Optimization

Saturday, 1 March 2025 2:55 PM

- Optimization is everywhere!
- Most Machine Learning techniques can be seen as optimization problems

Examples :-

Linear Regression

$$\min_{x \in R^d} \|Ax - b\|_2^2$$

Support Vector Machines (SVM)

$$\min_{w \in R^d,\ w_0 \in R} \|w\|_2^2$$

$$\text{s.t. } (w^T x_i + w_0) y_i \geq 1 \quad \forall i$$

k-Means Clustering

$$\min_{C = \{c_1, c_2, \ldots c_k\}} \sum_{i=1}^{k} \sum_{p \in U_i} \|p - c_i\|_2^2$$

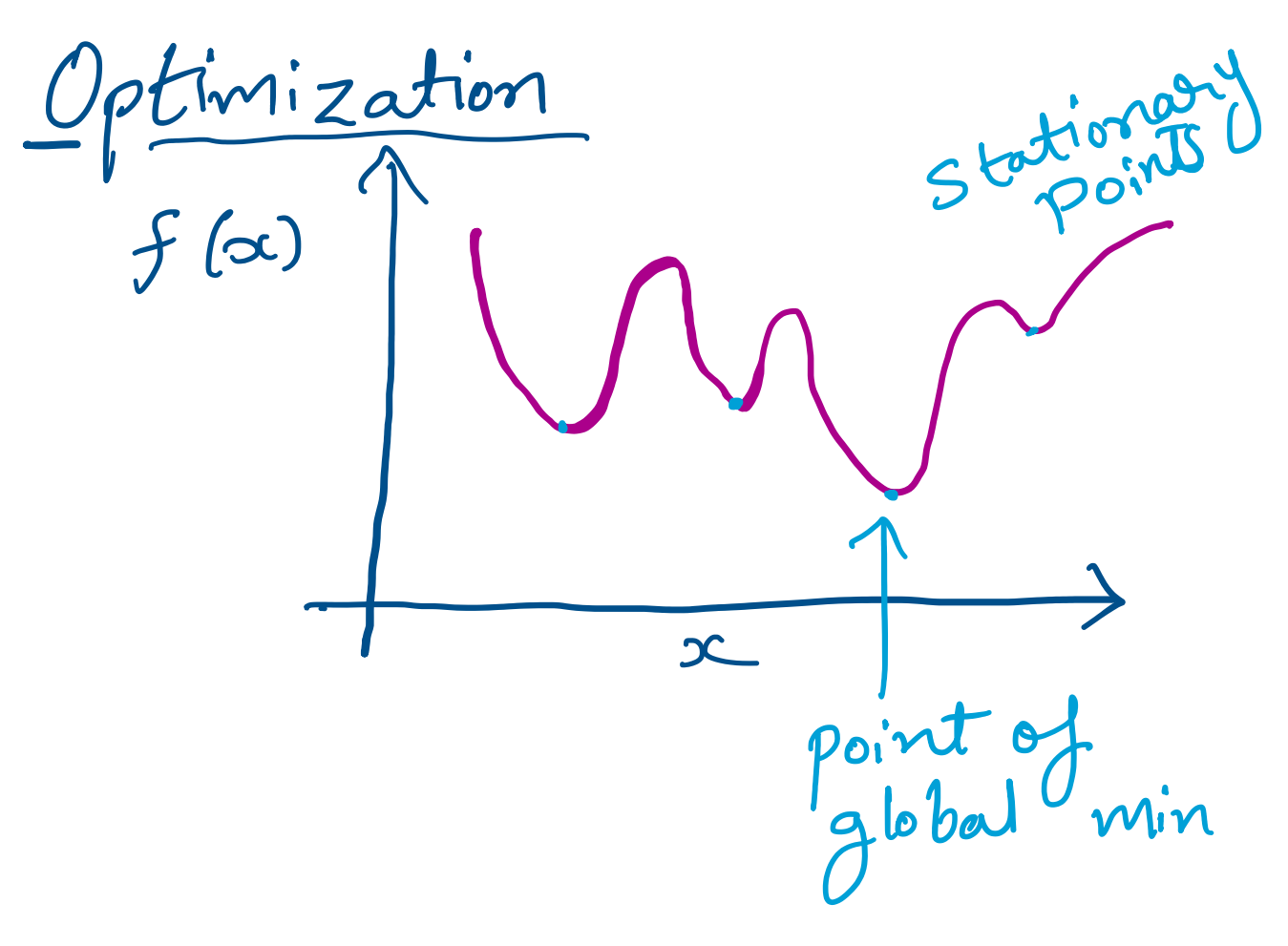
What is common?

ML Training ≈ Optimization — Some objective for given data

Lots of Avenues for optimization in ML/DS

- Modeling / Formulation
- Computational Aspect
- Combinatorial Optimization and . . . . .

Optimization



$f: R^n \to R$

Unconstrained optimization	Constrained
$\min f(x)$	$\min f(x)$
s.t. $x \in R^n$	s.t. $x \in X \subseteq R^n$

Some Important Results/Definitions

- Let $f: R^n \to R$ be continuously differentiable

x^* is pt. of local opt. $\Rightarrow \nabla f(x^*) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} = 0$

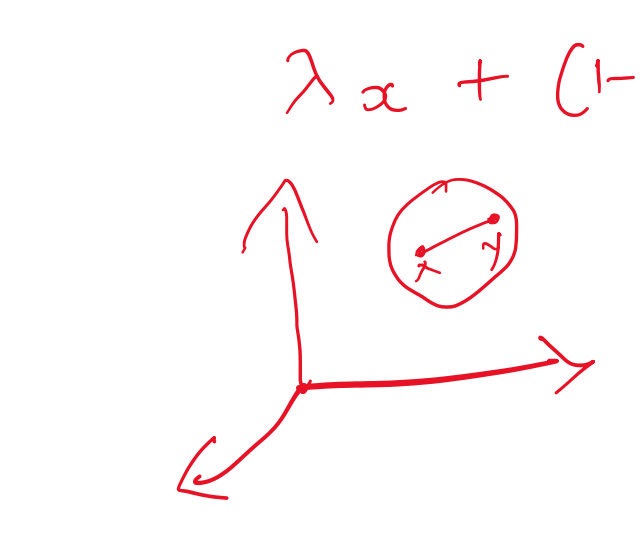
- Let f be twice differentiable

$$H(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ & & & \\ & & & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

then $\nabla f(x^*) = 0$ & $H(x^*)$ is P.D. $\Rightarrow$ x^* is a p.t. of local optimum (min)

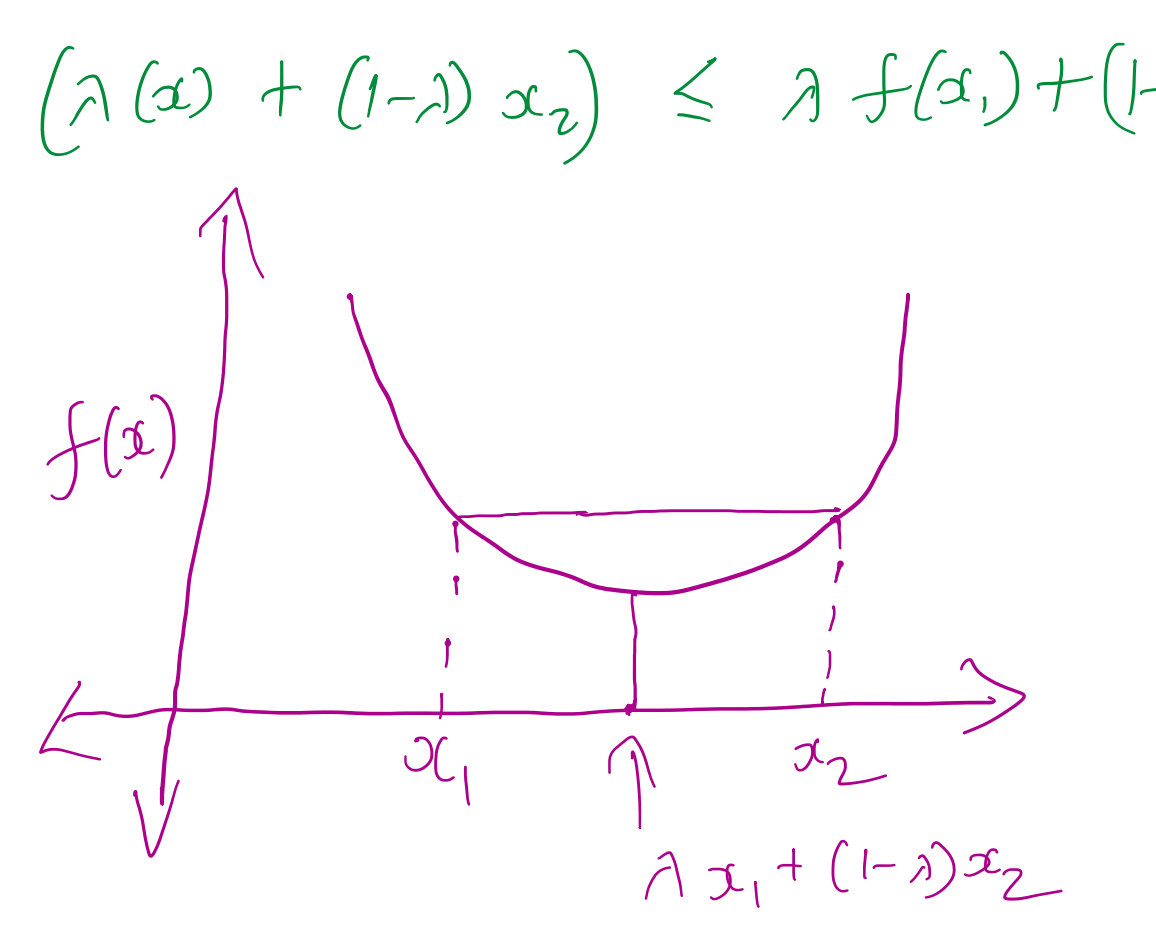
Convex Set :- $X \subseteq R^n$ is said to be convex if $\forall x, y \in X$

$$\lambda x + (1-\lambda) y \in X \quad \forall\ 0 \leq \lambda \leq 1$$



Convex fn :- $f: R^n \to R$ is defined over a convex set in following way $\forall x_1, x_2$

$$f(\lambda(x) + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$



$f(\cdot)$ is convex $\iff H(x)$ is P.S.D $\forall x \in X$

Why Convex fn are important?

- Many useful ML loss fn convex
- For Convex local Min $\Rightarrow$ Global Min

Gradient Descent Algorithm

Many ML algorithms are Sum structured

$$f(x) = \sum_{i=1}^{n} f_i(x)$$

$f(x)$ ↑ Total Loss; x ← parameter vector/Model; $f_i(x)$ ↑ Loss over i^{th} point

Idea :-

Want to minimize $f(x)$ i.e. try & find x^* s.t. $f(x^*)$ is minimum

Assumption : $f: R^d$ — $f(x) \Rightarrow R$ convex, differentiable, has global Min x^*

Output $x \in R^d$ such that $f(x^*) \leq \epsilon$

Algo :-

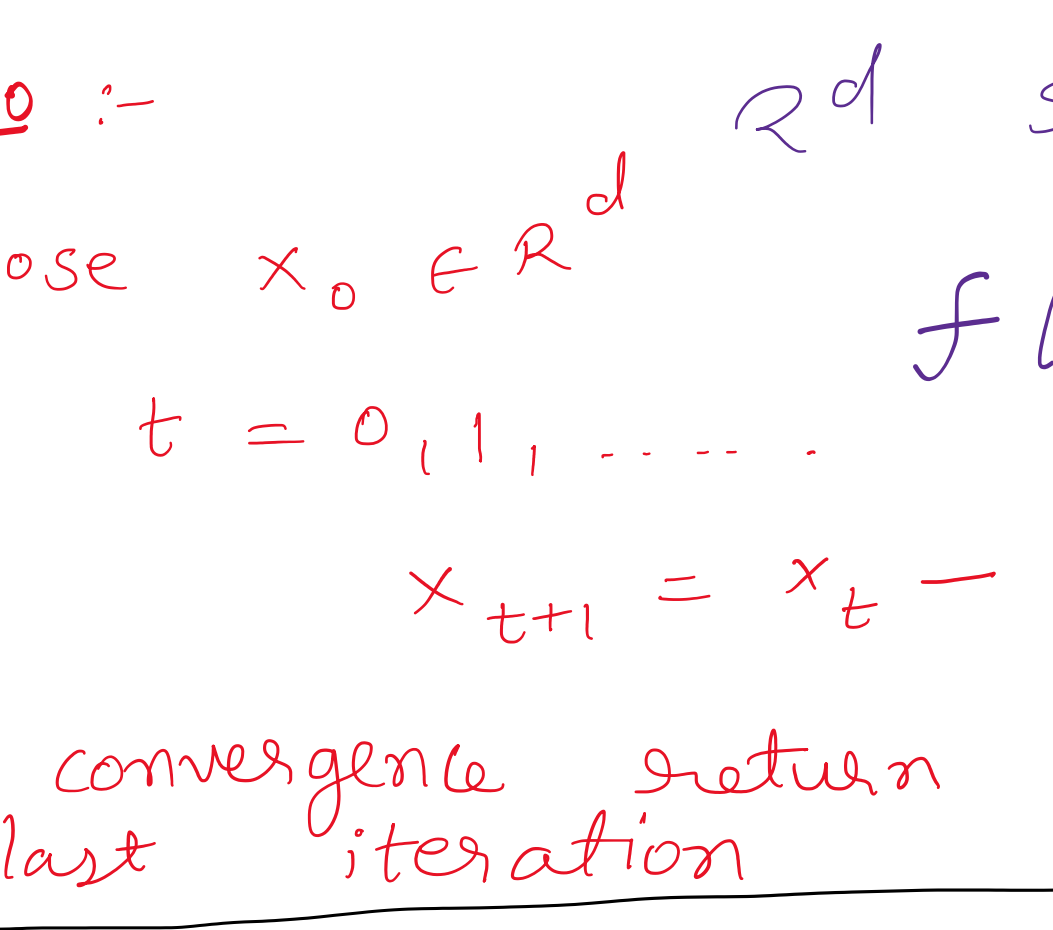
Choose $x_0 \in R^d$

For $t = 0, 1, \ldots$

$x_{t+1} = x_t -$

On convergence return in last iteration

Idea :-



Want to iterate FASTER

SGD (Stochastic Gradient Descent)

ALGO

Choose $x_0 \in R^d$

For $t = 0, 1, \ldots$ and $\gamma_t \geq 0$

sample $i \in [n]$ uniformly at random

$$x_{t+1} = x_t - \gamma_t \nabla f_i(x_t)$$

On convergence return x obtained in Last iteration

- Each iteration much faster

Effect of γ ? ← Discuss