LECTURE 16 (October 17th)

Dimensionality Reduction / Sketching

How do we deal with data in high dimensions?

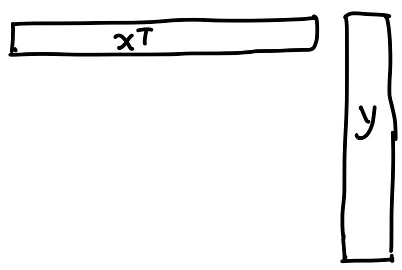
We often visualize data and algorithms in 1, 2 or 3 dimensions, e.g. a graph or 3D plot

But high dimensional space is not like low dimensional space, as we will see in the first part of this lecture, so such visualization is not very informative

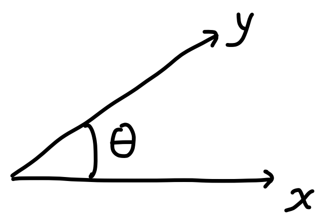
In the second part of the lecture, we are going to ignore our own advice and look at sketching, aka, dimensionality reduction techniques

High-dimensional Geometry

Recall that inner product of two vectors in d dimensions

$\langle x, y\rangle =$  $= \sum_{i=1}^{d} x_i y_i$

$= \|x\|_2 \cdot \|y\|_2 \cdot \cos\theta$



Q: How many $\underbrace{\text{mutually orthogonal}}$ unit vectors $x_1, \dots x_t$ can we find in d dimensions?

This means $|x_i^T x_j| = 0 \quad \forall\, i,j \in [t]$

Answer: We can find d such vectors

Q: How many $\underbrace{\text{nearly orthogonal}}$ unit vectors $x_1, \dots x_t$ can we find in d dimensions?

This means $|x_i^T x_j| \le 0.01 \quad \forall\, i,j \in [t]$ OR the vectors are far apart

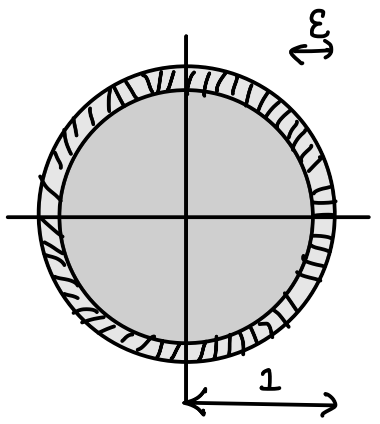
A: There can be $2^{\Theta(d)}$ such vectors. In general, if we want inner product to be atmost ε, then there can be $2^{\Theta(\varepsilon^2 d)}$ such vectors.

Curse of dimensionality Suppose we want to find nearest neighbors in high dimensions. We typically need an exponential amount of data before we see close points if our data is truly random.

The existence of lower dimensional structure in our data is often the only reason we can hope to learn

Let's look at another example in high-dimensional geometry

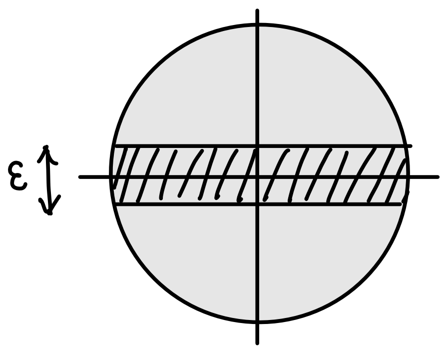
Consider the unit ball in d dimensions



What fraction of volume of B_d falls in the ε-shell around the boundary?

In 2- or 3-dimension, this is small, $O(\varepsilon)$ fraction

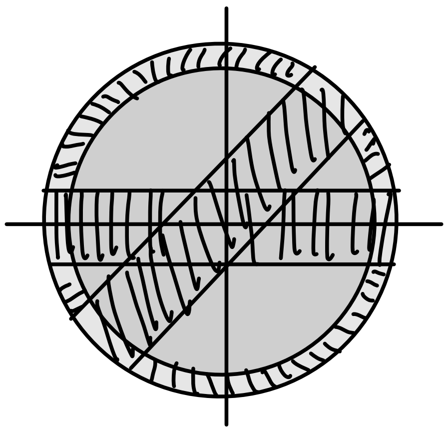
But in d-dimension, this fraction almost $\approx 1 - 2^{-\Theta(\varepsilon d)}$



What fraction of volume is close to the equator?

In 2 or 3 dimension, this is small

But in d-dimension, this fraction is $\approx 1 - 2^{-\Theta(\varepsilon^2 d)}$



Most of the volume lives in the shaded region

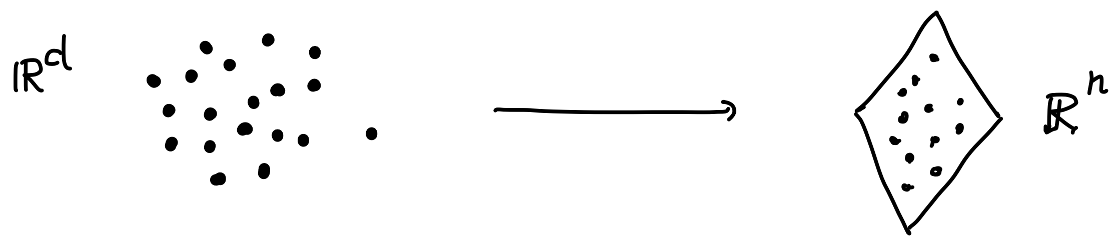
High dimensional ball looks nothing like the 2D-ball

Sketching or Dimensionality Reduction

Despite the fact that low dimensional space behaves nothing like high-dimensional space, we can still leverage its wierdness to our advantage

In particular, suppose we have data $x_1, \dots, x_N \in \mathbb{R}^d$

We want to find some way of making it low-dimensional, say in $\mathbb{R}^n$ where $n \ll d$



This is some sort of data compression

Of course, we should not expect lossless data compression but we would also like to preserve geometry of our data

For us, it will be pairwise distances between the points that is approximately preserved

How is this useful? Let's look at an example from computational geometry, where such a thing is very useful

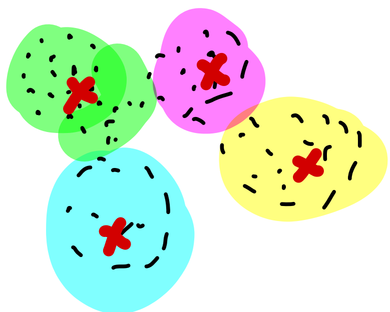
Consider the k-means clustering problem

<u>Input</u> $x_1, \dots x_N \in \mathbb{R}^d$ and an integer $k > 1$

<u>Output</u> Find $y_1, \dots y_k \in \mathbb{R}^d$ such that

$$\sum_{i=1}^{n} \min_{j \in [k]} \|x_i - y_j\|_2^2 \quad \text{is minimized}$$

Basically, we want to partition the input into k-clusters and y_j's are the centers of these clusters & we want to minimize the sum of distances of points from their closest center



<u>Note</u>: The fact that y_j's are the centers of the cluster requires a proof which we will not cover here

In particular, this problem only looks at pairwise distances between points, thus if we have a way of reducing the dimension while approximately preserving the distances, we can solve approximate k-means faster in low dimensions

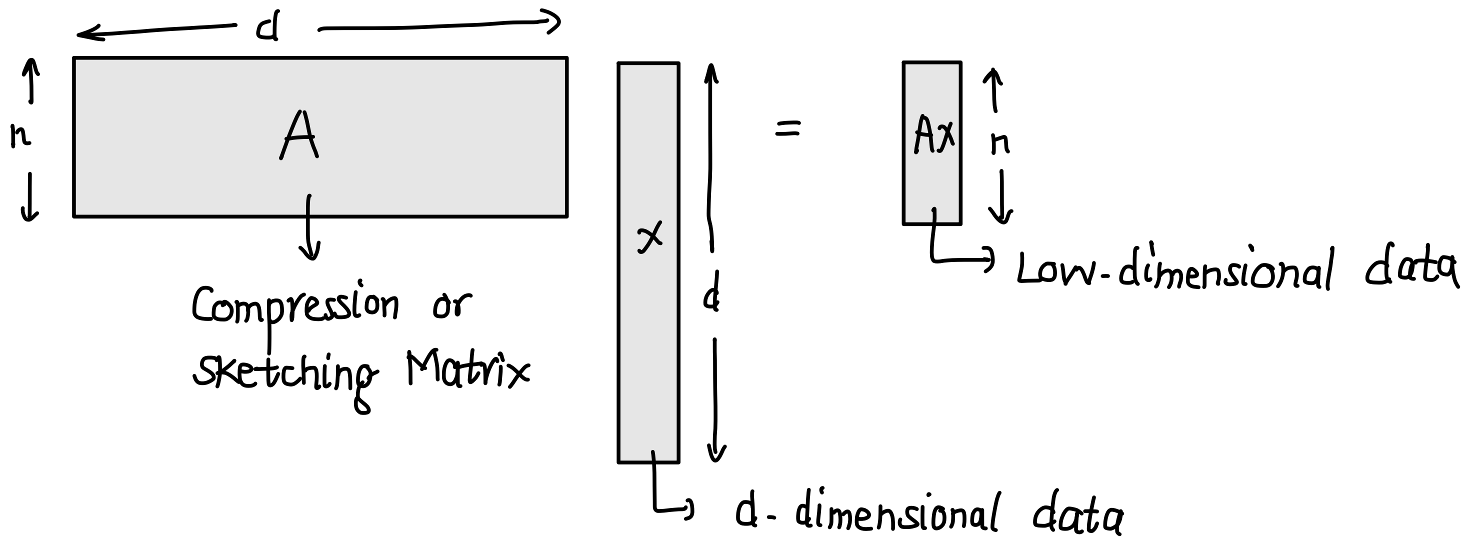
Similarly for other problems like nearest neighbour search and so on

Johnson - Lindenstrauss Lemma

This gives a way: data $x_1, \dots x_N \in \mathbb{R}^d \longrightarrow \mathbb{R}^n$ where $n \ll d$

In particular, $n = O(\log N)$ where N is the number of data points so, we get an exponential improvement

And the way to embed data is via a linear map or linear transformation, in other words a matrix



Theorem (Johnson-Lindenstrauss '84)

For all points $x_1, \ldots, x_N \in \mathbb{R}^d$, $\exists\ n = c \log N$ and a matrix $A \in \mathbb{R}^{n \times d}$ such that

$$0.99 \|x_i - x_j\|_2 \leq \| Ax_i - Ax_j \| \leq 1.01 \|x_i - x_j\| \quad \forall\ i,j \in [N]$$

How do we find such an A ? Just picking a matrix randomly would work with high probability

To prove this, we need some more probability tools so we take a small detour

Gaussian or Normal Distribution

We will work with continuous probability distributions for a bit, in particular distributions on the real line $\mathbb{R}$ or in d-dimensional real space $\mathbb{R}^d$

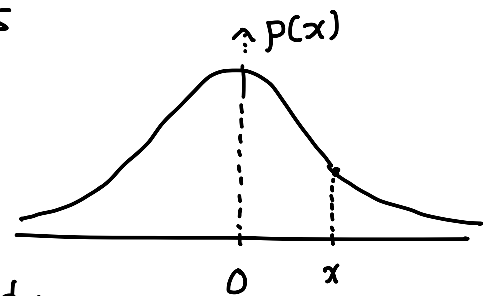
Continuous distributions have a probability density function (p.d.f.) which tells us the weight the distribution gives to a particular region

E.g. in 1-dimension $p : \mathbb{R} \to \mathbb{R}_{\geq 0}$

and probability of an interval $I = \int_I p(x)\,dx$

Gaussian distribution is one of the most useful distributions
The pdf of 1-dimensional standard Gaussian is

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$



The probability of an interval of size dx is $p(x)dx$

The mean is $\mu = \mathbb{E}[G] = \int_{\mathbb{R}} x\, p(x)\, dx$ analogous to the discrete case

$= 0$ $\qquad \sum x \cdot \mathbb{P}[X=x]$

Another quantity that is important is the variance

$$\sigma^2 = \mathbb{E}[(G-\mu)^2] = \int_{\mathbb{R}} x^2 p(x)\, dx = 1$$

The standard 1-D Gaussian or Normal distribution is denoted by $\mathcal{N}(0,1)$

One can have a Gaussian with mean μ & variance σ^2 denoted $N(\mu, \sigma^2)$ with the pdf $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$

Properties of the Gaussian Distribution

The normal distribution has a lot of unique properties

1 Tail Bounds

For example, suppose we toss n independent coins

$X_1, \ldots\ldots X_n \in \{\pm 1\}$, so $\mathbb{P}[X_i = +1] = \mathbb{P}[X_i = -1] = 1/2$

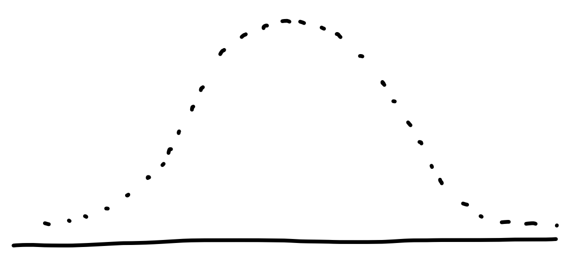
Let $X = \sum_{i=1}^{n} X_i$. Then, $\mathbb{E}[X] = 0$

And Chernoff bounds imply that $\mathbb{P}\left[\frac{|X|}{\sqrt{n}} \geq t\right] \leq e^{-t^2/2}$

so, $X \approx \mathbb{E}[X]$, since the decay is superexponential

But in fact something more is true, as $n \to \infty$

$\frac{X}{\sqrt{n}} \longrightarrow \mathcal{N}(0,1)$, so the distribution starts to look like a Gaussian



The tail inequality of the form $\mathbb{P}[\,|G| \geq t\,] \leq e^{-t^2/2}$ is called a Gaussian tail bound because it holds when G is $N(0,1)$

[Proof : Calculus]

2 Sum and scaling

Let G_1 be $N(\mu_1, \sigma_1^2)$ and G_2 be $N(\mu_2, \sigma_2^2)$

Then, $G_1 + G_2$ is $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ ← Sum of Gaussians is a Gaussian with different mean & variance

Note: This also holds for sum of many Gaussians

Similarly, if G is $N(\mu, \sigma^2)$

Then, αG is $N(\mu\alpha, \alpha^2\sigma^2)$ ← Variance scales by a factor of α^2 & mean by a factor of α

Multivariate Gaussian Distribution

A standard Gaussian distribution in d-dimensions is a vector

$G = (G_1, G_2, \ldots, G_d)$ where each coordinate G_i is an independent $N(0,1)$ random variable

The pdf of this is given by

$$p(x_1, \ldots x_d) = \underbrace{p(x_1) \cdots p(x_d)}$$

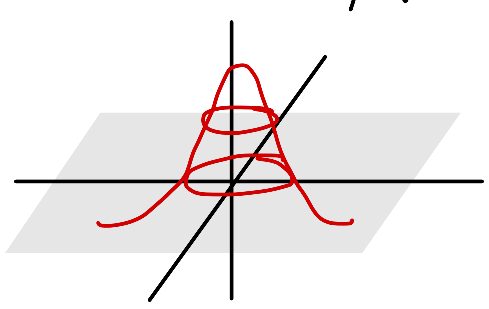
↳ pdf of 1-dimensional Gaussian

$$= \left(\frac{1}{\sqrt{2\pi}} e^{-x_1^2/2}\right) \cdots \left(\frac{1}{\sqrt{2\pi}} e^{-x_d^2/2}\right)$$

$$= \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{(x_1^2 + \cdots + x_d^2)}{2}} = \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{\|x\|^2}{2}}$$

where $x = (x_1, \ldots x_d) \in \mathbb{R}^d$

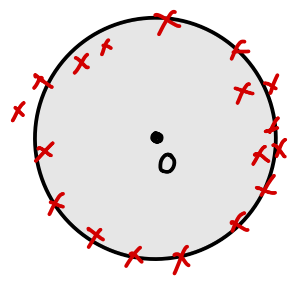
Pictorially, the 2-dimensional pdf looks like



One basic property of a high-dimensional Gaussian is the thin shell phenomenon

If we sample many points from a d-dimensional Gaussian most of them are close to the surface of a $\sqrt{d}$-radius ball even though the pdf has a higher value around 0.

This is similar to the fact mentioned before that most of the volume of the unit ball is near its surface.



Concretely, the thin shell theorem says that for a d-dimensional standard Gaussian $G = (G_1, \ldots G_d)$

$$\mathbb{P}\left[\, 0.99\sqrt{d} \le \|G\| \le 1.01\sqrt{d} \,\right] \ge 1 - e^{-cd}$$

for some constant c

Proof of Johnson-Lindenstrauss Lemma

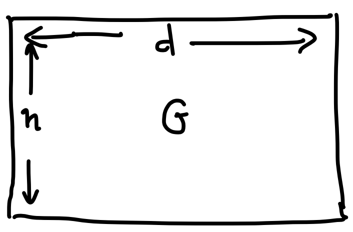
We now have all the tools to prove the Johnson-Lindenstrauss Lemma.

Theorem (Johnson-Lindenstrauss '84)

For all points $x_1, \ldots x_N \in \mathbb{R}^d$, $\exists\, n = c\log N$ and a matrix $A \in \mathbb{R}^{n\times d}$ such that

$$0.99\,\|x_i - x_j\|_2 \le \|Ax_i - Ax_j\| \le 1.01\,\|x_i - x_j\| \quad \forall\, i,j \in [N]$$

Proof Picking $A = \frac{G}{\sqrt{n}}$ to be a random Gaussian matrix will work with high probability



$n \times d$ Gaussian matrix

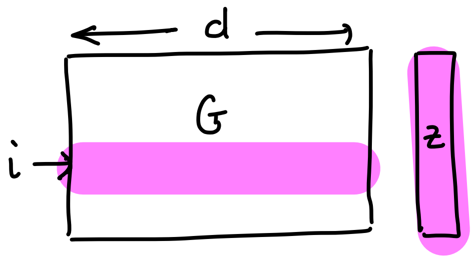
Each entry of the matrix is an independent $N(0,1)$ Gaussian, i.e., $G_{ij} \sim N(0,1)$

Let's first understand what this matrix does to a fixed vector $z \in \mathbb{R}^d$

$\boxed{\text{FACT}}$ $\forall\, z \in \mathbb{R}^d$, $\|z\| = 1$ (i.e., z is a unit vector)

$\underbrace{Gz}_{\text{Matrix-vector product}}$ is a n-dimensional standard Gaussian, i.e, each coordinate $(Gz)_i$ is $N(0,1)$ & independent

We have $(Gz)_i$ =



= product of i^{th} row of G and z

$$= \sum_{j=1}^{d} G_{ij} z_j$$

↳ scalar

Each G_{ij} is independent $N(0,1)$

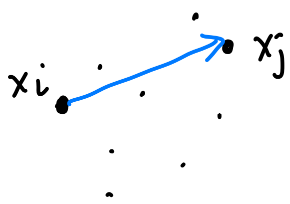
Each term $G_{ij} z_j$ is $N(0, z_j^2)$

sum of all terms is $N(0, z_1^2 + z_2^2 + \cdots + z_d^2)$

$$= N(0, \|z\|^2) = N(0,1) \text{ since } z \text{ is a unit vector}$$

All coordinates of Gz are also independent since each row of G is independent

We want to prove all pairwise distances are approximately preserved so, let us pick a pair of points



Consider $z = \dfrac{x_i - x_j}{\|x_i - x_j\|}$

Then, Gz is standard n-dimensional Gaussian and by Thin shell theorem

$$\mathbb{P}\left[0.99\sqrt{n} \le \frac{\|G(x_i - x_j)\|}{\|x_i - x_j\|} \le 1.01 \right] \geq 1 - e^{-cn}$$

$$\Leftrightarrow \mathbb{P}\left[0.99\|x_i - x_j\| \le \Big\| \underbrace{\frac{G}{\sqrt{n}}}_{\text{This is our matrix } A} x_i - \underbrace{\frac{G}{\sqrt{n}}}_{\text{This is our matrix } A} x_j \Big\| \le 1.01\|x_i - x_j\| \right] \geq 1 - e^{-cn}$$

Thus, the probability that the event $\|Ax_i - Ax_j\| \notin [0.99\|x_i - x_j\|, 1.01\|x_i - x_j\|]$, call it E_{ij} holds for a given pair (i,j) is e^{-cn}

What is the probability that there is some pair (i,j) where E_{ij} holds?

$$\mathbb{P}\left[\exists (i,j) \in \binom{N}{2} : E_{ij}\right] \le \sum_{ij} \mathbb{P}[E_{ij}] \le N^2 \cdot e^{-cn} \text{ by union bound}$$

So, if $n = c' \log N$ for a large enough c', the prob. is at most $\frac{1}{N^{100}}$

Thus, a random matrix works with high probability