Optimization

Mahdi Roozbahani
Georgia Tech

Outline

Motivation

Entropy

Conditional Entropy and Mutual Information

Cross-Entropy and KL-Divergence

Let's work on this subject in our Optimization lecture

Cross Entropy

Cross Entropy: The expected number of bits when a wrong distribution Q is assumed while the data actually follows a distribution P

Actual pd predicted pd

$$H(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x) = H(P) + KL[P][Q]$$

This is because:

$$H(p, q) = \mathrm{E}_p[l_i] = \mathrm{E}_p\left[\log \frac{1}{q(x_i)}\right]$$

$$H(p, q) = \sum_{x_i} p(x_i) \log \frac{1}{q(x_i)}$$

$$H(p, q) = -\sum_{x} p(x) \log q(x).$$

Labeling target values
Label encoding (ordinal) and One-hot encoding

Multi class classification

$$X = \begin{bmatrix} w & h & age \; \cdots \\ & & \\ & & \end{bmatrix}_{n\times d}$$

$$Y_a = \begin{bmatrix} \text{cat} \\ \text{fish} \\ \text{dog} \\ \text{fish} \\ \vdots \end{bmatrix}_{n\times 1} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \\ \vdots \end{bmatrix}$$

$$\hat{Y}_p = \begin{bmatrix} 0 \\ 1 \\ 1.5 \\ \vdots \end{bmatrix}_{n\times 1}$$

$$Y_a = \begin{array}{r} \text{cat} \\ \text{fish} \\ \\ \text{dog} \\ \\ \end{array} \begin{bmatrix} [1 & 0 & 0] \\ [0 & 1 & 0] \\ [0 & 0 & 1] \\ [0 & 1 & 0] \\ \vdots & & \end{bmatrix}$$

$$\hat{Y}_p = \begin{bmatrix} [0.8 & 0.1 & 0.1] \\ [0.2 & 0.7 & 0.1] \\ \vdots & & \end{bmatrix}$$

$\boxed{\text{Softmax}}$

Min

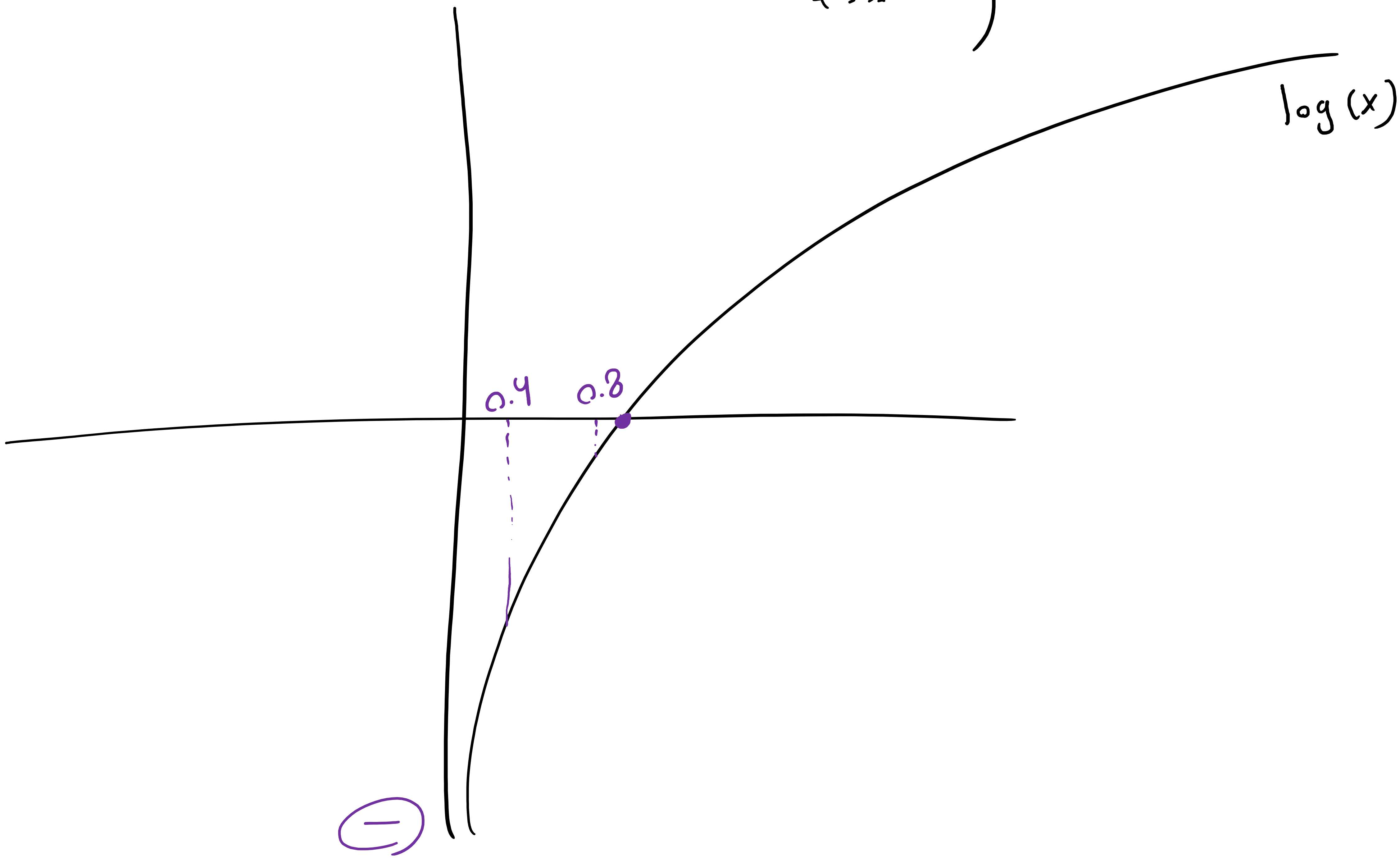
$$\|Y_a - Y_p\|_2^2 = \left[(1-0.8)^2 + (0-0.1)^2 + (0-0.1)^2\right] + \left[(0-0.2)^2 + (1-0.7)^2 + (0-0.1)^2\right] + \cdots$$

Max

$$\sum Y_a \cdot \hat{Y}_p = [1 \;\; 0 \;\; 0] \begin{bmatrix} 0.8 \\ 0.1 \\ 0.1 \end{bmatrix} + [0 \;\; 1 \;\; 0] \begin{bmatrix} 0.2 \\ 0.7 \\ 0.1 \end{bmatrix} + \cdots$$

Why Cross entropy and not simply use dot product?

$$H(p,q) = -\sum p(x) \log q(x) = -\left(\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \log 0.8 \\ \log 0.1 \\ \log 0.1 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \log 0.2 \\ \log 0.7 \\ \log 0.1 \end{bmatrix} + \dots \right)$$



Kullback-Leibler Divergence

Another useful information theoretic quantity measures the difference between two distributions.

$$\mathbf{KL}[P(S)\|Q(S)] = \sum_s P(s)\log\frac{P(s)}{Q(s)}$$

$$= \underbrace{\sum_s P(s)\log\frac{1}{Q(s)}}_{\text{cross entropy}} - \mathbf{H}[P] = H(P,Q) - H(P)$$

Excess cost in bits paid by encoding according to Q instead of P.

KL Divergence is a **KIND OF** distance measurement

$$-\mathbf{KL}[P\|Q] = \sum_s P(s)\log\frac{Q(s)}{P(s)}$$

log function is concave or convex?

$$\sum_s P(s)\log\frac{Q(s)}{P(s)} \le \log\sum_s P(s)\frac{Q(s)}{P(s)}$$

By Jensen Inequality

$$= \log\sum_s Q(s) = \log 1 = 0$$

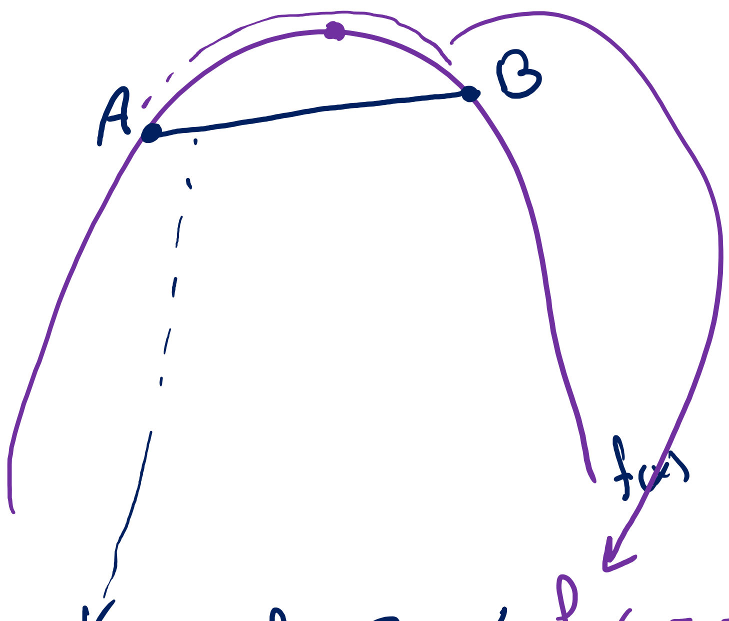
So $\mathbf{KL}[P\|Q] \ge 0$. Equality iff $P = Q$

When $P = Q,\ KL[P||Q] = 0$

Concave

$$f(x) = -x^2$$

$$f''(x) = -2$$

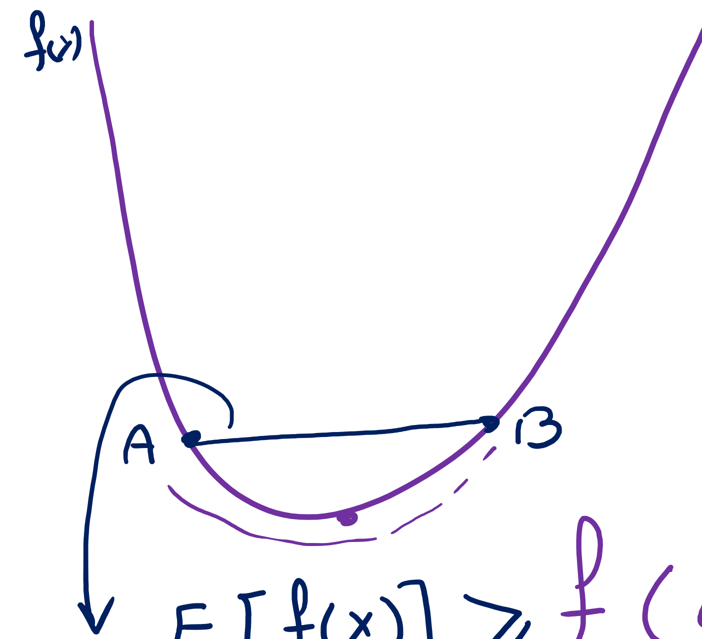


$$E[f(x)] \leq f(E[x])$$

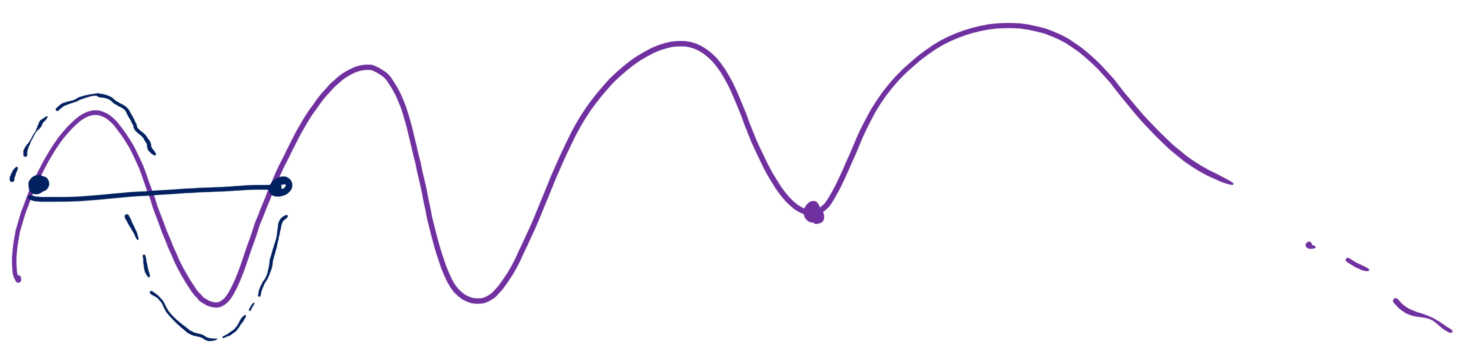
Convex

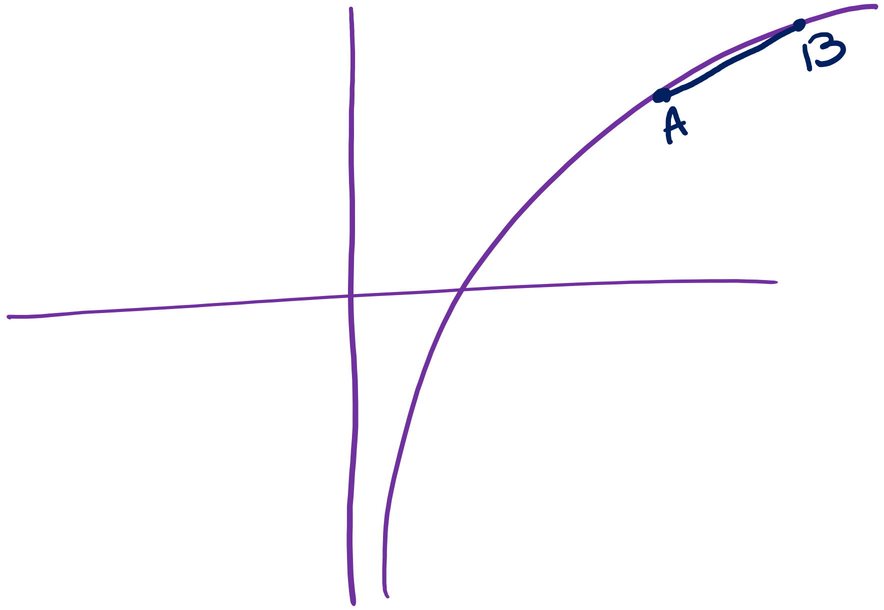
$$f(x) = x^2$$

$$f''(x) = 2$$



$$E[f(x)] \geq f(E[x])$$





$\log x = f(x)$

$$E[f(x)] \leq f(E[x])$$

$$E[\log(x)] \leq \log(E[x])$$

$$\frac{Q(s)}{P(s)} = g(s)$$

$$-KL[P][Q] = \sum P(s) \log \frac{Q(s)}{P(s)} = \sum \underbrace{P(s)} \underbrace{\log g(s)}$$

$$-KL[P][Q] = E[\log g(s)] \leq \log E[g(s)]$$

$$\text{〃} \quad \leq \log \sum P(s)\, g(s)$$

$$\text{〃} \quad \leq \log \sum \cancel{P(s)} \frac{Q(s)}{\cancel{P(s)}}$$

$$\text{〃} \quad \leq \log\left(\sum Q(s)\right)$$

$$-KL[P][Q] \leq \log 1$$

$$KL[P][Q] \geq 0$$

Optimization

- Probabilistic models $\longrightarrow N(X \mid \overset{a,b}{\mu, \Sigma}) \rightsquigarrow$ MLE
  - $\downarrow$
  - $l \rightsquigarrow \dfrac{\partial l}{\partial \cdot} = 0$

- Non-probabilistic models
  - No-constraint
    - $f(x,y) = x^2 + y^2$
    - $\dfrac{\partial f(x,y)}{\partial x} = 0$
    - $\dfrac{\partial f(x,y)}{\partial y} = 0$
  - Equality constraint
    - Objective function
    - $f(x,y) = x^2 + y^2$
    - s.t. $x - y = 8 \rightarrow g(x,y) = x - y - 8$
    - constraint function
    - $\downarrow$
    - lagrange function
    - $\dfrac{\partial L}{\partial \cdot} = 0$
  - Inequality constraint
    - $f(x,y) = x^2 + y^2$
    - s.t. $x - y \le 8$
    - $\rightsquigarrow$
    - $\dfrac{\partial L}{\partial \cdot} = 0$
    - & Satisfy KKT conditions

$$f(M,S) = 6M^2 + 3S^2$$

$$f'' = \begin{bmatrix} \frac{f^2}{\partial^2 M} & \frac{f^2}{\partial M \partial S} \\ \frac{f^2}{\partial S \partial M} & \frac{f^2}{\partial^2 S} \end{bmatrix}$$

M: # hours study ML

S: # " sleep

$$\frac{\partial f(M,S)}{\partial M} = 12M = 0 \Rightarrow M = 0$$

$$\frac{\partial f(M,S)}{\partial S} = 6S = 0 \Rightarrow S = 0$$

objective function

$$f(m,s) = 6m^2 + 3s^2$$

s.t. $m+s=24$

s.t. $m-s=10$

$\Rightarrow$

constraint function

$$g(m,s) = m+s-24=0$$

$$h(m,s) = m-s-10$$

$\boxed{\delta, \alpha, \beta}$

lagrange mulitiplier

$$L(m,s,\delta_1,\delta_2) = f(m,s) - \delta_1 g(m,s) - \delta_2 h(m,s)$$

$$f(m,s) = 6m^2+3s^2$$

s.t. $m+s=24$

$$L(m,s,\delta) = 6m^2+3s^2 - \delta(m+s-24)$$

$$\frac{\partial L}{\partial m}=0 \Rightarrow 12m-\delta=0 \Rightarrow m=\frac{\delta}{12}=\frac{96}{12}=8$$

$$m+s=24$$

$$8+16=24$$

$$\frac{\partial L}{\partial s}=0 \Rightarrow 6s-\delta=0 \Rightarrow s=\frac{\delta}{6}=\frac{96}{6}=16$$

$$\frac{\partial L}{\partial \delta}=0 \Rightarrow -(m+s-24)=0 \Rightarrow m+s=24 \Rightarrow \frac{\delta}{12}+\frac{\delta}{6}=24 \Rightarrow \delta=96$$

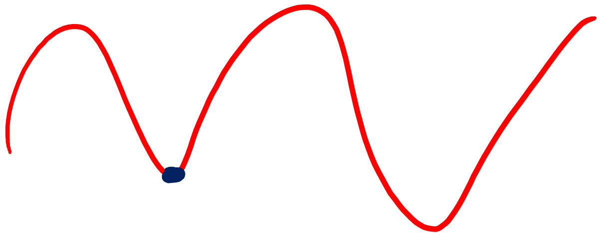
$$L(m,s,\delta) = f(m,s) - \delta\, g(m,s)$$

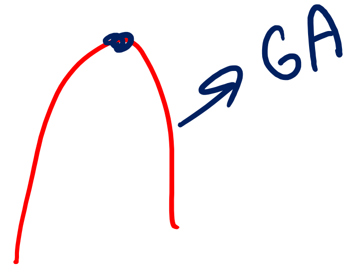
$$\nabla f(m,s) \sim \nabla g(m,s)$$
$$\Downarrow$$
$$\nabla f(m,s) = \delta \nabla g(m,s)$$

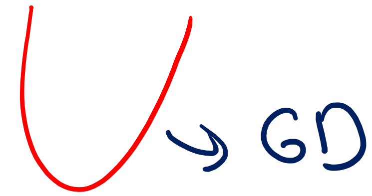
$$\nabla_{m,s,\delta} L(m,s,\delta) = 0$$

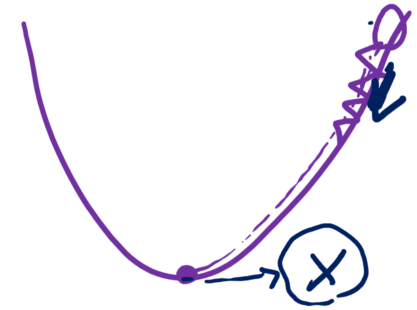
$$\nabla\left(f(m,s) - \delta\, g(m,s) \right) = 0$$

$$\nabla f(m,s) - \cancel{\delta} \nabla g(m,s) = 0$$









$$f(x) = x^2$$

$$\frac{\partial f(x)}{\partial x} = 2x$$

learning step = 0.002

$$x^{\{t+1\}} \leftarrow x^{\{t\}} - \alpha \frac{\partial f(x)}{\partial x} \rightarrow GD$$

$$x^{\{t+1\}} \leftarrow x^{\{t\}} - \alpha\, 2x^{\{t\}}$$

$$L(m, s, \delta) = f(m, s) - \delta\, g(m, s) \qquad m, s, \delta$$

$$m^{\{t+1\}} \leftarrow m^{\{t\}} - \alpha \frac{\partial L}{\partial m}$$

$$s^{\{t+1\}} \leftarrow s^{\{t\}} - \alpha \frac{\partial L}{\partial s}$$

$$\delta^{\{t+1\}} \leftarrow \delta^{\{t\}} - \alpha \frac{\partial L}{\partial \delta}$$

Min $f(M,S) = 6M^2 + 3S^2$

s.t. $M+S \leq 24 \Rightarrow \underline{M+S-24} \leq 0$

$g(M,S) \leq 0$

KKT conditions

1. Stationary condition $\Rightarrow$ Min $L = \underbrace{6M^2 + 3S^2}_{\text{Min}} + \delta(M+S-24)$

2. Primal feasibility $g(M,S) \leq 0$

3. Dual feasibility $\delta \geq 0$

4. Complementary slackness $g(M,S)\,\delta = 0$

$g(M,S) = 0 \rightarrow$ Active $\rightarrow \delta \geq 0$

$g(M,S) \neq 0 \rightarrow$ Inactive solution $\rightarrow \delta = 0$

Using Gradient Descent **as an alternative** to solve the equality constraint optimization example

This won't converge; do you know why?

```
import numpy as np

def minimize_gd():
    LEARNING_RATE = 0.01
    TOLERANCE = 1e-6
    # Initialize M, S, and lambda (make sure M + S = 24)
    M = 12.0
    S = 12.0
    lm = 0.0  # Lagrange Multiplier

    while True:
        # Calculate the gradients of the Lagrangian w.r.t. M, S, and lambda
        gradientM = 12 * M - lm
        gradientS = 6 * S - lm
        gradientLambda = - (M + S - 24)

        # Update M, S, and lambda
        newM = M - LEARNING_RATE * gradientM
        newS = S - LEARNING_RATE * gradientS
        newLambda = lm - LEARNING_RATE * gradientLambda

        # If the changes in M, S, and lambda are smaller than the tolerance, we break the loop
        if np.abs(newM - M) < TOLERANCE and np.abs(newS - S) < TOLERANCE and np.abs(newLambda - lm) < TOLERANCE:
            break

        M = newM
        S = newS
        lm = newLambda

    print("Minimum occurs at M = ", M , ", S = " , S , ", with lambda = ", lm)
    print("Minimum value of z = " , (6*M * M + 3*S * S))

minimize_gd()
```

Example 1:

https://www.geogebra.org/3d/srzmv8uh

Example 2:

https://www.geogebra.org/3d/syhkqpk7

1	0	-1
1	0	-1
1	0	-1

$$\mathrm{Sum}\left(\begin{bmatrix} 0 & 0 & -2 \\ 2 & 0 & -3 \\ 1 & 0 & -1 \end{bmatrix}\right) = |-3|$$

0	1	2	4	5	6	0	2
2	1	3	8	9	255	72	83
1	2	1	4	79	65	53	33
43	97	15	67	104	77	163	43
36	173	13	76	205	89	174	34
63	163	113	86	209	91	185	98
84	153	123	96	134	101	196	121
96	143	133	79	135	103	216	211

	3						

$[1 \quad 0 \quad -1 \quad 1 \quad 0 \quad -1 \quad 1 \quad 0 \quad -1]$

$[0 \quad 1 \quad 2 \quad 2 \quad 1 \quad 3 \quad 1 \quad 2 \quad 1]$