

CSCI B609: “Foundations of Data Science”

Lecture 19: L_0 -sampling, L_1 -sparse recovery, Count Sketch

Slides at <http://grigory.us/data-science-class.html>

Grigory Yaroslavtsev

<http://grigory.us>

Data Streams

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

$$\langle x_1, x_2, \dots, x_m \rangle = \langle 5, 8, 1, 1, 1, 4, 3, 5, \dots, 10 \rangle$$

- f_i = frequency of i in the stream = # of occurrences of value i

$$f = \langle f_1, \dots, f_n \rangle$$

Frequency Moments

- Define $F_k = \sum_i f_i^k$ for $k \in \{0,1,2, \dots\}$
 - $F_0 = \#$ number of distinct elements
 - $F_1 = \#$ elements
 - $F_2 =$ “Gini index”, “surprise index”

Proof of Lemma

- Let $j = \lceil \log \widetilde{F}_0 \rceil$ and note that $2F_0 < 2^j < 12 F_0$
- For any i , $\Pr[h_j(i) = 0] = \frac{1}{2^j}$
- Probability there exists a unique i such that $h_j(i) = 0$,

$$\begin{aligned} & \sum_i \Pr[h_j(i) = 0 \text{ and } \forall k \neq i, h_j(k) \neq 0] \\ &= \sum_i \Pr[h_j(i) = 0] \Pr[\forall k \neq i, h_j(k) \neq 0 \mid h_j(i) = 0] \\ &\geq \sum_i \Pr[h_j(i) = 0] \left(1 - \sum_{k \neq i} \Pr[h_j(k) = 0 \mid h_j(i) = 0] \right) \\ &= \sum_i \Pr[h_j(i) = 0] \left(1 - \sum_{k \neq i} \Pr[h_j(k) = 0] \right) \geq \sum_i \frac{1}{2^j} \left(1 - \frac{F_0}{2^j} \right) \geq \frac{1}{24} \end{aligned}$$

- Holds even if h_j are only 2-wise independent

Sparse Recovery

- **Goal:** Find g such that $\|f - g\|_1$ is minimized among g 's with at most k non-zero entries.
- **Definition:** $Err^k(f) = \min_{g: \|g\|_0 \leq k} \|f - g\|_1$
- **Exercise:** $Err^k(f) = \sum_{i \notin S} |f_i|$ where S are indices of k largest f_i
- Using $O(\epsilon^{-1} k \log n)$ space we can find \tilde{g} such that $\|\tilde{g}\|_0 \leq k$ and
$$\|\tilde{g} - f\|_1 \leq (1 + \epsilon) Err^k(f)$$

Count-Min Revisited

- Use Count-Min with $d = O(\log n)$, $w = 4k/\epsilon$
- For $i \in [n]$, let $\tilde{f}_i = c_{j, h_j(i)}$ for some row $j \in [d]$
- Let $S = \{i_1, \dots, i_k\}$ be the indices with max. frequencies. Let A_i be the event there doesn't exist $k \in S/i$ with $h_j(i) = h_j(k)$
- Then for $i \in [n]$:

$$\begin{aligned}
 & \Pr \left[|f_i - \tilde{f}_i| \geq \frac{\epsilon \text{Err}^k(f)}{k} \right] = \\
 & \Pr[\text{not } A_i] \times \Pr \left[|f_i - \tilde{f}_i| \geq \frac{\epsilon \text{Err}^k(f)}{k} \mid \text{not } A_i \right] + \\
 & \Pr[A_i] \times \Pr \left[|f_i - \tilde{f}_i| \geq \frac{\epsilon \text{Err}^k(f)}{k} \mid A_i \right] \\
 & \leq \Pr[\text{not } A_i] + \Pr \left[|f_i - \tilde{f}_i| \geq \frac{\epsilon \text{Err}^k(f)}{k} \mid A_i \right] \leq \frac{k}{w} + \frac{1}{4} \leq \frac{1}{2}
 \end{aligned}$$

- Because $d = O(\log n)$ w.h.p. all f_i 's approx. up to $\frac{\epsilon \text{Err}^k(f)}{k}$

Sparse Recovery Algorithm

- Use Count-Min with $d = O(\log n)$, $w = 4k/\epsilon$
- Let $f' = (\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_n)$ be frequency estimates:

$$|f_i - \tilde{f}_i| \leq \frac{\epsilon \text{Err}^k(f)}{k}$$

- Let \tilde{g} be f' with all but the k -th largest entries replaced by 0.
- **Lemma:** $\|\tilde{g} - f\|_1 \leq (1 + 3\epsilon) \text{Err}^k(f)$

$$\| \tilde{g} - f \|_1 \leq (1 + 3 \epsilon) Err^k(f)$$

- Let $S, T \subseteq [n]$ be indices corresponding to k largest values of f and f' .
- For a vector $x \in \mathbb{R}^n$ and $I \subseteq [n]$ denote as x_I the vector formed by zeroing out all entries of x except for those in I .

$$\begin{aligned}
 \|f - f'_T\|_1 &\leq \|f - f_T\|_1 + \|f_T - f'_T\|_1 \\
 &= \|f\|_1 - \|f_T\|_1 + \|f_T - f'_T\|_1 \\
 &= \|f\|_1 - \|f'_T\|_1 + (\|f'_T\|_1 - \|f_T\|_1) + \|f_T - f'_T\|_1 \\
 &\leq \|f\|_1 - \|f'_T\|_1 + 2 \|f_T - f'_T\|_1 \\
 &\leq \|f\|_1 - \|f'_S\|_1 + 2 \|f_T - f'_T\|_1 \\
 &\leq \|f\|_1 - \|f_S\|_1 + (\|f_S\|_1 - \|f'_S\|_1) + 2 \|f_T - f'_T\|_1 \\
 &\leq \|f - f_S\|_1 + \|f_S - f'_S\|_1 + 2 \|f_T - f'_T\|_1 \\
 &\leq Err^k(f) + k \epsilon \frac{Err^k(f)}{k} + 2k \epsilon \frac{Err^k(f)}{k} \\
 &\leq (1 + 3 \epsilon) Err^k(f)
 \end{aligned}$$

Count Sketch [Charikar, Chen, Farach-Colton]

- In addition to $H_i: [n] \rightarrow [w]$ use random signs $r_i[n] \rightarrow \{-1, 1\}$

$$c_{i,j} = \sum_{x: H_i(x)=j} r_i(x) f_x$$

- Estimate:

$$\hat{f}_x = \text{median}(r_1(x)c_{1,H_1(x)}, \dots, r_d(x)c_{d,H_d(x)})$$

- Parameters: $d = O\left(\log \frac{1}{\delta}\right)$, $w = \frac{3}{\epsilon^2}$

$$\Pr[|\hat{f}_x - f_x| + \epsilon \|f\|_2 \geq 1 - \delta]$$

- **Lemma:** $E[r_i(x)c_{i,H_i(x)}] = f_x$

- **Lemma:** $\text{Var}[r_i(x)c_{i,H_i(x)}] \leq \frac{F_2}{w}$

- By Chebyshev: $\Pr[|r_i(x)c_{i,H_i(x)} - f_x| \geq \epsilon \sqrt{F_2}] \leq 1/3$

- By Chernoff with $d = O\left(\log \frac{1}{\delta}\right)$ error prob. $1 - \delta$.

Count Sketch Analysis

- Fix i and x . Let $X_y = I[H(x) = H(y)]$:

$$r(x)C_{H(x)} = \sum_y r(x)r(y)f_y X_y$$

- Lemma:** $E[r_i(x)c_{i,H_i(x)}] = f_x$

$$E[r(x)C_{H(x)}] = E[f_x + \sum_{y \neq x} r(x)r(y)f(y)X_y] = f_x$$

- Lemma:** $\text{Var}[r_i(x)c_{i,H_i(x)}] \leq \frac{F_2}{w}$

$$\begin{aligned} \text{Var}[r(x)C_{H(x)}] &\leq E[(\sum_y r(x)r(y)f_y X_y)^2] \\ &= E[\sum_y f_y^2 X_y^2 + (\sum_{y \neq z} r(y)r(z)f_y f_z X_y X_z)] \\ &= F_2/w \end{aligned}$$