CS369G: Algorithmic Techniques for Big Data Spring 2015-2016

Lecture 9: Sparse Recovery and Threshold Sampling

Prof. Moses Charikar *Scribes: Cong Qiaoben*

1 Overview

Last lecture, we covered count-sketch for estimating the frequencies of the elements. In this lecture we will introduce an application of count-sketch in sparse recovery. Analysis will show that the estimate is quite close to the optimum. Then we will move on to sampling, and introduce threshold sampling as our first example.

2 Sparse Recovery

Recall that using Count-Sketch, we could estimate f_i, s.t. $|\hat{f}_i - f_i| \leq \epsilon\sqrt{F_2}$ with high probability. Sparse recovery only considers the heavy hitters: instead of recovering x from the linear sketch $S(x) = Ax$, it finds z, s.t. $||z||_0 \leq k$ and $||x - z||_p$ is minimized. Notice the minimum error will be $err_p^k(x) := (\sum_{t \notin S} |x_i|^p)^{\frac{1}{p}}$, where S is defined as the set of largest k coordinates of x.

3 Using Count-Sketch for Sparse Recovery

What we did last time was Count-Sketch with $w = \frac{3}{\epsilon^2}$, $d = O(\log n)$ and we proved with high probability $\forall i\ |\hat{x_i} - x_i| \leq \epsilon\sqrt{F_2} = \epsilon \cdot err_2^0(x)$. This time, we will use Count-sketch with $w = \frac{3k}{\epsilon^2}$, $d = O(\log n)$ [1].

Lemma 1. *With high probability,* $\forall i\ |\hat{x_i} - x_i| \leq \frac{\epsilon}{\sqrt{k}} err_2^k(x)$.

Proof. Let $\tilde{x_{ij}}$ be the estimator of x_i from the j-th row of $\{C_{ij}\}$. As was noted in the last lecture, $\tilde{x_{ij}} = C_{jh_j(i)}\sigma_j(i)$, where $\sigma_j(i)$ is the sign of i given by the hash function $\sigma_j : [n] \to \{\pm 1\}$. Let S be the largest k coordinates in x. Let A_i be the event that $\exists i' \in S \backslash i$, s.t. $h_j(i) = h_j(i')$.

$$\Pr\left[|x_i - \tilde{x_{ij}}| \geq \frac{\epsilon}{\sqrt{k}} err_2^k\right] = \Pr[A_i] \Pr\left[|x_i - \tilde{x_{ij}}| \geq \frac{\epsilon}{\sqrt{k}} err_2^k \middle| A_i\right] + Pr[\bar{A_i}] \Pr\left[|x_i - \tilde{x_{ij}}| \geq \frac{\epsilon}{\sqrt{k}} err_2^k \middle| \bar{A_i}\right]$$
$$\leq \Pr[A_i] + \Pr\left[|x_i - \tilde{x_{ij}}| \geq \frac{\epsilon}{\sqrt{k}} err_2^k \middle| \bar{A_i}\right]$$

We first obtain a bound on the first term. The probability of any index $i' \in S \setminus i$ colliding with i

is $\frac{1}{w}$. Applying union bound we get $Pr[A_i] \leq k\frac{1}{w} = \frac{\epsilon^2}{3}$.

$$\Pr\left[|x_i - \tilde{x_{ij}}| \geq \frac{\epsilon}{\sqrt{k}} err_2^k(x) \middle| \bar{A_i}\right] = \Pr\left[|x_i - \tilde{x_{ij}}| \geq \sqrt{\frac{err_2^k(x)^2}{w}} \middle| \bar{A_i}\right] \leq \frac{1}{3}$$

The last step is directly from the last lecture. Therefore:

$$\Pr\left[|x_i - \tilde{x_{ij}}| \geq \frac{\epsilon}{\sqrt{k}} err_2^k(x)\right] \leq \frac{1}{3} + \frac{\epsilon^2}{3}$$

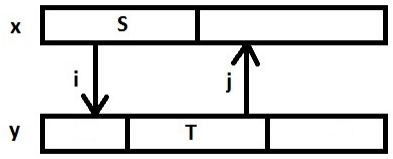
Use the median of $\tilde{x_{ij}}$'s, one can get with high probability:

$$|\hat{x_i} - x_i| \leq \frac{\epsilon}{\sqrt{k}} err_2^k(x)$$

□

Lemma 2. *Given a vector y such that $||x - y||_\infty \leq \frac{\epsilon}{\sqrt{k}} err_2^k(x)$, let T be the set of largest k coordinates in y and set $z := y_T$. Then it holds, $||x - z||_2 \leq (1 + 5\epsilon) error_2^k(x)$.*

Proof. We break down the error $\|x - z\|_2^2$ into three terms depending on whether a coordinate belongs in T, $S \setminus T$ or in the complement of $S \cup T$.



$$||x - z||_2^2 = ||(x - z)_T||^2 + ||x_{S\setminus T}||^2 + ||x_{[n]\setminus(S\cup T)}||^2$$

Let $E = err_2^k(x)$. Since, $||x - y||_\infty \leq \frac{\epsilon}{\sqrt{k}} err_2^k(x)$ we get:

$$||(x - z)_T||^2 \leq k\frac{\epsilon^2}{k}E^2 = \epsilon^2 E^2$$

Furthermore, $|x_i| \leq |y_i| + \frac{\epsilon}{\sqrt{k}}E \ \forall i \in S\setminus T$ and $|x_j| \leq -|y_j| + \frac{\epsilon}{\sqrt{k}}E \ \forall j \in T\setminus S$. Hence,

$$|x_i| - |x_j| \leq |y_i| - |y_j| + \frac{2\epsilon}{\sqrt{k}}E \leq \frac{2\epsilon}{\sqrt{k}}E \ \forall i \in S\setminus T, j \in T\setminus S$$

Let $a = \max_{i \in S \backslash T} |x_i|$, $b = \min_{j \in T \backslash S} |x_j|$. From the above we get $a - b \leq \frac{2\epsilon}{\sqrt{k}} E$. Therefore,

$$\begin{aligned}
||x_{S \backslash T}||_2^2 &\leq a^2 |S \backslash T| \\
&\leq (b + \frac{2\epsilon}{\sqrt{k}} E)^2 |S \backslash T| \\
&\leq (\frac{||X_{T \backslash S}||_2}{\sqrt{|S \backslash T|}} + \frac{2\epsilon}{\sqrt{k}} E)^2 |S \backslash T| \\
&\leq (||X_{T \backslash S}||_2 + 2\epsilon E)^2 \\
&\leq ||X_{T \backslash S}||_2^2 + 4\epsilon E ||X_{T \backslash S}||_2 + 4\epsilon^2 E^2 \\
&\leq ||X_{T \backslash S}||_2^2 + 8\epsilon E^2
\end{aligned}$$

By definition, $||X_{T \backslash S}||_2^2 + ||x_{[n] \backslash (S+T)}||^2 = E^2 \Rightarrow ||x - z||_2^2 \leq (1 + 10\epsilon) E^2$.

$$\Rightarrow ||x - z||_2 \leq (1 + 5\epsilon) E$$

□

Corollary 3. *With high probability, we can construct z, s.t.* $||z||_0 \leq k$, $||x - z||_2 \leq (1 + 5\epsilon) err_2^k(x)$.

4 Sampling and Threshold sampling

Sampling refers to the following problem: given n elements and each element i has weight w_i, one needs to store a sample S s.t. he could answer query Q "What is $\sum_{i \in I} w_i$ for $I \subseteq [n]$". The sequence of w_i's is given in one pass.

4.1 Threshold sampling

Threshold sampling is a special type of sampling. The algorithm is as follows:

If $w_i \leq \tau$:

include in S ($\hat{w}_i = \tau$) with probability $\frac{w_i}{\tau}$.

Otherwise:

include in S ($\hat{w}_i = w_i$).

It is easy to check $E[\hat{w}_i] = w_i$. Therefore, this gives an unbiased estimate of $\sum \hat{w}_i$. However, there are two major drawbacks of threshold sampling. Although the expected size is $\sum_i \min(1, \frac{w_i}{\tau})$, one could rarely have control over the sample size. Another problem is that threshold sampling relies on good τ. If τ is too large or too small, the sample will not be desirable.

References

[1] G. Cormode and S. Muthukrishnan. Combinatorial algorithms for Compressed Sensing. In Proc. 40th Ann. Conf. Information Sciences and Systems, Princeton, Mar. 2006.