Probabilistic & Unsupervised Learning

Expectation Maximisation

Maneesh Sahani
maneesh@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit, and
MSc ML/CSML, Dept Computer Science
University College London

Term 1, Autumn 2014
Log-likelihoods

- Exponential family models: $p(x|\theta) = f(x)e^{\theta^T T(x)} / Z(\theta)$

\[
\ell(\theta) = \theta^T \sum_n T(x_n) - N \log Z(\theta) \quad (+\text{ constants})
\]
Log-likelihoods

- Exponential family models: \( p(\mathbf{x}|\theta) = f(\mathbf{x})e^{\theta^T\mathbf{T}(\mathbf{x})}/Z(\theta) \)

\[
\ell(\theta) = \theta^T \sum_n T(\mathbf{x}_n) - N \log Z(\theta) \quad (+ \text{constants})
\]

- Concave function.
Log-likelihoods

- Exponential family models: \( p(x|\theta) = f(x)e^{\theta^T T(x)}/Z(\theta) \)

\[
\ell(\theta) = \theta^T \sum_n T(x_n) - N \log Z(\theta) \ (\text{+ constants})
\]

- Concave function.
- Maximum may be closed-form.
Log-likelihoods

- Exponential family models: 
  \[ p(x|\theta) = f(x)e^{\theta^T T(x)}/Z(\theta) \]

  \[ \ell(\theta) = \theta^T \sum_{n} T(x_n) - N \log Z(\theta) \quad (+ \text{constants}) \]

  - Concave function.
  - Maximum may be closed-form.
  - If not, numerical optimisation is still generally straightforward.
Log-likelihoods

- Exponential family models: 
  \[ p(x|\theta) = f(x)e^{\theta^T T(x)} / Z(\theta) \]
  \[ \ell(\theta) = \theta^T \sum_n T(x_n) - N \log Z(\theta) \]  (+ constants)

  - Concave function.
  - Maximum may be closed-form.
  - If not, numerical optimisation is still generally straightforward.

- Latent variable models: 
  \[ p(x|\theta_x, \theta_y) = \int dy \, f_x(x) \frac{e^{\phi(\theta_x, y)^T T_x(x)}}{Z_x(\phi(\theta_x, y))} \frac{f_y(y) e^{\theta_y^T T_y(y)}}{Z_y(\theta_y)} \]
  \[ \ell(\theta_x, \theta_y) = \sum \log \int dy \, f_x(x) \frac{e^{\phi(\theta_x, y)^T T_x(x)}}{Z_x(\phi(\theta_x, y))} \frac{f_y(y) e^{\theta_y^T T_y(y)}}{Z_y(\theta_y)} \]
Log-likelihoods

► Exponential family models: \( p(x|\theta) = f(x)e^{\theta^T T(x)} / Z(\theta) \)

\[ \ell(\theta) = \theta^T \sum_n T(x_n) - N \log Z(\theta) \text{ (+ constants)} \]

► Concave function.
► Maximum may be closed-form.
► If not, numerical optimisation is still generally straightforward.

► Latent variable models: \( p(x|\theta_x, \theta_y) = \int dy \, f_x(x) \frac{e^{\phi(\theta_x, y)^T T_x(x)}}{Z_x(\phi(\theta_x, y))} \frac{f_y(y) e^{\theta_y^T T_y(y)}}{Z_y(\theta_y)} \)

\[ \ell(\theta_x, \theta_y) = \sum_n \log \int dy \, f_x(x) \frac{e^{\phi(\theta_x, y)^T T_x(x)}}{Z_x(\phi(\theta_x, y))} f_y(y) \frac{e^{\theta_y^T T_y(y)}}{Z_y(\theta_y)} \]

► Usually no closed form optimum.
Log-likelihoods

- Exponential family models: 
  \[ p(x|\theta) = f(x)e^{\theta^T T(x)}/Z(\theta) \]
  \[ \ell(\theta) = \theta^T \sum_n T(x_n) - N \log Z(\theta) \text{ (with + constants)} \]

  - Concave function.
  - Maximum may be closed-form.
  - If not, numerical optimisation is still generally straightforward.

- Latent variable models: 
  \[ p(x|\theta_x, \theta_y) = \int dy f_x(x) \frac{e^{\phi(\theta_x, y)^T T_x(x)}}{Z_x(\phi(\theta_x, y))} f_y(y) \frac{e^{\theta_y^T T_y(y)}}{Z_y(\theta_y)} \]
  \[ \ell(\theta_x, \theta_y) = \sum_n \log \int dy f_x(x) \frac{e^{\phi(\theta_x, y)^T T_x(x)}}{Z_x(\phi(\theta_x, y))} f_y(y) \frac{e^{\theta_y^T T_y(y)}}{Z_y(\theta_y)} \]

  - Usually no closed form optimum.
  - Often multiple local maxima.
Log-likelihoods

- Exponential family models:
  \[ p(x|\theta) = f(x)e^{\theta^T T(x)} / Z(\theta) \]
  \[ \ell(\theta) = \theta^T \sum_n T(x_n) - N \log Z(\theta) \text{ (+ constants)} \]
  - Concave function.
  - Maximum may be closed-form.
  - If not, numerical optimisation is still generally straightforward.

- Latent variable models:
  \[ p(x|\theta_x, \theta_y) = \int dy \ f_x(x) \frac{e^{\phi(\theta_x, y)^T T_x(x)}}{Z_x(\phi(\theta_x, y))} \ f_y(y) \frac{e^{\theta_y^T T_y(y)}}{Z_y(\theta_y)} \]
  \[ \ell(\theta_x, \theta_y) = \sum_n \log \int dy \ f_x(x) \frac{e^{\phi(\theta_x, y)^T T_x(x)}}{Z_x(\phi(\theta_x, y))} \ f_y(y) \frac{e^{\theta_y^T T_y(y)}}{Z_y(\theta_y)} \]
  - Usually no closed form optimum.
  - Often multiple local maxima.
  - Direct numerical optimisation may be possible but infrequently easy.
Example: mixture of Gaussians

Data: \( \mathcal{X} = \{ \mathbf{x}_1 \ldots \mathbf{x}_N \} \)

Latent process:
\( s_i \overset{iid}{\sim} \text{Disc}[\pi] \)

Component distributions:
\( \mathbf{x}_i \mid (s_i = m) \sim \mathcal{P}_m[\theta_m] = \mathcal{N}(\mu_m, \Sigma_m) \)

Marginal distribution:
\[
P(\mathbf{x}_i) = \sum_{m=1}^{k} \pi_m P_m(\mathbf{x}; \theta_m)
\]

Log-likelihood:
\[
l(\{\mu_m\}, \{\Sigma_m\}, \pi) = \sum_{i=1}^{n} \log \sum_{m=1}^{k} \frac{\pi_m}{\sqrt{2\pi |\Sigma_m|}} e^{-\frac{1}{2}(\mathbf{x}_i - \mu_m)^\top \Sigma_m^{-1}(\mathbf{x}_i - \mu_m)}
\]
The joint-data likelihood

For many models, maximisation might be straightforward if $y$ were not latent, and we could just maximise the joint-data likelihood:

$$
\ell(\theta_x, \theta_y) = \sum_n \phi(\theta_x, y_n)^T T_x(x_n) + \theta_y^T T_y(y_n) - \sum_n \log Z_x(\phi(\theta_x, y_n)) - N \log Z_y(\theta_y)
$$
The joint-data likelihood

For many models, maximisation might be straightforward if $y$ were not latent, and we could just maximise the joint-data likelihood:

$$\ell(\theta_x, \theta_y) = \sum_n \phi(\theta_x, y_n)^T T_x(x_n) + \theta_y^T T_y(y_n) - \sum_n \log Z_x(\phi(\theta_x, y_n)) - N \log Z_y(\theta_y)$$

Conversely, if we knew $\theta$, we could compute (the posterior over) the values of $y$. 
The joint-data likelihood

- For many models, maximisation might be straightforward if $y$ were not latent, and we could just maximise the joint-data likelihood:

$$\ell(\theta_x, \theta_y) = \sum_n \phi(\theta_x, y_n)^T T_x(x_n) + \theta_y^T T_y(y_n) - \sum_n \log Z_x(\phi(\theta_x, y_n)) - N \log Z_y(\theta_y)$$

- Conversely, if we knew $\theta$, we could compute (the posterior over) the values of $y$.

- **Idea**: update $\theta$ and (the distribution on) $y$ in alternation, converging to a self-consistent answer.
The joint-data likelihood

- For many models, maximisation might be straightforward if $y$ were not latent, and we could just maximise the joint-data likelihood:

$$
\ell(\theta_x, \theta_y) = \sum_n \phi(\theta_x, y_n)^T T_x(x_n) + \theta_y^T T_y(y_n) - \sum_n \log Z_x(\phi(\theta_x, y_n)) - N \log Z_y(\theta_y)
$$

- Conversely, if we knew $\theta$, we could compute (the posterior over) the values of $y$.

- Idea: update $\theta$ and (the distribution on) $y$ in alternation, converging to a self-consistent answer.

- Will this yield the right answer?
The joint-data likelihood

- For many models, maximisation might be straightforward if \( y \) were not latent, and we could just maximise the joint-data likelihood:

\[
\ell(\theta_x, \theta_y) = \sum_n \phi(\theta_x, y_n)^T T_x(x_n) + \theta_y^T T_y(y_n) - \sum_n \log Z_x(\phi(\theta_x, y_n)) - N \log Z_y(\theta_y)
\]

- Conversely, if we knew \( \theta \), we could compute (the posterior over) the values of \( y \).

- Idea: update \( \theta \) and (the distribution on) \( y \) in alternation, converging to a self-consistent answer.

- Will this yield the right answer?

- Typically, it will (as we shall see). This is the Expectation Maximisation (EM) algorithm.
The Expectation Maximisation (EM) algorithm

The EM algorithm (Dempster, Laird & Rubin, 1977; but significant earlier precedents) finds a (local) maximum of a latent variable model likelihood. 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.
The Expectation Maximisation (EM) algorithm

The EM algorithm (Dempster, Laird & Rubin, 1977; but significant earlier precedents) finds a (local) maximum of a latent variable model likelihood. 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 hidden variables were, in fact, observed (e.g. MoGs).
The Expectation Maximisation (EM) algorithm

The EM algorithm (Dempster, Laird & Rubin, 1977; but significant earlier precedents) finds a (local) maximum of a latent variable model likelihood. 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 hidden variables were, in fact, observed (e.g. MoGs).
- Decomposes difficult problems into series of tractable steps.
The Expectation Maximisation (EM) algorithm

The EM algorithm (Dempster, Laird & Rubin, 1977; but significant earlier precedents) finds a (local) maximum of a latent variable model likelihood. 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 hidden variables were, in fact, observed (e.g. MoGs).
- Decomposes difficult problems into series of tractable steps.
- No learning rate.
The Expectation Maximisation (EM) algorithm

The EM algorithm (Dempster, Laird & Rubin, 1977; but significant earlier precedents) finds a (local) maximum of a latent variable model likelihood. 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 hidden variables were, in fact, observed (e.g. MoGs).
- Decomposes difficult problems into series of tractable steps.
- No learning rate.
- **Framework lends itself to principled approximations.**
The Expectation Maximisation (EM) algorithm

The EM algorithm (Dempster, Laird & Rubin, 1977; but significant earlier precedents) finds a (local) maximum of a latent variable model likelihood. 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 hidden variables were, in fact, observed (e.g. MoGs).
- Decomposes difficult problems into series of tractable steps.
- No learning rate.
- Framework lends itself to principled approximations.

How does it work?
Jensen’s inequality

One view: EM iteratively refines a lower bound on the log-likelihood.
Jensen’s inequality

One view: EM iteratively refines a lower bound on the log-likelihood.

\[
\log(x) = \alpha x_1 + (1 - \alpha) x_2
\]

In general:

For \( \alpha_i \geq 0, \sum \alpha_i = 1 \) (and \( \{x_i > 0\} \)):

\[
\log(\sum \alpha_i x_i) \geq \sum \alpha_i \log(x_i)
\]

For probability measure \( \alpha \) and concave \( f \):

\[
f(\mathbb{E}_\alpha[x]) \geq \mathbb{E}_\alpha[f(x)]
\]

Equality (if and) only if \( f(x) \) is almost surely constant or linear on (convex) support of \( \alpha \).
Jensen’s inequality

One view: EM iteratively refines a lower bound on the log-likelihood.
Jensen’s inequality

One view: EM iteratively refines a lower bound on the log-likelihood.

\[
\log(x) = \alpha x_1 + (1 - \alpha) x_2
\]

In general:

For \( \alpha_i \geq 0 \), \( \sum \alpha_i = 1 \) (and \( \{ x_i > 0 \} \)):

\[
\log \left( \sum_i \alpha_i x_i \right) \geq \sum_i \alpha_i \log(x_i)
\]

For probability measure \( \alpha \) and concave \( f \):

\[
\mathbb{E}_\alpha[f(x)] \geq f \left( \mathbb{E}_\alpha[x] \right)
\]

Equality (if and) only if \( f(x) \) is almost surely constant or linear on (convex) support of \( \alpha \).
Jensen’s inequality

One view: EM iteratively refines a lower bound on the log-likelihood.

In general:

For $\alpha_i \geq 0, \sum \alpha_i = 1$ (and $\{x_i > 0\}$):

$$\log(\sum \alpha_i x_i) \geq \sum \alpha_i \log(x_i)$$

For probability measure $\alpha$ and concave $f$:

$$f(\mathbb{E}_\alpha [x]) \geq \mathbb{E}_\alpha [f(x)]$$

Equality (if and) only if $f(x)$ is almost surely constant or linear on (convex) support of $\alpha$. 
Jensen’s inequality

One view: EM iteratively refines a lower bound on the log-likelihood.

In general:

\[
\log(\alpha x_1 + (1 - \alpha)x_2) \geq \alpha \log(x_1) + (1 - \alpha) \log(x_2)
\]
Jensen’s inequality

One view: EM iteratively refines a lower bound on the log-likelihood.

In general:

For \( \alpha_i \geq 0, \sum \alpha_i = 1 \) (and \( \{x_i > 0\} \)):

\[
\log \left( \sum_i \alpha_i x_i \right) \geq \sum_i \alpha_i \log(x_i)
\]
Jensen’s inequality

One view: EM iteratively refines a lower bound on the log-likelihood.

\[ \log(\alpha x_1 + (1 - \alpha)x_2) \]
\[ \alpha \log(x_1) + (1 - \alpha)\log(x_2) \]

In general:

For \( \alpha_i \geq 0, \sum \alpha_i = 1 \) (and \( \{x_i > 0\} \)):

\[ \log \left( \sum_i \alpha_i x_i \right) \geq \sum_i \alpha_i \log(x_i) \]

For probability measure \( \alpha \) and concave \( f \)

\[ f(\mathbb{E}_\alpha [x]) \geq \mathbb{E}_\alpha [f(x)] \]
Jensen’s inequality

One view: EM iteratively refines a lower bound on the log-likelihood.

In general:

For \( \alpha_i \geq 0, \sum \alpha_i = 1 \) (and \( \{x_i > 0\} \)): 

\[
\log \left( \sum_i \alpha_i x_i \right) \geq \sum_i \alpha_i \log(x_i)
\]

For probability measure \( \alpha \) and concave \( f \): 

\[
f(\mathbb{E}_\alpha [x]) \geq \mathbb{E}_\alpha [f(x)]
\]

Equality (if and) only if \( f(x) \) is almost surely constant or linear on (convex) support of \( \alpha \).
The lower bound for EM – “free energy”

Observed data \( \mathcal{X} = \{x_i\} \); Latent variables \( \mathcal{Y} = \{y_i\} \); Parameters \( \theta = \{\theta_x, \theta_y\} \).

Log-likelihood:
\[
\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int d\mathcal{Y} P(\mathcal{Y}, \mathcal{X}|\theta)
\]
The lower bound for EM – “free energy”

Observed data $\mathcal{X} = \{x_i\}$; Latent variables $\mathcal{Y} = \{y_i\}$; Parameters $\theta = \{\theta_x, \theta_y\}$.

Log-likelihood:
$$\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int d\mathcal{Y} P(\mathcal{Y}, \mathcal{X}|\theta)$$

By Jensen, any distribution, $q(\mathcal{Y})$, over the latent variables generates a lower bound:
$$\ell(\theta) = \log \int d\mathcal{Y} P(\mathcal{Y}, \mathcal{X}|\theta)$$
The lower bound for EM – “free energy”

Observed data $\mathcal{X} = \{x_i\}$; Latent variables $\mathcal{Y} = \{y_i\}$; Parameters $\theta = \{\theta_x, \theta_y\}$.

Log-likelihood:
$$\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int d\mathcal{Y} P(\mathcal{Y}, \mathcal{X}|\theta)$$

By Jensen, any distribution, $q(\mathcal{Y})$, over the latent variables generates a lower bound:
$$\ell(\theta) = \log \int d\mathcal{Y} \; q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} = \mathbb{E}_{q(\mathcal{Y})} \left[ \log P(\mathcal{Y}, \mathcal{X}|\theta) \right] + H[q(\mathcal{Y})]$$

where $H[q(\mathcal{Y})]$ is the entropy of $q(\mathcal{Y})$. So:
$$F(q, \theta) = \mathbb{E}_{q(\mathcal{Y})} \left[ \log P(\mathcal{Y}, \mathcal{X}|\theta) \right] + H[q(\mathcal{Y})]$$
The lower bound for EM – “free energy”

Observed data $\mathcal{X} = \{x_i\}$; Latent variables $\mathcal{Y} = \{y_i\}$; Parameters $\theta = \{\theta_x, \theta_y\}$.

Log-likelihood:
$$
\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int d\mathcal{Y} \; P(\mathcal{Y}, \mathcal{X}|\theta)
$$

By Jensen, any distribution, $q(\mathcal{Y})$, over the latent variables generates a lower bound:
$$
\ell(\theta) = \log \int d\mathcal{Y} \; q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \geq \int d\mathcal{Y} \; q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})}
$$
The lower bound for EM – “free energy”

Observed data $\mathcal{X} = \{x_i\}$; Latent variables $\mathcal{Y} = \{y_i\}$; Parameters $\theta = \{\theta_x, \theta_y\}$.

Log-likelihood:
\[
\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int d\mathcal{Y} P(\mathcal{Y}, \mathcal{X}|\theta)
\]

By Jensen, any distribution, $q(\mathcal{Y})$, over the latent variables generates a lower bound:
\[
\ell(\theta) = \log \int d\mathcal{Y} q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \geq \int d\mathcal{Y} q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \overset{\text{def}}{=} \mathcal{F}(q, \theta).
\]
The lower bound for EM – “free energy”

Observed data $\mathcal{X} = \{x_i\}$; Latent variables $\mathcal{Y} = \{y_i\}$; Parameters $\theta = \{\theta_x, \theta_y\}$.

Log-likelihood:
$$\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int d\mathcal{Y} P(\mathcal{Y}, \mathcal{X}|\theta)$$

By Jensen, any distribution, $q(\mathcal{Y})$, over the latent variables generates a lower bound:
$$\ell(\theta) = \log \int d\mathcal{Y} q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \geq \int d\mathcal{Y} q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \overset{\text{def}}{=} \mathcal{F}(q, \theta).$$

Now,
$$\int d\mathcal{Y} q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} = \int d\mathcal{Y} q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) - \int d\mathcal{Y} q(\mathcal{Y}) \log q(\mathcal{Y})$$
The lower bound for EM – “free energy”

Observed data $\mathcal{X} = \{x_i\}$; Latent variables $\mathcal{Y} = \{y_i\}$; Parameters $\theta = \{\theta_x, \theta_y\}$.

Log-likelihood:
\[
\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int d\mathcal{Y} P(\mathcal{Y}, \mathcal{X}|\theta)
\]

By Jensen, any distribution, $q(\mathcal{Y})$, over the latent variables generates a lower bound:
\[
\ell(\theta) = \log \int d\mathcal{Y} q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \geq \int d\mathcal{Y} q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \overset{\text{def}}{=} \mathcal{F}(q, \theta).
\]

Now,
\[
\int d\mathcal{Y} q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} = \int d\mathcal{Y} q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) - \int d\mathcal{Y} q(\mathcal{Y}) \log q(\mathcal{Y})
\]
\[
= \int d\mathcal{Y} q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) + H[q],
\]
where $H[q]$ is the entropy of $q(\mathcal{Y})$. 
The lower bound for EM – “free energy”

Observed data $\mathcal{X} = \{x_i\}$; Latent variables $\mathcal{Y} = \{y_i\}$; Parameters $\theta = \{\theta_x, \theta_y\}$.

Log-likelihood:
$$\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int d\mathcal{Y} P(\mathcal{Y}, \mathcal{X}|\theta)$$

By Jensen, any distribution, $q(\mathcal{Y})$, over the latent variables generates a lower bound:
$$\ell(\theta) = \log \int d\mathcal{Y} q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \geq \int d\mathcal{Y} q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \overset{\text{def}}{=} \mathcal{F}(q, \theta).$$

Now,
$$\int d\mathcal{Y} q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} = \int d\mathcal{Y} q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) - \int d\mathcal{Y} q(\mathcal{Y}) \log q(\mathcal{Y})$$
$$= \int d\mathcal{Y} q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) + H[q],$$

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

So:
$$\mathcal{F}(q, \theta) = \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{q(\mathcal{Y})} + H[q]$$
The E and M steps of EM

The lower bound on the log likelihood is given by:

\[ F(q, \theta) = \langle \log P(Y, X|\theta) \rangle_{q(Y)} + H[q], \]
The E and M steps of EM

The lower bound on the log likelihood is given by:

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

EM alternates between:

- **E step**: optimize \( F(q, \theta) \) wrt distribution over hidden variables holding parameters fixed:

  \[ q^{(k)}(Y) := \arg \max_{q(Y)} F(q(Y), \theta^{(k-1)}). \]
The E and M steps of EM

The lower bound on the log likelihood is given by:

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

EM alternates between:

- **E step**: optimize \( \mathcal{F}(q, \theta) \) wrt distribution over hidden variables holding parameters fixed:
  \[ q^{(k)}(Y) := \arg\max_{q(Y)} \mathcal{F}(q(Y), \theta^{(k-1)}) \]

- **M step**: maximize \( \mathcal{F}(q, \theta) \) wrt parameters holding hidden distribution fixed:
  \[ \theta^{(k)} := \arg\max_{\theta} \mathcal{F}(q^{(k)}(Y), \theta) = \arg\max_{\theta} \langle \log P(Y, \mathcal{X}|\theta) \rangle_{q^{(k)}(Y)} \]

The second equality comes from the fact \( H[q^{(k)}(Y)] \) does not depend directly on \( \theta \).
The E Step

The free energy can be re-written

\[ F(q, \theta) = \int q(Y) \log P(Y|X, \theta) q(Y) dY = \int q(Y) \log P(X|\theta) dY + \int q(Y) \log P(Y|X, \theta) q(Y) dY = \ell(\theta) - KL[q(Y) \parallel P(Y|X, \theta)] \]

The second term is the Kullback-Leibler divergence. This means that, for fixed \( \theta \), \( F \) is bounded above by \( \ell(\theta) \), and achieves that bound when \( KL[q(Y) \parallel P(Y|X, \theta)] = 0 \). But \( KL[q \parallel p] \) is zero if and only if \( q = p \) (see appendix.) So, the E step sets \( q^{(k)}(Y) = P(Y|X, \theta^{(k-1)}) \) and, after an E step, the free energy equals the likelihood.
The E Step

The free energy can be re-written

\[
\mathcal{F}(q, \theta) = \int q(Y) \log \frac{P(Y, X|\theta)}{q(Y)} \, dY
\]
The E Step

The free energy can be re-written

\[ \mathcal{F}(q, \theta) = \int q(Y) \log \frac{P(Y, X|\theta)}{q(Y)} \, dY \]

\[ = \int q(Y) \log \frac{P(Y|X, \theta)P(X|\theta)}{q(Y)} \, dY \]

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

But \( KL[q \parallel p] \) is zero if and only if \( q = p \) (see appendix.). So, the E step sets \( q(Y)^{(k)} = P(Y|X, \theta^{(k-1)}) \) and, after an E step, the free energy equals the likelihood.
The E Step

The free energy can be re-written

\[
\mathcal{F}(q, \theta) = \int q(Y) \log \frac{P(Y, X|\theta)}{q(Y)} dY
\]

\[
= \int q(Y) \log \frac{P(Y|X, \theta)P(X|\theta)}{q(Y)} dY
\]

\[
= \int q(Y) \log P(X|\theta) dY + \int q(Y) \log \frac{P(Y|X, \theta)}{q(Y)} dY
\]

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

But \(KL[q \parallel p]\) is zero if and only if \(q = p\) (see appendix.)

So, the E step sets \(q(Y)^{(k)} = P(Y|X, \theta^{(k-1)})\) and, after an E step, the free energy equals the likelihood.
The E Step

The free energy can be re-written

\[
\mathcal{F}(q, \theta) = \int q(Y) \log \frac{P(Y, X|\theta)}{q(Y)} \, dY
\]

\[
= \int q(Y) \log \frac{P(Y|X, \theta) P(X|\theta)}{q(Y)} \, dY
\]

\[
= \int q(Y) \log P(X|\theta) \, dY + \int q(Y) \log \frac{P(Y|X, \theta)}{q(Y)} \, dY
\]

\[
= \ell(\theta) - \text{KL}[q(Y) || P(Y|X, \theta)]
\]

The second term is the Kullback-Leibler divergence.
The E Step

The free energy can be re-written

\[ \mathcal{F}(q, \theta) = \int q(Y) \log \frac{P(Y, \mathcal{X} | \theta)}{q(Y)} \, dY \]

\[ = \int q(Y) \log \frac{P(Y|\mathcal{X}, \theta)P(\mathcal{X}|\theta)}{q(Y)} \, dY \]

\[ = \int q(Y) \log P(\mathcal{X}|\theta) \, dY + \int q(Y) \log \frac{P(Y|\mathcal{X}, \theta)}{q(Y)} \, dY \]

\[ = \ell(\theta) - \text{KL}[q(Y) \| P(Y|\mathcal{X}, \theta)] \]

The second term is the Kullback-Leibler divergence.

This means that, for fixed \( \theta \), \( \mathcal{F} \) is bounded above by \( \ell \), and achieves that bound when

\[ \text{KL}[q(Y) \| P(Y|\mathcal{X}, \theta)] = 0. \]
The E Step

The free energy can be re-written

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

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

$$
= \int q(Y) \log P(\mathcal{X}|\theta) \, dY + \int q(Y) \log \frac{P(Y|\mathcal{X}, \theta)}{q(Y)} \, dY
$$

$$
= \ell(\theta) - KL[q(Y)\|P(\mathcal{Y}|\mathcal{X}, \theta)]
$$

The second term is the Kullback-Leibler divergence.

This means that, for fixed $\theta$, $F$ is bounded above by $\ell$, and achieves that bound when

$$
KL[q(Y)\|P(\mathcal{Y}|\mathcal{X}, \theta)] = 0.
$$

But $KL[q\|p]$ is zero if and only if $q = p$ (see appendix.)
The E Step

The free energy can be re-written

\[ F(q, \theta) = \int q(Y) \log \frac{P(Y, X|\theta)}{q(Y)} \, dY \]

\[ = \int q(Y) \log \frac{P(Y|X, \theta)P(X|\theta)}{q(Y)} \, dY \]

\[ = \int q(Y) \log P(X|\theta) \, dY + \int q(Y) \log \frac{P(Y|X, \theta)}{q(Y)} \, dY \]

\[ = \ell(\theta) - KL[q(Y)\|P(Y|X, \theta)] \]

The second term is the Kullback-Leibler divergence.

This means that, for fixed \( \theta \), \( F \) is bounded above by \( \ell \), and achieves that bound when \( KL[q(Y)\|P(Y|X, \theta)] = 0 \).

But \( KL[q\|p] \) is zero if and only if \( q = p \) (see appendix.)

So, the E step sets

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

and, after an E step, the free energy equals the likelihood.
Coordinate Ascent in $\mathcal{F}$ (Demo)

To visualise, we consider a one parameter / one latent mixture:

\[
\begin{align*}
    s & \sim \text{Bernoulli}[\pi] \\
    x|s=0 & \sim \mathcal{N}[-1, 1] \\
    x|s=1 & \sim \mathcal{N}[1, 1].
\end{align*}
\]

Single 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)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $F$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $F$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
Coordinate Ascent in $\mathcal{F}$ (Demo)
EM Never Decreases the Likelihood

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

\[ \ell(\theta^{(k-1)}) \]
EM Never Decreases the Likelihood

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

\[ \ell(\theta^{(k-1)}) = \mathcal{F}(q^{(k)}, \theta^{(k-1)}) \]

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

- The M-step maximises the free energy \( \theta \).

\[ \mathcal{F} \leq \ell \text{ by Jensen} \] – 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 \( \theta \) iff the likelihood increases.

Can also show that fixed points of EM (generally) correspond to maxima of the likelihood (see appendices).
EM Never Decreases the Likelihood

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

\[
\ell(\theta^{(k-1)}) = \mathcal{F}(q^{(k)}, \theta^{(k-1)}) \leq \mathcal{F}(q^{(k)}, \theta^{(k)})
\]

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt \( \theta \).
EM Never Decreases the Likelihood

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

\[
\ell(\theta^{(k-1)}) = \mathcal{F}(q^{(k)}, \theta^{(k-1)}) \leq \mathcal{F}(q^{(k)}, \theta^{(k)}) \leq \ell(\theta^{(k)}),
\]

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt $\theta$.
- $\mathcal{F} \leq \ell$ 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:

\[
\ell(\theta^{(k-1)}) = \mathcal{F}(q^{(k)}, \theta^{(k-1)}) \leq \mathcal{F}(q^{(k)}, \theta^{(k)}) \leq \ell(\theta^{(k)}),
\]

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt \( \theta \).
- \( \mathcal{F} \leq \ell \) 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 \( \theta \) iff the likelihood increases.
EM Never Decreases the Likelihood

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

\[
\ell(\theta^{(k-1)}) = \mathcal{F}(q^{(k)}, \theta^{(k-1)}) \leq \mathcal{F}(q^{(k)}, \theta^{(k)}) \leq \ell(\theta^{(k)}),
\]

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt \( \theta \).
- \( \mathcal{F} \leq \ell \) 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 \( \theta \) iff the likelihood increases.

Can also show that fixed points of EM (generally) correspond to maxima of the likelihood (see appendices).
EM Summary

- An iterative algorithm that finds (local) maxima of the likelihood of a latent variable model.

\[ \ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int d\mathcal{Y} P(\mathcal{X}|\mathcal{Y}, \theta)P(\mathcal{Y}|\theta) \]

- Increases a variational lower bound on the likelihood by coordinate ascent.

\[ \mathcal{F}(q, \theta) = \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{q(\mathcal{Y})} + H[q] = \ell(\theta) - \text{KL}[q(\mathcal{Y})||P(\mathcal{Y}|\mathcal{X})] \leq \ell(\theta) \]

- E step:

\[ q^{(k)}(\mathcal{Y}) := \arg\max_{q(\mathcal{Y})} \mathcal{F}(q(\mathcal{Y}), \theta^{(k-1)}) = P(\mathcal{Y}|\mathcal{X}, \theta^{(k-1)}) \]

- M step:

\[ \theta^{(k)} := \arg\max_{\theta} \mathcal{F}(q^{(k)}(\mathcal{Y}), \theta) = \arg\max_{\theta} \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{q^{(k)}(\mathcal{Y})} \]

- After E-step \( \mathcal{F}(q, \theta) = \ell(\theta) \Rightarrow \) maximum of free-energy is maximum of likelihood.
Partial M steps and Partial E steps

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

In fact, immediately after an E step

$$
\frac{\partial}{\partial \theta} \bigg|_{\theta^{(k-1)}} \langle \log P(\mathcal{X}, \mathcal{Y}|\theta) \rangle_{q^{(k)}(\mathcal{Y})[=P(\mathcal{Y}|\mathcal{X}, \theta^{(k-1)})]} = \frac{\partial}{\partial \theta} \bigg|_{\theta^{(k-1)}} \log P(\mathcal{X}|\theta)
$$

So E-step (inference) can be used to construct other gradient-based optimisation schemes (e.g. “Expectation Conjugate Gradient”, Salakhutdinov et al. *ICML* 2003).

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

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. One might also update the posterior over a subset of the hidden variables, while holding others fixed...
EM for MoGs

- Evaluate responsibilities

\[ r_{im} = \frac{P_m(x) \pi_m}{\sum_{m'} P_{m'}(x) \pi_{m'}} \]

- Update parameters

\[ \mu_m \leftarrow \frac{\sum_i r_{im} x_i}{\sum_i r_{im}} \]
\[ \Sigma_m \leftarrow \frac{\sum_i r_{im} (x_i - \mu_m)(x_i - \mu_m)^T}{\sum_i r_{im}} \]
\[ \pi_m \leftarrow \frac{\sum_i r_{im}}{N} \]
The Gaussian mixture model (E-step)

In a univariate Gaussian mixture model, the density of a data point $x$ is:

$$p(x|\theta) = \sum_{m=1}^{k} p(s = m|\theta)p(x|s = m, \theta) \propto \sum_{m=1}^{k} \pi_m \exp \left\{ - \frac{1}{2\sigma_m^2} (x - \mu_m)^2 \right\},$$

where $\theta$ is the collection of parameters: means $\mu_m$, variances $\sigma_m^2$ and mixing proportions $\pi_m = p(s = m|\theta)$.

The hidden variable $s_i$ indicates which component generated observation $x_i$. 
The Gaussian mixture model (E-step)

In a univariate Gaussian mixture model, the density of a data point $x$ is:

$$p(x|\theta) = \sum_{m=1}^{k} p(s = m|\theta)p(x|s = m, \theta) \propto \sum_{m=1}^{k} \frac{\pi_m}{\sigma_m} \exp \left\{ -\frac{1}{2\sigma_m^2} (x - \mu_m)^2 \right\},$$

where $\theta$ is the collection of parameters: means $\mu_m$, variances $\sigma_m^2$ and mixing proportions $\pi_m = p(s = m|\theta)$.

The hidden variable $s_i$ indicates which component generated observation $x_i$.

The E-step computes the posterior for $s_i$ given the current parameters:

$$q(s_i) = p(s_i|x, \theta) \propto p(x_i|s_i, \theta)p(s_i|\theta)$$
The Gaussian mixture model (E-step)

In a univariate Gaussian mixture model, the density of a data point $x$ is:

$$p(x|\theta) = \sum_{m=1}^{k} p(s = m|\theta)p(x|s = m, \theta) \propto \sum_{m=1}^{k} \frac{\pi_m}{\sigma_m} \exp \left\{ -\frac{1}{2\sigma_m^2} (x - \mu_m)^2 \right\},$$

where $\theta$ is the collection of parameters: means $\mu_m$, variances $\sigma_m^2$ and mixing proportions $\pi_m = p(s = m|\theta)$.

The hidden variable $s_i$ indicates which component generated observation $x_i$.

The E-step computes the posterior for $s_i$ given the current parameters:

$$q(s_i) = p(s_i|x_i, \theta) \propto p(x_i|s_i, \theta)p(s_i|\theta)$$

$$q(s_i = m) \propto \frac{\pi_m}{\sigma_m} \exp \left\{ -\frac{1}{2\sigma_m^2} (x_i - \mu_m)^2 \right\}$$

with the normalization such that $\sum_m r_{im} = 1$. 
The Gaussian mixture model (E-step)

In a univariate Gaussian mixture model, the density of a data point $x$ is:

$$p(x|\theta) = \sum_{m=1}^{k} p(s = m|\theta)p(x|s = m, \theta) \propto \sum_{m=1}^{k} \frac{\pi_m}{\sigma_m} \exp \left\{ -\frac{1}{2\sigma_m^2} (x - \mu_m)^2 \right\},$$

where $\theta$ is the collection of parameters: means $\mu_m$, variances $\sigma_m^2$ and mixing proportions $\pi_m = p(s = m|\theta)$.

The hidden variable $s_i$ indicates which component generated observation $x_i$.

The E-step computes the posterior for $s_i$ given the current parameters:

$$q(s_i) = p(s_i|x_i, \theta) \propto p(x_i|s_i, \theta)p(s_i|\theta)$$

$$r_{im} \overset{\text{def}}{=} q(s_i = m) \propto \frac{\pi_m}{\sigma_m} \exp \left\{ -\frac{1}{2\sigma_m^2} (x_i - \mu_m)^2 \right\} \quad \text{(responsibilities)}$$

with the normalization such that $\sum_m r_{im} = 1$. 
The Gaussian mixture model (E-step)

In a univariate Gaussian mixture model, the density of a data point $x$ is:

$$p(x|\theta) = \sum_{m=1}^{k} p(s = m|\theta)p(x|s = m, \theta) \propto \sum_{m=1}^{k} \frac{\pi_m}{\sigma_m} \exp\left\{ -\frac{1}{2\sigma_m^2}(x - \mu_m)^2 \right\},$$

where $\theta$ is the collection of parameters: means $\mu_m$, variances $\sigma_m^2$ and mixing proportions $\pi_m = p(s = m|\theta)$.

The hidden variable $s_i$ indicates which component generated observation $x_i$.

The E-step computes the posterior for $s_i$ given the current parameters:

$$q(s_i) = p(s_i|x_i, \theta) \propto p(x_i|s_i, \theta)p(s_i|\theta)$$

$$r_{im} \stackrel{\text{def}}{=} q(s_i = m) \propto \frac{\pi_m}{\sigma_m} \exp\left\{ -\frac{1}{2\sigma_m^2}(x_i - \mu_m)^2 \right\} \quad \text{(responsibilities)} \leftarrow \langle \delta_{s_i=m} \rangle_q$$

with the normalization such that $\sum_m r_{im} = 1$. 
The Gaussian mixture model (M-step)

In the M-step we optimize the sum (since $s$ is discrete):

\[
E = \langle \log p(x, s|\theta) \rangle_{q(s)} = \sum q(s) \log[p(s|\theta) \ p(x|s, \theta)]
\]

\[
= \sum_{i,m} r_{im} \left[ \log \pi_m - \log \sigma_m - \frac{1}{2\sigma_m^2} (x_i - \mu_m)^2 \right].
\]

Optimum is found by setting the partial derivatives of $E$ to zero:
The Gaussian mixture model (M-step)

In the M-step we optimize the sum (since s is discrete):

\[
E = \langle \log p(x, s|\theta) \rangle_{q(s)} = \sum q(s) \log[p(s|\theta) p(x|s, \theta)] \\
= \sum_{i,m} r_{im} \left[ \log \pi_m - \log \sigma_m - \frac{1}{2\sigma_m^2} (x_i - \mu_m)^2 \right].
\]

Optimum is found by setting the partial derivatives of \( E \) to zero:

\[
\frac{\partial}{\partial \mu_m} E = \sum_i r_{im} \frac{(x_i - \mu_m)}{2\sigma_m^2} = 0 \quad \Rightarrow \quad \mu_m = \frac{\sum_i r_{im} x_i}{\sum_i r_{im}},
\]
The Gaussian mixture model (M-step)

In the M-step we optimize the sum (since $s$ is discrete):

$$E = \langle \log p(x, s|\theta) \rangle_{q(s)} = \sum q(s) \log[p(s|\theta) p(x|s, \theta)]$$

$$= \sum r_{im} \left[ \log \pi_m - \log \sigma_m - \frac{1}{2\sigma_m^2} (x_i - \mu_m)^2 \right].$$

Optimum is found by setting the partial derivatives of $E$ to zero:

$$\frac{\partial}{\partial \mu_m} E = \sum_i r_{im} \frac{x_i - \mu_m}{2\sigma_m^2} = 0 \implies \mu_m = \frac{\sum_i r_{im} x_i}{\sum_i r_{im}},$$

$$\frac{\partial}{\partial \sigma_m} E = \sum_i r_{im} \left[ -\frac{1}{\sigma_m} + \frac{(x_i - \mu_m)^2}{\sigma_m^3} \right] = 0 \implies \sigma_m^2 = \frac{\sum_i r_{im} (x_i - \mu_m)^2}{\sum_i r_{im}},$$

where $\lambda$ is a Lagrange multiplier ensuring that the mixing proportions sum to unity.
The Gaussian mixture model (M-step)

In the M-step we optimize the sum (since s is discrete):

$$E = \langle \log p(x, s|\theta) \rangle_{q(s)} = \sum q(s) \log[p(s|\theta) p(x|s, \theta)]$$

$$= \sum_i r_{im} \left[ \log \pi_m - \log \sigma_m - \frac{1}{2\sigma_m^2} (x_i - \mu_m)^2 \right].$$

Optimum is found by setting the partial derivatives of $E$ to zero:

$$\frac{\partial}{\partial \mu_m} E = \sum_i r_{im} \frac{x_i - \mu_m}{2\sigma_m^2} = 0 \Rightarrow \mu_m = \frac{\sum_i r_{im} x_i}{\sum_i r_{im}},$$

$$\frac{\partial}{\partial \sigma_m} E = \sum_i r_{im} \left[ -\frac{1}{\sigma_m} + \frac{(x_i - \mu_m)^2}{\sigma_m^3} \right] = 0 \Rightarrow \sigma_m = \left( \frac{\sum_i r_{im} (x_i - \mu_m)^2}{\sum_i r_{im}} \right),$$

$$\frac{\partial}{\partial \pi_m} E = \sum_i r_{im} \frac{1}{\pi_m}, \quad \frac{\partial E}{\partial \pi_m} + \lambda = 0 \Rightarrow \pi_m = \frac{1}{n} \sum_i r_{im},$$

where $\lambda$ is a Lagrange multiplier ensuring that the mixing proportions sum to unity.
EM for Factor Analysis

The model for $x$:

$$p(x|\theta) = \int p(y|\theta)p(x|y, \theta)dy = \mathcal{N}(0, \Lambda\Lambda^T + \Psi)$$

Model parameters: $\theta = \{\Lambda, \Psi\}$.

**E step:** For each data point $x_n$, compute the posterior distribution of hidden factors given the observed data: $q_n(y_n) = p(y_n|x_n, \theta_t)$.

**M step:** Find the $\theta_{t+1}$ that maximises $\mathcal{F}(q, \theta)$:

$$\mathcal{F}(q, \theta) = \sum_n \int q_n(y_n) \left[ \log p(y_n|\theta) + \log p(x_n|y_n, \theta) - \log q_n(y_n) \right] dy_n$$

$$= \sum_n \int q_n(y_n) \left[ \log p(y_n|\theta) + \log p(x_n|y_n, \theta) \right] dy_n + c.$$
**The E step for Factor Analysis**

**E step:** For each data point $x_n$, compute the posterior distribution of hidden factors given the observed data: $q_n(y_n) = p(y_n|x_n, \theta) = p(y_n, x_n|\theta)/p(x_n|\theta)$

**Tactic:** write $p(y_n, x_n|\theta)$, consider $x_n$ to be fixed. What is this as a function of $y_n$?

$$p(y_n, x_n) = p(y_n)p(x_n|y_n) = (2\pi)^{-\frac{K}{2}} \exp\{-\frac{1}{2} y_n^T y_n\} \cdot 2\pi \Psi^{\frac{1}{2}} \exp\{-\frac{1}{2} (x_n - \Lambda y_n)^T \Psi^{-1} (x_n - \Lambda y_n)\}$$

$$= c \times \exp\{-\frac{1}{2} [y_n^T y_n + (x_n - \Lambda y_n)^T \Psi^{-1} (x_n - \Lambda y_n)]\}$$

$$= c' \times \exp\{-\frac{1}{2} [y_n^T (I + \Lambda^T \Psi^{-1} \Lambda) y_n - 2y_n^T \Lambda^T \Psi^{-1} x_n]\}$$

$$= c'' \times \exp\{-\frac{1}{2} [y_n^T \Sigma^{-1} y_n - 2y_n^T \Sigma^{-1} \mu_n + \mu_n^T \Sigma^{-1} \mu_n]\}$$

So $\Sigma = (I + \Lambda^T \Psi^{-1} \Lambda)^{-1} = I - \beta \Lambda$ and $\mu_n = \Sigma \Lambda^T \Psi^{-1} x_n = \beta x_n$. Where $\beta = \Sigma \Lambda^T \Psi^{-1}$. Note that $\mu_n$ is a linear function of $x_n$ and $\Sigma$ does not depend on $x_n$. 

The M step for Factor Analysis

**M step:** Find $\theta_{t+1}$ by maximising

$$F = \sum_n \langle \log p(y_n | \theta) + \log p(x_n | y_n, \theta) \rangle_{q_n(y_n)} + c$$
The M step for Factor Analysis

**M step:** Find $\theta_{t+1}$ by maximising $\mathcal{F} = \sum_n \langle \log p(y_n|\theta) + \log p(x_n|y_n, \theta) \rangle_{q_n(y_n)} + c$

$$\log p(y_n|\theta) + \log p(x_n|y_n, \theta)$$
The M step for Factor Analysis

M step: Find $\theta_{t+1}$ by maximizing $\mathcal{F} = \sum_n \langle \log p(y_n|\theta) + \log p(x_n|y_n, \theta) \rangle_{q_n(y_n)} + c$

\[
\log p(y_n|\theta) + \log p(x_n|y_n, \theta) = c - \frac{1}{2} y_n^\top y_n - \frac{1}{2} \log |\Psi| - \frac{1}{2} (x_n - \Lambda y_n)^\top \Psi^{-1} (x_n - \Lambda y_n)
\]
The M step for Factor Analysis

M step: Find $\theta_{t+1}$ by maximising $\mathcal{F} = \sum_n \langle \log p(y_n|\theta) + \log p(x_n|y_n, \theta) \rangle_{q_n(y_n)} + c$

$$\log p(y_n|\theta) + \log p(x_n|y_n, \theta)$$

$$= c - \frac{1}{2} y_n^T y_n - \frac{1}{2} \log |\psi| - \frac{1}{2} (x_n - \Lambda y_n)^T \psi^{-1} (x_n - \Lambda y_n)$$

$$= c' - \frac{1}{2} \log |\psi| - \frac{1}{2} \left[ x_n^T \psi^{-1} x_n - 2x_n^T \psi^{-1} \Lambda y_n + y_n^T \Lambda^T \psi^{-1} \Lambda y_n \right]$$
The M step for Factor Analysis

**M step:** Find $\theta_{t+1}$ by maximising $F = \sum_n \langle \log p(y_n|\theta) + \log p(x_n|y_n, \theta) \rangle_{q_n(y_n)} + c$

\[
\log p(y_n|\theta) + \log p(x_n|y_n, \theta) \\
= c - \frac{1}{2} y_n^T y_n - \frac{1}{2} \log |\psi| - \frac{1}{2} (x_n - \Lambda y_n)^T \Psi^{-1} (x_n - \Lambda y_n) \\
= c' - \frac{1}{2} \log |\psi| - \frac{1}{2} \left[ x_n^T \Psi^{-1} x_n - 2 x_n^T \Psi^{-1} \Lambda y_n + y_n^T \Lambda^T \Psi^{-1} \Lambda y_n \right] \\
= c' - \frac{1}{2} \log |\psi| - \frac{1}{2} \left[ x_n^T \Psi^{-1} x_n - 2 x_n^T \Psi^{-1} \Lambda y_n + \text{Tr} \left[ \Lambda^T \Psi^{-1} \Lambda y_n y_n^T \right] \right]
\]
The M step for Factor Analysis

M step: Find $\theta_{t+1}$ by maximising $\mathcal{F} = \sum_n \langle \log p(y_n|\theta) + \log p(x_n|y_n, \theta) \rangle_{q_n(y_n)} + c$

\[
\log p(y_n|\theta) + \log p(x_n|y_n, \theta)
= c - \frac{1}{2} y_n^T y_n - \frac{1}{2} \log |\Psi| - \frac{1}{2} (x_n - \Lambda y_n)^T \Psi^{-1} (x_n - \Lambda y_n)
= c' - \frac{1}{2} \log |\Psi| - \frac{1}{2} \left[ x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda y_n + y_n^T \Lambda^T \Psi^{-1} \Lambda y_n \right]
= c' - \frac{1}{2} \log |\Psi| - \frac{1}{2} \left[ x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda y_n + \text{Tr} \left[ \Lambda^T \Psi^{-1} \Lambda y_n y_n^T \right] \right]
\]

Taking expectations wrt $q_n(y_n)$:
The M step for Factor Analysis

**M step:** Find $\theta_{t+1}$ by maximising $\mathcal{F} = \sum_n \langle \log p(y_n|\theta) + \log p(x_n|y_n, \theta) \rangle_{q_n(y_n)} + c$

$$\log p(y_n|\theta) + \log p(x_n|y_n, \theta)$$

$$= c - \frac{1}{2} y_n^T y_n - \frac{1}{2} \log |\psi| - \frac{1}{2} (x_n - \Lambda y_n)^T \psi^{-1} (x_n - \Lambda y_n)$$

$$= c' - \frac{1}{2} \log |\psi| - \frac{1}{2} \left[ x_n^T \psi^{-1} x_n - 2x_n^T \psi^{-1} \Lambda y_n + y_n^T \Lambda^T \psi^{-1} \Lambda y_n \right]$$

$$= c' - \frac{1}{2} \log |\psi| - \frac{1}{2} \left[ x_n^T \psi^{-1} x_n - 2x_n^T \psi^{-1} \Lambda y_n + \text{Tr} \left[ \Lambda^T \psi^{-1} \Lambda y_n y_n^T \right] \right]$$

Taking expectations wrt $q_n(y_n)$:

$$= c' - \frac{1}{2} \log |\psi| - \frac{1}{2} \left[ x_n^T \psi^{-1} x_n - 2x_n^T \psi^{-1} \Lambda \mu_n + \text{Tr} \left[ \Lambda^T \psi^{-1} \Lambda (\mu_n \mu_n^T + \Sigma) \right] \right]$$
The M step for Factor Analysis

**M step:** Find $\theta_{t+1}$ by maximising $\mathcal{F} = \sum_n \langle \log p(y_n|\theta) + \log p(x_n|y_n, \theta) \rangle_{q_n(y_n)} + c$

\[
\log p(y_n|\theta) + \log p(x_n|y_n, \theta) \\
= c - \frac{1}{2} y_n^T y_n - \frac{1}{2} \log |\psi| - \frac{1}{2} (x_n - \Lambda y_n)^T \psi^{-1} (x_n - \Lambda y_n) \\
= c' - \frac{1}{2} \log |\psi| - \frac{1}{2} \left[ x_n^T \psi^{-1} x_n - 2 x_n^T \psi^{-1} \Lambda y_n + y_n^T \Lambda^T \psi^{-1} \Lambda y_n \right] \\
= c' - \frac{1}{2} \log |\psi| - \frac{1}{2} \left[ x_n^T \psi^{-1} x_n - 2 x_n^T \psi^{-1} \Lambda y_n + \text{Tr} \left[ \Lambda^T \psi^{-1} \Lambda y_n y_n^T \right] \right]
\]

Taking expectations wrt $q_n(y_n)$:

\[
= c' - \frac{1}{2} \log |\psi| - \frac{1}{2} \left[ x_n^T \psi^{-1} x_n - 2 x_n^T \psi^{-1} \Lambda \mu_n + \text{Tr} \left[ \Lambda^T \psi^{-1} \Lambda (\mu_n \mu_n^T + \Sigma) \right] \right]
\]

Note that we don’t need to know everything about $q(y_n)$, just the moments $\langle y_n \rangle$ and $\langle y_n y_n^T \rangle$. These are the expected sufficient statistics.
The M step for Factor Analysis (cont.)

\[
\mathcal{F} = c' - \frac{N}{2} \log|\Psi| - \frac{1}{2} \sum_n \left[ x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda \mu_n + \text{Tr} \left[ \Lambda^T \Psi^{-1} \Lambda (\mu_n \mu_n^T + \Sigma) \right] \right]
\]
The M step for Factor Analysis (cont.)

\[ F = c' - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum n \left[ x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda \mu_n + \text{Tr} \left[ \Lambda^T \Psi^{-1} \Lambda (\mu_n \mu_n^T + \Sigma) \right] \right] \]

Taking derivatives wrt \( \Lambda \) and \( \Psi^{-1} \), using \( \frac{\partial \text{Tr}[AB]}{\partial B} = A^T \) and \( \frac{\partial \log |A|}{\partial A} = A^{-T} \):
The M step for Factor Analysis (cont.)

\[ \mathcal{F} = c' - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n \left[ x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda \mu_n + \text{Tr} \left[ \Lambda^T \Psi^{-1} \Lambda (\mu_n \mu_n^T + \Sigma) \right] \right] \]

Taking derivatives wrt \( \Lambda \) and \( \Psi^{-1} \), using \( \frac{\partial \text{Tr}(AB)}{\partial B} = A^T \) and \( \frac{\partial \log |A|}{\partial A} = A^{-T} \):

\[ \frac{\partial \mathcal{F}}{\partial \Lambda} = \Psi^{-1} \sum_n x_n \mu_n^T - \Psi^{-1} \Lambda \left( N \Sigma + \sum_n \mu_n \mu_n^T \right) = 0 \]
The M step for Factor Analysis (cont.)

\[ F = c' - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n \left[ x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda \mu_n + \text{Tr} \left[ \Lambda^T \Psi^{-1} \Lambda (\mu_n \mu_n^T + \Sigma) \right] \right] \]

Taking derivatives wrt \( \Lambda \) and \( \Psi^{-1} \), using \( \frac{\partial \text{Tr}[AB]}{\partial B} = A^T \) and \( \frac{\partial \log |A|}{\partial A} = A^{-T} \):

\[ \frac{\partial F}{\partial \Lambda} = \Psi^{-1} \sum_n x_n \mu_n^T - \Psi^{-1} \Lambda \left( N\Sigma + \sum_n \mu_n \mu_n^T \right) = 0 \]

\[ \Rightarrow \hat{\Lambda} = \left( \sum_n x_n \mu_n^T \right) \left( N\Sigma + \sum_n \mu_n \mu_n^T \right)^{-1} \]
The M step for Factor Analysis (cont.)

\[ F = c' - \frac{N}{2} \log |\psi| - \frac{1}{2} \sum_n \left[ x_n^T \psi^{-1} x_n - 2x_n^T \psi^{-1} \Lambda \mu_n + \text{Tr} \left[ \Lambda^T \psi^{-1} (\mu_n \mu_n^T + \Sigma) \right] \right] \]

Taking derivatives wrt \( \Lambda \) and \( \psi^{-1} \), using \( \frac{\partial \text{Tr}[AB]}{\partial B} = A^T \) and \( \frac{\partial \log |A|}{\partial A} = A^{-T} \):

\[ \frac{\partial F}{\partial \Lambda} = \psi^{-1} \sum_n x_n \mu_n^T - \psi^{-1} \Lambda \left( N\Sigma + \sum_n \mu_n \mu_n^T \right) = 0 \]

\[ \Rightarrow \widehat{\Lambda} = \left( \sum_n x_n \mu_n^T \right) \left( N\Sigma + \sum_n \mu_n \mu_n^T \right)^{-1} \]

\[ \frac{\partial F}{\partial \psi^{-1}} = \frac{N}{2} \psi - \frac{1}{2} \sum_n \left[ x_n x_n^T - \Lambda \mu_n x_n^T - x_n \mu_n^T \Lambda^T + \Lambda (\mu_n \mu_n^T + \Sigma) \Lambda^T \right] \]

Note: we should actually only take derivatives w.r.t. \( \Psi_{dd} \) since \( \Psi \) is diagonal.
The M step for Factor Analysis (cont.)

\[ F = c' - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n \left[ x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda \mu_n + \text{Tr} \left[ \Lambda^T \Psi^{-1} \Lambda (\mu_n \mu_n^T + \Sigma) \right] \right] \]

Taking derivatives wrt \( \Lambda \) and \( \Psi^{-1} \), using \( \frac{\partial \text{Tr}[AB]}{\partial B} = A^T \) and \( \frac{\partial \log |A|}{\partial A} = A^{-T} \):

\[ \frac{\partial F}{\partial \Lambda} = \Psi^{-1} \sum_n x_n \mu_n^T - \Psi^{-1} \Lambda \left( N\Sigma + \sum_n \mu_n \mu_n^T \right) = 0 \]

\[ \Rightarrow \hat{\Lambda} = \left( \sum_n x_n \mu_n^T \right) \left( N\Sigma + \sum_n \mu_n \mu_n^T \right)^{-1} \]

\[ \frac{\partial F}{\partial \Psi^{-1}} = \frac{N}{2} \Psi - \frac{1}{2} \sum_n \left[ x_n x_n^T - \Lambda \mu_n x_n^T - x_n \mu_n^T \Lambda^T + \Lambda (\mu_n \mu_n^T + \Sigma) \Lambda^T \right] \]

\[ \Rightarrow \hat{\Psi} = \frac{1}{N} \sum_n \left[ x_n x_n^T - \Lambda \mu_n x_n^T - x_n \mu_n^T \Lambda^T + \Lambda (\mu_n \mu_n^T + \Sigma) \Lambda^T \right] \]

Note: we should actually only take derivatives w.r.t. \( \Psi_{dd} \) since \( \Psi \) is diagonal.
The M step for Factor Analysis (cont.)

\[ F = c' - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n \left[ x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda \mu_n + \text{Tr} \left[ \Lambda^T \Psi^{-1} (\mu_n \mu_n^T + \Sigma) \right] \right] \]

Taking derivatives wrt \( \Lambda \) and \( \Psi^{-1} \), using \( \frac{\partial \text{Tr}[AB]}{\partial B} = A^T \) and \( \frac{\partial \log |A|}{\partial A} = A^{-T} \):

\[ \frac{\partial F}{\partial \Lambda} = \Psi^{-1} \sum_n x_n \mu_n^T - \Psi^{-1} \Lambda \left( N\Sigma + \sum_n \mu_n \mu_n^T \right) = 0 \]

\[ \Rightarrow \hat{\Lambda} = \left( \sum_n x_n \mu_n^T \right) \left( N\Sigma + \sum_n \mu_n \mu_n^T \right)^{-1} \]

\[ \frac{\partial F}{\partial \Psi^{-1}} = \frac{N}{2} \Psi - \frac{1}{2} \sum_n \left[ x_n^T \Psi^{-1} x_n - \Lambda \mu_n x_n^T - x_n \mu_n^T \Lambda + \Lambda (\mu_n \mu_n^T + \Sigma) \Lambda^T \right] \]

\[ \Rightarrow \hat{\Psi} = \frac{1}{N} \sum_n \left[ x_n^T \Psi^{-1} x_n - \Lambda \mu_n x_n^T - x_n \mu_n^T \Lambda + \Lambda (\mu_n \mu_n^T + \Sigma) \Lambda^T \right] \]

\[ \hat{\Psi} = \Lambda \Sigma \Lambda^T + \frac{1}{N} \sum_n (x_n - \Lambda \mu_n)(x_n - \Lambda \mu_n)^T \]  

(squared residuals)

Note: we should actually only take derivatives w.r.t. \( \Psi_{dd} \) since \( \Psi \) is diagonal.
The M step for Factor Analysis (cont.)

\[
F = c' - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n \left[ x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda \mu_n + \text{Tr} \left[ \Lambda^T \Psi^{-1} \Lambda (\mu_n \mu_n^T + \Sigma) \right] \right]
\]

Taking derivatives wrt \( \Lambda \) and \( \Psi^{-1} \), using \( \frac{\partial \text{Tr}[AB]}{\partial B} = A^T \) and \( \frac{\partial \log |A|}{\partial A} = A^{-T} \):

\[
\frac{\partial F}{\partial \Lambda} = \Psi^{-1} \sum_n x_n \mu_n^T - \Psi^{-1} \Lambda \left( N \Sigma + \sum_n \mu_n \mu_n^T \right) = 0
\]

\[
\Rightarrow \hat{\Lambda} = \left( \sum_n x_n \mu_n^T \right) \left( N \Sigma + \sum_n \mu_n \mu_n^T \right)^{-1}
\]

\[
\frac{\partial F}{\partial \Psi^{-1}} = \frac{N}{2} \Psi - \frac{1}{2} \sum_n \left[ x_n x_n^T - \Lambda \mu_n x_n^T - x_n \mu_n^T \Lambda^T + \Lambda (\mu_n \mu_n^T + \Sigma) \Lambda^T \right]
\]

\[
\Rightarrow \hat{\Psi} = \frac{1}{N} \sum_n \left[ x_n x_n^T - \Lambda \mu_n x_n^T - x_n \mu_n^T \Lambda^T + \Lambda (\mu_n \mu_n^T + \Sigma) \Lambda^T \right]
\]

\[
\hat{\Psi} = \Lambda \Sigma \Lambda^T + \frac{1}{N} \sum_n (x_n - \Lambda \mu_n) (x_n - \Lambda \mu_n)^T \quad \text{(squared residuals)}
\]

Note: we should actually only take derivatives w.r.t. \( \Psi_{dd} \) since \( \Psi \) is diagonal. As \( \Sigma \to 0 \) these become the equations for ML linear regression.
Mixtures of Factor Analysers

Simultaneous clustering and dimensionality reduction.

\[
p(x|\theta) = \sum_k \pi_k \mathcal{N}(\mu_k, \Lambda_k \Lambda_k^T + \Psi)
\]

where \(\pi_k\) is the mixing proportion for FA \(k\), \(\mu_k\) is its centre, \(\Lambda_k\) is its “factor loading matrix”, and \(\Psi\) is a common sensor noise model. \(\theta = \\{\{\pi_k, \mu_k, \Lambda_k\}_{k=1}^{K}, \Psi\}\)

We can think of this model as having \textit{two} sets of hidden latent variables:

- A discrete indicator variable \(s_n \in \{1, \ldots, K\}\)
- For each factor analyzer, a continuous factor vector \(y_{n,k} \in \mathcal{R}^{D_k}\)

\[
p(x|\theta) = \sum_{s_n=1}^{K} p(s_n|\theta) \int p(y|s_n, \theta) p(x_n|y, s_n, \theta) \, dy
\]

As before, an EM algorithm can be derived for this model:

**E step**: We need moments of \(p(y_n, s_n|x_n, \theta)\), specifically: \(\langle \delta_{s_n=m} \rangle, \langle \delta_{s_n=m} y_n \rangle\) and \(\langle \delta_{s_n=m} y_n y_n^T \rangle\).

**M step**: Similar to M-step for FA with responsibility-weighted moments.

See \url{http://www.learning.eng.cam.ac.uk/zoubin/papers/tr-96-1.pdf}
EM for exponential families

EM is often applied to models whose joint over \( z = (y, x) \) has exponential-family form:

\[
p(z|\theta) = f(z) \exp\{\theta^T T(z)\} / Z(\theta)
\]

(with \( Z(\theta) = \int f(z) \exp\{\theta^T T(z)\} dz \) but whose marginal \( p(x) \not\in \text{ExpFam} \).
EM for exponential families

EM is often applied to models whose joint over \( z = (y, x) \) has exponential-family form:

\[
p(z|\theta) = f(z) \exp\{\theta^T T(z)\} / Z(\theta)
\]

(with \( Z(\theta) = \int f(z) \exp\{\theta^T T(z)\} dz \)) but whose marginal \( p(x) \not\in \text{ExpFam} \).

The free energy dependence on \( \theta \) is given by:

\[
\mathcal{F}(q, \theta) = \int q(y) \log p(y, x|\theta) dy - H[q]
\]
EM for exponential families

EM is often applied to models whose **joint** over \( z = (y, x) \) has exponential-family form:

\[
p(z|\theta) = f(z) \exp{\theta^T T(z)} / Z(\theta)
\]

(with \( Z(\theta) = \int f(z) \exp{\theta^T T(z)} \, dz \)) but whose marginal \( p(x) \not\in \text{ExpFam} \).

The free energy dependence on \( \theta \) is given by:

\[
\mathcal{F}(q, \theta) = \int q(y) \log p(y, x|\theta) \, dy - H[q]
\]

\[
= \int q(y) [\theta^T T(z) - \log Z(\theta)] \, dy + \text{const wrt } \theta
\]
EM for exponential families

EM is often applied to models whose joint over $z = (y, x)$ has exponential-family form:

$$p(z|\theta) = f(z) \exp\{\theta^T T(z)\} / Z(\theta)$$

(with $Z(\theta) = \int f(z) \exp\{\theta^T T(z)\} dz$) but whose marginal $p(x) \not\in \text{ExpFam}$. 

The free energy dependence on $\theta$ is given by:

$$\mathcal{F}(q, \theta) = \int q(y) \log p(y, x|\theta) dy - H[q]$$

$$= \int q(y) [\theta^T T(z) - \log Z(\theta)] dy + \text{const wrt } \theta$$

$$= \theta^T \langle T(z) \rangle_{q(y)} - \log Z(\theta) + \text{const wrt } \theta$$

So, in the E step all we need to compute are the expected sufficient statistics under $q$. 

EM for exponential families

EM is often applied to models whose joint over \( z = (y, x) \) has exponential-family form:

\[
p(z|\theta) = f(z) \exp\{\theta^T T(z)\} / Z(\theta)
\]

(with \( Z(\theta) = \int f(z) \exp\{\theta^T T(z)\} dz \)) but whose marginal \( p(x) \not\in \text{ExpFam} \).

The free energy dependence on \( \theta \) is given by:

\[
\mathcal{F}(q, \theta) = \int q(y) \log p(y, x|\theta) dy - H[q]
\]

\[
= \int q(y) [\theta^T T(z) - \log Z(\theta)] dy + \text{const wrt } \theta
\]

\[
= \theta^T \langle T(z) \rangle_{q(y)} - \log Z(\theta) + \text{const wrt } \theta
\]

So, in the **E step** all we need to compute are the expected sufficient statistics under \( q \).

We also have:

\[
\frac{\partial}{\partial \theta} \log Z(\theta) = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} Z(\theta) = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} \int f(z) \exp\{\theta^T T(z)\}
\]
EM for exponential families

EM is often applied to models whose **joint** over \( z = (y, x) \) has exponential-family form:

\[
p(z|\theta) = f(z) \exp\{\theta^T T(z)\} / Z(\theta)
\]

(with \( Z(\theta) = \int f(z) \exp\{\theta^T T(z)\} dz \)) but whose marginal \( p(x) \not\in \text{ExpFam} \).

The free energy dependence on \( \theta \) is given by:

\[
\mathcal{F}(q, \theta) = \int q(y) \log p(y, x|\theta) dy - H[q]
\]

\[
= \int q(y) [\theta^T T(z) - \log Z(\theta)] dy + \text{const wrt } \theta
\]

\[
= \theta^T \langle T(z) \rangle_{q(y)} - \log Z(\theta) + \text{const wrt } \theta
\]

So, in the **E step** all we need to compute are the **expected sufficient statistics** under \( q \). We also have:

\[
\frac{\partial}{\partial \theta} \log Z(\theta) = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} Z(\theta) = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} \int f(z) \exp\{\theta^T T(z)\}
\]

\[
= \int \frac{1}{Z(\theta)} f(z) \exp\{\theta^T T(z)\} \cdot T(z)
\]
EM for exponential families

EM is often applied to models whose joint over $z = (y, x)$ has exponential-family form:

$$p(z|\theta) = f(z) \exp\{\theta^T T(z)\} / Z(\theta)$$

(with $Z(\theta) = \int f(z) \exp\{\theta^T T(z)\} \, dz$) but whose marginal $p(x) \not\in ExpFam$.

The free energy dependence on $\theta$ is given by:

$$\mathcal{F}(q, \theta) = \int q(y) \log p(y, x|\theta) \, dy - H[q]$$

$$= \int q(y) [\theta^T T(z) - \log Z(\theta)] \, dy + \text{const wrt } \theta$$

$$= \theta^T \langle T(z) \rangle_{q(y)} - \log Z(\theta) + \text{const wrt } \theta$$

So, in the E step all we need to compute are the expected sufficient statistics under $q$.

We also have:

$$\frac{\partial}{\partial \theta} \log Z(\theta) = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} Z(\theta) = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} \int f(z) \exp\{\theta^T T(z)\}$$

$$= \int \frac{1}{Z(\theta)} f(z) \exp\{\theta^T T(z)\} \cdot T(z) = \langle T(z) | \theta \rangle$$
EM for exponential families

EM is often applied to models whose \textbf{joint} over \( z = (y, x) \) has exponential-family form:

\[
p(z|\theta) = f(z) \exp\{\theta^T T(z)\} / Z(\theta)
\]

(with \( Z(\theta) = \int f(z) \exp\{\theta^T T(z)\} \, dz \)) but whose marginal \( p(x) \not\in ExpFam \).

The free energy dependence on \( \theta \) is given by:

\[
\mathcal{F}(q, \theta) = \int q(y) \log p(y, x|\theta) \, dy - H[q]
\]

\[
= \int q(y) [\theta^T T(z) - \log Z(\theta)] \, dy + \text{const wrt } \theta
\]

\[
= \theta^T \langle T(z) \rangle_{q(y)} - \log Z(\theta) + \text{const wrt } \theta
\]

So, in the \textbf{E step} all we need to compute are the \textbf{expected sufficient statistics} under \( q \).

We also have:

\[
\frac{\partial}{\partial \theta} \log Z(\theta) = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} Z(\theta) = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} \int f(z) \exp\{\theta^T T(z)\}
\]

\[
= \int \frac{1}{Z(\theta)} f(z) \exp\{\theta^T T(z)\} \cdot T(z) = \langle T(z) | \theta \rangle
\]

Thus, the \textbf{M step} solves:

\[
\frac{\partial \mathcal{F}}{\partial \theta} = \langle T(z) \rangle_{q(y)} - \langle T(z) | \theta \rangle = 0
\]


Proof of the Matrix Inversion Lemma

\[(A + XBX^T)^{-1} = A^{-1} - A^{-1}X(B^{-1} + X^TA^{-1}X)^{-1}X^TA^{-1}\]

Need to prove:

\[\left( A^{-1} - A^{-1}X(B^{-1} + X^TA^{-1}X)^{-1}X^TA^{-1} \right) (A + XBX^T) = I \]

Expand:

\[I + A^{-1}XBX^T - A^{-1}X(B^{-1} + X^TA^{-1}X)^{-1}X^T - A^{-1}X(B^{-1} + X^TA^{-1}X)^{-1}X^TA^{-1}XBX^T\]

Regroup:

\[= I + A^{-1}X \left( BX^T - (B^{-1} + X^TA^{-1}X)^{-1}X^T - (B^{-1} + X^TA^{-1}X)^{-1}X^TA^{-1}XBX^T \right)\]

\[= I + A^{-1}X \left( BX^T - (B^{-1} + X^TA^{-1}X)^{-1}B^{-1}BX^T - (B^{-1} + X^TA^{-1}X)^{-1}X^TA^{-1}XBX^T \right)\]

\[= I + A^{-1}X \left( BX^T - (B^{-1} + X^TA^{-1}X)^{-1}(B^{-1} + X^TA^{-1}X)BX^T \right)\]

\[= I + A^{-1}X(BX^T - BX^T) = I\]
**KL** \([q(x)\|p(x)] \geq 0, \text{ with equality iff } \forall x : p(x) = q(x)\)

First consider discrete distributions; the Kullback-Liebler divergence is:

\[
\text{KL}[q\|p] = \sum_i q_i \log \frac{q_i}{p_i}.
\]

To minimize wrt distribution \(q\) we need a Lagrange multiplier to enforce normalisation:

\[
E \overset{\text{def}}{=} \text{KL}[q\|p] + \lambda (1 - \sum_i q_i) = \sum_i q_i \log \frac{q_i}{p_i} + \lambda (1 - \sum_i q_i)
\]

Find conditions for stationarity

\[
\frac{\partial E}{\partial q_i} = \log q_i - \log p_i + 1 - \lambda = 0 \Rightarrow q_i = p_i \exp(\lambda - 1)
\]

\[
\frac{\partial E}{\partial \lambda} = 1 - \sum_i q_i = 0 \Rightarrow \sum_i q_i = 1
\]

Check sign of curvature (Hessian):

\[
\frac{\partial^2 E}{\partial q_i \partial q_i} = \frac{1}{q_i} > 0, \quad \frac{\partial^2 E}{\partial q_i \partial q_j} = 0,
\]

so unique stationary point \(q_i = p_i\) is indeed a minimum. Easily verified that at that minimum,

\[
\text{KL}[q\|p] = \text{KL}[p\|p] = 0.
\]

A similar proof holds for continuous densities, using functional derivatives.
Fixed Points of EM are Stationary Points in $\ell$

Let a fixed point of EM occur with parameter $\theta^*$. Then:

$$
\frac{\partial}{\partial \theta} \left< \log P(Y, X \mid \theta) \right>_P(Y \mid X, \theta^*) \bigg|_{\theta^*} = 0
$$

The second term is 0 at $\theta^*$ if the derivative exists (minimum of $\mathcal{KL}[\cdot \parallel \cdot]$), and thus:

$$
\frac{\partial}{\partial \theta} \ell(\theta) \bigg|_{\theta^*} = \frac{\partial}{\partial \theta} \left< \log P(Y, X \mid \theta) \right>_P(Y \mid X, \theta^*) \bigg|_{\theta^*} = 0
$$

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

Fixed Points of EM are Stationary Points in $\ell$

Let a fixed point of EM occur with parameter $\theta^*$. Then:

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

Now, $\ell(\theta) = \log P(X \mid \theta) = \langle \log P(X \mid \theta) \rangle_{P(Y \mid X, \theta^*)}$
Fixed Points of EM are Stationary Points in $\ell$

Let a fixed point of EM occur with parameter $\theta^*$. Then:

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

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

$$= \left\langle \log \frac{P(Y, X \mid \theta)}{P(Y \mid X, \theta)} \right\rangle_{P(Y \mid X, \theta^*)}$$
Fixed Points of EM are Stationary Points in $\ell$

Let a fixed point of EM occur with parameter $\theta^*$. Then:

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

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

$$= \left\langle \log \frac{P(Y, X \mid \theta)}{P(Y \mid X, \theta)} \right\rangle_{P(Y \mid X, \theta^*)}$$

$$= \langle \log P(Y, X \mid \theta) \rangle_{P(Y \mid X, \theta^*)} - \langle \log P(Y \mid X, \theta) \rangle_{P(Y \mid X, \theta^*)}$$
Fixed Points of EM are Stationary Points in $\ell$

Let a fixed point of EM occur with parameter $\theta^*$. Then:

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

Now,

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

$$
= \left\langle \log \frac{P(Y, X \mid \theta)}{P(Y \mid X, \theta)} \right\rangle_{P(Y \mid X, \theta^*)}
$$

$$
= \langle \log P(Y, X \mid \theta) \rangle_{P(Y \mid X, \theta^*)} - \langle \log P(Y \mid X, \theta) \rangle_{P(Y \mid X, \theta^*)}
$$

so,

$$
\frac{d}{d\theta} \ell(\theta) = \frac{d}{d\theta} \langle \log P(Y, X \mid \theta) \rangle_{P(Y \mid X, \theta^*)} - \frac{d}{d\theta} \langle \log P(Y \mid X, \theta) \rangle_{P(Y \mid X, \theta^*)}
$$
Fixed Points of EM are Stationary Points in $\ell$

Let a fixed point of EM occur with parameter $\theta^*$. Then:

$$\frac{\partial}{\partial \theta} \left. \frac{\langle \log P(Y, X \mid \theta) \rangle}{P(Y \mid X, \theta^*)} \right|_{\theta^*} = 0$$

Now,

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

$$= \left\langle \log \frac{P(Y, X \mid \theta)}{P(Y \mid X, \theta)} \right\rangle_{P(Y \mid X, \theta^*)}$$

$$= \langle \log P(Y, X \mid \theta) \rangle_{P(Y \mid X, \theta^*)} - \langle \log P(Y \mid X, \theta) \rangle_{P(Y \mid X, \theta^*)}$$

So,

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

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

$$\left. \frac{d}{d\theta} \ell(\theta) \right|_{\theta^*} = \left. \frac{d}{d\theta} \langle \log P(Y, X \mid \theta) \rangle_{P(Y \mid X, \theta^*)} \right|_{\theta^*} = 0$$
Fixed Points of EM are Stationary Points in $\ell$

Let a fixed point of EM occur with parameter $\theta^*$. Then:

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

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

$$= \left\langle \log \frac{P(Y, X \mid \theta)}{P(Y \mid X, \theta)} \right\rangle_{P(Y \mid X, \theta^*)}$$

$$= \langle \log P(Y, X \mid \theta) \rangle_{P(Y \mid X, \theta^*)} - \langle \log P(Y \mid X, \theta) \rangle_{P(Y \mid X, \theta^*)}$$

So,

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

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

$$\left. \frac{d}{d\theta} \ell(\theta) \right|_{\theta^*} = \left. \frac{d}{d\theta} \langle \log P(Y, X \mid \theta) \rangle_{P(Y \mid X, \theta^*)} \right|_{\theta^*} = 0$$

So, EM converges to a stationary point of $\ell(\theta)$. 
Maxima in $\mathcal{F}$ correspond to maxima in $\ell$

Let $\theta^*$ 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 $\ell$

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

Differentiating the previous expression wrt $\theta$ again we find

$$\frac{d^2}{d\theta^2} \ell(\theta) = \frac{d^2}{d\theta^2} \langle \log P(\mathbf{y}, \mathbf{x}|\mathbf{0}) \rangle_{P(\mathbf{y}|\mathbf{x}, \theta^*)} - \frac{d^2}{d\theta^2} \langle \log P(\mathbf{y}|\mathbf{x}, \theta) \rangle_{P(\mathbf{y}|\mathbf{x}, \theta^*)}$$

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 $\theta^*$ is a maximum of $\ell$.

[. . . as long as the derivatives exist. They sometimes don’t (zero-noise ICA).]
Maxima in $F$ correspond to maxima in $\ell$

Let $\theta^*$ now be the parameter value at a local maximum of $F$ (and thus at a fixed point).

Differentiating the previous expression wrt $\theta$ again we find

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

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 $\ell$

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

Differentiating the previous expression wrt $\theta$ again we find

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

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

$\theta^*$ is a maximum of $\ell$. 

**Maxima in \( \mathcal{F} \) correspond to maxima in \( \ell \)**

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

Differentiating the previous expression wrt \( \theta \) again we find

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

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

\[ \theta^* \text{ is a maximum of } \ell. \]

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