CS 395T: Sublinear Algorithms, Fall 2020 21st October 2020

Lecture 16: Lower Bounds for Compressed Sensing

Prof. Eric Price *Scribe: Aditya Parulekar, Advait Parulekar*

NOTE: THESE NOTES HAVE NOT BEEN EDITED OR CHECKED FOR CORRECTNESS

1 Overview

In the last lecture we talked about RIP matrices, and the fact that you can make do with $\frac{1}{\varepsilon^2} k \log \frac{n}{k}$ rows.

In this lecture we will show that you can't do better than this.

2 Lower Bound on Compressed Sensing

2.1 $k = 1$

Consider, for simplicity, the $k = 1$ case.

Claim 1. *There exists $\mathcal{X} \subset \mathbb{R}^n$, and some noise distribution $\mathcal{W}$, such that if an algorithm observes Ax' for $x \in \mathcal{X}, w \sim \mathcal{W}, x' = x + w$ and outputs $\hat{x}'$ such that $\|\hat{x}' - x'\|_2 \leq 5 \min_{y \in \mathcal{X}} \|x' - y\|_2$ with probability 0.9, then A must have $\Omega(\log n)$ rows.*

Proof. We take $\mathcal{X} = \{e_i : i \in [n]\}$ as the set of standard basis vectors, and set $\mathcal{W} = \mathcal{N}\left(0, \frac{1}{1000n}\right)$ so we have $\mathbb{E}[\|w\|_2^2] = \frac{1}{1000}$. By Markov's inequality,

$$\mathbb{P}[\|w\|^2 > \frac{1}{100}] \leq \frac{\mathbb{E}[\|w\|_2^2]}{\frac{1}{100}} = \frac{1}{10},$$

and so with 0.9 probability, $\|w\|_2 \leq \frac{1}{10}$. This means that with probability 0.8,

$$\|\hat{x}' - x'\| \leq 5 \min_{y \in \mathcal{X}} \|x' - y\|_2 \leq 5\|x' - x\|_2 = 5\|w\| \leq \frac{1}{2}.$$

Fano's inequality says that, if the number of possible values of x is $|\mathcal{X}|$, and the probability of error is $\mathbb{P}[\text{error}]$,

$$H(x|\hat{x}) \leq H(\mathbb{P}[\text{error}]) + \mathbb{P}[\text{error}] \cdot (\log(|\mathcal{X}|) - 1)$$

which we can weaken to

$$H(x|\hat{x}) \leq 1 + \mathbb{P}[\text{error}] \cdot \log(|\mathcal{X}|)$$

Then, we lower bound the mutual information needed:

$$I(x; \hat{x}) = H(x) - H(x|\hat{x}) \geq \log(|\mathcal{X}|) - \mathbb{P}[\text{error}] \cdot \log(|\mathcal{X}|) - 1 \geq \frac{8}{10} \log n - 1 = \Omega(\log(n))$$

Where we use that $|\mathcal{X}| = n$, since the choices of x are the standard basis vectors. Also $H(x) = \log(|\mathcal{X}|)$, since any of the possible x are equally likely before we see anything. This is just saying that given the noise we assumed, if we are to figure out a random location from 1 to n, then we need roughly $\log(n)$ bits of information.

Now, each step of the process $x \to Ax \to A(x+w) \to \hat{x}$ only depends on the previous step, so the data processing inequality gives $I(x; \hat{x}) \leq I(Ax; Ax + w)$.

Consider $m = 1$, one row. Then, $I(Ax; A(x+w)) = I(\langle a, x\rangle; \langle a, x\rangle + \langle a, w\rangle) = I(a_i; a_i + \langle a, w\rangle)$. We notice that $\langle a, w\rangle$ is just Additive White Gaussian Noise, and so we can use the following theorem to bound this mutual information:

Theorem 2. *(Capacity of Additive White Gaussian Noise channel):*

$$I(a; a+z) \leq \frac{1}{2}\log(1 + SNR) = \frac{1}{2}\log\left(1 + \frac{\mathbb{E}[a^2]}{\mathbb{E}[z^2]}\right)$$

for all distributions a if z is an independent Gaussian

Proof of AWGN Capacity Theorem.

$$\begin{aligned} I(a; a+z) &= H(a+z) - H(a+z|a) \\ &= H(a+z) - H(z) \\ &\leq \frac{1}{2}\ln(2\pi e\mathbb{E}[(a+z)^2]) - \frac{1}{2}\ln(2\pi e\mathbb{E}[z^2]) \\ &= \frac{1}{2}\ln\left(1 + \mathbb{E}[a^2]/\mathbb{E}[z^2]\right) \end{aligned}$$

which follows because entropy of distribution of variance σ^2 is less than entropy of $N(0, \sigma^2)$, and from the entropy of a gaussian. □

In our case, $\langle a, w\rangle \sim N(0, \frac{\|a\|_2^2}{1000n})$. Further, since $\mathbb{E}[a_i^2] = \sum_{i=1}^{n} \mathbb{P}[i] \cdot a_i^2 = \frac{\|a\|_2^2}{n}$, and so

$$I(a_i; a_i + \langle a, w\rangle) \leq \frac{1}{2}\ln\left(1 + \frac{\mathbb{E}[a_i^2]}{\|a\|_2^2/1000n}\right) = \frac{1}{2}\ln(1 + 1000)$$

That is to say, the one measurement only gives a constant amount of information. To bound how much information is in many measurements, we need following lemma:

Lemma 3. *If $y = \bar{y} + w \in R^n$, w_i is independent of all other w_j and $\bar{y}$. Then, $I(y; \bar{y}) \leq \sum_{i=1}^{m} I(y_i; \bar{y}_i)$.*

Proof.

$$\begin{aligned} I(y; \bar{y}) &= h(y) - h(y|\bar{y}) \\ &= h(y) - h(w|\bar{y}) \\ &= h(y) - h(w) \\ &= \sum_{i=1}^{m} h(y_i|y_1, \ldots, y_{i-1}) - h(w_i|w_1, \ldots, w_{i-1}) \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{m} h(y_i) - h(w_i|w_1,\ldots,w_{i-1}) && \text{conditioning decreases entropy} \\
&= \sum_{i=1}^{m} h(y_i) - h(w_i) && h(w_i) = h(w_i|w_1,\ldots,w_{i-1}) \text{ from independence} \\
&= \sum_{i=1}^{m} h(y) - h(y_i|\bar{y}_i) && y_i = \bar{y}_i + w_i \\
&= \sum_{i=1}^{m} I(y_i;\bar{y}_i)
\end{aligned}
$$

□

Using this, we get

$$I(Ax; A(x+w)) \leq \sum_{i=1}^{m} I(a_i; a_i + \langle a, w\rangle) = m\frac{1}{2}\ln 1001$$

Combining this with our lower bound $I(Ax; A(x+w)) = \Omega(\log(n))$, we find that $m = O(\log(n))$. □

2.2 Extension to $k > 1$

Here is a summary of what we just did in the $k = 1$ case:

- Pick distribution over $\mathcal{X}$
- Pick the noise to be i.i.d. Gaussian.
- If $\|w\|_2$ is small, you can correctly recover x. This meant that the mutual information between y and Ax is at least $\log(|\mathcal{X}|)$.
- SNR is small $\implies$ mutual information from each sample is small, and mutual information from all samples is $\log(n)$.

For larger k, we do the same thing, but instead of picking x uniformly over standard basis vectors, pick x uniformly over a "code" $\mathcal{C}$ with the following properties:

1. $|\mathcal{C}| \geq 2^{\Omega(k \log \frac{n}{k})} = \binom{n}{k}^{O(k)}$
2. Good distance: $\|x\|_2 \leq 1 \forall x \in \mathcal{C}$, and $\|x - y\|_2 \geq \frac{1}{4} \forall x \neq y \in \mathcal{C}$.
3. $x \in \mathcal{C}$ is k-sparse
4. $\mathbb{E}[x_i^2] = \frac{1}{n} \forall i$
5. $\mathbb{E}[x_i x_j] = 0$.

Similar to the $k = 1$ case, the second property says that that the noise is typically low enough that an algorithm that correctly recovers $x+w$ can also correctly recover x (specifically, the closest $k-$sparse vector to $x + w$ is x). That is, $\|\hat{x} - (x + w)\|$. The first property is sufficient to show that any algorithm that recovers x with constant probability requires $I(y; Ax) \geq \Omega(k \log \frac{n}{k})$. In particular, from Fano's inequality we similarly have

$$H(x|\hat{x}) \leq 1 + \mathbb{P}[\text{error}] \log |\mathcal{C}| \implies I(y; Ax) \geq I(x; \hat{x}) \geq (1 - \mathbb{P}[\text{error}]) \log |\mathcal{C}| - 1.$$

The second two properties are used to upper bound the mutual information gained per sample.

Let $a \in R^n$ be jth row of A. The capacity of an AWGN channel gives us that

$$I((Ax)_j; (Ax + Aw)_j) = I(\langle a, x\rangle ; \langle a, x\rangle + \langle a, w\rangle) \leq \frac{1}{2} \log \left(1 + \frac{\mathbb{E}[\langle a, x\rangle^2]}{\mathbb{E}[\langle a, w\rangle^2]}\right)$$

The denominator is $\|a\|_2^2/1000n$, just as before. The numerator is

$$\begin{aligned}
\mathbb{E}_x[\langle a, x\rangle^2] &= \mathbb{E}_x[\sum_i a_i^2 x_i^2 + \sum_{i \neq j} a_i a_j x_i x_j] \\
&= \sum_{i=1}^{n} a_i^2 \mathbb{E}[x_i^2] + \sum_{i \neq j} a_i a_j \mathbb{E}_x[x_i x_j] \\
&= \|a\|_2^2/n
\end{aligned}$$

where we plugged in what we know from properties 4. and 5.

Now, all that remains is to find a code that satisfies properties 1-5. For this, we turn to Reed Solomon codes.

2.3 Reed Solomon Codes

The code is generated by evaluating a degree $k - d$ polynomial over F_q on k points. Any two such polynomials can agree on at most $k - d$ points. This means that any two codewords (which are of size k) have to disagree on at least d coordinates, and so have distance d.

This gives codewords in F_q^k. However, we need k-sparse vectors in $\mathbb{R}^n$. To make this transformation, set $q = n/k$. Given a word $z \in F_q^k$, set y to be the concatenation of $e^q_{z_i}$, which denotes the standard basis vector of size q with a 1 at the z_i'th index. Then, set $x = \frac{q}{\sqrt{k}} y$.

Let's show that all of the required properties are satisfied:

1. There are $q^{k-d} = \left(\frac{n}{k}\right)^{k/16}$ polynomials of degree $k - d$ over F_q.
2. $\|x\|_2 = 1$ since we scaled down by the square root of the number of ones. Any two x, x' satisfy $\|x - x'\|_2 \geq \frac{1}{\sqrt{k}} \sqrt{d}$, since they disagree on d. Setting $d = \frac{k}{16}$ gives the desired bound.
3. Clearly, x all k-sparse, since it is made of k standard basis vectors.
4. Each x_i is $1/\sqrt{k}$ w.p. k/n, and so $\mathbb{E}[x_i^2] = \frac{1}{n}$.
5. Give each coordinate a random sign to make $\mathbb{E}[x_i x_j] = 0$.