CS 559: Algorithmic Aspects of Computer Networks **Fall 2007**

Lecture 9 — October 3,8

Lecturer: John Byers *BOSTON UNIVERSITY* *Scribe: Samuel Epstein*

In these lectures, we learned about how to use Count-Min sketches to efficiently approximate data streams. The first section of the notes describe fundamentals aspects of bounding probability distributions. The second section describes the algorithm and analysis of the Count-Min sketch algorithm.

9.1 Bounds on Distributions

9.1.1 Expectation and Variance

The expectation of a random variable X is the weighted average of the values it assumes:

$$\mathbf{E}[X] = \sum_i i * \Pr(X = i)$$

The expectation of the sum of random variables is equal to the sum of the expectations of the random variables. This is known as the *linearity of expectations.* Given n random variables $X_1, X_2, \ldots, X_n$ with finite expectations:

$$\mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i}^{n} \mathbf{E}[X_i]$$

The variance of a random variable is a measure of how far the random variable is likely to be away from its expectation:

$$\mathbf{Var}[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

9.1.2 Markov Inequality

Let X be a random variable that assumes only nonnegative values. Then for all $a \geq 0$:

$$\Pr(X \geq a) \leq \frac{\mathbf{E}(X)}{a}$$

Markov's inequality gives the best tail bound possible when the only information available is the expectation of the random variable, and the guarantee that X is not negative.

9.1.3 Chebysev Inequality

A tighter bound can be made if the variance of a probability distribution is known. Let X be a random variable, for any $a \geq 0$.

$$\Pr(|X - \mathbf{E}[X]| \geq a) \leq \frac{\mathbf{Var}[X]}{a^2}$$

9.1.4 Chernoff Bounds

Chernoff bounds are extremely powerful, giving exponentially decreasing bounds on the tail distributions. Let X be a random variable, and let $0 \leq \delta \leq 1$. The Chernoff Bound is as follows:

$$\Pr(X \geq (1+\delta)\mathbf{E}[X]) \leq 2e^{\mathbf{E}[X]\delta^2/3}$$

9.2 Datastream Summarization Problem

Consider a vector A of dimension size n. The vector's current state at time t is represented by $A(t) = [A_1(t), A_2(t), \ldots, A_n(t)]$. A is initially the zero vector $A_i[0] = 0$ for all i. The tth update is represented by $< i_t, c_t >$, meaning that position i_t is increased by c_t:

$$\begin{aligned} A_{i_t}(t) &= A_{i_t}(t-1) + c_t \\ A_{i'}(t) &= A_{i'}(t-1), \text{ for all } i' \neq i_t \end{aligned}$$

This vector A is a formalization of a datastream. The *Datastream Summarization Problem* aims to use polylogithrmic space on n to approximate three types of queries for vectors A and B.

- *point query* - Approximate $A_i(t)$
- *range query* - Approximate $\sum_{i \leq j \leq k} A_i(t)$
- *inner product query* - Approximate $A_i(t) \bullet B_i(t) = \sum_i A_i(t) * B_i(t)$

9.3 Cash Register Problem

The datastream A is said to be a cash register, or *monotone* if $c_t \geq 0$. Thus $A_i(t) \geq A_i(s)$ for all $t \geq s$. The Count-Min, or CM, sketch (fig. 9.2) is a two dimensional array of width w

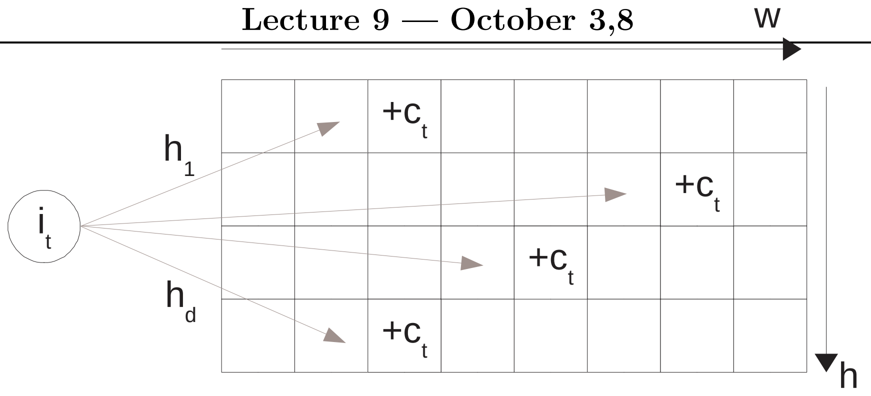

Figure 9.1. Count Min Sketch

and height h and d hash functions $h_1, \dots, h_d$ of range w. The two dimensional array consists of wd counts. When an update $< i_t, c_t >$ arrives, for each $1 \leq i \leq d$:

$$count[i, h_i(i_t)] \leftarrow count[i, h_i(i_t)] + c_t$$

CM-Sketch can produce an estimate of $A_t(i)$, denoted by $\hat{A}_t(i)$. When a query for $A_t(i)$ is received, CM-Sketch returns:

$$\hat{A}_t(i) = \min_{1 \leq j \leq d} count[j, h_j(i)]$$

9.3.1 Correctness Analysis of Cash Register Problem

The CM-Sketch can be written as CM-Sketch(δ, ε), where $w = e/\varepsilon$ and $h = \lceil \ln \frac{1}{\delta} \rceil$. It can be proven using *linearity of expectations* and the *Markov Inequality*:

$$\Pr\left[\hat{A}_t(i) > A_t(i) + \varepsilon * ||A||_1\right] \leq \delta$$

,where $||A||_1 = \sum_{i=1}^{n} |a_i(t)|$. $\hat{A}_t(i)$ will always be greater than $A_t(i)$, and the above equation gives probabilistic upper bounds on the approximation.

9.4 Turnstile Problem

The datastream A is said to be an Turnstile, if c_t can be negative. For the *Turnstile Summarization Problem*, the approximation returned is equal to the median of the count cells:

$$\hat{A}_t(i) = \text{median}_{1 \leq j \leq d}\, count[j, h_j(i)]$$

9.5 Space Analysis of CM-Sketch

Previous sketches required $\Omega\left(\frac{1}{\varepsilon^2}\ln\frac{1}{\delta}\right)$ space, whereas CM-Sketch requires $O\left(\frac{1}{\varepsilon}\ln\frac{1}{\delta}\right)$ space.