High-dimensional spaces

Consider $X_1, \dots, X_d$ independent, $\mathbb{E}(X_i) = 0$, $\mathrm{Var}(X_i) = 1$

$$\text{object} = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} =: X$$

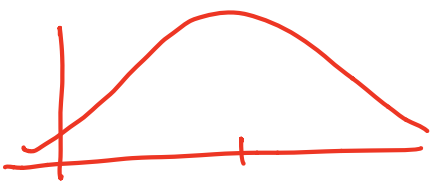
① **Norms of high-dim vectors are concentrated:**

$$\mathbb{E}(\|X\|^2) = \mathbb{E}\left(\sum_{i=1}^{d} X_i^2\right) = \sum_{i=1}^{d} \underbrace{\mathbb{E}X_i^2}_{1} = d$$

By Bernstein inequality one can prove:

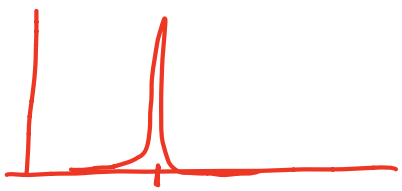
$$P\left(\left|\|X\| - \sqrt{d}\right| > t\right) \le 2\exp(-ct^2)$$

2-dim



$\sqrt{2}$

d-dim



$\sqrt{d}$

$X \sim N(0, I_2)$

$X \sim N(0, I_d)$

$\sqrt{d}$

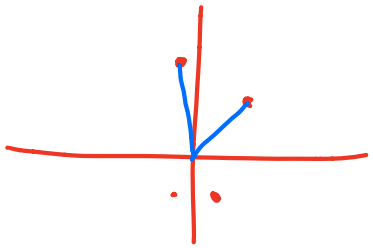
(2) **Two random vectors are nearly orthogonal**

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix}$$

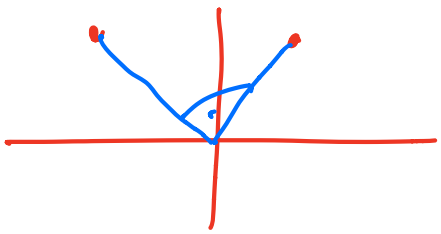
$$\mathbb{E}\left(\langle X, Y \rangle\right) = 1/d .$$

But again: strong concentration.

2 dim normal



d-dim



(3) **Distances in high-dim spaces are dominated by noise**

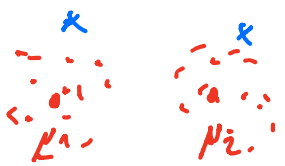
$$X \sim N(\mu_1, \sigma_1^2 I_d), \quad Y \sim N(\mu_2, \sigma_2^2 I_d)$$

$$\mathbb{E}\left(\|X - Y\|^2\right) = \mathbb{E}\left(\sum_{i=1}^{d} |x_i - y_i|^2\right)$$

$$= \sum_{i=1}^{d} \left(\mathrm{var}(x_i - y_i) + \left(\mathbb{E}(x_i - y_i)\right)^2\right)$$

$$= \underbrace{d \cdot \left(\sigma_1^2 + \sigma_2^2\right)}_{\text{noise}} + \underbrace{\|\mu_1 - \mu_2\|^2}_{\text{signal}}$$

2 dim



d-dim

noise >> signal

$dist(x,y) = \|\mu_1 + \mu_2\| + \text{small noise}$

Even worse: $x, x' \sim N(\mu_1, \sigma_1^2 I_d)$, $y, y' \sim N(\mu_2, \sigma_2^2 I_d)$

$E(\|x - x'\|^2) = 2d\sigma_1^2$ within-class distance

$E(\|x - y\|^2) = \|\mu_1 - \mu_2\|^2 + d(\sigma_1^2 + \sigma_2^2)$ between class distance

$E(\|y - y'\|^2) = 2d\sigma_2^2$ within-class distance in second class

If $\sigma_1 \ll \sigma_2$, then

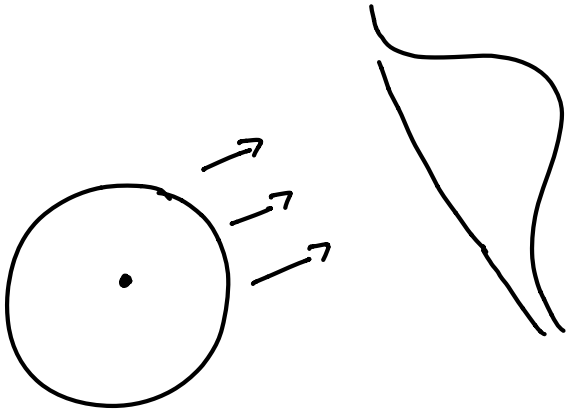
$\sigma_1^2 + \sigma_2^2 \ll \sigma_2^2$

between $\ll$ within-class 2

(4) Projections always look like Gaussians

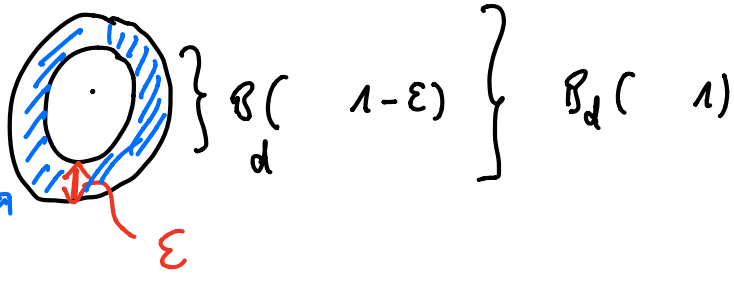
$$X \sim \text{Unif}\left(S_{d-1}\sqrt{d}\right)$$

Project X on a fixed vector v:



$$\langle X, v \rangle \longrightarrow N(0,1) \quad \text{in distribution}$$

(5) Volume is concentrated along surface:



$$\text{vol}\left(B_d(1) - B_d(1-\varepsilon)\right) =$$

$$= 1^d \cdot \text{vol}\left(B_d(1)\right) - (1-\varepsilon)^d \,\text{vol}\left(B_d(1)\right)$$

$$= \underbrace{\left(1 - (1-\varepsilon)^d\right)}_{\to 1 \text{ as } d \to \infty} \text{vol}\, B_d(1)$$

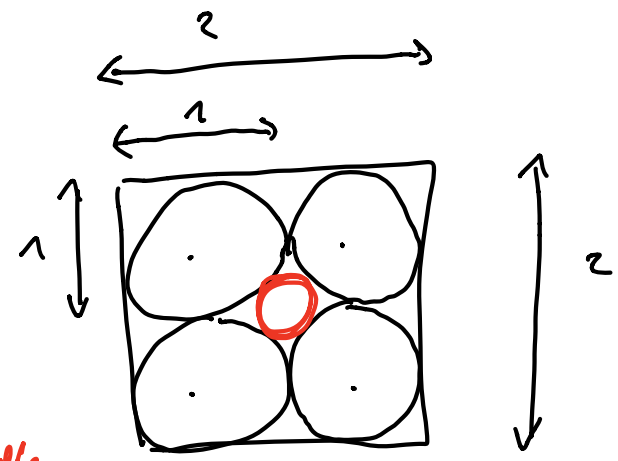
In particular, if you sample from a normal distribution:

2-dim

d-dim

Intuition is often wrong

Riddle



2-dim:

- Square of side length 2
- Circles of diam 1
- small circle in the middle

Radius of small circle:

$\frac{1}{2}$ (diagonal of square − 2 diam circles)

$\frac{1}{2}(2\sqrt{2} - 2) = \sqrt{2} - 1$

d-dim

outer circle diam = 2

diam cube $2\sqrt{d}$

inner circle $\sqrt{d} - 1$