Log-likelihoods

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

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

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

- Latent variable models: \( p(x|\theta_x, \theta_y) = \int dy f(y) e^{\phi(\theta_x, y)^T T_x(x)/Z_x(\phi(\theta_x, y))} f_m(y) e^{\phi(\theta_y, y)^T T_y(y)/Z_y(\theta_y)} \)

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

- Usually no closed form optimum.
- Often multiple local maxima.
- Direct numerical optimisation may be possible but infrequently easy.

**Example: mixture of Gaussians**

Data: \( X = \{ x_1, \ldots, x_n \} \)

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

Component distributions: \( x_i | (s_i = m) \sim P_m[\theta_m] = \mathcal{N}(\mu_m, \Sigma_m) \)

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

Log-likelihood:

\[
\ell(\{ \mu_m \}, \{ \Sigma_m \}, \pi) = \sum_{i=1}^n \log \sum_{m=1}^k \frac{\pi_m}{\sqrt{2\pi \Sigma_m}} e^{-\frac{1}{2} (x_i - \mu_m)^T \Sigma_m^{-1} (x_i - \mu_m)}
\]
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?

The lower bound for EM – “free energy”

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

Log-likelihood:

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

By Jensen, any distribution, $q(\mathcal{Y})$, over the latent variables generates a lower bound:

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

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:

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

Jensen’s inequality

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

![Jensen's inequality diagram](image)

In general:

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

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

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 $\alpha$.

The E and M steps of EM

The lower bound on the log likelihood is given by:

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

EM alternates between:

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

  $$\alpha^{(k)}(\mathcal{Y}) := \arg\max\limits_{q(\mathcal{Y})} \mathcal{F}(q(\mathcal{Y}), \theta^{(k-1)}).$$

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

  $$\theta^{(k)} := \arg\max\limits_{\theta} \mathcal{F}(\alpha^{(k)}(\mathcal{Y}), \theta) = \arg\max\limits_{\theta} \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{q^{(k)}(\mathcal{Y})}$$

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

The free energy can be re-written

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

The second term is the Kullback-Leibler divergence.

This means that, for fixed \( \theta \), \( F \) is bounded above by \( \ell \), 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.

Coordinate Ascent in \( F \) (Demo)

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

\[ s \sim \text{Bernoulli}[\pi] \]
\[ x|s = 0 \sim N[-1, 1] \quad x|s = 1 \sim N[1, 1]. \]

Single data point \( x_1 = 0.3 \).

\( q(s) \) is a distribution on a single binary latent, and so is represented by \( n \in [0, 1] \).

EM Never Decreases the Likelihood

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

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

\( \mathcal{F} \leq \ell \) 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.

\[ \ell(\theta) = \log P(X|\theta) = \log \int dY \, P(X|Y, \theta) P(Y|\theta) \]

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

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

- E step:

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

- M step:

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

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

EM for MoGs

- Evaluate responsibilities

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

- Update parameters

\[ \mu_m \leftarrow \frac{\sum_i r_m x_i}{\sum_i r_m} \]
\[ \Sigma_m \leftarrow \frac{\sum_i r_m (x_i - \mu_m)(x_i - \mu_m)^T}{\sum_i r_m} \]
\[ \pi_m \leftarrow \frac{\sum_i r_m}{N} \]

Partial M steps and Partial E steps

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

In fact, immediately after an E step

\[ \frac{\partial}{\partial \theta_{(k-1)}} \langle \log P(X, Y|\theta) \rangle_{q^{(k)}(Y)} - \text{KL}[q(Y)||P(Y|X)]_{(k-1)} \rangle = \frac{\partial}{\partial \theta_{(k-1)}} \log P(X|\theta) \]

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 \( \text{increase } \mathcal{F} \text{ wrt some of the } q s \).

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...

The Gaussian mixture model (E-step)

In a univariate Gaussian mixture model, the density of a data point \( x \) is:

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

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:

\[ q(s_i = m) = \frac{\pi_m \frac{1}{\sigma_m} \exp \left\{ - \frac{1}{2 \sigma_m^2} \left( x_i - \mu_m \right)^2 \right\} \text{ (responsibilities)}}}{\sum_m \pi_m \frac{1}{\sigma_m} \exp \left\{ - \frac{1}{2 \sigma_m^2} \left( x_i - \mu_m \right)^2 \right\} \text{ (normalization)}} \]

with the normalization such that \( \sum_m r_m = 1 \).
The E step for Factor Analysis

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

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

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

\[ \frac{\partial}{\partial \mu_m} E = \sum_i r_m \frac{x_i - \mu_m}{2 \sigma_m^2} = 0 \Rightarrow \mu_m = \frac{\sum_i r_m x_i}{\sum_i r_m}, \]
\[ \frac{\partial}{\partial \sigma_m} E = \sum_i r_m \left[ - \frac{1}{\sigma_m} + \frac{(x_i - \mu_m)^2}{\sigma_m^2} \right] = 0 \Rightarrow \sigma_m^2 = \frac{\sum_i r_m (x_i - \mu_m)^2}{\sum_i r_m}, \]
\[ \frac{\partial}{\partial \pi_m} E = \sum_i r_m \left( \frac{1}{\pi_m} \cdot \frac{\partial E}{\partial \pi_m} + \lambda = 0 \Rightarrow \pi_m = \frac{1}{n} \sum_i r_m, \right. \]

where \( \lambda \) is a Lagrange multiplier ensuring that the mixing proportions sum to unity.

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: \( q_n(y_n) = p(y_n \mid x_n, \theta) \).

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

\[ p(y_n, x_n) = p(y_n) p(x_n \mid y_n) \]
\[ = (2\pi)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} y_n^T \Sigma^{-1} y_n \right\} 2\pi \Psi^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (x_n - \Lambda y_n)^T \Psi^{-1} (x_n - \Lambda y_n) \right\} \]
\[ = c \times \exp \left\{ -\frac{1}{2} y_n^T (I + \Lambda^T \Psi^{-1} \Lambda) y_n - 2 y_n^T \Lambda^T \Psi^{-1} x_n \right\} \]
\[ = c^* \times \exp \left\{ -\frac{1}{2} y_n^T (I + \Lambda^T \Psi^{-1} \Lambda)^{-1} y_n - 2 y_n^T \Lambda^T \Psi^{-1} x_n + \mu_n^T \Lambda^T \Psi^{-1} \mu_n \right\}. \]

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 \( \theta_{n+1} \) by maximising \( \mathcal{F}(q, \theta) \): 

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

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

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

\[ \mathcal{E}(\theta, \theta') = \sum_n \left( \log p(y_n \mid \theta) + \log p(x_n \mid y_n, \theta) - \log q_n(y_n) \right) dy_n \]
\[ = \sum_n \left( \log p(y_n \mid \theta) + \log p(x_n \mid y_n, \theta) \right) dy_n + c. \]

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.
EM for exponential families

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

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

(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:

\[
F(q, \theta) = \int q(y) \log p(y, x; \theta) \, dy - H[q]
\]

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

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

So, in the E step all we need to compute are the expected sufficient statistics \( \langle T(z) \rangle_{q(y)} \). We also have:

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

\[
= \int \frac{1}{Z(\theta)} f(z) \exp(\theta^T T(z)) \cdot T(z) = \langle T(z) \rangle_{q(y)} - \langle T(z) \rangle_{q(y)}
\]

Thus, the M step solves:

\[
\frac{\partial F}{\partial \theta} = \langle T(z) \rangle_{q(y)} - \langle T(z) \rangle_{q(y)} = 0
\]
Proof of the Matrix Inversion Lemma

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

Need to prove:

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

Expand:

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

Regroup:

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

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

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

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

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

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

\[\frac{\partial}{\partial \theta} \left( \log P(Y \mid X \mid \theta) \right)_{P(Y \mid X, \theta^*)} \bigg|_{\theta^*} = 0\]

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

\[= \log P(Y, X \mid \theta)_{P(Y \mid X, \theta^*)} - \log P(Y \mid X, \theta)_{P(Y \mid X, \theta^*)}\]

so,

\[\frac{d}{d\theta} \ell(\theta) = \frac{d}{d\theta} \log P(Y, X \mid \theta)_{P(Y \mid X, \theta^*)} - \frac{d}{d\theta} \log P(Y \mid X, \theta)_{P(Y \mid X, \theta^*)}\]

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

\[\frac{d}{d\theta} \ell(\theta) \bigg|_{\theta^*} = \frac{d}{d\theta} \log P(Y, X \mid \theta)_{P(Y \mid X, \theta^*)} \bigg|_{\theta^*} = 0\]

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

KL $[q(x)||p(x)] \geq 0$, with equality iff $\forall x : p(x) = q(x)$

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

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

To minimize wrt distribution $q$ we need a Lagrange multiplier to enforce normalisation:

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

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):

\[\frac{d^2 E}{dq_i dq_j} = \frac{1}{q_i} > 0, \quad \frac{d^2 E}{dq_i dq_j} = 0,\]

so unique stationary point $q_i = p_i$ is indeed a minimum. Easily verified that at that minimum, $\text{KL}[q \parallel p] = \text{KL}[p \parallel p] = 0$.

A similar proof holds for continuous densities, using functional derivatives.

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

\[\frac{d^2}{d\theta^2} \ell(\theta) = \frac{d^2}{d\theta^2} \log P(Y, X \mid \theta)_{P(Y \mid X, \theta^*)} - \frac{d^2}{d\theta^2} \log P(Y \mid X, \theta)_{P(Y \mid X, \theta^*)}\]

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)].