Georgia Tech

Optimization

Mahdi Roozbahani
Georgia Tech

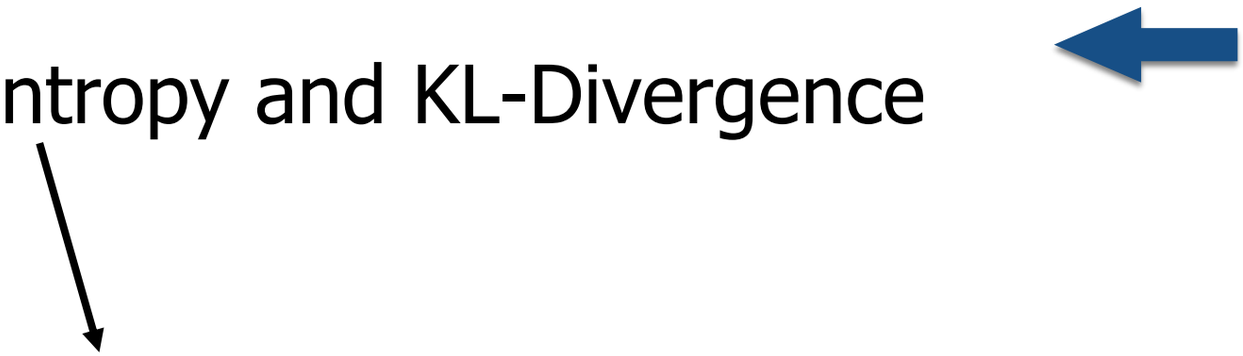
Outline

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

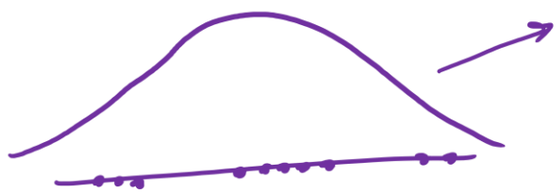
$$H(p,q) = -\sum_{x \in \mathcal{X}} p(x)\ \log q(x) = H(P) + KL[P][Q]$$

This is because:

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

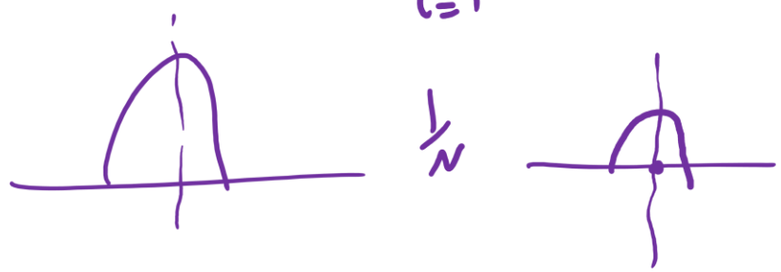
$$H(p,q) = \sum_{x_i} p(x_i)\ \log \frac{1}{q(x_i)}$$

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



$$f(x|\mu,\sigma) = f(x|\theta)$$

$$\text{Max } \ell(\theta|X) = \sum_{i=1}^{N} \log f(x^{[i]}|\theta)$$



$\frac{1}{N}$

p → True actual pdf

Q → predicted pdf model

$$\text{Min } -\frac{1}{N}\ell(\theta|x) = -\frac{1}{N}\sum_{i=1}^{N}\log \underbrace{f(x^{[i]}|\theta)}_{Q}$$

$$\frac{1}{N}\sum(\cdot) = E[\cdot]$$

$$= \frac{1}{N}\sum \log Q = E[-\log Q] = \overset{\text{avg}}{\text{Negative}} \overset{\log}{\text{likelihood}}$$

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

$$E\left[-\log P + \log P - \log Q\right]$$

$$\log a - \log b = \log \frac{a}{b} = -\log \frac{b}{a}$$

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

$$E[g(x)] = \sum_{i=1}^{N} P(x)\, g(x)$$

$$E\left[-\log P\right] + E\left[-\log \frac{Q}{P}\right]$$

$$\sum_{i=1}^{N} P(x) \log \frac{1}{P(x)} + \sum P(x) * \left(-\log \frac{Q(x)}{P(x)}\right)$$

$$H(x) + KL[P][Q] = CE = \text{Negative log likelihood}^{\text{avg}}$$

Labeling target values

Label encoding (ordinal) and One-hot encoding

$$X = \begin{bmatrix} h & w & a & \cdots \\ & & & \\ & & & \end{bmatrix}_{n \times d} \qquad y_a = \begin{bmatrix} \text{cat} \\ \text{fish} \\ \text{dog} \\ \text{cat} \\ \vdots \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \\ \vdots \end{bmatrix} \rightsquigarrow \text{ML} \quad \hat{y}_p = \begin{bmatrix} 0 \\ 0.5 \\ 1.2 \\ \vdots \end{bmatrix}$$

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

$$y_a = \begin{bmatrix} [1 & 0 & 0] \\ [0 & 1 & 0] \\ [0 & 0 & 1] \\ [1 & 0 & 0] \\ \vdots \end{bmatrix} \rightsquigarrow \text{ML} \Rightarrow \hat{y}_p = \begin{bmatrix} [70 & 20 & 7] \\ [20 & 137 & 2] \\ [5 & 6 & 80] \\ \vdots \end{bmatrix} \rightsquigarrow \text{Softmax} \quad \hat{y}_p = \begin{bmatrix} [0.7 & 0.2 & 0.1] \\ [0.3 & 0.6 & 0.1] \\ [0.1 & 0.1 & 0.8] \\ \vdots \end{bmatrix}$$

$$\text{Min} \; \| y_a - \hat{y}_p \|_2^2 = [(1-0.7)^2 + (0-0.2)^2 + (0-0.1)^2] + [(0-0.3)^2 + (1-0.6)^2 + (0-0.1)^2] + \cdots$$

$\nearrow x^{[2]}$

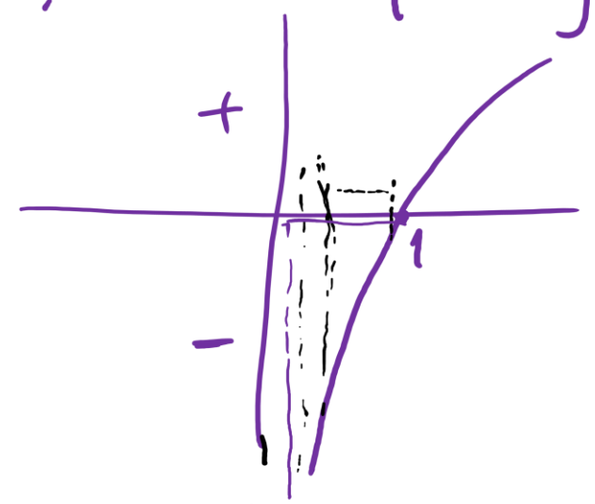
$$\text{Max} \; y_a \hat{y}_p^T = [1 \; 0 \; 0] \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix} + [0 \; 1 \; 0] \begin{bmatrix} 0.3 \\ 0.6 \\ 0.1 \end{bmatrix} + \cdots = N$$

Why Cross entropy and not simply use dot product?

$$CE = -\sum P(x) \log Q(x) = -\sum y_a \log \hat{y}_p$$

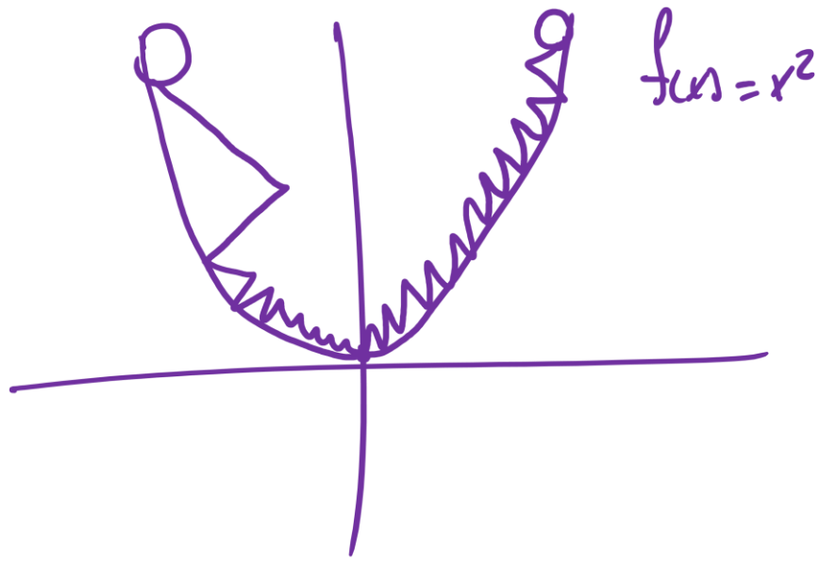
$$y_a = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \hat{Y}_p = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.6 & 0.1 \end{bmatrix}$$

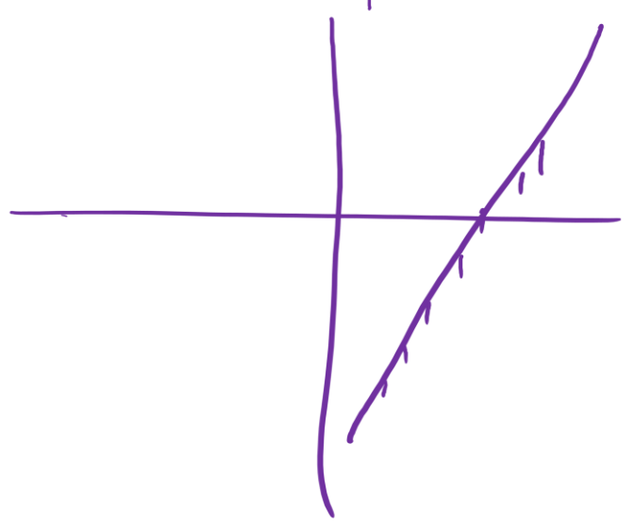
$$CE = -\left[\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \log 0.7 \\ \log 0.2 \\ \log 0.1 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \log 0.3 \\ \log 0.6 \\ \log 0.1 \end{bmatrix} + \dots \right]$$



① Min CE $\quad -y_a \log \hat{y}_p^T$

② Min $\quad -y_a \hat{y}_p^T$





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?

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

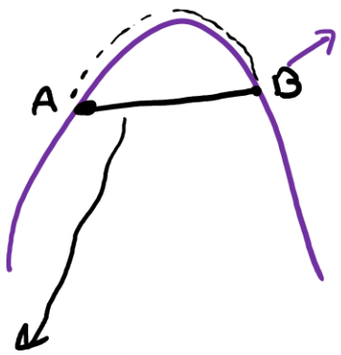
By Jensen Inequality

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

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

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

Strictly Concave

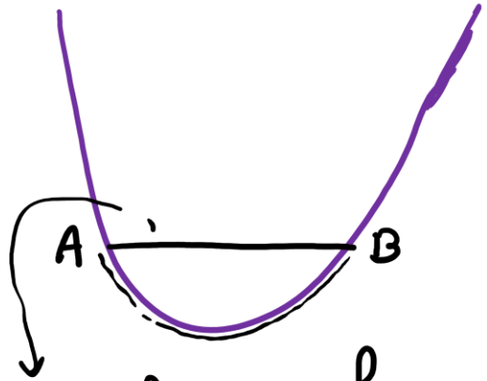


$f(x) = -x^2$ concave

$f''(x) = -2$

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

Convex



$f(x) = x^2$

$f''(x) = 2$

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

$f(x) = x^2 \quad x = [1 \quad 2]$

$P(x) = [\frac{1}{2} \quad \frac{1}{2}]$

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

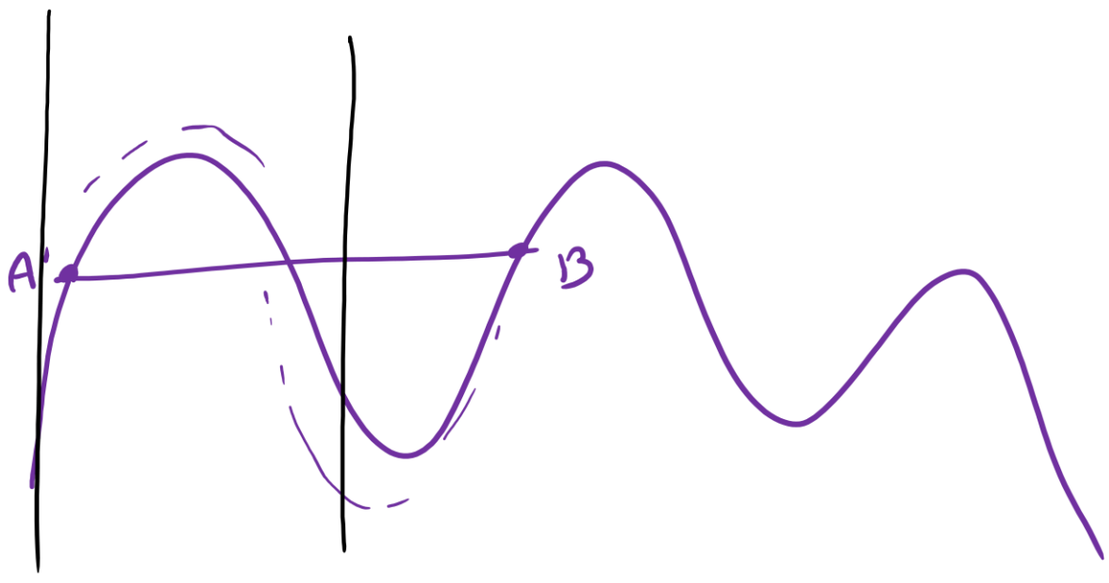
- $f(x=1) = 1 \quad f(x=2) = 4$

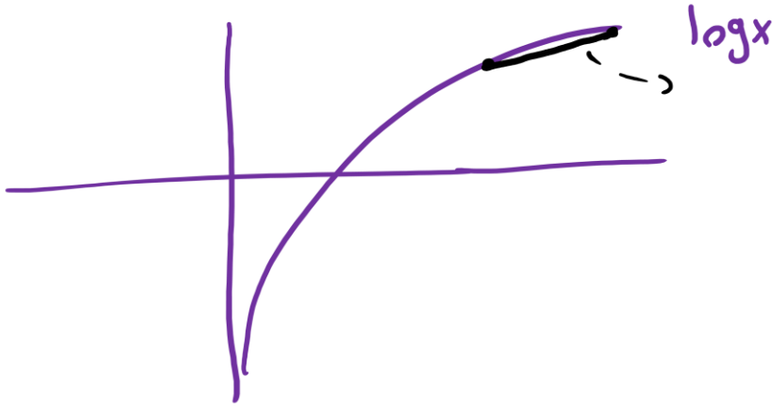
$E[f(x)] = \frac{1}{2} * 1 + \frac{1}{2} * 4 = 2.5$

$f(E[x]) = ? \quad E[x] = \sum P(x) x$

$E[x] = \frac{1}{2} * 1 + \frac{1}{2} * 2 = 1.5$

$f(E[x] = 1.5) = 2.25$





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

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

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

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

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

$$\leq \log \sum p(x)\, g(x)$$

$$\leq \log \sum \cancel{p(x)} \frac{Q(x)}{\cancel{P(x)}}$$

$$\leq \log \sum Q(x)$$

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

Optimization $\min f(x)$ → objective function

s.t $g(x) = c$ → Equality constraint (→ constraint function)

$g(x) \leq c$, $g(x) \geq c$ } ⇒ Inequality constraint ⇝ KKT conditions

Lagrange function

↘ $f(x,y)$

s.t $x+y=2$ $\quad g(x,y) = x+y-2$

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

① Linear programming ⇝ $f(x,y) = x+y$, s.t. $x+y=2$

② Quadratic programming ⇝ $f(x,y) = x^2+y$, s.t. $x+y=2$

③ Non linear programming ⇝ $f(x,y) = x^3+y^2$, s.t. $x^2+y=2$

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

M = # hours you study ML/day

S = # hours you sleep/day

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

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

δ, α, β

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

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

$\rightsquigarrow$ Lagrange multiplier

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

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

$$\frac{\partial L}{\partial \delta} = 0 \Rightarrow 0+0-(M+S-24) = 0 \Rightarrow M+S = 24 \Rightarrow \frac{\delta}{12} + \frac{\delta}{6} = 24 \Rightarrow$$

$$\delta = 96$$

$$\frac{\partial L}{\partial M} = 0 \Rightarrow 12M + 0 - \delta(1) = 0 \Rightarrow M = \frac{\delta}{12} = \frac{96}{12} = 8$$

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

$M+S = 24 \qquad 8+16 = 24$

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

s.t. $M + S = 24 \Rightarrow g(M,S) = M + S - 24 \rightarrow \delta_1$

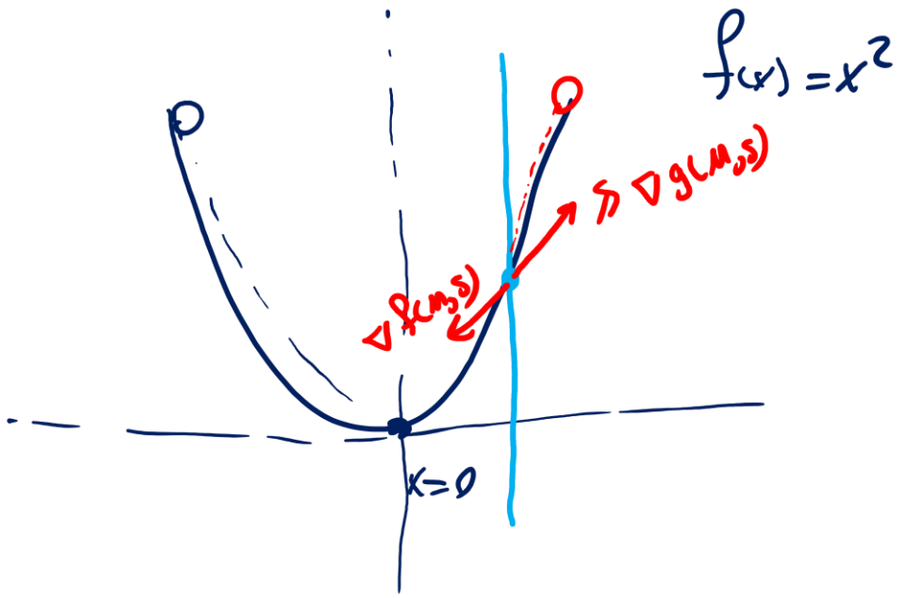
s.t. $M - S = 12 \Rightarrow h(M,S) = M - S - 12 \rightsquigarrow \delta_2$

$$L(M,S,\delta_1,\delta_2) = f(M,S) - \delta_1 g(M,S) - \delta_2 h(M,S)$$

$$L(x, \lambda) = f(x) - \lambda g(x)$$

$$\nabla L(x, \lambda) = 0 \implies \nabla (f(x) - \lambda g(x)) = 0$$

$$\nabla f(x) = \lambda \nabla g(x)$$



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

s.t. $M + S \leq 24$

$g(M,S) = M + S - 24 \leq 0$ → $g(M,S) = 0$ ⇝ Active solution

→ $g(M,S) < 0$ ⇝ Inactive solution

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

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

② Primal feasibility $g(M,S) \leq 0$

③ Dual feasibility $\lambda \geq 0$

④ Complementary slackness $g(M,S)\,\lambda = 0$ →

- $g(M,S) \neq 0 \rightsquigarrow \lambda = 0$ (Inactive)
- $\lambda > 0 \rightsquigarrow g(M,S) = 0$ (Active)

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

Convex $\rightsquigarrow$ min $\rightsquigarrow$ primal form

$$M + S = 24$$

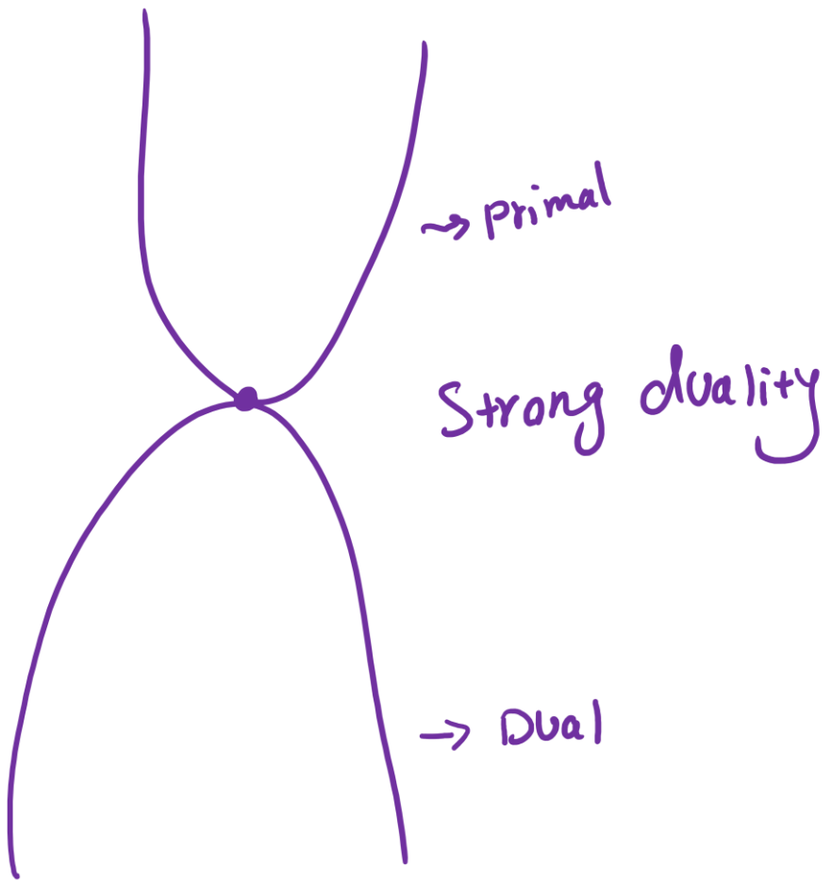
$$L(M,S,\delta) = \frac{1}{2}M^2 + \frac{1}{2}S^2 - \delta(M+S-24)$$

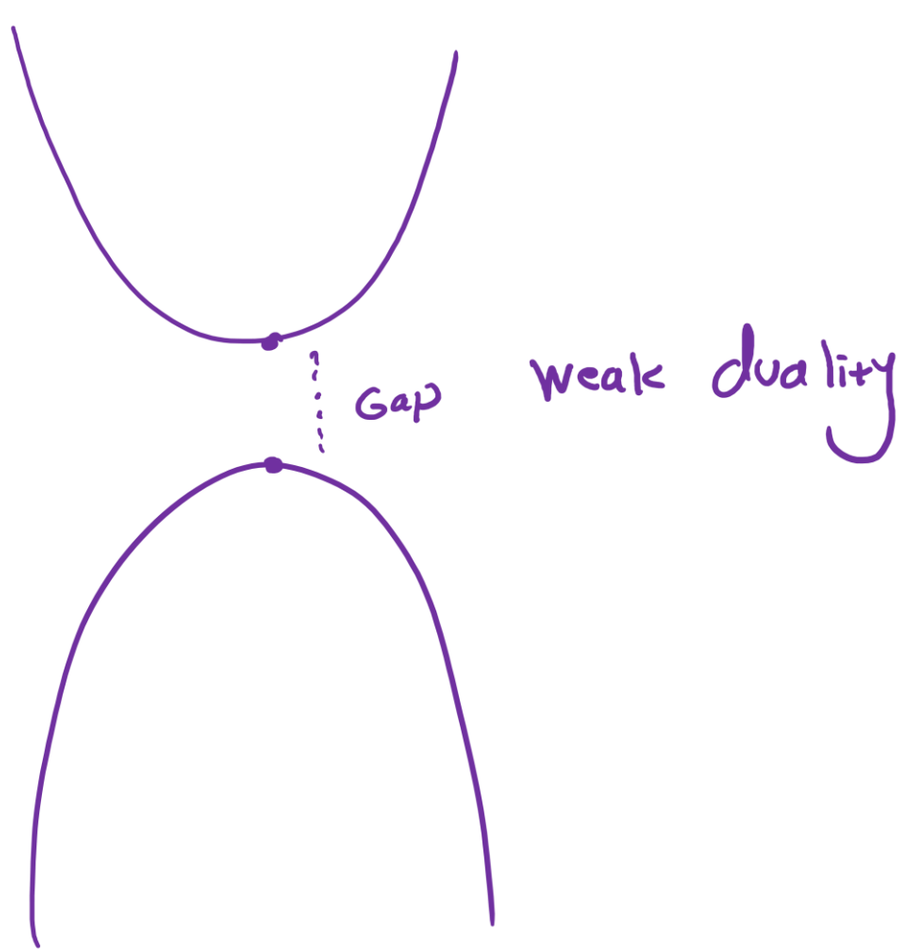
$$\frac{\partial L}{\partial M} = 0 \Rightarrow M - \delta = 0 \Rightarrow M = \delta \qquad \frac{\partial L}{\partial S} = 0 \Rightarrow S = \delta$$

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

Dual form $\Rightarrow$ concave $\rightarrow$ max

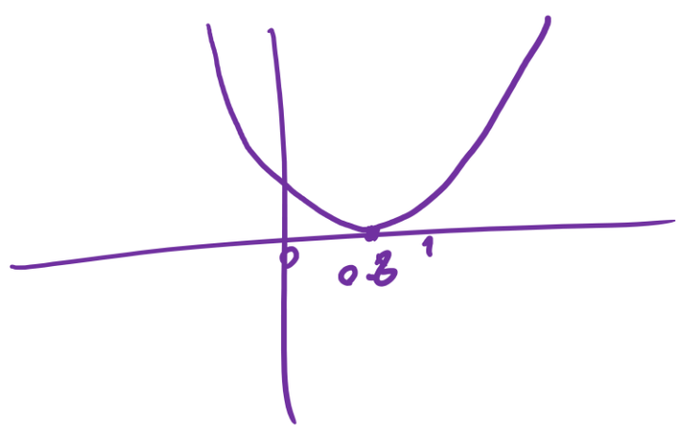
$$\frac{\partial L(\delta)}{\partial \delta} = 0 \Rightarrow -2\delta + 24 = 0 \Rightarrow \delta = 12$$


→ primal
Strong duality
→ Dual


Gap
Weak duality

$f(x) = e^x + x^2 \qquad f''(x) = e^x + 2 > 0$

$f'(x) = e^x + 2x = 0$



$f(x) = x^2$

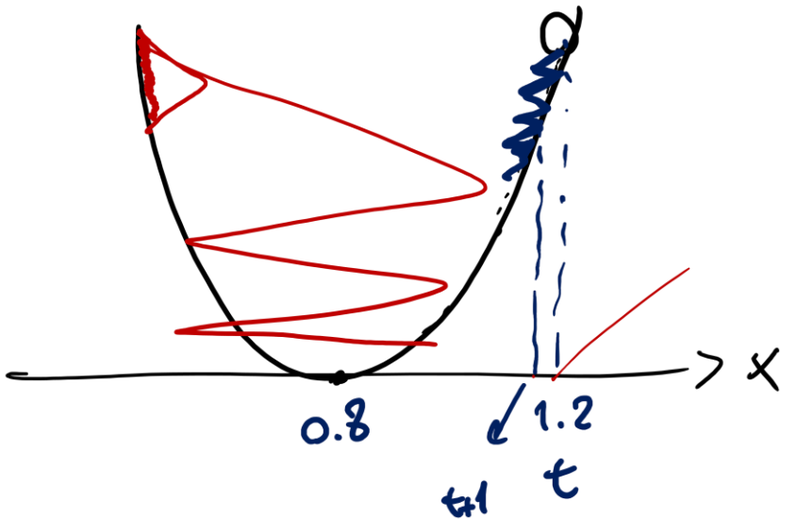
$f'(x) = 2x = 0$

$x = 0$

Gradient descent ← $\{t+1\}$ $\{t\}$

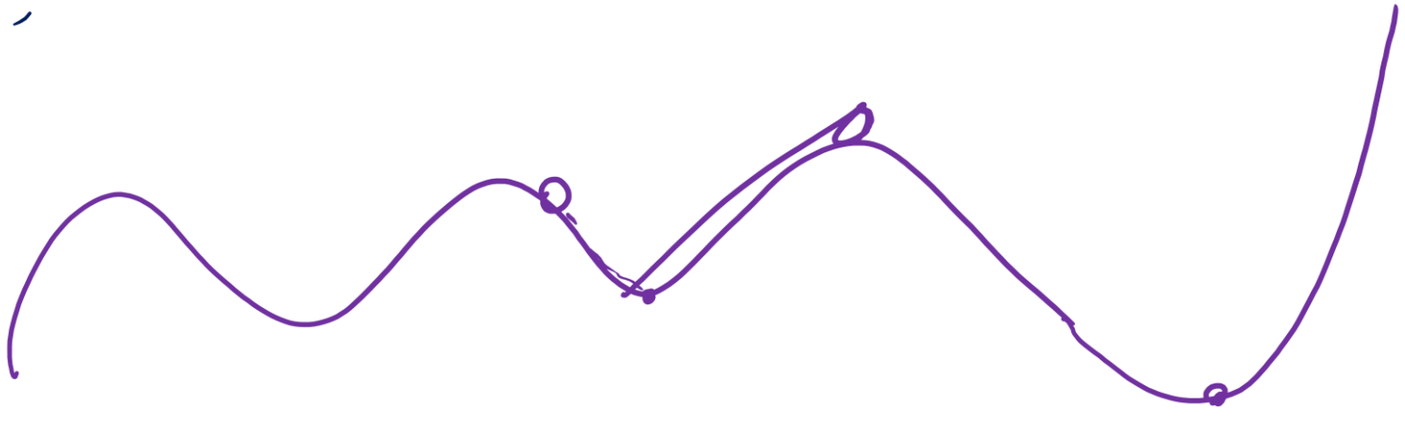
Gradient Ascent → $X \leftarrow X - \alpha f'(x)$ → Dynamic Adam optimizer

α → learning step 0.01



$\{t+1\}$

$X \leftarrow 1.2 - 0.01\left(e^{1.2} + 2 \times 1.2\right)$



Derivation rules:

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

Product rule: $(fg)' = f'g + fg'$

Chain rule: $(u^n)' = nu'u^{n-1}$