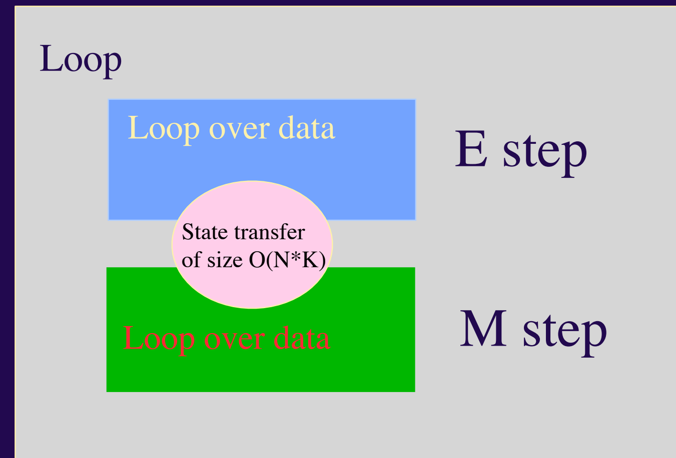
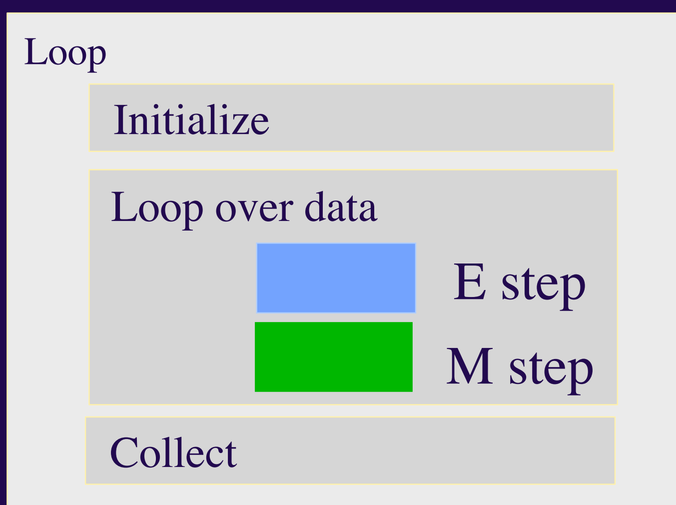
EM in practice (continued)

- Memory problems ---> we can compute means, etc., as running totals so that we do not need to store responsibilities for all points over all clusters.

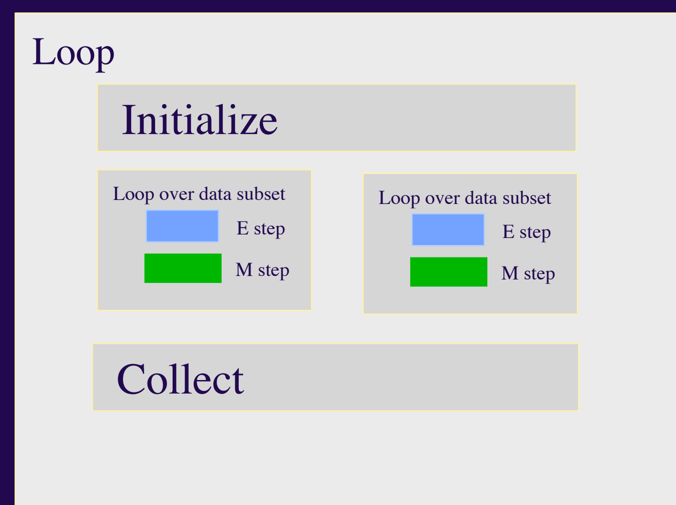
EM (Straight Forward Implementation)



EM (scalable)



EM (parallelized)



Analysis of EM

- Maximizing the Q function provided a new parameter estimate which increased the likelihood
- Showing this typically uses Jensen's inequality
  - Bishop (§9.4), instead, uses the fact that the KL divergence between two distributions is non-negative, but showing this also uses Jensen's.
- Given a bounded likelihood, this means the algorithm converges to a stationary point
  - Typically a local maximum but examples where it is a saddle point can be constructed.

Analysis of EM

- We will sketch the summary provided in the online resource "The Expectation Maximization Algorithm: A short tutorial" by Sean Borman
- This follows "The EM Algorithm and Extensions" by Geoffrey McLachlan and Thriyambakam Krishnan.
- See also Bishop (§9.4)

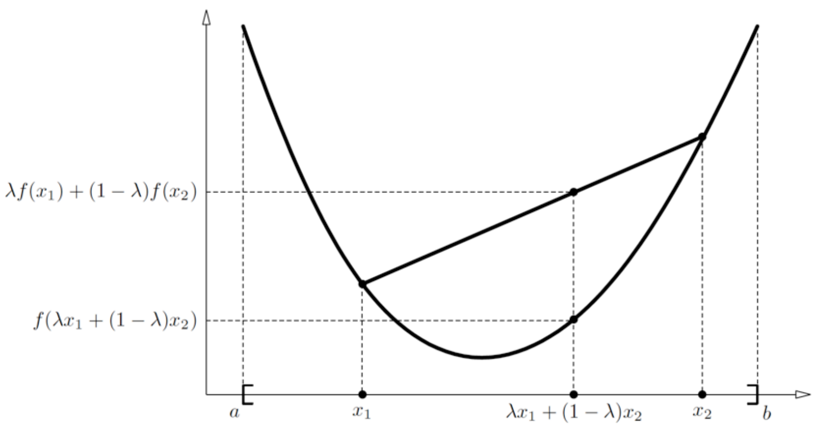


Figure 1: f is *convex* on $[a, b]$ if $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ $\forall x_1, x_2 \in [a, b]$, $\lambda \in [0, 1]$.

From "The Expectation Maximization Algorithm: A short tutorial" by Sean Borman

More generally, if f is convex, then, for

$\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$

we have

$$f\left(\sum_i \lambda_i x_i\right) \leq \sum_i \lambda_i f(x_i)$$

(Jensen's inequality)

Result from calculas (prove via mean value theorem)

If f is twice differentiable on $[a,b]$ and $f'' \geq 0$ on $[a,b]$, then $f(x)$ is convex on $[a,b]$.

Notice that $f(x) = -\log(x)$ is convex

Proof?

$$f'(x) = -\frac{1}{x}$$

$$f''(x) = \frac{1}{x^2}$$

$$f\left(\sum_i \lambda_i x_i\right) \leq \sum_i \lambda_i f(x_i) \quad \text{(Jensen's inequality)}$$

$$\log\left(\sum_i \lambda_i x_i\right) \geq \sum_i \lambda_i \log(x_i) \quad (-\log(x) \text{ is convex})$$

Working with bounds allows us to do something with the nasty sum in the log().

In EM, we seek θ to maximize $L(\theta) = \ln P(\mathbf{X}|\theta)$

Suppose at step n we have $L(\theta_n)$

$$L(\theta) - L(\theta_n) = \ln\left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)\right) - \ln \mathcal{P}(\mathbf{X}|\theta_n).$$

From "The Expectation Maximization Algorithm: A short tutorial" by Sean Borman

$$
\begin{aligned}
L(\theta) - L(\theta_n) &= \ln\left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)\right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \\
&= \ln\left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)\cdot\frac{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)}\right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \\
&= \ln\left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)}\right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \\
&\geq \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\ln\left(\frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)}\right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \qquad \text{(Jensen's)} \\
&= \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\ln\left(\frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\mathcal{P}(\mathbf{X}|\theta_n)}\right) \\
&\triangleq \Delta(\theta|\theta_n).
\end{aligned}
$$

Notice that this is our Q function

$\ln P(\mathbf{X}|\theta_n) = \sum_z \ln P(\mathbf{X}|\theta_n) P(z|X,\theta_n)$

because $P(\mathbf{X}|\theta_n)$ does not depend on z,

and $\sum_z P(z|X,\theta_n) = 1$

From "The Expectation Maximization Algorithm: A short tutorial" by Sean Borman

$$L(\theta) \geq L(\theta_n) + \Delta(\theta|\theta_n)$$

$$l(\theta|\theta_n) \triangleq L(\theta_n) + \Delta(\theta|\theta_n)$$

$$L(\theta) \geq l(\theta|\theta_n).$$

(We will see that maximizing $l(\theta|\theta_n)$ will give the same θ_{n+1} as maximizing Q (proof in a few slides))

Note that $\Delta(\theta_n|\theta_n) = 0$

and that $l(\theta_n|\theta_n) = L(\theta_n)$

(proof on next slide)

From "The Expectation Maximization Algorithm: A short tutorial" by Sean Borman

Proof that $\Delta(\theta_n|\theta_n) = 0$ and thus $l(\theta_n|\theta_n) = L(\theta_n)$

$$
\begin{aligned}
l(\theta_n|\theta_n) &= L(\theta_n) + \Delta(\theta_n|\theta_n) \\
&= L(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\ln\frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta_n)\mathcal{P}(\mathbf{z}|\theta_n)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\mathcal{P}(\mathbf{X}|\theta_n)} \\
&= L(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\ln\frac{\mathcal{P}(\mathbf{X},\mathbf{z}|\theta_n)}{\mathcal{P}(\mathbf{X},\mathbf{z}|\theta_n)} \\
&= L(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\ln 1 \\
&= L(\theta_n),
\end{aligned}
$$

From "The Expectation Maximization Algorithm: A short tutorial" by Sean Borman

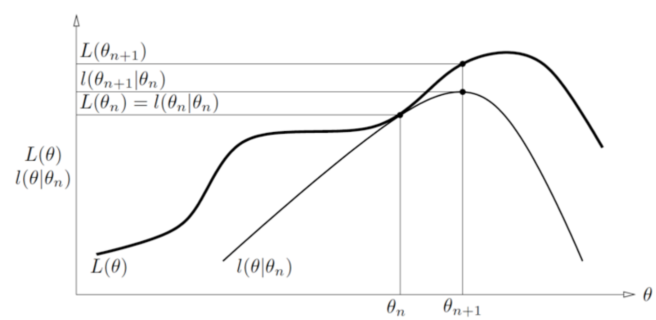


Figure 2: Graphical interpretation of a single iteration of the EM algorithm: The function $l(\theta|\theta_n)$ is bounded above by the likelihood function $L(\theta)$. The functions are equal at $\theta = \theta_n$. The EM algorithm chooses θ_{n+1} as the value of θ for which $l(\theta|\theta_n)$ is a maximum. Since $L(\theta) \geq l(\theta|\theta_n)$ increasing $l(\theta|\theta_n)$ ensures that the value of the likelihood function $L(\theta)$ is increased at each step.

From "The Expectation Maximization Algorithm: A short tutorial" by Sean Borman

Proof that maximizing $l(\theta|\theta_n)$ gives the same θ_{n+1} as maximizing Q

$$\theta_{n+1} = \arg\max_{\theta} \{ l(\theta|\theta_n) \}$$

$$= \arg\max_{\theta} \left\{ L(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{X}|\theta_n)\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)} \right\}$$

Now drop terms which are constant w.r.t. θ → details on next slide

$$= \arg\max_{\theta} \left\{ \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta) \right\}$$

(def'n conditional probability)

$$= \arg\max_{\theta} \left\{ \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \frac{\mathcal{P}(\mathbf{X},\mathbf{z},\theta)}{\cancel{\mathcal{P}(\mathbf{z},\theta)}} \frac{\cancel{\mathcal{P}(\mathbf{z},\theta)}}{\mathcal{P}(\theta)} \right\}$$

$$= \arg\max_{\theta} \left\{ \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \mathcal{P}(\mathbf{X},\mathbf{z}|\theta) \right\}$$

$$= \arg\max_{\theta} \left\{ \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta_n} \{ \ln \mathcal{P}(\mathbf{X},\mathbf{z}|\theta) \} \right\}$$

Q-function

From "The Expectation Maximization Algorithm: A short tutorial" by Sean Borman

Algebra for "drop terms which are constant w.r.t. θ"

$$\sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{X}|\theta_n)\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)}$$

$$= \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta) - \underbrace{\sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \mathcal{P}(\mathbf{X}|\theta_n)\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)}_{\text{These ones do not matter for } \arg\max_{\theta}()}$$

Summary

Jensen's provides a lower bound for the likelihood, L(θ) in terms of a current θ_n, namely $l(\theta_n|\theta_n)$.

The max of $l(\theta|\theta_n)$ is reached at the same θ_{n+1} found by maximizing the Q function

Since $l(\theta|\theta_n)$ is a lower bound for θ,

$$\mathrm{L}(\theta_{n+1}) \geq \underbrace{l(\theta_{n+1}|\theta_n) \geq l(\theta_n|\theta_n)}_{\text{because we maximized}} = \mathrm{L}(\theta_n)$$

In other words, L(θ) goes up (or stays the same) with the successive θ found by EM.