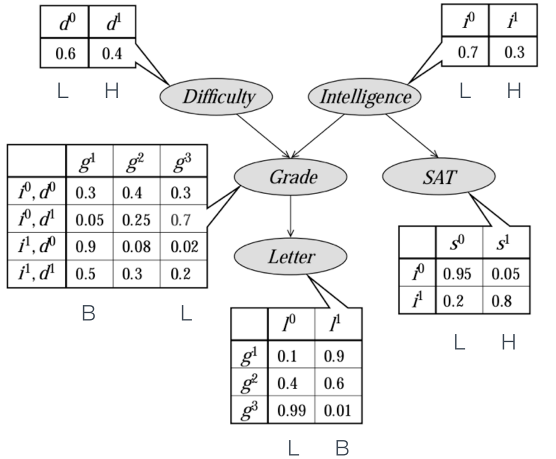
Royal Institute of Technology

MACHINE LEARNING 2- EM ALGORITHM

LECTURE 2

EXTENDED STUDENT EXAMPLE



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

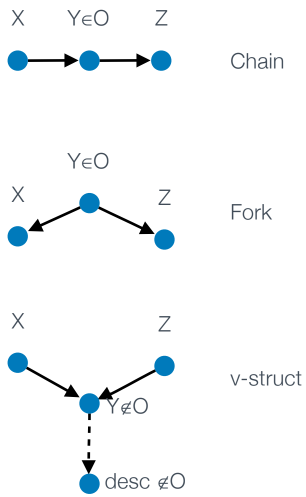
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

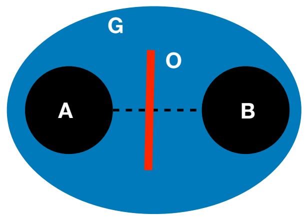
B - better
H - higher
L - less

D-SEPARATION

★ A path is d-separated by O if it has

- a chain X → Y → Z where Y ∊ O
- a fork X ← Y → Z where Y ∊ O
- a v-structure X → Y ← Z where (Y ∪ desc(Y)) ∩ O = ∅





D-SEPARATION SETS AND CI OF DAGS

★ A is d-separated from B given O if every undirected path between A and B is d-separated by O

★ In a DAG G,

$$x_A \perp_G x_B | x_O$$

↕

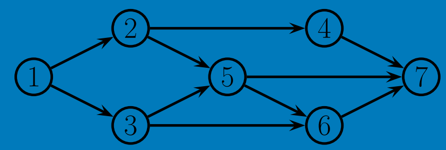
A is d-separated from B given O

- ★ Global (G): d-separation
- ★ Local (L): $\boldsymbol{X}_t \perp \boldsymbol{X}_{V\backslash \mathrm{desc}(t)} | \boldsymbol{X}_{\mathrm{pa}(t)}$
- ★ Ordered (O): $\boldsymbol{X}_t \perp \boldsymbol{X}_{\mathrm{pred}(t)} | \boldsymbol{X}_{\mathrm{pa}(t)}$

  where pred is according to a topological order
- ★ Factorized (F): can be family-factorized
- ★ Theorem: G ⇔ L ⇔ O ⇔ F

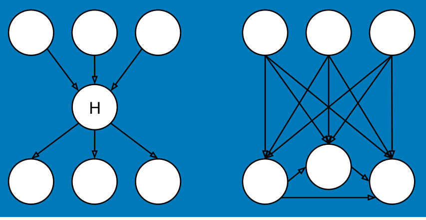
EQUIVALENCE OF INDEPENDENCE DEFINITIONS

MARKOV BLANKET



- ★ A minimal set B s/t X_t is independent from $X_{V\backslash(B\cup t)}$ given X_B is a Markov blanket
- ★ For t, pa(t) ∪ c(t) ∪ pa(c(t)) is a Markov blanket – i.e., parents, children, and co-parents (necessary due to v-structures)

LATENT = HIDDEN



17 parameters 59 parameters

- ★ Can reduce #parameters
- ★ Can represent common causes

LEARNING PARAMETERS COMPLETE DATA

- ★ "...Bayesian view, the parameters are unknown variables and should also be inferred"
- ★ Learning from complete data $\mathcal{D} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$, $\boldsymbol{x}_n = \{\boldsymbol{x}_{n1}, \dots, \boldsymbol{x}_{nV}\}$
- ★ Likelihood

$$P(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} P(\boldsymbol{x}_n|\boldsymbol{\theta}) = \prod_{n=1}^{N}\prod_{v=1}^{V} P(\boldsymbol{x}_{nv}|\boldsymbol{x}_{n,\mathrm{pa}(v)}, \boldsymbol{\theta})$$
$$= \prod_{v=1}^{V}\prod_{n=1}^{N} P(\boldsymbol{x}_{nv}|\boldsymbol{x}_{n,\mathrm{pa}(v)}, \boldsymbol{\theta}) = \prod_{v=1}^{V} P(\mathcal{D}_v|\boldsymbol{\theta}_v)$$

where D_v is values of v together with its parents and θ_v is v's CPD

MLE FOR CAT CPDS

- ★ Each $P(\mathcal{D}_v|\boldsymbol{\theta}_v)$, i.e., here each $\boldsymbol{\theta}_{v\mathbf{c}}$

  can be maximized independently

- ★ So, MLE is

$$\boldsymbol{\theta}_{v\mathbf{c}k} = N_{v\mathbf{c}k}/N_{v\mathbf{c}}$$

- ★ where

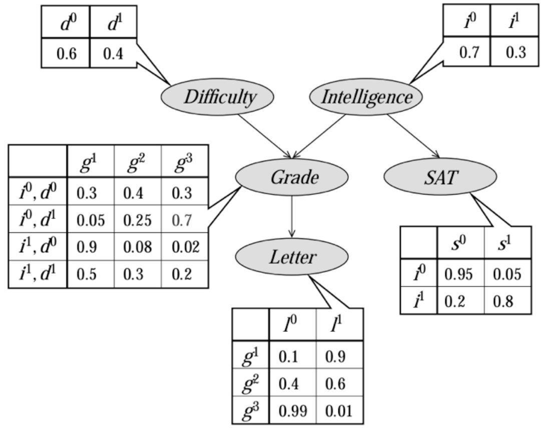
$$N_{v\mathbf{c}k} = \sum_{n=1}^{N} I(x_{nv} = k, x_{n,\mathrm{pa}(v)} = \mathbf{c})$$

$$N_{v\mathbf{c}} = \sum_{n=1}^{N} I(x_{n,\mathrm{pa}(v)} = \mathbf{c})$$

SUFFICIENT STATISTICS

c

EXTENDED STUDENT EXAMPLE



d^0	d^1
0.6	0.4

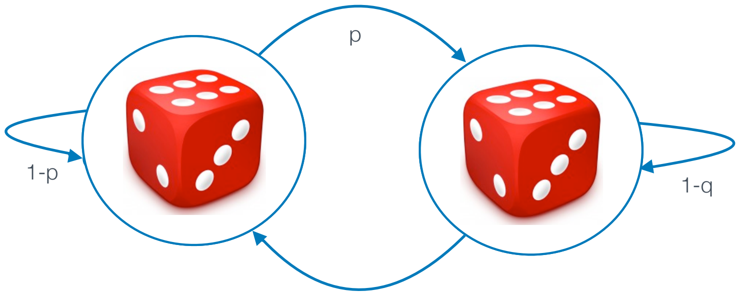
i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

GIVEN DATA AND NON-OPTIMAL PARAMETERS



- ★ We observe the sequence of dice outcomes of visited vertices

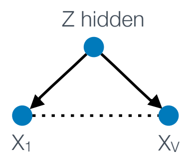
```
Rolls:    664153216162115234653214356634261655234232315142464156663246
Die:      LLLLLLLLLLLLLLFFFFFFLLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFLLLLLLLL
```

MIXTURE MODELS AND THE EXPECTATION MAXIMIZATION (EM) ALGORITHM

EM iteratively improves parameters

E-step: compute expected sufficient statistics (ESS) w.r.t. $p(z_k|\boldsymbol{x}_k, \boldsymbol{\theta})$

M-step: find optimal θ´ using the ESS



Used for MLE of

- ★ mixture models
- ★ parameters of DGMs, HMMs
- ★ DAG in DGMs (the structural EM by Nir Friedman)

Chapter 3

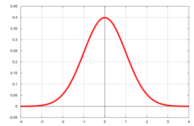
EXPECTATION-MAXIMIZATION THEORY

3.1 Introduction

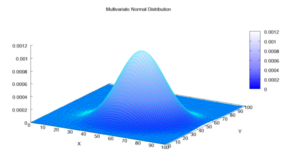
Learning networks are commonly categorized in terms of supervised and unsupervised networks. In unsupervised learning, the training set consists of input training patterns only. In contrast, in supervised learning networks, the training data consist

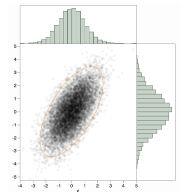
---

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

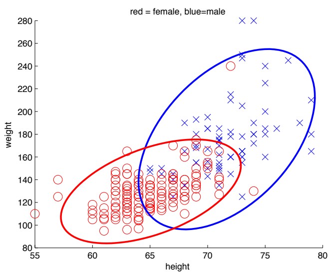


$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$





GAUSSIAN – MVN



TWO DIMENSIONAL NORMAL

GAUSSIAN MIXTURE MODELS (GMM)

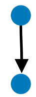
$\mathcal{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$

$\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nD})$

$Z \sim \mathrm{Cat}(\boldsymbol{\pi})$

$p(\boldsymbol{X}|Z=c) = p_c(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}|\boldsymbol{\mu}_c, \Sigma_c)$

Z hidden $\sim \mathrm{Cat}(\boldsymbol{\pi})$



$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_c, \Sigma_c)$

1-DIM GAUSSIAN MIXTURE MODELS

$\mathcal{D} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$

$Z \sim \text{Cat}(\boldsymbol{\pi})$

$p(\boldsymbol{X}|Z = c) = p_c(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}|\boldsymbol{\mu}_c, \sigma_c)$

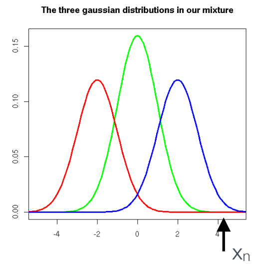
$\boldsymbol{\theta}_c = (\boldsymbol{\mu}_c, \sigma_c)$

Z hidden $\sim \text{Cat}(\boldsymbol{\pi})$

$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_c, \sigma_c)$

EXAMPLE

z_n is red with probability 1/2, green with probability 3/10, blue with probability 1/5



z_n = blue

x_n is generated from the Gaussian indicated by z_n

We get $x_1, \dots, x_N$

GMM

So,

$$p(x_n, z_n) = p(z_n)p(x_n|z_n)$$

and

$$p(x_n) = \sum_{c=1}^{C} p(z_n = c)p(x_n|z_n = c) = \sum_{c=1}^{C} \pi_c p(x_n|z_n = c)$$

and

$$p(z_n = c|x_n) = \frac{p(z_n = c, x_n)}{p(x_n)} = \frac{\pi_c p(x_n|z_n = c)}{\sum_{c=1}^{C} \pi_c p(x_n|z_n = c)}$$

COMPLETE LOG LIKELIHOOD (KNOWING ALL Z)

All paremeters

$$\begin{aligned}
l(\theta'; \mathcal{D}) &= \log \prod_n p(\boldsymbol{x}_n, z_n|\theta') \\
&= \sum_n \log \prod_c (\pi'_c p(\boldsymbol{x}_n|Z_n = c, \theta')^{I(z_n = c)} \\
&= \sum_n \sum_c I(z_n = c) \log(\pi'_c p(\boldsymbol{x}_n|\theta'_c)) \\
&= \sum_n \sum_c I(z_n = c) \log \pi'_c + \sum_n \sum_c I(z_n = c) \log p(\boldsymbol{x}_n|\theta'_c) \\
&= \sum_c \sum_{n: I(z_n = c)} \log \pi'_c + \sum_c \sum_{n: I(z_n = c)} \log p(\boldsymbol{x}_n|\theta'_c) \\
&= \sum_c N_c \log \pi'_c + \sum_c \sum_{n: I(z_n = c)} \log p(\boldsymbol{x}_n|\theta'_c)
\end{aligned}$$

$N_c = \sum_n I(z_n = c)$

MLE FOR GAUSSIAN

- Maximizeing

$$l(\theta';\mathcal{D}) = \sum_c N_c \log \pi'_c + \sum_c \sum_{n:I(z_n=c)} \log p(\boldsymbol{x}_n|\theta'_c)$$

- Boils down to maximizing

$$\sum_{n:I(z_n=c)} \log p(\boldsymbol{x}_n|\theta'_c)$$

that is $\displaystyle \sum_{n:I(z_n=c)} \log \frac{1}{\sqrt{2\pi\sigma_c'^2}} \exp\left(-\frac{1}{2\sigma_c'^2}(x_n-\mu_c)^2\right)$

MLE FOR GAUSSIAN

$$f(\sigma'_c,\mu'_c) = \sum_{n:I(z_n=c)} \log \frac{1}{\sqrt{2\pi\sigma_c'^2}} \exp\left(-\frac{1}{2\sigma_c'^2}(x_n-\mu'_c)^2\right)$$
$$= \sum_{n:I(z_n=c)} \log \frac{1}{\sqrt{2\pi\sigma_c'^2}} - \sum_{n:I(z_n=c)} \frac{1}{2\sigma_c'^2}(x_n-\mu'_c)^2$$

Derivation, $\displaystyle \frac{\partial f}{\partial \mu'_c} = \sum_{n:I(z_n=c)} \frac{2}{2\sigma_c'^2}(x_n-\mu'_c)$

Solving, $\displaystyle \frac{\partial f}{\partial \mu'_c} = 0 \Rightarrow \sum_{n:I(z_n=c)} x_n = \sum_{n:I(z_n=c)} \mu'_c = N_c\mu'_c$

So, $\displaystyle \mu'_c = \frac{\sum_{n:I(z_n=c)} x_n}{N_c}$ where $\displaystyle N_c = \sum_n I(z_n=c)$

K-MEANS

★ Data vectors D={$x_1,\ldots,x_N$}

★ Randomly selected classes $z_1,\ldots,z_N$

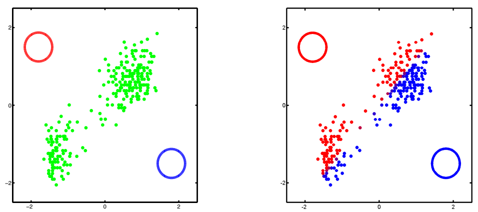
★ Iteratively do

$$\boldsymbol{\mu}_c = \frac{1}{N_c}\sum_{n:z_n=c} \boldsymbol{x}_n, \qquad \text{where } N_c = |\{n : z_n = c\}|$$

$$z_n = \mathrm{argmin}_c ||\boldsymbol{x}_n - \boldsymbol{\mu}_c||_2$$

★ One step O(NKD), can be improved

ASSIGN POINT TO MEANS



K-MEANS AS GMM

★ Fixed variance, only means must be estimated $\theta_c = (\mu_c, \sigma^2)$

★ Idea: each point can belong to several means (clusters)

★ Use responsibilities to find means

$$r_{nc} = p(z_n = c|\boldsymbol{x}_n, \boldsymbol{\theta}) = \frac{p(z_n = c|\boldsymbol{\theta})p(\boldsymbol{x}_n|z_n = c, \boldsymbol{\theta})}{\sum_{c=1}^{C} p(z_n = c|\boldsymbol{\theta})p(\boldsymbol{x}_n|z_n = c, \boldsymbol{\theta})}$$

$$\boldsymbol{\mu}_c = \frac{1}{N_c}\sum_n r_{nc}\boldsymbol{x}_n, \qquad \text{where } N_c = \sum_n r_{nc}$$

EM FOR GMM

★ Problem with previous most similar approach (MLE by derivation)

$$L(\boldsymbol{\theta}') = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta}') = \prod_{n=1}^{N}\sum_{z_n} p(\boldsymbol{x}_n, z_n|\boldsymbol{\theta}')$$

$$l(\boldsymbol{\theta}') = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}') = \sum_{n=1}^{N} \log \sum_{z_n} p(\boldsymbol{x}_n, z_n|\boldsymbol{\theta}')$$

★ How to handle the sum? the log cannot be pushed inside.

★ Idea: use z_i from $p(z_n|\boldsymbol{x}_n, \boldsymbol{\theta})$

Current parameters

EM & EXPECTED LOG LIKELIHOOD (Q-TERM)

- Iteratively maximizing the expected log likelihood in practice always leads to a local maxima
- The expectation is over latent variables given data and current parameters
- We maximize the expression by choosing new parameters.

LOG LIKELIHOOD

$$\begin{aligned}
l(\theta'; \mathcal{D}) &= \log \prod_n p(\boldsymbol{x}_n, z_n|\theta') \\
&= \sum_n \log \prod_c (\pi'_c p(\boldsymbol{x}_n|Z_n = c, \theta')^{I(z_n=c)} \\
&= \sum_n \sum_c I(z_n = c) \log(\pi'_c p(\boldsymbol{x}_n|\theta'_c) \\
&= \sum_n \sum_c I(z_n = c) \log \pi'_c + \sum_n \sum_c I(z_n = c) \log p(\boldsymbol{x}_n|\theta'_c)
\end{aligned}$$

EXPECTED LOG LIKELIHOOD (Q-TERM)

$$E_{p(z_n|\boldsymbol{x}_n,\boldsymbol{\theta})}\left[l(\theta';\mathcal{D})\right] = E_{p(z_n|\boldsymbol{x}_n,\boldsymbol{\theta})}\left[\log\prod_n p(\boldsymbol{x}_n, z_n|\theta')\right]$$
$$= \sum_n E\left[\log\prod_c (\pi'_c p(\boldsymbol{x}_n|z_n = c, \theta')^{I(z_n=c)}\right]$$
$$= \sum_n\sum_c E\left[I(z_n = c)\log(\pi'_c p(\boldsymbol{x}_n|\theta'_c)\right]$$
$$= \sum_n\sum_c p(z_n = c|\boldsymbol{x}_n, \boldsymbol{\theta})\log(\pi'_c p(\boldsymbol{x}_n|\theta'_c))$$
$$= \sum_n\sum_c r_{nc}\log\pi'_c + \sum_n\sum_c r_{nc}\log p(\boldsymbol{x}_n|\theta'_c)$$

where for a one dim Gaussian $\theta_c = (\mu_c, \sigma_c^2)$

HOW TO FIND θ´?

- ★ We want $\operatorname{argmax}_{\theta'} E_{p(z_n|\boldsymbol{x}_n,\boldsymbol{\theta})}\left[l(\theta';\mathcal{D})\right]$
- ★ The 2 sums $\sum_c\left(\sum_n r_{nc}\right)\log\pi'_c$ & $\sum_c\sum_n r_{nc}\log p(\boldsymbol{x}_n|\theta'_c)$ are independent
- ★ So, $\pi'_c = \sum_n r_{nc}/N = r_c/N$
- ★ In the second, different c indices are independent
- ★ So, we want to maximize each

$$\sum_n r_{nc}\log\frac{1}{\sqrt{2\pi\sigma_c'^2}}\exp\left(-\frac{1}{2\sigma_c'^2}(x_n - \mu_c)^2\right)$$

WEIGHTED GAUSS - MEAN

$$f(\sigma'_c, \mu'_c) = \sum_n r_{nc}\log\frac{1}{\sqrt{2\pi\sigma_c'^2}}\exp\left(-\frac{1}{2\sigma_c'^2}(x_n - \mu'_c)^2\right)$$
$$= \sum_n r_{nc}\log\frac{1}{\sqrt{2\pi\sigma_c'^2}} - \sum_n r_{nc}\frac{1}{2\sigma_c'^2}(x_n - \mu'_c)^2$$

Derivation, $\frac{\partial f}{\partial\mu'_c} = \sum_n r_{nc}\frac{2}{2\sigma_c'^2}(x_n - \mu'_c)$

Solving, $\frac{\partial f}{\partial\mu'_c} = 0 \Rightarrow \sum_n r_{nc}x_n = \left(\sum_n r_{nc}\right)\mu'_c = r_c\mu'_c$

So, $\mu'_c = \frac{\sum_n r_{nc}x_n}{r_c}$

VARIANCE

Let $\alpha'_c = 1/\sigma'_c$

$$f(\sigma'_c, \mu'_c) = \sum_n r_{nc}\log\frac{1}{\sqrt{2\pi\sigma_c'^2}} - \sum_n r_{nc}\frac{1}{2\sigma_c'^2}(x_n - \mu'_c)^2$$
$$= \sum_n r_{nc}\log\frac{\alpha'_c}{\sqrt{2\pi}} - \sum_n r_{nc}\frac{\alpha_c'^2}{2}(x_n - \mu'_c)^2$$

Derivation, $\frac{\partial f}{\partial\alpha'_c} = \sum_n\frac{r_{nc}}{\alpha'_c} - \sum_n r_{nc}\alpha'_c(x_n - \mu'_c)^2$

Solving, $\frac{\partial f}{\partial\alpha'_c} = 0 \Rightarrow \frac{r_c}{\alpha'_c} = \sum_n r_{nc}\alpha_c'^2(x_n - \mu'_c)^2$

So, $\sigma_c'^2 = \frac{1}{\alpha_c'^2} = \sum_n r_{nc}(x_n - \mu'_c)^2/r_c$

THEORETICAL BASIS FOR EM

★ 11.4.7 stared but read it.

★ Three prerequisites

- Jensen's inequality – do not read
- Entropy – 2.8.1, read
- Kullback-Leibler (KL) divergence – 2.8.2, read

Or make sure to understand these slides or read the EM theory text

Entropy and KL are interesting and included, but not necessary for the slides

JENSEN'S INEQUALITY

Let

$s \in [0,1]$

$c = a + s(b-a) = (1-s)a + sb$

$c' = (1-s)\log a + s \log b$

Then

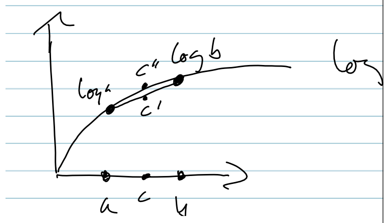
$$c'' = \log c = \log[(1-s)a + sb] \geq (1-s)\log a + s\log b = c'$$

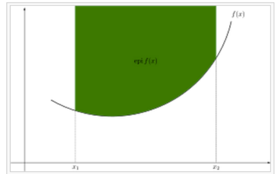
In general,

$$\log \sum_x p(x) f(x) \geq \sum_x p(x) \log f(x)$$

i.e.,

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





A function (in black) is convex if and only if the region above its graph (in green) is a convex set.

a,b,c can be values that f takes on

Expected complete log-likelihood – Q

$$\begin{aligned}
&\log p(\boldsymbol{x}|\boldsymbol{\theta}') \\
&= \log \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}') \\
&= \log \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}) \frac{p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}')}{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})} \\
&= \log E_{\boldsymbol{z}} \left(\frac{p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}')}{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})} \mid \boldsymbol{x}, \boldsymbol{\theta} \right) \\
&\geq^{\text{Jensen}} E_{\boldsymbol{z}} \left(\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}')}{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})} \mid \boldsymbol{x}, \boldsymbol{\theta} \right) \\
&= \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}')}{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})} \\
&= \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}) \log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}') - \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}) \log p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}) \\
&= Q(\boldsymbol{\theta}'; \boldsymbol{\theta}) - R(\boldsymbol{\theta}; \boldsymbol{\theta})
\end{aligned}$$

Expected complete: notation

$$\begin{aligned}
&\log p(\boldsymbol{x}_n|\boldsymbol{\theta}') \\
&= \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}') \\
&= \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \\
&= \log E_{\boldsymbol{z}_n} \left(\frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \mid \boldsymbol{x}_n, \boldsymbol{\theta} \right) \\
&\geq^{\text{Jensen}} E_{\boldsymbol{z}_n} \left(\log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \mid \boldsymbol{x}_n, \boldsymbol{\theta} \right) \\
&= \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \\
&= \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}') - \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \\
&= Q_n(\boldsymbol{\theta}'; \boldsymbol{\theta}) - R_n(\boldsymbol{\theta}; \boldsymbol{\theta})
\end{aligned}$$

Expected complete: fewer steps

$$
\begin{aligned}
\log p(\boldsymbol{x}_n|\boldsymbol{\theta}') \\
&= \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}') \\
&= \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \\
&\geq \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \\
&= \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}') - \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \\
&= Q_n(\boldsymbol{\theta}'; \boldsymbol{\theta}) - R_n(\boldsymbol{\theta}; \boldsymbol{\theta})
\end{aligned}
$$

Expected complete: of all data

$$
\begin{aligned}
\log p(\mathcal{D}|\boldsymbol{\theta}') \\
&= \sum_n \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}') \\
&= \sum_n \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \\
&\geq \sum_n \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \\
&= \sum_n \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}') - \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{z}_n|\boldsymbol{x}_n \\
&= \underbrace{\sum_n Q_n(\boldsymbol{\theta}'; \boldsymbol{\theta})}_{Q(\boldsymbol{\theta}';\boldsymbol{\theta})} - \underbrace{\sum_n R_n(\boldsymbol{\theta}; \boldsymbol{\theta})}_{R_n(\boldsymbol{\theta};\boldsymbol{\theta})}
\end{aligned}
$$

Expected complete: for Θ

$$
\begin{aligned}
\log p(\boldsymbol{x}_n|\boldsymbol{\theta}) \\
&= [\log p(\boldsymbol{x}_n|\boldsymbol{\theta})] \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \\
&= \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{x}_n|\boldsymbol{\theta}) \\
&= \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{z}_n, \boldsymbol{x}_n|\boldsymbol{\theta})}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \\
&= \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{z}_n, \boldsymbol{x}_n|\boldsymbol{\theta}) - \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \\
&= Q_n(\boldsymbol{\theta}, \boldsymbol{\theta}) - R_n(\boldsymbol{\theta}, \boldsymbol{\theta})
\end{aligned}
$$

RELATIONS BETWEEN LOG-LIKELIHOODS AND Q-TERMS

Theorem: for $\boldsymbol{\theta}' = \mathrm{argmax}_{\boldsymbol{\theta}''} Q(\boldsymbol{\theta}'', \boldsymbol{\theta})$

$$\log p(\mathcal{D}|\boldsymbol{\theta}') \geq Q(\boldsymbol{\theta}', \boldsymbol{\theta}) - R(\boldsymbol{\theta}, \boldsymbol{\theta}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}) - R(\boldsymbol{\theta}, \boldsymbol{\theta}) = \log p(\mathcal{D}|\boldsymbol{\theta})$$

So by maximizing Q-term (through ESS) we monotonically increase the likelihood.

The Q-term may not increase in every step!

- ★ Starting points
- ★ Number of starting points
- ★ Sieving starting points
- ★ The competition
  - The first iterations of EM show huge improvement in the likelihood. These are then followed by many iterations that slowly increase the likelihood. Conjugate gradient shows the opposite behaviour.

PRACTICAL ISSUES

THE END

$$\hat{\boldsymbol{\mu}}_{mle} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \triangleq \overline{\mathbf{x}}$$

$$\hat{\boldsymbol{\Sigma}}_{mle} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T = \frac{1}{N}(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T) - \overline{\mathbf{x}}\,\overline{\mathbf{x}}^T$$

MVN MLE

FOR MVN

$$\pi'_c = r_c/N$$

$$\boldsymbol{\mu}'_c = \frac{\sum_n r_{nc}\boldsymbol{x}_n}{r_c}$$

$$\Sigma'_c = \sum_n \frac{r_{nc}\boldsymbol{x}_n\boldsymbol{x}_n^T}{r_c} - \boldsymbol{\mu}_c\boldsymbol{\mu}_c^T$$

Z hidden $\sim \mathrm{Cat}(\boldsymbol{\pi})$

$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_c, \Sigma_c)$

ENTROPY

★ Definition

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

★ q a fair coin

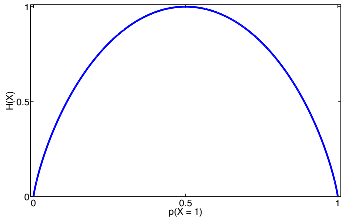
$$H(q) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}$$
$$= -1(-1) = 1$$

★q uniform on [K]

$$H(q) = -\sum \frac{1}{K}\log\frac{1}{K} = \log K$$

★q on [K], q(1)=1

$$H(q) = -1\log 1 - \sum_{k=2}^{K} 0 \log 0$$
$$= 0$$



ENTROPY VS EXPECTATION

value	1	2	3	μ= 1/2 +1/2+3/4 = 7/4
value	1	9	27	μ= 1/2 +9/4+27/4 = 38/4
probability	1/2	1/4	1/4	H= -(log 1/2)/2-2(log 1/4)/4=3/2

★ Entropy depends on the distribution only

KL-divergence

★ Definition

$$\mathrm{KL}(q||p) = \sum_{k=1}^{K} p_k \log\frac{p_k}{q_k}$$

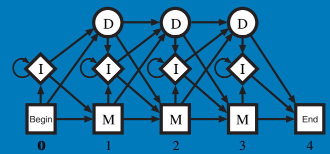
★ Alternative

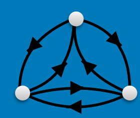
$$\mathrm{KL}(q||p) = \sum_x p(x) \log\frac{p(x)}{q(x)}$$

★ Theorem (you do not have to read the proof)

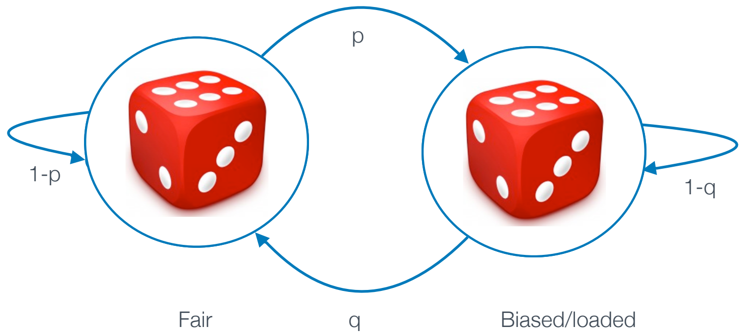
$$\mathrm{KL}(q||p) \geq 0 \text{ with equality iff } p = q$$

HMMS (LAYERED OR NOT)





GIVEN DATA AND NON-OPTIMAL PARAMETERS



★ We observe the sequence of dice outcomes of visited vertices

```
Rolls:     664153216162115234653214356634261655234232315142464156663246
Die:       LLLLLLLLLLLLLFFFFFFLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFLLLLLLLL
```