Probabilistic and Bayesian Machine Learning

Day 3: The EM Algorithm

Yee Whye Teh
ywteh@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit
University College London

http://www.gatsby.ucl.ac.uk/~ywteh/teaching/probmodels

The Expectation Maximisation (EM) algorithm

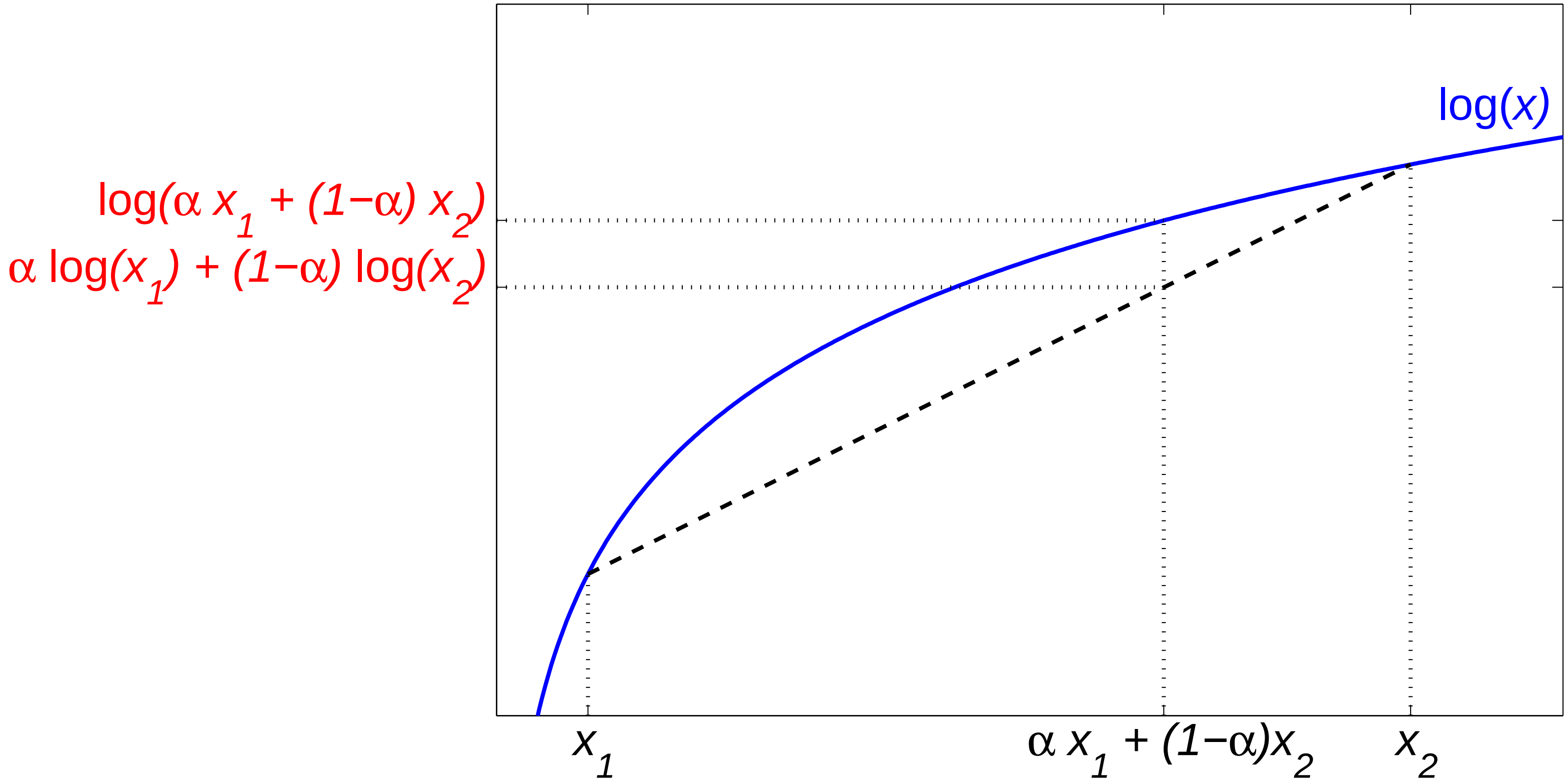
The EM algorithm finds a (local) maximum of a latent variable model likelihood $P(\mathbf{X}, \mathbf{Y}|\theta)$. It starts from arbitrary values of the parameters, and iterates two steps:

E step: Fill in values of latent variables according to posterior given data.

M step: Maximise likelihood as if latent variables were not hidden.

- Useful in models where learning would be easy if unobserved variables were, in fact, observed (e.g. MoGs).
- Decomposes difficult problems into series of tractable steps.
- No gradients and learning rate.
- Framework lends itself to principled approximations.

Jensen's Inequality



For $\alpha_i \geq 0$, $\sum \alpha_i = 1$ and any $\{x_i > 0\}$

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

Equality if and only if $\alpha_i = 1$ for some i (and therefore all others are 0).

The Free Energy for a Latent Variable Model

Observed data $\mathbf{X} = \{X_i\}$; Latent variables $\mathbf{Y} = \{Y_i\}$; Parameters θ.

Goal: Maximize the log likelihood (i.e. ML learning) wrt θ:

$$\ell(\theta) = \log P(\mathbf{X}|\theta) = \log \int P(\mathbf{Y}, \mathbf{X}|\theta) d\mathbf{Y},$$

The Free Energy for a Latent Variable Model

Observed data $\mathbf{X} = \{X_i\}$; Latent variables $\mathbf{Y} = \{Y_i\}$; Parameters θ.

Goal: Maximize the log likelihood (i.e. ML learning) wrt θ:

$$\ell(\theta) = \log P(\mathbf{X}|\theta) = \log \int P(\mathbf{Y}, \mathbf{X}|\theta) d\mathbf{Y},$$

Any distribution, $q(\mathbf{Y})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen's inequality:

$$\ell(\theta) = \log \int q(\mathbf{Y}) \frac{P(\mathbf{Y}, \mathbf{X}|\theta)}{q(\mathbf{Y})} \, d\mathbf{Y} \geq \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}, \mathbf{X}|\theta)}{q(\mathbf{Y})} \, d\mathbf{Y} \qquad .$$

The Free Energy for a Latent Variable Model

Observed data $\mathbf{X} = \{X_i\}$; Latent variables $\mathbf{Y} = \{Y_i\}$; Parameters θ.

Goal: Maximize the log likelihood (i.e. ML learning) wrt θ:

$$\ell(\theta) = \log P(\mathbf{X}|\theta) = \log \int P(\mathbf{Y}, \mathbf{X}|\theta) d\mathbf{Y},$$

Any distribution, $q(\mathbf{Y})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen's inequality:

$$\ell(\theta) = \log \int q(\mathbf{Y}) \frac{P(\mathbf{Y}, \mathbf{X}|\theta)}{q(\mathbf{Y})} \, d\mathbf{Y} \geq \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}, \mathbf{X}|\theta)}{q(\mathbf{Y})} \, d\mathbf{Y} \stackrel{\text{def}}{=} \mathcal{F}(q, \theta).$$

The Free Energy for a Latent Variable Model

Observed data $\mathbf{X} = \{X_i\}$; Latent variables $\mathbf{Y} = \{Y_i\}$; Parameters θ.

Goal: Maximize the log likelihood (i.e. ML learning) wrt θ:

$$\ell(\theta) = \log P(\mathbf{X}|\theta) = \log \int P(\mathbf{Y}, \mathbf{X}|\theta) d\mathbf{Y},$$

Any distribution, $q(\mathbf{Y})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen's inequality:

$$\ell(\theta) = \log \int q(\mathbf{Y}) \frac{P(\mathbf{Y}, \mathbf{X}|\theta)}{q(\mathbf{Y})} \, d\mathbf{Y} \geq \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}, \mathbf{X}|\theta)}{q(\mathbf{Y})} \, d\mathbf{Y} \stackrel{\text{def}}{=} \mathcal{F}(q, \theta).$$

Now,

$$\begin{aligned} \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}, \mathbf{X}|\theta)}{q(\mathbf{Y})} \, d\mathbf{Y} &= \int q(\mathbf{Y}) \log P(\mathbf{Y}, \mathbf{X}|\theta) \, d\mathbf{Y} - \int q(\mathbf{Y}) \log q(\mathbf{Y}) \, d\mathbf{Y} \\ &= \int q(\mathbf{Y}) \log P(\mathbf{Y}, \mathbf{X}|\theta) \, d\mathbf{Y} + \mathbf{H}[q], \end{aligned}$$

where $\mathbf{H}[q]$ is the entropy of $q(\mathbf{Y})$.

The Free Energy for a Latent Variable Model

Observed data $\mathbf{X} = \{X_i\}$; Latent variables $\mathbf{Y} = \{Y_i\}$; Parameters θ.

Goal: Maximize the log likelihood (i.e. ML learning) wrt θ:

$$\ell(\theta) = \log P(\mathbf{X}|\theta) = \log \int P(\mathbf{Y}, \mathbf{X}|\theta) d\mathbf{Y},$$

Any distribution, $q(\mathbf{Y})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen's inequality:

$$\ell(\theta) = \log \int q(\mathbf{Y}) \frac{P(\mathbf{Y}, \mathbf{X}|\theta)}{q(\mathbf{Y})} \, d\mathbf{Y} \geq \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}, \mathbf{X}|\theta)}{q(\mathbf{Y})} \, d\mathbf{Y} \stackrel{\text{def}}{=} \mathcal{F}(q, \theta).$$

Now,

$$\begin{aligned}\int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}, \mathbf{X}|\theta)}{q(\mathbf{Y})} \, d\mathbf{Y} &= \int q(\mathbf{Y}) \log P(\mathbf{Y}, \mathbf{X}|\theta) \, d\mathbf{Y} - \int q(\mathbf{Y}) \log q(\mathbf{Y}) \, d\mathbf{Y} \\ &= \int q(\mathbf{Y}) \log P(\mathbf{Y}, \mathbf{X}|\theta) \, d\mathbf{Y} + \mathsf{H}[q],\end{aligned}$$

where $\mathsf{H}[q]$ is the entropy of $q(\mathbf{Y})$.
So:

$$\mathcal{F}(q, \theta) = \langle \log P(\mathbf{Y}, \mathbf{X}|\theta) \rangle_{q(\mathbf{Y})} + \mathsf{H}[q]$$

The E and M steps of EM

The lower bound on the log likelihood is given by:

$$\mathcal{F}(q,\theta) = \langle \log P(\mathbf{Y},\mathbf{X}|\theta)\rangle_{q(\mathbf{Y})} + \mathbf{H}[q],$$

The E and M steps of EM

The lower bound on the log likelihood is given by:

$$\mathcal{F}(q, \theta) = \langle \log P(\mathbf{Y}, \mathbf{X}|\theta) \rangle_{q(\mathbf{Y})} + \mathbf{H}[q],$$

EM alternates between:

E step: optimize $\mathcal{F}(q, \theta)$ wrt distribution over hidden variables holding parameters fixed:

$$q^{(k)}(\mathbf{Y}) := \underset{q(\mathbf{Y})}{\text{argmax}} \; \mathcal{F}\big(q(\mathbf{Y}), \theta^{(k-1)}\big).$$

The E and M steps of EM

The lower bound on the log likelihood is given by:

$$\mathcal{F}(q, \theta) = \langle \log P(\mathbf{Y}, \mathbf{X}|\theta) \rangle_{q(\mathbf{Y})} + \mathbf{H}[q],$$

EM alternates between:

E step: optimize $\mathcal{F}(q, \theta)$ wrt distribution over hidden variables holding parameters fixed:

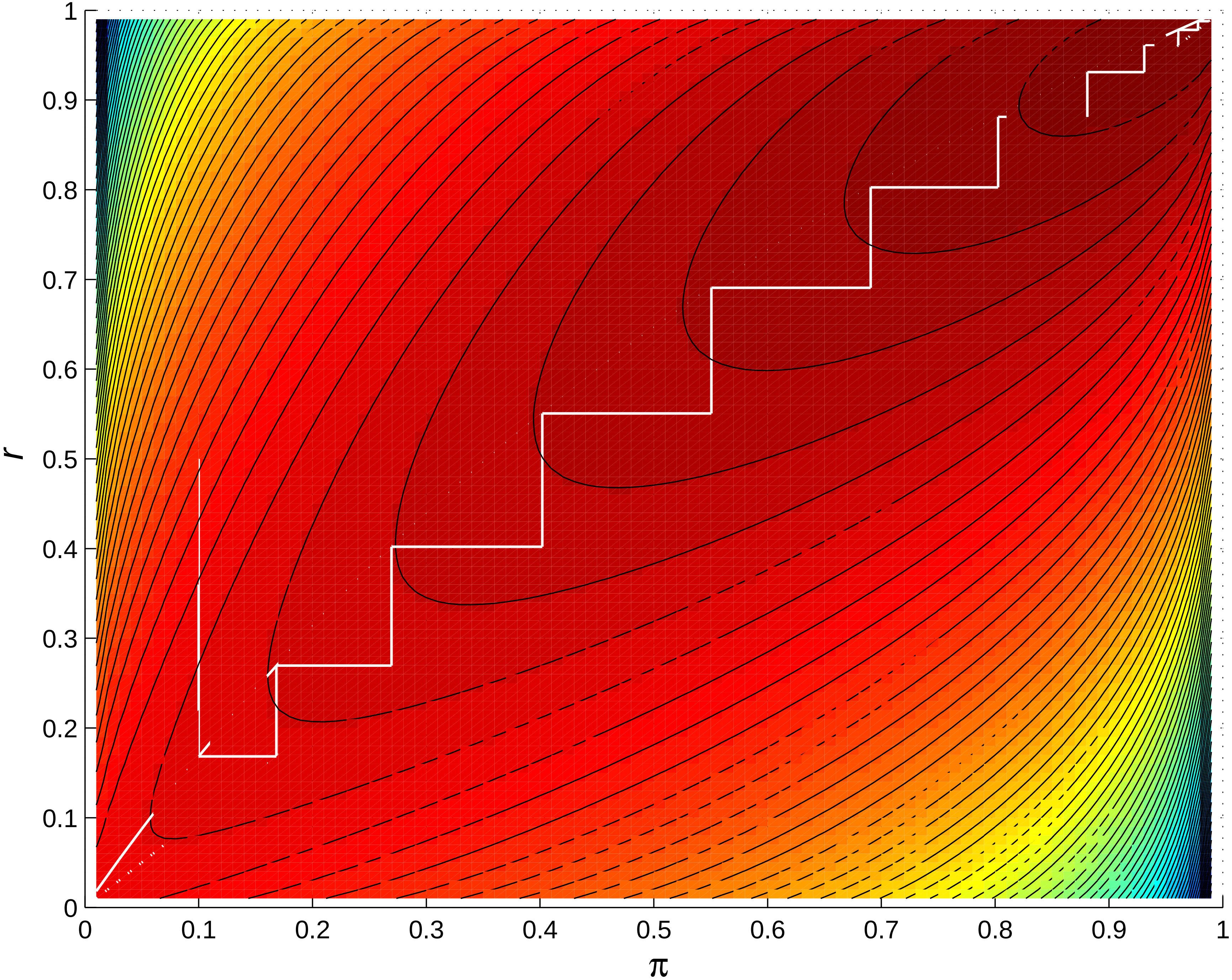
$$q^{(k)}(\mathbf{Y}) := \underset{q(\mathbf{Y})}{\text{argmax}} \; \mathcal{F}\big(q(\mathbf{Y}), \theta^{(k-1)}\big).$$

M step: maximize $\mathcal{F}(q, \theta)$ wrt parameters holding hidden distribution fixed:

$$\theta^{(k)} := \underset{\theta}{\text{argmax}} \; \mathcal{F}\big(q^{(k)}(\mathbf{Y}), \theta\big) = \underset{\theta}{\text{argmax}} \; \langle \log P(\mathbf{Y}, \mathbf{X}|\theta) \rangle_{q^{(k)}(\mathbf{Y})}$$

The second equality comes from the fact that the entropy of $q(\mathbf{Y})$ does not depend directly on θ.

EM as Coordinate Ascent in $\mathcal{F}$



The E Step

The free energy can be re-written

$$\mathcal{F}(q,\theta) = \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y},\mathbf{X}|\theta)}{q(\mathbf{Y})} \, d\mathbf{Y}$$

The E Step

The free energy can be re-written

$$\begin{aligned}\mathcal{F}(q,\theta) &= \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y},\mathbf{X}|\theta)}{q(\mathbf{Y})}\, d\mathbf{Y} \\ &= \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}|\mathbf{X},\theta)P(\mathbf{X}|\theta)}{q(\mathbf{Y})}\, d\mathbf{Y}\end{aligned}$$

The E Step

The free energy can be re-written

$$\begin{aligned}\mathcal{F}(q,\theta) &= \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y},\mathbf{X}|\theta)}{q(\mathbf{Y})}\, d\mathbf{Y} \\ &= \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}|\mathbf{X},\theta)P(\mathbf{X}|\theta)}{q(\mathbf{Y})}\, d\mathbf{Y} \\ &= \int q(\mathbf{Y}) \log P(\mathbf{X}|\theta)\, d\mathbf{Y} + \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}|\mathbf{X},\theta)}{q(\mathbf{Y})}\, d\mathbf{Y}\end{aligned}$$

The E Step

The free energy can be re-written

$$
\begin{aligned}
\mathcal{F}(q,\theta) &= \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y},\mathbf{X}|\theta)}{q(\mathbf{Y})}\, d\mathbf{Y} \\
&= \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}|\mathbf{X},\theta)P(\mathbf{X}|\theta)}{q(\mathbf{Y})}\, d\mathbf{Y} \\
&= \int q(\mathbf{Y}) \log P(\mathbf{X}|\theta)\, d\mathbf{Y} + \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}|\mathbf{X},\theta)}{q(\mathbf{Y})}\, d\mathbf{Y} \\
&= \ell(\theta) - \textbf{KL}[q(\mathbf{Y}) \| P(\mathbf{Y}|\mathbf{X},\theta)]
\end{aligned}
$$

The second term is the Kullback-Leibler divergence.

The E Step

The free energy can be re-written

$$\begin{aligned}\mathcal{F}(q,\theta) &= \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y},\mathbf{X}|\theta)}{q(\mathbf{Y})}\, d\mathbf{Y} \\ &= \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}|\mathbf{X},\theta)P(\mathbf{X}|\theta)}{q(\mathbf{Y})}\, d\mathbf{Y} \\ &= \int q(\mathbf{Y}) \log P(\mathbf{X}|\theta)\, d\mathbf{Y} + \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}|\mathbf{X},\theta)}{q(\mathbf{Y})}\, d\mathbf{Y} \\ &= \ell(\theta) - \mathbf{KL}[q(\mathbf{Y}) \| P(\mathbf{Y}|\mathbf{X},\theta)]\end{aligned}$$

The second term is the Kullback-Leibler divergence.

This means that, for fixed θ, $\mathcal{F}$ is bounded above by ℓ, and achieves that bound when $\mathbf{KL}[q(\mathbf{Y}) \| P(\mathbf{Y}|\mathbf{X},\theta)] = 0$.

The E Step

The free energy can be re-written

$$\begin{aligned}\mathcal{F}(q,\theta) &= \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y},\mathbf{X}|\theta)}{q(\mathbf{Y})}\, d\mathbf{Y} \\ &= \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}|\mathbf{X},\theta)P(\mathbf{X}|\theta)}{q(\mathbf{Y})}\, d\mathbf{Y} \\ &= \int q(\mathbf{Y}) \log P(\mathbf{X}|\theta)\, d\mathbf{Y} + \int q(\mathbf{Y}) \log \frac{P(\mathbf{Y}|\mathbf{X},\theta)}{q(\mathbf{Y})}\, d\mathbf{Y} \\ &= \ell(\theta) - \textbf{KL}[q(\mathbf{Y})\|P(\mathbf{Y}|\mathbf{X},\theta)]\end{aligned}$$

The second term is the Kullback-Leibler divergence.

This means that, for fixed θ, $\mathcal{F}$ is bounded above by ℓ, and achieves that bound when $\textbf{KL}[q(\mathbf{Y})\|P(\mathbf{Y}|\mathbf{X},\theta)] = 0$.

But $\textbf{KL}[q\|p]$ is zero if and only if $q = p$. So, the E step simply sets

$$q^{(k)}(\mathbf{Y}) = P(\mathbf{Y}|\mathbf{X},\theta^{(k-1)})$$

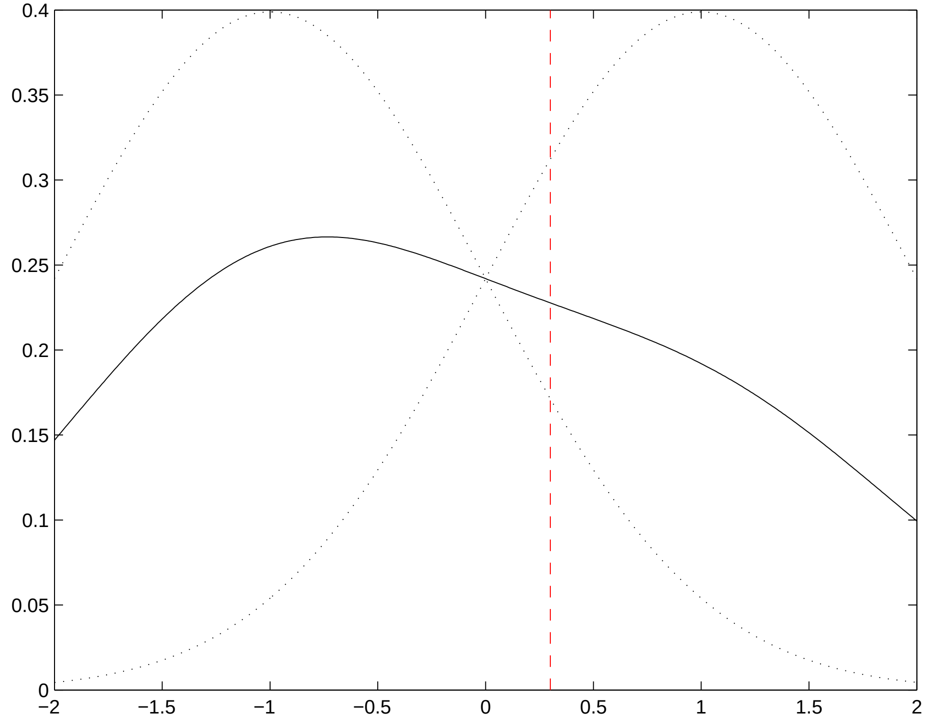
and, after an E step, the free energy equals the likelihood.

Coordinate Ascent in $\mathcal{F}$ (Demo)

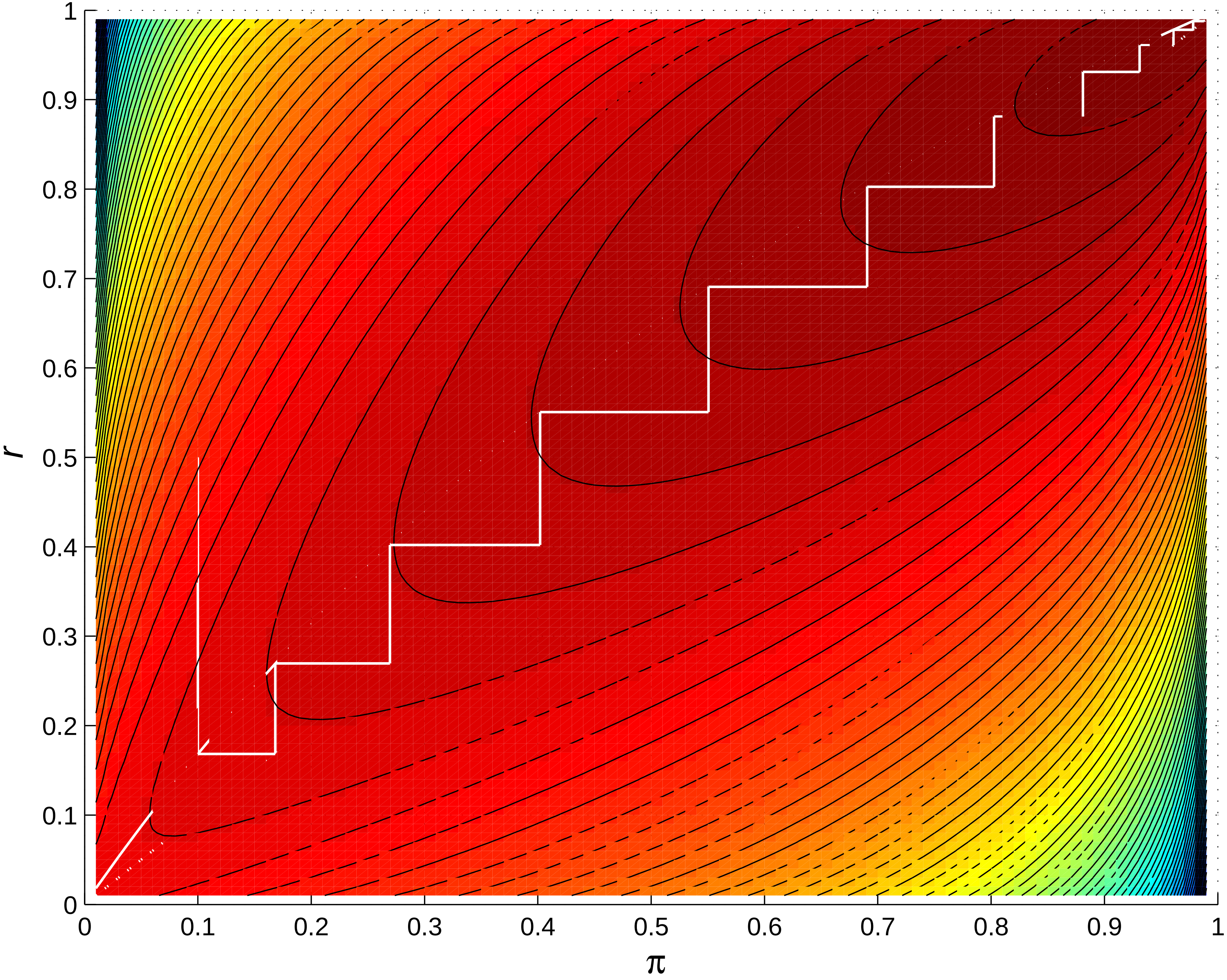
One parameter mixture:

$$s \sim \text{Bernoulli}[\pi]$$
$$x|s = 0 \sim \mathcal{N}[-1, 1] \qquad x|s = 1 \sim \mathcal{N}[1, 1]$$

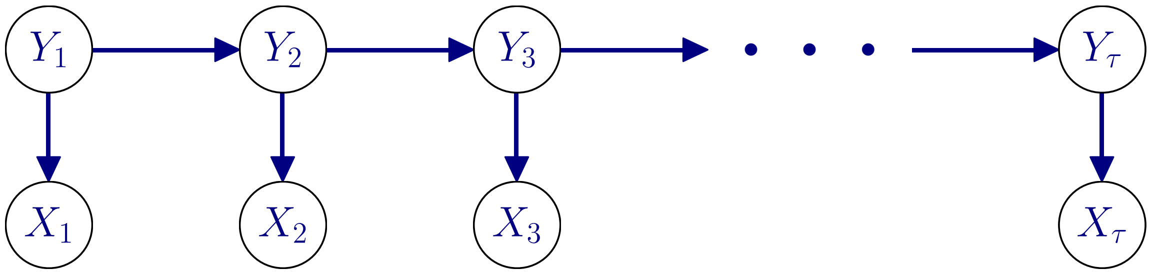
and one data point $x_1 = .3$.
$q(s)$ is a distribution on a single binary latent, and so is represented by $r_1 \in [0, 1]$.



Coordinate Ascent in $\mathcal{F}$ (Demo)



EM for Learning HMMs



Parameters: $\theta = \{\pi, T, A\}$

Free energy:

$$\mathcal{F}(q, \theta) = \sum_{Y_{1:\tau}} q(Y_{1:\tau})(\log P(X_{1:\tau}, Y_{1:\tau}|\theta) - \log q(Y_{1:\tau}))$$

E-step: Maximise $\mathcal{F}$ w.r.t. q with θ fixed: $q^*(Y_{1:\tau}) = P(Y_{1:\tau}|X_{1:\tau}, \theta)$
We will only need the marginal probabilities $q^*(Y_t, Y_{t+1})$, which can also be obtained from the forward-backward algorithm.
M-step: Maximize $\mathcal{F}$ w.r.t. θ with q fixed.
We can re-estimate the parameters by computing the expected number of times the HMM was in state i, emitted symbol k and transitioned to state j.

This is the Baum-Welch algorithm and it predates the (more general) EM algorithm.

M step: Parameter updates are given by just ratios of expected counts

We can derive the following updates by taking derivatives of $\mathcal{F}$ w.r.t. θ.

- Let the posterior marginals be:

$$\gamma_t(i) = P(Y_t = i | X_{1:\tau}) \propto \alpha_t(i)\beta_t(i)$$
$$\xi_t(ij) = P(Y_t = i, Y_{t+1} = j | X_{1:\tau}) \propto \alpha_i(i) P(Y_{t+1} = j | Y_t = i) P(X_{t+1} | Y_{t+1} = j) \beta_{t+1}(j)$$

- The initial state distribution is the expected number of times in state i at $t = 1$:

$$\hat{\pi}_i = \gamma_1(i)$$

- The estimated transition probabilities are:

$$\hat{T}_{ij} = \frac{\sum_{t=1}^{\tau-1} \xi_t(ij)}{\sum_{t=1}^{\tau-1} \gamma_t(i)}$$

- The output distributions are the expected number of times we observe a particular symbol in a particular state:

$$\hat{A}_{ik} = \frac{\sum_{t:X_t=k} \gamma_t(i)}{\sum_{t=1}^{\tau} \gamma_t(i)}$$

(or the state-probability-weighted sufficient statistics for exponential family observation models).

EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

$$\ell\Big(\theta^{(k-1)}\Big)$$

EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

$$\ell\big(\theta^{(k-1)}\big) \underset{\text{E step}}{=} \mathcal{F}\big(q^{(k)}, \theta^{(k-1)}\big)$$

- The E step brings the free energy to the likelihood.

EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

$$\ell\big(\theta^{(k-1)}\big) \underset{\text{E step}}{=} \mathcal{F}\big(q^{(k)}, \theta^{(k-1)}\big) \underset{\text{M step}}{\leq} \mathcal{F}\big(q^{(k)}, \theta^{(k)}\big)$$

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt θ.

EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

$$\ell\big(\theta^{(k-1)}\big) \underset{\text{E step}}{=} \mathcal{F}\big(q^{(k)}, \theta^{(k-1)}\big) \underset{\text{M step}}{\leq} \mathcal{F}\big(q^{(k)}, \theta^{(k)}\big) \underset{\text{Jensen}}{\leq} \ell\big(\theta^{(k)}\big),$$

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt θ.
- $\mathcal{F} \leq \ell$ by Jensen – or, equivalently, from the non-negativity of KL

EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

$$\ell\big(\theta^{(k-1)}\big) \underset{\text{E step}}{=} \mathcal{F}\big(q^{(k)}, \theta^{(k-1)}\big) \underset{\text{M step}}{\leq} \mathcal{F}\big(q^{(k)}, \theta^{(k)}\big) \underset{\text{Jensen}}{\leq} \ell\big(\theta^{(k)}\big),$$

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt θ.
- $\mathcal{F} \leq \ell$ by Jensen – or, equivalently, from the non-negativity of KL

If the M-step is executed so that $\theta^{(k)} \neq \theta^{(k-1)}$ iff $\mathcal{F}$ increases, then the overall EM iteration will step to a new value of θ iff the likelihood increases.

Fixed Points of EM are Stationary Points in ℓ

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial \theta} \langle \log P(\mathbf{Y}, \mathbf{X} \mid \theta) \rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} \bigg|_{\theta^*} = 0$$

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial\theta}\langle\log P(\mathbf{Y},\mathbf{X}\mid\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\bigg|_{\theta^*}=0$$

Now, $$\ell(\theta)=\log P(\mathbf{X}|\theta)=\langle\log P(\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}$$

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial\theta}\langle \log P(\mathbf{Y},\mathbf{X} \mid \theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\bigg|_{\theta^*} = 0$$

Now,

$$\begin{aligned}\ell(\theta) &= \log P(\mathbf{X}|\theta) = \langle \log P(\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} \\ &= \left\langle \log \frac{P(\mathbf{Y},\mathbf{X}|\theta)}{P(\mathbf{Y}|\mathbf{X},\theta)} \right\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\end{aligned}$$

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial\theta}\langle \log P(\mathbf{Y},\mathbf{X} \mid \theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\bigg|_{\theta^*} = 0$$

Now,

$$\begin{aligned}
\ell(\theta) &= \log P(\mathbf{X}|\theta) = \langle \log P(\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} \\
&= \left\langle \log \frac{P(\mathbf{Y},\mathbf{X}|\theta)}{P(\mathbf{Y}|\mathbf{X},\theta)} \right\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} \\
&= \langle \log P(\mathbf{Y},\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} - \langle \log P(\mathbf{Y}|\mathbf{X},\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}
\end{aligned}$$

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial\theta}\langle \log P(\mathbf{Y},\mathbf{X} \mid \theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\bigg|_{\theta^*} = 0$$

Now,

$$\begin{aligned}\ell(\theta) &= \log P(\mathbf{X}|\theta) = \langle \log P(\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\\ &= \left\langle \log \frac{P(\mathbf{Y},\mathbf{X}|\theta)}{P(\mathbf{Y}|\mathbf{X},\theta)}\right\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\\ &= \langle \log P(\mathbf{Y},\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} - \langle \log P(\mathbf{Y}|\mathbf{X},\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\end{aligned}$$

so,

$$\frac{d}{d\theta}\ell(\theta) = \frac{d}{d\theta}\langle \log P(\mathbf{Y},\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} - \frac{d}{d\theta}\langle \log P(\mathbf{Y}|\mathbf{X},\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}$$

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial\theta}\langle \log P(\mathbf{Y},\mathbf{X} \mid \theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\Big|_{\theta^*} = 0$$

Now,

$$\begin{aligned}\ell(\theta) &= \log P(\mathbf{X}|\theta) = \langle \log P(\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} \\ &= \left\langle \log \frac{P(\mathbf{Y},\mathbf{X}|\theta)}{P(\mathbf{Y}|\mathbf{X},\theta)} \right\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} \\ &= \langle \log P(\mathbf{Y},\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} - \langle \log P(\mathbf{Y}|\mathbf{X},\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\end{aligned}$$

so,

$$\frac{d}{d\theta}\ell(\theta) = \frac{d}{d\theta}\langle \log P(\mathbf{Y},\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} - \frac{d}{d\theta}\langle \log P(\mathbf{Y}|\mathbf{X},\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}$$

The second term is 0 at θ^* if the derivative exists (minimum of **KL**$[\cdot\|\cdot]$), and thus:

$$\frac{d}{d\theta}\ell(\theta)\Big|_{\theta^*} = \frac{d}{d\theta}\langle \log P(\mathbf{Y},\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\Big|_{\theta^*} = 0$$

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial\theta}\langle\log P(\mathbf{Y},\mathbf{X}\mid\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\Big|_{\theta^*} = 0$$

Now,

$$\begin{aligned}\ell(\theta) &= \log P(\mathbf{X}|\theta) = \langle\log P(\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} \\ &= \left\langle\log\frac{P(\mathbf{Y},\mathbf{X}|\theta)}{P(\mathbf{Y}|\mathbf{X},\theta)}\right\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} \\ &= \langle\log P(\mathbf{Y},\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} - \langle\log P(\mathbf{Y}|\mathbf{X},\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\end{aligned}$$

so,

$$\frac{d}{d\theta}\ell(\theta) = \frac{d}{d\theta}\langle\log P(\mathbf{Y},\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} - \frac{d}{d\theta}\langle\log P(\mathbf{Y}|\mathbf{X},\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}$$

The second term is 0 at θ^* if the derivative exists (minimum of **KL**$[\cdot\|\cdot]$), and thus:

$$\frac{d}{d\theta}\ell(\theta)\Big|_{\theta^*} = \frac{d}{d\theta}\langle\log P(\mathbf{Y},\mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}\Big|_{\theta^*} = 0$$

So, EM converges to a stationary point of $\ell(\theta)$.

Maxima in $\mathcal{F}$ correspond to maxima in ℓ

Let θ^* now be the parameter value at a local maximum of $\mathcal{F}$ (and thus at a fixed point)

Maxima in $\mathcal{F}$ correspond to maxima in ℓ

Let θ^* now be the parameter value at a local maximum of $\mathcal{F}$ (and thus at a fixed point)

Differentiating the previous expression wrt θ again we find

$$\frac{d^2}{d\theta^2}\ell(\theta) = \frac{d^2}{d\theta^2}\langle \log P(\mathbf{Y}, \mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} - \frac{d^2}{d\theta^2}\langle \log P(\mathbf{Y}|\mathbf{X}, \theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}$$

Maxima in $\mathcal{F}$ correspond to maxima in ℓ

Let θ^* now be the parameter value at a local maximum of $\mathcal{F}$ (and thus at a fixed point)

Differentiating the previous expression wrt θ again we find

$$\frac{d^2}{d\theta^2}\ell(\theta) = \frac{d^2}{d\theta^2}\langle \log P(\mathbf{Y}, \mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} - \frac{d^2}{d\theta^2}\langle \log P(\mathbf{Y}|\mathbf{X}, \theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}$$

The first term on the right is negative (a maximum) and the second term is positive (a minimum).

Maxima in $\mathcal{F}$ correspond to maxima in ℓ

Let θ^* now be the parameter value at a local maximum of $\mathcal{F}$ (and thus at a fixed point)

Differentiating the previous expression wrt θ again we find

$$\frac{d^2}{d\theta^2}\ell(\theta) = \frac{d^2}{d\theta^2}\langle \log P(\mathbf{Y}, \mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} - \frac{d^2}{d\theta^2}\langle \log P(\mathbf{Y}|\mathbf{X}, \theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}$$

The first term on the right is negative (a maximum) and the second term is positive (a minimum). Thus the curvature of the likelihood is negative and

θ^* is a maximum of ℓ.

Maxima in $\mathcal{F}$ correspond to maxima in ℓ

Let θ^* now be the parameter value at a local maximum of $\mathcal{F}$ (and thus at a fixed point)

Differentiating the previous expression wrt θ again we find

$$\frac{d^2}{d\theta^2}\ell(\theta) = \frac{d^2}{d\theta^2}\langle \log P(\mathbf{Y}, \mathbf{X}|\theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)} - \frac{d^2}{d\theta^2}\langle \log P(\mathbf{Y}|\mathbf{X}, \theta)\rangle_{P(\mathbf{Y}|\mathbf{X},\theta^*)}$$

The first term on the right is negative (a maximum) and the second term is positive (a minimum). Thus the curvature of the likelihood is negative and

θ^* is a maximum of ℓ.

[. . . as long as the derivatives exist. They sometimes don't (zero-noise ICA)].

Partial M steps and Partial E steps

Partial M steps: The proof holds even if we just *increase* $\mathcal{F}$ wrt θ rather than maximize. (Dempster, Laird and Rubin (1977) call this the generalized EM, or GEM, algorithm).

Partial E steps: We can also just *increase* $\mathcal{F}$ wrt to some of the qs.

For example, sparse or online versions of the EM algorithm would compute the posterior for a subset of the data points or as the data arrives, respectively. You can also update the posterior over a subset of the hidden variables, while holding others fixed...

Failure Modes of EM

EM can fail under a number of degenerate situations:

- EM may converge to a bad local maximum.
- Likelihood function may not be bounded above. E.g. a cluster responsible for a single data item can given arbitrarily large likelihood if variance $\sigma_m \to 0$.
- Free energy may not be well defined (or is $-\infty$).

EM for Exponential Families

Defn: P is in the exponential family for Y, X if it can be written:

$$P(Y, X|\theta) = h(Y, X) \exp\{\theta^\top \mathbf{T}(Y, X)\}/Z(\theta)$$

where $Z(\theta) = \int h(Y, X) \exp\{\theta^\top \mathbf{T}(Y, X)\} d(Y, X)$

E step: $q(Y) = P(Y|X, \theta)$

M step: $\theta^{(k)} := \underset{\theta}{\text{argmax}} \ \mathcal{F}(q, \theta)$

$$\begin{aligned} \mathcal{F}(q, \theta) &= \int q(Y) \log P(Y, X|\theta) dY - \mathbf{H}[q] \\ &= \int q(Y)[\theta^\top \mathbf{T}(Y, X) - \log Z(\theta)] dY + \text{const} \end{aligned}$$

It is easy to verify that: $\dfrac{\partial \log Z(\theta)}{\partial \theta} = E_{P(Y,X|\theta)}[\mathbf{T}(Y, X)]$

Therefore, M step solves: $\dfrac{\partial \mathcal{F}}{\partial \theta} = E_{q(Y)}[\mathbf{T}(Y, X)] - E_{P(Y,X|\theta)}[\mathbf{T}(Y, X)] = 0$

The Central Role of the Partition Function

The partition function $Z(\theta)$ of exponential families plays an important role in inference and learning of such models.

- Undirected graphical models are exponential families if each factor in the model has an exponential family form:

$$f_i(Y_{C_i}, X_{C_i}) = h_i(Y_{C_i}, X_{C_i}) \exp\{\theta_i^\top \mathbf{T}_i(Y_{C_i}, X_{C_i})\}$$

$$P(Y, X|\theta) = \frac{1}{Z(\theta)} \prod_i h_i(Y_{C_i}, X_{C_i}) \exp\left\{\sum_i \theta_i^\top \mathbf{T}_i(Y_{C_i}, X_{C_i})\right\}$$

- Likelihoods $P(X|\theta)$ are basically partition functions of undirected graphical models.
- Derivatives give the sufficient statistics of the models:

$$\nabla \log Z(\theta) = \mu = E_{P(Y,X|\theta)}[\mathbf{T}(Y, X)]$$

- Second derivatives give the covariance of sufficient statistics:

$$\nabla^2 \log Z(\theta) = E_{P(Y,X|\theta)}[(\mathbf{T}(Y, X) - \mu)(\mathbf{T}(Y, X) - \mu)^\top]$$

- Higher order derivatives give all cumulants, so $\log Z(\theta)$ is the cumulant generative function of the exponential family distribution.
- Many approximate inference techniques are based on approximating $\log Z(\theta)$.

End Notes

A. P. Dempster, N. M. Laird and D. B. Rubin (1977).
Maximum Likelihood from Incomplete Data via the EM Algorithm.
Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1 (1977), pp. 1-38.
http://www.jstor.org/stable/2984875

R. M. Neal and G. E. Hinton (1998).
A view of the EM algorithm that justifies incremental, sparse, and other variants.
In M. I. Jordan (editor) Learning in Graphical Models, pp. 355-368, Dordrecht: Kluwer Academic Publishers.
http://www.cs.utoronto.ca/ radford/ftp/emk.pdf

Z. Ghahramani and G. E. Hinton (1996).
The EM Algorithm for Mixtures of Factor Analyzers.
University of Toronto Technical Report CRG-TR-96-1.
http://learning.eng.cam.ac.uk/zoubin/papers/tr-96-1.pdf

End Notes