA'2015. Here is the link : http://www.cs.utexas.edu/ ecprice/slides/concentration-slides.pdf.

Theorem 4.1. *Let z be symmetric random variable and $E[z^2] = 1$, if $\mathcal{F}_z(t) \geq 0, \forall t$, then*

$$Pr[|z| \leq \varepsilon] \gtrsim \varepsilon$$

Theorem 4.2. *Assume that h and s are fully independent hash functions, and consider the output $\hat{x}$ of Count-Sketch. Then $\forall t \leq R$, we have*

$$|\hat{x}_i - x_i| \leq \sqrt{\frac{t}{R}} \cdot \frac{\|x_T\|_2}{\sqrt{k}}$$

with probability $1 - e^{-\Omega(t)}$.

This implies that $\mathbb{E}[\hat{x}_i - x_i] \leq \frac{1}{\sqrt{R}} \cdot \frac{\|x_T\|_2}{\sqrt{k}}$ after excluding $e^{-\Omega(R)}$ events.

Theorem 4.3. *For any set S of size K,*

$$Pr\big[\|\hat{x}_s - x_s\|_2 \geq O(c) \cdot \sqrt{\frac{K}{R}} \cdot \sigma\big] \leq \frac{1}{K^c} \text{ for any } c > 0.$$

References

[CCF02] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding Frequent Items in Data Streams. *ICALP*, 2002.

[CM05] Graham Cormode and S. Muthukrishnan. Summarizing and Mining Skewed Data Streams. *SDM*, 2005.

[MP14] Gregory T. Minton and Eric Price. Improved Concentration Bounds for Count-Sketch *SODA(best student paper)* 2014.