Sketching Algorithms for Big Data	Fall 2017

Lecture 14 — October 19, 2017

Prof. Jelani Nelson — *Scribe: Ali Vakilian*

1 Overview

In the last lecture we started the topic of *Compressive Sensing* and in particular we described the *Basis Pursuit (BP)* algorithm. In compressive sensing, given measurement matrix $\Pi \in \mathbb{R}^{m\times n}$ and measurement vector $y = \Pi x$, the goal is to recover the vector x which is known to be either exactly or nearly k-sparse.

Basis Pursuit(Π, y):

Min $||z||_1$
s.t. $\Pi z = y$

Remark 1. *More generally, we can consider the case in which there is some* post measurement *noise e such that $||e||_2 \leq \alpha$. Then, we can adjust the linear program as follows:*

Basis Pursuit(Π, y, α):

Min $||z||_1$
s.t. $||\Pi z - y||_2 \leq \alpha$

The main result we proved in the last lecture is the following.

Theorem 2. *If $\hat{x}$ is output of* **Basis Pursuit**(Π, y) *and Π satisfies (ε, Ck)-***RIP** *for sufficiently small constant $\varepsilon > 0$, and sufficiently large constant $C > 1$, then*

$$||\hat{x} - x||_2 \leq O(\frac{1}{\sqrt{k}}).||x_{\mathsf{tail}(k)}||_1 \tag{1}$$

Corollary 3. *If x is actually k-sparse, the is no error in the output of the recovery;* **Basis Pursuit**(Π, y) *returns x.*

Though the **BP** works with a single measurement matrix Π that works for (recovering) all nearly k-sparse vectors x, it is not fast enough. The reason is that solving LP generally requires polynomial time in n and is not very fast.

In this lecture we describe an *iterative fast* approach for the sparse recovery task that has a running time which is nearly linear (if the measurement matrix supports nearly linear time matrix-vector multiplication). This approach was first used by Needell and Tropp [NT08] (CoSAMP). The algorithm we cover here is called **Iterative Hard Thresholding (IHT)** and it is due to Blumensath and Davies [BD09]

2 Iterative Hard Thresholding (IHT) for Compressed Sensing

Roughly speaking, the algorithm starts with some guess on vector x (which is the all zero vector) and goes through T iterations of updating the vector. The goal is to show that by these updates, the sequence converges to the *true* x. More formally, assuming $x^{[1]}, \cdots, x^{[T]}$ are the vectors produced over the T iterations, here is the main theorem of **IHT**:

Theorem 4 ([BD09])**.** *If* Π *satisfies* $(\varepsilon, 3k)$*-RIP for* $\varepsilon < \frac{1}{4\sqrt{2}}$*, then* $\forall\ T \geq 1$

$$||x^{[T+1]} - x||_2 \lesssim 2^{-T}||x||_2 + ||x_{\mathsf{tail}(k)}||_2 + \frac{1}{\sqrt{k}}||x_{\mathsf{tail}(k)}||_1 + ||e||_2 \tag{2}$$

Comparing to the guarantee of **BP** approach (Theorem 2), in (2) we have three extra terms: $2^{-T}||x||_2$, $||x_{\mathsf{tail}(k)}||_2$ and $||e||_2$. Note that the last term corresponds to the post-measurement noise and it is unavoidable. For the second term, $||x_{\mathsf{tail}(k)}||_2$, we shortly shows that it is dominated by $||x_{\mathsf{tail}(k)}||_1/\sqrt{k}$. Hence, the only difference is the exponentially decaying term $2^{-T}||x||_2$. In turn, the **IHT** algorithm is much faster than **BP**.

Claim 5. $||x_{\mathsf{tail}(2k)}||_2 \leq \frac{1}{\sqrt{k}}||x_{\mathsf{tail}(k)}||_1$.

Proof. (shelling method) WLOG, let us assume that the coordinate of x are sorted in a decreasing order of their absolute values: $|x_1| \geq |x_2| \geq \cdots \geq |x_n|$. Moreover, we partition the coordinates of x into blocks of size k as follows: $B_1, \cdots, B_{n/k}$.

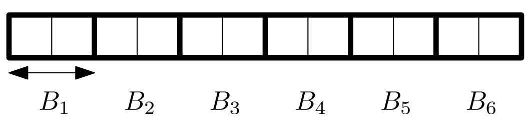

Figure 1: In this example, $k = 2$ and $n = 12$.

Now, we apply the shelling method. Since coordinates of x are sorted by their absolute values, for each coordinate $j \in B_t$, $|x_j| \leq \frac{1}{k}\sum_{i \in B_{t-1}} |x_i| = \frac{1}{k}||x_{B_{t-1}}||_1$.

$$||x_{\mathsf{tail}(2k)}||_2^2 = \sum_{t=3}^{n/k} ||x_{B_t}||_2^2 \leq \sum_{t=3}^{n/k} k \cdot (\frac{||x_{B_{t-1}}||_1}{k})^2 = \frac{1}{k}\sum_{t=2}^{n/k} ||x_{B_t}||_1^2$$

Finally, using the fact that for positive values $A_1, \cdots, A_\ell$, $\sqrt{A_1 + \cdots + A_\ell} \leq \sqrt{A_1} + \cdots + \sqrt{A_\ell}$:

$$||x_{\mathsf{tail}(2k)}||_2 \leq \frac{1}{k} \cdot \sqrt{\sum_{t=2}^{n/k} ||x_{B_t}||_1^2} \leq \frac{1}{\sqrt{k}}||x_{\mathsf{tail}(k)}||_1$$

$\square$

Now lets focus on the proof of the convergence of **IHT** algorithm (proof of Theorem 4). Note that, in the analysis we can assume that x is *exactly* k-sparse. More precisely, we can include the $\mathsf{tail}_k(x)$

term in the noise term and denote the new noise as $\tilde{e}$.

$$\Pi x + e = \Pi(x_{\mathsf{head}(k)} + x_{\mathsf{tail}(k)}) + e = \Pi x_{\mathsf{head}(k)} + \underbrace{(\Pi x_{\mathsf{tail}(k)} + e)}_{\tilde{e}} \tag{3}$$

Setting $\tilde{e} = \Pi x_{\mathsf{tail}(k)} + e$, then we have $||\tilde{e}||_2$ in the error term which is less than:

$$\begin{aligned} ||\tilde{e}||_2 \underset{\triangle\text{-inequality}}{\leq} ||e||_2 + ||\Pi x_{\mathsf{tail}(k)}||_2 = ||e||_2 + ||\sum_{t=2} \Pi x_{B_t}||_2 &\leq ||e||_2 + \sum_{t=2} ||\Pi x_{B_t}||_2 \\ &\overset{\textbf{RIP}}{\leq} ||e||_2 + (1+\varepsilon)\sum_{t=2} ||x_{B_t}||_2 \\ &\leq ||e||_2 + \frac{1+\varepsilon}{\sqrt{k}} ||x_{\mathsf{tail}(k)}||_1 \end{aligned}$$

Hence, it does not change the performance guarantee of **IHT** by more than an ε-factor. In the rest of this section, we assume that the input vector x is k-sparse.

Algorithm 1 Iterative Hard Thresholding (IHT).

```
1: function IHT(Π, y(= Πx + e), k, T)
2:     x^[1] ← 0
3:     for t = 1 ··· T do
4:         x^{t+1} ← H_k(x^[t] + Π^⊤(y − Πx^[t]))   ▷ Hard thresholding operator (project a^[t+1] on x_head(k))
5:     end for
6:     return x^{T+1}
7: end function
```

The formal definition of H_k operator is as follows: $H_k(z) := \underset{k\text{-sparse } \hat{z}}{\mathsf{argmin}} ||z - \hat{z}||_2$ which is the projection on $\mathsf{head}(k)$ coordinates of z.

Proof sketch of Theorem 4. We measure the progress of **IHT** algorithm based on the residual vector $r^{[t]} := x - x^{[t]}$. The hope is to show that r decreases at some rate. For analysis purpose, we define $a^{[t+1]} := x^{[t]} + \Pi^\top(y - \Pi x^{[t]})$ (note that $x^{[t+1]} = H_k(a^{[t+1]})$).

$$\begin{aligned} a^{[t+1]} = x^{[t]} + \Pi^\top(y - \Pi x^{[t]}) &= x^{[t]} + \Pi^\top(\Pi x + e - \Pi x^{[t]}) \\ &= x^{[t]} + \underbrace{\Pi^\top(\Pi}_{\approx \mathbf{I}} r^{[t]} + e) \approx x^{[t]} + r^{[t]} + \Pi^T e \approx x^{[t]} + r^{[t]} + e. \end{aligned}$$

Intuitively, assuming $r^{[t]}$ is decaying, $a^{[t]}$ converges to x. The role of *hard threshold operator* H_k is to make sure that all vectors are sparse so that Π behaves well on them.

Notation. To analyze the **IHT** algorithm, we setup the following notations:

- $\Gamma_k^* = \mathsf{supp}(x)$,
- $\Gamma^{[t]} = \mathsf{supp}(x^{[t]})$, and
- $B^{[t]} = \Gamma_k^* \cup \Gamma^{[t]}$.

As we mentioned, the goal is to bound the residual vector $r^{[t+1]}$. In particular, we need to show that $r^{[t+1]}$ is decaying.

$$\begin{aligned}||r^{[t+1]}||_2 &= ||x - x^{[t+1]}||_2 = ||x_{B^{[t+1]}} - x^{[t+1]}_{B^{[t+1]}}||_2 \\ &\underset{\triangle\text{-ineq}}{\leq} \underbrace{||x_{B^{[t+1]}} - a^{[t+1]}_{B^{[t+1]}}||_2}_{I} + \underbrace{||a^{[t+1]}_{B^{[t+1]}} - x^{[t+1]}_{B^{[t+1]}}||_2}_{II}\end{aligned}$$

Claim 6. *$||a^{[t+1]}_{B^{[t+1]}} - x^{[t+1]}_{B^{[t+1]}}||_2 \leq ||x_{B^{[t+1]}} - a^{[t+1]}_{B^{[t+1]}}||_2$ (or $II \leq I$).*

Proof. By definition of *hard threshold operator* H_k, $x^{[t+1]}$ is the best k-sparse approximate of $a^{[t+1]}$. Since x is also a k-sparse vector, $II \leq I$. □

For brevity, in the rest of proof, we use B to denote $B^{[t+1]}$ and B' to denote $B^{[t]}$.

$$\begin{aligned}||r^{[t+1]}||_2 &= ||x - x^{[t+1]}||_2 = ||x_B - x^{[t+1]}_B||_2 \\ &\underset{\triangle\text{-ineq}}{\leq} ||x_B - a^{[t+1]}_B||_2 + ||a^{[t+1]}_B - x^{[t+1]}_B||_2 \\ &\underset{\text{Claim 6}}{\leq} 2||x_B - a_B||_2 = 2||\underbrace{x_B - x^{[t]}_B}_{r^{[t]}} - \Pi_B^\top(y - \Pi x^{[t]})||_2 \end{aligned} \tag{4}$$

Note that Π_B is equal to Π but columns in $\bar{B}$ are zero out. Next, by expanding y, we have:

$$\begin{aligned}&\overset{(4)}{=} 2||r^{[t]}_B - \Pi_B^\top(\Pi r^{[t]} + e)||_2 \quad (\text{write } r^{[t]} = r^{[t]}_B + r^{[t]}_{B'\setminus B}) \\ &= 2||\underbrace{r^{[t]}_B}_{\mathbf{I}_B r^{[t]}} - \Pi_B^\top \underbrace{\Pi r^{[t]}_B}_{\Pi_B r^{[t]}_B} - \Pi_B^\top \underbrace{\Pi r^{[t]}_{B'\setminus B}}_{\Pi_{B'\setminus B} r^{[t]}_{B'\setminus B}} - \Pi_B^\top e||_2 \\ &= 2||(\mathbf{I}_B - \Pi_B^\top \Pi_B) r^{[t]}_B - \Pi_B^\top \Pi_{B'\setminus B} r^{[t]}_{B'\setminus B} - \Pi_B^\top e||_2 \\ &\underset{\triangle\text{-ineq}}{\leq} 2[||\mathbf{I}_B - \Pi_B^\top \Pi_B|| \cdot ||r^{[t]}_B||_2 && (5) \\ &\quad + ||\Pi_B^\top \Pi_{B'\setminus B}|| \cdot ||r^{[t]}_{B'\setminus B}||_2 && (6) \\ &\quad + ||\Pi_B|| \cdot ||e||_2] && (7)\end{aligned}$$

By the following claims, we upper bound terms (5), (6) and (7).

Claim 7. *$||\mathbf{I}_B - \Pi_B^\top \Pi_B|| \leq \varepsilon$.*

Proof. Π is an ε-subspace embedding (ε-s.e.) for colspan(U) if $||(\Pi U)^\top \Pi U - \mathbf{I}|| \leq \varepsilon$. Since the measurement matrix Π is $(\varepsilon, 3k)-$RIP, it is ε-s.e. for all $\binom{n}{k}$ k-dim subspaces (For more details refer to Definition 4 in Lecture 11). □

Claim 8. *$||\Pi_B^\top \Pi_{B'\setminus B}|| \leq \varepsilon$.*

Proof. By definition of operator norm, $||\Pi_B^\top \Pi_{\underbrace{B' \setminus B}_{D}}|| = \sup_{||a||,||s||=1} \langle \Pi_B a_B, \Pi_D s_D \rangle = \sup_{||a||,||s||=1} \langle \Pi a_B, \Pi s_D \rangle$. Since Π satisfies **JL** property, it preserves the dot product. Moreover, since $D \cap B = \emptyset$, $\langle a_B, s_D \rangle = 0$; hence, $\langle \Pi a_B, \Pi s_D \rangle \leq \varepsilon$ (note that $a_B + s_D$ and $a_B - s_D$ are $3k$-sparse and Π is a $(\varepsilon, 3)$-RIP matrix). ☐

Claim 9. $||\Pi_B^\top|| = ||\Pi_B|| \leq \sqrt{1+\varepsilon}$.

Proof. Note that Π satisfies **JL** properties and in particular preserves the ℓ_2 norm. Then,

$$||\Pi_B|| = \sup_{||a||_2=1} ||\Pi_B a_B||_2 \underset{\textbf{JL}\text{ property}}{\leq} \sqrt{1+\varepsilon}||a_B||_2 = \sqrt{1+\varepsilon}.$$

☐

Then, using above three claims, we bound $r^{[t+1]}$ as follows:

$$\begin{aligned}
||r^{[t+1]}||_2 &\leq 2[||\mathbf{I}_B - \Pi_B^\top \Pi_B|| \cdot ||r_B^{[t]}||_2 + ||\Pi_B^\top \Pi_{B' \setminus B}|| \cdot ||r_{B' \setminus B}^{[t]}||_2 + ||\Pi_B|| \cdot ||e||_2] \\
&\leq 2\varepsilon(\underbrace{||r_B^{[t]}||_2 + ||r_{B' \setminus B}^{[t]}||_2}_{\text{by Claim 11}}) + 3||e||_2 \qquad \text{By Claims 7, 8 and 9} \\
&\leq 2\sqrt{2}\varepsilon||r^{[t]}||_2 + 3||e||_2 \qquad \text{For sufficiently small } \varepsilon \\
&\leq \frac{1}{2}||r^{[t]}||_2 + 3||e||_2
\end{aligned} \tag{8}$$

Corollary 10. $||r^{[T+1]}||_2 \leq 2^{-T}||x||_2 + 6||e||_2$

Proof. Using (8) and by induction,

$$\begin{aligned}
||r^{[T+1]}||_2 &\leq \frac{1}{2^T}||r^{[1]}||_2 + 3(1 + 1/2 + \cdots 1/2^T)||e||_2 \\
&\leq 2^{-T}||x||_2 + 6||e||_2
\end{aligned}$$

☐

Claim 11. $||r_B^{[t]}||_2 + ||r_{B' \setminus B}^{[t]}||_2 \leq \sqrt{2} \cdot ||r^{[t]}||_2$.

Proof. Define $z = r_{B \cup B'}$, $x = r_B$ and $y = r_{B'}$. Then, $||z||_2^2 = ||x||_2^2 + ||y||_2^2$. By (**AM-GM**) inequality,

$$\sqrt{||x||_2^2} + \sqrt{||y||_2^2} \leq \sqrt{2}\sqrt{||x||_2^2 + ||y||_2^2} \leq \sqrt{2}\sqrt{||z||_2^2}.$$

☐

3 Model Based Compressed Sensing

In standard *compressed sensing*, the assumption is that x is an approximately k-sparse vector. This implies that there exists $S \in \Omega_{n,k}$ such that $||x - x_S||_2$ is small where $\Omega_{n,k} = \binom{[n]}{k}$. Then, to do k-sparse recovery, enough for Π to be ε-subspace embedding for all k-dim coordinates indexed by $\Omega_{n,k}$. This led Π to have $\frac{k+\lg|\Omega_{n,k}|}{\varepsilon^2} = \lg(1/\delta)\varepsilon^2$ ($\delta \ll \frac{1}{C^k|\Omega_{n,k}|}$). Note that k is required for preserving a single k-dim subspace and the second term is for preserving all k-dim coordinates subspaces in $\Omega_{n,k}$. But, what if we know more about the structure of x? This leads to the *model based compressed sensing*.

In model based compressed sensing, $\Omega_{n,k}$ will be replaced by $\mathcal{M}$ and then it only required to blow up the number of rows in Π by a factor of $\lg(\mathcal{M})$ which can be much smaller than $k\lg(n/k)$ ($\lg|\Omega_{n,k}|$).

The model based **RIP** studied by Baraniuk et al. [BCDH10]. Using model based **RIP**, we can adopt the **IHT** algorithm slightly to obtain *model based* **IHT**. It only suffices to instead of projecting on $\Omega_{n,k}$ (using H_k operator), in each iteration project $x^{[t]}$ to $\mathcal{M}$ via $P_{\mathcal{M}}$ operator: $P_{\mathcal{M}}(z) := \underset{\hat{z}\in\mathcal{M}}{\mathsf{argmin}}||z - \hat{z}||_2$.

Algorithm 2 Model Based Iterative Hard Thresholding (MB-IHT).

```
function IHT(Π, y(= Πx + e), k, T)
    x^[1] ← 0
    for t = 1 ··· T do
        x^[t+1] ← P_M(x^[t] + Π^T(y − Πx^[t]))    ▷ Hard thresholding operator (project a^[t+1] on M)
    end for
    return x^(T+1)
end function
```

Similarly to the standard compressed sensing in which Π is required to be $(\varepsilon, 3k)$-**RIP**, in the model based compressed sensing, we need the measurement matrix Π to be RIP for $\mathcal{M}^3 = \{A \cup B \cup C | A, B, C \in \mathcal{M}\}$ (to show similar results to those in Claim 7, 8 and 9).

This approach (model based compressed sensing) improves the guarantees of the standard compressed sensing for signals with structured sparsity such as wavelet and block models [BCDH10] and tree sparsity [HIS15, HIS14a, HIS14b, BIS17].

References

[BCDH10] Richard G. Baraniuk, Volkan Cevher, Marco F. Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2010.

[BD09] Thomas Blumensath and Mike E. Davies. A simple, efficient and near optimal algorithm for compressed sensing. In *ICASSP*, 2009.

[BIS17] Arturs Backurs, Piotr Indyk, and Ludwig Schmidt. Better approximations for tree sparsity in nearly-linear time. In *SODA*, pages 2215–2229, 2017.

[HIS14a] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. A fast approximation algorithm for tree-sparse recovery. In *ISIT*, pages 1842–1846, 2014.

[HIS14b] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. Nearly linear-time model-based compressive sensing. In *ICALP*, pages 588–599, 2014.

[HIS15] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. Approximation algorithms for model-based compressive sensing. *IEEE Transactions on Information Theory*, 2015.

[NT08] Deanna Needell and Joel A. Tropp. CoSAMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 2008.