Randomized Algorithms 2013A
Lecture 2 – The second moment and data-stream algorithms*

Robert Krauthgamer

1 The second moment

Chebychev's inequality: Let X be a random variable with finite variance $\sigma^2 > 0$. Then

$$\forall t \geq 1, \quad \Pr\Big[|X - \mathbb{E}X| \geq t\sigma\Big] \leq \tfrac{1}{t^2}.$$

Intuition: Such a random variable is WHP in the range $\mu \pm \sigma$.

Proof: seen in class based on Markov's inequality.

Exer: Prove Markov's inequality. (Hint: use the law of total expectation.)

2 More occupancy problems

2.1 Empty bins for $m = n$ balls

Let Z_i be an indicator for the event that bin i is empty, which in the languange of previous class is just $I_{\{X_i=0\}}$. Denote the number of empty bins by $Z = \sum_i Z_i$, then we saw last week $\mathbb{E}[Z] \approx n/e$.

Can we give a high probability bound on the value of Z?

$$\mathbb{E}[Z^2] = \mathbb{E}[\sum_{i,j} Z_i Z_j] = \sum_{i,j} \Pr[Z_i = Z_j = 1] = \sum_{i \neq j} (1-2/n)^n + \sum_i (1-1/n)^n \approx \tfrac{n(n-1)}{e^2} + \tfrac{n}{e} \approx \tfrac{n^2}{e^2}.$$

Thus, when analyzing $\mathrm{Var}(Z) = \mathbb{E}[Z^2] - (\mathbb{E}Z)^2 \approx \frac{n^2}{e^2} - \frac{n^2}{e^2}$ requires going into lower order terms...

Exer: Prove that $\mathrm{Var}(Z) \leq O(n)$.

Using the exercise, we can conclude that WHP $Z = \frac{n}{e} \pm O(\sqrt{n})$.

*These notes summarize the material covered in class, usually skipping proofs, details, examples and so forth, and possibly adding some remarks, or pointers. The exercises are for self-practice and need not be handed in. In the interest of brevity, most references and credits were omitted.

2.2 Hitting all bins (coupon collector)

Let Y_i be the number balls thrown until i distinct bins are hit. We are interested in Y_n, and by definition $Y_1 = 1$. Observe that $Z_i = Y_i - Y_{i-1}$ has geometric distribution $G(p = \frac{n-(i-1)}{n})$. Thus,

$$\mathbb{E}[Z_i] = \tfrac{1}{p} = \tfrac{n}{n-i+1}, \quad \mathrm{Var}(Z_i) = (1-p)/p^2 = \tfrac{i-1}{n} \cdot \tfrac{n^2}{(n-i+1)^2} = \tfrac{(i-1)n}{(n-i+1)^2}.$$

Since we can write $Y_n = \sum_{i=1}^n Z_i$ (by convention $Z_1 = 1$), we can easily see that $\mathbb{E}[Y_n] \approx n \ln n$ and $\mathrm{Var}(Y_n) \leq O(n^2)$. Thus, using Chebyshev's inequality,

$$\Pr[Y_n > 3n \ln n] \leq \Pr[Y_n - \mathbb{E}Y_n \geq 2n \ln n] \leq O(1/\ln^2 n).$$

But we can get a stronger bound using a direct calculation:

$$\Pr[X_1 = 0] \leq (1 - 1/n)^m \leq e^{-m/n} = 1/n^3,$$

hence

$$\Pr[\exists i, X_i = 0] \leq n \Pr[X_1 = 0] \leq 1/n^2.$$

2.3 Collisions for $m = c\sqrt{n}$ (birthday paradox)

We shall use Chebyshev's inequality, although it's also possible to analyze via a direct computation.

Exer: Show that if $c > 0$ is a sufficiently small constant, then with high (constant) probability there are no collisions, i.e., the maximum load is $max_i X_i \leq 1$. (Hint: Look at every pair of balls.)

Exer: Show that if $c > 0$ is a sufficiently large constant, then with high (constant) probability there is at least one collision, i.e., $\max_i X_i \geq 2$. (Hint: Look at every pair of balls.)

3 AMS algorithm for ℓ_2-norm of a data stream

Data stream model:

Input: a vector $x \in \mathbb{R}^n$, given as a stream (sequence) of m updates of the form (i, a), meaning $x_i \leftarrow x_i + a$.

Motivation: We receive a stream of m items, each in the range $[n]$, and we let x_i is the frequency of item i. Upon seeing an item $i \in [n]$, we update $(i, +1)$. Then the second frequency moment F_2 is just $\|x\|_2^2$.

ℓ_p-norm problem:

Assumption: updates a are integral and $|x_i| \leq \mathrm{poly}(n)$.

Goal: estimate its ℓ_p-norm $\|x\|_p$. It's usually more convenient to work with its p-th power $(\|x\|_p)^p = \sum_{i=1}^n |x_i|^p$.

We focus here on $p = 2$. Note that we could have $a < 0$ (deletions) and maybe even $x_i < 0$.

Linear sketch: We shall use a randomized linear map $L : \mathbb{R}^n \to \mathbb{R}^s$ for small $s > 0$. The algorithm will only maintain Lx, which is easy to update since:

$$L(x + ae_i) = Lx + a(Le_i).$$

Of course, one has to choose L that somehow "stores" $\|x\|_2$. Note that L is essentially an $s \times n$ (real) matrix.

The memory requirement depends on the dimension s, the accuracy needed for each coordinate, and the representation of L (more precisely, storing a few random bits that suffice to produce L_{ij} on the fly).

Theorem 1 [Alon-Matthias-Szegedy'96]: One can estimate the ℓ_2 norm within factor $1+\varepsilon$ using a linear sketch of $s = O(\varepsilon^{-2}\log n)$ memory words.

Algorithm A:

1. Choose initially $r_1, \ldots, r_m$ independently and uniformly at random from $\{-1, +1\}$.
2. Maintain $Z = \sum_i r_i x_i$ (a linear sketch, hence can be updated as above).
3. Output: Z^2.

Analysis of expectation: As seen in class, $\mathbb{E}[Z^2] = \sum_i x_i^2 = \|x\|_2^2$.

We aren't done yet since we want to get $1+\varepsilon$ accuracy...

Analysis of second moment: As seen in class, $\mathrm{Var}(Z^2) \le \mathbb{E}[Z^4] \le 3(\mathbb{E}[Z^2])^2$.

Algorithm B: Execute $t = O(1/\varepsilon^2)$ independent copies of Algorithm A, denoting their estimates by $Y_1, \ldots, Y_t$, and output their mean $\tilde{Y} = \sum_j Y_j/t$.

Observe that the sketch $(Y_1, \ldots, Y_t) \in \mathbb{R}^t$ is still linear.

Analysis: As seen in class, using Chebychev's inequality and an appropriate $t = O(1/\varepsilon^2)$

$$\Pr[\tilde{Y} \neq (1 \pm \varepsilon)\|x\|_2^2] \le \tfrac{3}{t\varepsilon^2} \le 1/3.$$

Space requirement: $t = O(1/\varepsilon^2)$ words (for constant success probability), without counting memory used to represent/store L.

Concern: How do we store the n values $r_1, \ldots, r_n$?

Exer: For what value of k would the basic analysis work assuming that $r_1, \ldots, r_n$ are k-wise independent?

Exer: What would happen (to accuracy analysis) if the r_i's were chosen as standard gaussians $N(0, 1)$?

High probability bound:

Lemma: Let B' be a randomized algorithm to approximate some function $f(x)$, i.e.,

$$\forall x, \quad \Pr[B'(x) = (1 \pm \varepsilon)f(x)] \ge 2/3.$$

Let algorithm C output the median of $O(\log \frac{1}{\delta})$ independent executions of algorithm B'. Then

$$\forall x, \quad \Pr[C(x) = (1 \pm \varepsilon) f(x)] \geq 1 - \delta.$$

Exer: prove this lemma. (Hint: Use the Chernoff-Hoeffding bound.)

4 Count-min sketch for ℓ_1 point queries

ℓ_p point query problem:

Goal: at the end of the stream, given query i, report, for a parameter $\alpha \in (0, 1)$,

$$\tilde{x}_i = x_i \pm \alpha \|x\|_p.$$

Observe: $\|x\|_1 \geq \|x\|_2 \geq \ldots \geq \|x\|_\infty$, hence higher norms (larger p) gives better accuracy.

Exer: Show that the ℓ_1 and ℓ_2 norms differ by at most a factor of $\sqrt{n}$, and that this is tight. Do the same for ℓ_2 and ℓ_∞.

Theorem 2 [Cormode-Muthukrishnan'05]: One can answer ℓ_1 point queries within error α with probability $1 - 1/n^2$ using a linear sketch of $O(\alpha{-}1 \log n)$ memory words.

Algorithm D: (We assume for now $x_i \geq 0$ for all i.)

1. Set $w = 2/\alpha$ and choose a random function $h : [m] \to [w]$ (actually, a hash function).
2. Maintain a table $Z = [Z_1, \ldots, Z_w]$ where each $Z_j = \sum_{i:h(i)=j} x_i$ (which is a linear sketch).
3. When asked to estimate x_i, output $\tilde{x}_i = Z_{h(i)}$.

Analysis (correctness): As seen in class, $\tilde{x}_i \geq x_i$ holds always, and using Markov's inequality, $\Pr[\tilde{x}_i - x_i \geq \alpha \|x\|_1] \leq 1/2$.

Algorithm E: Execute $t = O(\log n)$ independent copies of algorithm D, i.e., maintain vectors $Z^1, \ldots, Z^t$ and functions $h^1, \ldots, h^t$. When asked to estimate, output the minimum among the t estimates, i.e., $\hat{x}_i = \min_l Z^l_{h^l(i)}$.

Analysis (correctness): Setting $t = O(\log n)$ we have

$$\Pr[|\hat{x}_i - x_i| \geq \alpha \|x\|_1] \leq (1/2)^t = 1/n^2.$$

Space requirement: $O(\alpha^{-1} \log n)$ words (for success probability $1 - 1/n^2$), without counting memory used to represent the hash functions.

Exer: Extend the algorithm to general x. (Hint: replace the min operator by median.)