MIT 6.854/18.415: Advanced Algorithms Spring 2016

Lecture 5 – February 17, 2016

Prof. Ankur Moitra *Scribe: Jiaming Luo, Mingmin Zhao*

This lecture is on dimensionality reduction, which aims at 'squashing' down the dimensionality while still preserving some geometric properties. The motivation behind this technique is that many types of data are high-dimensional, and it will be operationally much easier to manipulate these data in a lower dimension. For instance, the bag-of-words representation of documents, which treats every document as a vector of word counts, can easily have tens of thousands of dimensions. In domains where images or videos are the subjects of interest, the dimension is even greater.

1 Setup

The geometric property of interest in this lecture is Euclidean distance: we want to preserve all-pairs distances of n vectors of length d. Naively, computing all distances would take $O(n^2 d)$ time, spending $O(d)$ time to compute the distance for each of $\binom{n}{2} = O(n^2)$ pairs of vectors. The Johnson-Lindenstrauss lemma, the main result from this lecture, can immediately brings the runtime down to $O(dnk + n^2 k)$ for $k = O(\frac{\log n}{\epsilon^2})$ to get ϵ multiplicative error with high probability. The basic idea is that we first embed all the vectors into a space of dimensionality $k << d$. This embedding takes $O(dnk)$ time, and we then only need to compute $\binom{n}{2}$ distances between k-dimensional vectors, taking an additional $O(n^2 k)$ time .

2 Johnson-Lindenstrauss Theorem

To answer the main question, we have the following theorem.

2.1 JL Theorem

Theorem 1 (Johnson; Lindenstrauss [1])**.** *For any set S of n-points in d-dimensions, there is a matrix $A \in \mathbb{R}^{k \times d}$ with*

$$\forall u, v \in S, \qquad (1-\epsilon)\|u-v\|_2^2 \leq \|Au - Av\|_2^2 \leq (1+\epsilon)\|u-v\|_2^2,$$

with $k = O\left(\frac{\log n}{\epsilon^2}\right)$.

Interestingly, this result has only a logarithmic dependence on n and *no* dependence on d.

Another comment is that as we shall see promptly, we can actually choose the matrix A from a random distribution (in particular, independent Gaussian entries) that does not depend on the actual point-set S. This also implies that we can (with high probability) find such an A efficiently.

2.2 Norm preservation

A key observation is that JL can reduce to a randomized norm preservation property. Suppose that A is a random matrix, which satisfies, for any *fixed* x,

$$\mathbb{P}[(1-\epsilon)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\epsilon)\|x\|_2^2] \geq 1-\delta \tag{1}$$

Then, for $\delta = \frac{1}{n^3}$, union bounding over $x = u - v$ for all $u, v \in S$ gives us JL. The number of u, v pairs is $< n^2$, so $\delta = \frac{1}{n^3}$ makes this satisfied with probability at least $1 - \frac{1}{n}$ (and shrinking δ boosts the probability further).

We will obtain this norm preservation for $k = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, obtaining the desired $O\left(\frac{\log n}{\epsilon^2}\right)$ to get high probability for JL.

2.3 Digress

Before delving into the main proof, we include a little side note on the Gaussian distribution:

$$\mathcal{N}(\mu, \sigma^2, x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where μ is the mean, and σ^2 is the variance.

The central limit theorem has contributed to the popularity of Gaussian distributions, which states

Theorem 2 (Central limit theorem)**.** *Sums of independent random variables (with the right normalization), converge to Gaussian distributions.*

Another particularly handy fact is that for independent Gaussian random variables z_1, z_2,

Fact 3. *If $z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then $z_1 + z_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

2.4 The proof

Now let's head back to the proof. Choose entries of A independently from $\mathcal{N}(0, \frac{1}{k})$. Then

$$((Ax)_1)^2 = (\sum_{i=1}^{d} A_{1i} x_i)^2)$$

and the same for all other entries of Ax. Note that by Fact 3 above, the summation inside the bracket follows the Gaussian distribution $\mathcal{N}(0, \frac{\sum_{i=1}^{d} x_i^2}{k})$, so

$$\mathbb{E}[\|Ax\|_2^2] = k\mathbb{E}[((Ax)_1)^2] = k \cdot \frac{\|x\|_2^2}{k} = \|x\|_2^2$$

This implies that this embedding provides an unbiased estimator for the norm of x. However, we actually care about the deviation probabilities from this mean.

Specifically, we have:

$$\mathbb{P}[\|Ax\|_2^2 > (1+\epsilon)\|x\|_2^2] = \mathbb{P}[\sum_{i=1}^{k} z_i^2 > (1+\epsilon)\|x\|_2^2]$$

where $z_i \sim \mathcal{N}(0, \frac{\|x\|_2^2}{k})$. and similarly for the lower tail.

By dividing through by $\frac{\|x\|_2^2}{k}$, we get $\mathbb{P}[\sum_{i=1}^{k} Y_i^2 > (1+\epsilon)k]$, where $Y_i \sim \mathcal{N}(0,1)$.

Lemma 4 (Chernoff bound for chi-square distributions)**.**

$$\mathbb{P}[\sum_{i=1}^{k} Y_i^2 > (1+\epsilon)k] \leq e^{-\frac{k}{4}(\epsilon^2-\epsilon^3)}$$
$$\mathbb{P}[\sum_{i=1}^{k} Y_i^2 < (1-\epsilon)k] \leq e^{-\frac{k}{4}(\epsilon^2-\epsilon^3)}$$

Now, we may set $k = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ and get $\|Ax\|_2^2 \underset{1\pm\epsilon}{\sim} \|x\|_2^2$ with probability at least $1-\delta$.

3 Extensions

What really happened in the proof is that for i.i.d, mean-zero, variance-one $z_1, z_2, \ldots, z_d$, the quantity $(\sum_{i=1}^{d} z_i x_i)^2$ is an unbiased estimator for $\|x\|_2^2$. We are going to list out some interesting extensions of JL lemma in this section.

Theorem 5 (Achlioptas [2])**.** *If the entries of A are chosen as, for all $1 \leq i \leq k, 1 \leq j \leq d$,*

$$R_{ij} = \sqrt{3} \cdot \begin{cases} +1 & \text{with probability } \frac{1}{6} \\ 0 & \text{with probability } \frac{2}{3} \\ -1 & \text{with probability } \frac{1}{6} \end{cases}$$

Then $A = \frac{1}{k}R$ has the JL property. In other words, fully dense matrices are not actually needed for JL embedding.

The idea of finding sparse dimension reduction can be taken further, as illustrated by the following theorem.

Theorem 6 (Kane, Nelson [3])**.** *There are distributions on $A \in \mathcal{R}^{k\times d}$, with $k = O(\frac{\log n}{\epsilon^2})$ and $s = O(\frac{\log n}{\epsilon})$ non-zeros per column, satisfying the JL property.*

This value of s is tight (without increasing k by more than a constant factor), and obtaining it requires using a matrix with entries that are not i.i.d.

A fast way to compute embeddings is given by

Theorem 7 (Fast Johnson-Lindenstrauss transformation (FJLT), Ailon, Chazelle, 2006)**.**

$$A = PHD$$

P is sparse, H is a Hadamard, and D is diagonal.

$$H_1 = \begin{bmatrix}1\end{bmatrix}, H_2 = \begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}, H_{2^r} = \begin{bmatrix}H_{2^{r-1}} & H_{2^{r-1}}\\ H_{2^{r-1}} & -H_{2^{r-1}}\end{bmatrix}, D = \begin{bmatrix}\pm 1 & & & \\ & \pm 1 & & \\ & & \ddots & \\ & & & \pm 1\end{bmatrix}$$

Overall time complexity is $O(d \log d + \frac{d \log n}{\epsilon^2})$ for FJLT. The idea is that multiplying a sparse vector by a sparse matrix can result in a poor embedding, so instead we make vectors dense. FJLT relies on the fact that multiplying by H can be done by divide-and-conquer, and *uncertainty principles*.

References

[1] Johnson, W. B. and Lindenstrauss, J. 1984, Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26:189-206.

[2] Achlioptas, D., 2003. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4), pp.671-687.

[3] Kane, D.M. and Nelson, J., 2014. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1), p.4.