Concentration Bounds

Theorem (Markov)

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

$$\Pr[X \geq t\mu] \leq \frac{1}{t}$$

Theorem (Chebyshev)

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

$$\Pr[|X - \mu| \geq \delta \mu] \leq \frac{\text{Var}[X]}{(\delta \mu)^2}$$

Theorem (Chernoff)

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

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

Corollary (Chernoff)

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

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

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

$$P[|X - \mu| \geq \delta \mu] = P[|Y - \mu/c| \geq \delta \mu/c] \leq 2 \exp \left( \frac{-\mu t \delta^2}{3c} \right)$$
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.,
  \[
  \langle x_1, x_2, \ldots, x_m \rangle = \langle 3, 5, 103, 17, 5, 4, \ldots, 1 \rangle
  \]
  
- Let $f_i$ be the frequency of $i$ in the stream. The “frequency vector” is
  \[
  f = (f_1, \ldots, f_n)
  \]
Outline

Warm-Up: Median Approximation

Reservoir Sampling

AMS Sampling
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 $\epsilon$-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 $\epsilon$-median with probability $1 - \delta$.

- We’ll later present an algorithm with smaller space.
Median Analysis

- Partition $S$ into 3 groups:
  \[ S_L = \{ x \in S : \text{rank}(x) \leq m/2 - \epsilon m \} \]
  \[ S_M = \{ x \in S : m/2 - \epsilon m < \text{rank}(x) < m/2 + \epsilon m \} \]
  \[ S_U = \{ x \in S : \text{rank}(x) \geq m/2 + \epsilon m \} \]

- 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 $\epsilon$-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 $\leq \delta/2$.

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

8/14
Outline

Warm-Up: Median Approximation

Reservoir Sampling

AMS Sampling
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 $t \geq i$?
    \[ P[s = x_i] = \frac{1}{i} \times \left(1 - \frac{1}{i+1}\right) \times \ldots \times \left(1 - \frac{1}{t}\right) = \frac{1}{t} \]
  - To get $k$ samples we use $O(k \log n)$ bits of space.
Outline

Warm-Up: Median Approximation

Reservoir Sampling

AMS Sampling
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 $J \in \mathbb{R} [m]$ and compute

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

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

- **Correct Expectation:**

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

- **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 $k \in \{1, 2, 3, \ldots\}$
- Use AMS estimator with $X = m(r^k - (r - 1)^k)$.

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

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

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

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

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

- **Thm:** In $O(k n^{1-1/k} \epsilon^{-2} \log \delta^{-1} \log(n m))$ space we find an ($\epsilon, \delta$) approximation for $F_k$. 

13/14
Example: Frequency Moments (b)

Lemma
\[ mf_*^{k-1} / F_k \leq n^{1-1/k}. \]

Proof.

- **Exercise:** \( F_k \geq n(m/n)^k \). (Hint: Use convexity of \( g(x) = x^k \).)
- **Case 1:** Suppose \( f_*^k \leq n(m/n)^k \). Then,
  \[
  \frac{mf_*^{k-1}}{F_k} \leq \frac{mn^{1-1/k}(m/n)^{k-1}}{n(m/n)^k} = n^{1-1/k}
  \]
- **Case 2:** Suppose \( f_*^k \geq n(m/n)^k \). Then,
  \[
  \frac{mf_*^{k-1}}{F_k} \leq \frac{mf_*^{k-1}}{f_*^k} = \frac{m}{f_*} \leq \frac{m}{n^{1/k}(m/n)} = n^{1-1/k}
  \]