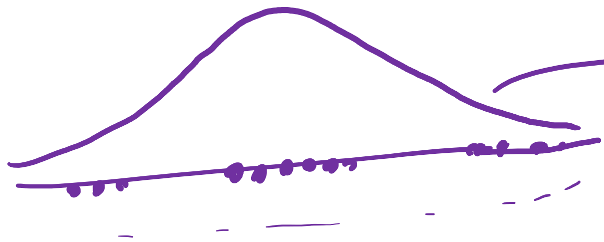
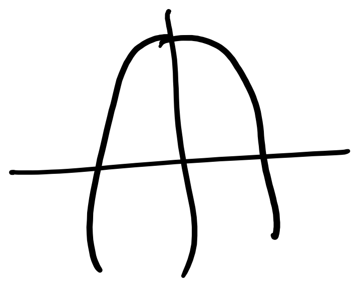
$$X = \overset{n}{\begin{bmatrix} \quad \end{bmatrix}}$$



$$\underbrace{f(x^{\{i\}}|\theta)} = q(x) \rightsquigarrow \text{predicted} = f(x^{\{i\}}|\mu, \sigma)$$

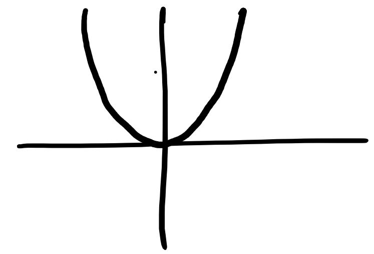
max

$$\underbrace{f(x)}_{\text{objective function}} = -x^2$$

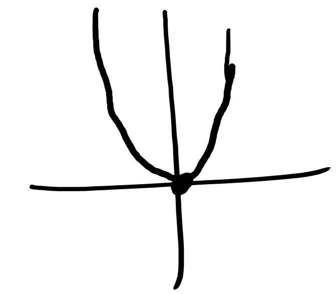


min

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



$$f(x) = 6x^2$$



$$L(\theta|x) = \prod_{i=1}^{N} f(x^{\{i\}}|\theta) \Rightarrow \max \ \ell(\theta|x) = \sum_{i=1}^{N} \log f(x^{\{i\}}|\theta) \Rightarrow \min \ \ell(\theta|x) = -\sum_{i=1}^{N} \log f(x^{\{i\}}|\theta)$$

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

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

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

$$E[-\log p(x) + \log p(x) - \log q(x)] = E[\underbrace{-\log p(x)}_{a} + \underbrace{\log \frac{p(x)}{q(x)}}_{b}]$$

$$= E[-\log p(x)] + E[\log \frac{p(x)}{q(x)}]$$

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

$$= \sum p(x)(-\log p(x)) + \sum p(x) \log \frac{p(x)}{q(x)}$$

$$= \sum p(x) \log \frac{1}{p(x)} + \sum p(x) \log \frac{p(x)}{q(x)}$$

$$= H(p) + KL[P][Q] = CE = H(p, q)$$

$-$ average log likelihood = cross entropy

$$KL[P][Q] = CE - H(p) = \boxed{-\sum p(x)\, q(x)} - H(p)$$

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:

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

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

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

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

$$X = \begin{bmatrix} h & w & a \\ & & \\ & & \end{bmatrix} \quad Y_a = \begin{bmatrix} cat \\ fish \\ dog \\ \vdots \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ \vdots \end{bmatrix} \xrightarrow{ML} \hat{Y}_p = \begin{bmatrix} 0.2 \\ 0.5 \\ 1.3 \\ \vdots \end{bmatrix}$$

loss function compares

Y_a & $\hat{Y}_p \Rightarrow \in \mathbb{R}$

pmf

$Cat = [1 \quad 0 \quad 0]$ $\quad fish = [0 \quad 1 \quad 0]$ $\quad dog = [0 \quad 0 \quad 1]$ ⇝ One-hot encoding

$$Y_a = \begin{bmatrix} [1 & 0 & 0] \\ [0 & 1 & 0] \\ [0 & 0 & 1] \\ & \vdots & \end{bmatrix} \rightarrow p_i \quad \xrightarrow{ML} \quad \hat{Y}_p = \begin{bmatrix} [74 & 5 & 6] \\ [2 & 85 & 3] \\ [3 & 5 & 66] \\ & \vdots & \end{bmatrix} \xrightarrow{Softmax} = \begin{bmatrix} [0.7 & 0.2 & 0.1] \\ [0.3 & 0.6 & 0.1] \\ & \vdots & \end{bmatrix} \rightarrow q_i$$

pmf

$$\text{Min} \sum_{i=1}^{N} \| y_a^{\{i\}} - \hat{y}_p^{\{i\}} \|^2 = \left((1-0.7)^2 + (0-0.2)^2 + (0-0.1)^2 \right) + \dots\dots + \text{last data point} = 0$$

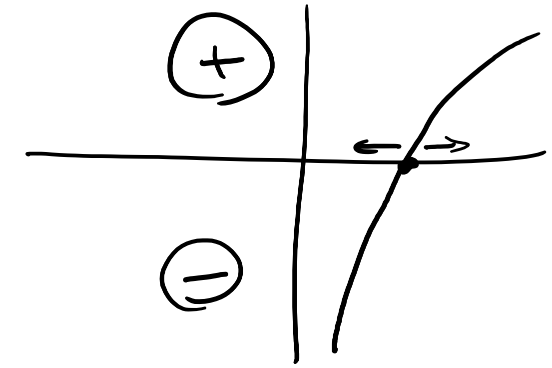
perfect result

$$\text{Max} \sum_{j=1}^{N} y_a^{\{i\}} \hat{y}_p^{T\{i\}} = \left([1 \quad 0 \quad 0] \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix} \right) + \dots\dots + \text{last data point} = N$$

perfect result

Why Cross entropy and not simply use dot product?

$$CE = -\sum_{i=1}^{N} p(x) \log q(x) = -\sum_{i=1}^{N} y_a^{\{i\}} \cdot \log \hat{y}_p^{\{i\}T}$$



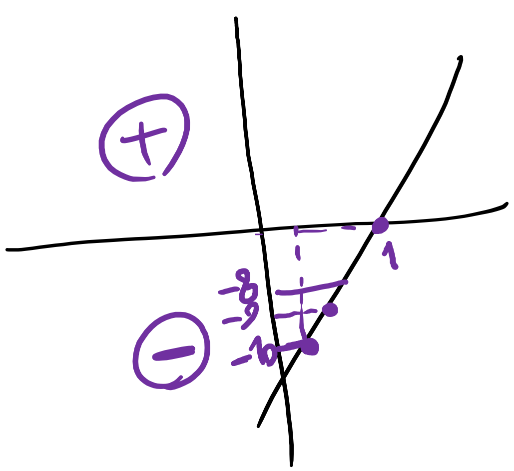
$$y_a = \begin{bmatrix} [1 & 0 & 0] \\ \\ \end{bmatrix} \qquad \hat{y}_p = \begin{bmatrix} [0.7 & 0.2 & 0.1] \\ \\ \end{bmatrix}$$

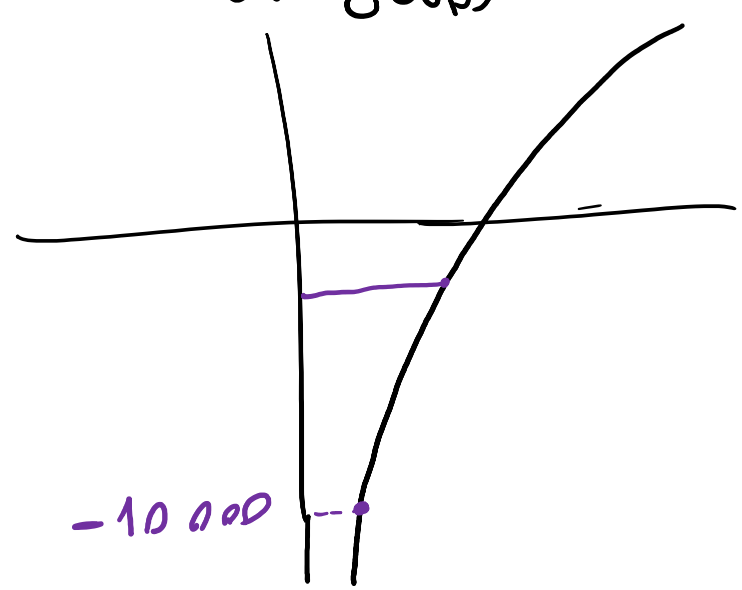
$$CE = -\left([1 \quad 0 \quad 0] \begin{bmatrix} \log 0.7 \\ \log 0.2 \\ \log 0.1 \end{bmatrix} + \cdots \right)$$

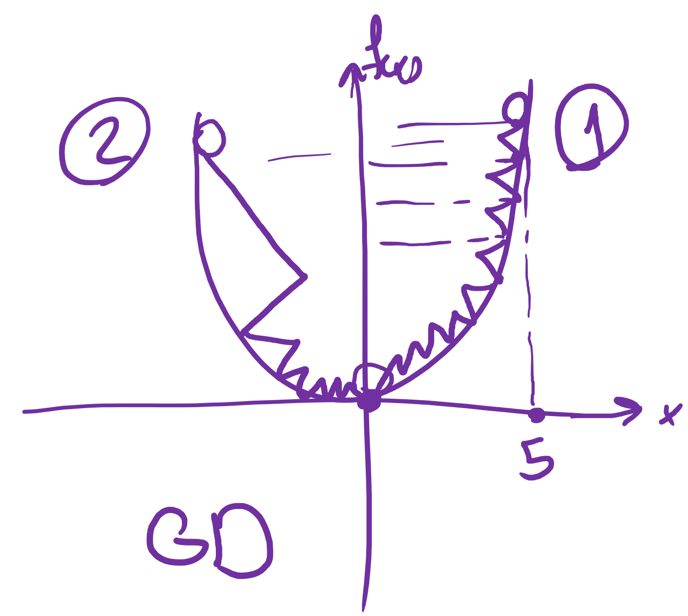
$f'(x) = 2x = 0$

$f(x) = x^2$

$$-\sum y_a^{\{i\}} \hat{y}_p^{\{i\}T} \qquad -\sum y_a^{\{i\}} \log(\hat{y}_p)^{\{i\}T}$$







Kullback-Leibler Divergence

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

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

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

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?

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

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

By Jensen Inequality

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

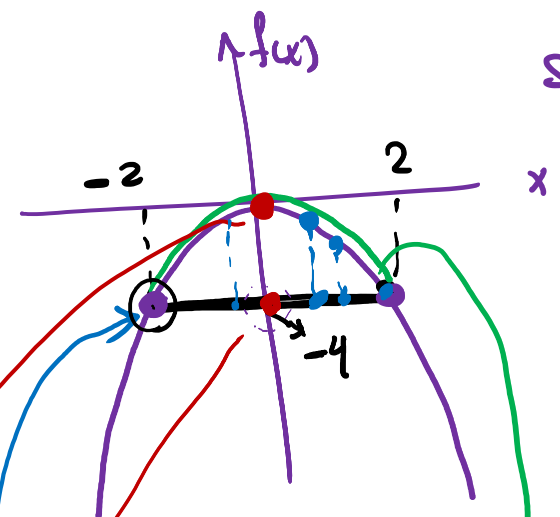
When $P = Q,\ KL[P||Q] = 0$

$f(x) = -x^2$

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

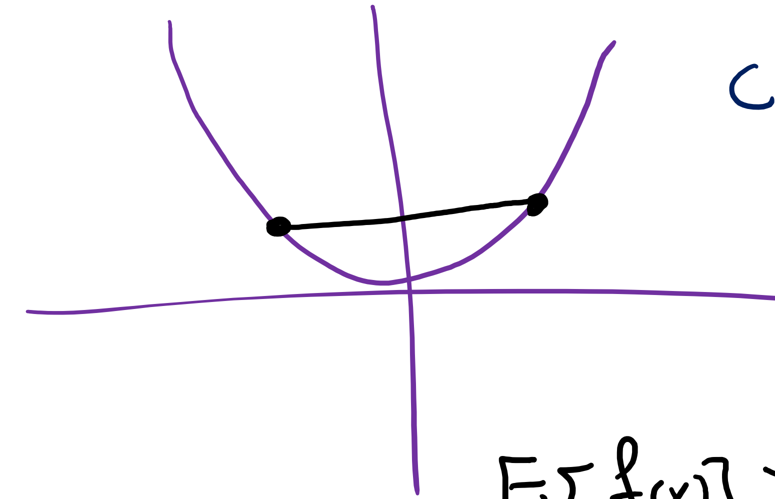
$f''(x) = -2$

concave



Strictly concave

concave



Convex

minimize

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

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

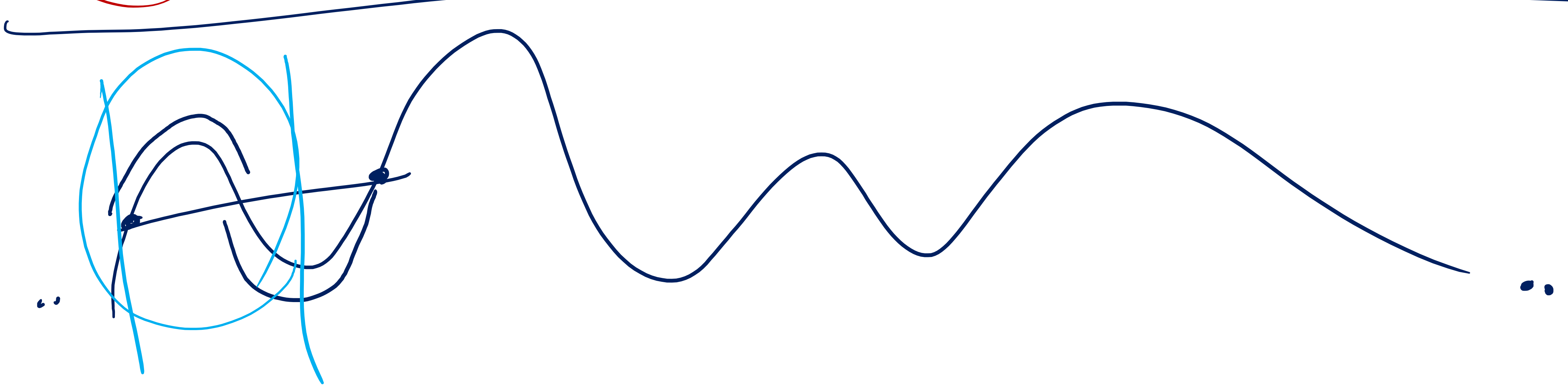
Jensen Inequality

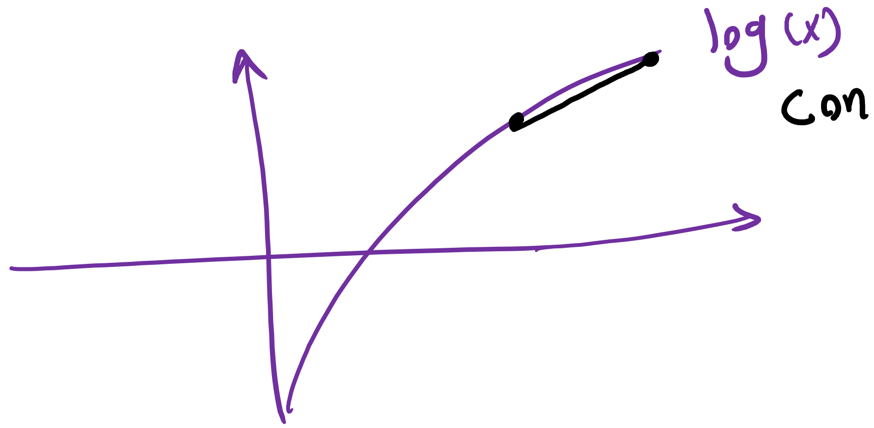
$x = [-2, 2]$ $\quad f(x) = [-4, -4]$

$p(x) = [0.5, 0.5] \to [0.4, 0.6]$

$$E[f(x)] = \sum_{-2,2} p(x) f(x) = 0.5 * -4 + 0.5 * -4 = -4$$

$$f(E[x]) = ? \Rightarrow E[x] = \sum_{-2,2} p(x)\, x = 0.5 * -2 + 0.5 * 2 = 0$$





concave

$$E[f(x)] \le f(E[x]) \Rightarrow \underbrace{E[\log(x)]}_{lb} \le \underbrace{\log(E[x])}_{ub}$$

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

$$-KL[P][Q] = E[g(x)] = E\left[\log\frac{Q(x)}{P(x)}\right] \le \log\left(E\left[\frac{Q(x)}{P(x)}\right]\right)$$

$$\le \log\sum \cancel{P(x)}\,\frac{Q(x)}{\cancel{P(x)}}$$

$$-KL[P][Q] \le \log \sum Q(x)^{=1}$$

$$Q(x) = [0.7\ 0.2\ 0.1]$$

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

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

Optimization → Objective function with no constraints

→ 〃 〃 with constraints

Equality — Inequality

→ Lagrange function

→ KKT conditions

gaussian

pdf

$f(x) = x^2$

$f(x) = \prod_{i=1}^{N} f(x^{(i)} | \theta)$

$f(x, y) = x^2 + y^2$

LP

① Linear programming $\quad f(x,y) = 2x + y$

s.t. $\quad x + y = 20$

② Quadratic programming QP $\quad f(x,y) = 2x^2 + y$

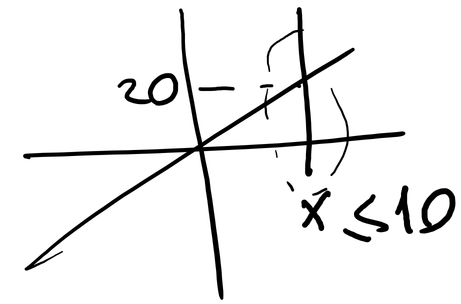
s.t. $\quad x + y = 20$

Strong duality

③ Non-linear 〃 $\quad f(x) = 2\exp(x^2) + \sqrt{y}$

s.t. $\quad x^3 + \log(y) = 20$

$f(x) = 2x$



$f(x) = x^2 \Rightarrow f''(x) = 2$ U

$f(x, y) = ?$

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

Hessian

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

M: # hours you study ML in a day

S: # " " sleep " " "

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

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

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

$$\text{s.t.} \quad M + S = 24 \rightsquigarrow \text{Constraint function} \quad g(M,S) = M + S - 24$$

lagrang multiplier $\rightsquigarrow$ δ, α, β

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

$$\frac{\partial L}{\partial \delta} = 0 \Rightarrow M+S-24=0 \Rightarrow M+S=24 \Rightarrow \frac{\delta}{12} + \frac{\delta}{6} = 24 \Rightarrow \delta = 96$$

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

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

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

$f(M,S)$

s.t $M+S=24 \rightsquigarrow \lambda_1 \rightsquigarrow g(M,S) = M+S-24$

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

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

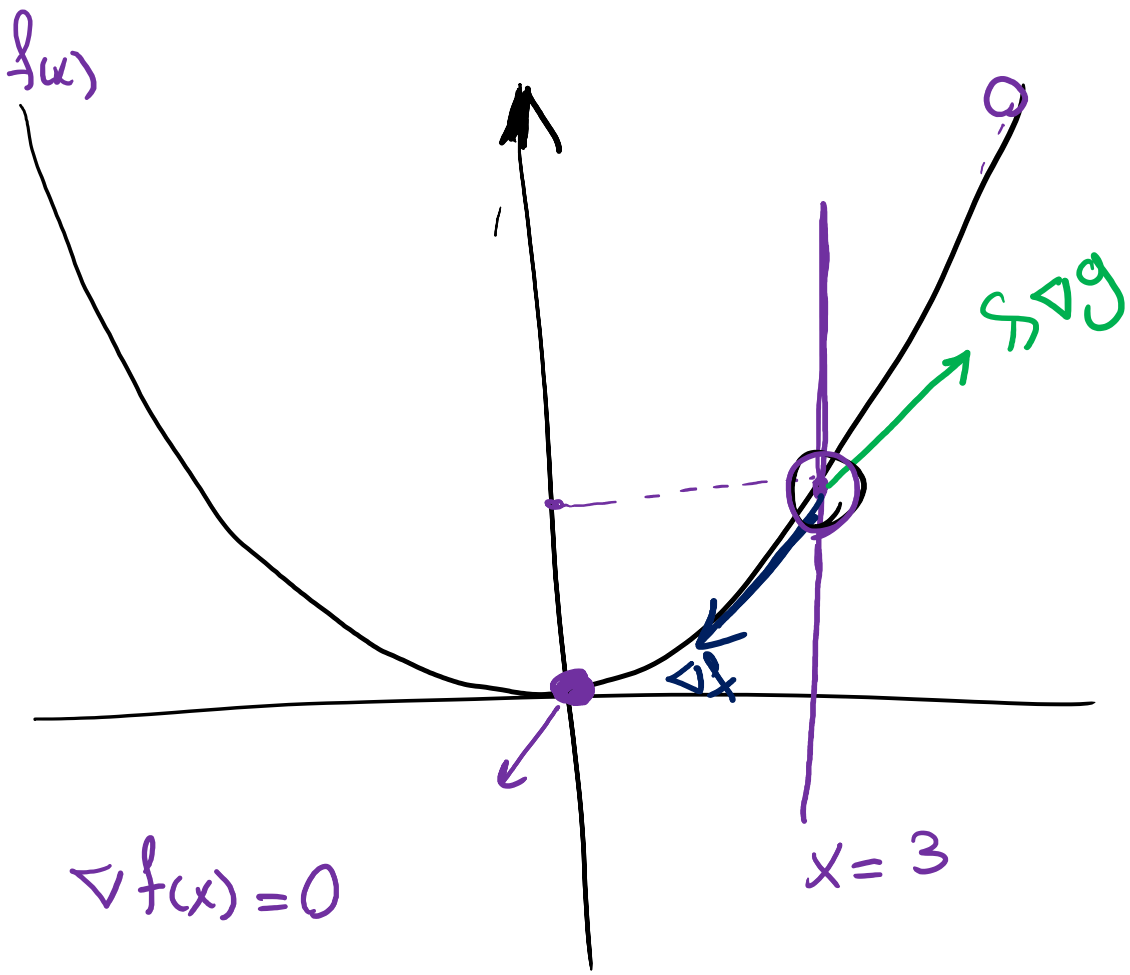
$f(m,s)$ $\qquad$ $L = f(m,s) - \delta\, g(m,s)$

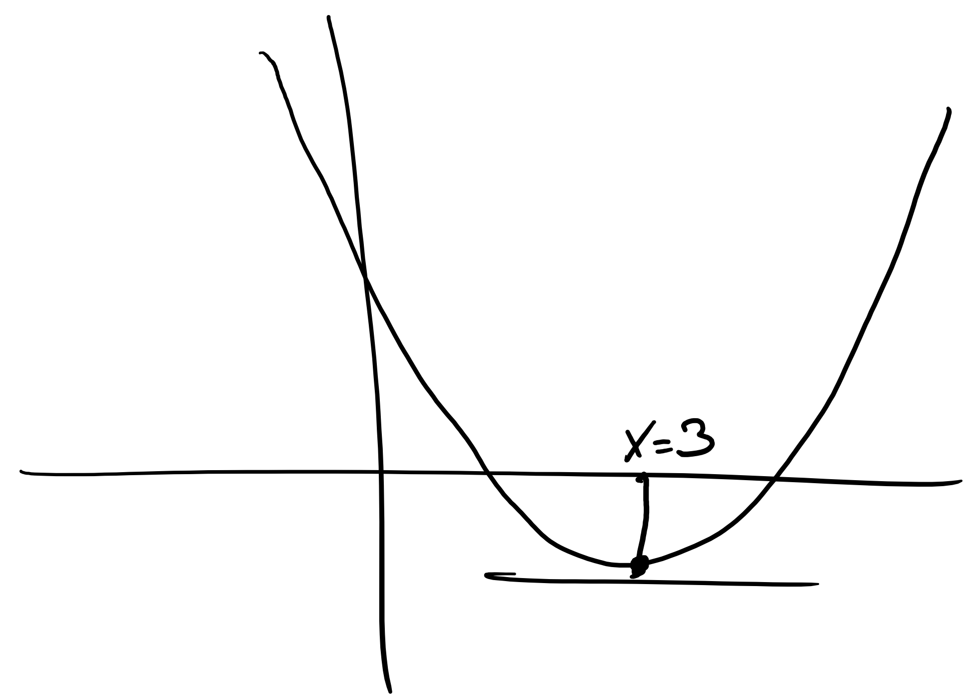
s.t. $g(m,s)$ $\qquad$ $\nabla L = 0 \Rightarrow \nabla(f - \delta g) = 0$

$\nabla f - \nabla(\delta g) = 0 \Rightarrow \nabla f - \delta \nabla g = 0 \Rightarrow \nabla f = \delta \nabla g$

$\nabla f \approx \nabla g$

$\nabla f = \delta \nabla g$





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

s.t. $m+s \le 24 \Rightarrow g(m,s) = m+s-24$

- Active solution $m+s=24$
- InActive solution $m+s<24$

Min

① Stationary condition $\Rightarrow L(m,s,\delta) = f(m,s) \overset{+}{+} \overset{-}{\delta}\, g(m,s) \Rightarrow \nabla L = 0$

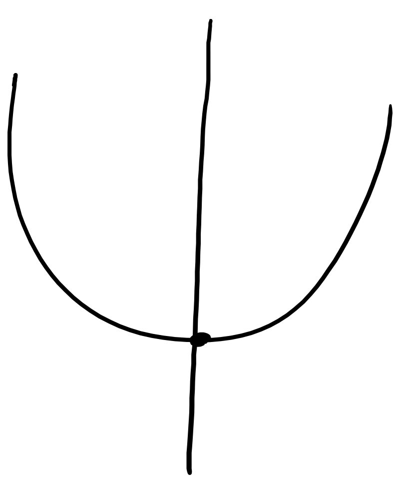
② primal feasibility $g(m,s) \le 0$

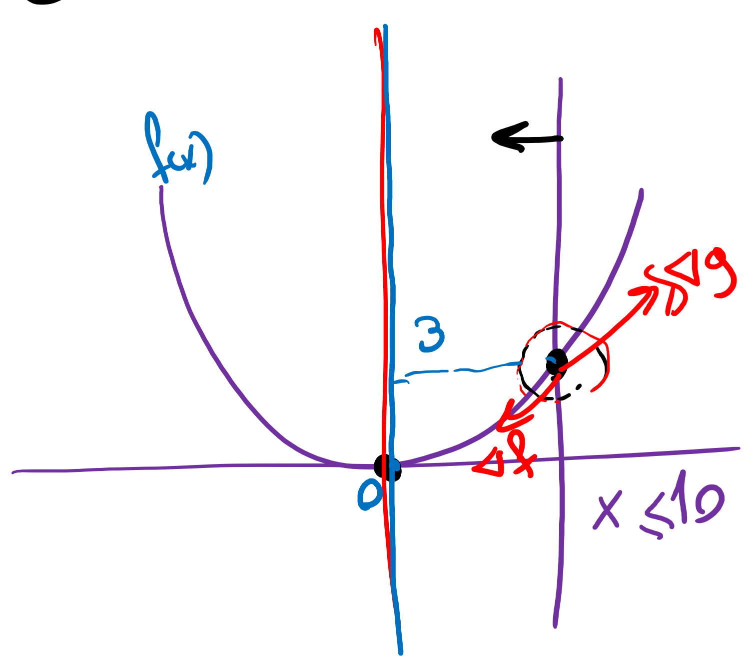
③ Dual feasibility $\delta \ge 0$

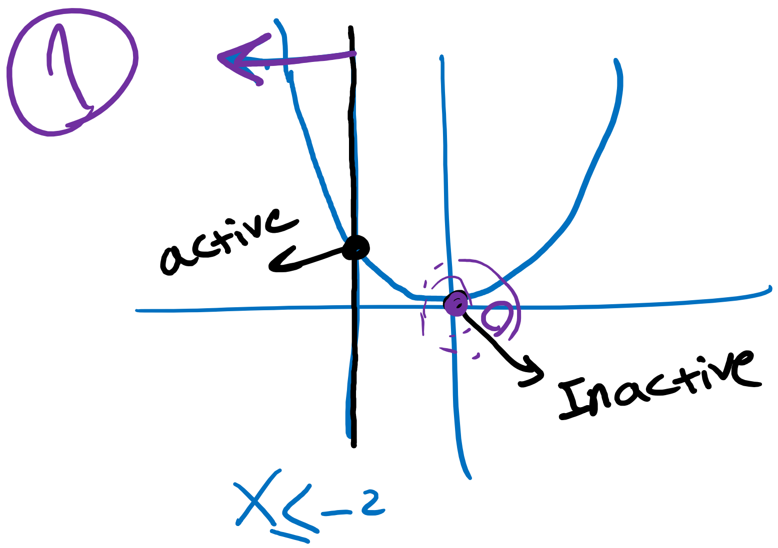
④ complementary slackness $\delta\, g(m,s) = 0$

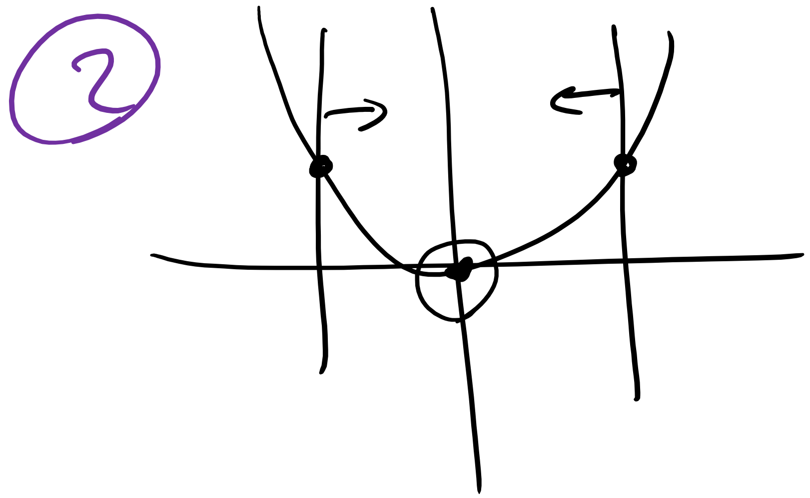
- $\delta = 0 \Rightarrow g(m,s) \neq 0$
- $\delta \neq 0 \Rightarrow g(m,s) = 0$

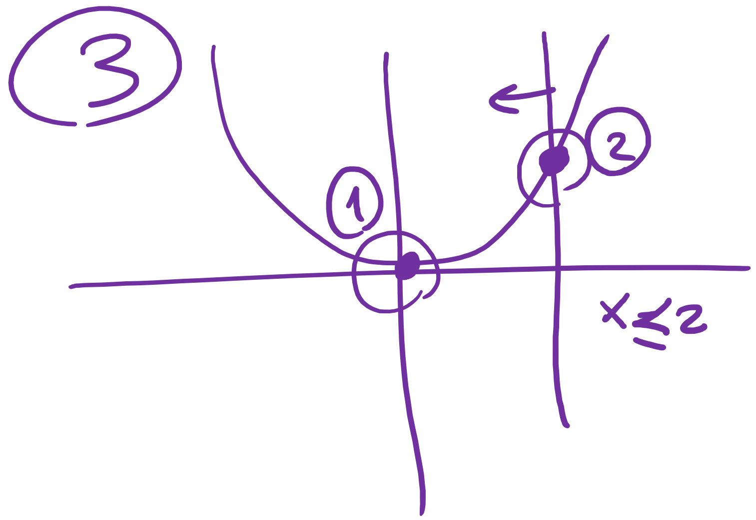
$\delta > 0$

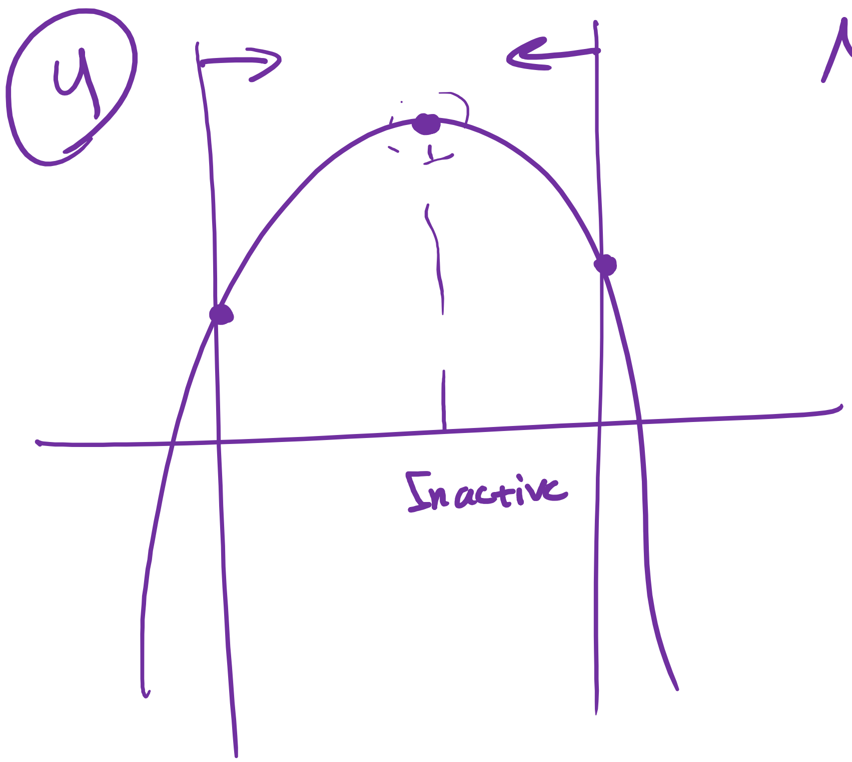












Min

$g(M,S) = \frac{1}{2}M^2 + \frac{1}{2}S^2$ Minimization convex

QP

Primal form

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

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

①

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

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

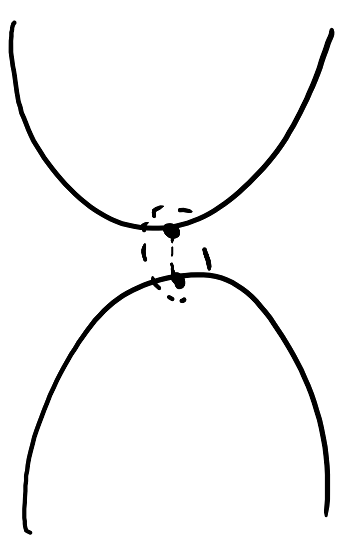
Dual form concave

②

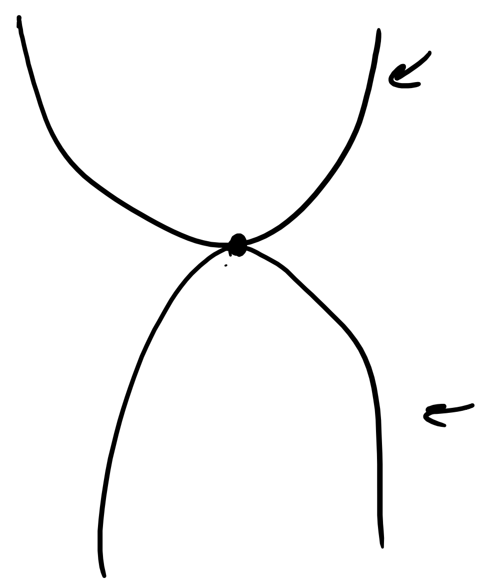
$L(\delta) = \frac{\delta^2}{2} + \frac{\delta^2}{2} - \delta(2\delta - 24) = \delta^2 - 2\delta^2 + 24\delta = -\delta^2 + 24\delta$

maximization

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



weak duality

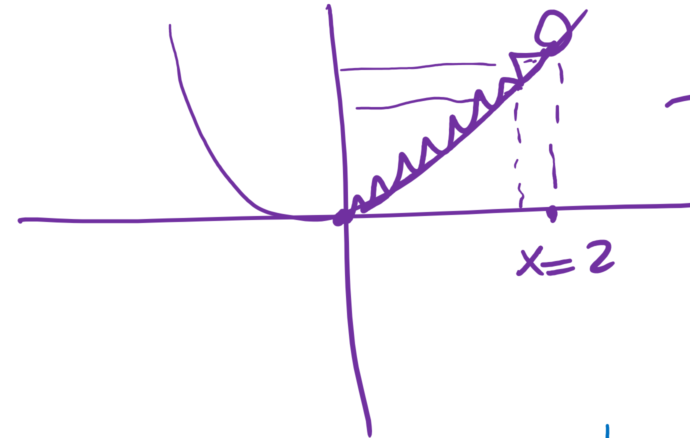


Strong duality

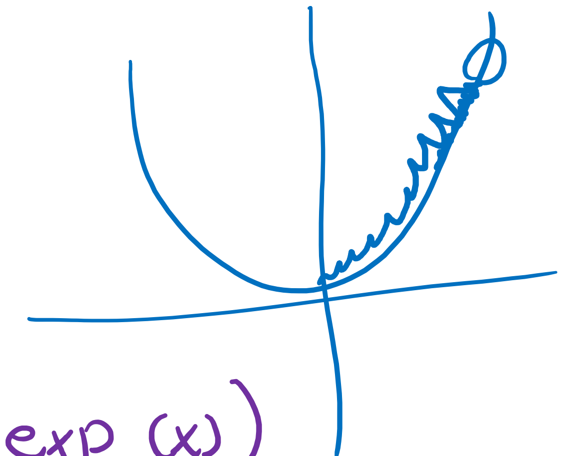
$$f(x) = x^2 + \exp(x)$$

$$f'(x) = 2x + \exp(x) = 0 \Rightarrow x = ?$$

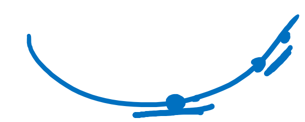
No closed form solution



$\rightarrow$ GD



$$x^{\{t+1\}} \leftarrow x^{\{t\}} - \alpha f'(x) = x^{\{t\}} - \alpha\,(2x + \exp(x))$$



$$x^{\{t+1\}} = 2 - \alpha\,(2 * 2 + \exp(2))$$

learning rate

0.002

hyper Parameter

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

$$f(x) = 2x = 0 \Rightarrow x = 0$$

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

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

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

$$\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} * \begin{bmatrix} 0 & 1 & 2 \\ 2 & 1 & 3 \\ 1 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -2 \\ 2 & 0 & -3 \\ 1 & 0 & -1 \end{bmatrix}$$

$$\mathrm{Sum}\left(\begin{bmatrix} \; & \; & \; \\ \; & \; & \; \\ \; & \; & \; \end{bmatrix}\right) =$$

$$0 + 0 + -2 + 2 + \cdots = -3$$

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