CS460/626 : Natural Language Processing/Speech, NLP and the Web

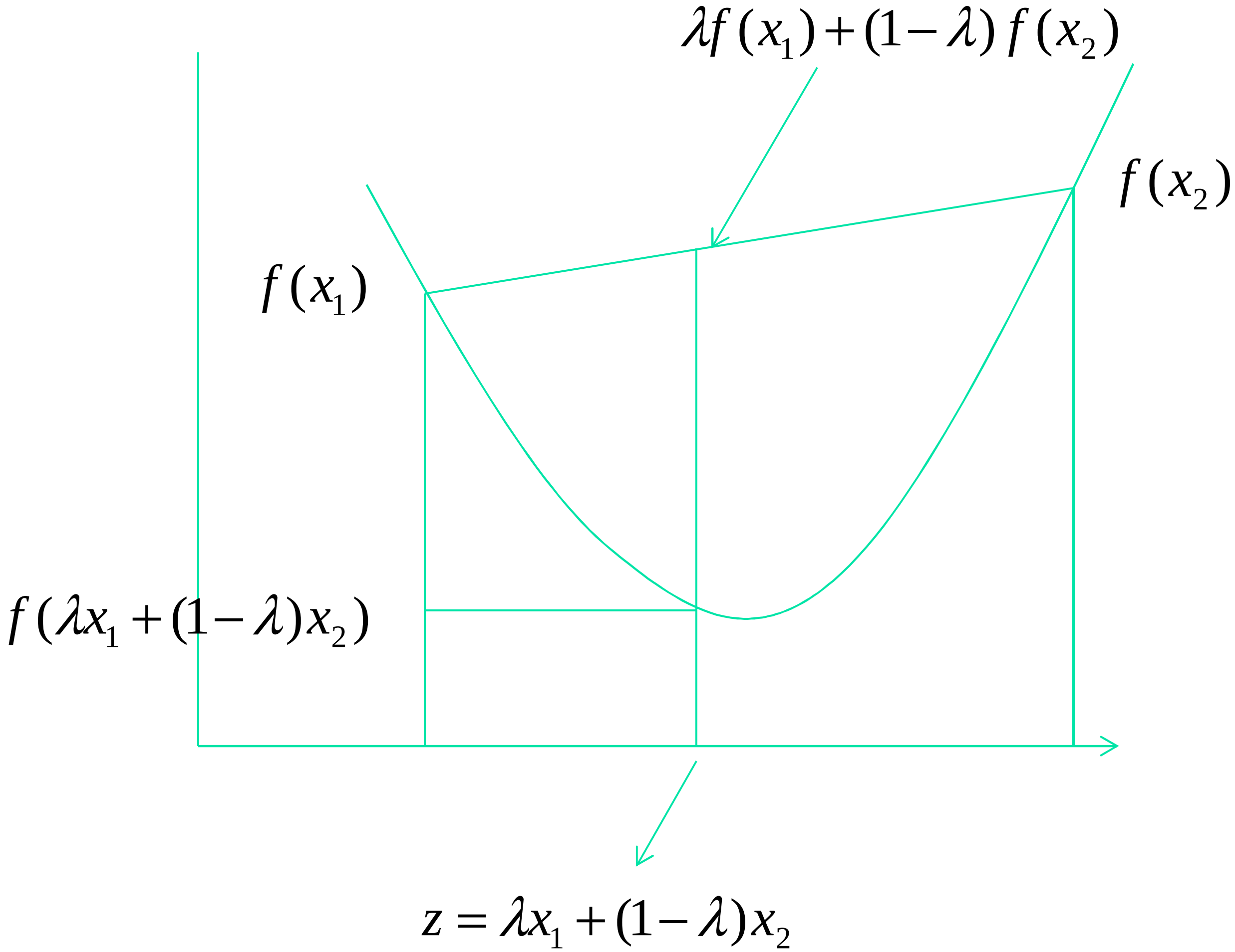
Lecture 31-32: Expectation Maximisation

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

5th and 6th Nov, 2012

Some Useful mathematical concepts

- Convex/ concave functions
- Jensen's inequality
- Kullback–Leibler distance/divergence



$\lambda f(x_1)+(1-\lambda)f(x_2)$
$f(x_2)$
$f(x_1)$
$f(\lambda x_1+(1-\lambda)x_2)$
$z=\lambda x_1+(1-\lambda)x_2$

Criteria for convexity

- A function f(x) is said to be convex in the interval [a,b] iff

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

$$x_1 < x_2$$

$$\forall x_1, x_2 \in [a,b]$$

Jensen's inequality

- For any convex function f(x)

$$f(\sum_{i=1}^{n}\lambda_i x_i) \le \sum_{i=1}^{n}\lambda_i f(x_i)$$

Where $\sum_{i=1}^{n}\lambda_i = 1$ and $\forall i, 0 \le \lambda_i \le 1$

Proof of Jensen´s inequality

- Method:- By induction on N
- Base case:-

$$N = 1$$

$$f(\lambda x) \leq \lambda f(x)$$

$$\sum \lambda_i = 1 \Rightarrow \lambda = 1$$

$$\therefore f(x) \leq f(x), \text{trivially true}$$

Another base case

- <u>N = 2</u>

$$f(\lambda_1 x_1 + \lambda_2 x_2)$$

$$= f(\lambda_1 x_1 + (1-\lambda_1) x_2) \qquad \text{since } \lambda_1 + \lambda_2 = 1$$

$$\leq \lambda_1 f(x_1) + (1-\lambda_1) f(x_2) \qquad \text{since f(x) is convex}$$

Hypothesis

Suppose true for $N = k$

i.e $f(\sum_{i=1}^{n} \lambda_i x_i) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$

Induction Step

Show that

$$f\left(\sum_{i=1}^{k+1} \lambda_i x_i\right) \leq \sum_{i=1}^{k+1} \lambda_i f(x_i)$$

given

$$f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leq \sum_{i=1}^{k} \lambda_i f(x_i)$$

$$\lambda_1 + \lambda_2 + \lambda_3 + \ldots\ldots + \lambda_k + \lambda_{k+1} = 1$$

Proof

$$f(\lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \ldots\ldots + \lambda_{k+1} x_{k+1})$$

$$= f((1-\lambda_{k+1})\sum_{i=1}^{k} \frac{\lambda_i x_i}{(1-\lambda_{k+1})} + \lambda_{k+1} x_{k+1})$$

$$\leq (1-\lambda_{k+1}) f(\sum_{i=1}^{k} \frac{\lambda_i x_i}{(1-\lambda_{k+1})}) + \lambda_{k+1} f(x_{k+1}) \quad \text{By convexity}$$

$$= (1-\lambda_{k+1}) f(\sum_{i=1}^{k} \mu_i x_i) + \lambda_{k+1} f(x_{k+1}) \qquad \text{where } \mu_i = \frac{\lambda_i}{(1-\lambda_{k+1})}$$

Continued...

- Examine each μ_i

$$\sum_{i=1}^{k} \mu_i = \mu_1 + \mu_2 + \mu_3 \ldots\ldots + \mu_k$$

$$= \frac{\lambda_1}{(1-\lambda_{k+1})} + \frac{\lambda_2}{(1-\lambda_{k+1})} + \frac{\lambda_3}{(1-\lambda_{k+1})} + \ldots\ldots + \frac{\lambda_k}{(1-\lambda_{k+1})}$$

$$= \frac{\lambda_1 + \lambda_2 + \lambda_3 + \ldots\ldots + \lambda_k}{(1-\lambda_{k+1})} = \frac{(1-\lambda_{k+1})}{(1-\lambda_{k+1})}$$

Continued…

- Therefore,

$$(1-\lambda_{k+1})\, f\,(\sum_{i=1}^{k} \mu_i x_i) + \lambda_{k+1}\, f\,(x_{k+1})$$

$$\leq (1-\lambda_{k+1})\sum_{i=1}^{k} \mu_i\, f\,(x_i) + \lambda_{k+1}\, f\,(x_{k+1})$$

$$= \sum_{i=1}^{k} \lambda_i\, f\,(x_i) + \lambda_{k+1}\, f\,(x_{k+1})$$

Finally at the induction step

$$f\,(\sum_{i=1}^{k+1} \lambda_i x_i) \leq \sum_{i=1}^{k+1} \lambda_i\, f\,(x_i)$$

Thus Jensen´s inequality is proved

KL -divergence

- We will do the discrete form of probability distribution.
- Given two probability distribution P,Q on the random variable
  - X : $x_1, x_2, x_3 \ldots x_N$
  - P:$p_1=p(x_1)$, $p_2=p(x_2)$, … $p_n=p(x_n)$
  - Q:$q_1=q(x_1)$, $q_2=q(x_2)$, … $q_n=q(x_n)$

KLD definition

$$\mathrm{KL}(\mathrm{P},\mathrm{Q}) = \mathrm{D} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{p}_\mathrm{i} \log \frac{\mathrm{p}_\mathrm{i}}{\mathrm{q}_\mathrm{i}} \qquad \sum \mathrm{p}_\mathrm{i} = 1, \sum \mathrm{q}_\mathrm{i} = 1$$

D is assymmetric and $\mathrm{D} \geq 0$

also written as

$$\mathrm{KL}(\mathrm{P},\mathrm{Q}) = \mathrm{D}$$

$$= E_p(\log P) - E_p(\log Q)$$

Proof: KLD>=0

$$\mathrm{KL}(P,Q) = \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i} \geq 0$$

Proof :-

$$\sum_{i=1}^{N} p_i \log \frac{p_i}{q_i} = \sum_{i=1}^{N} p_i \left(-\log \frac{q_i}{p_i} \right)$$

$$-\log x \text{ is convex in } [0, \infty]$$

$$\text{So } -\log\left(\sum_{i=1}^{N} p_i x_i \right) \leq \sum_{i=1}^{N} p_i (-\log x_i)$$

Proof cntd.

- Apply Jensen's inequality

$$\text{So}\quad -\log\left(\sum_{i=1}^{N} p_i \frac{q_i}{p_i}\right) \le \sum_{i=1}^{N} p_i (-\log \frac{q_i}{p_i})$$

$$\Rightarrow -\log\left(\sum_{i=1}^{N} q_i\right) \le \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i}$$

$$\Rightarrow \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i} \ge 0 \qquad\qquad \sum_{i=1}^{N} q_i = 1$$

Convexity of –log x

$$-\log(\lambda x_1 + (1-\lambda)x_2) \le \lambda(-\log x_1) + (1-\lambda)(-\log x_2)$$

i.e.

$$\log(\lambda x_1 + (1-\lambda)x_2) \ge \lambda \log x_1 + (1-\lambda)\log x_2$$

$$\Rightarrow \lambda x_1 + (1-\lambda)x_2 \ge x_1{}^{\lambda} x_2{}^{1-\lambda}$$

$$\Rightarrow \lambda\left(\frac{x_1}{x_2}\right)^{1-\lambda} + (1-\lambda)\frac{x_2{}^{1-1-\lambda}}{x_1{}^{\lambda}} \ge 1$$

$$\Rightarrow \lambda\left(\frac{x_1}{x_2}\right)^{1-\lambda} + (1-\lambda)\left(\frac{x_2}{x_1}\right)^{\lambda} \ge 1$$

$$\Rightarrow \lambda y^{1-\lambda} + \frac{(1-\lambda)}{y^{\lambda}} \ge 1 \qquad y = \frac{x_1}{x_2} \le 1$$

Interesting problem

- Try to prove:-

$$\frac{w_1 x_1 + w_2 x_2}{w_1 + w_2} \geq \sqrt[w_1+w_2]{x_1^{\,w_1} x_2^{\,w_2}}$$

2nd definition of convexity

- Theorem:

If $f(x)$ is twice differentiable in $[a,b]$ and

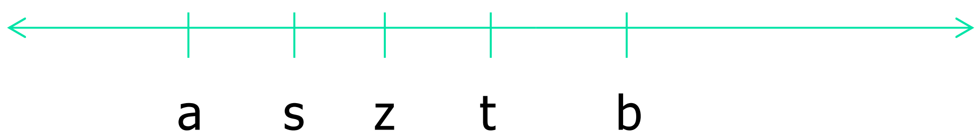
$f''(x) \geq 0 \; \forall \; x \in [a,b]$, then f(x) is convex in $[a,b]$.

So - log x is convex .

Lemma 1

If $f^{''}(x) \geq 0 \text{ in } [a,b]$

then $f^{'}(t) > f^{'}(s), \quad \forall\, s,t \qquad \text{t} > \text{s and t}, \text{s} \in [a,b]$



Mean Value Theorem

$$f(z) - f(a) = (z-a) f^{'}(s) \quad \exists s \in (\mathrm{z}, \mathrm{a})$$

For any function $f(x)$

$$f(n) - f(m) = (n-m) f^{'}(p) \quad \text{where } \mathrm{m} \leq \mathrm{p} \leq \mathrm{n}$$

Alternative form of z

$$z = \lambda x_1 + (1-\lambda)x_2$$

Add –λz to both sides

$$(1-\lambda)z = \lambda(x_1 - z) + (1-\lambda)x_2$$

$$(1-\lambda)(x_2 - z) = \lambda(z - x_1)$$

Alternative form of convexity

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda)f(x_2)$$

Add –λf(z) to both sides

$$\Rightarrow f(z) - \lambda f(z) \le \lambda f(x_1) + (1-\lambda)f(x_2) - \lambda f(z)$$

$$\Rightarrow (1-\lambda)f(z) \le \lambda\left(f(x_1) - f(z)\right) + (1-\lambda)f(x_2)$$

$$\Rightarrow (1-\lambda)f(z) \le \lambda\left(f(x_1) - f(z)\right) + (1-\lambda)f(x_2)$$

Proof: second derivative >=0 implies convexity (1/2)

We have that

$$z \triangleq \lambda x_1 + (1-\lambda)x_2$$

$$f(z) \triangleq \lambda f(x_1) + (1-\lambda)f(x_2)$$

$$(1-\lambda)[f(x_2) - f(z)] \geq \lambda[f(z) - f(x_1)] \qquad (1)$$

$$(1-\lambda)[x_2 - z] = \lambda(z - x_1) \qquad (2)$$

Second derivative >=0 implies convexity (2/2)

(2) Is equivalent to

$$(1-\lambda)f'(t).(x_2-\lambda) \geq \lambda f'(s)(z-x_1)$$

For some *s* and *t*, where

$$x_1 < s < z < t < x_2$$

Now since *f''(x) >=0*

$$f'(t) > f'(s)$$

Combining this with (1), the result is proved

Why all this

- In EM, we maximize the *expectation* of log likelihood of the data
- Log is a concave function
- We have to take iterative steps to get to the maximum
- There are two unknown values: Z (unobserved data) and θ (parameters)
- From θ, get new value of Z (E-step)
- From Z, get new value of θ (M-step)

How to change θ

- How to choose the next θ?
- Take

$$argmax_{\theta}(LL(X,Z:\theta) - LL(X,Z:\theta_n))$$

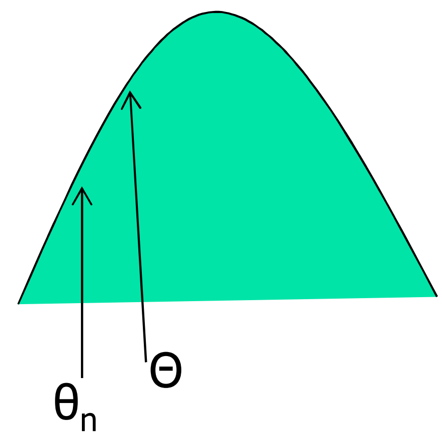
Where,

X: observed data

Z: unobserved data

Θ: parameter

$LL(X,Z:\theta_n)$: log likelihood of complete data with parameter value at θ_n



At every step *LL(.)* will **Increase,** ultimately reaching local/global maximum

This is in lieu of, for example, gradient ascent

Why expectation of log likelihood? (1/2)

- $P(X:\theta)$ may not be a convenient mathematical expression
- Deal with $P(X,Z:\theta)$, marginalized over Z
- $Log(\Sigma_Z P(X,Z:\theta))$ is mathematically processed with multiplying by $P(Z|X: \theta_n)$ which for each Z is between 0 and 1 and sums to 1
- Then Jensen inequality will give

$$Log(\Sigma_Z P(X,Z:\theta))$$

$$>= Log(\Sigma_Z P(Z|X: \theta_n)P(X,Z:\theta)/P(Z|X: \theta_n))$$

$$= \Sigma_Z P(Z|X: \theta_n)Log(P(X,Z:\theta)/P(Z|X: \theta_n))$$

Why expectation of log likelihood? (2/2)

$LL(X:\theta) - LL(X:\theta_n)$

$= Log(\Sigma_Z P(X,Z:\theta)) - Log(P(X:\theta_n))$

$>= Log(\Sigma_Z P(Z|X: \theta_n) P(X,Z:\theta)/P(Z|X: \theta_n)) - Log(P(X:\theta_n))$

$= \Sigma_Z P(Z|X: \theta_n) Log(P(X,Z:\theta)/(P(Z|X: \theta_n) . P(X:\theta_n))$

since $\Sigma_Z P(Z|X: \theta_n)=1$

$= \Sigma_Z P(Z|X: \theta_n) Log((P(X,Z:\theta)/(P(X,Z:\theta_n))$

So, $argmax_\theta (LL(X:\theta) - LL(X:\theta_n))$

$= \Sigma_Z P(Z|X: \theta_n) Log(P(X,Z:\theta)$

$= E_Z(Log(P(X,Z:\theta))$, where $E_Z(.)$ is the expectation of log likelihood of complete data wrt Z

Why expectation of *Z*?

- If the log likelihood is a linear function of Z, then the expectation can be carries inside of the log likelihood and E(Z) is computed
- The above is true when the hidden variables form a mixture of distributions (e..g, in tosses of two coins), and
- Each distribution is an exponential distribution like multinomial/normal/poisson