CS265/CME309: Randomized Algorithms and Probabilistic Analysis

Lecture #9: Compressed Sensing and the Restricted Isometry Property

Mary Wootters (edited by Greg Valiant)

February 1, 2022

1 Sparsity

A vector $x \in \mathbb{R}^n$ is *sparse* if it has only a few large components. Sparsity shows up all over the place in real-world data. For example, we hope that vectors indicating errors or anomolies will be sparse. Even things that we might not think of as sparse—like natural images or human speech—tend to be pretty sparse when they are transformed into the correct basis. In this lecture, we'll build on the ideas we saw in the previous lecture about the Johnson-Lindenstrauss transform to see how to do dimension reduction for the set of sparse vectors, and an application of this called *compressed sensing.*

2 Compressed Sensing

The mathematical set-up for the *compressed sensing* problem, introduced[1] by Candés, Romberg and Tao [1] and Donoho [2], is the following. Let $A \in \mathbb{R}^{m \times n}$ be a matrix, with $m \leq n$, and suppose that $x \in \mathbb{R}^n$ is an (approximately) k-sparse vector. That is, x has only k nonzero entries, or perhaps only k entries that are very large. The goal is, given $y = Ax$, recover x. If it weren't for the assumption of sparsity, this would be impossible: $Ax = y$ is an under-determined linear system, and there's not a unique solution x. However, it turns out that, subject some conditions on A, there is a unique *sparse* solution x, and moreover we can find it efficiently.

The motivation for compressed sensing comes from signal processing, in cases where it's natural to recover some linear function of a signal. A motivational example is MRIs. Glossing over a lot of details, each measurement that you take in an MRI is basically a Fourier coefficient of some sparse vector. That is, if F is the $n \times n$ discrete Fourier transform (so, $F_{kj} = e^{-2\pi ikj/n}$), each

[1] The question of "sparse approximation" which is similar but is after slightly different guarantees and comes up in different contexts, had already been around for a while, as had related problems like sparse regression in statistics.

measurement looks like a component of the vector Fx, where x is sparse. The goal at the end of the day is to recover x, which will allow us to recover the MRI image. One thing we could do at this point is invert F to find x, but this is wasteful. Why take all those n measurements[2] if we are just after the $k \ll n$ nonzero entries of x? Taking fewer measurements corresponds to acquiring the vector Ax, where $A \in \mathbb{R}^{m \times n}$ consists of m rows of F. Then the problem is: can we recover x (or something approximately like it) given the observations Ax? And this is the compressed sensing problem described above. There are also examples of applications (like the "single-pixel camera") where we get to design the matrix A.

2.1 The Restricted Isometry Property

When can we recover a sparse vector x given Ax? Not every matrix A will work. For example, suppose that $A = [I|0]$ has an $m \times m$ identity matrix on the left and a bunch of zeros on the right. Then there are plenty of sparse vectors in the kernel of A, so A would not work. The following property is not only enough to guarantee that we can, in theory, recover x from Ax, but also that we can do it efficiently.

Definition 1 (Restricted Isometry Property)**.** *A matrix $A \in \mathbb{R}^{m \times n}$ has the* restricted isometry property *(RIP) with parameters k and ε if, for every k-sparse vector $x \in \mathbb{R}^n$,*

$$(1 - \varepsilon)\|x\|_2 \leq \|Ax\|_2 \leq (1 + \varepsilon)\|x\|_2.$$

Above, we defined A and x above as being real, but they could be complex too and the definition would be the same.

The RIP might look somewhat familiar: it's a bit like the guarantee from the Johnson-Lindenstrauss lemma. In that language, if A has the RIP with parameters $2k$ and ε, then multiplication by A is a $(1 + O(\varepsilon))$-distortion embedding of the (infinite) set X of k-sparse vectors into $\mathbb{R}^m$.

If A has the RIP with parameter $2k$, then we can always recover a k-sparse vector x given Ax. Indeed, suppose that x and x' are two k-sparse vectors with $Ax = Ax'$. Then the RIP implies that $0 = \|A(x - x')\| \geq (1 - \varepsilon)\|x - x'\|_2$, which implies that $x = x'$. But it also proves something stronger, which is that there's an efficient algorithm to find x. We won't go into that in these notes, but the idea is that instead of solving the problem "find the sparsest x so that $Ax = y$" (which is computationally intractable), we can solve the problem "find the x with the smallest ℓ_1 norm so that $Ax = y$" (which can be solved by linear programming). It turns out that if A has the RIP, then the solutions to these two problems are the same, so we can efficiently recover an exactly k-sparse vector. More generally, the RIP allows us to efficiently recover a nearly-k-sparse vector, up to some error.

Theorem 1. *Suppose that A has the RIP with parameters $2k$ and $\delta = \Theta(1)$. Let $x \in \mathbb{R}^n$ be any vector (not necessarily sparse). Then there is an efficient algorithm that, given A and Ax, can recover $\hat{x}$ so that*

$$\|x - \hat{x}\|_1 \leq C \|x - x_k\|_1$$

[2] In this example, more measurements correspond to more time you actually spend in the MRI. If you have ever had an MRI, you understand why it's desirable to minimize this.

where C is some constant and x_k is the vector consisting of the largest k components of x.

We won't prove that theorem here, but see [3] for a nice textbook with some more details about compressed sensing, or the CS168 lecture notes (https://web.stanford.edu/class/cs168/l/l17.pdf) for some intuition.

Thus, our goal is to find matrices with the RIP, where m is as small as possible. And this is where randomness comes in. There's no probability in the definition of the RIP, so can't we just find some deterministic A that does the job and call it a day? Perhaps we could, but giving a deterministic construction with close to the best possible m is still an open question! On the other hand, there are many randomized constructions that work.

2.2 A random Gaussian matrix has the RIP with high probability

A natural matrix to study—motivated by the MRI example above as well as many other applications—is the "rows-of-a-DFT" matrix mentioned above. It turns out that this does have the RIP with high probability if you take random rows, but that proof is just a bit beyond the tools we have right now. Check out [4] for one nice proof if you're curious. Instead, we'll prove that a matrix A with independent Gaussian entries $A_{ij} \sim N(0, 1/m)$ has the RIP.

Theorem 2. *Let $\delta \in (0,1)$ and choose integers k and n. There is some $m = O\left(\frac{k \log n}{\delta^2}\right)$ so that, with probability at least $1 - o(1)$, a matrix $A \in \mathbb{R}^{m \times n}$ with independent entries $A_{ij} \sim N(0, 1/m)$ has the RIP with parameters k and δ.*

First, notice that it suffices to prove the theorem for k-sparse x with $\|x\|_2 = 1$. Thus, we will restrict our attention to only these unit-norm vectors x from now on. Let

$$\Sigma_k = \{x \in \mathbb{R}^n \,:\, \|x\|_2 = 1 \text{ and } x \text{ is } k\text{-sparse}\}$$

be the set of unit-norm k-sparse vectors.

Our first thought might be to do something like we did when we proved the Johnson-Lindenstrauss lemma. First, show that for any *fixed* k-sparse x, $\|Ax\|_2 \approx \|x\|_2$ with really high probablity. Then, union bound over all possible x's. The obvious catch here is that, in this case, we have infinitely many x's to union bound over! Instead, we'll use a technique called an *ε-covering* to turn this into only a finite union bound.

Definition 2 (ε-covering)**.** *An ε-covering of a set $X \subseteq \mathbb{R}^n$ is a finite set $\mathcal{N}$ so that for every $x \in X$, there is some $y \in \mathcal{N}$ so that $\|x - y\|_2 \leq \varepsilon$.*

First, we'll show that Σ_k has a small ε-covering, for $\varepsilon = \delta/4$. We'll start by showing that there's a small ε-covering for all of the vectors with a particular support S of size k, and then we'll take the union of all $\binom{n}{k}$ such coverings to get our final covering.

Lemma 3.

$$X = \left\{x \in \mathbb{R}^k \,:\, \|x\|_2 = 1\right\}$$

has an ε-covering of size at most $(3/\varepsilon)^k$.

Proof. For a point y, let the set $B(y;\varepsilon)$ denotes the ball of radius ε about y. We will construct the set $\mathcal{N} \subset X$ greedily as follows. While the set

$$X \setminus \left(\bigcup_{y \in \mathcal{N}} B(y;\varepsilon) \right),$$

is not empty, take any point $z \in X \setminus \left(\bigcup_{y \in \mathcal{N}} B(y;\varepsilon) \right)$ and add it to $\mathcal{N}$.

The set that we end up with at the end of this algorithm is definitely an ε-covering, so we just need to show that it is not too big. Notice that for all $y, y' \in \mathcal{N}$, we have $\|y - y'\|_2 \geq \varepsilon$. This is because if $\|y - y'\|_2 < \varepsilon$, we wouldn't have added both of them to our covering. Therefore, the set of balls $B(y;\varepsilon/2)$ for $y \in \mathcal{N}$ are disjoint. Moreover, all of these balls lie in the larger ball $B(0, 1+\varepsilon/2)$. This is because all of the centers $y \in \mathcal{N}$ have $\|y\|_2 = 1$. Then, we have

$$|\mathcal{N}| \cdot \mathrm{Vol}(B(0;\varepsilon/2)) \leq \mathrm{Vol}(B(0;1+\varepsilon/2)),$$

where Vol denotes the k-dimensional volume. If you've forgotten the formula for the k-dimensional volume of a ball in $\mathbb{R}^k$, the important thing is that it looks like $\mathrm{Vol}(B(0;\rho)) = C_k \cdot \rho^k$, for some constant C_k. Thus, the above implies that

$$|\mathcal{N}| \leq \frac{\mathrm{Vol}(B(0;1+\varepsilon/2))}{\mathrm{Vol}(B(0;\varepsilon/2))} = \left(\frac{1+\varepsilon/2}{\varepsilon/2} \right)^k \leq \left(\frac{3}{\varepsilon} \right)^k$$

for $\varepsilon < 1$. □

Corollary 4. *Let $\varepsilon \in (0,1)$. There is an ε-covering of Σ_k of size at most $\binom{n}{k}\left(\frac{3}{\varepsilon}\right)^k$.*

Let $\mathcal{N}$ be the ε-covering from Corollary 4, for $\varepsilon = \delta/4$. We'll use a union bound to show that with high probability, for any $y \in \mathcal{N}$, $\|y\|_2 \approx \|Ay\|_2$. More precisely, we can re-use the argument from the proof of the Johnson-Lindenstrauss lemma to say that, for any fixed $y \in \mathbb{R}^n$,

$$\Pr[\,|\,\|y\|_2 - \|Ay\|_2\,| > \varepsilon\,\|y\|_2] \leq 2\exp(-c \cdot \varepsilon^2 m),$$

for some constant c. (We won't write it out again here, but go back to those notes and check that this still applies). Thus, by a union bound over all $y \in \mathcal{N}$, we see that for any $y \in \mathcal{N}$, we have

$$(1-\varepsilon)\|y\|_2 \leq \|Ay\|_2 \leq (1+\varepsilon)\|y\|_2 \tag{1}$$

with probability at least

$$\begin{aligned}\binom{n}{k} \cdot (3/\varepsilon)^k \cdot 2\exp(-c\varepsilon^2 m) &\leq \exp\left(k\log n + k\log(3/\varepsilon) - c\varepsilon^2 m\right) \\ &\leq \exp(-\Omega(k\log n)),\end{aligned}$$

for some appropriate choice of $m = O\left(\frac{k\log n}{\varepsilon^2}\right)$. Let's assume that this event occurs, so that

$$|\,\|y\|_2 - \|Ay\|_2\,| \leq \varepsilon\,\|y\|_2 = \varepsilon \qquad \forall y \in \mathcal{N}. \tag{2}$$

Finally, we need to show that $\|Ay\|_2 \approx \|y\|_2$ for any $y \in \Sigma_k$, not just for $y \in \mathcal{N}$.

Suppose that δ^* is the smallest δ' so that $|\, \|Az\|_2 - \|z\|_2 \,| \leq \delta' \|z\|_2$ for any k-sparse vector z. (That is, our eventual goal is to show that $\delta^* \leq \delta$). Choose any $x \in \Sigma_k$, and let $y \in \mathcal{N}$ be such that $\|x - y\|_2 \leq \varepsilon$. Notice that we can choose y so that $x - y$ is k-sparse, using our construction of $\mathcal{N}$. Then,

$$\begin{aligned} |\, \|Ax\|_2 - \|x\|_2 \,| &\leq \|A(x-y)\|_2 + \|x-y\|_2 + |\, \|Ay\|_2 - \|y\|_2 \,| \\ &\leq (1+\delta^*)\varepsilon + \varepsilon + \varepsilon, \end{aligned}$$

using the triangle inequality in the first line, and the definition of δ^*, as well as (2) in the second line.

But, since δ^* is the *smallest* value that satisfies this, we have

$$\delta^* \leq (1+\delta^*)\varepsilon + 2\varepsilon,$$

and solving for δ^* this implies that

$$\delta^* \leq \frac{3\varepsilon}{1-\varepsilon} \leq 4\varepsilon = \delta$$

provided that ε is sufficiently small. This proves the theorem.

References

[1] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[2] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[3] Simon Foucart and Holger Rauhut. An invitation to compressive sensing. In *A mathematical introduction to compressive sensing*, pages 1–39. Springer, 2013.

[4] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 61(8):1025–1045, 2008.