High-dimensional vectors are ubiquitous in applications (gene expression data, set of movies watched by Netflix customer, etc.) and this lecture seeks to introduce some common properties of these vectors. We encounter the so-called *curse of dimensionality* which refers to the fact that algorithms are simply harder to design in high dimensions and often have a running time exponential in the dimension. We also encounter the *blessings of dimensionality*, which allows us to reason about higher dimensional geometry using tools like Chernoff bounds. We also show the possibility of *dimension reduction* — it is sometimes possible to reduce the dimension of a dataset, for some purposes.

Notation: For a vector $x \in \Re^d$ its ℓ_2-*norm* is $|x|_2 = (\sum_i x_i^2)^{1/2}$ and the ℓ_1-norm is $|x|_1 = \sum_i |x_i|$. For any two vectors x, y their Euclidean distance refers to $|x - y|_2$ and Manhattan distance refers to $|x - y|_1$.

High dimensional geometry is inherently different from low-dimensional geometry.

EXAMPLE 1 Consider how many *almost orthogonal* unit vectors we can have in space, such that all pairwise angles lie between 88 degrees and 92 degrees.

In $\Re^2$ the answer is 2. In $\Re^3$ it is 3. (Prove these to yourself.)
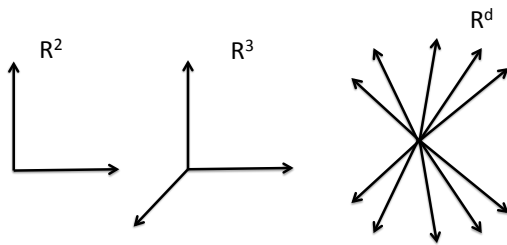
In $\Re^d$ the answer is $\exp(cd)$ where $c > 0$ is some constant.

Figure 1: Number of almost-orthogonal vectors in $\Re^2, \Re^3, \Re^d$

EXAMPLE 2 Another example is the ratio of the the volume of the unit sphere to its circumscribing cube (i.e. cube of side 2). In $\Re^2$ it is $\pi/4$ or about 0.78. In $\Re^3$ it is $\pi/6$ or about 0.52. In d dimensions it is $\exp(-c \cdot d \log d)$.

Let's start with useful generalizations of some geometric objects to higher dimensional geometry:

- The *n-cube* in $\Re^n$: $\{(x_1...x_n) : 0 \le x_i \le 1\}$. To visualize this in $\Re^4$, think of yourself as looking at one of the faces, say $x_1 = 1$. This is a cube in $\Re^3$ and if you were able to look in the fourth dimension you would see a parallel cube at $x_1 = 0$. The visualization in $\Re^n$ is similar.

  The volume of the n-cube is 1.

- The unit *n-ball* in $\Re^d$: $B_d := \{(x_1...x_d) : \sum x_i^2 \le 1\}$. Again, to visualize the ball in $\Re^4$, imagine you have sliced through it with a hyperplane, say $x_1 = 1/2$. This slice is a ball in $\Re^3$ of radius $\sqrt{1 - 1/2^2} = \sqrt{3}/2$. Every parallel slice also gives a ball.

  The volume of B_d is $\frac{\pi^{d/2}}{(d/2)!}$ (assume d is even if the previous expression bothers you), which is $\frac{1}{d^{\Theta(d)}}$.

- In $\Re^2$, if we slice the unit ball (i.e., disk) with a line at distance $1/2$ from the center then a significant fraction of the ball's volume lies on each side. In $\Re^d$ if we do the same with a hyperplane, then the radius of the $d-1$ dimensional ball is $\sqrt{3}/2$, and so the volume on the other side is negligible. In fact a constant fraction of the volume lies within a slice at distance $1/\sqrt{d}$ from the center, and for any $c > 1$, a $(1-1/c)$-fraction of the volume of the d-ball lies in a strip of width $O(\sqrt{\frac{\log c}{d}})$ around the center.

- A good approximation to picking a random point on the surface of B_n is by choosing random $x_i \in \{-1, 1\}$ independently for $i = 1..n$ and normalizing to get $\frac{1}{\sqrt{n}}(x_1, ..., x_n)$.

  An exact way to pick a random point on the surface of B^n is to choose x_i from the standard normal distribution for $i = 1..n$, and to normalize: $\frac{1}{l}(x_1, ..., x_n)$, where $l = (\sum_i x_i^2)^{1/2}$.

1   Number of almost-orthogonal vectors

Now we show there are $\exp(d)$ vectors in $\Re^d$ that are almost-orthogonal. Recall that the angle between two vectors x, y is given by $\cos(\theta) = \langle x, y \rangle / |x|_2 |y|_2$.

LEMMA 1
Suppose a is a unit vector in $\Re^n$. Let $x = (x_1, ..., x_n) \in R^n$ be chosen from the surface of B_n by choosing each coordinate at random from $\{1, -1\}$ and normalizing by factor $\frac{1}{\sqrt{n}}$. Denote by X the random variable $a \cdot x = \sum a_i x_i$. Then:

$$Pr(|X| > t) < e^{-nt^2}$$

PROOF: We have:

$$\mu = E(X) = E(\sum a_i x_i) = 0$$

$$\sigma^2 = E[(\sum a_i x_i)^2] = E[\sum a_i a_j x_i x_j] = \sum a_i a_j E[x_i x_j] = \sum \frac{a_i^2}{n} = \frac{1}{n}.$$

Using the Chernoff bound, we see that,

$$Pr(|X| > t) < e^{-(\frac{t}{\sigma})^2} = e^{-nt^2}.$$

□

COROLLARY 2

If x, y are chosen at random from $\{-1, 1\}^n$, and the angle between them is $\theta_{x,y}$ then

$$Pr\left[|cos(\theta_{x,y})| > \sqrt{\frac{\log c}{n}}\right] < \frac{1}{c}.$$

Hence by if we pick say $\sqrt{c}/2$ random vectors in $\{-1, 1\}^n$, the union bound says that the chance that they all make a pairwise angle with cosine less than $\sqrt{\frac{\log c}{n}}$ is less than $1/2$. Hence we can make $c = \exp(0.01n)$ and still have the vectors be almost-orthogonal (i.e. cosine is a very small constant).

2 Curse of dimensionality

Suppose we have a set of vectors in d dimensions and given another vector we wish to determine its closest neighbor (in ℓ_2 norm) to it. Designing fast algorithms for this in the plane (i.e., $\Re^2$) uses the fact that in the plane there are only $O(1)$ distinct points whose pairwise distance is about $1 \pm \varepsilon$. In $\Re^d$ there can be $\exp(d)$ such points.

Thus most algorithms —nearest neighbor, minimum spanning tree, point location etc.— have a running time depending upon $\exp(d)$. This is the *curse of dimensionality* in algorithms. (The term was invented by R. Bellman, who as we saw earlier had a knack for giving memorable names.)

In machine learning and statistics sometimes the term refers to the fact that available data is too sparse in high dimensions; it takes $\exp(d)$ amount of data (say, points on the sphere)to ensure that each new sample is close to an existing sample. This is a different take on the same underlying phenomenon.

I hereby coin a new term: *Blessing of dimensionality.* This refers to the fact that many phenomena become much clearer and easier to think about in high dimensions because one can use simple rules of thumb (e.g., Chernoff bounds, measure concentration) which don't hold in low dimensions.

3 Dimension Reduction

Now we describe a central result of high-dimensional geometry (at least when distances are measured in the ℓ_2 norm). Problem: Given n points $z^1, z^2, ..., z^n$ in $\Re^n$, we would like to find n points $u^1, u^2, ..., u^n$ in $\Re^m$ where m is of low dimension (compared to n) and the metric restricted to the points is almost preserved, namely:

$$\|z^i - z^j\|_2 \leq \|u^i - u^j\|_2 \leq (1 + \varepsilon)\|z^j - z^j\|_2 \; \forall i, j. \tag{1}$$

The following main result is by Johnson & Lindenstrauss :

THEOREM 3

In order to ensure (1), $m = O(\frac{\log n}{\varepsilon^2})$ suffices, and in fact the mapping can be a linear mapping.

The following ideas do not work to prove this theorem (as we discussed in class): (a) take a random sample of m coordinates out of n. (b) Partition the n coordinates into m subsets of size about n/m and *add* up the values in each subset to get a new coordinate.

PROOF: Choose m vectors $x^1, ..., x^m \in \Re^n$ at random by choosing each coordinate randomly from $\{\sqrt{\frac{1+\varepsilon}{m}}, -\sqrt{\frac{1+\varepsilon}{m}}\}$. Then consider the mapping from $\Re^n$ to $\Re^m$ given by

$$z \longrightarrow (z \cdot x^1, z \cdot x^2, \ldots, z \cdot x^m).$$

In other words $u^i = (z^i \cdot x^1, z^i \cdot x^2, ..., z^i \cdot x^m)$ for $i = 1, \ldots, k$. We want to show that with positive probability, $u^1, ..., u^k$ has the desired properties. This would mean that there exists at least one choice of $u^1, ..., u^k$ satisfying inequality 1. To show this, first we write the expression $\|u^i - u^j\|$ explicitly:

$$\|u^i - u^j\|^2 = \sum_{k=1}^{m} \left(\sum_{l=1}^{n} (z_l^i - z_l^j) x_l^k \right)^2.$$

Denote by z the vector $z^i - z^j$, and by u the vector $u^i - u^j$. So we get:

$$\|u\|^2 = \|u^i - u^j\|^2 = \sum_{k=1}^{m} \left(\sum_{l=1}^{n} z_l x_l^k \right)^2.$$

Let X_k be the random variable $(\sum_{l=1}^{n} z_l x_l^k)^2$. Its expectation is $\mu = \frac{1+\varepsilon}{m} \|z\|^2$ (can be seen similarly to the proof of lemma 1). Therefore, the expectation of $\|u\|^2$ is $(1 + \varepsilon)\|z\|^2$. If we show that $\|u\|^2$ is concentrated enough around its mean, then it would prove the theorem. More formally, this is done in the following Chernoff bound lemma. $\square$

LEMMA 4
There exist constants $c_1 > 0$ and $c_2 > 0$ such that:

1. *$Pr[\|u\|^2 > (1 + \beta)\mu] < e^{-c_1 \beta^2 m}$*

2. *$Pr[\|u\|^2 < (1 - \beta)\mu] < e^{-c_2 \beta^2 m}$*

Therefore there is a constant c such that the probability of a "bad" case is bounded by:

$$Pr[(\|u\|^2 > (1 + \beta)\mu) \vee (\|u\|^2 < (1 - \beta)\mu)] < e^{-c\beta^2 m}$$

Now, we have $\binom{n}{2}$ random variables of the type $\|u_i - u_j\|^2$. Choose $\beta = \frac{\varepsilon}{2}$. Using the union bound, we get that the probability that any of these random variables is not within $(1 \pm \frac{\varepsilon}{2})$ of their expected value is bounded by

$$\binom{n}{2} e^{-c\frac{\varepsilon^2}{4} m}.$$

So if we choose $m > \frac{8(\log n + \log c)}{\varepsilon^2}$, we get that with positive probability, all the variables are close to their expectation within factor $(1 \pm \frac{\varepsilon}{2})$. This means that for all i,j:

$$(1 - \frac{\varepsilon}{2})(1 + \varepsilon)\|z^i - z^j\|^2 \leq \|u^i - u^j\|^2 \leq (1 + \frac{\varepsilon}{2})(1 + \varepsilon)\|z^i - z^j\|^2$$

Therefore,

$$\|z_i - z_j\|^2 \leq \|u^i - u^j\|^2 \leq (1 + \varepsilon)^2\|z^i - z^j\|^2,$$

and taking square root:

$$\|z^i - z^j\| \leq \|u^i - u^j\| \leq (1 + \varepsilon)\|z^i - z^j\|,$$

as required.

Question: The above dimension reduction preserves (approximately) ℓ_2-distances. Can we do dimension reduction that preserves ℓ_1 distance? This was an open problem for many years until Brinkman and Charikar showed in 2004 that no such dimension reduction is possible.

Question: Is the theorem tight, or can we reduce the dimension even further below $O(\log n/\varepsilon^2)$? Alon has shown that this is essentially tight.

Finally, we note that there is a now-extensive literature on more efficient techniques for JL-style dimension reduction, with a major role played by a 2006 paper of Ailon and Chazelle. Do a google search for "Fast Johnson Lindenstrauss Transforms."

3.1 Locality preserving hashing

Suppose we wish to hash high-dimensional vectors so that nearby vectors tend to hash into the same bucket. To do this we can do a random projection into say the cube in 5 dimensions. We discretise the cube into smaller cubes of size ε. Then there are $1/\varepsilon^5$ smaller cubes; these can be the buckets.

This is simplistic; more complicated schemes have been constructed. Things get even more interesting when we are interested in ℓ_1-distance.

3.2 Dimension reduction for efficiently learning a linear classifier

Suppose we are given a set of m data points in $\Re^d$, each labeled with 0 or 1. For example the data points may represent emails (represented by the vector giving frequencies of various words in them) and the label indicates whether or not the user labeled them as spam. We are trying to learn the rule (or "classifier") that separates the 1's from 0's.

The simplest classifier is a halfspace. Finding whether there exists a halfspace $\sum_i a_i x_i \geq b$ that separates the 0's from 1's is solvable via Linear Programming. This LP has $n + 1$ variables and m constraints.

However, there is no guarantee in general that the halfspace that separates the training data will generalize to new examples? ML theory suggests conditions under which the classifier does generalize, and the simplest is *margin*. Suppose the data points are unit vectors. We say the halfspace has *margin ε* if every datapoint has distance at least ε to the halfspace.

In the next homework you will show that if such a margin exists then dimension reduction to $O(\log n/\varepsilon^2)$ dimensions at most halves the margin. Hence the LP to find it only has $O(\log n/\varepsilon^2)$ variables instead of $n + 1$.
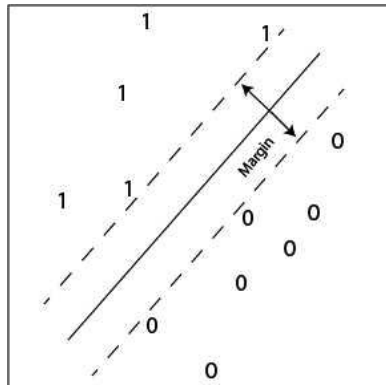
Figure 2: Margin of a linear classifier with respect to some labeled points

Bibliography:

1. W. Brinkman and M. Charikar. On the impossibility of Dimension Reduction in ℓ_1. IEEE FOCS 2003.