CSCE 689: Special Topics in Modern Algorithms for Data Science

Lecture 14

Samson Zhou

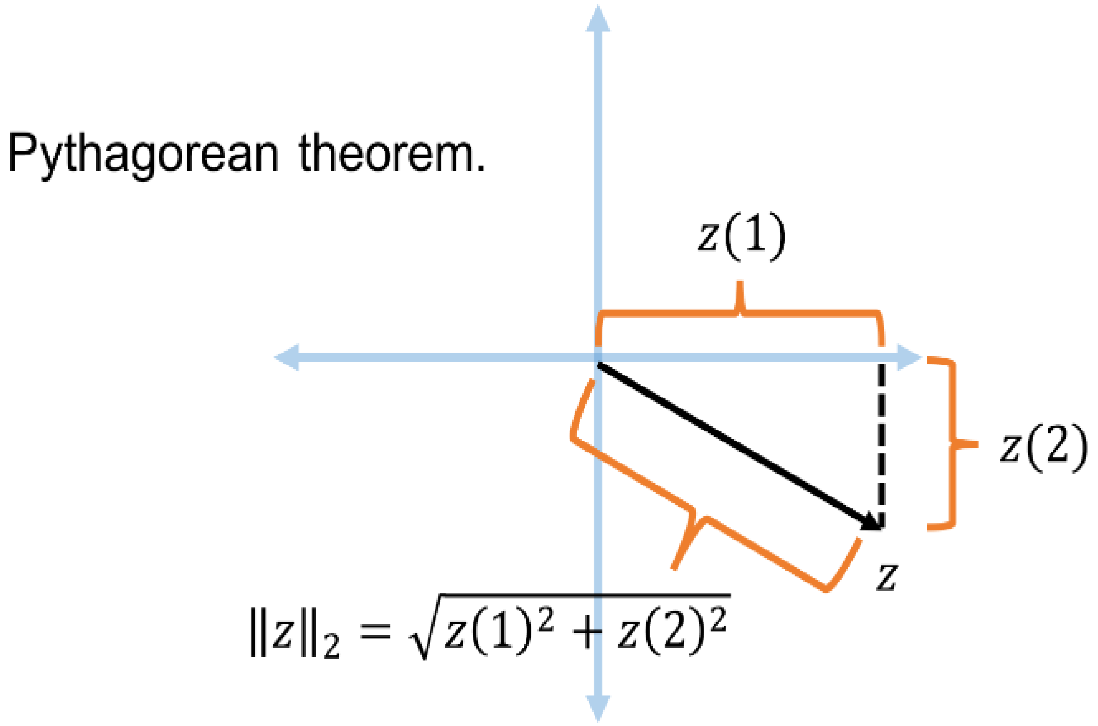
Presentation Schedule

- September 25: Team DAP, Team Bokun, Team Jason
- September 27: Galaxy AI, Team STMI
- September 29: Jung, Anmol, Chunkai

Recall: Euclidean Space and L_2 Norm

- For $z \in R^n$, the L_2 norm of z is denoted by $\|z\|_2$ and defined as:

$$\|z\|_2 = \sqrt{z_1^2 + z_2^2 + \cdots + z_n^2}$$



- For $x, y \in R^n$, the distance function D is denoted by $\|\cdot\|_2$ and defined as $\|x - y\|_2$

Recall: CountSketch Summary

- CountSketch solves the L_2 heavy-hitters problem: Given a set S of m elements from $[n]$ that induces a frequency vector $f \in R^n$ and a threshold parameter $\varepsilon \in (0, 1)$, output a list that includes:
  - The items from $[n]$ that have frequency at least $\varepsilon \cdot \|f\|_2$
  - No items with frequency less than $\frac{\varepsilon}{2} \cdot \|f\|_2$

- Space usage: $O\left(\frac{1}{\varepsilon^2} \log^2 n\right)$ bits of space

L_2 Estimation

- Goal: Given a set S of m elements from $[n]$ that induces a frequency vector $f \in R^n$ and an accuracy parameter $\varepsilon \in (0, 1)$, output a $(1+\varepsilon)$-approximation to $\|f\|_2$

- Find Z such that $(1-\varepsilon) \cdot \|f\|_2 \leq Z \leq (1+\varepsilon) \cdot \|f\|_2$

- Find Z' such that $(1-\varepsilon) \cdot \|f\|_2^2 \leq Z' \leq (1+\varepsilon) \cdot \|f\|_2^2$

F_2 Moment Estimation

- Goal: Find Z' such that $(1-\varepsilon)\cdot\|f\|_2^2 \leq Z' \leq (1+\varepsilon)\cdot\|f\|_2^2$

F_2 Moment Estimation

- Goal: Find Z' such that $(1-\varepsilon)\cdot\|f\|_2^2 \leq Z' \leq (1+\varepsilon)\cdot\|f\|_2^2$

1 7 7 7 3 7 7 1 4 1 1 1 1 5 1 1 7 1 7 5 1 7 7

F_2 Moment Estimation

- Goal: Find Z' such that $(1-\varepsilon)\cdot\|f\|_2^2 \leq Z' \leq (1+\varepsilon)\cdot\|f\|_2^2$

1 7 7 7 3 7 7 1 4 1 1 1 1 5 1 1 7 1 7 5 1 7 7

f_1	f_2	f_3	f_4	f_5	f_6	f_7
10	0	1	1	2	0	9

Johnson-Lindenstrauss Lemma

- Distributional Johnson-Lindenstrauss Lemma: Given $\Pi \in R^{m \times n}$ with $m = O\left(\frac{\log 1/\delta}{\varepsilon^2}\right)$ and each entry drawn from $\frac{1}{\sqrt{m}} N(0,1)$, then for any $x \in R^n$ and setting $y = \Pi x$, then with probability at least $1 - \delta$

$$(1 - \varepsilon)\|x\|_2 \leq \|y\|_2 \leq (1 + \varepsilon)\|x\|_2$$

F_2 Moment Estimation

- Algorithm: Generate $\Pi \in R^{m\times n}$ with $m = O\left(\frac{\log 1/\delta}{\varepsilon^2}\right)$ and each entry drawn from $\frac{1}{\sqrt{m}}N(0,1)$. Set $g = \Pi \cdot f$

- Whenever there is an update to a coordinate of f, update g

F_2 Moment Estimation

- Algorithm: Generate $\Pi \in R^{m \times n}$ with $m = O\left(\frac{\log 1/\delta}{\varepsilon^2}\right)$ and each entry drawn from $\frac{1}{\sqrt{m}} N(0,1)$. Set $g = \Pi \cdot f$

1 7 7 7 3 7 7 1 4 1 1 1 1 5 1 1 7 1 7 5 1 7 7

- Whenever there is an update to a coordinate of f, update g

F_2 Moment Estimation

- Algorithm: Generate $\Pi \in R^{m\times n}$ with $m = O\left(\frac{\log 1/\delta}{\varepsilon^2}\right)$ and each entry drawn from $\frac{1}{\sqrt{m}}N(0,1)$. Set $g = \Pi \cdot f$

1 7 7 7 3 7 7 1 4 1 1 1 1 5 1 1 7 1 7 5 1 7 7

- Whenever there is an update to a coordinate of f, update g
- $f = f + e_1$
- $f = f + e_7$
- $f = f + e_7$

F_2 Moment Estimation

- Algorithm: Generate $\Pi \in R^{m \times n}$ with $m = O\left(\frac{\log 1/\delta}{\varepsilon^2}\right)$ and each entry drawn from $\frac{1}{\sqrt{m}} N(0,1)$. Set $g = \Pi \cdot f$

1 7 7 7 3 7 7 1 4 1 1 1 1 5 1 1 7 1 1 7 5 1 7 7

- Whenever there is an update to a coordinate of f, update g
- $f = f + e_1$, $g = g + \Pi e_1$
- $f = f + e_7$, $g = g + \Pi e_7$
- $f = f + e_7$, $g = g + \Pi e_7$

AMS Algorithm

- Generate a random sign vector $s \in \{-1, +1\}^n$
- Maintain $Z = \langle s, f \rangle$
- Output $W := Z^2$

1

1

2

1

2

1

1

1

2

1

1

2

2

2

1

AMS Algorithm

- What values of Z did you get?
- $Z = \langle s, f \rangle = s_1 f_1 + s_2 f_2 + \cdots + s_n f_n$
- What values of W did you get?
- $W = Z^2 = \sum_{i,j} s_i s_j f_i f_j$

AMS Algorithm

- What values of W did you get?

- $W = Z^2 = \sum_{i,j} s_i s_j f_i f_j$

f_1	f_2	f_3	f_4	f_5	f_6	f_7
9	6	0	0	0	0	0

AMS Algorithm

- What is $\mathrm{E}[W]$?
- $Z = \langle s, f \rangle = s_1 f_1 + s_2 f_2 + \cdots + s_n f_n$
- $W = Z^2 = \sum_{i,j} s_i s_j f_i f_j$
- $\mathrm{E}[W] = \sum_{i,j} \mathrm{E}\left[s_i s_j f_i f_j\right] = \sum_i \mathrm{E}\left[f_i^2\right] = \|f\|_2^2$

AMS Algorithm

- What is $\mathrm{Var}[W]$?
- $Z = \langle s, f \rangle = s_1 f_1 + s_2 f_2 + \cdots + s_n f_n$
- $W^2 = Z^4 = \sum_{a,b,c,d} s_a s_b s_c s_d f_a f_b f_c f_d$
- $\mathrm{E}[W^2] = \sum_{a,b,c,d} \mathrm{E}[s_a s_b s_c s_d f_a f_b f_c f_d] = \sum_i \mathrm{E}\left[f_i^4\right] + 6 \sum_{i \neq j} \mathrm{E}\left[f_i^2 f_j^2\right] \leq 6\|f\|_2^4$

AMS Algorithm

- By Chebyshev's inequality, W will be a 9-approximation to $\|f\|_2^2$ with probability $\frac{2}{3}$

AMS Algorithm

- How to get $(1+\varepsilon)$-approximation?

- Repeat $O\left(\frac{1}{\varepsilon^2}\right)$ times and take the average

AMS Algorithm

- Space of algorithm: $O\left(\frac{1}{\varepsilon^2}\right)$ words of space or $O\left(\frac{1}{\varepsilon^2}\log m\right)$ bits of space