The **EM algorithm** is a method for fitting latent variable models. Let:

- $X = \{x_1, \ldots, x_N\}$ be the collection of observed data points
- θ be a collection of model parameters
- $T = \{t_1, \ldots t_N\}$ be the collection of latent variables associated with each data point

Variational Lower Bound on Marginal Likelihood

In fitting our model we will attempt to find the setting of the parameters θ that maximizes the **marginal likelihood** of the data:

$$\begin{aligned} P(X|\theta) &= \prod_{i=1}^{N} P(x_i|\theta) && \text{Assume data are iid} \\ &= \prod_{i=1}^{N} \sum_{c=1}^{T} P(x_i, t_i = c|\theta) \end{aligned}$$

As always, it is generally easier to maximize the **marginal log likelihood** rather than the likelihood directly:

$$\begin{aligned} \log P(X|\theta) &= \log \prod_{i=1}^{N} P(x_i|\theta) && \text{Assume data are iid.} \\ &= \sum_{i=1}^{N} \log P(x_i|\theta) \\ &= \sum_{i=1}^{N} \log \left[\sum_{c=1}^{T} P(x_i, t_i = c|\theta) \right] \end{aligned}$$

The problem is that this expression is still difficult to optimize directly (e.g., via SGD). In EM, we opt to instead try to maximize a **lower bound**, $\mathcal{L}$ on the marginal log likelihood instead:

$$\overbrace{\log P(X|\theta)}^{\text{Marginal log likelihood}} \quad \geq \quad \underbrace{\mathcal{L}}_{\text{Variational lower bound}}$$

The issue is that there is no reason to expect a single lower bound to be useful for finding a local maxima of the marginal log likelihood. What we really want is a *family* of lower bounds, which we can tune to get better and better local approximations to the marginal log likelihood at θ. To achieve this, we introduce a new parameter to the lower bound, a distribution over the latent variable classes:

$$q(t_i = c)$$

This distribution will be used as a flexible parameter of our family of lower bounds, $\mathcal{L}$, allowing us modify the form of the lower bound over the course of optimization.

We derive the form for the family of lower bounds using **Jensen's inequality**:

$$\log(\mathbb{E}[X]) \geq \mathbb{E}[\log X] \tag{1}$$

or, if we assume X is a discrete random variable:

$$\log \sum_i \alpha_i x_i \geq \sum_i \alpha_i \log x_i \tag{2}$$

where $\alpha_i \geq 0 \; \forall i$ and $\sum_i \alpha_i = 1$. Using this inequality, we can derive a lower bound on the marginal log likelihood:

$$\begin{aligned}
\log P(X|\theta) &= \sum_{i=1}^{N} \log \left[\sum_{c=1}^{T} P(x_i, t_i = c|\theta) \right] \\
&= \sum_{i=1}^{N} \log \left[\sum_{c=1}^{T} \underbrace{\frac{q(t_i = c)}{q(t_i = c)}}_{\text{this is just 1}} \times P(x_i, t_i = c|\theta) \right]
\end{aligned}$$

At this point, notice that we can rewrite the last line as

$$\log P(X|\theta) = \sum_{i=1}^{N} \log \mathbb{E}_q \left[\frac{P(x_i, T|\theta)}{q(T)} \right]$$

This allows us to apply Jensen's inequality (Eq. 1), to define a family of lower bounds:

$$\begin{gathered}
\log P(X|\theta) \geq \mathcal{L}(\theta, q) \\
\sum_{i=1}^{N} \log \mathbb{E}_q \left[\frac{P(x_i, T|\theta)}{q(T)} \right] \geq \sum_{i=1}^{N} \mathbb{E}_q \left[\log \frac{P(x_i, T|\theta)}{q(T)} \right] \\
\sum_{i=1}^{N} \log \left[\sum_{c=1}^{T} \frac{q(t_i = c)}{q(t_i = c)} \times P(x_i, t_i = c|\theta) \right] \geq \underbrace{\sum_{i=1}^{N} \sum_{c=1}^{T} q(t_i = c) \log \frac{P(x_i, t_i = c|\theta)}{q(t_i = c)}}_{\mathcal{L}(\theta, q)}
\end{gathered}$$

Summary *Variational Lower Bound*
We have now derived a *family* of lower bounds on the marginal log likelihood, $\log P(X|\theta)$. The functions in this family are of the form

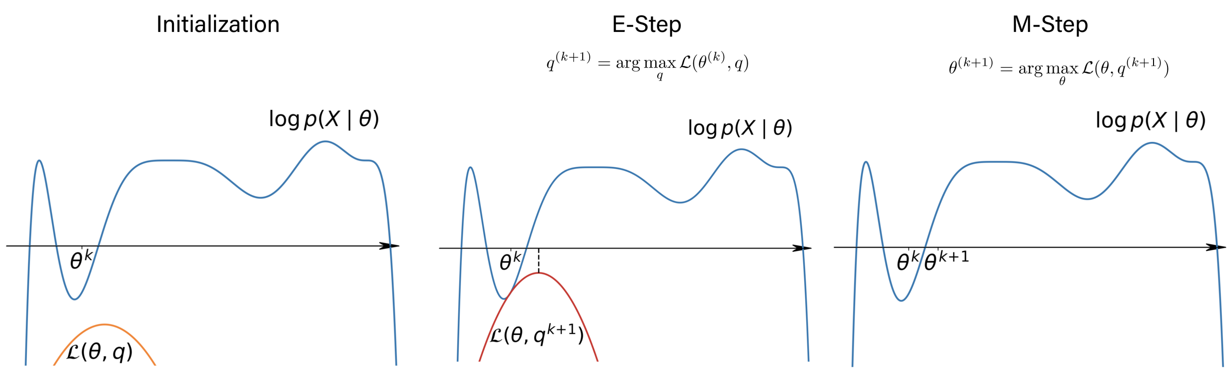
$$\begin{aligned}\mathcal{L}(\theta, q) &= \sum_{i=1}^{N} \sum_{c=1}^{T} q(t_i = c) \log \left[\frac{P(x_i, t_i = c|\theta)}{q(t_i = c)} \right] \\ &= \sum_{i=1}^{N} \mathbb{E}_q \left[\log \frac{P(x_i, T|\theta)}{q(T)} \right]\end{aligned}$$

During optimization, we can modify the distribution q to achieve lower bounds that provide better local approximations to $\log P(X|\theta)$ at θ.

EM Algorithm Overview

The EM algorithm is an iterative approach to coordinate ascent on the marginal likelihood, $P(X|\theta)$. It consists of two steps, which are repeated until convergence:

1. **E-Step**: Given a starting value for θ, find the distribution q^* that maximizes $\mathcal{L}(\theta, q)$.
2. **M-Step**: Given the distribution q identified during the E-step, find the value of θ^* that maximizes $\mathcal{L}(\theta, q^*)$.



E-Step Details

During the E-step, we fix the current value for the parameters, θ, and try to maximize the variational lower bound, $\mathcal{L}$, with respect to the distribution q:

$$q^{(k+1)} = \arg\max_q \mathcal{L}(\theta^{(k)}, q)$$

This maximization problem is equivalent to *minimizing* the gap between $\log P(X|\theta^{(k)})$ and $\mathcal{L}(\theta^{(k)}, q)$:

$$q^{(k+1)} = \arg\min_q \log P(X|\theta^{(k)}) - \mathcal{L}(\theta^{(k)}, q)$$

We can rewrite the gap between $\log P(X|\theta^{(k)})$ and $\mathcal{L}(\theta^{(k)}, q)$ as:

$$\begin{aligned}
&\log P(X|\theta) - \mathcal{L}(\theta, q) \\
&= \sum_{i=1}^{N} \log P(x_i|\theta) - \sum_{i=1}^{N}\sum_{c=1}^{T} q(t_i = c) \log\left[\frac{P(x_i, t_i = c|\theta)}{q(t_i = c)}\right] \\
&= \sum_{i=1}^{N} \left(\log P(x_i|\theta) \underbrace{\sum_{c=1}^{T} q(t_i = c)}_{\text{this is just 1}} - \sum_{c=1}^{T} q(t_i = c) \log\left[\frac{P(x_i, t_i = c|\theta)}{q(t_i = c)}\right] \right) \\
&= \sum_{i=1}^{N}\sum_{c=1}^{T} q(t_i = c) \left(\log P(x_i|\theta) - \log \frac{P(x_i, t_i = c|\theta)}{q(t_i = c)} \right) \\
&= \sum_{i=1}^{N}\sum_{c=1}^{T} q(t_i = c) \log\left[\frac{P(x_i|\theta) q(t_i = c)}{P(x_i, t_i = c|\theta)}\right] \\
&= \sum_{i=1}^{N}\sum_{c=1}^{T} q(t_i = c) \log\left[\frac{P(x_i|\theta) q(t_i = c)}{P(t_i = c|x_i, \theta) P(x_i|\theta)}\right] \\
&= \sum_{i=1}^{N}\sum_{c=1}^{T} q(t_i = c) \log\left(\frac{q(t_i = c)}{P(t_i = c|x_i, \theta)}\right) \\
&= \sum_{i=1}^{N} \mathbb{KL}(q(t_i) \;||\; P(t_i|x_i, \theta))
\end{aligned}$$

Thus we have that during the E-step,

$$\begin{aligned}
q^{(k+1)} &= \arg\min_q \log P(X|\theta^{(k)}) - \mathcal{L}(\theta^{(k)}, q) \\
&= \arg\min_q \sum_{i=1}^{N} \mathbb{KL}(q(t_i) \;||\; P(t_i|x_i, \theta))
\end{aligned}$$

Because the smallest KL-divergence is achieved when $q(t_i) = P(t_i|x_i, \theta)$, we conclude that the update for the **E-step** is simply:

$$q^{(k+1)} = \underbrace{P(T|X, \theta^{(k)})}_{\text{Posterior over latent classes}} \tag{3}$$

Summary *E-Step*
During the **E-step**, we fix the current value for the parameters, θ, and try to maximize the variational lower bound, $\mathcal{L}$, with respect to the distribution q.

Above, we demonstrate that maximizing $\mathcal{L}$ wrt q is equivalent to minimizing the gap between $\log P(X|\theta)$ and $\mathcal{L}$, which is in turn equivalent to minimizing the sum of the KL divergences between $q(t_i)$ and $P(t_i|x_i, \theta)$.

This observation implies that the update during the **E-step** should simply be:

$$q^{(k+1)} = P(T|X, \theta^{(k)})$$

The caveat is that often it is intractable to compute the posterior over latent classes, $P(T|X, \theta)$, exactly. In these cases, it is necessary to use a variational approximation to $P(T|X, \theta)$ (e.g., a mean field approximation), and minimize the KL divergence between q and the variational approximation. This approach is known as **variational EM**.

M-Step Details

During the M-step we fix q to the value we computed during the E-step and try to find $\theta^{(k+1)}$ that maximizes $\mathcal{L}$:

$$\theta^{(k+1)} = \arg\max_{\theta} \mathcal{L}(\theta, q^{(k+1)})$$

Here, we decompose the variational lower bound into terms that depend on θ:

$$\begin{aligned}
\mathcal{L}(q, \theta) &= \sum_{i=1}^{N} \sum_{c=1}^{T} q(t_i = c) \log \left[\frac{P(x_i, t_i = c|\theta)}{q(t_i = c)} \right] \\
&= \sum_{i=1}^{N} \sum_{c=1}^{T} q(t_i = c) \log P(x_i, t_i = c|\theta) - \underbrace{\sum_{i=1}^{N} \sum_{c=1}^{T} q(t_i = c) \log q(t_i = c)}_{\text{does not depend on } \theta} \\
&\propto \sum_{i=1}^{N} \sum_{c=1}^{T} q(t_i = c) \log P(x_i, t_i = c|\theta) \\
&\propto \mathbb{E}_q[\log P(X, T|\theta)]
\end{aligned}$$

This expectation, $\mathbb{E}_q[\log P(X, T|\theta)]$ is usually concave and tends to be relatively easy to maximize with respect to θ for most models.

Summary *M-Step*
During the **M-step** we fix q to the value we computed during the previous E-step and try to find $\theta^{(k+1)}$ that maximizes $\mathcal{L}$:

$$\theta^{(k+1)} = \arg\max_{\theta} \mathcal{L}(\theta, q^{(k+1)})$$

In the derivation above, we showed that this is equivalent to finding θ that maximizes the following expected value:

$$\theta^{(k+1)} = \arg\max_{\theta} \mathbb{E}_q[\log P(X, T|\theta)] \tag{4}$$

This is the M-step update, typically achieved by taking the partial derivative of the above expectation wrt each parameter, setting it to 0, and solving.