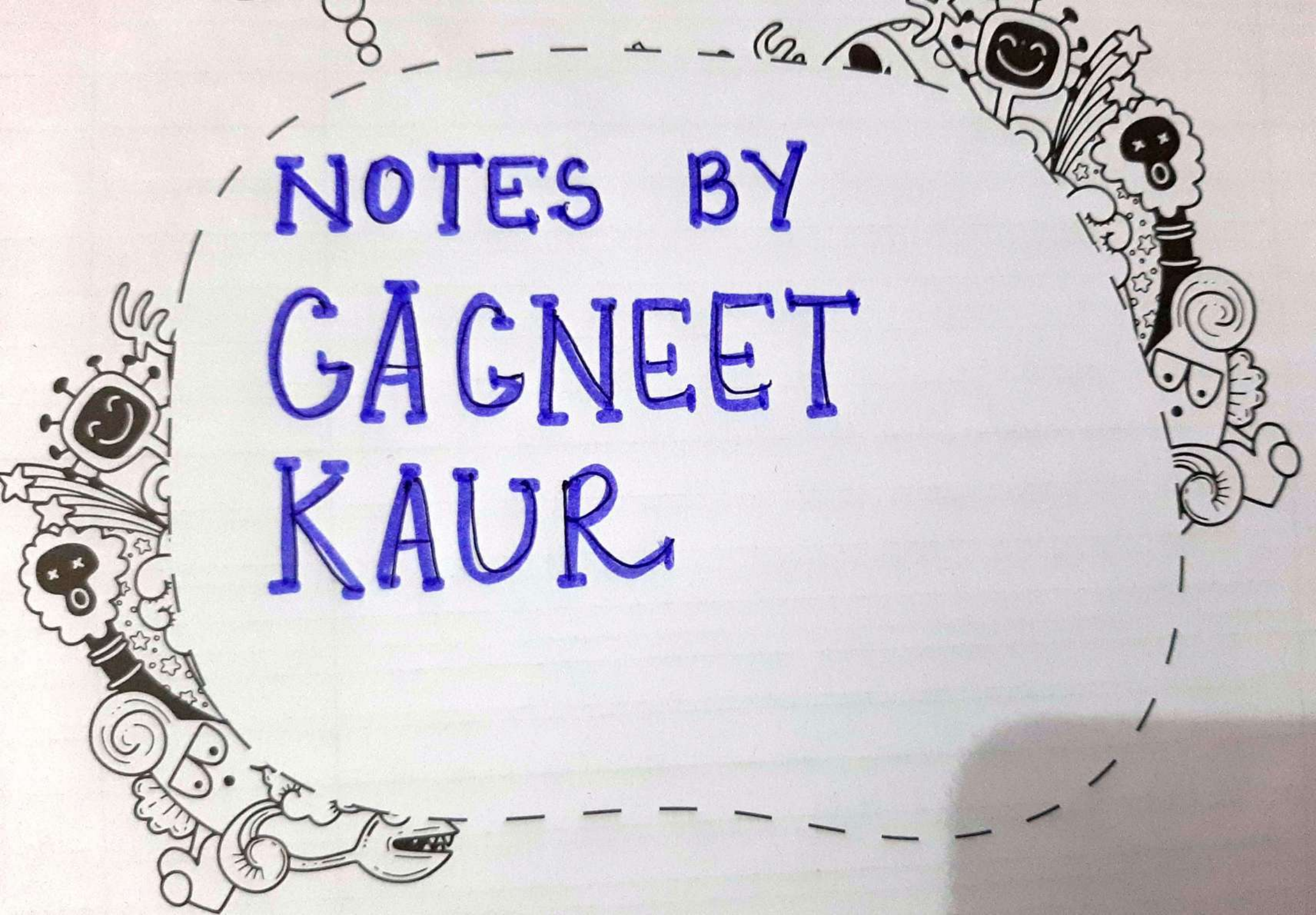

NOTES BY
GAGNEET
KAUR

Week 8 Optimization

Pillars Of Machine Learning

- Linear Algebra : to find the structure/relationship between the data
- Probability : to model noise in the data
- Optimization : to get the best model out of all possible models.

Introduction to Optimization

- WHY OPTIMIZATION ?

→ we care about finding the "best" classifier !

"Least loss" ⟷ "Maximum" reward

- GENERAL FORM OF OPTIMIZATION

OBJECTIVE

$$\min_{x \in \mathbb{R}^d} f(x)$$

$f(x)$ ← objective fn

$x \in \mathbb{R}^d$ ← variable/parameter

Constraints →

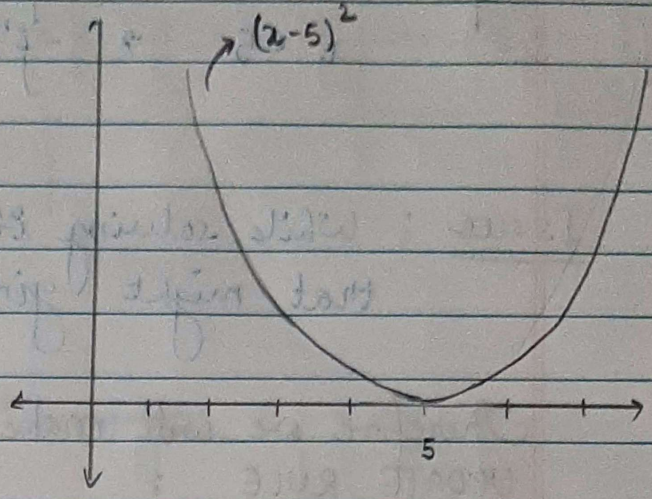
$$\begin{cases} g_i(x) \leq 0 & \forall\ i = 1, \ldots, k \\ h_j(x) = 0 & \forall\ j = 1, \ldots, l \end{cases}$$

$g_i(x) \leq 0$: Inequality constraints

$h_j(x) = 0$: Equality constraint

Solving an unconstrained optimization problem

$\min_{x \in \mathbb{R}} (x-5)^2$



Approach :

(i) Arbitrary Choice

Start with $x_0 \in \mathbb{R}$

(ii) Update Rule

for $t = 1, \ldots, T$

update, $\boxed{x_{t+1} = x_t + d}$ → direction — "good direction to move"

(iii) end

⊙ Given x, what is a good direction ?

		want
if	$x > 5$,	then more left $d < 0$
if	$x < 5$,	then more right $d > 0$

⇒ direction d will depend on x, i.e, d must be a function of x.

Using derivatives, we have : if $x > 5 \Rightarrow f'(x) > 0$

if $x < 5 \Rightarrow f'(x) < 0$

Thus, we will $\boxed{\text{CHOOSE } d = -f'(x)}$ as it satisfies the conditions we want.

∴ We have,

if $x > 5 \Rightarrow -f'(x) < 0$

$x < 5 \Rightarrow -f'(x) > 0$

Issue : While solving it, we could encounter some x_t that might give oscillating values.

Therefore we will make a small change in our UPDATE RULE :

$$x_{t+1} = x_t + \eta_t \left(-f'(x)\right)$$

η_t → step size (scalar +ve quantity)

- How to choose STEP SIZE ? (as a fn of t)

$$\boxed{\eta_t = \frac{1}{t+1}}$$ ← good step size sequence

$$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots, \frac{1}{t+1}$$

η_0 η_1 ....

Basic Algorithm for Unconstrained Optimization

GRADIENT DESCENT ALGORITHM

GOAL : $\min_{x \in \mathbb{R}} f(x)$

ALGORITHM : (1) Initialize at $x_0 \in R$

(2) for $t = 1, 2, \ldots$

$$x_{t+1} = x_t - \eta_t f'(x_t)$$

where $\eta_t = \frac{1}{t+1}$

(3) end

Properties of Gradient Descent

- If $\eta_t = \frac{1}{t+1}$, the algorithm converges
- Gradient descent converges to "Local Minimum"

LOCAL MINIMUM : $\exists\ \epsilon > 0$ such that $f(\hat{x}) \leq f(x) \quad \forall\ x \in [\hat{x} - \epsilon, \hat{x} + \epsilon]$

GLOBAL MINIMUM : $\forall\ x,\ f(x') \leq f(x)$

Gradient Descent and Taylor Series

Objective → $\min_{x \in R^d} f(x)$

Gradient Descent Update Rule : $x_{t+1} = x_t + \eta_t [-f'(x)]$

↓ what is so special about $-f'(x)$?

TAYLOR SERIES

$$f(\underbrace{x+\eta d}_{\bar{x}}) = f(x) + \eta d\, f'(x) + \eta^2 \frac{d^2}{2} f''(x) + \ldots\ldots$$

evaluations are all at x

Local Information gives Global Information

$$f(x+\eta d) = f(x) + \eta d f'(x) + \underbrace{f''(x)\frac{\eta^2 d^2}{2} + \ldots\ldots}_{\text{higher order terms}}$$

η: small the step size; d: "direction"

For small enough η,

$$f(x+\eta d) \approx f(x) + \eta d f'(x)$$

$$\Rightarrow \boxed{f(x+\eta d) - f(x) \approx \eta d f'(x)}$$

$f(x+\eta d)$: function evaluation at updated point along 'direction' d

$f(x)$: function evaluation at the current point

→ Want to choose a direction such that

$$\boxed{f(x+\eta d) - f(x) < 0}$$

⇒ want 'd' such that $\boxed{\eta d f'(x) < 0}$

η: small the constant ⇒ η doesn't change sign of this term

⇒ want 'd' such that $\boxed{d f'(x) < 0}$

For the choice of $d = -f'(x)$,

$$df'(x) = -[f'(x)]^2 < 0$$

Gradient Descent For Multivariate Function

Higher Dimensions : $f(x_1, x_2) = x_1^2 + 4x_2 + 8x_2^2$

Derivative $\Leftrightarrow$ Gradient $\rightarrow$ (vector of partial derivatives)
(for one variable) (for multi-variables)

$$\nabla f\left(\begin{bmatrix} a \\ b \end{bmatrix}\right) = \begin{bmatrix} \left.\frac{\partial f}{\partial x_1}\right|_{x_1 = a} \\ \left.\frac{\partial f}{\partial x_2}\right|_{x_2 = b} \end{bmatrix}$$

Update Rule for Gradient Descent in Higher Dimensions :

$$\vec{x}_{t+1} = \vec{x}_{t+1} + \eta\left(-\nabla f(x_t)\right)$$

$\vec{x}_{t+1}$: vector, $\vec{x}_{t+1}$: vector, η: scalar, $\nabla f(x_t)$: vector

Taylor Series in Higher Dimensions

$$f(\underbrace{x + \eta d}_{\text{vector}}) = \overbrace{f(x)}^{\text{scalar}} + \underbrace{\eta d^T \nabla f(x)}_{\text{scalar}} + \dots\dots$$

$f(x)$: vector

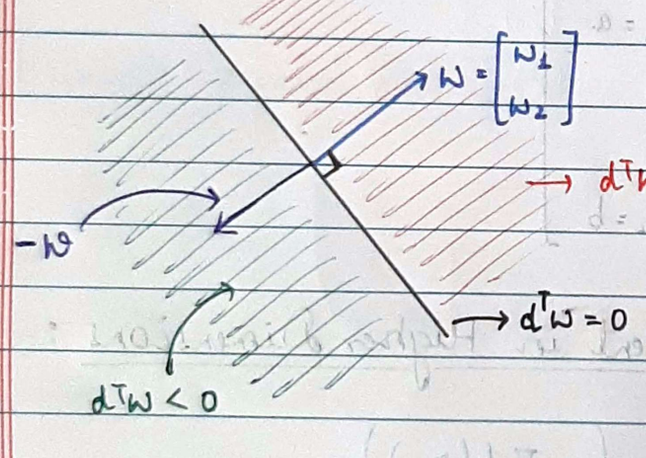
What is a good 'd' to choose ?

$$f(x+\eta d) - f(x) \approx \eta d^T \nabla f(x)$$

want d such that $\boxed{d^T \nabla f(x) < 0}$

NOTE: If $d = -\nabla f(x)$, $\boxed{(-\nabla f(x))^T (\nabla f(x)) = -\|\nabla f(x)\|_2^2 < 0}$

- There can be other d's in which if you move you will still reduce your function !!



$\boxed{d^T w = ?}$

$\boxed{-\nabla f(x) \text{ gives steepest descent}}$