CS 395T: Sublinear Algorithms Fall 2014

Lecture 10&11 — Sept 30 & Oct 2, 2014

Prof. Eric Price *Scribe: Xue Chen*

1 Overview

These notes describe two lectures. The first gives an introduction to ϵ-covers, ϵ-packings, and RIP matrices. The second describes compressed sensing and iterative hard thresholding.

2 ϵ-cover and ϵ-packing

First, we define metric spaces, ϵ-covers and covering number.

Definition 1 (Metric Space)**.** *A metric space is an ordered pair (X, d) where X is a space and d is a metric on X such that $\forall x, y \in X$:*

1. $d(x, y) \geq 0$.
2. $d(x, y) = 0 \Leftrightarrow x = y$.
3. $d(x, y) = d(y, x)$.
4. $d(x, y) \leq d(x, z) + d(z, y), \forall z \in X$.

Definition 2. *An ϵ-cover of X with respect to d is a collection of points $\{x_1, \cdots, x_n\} \subseteq X$ such that $\forall y \in X, min_{1 \leq i \leq n} d(y, x_i) \leq \epsilon$.*

Definition 3. *The covering number $N(\epsilon, X, d)$ is the minimal number of points of all ϵ-cover of X w.r.t. d.*

We abuse the notation N to denote $N(\epsilon, X, d)$ if X, d is clear. We use $\log N(\epsilon, X, d)$ to denote the metric entropy of (X, d) which indicates the information by knowing the positions of a point to ϵ distance in d.

Example 4. *$X = [-1, 1], d(x, y) = |x - y|$. Then $\{0, \pm 2\epsilon, \pm 4\epsilon, \cdots\}$ is an ϵ-cover of X, so $N(\epsilon, X, d) \leq \frac{2}{2\epsilon} + 1 = 1 + 1/\epsilon$.*

Example 5. *$X = [-1, 1]^m, d(x, y) = |x - y|_\infty$. From the above example, $\{0, \pm 2\epsilon, \pm 4\epsilon, \cdots\}^m$ is an ϵ-cover of X. So $N(\epsilon, X, d) \leq (1 + 1/\epsilon)^m$ and the metric entropy is $\log N = \Theta(m \log(\frac{1}{\epsilon}))$.*

A closely related concept of covering number is packing number.

Definition 6. *An ϵ-packing of X w.r.t. d is a collection of points $\{x_1, \cdots, x_n\} \subseteq X$ such that $\forall i, j (i \neq j), d(x_i, x_j) \geq \epsilon$.*

The packing number $M(\epsilon, X, d)$ is the maximal number of points of all ϵ-packings of (X, d).

Lemma 7.

$$M(2\epsilon, X, d) \leq N(\epsilon, X, d) \leq M(\epsilon, X, d).$$

In general, the difference between ϵ and 2ϵ will only affect constants, which we will not care about in this course. So we will freely switch between the packing number and the covering number of (X, d).

3 $N(\epsilon, B_q^d, \|\cdot\|_p)$

Now let us consider the covering of L_q balls in L_p norm.

Definition 8. $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ *for any* $p > 0$. *The unit ball of dimension* d *in* L_p *is* $B_p^d = \{x \in R^d | \|x\|_p \leq 1\}$.

One can verify $\|\cdot\|_p$ is a metric by Holder's inequality for any $p > 0$. We abuse the notation $\|\cdot\|$ if p is clear. One of the basic property of L_p we will use in this lecture is $\|x\|_{p_0} \leq \|x\|_{p_1}$ for $p_0 \geq p_1$. We also use $Vol(S)$ to denote the volume of S. Another property of ball volume often used in this lecture is that $\frac{\alpha B_p^d}{\beta B_p^d} = (\frac{\alpha}{\beta})^d$ – scaling a ball by α in each of d dimensions increases its volume by a factor of α^d.

Fact 9. $\frac{1}{\epsilon^d} \frac{Vol(B_q^d)}{Vol(B_p^d)} \leq N(\epsilon, B_q^d, \|\cdot\|_p) \leq (\frac{2}{\epsilon})^d \frac{Vol(B_q^d + \frac{\epsilon}{2} B_p^d)}{Vol(B_p^d)}$.

Proof. Lower bound: Let $\{x_1, x_2, \cdots, x_N\}$ be an ϵ-cover of B_q^d. Because $B_q^d \subseteq \cup_i (x_i + \epsilon B_p^d)$, $Vol(B_q^d) \leq N \cdot Vol(\epsilon B_p^d)$.

Upper Bound: Let $\{x_1, x_2, \cdots, x_N\}$ be an ϵ-packing of B_q^d. Because all the balls of $x_i + \frac{\epsilon}{2} B_p^d$ are disjoint, $\cup_i (x_i + \frac{\epsilon}{2} B_p^d) \subseteq B_q^d + \frac{\epsilon}{2} B_p^d$(some x_i may be on the surface). Therefore $N \cdot Vol(\frac{\epsilon}{2} B_p^d) \leq Vol(B_q^d + \frac{\epsilon}{2} B_p^d)$ and a upper bound of packing number is also a upper bound of covering number by Lemma 7. □

To make this simpler, let's look at a couple cases.

Same norm. If $p = q$, the upper bound $\frac{Vol(B_q^d + \frac{\epsilon}{2} B_p^d)}{Vol(B_p^d)} = \frac{Vol((1+\frac{\epsilon}{2}) B_p^d)}{Vol(B_p^d)} = (1 + \frac{\epsilon}{2})^d$. Therefore $\frac{1}{\epsilon^d} \leq N \leq (1 + \frac{2}{\epsilon})^d$.

When $q = 1$ and $p = 2$. Because $B_1^d \leq B_2^d$ from the property of L_p, one upper bound is $\frac{Vol(B_q^d + \frac{\epsilon}{2} B_p^d)}{Vol(B_p^d)} \leq \frac{Vol((1+\frac{\epsilon}{2}) B_p^d)}{Vol(B_p^d)} \leq (1 + \frac{\epsilon}{2})^d$. For the lower bound, $Vol(B_1^d) = \frac{2^d}{d!}$ because there are two signs for each dimension and the volume of each d-simplex is $\frac{1}{d!}$. $Vol(B_2^d) = \frac{\pi^{d/2}}{(d/2)!}$ for even d[Sphere]. Therefore $\frac{1}{\epsilon^d} \cdot \frac{2^d / d!}{\pi^{d/2}/(d/2)!} \leq N$ and $d \log(1/\epsilon) - \frac{d}{2} \log d \leq \log N \leq d \log(2/\epsilon)$.

This gives a tight bound of $N = \Theta(d \log(1/\epsilon))$ when $\epsilon < 1/d$. For $\epsilon > 1/\sqrt{d}$, however, the lower bound is trivial.

In fact, the volume argument is loose in the "large ϵ" setting. We can also show that $\log N \leq O(\frac{\log d}{\epsilon^2})$ by Maurey's empirical method (see, for example, [NPW12]):

1. For any $\vec{x} = (x_1, \cdots, x_d) \leq B_1^d$, we consider the following experiment($e_1, e_2, \cdots, e_d$ is the standard basis of $\mathbb{R}^d$):
2. Randomly sampling z_i from $\{e_1, e_2, \cdots, e_d\}$ according to $(|x_1|, |x_2|, \cdots, |x_n|)$ (or 0 for the remainder) for $i = 1, \cdots, t$ independently.
3. Let $z = \frac{1}{t}\sum_i z_i$. Then $E[z] = x$ and
$$E[\|x - z\|_2^2] = \sum_{j=1}^{d}(x_j - \frac{1}{t}\sum_i 1_{z_i = e_j})^2 = \frac{1}{t}\sum_{j=1}^{d} x_j(1 - x_j) \leq \frac{1}{t}\sum_j x_j \leq 1/t.$$
4. So there exists a z such that $E[\|z - x\|_2] \leq \epsilon$ after choosing $t = 1/\epsilon^2$. Therefore $N \leq (2d+1)^t$ and $\log N \leq O(\frac{\log d}{\epsilon^2})$.

4 Sparse Vectors and RIP matrix

Let us start with some definitions.

Definition 10. *We use $supp(x) = \{i | x_i \neq 0, 1 \leq i \leq d\}$ to denote the support of vector x. And we define $\|x\|_0 = |supp(x)|$ and k-sparse space*
$$T_k = \{x : \|x\|_2 \leq 1, \|x\|_0 = k\}.$$

It is not difficult to see $N(\epsilon, T_k, \|\cdot\|_2) \leq \binom{d}{k}(1 + \frac{2}{\epsilon})^k$ by a union bound over all k-dimensional subspaces. Therefore $\log N(\epsilon, T_k, \|\cdot\|_2) \leq O(k \log \frac{d}{k\epsilon})$.

Now we are interested in finding a matrix A with "few" rows that preserves the norm of *every vector* $x \in T_k$, i.e. has bounded $\max_{x:x\in T_k} \frac{\|Ax\|_2}{\|x\|_2}$ over nonzero vector x. Recall what we proved in the construction of a JL matrix: if we sample a matrix $A \in \mathbb{R}^{m\times n}$ by independently sampling each entry $a_{i,j} \sim N(0, 1/m)$, then $\forall x \in \mathbb{R}^n, \|Ax\|_2^2 = (1 \pm \epsilon)\|x\|_2^2$ with prob. at least $1 - 2e^{-\epsilon^2 m/C}$ for some constant C(see lecture note 2). If T_k is finite, we could take a union bound to argue the same way to generate A also works here for every vector $x \in T_k$. However, T_k is infinite and we need another argument to bound the error.

Instead of union bound, let $S = \{x^{(1)}, \cdots, x^{(N)}\}$ be an ϵ-cover(a.k.a."net") of T_k with size $N \leq \binom{d}{k}(1 + \frac{2}{\epsilon})^d$. We can decompose any $x \in T_k$ in this way:

1. Find $x_1 \in S$ such that $x = x_1 + \epsilon x'$ for $\|x'\|_2 \leq 1$ and $|supp(x')| \leq k$. Since S is an ϵ-cover, there always exists $x_1 \in S$ such that $\|x - x_1\|_2 < \epsilon$ by definition. $|supp(x')| \leq k$ follows from a special property of our net: because our net is a union bound over all k-dimensional subspaces, we can choose x_1 from the same k-dimensional subspace as x.
2. If $\|x'\| \neq 0$, then applying the above procedure on x' again to get $x' = x_2 + \epsilon x''$ such that $\|x''\|_2 \leq 1$ and $|supp(x'')| \leq k$, and so on.

3. Eventually, we have $x = x_1 + \epsilon x_2 + \epsilon^2 x_3 + \cdots + \epsilon^{i-1} x_i + \cdots$ so all the $x_i \in S$.

Now we choose $m = C_0 \log N/\epsilon^2 = O(\frac{d}{\epsilon^2} \log \frac{d}{\epsilon k})$ for a large constant C_0 and $\epsilon < \frac{1}{2}$, then $\|Ax\|_2 = (1 \pm \epsilon)\|x\|_2$ for all $x \in S$ with high probability by union bound. This concludes

$$\begin{aligned}
\forall x \in T, \|Ax\|_2 &= \|A(x_1 + \epsilon x_2 + \epsilon^2 x_3 + \cdots + \epsilon^{i-1} x_i \cdots)\|_2 \\
&\leq \|Ax_1\|_2 + \epsilon\|Ax_2\|_2 + \epsilon^2\|Ax_3\|_2 + \cdots + \epsilon^{i-1}\|Ax_i\|_2 + \cdots \\
&\leq (1+\epsilon)(1 + \epsilon + \epsilon^2 + \cdots + \epsilon^{i-1} + \cdots) \\
&\leq (1+\epsilon)\frac{1}{1-\epsilon} \\
&\leq 1 + O(\epsilon)
\end{aligned}$$

Definition 11 (Restricted Isometry Property)**.** *An $m \times n$ matrix A has restricted isometry property(RIP) of (k, ϵ) if $\forall x$ with $\|x\|_0 \leq k, (1-\epsilon)\|x\|_2 \leq \|Ax\|_2 \leq (1+\epsilon)\|x\|_2$.*

An equivalent way to define RIP is for any subset S of size $k, \|(A_S)^T A_S - I\|_2 \leq \epsilon$ where A_S is the matrix by picking column in S and $\|A\|_p = \sup_{x \neq \vec{0}} \frac{\|Ax\|_p}{\|x\|_p}$.

From now on, we assume ϵ is a small constant such as $1/10$ and $1/100$ without further specification. In general, we are interested in the construction of an $m \times n$ RIP matrix A with the following properties:

1. It is easy to check A satisfies RIP or not.
2. A can be stored in $o(mn)$ space.
3. The multiplication Ax can be computed in $o(mk)$ time for $x \in T_k$.
4. m is as small as possible.

Example 12 (Random (sub)Gaussian Matrix)**.** *We generate A by independently sampling a (sub)Gaussian variable in every entry. With overwhelming probability ($\geq 1 - e^{\Omega(m)}$), A is an RIP matrix for sufficiently large $m = \Omega(\frac{1}{\epsilon^2} klog(n/k))$. However, we do not know how to verify A satisfies RIP even though it happens with very high prob., and it is also bad in storage and multiplication.*

Example 13 (Coherent Matrix)**.** *A matrix A with n columns $\{a_1, a_2, \cdots, a_n\}$ is defined to be α-coherent iff $\frac{<a_i, a_j>}{\sqrt{\|a_i\|_2 \cdot \|a_j\|_2}} \leq \alpha$. Let A' be the normalized matrix of A(normalize every column to a unit vector). We can show A' is a $(k, \alpha \cdot k)$-RIP matrix: for any k-sparse vector x,*

$$\|Ax\|_2^2 = \sum_{i \in supp(x)} \sum_{j \in supp(x)} < a_i', a_j' > x_i x_j = \sum_{i \in supp(x)} x_i^2 + \sum_{i \neq j} \alpha \cdot |x_i x_j| \leq \|x\|_2^2 + \alpha k \|x\|_2^2.$$

Coherent matrix is easy to verify but it need a large m if we want to use it as a RIP matrix. For example, suppose we generate every a_i by independently choosing $\{\pm 1\}$ in every entry. It is not to difficult to see $m \geq 1/\alpha^2 \geq \Omega(k^2)$ if we want a_i is α-coherent with high prob.(compute the variance of $< a_i, a_j >$). Finding an explicit RIP matrix with m much smaller than k^2 is a challenging open problem. Some progress was made by Bourgain et. al.[BDFKK11], who obtained $m = k^{2-\delta}$ for a universal constant δ, but δ is tiny.

Example 14 (Fourier Matrix)**.** *[FM] An $n \times n$ square Fourier matrix F is defined as $F_{i,j} = \omega^{ij}$ where $\omega = e^{2\pi i/n}$(i is the imaginary unit here). Another way to construct a (k, ϵ)-RIP matrix is sampling a row subset S with size $m = O(k \log n \log^3 \log k)$ to get a $|S| \times n$ matrix A[CGV2013], and A is a RIP matrix w.h.p.*

There exists an algorithm multiplies A to x in time $\tilde{O}(m)$ and stores A in space $O(k \log^2 n \log^3 \log k)$. However, we do not know how to verify a matrix A that is generated this way is a RIP matrix. The same construction also works for Hadamard matrix and a similar one for circulant matrices.

One interesting question for RIP matrices is how sparse can a RIP matrix be. The negative result is that there are at least $\Omega(k)$ nonzero entries in every column (see HW3).

5 Compressed Sensing

Let $A \in \mathbb{R}^{m \times n}$. We normalize every column of A to be roughly 1. Given $y = Ax + e$ for some $x \in \mathbb{R}^n$ with $\|x\|_0 \leq k$ and e is a noise vector. The goal of compressed sensing is to efficiently recover $\hat{x}$ given y such that $\|x - \hat{x}\|_2 \leq C\|e\|_2$ for some universal constant C.

Compressed sensing has widely applications in industry. For example, it has been used in imaging processing, magnetic resonance imaging(MRI), oil expolation, spectrum sensing and feature testing. It takes advantage of the data's sparseness in some basis. We first discuss the differences between compressed sensing and sparse recovery, then introduce Iterative Hard Thresholding to recover $\hat{x}$.

Compressed sensing is very similar to "sparse recovery" or "heavy hitters", problems we saw earlier in class with Count-Min and Count-Sketch. "Compressed sensing" and "sparse recovery" are terms for essentially the same problem that grew out of different communities: "compressed sensing" from math/statistics/signal processing, and "sparse recovery" from streaming algorithms in computer science. That said, there are noticeable differences in problem formulation and approaches between the two communities, so it makes sense to preserve the distinction.

Note that this list isn't formal or definitive; it's a sense of differences between the two communities working in the same area, but there's work that blurs the lines.

1. In sparse recovery, it is allowed to choose matrix A after giving x and the algorithm only needs to be able to find the correct answer w.h.p. over A such as Count-Sketch. However, we have to choose the matrix A before reading y in compressed sensing, and the algorithm should be able to recover $\hat{x}$ for every $y = Ax + e$ where x, e satisfy the requirements.

2. In compressed sensing, there is a noise vector e and one often assumes that x is exactly k-sparse, while in sparse recovery one generally assumes that x is not k-sparse but you observe Ax exactly. This distinction is generally not too important, and algorithms that work in either noise model typically also work in the other one.

3. Sparse recovery cares more about the running time. Sparse recovery algorithms strive for $n \log^c n$ or ideally $k \log^c n$ time, while compressed sensing algorithms are often happy with n^c time.

4. In sparse recovery, algorithms is closed tied to the matrices we used in it. In compressed sensing, algorithms works well as long as the matrix A has some property P. For example,

if A satisfies RIP, solving $argmin_{\hat{x}:\|A\hat{x}-y\|\leq\epsilon}\|\hat{x}\|_1$ will give a good recovery $\hat{x}$ of x by convex optimization, L_1 minimization and iterative methods.

6 Iterative Hard Thresholding

We are focusing on Iterative Hard Thresholding in the rest of this notes. We first describe its algorithm and go to the analysis later. Let A be a $(C\cdot k,\epsilon)$-RIP matrix and $H_k(x)$ denote the projection of x on its top k elements. Given $y=Ax+e$ for $\|x\|_0\leq k$ and $\|e\|_2$ is small, the algorithm works as following with an appropriate choice t:

1. $x^{(1)}=\vec{0}$.
2. For $i=1,2,\cdots,t$
3. $\quad x^{(i+1)}=H_k(x^{(i)}+A^T(y-Ax^{(i)}))$.

We are going to prove $\|x^{(t+1)}-x\|_2\leq O(\|e\|)$ for $t=O(\log\frac{\|x\|}{\|e\|})$. The intuition of the algorithm is $A^T(y-Ax^{(i)})=A^TA(x-x^{(i)})+A^Te$. Because $\|A_S^TA_S-I_k\|_2\leq\epsilon$ for any column subset S with size $C\cdot k$, we can think $A^TA\approx I_n$ and $A^TA(x-x^{(i)})+A^Te\approx x-x^{(i)}$.

Proof. Let $x_0=x-x^{(i)},z=A^TAx_0+A^Te$. We use H be the support set of x_0. Because $x^{(i+1)}=H_k(x^{(i)}+z)$, we try to bound $\|z-x_0\|$ at first. For any column subset S with size at most $(C-2)k$, we bound $\|(z-x_0)_S\|$ as this way:

$$\begin{aligned}\|(z-x_0)_S\|_2&=\|\big((A^TA-I)x_0+A^Te\big)_S\|_2\\&\leq\|\big((A^TA-I)x_0\big)_{S\cup H}\|_2+\|(A^Te)_S\|_2\\&\leq\|(A^TA-I)_{(S\cup H)\times(S\cup H)}\|_2\cdot\|x_0\|_2+\|A_S\|_2\cdot\|e\|_2\\&\leq\epsilon\|x_0\|_2+(1+\epsilon)\|e\|_2\end{aligned}$$

However, our goal is to prove $\|x-x^{(i+1)}\|$ is small after enough steps which is equivalent to prove $\|z_S-x_0\|_2/\|x_0\|_2<1$ for the top k elements subset S in x^{i+1}. We need the following Lemma to prove the bound of $\|(z-x_0)_S\|_2$ can be used to bound $\|z_S-x_0\|_2$.

Lemma 15. *Let $x,z\in\mathbb{R}^n$, x is k-sparse with support set H and S is the top k elements subset of z. Then*

$$\|x-z_S\|_2^2\leq 5\|(x-z)_{H\cup S}\|_2^2.$$

Proof. Pairing up $i\in H\setminus S$ and $j\in S\setminus H$(Recall $|S|=|H|=k$ and $z_j\geq z_i$ by definition), it is enough to prove

$$z_j^2+x_i^2\leq 5\big((z_i-x_i)^2+z_j^2\big).$$

We discuss it by two cases:

1. $|z_i|>|x_i|/2:x_i^2\leq 4z_i^2\leq 4z_j^2$.
2. $|z_i|<|x_i|/2:x_i^2\leq 4(x_i-z_i)^2$.

□

Continue to the proof of the convergence, Taking $(x, x^{(i+1)})$ into the lemma, if $\epsilon < 0.1$ and $\|x^{(i)} - x\| \geq 12\|e\|$ we have

$$\begin{aligned}
\|x - x^{(i+1)}\| &\leq \sqrt{5}\|(x - x^{(i+1)})_{S\cup H}\| \\
&\leq \sqrt{5}\|(x_0 - z)_{S\cup H}\| \\
&\leq \sqrt{5}\epsilon\|x_0\| + \sqrt{5}(1+\epsilon)\|e\| \\
&\leq \|x^{(i)} - x\|/4 + 3\|e\| \\
&\leq \|x^{(i)} - x\|/2.
\end{aligned}$$

For $t = O(\log \frac{\|x\|}{\|e\|})$, we have $x^{(t+1)} = O(\|e\|)$. □

References

[Sphere] https://en.wikipedia.org/wiki/Sphere.

[NPW12] Jelani Nelson, Eric Price, Mary Wooters. New constructions of RIP matrices with fast multiplication and fewer rows. *SODA 2014*: 1515-1528.

[CGV2013] Mahdi Cheraghchi ,Venkatesan Guruswami,Ameya Velingker. Restricted isometry of fourier matrices and list decodability of random linear codes em SODA 2013:Pages 432-442

[BDFKK11] Jean Bourgain, Stephen Dilworth, Kevin Ford, Sergei Konyagin, Denka Kutzarova. Breaking the k^2 Barrier for Explicit RIP Matrices. *STOC*:pages 637644, 2011.

[FM] http://mathworld.wolfram.com/FourierMatrix.html.