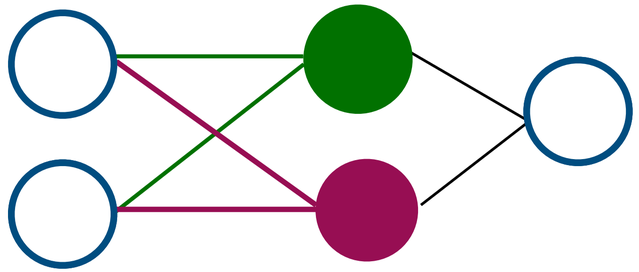
Neural Networks: A Theory Lab

Shallow neural networks

Optimization with gradient descent

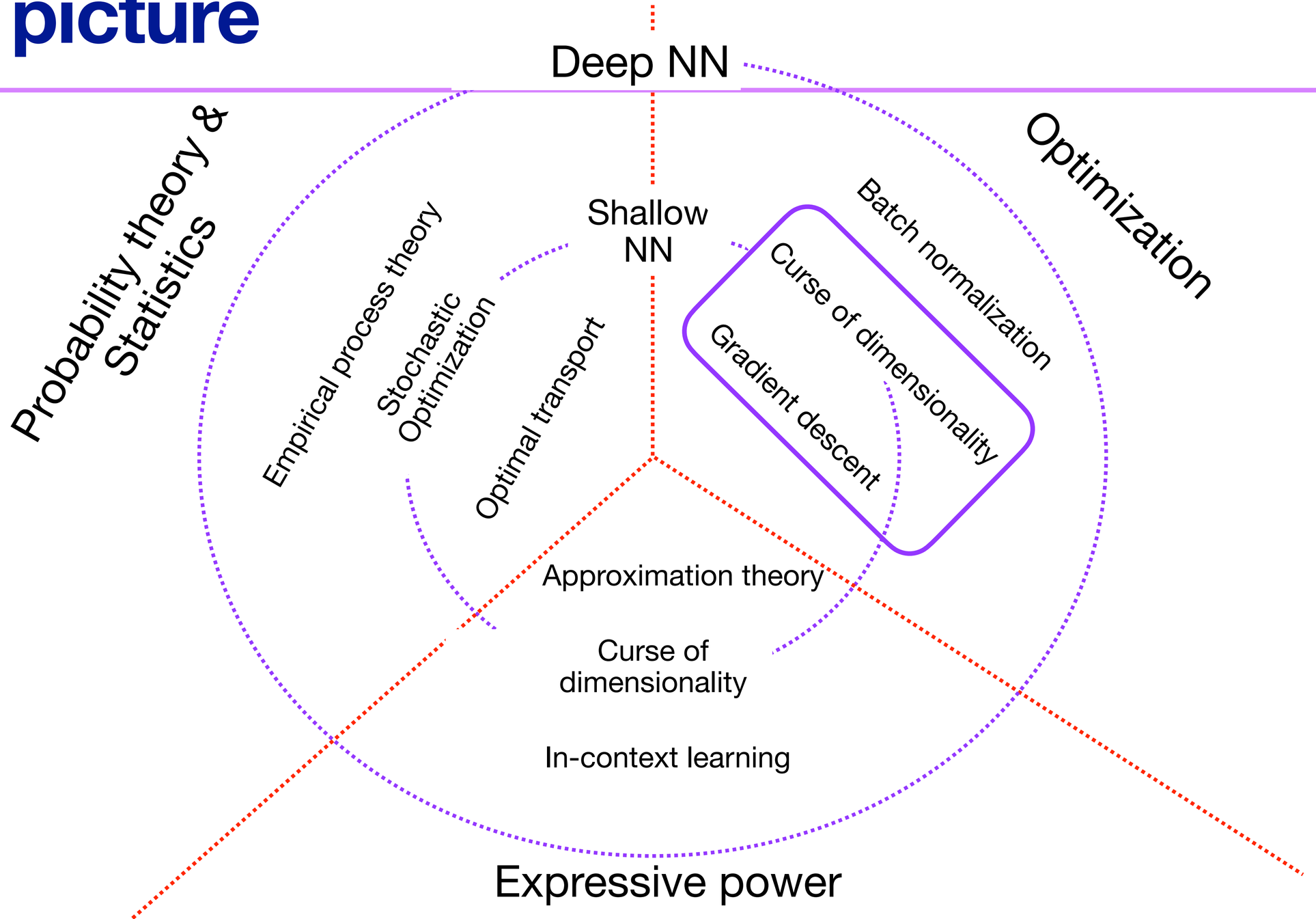


Hadi Daneshmand

University of Virginia

HW1 Questions?

Big picture


Deep NN
Probability theory & Statistics
Optimization
Shallow NN
Empirical process theory
Stochastic Optimization
Optimal transport
Batch normalization
Curse of dimensionality
Gradient descent
Approximation theory
Curse of dimensionality
In-context learning
Expressive power

Curses of dimensionality

Function approximation

$$\mathbb{E}(f(x) - f_n(x))^2 \geq c\frac{C_f}{d}\left(\frac{1}{n}\right)^{1/d}$$

$$n = \Omega((\frac{1}{\epsilon})^d) \text{ for } \epsilon-\text{accuracy}$$

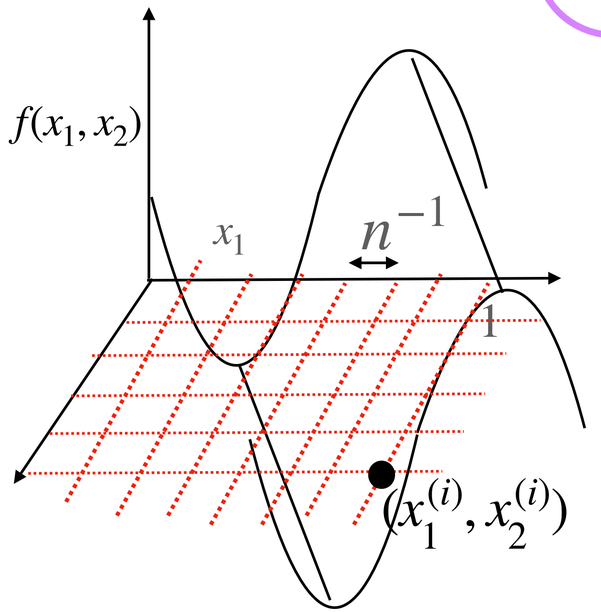
Optimization

$$f(x_0) - \min f(x) \leq \epsilon$$

$$\frac{1}{\epsilon^d} \text{ for } \epsilon-\text{accuracy}$$

Curse of dimensionality for optimization

	1D	2D	d-D
Grid search	$\frac{1}{\epsilon}$	$\frac{1}{\epsilon^2}$	$\frac{1}{\epsilon^d}$



To find x_0 such that $f(x_0) - \min_{x \in [0,1]^d} f(x)$, we need to evaluate $f(x_i)$ $O(\frac{L^d}{\epsilon^d})$ times

Gradient Descent (GD)

- Assume: f is twice differentiable and $\|\nabla^2 f(x)\| \leq \beta$

$$\nabla^2 f(x) \in \mathbb{R}^{d\times d}, [\nabla^2 f(x)]_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f|_x$$

- GD: $x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$

- **Descent:** $f(x_{k+1}) \leq f(x_k) - \frac{1}{\beta}\|\nabla^2 f(x_k)\|^2$

Does GD suffer from curse of dim for optimization?

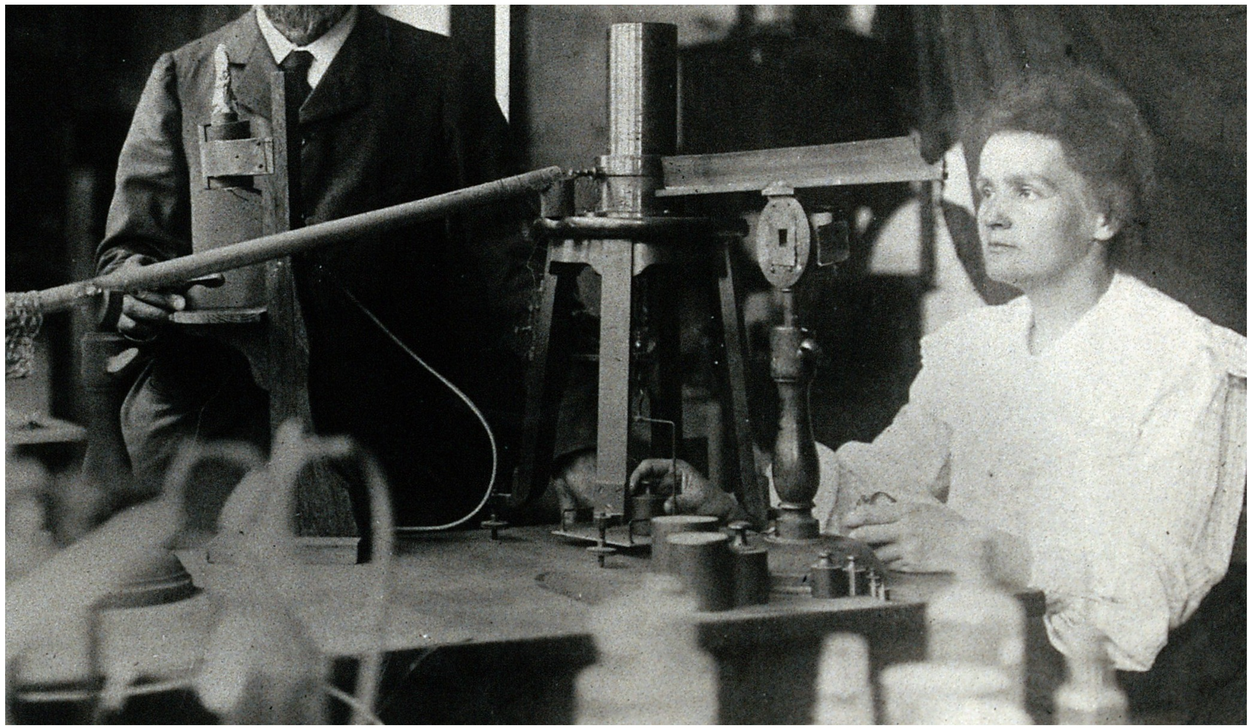
	1D	2D	d-D
Grid search	$\frac{1}{\epsilon}$	$\frac{1}{\epsilon^2}$	$\frac{1}{\epsilon^d}$

GD: $x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$

Can GD break the curse of dimensionality?

Warm-up

- GD: $x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$
- Descent: $f(x_{k+1}) \leq f(x_k) - \frac{1}{\beta}\|\nabla^2 f(x_k)\|^2$
- Question: What are limits $\lim_{k\to\infty} f(x_k), \quad \lim_{k\to\infty} x_k$?
- $f(x_k) \leq f(x_{k-1}) - \frac{1}{\beta}\|\nabla f(x_{k-1})\|^2 \leq f(x_0) - \frac{1}{\beta}\sum_{i=1}^{k}\|\nabla f(x_{i-1})\|^2$



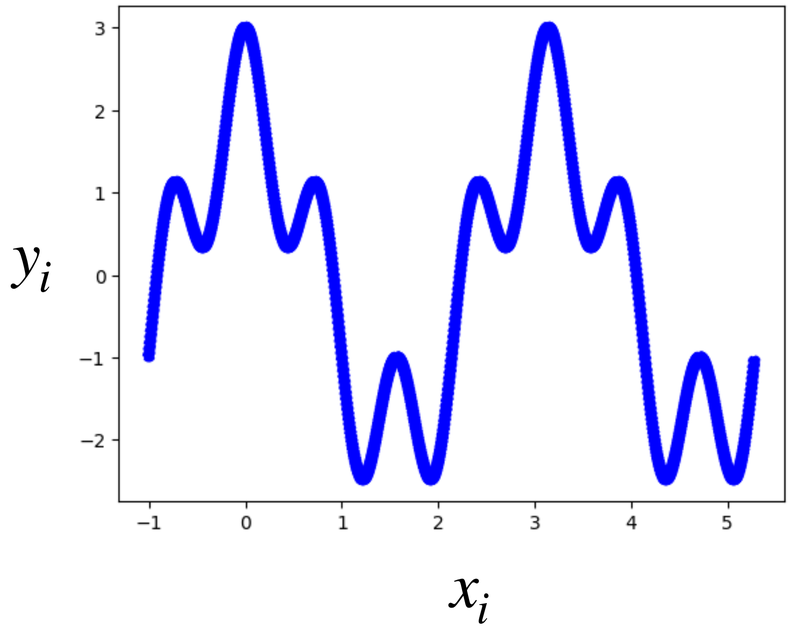
wikipedia: Marie Curie

Experiments

Get ready for hands-on group activity

Settings

- $f(x) = 2\cos(2x) + \frac{1}{2}\cos(8x)$ where
- Generating $\{(x_i, y_i = f(x_i))\}_{i=1}^{1000}$



Task 1: please plot the norm of gradient

Training Loss: $f(w, \alpha, b) := \frac{1}{1000} \sum_{i=1}^{1000} (y_i - \sum_{i=1}^{m} \alpha_i \cos(w_i x + b_i))^2$

#Neurons: $m = 2$

GD: $\theta_{k+1} = \theta_k - \gamma \nabla f(\theta_k), \theta := (w, b, \alpha)$

Colab: https://shorturl.at/xr3l9

Plot: $\|\nabla f(w_k, b_k, \alpha_k)\|^2$

Task 2: please plot the norm of gradient

Training Loss: $f(w, \alpha, b) := \frac{1}{1000} \sum_{i=1}^{1000} (y_i - \sum_{i=1}^{m} \alpha_i \cos(w_i x + b_i))^2$

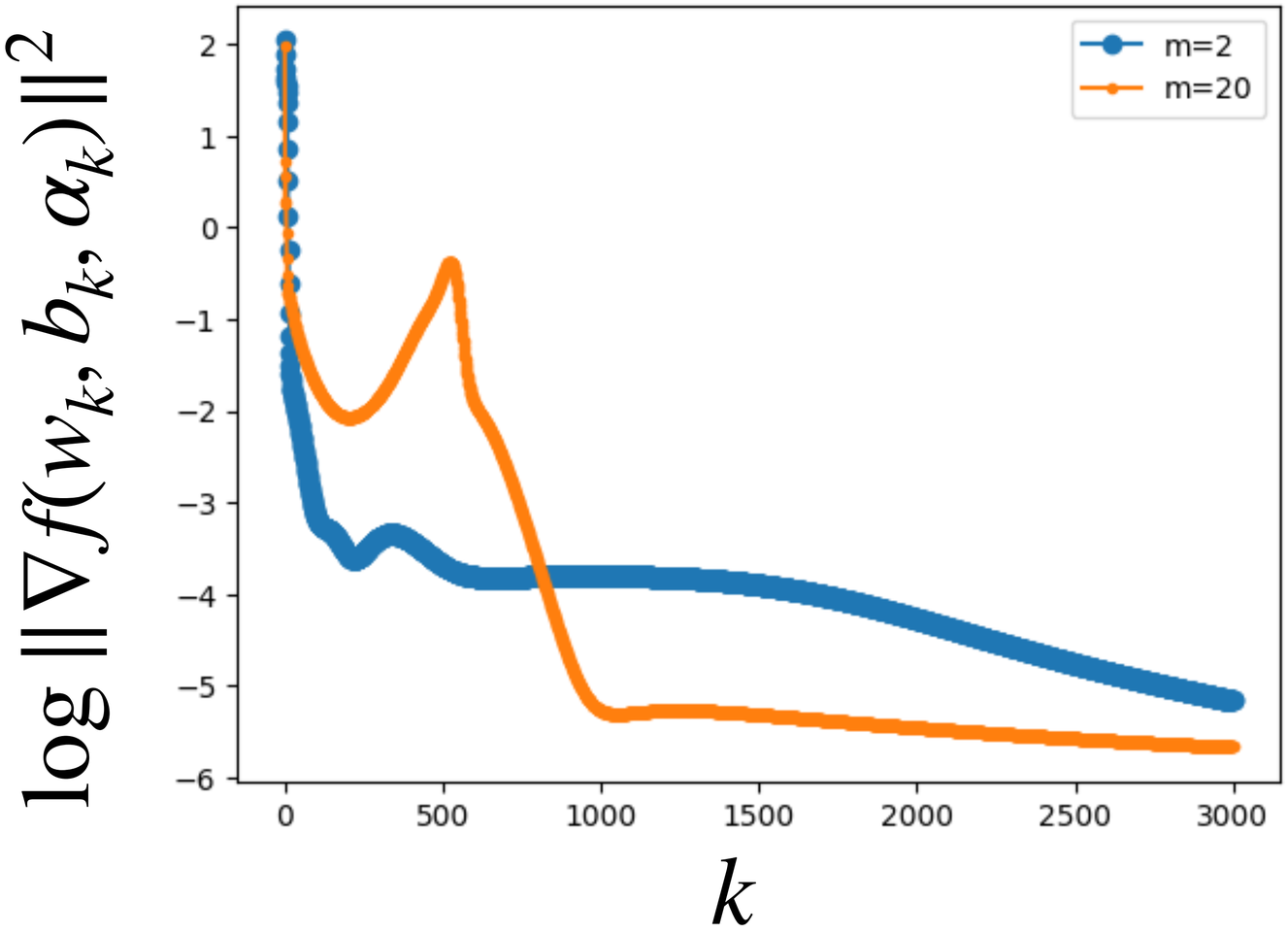
#Neurons: ~~$m = 2$~~ $m = 20$

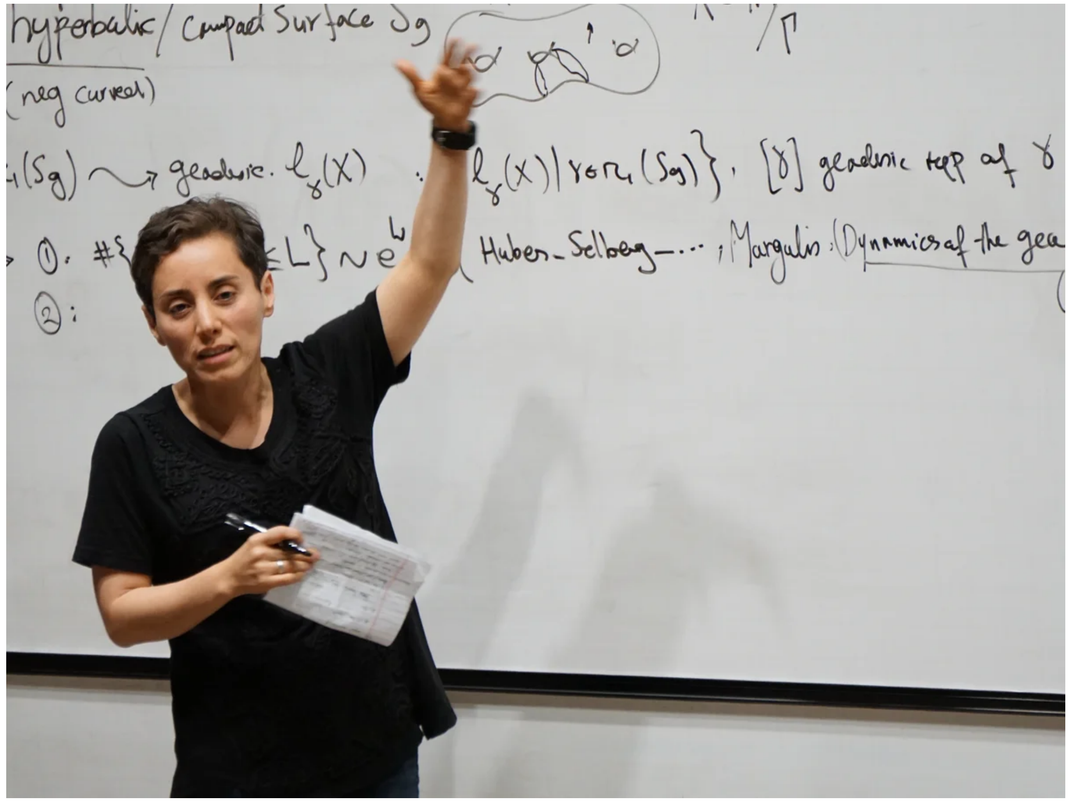
GD: $\theta_{k+1} = \theta_k - \gamma \nabla f(\theta_k), \theta := (w, b, \alpha)$

Colab: https://shorturl.at/xr3l9

Plot: $\|\nabla f(w_k, b_k, \alpha_k)\|^2$

Observation: Decay rate is dimension-independent





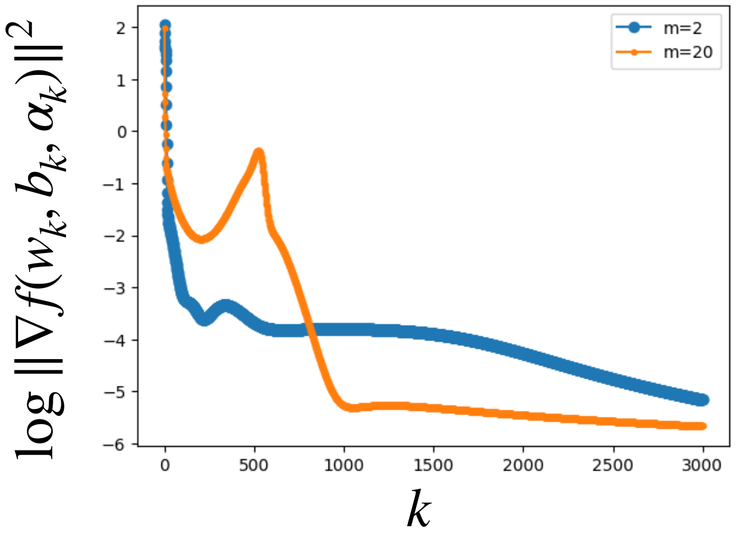
theguardian: Maryam Mirzakhani

Theory

Get ready for math

GD Convergence

- GD: $x_{k+1} = x_k - \gamma f'(x_k)$
- Assume: f is twice differentiable and $\|\nabla^2 f(x)\| \leq \beta$
- Statement: $\min_{k \leq n} \|\nabla f(x_k)\|^2 \leq \dfrac{f(x_0) - \min f(x)}{n}$ when $\gamma = \dfrac{1}{\beta}$



Convergence of gradient descent

- GD: $x_{k+1} = x_k - \gamma f'(x_k)$
- Statement: $\min_{k \le n} \|\nabla f(x_k)\|^2 \le \dfrac{\beta(f(x_0) - \min f(x))}{n}$ when $\gamma = \dfrac{1}{\beta}$
- Conclusion: GD breaks the curse of dimensionality!

Finding $\|\nabla f(x_k)\|^2 \le \epsilon$	**1D**	**2D**	**d-D**
Grid search	$\dfrac{1}{\epsilon}$	$\dfrac{1}{\epsilon^2}$	$\dfrac{1}{\epsilon^d}$
GD	$\dfrac{1}{\epsilon}$	$\dfrac{1}{\epsilon}$	$\dfrac{1}{\epsilon}$

No d

Comparing GD with grid search

- To optimize ChatGPT, $d = 10^{10}$; thus,

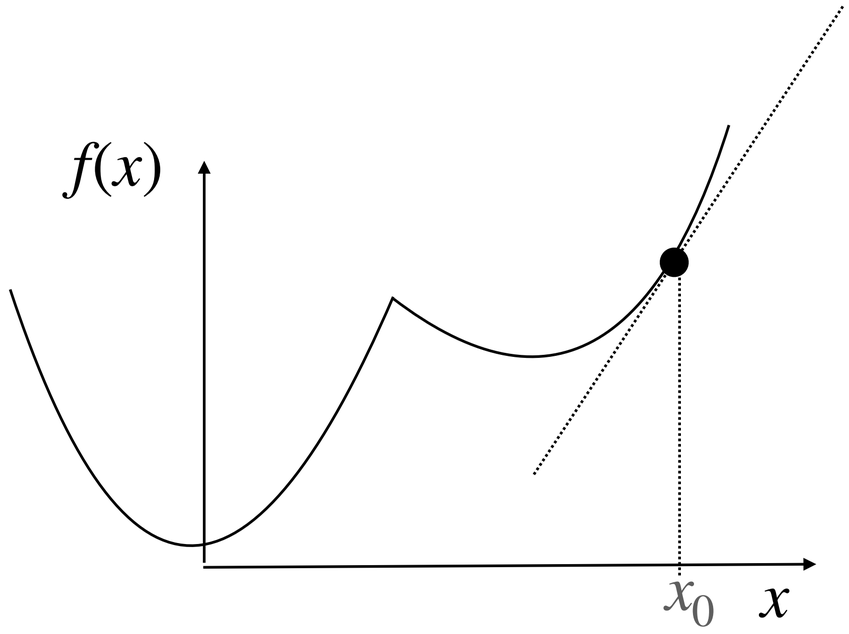
Finding $\|\nabla f(x_k)\|^2 \leq \frac{1}{2}$	**d-D**
Grid search	$O(2^{10000000000})$
GD	$O(1)$

Let's prove

- GD: $x_{k+1} = x_k - \gamma f'(x_k)$, f is twice differentiable and $\|\nabla^2 f(x)\| \leq \beta$
- Prove: $\min_{k \leq n} \|\nabla f(x_k)\|^2 \leq \frac{\beta(f(x_0) - \min f(x))}{n}$ when $\gamma = \frac{1}{\beta}$
- Hint: Use $f(x_k) \leq f(x_{k-1}) - \frac{1}{\beta}\|\nabla f(x_{k-1})\|^2$ from the last lecture
- Solution: $\frac{1}{\beta}\sum_{k=1}^{n} \|\nabla f(x_k)\| \leq \sum_{k=1}^{n} f(x_k) - f(x_{k+1})$
- $\frac{1}{\beta}\sum_{k=1}^{n} \|\nabla f(x_k)\| \leq f(x_0) - f(x_{n+1})$(Telescoping series) $\leq f(x_0) - \min f$
- $\frac{1}{\beta} n \min_{k \leq n} \|\nabla f(x_k)\|^2 \leq \frac{1}{\beta}\sum_{k=1}^{n} \|\nabla f(x_k)\| \leq f(x_0) - f(x_{n+1})$(Telescoping series) $\leq f(x_0) - \min f$

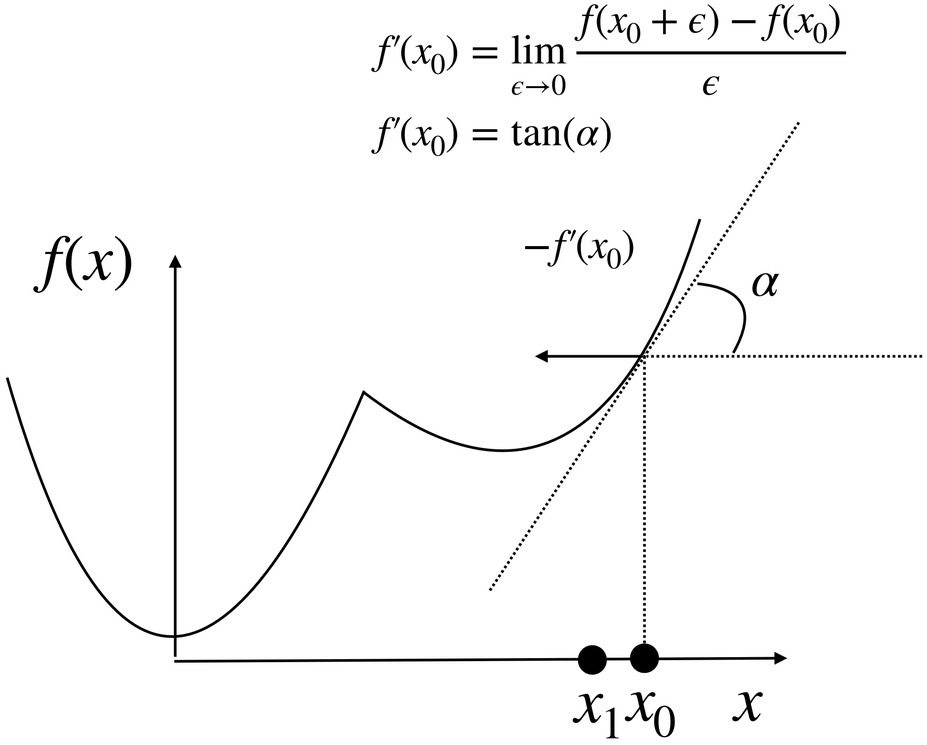
Example: GD convergence

- GD $x_{k+1} = x_k - \gamma f'(x_k)$
  - $\gamma > 0$ is called GD stepsize
- Question 1: predict the location of x_k for very large k



Example: GD convergence

- GD $x_{k+1} = x_k - \gamma f'(x_k)$
  - $\gamma > 0$ is called GD stepsize
- Question 1: predict the location of x_k for very large k
  - Where is x_1? left side of x_1

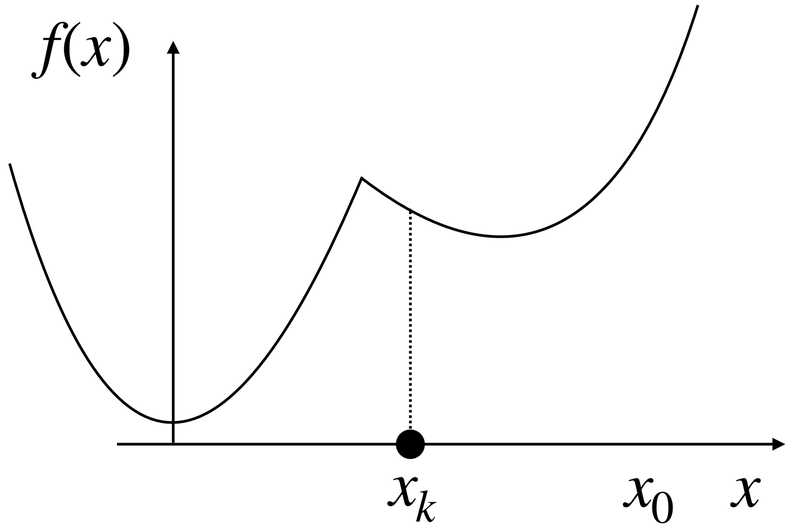


Example: GD convergence

- GD $x_{k+1} = x_k - \gamma f'(x_k)$
  - $\gamma > 0$ is called GD stepsize
- Question 1: predict the location of x_k for very large k
  - If x_k is here, where x_{k+1} will be?

$$f'(x_0) = \lim_{\epsilon \to 0} \frac{f(x_0 + \epsilon) - f(x_0)}{\epsilon}$$

$$f'(x_0) = \tan(\alpha)$$

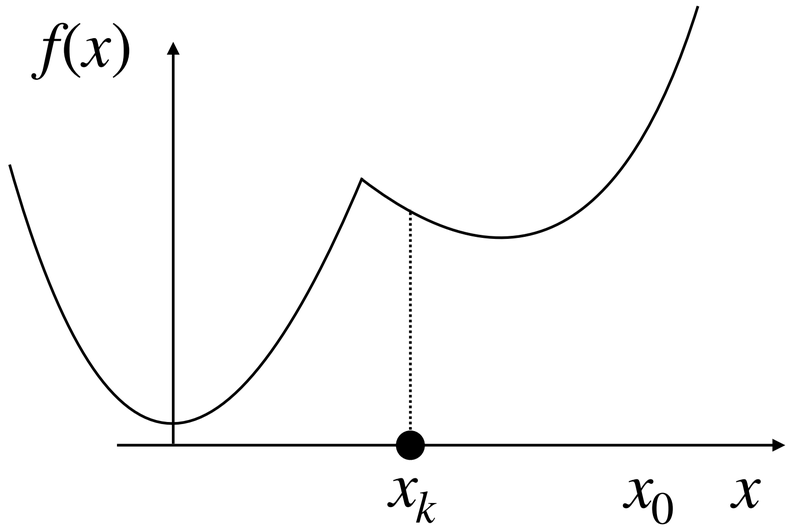


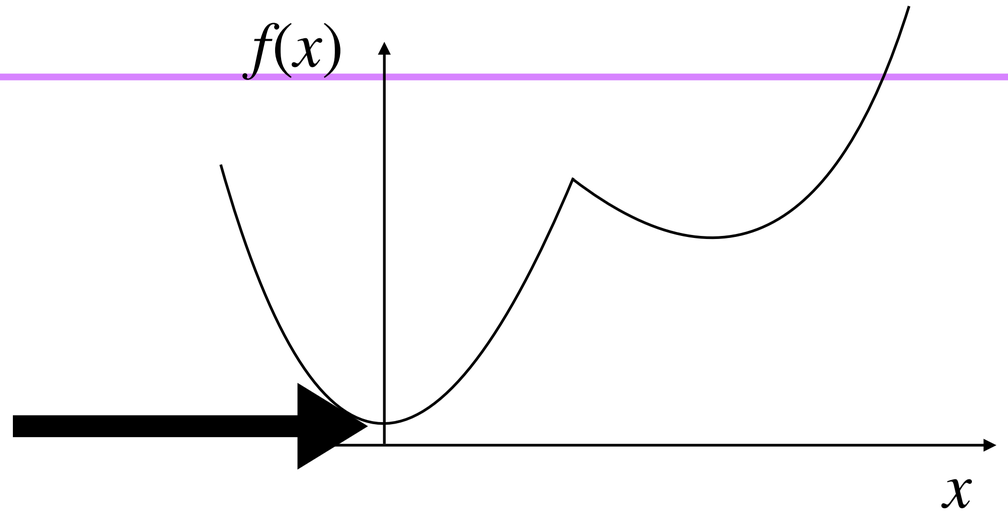
Example: GD convergence

- GD $x_{k+1} = x_k - \gamma f'(x_k)$
  - $\gamma > 0$ is called GD stepsize
- Question 1: predict the location of x_k for very large k
  - If x_k is here, where x_{k+1} will be?
  - Thus, x_k can not converge to the global minimum $x = 0$

$$f'(x_0) = \lim_{\epsilon \to 0} \frac{f(x_0 + \epsilon) - f(x_0)}{\epsilon}$$

$$f'(x_0) = \tan(\alpha)$$





When does GD converge to the global minimum?

Convergence to the minimizer

- When GD can recover $\arg\min f(x)$?
- Let C is a function class such that
  - All affine functions belongs to C
  - $\forall f \in C\colon \nabla f(x_*) = 0 \rightarrow x_* = \arg\min_x f(x)$
  - $\forall f, g \in C \;\&\; \alpha > 0, \beta > 0 \rightarrow \alpha f + \beta g \in C$
- Then GD, on $f(x) \in C$ recovers $\min f$ with no curse of dimensionality

Convex functions

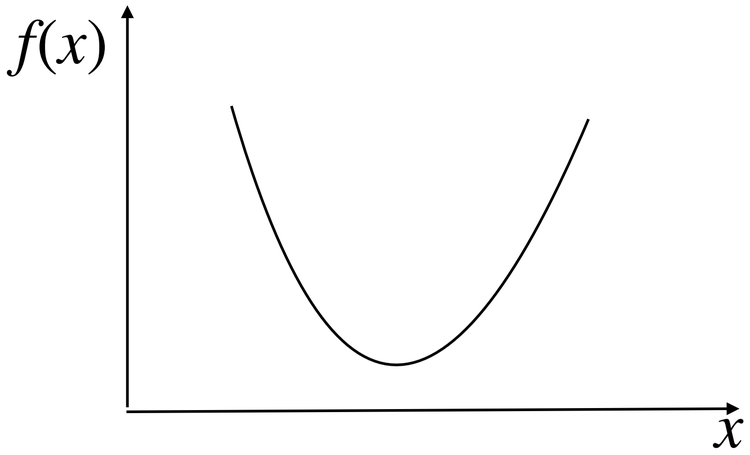
- Let C is a function class such that
  - All affine functions belongs to C
  - $\forall f \in C\colon\ \nabla f(x_*) = 0 \rightarrow x_* = \arg\min_x f(x)$
  - $\forall f, g \in C\ \&\ \alpha > 0, \beta > 0 \rightarrow \alpha f + \beta g \in C$
- C contains convex smooth functions
  - $f \in C \implies f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$

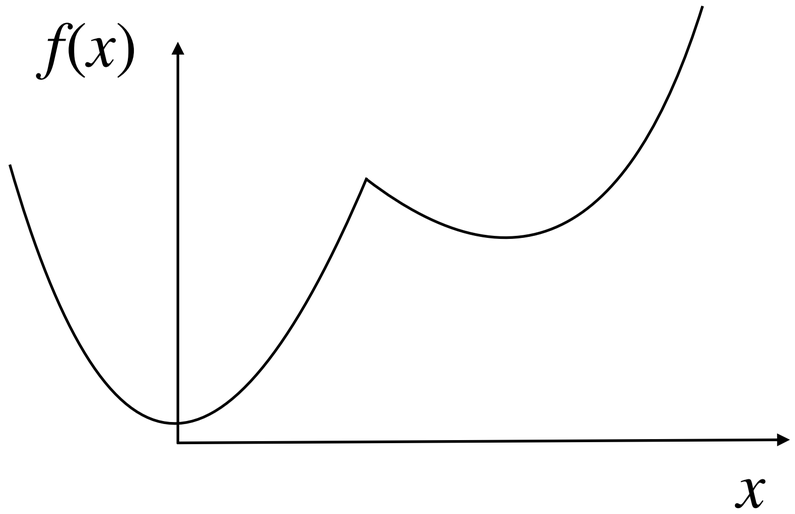
Take home exercise

- Prove: $f \in C \implies f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$

Geometric view

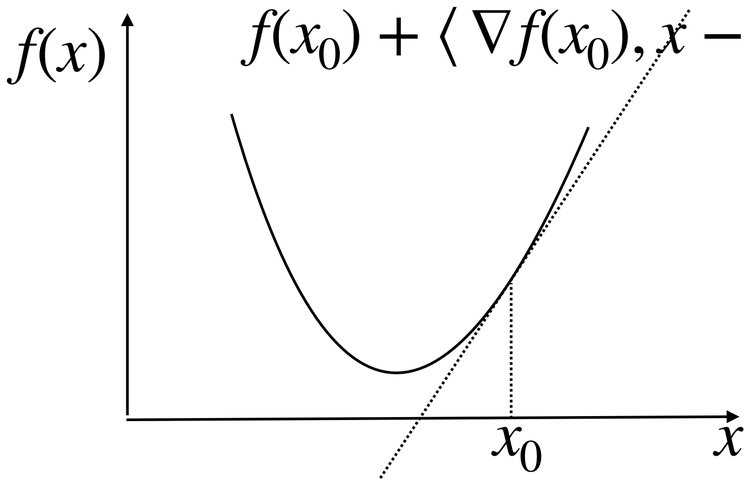
- For a convex $f(x)$: $f(x) \geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle$
- Question: Are the following function convex?



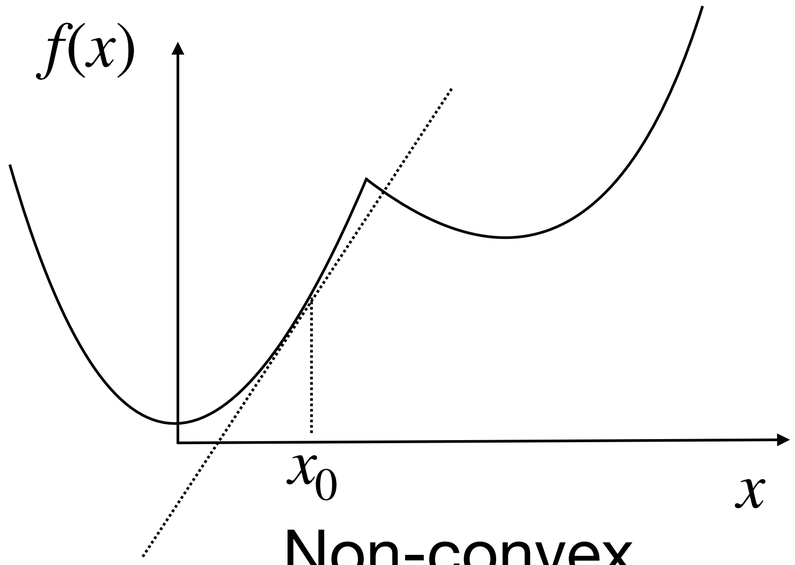


Geometric view

- For a convex $f(x)$: $f(x) \geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle$
- Question: Are the following function convex?



convex



Non-convex

Convexity and breaking the curse of dimensionality

- GD: $x_{k+1} = x_k - \gamma f'(x_k)$
- To find $f(x_k) - \min f(x) \leq \epsilon$, how large k has to be?

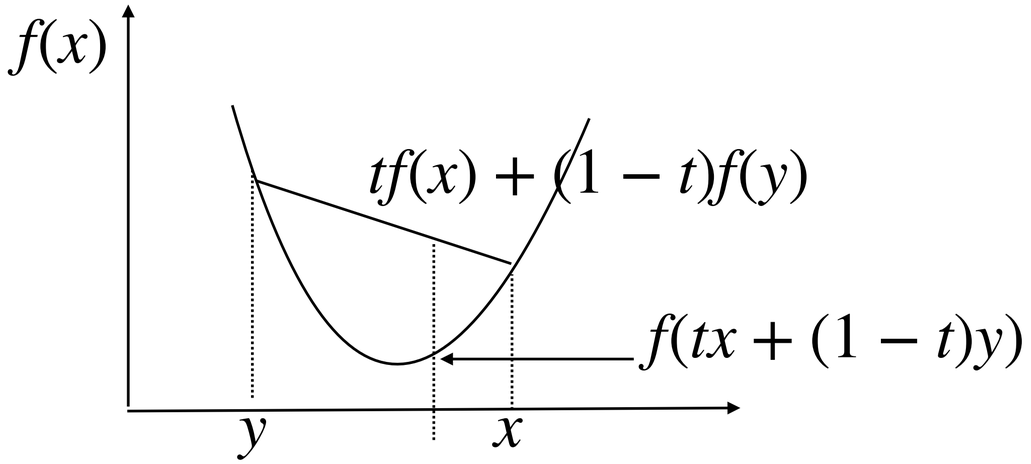
	1D	2D	d-D
Grid search	$\frac{1}{\epsilon}$	$\frac{1}{\epsilon^2}$	$\frac{1}{\epsilon^d}$
GD	$\frac{1}{\epsilon}$	$\frac{1}{\epsilon}$	$\frac{1}{\epsilon}$

No d

- Convergence rate can be improved depending on f

Convexity: general definition

- $\forall t \in [0,1] : f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$
- If f is also differentiable, then
  - $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ holds



convex

Next week lectures: optimization for neural nets

Language models and convex analysis

Non-convexity and neural nets

Thank you very much!