Chernoff bounds: Let $X_1, \ldots, X_n$ be indep. r.v.'s where $0 \le X_i \le 1$. Let $X = \sum_{i=1}^{n} X_i$ & $\mu = E[X]$.

For $0 \le \delta \le 1$,

$$\Pr(X \ge (1+\delta)\mu) \le e^{-\delta^2 \mu/3}$$
$$\Pr(X \le (1-\delta)\mu) \le e^{-\delta^2 \mu/2}$$

Warm-up example: Median estimate

Let $S = [x_1, \ldots, x_m]$

for simplicity assume x_i's are distinct

& let $\text{rank}(y) = |\{x \in S : x \le y\}|$

Goal: Find ϵ-approx. median(S),

more precisely, given $\epsilon > 0$, find y where:

$$\frac{m}{2} - \epsilon m = \frac{m}{2}\left(1 - 2\epsilon\right) < \text{rank}(y) < \frac{m}{2}\left(1 + 2\epsilon\right) = \frac{m}{2} + \epsilon m$$

Alg.: Choose t random elements from S.
Let $R = [r_1, \ldots, r_t]$
Return median(R).

Claim: if $t \geq \frac{2}{\epsilon^2} \log(\frac{2}{\delta})$, then the alg. returns an ϵ-approx. median with prob. $\geq 1-\delta$.
This is an (ϵ, δ)-approximation of the median.

Proof: Let $S_L = \{x \in S : \text{rank}(x) \leq \frac{m}{2} - \epsilon m\}$

$S_M = \{x \in S : \frac{m}{2} - \epsilon m \leq \text{rank}(x) < \frac{m}{2} + \epsilon m\}$

$S_U = \{x \in S : \text{rank}(x) \geq \frac{m}{2} + \epsilon m\}$

If $< \frac{t}{2}$ elements of S_L & $< \frac{t}{2}$ of S_U are in R then the median of R is in S_M, and it's an ϵ-approx. median.

Let $X_i = \begin{cases} 1 & \text{if } r_i \in S_L \\ 0 & \text{o/w} \end{cases}$

& $X = \sum_{i=1}^{t} X_i$,

Then, $E[X] = t\left(\frac{n}{2} - \epsilon n\right) = \frac{t}{2} - \epsilon t$

$$\Pr\left(X \geq \frac{t}{2}\right) \leq \Pr\left(X \geq E[X]\left(1 + 2\epsilon\right)\right)$$

$$\leq e^{-4\epsilon^2\left(\frac{1}{2} - \epsilon\right)t/3}$$

$$\leq e^{-\frac{4\epsilon^2}{7}t} \leq \frac{\delta}{2} \quad \text{for } t \geq \frac{2}{\epsilon^2}\log\left(\frac{2}{\delta}\right).$$

& $\Pr\left(\geq \frac{t}{2} \text{ elements from } S_U \text{ in } R\right) \leq \frac{\delta}{2}$ by analogous argument.

Hence, $\Pr(\text{alg. outputs } \epsilon\text{-approx. median}) \geq 1 - \delta$. ∎

Streaming: we get one-by-one m elements

$$X_1, X_2, \ldots, X_m$$

where $X_i \in \{1, 2, \ldots, n\}$.

Think of m as huge so can't store entire stream.

Let f_i be the frequency of number i in the stream.

& $f = (f_1, \ldots, f_n)$.

Reservoir Sampling:

We don't know n beforehand, but we want to choose an element s uniformly at random from $x_1, \ldots, x_n$.

Alg.:

Set $s = x_1$

for $t > 1$:

- upon seeing t^{th} element x_t, with prob. $\frac{1}{t}$ set $s = x_t$

What's prob. that $s = x_i$ for some time $t \geq i$?

$$\Pr(s = x_i) = \frac{1}{i} \times \left(1 - \frac{1}{i+1}\right) \times \left(1 - \frac{1}{i+2}\right) \times \cdots \times \left(1 - \frac{1}{t}\right)$$

$$= \frac{1}{i} \times \left(\frac{i}{i+1}\right) \times \left(\frac{i+1}{i+2}\right) \times \cdots \times \left(\frac{t-1}{t}\right)$$

$$= \frac{1}{t}$$

— Takes $O(\log n)$ bits of space to get s & $O(k \log n)$ bits to get k samples.

AMS = [Alon, Matias, Szegedy '99]

Have function g where $g(0)=0$.

Goal: Compute $\sum_{i=1}^{\hat{n}} g(f_i)$

Unbiased estimator: r.v. X where $E[X] = \sum_{i=1}^{\hat{n}} g(f_i)$

Alg.: First choose random ~~sample~~ index $J \in_R \{1,\dots,m\}$ → do this using reservoir sampling scheme

& compute

$r = |\{j \geq J : x_j = x_J\}|$

= # of occurrences of x_J after time J.

Then, output $X = m \times (g(r) - g(r-1))$

Claim: $E[X] = \sum_{i=1}^{\hat{n}} g(f_i)$

Proof: $E[X] = \sum_i \Pr(X_J = i) E[X \mid X_J = i] = \sum_i \frac{f_i}{m} \sum_{r=1}^{f_i} \frac{m(g(r)-g(r-1))}{f_i}$

(f_i elements $= i$; elements of value i afterwards)

$= \sum_i g(f_i)$

How to ensure the output is (ϵ, δ)-approx.?

idea: Take t estimators & average?

Example: Frequency moments

For integer $k \geq 1$, let $F_k = \sum_{i=1}^{n} f_i^k$

So $g(r) = r^k$

Run AMS alg. & we have $X = m(r^k - (r-1)^k)$

We know: $E[X] = F_k$

Idea: Do t indep. trials to get $X_1, \ldots, X_t$
& output $\frac{1}{t}\sum_{i=1}^{t} X_i$

Hope to use Chernoff to show it's (ϵ, δ) approx.

But: X_i can be huge.

So show $Var(X)$ is small & apply Chebyshev's.

Lemma: $Var(X) \leq k n^{1-1/k} F_k^2$

For $\ell = \frac{3Var(X)}{\epsilon^2 E[X]^2} = \frac{3k n^{1-1/k} F_k^2}{\epsilon^2 F_k^2} = 3k n^{1-1/k} \epsilon^{-2}$,

let $Y = \frac{1}{\ell}\sum_{i=1}^{\ell} X_i$ = mean of ℓ trials.

$$E[Y] = E[X_i] = F_k$$

$$\mathrm{Var}(Y) = \frac{1}{\ell^2}\sum_{i=1}^{\ell}\mathrm{Var}(X_i) = \frac{\mathrm{Var}(X)}{\ell}$$

By Chebyshev's ineq.:

$$\Pr\left(|Y - E[Y]| \geq \epsilon E[Y]\right)$$
$$= \Pr\left(|Y - F_k| \geq \epsilon F_k\right)$$
$$\leq \frac{\mathrm{Var}(Y)}{(\epsilon F_k)^2} \leq \frac{\mathrm{Var}(X)}{\ell \epsilon^2 F_k^2} \leq \frac{k n^{1-\frac{1}{k}} F_k^2}{3k n^{1-\frac{1}{k}} \epsilon^{-2} \epsilon^2 F_k^2} = \frac{1}{3}$$

So with prob. $\geq \frac{2}{3}$ it's an ϵ-approx. of F_k

How to boost this prob. to $1-\delta$?

Repeat t times & take the median.

Now we do this $t = c\log(1/\delta)$ times, call them $Y_1, \ldots, Y_t$

Let $Z_j = \begin{cases} 1 & \text{if } |Y_j - F_k| \leq \epsilon F_k \text{ so it's an } \epsilon\text{-approx} \\ 0 & \text{o/w} \end{cases}$

Then, for $Z = \sum Z_j$, $E[Z] \geq \frac{2}{3}t$

If $Z \geq \frac{t}{2}$ then the median is an ϵ-approx.

$$\Pr\left(Z \leq \tfrac{t}{2}\right) \leq \Pr\left(Z \leq E[Z]\left(1 - \tfrac{1}{6}\right)\right) \leq e^{-\frac{2}{3}t/36^2} \leq \delta$$

for big enough c.

Proof of Lemma:

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

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

$$= m \sum_j \sum_r \left(r^k - (r-1)^k\right)^2$$

see claim on next page $\rightarrow \overset{?}{\le} m \sum_j \sum_r k r^{k-1} \left(r^k - (r-1)^{k-1}\right)$

$$\le m \sum_j k f_j^{k-1} \sum_r \left(r^k - (r-1)^{k-1}\right)$$

$$= m \sum_j k f_j^{k-1} f_j^k$$

$$= k F_1 F_{2k-1}$$

$$\overset{?}{\le} k n^{1-\frac{1}{k}} F_k^2$$

See claim 2 on Page 11

Claim: $r^k - (r-1)^k \leq kr^{k-1}$

Mean-value theorem:

for function f which is continuous on $[a,b]$ & differentiable on (a,b)

then $\exists\, c \in (a,b)$ where:

$$f'(c) = \frac{f(b)-f(a)}{b-a}$$

Thus, $\exists\, l \in [r-1, r]$ where: for $f(z) = z^k$,

$$f'(l) = \frac{f(r)-f(r-1)}{1}$$

$$kl^{k-1} = r^k - (r-1)^{k-1}$$

Therefore,

$$r^k - (r-1)^{k-1} = kl^{k-1} \leq kr^{k-1}$$

∎

$$F_1 F_{2k-1} = \left(\sum_i f_i\right)\left(\sum_i f_i^{2k-1}\right)$$

$$\leq \left(\sum_i f_i\right)\left(f_*^{k-1} \sum_i f_i^k\right) \quad \text{where } f_* = \max_i f_i$$

$$\leq \left(\sum_i f_i\right)\left(f_*^k\right)^{k-1/k}\left(\sum_i f_i^k\right)$$

$$\leq \left(\sum_i f_i\right)\left(\sum_i f_i^k\right)^{k-1/k}\left(\sum_i f_i^k\right)$$

$$\overset{?}{\leq} n^{1-\frac{1}{k}}\left(\sum_i f_i^k\right)^{1/k}\left(\sum_i f_i^k\right)^{k-1/k}\left(\sum_i f_i^k\right)$$

$$= n^{1-\frac{1}{k}} F_k^2$$

$$\frac{1}{n}\sum_i x_i \leq \left(\frac{1}{n}\sum_i x_i^k\right)^{1/k}$$

AM $\leq$ k^{th}-power mean

Power-mean inequality.