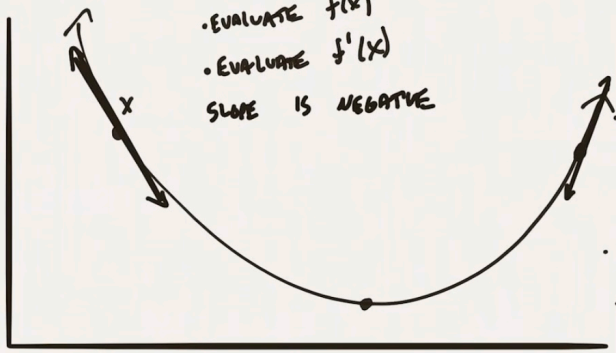
GRADIENT DESCENT

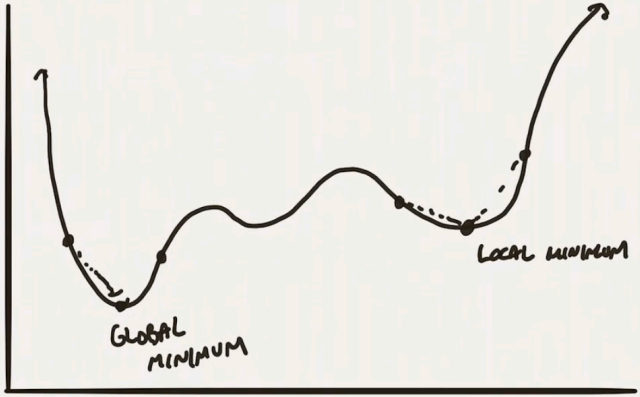
GOAL: FIND MINIMUM OF THIS FUNCTION



- EVALUATE $f(x)$
- EVALUATE $f'(x)$

SLOPE IS NEGATIVE

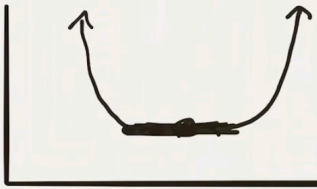
- IF $f'(x) < 0$ MOVE A BIT TO RIGHT
- IF $f'(x) > 0$ MOVE A BIT TO THE LEFT
- IF $f'(x) \approx 0$ STOP OUTPUT X.



$f'(x) > 0$ MOVE LEFT

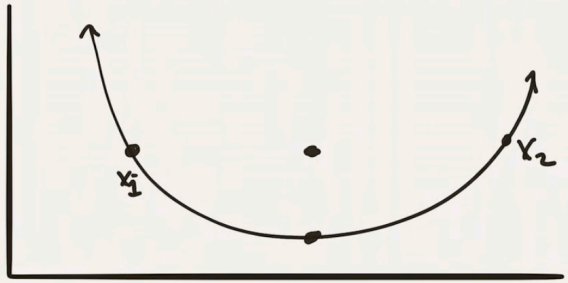
$f'(x) < 0$ MOVE RIGHT

O.W. STOP.



(CONVEX ?)

A FUNCTION IS CONVEX IF THE CHORD CONNECTING ANY 2 POINTS OF THE GRAPH LIES ABOVE THE FUNCTION.



$$f\left(\tfrac{1}{2}x_1 + \tfrac{1}{2}x_2\right) \leq \tfrac{1}{2}f(x_1) + \tfrac{1}{2}f(x_2)$$

EQUIVALENT DEFINITION OF CONVEXITY: A FUNCTION IS CONVEX IF $f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha)f(x_2)$

Let's say x^* is global min
WE'RE CURRENTLY AT x

$$f(\alpha x + (1-\alpha)x^*) \leq \alpha f(x) + (1-\alpha) f(x^*)$$

THUS FAR: OUR IDEA HAS BEEN TO LOOK AT TANGENT LINES AND THIS IDEA WORKS FOR SAY LINEAR FUNCTIONS AND SIMPLE CONVEX FUNCTIONS.

EVEN IF WE WANT TO MINIMIZE MORE COMPLICATED FUNCTIONS, ASSUME THEY ARE "LOCALLY" LINEAR.

f AT FIXED POINT x

$$f(x+\epsilon) = f(x) + \epsilon \cdot f'(x) + \frac{\epsilon^2}{2!} f''(x) + \frac{\epsilon^3}{3!} f'''(x) + \cdots$$

EXPRESSION IN TERMS OF ϵ.

LINEAR FUNCTION OF ϵ

WHEN ϵ IS SMALL, THESE TERMS ARE NEGLIGIBLE.

TAYLOR'S THM ALSO HOLDS IN d DIMENSIONS

INSTEAD OF TAKING DERIVATIVES (UNIVARIATE CASE)

FOR HIGHER DIMENSIONS WE MUST LOOK AT GRADIENTS.

THE GRADIENT OF f AT POINT x

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_d}(x) \right)$$

↙ d-dimensional vector.

$$f = w^T x + b \qquad \frac{\partial f}{\partial x_i} = w_i \qquad \nabla f = w$$

$f(x) = x^T A x - b^T x$ $\quad (n - \text{dimensions})$

$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j - \sum_{i=1}^{n} b_i x_i$

(*)

$$\frac{\partial f}{\partial x_k} = \sum_{j=1}^{n} a_{kj} x_j + \sum_{i=1}^{n} a_{ik} x_i - b_k$$

a_{kj}: k^{th} ROW OF A — INNER WITH x

a_{ik}: k^{TH} COL OF A — INNER PROD WITH x

CASE 1: $i = K$ (WHEN $i=K$ AND $j=K$ IN (*)) $a_{KK} x_K^2$ $\quad \frac{\partial (a_{KK} x_K^2)}{\partial x_K} = 2 a_{KK} x_K$

CASE 2: $i \neq K$ WE WONT HAVE $a_{KK} x_K$ TERM BECAUSE $i \neq K$

ANSWER $\boxed{Ax + A^T x - b}$

← GRADIENT AT PUNT X

IF A IS SYMMETRIC

$2Ax - b$

DEFINE GRADIENT DESCENT

INITIALLY WE'LL CHOOSE w RANDOMLY

(WANT TO MINIMIZE $f(w)$)

IF $\|\nabla f(w)\|_2 < \epsilon$ STOP OUTPUT w

OTHERWISE $w_{NEW} = w_{OLD} - \eta \nabla f(w)$

↳ STEP-SIZE PARAMETER

$$w_j^{NEW} = w_j^{OLD} - \eta \frac{\partial f}{\partial w_j}(w)$$

IS USUALLY SET TO BE RELATIVELY SMALL.

APPLY GRADIENT DESCENT TO LINEAR REGRESSION

$h(x) = w^T x + b$ (SEARCHING FOR THIS FUNCTION)

(WE HAVE A TRAINING SET OF SIZE m)

$$M.S.E.(w) = \frac{1}{m} \sum_{j=1}^{m} \underbrace{\left(w^T x^j + b - y^j \right)}_{g_j}^2$$

$$\frac{\partial g_j}{\partial w_i}$$

$$\frac{d(h_1 + h_2)}{dz} = \frac{dh_1}{dz} + \frac{dh_2}{dz}$$

APPLY GRADIENT DESCENT TO LINEAR REGRESSION

$h(x) = w^T x + b$ (SEARCHING FOR THIS FUNCTION)

(WE HAVE A TRAINING SET OF SIZE m)

$$M.S.E.(w) = \frac{1}{m} \sum_{j=1}^{m} \underbrace{\left(w^T x^j + b - y^j \right)^2}_{g_j}$$

M.S.E.(w) IS A CONVEX FUNCTION.

$$\frac{\partial g_j}{\partial w_i} = 2 \cdot \left(w^T x^j + b - y^j \right) x_i^j$$

RUNNING TIME $O(m \cdot n)$

$$\nabla g_j(w) = 2 \left(w^T x^j + b - y^j \right) x^j$$

$$\nabla M.S.E.(w) = \frac{2}{m} \cdot \sum_{j=1}^{m} \left(w^T x^j + b - y^j \right) \cdot x^j$$

STOCHASTIC GRADIENT DESCENT

- PREVIOUSLY IN LINEAR REGRESSION EXAMPLE WE SUMMED OVER ALL POINTS IN TRAINING SET.

- CHOOSE AN INDEX j AT RANDOM; COMPUTE GRADIENT W.R.T. THIS POINT <u>ONLY</u>

$$W_{NEW} = W_{OLD} - 2\cdot\eta\,(w^T x^j + b - y^j)x^j$$

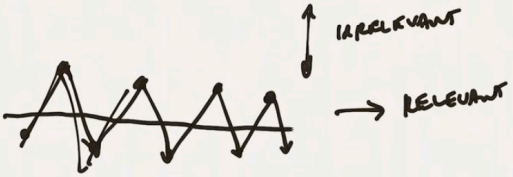
$$\mathbb{E}[W_{NEW}] = W_{OLD} - 2\eta\cdot\frac{1}{m}\sum_{j=1}^{m}(w^T x^j + b - y^j)x^j$$

- USE "BATCHES" TO INTERPOLATE BTWN GRADIENT DESCENT AND PURE S.G.D. (SGD)

- HOW TO CHOOSE η THE STEP-SIZE

- MORE ART THAN SCIENCE; USE CROSS VALIDATION TO PICK η
- MANY TECHNIQUES FOR ADAPTIVELY CHOOSING η

- MOMENTUM



MOMENTUM HAS A "VELOCITY" VARIABLE V

$$V_0 = 0 \qquad V_1 = -\eta g_i$$

$$V_i = \alpha \cdot V_{i-1} - \eta g_i$$

THIS TAKES A WEIGHTED MOVING AVERAGE OF $-\eta g_i$'s

EXPONENTIAL " "

$$W_{NEW} = W_{OLD} + V_i$$

ACCELERATED GRADIENT DESCENT