Lecture 2

Barna Saha

AT&T-Labs Research

September 12, 2013

Outline

First and Second Moment Bounds

- **Markov Inequality** For any positive random variable X and $t > 0$
$$\Pr[X > t] \leq \frac{\mathrm{E}[X]}{t}$$
- **Chebyshev Inequality** For any random variable X and $t > 0$
$$\Pr[|X - \mathrm{E}[X]| > t] \leq \frac{\mathrm{Var}[X]}{t^2}$$

The Chernoff Bound

- Let $X_1, X_2 \ldots X_n$ be n independent Bernoulli random variables with $\Pr(X_i = 1) = p_i$. Let $X = \sum X_i$. Hence,

$$E[X] = E\left[\sum X_i\right] = \sum E[X_i] = \sum \Pr(X_i = 1) = \sum p_i = \mu(\text{say}).$$

Then the Chernoff Bound says for any $\epsilon > 0$

$$\begin{aligned} \Pr(X > (1+\epsilon)\mu) &\leq \left(\frac{e^{\epsilon}}{(1+\epsilon)^{\epsilon}}\right)^{\mu} \text{ and} \\ \Pr(X < (1-\epsilon)\mu) &\leq \left(\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right)^{\mu} \end{aligned}$$

When $0 < \epsilon < 1$ the above expression can be further simplified to

$$\begin{aligned} \Pr(X > (1+\epsilon)\mu) &\leq e^{\frac{-\mu\epsilon^2}{3}} \text{ and} \\ \Pr(X < (1-\epsilon)\mu) &\leq e^{\frac{-\mu\epsilon^2}{2}} \end{aligned}$$

Hence

$$\Pr(|X - \mu| > \epsilon\mu) \leq 2e^{\frac{-\mu\epsilon^2}{3}}$$

Universal Hash Family

A family of hash functions $\mathcal{H} = \{h \mid h : [N] - - > [M]\}$ is called a pairwise independent family of hash functions if for all $i \neq j \in [N]$ and any $k, l \in [M]$

$$\Pr_{h \leftarrow \mathcal{H}}\left[h(i) = k \wedge h(j) = l\right] = \frac{1}{M^2} \text{strongly universal hash family} \quad (1)$$

Hash functions are uniform over $[M]$,

$$\Pr_{h \leftarrow \mathcal{H}}\left[h(i) = k\right] = \frac{1}{M} \quad (2)$$

$$\Pr_{h \leftarrow \mathcal{H}}\left[h(i) = h(j)\right] = \frac{1}{M} \text{weakly universal hash family} \quad (3)$$

- Construction Let p be a prime. For any $a, b \in \mathbb{Z}_p = \{0, 1, 2, .., p-1\}$, define $h_{a,b} : \mathbb{Z}_p \rightarrow \mathbb{Z}_p$ by $h_{a,b}(x) = ax + b \bmod p$. Then the collection of functions $\mathcal{H} = \{h_{a,b} | a, b \in \mathbb{Z}_p\}$ is a pairwise independent hash family.

Counting Distinct Items

Algorithm 1 $[\mathbf{a}, \epsilon, \delta]$

$\epsilon' = \epsilon/2$

for $t = 1, \lceil(1+\epsilon')\rceil, \lceil(1+\epsilon')^2\rceil, \ldots \lceil(1+\epsilon')^{\log_{1+\epsilon'} n}\rceil$ **do**

$\delta' = \frac{\epsilon'\delta}{\log n}$ {Run in parallel}

b_t = ESTIMATE$(\mathbf{a}, t, \epsilon', \delta')$ {b_t is a boolean variable YES/NO}

end for

return the smallest value of t such that b_{t-1} =YES and b_t =NO, if no such t exists, return n

Counting Distinct Items

Algorithm 2 [ESTIMATE($\mathbf{a}, t, \epsilon', \delta'$)]

$count \leftarrow 0$
for $i = 1$ to $\frac{c}{\epsilon'^2} \log \frac{1}{\delta'}$ **do**
  Select a hash function h_i uniformly and randomly from a fully-independent hash family $\mathcal{H}$ {run in parallel}
  $b_t^i \leftarrow$ NO
  repeat
    Consider the current element in the stream **a**, say $a_l = (j, \nu)$
    if $h_i(j) == 1$ **then**
      $b_t^i \leftarrow$ YES, BREAK
    end if
  until a is exhausted
  if $b_t^i ==$ NO **then**
    $count = count + 1$
  end if
end for

Counting Distinct Items

Algorithm 3 [ESTIMATE($\mathbf{a}, t, \epsilon', \delta'$)]continued

if *count* $\geq \frac{1}{e}\frac{c}{\epsilon'^2} \log \frac{1}{\delta'}$ **then**
  return NO
else
  return YES
end if

- **Space Complexity:** $O(\frac{1}{\epsilon^3} \log n(\log \frac{1}{\delta} + \log \log n + \log \frac{1}{\epsilon}))$
- **Time Complexity:** $O(\frac{1}{\epsilon^3} \log n(\log \frac{1}{\delta} + \log \log n + \log \frac{1}{\epsilon}))$

Counting Distinct Items

- Lemma
  Consider the ith round of ESTIMATE($\mathbf{a}, t, \epsilon', \delta'$) for any $i \in [\frac{C}{\epsilon^2} \log \frac{1}{\delta'}]$
  - *If* $DE > (1+\epsilon')t$ *then* $\Pr\left[b_t^i == NO\right] \leq \frac{1}{e} - \frac{\epsilon}{2e}$.
  - *If* $DE < (1-\epsilon')t$ *then* $\Pr\left[b_t^i == NO\right] \geq \frac{1}{e} + \frac{\epsilon}{2e}$.

- Lemma
  - *If* $DE > (1+\epsilon')t$ *then* $\Pr\left[b_t == NO\right] \leq \frac{\delta'}{2}$.
  - *If* $DE < (1-\epsilon')t$ *then* $\Pr\left[b_t == YES\right] \leq \frac{\delta'}{2}$.

- Lemma
  - *If* $|DE - t| > \epsilon' t$ *then* $\Pr\left[ERROR\right] \leq \delta'$.

Counting Distinct Items

- Lemma
  For all t such that $|DE - t| > \epsilon' t$ $\Pr\big[ERROR\big] \leq \delta$.
- Theorem
  Algorithm 1 returns an estimate of DE within $(1 \pm \epsilon)$ *with probability* $\geq (1 - \delta)$.

AMS Sketch for Counting Distinct Element

- Uses pair-wise independent hash function
- Improved space and time complexity
- Worse approximation

Algorithm 4 AMS Counting Distinct Items

Initialize
$z \leftarrow 0$
End Initialize
Process($a_l = (j, \nu)$)
if $zeros(h(j)) > z$ **then**
  $z \leftarrow zeros(h(j))$
end if
End Process
Estimate
return $2^{z+\frac{1}{2}}$
End Estimate

AMS Sketch for Counting Distinct Element

- Define $X_j^r = 1$ if $zeros(h(j)) \geq r$ and 0 otherwise. Define $Y^r = \sum_j X_j^r$.
- Lemma
  - $\mathrm{E}\left[X_j^r\right] = \frac{1}{2^r}$
  - $\mathrm{E}\left[Y^r\right] = \frac{DE}{2^r}$
  - $\mathrm{Var}\left[Y^r\right] \leq \frac{1}{2^r}$
- Lemma
  - *Consider the largest level a such that* $2^{a+\frac{1}{2}} < \frac{DE}{3}$. $\Pr\left[z \leq a\right] < \frac{\sqrt{2}}{3}$.
  - *Consider the smallest level b such that* $2^{b+\frac{1}{2}} > 3DE$. $\Pr\left[z \geq b\right] < \frac{\sqrt{2}}{3}$.
  - $\Pr\left[\frac{DE}{3} < 2^{z+\frac{1}{2}} < 3DE\right] \geq 1 - \frac{2\sqrt{2}}{3}$.

AMS Sketch for Counting Distinct Element

- **Boosting the confidence** Median Trick.
  Keep $C \log \frac{1}{\delta}$ copies and return the median estimate

- Theorem
  There exists a randomized algorithm that returns an estimate of DE satisfying $\Pr\left[\frac{DE}{3} < 2^{z+\frac{1}{2}} < 3DE\right] \geq 1-\delta$ *using space* $O(\log \frac{1}{\delta} \log n)$