CMPUT 675: Algorithms for Streaming and Big Data **Fall 2019**

Lecture 5 (Sep 18, 2019): Dimensionality reduction and F_p-estimator

Lecturer: Mohammad R. Salavatipour *Scribe: Ramin Mousavi*

5.1 Introduction

Recall the AMS sketching algorithm for estimating the second frequency moment F_2 of a stream from last lecture. We can represent the algorithm by a matrix $M_{t\times n}$[1] where $M_{ij} = h_i(j)$, i.e., the ij-th entry is the value of the hash function chosen in the i-th copy of the basic algorithm with input $j \in [n]$. Let $x \in \mathbb{N}^n$ be the frequency vector of the stream σ, then the output of the algorithm equals to $\frac{\|Mx\|_2^2}{t}$.

In this lecture we will see another sketching algorithm known as Johnson-Lindenstrauss algorithm for dimensionality reduction [JL84]. This algorithm inspires another F_2-estimator. However, this algorithm is not space efficient in reality, please see Section 5.3. In the last section, we start the discussion on how to estimate F_p for $p \in (0, 2]$.

5.2 Johnson-Lindenstrauss Dimensionality Reduction

5.2.1 Normal Distribution

Definition 1 *We say X is a normal random variable denoted by $N(\mu, \sigma)$ with mean μ and variance σ^2. The density function of X is defined as follows:*

$$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Let X be a normal random variable with density function ϕ, then we have $\Pr[a \le X \le b] = \int_a^b \phi(x)dx$. The standard normal distribution is denoted by $N(0,1)$ and its density function is $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. We need the following two facts about standard normal distributions in the analysis of our algorithms.

Lemma 1 (Theorem 9.2 from [MU18]) *Let $X_1, ..., X_k$ be k independent standard normal random variables and let $Y = a_1X_1 + ... + a_kX_k$, then Y has a normal distribution with mean 0 and variance $a_1^2 + ... + a_k^2$, i.e., Y has the same distribution as $N(0, \sqrt{\sum_{i=1}^{k} a_i^2})$.*

Lemma 2 *Let $X_1, ..., X_k$ be k independent standard normal random variables and let $Y = \frac{1}{k}\sum_{i=1}^{k} X_i^2$. Then, for any $0 < \epsilon < \frac{1}{2}$ we have*

$$\Pr[(1-\epsilon)^2 \le Y \le (1+\epsilon)^2] \ge 1 - e^{-c\epsilon^2 k},$$

[1] M is implicit and the algorithm does not need to store the matrix as it will be too expensive to store, please see the last lecture for more details.

where c is a constant.

Proof. We give a sketch of the proof, after all we are talking about sketching algorithms!

First we need to compute the moment generating function of X^2 where $X \sim N(0,1)$.

$$M_{X^2}(t) := \mathbb{E}[e^{tX^2}] = \int_{-\infty}^{\infty} e^{tx^2} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{1-2t}}.$$

Then, following the same proof for Chernoff bounds we get

$$\begin{aligned}
\Pr[Y \geq 1+\epsilon] &= \Pr[e^{tkY} \geq e^{tk(1+\epsilon)}] && (5.1)\\
&\leq \frac{\mathbb{E}[e^{tkY}]}{e^{tk(1+\epsilon)}} && (5.2)\\
&= \frac{\prod_{i=1}^{k} \mathbb{E}[e^{tkX_i^2}]}{e^{tk(1+\epsilon)}} && (5.3)\\
&= \frac{\prod_{i=1}^{k} M_{X_i^2}(tk)}{e^{tk(1+\epsilon)}} && (5.4)\\
&= \frac{\prod_{i=1}^{k} \frac{1}{\sqrt{1-2kt}}}{e^{tk(1+\epsilon)}}, && (5.5)
\end{aligned}$$

where the inequality follows from Markov inequality. The lemma follows by optimizing (5.5), i.e., find t such that (5.5) is minimized. Similarly, we can upper bound $\Pr[Y \leq 1-\epsilon]$. ■

5.2.2 Johnson-Lindenstrauss Algorithm

Theorem 1 (Johnson-Lindenstrauss) *Let $v_1, ..., v_n$ be any set of n points in $\mathbb{R}^d$. For any $0 < \epsilon < \frac{1}{2}$, there is a linear mapping $f : \mathbb{R}^d \to \mathbb{R}^k$ for $k = O(\alpha \frac{\log n}{\epsilon^2})$, where α is the precision parameter, such that for all $1 \leq i, j \leq n$ we have*

$$\Pr\left[(1-\epsilon)\left\|v_i - v_j\right\|_2 \leq \left\|f(v_i) - f(v_j)\right\|_2 \leq (1+\epsilon)\left\|v_i - v_j\right\|_2\right] \geq 1 - \frac{1}{n^\alpha}.$$

Furthermore, f can be computed in polytime.

Proof. Let M be a $k \times d$ matrix where $M_{ij} \sim N(0,1)$, i.e., each entry is drawn from standard normal random variable. We define f in the theorem's statement to be $f(v) := \frac{1}{\sqrt{k}} M v$.

The next lemma shows that we only need to work with unit vectors in $\mathbb{R}^d$.

Lemma 3 *To prove Theorem 1 it is enough to show that for an arbitrary unit vector $v \in \mathbb{R}^d$ we have*

$$\Pr[1-\epsilon \leq \|f(v)\|_2 \leq 1+\epsilon] \geq 1 - \frac{1}{n^{\alpha+2}}. \qquad (5.6)$$

Proof of Lemma 3. Let $v := \frac{v_i - v_j}{\|v_i - v_j\|_2}$. Since v is a unit vector by (5.6) we have

$$\Pr[1-\epsilon \leq \left\| f\big(\frac{v_i - v_j}{\|v_i - v_j\|_2}\big) \right\|_2 \leq 1+\epsilon] \geq 1 - \frac{1}{n^{\alpha+2}}.$$

Since f is linear we get that

$$\Pr\left[(1-\epsilon)\left\|v_i - v_j\right\|_2 \leq \left\|f(v_i) - f(v_j)\right\|_2 \leq (1+\epsilon)\left\|v_i - v_j\right\|_2\right] \geq 1 - \frac{1}{n^{\alpha+2}},$$

in other words, the probability of failure for one pair (v_i, v_j) is at most $\frac{1}{n^{\alpha+2}}$. By the union bound the total probability of failure is at most $n^2 \cdot \frac{1}{n^{\alpha+2}} = \frac{1}{n^\alpha}$ and thus Theorem 1 follows. ■

Let us first show that in expectation the square of the norm of v will be preserved under f.

Lemma 4 $\mathbb{E}\left[\,\|f(v)\|_2^2\,\right] = \|v\|_2^2 = 1.$

Proof of Lemma 4. Define δ_{jk} to be 1 if $j = k$ and 0 otherwise. Then, we can write

$$\begin{aligned}
\mathbb{E}\left[\,\|f(v)\|_2^2\,\right] &= \mathbb{E}\left[\,\left\|\frac{1}{\sqrt{k}}Mv\right\|_2^2\,\right] && (5.7)\\
&= \sum_{i=1}^{k} \frac{1}{k}\,\mathbb{E}\left[(\sum_{j=1}^{d} M_{ij}v_j)^2\right] && (5.8)\\
&= \sum_{i=1}^{k} \frac{1}{k} \sum_{1\leq j,k\leq d} v_j v_k\,\mathbb{E}[M_{ij}M_{ik}] && (5.9)\\
&= \sum_{i=1}^{k} \frac{1}{k} \sum_{1\leq j,k\leq d} v_j v_k \delta_{jk} && (5.10)\\
&= \sum_{i=1}^{k} \frac{1}{k} \sum_{j=1}^{d} v_j^2 && (5.11)\\
&= \|v\|_2^2 && (5.12)\\
&= 1, && (5.13)
\end{aligned}$$

where (5.10) follows from the fact that M_{ij} and M_{ik} are drawn independently from $N(0,1)$. ■

To prove Theorem 1, by Lemma 3 it is enough to prove (5.6) for an arbitrary unit vector $v \in \mathbb{R}^d$. From now on we fix a unit vector $v \in \mathbb{R}^d$.

Lemma 5 *Let* $0 < \epsilon < \frac{1}{2}$ *and let* $k := \frac{(\alpha+2)\log n}{c\epsilon^2}$*, where* c *is the constant in Lemma 2. Then, we have*

$$\Pr[1-\epsilon \leq \|f(v)\|_2 \leq 1+\epsilon] \geq 1 - \frac{1}{n^{\alpha+2}}.$$

Proof of Lemma 5. We know that $(Mv)_i = \sum\limits_{j=1}^{d} M_{ij}v_j$. Since M_{ij} for $1 \leq j \leq d$ are independently drawn from $N(0,1)$, by Lemma 1 we know that $(Mv)_i = N(0, \sqrt{\sum\limits_{j=1}^{d} v_j^2}) = N(0,1)$ since $\|v\|_2^2 = 1$. Therefore, we can apply Lemma 2 to $\frac{1}{k}\sum\limits_{i=1}^{k}(Mv)_i^2 = \frac{1}{k}\|Mv\|_2^2$ and conclude that

$$\Pr[(1-\epsilon)^2 \leq \frac{\|Mv\|_2^2}{k} \leq (1+\epsilon)^2] \geq 1 - e^{-c\epsilon^2 k} = 1 - e^{-(\alpha+2)\log n} = 1 - \frac{1}{n^{\alpha+2}}, \tag{5.14}$$

where the first equality follows from substituting the value of k. From (5.14), we get that

$$\Pr[1-\epsilon \le \|f(v)\|_2 \le 1+\epsilon] \ge 1 - \frac{1}{n^{\alpha+2}},$$

as desired. ■

By putting together Lemma 3 and Lemma 5, we get Theorem 1. ■

5.3 Estimating F_p

Indyk [I06] gave a sketching algorithm for estimating F_p for $p \in (0,2]$. This algorithm also directly translates into dimensionality reduction for l_p norm for $p \in (0,2]$.

Before we present the sketching algorithm for F_p, we need some definition.

Definition 2 (Stable Distributions) *Let p be a positive real number. A probability distribution D_p is called p-stable if for n independently random variables $X_1, ..., X_n$ that are drawn from D_p and $c \in \mathbb{R}^n$, the random variable $\sum\limits_{i=1}^{n} c_i X_i$ has the same distribution as $\|c\|_p \cdot X$ where $X \sim D_p$.*

For example, $N(0,1)$ is a 2-stable distribution. Another example is Cauchy distribution $\mathcal{D}_C$ defined by the density function $c(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$, is 1-stable.

We have the following method for generating a random variable from D_p.

Theorem 2 (Chambers-Mallows-Stuck'76 [CMS76]) *Given a p-stable probability distribution D_p, we can generate a random variable X from D_p as follows:*

- *sample (θ, r) from $[-\frac{\pi}{2}, \frac{\pi}{2}] \times [0,1]$ uniformly at random,*
- *return $X = \frac{\sin(p\theta)}{(\cos\theta)^{\frac{1}{p}}} (\frac{\cos((1-p)\theta)}{-\ln\frac{1}{r}})^{\frac{1-p}{p}}$.*

Now we can present a JL based F_2-estimator.

JL based F_2-estimator

```
1. Let $Y_1, ..., Y_n$ be independent random variables drawn from $N(0,1)$
2. $x \leftarrow 0$
3. While there is a new element $a_j = i \in [n]$ in the stream do
       $x \leftarrow x + Y_i$
4. Return $x^2$
```

First note that we need to generate all n random numbers $Y_1, ..., Y_n$ and store them (because we have to apply the same random number to the same value in the stream). Hence, this algorithm is not efficient, however by using pseudorandom generators we can address this issue.

The analysis of the above algorithm is like before.

Theorem 3 *The JL based F_2-estimator is an (ϵ, δ)-estimator.*

Proof. Let $\bar{x}$ be the value of x at the end of the algorithm. Thus, $\bar{x} = \sum_{i=1}^{n} f_i Y_i$. Since $Y_i \sim N(0,1)$ for $1 \leq i \leq n$ and $N(0,1)$ is a 2-stable distribution, we know that $\bar{x} = \|f\|_2 \cdot X = \sqrt{F_2} \cdot X$ where $X \sim N(0,1)$. So $\mathbb{E}[\bar{x}^2] = \|f\|_2^2 = F_2$. Also note that $\mathrm{Var}[\bar{x}^2] = F_2^2 \cdot \mathrm{Var}[X^2] = 2F_2^2$ since $\mathrm{Var}[X^2] = 2$. Then, we can apply the Chebyshev's bound and using the median of mean trick we can reduce the error probability. ■

The idea of the F_p-estimator is the same as the JL based F_2-estimator with this difference that instead of $N(0,1)$ we sample from a p-stable distribution D_p. Before presenting the algorithm we need two more definitions.

Definition 3 *The median of a probability distribution D is μ if for $X \sim D$, we have $\Pr[X \leq \mu] = \frac{1}{2}$. In other words, if $\phi(x)$ is the density function of D, then we have $\int_{-\infty}^{\mu} \phi(x)dx = \frac{1}{2}$.*

Definition 4 *Let D be a probability distribution with density function $\phi(x)$ such that $\phi(x) = \phi(-x)$. Then, $|D|$ is a probability distribution with the following density function $\psi(x)$:*

$$\psi(x) = \begin{cases} 2\phi(x), & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Finally, we can present the F_p-estimator.

An (ϵ, δ) F_p-estimator

```
1. $t \leftarrow O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$
2. $x \leftarrow 0$
3. Let $M$ be a $t \times n$ matrix where $M_{ij} \sim D_p$
4. While there are new element $a_j = i \in [n]$ in the stream do
       For $i = 1$ to $t$ do
           $x_i \leftarrow x_i + M_{ia_j}$
5. Return $\frac{\text{median}\,(|x_1|,\ldots,|x_t|)}{\text{median}\,(|D_p|)}$
```

We will analyze the above algorithm in the next lecture.

References

CMS76 J. M. Chambers, C. L. Mallows, and B. W. Stuck, A method for simulating stable random variables. *J. Amer. Statist. Assoc.*, 71, 340-344, 1976.

I06 P. Indyk, Stable distributions, pseudorandom generators, embeddings and data stream computation. *Journal of the ACM (JACM)*, 53.3: 307-323, 2006.

JL84 W. Johnson, and J. Lindenstrauss, Extension of Lipshitz mapping into Hilbert space. *Contemp. Math.*, 26: 189-206, 1984.

MU17 M. Mitzenmacher, and E. Upfal, Probability and computing: randomization and probabilistic techniques in algorithms and data analysis. *Cambridge university press*, 2017.