COMS E6998-9: Algorithms for Massive Data (Spring'19) Feb 21, 2019

Lecture 10: Compressed Sensing: Iterative Hard Thresholding

Instructor: *Alex Andoni* Scribes: *Xi Yang, Kang Sun*

1 Review

Given $x \in \mathbb{R}^n$, think of it as large as a signal of an image, and the main idea to measure it is to use the "measurement matrix" A which greatly reduces dimensions. A matches something from n-dimension to m-dimension where $m \ll n$. If we want to recover the information of x using just a few measurements, we need the structure with k-sparsity. Think of x to be k-sparse, or 'almost'. We can define $Err_1^k(x) = \min_{k\text{-sparse } x' \in \mathbb{R}^n} ||x - x'||_1$.

$y = Ax \in \mathbb{R}^m, m \ll n$,which is the 'measurement'.

$\ell_1\text{-}\min: \min_{x^* \in \mathbb{R}^n} ||x^*||_1 \quad$ s.t. $Ax^* = y$, where x^* is the approximation of x. This problem is in fact a linear programming problem. We can show it in the LP form:

$$\min \sum_{i=1}^{n} t_i$$

$$\text{s.t. } Ax^* = y, -t_i \le x_i^* \le t_i$$

$$\text{in unknown variables } \{x_i^*\}_i, \{t_i\}_i$$

Minimizing the ℓ_1 norm is equivalent to finding the optimal solution to LP problem. t_i will be the absolute value of x_i^*.

Linear programming problem can be solved in polynomial time $n^{O(1)}$. The goal today is to find faster algorithms.

2 Completing the proof from last lecture

Definition 1. *A is (k, ϵ)-RIP if $\forall x \in k$-sparse:*

$$||Ax||_2 = (1 \pm \epsilon)||x||_2$$

We are trying to recover x, which is k-sparse. If x can not be approximated by k-sparse vectors, the error will be sufficiently large.

Theorem 2. *$\exists A$ is (k, ϵ)-RIP for $m = O(k \log \frac{n}{k})$*

Note that a matrix A, with a weaker bound on m, already follows from existence of OSE matrices (as proved last lecture); existence of good OSE matrices follows by the problem on problem set 2.

Theorem 3. *If A is $(4k, \epsilon)$-RIP, then the x^* from $\ell_1\text{-}\min$ satisfies:*

$$||x - x^*||_1 \le C \cdot Err_1^k(x)$$

where C is a constant.

x^* is the optimal solution from ℓ_1-min. $||x - x^*||_1$ is vector difference meansured in ℓ_1 norm. This ℓ_1 norm is as good as whatever the best k-sparse vector can recover in ℓ_1 norm. So this is called ℓ_1/ℓ_1 guarantee.

We can use ℓ_2/ℓ_1 guarantee to get a tighter bound.

$$||x - x^*||_2 \le C \cdot \frac{Err_1^k(x)}{\sqrt{k}}$$

Note: ℓ_2/ℓ_1 guarantee $\Rightarrow$ ℓ_1/ℓ_1 guarantee.

Definition 4. *A satisfies (k, ϵ) - null space property if $\forall \eta \in \mathbb{R}^n$, $T \subset \{n\}$ of size k*

$$\begin{aligned} A\eta = 0 \Rightarrow ||\eta||_1 &\le (1+\epsilon)||\eta_{-T}||_1 \\ \Leftrightarrow ||\eta_T||_1 &\le \epsilon \cdot ||\eta_{-T}||_1 \end{aligned}$$

Lemma 5. *If A is $((2+r)k, \epsilon)$ - RIP $\Rightarrow$ A satisfies $(2k, \sqrt{2/r} \cdot \frac{1+\epsilon}{1-\epsilon})$ - null space property.*

For the proof of this lemma, see the link on the class website.

Lemma 6. *If A satisfies $(2k, \epsilon)$ - null space property, for $\epsilon < 1/2$, then:*

$$||x - x^*||_1 \le 2\,\frac{1+\epsilon}{1-\epsilon}\,Err_1^k(x)$$

Proof. Let $\eta = x - x^*$, i.e. "residual" vector. We want $||\eta||_1$ to be small.

$A\eta = Ax - Ax^* = 0$ since x^* satisfies l_1-min $\Rightarrow ||\eta_T||_1 \le \epsilon \cdot ||\eta_{-T}||_1$, for $T = k$ largest entries of x.

Since x^* is the optimal solution of ℓ_1-min problem,

$$\begin{aligned} \Rightarrow ||x^*||_1 \le ||x||_1 \Leftrightarrow ||x^*_T||_1 + ||x^*_{-T}||_1 &\le ||x_T||_1 + ||x_{-T}||_1 \\ \text{where } ||x^*_T||_1 = ||x_T - (x_T - x^*_T)||_1 &\ge ||x_T||_1 - ||\eta_T||_1 \\ \text{and } ||x^*_{-T}|| = ||x_{-T} - x^*_{-T} - x_{-T}||_1 &\ge ||\eta_{-T}||_1 - ||x_{-T}||_1 \end{aligned}$$

$$\begin{aligned} \text{So } ||x^*_T||_1 + ||x^*_{-T}||_1 &\le ||x_T||_1 + ||x_{-T}||_1 \\ \Rightarrow ||x_T||_1 - ||\eta_T||_1 + ||\eta_{-T}||_1 - ||x_{-T}||_1 &\le ||x_T||_1 + ||x_{-T}||_1 \end{aligned}$$

Notice that $||x_T||_1$ is the recovered signal and it is canceled in both sides. $||x_{-T}||_1$ is $Err_1^k(x)$. So we can get

$$\begin{aligned} &||\eta_{-T}||_1 \le ||\eta_T||_1 + 2 \cdot Err_1^k(x) \le \epsilon \cdot ||\eta_{-T}||_1 + 2 \cdot Err_1^k(x) \\ &\Rightarrow ||\eta_{-T}||_1 \le \frac{2}{1-\epsilon} Err_1^k(x) \end{aligned}$$

Relating $||\eta||_1$ to $||\eta_{-T}||_1$, we get

$$||\eta||_1 \leq (1+\epsilon) \cdot \frac{2}{1-\epsilon} Err_1^k(x) = 2\frac{1+\epsilon}{1-\epsilon} Err_1^k(x)$$

$\square$

3 Iterative Hard Thresholding

Note: In the section, we adopt the notation of $\perp$ for matrices and vectors as transpose symbol, mainly because transpose T would conflict with the iteration number T.

So far, we have been approaching Compressed Sensing via $\ell_1 - \min$, which is also refered to as "Basis Pursuit". $\ell_1 - \min$ has a time complexity of $n^{O(1)}$, and our hope is to reduce the time complexity to roughlythe size of matrix A, which is $O(nm)$.

Here we introduce the Iterative Hard Thresholding algorithm (1). The main idea of this algorithm is to try to determine which coordinates matter when producing k-sparse vector. Note that the function

$$P_k(z) = \underset{z':k-\text{sparse}\in\mathbb{R}^n}{\arg\min} ||z - z'||_1$$

picks the k most significant coordinates from vector z and set the values of remaining coordinates to 0.

Algorithm 1 Iterative Hard Thresholding

Require: $y\ (= Ax), T$
Ensure: x^{T+1}, which is a k-sparse approximation of x
  $x^1 \leftarrow 0^n$
  for $t \leftarrow 1$ to T **do**
    $x^{t+1} \leftarrow P_k(x^t + A^\perp(y - Ax^t))$
  end for
  return x^{T+1}

Here's the intuition why we use $P_k(z)$. If we define $a^{t+1} = x^t + A^\perp(y - Ax^t)$, we can see although x^t is guaranteed to be k-sparse (by induction), $A^\perp(y - Ax^t)$ could introduce nonzero values on other coordinates. By using $P_k(z)$ function, we can restore the k-sparsity and project a^{t+1} to x^{t+1}.

In a sense, we can see an analogy between IHT algorithm and projection methods in constrained optimization. In constrained optimization problem, we start with a feasible solution and find another solution that yields slightly better objective function value; if that solution violates the constraints, we project it back to the allowed region defined by the constraints.

Now we proceed to proving IHT algorithm works.

Theorem 7. ***[Blumensath-Davies '09]***
A is $(3k, \epsilon)$-RIP matrix, where $\epsilon < \frac{1}{8}$. Let $y = Ax + e$ (e as the error term), then $\forall T \geq 1$, IHT iterate x^{T+1} satisfies:

$$||x^{T+1} - x||_2 \leq O(1) \cdot \left[2^{-T} \cdot ||x||_2 + \frac{Err_1^k(x)}{\sqrt{k}} + ||e||_2\right]$$

Today we prove a simpler version of the theorem, where we assume zero error and the x is precisely k-sparse.

Theorem 8. *Following the notations from Theorem 7, suppose $x \in \mathbb{R}^n$ is exactly k-sparse, $e = 0$, then*

$$||x^{T+1} - x||_2 \leq O(2^{-T}) \cdot ||x||_2$$

Proof.

Observation 9. *Fix k-sparse vector $z \in \mathbb{R}^n$, $||Az||_2^2 \in (1 \pm \epsilon)||z||_2^2$, which means $z^\perp A^\perp Az \in (1 \pm \epsilon) z^T z \implies z^\perp(A^\perp A - I)z \in (-\epsilon, \epsilon) \cdot \|z\|^2$. We can see $A^\perp A \approx I$ when operating on k-sparse vector z.*

We first define $r^t = x - x^t$ to be the residual, our goal is to prove $||r^t||_2$ decreases exponentially when t iterates from 1 to T.

We have $a^{t+1} = x^t + A^\perp(y - Ax^t)$ (note the second term might not be k-sparse). Intuitively, $a^{t+1} = x^t + A^\perp Ar^t \overset{\text{Observation 9}}{\approx} x^t + r^t = x$.

Define $B^t = \text{support}(x) \cup \text{support}(x^t)$. Sicne both x and x^t are k-sparse vectors, $|B^t| \leq 2k$.

Let $B = B^{t+1}$ and $B^- = B^t$.

$$\begin{aligned}
||r^{t+1}||_2 &= ||x - x^{t+1}||_2 \\
&= ||x_B - x_B^{t+1}||_2 \\
&\overset{\text{triangle ineq}}{\leq} ||x_B - a_B||_2 + ||a_B^{t+1} - x_B^{t+1}||_2
\end{aligned}$$

Since by definition $x_B^{t+1} = \arg\min_{x^{t+1}:k\text{-sparse}} ||a^{t+1} - x^{t+1}||_2$, we have $||a_B^{t+1} - x_B^{t+1}||_2 \leq ||x_B - a_B^{t+1}||_2$. Therefore $||r^{t+1}||_2 \leq 2||x_B - a_B^{t+1}||_2$.

$a_B^{t+1} = [x^t + A^T Ar^t]_B = x_B^t + (A^\perp Ar^t)_B = x_B^t + A_B^\perp Ar^t$. Here we define A_B to be the matrix where all columns not in set B zeroed out while entries with columns in B have the same value as in A. Then we can continue work on

$$\begin{aligned}
||x_B - a_B^{t+1}||_2 &= ||x_B - x_B^t - A_B^\perp Ar^t||_2 \\
&= ||r_B^t - A_B^\perp Ar^t||_2 \\
&= ||r_B^t - A_B^\perp A_B r_B^t - A_B^\perp Ar_{-B}^t||_2 \\
&\overset{\text{triangle ineq}}{\leq} ||(I - A_B^\perp A_B)r_B^t||_2 + ||A_B^\perp Ar_{-B}^t||_2 \\
&\overset{\text{Observation 9}}{\leq} \epsilon||r_B^t||_2 + ||A_B^\perp Ar_{-B}^t||_2 \\
&= \epsilon||r_B^t||_2 + ||A_B^\perp A_{B^-\setminus B} r^t||_2
\end{aligned}$$

note the last equation comes from the fact that we have $r_{-B}^t = r_{B^-\setminus B}^t$ and $Ar_{B^-\setminus B}^t = A_{B^-\setminus B} r^t$.

Claim 10.

$$||A_B^\perp A_{B^-\setminus B}||_2 \leq 2\epsilon$$

Not done during the lecture. Let $C = B^- \setminus B$. Consider unit-norm p_B, supported on B, and unit-norm q_C, supported on C. Then $||A_B^\perp A_{B^-\setminus B}||_2 = \max_{p_B, q_C} |p_B A^\perp A q_C|$.

Consider the quantity $\|p_B - q_C\|^2 = 2$. Then $\|A(p_B - q_C)\|^2 \geq 2 - 2\epsilon$ by RIP property of A. At the same time, $\|A(p_B - q_C)\|^2 = \|Ap_B\|^2 + \|Aq_C\|^2 - 2p_B A^{\perp} Aq_C \leq 2 + 2\epsilon - 2p_b A^{\perp} Aq_C$ (again, by RIP). Hence, $p_B A^{\perp} Aq_C \leq 2\epsilon$ and the claim follows. □

Therefore we get

$$\begin{aligned} ||r_B^{t+1}||_2 &\leq 2||x_B - a_B^{t+1}||_2 \\ &\leq 2(\epsilon + ||A_B^{\perp} A_{B-\backslash B}||_2) \cdot ||r^t||_2 \\ &\stackrel{\text{Claim 10}}{\leq} 6\epsilon ||r^t||_2 \\ &\leq \frac{1}{2} ||r^t||_2 \end{aligned}$$

so

$$||r^t||_2 \leq O(2^{-t}) \cdot ||r^1||_2 = O(2^{-t})||x||_2$$

□

Theorem 7 and Theorem 8 tell us that the error of the k-sparse approximation x^t decreases exponentially with T.