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

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

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

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

$$y = \underset{\text{actual}}{\downarrow} \begin{bmatrix} cat \\ dog \\ fish \\ cat \\ \vdots \end{bmatrix}_{n \times 1} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \\ \vdots \end{bmatrix} \xrightarrow{ML} \underset{\text{Predicted}}{\hat{y}} = \begin{bmatrix} 1 \\ 2 \\ 2.5 \\ 1 \\ \vdots \end{bmatrix}$$

one-hot encoding → Pdf

$$\underset{\text{actual}}{Y} = \begin{bmatrix} [1 & 0 & 0] \\ [0 & 1 & 0] \\ [0 & 0 & 1] \\ [0 & 0 & 1] \\ [1 & 0 & 0] \end{bmatrix} \rightsquigarrow ML \rightsquigarrow Softmax \rightsquigarrow \hat{Y} = \begin{bmatrix} [0.8 & 0.1 & 0.1] \\ [0.3 & 0.6 & 0.1] \\ [\dots & \dots & \dots] \\ \vdots \end{bmatrix}$$ ← Pdf

① $\| y - \hat{y} \|_2 = \sqrt{(1-0.8)^2 + (0-0.1)^2 + (0-0.1)^2} + \sqrt{(0-0.3)^2 + (1-0.6)^2 + \dots} + \dots$ Minimize

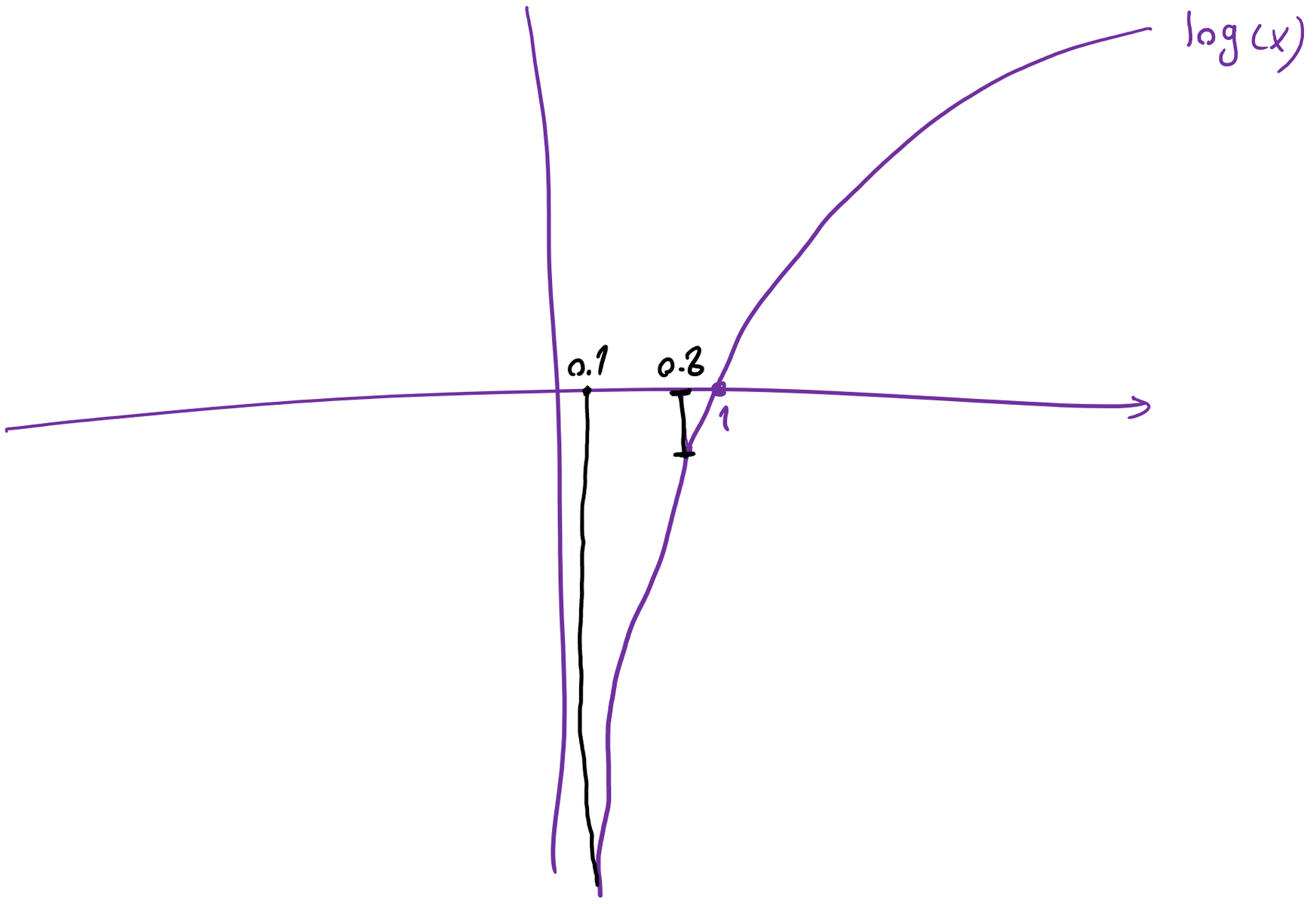
② $\overset{\text{sum}}{|y \cdot \hat{y}|} = \begin{bmatrix} 0.8 \\ 0.6 \\ \vdots \end{bmatrix} = \dots \checkmark$ maximize

Why Cross entropy and not simply use dot product?

$$CE = -\sum P(x) \log Q(x)$$

Minmization

$$CE = -\ (1 \times \log 0.8 + 0 \times \log 0.1 + 0 \times \log 0.1) - (0 \times \log 0.3 + \cdots) - \cdots$$



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?

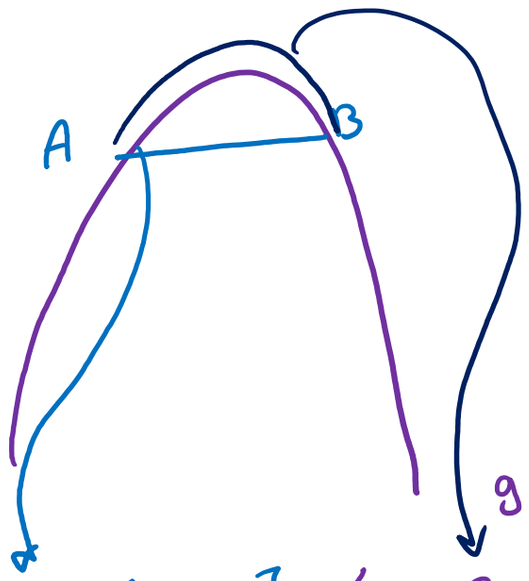
$$\sum_s P(s)\log\frac{Q(s)}{P(s)} \leq \log\sum_s P(s)\frac{Q(s)}{P(s)}$$ By Jensen Inequality

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

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

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

Concave

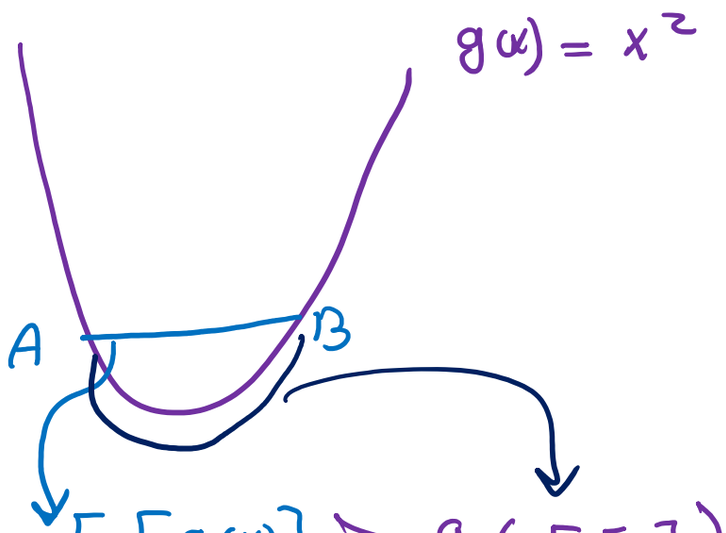


$g(x) = -x^2$

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

Jensen Inequality

Convex

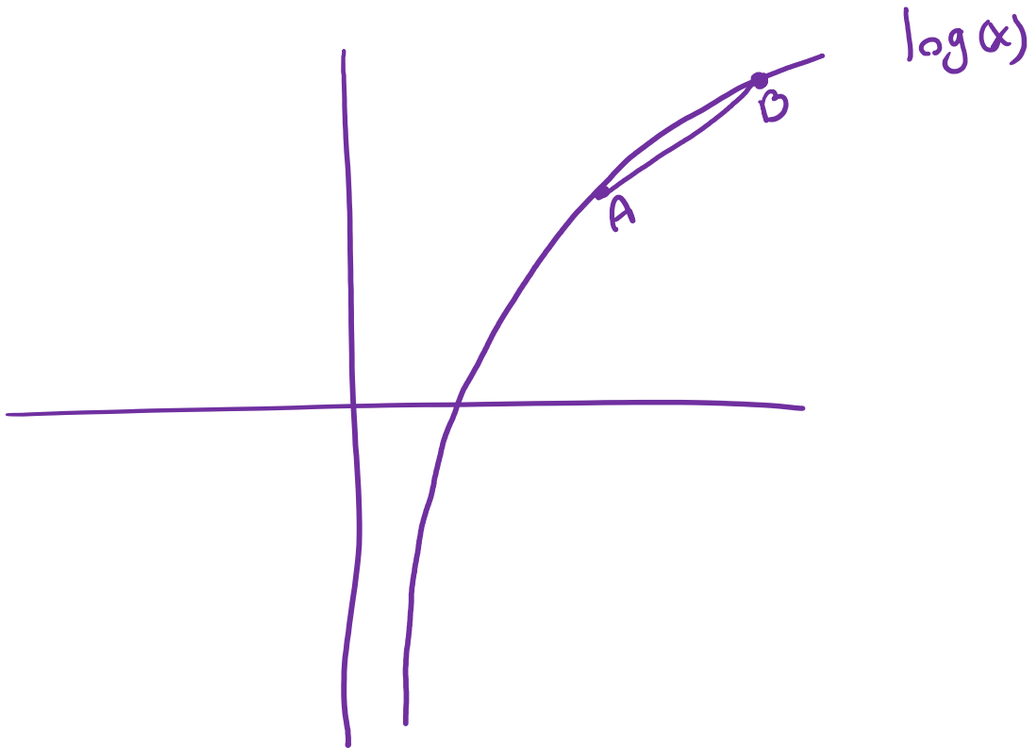


$g(x) = x^2$

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

$g'(x) = 2x$

$g''(x) = 2$



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

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

($\frac{Q(x)}{P(x)}$: pdf)

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

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

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

$$\leq \log \sum Q(x) = \log 1 = 0$$

$$-KL[P][Q] \leq 0 \quad \text{or} \quad KL[P][Q] \geq 0$$

Optimization

→ Probabilistic models → Gaussian distribution $\theta \in \{\mu, \sigma\}$

MLE ⇝ derivative

→ Non-probabilistic models ⇝ NO-constraint ⇝ derivative

Equality constraint

↓

Lagrange function

↓

derivative

Inequality constraint

⇝

KKT → 4 conditions

⇝

derivative

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

M # hours you study ML per day

S # hours you sleep per day

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

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

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

s.t $M+S = 24$ ⇒ $g(M,S) = M+S-24$

s.t $M-S = 8$

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

$L(M,S,\delta_1,\delta_2) =$

↳ $f(M,S) - \delta_1 g_1(M,S) - \delta_2 g_2(M,S)$

$\delta, \alpha, \beta, \mu$

$\nabla L(M,S,\delta) = \nabla\left(f(M,S) - \delta\, g(M,S)\right) = 0$

$\nabla f(M,S) = \delta \nabla g(M,S)$

$\nabla f(M,S) \approx \nabla g(M,S)$

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

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

$\delta = 96$

$\frac{\partial L(M,S,\delta)}{\partial M} = 0 \Rightarrow 12M - \delta = 0 \Rightarrow M = \frac{\delta}{12} = 8$

$M+S = 24$

$\frac{\partial L(M,S,\delta)}{\partial S} = 0 \Rightarrow 6S - \delta = 0 \Rightarrow S = \frac{\delta}{6} = 16$

$8+16 = 24$

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

$$M+S \leq 24 \Rightarrow g(M,S) = M+S-24 \qquad M+S-24 \leq 0$$

$$g(M,S) \leq 0$$

KKT conditions

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

① Stationarity condition $\nabla L(M,S,\lambda) = 0$

② primal feasibility $g(M,S) \leq 0$

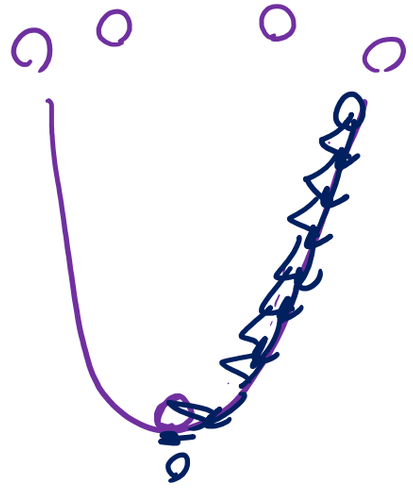
③ Dual feasibility $\lambda \geq 0$

④ Complementary slackness

$g(M,S) \neq 0 \quad \lambda = 0$

$g(M,S) = 0 \quad \lambda \geq 0$

$f(x) = x^2$



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

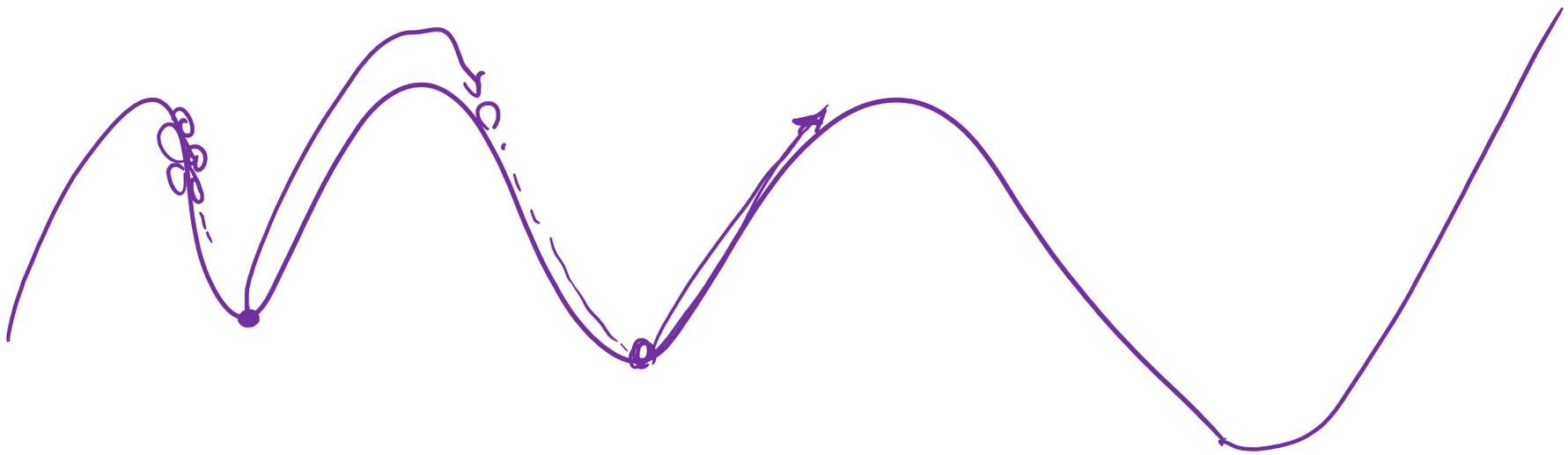
gradient direction

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

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

learning rate

$\alpha = 0.02$



$\mu, m, s \rightsquigarrow$ ① initialize randomly μ, m, s

$$\mu^{\{t+1\}} \leftarrow \mu^{\{t\}} - \alpha \frac{\partial L(\mu, m, s)}{\partial \mu}$$

$$m^{\{t+1\}} \leftarrow m^{\{t\}} - \alpha \frac{\partial L(\mu, m, s)}{\partial m}$$

$$s^{\{t+1\}} \leftarrow s^{\{t\}} - \alpha \frac{\partial L(\mu, m, s)}{\partial s}$$

Example 1:

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

Example 2:

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