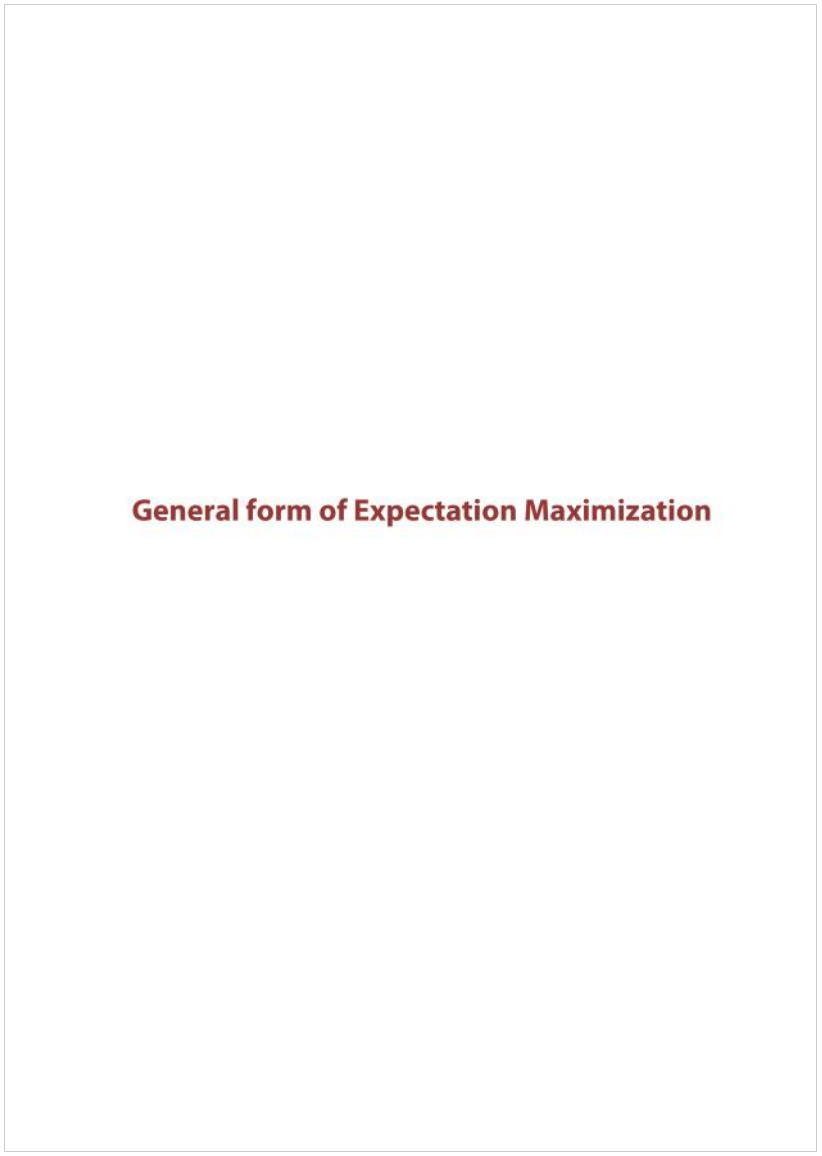
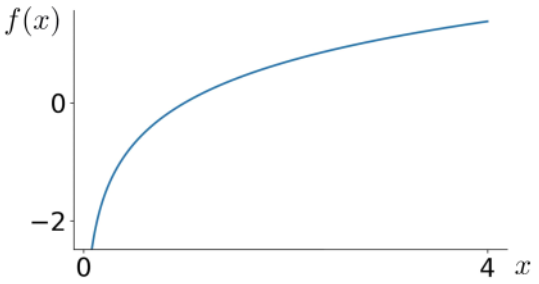
EM algorithm

Saturday, June 23, 2018 4:55 PM

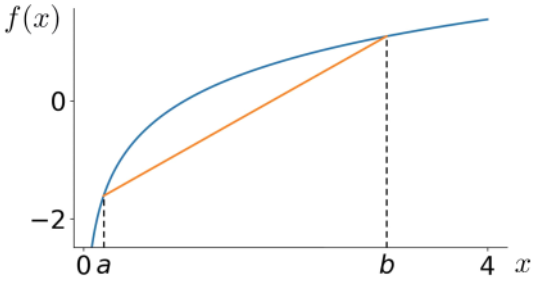
em1



Concave functions

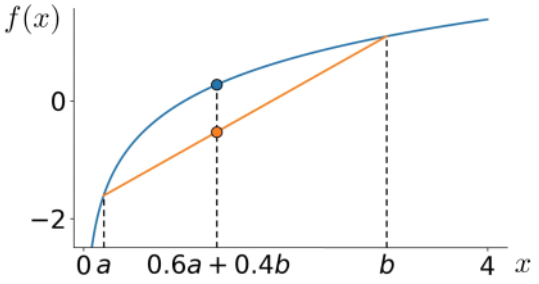


Concave functions



Concave functions

Concave function



Def.: $f(x)$ is concave if

for any a, b, α : $f(\alpha a + (1-\alpha)b) \geq \alpha f(a) + (1-\alpha)f(b)$

$$0 \leq \alpha \leq 1$$

Jensen's inequality

α - probability

$$f(\alpha a + (1-\alpha)b) \geq \alpha f(a) + (1-\alpha) f(b)$$

Jensen's inequality

If $f(\alpha a + (1-\alpha)b) \geq \alpha f(a) + (1-\alpha)f(b)$

Then $\alpha_1 + \alpha_2 + \alpha_3 = 1;\ \alpha_k \geq 0.$

$$f(\alpha_1 a_1 + \alpha_2 a_2 + \alpha_3 a_3) \geq \alpha_1 f(a_1) + \alpha_2 f(a_2) + \alpha_3 f(a_3)$$

Jensen's inequality

If $f(\alpha a + (1-\alpha)b) \geq \alpha f(a) + (1-\alpha)f(b)$

Then $\alpha_1 + \alpha_2 + \alpha_3 = 1;\ \alpha_k \geq 0.$

$$f(\alpha_1 a_1 + \alpha_2 a_2 + \alpha_3 a_3) \geq \alpha_1 f(a_1) + \alpha_2 f(a_2) + \alpha_3 f(a_3)$$

$$\begin{aligned} p(t = a_1) &= \alpha_1, \\ p(t = a_2) &= \alpha_2, \\ p(t = a_3) &= \alpha_3 \end{aligned}$$

Jensen's inequality

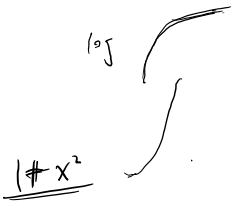
If $f(\alpha a + (1-\alpha)b) \geq \alpha f(a) + (1-\alpha)f(b)$

Then $\alpha_1 + \alpha_2 + \alpha_3 = 1;\ \alpha_k \geq 0.$

$$f(\underbrace{\alpha_1 a_1 + \alpha_2 a_2 + \alpha_3 a_3}_{\mathbb{E}_{p(t)} t}) \geq \underbrace{\alpha_1 f(a_1) + \alpha_2 f(a_2) + \alpha_3 f(a_3)}_{\mathbb{E}_{p(t)} f(t)}$$

$$\begin{aligned} p(t = a_1) &= \alpha_1, \\ p(t = a_2) &= \alpha_2, \\ p(t = a_3) &= \alpha_3 \end{aligned}$$

Jensen's inequality



If $f(\alpha a + (1-\alpha)b) \geq \alpha f(a) + (1-\alpha)f(b)$

Then Jensen's inequality:

$$f\left(\mathbb{E}_{p(t)} t\right) \geq \mathbb{E}_{p(t)} f(t)$$

Infinitely many points

Random variable x follows standard Gaussian distribution: $x \sim \mathcal{N}(0, 1)$. A random variable y is a deterministic function of x: $y = 1 + x^2$. Apply Jensen's inequality to $\log(y)$ (logarithm is a concave function of its argument y) and choose all the correct inequalities from the options below:

- [ ] $\mathbb{E}\log(1+x^2) \geq \log(\mathbb{E}(1+x^2))$

  Un-selected is correct

- [ ] Jensen's inequality is not applicable, since $\log(1+x^2)$ is not a concave function of x

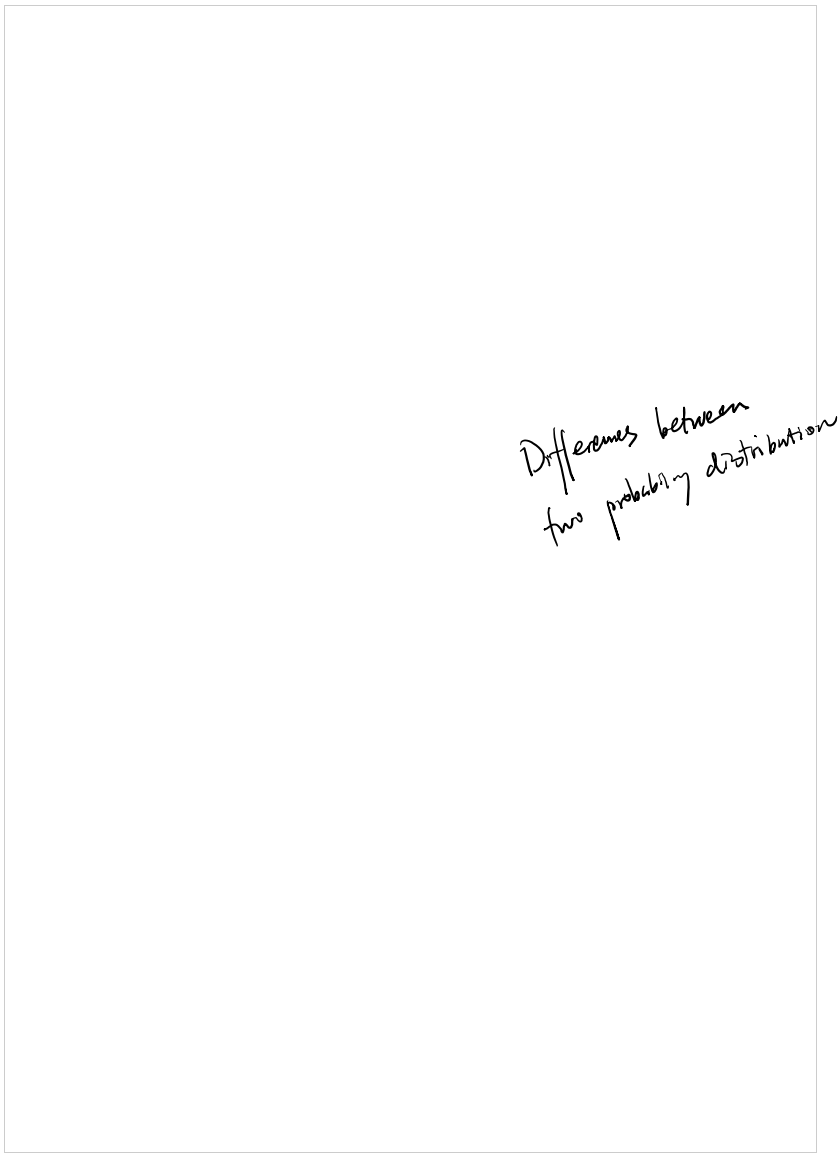
  Un-selected is correct

- [x] $\mathbb{E}\log(1+x^2) \leq \log(\mathbb{E}(1+x^2))$

  Correct
  Here the order is changed, but the inequality sign is reversed as well, so it doesn't change anything. Also, instead of y we write $1+x^2$, but they are identical.

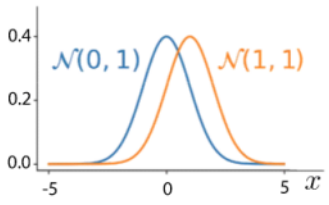
- [x] $\log(\mathbb{E}y) \geq \mathbb{E}\log(y)$

Kullback–Leibler divergence

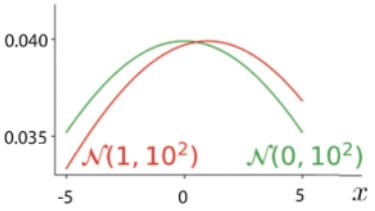

Differences between
two probability distributions

Kullback–Leibler divergence

Parameters difference: 1



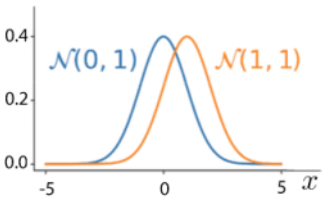
Parameters difference: 1



Kullback–Leibler divergence

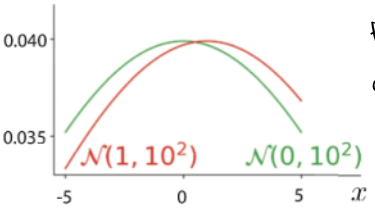
Parameters difference: 1

$$\mathcal{KL}(q_1 \parallel p_1) = 0.5$$



Parameters difference: 1

$$\mathcal{KL}(q_2 \parallel p_2) = 0.005$$

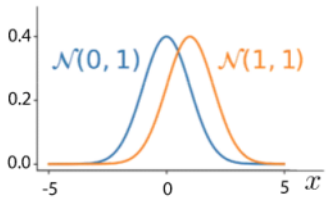


the second set is closer to each other with this measure

Kullback–Leibler divergence

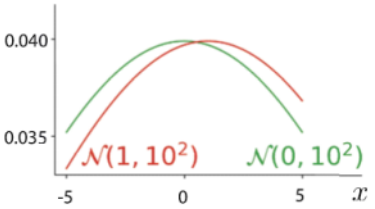
Parameters difference: 1

$\mathcal{KL}(q_1 \parallel p_1) = 0.5$



Parameters difference: 1

$\mathcal{KL}(q_2 \parallel p_2) = 0.005$



$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

the expected value of the difference between the two distribution

Property: It's non-symmetric. if change p, q, things will be different

Kullback–Leibler divergence

$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

Kullback–Leibler divergence

$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

1. $\mathcal{KL}(q \parallel p) \neq \mathcal{KL}(p \parallel q)$ asymmetric

Kullback–Leibler divergence

$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

1. $\mathcal{KL}(q \parallel p) \neq \mathcal{KL}(p \parallel q)$
2. $\mathcal{KL}(q \parallel q) = 0$ $= \int q(x) \log 1 \, dx = 0$

Kullback–Leibler divergence

$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

1. $\mathcal{KL}(q \parallel p) \neq \mathcal{KL}(p \parallel q)$
2. $\mathcal{KL}(q \parallel q) = 0$
3. $\mathcal{KL}(q \parallel p) \geq 0$

Kullback–Leibler divergence

$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

1. $\mathcal{KL}(q \parallel p) \neq \mathcal{KL}(p \parallel q)$
2. $\mathcal{KL}(q \parallel q) = 0$
3. $\mathcal{KL}(q \parallel p) \geq 0$

Proof: $-\mathcal{KL}(q \parallel p) = \mathbb{E}_q \left(-\log \frac{q}{p} \right)$

Kullback–Leibler divergence

$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

1. $\mathcal{KL}(q \parallel p) \neq \mathcal{KL}(p \parallel q)$
2. $\mathcal{KL}(q \parallel q) = 0$
3. $\mathcal{KL}(q \parallel p) \geq 0$

Proof: $-\mathcal{KL}(q \parallel p) = \mathbb{E}_q \left(-\log \frac{q}{p} \right) = \mathbb{E}_q \left(\log \frac{p}{q} \right)$

Kullback–Leibler divergence

$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

1. $\mathcal{KL}(q \parallel p) \neq \mathcal{KL}(p \parallel q)$
2. $\mathcal{KL}(q \parallel q) = 0$
3. $\mathcal{KL}(q \parallel p) \geq 0$

Proof: $-\mathcal{KL}(q \parallel p) = \mathbb{E}_q \left(-\log \frac{q}{p} \right) = \mathbb{E}_q \left(\log \frac{p}{q} \right)$

$\leq \log(\mathbb{E}_q \frac{p}{q})$

Jensen Inequality.

concave function

Kullback–Leibler divergence

$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

1. $\mathcal{KL}(q \parallel p) \neq \mathcal{KL}(p \parallel q)$
2. $\mathcal{KL}(q \parallel q) = 0$
3. $\mathcal{KL}(q \parallel p) \geq 0$

Proof: $-\mathcal{KL}(q \parallel p) = \mathbb{E}_q \left(-\log \frac{q}{p} \right) = \mathbb{E}_q \left(\log \frac{p}{q} \right)$

$$\leq \log(\mathbb{E}_q \frac{p}{q}) = \log \int q(x) \frac{p(x)}{q(x)} dx = 0$$

$q(x) = N(\mu_1, \sigma_1)$ $\quad p(x) = N(\mu_2, \sigma_2)$

$$KL(q \| p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

$$= \int \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} \log \frac{\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\}}{\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left\{\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}} dx$$

$$= \int \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} \log\left[\frac{\sigma_2}{\sigma_1} \cdot \exp\frac{(x-\mu_1)^2(\sigma_2^2-\sigma_1^2)}{2\sigma_1^2\sigma_2^2}\right] dx.$$

$$= \int \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} \left[\log\frac{\sigma_2}{\sigma_1} + \frac{(x-\mu_1)^2(\sigma_2+\sigma_1)(\sigma_2-\sigma_1)}{2\sigma_1^2\sigma_2^2}\right] dx$$

$$= \log\frac{\sigma_2}{\sigma_1} \cdot 1 + \int \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} (x-\mu_1)^2 dx \times \frac{\sigma_2^2-\sigma_1^2}{2\sigma_1^2\sigma_2^2}$$

(integral = variance $= \sigma_1^2$)

$$= \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2(\sigma_2^2-\sigma_1^2)}{2\sigma_1^2\sigma_2^2} = \log\frac{\sigma_2}{\sigma_1} + \frac{1}{2} - \frac{\sigma_1^2}{2\sigma_2^2}$$

Kullback–Leibler divergence

$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

Summary

A way to compare distributions
not a proper distance

1. $\mathcal{KL}(q \parallel p) \neq \mathcal{KL}(p \parallel q)$
2. $\mathcal{KL}(q \parallel q) = 0$
3. $\mathcal{KL}(q \parallel p) \geq 0$

em2

General form of Expectation Maximization

General form of Expectation Maximization

General form of Expectation Maximization

$$p(x_i \mid \theta) = \sum_{c=1}^{3} p(x_i \mid t_i = c, \theta)\, p(t_i = c \mid \theta)$$

General form of Expectation Maximization

$$\max_{\theta} \quad p(X \mid \theta)$$

General form of Expectation Maximization

$$\max_{\theta}. \log p(X \mid \theta)$$

General form of Expectation Maximization

$$\max_{\theta}. \log p(X \mid \theta) = \log \prod_{i=1}^{N} p(x_i \mid \theta)$$

General form of Expectation Maximization

$$\max_{\theta}. \log p(X \mid \theta) = \log \prod_{i=1}^{N} p(x_i \mid \theta)$$
$$= \sum_{i=1}^{N} \log p(x_i \mid \theta)$$

General form of Expectation Maximization

$$\log p(X \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$

General form of Expectation Maximization

$$\log p(X \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$
$$= \sum_{i=1}^{N} \log \sum_{c=1}^{3} p(x_i, t_i = c \mid \theta)$$

General form of Expectation Maximization

$$\log p(X \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$

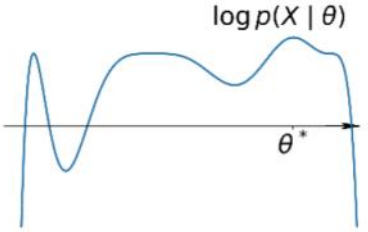
$$= \sum_{i=1}^{N} \log \sum_{c=1}^{3} p(x_i, t_i = c \mid \theta) \geq \mathcal{L}(\theta)$$

Jensen's inequality

General form of Expectation Maximization

$$\log p(X \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$

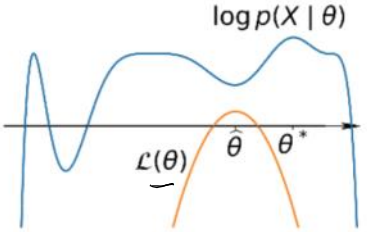
$$= \sum_{i=1}^{N} \log \sum_{c=1}^{3} p(x_i, t_i = c \mid \theta) \geq \mathcal{L}(\theta)$$



General form of Expectation Maximization

$$\log p(X \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$

$$= \sum_{i=1}^{N} \log \sum_{c=1}^{3} p(x_i, t_i = c \mid \theta) \geq \mathcal{L}(\theta)$$



instead of one lower bound
⇒ choose a family of lower bound.

General form of Expectation Maximization

$$\log p(X \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$
$$= \sum_{i=1}^{N} \log \sum_{c=1}^{3} p(x_i, t_i = c \mid \theta)$$

General form of Expectation Maximization

$$\log p(X \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$
$$= \sum_{i=1}^{N} \log \sum_{c=1}^{3} \frac{q(t_i = c)}{q(t_i = c)} p(x_i, t_i = c \mid \theta)$$

General form of Expectation Maximization

$$\log p(X \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$

$$= \sum_{i=1}^{N} \log \sum_{c=1}^{3} \frac{q(t_i = c)}{q(t_i = c)} p(x_i, t_i = c \mid \theta)$$

treat q as weights of Jensen inequality

Jensen's inequality

$$\log \left(\sum_c \alpha_c v_c \right) \geq \sum_c \alpha_c \log(v_c)$$

General form of Expectation Maximization

$$\log p(X \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$

$$= \sum_{i=1}^{N} \log \sum_{c=1}^{3} \frac{q(t_i = c)}{q(t_i = c)} p(x_i, t_i = c \mid \theta)$$

$$\geq \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i = c) \log \frac{p(x_i, t_i = c \mid \theta)}{q(t_i = c)}$$

Jensen's inequality

$$\log \left(\sum_c \alpha_c v_c \right) \geq \sum_c \alpha_c \log(v_c)$$

weights

General form of Expectation Maximization

$$\log p(X \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$

$$= \sum_{i=1}^{N} \log \sum_{c=1}^{3} \frac{q(t_i = c)}{q(t_i = c)} p(x_i, t_i = c \mid \theta)$$

weights

$$\geq \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i = c) \log \frac{p(x_i, t_i = c \mid \theta)}{q(t_i = c)}$$

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

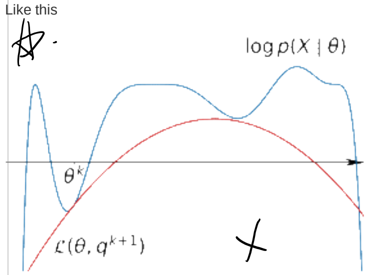
let

这个L是啥意思

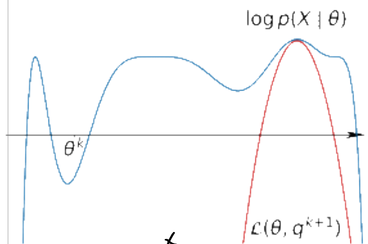
How do you think the optimal lower bound L(\theta, q^{k+1}) $L(\theta, q^{k+1})$ will look like? Assume that qq is the solution of the following optimization problem:

\max_q \mathcal{L}(\theta^{k}, q) $\max_q L(\theta^k, q)$

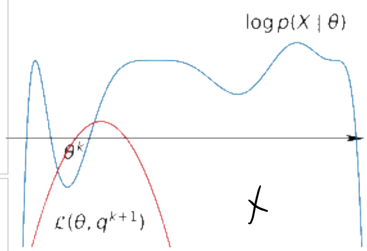
Like this



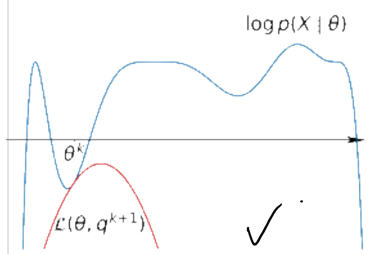
Like this



Like this



Like this



Correct

From <https://www.coursera.org/learn/bayesian-methods-in-machine-learning/lecture/Fm3mY/expectation-maximization-algorithm>

$\log \sum q(t_i = c) \frac{p}{q}$

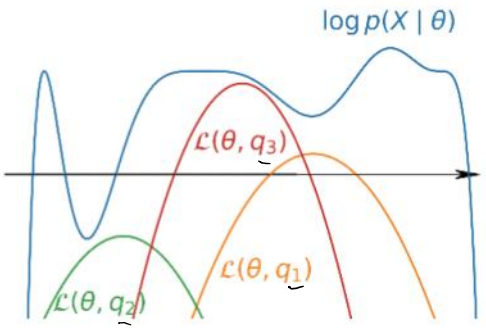
$\log q_1 \cdot \frac{p_1}{q_1} + \log q_2 \cdot \frac{p_2}{q_2} + \log q_3 \cdot \frac{p_3}{q_3}$

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

let

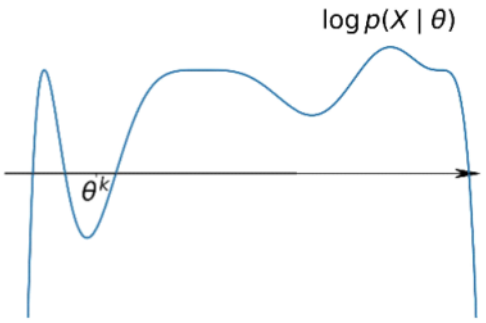
General form of Expectation Maximization

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$



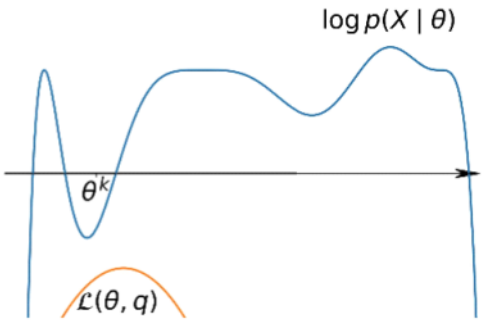
General form of Expectation Maximization

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$



General form of Expectation Maximization

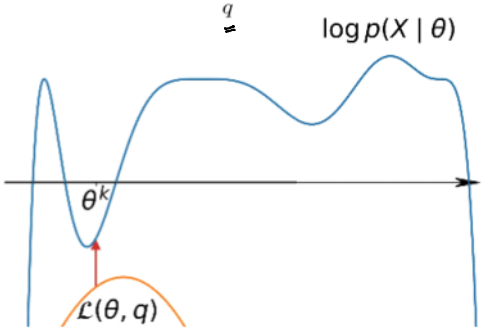
$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$



General form of Expectation Maximization

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$

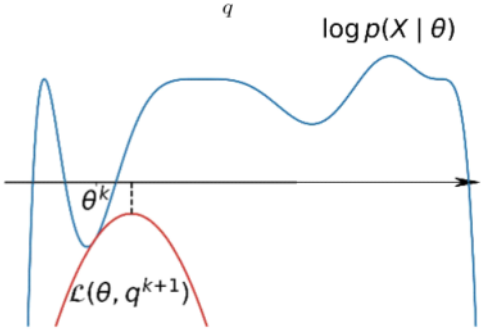
$$q^{k+1} = \arg\max_{q} \mathcal{L}(\theta^k, q)$$



General form of Expectation Maximization

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$

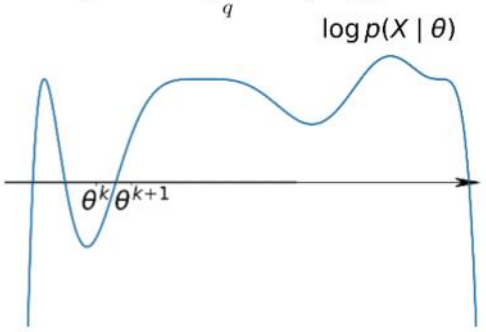
$$q^{k+1} = \arg\max_{q} \mathcal{L}(\theta^k, q)$$



General form of Expectation Maximization

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$

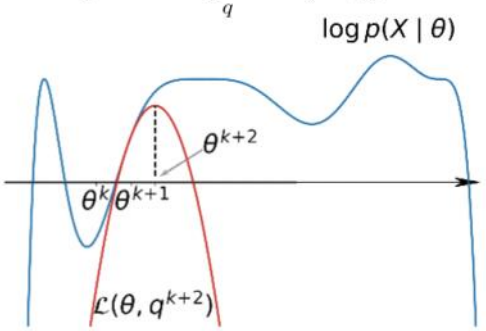
$$q^{k+1} = \arg\max_{q} \mathcal{L}(\theta^k, q)$$



General form of Expectation Maximization

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$

$$q^{k+1} = \arg\max_{q} \mathcal{L}(\theta^k, q)$$



↓. variational lower bound

Summary of Expectation Maximization

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$

Variational lower bound

E-step

$$q^{k+1} = \arg\max_q \mathcal{L}(\theta^k, q)$$

fix θ, find q (distribution)

M-step

$$\theta^{k+1} = \arg\max_\theta \mathcal{L}(\theta, q^{k+1})$$

fix q, find θ

em-e

E-step details

$$\text{GAP} = \log p(X|\theta) - \mathcal{L}(\theta, q) \quad \max_q$$

$$\sum_{i=1}^{N} KL(q(t_i) \| p(t_i = c | x_i, \theta))$$

Want $\min_q$ GAP

$$\text{Gap} = \sum_{i=1}^{N} \log p(x_i|\theta) - \sum_{i=1}^{N}\sum_{c=1}^{3} q(t_i=c) \log \frac{p(x_i, t_i=c|\theta)}{p(t_i=c)}$$

$$= \sum_{i=1}^{N} \left(\log p(x_i|\theta) \cdot \sum_{c=1}^{3} q(t_i=c) \right) - \sum_{c=1}^{3} q(t_i=3) \log \frac{p(x_i, t_i=c|\theta)}{q(t_i=c)}$$

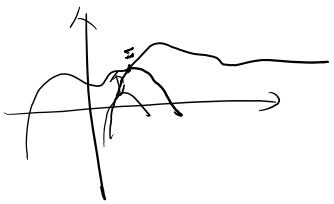
$$= \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i=c) \left[\log p(x_i|\theta) - \log \frac{p(x_i, t_i=c|\theta)}{q(t_i=c)} \right]$$

$$= \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i=c) \log \frac{p(x_i|\theta)\, q(t_i=c)}{p(x_i, t_i=c|\theta)}$$

want min GAP
q

minimize the sum of KL divergence $\Rightarrow$ $\min_q \sum_{i=1}^{N} KL(q(t_i) \| p(t_i = c | x_i, \theta))$

use property of KL $\Rightarrow$ want $q(t_i) = p(t_i | x_i, \theta)$



$$= \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i = c) \log \frac{p(x_i|\theta)\, q(t_i = c)}{p(x_i, t_i = c | \theta)} \geq 0$$

$$p(t_i = c | x_i, \theta)\, p(x_i | \theta)$$

$$= \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i = c) \log \frac{q(t_i = c)}{p(t_i = c | x_i, \theta)}$$

$$KL(q(t_i) \| p(t_i = c | x_i, \theta))$$

E-step details

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q)$$

E-step details

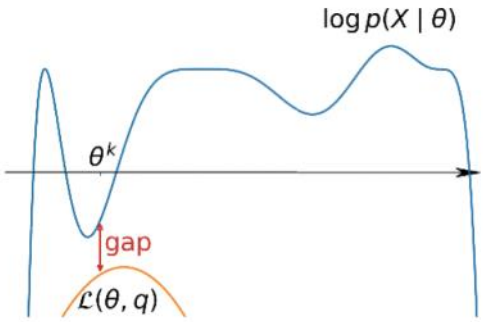
$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q)$$

E-step: $\max_q \mathcal{L}(\theta^k, q)$

E-step details

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q)$$

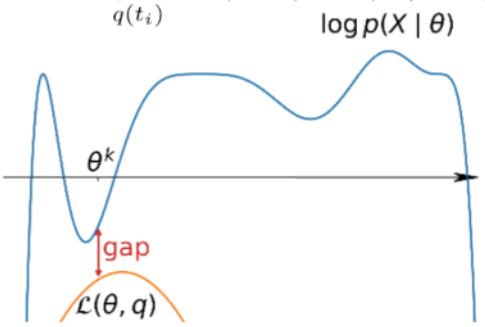
E-step: $\max_q \mathcal{L}(\theta^k, q)$



E-step summary

$$\log p(X \mid \theta) - \mathcal{L}(\theta, q) = \sum_i \mathcal{KL}\left(q(t_i) \parallel p(t_i \mid x_i, \theta)\right)$$

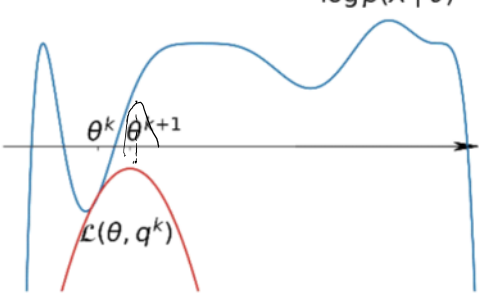
E-step: $\arg\max_{q(t_i)} \mathcal{L}(\theta^k, q) = p(t_i \mid x_i, \theta)$



E-step summary

$$\log p(X \mid \theta) - \mathcal{L}(\theta, q) = \sum_i \mathcal{KL}\left(q(t_i) \parallel p(t_i \mid x_i, \theta)\right)$$

E-step: $\arg\max_{q(t_i)} \mathcal{L}(\theta^k, q) = p(t_i \mid x_i, \theta)$



M step

$$\mathcal{L}(\theta, q) = \sum_i \sum_c q(t_i = c) \left(\log \frac{p(x_i, t_i = c \mid \theta)}{p(t_i = c)}\right)$$

$$= \sum_i \sum_c q(t_i = c) \log p(x_i, t_i = c \mid \theta)$$

$$- \sum_i \sum_c q(t_i = c) \log q(t_i = c)$$

$$= \mathbb{E}_q \log p(X, T \mid \theta) + \text{constant}$$ ← No. θ.

concave function wrt. θ. easy to optimize

E-step:

$$q^{k+1} = \arg\min_q KL[q(T) \,\|\, p(T \mid X, \theta^k)]$$

$$\Rightarrow q^{k+1}(t_i) = p(t_i \mid x_i, \theta^k)$$

$q^{k+1} = \arg\max_q \mathcal{L}(\theta^k, q)$

$\Leftrightarrow q^{k+1}(t_i) = p(t_i | x_i, \theta^k)$

M-step

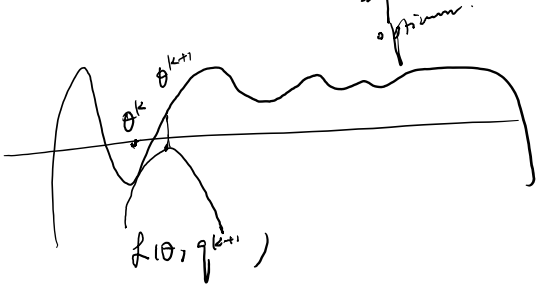
$\theta^{k+1} = \arg\max_\theta \mathbb{E}_{q^{k+1}} \log p(X, T | \theta)$

Convergence guarantees. 收敛保证

converge 到 local optimum.

$\star$ $\log p(X | \theta^{k+1}) \geq \mathcal{L}(\theta^{k+1}, q^{k+1}) \geq \mathcal{L}(\theta^k, q^{k+1})$

$= \log p(X | \theta^k)$



Marginal likely never decreases in EM algorithm — never make things worse!

Always improves

A debugging tool. If decreases $\Rightarrow$ wrong.

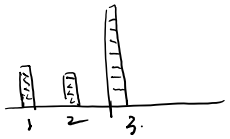
Can we compare value of $\mathcal{L}(\theta^k, q^{k-1})$ with value of log-likelihood $\log p(X \mid \theta^k)$ in the general case?

- $\mathcal{L}(\theta^k, q^{k-1}) \geq \log p(X \mid \theta^k)$
- (•) $\mathcal{L}(\theta^k, q^{k-1}) = \log p(X \mid \theta^k)$

Correct
Yes, since for any q the following inequality holds true $\mathcal{L}(\theta^k, q) \leq \log p(X \mid \theta^k)$, but for q^{k-1} is the maximizer of the lower bound $\mathcal{L}(\theta^k, q^{k-1}) = \max_q \mathcal{L}(\theta^k, q)$ and in the best case it became as large as $\log p(X \mid \theta^k)$

- $\mathcal{L}(\theta^k, q^{k-1}) \leq \log p(X \mid \theta^k)$
- No, the inequality sign depends on the particular model and the data at hand

EM example E-step



data: discrete distribution, 1, 2, 3

know: the data is generated from a mixture of distributions

$p(x_i) = \gamma \cdot p_1(x_i) + (1-\gamma)\, p_2(x_i)$

	1	2	3
p_1	α	$1-\alpha$	0
p_2	0	$1-\beta$	β

Task: estimate α and β. with the EM algorithm

Initiate: $\alpha_0 = \beta_0 = \gamma_0 = 0.5$

$x_i \leftarrow t_i$ Define latent variable.

$p(t_i = 1) = \gamma$, $p(x_i | t_i = 2) = p_2(x_i)$

$(x_i) \longleftarrow (t_i)$ Define latent variable

$p(t_i = 1) = \gamma$. $\quad p(x_i \mid t_i = 2) = P_2(x_i)$

E-step: Task: find the posterior distribution of $q(t_i)$

$q(t_i = c) = p(t_i = c \mid x_i)$

$$p(t_i = 1 \mid x_i = 1) = \frac{p(x_i = 1 \mid t_i = 1) \times p(t_i = 1)}{p(x_i = 1 \mid t_i = 1)\, p(t_i = 1) + p(x_i = 1 \mid t_i = 2)\, p(t_i = 2)}$$

$$= \frac{\alpha\gamma}{\alpha\gamma + 0 \cdot (1-\gamma)} = 1$$

if observe $x_i = 1$, it's certainly generated by $t_i = 1$

$$p(t_i = 1 \mid x_i = 2) = \frac{p(x_i = 2 \mid t_i = 1) \times p(t_i = 1)}{p(x_i = 2 \mid t_i = 1)\, p(t_i = 1) + p(x_i = 2 \mid t_i = 2)\, p(t_i = 2)}$$

$$= \frac{(1-\alpha)\gamma}{(1-\alpha)\gamma + (1-\beta)(1-\gamma)} = \frac{0.5 \times 0.5}{0.5 \times 0.5 + 0.5 \times 0.5} = 0.5$$

$$q(t_i = 1) = p(t_i = 1 \mid x_i) = \begin{cases} 1 & x_i = 1 \\ 0.5 & x_i = 2 \\ 0 & x_i = 3 \end{cases}$$

$q(t_i = 2) = 1 - q(t_i = 1)$

M-step Want to maximize the expected value

$$\max_{\alpha, \beta, \gamma} \sum_{i=1}^{N} \mathbb{E}_{q(t_i)}\left[\log p(x_i \mid t_i) p(t_i)\right]$$

$$= \sum_{i=1}^{N} q(t_i = 1) \log \underbrace{p_1(x_i \mid t_i)}_{\alpha,\, 1-\alpha,\, 0} \underbrace{p(t_i)}_{\gamma} + \sum_{i=1}^{N} q(t_i = 2) \log \underbrace{P_2(x_i \mid t_i)}_{0,\, \beta,\, 1-\beta} \underbrace{p(t_i)}_{(1-\gamma)}$$

Assume data: $N_1 = 30$, $N_2 = 20$, $N_3 = 60$ [observation. # of values in each category]

$$= 30 \times p(t_i = 1 \mid x_i = 1) \log \alpha\, \gamma + 20 \times 0.5 \times \log (1-\alpha)\gamma$$

$$+ 60 \cdot 0 \cdot \log 0 \quad (0)$$

$$+ 30 \times p(t_i = 2 \mid x_i = 1) \log \cdots \quad (0) \quad + 20 \times 0.5 \log (1-\beta)(1-\gamma)$$

$$+ 60 \times 1 \times \log \beta (1-\gamma)$$

$$\propto 30 \log \alpha + 10 \log (1-\alpha) + \text{const}(\alpha)$$

$$\frac{\partial}{\partial \alpha} = 30 \frac{1}{\alpha} + 10 \frac{1 \times (-1)}{1-\alpha} = 0 \qquad \alpha = \frac{30}{40} = \frac{3}{4}$$

$$\beta = \frac{6}{7} \qquad \gamma = \frac{4}{11}$$

$$\frac{\partial}{\partial \alpha} = 30 \frac{1}{\alpha} + 10 \frac{1 \times (-1)}{1-\alpha} = 0 \qquad \alpha = \frac{30}{40} = \frac{3}{4}$$

$$\beta = \frac{6}{7} \qquad \gamma = \frac{4}{11}$$

Summary of the Expectation - Maximization Algorithm

em-summary

Summary of Expectation Maximization

Summary of Expectation Maximization

- Method for training Latent Variable Models

Summary of Expectation Maximization

- Method for training Latent Variable Models
- Handles missing data

	High school grade	University grade	IQ score	Phone Interview
John	4.0	4.0	120	3/4
Helen	3.7	3.6	N/A	4/4
Jack	3.2	N/A	112	2/4
Emma	2.9	3.2	N/A	3/4

Summary of Expectation Maximization

- Method for training Latent Variable Models
- Handles missing data
- Sequence of simple task instead of one hard

Summary of Expectation Maximization

- Method for training Latent Variable Models
- Handles missing data
- Sequence of simple task instead of one hard
- Guaranties to converge

Summary of Expectation Maximization

- Method for training Latent Variable Models
- Handles missing data
- Sequence of simple task instead of one hard
- Guaranties to converge
- Helps with complicated parameter constraints

$$\Sigma_c \succ 0$$

Would you use EM algorithm to fit a Gaussian to one-dimensional data (i.e. estimate the parameters μ and σ)? What about multidimensional data (estimating the mean vector μ and the covariance matrix σ)?

- Yes, both for one-dimensional and multidimensional data
- Only for multi-dimensional data

Correct

We can treat missing values as latent variables and still estimate the Gaussian parameters. But with one-dimensional data, if a data point has missing values, it means that we don't know anything about it (its only dimension is missing) and not even the smartest latent variable model can extract information from a point like this. So the only thing that is left is to throw away points with missing data.

Note that we also don't need EM to estimate the mean vector (e.i. we need it only for the covariance matrix) in the multi-dimensional case: since each coordinate of the mean vector can be treated independently, we can treat each coordinate as one-dimensional case and just throw away missing values.

?

Summary of Expectation Maximization

- Method for training Latent Variable Models
- Handles missing data
- Sequence of simple task instead of one hard
- Guaranties to converge
- Helps with complicated parameter constraints can solve M step analytically.
- Numerous extensions:
  - Variational E-step: restrict the set of possible q Approximation.
    (week 3 and 5)
  - Sampling on M-step (week 4) for complicated distributions

Summary of Expectation Maximization

- Method for training Latent Variable Models
- Handles missing data
- Sequence of simple task instead of one hard
- Guaranties to converge
- Helps with complicated parameter constraints
- Numerous extensions

Cons

- Only local maximum (or saddle point)
- Requires math :)