CMPSCI 711: More Advanced Algorithms

Section 1-1: Sampling

Andrew McGregor

Last Compiled: April 29, 2012

Concentration Bounds

Theorem (Markov)

Let X be a non-negative random variable with expectation μ. For $t > 0$,

$$\mathbb{P}[X \geq t\mu] \leq 1/t$$

Concentration Bounds

Theorem (Markov)

Let X be a non-negative random variable with expectation μ. For $t > 0$,

$$\mathbb{P}[X \geq t\mu] \leq 1/t$$

Theorem (Chebyshev)

Let X be a random variable with expectation μ. Then for $t > 0$,

$$\mathbb{P}[|X - \mu| \geq \delta\mu] \leq \frac{\mathbb{V}[X]}{(\delta\mu)^2}$$

Concentration Bounds

Theorem (Markov)

Let X be a non-negative random variable with expectation μ. For $t > 0$,

$$\mathbb{P}\left[X \geq t\mu\right] \leq 1/t$$

Theorem (Chebyshev)

Let X be a random variable with expectation μ. Then for $t > 0$,

$$\mathbb{P}\left[|X - \mu| \geq \delta\mu\right] \leq \frac{\mathbb{V}\left[X\right]}{(\delta\mu)^2}$$

Theorem (Chernoff)

Let $X_1, \ldots, X_t$ be i.i.d. random variables with range $[0, 1]$ and expectation μ. Then, if $X = \frac{1}{t}\sum_i X_i$ and $1 > \delta > 0$,

$$\mathbb{P}\left[|X - \mu| \geq \delta\mu\right] \leq 2\exp\left(\frac{-\mu t\delta^2}{3}\right)$$

Chernoff Corollary

Corollary (Chernoff)

Let $X_1, \ldots, X_t$ be i.i.d. random variables with range $[0, c]$ and expectation μ. Then, if $X = \frac{1}{t}\sum_i X_i$ and $1 > \delta > 0$,

$$\mathbb{P}\left[|X - \mu| \geq \delta\mu\right] \leq 2\exp\left(\frac{-\mu t \delta^2}{3c}\right)$$

Chernoff Corollary

Corollary (Chernoff)

Let $X_1, \ldots, X_t$ be i.i.d. random variables with range $[0, c]$ and expectation μ. Then, if $X = \frac{1}{t}\sum_i X_i$ and $1 > \delta > 0$,

$$\mathbb{P}\left[|X - \mu| \geq \delta\mu\right] \leq 2\exp\left(\frac{-\mu t \delta^2}{3c}\right)$$

- For $i \in [t]$, let $Y_i = X_i/c$. Note that Y_i has expectation μ/c.

Chernoff Corollary

Corollary (Chernoff)

Let $X_1, \ldots, X_t$ be i.i.d. random variables with range $[0, c]$ and expectation μ. Then, if $X = \frac{1}{t}\sum_i X_i$ and $1 > \delta > 0$,

$$\mathbb{P}\left[|X - \mu| \geq \delta\mu\right] \leq 2\exp\left(\frac{-\mu t \delta^2}{3c}\right)$$

- For $i \in [t]$, let $Y_i = X_i/c$. Note that Y_i has expectation μ/c.
- Then,

$$\mathbb{P}\left[|X - \mu| \geq \delta\mu\right] = \mathbb{P}\left[|Y - \mu/c| \geq \delta\mu/c\right] \leq 2\exp\left(\frac{-\mu t \delta^2}{3c}\right)$$

Outline

Warm-Up: Median Approximation

Reservoir Sampling

AMS Sampling

Today's Set-Up

- *Stream:* m elements from universe $[n] = \{1, 2, \ldots, n\}$, e.g.,

$$\langle x_1, x_2, \ldots, x_m \rangle \quad = \quad \langle 3, 5, 103, 17, 5, 4, \ldots, 1 \rangle$$

- Let f_i be the frequency of i in the stream. The "frequency vector" is

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

Outline

Approximate Median

- Let $S = \{x_1, x_2, \ldots, x_m\}$ and define $\text{rank}(y) = |\{x \in S : x \leq y\}|$. For simplicity suppose elements in S are distinct.

Approximate Median

- Let $S = \{x_1, x_2, \ldots, x_m\}$ and define $\text{rank}(y) = |\{x \in S : x \leq y\}|$. For simplicity suppose elements in S are distinct.
- *Problem:* Find an ϵ-approximate median of S, i.e., y such that

$$m/2 - \epsilon m < \text{rank}(y) < m/2 + \epsilon m$$

Approximate Median

- Let $S = \{x_1, x_2, \ldots, x_m\}$ and define $\mathrm{rank}(y) = |\{x \in S : x \leq y\}|$. For simplicity suppose elements in S are distinct.
- *Problem:* Find an ϵ-approximate median of S, i.e., y such that

$$m/2 - \epsilon m < \mathrm{rank}(y) < m/2 + \epsilon m$$

- *Algorithm:* Sample t values from S (with replacement) and return the median of the sampled values.

Approximate Median

- Let $S = \{x_1, x_2, \ldots, x_m\}$ and define $\text{rank}(y) = |\{x \in S : x \leq y\}|$. For simplicity suppose elements in S are distinct.
- *Problem:* Find an ϵ-approximate median of S, i.e., y such that

$$m/2 - \epsilon m < \text{rank}(y) < m/2 + \epsilon m$$

- *Algorithm:* Sample t values from S (with replacement) and return the median of the sampled values.
- *Lemma:* If $t = 7\epsilon^{-2}\log(2\delta^{-1})$ then the algorithm returns an ϵ-median with probability $1 - \delta$.

Approximate Median

- Let $S = \{x_1, x_2, \ldots, x_m\}$ and define $\text{rank}(y) = |\{x \in S : x \leq y\}|$. For simplicity suppose elements in S are distinct.
- *Problem:* Find an ϵ-approximate median of S, i.e., y such that

$$m/2 - \epsilon m < \text{rank}(y) < m/2 + \epsilon m$$

- *Algorithm:* Sample t values from S (with replacement) and return the median of the sampled values.
- *Lemma:* If $t = 7\epsilon^{-2}\log(2\delta^{-1})$ then the algorithm returns an ϵ-median with probability $1 - \delta$.
- We'll later present an algorithm with smaller space.

Median Analysis

- Partition S into 3 groups:

$$
\begin{aligned}
S_L &= \{x \in S : \mathrm{rank}(x) \leq m/2 - \epsilon m\} \\
S_M &= \{x \in S : m/2 - \epsilon m < \mathrm{rank}(x) < m/2 + \epsilon m\} \\
S_U &= \{x \in S : \mathrm{rank}(x) \geq m/2 + \epsilon m\}
\end{aligned}
$$

Median Analysis

- Partition S into 3 groups:

$$\begin{aligned} S_L &= \{x \in S : \text{rank}(x) \leq m/2 - \epsilon m\} \\ S_M &= \{x \in S : m/2 - \epsilon m < \text{rank}(x) < m/2 + \epsilon m\} \\ S_U &= \{x \in S : \text{rank}(x) \geq m/2 + \epsilon m\} \end{aligned}$$

- If less than $t/2$ elements from both S_L and S_U are present in sample then the median of the sample is an ϵ-approximate median.

Median Analysis

- Partition S into 3 groups:

$$
\begin{aligned}
S_L &= \{x \in S : \text{rank}(x) \leq m/2 - \epsilon m\} \\
S_M &= \{x \in S : m/2 - \epsilon m < \text{rank}(x) < m/2 + \epsilon m\} \\
S_U &= \{x \in S : \text{rank}(x) \geq m/2 + \epsilon m\}
\end{aligned}
$$

- If less than $t/2$ elements from both S_L and S_U are present in sample then the median of the sample is an ϵ-approximate median.
- Let $X_i = 1$ if i-th sample if in S_L and 0 otherwise. Let $X = \sum_i X_i$. Assume $\epsilon < 1/10$. By Chernoff bound, if $t > 7\epsilon^{-2} \log(2\delta^{-1})$

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

Median Analysis

- Partition S into 3 groups:

$$\begin{aligned} S_L &= \{x \in S : \text{rank}(x) \leq m/2 - \epsilon m\} \\ S_M &= \{x \in S : m/2 - \epsilon m < \text{rank}(x) < m/2 + \epsilon m\} \\ S_U &= \{x \in S : \text{rank}(x) \geq m/2 + \epsilon m\} \end{aligned}$$

- If less than $t/2$ elements from both S_L and S_U are present in sample then the median of the sample is an ϵ-approximate median.
- Let $X_i = 1$ if i-th sample if in S_L and 0 otherwise. Let $X = \sum_i X_i$. Assume $\epsilon < 1/10$. By Chernoff bound, if $t > 7\epsilon^{-2}\log(2\delta^{-1})$

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

- Similarly, there are $\geq t/2$ elements from S_U with probability $\leq \delta/2$.

Median Analysis

- Partition S into 3 groups:

$$\begin{aligned} S_L &= \{x \in S : \text{rank}(x) \le m/2 - \epsilon m\} \\ S_M &= \{x \in S : m/2 - \epsilon m < \text{rank}(x) < m/2 + \epsilon m\} \\ S_U &= \{x \in S : \text{rank}(x) \ge m/2 + \epsilon m\} \end{aligned}$$

- If less than $t/2$ elements from both S_L and S_U are present in sample then the median of the sample is an ϵ-approximate median.
- Let $X_i = 1$ if i-th sample if in S_L and 0 otherwise. Let $X = \sum_i X_i$. Assume $\epsilon < 1/10$. By Chernoff bound, if $t > 7\epsilon^{-2}\log(2\delta^{-1})$

$$\mathbb{P}[X \ge t/2] \le \mathbb{P}[X \ge (1+\epsilon)\mathbb{E}[X]] \le e^{-\epsilon^2(1/2-\epsilon)t/3} \le \delta/2$$

- Similarly, there are $\ge t/2$ elements from S_U with probability $\le \delta/2$.
- By the union bound, with probability at least $1-\delta$ there are less than $t/2$ elements chosen from both S_L and S_U.

Outline

Reservoir Sampling

- *Problem:* Find uniform sample s from a stream if we don't know m

Reservoir Sampling

- *Problem:* Find uniform sample s from a stream if we don't know m
- *Algorithm:*
  - Initially $s = x_1$
  - On seeing the t-th element, $s \leftarrow x_t$ with probability $1/t$

Reservoir Sampling

- *Problem:* Find uniform sample s from a stream if we don't know m
- *Algorithm:*
  - Initially $s = x_1$
  - On seeing the t-th element, $s \leftarrow x_t$ with probability $1/t$
- *Analysis:*
  - What's the probability that $s = x_i$ at some time $t \geq i$?

Reservoir Sampling

- *Problem:* Find uniform sample s from a stream if we don't know m
- *Algorithm:*
  - Initially $s = x_1$
  - On seeing the t-th element, $s \leftarrow x_t$ with probability $1/t$
- *Analysis:*
  - What's the probability that $s = x_i$ at some time $t \geq i$?

$$\mathbb{P}\left[s = x_i\right] = \frac{1}{i} \times \left(1 - \frac{1}{i+1}\right) \times \ldots \times \left(1 - \frac{1}{t}\right) = \frac{1}{t}$$

Reservoir Sampling

- *Problem:* Find uniform sample s from a stream if we don't know m
- *Algorithm:*
  - Initially $s = x_1$
  - On seeing the t-th element, $s \leftarrow x_t$ with probability $1/t$
- *Analysis:*
  - What's the probability that $s = x_i$ at some time $t \geq i$?

$$\mathbb{P}\left[s = x_i\right] = \frac{1}{i} \times \left(1 - \frac{1}{i+1}\right) \times \ldots \times \left(1 - \frac{1}{t}\right) = \frac{1}{t}$$

  - To get k samples we use $O(k \log n)$ bits of space.

Outline

AMS Sampling

- *Problem:* Estimate $\sum_{i\in[n]} g(f_i)$ for any function g with $g(0) = 0$

AMS Sampling

- *Problem:* Estimate $\sum_{i\in[n]} g(f_i)$ for any function g with $g(0)=0$
- *Basic Estimator:* Sample x_J where $J \in_R [m]$ and compute

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

AMS Sampling

- *Problem:* Estimate $\sum_{i \in [n]} g(f_i)$ for any function g with $g(0) = 0$
- *Basic Estimator:* Sample x_J where $J \in_R [m]$ and compute

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

Output $X = m(g(r) - g(r-1))$

AMS Sampling

- *Problem:* Estimate $\sum_{i\in[n]} g(f_i)$ for any function g with $g(0)=0$
- *Basic Estimator:* Sample x_J where $J \in_R [m]$ and compute

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

  Output $X = m(g(r) - g(r-1))$
- *Correct Expectation:*

$$\mathbb{E}[X]$$

AMS Sampling

- *Problem:* Estimate $\sum_{i\in[n]} g(f_i)$ for any function g with $g(0)=0$
- *Basic Estimator:* Sample x_J where $J \in_R [m]$ and compute

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

  Output $X = m(g(r) - g(r-1))$
- *Correct Expectation:*

$$\mathbb{E}[X] \quad = \quad \sum_i \mathbb{P}[x_J = i]\,\mathbb{E}[X | x_J = i]$$

AMS Sampling

- *Problem:* Estimate $\sum_{i\in[n]} g(f_i)$ for any function g with $g(0)=0$
- *Basic Estimator:* Sample x_J where $J \in_R [m]$ and compute

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

  Output $X = m(g(r) - g(r-1))$
- *Correct Expectation:*

$$\begin{aligned}
\mathbb{E}[X] &= \sum_i \mathbb{P}[x_J = i]\,\mathbb{E}[X|x_J = i] \\
&= \sum_i \frac{f_i}{m}\left(\sum_{r=1}^{f_i} \frac{m(g(r)-g(r-1))}{f_i}\right)
\end{aligned}$$

AMS Sampling

- *Problem:* Estimate $\sum_{i\in[n]} g(f_i)$ for any function g with $g(0)=0$
- *Basic Estimator:* Sample x_J where $J \in_R [m]$ and compute

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

  Output $X = m(g(r) - g(r-1))$
- *Correct Expectation:*

$$\begin{aligned}
\mathbb{E}[X] &= \sum_i \mathbb{P}[x_J = i]\,\mathbb{E}[X|x_J = i] \\
&= \sum_i \frac{f_i}{m}\left(\sum_{r=1}^{f_i} \frac{m(g(r)-g(r-1))}{f_i}\right) \\
&= \sum_i g(f_i)
\end{aligned}$$

AMS Sampling

- *Problem:* Estimate $\sum_{i \in [n]} g(f_i)$ for any function g with $g(0) = 0$
- *Basic Estimator:* Sample x_J where $J \in_R [m]$ and compute

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

  Output $X = m(g(r) - g(r-1))$
- *Correct Expectation:*

$$\begin{aligned}
\mathbb{E}[X] &= \sum_i \mathbb{P}[x_J = i] \, \mathbb{E}[X | x_J = i] \\
&= \sum_i \frac{f_i}{m} \left(\sum_{r=1}^{f_i} \frac{m(g(r) - g(r-1))}{f_i} \right) \\
&= \sum_i g(f_i)
\end{aligned}$$

- *For high confidence:* Compute t estimators in parallel and average.

Example: Frequency Moments (a)

- *Frequency Moments:* Define $F_k = \sum_i f_i^k$ for $k \in \{1, 2, 3, \ldots\}$

Example: Frequency Moments (a)

- *Frequency Moments:* Define $F_k = \sum_i f_i^k$ for $k \in \{1, 2, 3, \ldots\}$
- Use AMS estimator with $X = m(r^k - (r-1)^k)$.

$$\mathbb{E}[X] = F_k$$

Example: Frequency Moments (a)

- *Frequency Moments:* Define $F_k = \sum_i f_i^k$ for $k \in \{1, 2, 3, \ldots\}$
- Use AMS estimator with $X = m(r^k - (r-1)^k)$.

$$\mathbb{E}[X] = F_k$$

- *Exercise:* $0 \leq X \leq mk{f_*}^{k-1}$ where $f_* = \max_i f_i$.

Example: Frequency Moments (a)

- *Frequency Moments:* Define $F_k = \sum_i f_i^k$ for $k \in \{1, 2, 3, \ldots\}$
- Use AMS estimator with $X = m(r^k - (r-1)^k)$.

$$\mathbb{E}[X] = F_k$$

- *Exercise:* $0 \le X \le mk{f_*}^{k-1}$ where $f_* = \max_i f_i$.
- Repeat t times and let $\hat{X}$ be the average value. By Chernoff,

$$\mathbb{P}\left[|\hat{X} - F_k| \ge \epsilon F_k\right] \le 2\exp\left(-\frac{tF_k\epsilon^2}{3mk{f_*}^{k-1}}\right)$$

Example: Frequency Moments (a)

- *Frequency Moments:* Define $F_k = \sum_i f_i^k$ for $k \in \{1, 2, 3, \ldots\}$
- Use AMS estimator with $X = m(r^k - (r-1)^k)$.

$$\mathbb{E}[X] = F_k$$

- *Exercise:* $0 \leq X \leq mk{f_*}^{k-1}$ where $f_* = \max_i f_i$.
- Repeat t times and let $\hat{X}$ be the average value. By Chernoff,

$$\mathbb{P}\left[|\hat{X} - F_k| \geq \epsilon F_k\right] \leq 2\exp\left(-\frac{tF_k\epsilon^2}{3mk{f_*}^{k-1}}\right)$$

- Hence, taking $t = \frac{3mk{f_*}^{k-1}\log(2\delta^{-1})}{\epsilon^2 F_k}$ ensures $\mathbb{P}\left[|\hat{X} - F_k| \geq \epsilon F_k\right] \leq \delta$.

Example: Frequency Moments (a)

- *Frequency Moments:* Define $F_k = \sum_i f_i^k$ for $k \in \{1, 2, 3, \ldots\}$
- Use AMS estimator with $X = m(r^k - (r-1)^k)$.

$$\mathbb{E}[X] = F_k$$

- *Exercise:* $0 \le X \le mk{f_*}^{k-1}$ where $f_* = \max_i f_i$.
- Repeat t times and let $\hat{X}$ be the average value. By Chernoff,

$$\mathbb{P}\left[|\hat{X} - F_k| \ge \epsilon F_k\right] \le 2\exp\left(-\frac{tF_k\epsilon^2}{3mk{f_*}^{k-1}}\right)$$

- Hence, taking $t = \frac{3mk{f_*}^{k-1}\log(2\delta^{-1})}{\epsilon^2 F_k}$ ensures $\mathbb{P}\left[|\hat{X} - F_k| \ge \epsilon F_k\right] \le \delta$.
- *Lemma:* $m{f_*}^{k-1}/F_k \le n^{1-1/k}$.

Example: Frequency Moments (a)

- *Frequency Moments:* Define $F_k = \sum_i f_i^k$ for $k \in \{1, 2, 3, \ldots\}$
- Use AMS estimator with $X = m(r^k - (r-1)^k)$.

$$\mathbb{E}[X] = F_k$$

- *Exercise:* $0 \le X \le mk{f_*}^{k-1}$ where $f_* = \max_i f_i$.
- Repeat t times and let $\hat{X}$ be the average value. By Chernoff,

$$\mathbb{P}\left[|\hat{X} - F_k| \ge \epsilon F_k\right] \le 2\exp\left(-\frac{tF_k\epsilon^2}{3mk{f_*}^{k-1}}\right)$$

- Hence, taking $t = \frac{3mk{f_*}^{k-1}\log(2\delta^{-1})}{\epsilon^2 F_k}$ ensures $\mathbb{P}\left[|\hat{X} - F_k| \ge \epsilon F_k\right] \le \delta$.
- *Lemma:* $m{f_*}^{k-1}/F_k \le n^{1-1/k}$.
- *Thm:* In $O(kn^{1-1/k}\epsilon^{-2}\log\delta^{-1}\log(nm))$ space we find an (ϵ, δ) approximation for F_k.

Example: Frequency Moments (b)

Lemma

$m{f_*}^{k-1}/F_k \leq n^{1-1/k}$.

Example: Frequency Moments (b)

Lemma

${mf_*}^{k-1}/F_k \leq n^{1-1/k}$.

Proof.

Example: Frequency Moments (b)

Lemma
$m{f_*}^{k-1}/F_k \le n^{1-1/k}$.

Proof.

- *Exercise:* $F_k \ge n(m/n)^k$. (Hint: Use convexity of $g(x) = x^k$.)

Example: Frequency Moments (b)

Lemma

${mf_*}^{k-1}/F_k \le n^{1-1/k}$.

Proof.

- *Exercise:* $F_k \ge n(m/n)^k$. (Hint: Use convexity of $g(x) = x^k$.)
- Case 1: Suppose ${f_*}^k \le n(m/n)^k$. Then,

$$\frac{m{f_*}^{k-1}}{F_k} \le \frac{mn^{1-1/k}(m/n)^{k-1}}{n(m/n)^k} = n^{1-1/k}$$

Example: Frequency Moments (b)

Lemma

${mf_*}^{k-1}/F_k \le n^{1-1/k}$.

Proof.

- *Exercise:* $F_k \ge n(m/n)^k$. (Hint: Use convexity of $g(x) = x^k$.)
- Case 1: Suppose ${f_*}^k \le n(m/n)^k$. Then,

$$\frac{m{f_*}^{k-1}}{F_k} \le \frac{mn^{1-1/k}(m/n)^{k-1}}{n(m/n)^k} = n^{1-1/k}$$

- Case 2: Suppose ${f_*}^k \ge n(m/n)^k$. Then,

$$\frac{m{f_*}^{k-1}}{F_k} \le \frac{m{f_*}^{k-1}}{{f_*}^k} = \frac{m}{f_*} \le \frac{m}{n^{1/k}(m/n)} = n^{1-1/k}$$

$\square$