

# Machine Learning

## Lecture 13: High-dimensional statistics

Hao Tang

October 20, 2022

# High-dimensional objects in machine learning

- A  $28 \times 28$  image
- A one-hot vector that represents a word in a 20000-word vocabulary set

# Intuitions

- We tend to extrapolate our intuitions from two- or three-dimensional space to high-dimensional spaces.
- Many intuitions in two- and three-dimensional spaces fail to hold in high dimensions.

## Volume concentration

- The volume of a ball with radius  $r$  is

$$V(r) = \frac{\pi^{d/2}}{\Gamma(d/2 + 1)} r^d. \quad (1)$$

- If we shrink the radius by a small amount  $\epsilon$ , the volume shrinks by

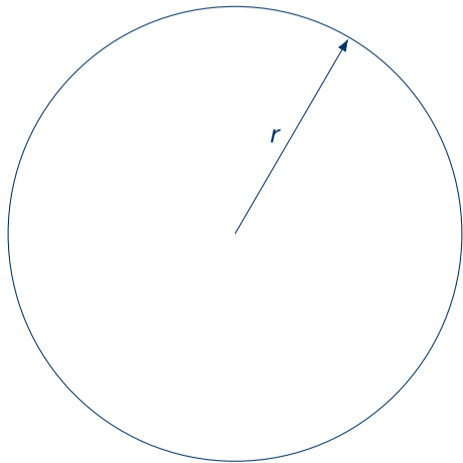
$$\frac{V((1 - \epsilon)r)}{V(r)} = \frac{(1 - \epsilon)^d V(r)}{V(r)} = (1 - \epsilon)^d \leq e^{-\epsilon d}. \quad (2)$$

- The last inequality uses  $1 - x \leq e^{-x}$ .

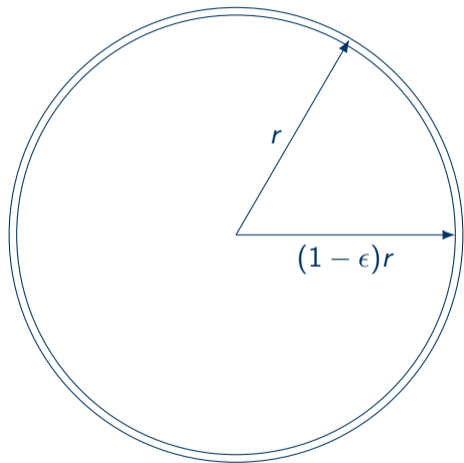
# Volume concentration

- As  $d$  becomes large,  $V((1 - \epsilon)r)$  is only a small fraction of  $V(r)$ .
- The volume of a ball in high dimensions concentrates on the thin crust of the ball.
- If you uniformly sample a point, you will likely end up at the crust of the ball.

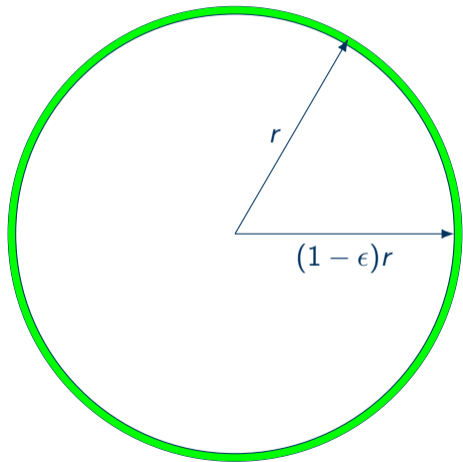
## Volume concentration



## Volume concentration

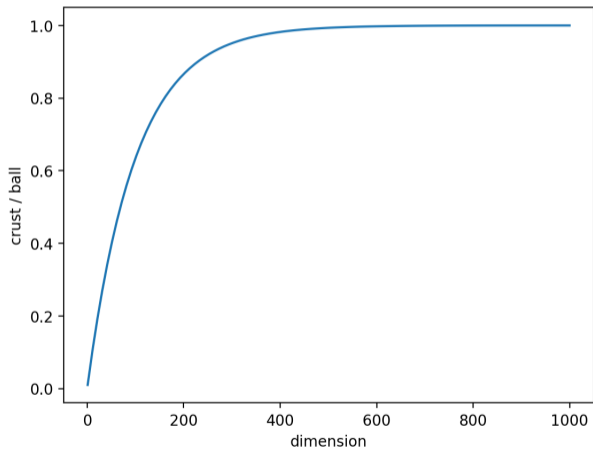


## Volume concentration





# Volume concentration



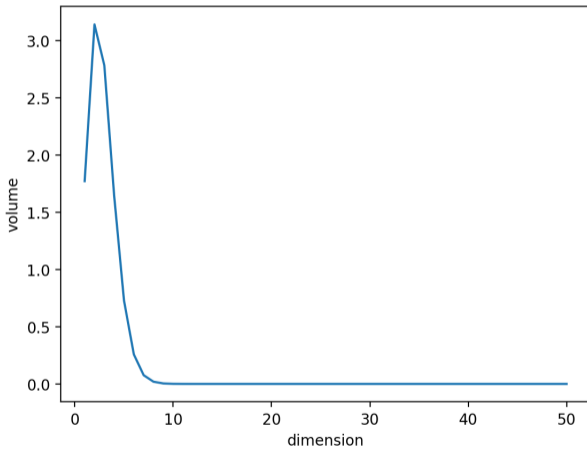
## Volume of a unit ball

- The volume of a unit ball ( $r = 1$ ) is

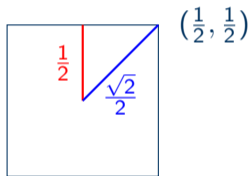
$$V(1) = \frac{\pi^{d/2}}{\Gamma(d/2 + 1)}. \quad (3)$$

- $\Gamma(x + 1) \sim \sqrt{2\pi x} \left(\frac{x}{e}\right)^x$
- $V(1) \rightarrow 0$  when  $d \rightarrow \infty$ .

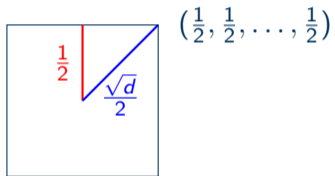
# Volume of a unit ball



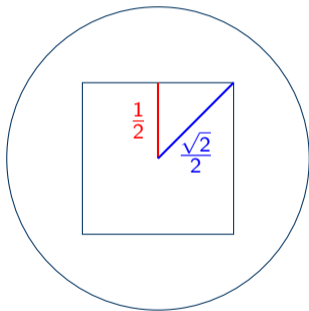
## Corners of the unit cube



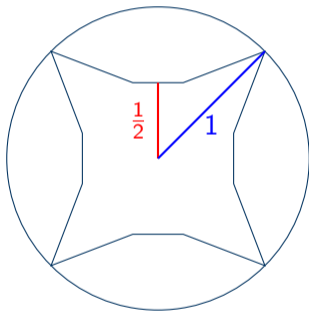
## Corners of the unit cube



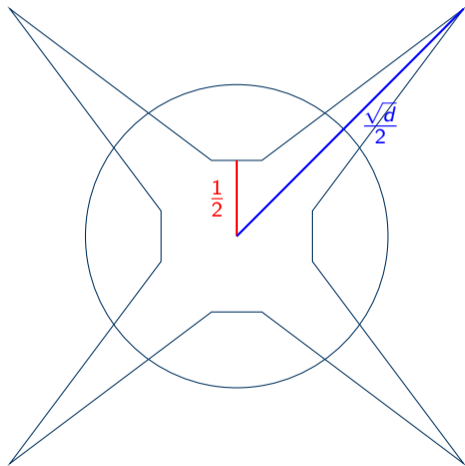
## Corners of the unit cube ( $d = 2$ )



## Corners of the unit cube ( $d = 4$ )

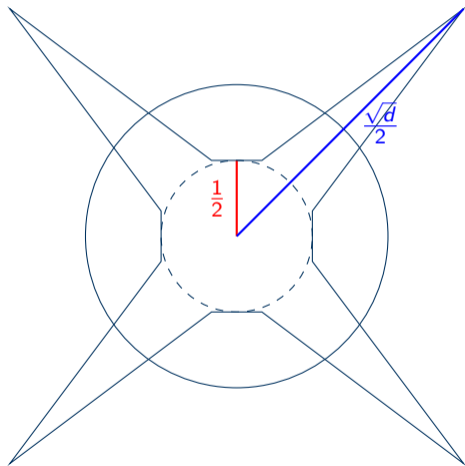


## Corners of the unit cube





## Corners of the unit cube



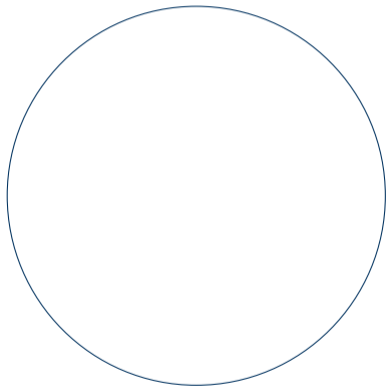
## Corners of the unit cube

- The distance between the origin and the corner is  $\frac{\sqrt{d}}{2}$ .
- For example, one corner is  $(1/2, 1/2, \dots, 1/2)$ .
- The distance between the origin and one of the faces is  $\frac{1}{2}$ .
- The distances to the faces stay the same, while the distances to the corners becomes large, when  $d$  is large.

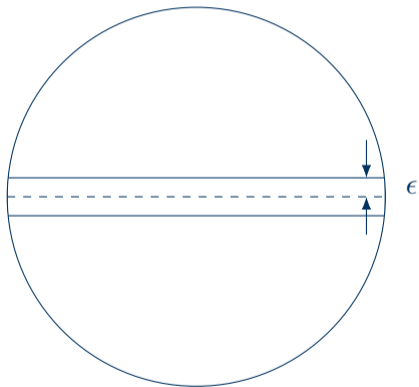
## Volume near the equator

- Pick a north direction.
- Pick  $\epsilon > 0$ , the width of a slab. The volume above is about  $\frac{2}{\epsilon\sqrt{d}} e^{-d\epsilon^2/2}$ .
- The volume is concentrated at the equator.

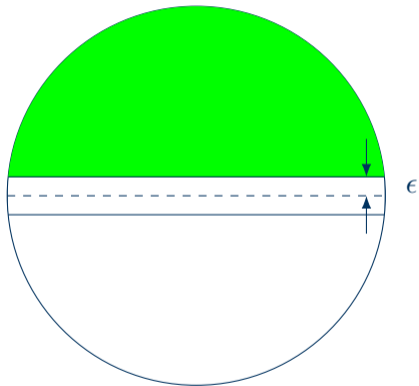
## Volume near the equator



## Volume near the equator



## Volume near the equator



## Two random vectors inside a unit ball

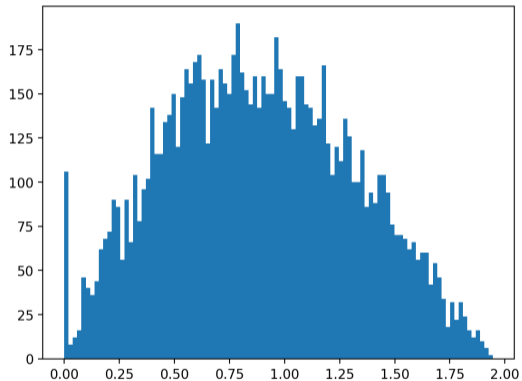
- Pick a random vector  $u$  inside the unit ball.
- Set  $u$  to be the north.
- Since most of the volume is concentrated at the equator, another random vector  $v$  will likely lie on the equator.
- The dot product  $u^T v$  will likely be close to 0.

## Distances of two random vectors

- Any two random vectors  $u$  and  $v$  in a unit ball are likely to be orthogonal, meaning  $u^\top v$  is likely to be small.
- The volume is concentrated at the crust, so  $\|u\|$  and  $\|v\|$  are likely to be close to 1.
- The distance of any two random vectors  $u$  and  $v$  is likely to be about  $\sqrt{2}$ .

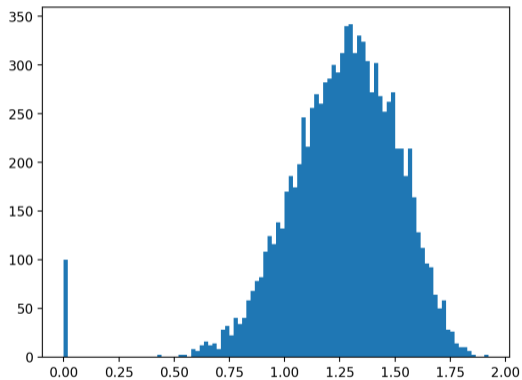


## Distances of two random vectors



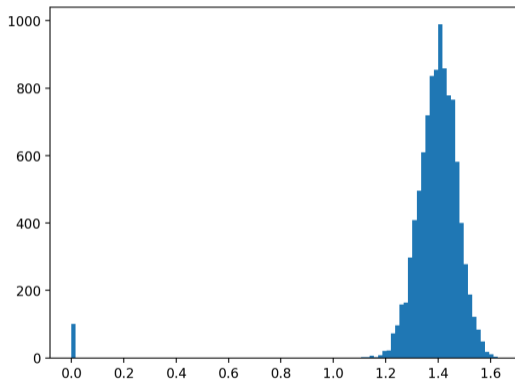
$$d = 2$$

## Distances of two random vectors



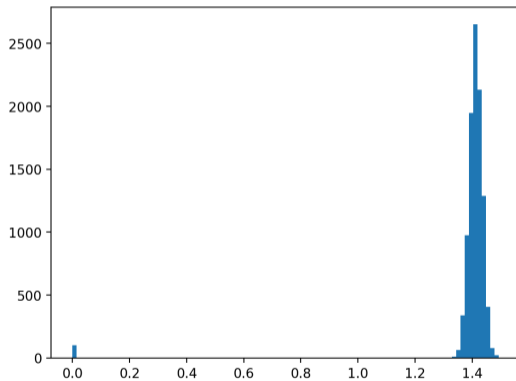
$d = 10$

## Distances of two random vectors



$d = 100$

## Distances of two random vectors



$d = 1000$

## Distances of two random vectors

- Distances of two random vectors in a unit ball concentrate at  $\sqrt{2}$ .
- Most points have similar distances!

## Norm of random Gaussian vectors

- For any  $x \sim \mathcal{N}(0, I)$ ,

$$\mathbb{P}\left(\frac{\|x\|^2}{d} - 1 > \epsilon\right) \leq \exp\left(\frac{-d\epsilon^2}{8}\right). \quad (4)$$

- In words, for any  $x \sim \mathcal{N}(0, I)$ ,  $\|x\|$  is about  $\sqrt{d}$ .
- “High-dimension Gaussian is like a soap bubble.”