Lecture Scribe October 7

Prathamesh Swar ps165
Tanishka Sonar ts158

1 Introduction

In streaming applications, storing all elements is infeasible. A stream is a sequence of updates $(i_1, \Delta_1), (i_2, \Delta_2), \ldots$, where i_j is an item and Δ_j is an increment (possibly negative). We want an approximate count c_i for each item i using sublinear memory.

2 Count Sketch

2.1 Motivation

Count Sketch handles negative increments and provides an unbiased estimator using a **sign hash function**.

2.2 Data Structure

- d hash functions $h_1, \ldots, h_d : [1..N] \to [1..R]$
- d sign functions $g_1, \ldots, g_d : [1..N] \to \{-1, 1\}$
- $d \times R$ array S initialized to 0

2.3 Operations

Update: For item i with increment Δ_i:

$$S[j, h_j(i)] \leftarrow S[j, h_j(i)] + \Delta_i \cdot g_j(i), \quad j = 1 \ldots d$$

Query: Estimate count:

$$\hat{c}_i = \text{median}_{j=1}^{d} \left[S[j, h_j(i)] \cdot g_j(i) \right]$$

2.4 Derivation of Estimate

For a single row j:

$$\hat{c}_i^{(j)} = g_j(i) \sum_{k=1}^{N} g_j(k) c_k \mathbf{1}_{\{h_j(k)=h_j(i)\}}$$

Split summation:

$$\hat{c}_i^{(j)} = g_j(i)\big[g_j(i)c_i\mathbf{1}_{\{h_j(i)=h_j(i)\}} + \sum_{k\neq i} g_j(k)c_k\mathbf{1}_{\{h_j(k)=h_j(i)\}}\big]$$

Since $g_j(i)^2 = 1$ and $\mathbf{1}_{\{h_j(i)=h_j(i)\}} = 1$:

$$\hat{c}_i^{(j)} = c_i + \sum_{k\neq i} g_j(i)g_j(k)c_k\mathbf{1}_{\{h_j(k)=h_j(i)\}}$$

2.4.1 Expectation

$$\mathbb{E}[\hat{c}_i^{(j)}] = c_i + \sum_{k\neq i} c_k \cdot \mathbb{E}[g_j(i)g_j(k)] \cdot \mathbb{E}[\mathbf{1}_{\{h_j(k)=h_j(i)\}}]$$

- $g_j(i)$ and $g_j(k)$ independent: $\mathbb{E}[g_j(i)g_j(k)] = 0$ - Uniform hash: $\mathbb{E}[\mathbf{1}_{\{h_j(k)=h_j(i)\}}] = 1/R$
Hence:

$$\mathbb{E}[\hat{c}_i^{(j)}] = c_i$$

Unbiased estimator.

2.4.2 Variance

$$\mathrm{Var}(\hat{c}_i^{(j)}) = \mathbb{E}\left[\left(\sum_{k\neq i} g_j(i)g_j(k)c_k\mathbf{1}_{\{h_j(k)=h_j(i)\}}\right)^2\right]$$

Expanding cross terms:

$$\mathrm{Var}(\hat{c}_i^{(j)}) = \sum_{k\neq i} c_k^2 \cdot \mathbb{E}[g_j(i)^2 g_j(k)^2\mathbf{1}_{\{h_j(k)=h_j(i)\}}] + \sum_{k\neq l, k, l\neq i} c_k c_l \mathbb{E}[g_j(i)^2 g_j(k)g_j(l)\mathbf{1}_{\{h_j(k)=h_j(i)\}}\mathbf{1}_{\{h_j(l)=h_j(i)\}}]$$

- Cross terms vanish since $\mathbb{E}[g_j(k)g_j(l)] = 0$ for $k \neq l$ - $g_j(i)^2 g_j(k)^2 = 1$
Thus:

$$\mathrm{Var}(\hat{c}_i^{(j)}) = \sum_{k\neq i} c_k^2 \cdot \Pr(h_j(k) = h_j(i)) \leq \frac{\Sigma_2}{R}, \quad \Sigma_2 = \sum_k c_k^2$$

2.4.3 Chebyshev Bound

$$\Pr(|\hat{c}_i^{(j)} - c_i| \geq k\sigma) \leq \frac{1}{k^2}, \quad \sigma^2 = \mathrm{Var}(\hat{c}_i^{(j)})$$

3 Count-Min Sketch

3.1 Motivation

CMS is simpler, faster, but:

- Counts always ≥ 0
- Overestimates true counts
- Cannot handle negative increments

3.2 Data Structure

- d hash functions $h_1, \ldots, h_d : [1..N] \to [1..R]$ - $d \times R$ array CMS initialized to 0

3.3 Operations

Increment:

$$CMS[i][h_i(x)] \leftarrow CMS[i][h_i(x)] + \Delta, \quad i = 1 \ldots d$$

Query:

$$\hat{c}_x = \min_{i=1}^{d} CMS[i][h_i(x)]$$

3.4 Derivation of Estimate

Single row i:

$$\hat{c}_i = CMS[h(i)] = c_i + \sum_{j \neq i} c_j \mathbf{1}_{\{h(j)=h(i)\}}$$

3.4.1 Expectation

$$\mathbb{E}[\hat{c}_i] = c_i + \sum_{j \neq i} c_j \cdot \frac{1}{R} = c_i + \epsilon\Sigma$$

- Always overestimates by at most $\epsilon\Sigma$ in expectation

3.4.2 Markov Bound

$$\Pr(\hat{c}_i - c_i > 2\epsilon\Sigma) \leq \frac{\mathbb{E}[\hat{c}_i - c_i]}{2\epsilon\Sigma} \leq \frac{1}{2}$$

d independent hash functions:

$$\Pr(\text{all rows unlucky}) \leq 0.5^d$$

3.5 Memory Requirement

$$R = \frac{1}{\epsilon}, \quad d = \log_2 \frac{N}{\delta} \implies \text{Memory} = O\left(\frac{1}{\epsilon} \log \frac{N}{\delta}\right)$$

3.6 Top-K Heavy Hitters

Maintain a min-heap of size K:

- Update CMS
- Query $\hat{c}_x$
- If $\hat{c}_x$ exceeds heap minimum, replace

Worst-case update: $O(d \log K)$

4 Proofs: Markov and Chebyshev

4.1 Markov Inequality

For $X \geq 0$, $a > 0$:

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Proof: Let $I = \mathbf{1}_{X \geq a}$. Then $X \geq aI \implies \mathbb{E}[X] \geq a \Pr(X \geq a)$.

4.2 Chebyshev Inequality

For X with mean μ, variance σ^2:

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Proof: Let $Y = (X - \mu)^2$. Then $\Pr(Y \geq k^2\sigma^2) \leq \frac{\mathbb{E}[Y]}{k^2\sigma^2} = \frac{1}{k^2}$

5 Summary

- Count Sketch: unbiased, handles negative counts, variance bounded by Σ_2/R, Chebyshev bounds.
- Count-Min Sketch: overestimates, fast, Markov bounds.
- Memory vs. accuracy tradeoff: increase R to reduce error, increase d to reduce probability of large error.
- Heavy hitters: maintain min-heap of size K, update $O(d \log K)$.