CMPSCI 611: Advanced Algorithms

Lecture 16: Chernoff and Balls & Bins

Andrew McGregor

Last Compiled: November 13, 2009

Probability Refresher

- For arbitrary events A and B,

$$\mathbb{P}[A \text{ and } B] = \mathbb{P}[A \text{ given } B]\,\mathbb{P}[B]$$

  and A and B are *independent* if $\mathbb{P}[A \text{ and } B] = \mathbb{P}[A]\,\mathbb{P}[B]$.
- Union Bound: $\mathbb{P}[A \text{ or } B] \leq \mathbb{P}[A] + \mathbb{P}[B]$

- Expectation: $\mathbb{E}[X] = \sum_r r\mathbb{P}[X = r]$
- Linearity of expectation: $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- Variance random variable: $\mathbb{V}[X] = \sigma_X^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$
- Linearity of variance if X and Y are independent:

$$\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$$

Examples of Random Variables

Example

Let X have the binomial distribution $Bin(n,p)$:

$$\mathbb{P}[X = i] = \binom{n}{i} p^i (1-p)^{n-i}$$

"How many heads do we see when we toss a coin with probability p of heads n times?" $\mathbb{E}[X] = np$ and $\mathbb{V}[X] = np(1-p)$.

Example

Let X have the binomial distribution $Geom(p)$:

$$\mathbb{P}[X = i] = (1-p)^{i-1} p$$

"How many times do we toss a coin with probability p of heads until we see a heads." $\mathbb{E}[X] = 1/p$, $\mathbb{V}[X] = (1-p)/p^2$.

Outline

Lazy Select

Chernoff Bounds

Balls and Bins (and Birthdays and Coupons!)

Lazy Select

S contains $n = 2k$ distinct values. We want to find k-th smallest value.

Algorithm

1. *Let R be a set of $n^{3/4}$ elements chosen uniformly at random with replacement from S.*
2. *Sort R and find a and b such that*

   $$\text{rank}_R(a) = kn^{-1/4} - \sqrt{n} \textit{ and } \text{rank}_R(b) = kn^{-1/4} + \sqrt{n}$$

   where $\text{rank}_X(x) = t$ *if x is the t-th smallest element in X.*
3. *Compute* $\text{rank}_S(a)$ *and* $\text{rank}_S(b)$*: Output FAIL if*

   $$k < \text{rank}_S(a) \quad \textit{or} \quad k > \text{rank}_S(b)$$

4. *Compute $P = \{i \in S : a \leq y \leq b\}$: Output FAIL if $|P| \geq 4n^{3/4}$*
5. *Return $(k - \text{rank}_S(a) + 1)$-th smallest elt of P by first sorting P.*

Lazy Select: Running Time

Theorem

Running time of Lazy Select is $O(n)$.

Theorem

With probability $1 - O(n^{-1/4})$, algorithm finds the median.

Proof.

- If we don't output FAIL, then we get the answer correct.
- Only three ways in which we fail and we'll show
  1. $\mathbb{P}[k < \text{rank}_S(a)] \leq O(n^{-1/4})$
  2. $\mathbb{P}[k > \text{rank}_S(b)] \leq O(n^{-1/4})$
  3. $\mathbb{P}\left[|P| \geq 4n^{3/4}\right] \leq O(n^{-1/4})$
- Result follows by the Union Bound

□

Lazy Select: Probability of Being Correct (2/3)

Claim

$\mathbb{P}[k < \text{rank}_S(a)] \leq O(n^{-1/4})$

Proof.

- Let u be the k-th smallest element in S
- Let $X_i = 1$ if i-th sample is $\leq u$ and $X_i = 0$ otherwise.

$$\mathbb{P}[X_i = 1] = k/n \text{ and } \mathbb{P}[X_i = 0] = 1 - k/n$$

- $X = \sum_{i \in [n^{3/4}]} X_i =$ number of elements in R that are at most u.
- $k < \text{rank}_S(a)$ implies $X < kn^{-1/4} - \sqrt{n}$
- X has binomial distribution:

$$\mathbb{E}[X] = kn^{-1/4} \text{ and } \mathbb{V}[X] = n^{3/4}(k/n)(1 - k/n) = n^{3/4}/4$$

- Apply Chebyshev bound: $\mathbb{P}\left[X < kn^{-1/4} - \sqrt{n}\right]$ is at most

$$\mathbb{P}\left[|X - \mathbb{E}[X]| > \sqrt{n}\right] \leq \mathbb{P}\left[|X - \mathbb{E}[X]| > 2n^{1/8}\sigma_X\right] = O(n^{-1/4})$$

Outline

Lazy Select

Chernoff Bounds

Balls and Bins (and Birthdays and Coupons!)

Chernoff Bound

Theorem

Let $X_1, \ldots, X_n$ be independent boolean random variables such that $\mathbb{P}[X_i = 1] = p_i$. Then, for $X = \sum_i X_i$, $\mu = \mathbb{E}[X]$, and $\delta > 0$,

$$\mathbb{P}[X > (1+\delta)\mu] < \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu}$$

Other versions: For $0 < \delta \leq 1$

$$\mathbb{P}[X \geq (1+\delta)\mu] \leq e^{-\delta^2 \mu/3}$$

$$\mathbb{P}[X \leq (1-\delta)\mu] \leq e^{-\delta^2 \mu/2}$$

Chernoff Bound: Proof of Upper Tail (1/2)

Proof.

- For any $t > 0$: $\mathbb{P}[X > (1+\delta)\mu] = \mathbb{P}\left[e^{tX} > e^{t(1+\delta)\mu}\right]$
- Apply Markov inequality:

$$\mathbb{P}\left[e^{tX} > e^{t(1+\delta)\mu}\right] \geq \mathbb{E}\left[e^{tX}\right]/e^{t(1+\delta)\mu}$$

- By independence:

$$\mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[e^{t\sum_i X_i}\right] = \mathbb{E}\left[\prod_i e^{tX_i}\right] = \prod_i \mathbb{E}\left[e^{tX_i}\right]$$

- We will prove $\prod_i \mathbb{E}\left[e^{tX_i}\right] \leq e^{(e^t-1)\mu}$ in a sec.
- For $t = \ln(1+\delta)$:

$$\mathbb{E}\left[e^{tX}\right]/e^{t(1+\delta)\mu} \leq e^{(e^t-1)\mu}/e^{t(1+\delta)\mu} = \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu}$$

Chernoff Bound: Proof of Upper Tail (2/2)

Lemma

$\prod_i \mathbb{E}\left[e^{tX_i}\right] \leq e^{(e^t-1)\mu}$

Proof.

- Using $1 + x \leq e^x$:

$$\mathbb{E}\left[e^{tX_i}\right] = p_i e^t + (1 - p_i) = 1 + p_i(e^t - 1) \leq \exp(p_i(e^t - 1))$$

- Using $\mu = \mathbb{E}\left[\sum_i X_i\right] = \sum_i p_i$:

$$\prod_i \exp(p_i(e^t - 1)) = \exp(\sum_i p_i(e^t - 1)) = \exp((e^t - 1)\mu)$$

$\square$

Example 1

- Your sports team wins each game (independently) with prob. $1/4$.
- What's probability that they win at least half of their n games.

Outline

Lazy Select

Chernoff Bounds

Balls and Bins (and Birthdays and Coupons!)

Balls and Bins

Throw m balls into n bins where each throw is independent.

- The maximum number of balls that fall into the same bin?
- How large must m be such that there exists a bin with at least two balls? (Birthday Paradox)
- How large must m be such that all bins get at least one ball? (Coupon Collecting)

Heaviest Bin (1/2)

Assume $m = n$. Let Y_i be number of balls that fall in i-th bin.

Lemma

Let $k \geq (3 \ln n)/\ln \ln n$. *Then* $\mathbb{P}[Y_i \geq k] \leq n^{-2}$

Proof.

- $\mathbb{P}[Y_i = j] = \binom{n}{j}(1/n)^j(1 - 1/n)^{n-j}$
- Using the bound $\binom{n}{j} \leq (ne/j)^j$:

$$\mathbb{P}[Y_i = j] = \binom{n}{j}(1/n)^j(1 - 1/n)^{n-j} \leq (e/j)^j$$

- By summing up a geometric series:

$$\mathbb{P}[Y_i \geq k] = \sum_{j \geq k}(e/j)^j \leq (e/k)^k \frac{1}{1 - e/k}$$

□

Heaviest Bin (2/2)

Assume $m = n$. Let Y_i be number of balls that fall in i-th bin.

Lemma

Let $k \geq (3 \ln n)/\ln\ln n$. *Then* $\mathbb{P}[Y_i \geq k] \leq n^{-2}$

Theorem

$\mathbb{P}[Y_i < k \text{ for all } i] \geq 1 - 1/n$.

Proof.

Use union bound:

$$\mathbb{P}[Y_i \geq k \text{ for some } i] \leq \sum_i \mathbb{P}[Y_i \geq k] \leq 1/n$$

□

Birthday Paradox

Lemma

$\mathbb{P}[\text{first } m \text{ balls fall in distinct bins}] \leq e^{-m(m-1)/(2n)}$.

Proof.

- Let A_i be event that the i-th ball lands in a bin not containing any of the first $i-1$ balls.
- $\mathbb{P}[\cap_{1 \leq i \leq m} A_i] = \mathbb{P}[A_1]\,\mathbb{P}[A_2|A_1] \ldots \mathbb{P}[A_m| \cap_{1 \leq i \leq m-1} A_i]$
- $\mathbb{P}[A_i| \cap_{1 \leq j \leq i-1} A_j] = 1 - (i-1)/n$
- Putting it together and using $\sum_{1 \leq i \leq a} i = (a+1)a/2$:

$$\mathbb{P}[\cap_{1 \leq i \leq m} A_i] = \prod_{1 \leq i \leq m} \left(1 - \frac{i-1}{n}\right) \leq e^{-m(m-1)/(2n)}$$

□

With $n = 365$ and $m = 29$, probability $< e^{-1}$. Tighter analysis possible.

For Next Time. . .

- Finish reading Chapter 6.