Variational Inference

Probabilistic Decision Making — Lecture 9

26th November 2025

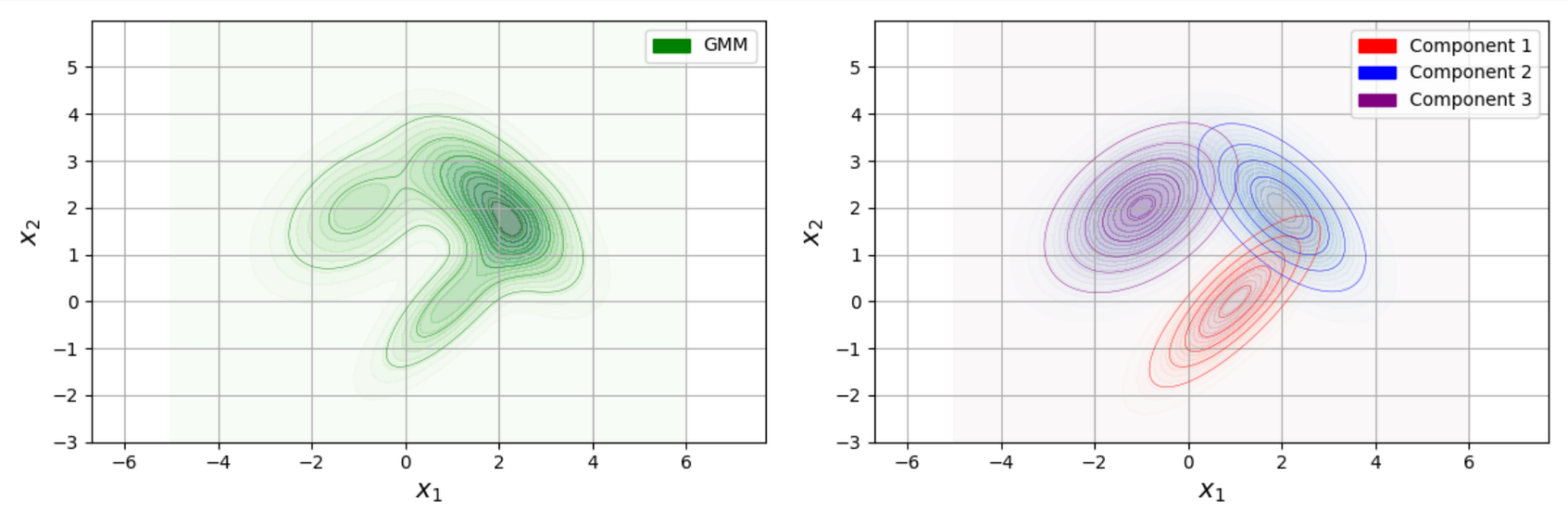
Robert Peharz
Institute of Machine Learning and Neural Computation
Graz University of Technology

$Z \in \{1, \ldots, K\}$ **latent variable**

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \, p(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \Sigma_k)$$

$$p(\boldsymbol{x}, z) = \overbrace{p(z)}^{w_z} \overbrace{p(\boldsymbol{x} \mid z)}^{p(\boldsymbol{x} \mid \boldsymbol{\mu}_z, \Sigma_z)}$$

$$p(\boldsymbol{x}) = \sum_{z} p(z) \, p(\boldsymbol{x} \mid z)$$



Expectation Maximization for GMMs

input: Data $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}\}$

output: Learned GMM parameters $\boldsymbol{\theta} = \{w_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^{K}$

Initialize $\boldsymbol{\theta} = \{w_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^{K}$ randomly

for *counter* $= 1 \ldots$ *max number of iterations* **do**

  E-Step

  for $i = 1 \ldots N$, $k = 1 \ldots K$ **do**

    $\gamma_{i,k} \leftarrow p(z^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\mu}_k, \Sigma_k)\, w_k}{\sum_{k'} p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\mu}_{k'}, \Sigma_{k'})\, w_{k'}}$ ▷ responsibilities

  end

  M-Step

  for $k = 1 \ldots K$ **do**

    $\boldsymbol{\mu}_k \leftarrow \frac{\sum_{i=1}^{N} \gamma_{i,k}\, \boldsymbol{x}^{(i)}}{\sum_{i=1}^{N} \gamma_{i,k}}$ ▷ weighted mean

    $\Sigma_k \leftarrow \frac{\sum_{i=1}^{N} \gamma_{i,k} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_k)(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^{N} \gamma_{i,k}}$ ▷ weighted covariance

    $w_k \leftarrow \frac{\sum_{i=1}^{N} \gamma_{i,k}}{N}$ ▷ weighted empirical frequencies

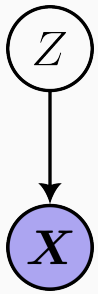
  end

end

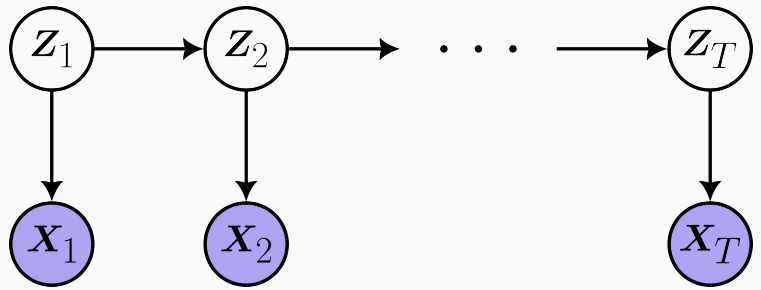
Latent Variable Models

- **latent variables** are a powerful modeling concept in probabilistic ML
- **mixture models** are one example of this concept
- but the concept is much broader, e.g. we can use Bayesian networks with latent variables
- famous example: **hidden Markov models** for sequential data

(Gaussian) Mixture Model



Hidden Markov Model



How to learn models with latent variables or missing data?

EM = maximum likelihood under unobserved data

Given: joint model $p(\boldsymbol{x}, \boldsymbol{z} \mid \theta)$ over observed $\boldsymbol{X}$ and unobserved $\boldsymbol{Z}$

- **initialize** parameters $\boldsymbol{\theta}_{\text{cur}}$
- **iterate**
  - **step 1** — using the current model, **infer** posterior over latents for each sample $i = 1 \ldots N$:

$$p(\boldsymbol{z}^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}_{\text{cur}})$$

  - **step 2** — compute **expected complete log-likelihood**:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{cur}}) = \sum_{i=1}^{N} \mathbb{E}_{p(\boldsymbol{z}^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}_{\text{cur}})} \left[\log p(\boldsymbol{x}^{(i)}, \boldsymbol{Z}^{(i)} \mid \boldsymbol{\theta}) \right]$$

  - **step 3** — maximize:

$$\boldsymbol{\theta}_{\text{new}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{cur}}) \quad \text{and let } \boldsymbol{\theta}_{\text{cur}} \leftarrow \boldsymbol{\theta}_{\text{new}}$$

Why does EM work?

- EM increases monotonically the log-likelihood
- meaning, in each iteration, the marginal log-likelihood can only increase or stay the same—**it can never decrease**
- applicable to missing data in general, not only mixture models
- **but, how and why does EM work?**

Key concept: the **evidence lower bound (ELBO)**, also called **variational lower bound**, **variational objective**, **variational free energy**

Evidence Lower Bound

Setting

- let a parametric joint model $p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})$ be given, where
  - $\boldsymbol{X}$ are observed
  - $\boldsymbol{Z}$ are unobserved
- **goal:** maximize **marginal log-likelihood** aka **evidence**:

$$\begin{aligned}\arg\max_{\boldsymbol{\theta}} \log p(\mathcal{D} \mid \boldsymbol{\theta}) &= \sum_{i=1}^{N} \log p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \log \int p(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)} \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{z}^{(i)}\end{aligned}$$

- **just standard maximum likelihood for latent variable models, i.e. models that are defined as marginal of a bigger model including latent RVs.**
- for discrete $\boldsymbol{Z}$, the integral is replaced by a sum

In GMMs, each sample has a single latent variable with K states:

$$\begin{aligned}\arg\max_{\boldsymbol{\theta}} \quad \log p(\mathcal{D} \mid \boldsymbol{\theta}) &= \sum_{i=1}^{N} \log p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \log \sum_{z^{(i)}} p(\boldsymbol{x}^{(i)}, z^{(i)} \mid \boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \log \sum_{z^{(i)}} w_{z^{(i)}}\, p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\mu}_{z^{(i)}}, \Sigma_{z^{(i)}})\end{aligned}$$

where $\boldsymbol{\theta} = \{w_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^{K}$ are all params of the GMM.

- for each sample i, introduce a **variational distribution** q_i, which can be **any arbitrary** distribution over $\boldsymbol{Z}^{(i)}$, the unobserved variables in the i^{th} sample
- often, the variational distribution might be **parametric**, but for now we treat them as **non-parametric objects**, without worrying how they are represented
- let $\boldsymbol{q} = \{q_i\}_{i=1}^{N}$ be the set of all variational distributions

Evidence Lower Bound (ELBO)

Recall the **marginal log-likelihood** aka **evidence**:

$$\log p(\mathcal{D} \mid \theta) = \sum_{i=1}^{N} \overbrace{\log \int p(x^{(i)}, z^{(i)} \mid \theta) \, \mathrm{d}z^{(i)}}^{\text{sample-wise log-likelihood } \log p(x^{(i)} \mid \theta)}$$

We derive the **sample-wise evidence lower bound (ELBO)**:

$$\begin{aligned}
\log p(x^{(i)} \mid \theta) &= \log \int \overbrace{\frac{q_i(z^{(i)})}{q_i(z^{(i)})}}^{=1} p(x^{(i)}, z^{(i)} \mid \theta) \, \mathrm{d}z^{(i)} \\
&= \log \int q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)} \mid \theta)}{q_i(z^{(i)})} \, \mathrm{d}z^{(i)} \\
&= \log \mathbb{E}_{q_i} \left[\frac{p(x^{(i)}, Z^{(i)} \mid \theta)}{q_i(Z^{(i)})} \right] \\
&\overset{\text{Jensen}}{\geq} \mathbb{E}_{q_i} \left[\log \frac{p(x^{(i)}, Z^{(i)} \mid \theta)}{q_i(Z^{(i)})} \right] =: \mathrm{ELBO}(\theta, q_i)
\end{aligned}$$

Interlude: Jensen's Inequality

In the last inequality, we were using

$$\log \mathbb{E}[\dots] \geq \mathbb{E}[\log \dots]$$

This is due to **Jensen's inequality**: Let $\boldsymbol{X}$ be any (set of) random variable(s) and g be any **concave** function (e.g. the log). Then

$$g(\mathbb{E}[\boldsymbol{X}]) \geq \mathbb{E}[g(\boldsymbol{X})]$$

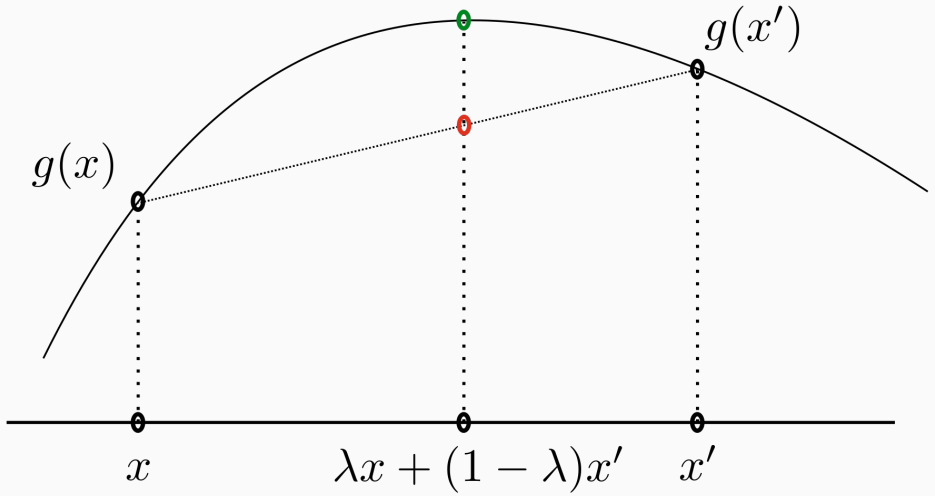
By the way, if g is a **convex** function then Jensen's inequality reverses:

$$g(\mathbb{E}[\boldsymbol{X}]) \leq \mathbb{E}[g(\boldsymbol{X})]$$

Memorizing Jensen's Inequality

Example

$$g(\lambda x + (1-\lambda)x') \geq \lambda g(x) + (1-\lambda)g(x')$$



- inequality above is the **definition of concavity**
- $\lambda x + (1-\lambda)x'$ is a **convex combination** (weighted average) of x and x', for $\lambda \in [0, 1]$
- $\mathbb{E}$ **operator is also a convex combination!**
- Jensen's inequality: definition of concavity (and convexity) generalizes to expectations

Evidence Lower Bound (ELBO) cont'd

Sample-wise ELBO:

$$\log p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}) \geq \mathbb{E}_{q_i}\left[\log \frac{p(\boldsymbol{x}^{(i)}, \boldsymbol{Z}^{(i)} \mid \boldsymbol{\theta})}{q_i(\boldsymbol{Z}^{(i)})}\right] =: \text{ELBO}(\boldsymbol{\theta}, q_i)$$

The **total ELBO** (or just **ELBO**) is

$$\log p(\mathcal{D} \mid \boldsymbol{\theta}) = \sum_i \log p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}) \geq \sum_i \text{ELBO}(\boldsymbol{\theta}, q_i) =: \text{ELBO}(\boldsymbol{\theta}, \boldsymbol{q})$$

Note that (total) ELBO is a function of both model params $\boldsymbol{\theta}$ and the set of all variational distributions $\boldsymbol{q} = \{q_i\}_{i=1}^N$. Of course, it also depends on the observed data, but this dependency is usually omitted in notation.

What is the ELBO for?

For **any** set of variational distributions $\boldsymbol{q} = \{q_i\}_{i=1}^N$ the ELBO is a lower bound of the marginal log-likelihood:

$$\log p(\mathcal{D} \mid \boldsymbol{\theta}) \geq \text{ELBO}(\boldsymbol{\theta}, \boldsymbol{q})$$

Since we aim to

$$\max_{\boldsymbol{\theta}} \log p(\mathcal{D} \mid \boldsymbol{\theta})$$

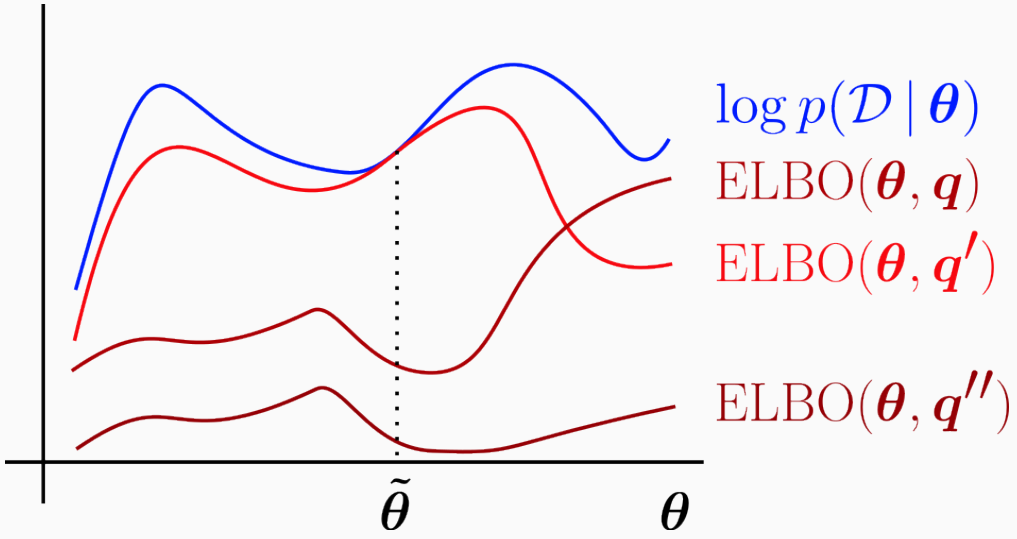
it is also meaningful to

$$\max_{\boldsymbol{\theta}} \text{ELBO}(\boldsymbol{\theta}, \boldsymbol{q})$$

Maximizing ELBO might sometimes be easier than maximizing the log-likelihood.

Proof for Expectation-Maximization

Illustrating log-likelihood and three ELBOs, for different sets of variational distributions $\boldsymbol{q}, \boldsymbol{q'}, \boldsymbol{q''}$



For any $\boldsymbol{q}$, it is always true that $\mathrm{ELBO}(\boldsymbol{\theta}, \boldsymbol{q}) \leq p(\mathcal{D} \mid \boldsymbol{\theta})$.
Moreover, for some $\boldsymbol{q}$ and some $\tilde{\boldsymbol{\theta}}$, it might happen that

$$\mathrm{ELBO}(\tilde{\boldsymbol{\theta}}, \boldsymbol{q}) = \log p(\mathcal{D} \mid \tilde{\boldsymbol{\theta}})$$

In this case we say that the ELBO is **tight** at $\tilde{\boldsymbol{\theta}}$.

In the example above, the ELBO corresponding to $\boldsymbol{q'}$ is tight at $\tilde{\boldsymbol{\theta}}$.

Interlude: Minorization-Maximization

Forgetting about ELBOs for a moment, assume we want to maximize some function

$$\max_{\boldsymbol{x}} \; f(\boldsymbol{x})$$

but f is "hard to optimize" (not differentiable, expensive to evaluate, highly non-convex, etc.)

Assume that for each $\tilde{\boldsymbol{x}}$ we can construct a function g

- that is a lower bound, i.e. for all $\boldsymbol{x}$ it holds that $g(\boldsymbol{x}) \leq f(\boldsymbol{x})$
- is **tight** at $\tilde{\boldsymbol{x}}$, that is $g(\tilde{\boldsymbol{x}}) = f(\tilde{\boldsymbol{x}})$

Such g is called a **minorization** of f at point $\tilde{\boldsymbol{x}}$. The idea is to construct a g that is easier to maximize than f.

Interlude: Minorization-Maximization

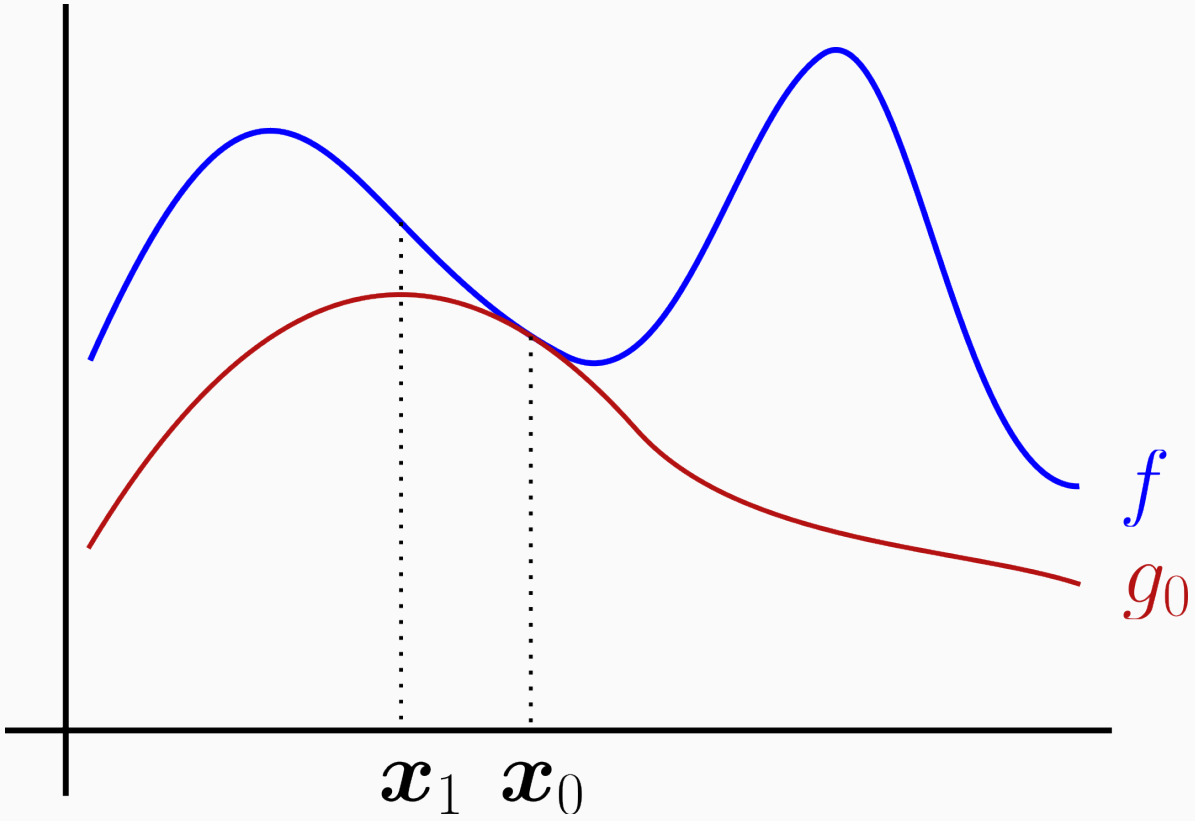
The **minorization-maximization** (MM) principle proceeds as follows:

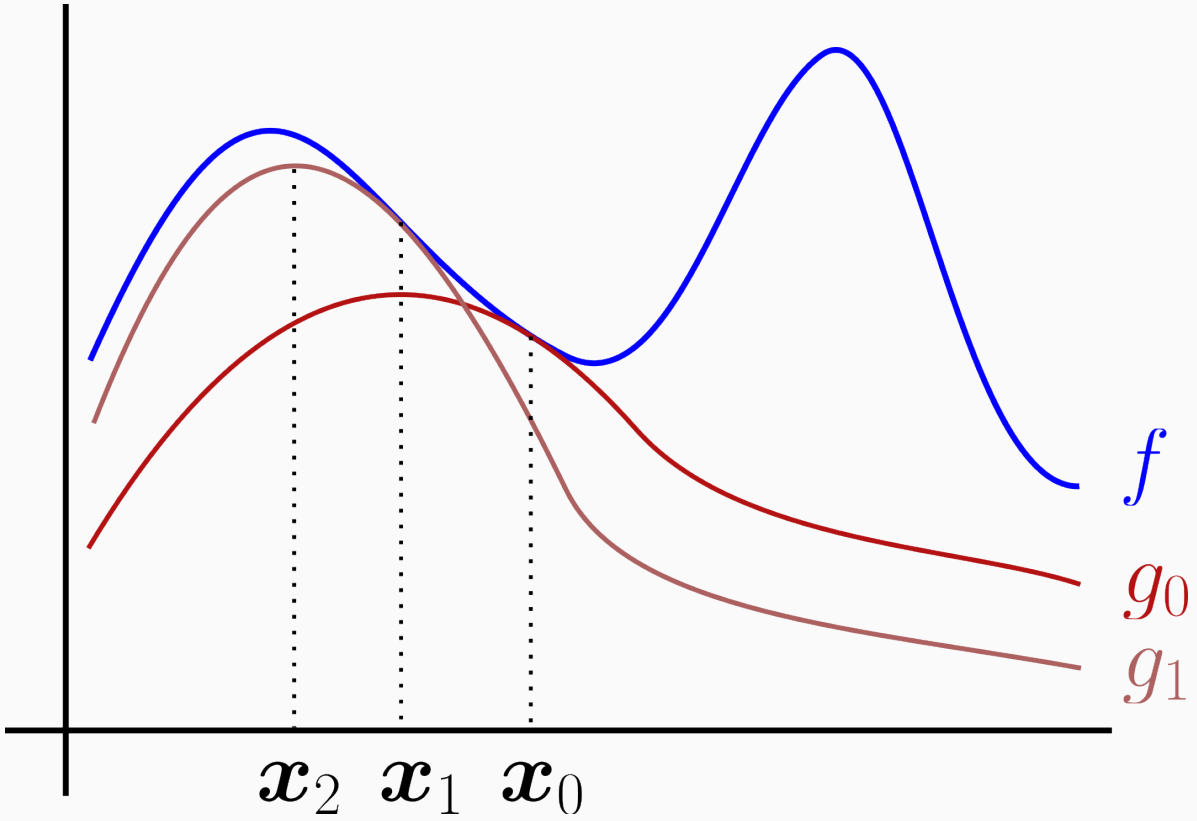
- initialize $\boldsymbol{x}$
- iterate:
  - construct minorization g at $\boldsymbol{x}$
  - compute $\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} g(\boldsymbol{x})$
  - $\boldsymbol{x} \leftarrow \boldsymbol{x}^*$

It follows

$$f(\boldsymbol{x}^*) \overset{\text{lower bound}}{\geq} g(\boldsymbol{x}^*) \overset{\text{maximization}}{\geq} g(\boldsymbol{x}) \overset{\text{tightness}}{=} f(\boldsymbol{x})$$

That is, MM monotonically increases f over iterations.

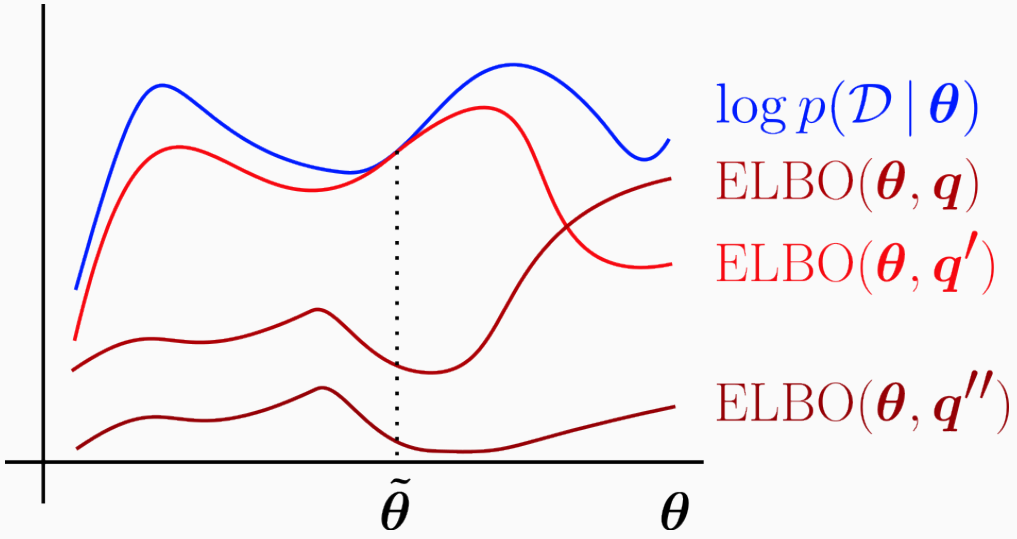

f
g_0
$\boldsymbol{x}_1$
$\boldsymbol{x}_0$



Note: In general, only **local** maximum can be found (MM per se does not even guarantee convergence to local maximum). But, monotonicity over iterations is guaranteed.

ELBO for Minorization-Maximization

Can we use the ELBO for a minorization-maximizaiton scheme?



We'd like to iterate:

- for current model parameters $\boldsymbol{\theta}_{\text{cur}}$ find $\boldsymbol{q}^*$ such that the ELBO becomes tight at $\boldsymbol{\theta}_{\text{cur}}$:

$$\log p(\mathcal{D} \mid \boldsymbol{\theta}_{\text{cur}}) = \mathrm{ELBO}(\boldsymbol{\theta}_{\text{cur}}, \boldsymbol{q}^*)$$

- $\boldsymbol{\theta}_{\text{new}} = \arg\max_{\boldsymbol{\theta}} \ \mathrm{ELBO}(\boldsymbol{\theta}, \boldsymbol{q}^*)$
- $\boldsymbol{\theta}_{\text{cur}} \leftarrow \boldsymbol{\theta}_{\text{new}}$

Finding Tight ELBO

Recall the ELBO for the i^{th} sample:

$$\begin{aligned}\text{ELBO}(\boldsymbol{\theta}, q_i) &= \mathbb{E}_{q_i}\left[\log \frac{p(\boldsymbol{x}^{(i)}, \boldsymbol{Z}^{(i)} \mid \boldsymbol{\theta})}{q_i(\boldsymbol{Z}^{(i)})}\right] \\ &= \mathbb{E}_{q_i}\left[\log p(\boldsymbol{x}^{(i)}, \boldsymbol{Z}^{(i)} \mid \boldsymbol{\theta}) - \log q_i(\boldsymbol{Z}^{(i)})\right]\end{aligned}$$

Using the chain rule, we can write

$$\begin{aligned}p(\boldsymbol{x}^{(i)}, \boldsymbol{Z}^{(i)} \mid \boldsymbol{\theta}) &= p(\boldsymbol{Z}^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta})\, p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}) \\ \log p(\boldsymbol{x}^{(i)}, \boldsymbol{Z}^{(i)} \mid \boldsymbol{\theta}) &= \log p(\boldsymbol{Z}^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}) + \log p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta})\end{aligned}$$

Hence

$$\begin{aligned}\text{ELBO}(\boldsymbol{\theta}, q_i) &= \mathbb{E}_{q_i}\left[\log p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{Z}^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}) - \log q_i(\boldsymbol{Z}^{(i)})\right] \\ &= \log p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}) + \mathbb{E}_{q_i}\left[\log p(\boldsymbol{Z}^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}) - \log q_i(\boldsymbol{Z}^{(i)})\right]\end{aligned}$$

Finding Tight ELBO

$$\text{ELBO}(\boldsymbol{\theta}, q_i) = \log p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}) + \mathbb{E}_{q_i}\left[\log p(\boldsymbol{Z}^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}) - \log q_i(\boldsymbol{Z}^{(i)})\right]$$

The second term is

$$\begin{aligned}\mathbb{E}_{q_i}\left[\log p(\boldsymbol{Z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}) - \log q_i(\boldsymbol{Z}^{(i)})\right] &= -\mathbb{E}_{q_i}\left[\log q_i(\boldsymbol{Z}^{(i)}) - \log p(\boldsymbol{Z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})\right] \\ &= -\mathbb{E}_{q_i}\left[\log \frac{q_i(\boldsymbol{Z}^{(i)})}{p(\boldsymbol{Z}^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta})}\right] \\ &= -\text{KL}(q_i||p(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}))\end{aligned}$$

that is, the **negative Kullback-Leiber divergence** between the variational distribution q_i and the posterior of $\boldsymbol{Z}^{(i)}$ given $\boldsymbol{x}^{(i)}$ under model $\boldsymbol{\theta}$!

Finding Tight ELBO

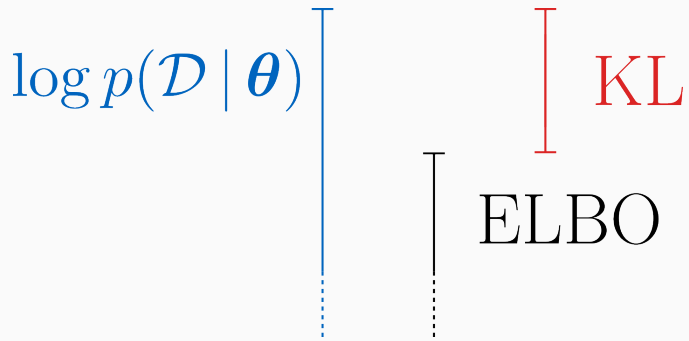
$$\log p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}) = \text{ELBO}(\boldsymbol{\theta}, q_i) + \text{KL}(q_i || p(\boldsymbol{z}^{(i)} | \boldsymbol{x}^{(i)}, \boldsymbol{\theta}))$$

$$\sum_i \log p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}) = \sum_i \text{ELBO}(\boldsymbol{\theta}, q_i) + \sum_i \text{KL}(q_i || p(\boldsymbol{z}^{(i)} | \boldsymbol{x}^{(i)}, \boldsymbol{\theta}))$$

$$\log p(\mathcal{D} \mid \boldsymbol{\theta}) = \text{ELBO}(\boldsymbol{\theta}, \boldsymbol{q}) + \underbrace{\sum_i \text{KL}(q_i || p(\boldsymbol{z}^{(i)} | \boldsymbol{x}^{(i)}, \boldsymbol{\theta}))}_{\textbf{(total) variational gap}}$$

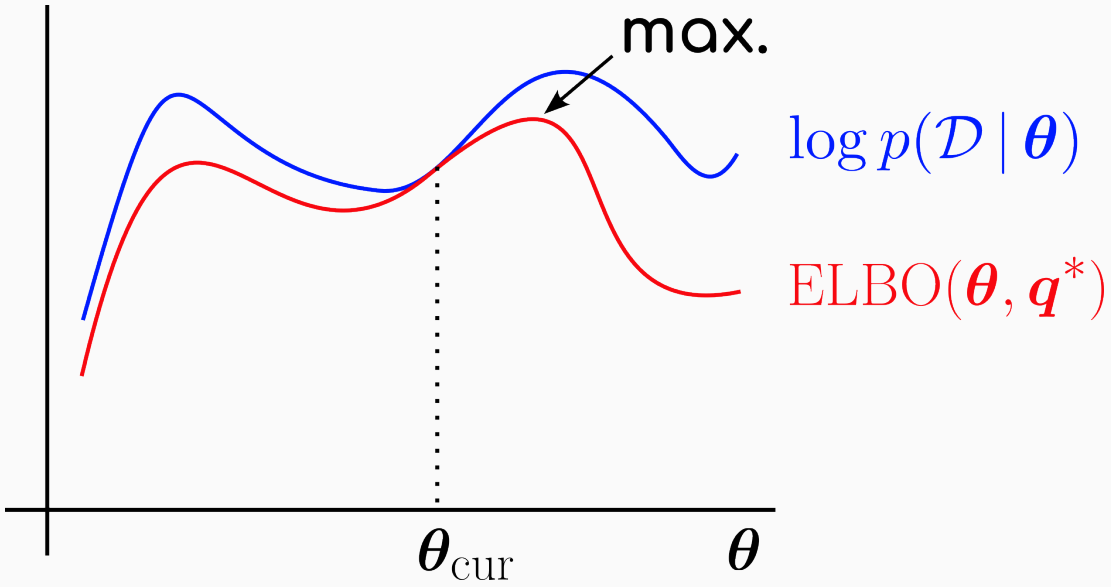
When for each sample q_i is the exact posterior, then KL $= 0$ and ELBO is tight!

The KL between q_i and $p(\boldsymbol{z}^{(i)} | \boldsymbol{x}^{(i)}, \boldsymbol{\theta})$ is called **variational gap** for i^{th} sample. Their sum is the **(total) variational gap**.



ELBO for Minorization-Maximization

- let a parameter $\boldsymbol{\theta}_{\text{cur}}$ of the joint model $p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})$ be given
- set $q_i^* \equiv p(\boldsymbol{z}^{(i)} | \boldsymbol{x}^{(i)}, \boldsymbol{\theta}_{\text{cur}})$ for all i and $\boldsymbol{q}^* = \{q_i^*\}_{i=1}^N$
- we have established that $\text{ELBO}(\boldsymbol{\theta}, \boldsymbol{q}^*)$ is a tight lower bound of $\log p(\mathcal{D} \mid \boldsymbol{\theta})$ at $\boldsymbol{\theta}_{\text{cur}}$
- hence, maximizing $\text{ELBO}(\boldsymbol{\theta}, \boldsymbol{q}^*)$ with respect to $\boldsymbol{\theta}$ will increase $\log p(\mathcal{D} \mid \boldsymbol{\theta})$ (minorization-maximization)



ELBO for Minorization-Maximization cont'd

Goal: $\max_{\boldsymbol{\theta}} \text{ ELBO}(\boldsymbol{\theta}, \boldsymbol{q}^*)$

Recall that

$$\text{ELBO}(\boldsymbol{\theta}, \boldsymbol{q}^*) = \sum_{i=1}^{N} \text{ELBO}(\boldsymbol{\theta}, q_i^*)$$

The sample-wise ELBO's can be written as

$$\begin{aligned}
\text{ELBO}(\boldsymbol{\theta}, q_i^*) &= \mathbb{E}_{q_i^*}\left[\log \frac{p(\boldsymbol{x}^{(i)}, \boldsymbol{Z}^{(i)} \mid \boldsymbol{\theta})}{q_i^*(\boldsymbol{Z}^{(i)})}\right] \\
&= \mathbb{E}_{q_i^*}\left[\log p(\boldsymbol{x}^{(i)}, \boldsymbol{Z}^{(i)} \mid \boldsymbol{\theta})\right] - \mathbb{E}_{q_i^*}\left[q_i^*(\boldsymbol{Z}^{(i)})\right] \\
&= \mathbb{E}_{q_i^*}\left[\log p(\boldsymbol{x}^{(i)}, \boldsymbol{Z}^{(i)} \mid \boldsymbol{\theta})\right] + \underbrace{H[q_i^*]}_{\text{const.}}
\end{aligned}$$

Recall that $H[q_i^*]$ is the entropy of q_i^*, which does **not** depend on $\boldsymbol{\theta}$ and can be ignored for maximization (it **does** depend on $\boldsymbol{\theta}_{\text{cur}}$, which is not optimized though).

ELBO for Minorization-Maximization cont'd

Goal: $\max_{\boldsymbol{\theta}}\ \text{ELBO}(\boldsymbol{\theta}, \boldsymbol{q}^*)$

Hence, dropping the entropy, maximizing $\text{ELBO}(\boldsymbol{\theta}, \boldsymbol{q}^*)$ is equivalent to

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \mathbb{E}_{q_i^*} \left[\log p(\boldsymbol{x}^{(i)}, \boldsymbol{Z}^{(i)} \mid \boldsymbol{\theta}) \right]$$

But, we have set $q_i^* \equiv p(\boldsymbol{z}^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}_{\text{cur}})$ and hence we get

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \mathbb{E}_{p(\boldsymbol{z}^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}_{\text{cur}})} \left[\log p(\boldsymbol{x}^{(i)}, \boldsymbol{Z}^{(i)} \mid \boldsymbol{\theta}) \right]$$

Compare this with slide 5.

ELBO for Minorization-Maximization cont'd

Goal: $\max_{\boldsymbol{\theta}} \text{ ELBO}(\boldsymbol{\theta}, \boldsymbol{q}^*)$

Hence, dropping the entropy, maximizing $\text{ELBO}(\boldsymbol{\theta}, \boldsymbol{q}^*)$ is equivalent to

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \mathbb{E}_{q_i^*} \left[\log p(\boldsymbol{x}^{(i)}, \boldsymbol{Z}^{(i)} \mid \boldsymbol{\theta}) \right]$$

But, we have set $q_i^* \equiv p(\boldsymbol{z}^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}_{\text{cur}})$ and hence we get

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \mathbb{E}_{p(\boldsymbol{z}^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}_{\text{cur}})} \left[\log p(\boldsymbol{x}^{(i)}, \boldsymbol{Z}^{(i)} \mid \boldsymbol{\theta}) \right]$$

Compare this with slide 5. This is exactly EM!

- ELBO is just a lower bound of the log-likelihood, following form Jensen's inequality
- it depends on the variational distributions q_i, hence it is actually a family of lower bounds
- generally, when dealing with maximization problems, maximizing lower bounds is meaningful
- when q_i is the posterior of $\boldsymbol{Z}^{(i)}$ given $\boldsymbol{x}^{(i)}$ for a given model $\boldsymbol{\theta}_{\text{cur}}$, the ELBO is tight at $\boldsymbol{\theta}_{\text{cur}}$
- tight ELBO leads to a minorization-maximization approach, also known as EM

Variational EM

ELBO and Posterior Inference

Recap relation between log-likelihood, ELBO and variational gap:

$$\underbrace{\log p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta})}_{\text{const. w.r.t. } q_i} = \text{ELBO}(\boldsymbol{\theta}, q_i) + \text{KL}(q_i || p(\boldsymbol{z}^{(i)} | \boldsymbol{x}^{(i)}, \boldsymbol{\theta}))$$

- ELBO is a function of both $\boldsymbol{\theta}$ and q_i
- when fixing $\boldsymbol{\theta}$, then $\log p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta})$ is constant
- hence, increasing $\text{ELBO}(\boldsymbol{\theta}, q_i)$ with respect to q_i will decrease $\text{KL}(q_i || p(\boldsymbol{z}^{(i)} | \boldsymbol{x}^{(i)}, \boldsymbol{\theta}))$
- **maximizing ELBO means performing posterior inference!**

ELBO and Posterior Inference cont'd

- global maximum of ELBO at $q_i \equiv p(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})$
- for GMMs, this is closed form, but for many other models computing the posterior is **hard**
- **Idea:** introduce **parametric variational distributions** $q_{\phi_i}(\boldsymbol{z}^{(i)})$, where ϕ_i are called **variational parameters**
- **maximizing the ELBO**:

$$\max_{\phi_i} \text{ ELBO}(\boldsymbol{\theta}, \phi_i) =: \text{ELBO}(\boldsymbol{\theta}, q_i)$$

  is equivalent to minimizing the KL to the posterior:

$$\min_{\phi_i} \text{ KL}(q_{\phi_i}||p(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}))$$

- when the variational distributions do not contain the true posterior, the KL cannot become 0 $\Rightarrow$ **approximate posterior**

Variational EM

- if the variational distribution is restricted, the ELBO might not be tight—nevertheless, the ELBO is always a lower bound of the log-likelihood, so maximizing it is meaningful
- we might optimizing the ELBO both with respect to $\boldsymbol{\theta}$ **and** $\boldsymbol{\phi}$:

$$\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \text{ELBO}(\boldsymbol{\theta}, \boldsymbol{\phi})$$

  where $\boldsymbol{\phi} = \{\boldsymbol{\phi}_i\}_{i=1}^N$ is the set of all variational parameters
  - maximize w.r.t. $\boldsymbol{\phi}$: improve lower bound
  - maximize w.r.t. $\boldsymbol{\theta}$: learn joint model with ELBO as a surrogate of the log-likelihood
- this scheme is called **variational EM**
- EM is just a special case where the ELBO is made tight

Let us demonstrate variational EM on GMMs. This is not very meaningful, since exact EM can be done here, but we aim to illustrate the concept.

- variational distributions are naturally Categoricals:

$$q(z^{(i)} = k \mid \boldsymbol{\phi}_i) = \phi_{i,k}, \quad \text{where } \phi_{i,k} \geq 0, \ \sum_{k=1}^{K} \phi_{i,k} = 1$$

- the set of all variational distributions might be interpreted as a non-negative $N \times K$ matrix whose rows sum to one:

$$\boldsymbol{\phi} = \begin{pmatrix} \phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,K} \\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{N,1} & \phi_{N,2} & \cdots & \phi_{N,K} \end{pmatrix}$$

We reparametrize ϕ by an unconstrained matrix

$$\boldsymbol{a} = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,K} \\ a_{2,1} & a_{2,2} & \dots & a_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \dots & a_{N,K} \end{pmatrix}$$

and define

$$\phi = \text{softmax}(\boldsymbol{a})$$

where the softmax acts row-wise. Specifically:

$$\phi_{i,j} = \frac{exp(a_{i,j})}{\sum_{j=1}^{K} \exp(a_{i,j})}$$

The exact **E-step**

$$\gamma_{i,k} = p(z^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta})$$

is replaced by a few gradient ascent steps:

$$\boldsymbol{a} \leftarrow \boldsymbol{a} + \eta \nabla_{\boldsymbol{a}} \text{ELBO}(\boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{a}))$$

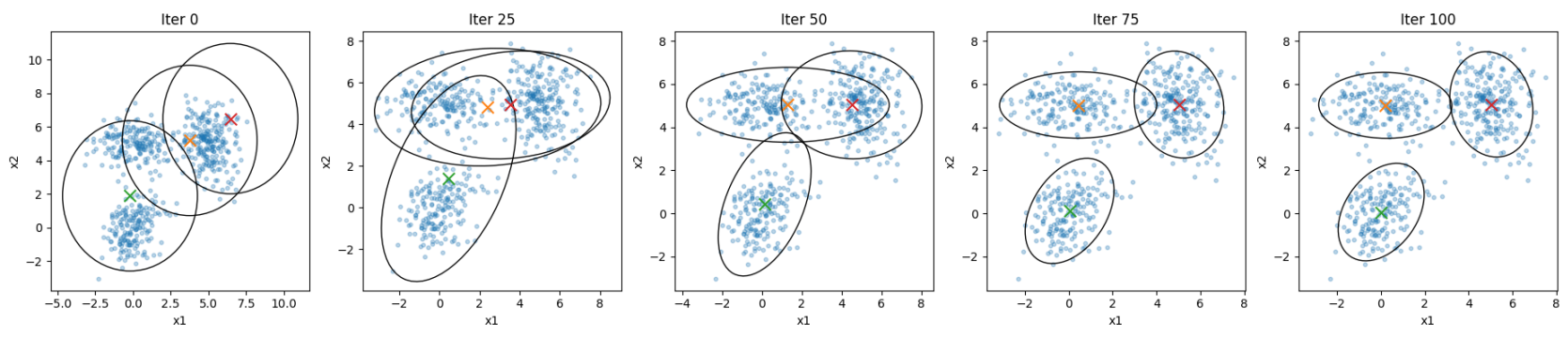
This is fairly easy to implement, either by hand or using autodiff.

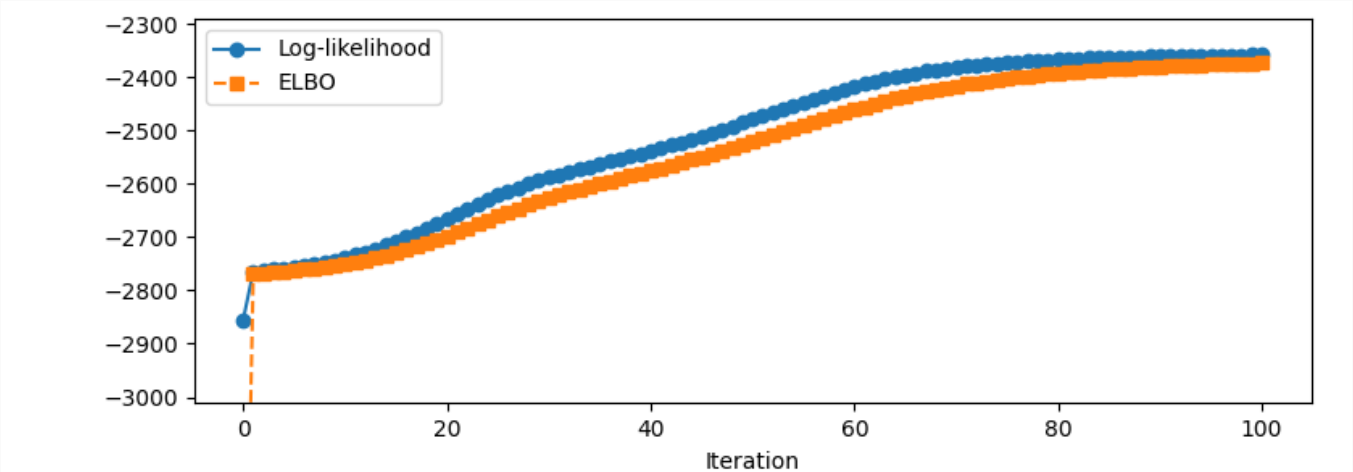
The **M-step** is exactly as before, but using $\phi_{i,j}$ instead of $\gamma_{i,j}$

- $\boldsymbol{\mu}_k \leftarrow \frac{\sum_{i=1}^N \phi_{i,k} \boldsymbol{x}^{(i)}}{\sum_{i=1}^N \phi_{i,k}}$
- $\boldsymbol{\Sigma}_k \leftarrow \frac{\sum_{i=1}^N \phi_{i,k} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_k)(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^N \phi_{i,k}}$
- $w_k \leftarrow \frac{\sum_{i=1}^N \phi_{i,k}}{N}$

Variational EM in GMMs cont'd

Example





Here 5 gradient ascend steps are done in each EM iteration. Convergence is slower than for exact EM but achieves a similar solution (compare with last lecture).