CS 395T: Sublinear Algorithms, Fall 2020 Oct. 13, 2020

Lecture 14: Iterative Hard Thresholding

Prof. Eric Price *Scribe: Hongru Yang*

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

Theorem 0.0.1. *If A satisfies $(O(k), \epsilon(= 0.001))$-RIP, then we can perform compressed sensing: Given $y = Ax + e$ and $\|x\|_0 \leqslant k$, recover $\hat{x}$ such that*

$$\|\hat{x} - x\|_2 \leqslant O(\|e\|_2)$$

therefore, we can get an upper bound if x is not k-sparse by

$$x = x_k + (x - x_k)$$
$$Ax = Ax_k + \underbrace{A(x - x_k)}_{e}$$

1 Iterative Hard Thresholding (IHT)

1.1 The idea

Given a matrix A satisfies the RIP, then for all $O(k)$-sparse x, we have

$$\begin{aligned}
&\Leftrightarrow \|Ax\|_2^2 = (1 \pm \epsilon)\|x\|_2^2 \\
&\Leftrightarrow x^T A^T A x = (1 \pm \epsilon) x^T x \\
&\Leftrightarrow |x^T (A^T A - I) x| \leqslant \epsilon x^T x \\
&\Leftrightarrow \|(A^T A - I)_{S \times S}\|_2 \leqslant \epsilon, \ \forall S \subseteq [n] : |S| \leqslant O(k)
\end{aligned}$$

Given y where $y = Ax + e$, we want to estimate x.

Idea: let

$$z = A^T y = \underbrace{A^T A}_{\approx I \text{ on sparse sets}} x + \underbrace{A^T e}_{\text{small if } e \text{ is small}}$$

Then

$$z - x = (A^T A - I)x + A^T e$$

Let $S \subseteq [n]$ such that $|S| \leqslant O(k)$ and $S \supseteq \operatorname{supp}(x)$. Then

$$\|(z - x)_S\|_2 \leqslant \|[(A^A - I)x]_S\|_2 + \|[A^T e]_S\|_2$$

Now we need to bound the two terms on the right-hand side.

$$\begin{aligned}\left\|[(A^TA-I)x]_S\right\|_2 &= \left\|(A^TA-I)_{S\times S}x_S\right\|_2 \quad \text{since } \operatorname{supp}(x)\subseteq S\\ &\leqslant \left\|(A^TA-I)_{S\times S}\right\|_2 \cdot \|x_S\|_2\\ &\leqslant \epsilon\,\|x\|_2\end{aligned}$$

and

$$\begin{aligned}\left\|(A^Te)_S\right\|_2 &\leqslant \left\|(A^T)_S\right\|_2 \cdot \|e\|_2\\ &\leqslant \sqrt{\|(A^TA)_{S\times S}\|_2}\cdot \|e\|_2\\ &\leqslant \sqrt{1+\epsilon}\,\|e\|_2\\ &\leqslant 2\,\|e\|_2\end{aligned}$$

Thus,

$$\|(z-x)_S\|_2 \leqslant \epsilon\,\|x\|_2 + 2\,\|e\|_2$$

We want error: $\|\hat{x}-x\| \leqslant O(\|e\|_2)$.

Idea: In order to reduce the error, we can apply the same method to the vector $r = x - z$. We now know

$$y^{(2)} = A(x-z) + e = y - Az$$

We set

$$z^{(2)} = A^T y^{(2)}$$

and let

$$\hat{r} = z^{(2)} - z$$

Then $\hat{r}$ is an estimate of r and

$$\begin{aligned}\left\|(z^{(2)}-x)_S\right\|_2 &= \|(z+\hat{r}-x)_S\|_2\\ &= \|(\hat{r}-(x-z))_S\|_2\\ &\leqslant \epsilon\,\|x-z\|_2 + 2\,\|e\|_2\end{aligned}$$

However, there are two problems with the idea:

1. The vector $r = x-z$ is no longer k-sparse so the equality in our first argument $\left\|[(A^TA-I)x]_S\right\|_2 = \left\|(A^TA-I)_{S\times S}x_S\right\|_2$ no longer holds.

2. From the first argument we got a bound on $\|(z-x)_S\|_2$ but our idea bounds $\left\|(z^{(2)}-x)_S\right\|_2$ using $\|x-z\|_2$.

Algorithm 1 Iterative Hard Thresholding (IHT)

```
for r = 0, 1, ..., R − 1 do
    x^(r+1) = H_k(x^(r) + A^T(y − Ax^(r)))
end for
return x^(R)
```

1.2 The algorithm

It turns out there is a simple fix to our idea: at the end of each update, we restrict the vector to the largest k values. See Algorithm 1.

Lemma 1.2.1.

$$\left\|x^{(r+1)} - x\right\|_2 \leqslant O(\epsilon)\left\|x^{(r)} - x\right\|_2 + O(\|e\|_2)$$

Proof. Set $S = \mathrm{supp}(x) \bigcup \mathrm{supp}(x^{(r)}) \bigcup \mathrm{supp}(x^{(r+1)})$ and thus $|S| \leqslant 3k$. As before,

$$\left\|[A^T(A(x - x^{(r)}) + e) - (x - x^{(r)})]_S\right\|_2 \leqslant \epsilon\left\|x - x^{(r)}\right\|_2 + 2\|e\|_2$$

Note that in our algorithm:

$$A^T(y - Ax^{(r)}) = A^T A(x - x^{(r)}) + e$$

and if we let

$$z = x^{(r)} + A^T(y - Ax^{(r)})$$

then our algorithm becomes

$$x^{(r+1)} = H_k(z)$$

and the inequality becomes

$$\|(z - x)_S\|_2 \leqslant \left\|x - x^{(r)}\right\|_2 + 2\|e\|_2$$

We need the following lemma to proceed.

Lemma 1.2.2. *If $x, z \in \mathbb{R}^n$ and x is k-sparse with $S = \mathrm{supp}(x)$, and $T \subseteq [n]$ contains the indices of k largest values in z, then*

$$\|x - z_T\|_2^2 \leqslant O(1)\left\|(x - z)_{S \bigcup T}\right\|_2^2$$

This implies

$$\begin{aligned}\left\|x - x^{(r+1)}\right\|_2 &\leqslant O(1)\left\|(x - z)_{\mathrm{supp}(x) \bigcup \mathrm{supp}(x^{(r+1)})}\right\|_2 \\ &\leqslant O(\epsilon)\left\|x - x^{(r)}\right\|_2 + O(\|e\|_2)\end{aligned}$$

Proof of Lemma 1.2.2. Consider the following three sets: $S\bigcap T,\ S\backslash T, T\backslash S$.

$$\|x - z_T\|_2^2 = \underbrace{\left\|(x-z)_{S\bigcap T}\right\|_2^2}_{\leqslant \left\|(x-z)_{S\bigcup T}\right\|_2} + \left\|x_{S\backslash T}\right\|_2^2 + \underbrace{\left\|z_{T\backslash S}\right\|_2^2}_{\substack{=\left\|(x-z)_{T\backslash S}\right\|_2^2 \\ \leqslant \left\|(x-z)_{S\bigcup T}\right\|_2^2}}$$

Now we analyze $\left\|x_{S\backslash T}\right\|_2^2$. The trick is to notice that since $|S| = |T|$, then $|S\backslash T| = |T\backslash S|$. Therefore, for all $i \in S\backslash T$ we can pair with a $j \in T\backslash S$ such that $|z_i| \leqslant |z_j|$ (or we use the fact that $\left\|z_{S\backslash T}\right\|_2^2 \leqslant \left\|z_{T\backslash S}\right\|_2^2$). Thus,

$$\begin{aligned} x_i^2 &= (x_i - z_i + z_i)^2 \\ &\leqslant (|z_i| + |z_i - x_i|)^2 \\ &\leqslant (|z_j| + |z_i - x_i|)^2 \\ &\leqslant 2|z_j|^2 + 2|z_i - x_i|^2 \end{aligned}$$

and

$$\begin{aligned} \left\|x_{S\backslash T}\right\|_2^2 &\leqslant 2\left\|z_{T\backslash S}\right\|_2^2 + 2\left\|(z-x)_{S\backslash T}\right\|_2^2 \\ &\leqslant 2\left\|(x-z)_{S\bigcup T}\right\|_2^2 \end{aligned}$$

□

Thus, for small ϵ, we have

$$\begin{aligned} \left\|x^{(r+1)} - x\right\|_2 &\leqslant \frac{1}{4}\left\|x^{(r)} - x\right\|_2 + O(\|e\|_2) \\ &\leqslant \frac{1}{2}\left\|x^{(r)} - x\right\|_2 \quad \text{unless } \left\|x^{(r)} - x\right\|_2 \leqslant O(\|e\|) \end{aligned}$$

and we can converge geometrically to $O(\|e\|_2)$ if we pick

$$R = \log_2 \frac{\|x\|_2}{\|e\|_2}$$

□

2 A Brief Introduction to Model-based Compressed Sensing

Suppose x is k-sparse and $\mathrm{supp}(x) \in \mathcal{F}$ where $\mathcal{F}$ is a subset of the family of all $\binom{n}{k}$ subsets of size k.

- Block sparsity. If the non-zeros in the vector are concentrated in blocks with block size $B > \log n$, then

$$|\mathcal{F}| \leqslant \binom{n/B}{k/B} \lesssim 2^{\frac{k}{B}\log\frac{n}{k}} \lesssim 2^k$$

- Tree sparsity (often appears in wavelet). In tree sparsity, each large coordinate correspondingly has 3 children. We want to find a subset of tree that is connected with k nonzeros. Consider the Eulerian tour on the large coordinates. At each step we can choose to go downleft, downright and up. We make at most $2k$ choices. We have $|\mathcal{F}| \leqslant 3^{2k} \leqslant 2^{O(k)}$.

Model-based RIP for $\mathcal{F}$: if $\forall x$ with $\text{supp}(x) \in \mathcal{F} \oplus \mathcal{F} = \{S \bigcup T : S, T \in \mathcal{F}\}$,

$$\|Ax\|_2^2 = (1 \pm \epsilon) \|x\|_2^2$$

[For Gaussian matrices: we need $O(k)$ measurements since $N(\epsilon, T_{\mathcal{F}}, \|\cdot\|_2) \leqslant (1/\epsilon)^k 2^{O(k)}$].

If we run model-IHT, we need to change H_k to $H_{\mathcal{F}}$. However, $H_{\mathcal{F}}$ could be slow. ($O(nk)$ for tree by DP and $O(n \log n)$ for $O(1)$-approximation).