Georgia Tech

Optimization

Mahdi Roozbahani
Georgia Tech

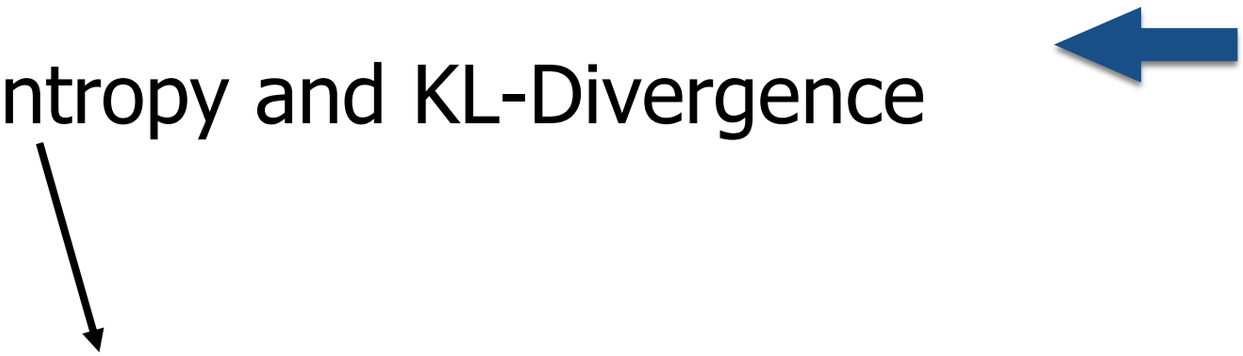
Outline

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

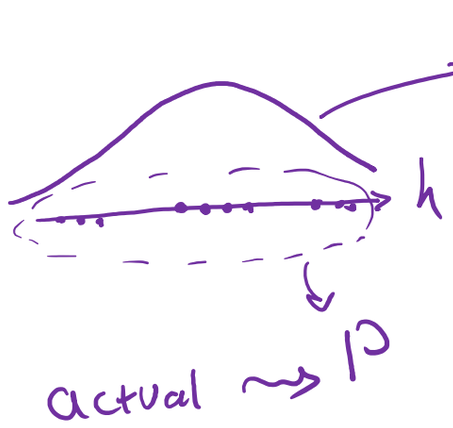
$$H(p,q) = -\sum_{x \in \mathcal{X}} p(x)\ \log q(x) = H(P) + KL[P][Q]$$

This is because:

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

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

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



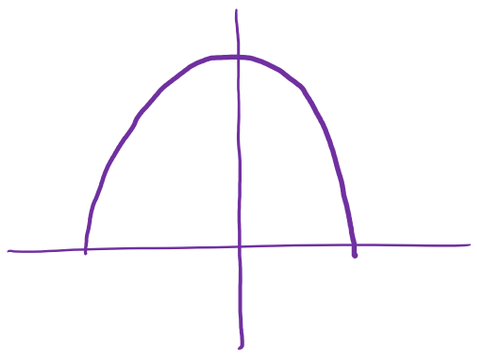
$$f(x^{[i]}|\theta) \rightsquigarrow \max \prod_{i=1}^{N} f(x^{[i]}|\theta) \Rightarrow \max l(\theta|x) = \sum_{i=1}^{N} \log f(x^{[i]}|\theta)$$

function

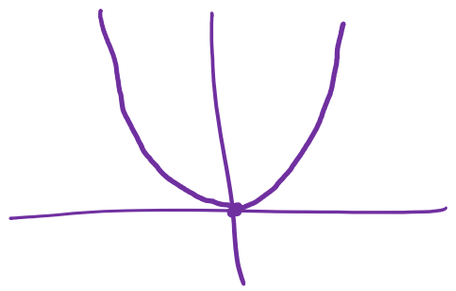
$\log L$

max

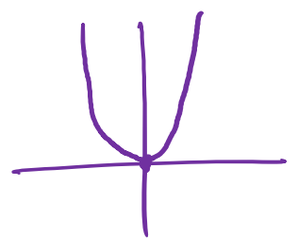
$f(x) = -x^2$



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



$f(x) = \frac{1}{6}x^2$



$$\left(\frac{1}{N}\sum(\cdot)\right) = E[\cdot] \rightsquigarrow \frac{1}{N}\sum g(x) = E[g(x)]$$

$$\min\ l(\theta|x) = \sum -\log f(x^{[i]}|\theta) \Rightarrow \left(\frac{1}{N}\sum\right) -\log f(x^{[i]}|\theta) = E[-\log f(x^{[i]}|\theta)]$$

$$= E[-\log Q]$$

$$E[-\log p + \log p - \log Q]$$

$$E[a+b+c] = E[a] + E[b+c]$$

$$E\underbrace{[-\log p]}_{g(x)} + E\underbrace{\left[\log \frac{p}{Q}\right]}_{g(x)} \rightarrow E[g(x)] = \sum p(x)\, g(x)$$

$$\frac{\sum p * (-\log p)}{\sum p \log \frac{1}{p}} + \frac{\sum p * \log \frac{p}{Q}}{-KL[P][Q]} = \sum p(-\log p) - \underbrace{\sum p \log \left(\frac{Q}{p}\right)}_{-KL[P][Q]}$$

$$H(p) + \left(-\sum p \log \frac{Q}{p}\right) = H(p) + KL[P][Q] = CE$$

CE = negative average log likelihood

$$E[g(x)] = \frac{1}{N}\sum (g(x)) = \sum p(x)\, g(x)$$

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

$$X = \begin{bmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \\ 4 & 8 \end{bmatrix}_{n \times d} \quad \text{height} = h, \quad \text{weight} = w$$

$$y_a = \begin{bmatrix} \text{cat} \\ \text{fish} \\ \text{dog} \\ \text{cat} \end{bmatrix}_{n \times 1} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \end{bmatrix} \xrightarrow{ML} \hat{y}_p = \begin{bmatrix} 0.1 \\ 0.5 \\ 2 \\ 0.2 \end{bmatrix}$$

cat = [1 0 0] fish = [0 1 0] dog = [0 0 1]

$$y_a = \begin{bmatrix} [1 & 0 & 0] \\ [0 & 1 & 0] \\ [0 & 0 & 1] \\ [1 & 0 & 0] \end{bmatrix} \rightsquigarrow ML \rightarrow \hat{y}_p = \begin{bmatrix} [23 & 2 & 1] \\ [72 & 120 & 23] \\ \vdots \end{bmatrix} \xrightarrow{\text{Soft max}} = \begin{bmatrix} [0.7 & 0.2 & 0.1] \\ [0.1 & 0.8 & 0.1] \\ \vdots \end{bmatrix}$$

$$Y_a * \hat{Y}_p = \begin{bmatrix} 0.7 \\ 0.8 \\ i \end{bmatrix} \Rightarrow \max \sum_{i=1}^{N} (Y_a * \hat{Y}_p)^{[i]} \quad \text{best result } 1$$

$$\min \|Y_a - \hat{Y}_p\|_2^2 = \sum_{i=1}^{N} \left[(1-0.7)^2 + (0-0.2)^2 + (0-0.1)^2 + \dots + \dots \right] \quad \text{best result is zero}$$

$$Y_a = \begin{bmatrix} [1 & 0 & 0] \\ [0 & 1 & 0] \\ [0 & 0 & 1] \\ [1 & 0 & 0] \end{bmatrix} \qquad \hat{Y}_p = \begin{bmatrix} [0.7 & 0.2 & 0.1] \\ [0.1 & 0.8 & 0.1] \\ [\quad & \quad & \quad] \\ [\quad & \quad & \quad] \end{bmatrix}$$

$$\text{Max} \sum_{i=1}^{N} \left(y_a^{\{i\}} \cdot \hat{y}_p^{\{i\}T} \right) = \Big(\underbrace{[1 \;\; 0 \;\; 0] \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix}}_{x^{\{1\}}} + \underbrace{[0 \;\; 1 \;\; 0] \begin{bmatrix} 0.1 \\ 0.8 \\ 0.1 \end{bmatrix}}_{x^{\{2\}}} + \cdots + x^{\{n\}} \Big)$$

$$\text{Min} \sum_{i=1}^{N} \| y_a^{\{i\}} - \hat{y}_p^{\{i\}} \|_2$$

Why Cross entropy and not simply use dot product?

Min

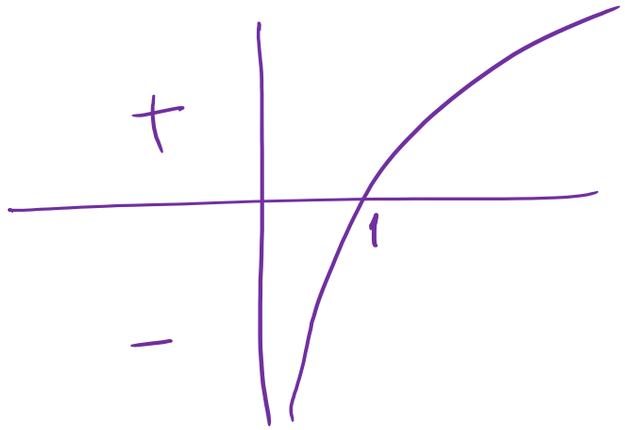
$$CE = -\sum p(x) \log q(x)$$

$$= -\sum y_a \log \hat{y}_p$$

$$p(x) = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

$$\log q(x) = \log(\hat{y}_p) =$$

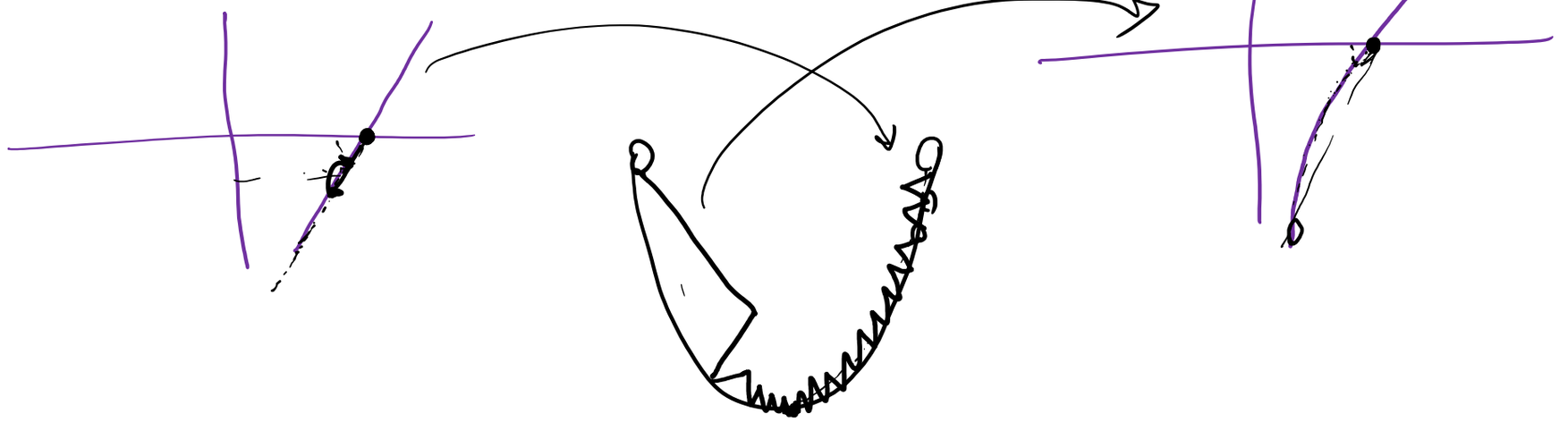
$$= \begin{bmatrix} \log .7 & \log .2 & \log .1 \\ \vdots & & \end{bmatrix}$$



$$CE = -([1\log .7 + 0\log .2 + 0\log .1 + \\ \dots + \dots)$$

$$CE = -\sum_{i=1}^{N} y_a^{\{i\}} \log \hat{y}_p^{\{i\}}$$

$$\text{Min} \; -\sum (y_a^{\{i\}} \hat{y}_p^{\{i\}T})$$



$$CE = -\sum y_a^{\{i\}} \log \hat{y}_p^{\{i\}}$$

Kullback-Leibler Divergence

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

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

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

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

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

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

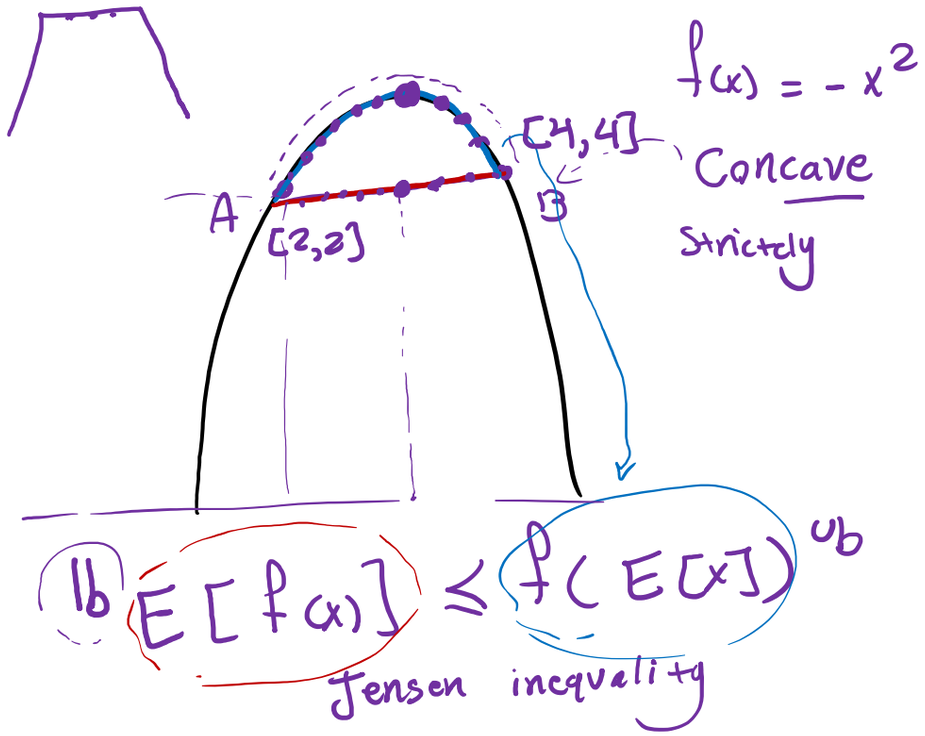
log function is concave or convex?

$$\sum_s P(s)\log\frac{Q(s)}{P(s)} \leq \log\sum_s P(s)\frac{Q(s)}{P(s)}$$ By Jensen Inequality

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

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

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



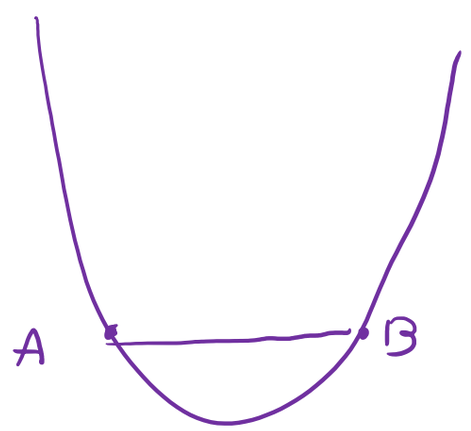
$f(x) = -x^2$

Concave

Strictly

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

Jensen inequality



Convex

$f(x) = x^2$

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

$E[f(x)] = \sum p(x) f(x)$

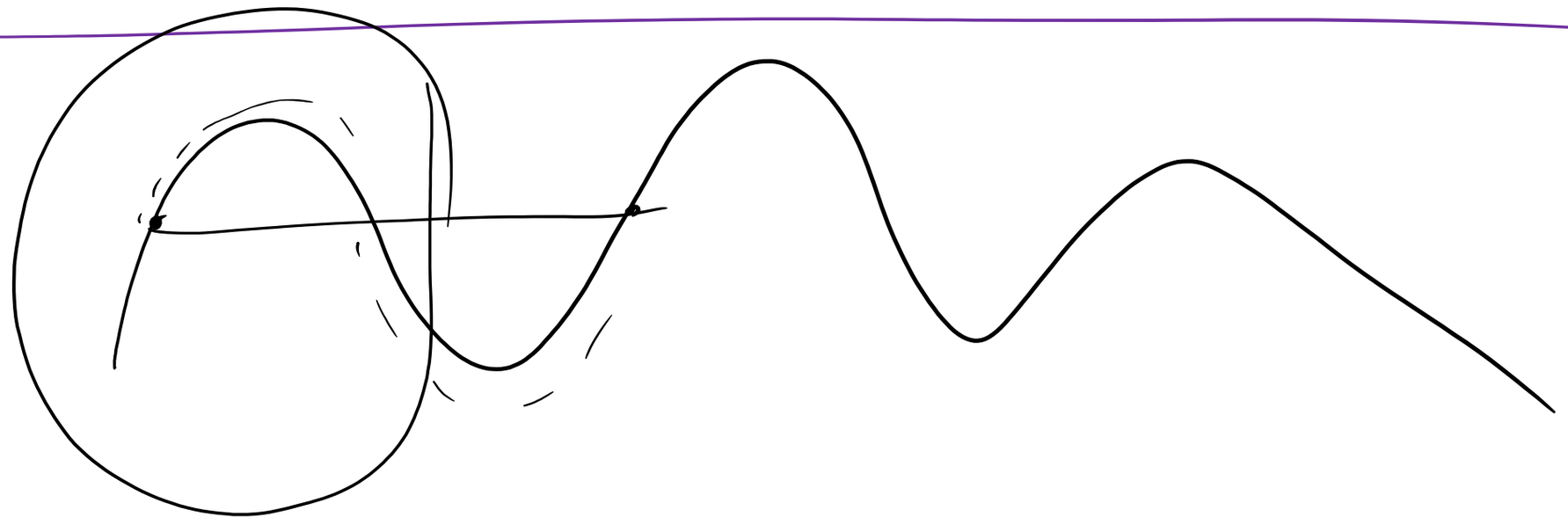
$= \frac{1}{2} f(x=A) + \frac{1}{2} f(x=b)$

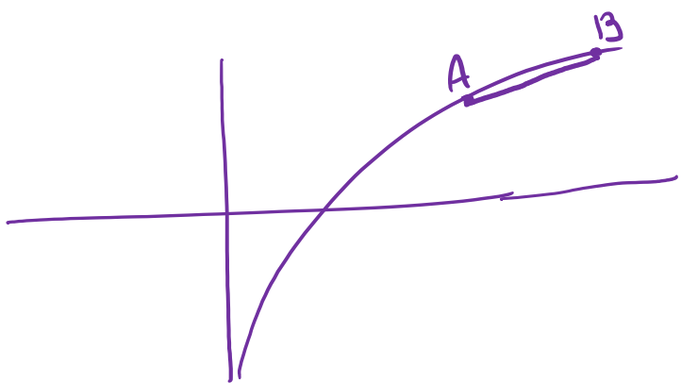
0.4 0.6

$f(E[x])$?

$E[x] = \sum p(x) x$

$\frac{1}{2} * 2 + \frac{1}{2} * 4 = 3$





Concave

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

$$f(x) = \log x$$

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

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

$$E[g(x)] = \sum p(x)\, g(x)$$

$$= \sum \cancel{P(x)} \frac{Q(x)}{\cancel{P(x)}}$$

$$\sum P \log g(x) = E[\log g(x)] \leq \log(E[g(x)])$$

$$\leq \log \sum Q(x) \quad (\sum Q(x) = 1)$$

$$\leq \log 1$$

$$-KL[P][Q] \leq 0$$

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

Objective function $f(x,y)$

s.t. $g(x,y) = 0$ Equality constraint ⇝ lagrange function

s.t. $h(x,y) > 2$ Inequality constraint ⇝ KKT conditions

↘

$$H = \begin{bmatrix} \frac{\partial^2 f(x,x)}{\partial x^2} & \frac{\partial^2 f(x,x)}{\partial x \partial y} \\ \frac{\partial^2 f(x)}{\partial y \partial x} & \frac{\partial^2 f(x,y)}{\partial y^2} \end{bmatrix}$$

① Linear Programming ⟶ $f(x,y) = x+y$

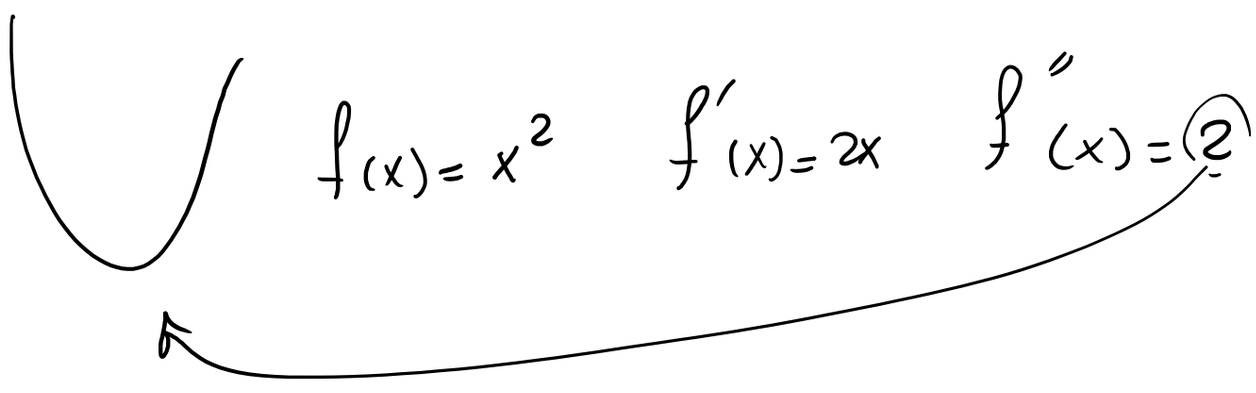
s.t. $x+y = 20$

② Quadratic Programming ⟶ $f(x) = x^2+y$

s.t. $x+y = 30$

③ Non linear programming ⇝ $f(x) = x^3+y$

s.t. $x^2+y = 20$



M = # hours you study ML / day

S = " " " sleep

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

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

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

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

λ, α, β → lagrange multiplier

s.t. $M + S = 24 \Rightarrow g(M,S) = M + S - 24$ constraint function

Objective

$$L(M,S,\lambda) = f(M,S) - \lambda\, g(M,S) = 6M^2 + 3S^2 - \lambda\,(M+S-24)$$

$$\frac{\partial L}{\partial \lambda} = 0 \Rightarrow -(M+S-24) = 0 \Rightarrow M + S = 24$$

$$\frac{\partial L}{\partial M} = 0 \Rightarrow 12M + 0 - \lambda = 0 \Rightarrow M = \frac{\lambda}{12} = \frac{96}{12} = 8$$

$$\frac{\partial L}{\partial S} = 0 \Rightarrow 6S - \lambda = 0 \Rightarrow S = \frac{\lambda}{6} = \frac{96}{6} = 16$$

$$\frac{\lambda}{12} + \frac{\lambda}{6} = 24 \Rightarrow \lambda = 96$$

$M + S = 24$

$8 + 16$

$$\min f(M,S)$$

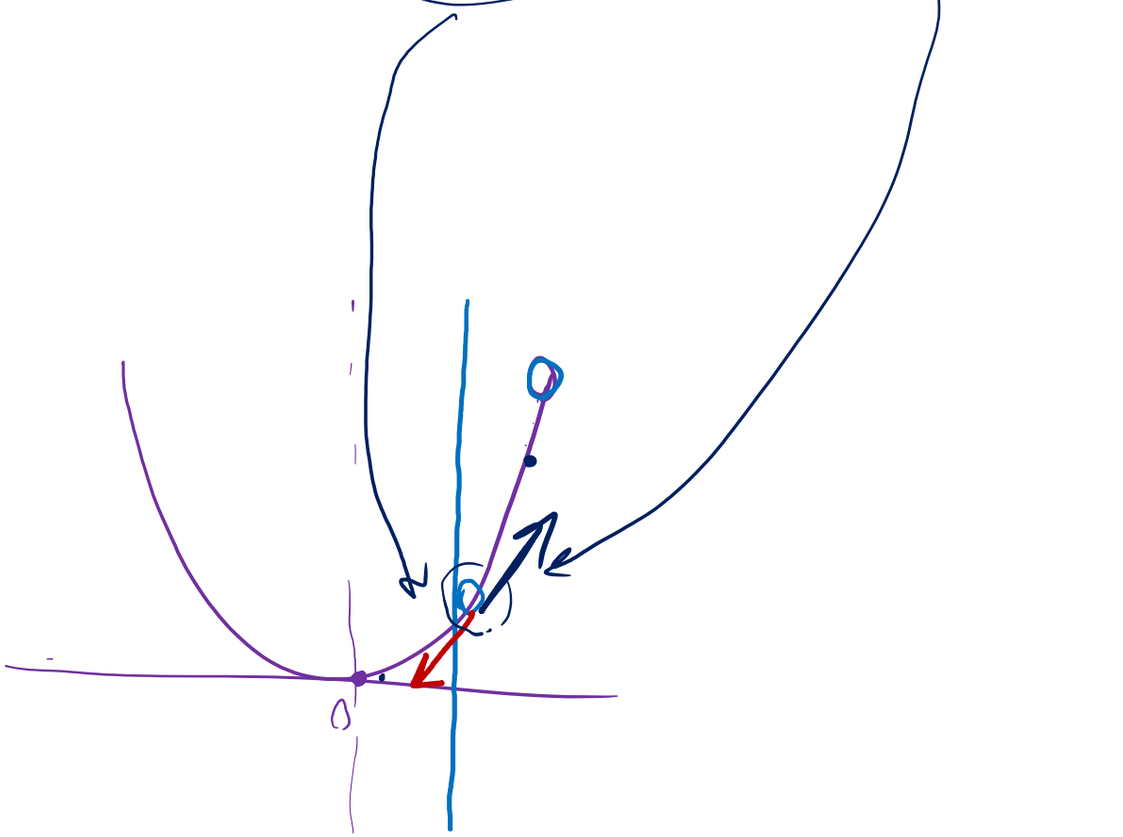
s.t. $M+S=20 \rightsquigarrow g(M,S)=M+S-20 \rightsquigarrow \lambda_1$

s.t. $M-S=10 \rightsquigarrow h(M,S)=M-S-10 \rightsquigarrow \lambda_2$

$$L(M,S,\lambda_1,\lambda_2) = f(M,S) - \lambda_1 g(M,S) - \lambda_2 h(M,S)$$

$$L(M,S,\lambda) = f(M,S) - \lambda g(M,S)$$

$$\nabla L = \nabla f(M,S) - \lambda \nabla g(M,S) = 0 \Rightarrow \nabla f(M,S) = \lambda \nabla g(M,S)$$



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

$M + S \leq 24$ ⇝ kkT condition ⇝ $g(M,S) = M + S - 24 \leq 0$ ⇝

- $g(M,S) = 0$ Active Solution
- $g(M,S) < 0$ Inactive Solution

$L(M,S,\lambda) = f(M,S) + \lambda g(M,S)$ (+, −)

kkT

① Stationary condition $\frac{\partial L}{\partial \cdot} = 0$

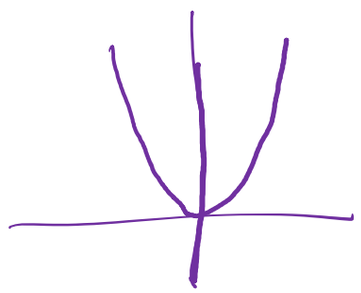
② primal feasibility $g(M,S) \leq 0$

③ Dual feasibility $\lambda \geq 0$

④ complementary slackness

$\lambda g(M,S) = 0 \Longrightarrow$

- $g(M,S) = 0 \rightsquigarrow \lambda \geq 0$
- $\lambda = 0 \rightsquigarrow g(M,S) \neq 0$



Min $f(M,S) = \frac{1}{2}M^2 + \frac{1}{2}S^2$ ⇝ Convex ⇝ minimize

Primal form

S.t. $M+S = 24 \Rightarrow g(M,S) = M+S-24$

x^2 ∪

$L(M,S,\delta) = \frac{1}{2}M^2 + \frac{1}{2}S^2 - \delta(M+S-24)$

$\frac{\partial L}{\partial M} = 0 \Rightarrow M - \delta = 0 \Rightarrow M = \delta$ $\quad$ $\frac{\partial L}{\partial S} = 0 \Rightarrow S = \delta$

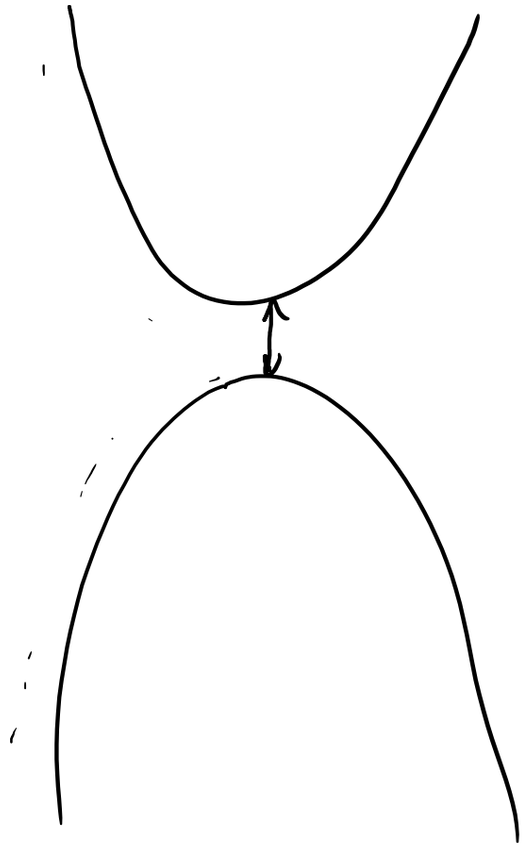
$L(\delta) = \frac{\delta^2}{2} + \frac{\delta^2}{2} - \delta(2\delta - 24) = \delta^2 - 2\delta^2 + 24\delta$

$= -\delta^2 + 24\delta$

Dual form

$-x^2$ ∩

Concave

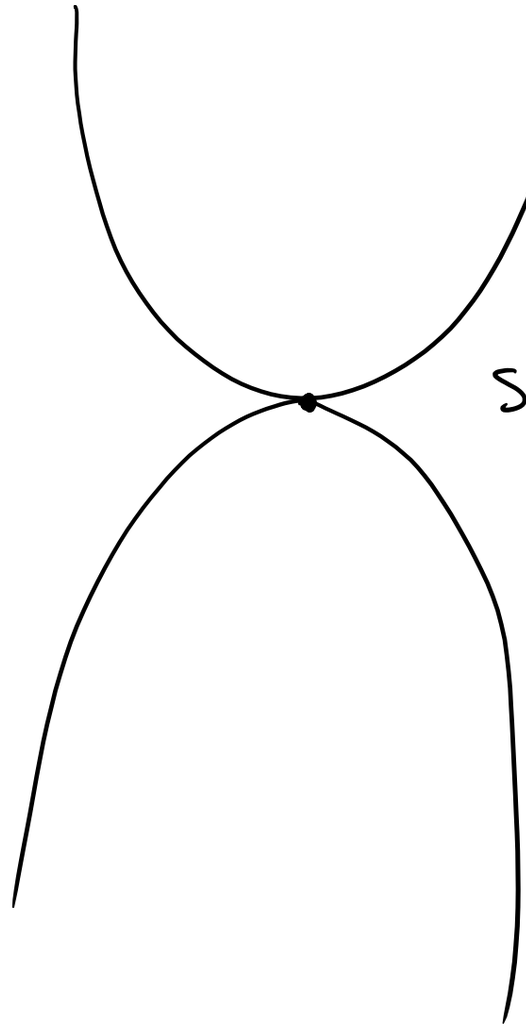


convex

primal

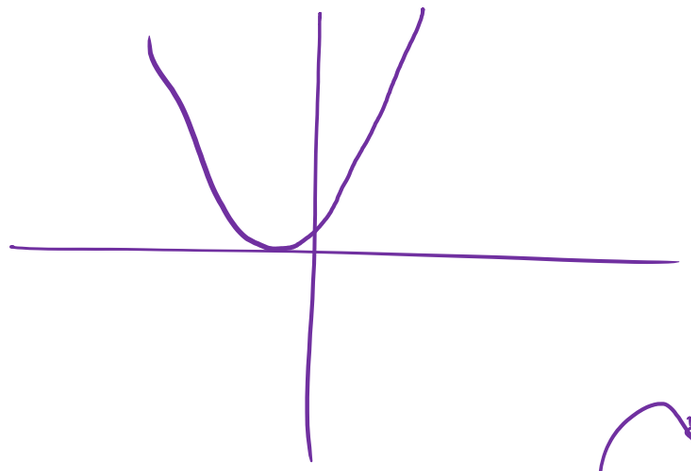
weak duality

Dual form



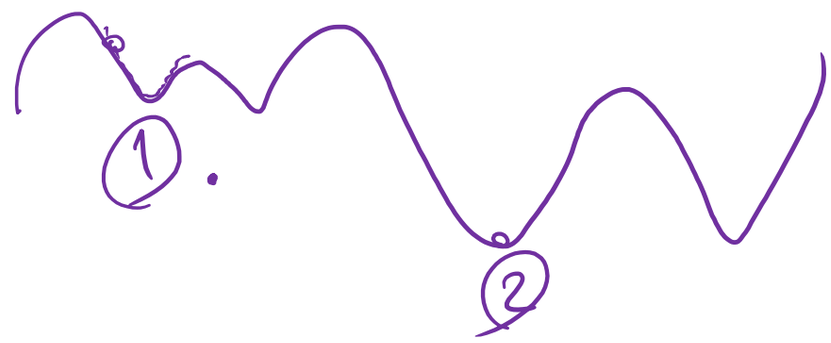
Strong duality

$$f(x) = \exp(x) + x^2$$

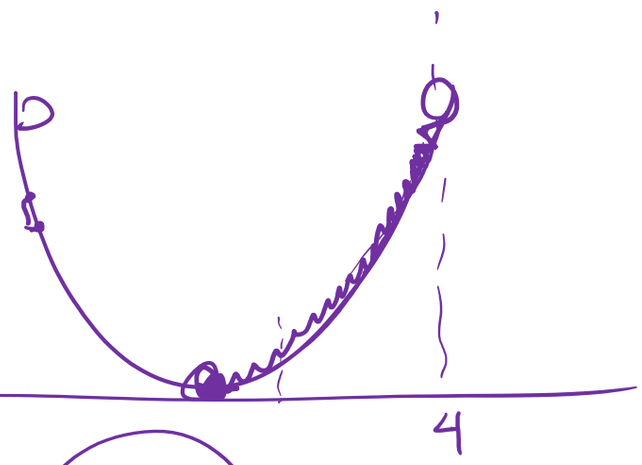


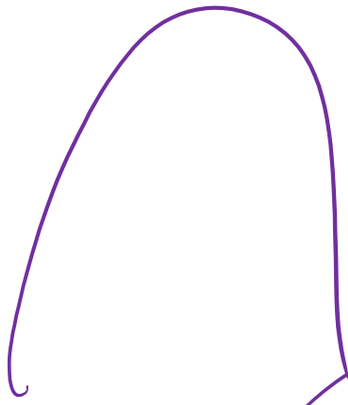
$$\frac{\partial f(x)}{\partial x} = 0 = \exp(x) + 2x = 0$$

$$\Rightarrow x = ?$$

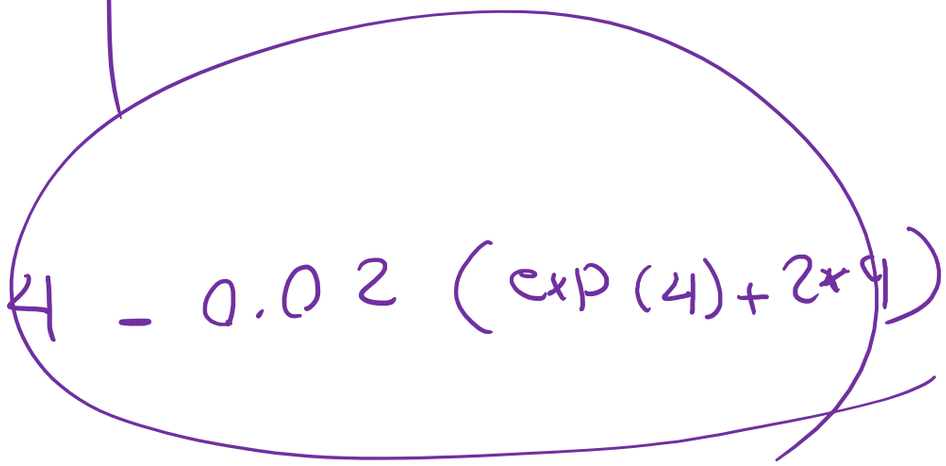


GD





GA



$$x^{\{t+1\}} \leftarrow x^{\{t\}} - \alpha \frac{\partial f(x)}{\partial x} = 4 - 0.02\,(\exp(4) + 2*4)$$

α = 0.02
hyper parameter

① Quotient rule: $\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$

② Chain rule: $(U)^{n'} = nU'U^{n-1}$

③ product rule: $(fg)' = f'g + fg'$

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

K = kernel

*

0	1	2
2	1	3
1	2	1

S = sub matrix

=

0	0	-2
2	0	-3
1	0	-1

Sum( ) =

0 + 0 + -2 + 2 + .. =

-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						

K = [1 0 -1 1 0 -1 1 0 -1]

S = [0 1 2 2 1 3 1 2 1]

K.S