EM Algorithm

EE698V - Machine Learning for Signal Processing

Vipul Arora

Announcements

- Coding Test
- Coding Tutorial

How to compare two pdf’s?

Entropy

X

$x \in \{1, 2, \ldots N\}$

- $H(p(x)) = -\sum_x p(x) \log(p(x))$ for discrete r.v. (Shannon)
- It is always ≥ 0
- When is it minimum?
- Is it upper bounded? When?
- $H(p(x)) = -\int p(x) \log(p(x)) dx$ for continuous r.v.
  - Called as continuous entropy
  - It is not lower bounded
  - We will discuss discrete case (Shannon entropy) only

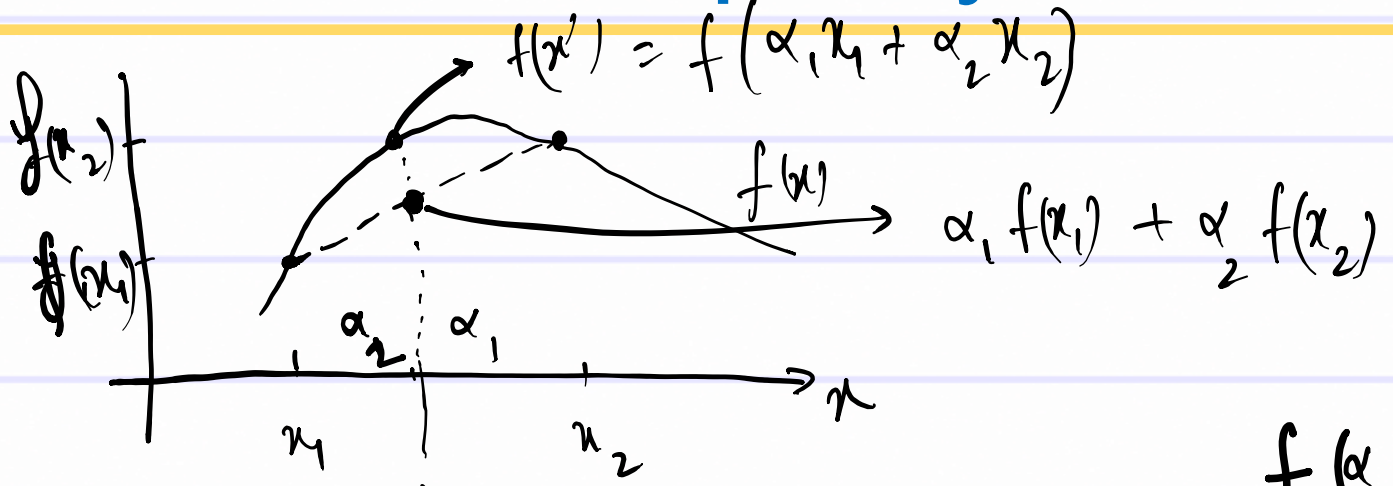
discr. r.v.

$p(x) \in [0, 1]$

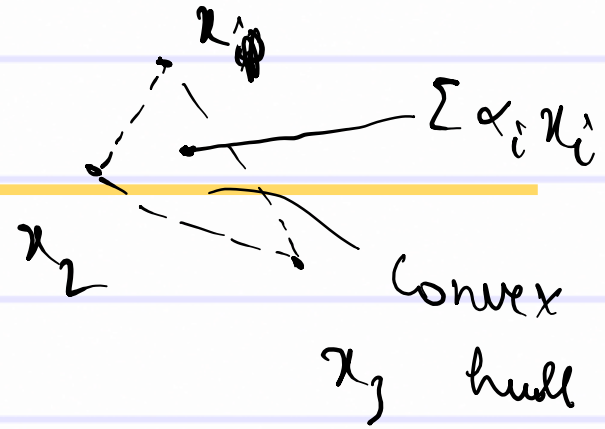
$\log p(x) \in (-\infty, 0]$

$-\log p(x) \in [0, \infty)$

Jensen's Inequality



$\alpha_1 f(x_1) + \alpha_2 f(x_2)$



$$x' = \frac{\alpha_1 x_1 + \alpha_2 x_2}{(\alpha_1 + \alpha_2)}$$

$$f(\alpha_1 x_1 + \alpha_2 x_2) \geq \alpha_1 f(x_1) + \alpha_2 f(x_2)$$

$\alpha_i \geq 0 \ \forall i$

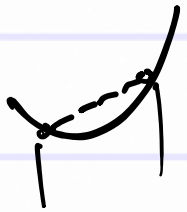
let $\sum_i \alpha_i = 1$

$$f\left(\sum_i \alpha_i x_i\right) \geq \sum_i \alpha_i f(x_i) \quad \text{if} \quad f'' \leq 0$$

and

$$f\left(\sum_i \alpha_i x_i\right) \leq \sum_i \alpha_i f(x_i) \quad \text{if} \quad f'' \geq 0$$

$y' =$



Lower Bound

$\log$ is $f'' \leq 0$

$$\therefore \quad \log\left(\sum_i \alpha_i x_i\right) \geq \sum_i \alpha_i \log(x_i)$$

$$H(p(x)) = -\sum_x \underbrace{p(x)}_{\alpha_i} \log p(x)$$

$$\leq \log \sum_i p(x_i)\, p(x_i)$$

$$\geq -\log \sum_i p^2(x_i)$$

$\alpha_i \geq 0$

$\sum_i \alpha_i = 1$

$$\sum_i p^2(x_i) \leq 1$$

$$\log(\downarrow) \leq 0$$

$$-\log\left(\sum_i p^2(x_i)\right) \geq 0$$

$$\Rightarrow H(p(x)) \geq 0$$

$H(p(x)) = 0$, when $p(x_i) = \delta_{ii^*}$

Upper Bound

$$H(p(x)) = -\sum_x p(x) \log p(x)$$

$$= \sum_x p(x) \log \frac{1}{p(x)}$$

Using Jensen's Ineq. $\log \sum_i \alpha_i x_i \geq \sum_i \alpha_i \log x_i$

$$H(p(x)) \leq \log\left(\sum_x p(x) \frac{1}{p(x)}\right)$$

$$= \log \sum_x 1$$

$$= \log N$$

N is the no. of possible values that r.v. x can take

Cross Entropy between 2 pdf's

- $H(p(x), q(x)) = -\sum_x p(x) \log(q(x))$ for discrete r.v.
- It is always ≥ 0
- When is it minimum?
- Is it upper bounded?

$H(p(x), p(x)) \geq 0$

$= 0$ when $\int -$

- $H(p(x), q(x)) = -\int p(x) \log\big(q(x)\big) dx$ for continuous r.v.
  - We will study the discrete case only

No Upper Bound

$$H(p(x), q(x)) = -\sum_{i=1}^{N} p(x_i) \log q(x_i)$$

unif, $p(x_i) = \frac{1}{N} \quad \forall i$

$$= -\sum_{i} \frac{1}{N} \log q(x_i)$$

for i^* $q(x_{i^*}) \to 0$

then $H \to \infty$

KL divergence between 2 pdf's

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

- $\mathrm{KL}(p(x)||q(x)) = -\sum_x p(x) \log \frac{q(x)}{p(x)}$ for discrete r.v. $= H(p(x), q(x)) - H(p(x))$
- $\mathrm{KL}(p(x)||q(x)) = -\int p(x) \log \frac{q(x)}{p(x)} dx$ for continuous r.v.
- It is always ≥ 0

if $p(x) = q(x)$, then $KL = 0$

- When is it minimum?
- Is it upper bounded? No
- Can you express it in terms of entropy?

- Can u use $H(p(x), q(x))$ as distance metric?
  - only if one is δ

- Can we use KL div. as distance metric?
  - Yes, but one should not be 0 when the other is non-zero.

Mutual Information between two r.v.'s

- $I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y)$
- Express it in terms of KL divergence
- It is always ≥ 0
- When is it minimum?
- Is it upper bounded?

Relation with KL divergence

$$I(x;y) = -\sum_x p(x) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y)$$

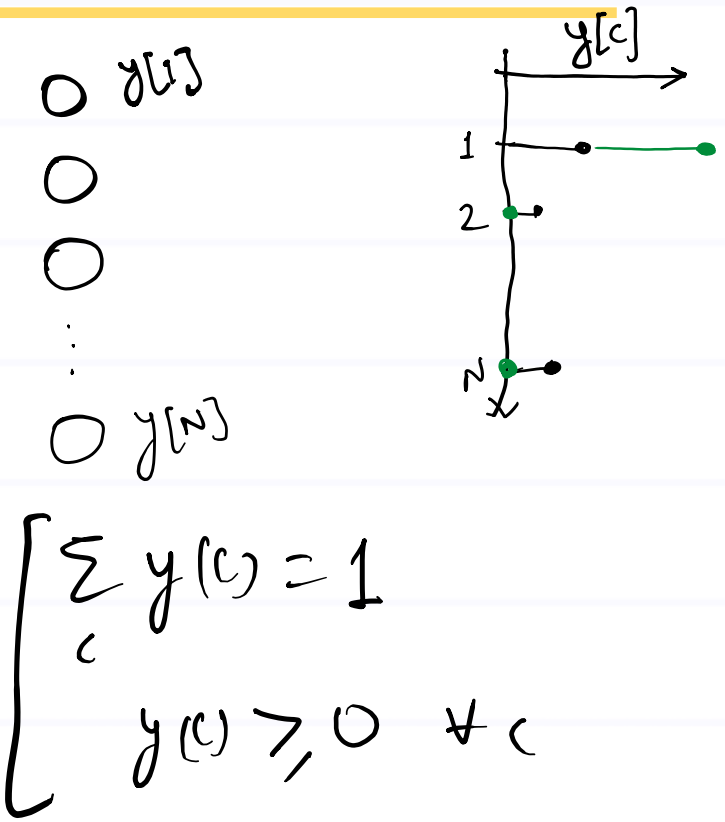
$$= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)}$$

$$= -\sum_{x,y} p(x,y) \log \left(\frac{p(x)\,p(y)}{p(x,y)} \right)$$

$$= KL\left[p(x,y) \,||\, p(x)\,p(y) \right]$$

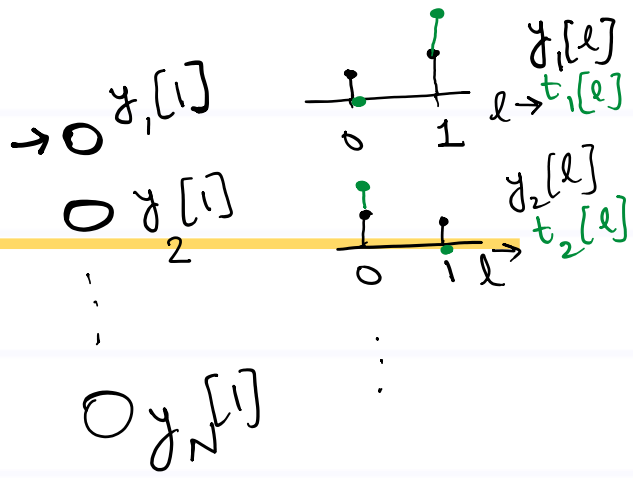
Loss Functions

- Multi class classification
  - The random variable is $c \in \{1, \dots, N\}$
  - The ground truth pdf is $t[c]$
  - The estimated pdf is $y[c]$
  - The metric to minimize is $H(t[c], y[c])$



$$\begin{cases} \sum_c y(c) = 1 \\ y(c) \geq 0 \quad \forall c \end{cases}$$

Loss Functions



- Multi-label classification
  - The random variables are $l_i \in \{0,1\}, \text{for } i \in \{1, \dots, N\}$
  - The ground truth pdf is $t_i[l_i]$
  - The estimated pdf is $y_i[l_i]$
  - The metric to minimize is $H(t_i[l_i], y_i[l_i])$
  - The output neuron may estimate $y_i[l_i = 1]$, the metric becomes
  - $H(t_i[l_i], y_i[l_i]) = -t_i[0] \log y_i[0] - t_i[1] \log y_i[1]$
  - Since, l_i can take only two values, $y_i[l_i = 0] = 1 - y_i[l_i = 1]$

EM Algorithm

Optimizing with Auxiliary Functions

Reference: PRML Section 9.4

Optimizing latent variable models

- We saw $p(X, Z|\theta)$ has a nice form, i.e., without summation. Hence, log likelihood is easy to maximize
- But Z is a latent variable, i.e., unobserved. So mostly we get $p(X|\theta)$ which we decompose as $\sum_Z p(X, Z|\theta)$. It is difficult to maximize log likelihood because of summation.
- Can we convert $\log p(X|\theta)$ into $\log p(X, Z|\theta)$?

Getting rid of summation inside log

- $\log p(X|\theta) = \log \sum_Z p(X, Z|\theta) \times \frac{q(Z)}{q(Z)}$

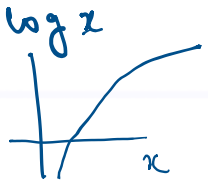
  v_i $\quad \alpha_i$

  or

- $\sum_i \log p(x_i|\theta) = \sum_i \log \sum_{z_i} p(x_i, z_i|\theta) \times \frac{q(z_i)}{q(z_i)} \times q(z_i) \geq \sum_i \sum_{z_i} q(z_i) \log \frac{p(x_i, z_i|\theta)}{q(z_i)}$

  v_i $\quad \alpha_i$

  $\mathcal{L}(q(Z), \theta)$



$\log \sum_i \alpha_i v_i \geq \sum_i \alpha_i \log v_i$

Jensen's Ineq.

$\sum_i \alpha_i = 1 \; ; \; \alpha_i \geq 0 \; \forall i$

Auxiliary Loss Function

- $\log p(X|\theta) \geq \sum_Z q(Z) \log \frac{p(X,Z|\theta)}{q(Z)} = \mathcal{L}(q(Z),\theta)$

Gap

- $\log p(X|\theta) - \mathcal{L}(q(Z), \theta) = \mathrm{KL}\left[q(Z)||p(Z|X,\theta)\right]$

q assigned

from data

$$\text{LHS} = \underbrace{\sum_z q(z)}_{=1} \cdot \log p(X|\theta) - \sum_z q(z) \log \frac{p(X,z|\theta)}{q(z)}$$

$$= \sum_z q(z) \log \underbrace{\frac{p(X|\theta)\cdot q(z)}{p(X,z|\theta)}}_{\frac{q(z)}{p(z|X,\theta)}}$$

$$= \text{RHS}$$

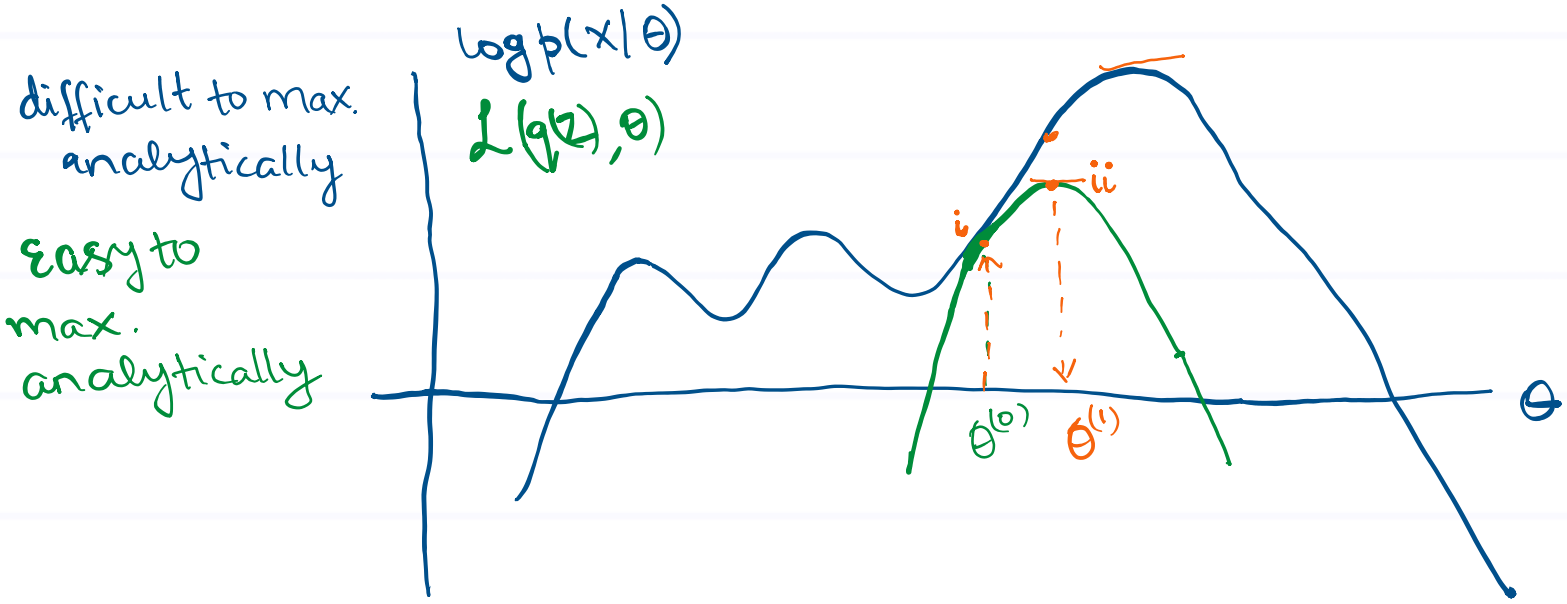
Auxiliary Loss Function

- $\log p(X|\theta) \geq \mathcal{L}(q(Z), \theta)$

max →

- $\log p(X|\theta) - \mathcal{L}(q(Z), \theta) = \mathrm{KL}\left[q(Z)||p(Z|X, \theta)\right]$

Equal when
$q(Z) = p(Z|X, \theta)$



i For given $\theta^{(0)}$, choose
$q(Z) = p(Z|X, \theta)$
This gives us $\mathcal{L}$

ii. For this $\mathcal{L}$, find $\max_{\theta} \mathcal{L}$ to get $\theta^{(1)}$.

EM Algorithm

- E-step
  - $q \leftarrow \text{argmin}_q \, \text{KL}\,[q(Z)||p(Z|X,\theta)]$ $q(z) = p(z|x,\theta)$
- M-step
  - $\theta \leftarrow \text{argmax}_\theta \, \mathbb{E}_q[\log p(X,Z|\theta)]$ analytically solvable