CSCE 689: Special Topics in Modern Algorithms for Data Science Fall 2023

Lecture 14 — September 22, 2023

Prof. Samson Zhou *Scribe: Chunkai Fu*

Previously we have learned that we can estimate the frequency of some element bounded by an error that is a certain fraction of the L_2 norm. The lecture today answers exactly this question. Here are some of the main points covered in class.

1 Why do we need to know L_2 norm?

The CountSketch algorithm 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 $\varepsilon \cdot \parallel f \parallel_2$, and no items with frequency less that $0.5\varepsilon \cdot \parallel f \parallel_2$. Notice that we would need to know the L_2 norm of the frequency vector to implement the CountSketch algorithm.

2 How do we apply the Johnson-Lindenstraus Lemma for finding L_2 norm?

We know that the distributional Johnson Lindenstrauss lemma applies a random d by n matrix with each entry drawn from a Gaussian distribution to multiple with a vector, then L_2 norm of the multiplied results will be bounded by $1 + / - \varepsilon$ of the L_2 norm of the vector. So to estimate the L_2 norm, we first generate a d by n matrix Π with each entry drawn from a Gaussian distribution. Let $g = \Pi f$ be the internal multiplication of the matrix and the frequency vector. Whenever there is an update to a coordinate of f, say the frequency count is changed, we will update g. Note that the update to g is basically adding a vector to f and then g. Then the L_2 norm of g would be bounded by a certain percentage of the frequency vector.

3 How does the AMS algorithm work?

The AMS algorithm can be used to estimate the L_2 norm of a frequency vector also. First, we will generate a random sign vector $s \in \{-1, +1\}$. We then set $W = \langle s, f \rangle$ and output $Z = W^2$ as the estimate for the squared L_2 norm.

Note that

$$\mathbb{E}[Z] = \mathbb{E}[W^2] = \mathbb{E}\left[\sum_{i,j} s_i s_j f_i f_j\right] = \sum f_i^2,$$

since s is a random sign vector. Hence, each dot product is an unbiased estimate of $\|f\|_2^2$. Moreover, the variance of Z is at most

$$\mathbb{E}[Z^2] = \mathbb{E}[W^4] = \mathbb{E}\left[\sum_{a,b,c,d} s_a s_b f_c f_d\right] \leq 6(\sum f_i^2)^2.$$

Hence by Chebyshev's inequality, if we take the mean of $O\left(\frac{1}{\varepsilon^2}\right)$ independent instances, we can obtain a $(1+\varepsilon)$-approximation to $\|f\|_2^2$.