Multiplicative vs. Additive Error

For point queries, we wanted to return, for any $i \in [n]$, an estimate $\hat{f}_i$ of f_i such that $|f_i - \hat{f}_i| \leq \epsilon \cdot m$. This is called an additive approximation with error $\epsilon \cdot m$.

Often, we will be interested in multiplicative error.

Definition: Let $A(\sigma)$ denote the output of a randomized streaming algorithm A on input σ. Let ϕ be the function it is supposed to compute. We say A outputs a (ϵ, δ)-multiplicative approximation to ϕ if we have for all streams σ:

$$\Pr\left[\left|A(\sigma) - \phi(\sigma)\right| \geq \epsilon \cdot \phi(\sigma)\right] \leq \delta$$

We also say a number N is a $(1 \pm \epsilon)$-multiplicative approximation to $\phi(\sigma)$ if $|N - \phi(\sigma)| \leq \epsilon \cdot \phi(\sigma)$.

Note: Multiplicative error means if the "right answer" is small, the additive error must also be small.

In particular, if $\phi(\sigma) = 0$, then the answer must be exact.

Sometimes this is impossible to achieve without storing the whole data stream (e.g. for point queries).

AMS Sampling for Frequency-Based Functions in the Vanilla Streaming model

Problem: Estimate $\sum_{i=1}^{n} g(f_i)$, where g is any function with $g(0)=0$. (Want an (ε,δ)-multiplicative approximation.)

Outline: Give an estimator that equals the right answer in expectation. Bound its variance. Average a bunch of copies of it to drive down the variance. Apply Chebyshev's inequality to conclude it's within a $(1\pm\varepsilon)$-factor of the right answer with probability 3/4. Do this a bunch of times and take the median to drive failure probability below δ.

Basic Estimator (unbiased, but high variance): from stream $\langle a_1, \ldots a_m \rangle$, sample a value j at random from $\{1,\ldots m\}$ and compute

$r = |\{\ell \geq j : a_\ell = a_j\}|$ = # of times a_j occurs in stream after update j.

Output $X := m \cdot (g(r) - g(r-1))$.

(1 a,m): $\mathbb{E}[X] = \sum_{i=1}^{n} g(f_i)$.

Proof: $\mathbb{E}[X] = \sum_{i=1}^{n} \Pr[a_j = i] \cdot \mathbb{E}[X \mid a_j = i]$

$= \sum_{i=1}^{n} \frac{f_i}{m} \cdot \sum_{r=1}^{f_i} \frac{m \cdot (g(r) - g(r-1))}{f_i}$

$= \sum_{i=1}^{n} \sum_{r=1}^{f_i} g(r) - g(r-1) = \sum_{i=1}^{n} g(f_i) - g(0) = \sum_{i=1}^{n} g(f_i)$ used that $g(0)=0$

To see this, note:

$\sum_{i=1}^{n} \Pr[a_j = i] \cdot \mathbb{E}[X \mid a_j = i]$

$= \sum_{i=1}^{n} \Pr[a_j = i] \cdot \sum_{z} z \Pr[X = z \mid a_j = i]$

$= \sum_{i=1}^{n} \Pr[a_j = i] \cdot \sum_{z} z \frac{\Pr[X = z \text{ and } a_j = i]}{\Pr[a_j = i]} = \sum_{i=1}^{n} \sum_{z} z \cdot \Pr[X = z \text{ and } a_j = i] = \sum_{z} z \cdot \sum_{i=1}^{n} \Pr[X = z \text{ and } a_j = i]$

$= \sum_{z} z \cdot \Pr[X = z] = \mathbb{E}[X]$

Let us now look at AMS sampling specifically, in the context of $g(r) = r^k$. That is, we want to compute $F_k := \sum_{i=1}^{n} f_i^k$.

F_k is called the k^{th} frequency moment of the stream.

E.g. if the frequency vector is interpreted as the empirical distribution of an underlying distribution τ over $[n]$ (i.e. all stream updates are interpreted as a random draw from τ), then F_2 is proportional to the sample variance i.e. $\sum_{i=1}^{n} f_i^2 - \left(\sum_{i=1}^{n} f_i\right)^2 = F_2 - m^2$.

F_2 also has applications to data bases (self-join size)

Claim 2: $Var[X] \le k \cdot n^{1-1/k} \cdot F_k^2$. Intuitively, this means a standard deviation for basic estimator is $\sqrt{k} n^{1-1/k}$ times larger than what we get (what if about F_k). So will need to average about $k \cdot n^{1-1/k}$ copies to drive down standard deviation

Before proving Claim 2, let us explain why it is enough to give an (ϵ, δ)-multiplicative approximation with space cost $O\left(\frac{1}{\epsilon^2} \cdot \log\left(\frac{1}{\delta}\right) \cdot k \cdot n^{1-1/k} \cdot (\log m + \log n)\right)$.

For constant k, and using $\tilde{O}$ notation to hide factors polylogarithmic in m, n, and $\frac{1}{\delta}$, this is $\tilde{O}\left(\frac{n^{1-1/k}}{\epsilon^2}\right)$. Later in the course we will see an algorithm achieving optimal space $\tilde{O}\left(\frac{n^{1-2/k}}{\epsilon^2}\right)$.

- Final estimator: First, drive down variance further by averaging several independent copies of the Basic Estimator. Call this the Mean Estimator. Second, output the median of several independent copies of the Mean Estimator.
- Analysis: Chebyshev's inequality implies the Mean Estimator is a $(1 \pm \epsilon)$-approximation with probability $\frac{3k}{\epsilon^2} \cdot n^{1-1/k}$. Chernoff Bounds imply the Median-of-Means estimator is $\Theta(\log(\frac{1}{\delta}))$

more precisely, we the following version of the Chernoff bound for additive error.

Let $e_1 \ldots e_t$ be i.i.d random variables taking values in $[0,1]$, and let $\mu = \sum_{i=1}^{t} \mathbb{E}[e_i]$ and $e = \sum_{i=1}^{t} e_i$. Then $\Pr[|X - \mu| \geq \lambda] \leq 2\exp\left(-\frac{2\lambda^2}{t}\right)$

So when $\lambda = \frac{t}{4}$, we get the probability is at most $2\exp\left(-\frac{2(\frac{t}{4})^2}{t}\right)$

$= \exp(-8t)$. For $t = \frac{1}{8}\log\left(\frac{1}{\delta}\right)$, this is δ.

Claim 3: There is a universal constant c such that the following holds. Let X be an unbiased estimator for a real quantity Q. Let $\{X_{ij}\}_{i \in [t], j \in [\ell]}$ be a collection of i.i.d. variables each distributed identically to X, where $t = c \cdot \log(\frac{1}{\delta})$ and $\ell = \frac{4 \, Var[X]}{\epsilon^2 \cdot \mathbb{E}[X]^2}$.

(by claim 2, this is at most $\frac{4 \cdot k \cdot n^{1-1/k}}{\epsilon^2}$)

(cancels all the F_k in Var)

Let $Z = \text{median}_{i \in [t]} \left(\frac{1}{\ell} \sum_{j=1}^{\ell} X_{ij} \right)$. Then $\Pr[|Z - Q| \geq \epsilon Q] \leq \delta$.

For each $i \in [t]$, let $Y_i := \frac{1}{\ell} \sum_{j=1}^{\ell} X_{ij}$.

Proof: By linearity of expectation,

$$\mathbb{E}[Y_i] = \frac{1}{\ell} \sum_{j=1}^{\ell} \mathbb{E}[X_{ij}] = \frac{1}{\ell} \sum_{j=1}^{\ell} Q = Q.$$

Since the variables X_{ij} are independent,

$$Var[Y_i] = \frac{1}{\ell^2} \sum_{j=1}^{\ell} Var[X_{ij}] = \frac{Var[X]}{\ell}.$$

By Chebyshev, $\Pr[|Y_i - Q| \geq \epsilon Q] \leq \frac{Var[Y_i]}{(\epsilon \cdot Q)^2} = \frac{Var[X]}{\ell \cdot \epsilon^2 \cdot \mathbb{E}[X]^2} = \frac{1}{4}$

to see this, note that

$$\epsilon \cdot Q = \frac{\epsilon \cdot Q \cdot \sigma(Y_i)}{\sqrt{Var(Y_i)}}$$

so Chebyshev says the LHS is at most the inverse of the boxed quantity squared.

Hence, Chernoff bounds tell us that the probability the median of the Y_i's has error more than $\epsilon \cdot Q$ is at most $\frac{1}{\delta}$.

(Apply Chernoff bounds to the variables $e_1, \ldots, e_t$ where $e_i = 1$ if $|Y_i - Q| \geq \epsilon Q$, and 0 otherwise.)

Proof of (Claim 2):

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \le \mathbb{E}[X^2]$$

$$= \sum_{j=1}^{n} \frac{f_j}{m} \sum_{r=1}^{f_j} \frac{1}{f_j} \cdot m^2 \left(r^k - (r-1)^k\right)^2 = m \sum_{j=1}^{n} \sum_{r=1}^{f_j} \left(r^k - (r-1)^k\right)^2.$$

By the mean value theorem, there exists a $\tau(z) \in [z-1, z]$ such that

$$z^k - (z-1)^k = k \cdot \tau(z)^{k-1} \le k \cdot z^{k-1}, \text{ where the last step uses } k \ge 1.$$

Hence $\le m \sum_{j=1}^{n} \sum_{r=1}^{f_j} k \cdot r^{k-1} \cdot \left(r^k - (r-1)^k\right)$

$$\le m \sum_{j=1}^{n} k \cdot f_j^{k-1} \sum_{r=1}^{f_j} \left(r^k - (r-1)^k\right)$$

$$= m \sum_{j=1}^{n} k f_j^{k-1} f_j^k = k \cdot m \cdot \sum_{j=1}^{n} f_j^{2k-1}.$$

$$= k \cdot \left(\sum_{j=1}^{n} f_j\right) \cdot \sum_{j=1}^{n} \left(f_j^{2k-1}\right).$$

Lemma: $\le n^{1-\frac{1}{k}} \cdot F_k^2.$

Proof: Let $f_{max} = \max_i f_i$. Then $f_{max}^{k-1} = \left(f_{max}^k\right)^{\frac{k-1}{k}} \le \left(\sum_{i=1}^{n} f_i^k\right)^{\frac{k-1}{k}}$.

By convexity of the function $x \mapsto x^k$, we have

$$\frac{1}{n} \sum_i f_i \le \left(\frac{1}{n} \sum_i f_i^k\right)^{1/k}. \quad \text{i.e.} \quad \left(\sum_i f_i\right) \le n^{1-1/k} \cdot \sum_i f_i^k$$

biggest gap between $\sum_i f_i$ & $\sum_i f_i^k$ is all f_i equal $\frac{1}{n}$ in $r=1$ and $r=n^{k-1}$

Hence $\left(\sum_i f_i\right)\left(\sum_i f_i^{2k-1}\right) \le \left(\sum_i f_i\right)\cdot\left(f_{max}^{k-1}\cdot\sum_i f_i^k\right)$

$\le \left(\sum_i f_i\right)\cdot\left(\sum f_i^k\right)^{\frac{k-1}{k}}\cdot\left(\sum_i f_i^k\right)$

$\le n^{1-1/k}\cdot\left(\sum_i f_i^k\right)\cdot$

$= n^{1-1/k}\cdot\left(\sum_i f_i^k\right)^2 = n^{1-1/k}\cdot F_k^2$.