University of Wrocław: Algorithms for Big Data (Fall'19) 14/10/2019

Lecture 2: AMS algorithm, Johnson-Lindenstrauss lemma

Lecturer: *Przemysław Uznański* Scribe: *Marcin Sidorowicz*

1 Models of streaming

1. **Sequence of values** - the input is simply a sequence of values x_i.
2. **Semi-turnstile** - an update consists of a pair (x_i, c_i), $c_i > 0$, where c_i is multiplicity of value x_i.
3. **Turnstile** - an update consists of a pair (x_i, c_i) with no bounds on c_i.

We can think of the state being an array A initially filled with 0's, and the update being the operation $A[x_i] \leftarrow A[x_i] + c_i$.

2 F_p moment

For a vector $\mathbf{x} = (x_1, \ldots, x_d)$ we define its p-th moment to be

$$F_p = \sum_{i=1}^{d} |x_i|^p$$

and p-th norm to be

$$L_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}}$$

Examples: F_0 - number of non-zero elements (which is nontrivial in turnstile model, because c_i can be less than 0), F_1/L_1 - Manhattan norm (trivial in semi-turnstile), L_2 - Euclidean norm.

2.1 AMS algorithm (Alon, Matias, Szegedy 1996) for F_2

2.1.1 Introduction

Let $A = (a_1, a_2, \ldots, a_n)$. We need a linear operation which will tell us the second moment of A, i.e. $\sum_{i=1}^{n} a_i^2$.
Consider a sequence of independent coefficients

$$r_i = \begin{cases} 1 & \text{with ppb } 1/2 \\ -1 & \text{with ppb } 1/2 \end{cases}$$

and let $Z = \sum_{i=1}^{n} r_i a_i$. $\mathbb{E}[Z]$ is obviously 0, so let us look at the variance:

$$\begin{aligned}
\mathrm{Var}(Z) = \mathbb{E}[Z^2] &= \mathbb{E}\left[\left(\sum_i r_i a_i\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i,j} r_i a_i r_j a_j\right] \\
&= \sum_{i,j} a_i a_j \mathbb{E}[r_i r_j] \\
&= \sum_i a_i^2 \mathbb{E}[r_i^2] \quad \text{(since for } i \neq j \text{ we have } \mathbb{E}[r_i r_j] = 0\text{)} \\
&= \sum_i a_i^2 = F_2,
\end{aligned}$$

so it's an unbiased estimator of F_2. How good is the approximation?

$$\begin{aligned}
\mathrm{Var}(Z^2) \leq \mathbb{E}[Z^4] &= \mathbb{E}\left[\left(\sum_i a_i r_i\right)^4\right] \\
&= \sum_i a_i^4 \mathbb{E}[r_i^4] + 3\sum_{i \neq j} a_i^2 a_j^2 \mathbb{E}[r_i^2 r_j^2] \\
&\leq 4\left(\sum_j a_i^2\right)^2 \\
&= 4\mathbb{E}[Z^2]^2,
\end{aligned}$$

while the transition from first to second line is because whenever the r_i are in odd power, their expectations cancel out to 0.
With such approximation the Chebyshev inequality gives us

$$\mathbb{P}\left(\left|Z^2 - \mathbb{E}[Z^2]\right| \geq 4\mathbb{E}[Z^2]\right) \leq \frac{1}{4}$$

A large, constant upper bound with constant probability isn't good, but we can apply the "median trick" again - repeat the process $O\left(\frac{1}{\varepsilon^2}\right)$ times, reducing the variance, and then repeat it $O\left(\log \frac{1}{\delta}\right)$ times to get an approximation with $1 - \delta$ probability. Overall complexity is $O\left(\frac{\log \frac{1}{\delta}}{\varepsilon^2}\right)$ (of words of memory) and we need 4-wise independence of the hashing functions.

2.2 Linearity of sketches

Rehashing the AMS sketches: fix $k = O(\varepsilon^{-2})$. Take a random matrix $R \in M_{k \times n}(\mathbb{R})$ with 4-wise independent coefficients $r_{ij} \in \{-1, 1\}$, calculate the vector $Z = Rx$ (sketch) and estimate $|x|_2^2$ with $|Z|_2^2 \cdot \frac{1}{k}$. The fact that our sketches are linear transformation is advantageous: if we compute sketch of vector x being Rx, and a sketch of vector y being Ry, then we can sketch $x - y$ as $R(x - y) = Rx - Ry$ - so e.g. we can estimate Euclidean distance between vectors. Similarly, linearity allows us to perform turnstile updates in the streaming setting.

The very idea of projecting $\mathbb{R}^n$ linearly to $\mathbb{R}^k$ for some small k is called *dimensionality reduction*.

2.3 Other distributions

Why did we choose a Bernoulli distribution for r_i? Since we only relied on the fact that $\mathbb{E}[r_i] = 0$ and $\mathrm{Var}(r_i) = 1$, in theory any symmetric distribution with variance 1 could work. Again, let $Z = \sum_{i=1}^{n} r_i a_i$. Since the mean and variance of r_i is finite, we know that

$$Z \xrightarrow{CLT} \mathcal{N}(\mu, \sigma^2),$$

so it might be advantageous to start with Gaussian distribution in the first place.

3 JL lemma

3.1 AMS with Gaussian coefficients

Let $R \in M_{k\times n}(\mathbb{R})$, $r_{ij} \sim \mathcal{N}(0,1)$, x and Z as before. Then

$$\mathbb{E}\left[|Rx|_2^2 \cdot \frac{1}{k}\right] = |x|_2^2.$$

We would like to end up with a bound for result deviation of

$$\mathbb{P}(|Z|_2^2 - k|x|_2^2 \le \varepsilon k|x|_2^2) \le \exp(-C\varepsilon^2 k),$$

since then for $k = O\left(\frac{\log\frac{1}{\delta}}{\varepsilon^2}\right)$ the probability of failure is $1-\delta$. WLOG assume that $|x|_2 = 1$.

$$\mathbb{P}(|Z|_2^2 \ge (1+\varepsilon)k \le \exp(-\varepsilon^2 k + O(k\varepsilon^3)).$$

Substituting $Y = |Z|_2^2$, $\alpha = k(1+\varepsilon)$, for any s from Markov inequality we get

$$\begin{aligned}
\mathbb{P}[Y > \alpha] = \mathbb{P}[e^{sY} > e^{s\alpha}] &\le \frac{\mathbb{E}[e^{sY}]}{e^{s\alpha}} \\
&= e^{-s\alpha}\mathbb{E}[e^{s\sum_{i=1}^{k} Z_i^2}] \\
&= e^{-s\alpha}\prod_{i=1}^{k}\mathbb{E}[e^{sZ_i^2}]
\end{aligned}$$

Since Z_i are all normal as a linear combination of independent normal variables, we have

$$\begin{aligned}
\mathbb{E}[Z_i] &= \sum_j \mathbb{E}[r_{ij}x_j] = 0 \\
\mathrm{Var}(Z_i) = \mathbb{E}[Z_i^2] &= \sum_j \mathbb{E}[r_{ij}^2 x_j^2] = |x|_2^2 = 1
\end{aligned}$$

So $Z_i \sim \mathcal{N}(0,1)$.

$$\begin{aligned}
\mathbb{E}[e^{sZ_i^2}] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left((s-\frac{1}{2})t^2\right)dt \\
&= \frac{1}{\sqrt{1-2s}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{u^2}{2}\right)du \\
&= \frac{1}{\sqrt{1-2s}}
\end{aligned}$$

Where we have used substitution $u^2 = (1-2s)t^2$ (then $dt = \frac{u}{t}\frac{1}{1-2s}du = \frac{1}{\sqrt{1-2s}}du$).

Minimizing this value, we set $s = \frac{1}{2} - \frac{k}{2\alpha}$, obtaining

$$\begin{aligned}\mathbb{P}(Y > \alpha) &= \exp(\frac{k-\alpha}{2})\left(\frac{k}{\alpha}\right)^{-\frac{k}{2}} \\ &= \exp(\frac{k}{2}(-\varepsilon + \ln(1+\varepsilon))) \\ &= \exp(k(-\frac{\varepsilon^2}{2}) + O(\varepsilon^3)).\end{aligned}$$

Lemma 1 (Johnson-Lindenstrauss)**.** *Let $P \subseteq \mathbb{R}^n$, $|P| = n$, $R \in M_{k\times n}(\mathbb{R})$, $r_{ij} \sim \mathcal{N}(0,1)$. Then for any $u, v \in P$*

$$|u-v|_2(1-\varepsilon) \leq \left|\frac{1}{\sqrt{k}}Ru - \frac{1}{\sqrt{k}}Rv\right|_2 \leq |u-v|_2(1+\varepsilon)$$

with prob. $1-\delta$, where $k = O\left(\frac{\log\frac{1}{\delta}}{\varepsilon^2}\right)$.

Informally: projection onto a "random" hyperplane (up to scaling) preserves distances with good accuracy and high probability.

Remark: r_{ij} from +1/-1 Bernoulli distribution like in AMS algorithm work too [c.f. Achlioptas 2003].

AMS analysis is much simpler and uses only variance, while JL is much more advanced. JL gives "fancier" convergence (no median trick) but essentially needs more independence for the random variables (k-independence).