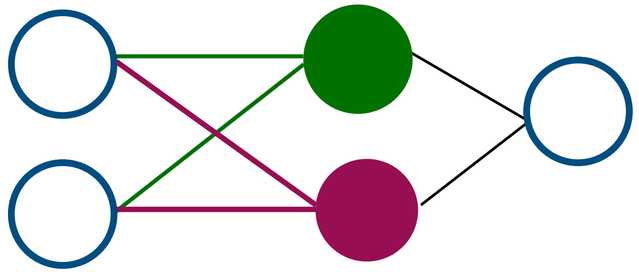
Neural Networks: A Theory Lab

Shallow neural networks

curse of dimensionality



Hadi Daneshmand

University of Virginia

News

- Public website will be updated with delays so please check Canvas
- Notes are available for the last lecture
- Please let me know if you want to help with scribing today lecture note
- Today lecture has theoretical group activity

Old vs new datasets

- Heart Disease dataset (1988) has 13 features

Variable Name	Role	Type
age	Feature	Integer
sex	Feature	Categorical
cp	Feature	Categorical
trestbps	Feature	Integer
chol	Feature	Integer
fbs	Feature	Categorical
restecg	Feature	Categorical
thalach	Feature	Integer
exang	Feature	Categorical
oldpeak	Feature	Integer

Deep learning: era of feature adequacy

- Examples: ImageNet

- Natural language processing
  - Each token has more than a billion features (if we formulate it properly)

Old vs new datasets

Old datasets

Example: Heart Disease dataset (1988)

13 features

Variable Name	Type
age	Integer
cp	Categorical
trestbps	Integer
chol	Integer
fbs	Categorical

New datasets

Example:ImageNet dataset

469 × 387 pixels = 181503 features

Neural Networks vs Random Features

Random Features

$$y \approx \sum_i \alpha_i \cos(v_i, x\rangle + b_i)$$

$v_i \sim p(w)$ depending on $k(x, y)$

Suffering from curse of dimensionality

Neural Networks

$$y \approx \sum_i \alpha_i \cos(\langle v_i, x\rangle + b_i)$$

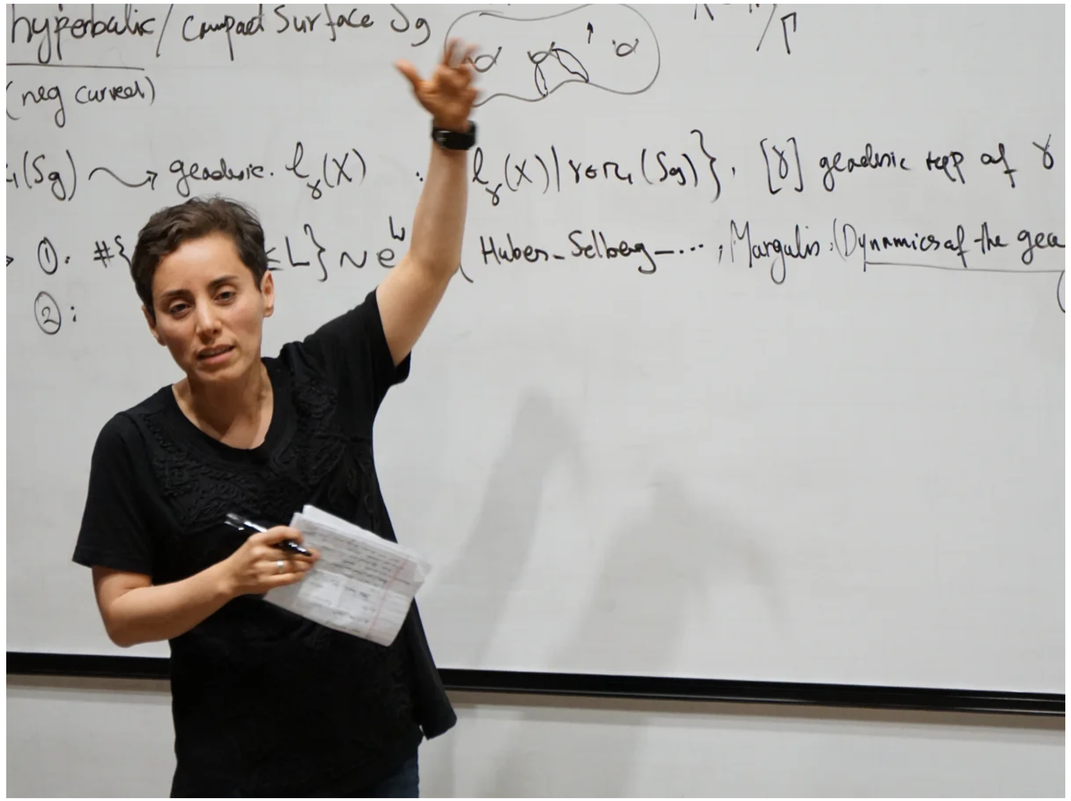
v_i are optimized since $p(w)$ is unknown

Breaking cures of dimensionality

Curses of dimensionality

- We will talk about two curses for
  - function approximation
  - optimization

- Intro
- Theory
- Lab



theguardian: Maryam Mirzakhani

Theory

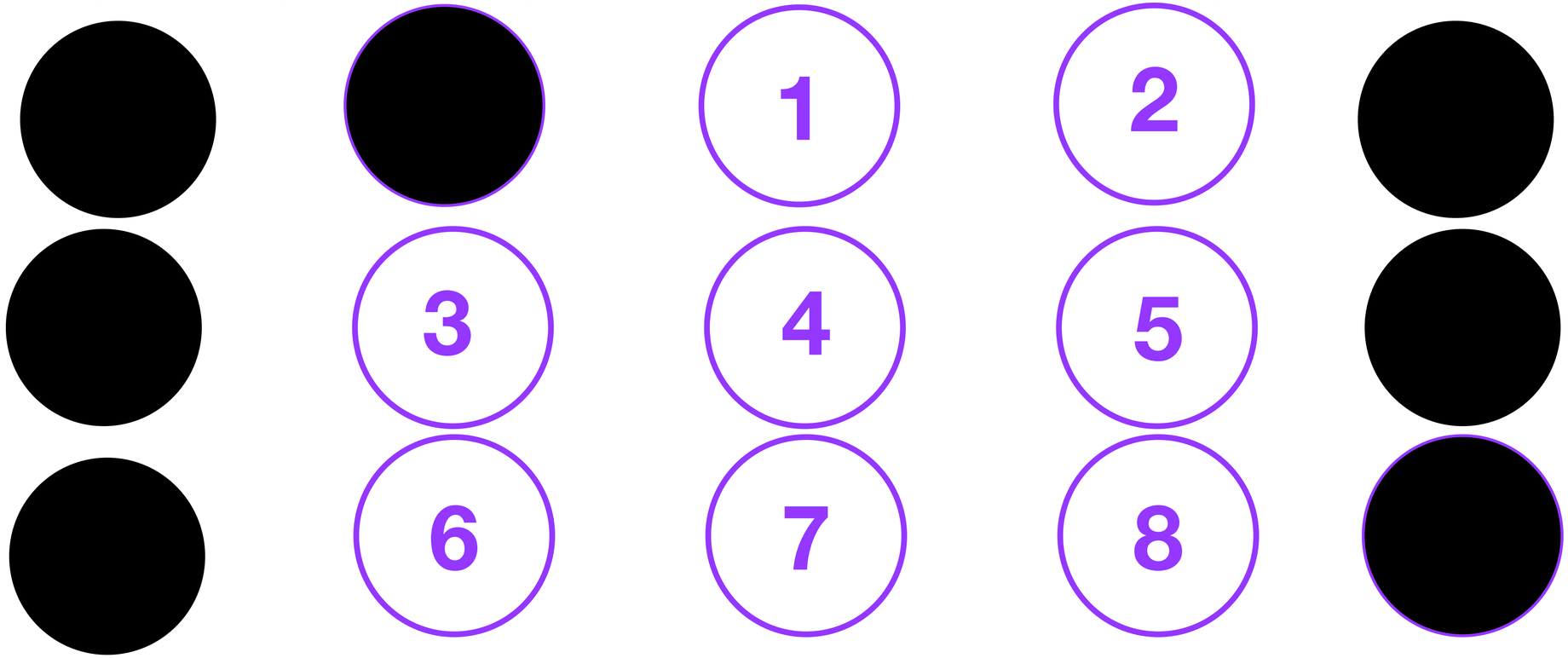
Get ready for math

Introducing curse of dimensionality

- Probabilistic derivation
- Expected value derivation

Group assignment

- Feel free to choose your own group if you want
- Thank you for your thoughtful approach to numbering the tables



I am here

Expected value calculation

- Let $w = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \ldots, 0] \in \mathbb{R}^d$
- v is d-dimensional whose elements are i.i.d. zero-mean normal with variance $1/d$
- **Question:** $\mathbb{E}\left[\langle w, v\rangle^2\right]$=?
- **Answer:** $\mathbb{E}\left[\langle w, v\rangle^2\right] = \frac{1}{2}\mathbb{E}\left[v_1^2 + v_2^2\right] = \frac{1}{d}$

Expectation of maximum of I.I.D. draw

- Settings: that $w = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \ldots, 0] \in \mathbb{R}^d$ and $v_1, \ldots, v_n \sim \mathcal{N}(0, \frac{1}{d} I_d)$
- Guess: $\mathbb{E} \max_{i \leq n} \langle v_i, w \rangle^2 = ?$

Order statistics

- Definition: Suppose that $x_1, \ldots, x_n$ are drawn from the same distribution $\mathscr{P}$. Let $x_{(k)}$ denotes the kth smallest element among $x_1, \ldots, x_n$. Then kth order statistics are moments of $x_{(k)}$.
- Example: $\mathbb{E}\left[\min_{i \leq n} x_i\right]$ is the first order statistics and $\mathbb{E}\left[\max_{i \leq n} x_i\right]$ is the n-th order statistics
- Reference: “Probability, Order Statistics and Sampling” by David Whitmer

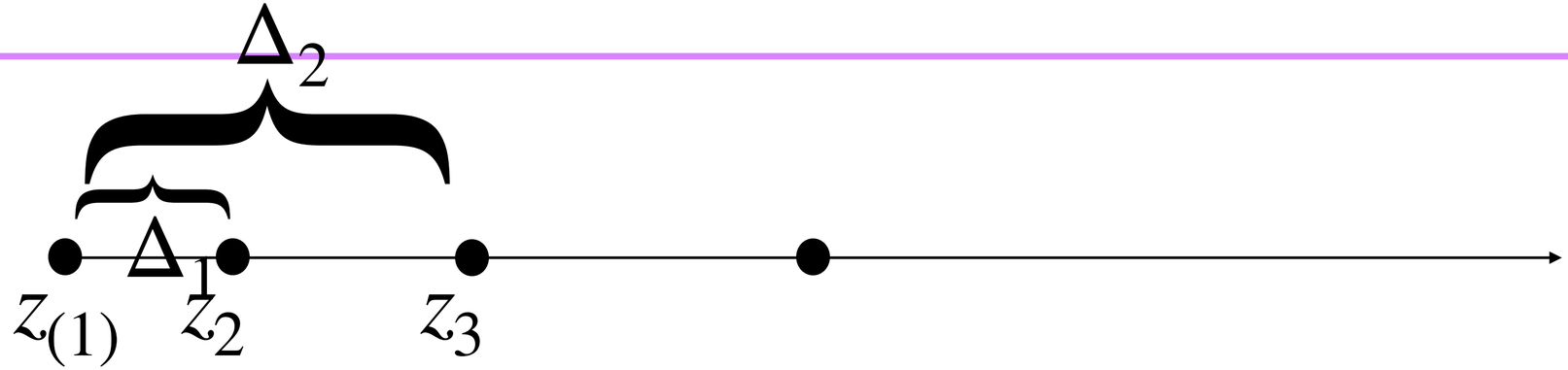
Subproblem 1

- Settings: Note that $z = \langle v, w \rangle^2 = \frac{1}{2}\left(v_1^2 + v_2^2\right)$
- Question:What is the distribution of z? Recall $v_i \sim N(0,1)$
- Solution: $\frac{1}{2d}\chi_2^2$ chi-square random variable of degree two
  - Alternatively, $2dz_i \sim Exp(\frac{1}{2})$

Subproblem 2

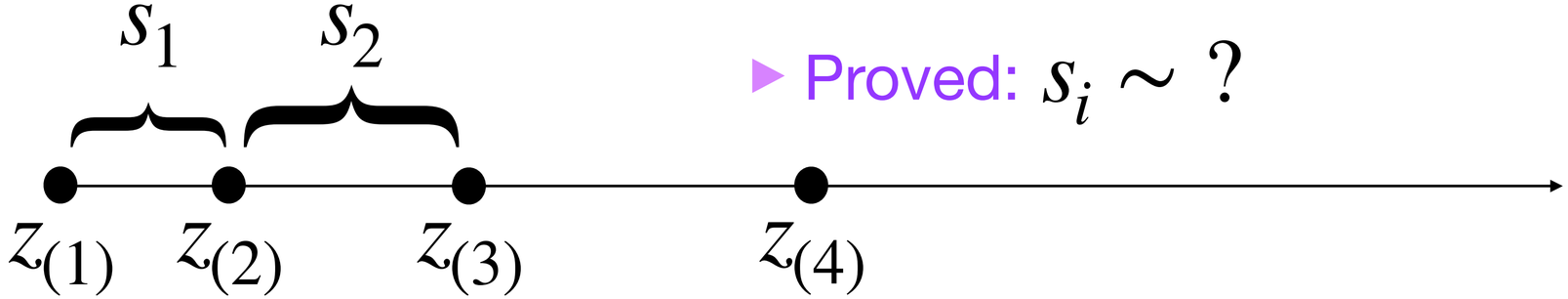
- Question: What is the distribution of $\min_{i \le n} z_i', z_i' \sim_{i.i.d} Exp(\lambda)$?
- Hint: Recall $p_\lambda(z) = \begin{cases} \lambda e^{-\lambda z} & z \ge 0 \\ 0 & \text{otherwise} \end{cases}, p(z \le a) = 1 - e^{-\lambda a}, \mathbb{E}[z] = \frac{1}{\lambda}$
- Solution: $P(\min z_i = a) = \sum_i p(\{z_j \ge a\}_{j \ne i}, z_{[i]} \le a)$
- i.i.d: $p(\min z_i = z) = n \prod_{i=1}^{n-1} p(z_i \ge z) p(z_1 = z) = n\lambda e^{-\lambda n z} = p_{n\lambda}(z)$
- An application: $\mathbb{E}\left[\min_{i \le n} z_i\right] = \frac{1}{n\lambda}$

Subproblem 3



- Question: Define $s_1 = z_{(2)} - z_{(1)}$; what is the distribution of s_1?
- Hint: exponential distribution is memory less: $p_\lambda(z \geq a + s \,|\, z \geq a) = p_\lambda(z \geq s)$
- Define: $\Delta_i := z_i - z_{(1)}$, $z_{(1)} = \min_{i \leq n} z_i$
- memoryless concludes $\Delta_i \sim Exp(\lambda)$ since $p_\lambda(z_i \geq z_{(1)} + b \,|\, z_i \geq z_{(1)}) = p_\lambda(z_i - z_{(1)} \geq b)$
- Solution:Thus $s_1 = \min_{i \leq n-1} \Delta_i$, $\Delta_i \sim Exp(\lambda) \implies \Delta_1 = Exp(\lambda(n-1))$

Subproblem 4

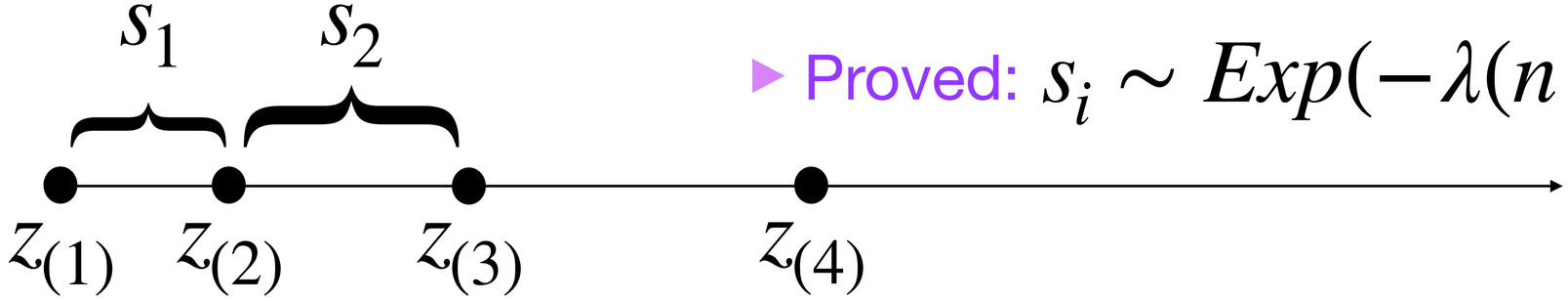


- Proved: $s_i \sim ?$

- Question: What is the distribution of s_i?

- Solution:$s_i \sim Exp(\lambda(n-i))$

Subproblem 5



- Proved: $s_i \sim Exp(-\lambda(n-i))$

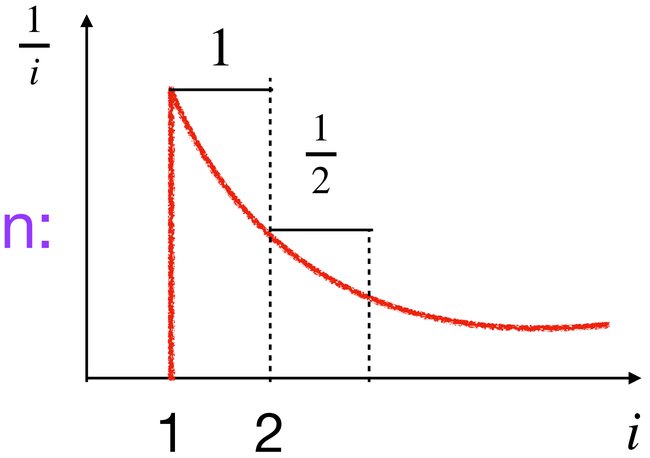
- Question: What is the distribution of $\mathbb{E}\left[z_{(n)}\right]$?

- Solution: $\mathbb{E}\left[z_{(n)}\right] = \mathbb{E}\left[z_{(1)} + \sum_{i=1}^{n-1} s_i\right] = \sum_{i=1}^{n} \frac{1}{i\lambda}$

Subproblem 6

- Question: Prove $\log(n) \leq \sum_{i=1}^{n} \frac{1}{i} \leq \log(n) + 1$

- Solution:



Area under the red curve $< \sum_{i=1}^{n} \frac{1}{i}$

Area: $\int_1^n \frac{1}{x} dx = \log(n) - \log(1) = \log(n) \leq \sum_{i=1}^{n} \frac{1}{i}$

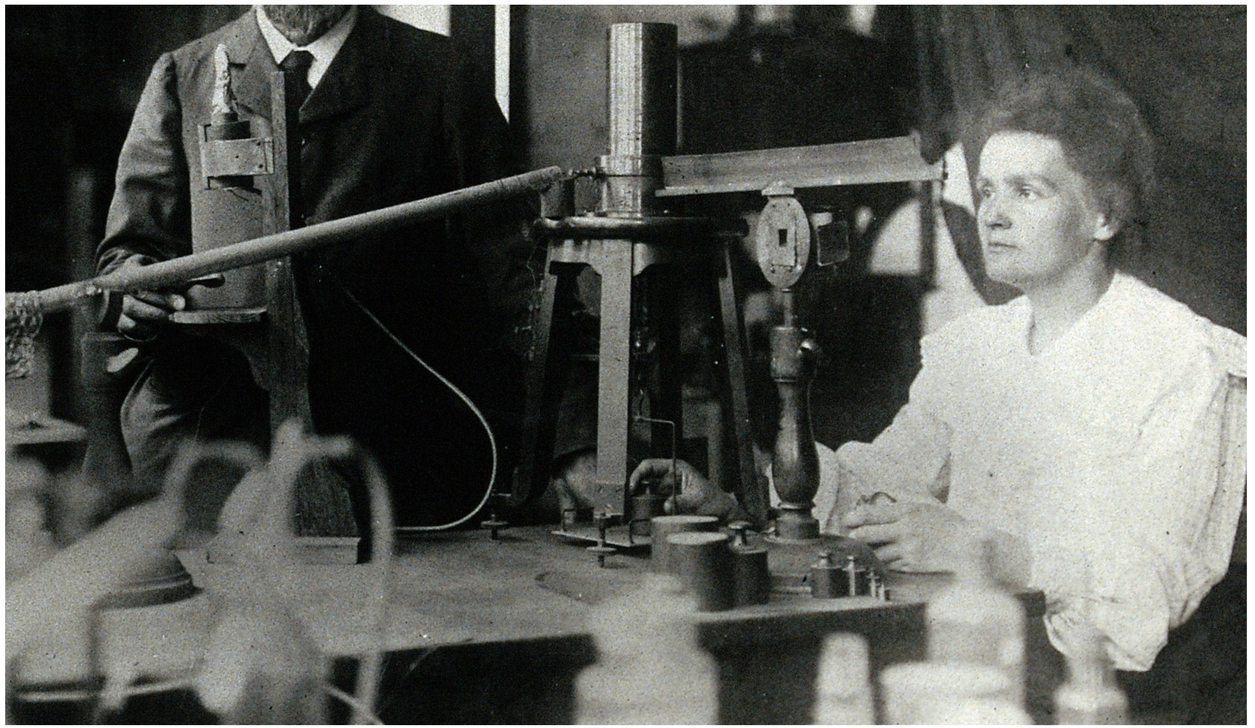
Putting all together

- P1: $2d\langle w, v_i\rangle^2 \sim Exp(1/2)$
- P2-5: $\mathbb{E}2d \max_{i\leq n}\langle w, v_i\rangle^2 = \sum_{i=1}^{n} \frac{2}{i}$
- P6: $\log(n) \leq \sum_{i=1}^{n} \frac{1}{i} \leq \log(n) + 1$
- Conclusion: $\frac{\log(n)}{d} \leq \mathbb{E}\left[\max_{i\leq n}\langle w, v_i\rangle^2\right] \leq \frac{\log(n)+1}{d}$

Curse of dimensionality

- Statement: $\mathbb{E}\left[\max_{i \le n} \langle v_i, w \rangle^2\right] \le \frac{\log(n)}{d}$
- Curse: To ensure $\mathbb{E}\left[\langle v_i, w \rangle^2\right] \ge \frac{1}{2}$, we need $n \ge e^{d/2}$.

- Intro
- Theory
- Lab

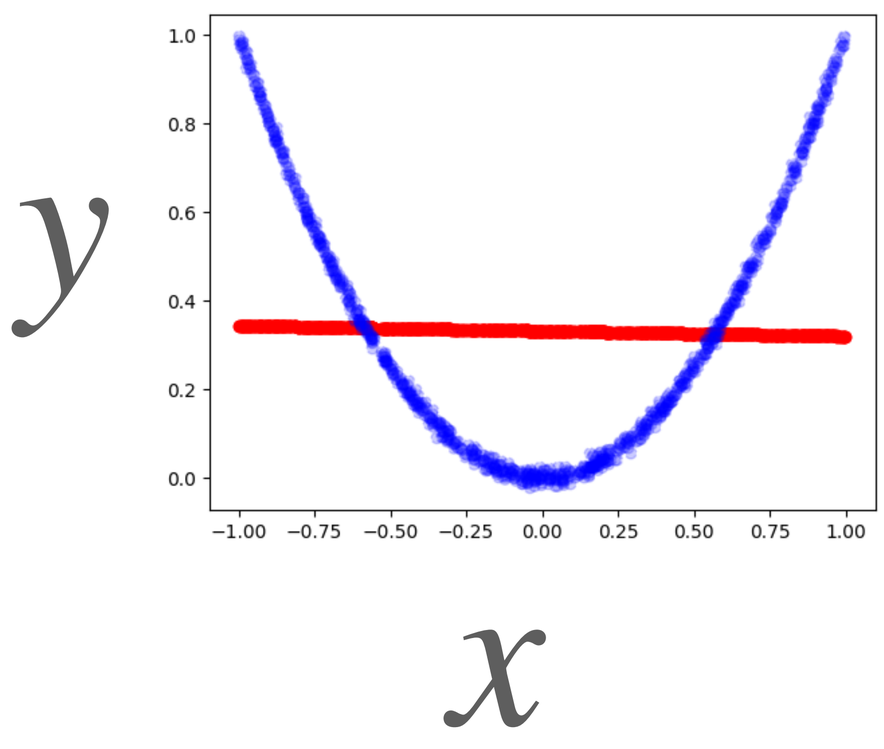


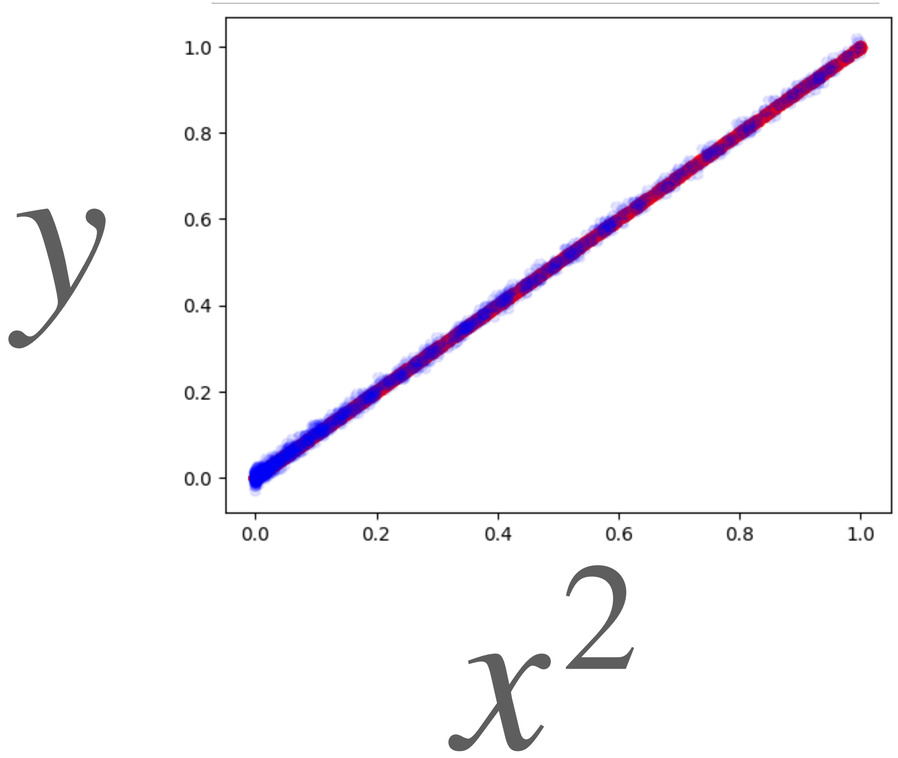
wikipedia: Marie Curie

Experiments

Get ready for coding group activity

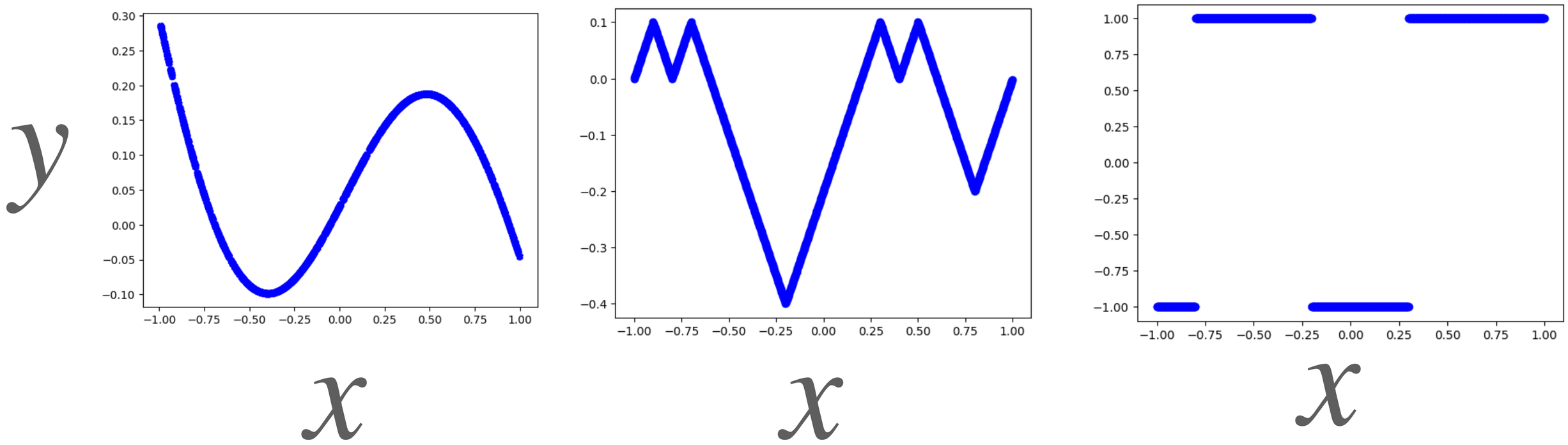
Recap: How to solve the non-linearity challenge





Recap

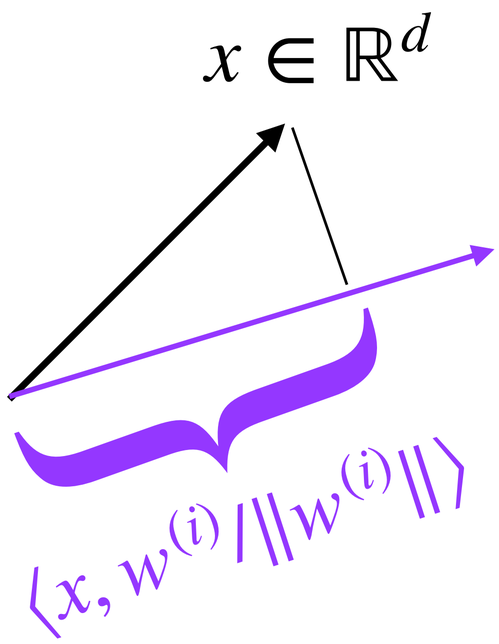
- Goal: We want to design **universal** non-linear features



Random features

- Given x, design features $\phi_1(x), \ldots, \phi_{20}(x)$ such that $y \approx \sum_{i=1}^{20} w_i \varphi_i(x)$

$$\phi_i(x) = \cos(\langle v_i, x \rangle + b_i)$$

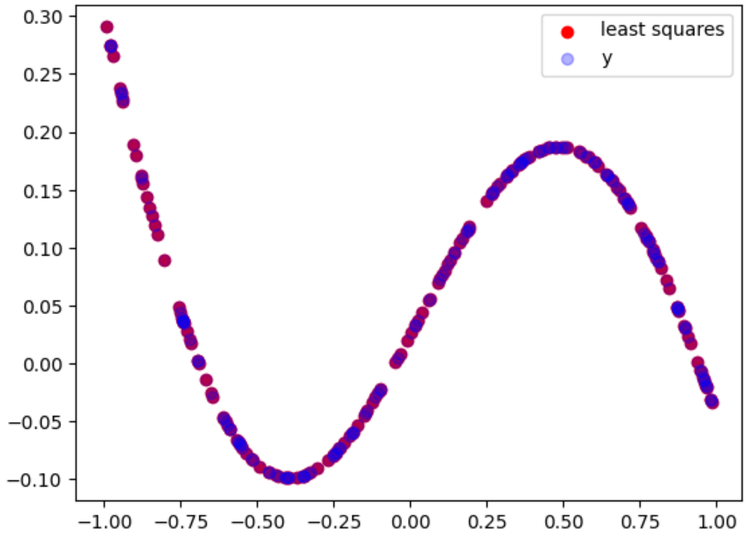


Recall: $\langle x, w \rangle = \sum_{i=1}^{d} x_i w_i = \|x\| \|w\| \cos(\theta)$

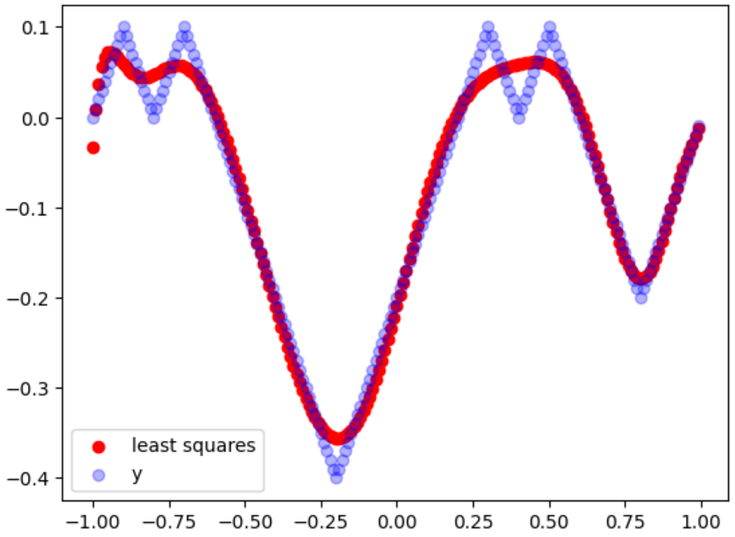
$b_i \sim \text{uniform}[0, 2\pi]$

Performance of random gaussian features

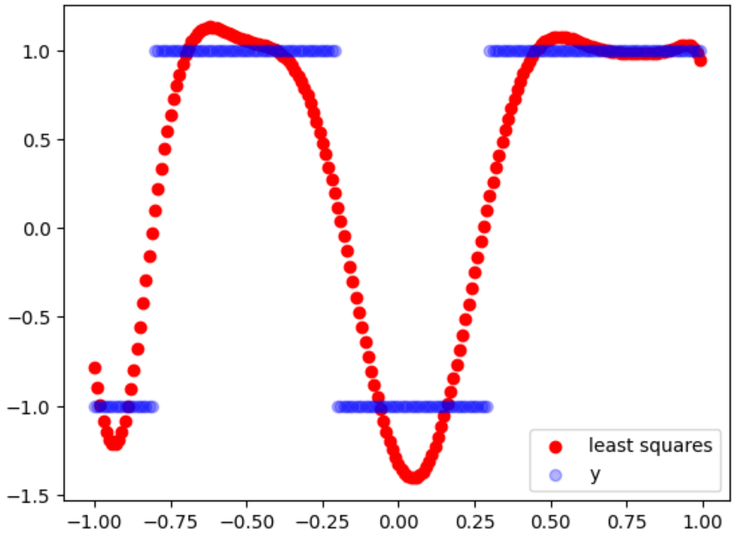
- Blue: y and red is the solution of regression on random features



Good approximation



Reasonable approximation



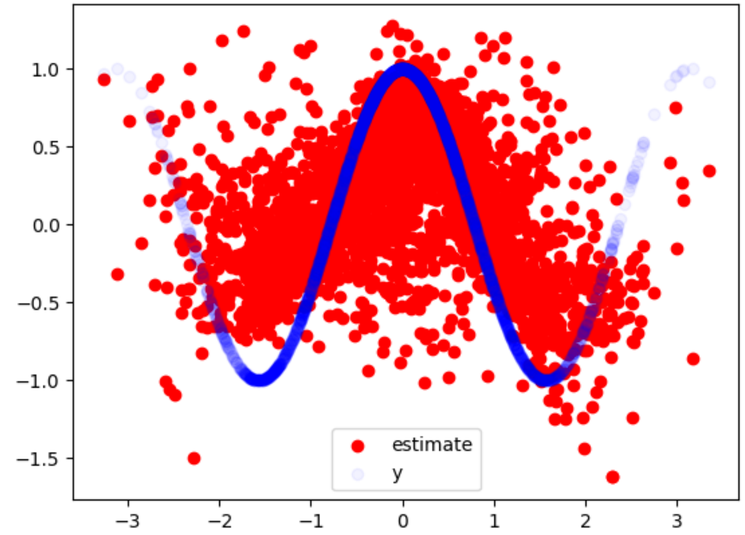
Poor approximation

Poor performance of random features

- $y = \cos(2 * \langle w, x \rangle)$ where
  - $x \in \mathbb{R}^4, x \sim \mathcal{N}(0, I_4)$
  - $w \sim uniform\{e_1, \ldots, e_d\},$
  - $e_j \in \mathbb{R}^d, [e_j]_k = \begin{cases} 1 & j = k \\ 0 & \text{otherwise} \end{cases}$
  - $f : \mathbb{R} \to \mathbb{R}$
- Estimate with 200 random feature

y

f_{100}

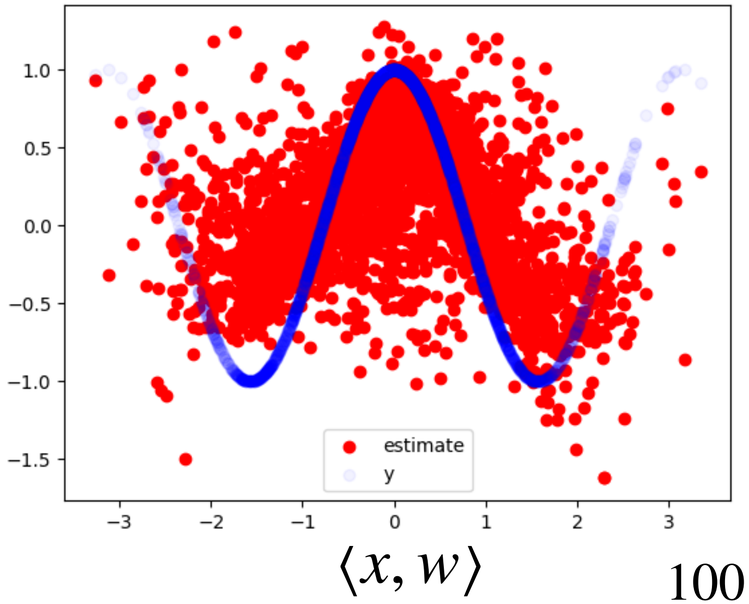


$\langle x, w \rangle$

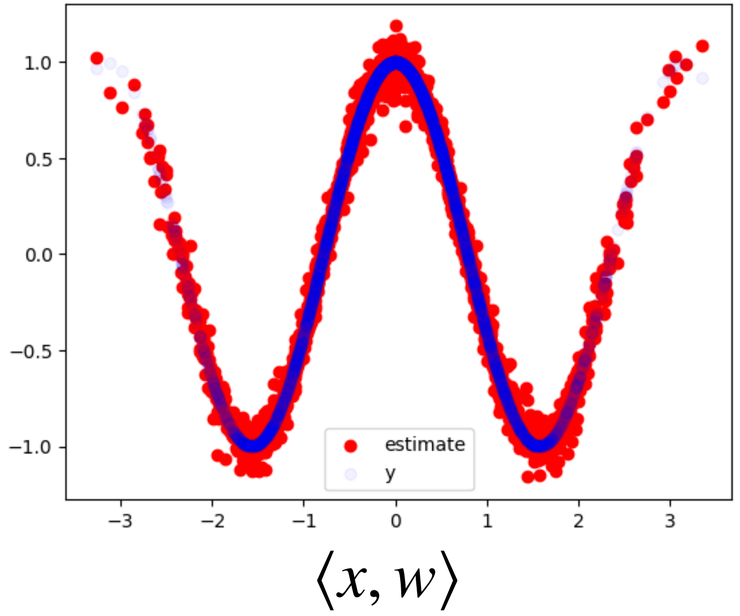
Task: improve random feature estimate

y

f_{100}



How?

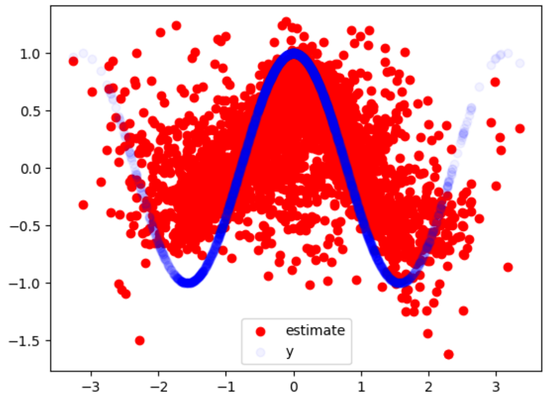


$$y \approx \underbrace{\sum_{i=1}^{100} \alpha_i \cos(\langle v_i, x \rangle + b_i), v_i, x \in \mathbf{R}^d}_{f_{100}(x)}$$

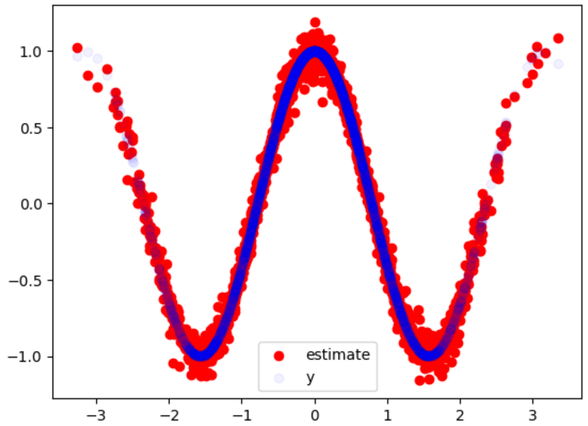
Task: improve random feature estimate

y

f_{100}



How?



$\langle x, w \rangle$

$$y \approx \underbrace{\sum_{i=1}^{100} \alpha_i \cos(\langle v_i, x \rangle + b_i)}_{f_{100}(x)}, v_i, x \in \mathbf{R}^d$$

Colab: https://shorturl.at/mWTLt

https://colab.research.google.com/drive/1SzLEF2EEMhjAXZkx86GuRxIwSkeTKbvs#scrollTo=V5xanwu8fDBt

- You can search or use ChatGPT

Let’s share our ideas

Solution

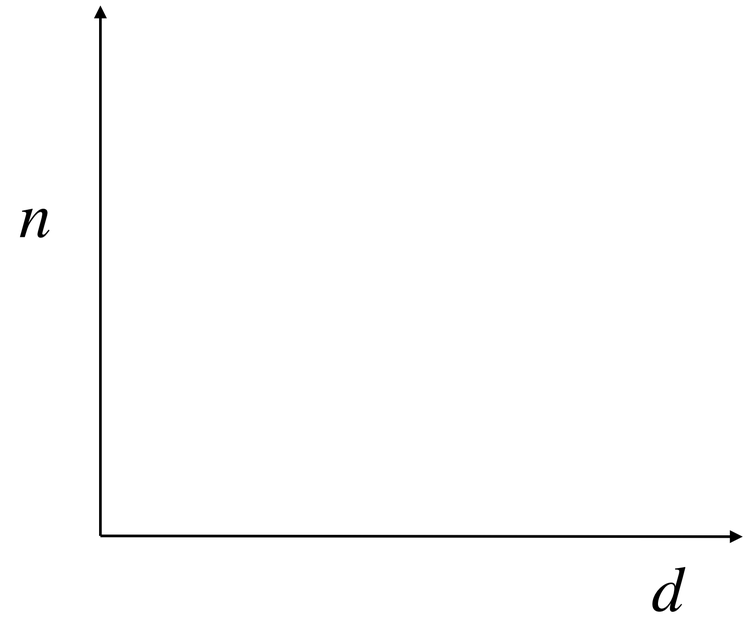
- Increasing the number of random features (n)

$$f(x) \approx \underbrace{\sum_{i=1}^{n} \alpha_i \cos(\langle v_i, x \rangle + b_i)}_{f_n(x)}, v_i, x \in \mathbf{R}^d$$

Task 2: observing the curse of dimensionality

$$f(x) \approx \underbrace{\sum_i \alpha_i \cos(\langle v_i, x \rangle + b_i)}_{f_n(x)}, v_i, x \in \mathbf{R}^d$$

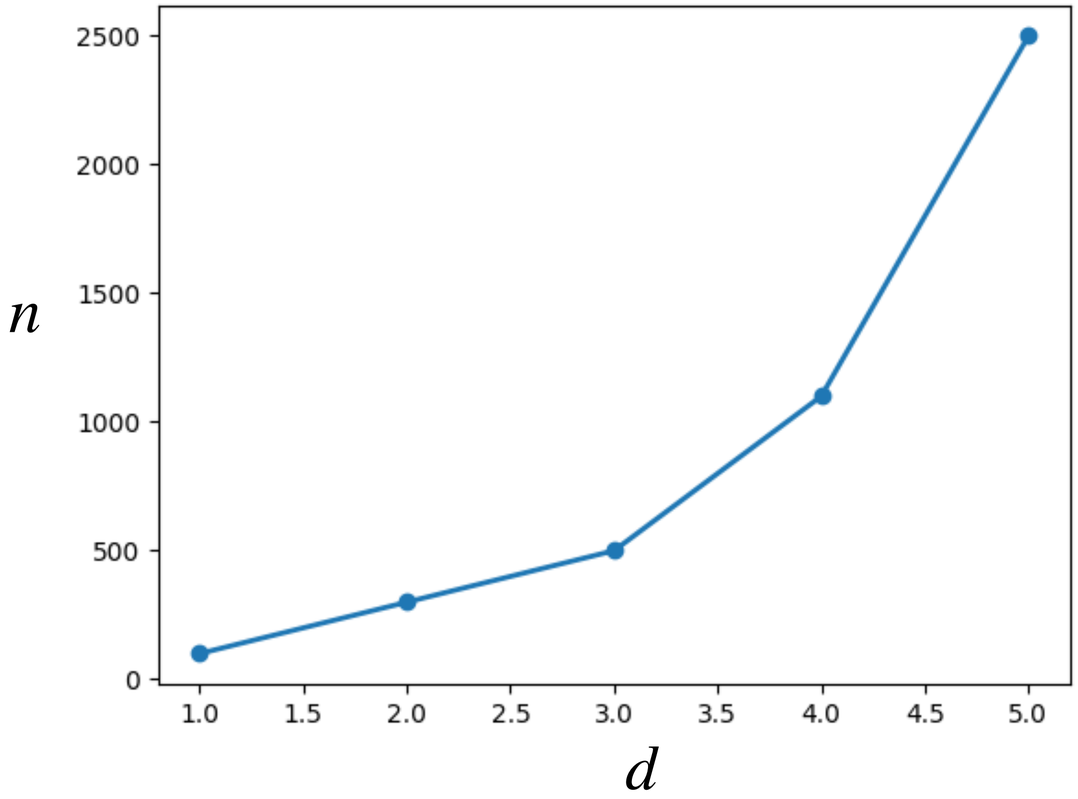
- $\mathbb{E}\left[(f(x) - f_n(x))^2\right] \leq 0.0005$
- How large n for $d = 1,2,3,4,5$



Solution

$$f(x) \approx \underbrace{\sum_i \alpha_i \cos(\langle v_i, x \rangle + b_i), v_i, x \in \mathbf{R}^d}_{f_n(x)}$$

- $\mathbb{E}\left[(f(x) - f_n(x))^2\right] \leq 0.0005$
- How large n for $d = 1,2,3,4,5$



Theory vs Lab

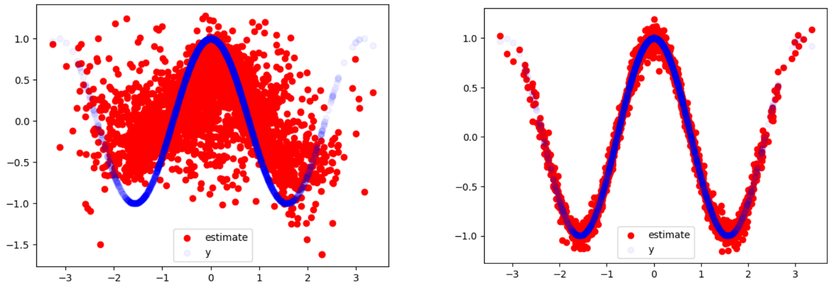
Theory

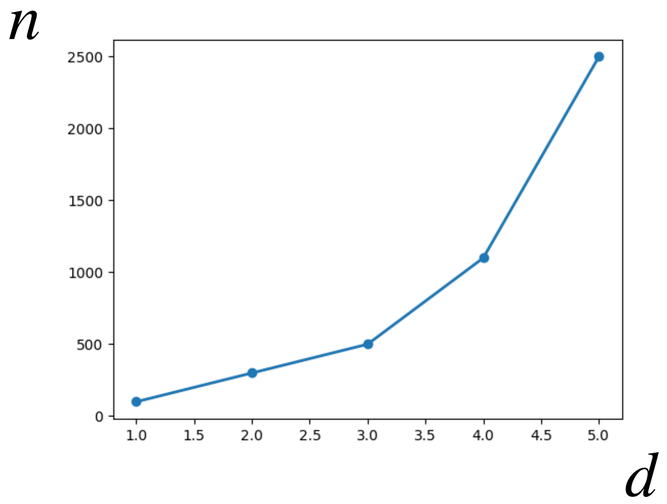
$$\mathbb{E}\left[\max_{i\le n}\langle v_i, w\rangle^2\right] \approx \frac{\log(n)}{d}$$

- Order statistics
- Memoryless property
- Integral bound for $\sum_i \frac{1}{i}$

Lab

$y = \cos(2 * \langle w, x \rangle)$





Connecting theory and lab

- $y = \cos(\langle w, x \rangle) \approx \sum_i \alpha_i \cos(\langle v_i, x \rangle)$
- To approximate y, there must exists a v_i such that $\|w - v_i\| \leq \epsilon$ Why?
  - $\cos(\langle v_i, x \rangle) - \cos(\langle w, x \rangle) \leq -2 \sin\left(\frac{\langle v_i + w, x \rangle}{2}\right) \sin\left(\frac{\langle v_i - w, x \rangle}{2}\right)$
  - $|\cos(\langle v_i, x \rangle) - \cos(\langle w, x \rangle)| \leq \|x\| \max\{\|w - v_i\|, \|w + v_i\|\}$
- $|\langle w, v_i \rangle| \geq 1 - \epsilon$

The next lecture

Random Features

$$y \approx \sum_i \alpha_i \cos(\langle v_i, x \rangle + b_i)$$

$v_i \sim p(w)$ depending on $k(x, y)$

Suffering from curse of dimensionality

Neural Networks

$$y \approx \sum_i \alpha_i \cos(\langle v_i, x \rangle + b_i)$$

v_i are optimized since $p(w)$ is unknown

Breaking curse of dimensionality

Thank you very much!