CS 15-859: Algorithms for Big Data Fall 2022

Lecture 10 Part 1 — November 17

Prof. David Woodruff *Scribe: Ying Feng*

1 Estimating the Number of Non-Zero Entries

Continuing our discussion on estimating the ℓ_0 norm of a vector:

1.1 Sparse x

Given a sparse vector x, i.e. $|x|_0 \in O(\frac{1}{\epsilon^2})$, its ℓ_0 norm can be found using the algorithm of recovering sparse vector, which outputs the subset of non-zero entries by maintaining the product of the vector with a Vandermonde matrix in the stream.

In the case that the vector is sparse, we can also apply the **Tail Guarantee** of CountSketch to find its ℓ_0 norm:

Tail Guarantee: CountSketch approximates every coordinates x_i of vector x simultaneously up to additive error $O(\frac{\|x_{-B/4}\|_2}{\sqrt{B}})$, with $x_{-B/4}$ being x after zeroing out its top B/4 coordinates in magnitude.

If x is $B/4$-sparse, after zeroing out its top B/4 coordinates produce, we get $\|x_{-B/4}\|_2 = 0$. Thus the tail guarantee says that with high probability, the approximation by CountSketch gives the exact values of the non-zero coordinates of the vector, and from that we can find the estimation of the ℓ_0 norm as desire.

1.2 Dense x

On the other hand, when $|x|_0 \gg O(\frac{1}{\epsilon^2})$, we need sampling for space-efficient estimation. Suppose we have a known estimator Z that provides a range $Z \leq |x|_0 \leq 2Z$, then we can independently sample each coordinate of x with probability $p = \frac{100}{Z\epsilon^2}$ and output the ℓ_0 norm of the sampled coordinates divided by p. We claim that this provides an approximation within a multiplicative factor of $1 \pm \epsilon$ with probability at least $\frac{49}{50}$.

Let Y_i be the indicator variable denoting if the i^{th} coordinate of the vector is sampled. And let y be the vector restricted to the coordinates i for which $Y_i = 1$.

Theorem 1. *$\frac{|y|_0}{p}$ is an $(1 \pm \epsilon)$ estimation of $|x|_0$ with probability at least $\frac{49}{50}$.*

Proof. We have the expectation and variance of the number of sampled non-zero coordinates:

$$\begin{aligned}
E(|y|_0) &= \sum_{i:\text{ s.t.} x_i \neq 0} E(Y_i) \\
&= p \cdot |x|_0 \\
&\geq \frac{100}{\epsilon^2} && (\because p|x|_0 = \frac{100}{z\epsilon^2}|x|_0 \geq \frac{100}{z\epsilon^2} z)
\end{aligned}$$

$$\begin{aligned}
Var(|y|_0) &= \sum_{i:\text{ s.t.} x_i \neq 0} Var(Y_i) && (\because \text{independent sampling}) \\
&\leq \frac{200}{\epsilon^2} && (\because |x|_0 \cdot p(1-p) \leq |x|_0 \cdot p \leq 2z \cdot \frac{100}{z\epsilon^2})
\end{aligned}$$

With those, we can apply the Chebyshev's inequality:

$$\begin{aligned}
Pr(|\ |y|_0 - E(|y|_0)\ | > \frac{100}{\epsilon}) &\leq Var(|y|_0) \cdot \frac{\epsilon^2}{100^2} \\
&\leq \frac{1}{50}
\end{aligned}$$

That is to say with high probability, $|y|_0$ deviates from $E(|y|_0)$ within an ϵ factor. Because $E(|y|_0) = p \cdot |x|_0$, outputting $\frac{|y|_0}{p}$ gives the desired estimation. Note that because y has a small number of non-zero entries, we can use sparse recovery or the tail guarantee of CountSketch as described in **1.1** to compute $|y|_0$ exactly.

■

However, such an estimator Z might not be given. In that case we need to guess Z in powers of 2. Within the possible value range $0 \leq |x|_0 \leq n$, there are $O(logn)$ possible guesses in total. We define the i^{th} guess $Z = 2^i$ to correspond to sampling the coordinates with probability $p = min(1, \frac{100}{2^i\epsilon^2})$.

Because we are guessing in powers of 2, among all guesses there must exist one satisfying the property $Z \leq |x|_0 \leq 2Z$. Therefore we can use the result from that guess as described above. To identify which guess satisfies the property, we run the previous algorithm to calculate the sampled norms $|y|_0$ for each guess.

Theorem 2. *Taking the largest $Z = 2^i$ for which $\frac{400}{\epsilon^2} \leq |y|_0 \leq \frac{3200}{\epsilon^2}$ guarantees $|y|_0 \in (1 \pm \epsilon)E(|y|_0)$ with constant probability.*

Proof. Consider by Chebyshev's inequality:

- $\frac{800}{\epsilon^2} \leq E(|y|_0) \leq \frac{1600}{\epsilon^2} \Rightarrow \frac{400}{\epsilon^2} \leq |y|_0 \leq \frac{3200}{\epsilon^2}$ with probability $1 - O(\epsilon^2)$
- $\frac{100}{\epsilon^2} \leq E(|y|_0) \leq \frac{200}{\epsilon^2} \Rightarrow |y|_0 < \frac{400}{\epsilon^2}$ with probability $1 - O(\epsilon^2)$

Note that if we sample from nested subsets when going down each layer of guesses, then the number of non-zero coordinates in the sample never increase. So the second bullet above also holds for deeper layers. Also, because $E(|y|_0)$ decreases by 1/2 between each pair of adjacent layers, we are guaranteed that a guess Z for which $\frac{400}{\epsilon^2} \leq |y|_0 \leq \frac{3200}{\epsilon^2}$ must exist.

Therefore by choosing the largest $Z = 2^i$ for which $\frac{400}{\epsilon^2} \leq |y|_0 \leq \frac{3200}{\epsilon^2}$, we find a guess such that $\frac{200}{\epsilon^2} \leq E(|y|_0) \leq \frac{1600}{\epsilon^2}$ with probability $1 - O(\epsilon^2)$. With that, following the same reasoning in the proof of Theorem 1, we are able to find $\frac{|y|_0}{p}$ as an $(1 \pm \epsilon)$ estimation of $|x|_0$ with probability at least $\frac{49}{50}$.

Remark 1. There are at most 4 layers of guesses that satisfies $\frac{400}{\epsilon^2} \leq |y|_0 \leq \frac{3200}{\epsilon^2}$. Since the error bound $|y|_0 \in (1 \pm \epsilon)E(|y|_0)$ holds simultaneously for all of them with probability $1 - \frac{4}{50}$ by the union bound, it doesn't matter which one we use.

Overall, the success probability is $1 - O(\epsilon^2) - \frac{4}{50} > \frac{4}{5}$.

■

Complexity:

If we use the k-sparse recovery algorithm with $k = O(\frac{1}{\epsilon^2})$-many hash buckets to compute $|y|_0$ at each layer, that requires $O(\frac{logn}{\epsilon^2})$ space for each layer, thus $O(\frac{log^2 n}{\epsilon})$ space in total.

However, this can be improved by observing that while the data structure recovers not only the non-zero coordinates but also their values, for our purpose of estimating ℓ_0 norm, we only need to count the number of non-zero coordinates.

Claim 1. We are able to reduce the counter size of each bucket from $O(logn)$ to $O(log(\frac{1}{\epsilon} + loglogn))$.

Proof. For any sampling layer, we have $O(\frac{1}{\epsilon^2})$ counters each taking $O(logn)$ bits. Therefore at most $O(\frac{logn}{\epsilon^2})$ prime numbers divide any of those counters. Thus if we choose a random prime $q \in O(\frac{logn \cdot (log(logn) + log(\frac{1}{\epsilon}))}{\epsilon^2})$, it is unlikely that q divides any counter.

That is to say, if we maintain the sparse recovery structure mod q instead of the original structure, it is unlikely that the estimated norm would change due to the modulo operation. So we total reduce the space complexity to $O(log(\frac{1}{\epsilon} + loglogn) \cdot \frac{1}{\epsilon^2} \cdot logn)$.

■

2 Information Theory Concepts

2.1 Discrete Distributions

Consider distributions p over a finite support of size n:

- $p = (p_1, p_2, \ldots, p_n)$
- $p_i \in [0, 1]$ for all i
- $\sum_i p_i = 1$

X is a random variable with distribution p if $Pr(X = i) = p_i$.

2.2 Entropy

Let X be a random variable with distribution p on n items.

Definition. (Entropy): $H(X) = \sum_i p_i \cdot log_2(\frac{1}{p_i})$

- If $p_i = 0$ then $p_i \cdot log_2(\frac{1}{p_i}) = 0$.
- $H(X) \leq log_2 n$. Equality holds when $p_i = \frac{1}{n}$ for all i.
- Entropy measures the "uncertainty" of X.

Example 1. (Binary Input): If B is a bit with bias p, then $H(B) = p \cdot log_2(\frac{1}{p}) + (1-p) \cdot log_2(\frac{1}{1-p})$. And the value of $H(B)$ is symmetric with respect to $p = \frac{1}{2}$.

2.2.1 Conditional and Joint Entropy

Let X and Y be random variables:

Definition. (Conditional Entropy): $H(X \mid Y) = \sum_y H(X \mid Y = y) Pr(Y = y)$

Definition. (Joint Entropy): $H(X, Y) = \sum_{x,y} Pr((X, Y) = (x, y)) \cdot log(1/Pr((X, Y) = (x, y)))$

2.2.2 Chain Rule for Entropy

Theorem 3. $H(X, Y) = H(X) + H(Y \mid X)$

Proof.

$$\begin{aligned}
H(X,Y) &= \sum_{x,y} Pr((X,Y) = (x,y)) \cdot log(\frac{1}{Pr((X,Y) = (x,y)})) \\
&= \sum_{x,y} Pr(X = x) \cdot Pr(Y = y \mid X = x) \cdot log(\frac{1}{Pr(X = x)Pr(Y = y \mid X = x)}) \\
&= \sum_{x,y} Pr(X = x) \cdot Pr(Y = y \mid X = x) \cdot (log(\frac{1}{Pr(X = x)}) + log(\frac{1}{Pr(Y = y \mid X = x)})) \\
&= H(X) + H(Y \mid X)
\end{aligned}$$

∎

With the chain rule for entropy, we can now show that conditioning never increase entropy.

Theorem 4. $H(X \mid Y) \leq H(X)$

Proof. Let f be a continuous, concave function; let $p_1, \ldots, p_n$ be non-negative reals that sum to 1. Then for any $x_1, \ldots, x_n$, by Jensen's inequality we have:

$$\sum_{i=1,\ldots,n} p_i \cdot f(x_i) \leq f(\sum_{i=1,\ldots,n} p_i \cdot x_i)$$

Then for the concave function log,

$$\sum_{x,y} Pr(X=x,Y=y) \cdot log(\frac{Pr(X=x)\cdot Pr(Y=y)}{Pr((X,Y)=(x,y))}) \leq log(\sum_{x,y} Pr(X=x,Y=y) \cdot \frac{Pr(X=x)\cdot Pr(Y=y)}{Pr((X,Y)=(x,y))})$$

Thus we now prove $H(X \mid Y) \leq H(X)$ by writing the joint entropy in two ways

$$\begin{aligned}
H(X \mid Y) - H(x) =& \sum_{x,y} Pr(Y=y) \cdot Pr(X=x \mid Y=y) \cdot log(\frac{1}{Pr(X=x \mid Y=y)}) \\
& - \sum_{x} Pr(X=x) \cdot log(\frac{1}{Pr(X-x)}) \cdot \sum_{y} Pr(Y=y \mid X=x) \\
=& \sum_{x,y} Pr(X=x,Y=y) \cdot log(\frac{Pr(X=x)}{Pr(X=x \mid Y=y)}) \\
=& \sum_{x,y} Pr(X=x,Y=y) \cdot log(\frac{Pr(X=x)\cdot Pr(Y=y)}{Pr((X,Y)=(x,y))}) \\
\leq & log(\sum_{x,y} Pr(X=x,Y=y) \cdot \frac{Pr(X=x)\cdot Pr(Y=y)}{Pr((X,Y)=(x,y))}) \qquad \text{(by the Jensen's Inequality)} \\
=& 0
\end{aligned}$$

Notice that if X and Y are independent then the equality holds. This is because by independence $log(\frac{Pr(X=x)\cdot Pr(Y=y)}{Pr((X,Y)=(x,y))}) = log(\frac{Pr(X=x)\cdot Pr(Y=y)}{Pr(X=x)\cdot Pr(Y=y)}) = log(1)$, thus we don't need the inequality step to derive 0.

■

2.3 Mutual Information

Definition. (Mutual Information):

$$\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) \\
&= H(Y) - H(Y \mid X) \\
&= I(Y;X)
\end{aligned}$$

Remark 2. $I(X;X) = H(X) - H(X \mid X) = H(X)$

Definition. (Conditional Mutual Information):

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z)$$

Remark 3. With the condition, the symmetric proof of mutual information still holds. i.e., $I(X;Y \mid Z) = I(Y;X \mid Z)$.

Conditioning can both decrease or increase the mutual information.

Claim 2. For certain X, Y, Z, we have $I(X;Y \mid Z) \leq I(X;Y)$

Intuitively, we can have Y only reveals information that Z has already revealed, and we condition on Z.

Example 2. Consider $X = Y = Z$:

- $I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z) = 0 - 0 = 0$
- $I(x;y) = H(X) - H(X \mid Y) = H(X) - 0 = H(X)$

Claim 3. For certain X, Y, Z, we have $I(X;Y \mid Z) \geq I(X;Y)$

Intuitively, it is possible that Y only reveals useful information about X after also conditioned on Z.

Example 3. Consider $X = (Y + Z) \bmod 2$:

- $I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z) = 1 - 0 = 1$
- $I(x;y) = H(X) - H(X \mid Y) = 1 - 1 = 0$

2.3.1 Chain Rule for Mutual Information

Theorem 5. $I(X,Y;Z) = I(X;Z) + I(Y;Z \mid X)$

Proof.

$$\begin{aligned}
I(X,Y;Z) &= H(X,Y) - H(X,Y \mid Z) \\
&= H(X) + H(Y \mid X) - H(X \mid Z) - H(Y \mid X, Z) \\
&= (H(X) - H(X \mid Z)) + (H(Y \mid X) - H(Y \mid X, Z)) \\
&= I(X;Z) + I(Y;Z \mid X)
\end{aligned}$$

Similarly, by induction, $I(X_1, \ldots, X_n; Z) = \sum_i I(X_i; Z \mid X_1, \ldots, X_{i-1})$.

■