Special Topics in Theoretical Computer Science	February 25, 2019

Lecture 13: Sketching for Linear Algebra Problems

Prof. Huy L. Nguyen — *Scribe: Lydia Zakynthinou*

First, we will finish the proof of the guarantee of the *Iterative Hard Thresholding* algorithm from Lecture 12.

1 Iterative Hard Thresholding: proof cont'd

Recall that the goal is to recover the k-sparse vector x from an observed measurement $y = \Pi x + e$ where e is the post-measurement noise and Π satisfies $(\varepsilon, 3k)$-RIP with $\varepsilon \leq \frac{1}{4\sqrt{2}}$.

In Lecture 12, we proved that the residual error $r^{(t)} = x - x^{(t)}$ satisfies the following inequality:

$$\|r^{(t+1)}\|_2 \leq 2 \left\| \left(I_{B^{(t+1)}} - \Pi^\top_{B^{(t+1)}} \Pi_{B^{(t+1)}} \right) r^{(t)}_{B^{(t+1)}} \right\|_2 + 2 \left\| \Pi^\top_{B^{(t+1)}} \Pi_{B^{(t)} \setminus B^{(t+1)}} r^{(t)}_{B^{(t)} \setminus B^{(t+1)}} \right\|_2 + 2 \left\| \Pi^\top_{B^{(t+1)}} e_{B^{(t+1)}} \right\|_2 \quad (1)$$

We bound each one of the three terms.

Claim 1 (Claim 3, Lecture 12)**.** $\left\| \left(I_{B^{(t+1)}} - \Pi^\top_{B^{(t+1)}} \Pi_{B^{(t+1)}} \right) r^{(t)}_{B^{(t+1)}} \right\|_2 \leq \varepsilon \|r^{(t)}_{B^{(t+1)}}\|_2$.

Claim 2. $\left\| \Pi^\top_{B^{(t+1)}} \Pi_{B^{(t)} \setminus B^{(t+1)}} r^{(t)}_{B^{(t)} \setminus B^{(t+1)}} \right\|_2 \leq \varepsilon \|r^{(t)}_{B^{(t)} \setminus B^{(t+1)}}\|_2$.

Proof. Similarly to Lemma 2 from Lecture 11, since Π satisfies the $(\varepsilon, 3k)$-RIP, for any $2k$-sparse vectors u and v with disjoint support:

$$\left| u \Pi^\top \Pi v \right| \leq \varepsilon \|u\|_2 \|v\|_2$$

In particular, if we consider arbitrary u with $\text{support}(u) \subseteq B^{(t+1)}$ and v with $\text{support}(v) \subseteq B^{(t)} \setminus B^{(t+1)}$:

$$\left\| \Pi^\top_{B^{(t+1)}} \Pi_{B^{(t)} \setminus B^{(t+1)}} \right\| = \sup_{\|u\|_2 = \|v\|_2 = 1} \left| u \Pi^\top \Pi v \right| \leq \varepsilon$$

Therefore,

$$\left\| \Pi^\top_{B^{(t+1)}} \Pi_{B^{(t)} \setminus B^{(t+1)}} r^{(t)}_{B^{(t)} \setminus B^{(t+1)}} \right\|_2 \leq \varepsilon \|r^{(t)}_{B^{(t)} \setminus B^{(t+1)}}\|_2$$

□

Claim 3. $\left\| \Pi^\top_{B^{(t+1)}} e_{B^{(t+1)}} \right\|_2 \leq (1 + \varepsilon) \|e\|_2$.

Proof. It holds that $\left\| \Pi^\top_{B^{(t+1)}} \right\| = \|\Pi_{B^{(t+1)}}\|$.

By the definition of the operator norm and for an arbitrary $2k$-sparse vector u with $\text{support}(u) \subseteq B^{(t+1)}$, $\|\Pi_{B^{(t+1)}}\| = \sup_{\|u\|_2 = 1} \|\Pi u\|_2 \leq (1 + \varepsilon)$.

It follows that $\left\|\Pi_{B^{(t+1)}}^{\top} e_{B^{(t+1)}}\right\|_2 \leq (1+\varepsilon)\|e_{B^{(t+1)}}\|_2 \leq (1+\varepsilon)\|e\|_2$. $\square$

By Claims **??**, **??**, **??**, and inequality (**??**), it follows that:

$$\begin{aligned}
\|r^{(t+1)}\|_2 &\leq 2\varepsilon\left(\|r^{(t)}_{B^{(t+1)}}\|_2 + \|r^{(t)}_{B^{(t)}\setminus B^{(t+1)}}\|_2\right) + 2(1+\varepsilon)\|e\|_2 \\
&\leq 2\sqrt{2}\varepsilon\|r^{(t)}\|_2 + 2(1+\varepsilon)\|e\|_2 && \text{(by the Pythagorean Theorem)} \\
&\leq \frac{\|r^{(t)}\|_2}{2} + 3\|e\|_2 && \text{(since } \varepsilon \leq \tfrac{1}{4\sqrt{2}}\text{)}
\end{aligned}$$

Thus,

$$\|r^{(t+1)}\|_2 \leq 2^{-1}\|r^{(t)}\|_2 + 3\|e\|_2 \tag{2}$$

By induction, we will show that

$$\|r^{(t+1)}\|_2 \leq 2^{-t}\|r^{(1)}\|_2 + 6\|e\|_2. \tag{3}$$

- *Base Step:* By (**??**), for $t = 1$, $\|r^{(2)}\|_2 \leq 2^{-1}\|r^{(1)}\|_2 + 3\|e\|_2 \leq 2^{-1}\|r^{(1)}\|_2 + 6\|e\|_2$.
- *Inductive Step:* By (**??**), it holds that, $\|r^{(t+1)}\|_2 \leq 2^{-1}\|r^{(t)}\|_2 + 3\|e\|_2$. By inductive hypothesis, $\|r^{(t)}\|_2 \leq 2^{-t+1}\|r^{(1)}\|_2 + 6\|e\|_2$. Therefore, $\|r^{(t+1)}\|_2 \leq 2^{-1}(2^{-t+1}\|r^{(1)}\|_2 + 6\|e\|_2) + 3\|e\|_2 = 2^{-t}\|r^{(1)}\|_2 + 6\|e\|_2$.

Hence, since $r^{(1)} = x - x^{(1)} = x$, we get that $\|r^{(t+1)}\|_2 \leq 2^{-t}\|x\|_2 + 6\|e\|_2$ and this concludes the proof of Theorem 2 from Lecture 12.

2 Model Based Compressive Sensing

So far, the model we have assumed for our signal x is the set of all vectors in $\mathbb{R}^n$ that are k-sparse. Would having more information about the structure of the signal help in its recovery? In the more general *Model Based Compressive Sensing* we assume some model $\mathcal{M}$. The number of rows of the matrix Π grows as the logarithm of the size of an ε-net of $\mathcal{M}$. If the model $\mathcal{M}$ is the set of all k-sparse vectors, as before, then this quantity would be $\sim \log\left(\binom{n}{k}\right)$.

In the general case, the Model Based Iterative Hard Thresholding algorithm is:

Algorithm 1 Model Based Iterative Hard Thresholding

$x^{(1)} \leftarrow 0$
for $t = 1$ to T **do**
  $a^{(t+1)} \leftarrow x^{(t)} + \Pi^{\top}\left(y - \Pi x^{(t)}\right)$
  $x^{(t+1)} \leftarrow P_{\mathcal{M}}(a^{(t+1)})$
end for

Here, instead of the operator H_k, we have $P_{\mathcal{M}}$, which projects $a^{(t+1)}$ to $\mathcal{M}$.

One example of such a model is the "block sparsity" model, where we assume there exist k non-zeros in the signal and each is in one of $\frac{k}{B}$ blocks of size B. Then, in order to recover the signal, one would have to guess the start of each of those blocks, so the number of measurements needed would be $\sim \frac{k}{B} \log\left(\frac{n}{k}\right)$. Another example is "tree sparsity". In the Tree Sparsity problem we are given a node-weighted tree of size n and aim to output a tree of size k with maximum weight ([**?**]). In this case, the number of measurements needed is $\sim k + \log(n)$.

3 Fast Algorithms for Linear Algebra Problems

3.1 Matrix Multiplication

Let $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{d\times p}$ be two matrices. Let a_i denote row i of A and b_j denote row j of B. The goal is to compute (approximately) the product $A^\top B$.

We can compute the product exactly in $O(ndp)$ time. Furthermore, if the matrices are in $\mathbb{R}^{n\times n}$, then this computation takes $O(n^\omega)$ time, where $\omega = \log_2(7)$ for Strassen's algorithm. The state of the art algorithm for this problem achieves $\omega = 2.3728639$.

We aim to compute a matrix C such that with probability at least $1-\delta$, $\|C - A^\top B\|_q \le \varepsilon \|A\|_p \|B\|_p$, for some norm p and q.

3.1.1 Sampling Technique

We will compute a matrix C as follows: we will sample the i-th term with probability p_i (to be defined later) and whenever the i-th term is picked, we add $\frac{1}{p_i} a_i b_i^\top$ to the sum.

Then, we have the following claim.

Claim 4. $\mathbb{E}[C] = A^\top B$.

Proof. $\mathbb{E}[C] = \sum_{i=1}^{n} p_i \left(\frac{1}{p_i} a_i b_i^\top\right) = \sum_{i=1}^{n} a_i b_i^\top = A^\top B$. □

Claim 5. $\mathbb{E}[\|A^\top B - C\|_F^2] = \sum_{i=1}^{n} \|a_i\|^2 \cdot \|b_i\|^2 \cdot \left(\frac{1}{p_i} - 1\right)$.

Proof. Let x_i be the indicator variable such that $x_i = 1$ if the i-th term is picked, and $x_i = 0$ otherwise.

$$
\begin{aligned}
\mathbb{E}\left[\left\|A^\top B - C\right\|_F^2\right] &= \mathbb{E}\left[\left\|\sum_{i=1}^n a_i b_i^\top \left(1 - \frac{x_i}{p_i}\right)\right\|_F^2\right] \\
&= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\left[\operatorname{trace}\left(\left(a_i b_i^\top \left(1 - \frac{x_i}{p_i}\right)\right)^\top \left(a_j b_j^\top \left(1 - \frac{x_j}{p_j}\right)\right)\right)\right] \\
&\qquad\qquad (\|M\|_F^2 = \operatorname{trace}\left(M^\top M\right) = \operatorname{trace}\left(MM^\top\right)) \\
&= \sum_{i=1}^n \mathbb{E}\left[\left(1 - \frac{x_i}{p_i}\right)^2 \cdot \operatorname{trace}\left(b_i a_i^\top a_i b_i^\top\right)\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\left(1 - \frac{x_i}{p_i}\right)^2 \cdot \|a_i\|^2 \cdot \operatorname{trace}\left(b_i b_i^\top\right)\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\left(1 - \frac{x_i}{p_i}\right)^2 \cdot \|a_i\|^2 \cdot \|b_i\|^2\right] \\
&= \sum_{i=1}^n \|a_i\|^2 \cdot \|b_i\|^2 \cdot \mathbb{E}\left[\left(1 - \frac{x_i}{p_i}\right)^2\right] \\
&= \sum_{i=1}^n \|a_i\|^2 \cdot \|b_i\|^2 \cdot \left(p_i \cdot \left(1 - \frac{1}{p_i}\right)^2 + (1 - p_i) \cdot 1\right) \\
&= \sum_{i=1}^n \|a_i\|^2 \cdot \|b_i\|^2 \cdot \left((p_i - 1)\left(1 - \frac{1}{p_i}\right) + (1 - p_i)\right) \\
&= \sum_{i=1}^n \|a_i\|^2 \cdot \|b_i\|^2 \cdot \left((p_i - 1)\left(1 - \frac{1}{p_i} - 1\right)\right) \\
&= \sum_{i=1}^n \|a_i\|^2 \cdot \|b_i\|^2 \cdot \left(\frac{1}{p_i} - 1\right)
\end{aligned}
$$

□

To minimize the expression, we set $p_i = \frac{\|a_i\|\|b_i\|}{\sum_{i=1}^n \|a_i\|\|b_i\|}$.

In the next lecture, we will present and prove the guarantee of the approximation matrix C.

References

[1] Arturs Backurs, Piotr Indyk, Ludwig Schmidt. Better Approximations for Tree Sparsity in Nearly-Linear Time. *SODA*, 2215–2229, 2017.