

# **Expectation Maximisation (EM)**

Erfan Loweimi

Speech Group, Machine Intelligence Lab, University of Cambridge

# Outline ... EM

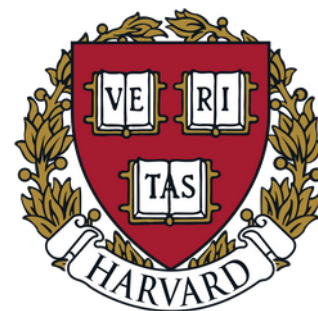
- Importance
- Goal
- Idea
- Derivation
- Visualisation

# Importance ...

## Maximum Likelihood from Incomplete Data via the *EM* Algorithm

By A. P. DEMPSTER, N. M. LAIRD and D. B. RUBIN

*Harvard University and Educational Testing Service*



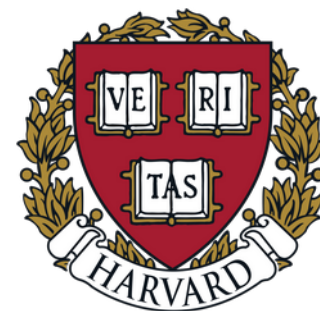
Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). *Maximum likelihood from incomplete data via the EM algorithm*. **Journal of the Royal Statistical Society**: Series B, 39, 1-38.

# Importance ...

## Maximum Likelihood from Incomplete Data via the *EM* Algorithm

By A. P. DEMPSTER, N. M. LAIRD and D. B. RUBIN

*Harvard University and Educational Testing Service*



<https://www.jstor.org> › stable

### Maximum Likelihood from Incomplete Data via the ... - JSTOR

by AP Dempster · 1977 · Cited by 69463 — A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality.

21, Feb, 2023

# Importance ...

## Keywords ...

Maximum Likelihood from Incomplete Data via the *EM* Algorithm

By A. P. DEMPSTER, N. M. LAIRD and D. B. RUBIN

*Harvard University and Educational Testing Service*



This seminal paper was **NOT** the first to discover EM but rather ...

- Generalised it beyond special circumstances/applications
- Sketched a convergence analysis

# Setup

$X$  : *observable* rv<sup>\*</sup>

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}, \quad \mathbf{x}_i \in \mathbb{R}^{D_1}$$

$Z$  : *latent* rv

$$\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N\}, \quad \mathbf{z}_i \in \mathbb{R}^{D_2}$$

$\mathbf{X}$  : *incomplete* data

$\{\mathbf{X}, \mathbf{Z}\}$  : *complete* data

$\theta$  : model parameters

\* rv: random variable

# Setup

independent

identically distributed

$$p(\mathbf{X}|\theta) = p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N|\theta) \stackrel{i.i.d}{=} \prod_{i=1}^N p(\mathbf{x}_i|\theta)$$

$$\log p(\mathbf{X}|\theta) = \sum_{i=1}^N \log p(\mathbf{x}_i|\theta)$$

$$p(\mathbf{X}, \mathbf{Z}|\theta) = p(\mathbf{x}_1, \mathbf{z}_1, \mathbf{x}_2, \mathbf{z}_2, \dots, \mathbf{x}_N, \mathbf{z}_N|\theta) \stackrel{i.i.d}{=} \prod_{i=1}^N p(\mathbf{x}_i, \mathbf{z}_i|\theta)$$

$$\log p(\mathbf{X}|\theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

# Setup

$P(\mathbf{X}|\theta)$  : *incomplete data* likelihood

$P(\mathbf{X}, \mathbf{Z}|\theta)$  : *complete data* likelihood

$P(\mathbf{Z}|\mathbf{X}, \theta)$  : *posterior* probability

Marginalisation

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

Chain rule (probability)

$$p(\mathbf{X}|\theta) = \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X}, \theta)}$$



# Goal ...

$$\theta_{ML}^* = \operatorname{argmax}_{\theta \in \Theta} p(\mathbf{X}|\theta)$$

Find  $\theta_{ML}$  such that the likelihood of data  $\mathbf{X}$ , being generated by Model  $\theta$ , is maximised.

# Goal ...

$$\theta_{ML}^* = \operatorname{argmax}_{\theta \in \Theta} p(\mathbf{X}|\theta) = \operatorname{argmax}_{\theta \in \Theta} \log p(\mathbf{X}|\theta)$$

- We prefer maximising **log**-likelihood ...
  - **Note:** log is *strictly increasing*  $\rightarrow \operatorname{argmax} f(x) = \operatorname{argmax} \log(f(x))$
  - **Advantages:**
    - ✓ Mathematical convenience  $\rightarrow \log(\exp[.]) = [.]$
    - ✓ Numerical stability

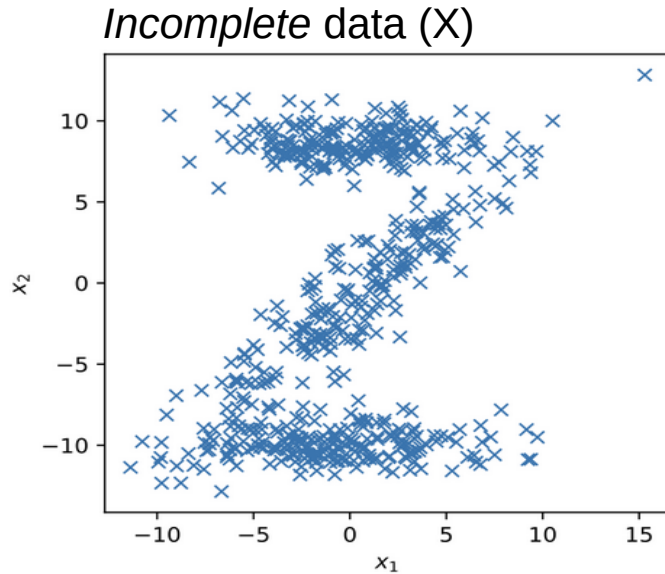
# Goal ...

$$\begin{aligned}\theta_{ML}^*(\mathbf{X}|\theta) &= \operatorname{argmax}_{\theta \in \Theta} \log p(\mathbf{X}|\theta) \\ &= \operatorname{argmax}_{\theta \in \Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)\end{aligned}$$

**EM assumes** model includes *latent variables* ( $\mathbf{Z}$ ).

# Latent Variable (Z)

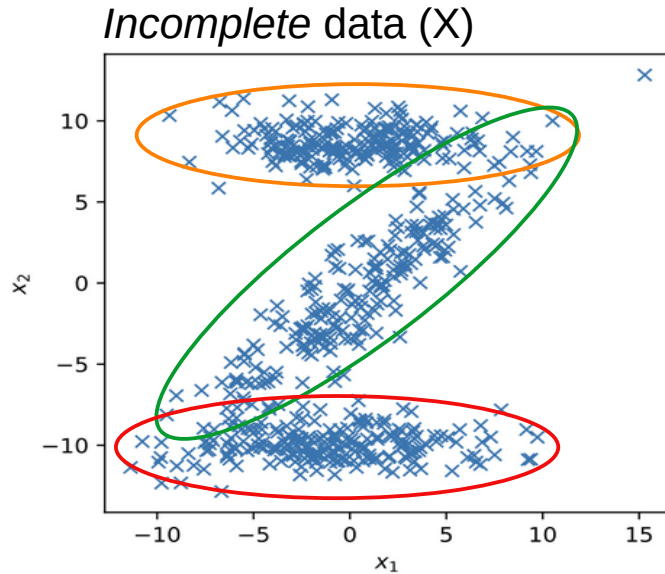
$$\mathbf{x} \sim p(\mathbf{x})$$



**Interpretation:** *latent variable* is a part of a model ... *explains X*.

# Latent Variable (Z)

$$\mathbf{x} \sim p(\mathbf{x})$$

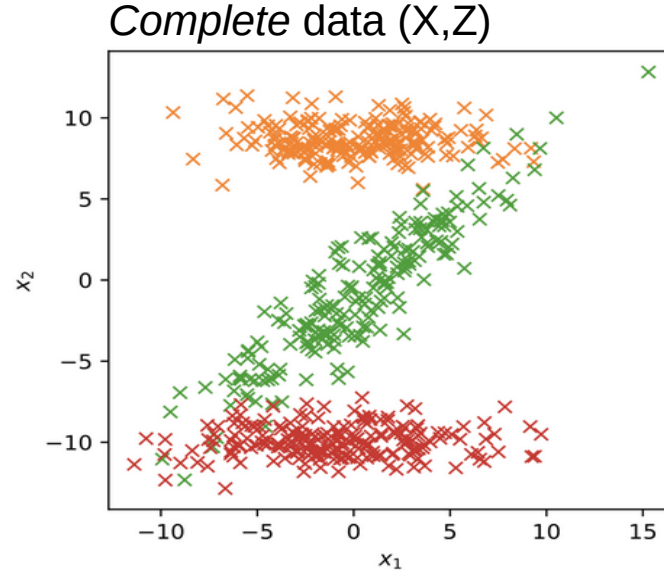
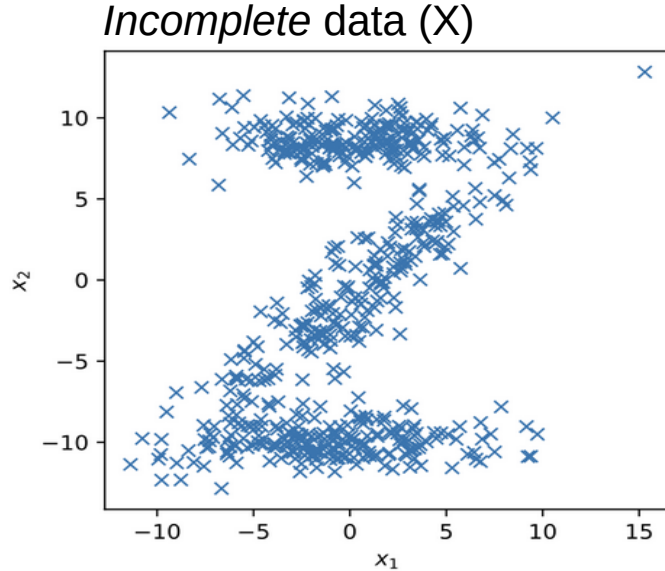


Consider clustering ...

**Interpretation:** *latent variable* is a part of a model ... *explains X*.

# Latent Variable (Z)

$$\mathbf{x} \sim p(\mathbf{x})$$



$$(\mathbf{x}, \mathbf{z}) \sim p(\mathbf{x}, \mathbf{z})$$

**Interpretation:** *latent variable* is a part of a model ... *explains* X.

# Direct Solution

$$\theta_{ML}^* = \operatorname{argmax}_{\theta \in \Theta} \underbrace{\log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} | \theta)}_{\log p(\mathbf{X} | \theta)}$$

$$\frac{\partial \log p(\mathbf{X} | \theta)}{\partial \theta} = 0$$

Step 1

Step 2

$$\left. \frac{\partial^2 \log p(\mathbf{X} | \theta)}{\partial \theta^2} \right|_{\theta = \theta_0} < 0$$

Step 3

$\theta_0$ : derivative roots

# Direct Solution

$$\theta_{ML}^* = \operatorname{argmax}_{\theta \in \Theta} \underbrace{\log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} | \theta)}_{\log p(\mathbf{X} | \theta)}$$

Intractable (no closed-form solution!)

$$\frac{\partial \log p(\mathbf{X} | \theta)}{\partial \theta} = 0$$

Step 1

Step 2

$$\left. \frac{\partial^2 \log p(\mathbf{X} | \theta)}{\partial \theta^2} \right|_{\theta = \theta_0} < 0$$

Step 3

$\theta_0$ : derivative roots



# Direction solution does not work!

$$\underbrace{\log \sum_z p(x, z|\theta)}_{\log p(x|\theta)} = \log(\dots + w_{z_i} e^{\frac{(x-\mu_i)^2}{2\sigma_i^2}} + \dots + w_{z_j} e^{\frac{(x-\mu_j)^2}{2\sigma_j^2}} + \dots)$$

Intractable

Consider a simple case ... Gaussian

$$\frac{\partial \log p(x|\theta)}{\partial \theta} = 0$$

How about *numerical methods*? Slow and do not scale!

# Direction solution does not work!

$$\log \sum_z p(x, z|\theta) = \log(\dots + w_{z_i} e^{\frac{(x-\mu_i)^2}{2\sigma_i^2}} + \dots + w_{z_j} e^{\frac{(x-\mu_j)^2}{2\sigma_j^2}} + \dots)$$

... = 0 → Intractable

---

# Direction solution does not work!

$$\log \sum_z p(x, z|\theta) = \log(\dots + w_{z_i} e^{\frac{(x-\mu_i)^2}{2\sigma_i^2}} + \dots + w_{z_j} e^{\frac{(x-\mu_j)^2}{2\sigma_j^2}} + \dots)$$

... = 0 → Intractable

If we could swap log &  $\Sigma$  ...

$$\begin{aligned} \sum_z \log p(x, z|\theta) &= (\dots + w_i \log e^{\frac{(x-\mu_i)^2}{2\sigma_i^2}} + \dots + w_j \log e^{\frac{(x-\mu_j)^2}{2\sigma_j^2}} + \dots) \\ &= (\dots + c_i(x - \mu_i)^2 + \dots + c_j(x - \mu_j)^2 + \dots) \end{aligned}$$

... = 0 → Tractable

Optimise ...  $\theta^*$

$$\log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} | \theta)$$

Intractable



Optimise ...  $\theta^*$

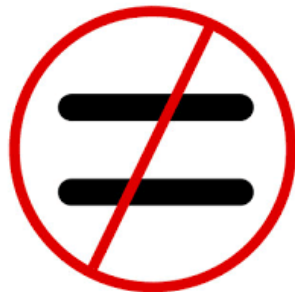
$$\sum_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z} | \theta)$$

Tractable

(when  $p$  belongs to the exponential family)

Optimise ...  $\theta^*$

$$\log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} | \theta)$$



Optimise ...  $\theta^*$

$$\sum_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z} | \theta)$$

Intractable

Tractable

*Your problem is to bridge the gap which exists between  
where you are now and the goal you intend to reach.*

*Earl Nightingale  
(1921-1989)*

EM

Optimise ...  $\theta^*$

$$\log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} | \theta)$$



Optimise ...  $\theta^*$

$$\sum_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z} | \theta)$$

Intractable

Tractable

*Your problem is to bridge the gap which exists between  
where you are now and the goal you intend to reach.*

*Earl Nightingale  
(1921-1989)*

# EM Derivation – Step 0

$$p(\mathbf{X}|\theta) = \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X}, \theta)}$$

**Step 0:** write  $p(\mathbf{X}|\theta)$  using *chain rule*

# EM Derivation – Step 1

$$\begin{aligned} p(\mathbf{X}|\theta) &= \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X}, \theta)} \frac{q(\mathbf{Z})}{q(\mathbf{Z})} \\ &= \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \end{aligned}$$

**Step 1:** Multiply right-hand side in a *special 1*



# EM Derivation – Step 2

$$\begin{aligned}\log p(\mathbf{X}|\theta) &= \log \left[ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right] \\ &= \log \left[ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right] + \log \left[ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right]\end{aligned}$$

**Step 2:** Take *log* from both sides

# EM Derivation – Step 3

$$q(\mathbf{Z}) \log p(\mathbf{X}|\theta) = q(\mathbf{Z}) \log \left[ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right] + q(\mathbf{Z}) \log \left[ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right]$$

**Step 3:** Multiply both sides by  $q(\mathbf{Z})$

# EM Derivation – Step 4

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right] + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right]$$

**Step 4:** *Marginalise* over  $\mathbf{Z}$

# EM Derivation – Step 4

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right] + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right]$$

$\log p(\mathbf{X}|\theta) = \dots$

**Step 4:** *Marginalise* over  $\mathbf{Z}$

# EM Derivation – Step 5

$$\log p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right] + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right]$$

# EM Derivation – Step 5.1

$$\log p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right] + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right]$$

# EM Derivation – Step 5.1

$$\log p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right] + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right]$$

$$D_{KL}(q \parallel p) \triangleq \sum_y q(y) \log \frac{q(y)}{p(y)}$$

- $D_{KL}$  (KL Divergence) properties:
  - ✓  $D_{KL}(q \parallel p) \geq 0$
  - ✓  $D_{KL}(q \parallel p) = 0 \iff q = p$

# EM Derivation – Step 5.2

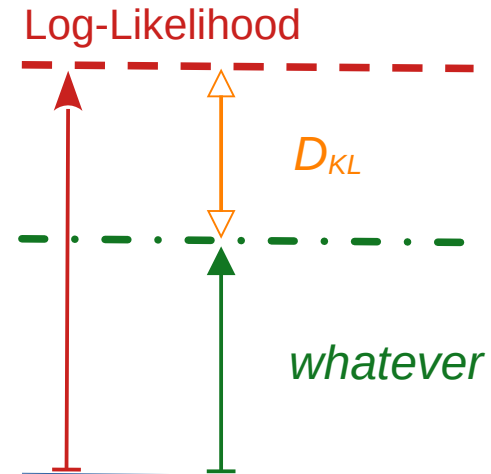
$$\log p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right] + \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right]}_{D_{KL}(q(\mathbf{Z}) \parallel p(\mathbf{Z}|\mathbf{X}, \theta))}$$

???



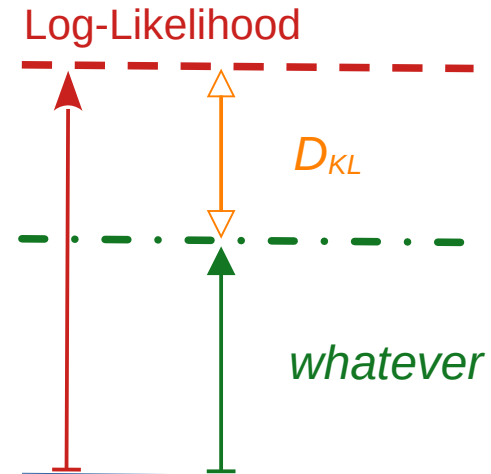
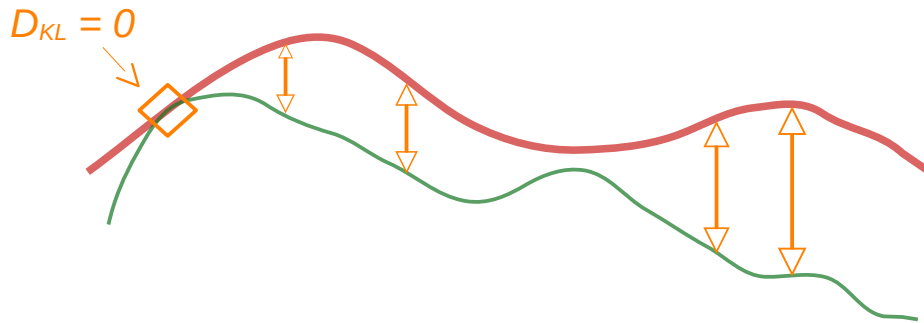
# EM Derivation – Step 5.2

$$\log p(\mathbf{X}|\theta) = \text{whatever} + \underbrace{D_{KL}(q(\mathbf{Z}) || p(\mathbf{Z}|\mathbf{X}, \theta))}_{\geq 0}$$



# EM Derivation – Step 5.2

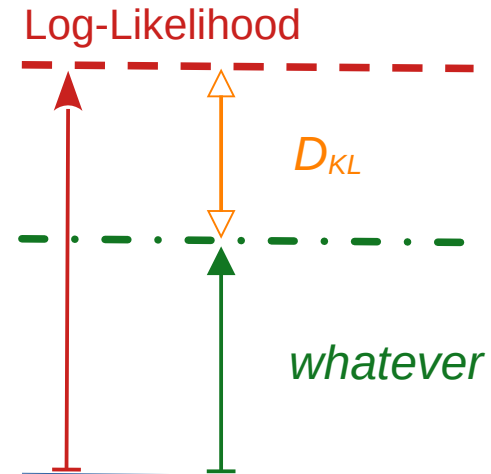
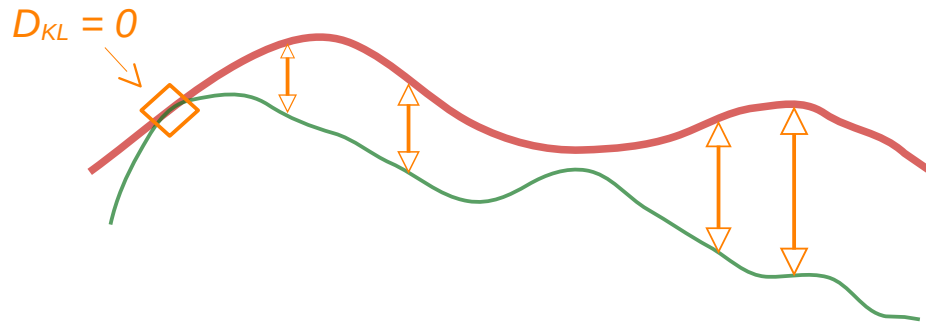
$$\log p(\mathbf{X}|\theta) = \text{whatever} + \underbrace{D_{KL}(q(\mathbf{Z}) || p(\mathbf{Z}|\mathbf{X}, \theta))}_{\geq 0}$$



# EM Derivation – Step 5.2

$$\log p(\mathbf{X}|\theta) = \text{whatever} + \underbrace{D_{KL}(q(\mathbf{Z}) || p(\mathbf{Z}|\mathbf{X}, \theta))}_{\geq 0}$$

$$\log p(\mathbf{X}|\theta) \geq \text{whatever}$$



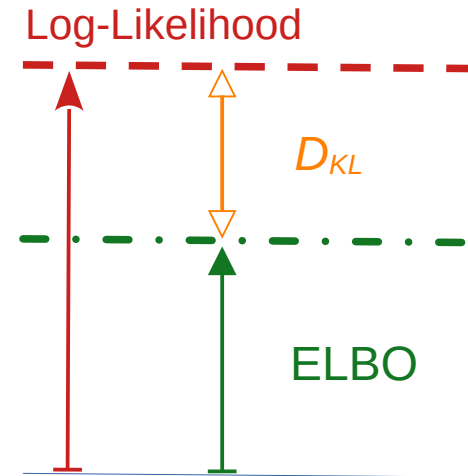
# EM Derivation – Step 5.2

$$\log p(\mathbf{X}|\theta) = \text{ELBO} + \underbrace{D_{KL}(q(\mathbf{Z}) || p(\mathbf{Z}|\mathbf{X}, \theta))}_{\geq 0}$$

$$\log p(\mathbf{X}|\theta) \geq \text{ELBO}$$

**E**vidence **L**ower **B**ound (ELBO)

Evidence  $\equiv$  log-likelihood



# EM Derivation – Step 5

$$\underbrace{\log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z})}_{\log p(\mathbf{X}|\theta)} = \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right]}_{ELBO(q, \theta)} + \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right]}_{D_{KL}(q, \theta)}$$

# EM Derivation – Step 6

$$\underbrace{\log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z})}_{\log p(\mathbf{X}|\theta)} = \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right]}_{ELBO(q, \theta)} + \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right]}_{D_{KL}(q, \theta)}$$

$$\theta_{ML}^* = \operatorname{argmax}_{\theta \in \Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

Intractable!

# Direction solution does not work!

$$\log \sum_z p(x, z|\theta) = \log(\dots + w_{z_i} e^{\frac{(x-\mu_i)^2}{2\sigma_i^2}} + \dots + w_{z_j} e^{\frac{(x-\mu_j)^2}{2\sigma_j^2}} + \dots)$$

... = 0 → Intractable

... = 0 → Tractable

RECALL slide 9

IFF we could swap log &  $\Sigma$  ...

$$\begin{aligned} \sum_z \log p(x, z|\theta) &= (\dots + w_i \log e^{\frac{(x-\mu_i)^2}{2\sigma_i^2}} + \dots + w_j \log e^{\frac{(x-\mu_j)^2}{2\sigma_j^2}} + \dots) \\ &= (\dots + c_i(x - \mu_i)^2 + \dots + c_j(x - \mu_j)^2 + \dots) \end{aligned}$$

# EM Derivation – Step 5

$$\underbrace{\log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z})}_{\log p(\mathbf{X}|\theta)} = \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right]}_{ELBO(q,\theta)} + \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right]}_{D_{KL}(q,\theta)}$$

Intractable!

$$\log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$



Tractable

$$\sum_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z}|\theta)$$



# EM Derivation – Step 6

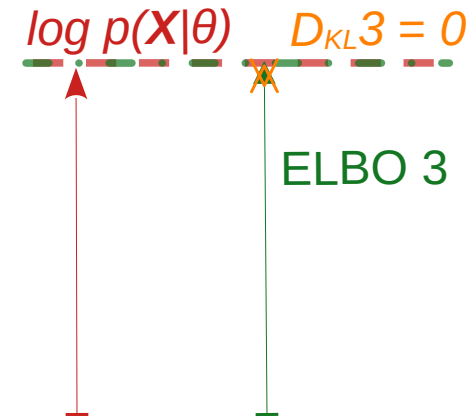
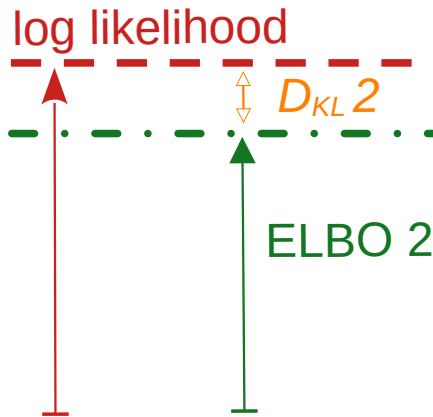
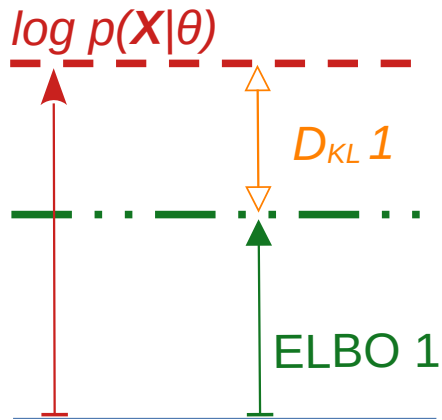
$$\underbrace{\log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z})}_{\log p(\mathbf{X}|\theta)} = \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right]}_{ELBO(q, \theta)} + \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \right]}_{D_{KL}(q, \theta)}$$

$$\theta^* = \operatorname{argmax}_{\theta \in \Theta} ELBO$$

Tractable

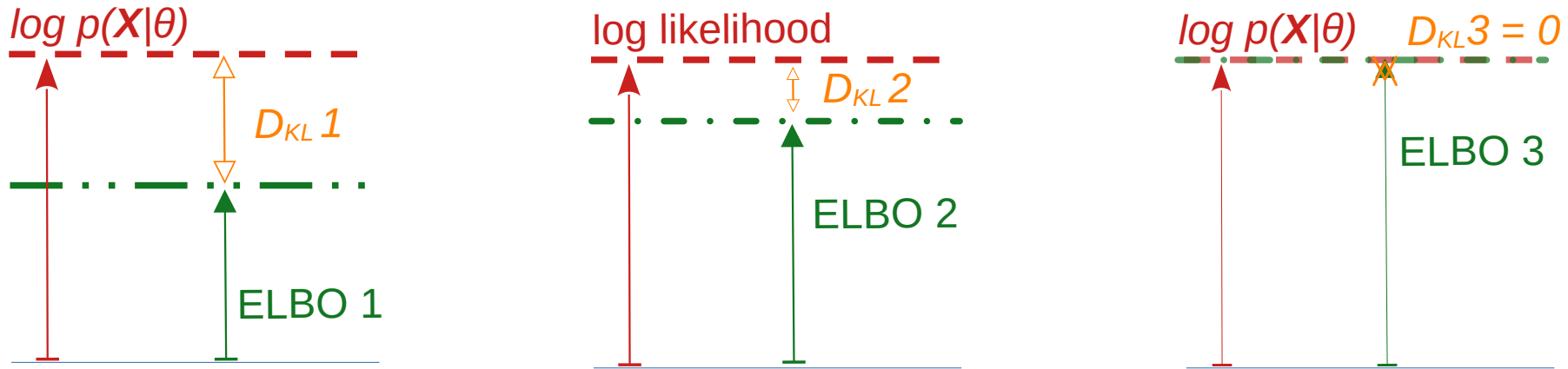
**EM** → Instead of **log-likelihood** ... maximise **ELBO** ...

# Best ELBO to optimise ...



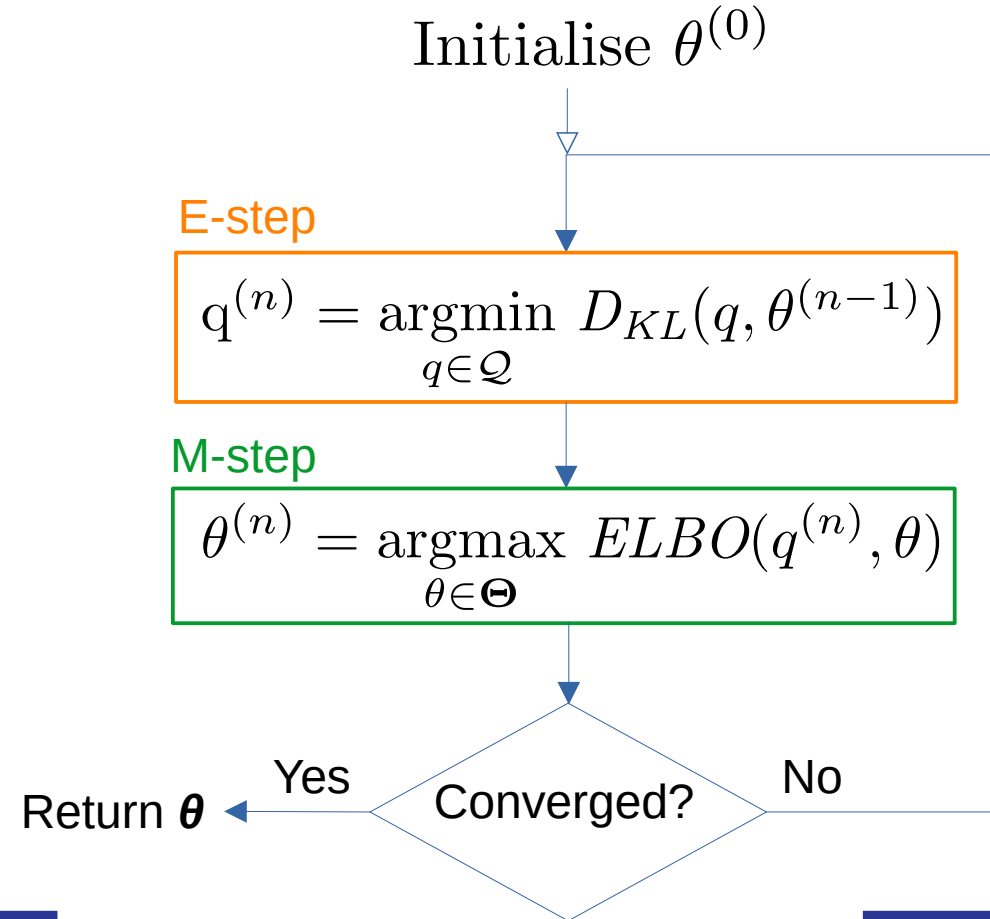
Which ELBO is better?

# Best ELBO to optimise ...



- Lower  $D_{KL}$   $\rightarrow$  Better ELBO (closer to  $\log p(\mathbf{X}|\theta)$ )
- **IDEAL:**  $D_{KL} = 0 \rightarrow$  Best ELBO =  $\log p(\mathbf{X}|\theta)$

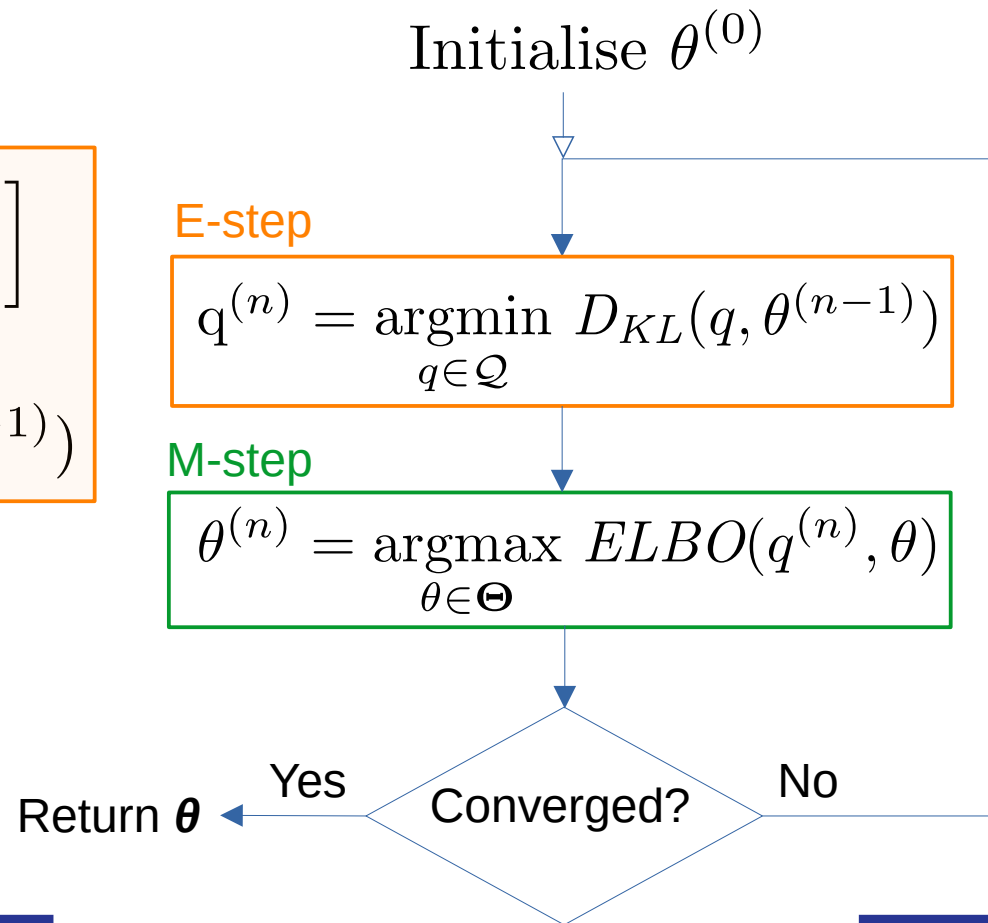
# EM Procedure



# E-Step

$$D_{KL}(q, \theta^{(n-1)}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta^{(n-1)})} \right]$$

$$\text{Min } D_{KL} = 0 \Leftrightarrow q^{(n)}(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{(n-1)})$$



# M-Step

$$\begin{aligned} ELBO(q^{(n)}, \theta) &= \sum_{\mathbf{Z}} q^{(n)}(\mathbf{Z}) \log \left[ \frac{p(\mathbf{X}, \mathbf{Z} | \theta)}{q^{(n)}(\mathbf{Z})} \right] \\ &= \underbrace{\sum_{\mathbf{Z}} q^{(n)}(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z} | \theta)}_{\text{constant}} - \underbrace{\sum_{\mathbf{Z}} q^{(n)}(\mathbf{Z}) \log q^{(n)}(\mathbf{Z})}_{\text{constant}} \end{aligned}$$

# M-Step

$$\begin{aligned} ELBO(q^{(n)}, \theta) &= \sum_{\mathbf{Z}} q^{(n)}(\mathbf{Z}) \log \left[ \frac{p(\mathbf{X}, \mathbf{Z} | \theta)}{q^{(n)}(\mathbf{Z})} \right] \\ &= \underbrace{\sum_{\mathbf{Z}} q^{(n)}(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z} | \theta)}_{\text{constant}} - \underbrace{\sum_{\mathbf{Z}} q^{(n)}(\mathbf{Z}) \log q^{(n)}(\mathbf{Z})}_{\text{constant}} \end{aligned}$$

---

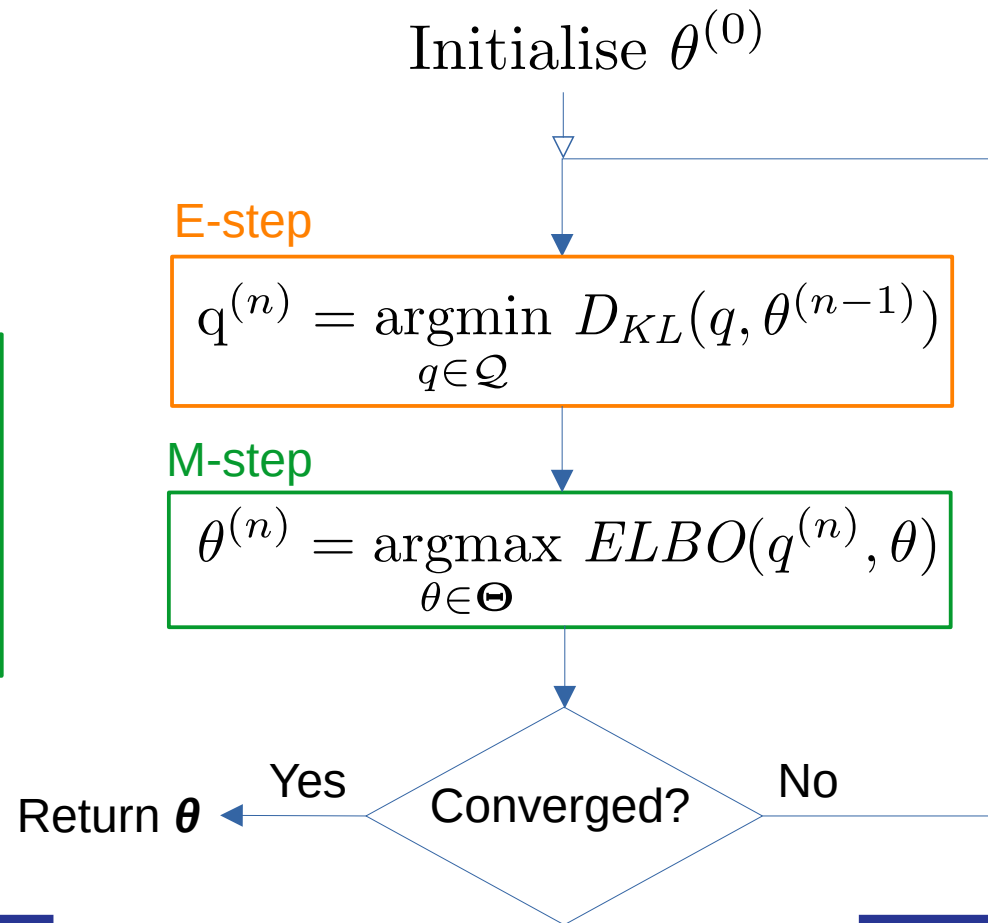
$$\theta^* = \operatorname{argmax}_{\theta \in \Theta} ELBO(q^{(n)}, \theta)$$

$$= \operatorname{argmax}_{\theta \in \Theta} \sum_{\mathbf{Z}} q^{(n)}(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z} | \theta)$$

# M-Step

M-step

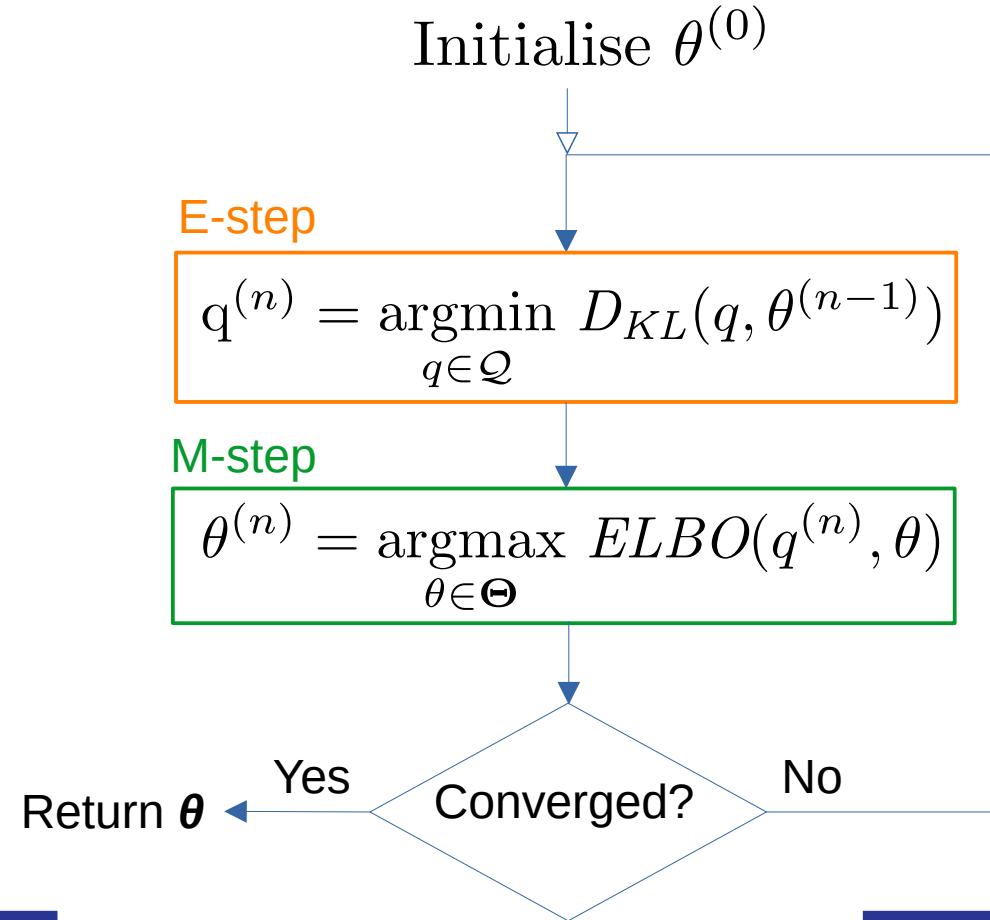
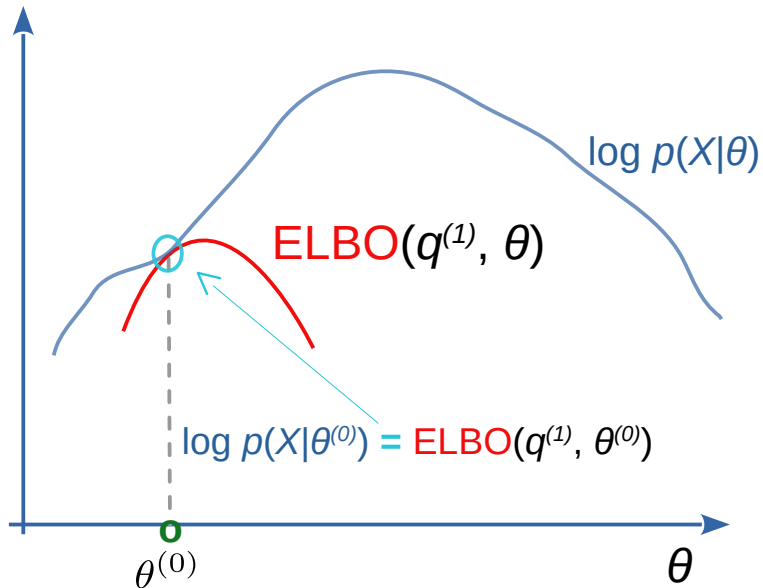
$$\begin{aligned}\theta^* &= \operatorname{argmax}_{\theta \in \Theta} ELBO(q^{(n)}, \theta) \\ &= \operatorname{argmax}_{\theta \in \Theta} \sum_{\mathbf{Z}} q^{(n)}(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z} | \theta)\end{aligned}$$





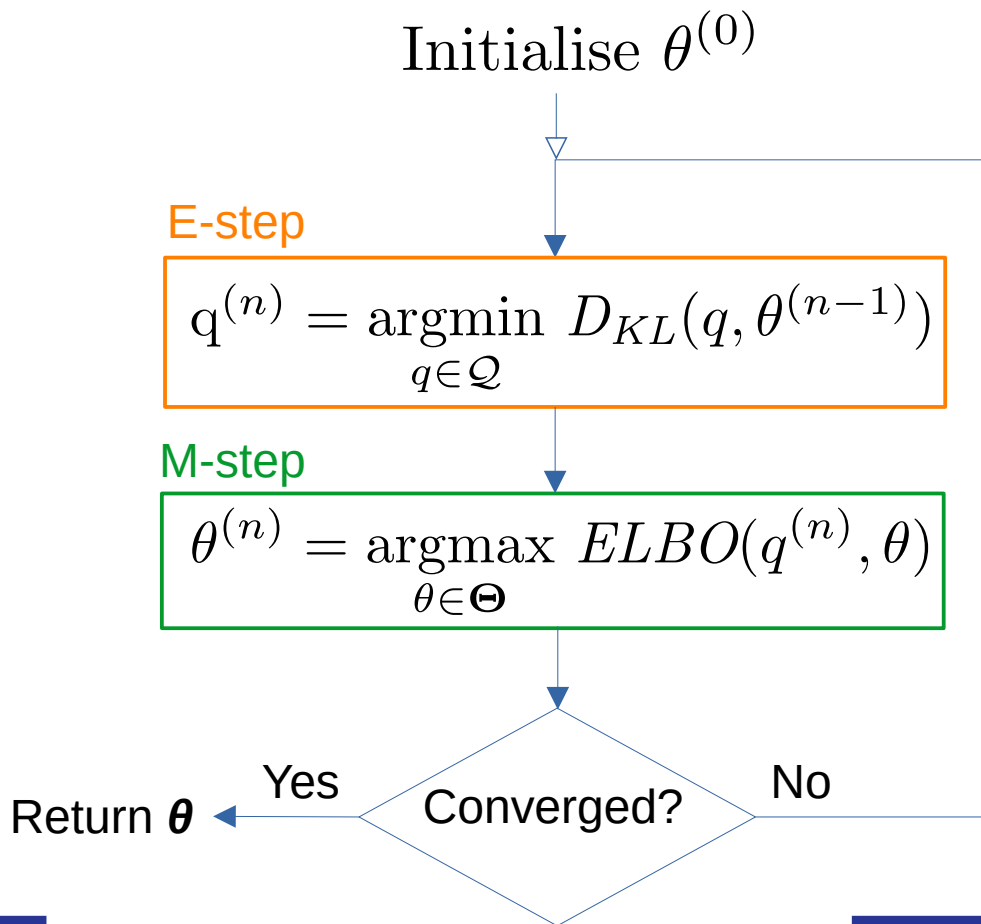
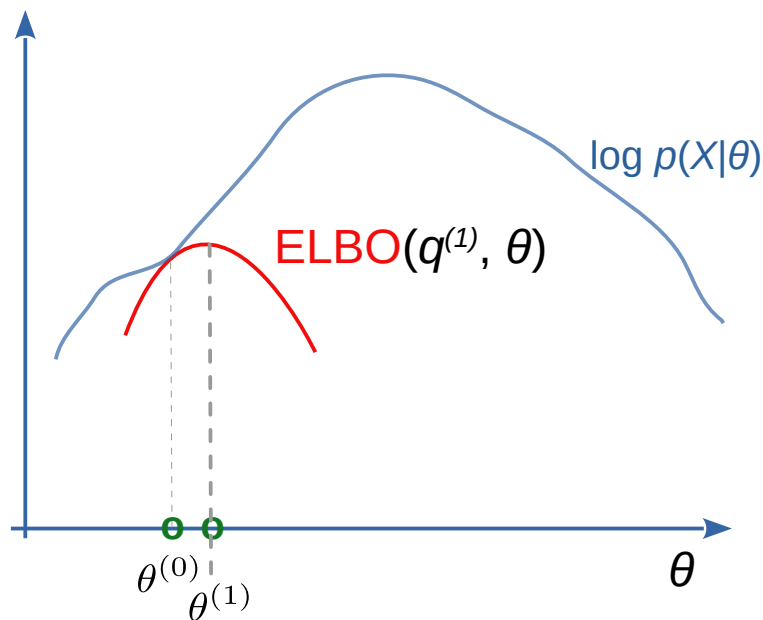
# Visualisation

*E-step:*  $q^{(1)} = \operatorname{argmin} D_{KL}(q, \theta^{(0)})$



# Visualisation

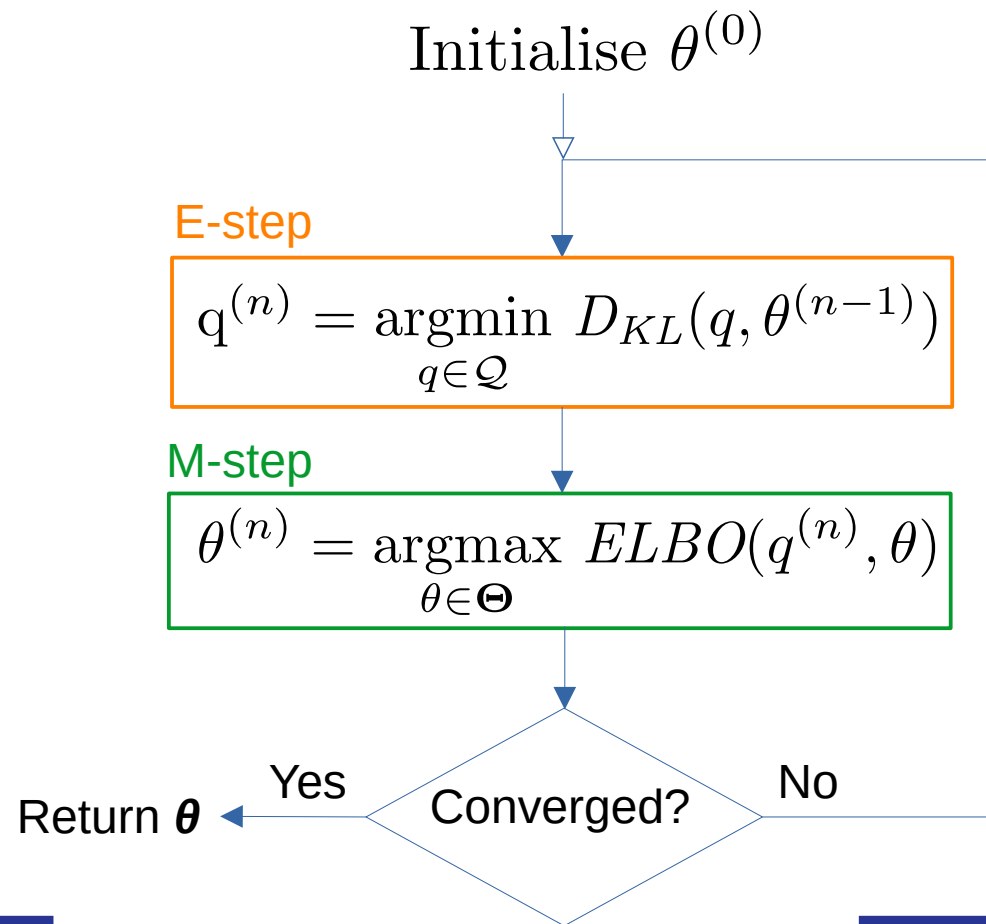
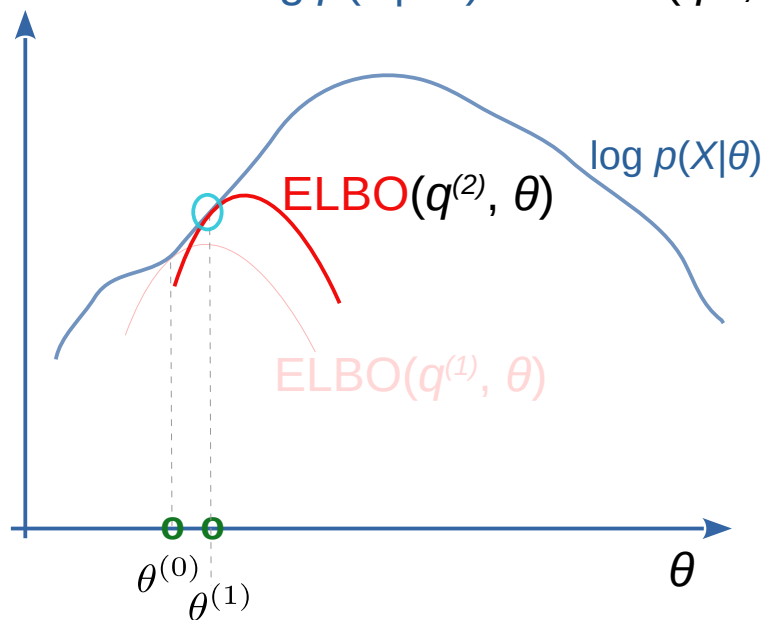
*M-step:*  $\theta^{(1)} = \operatorname{argmax} \text{ELBO}(q^{(1)}, \theta)$



# Visualisation

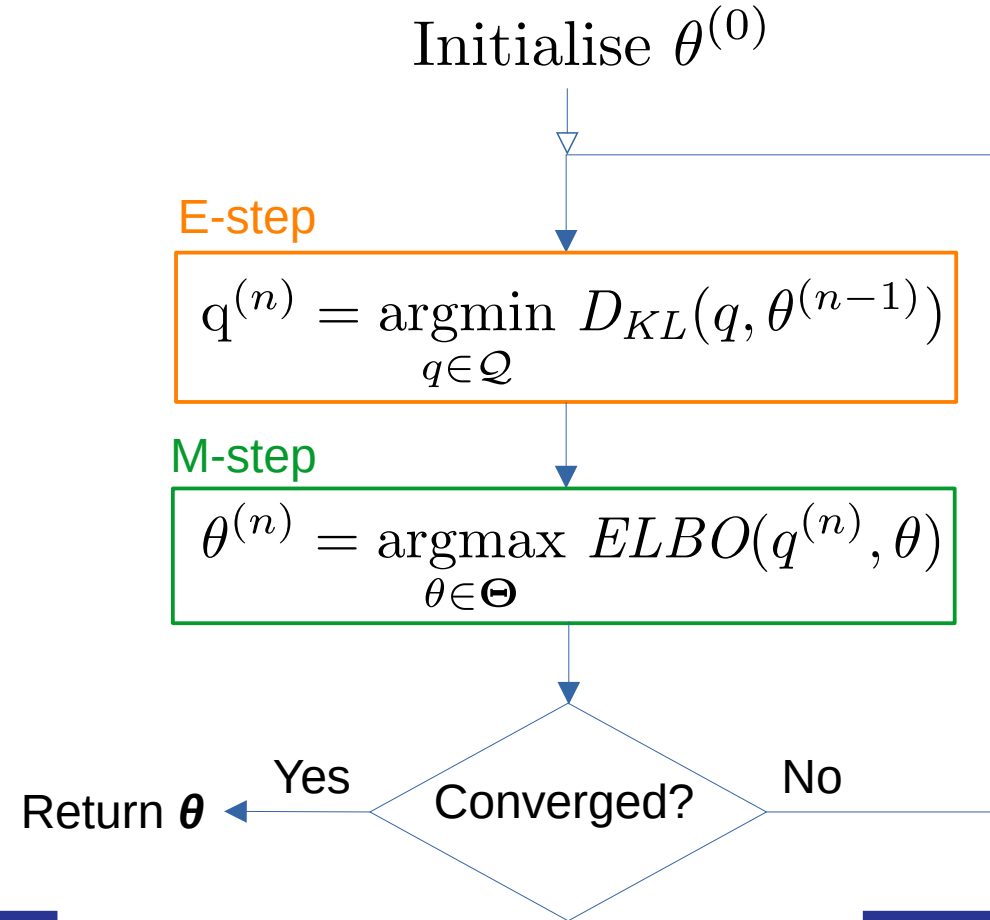
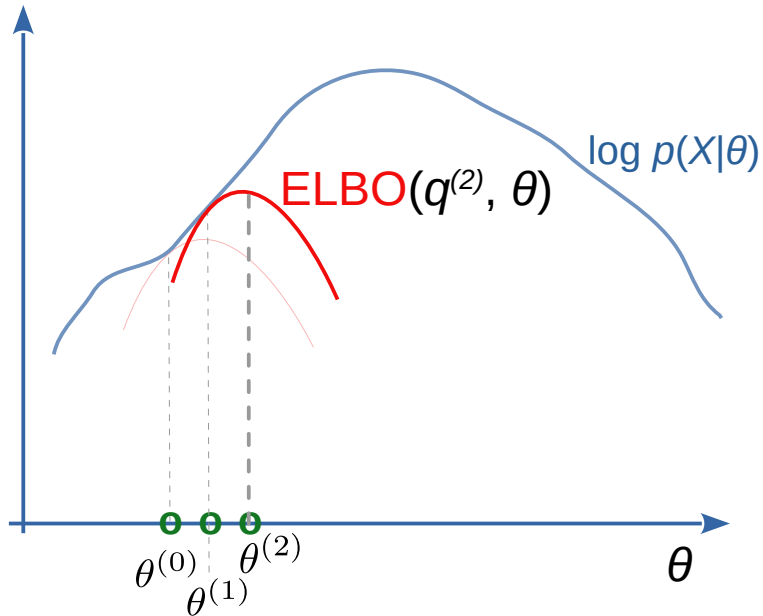
*E-step:*  $q^{(2)} = \operatorname{argmin} D_{KL}(q, \theta^{(1)})$

$$\log p(X|\theta^{(1)}) = \text{ELBO}(q^{(2)}, \theta^{(1)})$$



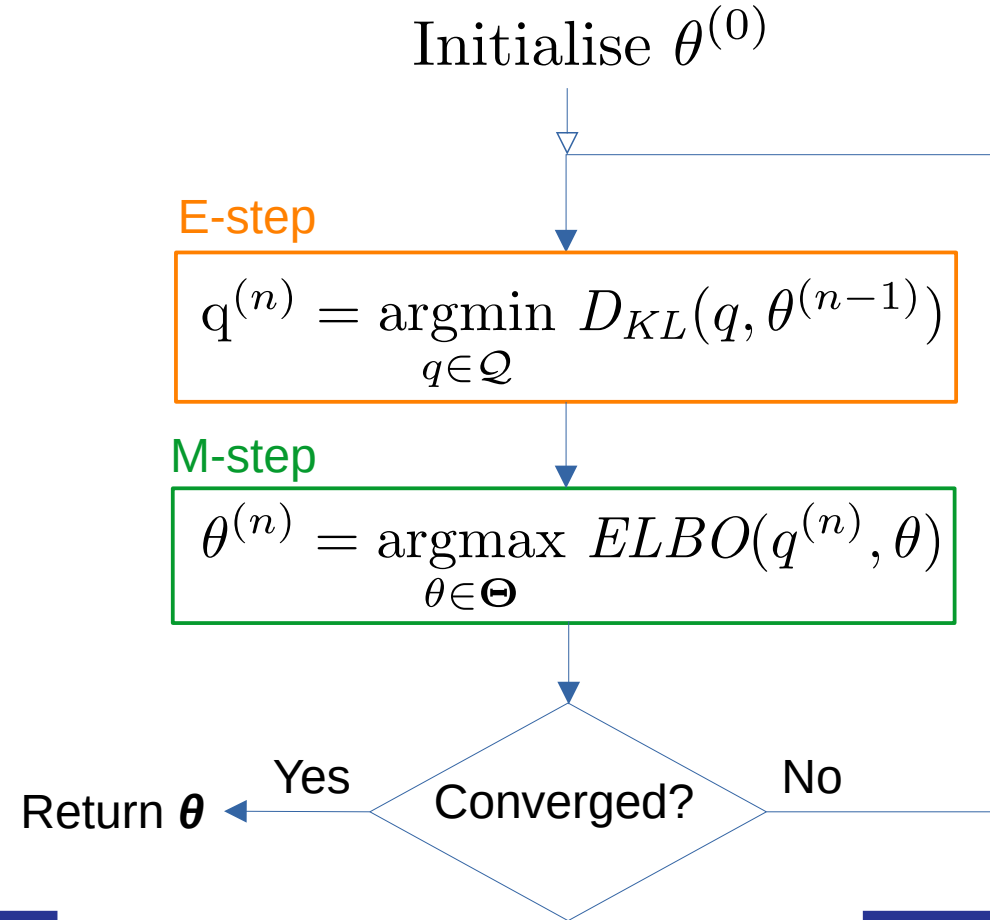
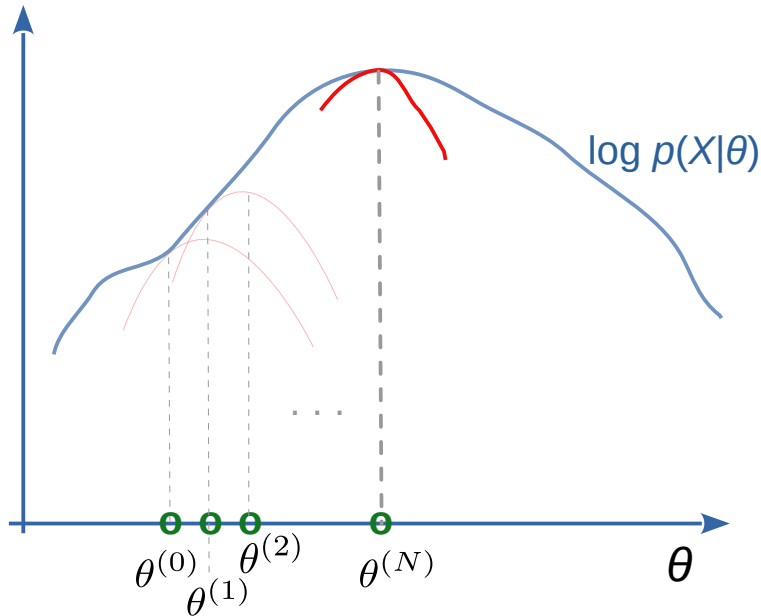
# Visualisation

*M-step:*  $\theta^{(2)} = \operatorname{argmax} \text{ELBO}(q^{(2)}, \theta)$

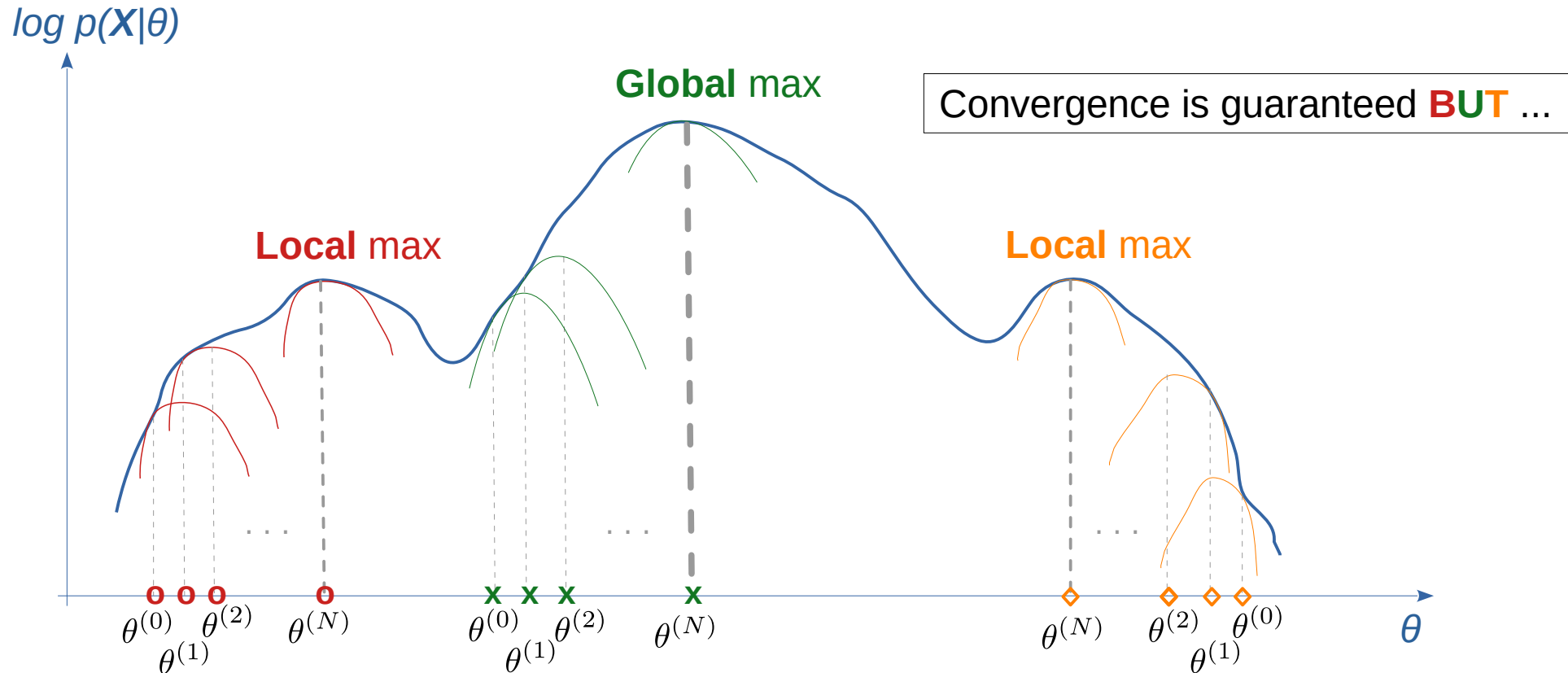


# Visualisation

*M-step:*  $\theta^{(N)} = \operatorname{argmax}_{\theta} \text{ELBO}(q^{(N)}, \theta)$

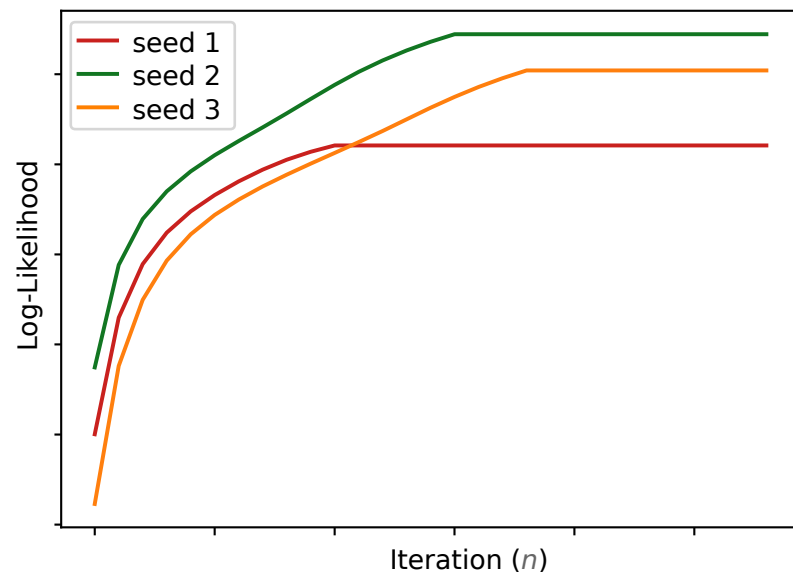


# Initialisation Matters (1)



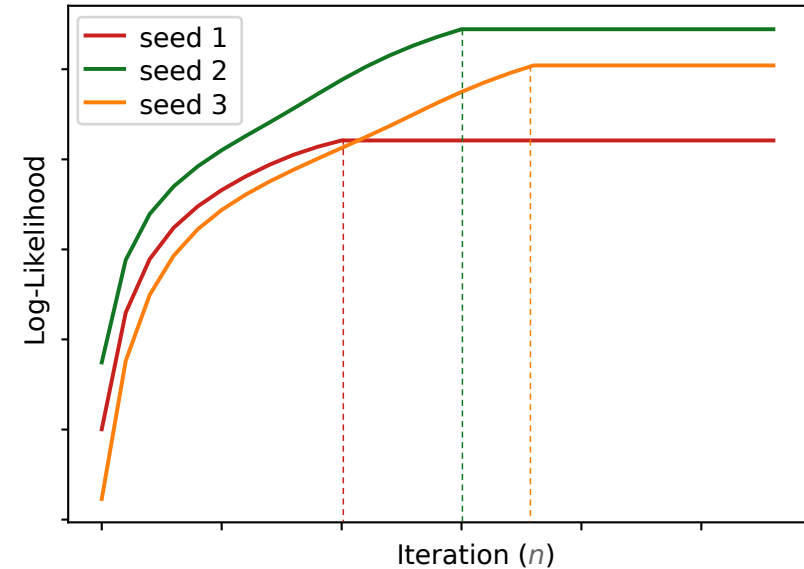
# Initialisation Matters (2)

- $p(\mathbf{X}|\theta^{(n)})$  is **ALWAYS** non-decreasing
  - ✓  $ELBO(q^{(n+1)}, \theta^{(n+1)}) \geq ELBO(q^{(n)}, \theta^{(n)})$
  - ✓ Convergence is guaranteed ... but to local optimum ...



# Initialisation Matters (3)

- $p(\mathbf{X}|\theta^{(n)})$  is **ALWAYS** non-decreasing
- Initialisation affects ... **convergence rate** and **final log-likelihood**

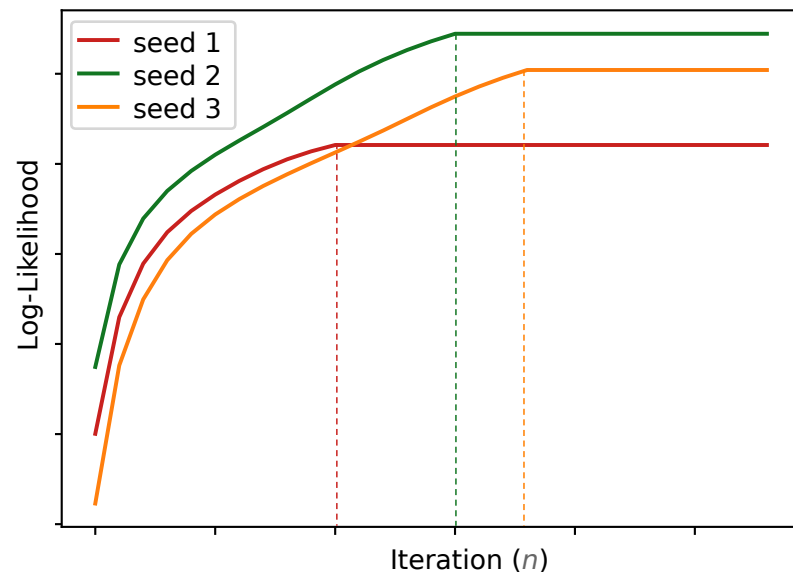




# Initialisation Matters (3)

- $p(X|\theta^{(n)})$  is **ALWAYS** non-decreasing
- Initialisation affects ... convergence rate and final log-likelihood

Try **multiple** initialisations and pick up the **best** local optimum (**best**  $\equiv$  **highest log-likelihood**).



# Other Considerations

- EM is closely related to **Coordinate Ascent** [App B]
  - E-step: fix  $\theta$ , optimise  $q$
  - M-step: fix  $q$ , optimise  $\theta$
- EM shines when M-step can be solved analytically
- Alternatives when M-step is intractable ...
  - Generalised EM (**GEM**)  $\rightarrow$  (conjugate) gradient ascent in M-Step
  - Expectation Conditional Maximisation (**ECM**)  $\rightarrow$  coordinate ascent in M-step

# Wrap-up ... EM ...

- **Goal:** estimate  $\theta_{ML}$  for probabilistic models with latent var
- **How:** an iterative two-stage (E-step, M-step) procedure
- **Applications:** GMM, HMM, Computational biology, ...
- **Assignment:** estimate  $\theta_{MAP}$  using EM
- **Appendices**
  - (A) Further Reading
  - (B) Coordinate Ascent

# (A) Further Reading

- **Murphy**, Chapter 8, Section 7.2, Pages 306-310
- **Bishop**, Chapter 9, Section 4, Pages 450-455
- **Andrew Ng's** [Lecture Notes](#), Chapter 11, Pages 142-147
- Others: [blog1](#) [blog2](#)

# (B) Coordinate Ascent

- **Iterative** optimisation method for **multi-variate** functions  $f(\mathbf{x})$
- **Idea:** **Maximise** over one variable (or a block of variables) at a time, assuming others are constants

## Procedure:

INITIALISE  $\mathbf{x}^{(0)} = [x_0, x_1, \dots, x_{D-1}]$

FOR  $i$  in range(**#iterations**):

FOR  $d$  in range( $D$ ):

$\mathbf{x}^{(i)}[d] \leftarrow \underset{x_d}{\operatorname{argmax}} f(\mathbf{x}^{(i)}[:d], x_d, \mathbf{x}^{(i-1)}[d+1:])$

constants

axis-aligned movements

