Dimension Reduction and Surprises in High Dimensions

Orestis Plevrakis

1 High Dimensional Data and Dimension Reduction

It is a typical case to have data which are stored as vectors with a large number of coordinates. For example, grayscale images are stored as matrices, where the (i, j) entry corresponds to how much white or black the pixel (i, j) is. An image with 1280×780 pixels is stored as a matrix in $\mathbb{R}^{1280 \times 780}$. If we view this matrix as a vector by concatenating its columns we will get a vector with around a million coordinates. Another example appears in bioinformatics, where data can be DNA sequences. These are very long sequences consisting of the 4 DNA bases (A,T,G,C).

```
CTGGGGCTTTACTGATGTCATACCGTCTTGCACGGGGATAGAAT
ATTTTCTGAAAGTTACAGACTTCGATTAAAAAGATCGGACTGCG
TTTTTCGACGTGTCAAGGACTCAAGGGAATAGTTTGGCGGGAGC
CGATAAAATTCAACTACTGGTTTCGGCCTAATAGGTCACGTTTT
CCCTGGGTGTTCTATGATAAGTCCTGCTTTATAACACGGGGCGG
ATCCAAGCGCCCGCTAATTCTGTTCTGTTAATGTTCATACCAAT
AGCCCAGTCGCAAGGGTCTGCTGCTGTTGTCGACGCCTCATGTT
GGTTAAGGCGTGTGATCGACGATGCAGGTATACATCGGCTCGGA
TCGCGGTTCGGCGCGTAGTTGAGTGCGATAACCCAACCGGTGGC
AGACAACCTAACTAATAGTCTCTAACGGGGAATTACCTTTACCA
CAATGATATCGCCCACAGAAAGTAGGGTCTCAGGTATCGCATAC
GACAGTAGAGAGCTATTGTGTAATTCAGGCTCAGCATTCATCGA
```

Figure 1: Example of a (short) DNA sequence

We often use numeric encodings for the bases, e.g., 0 for T, 1 for C, 2 for A, 3 for G, and so the sequence is represented as a large vector.

The high dimension creates running-time issues, so we would like to *compress* our data, i.e., reduce the dimension, while maintaining important information. In this lecture, we will focus on dimension-reduction that preserves the distances among the data. To check if this is possible to do, we first formulate it mathematically:

> **Question:** Is it true that for any $v_1, v_2, \ldots, v_n \in \mathbb{R}^d$, there exists an integer $\kappa < d$ and $\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_n \in \mathbb{R}^\kappa$ such that for all $i \neq j$, $\|\tilde{v}_i - \tilde{v}_j\| = \|v_i - v_j\|$?

The answer is no, and there is a very simple counterexample in every dimension d. Before presenting it, we first make a remark:

Remark 1. If the answer to the question was yes, then we could also add to the required properties of $\tilde{v}_i$'s that all norms and inner products are preserved. Here is why: consider the points $v_1, \ldots, v_n, 0 \in \mathbb{R}^d$. If there exist $\kappa < d$ and $\tilde{v}_1, \ldots, \tilde{v}_{n+1} \in \mathbb{R}^\kappa$ such that all distances are preserved, then we can update $\tilde{v}_i \leftarrow \tilde{v}_i - \tilde{v}_{n+1}$, and both distances and norms will be preserved. For the inner products, notice that for the new $\tilde{v}_i$'s we have $\|\tilde{v}_i - \tilde{v}_j\|^2 = \|v_i - v_j\|^2 \iff \|\tilde{v}_i\|^2 + \|\tilde{v}_j\|^2 - 2\tilde{v}_i \cdot \tilde{v}_j = \|v_i\|^2 + \|v_j\|^2 - 2v_i \cdot v_j$.

Thus, if the answer was yes, then for the standard basis $e_1, \ldots, e_d$, there exist $\kappa < d$, $\tilde{e}_1, \ldots, \tilde{e}_d \in \mathbb{R}^\kappa$ which are pairwise orthogonal unit vectors. This implies that they are linearly independent, which is a contradiction. Thus, even reducing the dimension by one ($\kappa = d - 1$), while preserving the distances, is impossible. Let's relax our goals: can we reduce the dimension while approximating the distances with up to 1% error? The answer is yes, and surprisingly, we can choose κ to be as small as $O(\log d)$! This is a theorem proved by Johnson and Lindenstrauss and it is called "JL lemma" (since the authors called it lemma in their paper).

Theorem 2. *Let $n \le poly(d)$ and $v_1, \dots, v_n \in \mathbb{R}^d$. Let $\epsilon \in (0,1)$. Then, there exists a $\kappa = O\left(\frac{\log d}{\epsilon^2}\right)$ and $\tilde{v}_1, \dots, \tilde{v}_n \in \mathbb{R}^\kappa$ such that for all $i \neq j$,*

$$(1-\epsilon)\|v_i - v_j\| \le \|\tilde{v}_i - \tilde{v}_j\| \le (1+\epsilon)\|v_i - v_j\| \tag{1}$$

As we will see in the proof, these $\tilde{v}_i$'s can also be computed efficiently.

Remark 3. Notice that by repeating the argument in Remark 1, we have that Theorem 2 could have been stated with the additional property that for all i, $(1-\epsilon)\|v_i\| \le \|\tilde{v}_i\| \le (1+\epsilon)\|v_i\|$.

2 A Special Case

As usual in problem-solving, we won't attack the theorem head-on; we will first prove a special case. Which special case? We focus on the one that showed that exact distance preservation is impossible: the standard basis!

Special case: Let $d \ge 1$ be an integer, and $\epsilon > 0$. There exists a $\kappa = O\left(\frac{\log d}{\epsilon^2}\right)$ and $\tilde{e}_1, \dots, \tilde{e}_d \in \mathbb{R}^\kappa$ such that for all $i \neq j$,

$$(1-\epsilon)\sqrt{2} \le \|\tilde{e}_i - \tilde{e}_j\| \le (1+\epsilon)\sqrt{2} \tag{2}$$

and for all i,

$$1-\epsilon \le \|\tilde{e}_i\| \le 1+\epsilon \tag{3}$$

In the proof of the special case, we will choose a $\kappa = \lfloor \frac{C \log d}{\epsilon^2} \rfloor$ where C will be a large constant. Before giving the proof, we want to highlight a surprising consequence. Take $\epsilon = 0.01$. Using cosine law, it can be shown that (2) and (3) imply that any pair $\tilde{e}_i, \tilde{e}_j$ (with $i \neq j$) forms an angle that is between 87° and 93°, i.e., almost 90°. At the same time, $d = e^{\Omega(k)}$. This implies the following (why?):

Exponentially many nearly orthogonal vectors in high dimensions: There exists an absolute constant $c > 0$, such that the following holds: for any integer $\kappa \ge 1$, there are at least $e^{\kappa/c}$ vectors in $\mathbb{R}^\kappa$ whose pairwise angles are all between 87° and 93°.

To compare with what happens in low dimensions, let A_κ be the maximum number of vectors in $\mathbb{R}^\kappa$ whose pairwise angles are all between 87° and 93°. It can be shown that A_2 and A_3 are just 2 and 3 respectively! Finally, note that for larger c, the interval around 90° will be smaller. Let's now prove the special case:

Proof. First, some notation. For a random variable X, we write $X \sim \{\pm 1\}$ to denote that $X = 1$ with probability 1/2 and $X = -1$ with probability 1/2. We will use probabilistic method. Let $C > 0$ be a large constant that we will choose later, and let $\kappa = \lfloor \frac{C \log d}{\epsilon^2} \rfloor$. We generate all $\tilde{e}_i$

randomly, by first sampling independent and indentically distributed random variables $g_{i\ell} \sim \{\pm 1\}$, for $i = 1, \dots, d$ and $\ell = 1, \dots, \kappa$, and then setting $\tilde{e}_i = \frac{1}{\sqrt{\kappa}}(g_{i1}, \dots, g_{i\kappa})$. Now, fix a pair $i \neq j$. We will show that $\tilde{e}_i, \tilde{e}_j$ are nearly orthogonal.

$$\tilde{e}_i \cdot \tilde{e}_j = \frac{1}{\kappa} \sum_{\ell=1}^{\kappa} g_{i\ell} \cdot g_{j\ell}$$

The products $W_\ell := g_{i\ell} \cdot g_{j\ell}$ are independent, identically distributed and $W_\ell \sim \{\pm 1\}$. From the law of large numbers, we expect that $\tilde{e}_i \cdot \tilde{e}_j \approx 0$. This is quantified by Hoeffding's inequality:

Theorem 4. *Let $W_1, \dots, W_\kappa$ random variables taking values in $[a, b]$. Let $M := \frac{1}{\kappa} \sum_{\ell=1}^{\kappa} W_\ell$. Then,*

$$\mathbb{P}(|M - \mathbb{E}[M]| \geq t) \leq 2 \exp\left(-\frac{2\kappa t^2}{(b-a)^2}\right)$$

for all $t > 0$.

By direct application, we get $\mathbb{P}(|\tilde{e}_i \cdot \tilde{e}_j| \geq \epsilon) \leq 2 \exp\left(-\frac{\kappa \epsilon^2}{2}\right)$. By union bound,

$$\mathbb{P}(\exists i \neq j : \ |\tilde{e}_i \cdot \tilde{e}_j| \geq \epsilon) \leq \binom{d}{2} \cdot 2 \exp\left(-\frac{\kappa \epsilon^2}{2}\right)$$

Since $\kappa = \lfloor \frac{C \log d}{\epsilon^2} \rfloor \geq \frac{C \log d}{2\epsilon^2}$, we have that the above bound is at most $\binom{d}{2} \cdot 2 \cdot d^{-C/4} \leq 1/d$ for $C = 1/12$. Thus, with high probability, for all $i \neq j$, $|\tilde{e}_i \cdot \tilde{e}_j| < \epsilon$, and since $\|\tilde{e}_i - \tilde{e}_j\|^2 = 2(1 - \tilde{e}_i \cdot \tilde{e}_j) \in [2(1-\epsilon), 2(1+\epsilon)]$ and $1 - \epsilon < \sqrt{1-\epsilon} < 1 < \sqrt{1+\epsilon} < 1 + \epsilon$, we are done. $\square$

Next time, we will use the special case, to prove the general theorem.