MAT 585: Johnson-Lindenstrauss, Group testing, and Compressed Sensing

Afonso S. Bandeira

April 9, 2015

1 The Johnson-Lindenstrauss Lemma

Suppose one has n points, $X = \{x_1, \ldots, x_n\}$, in $\mathbb{R}^d$ (with d very large). If $d > n$, since the points have to lie in a subspace of dimension n it is clear that one can consider the projection $f : \mathbb{R}^d \to \mathbb{R}^n$ of the points to that subspace without distorting the geometry of X. In particular, for every x_i and x_j, $\|f(x_i) - f(x_j)\|^2 = \|x_i - x_j\|^2$, meaning that f is an isometry in X.

Suppose now we allow a bit of distortion, and look for $f : \mathbb{R}^d \to \mathbb{R}^k$ that is an $\epsilon-$isometry, meaning that

$$(1 - \epsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1 + \epsilon)\|x_i - x_j\|^2. \tag{1}$$

Can we do better than $k = n$?

In 1984, Johnson and Lindenstrauss [9] showed a remarkable Lemma (below) that answers this question positively.

Theorem 1 (Johnson-Lindenstrauss Lemma [9]) *For any $0 < \epsilon < 1$ and for any integer n, let k be such that*

$$k \geq 4\frac{1}{\epsilon^2/2 - \epsilon^3/3}\log n.$$

Then, for any set X of n points in $\mathbb{R}^d$, there is a linear map $f : \mathbb{R}^d \to \mathbb{R}^k$ that is an $\epsilon-$isometry for X (see (1)). This map can be found in randomized polynomial time.

We borrow, from [5], an elementary proof for the Theorem. Before we need a few concentration of measure bounds, we will omit the proof of those but they are available in [5] and are essentially the same ideas as those used to show Hoeffding's inequality.

Lemma 2 (see [5]) *Let $y_1, \ldots, y_d$ be i.i.d standard Gaussian random variables and $Y = (y_1, \ldots, y_d)$. Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be the projection into the first k coordinates and $Z = g\left(\frac{Y}{\|Y\|}\right) = \frac{1}{\|Y\|}(y_1, \ldots, y_k)$ and $L = \|Z\|^2$. It is clear that $\mathbb{E}L = \frac{k}{d}$. In fact, L is very concentrated around its mean*

- *If $\beta < 1$,*
$$\Pr\left[L \leq \beta\frac{k}{d}\right] \leq \exp\left(\frac{k}{2}(1 - \beta + \log\beta)\right).$$
- *If $\beta > 1$,*
$$\Pr\left[L \geq \beta\frac{k}{d}\right] \leq \exp\left(\frac{k}{2}(1 - \beta + \log\beta)\right).$$

Proof. [of Johnson-Lindenstrauss Lemma]

We will start by showing that, given a pair x_i, x_j a projection onto a random subspace of dimension k will satisfy (after appropriate scaling) property (1) with high probability. WLOG, we can assume that $u = x_i - x_j$ has unit norm. Understanding what is the norm of the projection of u on a random subspace of dimension k is the same as understanding the norm of the projection of a (uniformly) random point on S^{d-1} the unit sphere in $\mathbb{R}^d$ on a specific $k-$dimensional subspace, let's say the one generated by the first k canonical basis vectors. This means that we are interested in the distribution of the norm of the first k entries of a random vector drawn from the uniform distribution over S^{d-1} – this distribution is the same as taking a standard Gaussian vector in $\mathbb{R}^d$ and normalizing it to the unit sphere.

Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be the projection on a random $k-$dimensional subspace and let $f : \mathbb{R}^d \to \mathbb{R}^k$ defined as $f = \frac{d}{k}g$. Then (by the above discussion), given a pair of distinct x_i and x_j, $\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2}$ has the same distribution as $\frac{d}{k}L$, as defined in Lemma 2. Using Lemma 2, we have, given a pair x_i, x_j,

$$\Pr\left[\frac{\|f(x_i) - f(x_j)\|^2}{\|x_i - x_j\|^2} \leq (1-\epsilon)\right] \leq \exp\left(\frac{k}{2}(1 - (1-\epsilon) + \log(1-\epsilon))\right),$$

since, for $\epsilon \geq 0$, $\log(1-\epsilon) \leq -\epsilon - \epsilon^2/2$ we have

$$\begin{aligned}\Pr\left[\frac{\|f(x_i) - f(x_j)\|^2}{\|x_i - x_j\|^2} \leq (1-\epsilon)\right] &\leq \exp\left(-\frac{k\epsilon^2}{4}\right) \\ &\leq \exp\left(-2\log n\right) = \frac{1}{n^2}.\end{aligned}$$

On the other hand,

$$\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \geq (1+\epsilon)\right] \leq \exp\left(\frac{k}{2}(1-(1+\epsilon)+\log(1+\epsilon))\right).$$

since, for $\epsilon \geq 0$, $\log(1+\epsilon) \leq \epsilon - \epsilon^2/2 + \epsilon^3/3$ we have

$$\begin{aligned}\text{Prob}\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \leq (1-\epsilon)\right] &\leq \exp\left(-\frac{k\left(\epsilon^2-2\epsilon^3/3\right)}{4}\right)\\ &\leq \exp\left(-2\log n\right) = \frac{1}{n^2}.\end{aligned}$$

By union bound it follows that

$$\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \notin [1-\epsilon, 1+\epsilon]\right] \leq \frac{2}{n^2}.$$

Since there exist $\binom{n}{2}$ such pairs, again, a simple union bound gives

$$\Pr\left[\exists_{i,j} : \frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \notin [1-\epsilon, 1+\epsilon]\right] \leq \frac{2}{n^2}\frac{n(n-1)}{2} = 1-\frac{1}{n}.$$

Therefore, choosing f as a properly scaled projection onto a random $k-$dimensional subspace is an $\epsilon-$ isometry on X (see (1)) with probability at least $\frac{1}{n}$. We can achieve any desirable constant probability of success by trying $\mathcal{O}(n)$ such random projections, meaning we can find an $\epsilon-$isometry in randomized polynomial time.

□

Note that by considering k slightly larger one can get a good projection on the first random attempt with very good confidence. In fact, it's trivial to adapt the proof above to obtain the following Lemma:

Lemma 3 *For any $0 < \epsilon < 1$, $\tau > 0$, and for any integer n, let k be such that*

$$k \geq (2+\tau)\frac{2}{\epsilon^2/2-\epsilon^3/3}\log n.$$

Then, for any set X of n points in $\mathbb{R}^d$, take $f : \mathbb{R}^d \to \mathbb{R}^k$ to be a suitably scaled projection on a random subspace of dimension k, then f is an $\epsilon-$isometry for X (see (1)) with probability at least $1-\frac{1}{n^\tau}$.

Lemma 3 is quite remarkable. Think about the situation where we are given a high-dimensional data set in a streaming fashion – meaning that we get each data point at a time, consecutively. To run a dimension-reduction technique like PCA or Diffusion maps we would need to wait until we received the last data point and then compute the dimension reduction map (both PCA and Diffusion Maps are, in some sense, data adaptive). Using Lemma 3 you can just choose a projection at random in the beginning of the process (all ones needs to know is an estimate of the log of the size of the data set) and just map each point using this projection matrix which can be done online – we don't need to see the next point to compute the projection of the current data point. Lemma 3 ensures that this (seemingly naïve) procedure will, with high probably, not distort the data by more than ϵ.

1.1 Fast Johnson-Lindenstrauss

(Disclaimer: the purpose of this section is just to provide a bit of intuition, there is lots of hand-waving!!)

Let's continue thinking about the high-dimensional streaming data. After we draw the random projection matrix, say M, for each data point x, we still have to compute Mx which, since M has $\mathcal{O}(\epsilon^{-2}\log(n)d)$ entries, has a computational cost of $\mathcal{O}(\epsilon^{-2}\log(n)d)$. In some applications this might be too expensive, can one do better? It turns out that there is no hope of (significantly) reducing the number of rows (see this paper of Noga Alon [2]). The only hope is to speed up the matrix-vector multiplication. If we were able to construct a sparse matrix M then we would definitely speed up the computation of Mx but it is possible to show that a sparse matrix will distort sparse vectors, so if the data set contains sparse vectors, then M will fail. Another option would be to exploit the Fast Fourier Transform and compute the FT of x (which takes $\mathcal{O}(d\log d)$ time) and then multiply the FT of x by a sparse matrix. However, this will again not work because x might have a sparse Fourier Transform. The solution comes from leveraging an uncertainty principle — it is impossible for both x and the FT of x to be sparse simultaneously. The idea is that if, before one takes the Fourier Transform of x, one flips (randomly) the signs of x, then the probably of obtaining a sparse vector is very small so a sparse matrix can be used for projection. In a nutshell the algorithm has M be a matrix of the form PHD, where D is a diagonal matrix that flips the signs of the vector randomly, H is a Fourier Transform (or Hadamard transform) and P a sparse matrix. This method was proposed and analysed in [1] and, roughly speaking, achieves a

complexity of $\mathcal{O}(d \log d)$, instead of the classical $\mathcal{O}(\epsilon^{-2} \log(n) d)$.

2 Group Testing

During the Second World War the United States was interested in weeding out all syphilitic soldiers called up for the army. However, syphilis testing back then was expensive and testing every soldier individually would have been very costly and inefficient. A basic breakdown of a test is: 1) Draw sample from a given individual, 2) Perform required tests, and 3) Determine presence or absence of syphilis.

If there are n soldiers, this method of testing leads to n tests. If a significant portion of the soldiers were infected then the method of individual testing would be reasonable. The goal however, is to achieve effective testing in the more likely scenario where it does not make sense to test n (say $n = 100,000$) people to get k (say $k = 10$) positives.

Let's say that it was believed that there is only one soldier infected, then one could mix the samples of half of the soldiers and with a single test determined in which half the infected soldier is, proceeding with a binary search we could pinpoint the infected individual in $\log n$ tests. If instead of one, one believes that there are at most k infected people, then one could simply run k consecutive binary searches and detect all of the infected individuals in $k \log n$ tests. Which would still be potentially much less than n.

For this method to work one would need to observe the outcome of the previous tests before designing the next test, meaning that the samples have to be prepared adaptively. This is often not practical, if each test takes times to run, then it is much more efficient to run them in parallel (at the same time). This means that one has to non-adaptively design T tests (meaning subsets of the n individuals) from which it is possible to detect the infected individuals, provided there are at most k of them. Constructing these sets is the main problem in (Combinatorial) Group testing, introduced by Robert Dorfman [7].

Let A_i be a subset of $[T] = \{1, \ldots, T\}$ that indicates the tests for which soldier i participates. Consider $\mathbb{A}$ the family of n such sets $\mathbb{A} = \{A_1, \ldots, A_n\}$. We say that $\mathbb{A}$ satisfies the k-disjunct property if no set in $\mathbb{A}$ is contained in the union of k other sets in $\mathbb{A}$. A test set designed in such a way will succeed at identifying the (at most k) infected individuals – the set of infected tests is also a subset of $[T]$ and it will be the union of the A_i's that correspond to the infected soldiers. If the set of infected tests contains a certain A_i then

this can only be explained by the soldier i being infected (provided that there are at most k infected people).

Theorem 4 *Given n and k, there exists a family $\mathbb{A}$ satisfying the k-disjunct property for a number of tests*

$$T = \mathcal{O}\left(k^2 \log n\right).$$

Proof. We will use the probabilistic method. We will show that, for $T = Ck^2 \log n$ (where C is a universal constant), by drawing the family $\mathbb{A}$ from a (well-chosen) distribution gives a $k-$disjunct family with positive probability, meaning that such a family must exist (otherwise the probability would be zero).

Let $0 \leq p \leq 1$ and let $\mathbb{A}$ be a collection of n random (independently drawn) subsets of $[T]$. The distribution for a random set A is such that each $t \in [T]$ belongs to A with probability p (and independently of the other elements).

Consider $k+1$ independent draws of this random variable, $A_0, \ldots, A_k$. The probability that A_0 is contained in the union of A_1 through A_k is given by

$$\Pr\left[A_0 \subseteq (A_1 \cup \cdots \cup A_k)\right] = \left(1 - p(1-p)^k\right)^T.$$

This is minimized for $p = \frac{1}{k+1}$. For this choice of p, we have

$$1 - p(1-p)^k = 1 - \frac{1}{k+1}\left(1 - \frac{1}{k+1}\right)^k$$

Given that there are n such sets, there are $(k+1)\binom{n}{k+1}$ different ways of picking a set and k others to test whether the first is contained in the union of the other k. Hence, using a union bound argument, the probability that $\mathbb{A}$ is k-disjunct can be bounded as

$$\Pr[k-\text{disjunct}] \geq 1 - (k+1)\binom{n}{k+1}\left(1 - \frac{1}{k+1}\left(1 - \frac{1}{k+1}\right)^k\right)^T.$$

In order to show that one of the elements in $\mathbb{A}$ is k-disjunct we show that this probability is strictly positive. That is equivalent to

$$\left(1 - \frac{1}{k+1}\left(1 - \frac{1}{k+1}\right)^k\right)^T \leq \frac{1}{(k+1)\binom{n}{k+1}}.$$

Note that $\left(1-\frac{1}{k+1}\right)^k \to e^{-1}\frac{1}{1-\frac{1}{k+1}} = e^{-1}\frac{k+1}{k}$, as $k \to \infty$. Thus, we only need

$$T \geq \frac{\log\left((k+1)\binom{n}{k+1}\right)}{-\log\left(1-\frac{1}{k+1}e^{-1}\frac{k+1}{k}\right)} = \frac{\log\left(k\binom{n}{k+1}\right)}{-\log\left(1-(ek)^{-1}\right)} = \mathcal{O}(k^2\log(n/k)),$$

where the last inequality uses the fact that $\log\left(\binom{n}{k+1}\right) = \mathcal{O}\left(k\log\left(\frac{n}{k}\right)\right)$ due to Stirling's formula and the Taylor expansion $-\log(1-x^{-1})^{-1} = \mathcal{O}(x)$ $\square$

This argument simply shows the existence of a family satisfying the k-disjunt property. However, it is easy to see that by having T slightly larger one can ensure that the probability that the random family satisfies the desired property can be made very close to 1.

Remarkably, the existence proof presented here is actually very close to the best known lower bound, which is $\Omega\left(\frac{k^2\log(n)}{\log k}\right)$ (meaning that it is impossible to build such a family for values of T smallers than this, see [8]). The tightest possible bound still remain an open problem.

2.1 In terms of ~~linear~~ Bernoulli algebra

We can describe the process above in terms of something similar to a linear system. let 1_{A_i} be the $t-$*dimensional* indicator vector of A_i, $1_{i:n}$ be the (unknown) $n-$dimensional vector of infected soldiers and $1_{t:T}$ the $T-$dimensional vector of infected (positive) tests. Then

$$\begin{bmatrix} | & & | \\ 1_{A_1} & \cdots & 1_{A_n} \\ | & & | \end{bmatrix} \otimes \begin{bmatrix} | \\ | \\ 1_{i:n} \\ | \\ | \end{bmatrix} = \begin{bmatrix} | \\ 1_{t:T} \\ | \end{bmatrix},$$

where $\otimes$ is matrix-vector multiplication in the Bernoulli algebra, basically the only thing that is different from the standard matrix-vector multiplications is that the addition operation is replaced by binary "or", meaning $1 \oplus 1 = 1$.

This means that we are essentially solving a linear system (with this non-standard multiplication). Since the number of rows is $T = \mathcal{O}(k^2\log(n/k))$ and the number or columns $n \gg T$ the system is highly underdetermined. What allows us to solve the system is the knowledge that the unknown vector, $1_{i:n}$ has only k non-zero components, meaning it is $k-$sparse.

3 Compressed Sensing

(Disclaimer: Once again, there is going to be a fair amount of hand-waving in this section!!)

The problem of sparse recovery consists of solving a (highly) underdetermined linear system

$$Ax = y, \tag{2}$$

where A is a $T \times n$ matrix, $y \in \mathbb{R}^T$ and x is an unknown vector in $\mathbb{R}^n$ that is known to be $k-$sparse, meaning that x has at most k non-zero entries.

In the early 2000's there were a few seminal papers by Emmanuel Candes, Terence Tao, and David Donoho [4, 6] that essentially showed that one can have $T = \mathcal{O}\left(k \log \frac{n}{k}\right) \ll n$ and still be able to efficiently solve the linear system. This has both inspired substantial mathematical and algorithmic development and also impacted a paradigm shift for many areas of application. Perhaps the most well known application is MRI imaging [10] with the goal of reducing the imaging time by exploiting sparsity of the signal in a suitable basis.

What follows is an intuitive explanation (that can, in fact, be made precise [3]) for why $T = \mathcal{O}\left(k \log \frac{n}{k}\right)$ linear measurements of a $k-$sparse vector suffice using the Johnson-Lindenstrauss Lemma.

For one to have hope to recover x from the linear measurements, meaning solve (2), then A must be injective on $k-$sparse vectors (otherwise the solution x may not be unique). If A is indeed injective on $k-$sparse vectors, then

$$\min \|x\|_0 \text{ subject to } Ax = y,$$

will recover x. Note that $\|\cdot\|_0$ is the ℓ_0 "norm" which is the number of non-zero components. However this minimization is known to be NP-hard. The idea is to substitute the ℓ_0 by ℓ_1 (the sum of the absolute values, which is known to promote sparsity) and to solve

$$\min \|x\|_1 \text{ subject to } Ax = y, \tag{3}$$

which is equivalent to a linear program.

One can show that in order for (3) to give the correct x it suffices that A not only is injective on $k-$sparse vectors but that it roughly maintains the distance between any pair of $k-$sparse vectors, which is known as the Restricted Isometry Property.

Definition 5 (Restricted Isometry Property) *A matrix $A \in \mathbb{R}^{T \times n}$ is said to satisfy the $(2k, \delta)$ restricted isometry property if*

$$(1-\delta)\|x_1 - x_2\|^2 \leq \|Ax_1 - Ax_2\|^2 \leq (1+\delta)\|x_1 - x_2\|^2,$$

for all two k-sparse vectors x_1 and x_2.

Using Johnson-Lindenstrauss one can show that such matrices exist for $T = \Omega\left(k \log \frac{n}{k}\right)$ (for the rigorous proof see [3]). More precisely, it can be shown that taking a matrix $A \in \mathbb{R}^{T \times n}$ with i.i.d. gaussian distributed entries for $T = \Omega\left(k \log \frac{n}{k}\right)$ it satisfies the Restricted Isometry Property with high probability. The idea is that it is enough to show that A roughly maintains the distances of unit-norm $k-$sparse vectors, meaning that for each sparsity pattern one only has to look at the unit sphere in k dimensions. One can discretize such a sphere using $(3/\delta)^k$ in a way that any point in the sphere is at most δ away from a point in the discrete set (the δ-net). One can further show that it suffices for A not to distort this finite set of points. Since there are $\binom{n}{k}$ different sparsity patterns the discretization has $\binom{n}{k}(3/\delta)^k$ points.

The Johnson-Lindenstrauss Lemma tells us that an A that does not distort by more than ϵ exists for

$$T \geq \mathcal{O}\left[\epsilon^{-2} \log\left(\binom{n}{k}(3/\delta)^k\right)\right] = \mathcal{O}\left[\epsilon^{-2} k \left(\log \frac{n}{k} + \log \frac{3}{\delta}\right)\right] = \mathcal{O}\left(k \log \frac{n}{k}\right),$$

if one considers ϵ and δ fixed.

Remark 6 *For Compressed Sensing, one can do non-adaptive sensing with $\mathcal{O}\left(k \log \frac{n}{k}\right)$ measurements but for Group Testing one (provably) needs $\mathcal{O}\left(k^2 \log \frac{n}{k}\right)$.*

References

[1] Nir Ailon and Bernard Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput*, pages 302–322, 2009.

[2] Noga Alon. Problems and results in extremal combinatorics, part I. *Discrete Math*, 273:2003.

[3] R. Baraniuk et al. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28:253–263, 2008.

[4] E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theor.*, 52(12):5406–5425, December 2006.

[5] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of the Johnson-Lindenstrauss Lemma. Technical Report TR-99-006, Berkeley, CA, 1999.

[6] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[7] R. Dorfman. The detection of defective members of large populations. 1943.

[8] Z. Füredi. On r-cover-free families. *Journal of Combinatorial Theory, Series A*, 1996.

[9] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.

[10] Wired Magazine. `http://www.wired.com/magazine/2010/02/ff_algorithm/`. 2010.