Day 2

June 16, 2024

Linear Regression:

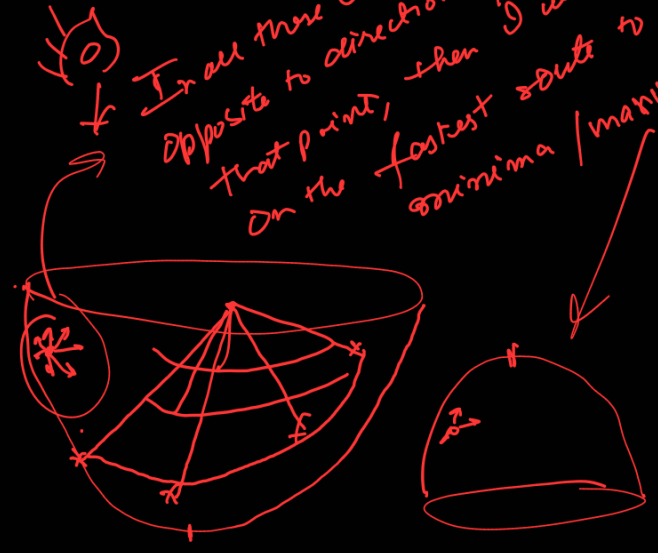
$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)})\right)^2 \quad \text{(MSE Loss)}$$

Gradient Descent (Generalized)

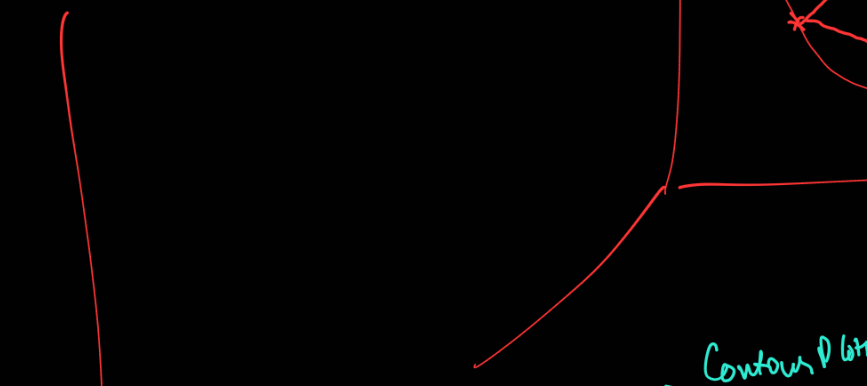
$$\theta_j^{(t+1)} \leftarrow \theta_j^{(t)} - \eta \cdot \nabla_{\theta_j} J(\theta)$$

$$\nabla_{\theta_j} J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} - h_\theta(x^{(i)})\right] x_j^{(i)}$$

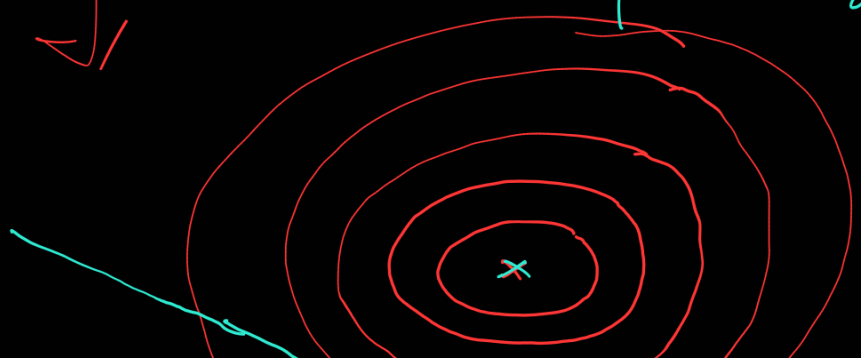
Multiple Directions



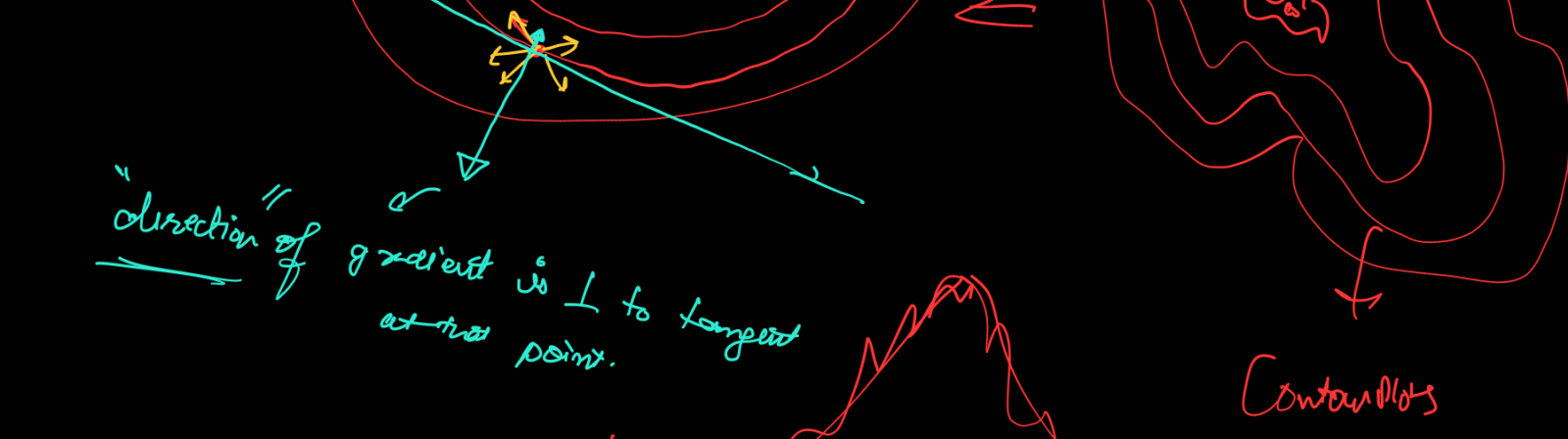
Contour Plots:



Contour Plot: Joining Points of same distance from a ref point.

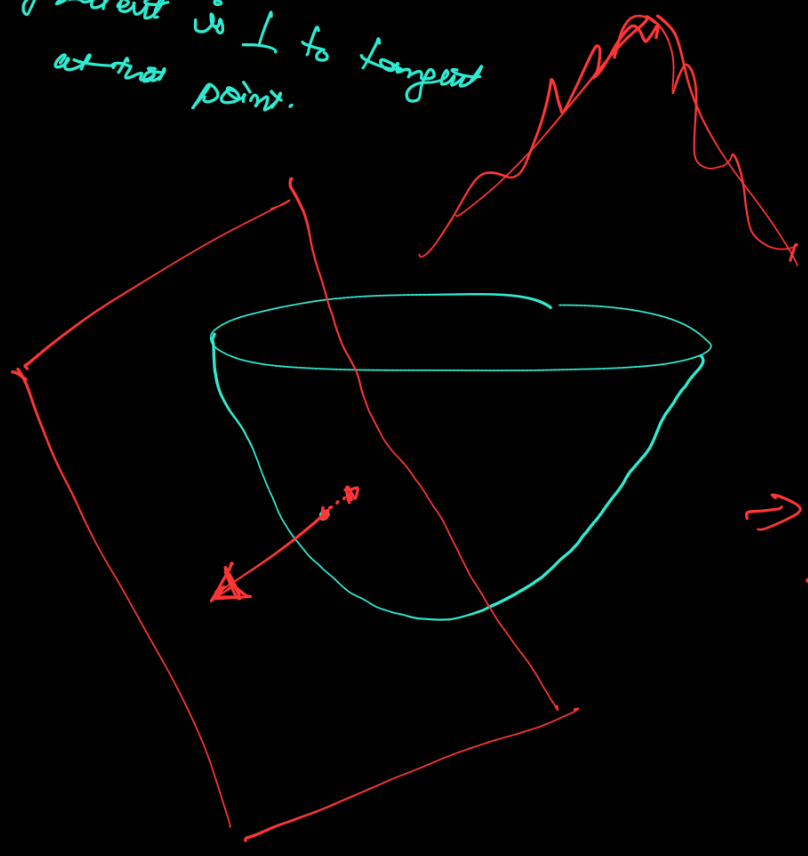


Depending on the curve.



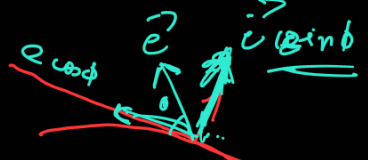
"direction" of gradient is ⊥ to tangent at that point.

Contour Plots



⇒ "3D"

Paper ← Katari°



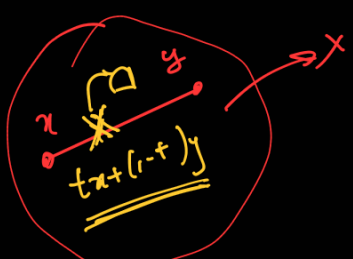
XII → Gaussian Surface

Convex Sets:

Relate to Section Formula (Interior)

Any set $X \subseteq \mathbb{R}^d$ is convex if,

$$tx + (1-t)y \in X, \quad x, y \in X, \quad t \in [0,1]$$



***

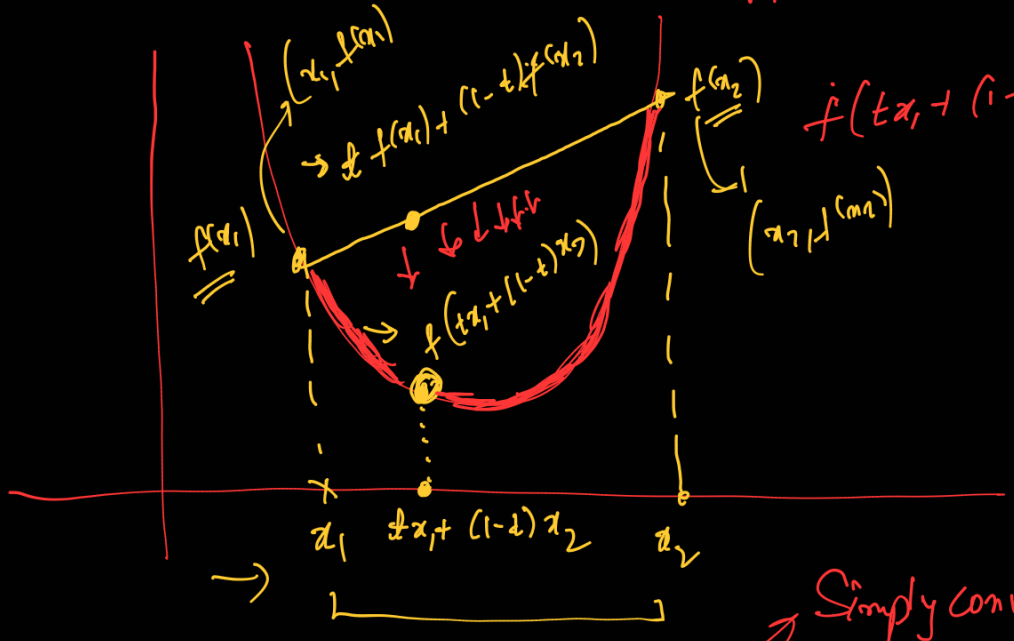
Convex functions: A function f is convex if: $\rightarrow$

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

$\forall x, y \in$ Domain (f) & $t \in [0, 1]$

↓

must also be a convex set.



$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

$$tP_1 + (1-t)P_2$$

$$t \in [0, 1]$$

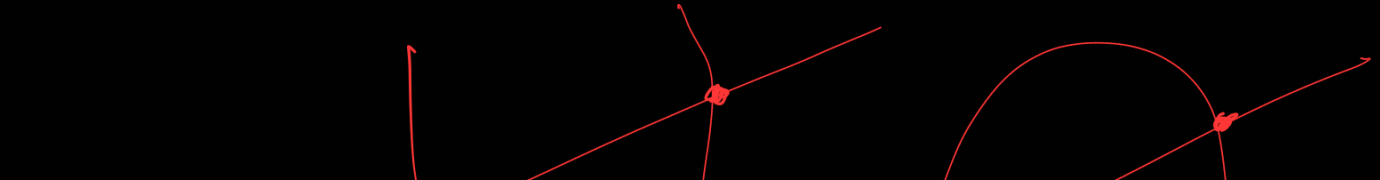
Simply convex

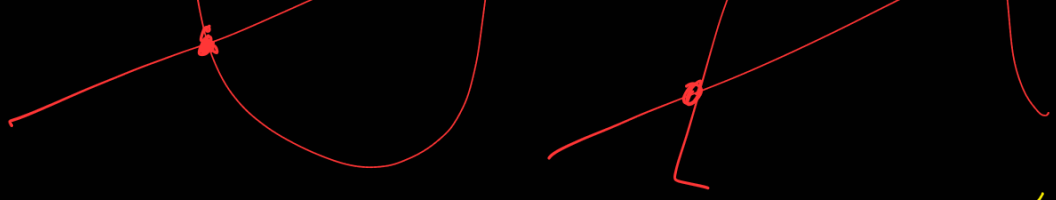
$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

Strictly convex

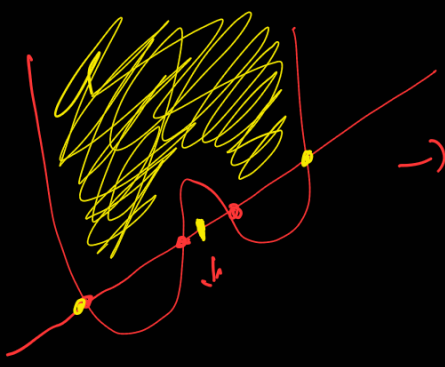
$$f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2)$$

$\rightarrow$ Convex functions: $\rightarrow$ Local minima / = Global minima / maxima maxima





✓ Set
✓ Function



→ X convex

$$l(\theta_1, \theta_2) = \theta_1^2 + \theta_2^2 + 2\theta_1\theta_2$$

$$\nabla_{\theta_1} l \rightarrow 2\theta_1 + 2\theta_2 \quad \left| \frac{\partial}{\partial \theta_1} \left(\theta_1^2 + \theta_2^2 + 2\theta_1\theta_2\right) \right.$$

$$\nabla_{\theta_2} l \rightarrow 2\theta_2 + 2\theta_1 \quad \left| \frac{\partial}{\partial \theta_2} \left(\theta_1^2 + \theta_2^2 + 2\theta_1\theta_2\right) \right.$$

$$\frac{\partial l(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial l(\theta_1, \theta_2)}{\partial \theta_1} \\ \frac{\partial l(\theta_1, \theta_2)}{\partial \theta_2} \end{bmatrix}$$

Whenever I have a multivariable function, it's derivative will be a vector i.e. take partial derivative w.r.t to every variable in the function.

$$l(\theta) = \theta_1^2 + \theta_2^2 + 2\theta_1\theta_2$$

$$\frac{\partial l(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial l(\theta_1, \theta_2)}{\partial \theta_1} \\ \frac{\partial l(\theta_1, \theta_2)}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} 2\theta_1 + 2\theta_2 \\ 2\theta_1 + 2\theta_2 \end{bmatrix} \leftarrow \nabla_{\theta} l(\theta)$$

$$\frac{\partial}{\partial\theta}\left(\frac{\partial \ell(\theta)}{\partial\theta}\right) = \frac{\partial \ell(\theta)}{\partial\theta^2} \rightarrow \frac{\partial}{\partial\theta}\begin{bmatrix}\frac{\partial \ell}{\partial\theta_1}\\ \frac{\partial \ell}{\partial\theta_2}\end{bmatrix}$$

$$\nabla_{k\times 1}\left(X_{m\times n}\right)$$

$$\begin{bmatrix}\frac{\partial \ell}{\partial\theta_1\theta_1} & \frac{\partial \ell}{\partial\theta_1\theta_2}\\ \frac{\partial \ell}{\partial\theta_2\theta_1} & \frac{\partial \ell}{\partial\theta_2\theta_2}\end{bmatrix}$$

$$D_{m\times n\times k\times \ell}$$

$$\frac{d}{dx}x^2 = 2x$$

$$\frac{d}{dx_1, dx_2}x^2 = \frac{dx^2}{dx_1}\;\frac{dx^2}{dx_2}$$

$$\nabla\begin{bmatrix}m_{11} & m_{12}\\ m_{21} & m_{22}\end{bmatrix}\rightarrow\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}\nabla\begin{bmatrix}m_{11} & m_{12}\\ m_{21} & m_{22}\end{bmatrix}\rightarrow\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}$$

4 X 4 = 16

$$= \nabla_{2\times 2}\left(2\times 2\right)$$

$$\begin{bmatrix}\begin{bmatrix}\frac{da_{11}}{dm_{11}} & \frac{da_{11}}{dm_{12}}\\ \frac{da_{11}}{dm_{21}} & \frac{da_{11}}{dm_{22}}\end{bmatrix} & \begin{bmatrix}\frac{da_{12}}{dm_{11}} & \frac{da_{12}}{dm_{12}}\\ \frac{da_{12}}{dm_{21}} & \frac{da_{12}}{dm_{22}}\end{bmatrix}\\ \begin{bmatrix}\frac{da_{21}}{dm_{11}} & \frac{da_{21}}{dm_{12}}\\ \frac{da_{21}}{dm_{21}} & \frac{da_{21}}{dm_{22}}\end{bmatrix} & \begin{bmatrix}\frac{da_{22}}{dm_{11}} & \frac{da_{22}}{dm_{12}}\\ \frac{da_{12}}{dm_{21}} & \frac{da_{22}}{dm_{22}}\end{bmatrix}\end{bmatrix}$$

a_{11} a_{12} a_{21} a_{22}

← matrix → tensor

$2\times 2\times 2\times 2$

$\nabla_{2\times 2}\left(2\times 2\right)$

$$\nabla_{2\times 2} \quad (2\times 2)(2\times 2)$$

$$\begin{bmatrix} \frac{df}{dm_{21}} & \frac{df}{dm_{22}} \end{bmatrix} \qquad \begin{bmatrix} \frac{df}{dm_{11}} & \frac{df}{dm_{12}} \end{bmatrix}$$

$$f = \theta_1^2 + \theta_2^2 + 2\theta_1\theta_2$$

$$\nabla_{1\times 2} \;(2\times 1) \rightarrow \begin{pmatrix} \frac{df}{d_{11}} & \frac{df}{d_{12}} \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} \frac{df}{d_{11}} \\ \frac{df}{d_{21}} \end{pmatrix} \quad \checkmark$$

$$(1\times 2)\times(1\times 2)$$
$$(1\times 2)\times(2\times 1)$$

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \rightarrow \nabla f = \begin{bmatrix} 2\theta_1 + 2\theta_2 \\ 2\theta_1 + 2\theta_2 \end{bmatrix}$$

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} (\theta_1 \; \theta_2) \qquad \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}$$

$$\rightarrow \nabla f = \begin{bmatrix} 2\theta_1 + 2\theta_2 & 2\theta_1 + 2\theta_2 \end{bmatrix}$$

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} (\theta_1 \; \theta_2)$$

$$\begin{bmatrix} e_1 \\ e_3 \end{bmatrix} \quad \begin{bmatrix} e_2 \\ e_4 \end{bmatrix} \quad \begin{bmatrix} e_1 & e_2 \\ e_3 & e_4 \end{bmatrix}$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)}) \right)^2 \leftarrow (\text{Loss})$$

$$\theta_j^{(t+1)} \leftarrow \theta_j^{(t)} - \eta \cdot \frac{dJ(\theta)}{d\theta_j} \leftarrow \text{Iterative}$$

Convergence Criteria: I can define myself.

I have 2 options: (i) Either I monitor the Loss value

(ii) either I monitor a param val of my interest.

(iii) either I monitor grad(J)

ML model → what we are trying to learn.

1. $|J(\theta^{t+1}) - J(\theta^{t})| < \delta$ (loss)

2. $|\theta^{t+1} - \theta^{t}| < \epsilon$ (Parameters)

3. $|\nabla_\theta J(\theta)|_{\theta^{(t+1)}} < \|\delta\| = 0$

Hyperparams

$[v_{11}\ v_{12}\ v_{13}]$

$L1$:

$L2$: $\sqrt{v_{11}^2 + v_{12}^2 + v_{13}^2}$

5

10,000

0.98

0.979912817

θ : 10

28th → 29th

$(\theta_9) : (0.008) \rightarrow (0.007954) \Rightarrow 0.0005$ (ϵ)

↳ Vector

$\left(\begin{bmatrix} \quad \end{bmatrix} - \begin{bmatrix} \quad \end{bmatrix} \right)$ gradient (vec) $\|[\quad]\|_N < \text{Threshold}$

(2 3)

$L2$: $\sqrt{2^2 + 3^2} = \sqrt{4+9} = (\sqrt{13}) < 4$ ✓

$y = mx + c$

Param1 Param2 → Trying to learn params of st. line.

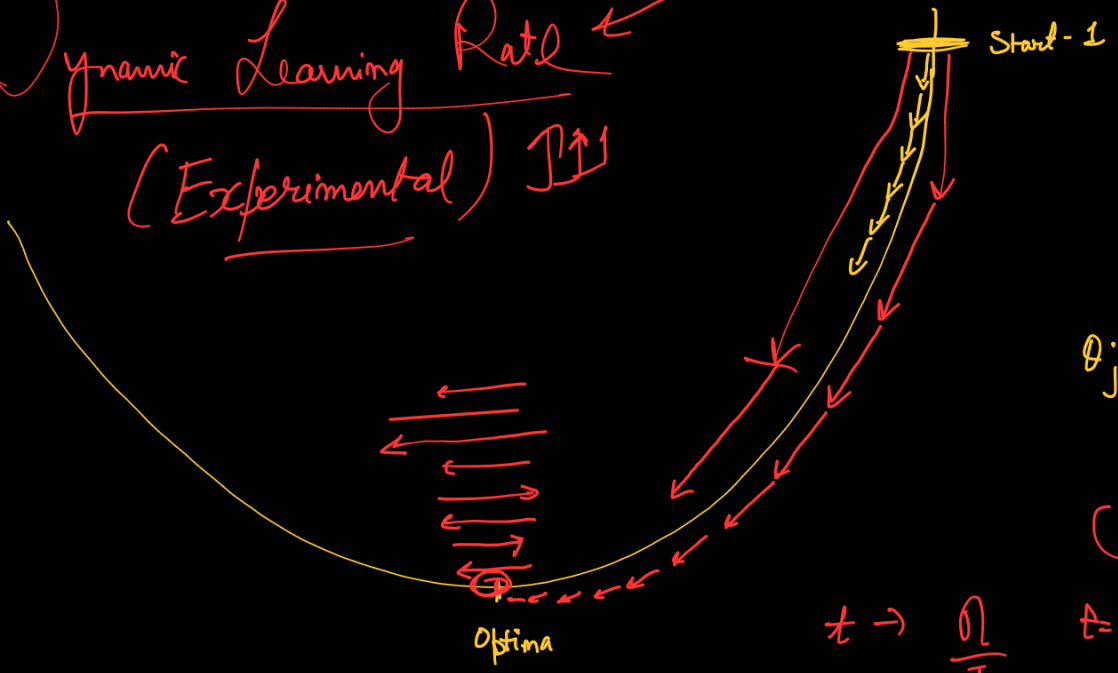
Good obs!!

Hyperparameter: We give them.

4. $\left| \nabla J(\theta)^{(t+1)} - \nabla J(\theta)^{(t)} \right| < \varepsilon$

$|\theta^{t+1} - \theta^{t}| < \eta$

Dynamic Learning Rate (Experimental) [1]



$\theta_j^{(t+1)} := \theta_j^{(t)} - \eta \cdot \frac{\partial J(\theta)}{\partial \theta_j}$

$t \rightarrow \frac{\eta}{t}$ $\quad$ (0.1)

$t=1 \quad 0.1$

$t=2 \quad \frac{0.1}{2} = 0.05$

$\eta \propto \frac{1}{t}$ or $\eta \propto \frac{1}{\sqrt{t}}$ $\quad \ll \frac{1}{t^2}$

Initial: $\eta_t = \frac{\eta_0}{t}$ $\quad$ $\eta_{t+1} = \eta_t$ $\quad$ $(\eta_{t+1} < \eta_t)$

Stochastic Gradient Descent (SGD)

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(y^{(i)} - \theta^T x^{(i)}\right)^2 \quad \text{(MSE)}$$

m # entire training data.

I add up m losses, then take average of m losses.

Advantage ⇒ I have seen my entire data.

$$\theta_j^{(t+1)} \rightarrow \theta_j^{(t)} - \eta \cdot \frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - \theta^T x^{(i)}\right) x_j^{(i)}$$

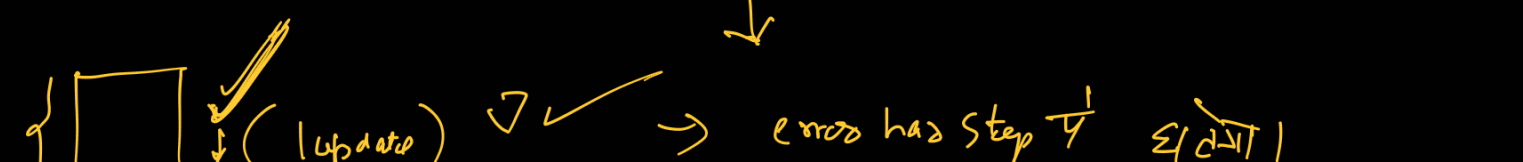
I have my 1 single update after seeing entire data.

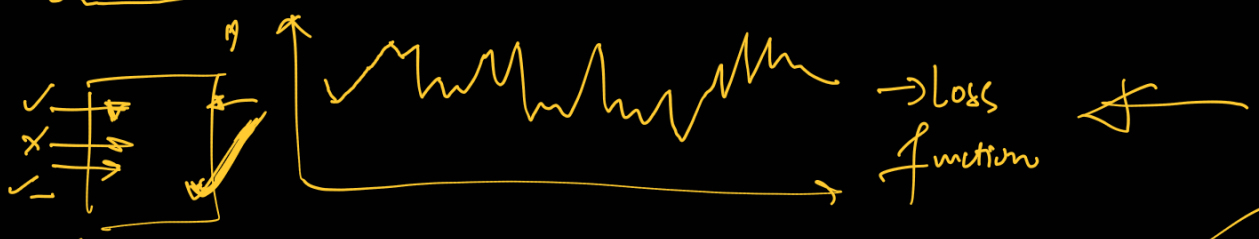
BATCH Gradient Descent

I can guarantee that my Cost will converge (reduce to acceptable criteria) with each passing iteration.

100 data points → 1000 → 1000x10 → .. 10^6, 10^7 millions, billions

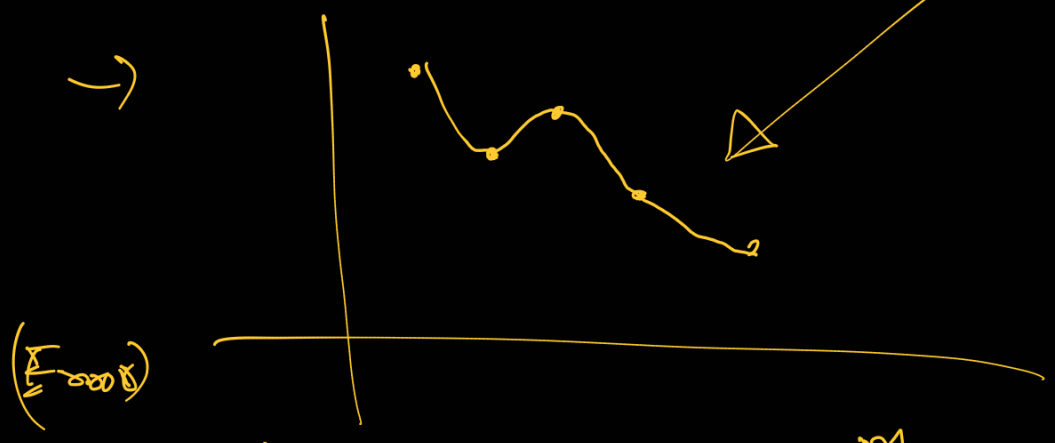
→ Other extreme → 1-1 data point





"1-1 example"

I am plotting magnitude of loss values (Loss Curve)



(Epochs)

If 1 → 0.9
2 → 0.7
3 → 0.8
4 → 0.6
5 → 0.5

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(y^{(i)} - \theta^T x^{(i)} \right)^2$$

we are plotting this → (Value of $J(\theta)$ at $(x^{(i)}, y^{(i)})$)

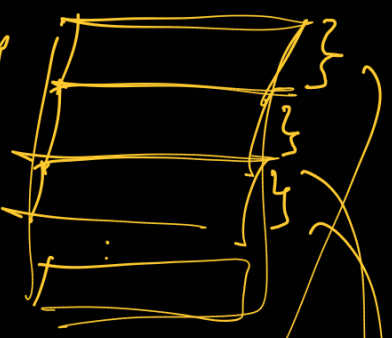
Mini-Batch Gradient Descent

b = 10

$$J^{(b)}(\theta) = \frac{1}{2b} \sum_{i=1}^{b} \left(y^{(i)} - h_\theta(x^{(i)}) \right)^2$$

(1000)

b is my batch-size



$$\theta_j^{(t+1)} = \theta_j^{(t)} - \eta \cdot \frac{1}{b} \sum_{i=1}^{b} \left(y^{(i)} - h_\theta(x^{(i)}) \right) q_j^{(i)}$$

Update

100/10 = 10 columns

m examples

b — batches

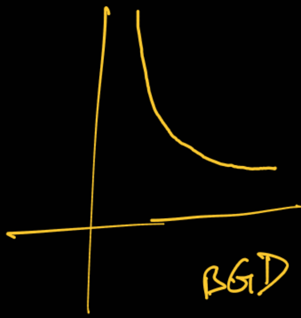
no of batch: $\lceil \frac{m}{b} \rceil$

10 10 .. 10 7 = 97

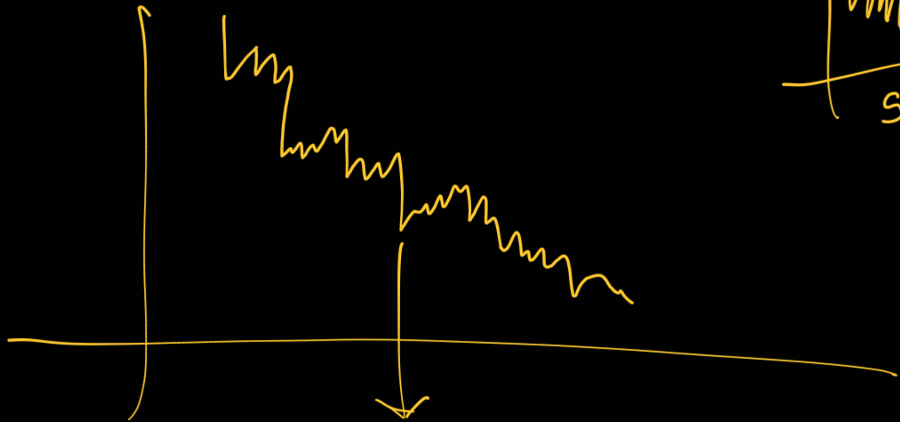
bs = 10

m = 100

SG



BGD



This I can proove Mathematically.

In expectation, SGD (MBGD) would converge in the long run to BGD.

(IIT-D) → exam question !!

$E[J^{(b)}(\theta)]$ → (Potential Exam Question!)

GD

- BGD
- * MBGD
- SGD (1-1) X

$\Rightarrow$ (SGD) ✓

Analytical Solution to Linear Regression

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)}) \right)^2, \quad h_\theta(x^{(i)}) = \theta^T x^{(i)}$$

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ \downarrow & \downarrow & & \downarrow \\ & & & \end{bmatrix}_{m \times n} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}_{(n+1)\times 1} \qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}_{(m \times 1)}$$

feature matrix.

$$+ \quad X = \begin{bmatrix} x_0 & x_1 & \cdots & x_n \\ \downarrow & \downarrow & & \downarrow \\ 1 & & & \\ \vdots & & & \\ 1 & & & \end{bmatrix}_{m \times (n+1)}$$

X θ Y

$(m \times (n+1))$ $((n+1)\times 1)$ $(m \times 1)$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)}) \right)^2$$

$$J(\theta) = \frac{1}{2m} \| Y - X\theta \|_2^2 \quad \text{(L2 norm)} \qquad \|x\|_2 \quad x^T x$$

$$= \frac{1}{2m} (Y - X\theta)^T (Y - X\theta)$$

$(m \times 1)$ $(m \times (n+1))$ $((n+1)\times 1)$

$(m \times 1) - (m \times 1)$

$((m \times 1))^T$

$(1 \times m) \quad (m \times 1)$

$= 1 \Rightarrow$ scalar

$$J(\theta) = \frac{1}{2m} (Y - X\theta)^T (Y - X\theta) \qquad (AB)^T = B^T A^T$$

$$J(\theta) = \frac{1}{2m} \left[(Y^T - \theta^T X^T)(Y - X\theta) \right]$$

$$J(\theta) = \frac{1}{2m} \left[Y^T (Y - X\theta) - \theta^T X^T (Y - X\theta) \right]$$

$$= \frac{1}{2m} \left[Y^T Y - Y^T X\theta - \theta^T X^T Y + \theta^T X^T X\theta \right]$$

$$= \frac{1}{2m} \left[Y^T Y - Y^T X\theta - (Y^T X\theta)^T + \theta^T X^T X\theta \right]$$

$(1 \times m)(m \times (n+1))((n+1) \times 1)$

(1×1) $\quad (1 \times 1)^T = 1 \times 1$

$=$

$\nabla_\theta (\theta^T A \theta) = 2A\theta$

$\nabla_\theta (a^T \theta) = a$

$\nabla_\theta (a\theta^T) = a$

$$J(\theta) = \frac{1}{2m} \left[Y^T Y - 2Y^T X\theta + \theta^T X^T X\theta \right]$$

$$J(\theta) = \frac{1}{2m} \left[Y^T Y - 2(X^T Y)\theta + \theta^T (X^T X)\theta \right]$$

$$\nabla_\theta J(\theta) = \frac{1}{2m} \left[0 - 2X^T Y + 2X^T X\theta \right] = 0$$

$X^T X \theta = X^T y$

$\theta = (X^T X)^{-1} X^T y$

We are not sure whether this inv exists.

↳ pinv

↳ Normal Equations Method!!

Gradient Desc.

Based Methods.

N. Eqns

Inv calculation is exp. consuming.

$O(n^2 m + n^3)$

$n = 1000$

$m = 1 \text{ Million}$

$\vec{v} = 2\hat{i} + 3\hat{j} + 4\hat{k}$

$|\vec{v}| = \left(\sqrt{2^2 + 3^2 + 4^2}\right)$ = L2 norm

$|2| + |3| + |4|$: L1

p-norm. $(v_1 \cdots v_n)$

$(v_1^p + v_2^p + \cdots v_n^p)^{1/p}$ → p norm

↳

$$\text{Error}_{(1)} = (y_p - y_q)^2$$

$$\text{Error}_{(2)} = (y_p - y_q)$$

$e_1 = e_2$

↓

Sum

When we take sum then 2 errors can cancel each other.

$$\sum (y_p - y_q) \cdot x$$

$$1 + 1 + 1 + 1 \rightarrow$$ L1 norm

Magnitude Cancellation ✗

Diff ✓

Squared error

MCX

Diff ✓