CS49/249 (Randomized Algorithms), Spring 2021 : Lecture 17

Topic: Streaming II : Count Sketch and Estimating F_2.

Disclaimer: These notes have not gone through scrutiny and in all probability contain errors. Please discuss in Piazza/email errors to deeparnab@dartmouth.edu

---

- **An Unbiased Estimate for Frequencies.** In the previous lecture, we saw that COUNTMIN gave an estimate for $\mathbf{f}_i$ which was biased. And yet, at least in the insertion only model, for any i, we could bound the error to within $\varepsilon \|\mathbf{f}\|_1$ with probability $\geq 1-\delta$. The space used was $O(\frac{1}{\varepsilon}\lg(1/\delta))$.

  In this lecture we will see an *unbiased* estimate to $\mathbf{f}_i$. Chronologically, this algorithm called COUNT-SKETCH due[1] to Moses Charikar, Kevin Chen, and Martin Farach-Colton predated the COUNT-MIN algorithm. We will see it is in some sense incomparable to COUNT-MIN. On the one-hand, the dependence on ε will be $\frac{1}{\varepsilon^2}$ instead of $\frac{1}{\varepsilon}$. On the other hand, the accuracy of the estimate will be better: we will have $|\widehat{\mathbf{f}}_i - \mathbf{f}_i| \leq \varepsilon \|\mathbf{f}\|_2$. And the 2-norm of a vector is always at most the 1-norm, and can be much smaller.

  The idea is still the same : one hashes $[n]$ to $[k]$, maintains k-counters, and updates only the counters of the hashes. Except, COUNT-SKETCH multiplies each entry with a random $\{-1,1\}$ factor. More precisely, upon receiving the update (i,c), the counter $C[h(i)]$ is updated to $C[h(i)] + c\cdot g(i)$ where $g(i) \in \{-1,+1\}$ with probability $1/2$. That is, g is another hash function drawn from a UHF with domain $[n]$ and range $\{-1,1\}$. Note that the counter values can now be negative. At the end, the final estimate is *corrected* by multiplying with the same $g(i)$; the estimate $\widehat{\mathbf{f}}_i := C[h(i)]\cdot g(i)$.

  The intuition is this : if nothing collided with i, then $C[h(i)]$ would precisely contain $g(i)\cdot \mathbf{f}_i$. And thus, when one multiplies with $g(i)$ again, because $g(i) \in \{\pm 1\}$, $\widehat{\mathbf{f}}_i$ would be $\mathbf{f}_i$. On the other hand for the collisions for $j \neq i$, the expected value of $g(i)g(j)$ is 0. This leads to an unbiased estimate. One can then do the median-of-means trick to obtain an (ε,δ)-estimate. The variance contains the ℓ_2-norm $\|\mathbf{f}\|_2$ of the frequency vector.

- *Algorithm.*

procedure COUNT-SKETCH(ε):
  Let H be a universal hash family with domain size $[n]$ and range $k = \lceil \frac{3}{\varepsilon^2} \rceil$.
  Let G be a strongly universal hash family[a] with domain size $[n]$ and range $\{-1,+1\}$.
  Maintain counters $C[1:k]$. Draw $h \sim H$ and $g \sim G$
  for stream update (i,c): **do** ▷ *Assume $c \geq 0$*
    Update $C[h(i)] \leftarrow C[h(i)] + c\cdot g(i)$.
  For $1 \leq i \leq n$, **return** $\widehat{\mathbf{f}}_i = g(i)C[h(i)]$.

[a] For $i \neq j$, probability $g(i) = a$ and $g(j) = b$ for any $a, b \in \{-1,1\}$ is $1/4$.

[1] Charikar, Moses, Kevin Chen, and Martin Farach-Colton. *"Finding frequent items in data streams."* International Colloquium on Automata, Languages, and Programming (ICALP), 2002.

Remark: *The algorithm works in the dynamic streaming or turnstile model which allows c's to be negative.*

- *Analysis.*

Theorem 1 (Basic Count Sketch Analysis.)**.** For any $1 \leq i \leq n$, with probability $\geq \frac{2}{3}$, we have $|\widehat{\mathbf{f}}_i - \mathbf{f}_i| \leq \varepsilon \|\mathbf{f}\|_2$. The space required is $O(\frac{1}{\varepsilon^2})$. Therefore, taking $O(\ln(1/\delta))$-parallel estimates and returning the median, gives the same result with probability $\geq 1 - \delta$, with a space blow up of $O(\ln(1/\delta))$.

  - As in the analysis of COUNT-MIN, we observe that for any $i \in [n]$,

$$C[h(i)] = g(i) \cdot \mathbf{f}_i \; + \sum_{j \neq i : h(j) = h(i)} g(j) \cdot \mathbf{f}_j \tag{1}$$

    And therefore,

$$\widehat{\mathbf{f}}_i = \mathbf{f}_i + \sum_{j \neq i : h(j) = h(i)} g(i)g(j) \cdot \mathbf{f}_j$$

    Taking expectations over the choices of both h and g, we get

$$\mathbf{Exp}[\widehat{\mathbf{f}}_i] = \mathbf{f}_i + \sum_{j \neq i} \mathbf{Pr}[h(j) = h(i)] \cdot \mathbf{f}_j \mathop{\mathbf{Exp}}_{g}[g(i)g(j)]$$

    Now, $\mathbf{Exp}_g[g(i)g(j)] = 1 \cdot \mathbf{Pr}[g(i) = g(j)] + (-1)\,\mathbf{Pr}[g(i) \neq g(j)] = 0$. This gives us that $\widehat{\mathbf{f}}_i$ is an unbiased estimate.
  - Let's now calculate the variance of $\widehat{\mathbf{f}}_i$. To do so, we first observe that $\mathbf{Exp}[\widehat{\mathbf{f}}_i^2] = \mathbf{Exp}[(C[h(i)])^2]$. Using (1), we get

$$(C[h(i)])^2 = \mathbf{f}_i^2 + \sum_{j \neq i : h(j) = h(i)} \mathbf{f}_j^2 + \sum_{j,k : j \neq k, h(j) = h(k) = h(i)} g(j)g(k)\mathbf{f}_j\mathbf{f}_k$$

    Taking expectations, we see

$$\mathbf{Exp}[\widehat{\mathbf{f}}_i^2] = \mathbf{f}_i^2 + \sum_{j \neq i} \frac{\mathbf{f}_j^2}{k} \quad \Rightarrow \quad \mathbf{Var}[\widehat{\mathbf{f}}_i] = \mathbf{Exp}[\widehat{\mathbf{f}}_i^2] - \left(\underbrace{\mathbf{Exp}[\widehat{\mathbf{f}}_i]}_{= \mathbf{f}_i} \right)^2 \leq \frac{1}{k} \cdot \|\mathbf{f}\|_2^2$$

    where again we have used the fact that for *any* $j \neq k$, we have $\mathbf{Exp}[g(j)g(k)] = 0$.
  - And therefore Chebyshev's inequality gives us that if $k \geq \frac{3}{\varepsilon^2}$, then

$$\mathbf{Pr}[|\widehat{\mathbf{f}}_i - \mathbf{f}_i| \geq \varepsilon \|\mathbf{f}\|_2] \leq \frac{\mathbf{Var}[\widehat{\mathbf{f}}_i]}{\varepsilon^2 \|\mathbf{f}\|_2^2} \leq \frac{1}{3}$$

- **Estimating F_2: the "Tug-of-War" Sketch.** Continuing on our theme of presenting things in a non-chronological order, let us look at the algorithm for estimating F_2. Recall, $F_2 := \|\mathbf{f}\|_2^2$ is the squared ℓ_2-norm of the frequency vector. This question, and more generally the question of studying other frequency moments, was initiated in a seminal paper[2] by Noga Alon, Yossi Matias, and Mario Szegedy. Although it was not the first paper on streaming algorithms, this arguably opened the floodgates in this area. Indeed, the algorithm shown below for estimating F_2 contains the main ideas for COUNT-SKETCH, which in turn contains the main ideas behind COUNT-MIN.

  The main idea is to pick a hash function $g : [n] \to \{-1, 1\}$ from a hash family. We want something stronger than UHFs here. For the lecture, you can keep the mental model of each $g(i)$ being ± 1 uniformly at random and *independent* of $g(j)$. Formally, what we need is that the family G has 4-wise independence. That is, we need

  For any four distinct $w, x, y, z \in [n]$, and any four $a, b, c, d \in \{-1, +1\}$,

  $$\mathop{\mathbf{Pr}}_{g\sim G}[g(w) = a,\ g(x) = b,\ g(y) = c,\ g(z) = d] = \frac{1}{16} \qquad \text{(4-way independence)}$$

  The algorithm maintains an estimate Z initialized to 0. Upon encountering update (i, c), it simply updates $Z \leftarrow Z + g(i) \cdot c$. At the end, it returns Z^2 as the estimate of $\|\mathbf{f}\|_2^2$. I hope you all see the similarity with COUNT-SKETCH. Alon, Mattias, and Szegedy called[3] this the "Tug-of-War" algorithm.

```
1: procedure AMS TUG-OF-WAR:
2:     Let G be a 4-wise independent universal hash family with domain size [n] and range
   {−1, +1}.
3:     Draw g ∼ G. Initialize Z ← 0.
4:     for update (i, c): do ▷ Assume c ≥ 0
5:         Update Z ← Z + c · g(i).
6:     return Z^2.
```

- *Analysis.*

  Theorem 2 (Tug-of-War Analysis.)**.** For any ε, δ, there is an algorithm taking $O(\frac{1}{\varepsilon^2}\ln(1/\delta))$-words of space which with probability $(1 - \delta)$ computes an $(1 \pm \varepsilon)$ estimate to $F_2 = \|\mathbf{f}\|_2^2$.

  – As in the COUNT-SKETCH analysis, we begin by observing that at the end of the stream,

  $$Z = \sum_{i=1}^{n} g(i)\mathbf{f}_i \quad \Rightarrow \quad Z^2 = \left(\sum_{i=1}^{n} g(i)\mathbf{f}_i\right)^2 = \sum_{i=1}^{n} \mathbf{f}_i^2 + 2 \sum_{1\le i<j\le n} g(i)g(j)\mathbf{f}_i\mathbf{f}_j$$

  And again using $\mathbf{Exp}_{g\sim G}[g(i)g(j)] = 0$, we get that $\mathbf{Exp}[Z^2] = \|\mathbf{f}\|_2^2$. That is, the algorithm returns an unbiased estimate.

[2] Noga Alon, Yossi Matias, and Mario Szegedy. *The space complexity of approximating the frequency moments.* J. Comput. Syst. Sci., 58(1):137–147, 1999.

[3] To be precise, this was named Tug-of-War in a follow-up paper of Noga Alon, Phillip B. Gibbons, Yossi Matias, Mario Szegedy: *Tracking Join and Self-Join Sizes in Limited Storage.* J. Comput. Syst. Sci. 64(3): 719-747 (2002)

– To calculate the variance of Z^2, we need to compute the expectation of Z^4. This is the fourth power of a sum of the $g(i)\mathbf{f}_i$'s. When one opens this up, one gets many terms. However *any* term that contains the product $g(i)g(j)$ for two distinct i, j will "vanish" when we take the expectation. In fact so will the expectation of $g(i)g(j)g(k)g(\ell)$ for four distinct i, j, k, ℓ, and this just needs (4-way independence) (and the only place where it is used). So, once we take the expectation of Z^4, all that will remain are the fourth powers of $\mathbf{f}_i$'s *and* the products of $\mathbf{f}_i^2\mathbf{f}_j^2$'s. More precisely,

$$\mathbf{Exp}[Z^4] = \sum_{i=1}^{n} \mathbf{f}_i^4 + 6 \sum_{1 \le i < j \le n} \mathbf{f}_i^2 \mathbf{f}_j^2$$

Now note that

$$\left(\mathbf{Exp}[Z^2]\right)^2 = \left(\sum_{i=1}^{n} \mathbf{f}_i^2\right)^2 = \sum_{i=1}^{n} \mathbf{f}_i^4 + 2 \sum_{1 \le i < j \le n} \mathbf{f}_i^2 \mathbf{f}_j^2$$

which implies

$$\mathbf{Var}[Z^2] = 4 \sum_{1 \le i < j \le n} \mathbf{f}_i^2 \mathbf{f}_j^2 \le 2 \left(\mathbf{Exp}[Z^2]\right)^2$$

where the inequality follows by simply comparing the above two equalities.

– Therefore, $\mathbf{Var}[Z^2]/(\mathbf{Exp}[Z^2])^2 \le 2$, which implies via the median-of-means theorem that taking $O\left(\frac{1}{\varepsilon^2} \ln(1/\delta)\right)$ parallel estimates of Z^2 can lead to an (ε, δ)-estimate for F_2.