CS 480/680
Introduction to Machine Learning

Lecture 12
Expectation Maximization and Gaussian Mixture Models

Kathryn Simone
24 October 2024

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

We know how to estimate parameters and make predictions

Problem Type 2:

Given: $\{x_1 = 1, x_2 = 2, x_3 = 0\}, x_i \sim \mathcal{N}(\mu, \sigma^2 = 1.0)$

Task: Estimate μ

Problem Type 1:

Given: $\{x_1 = 1, x_2 = 2, x_3\}, x_i \sim \mathcal{N}(\mu = 1.0, \sigma^2 = 1.0)$

Task: Predict x_3

Can we estimate parameters if data is missing?

Problem Type 3:

Given: $\{x_1 = 1, x_2 = 2, x_3\}, x_i \sim \mathcal{N}(\mu, \sigma^2 = 1.0)$

Task: Estimate (x_3, μ)

How could we solve it?

μ:

x_3:

KEY IDEA BEHIND EM ALGORITHM

Lecture Outline

I. How does the EM algorithm work in a special case?

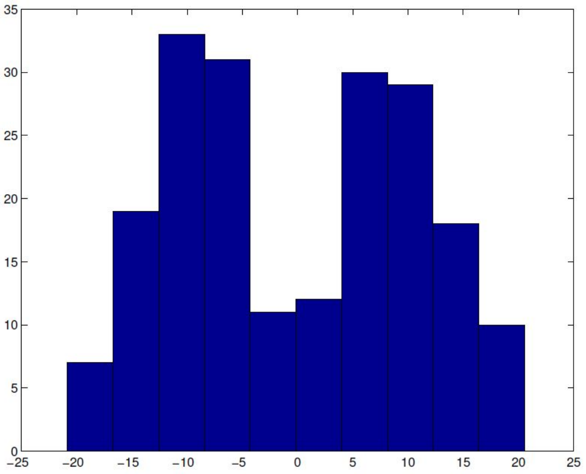
II. How does the EM algorithm work in general?

Lecture Outline

I. **How does the EM algorithm work in a special case?**

II. How does the EM algorithm work in general?

Estimating the parameters of a *mixture* of Gaussians



$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$
$$X = (1-\Delta)\cdot X_1 + \Delta \cdot X_2$$

Where Δ is a binary random variable:

$$\Delta \in \{0,1\}$$

Let π denote the probability of Δ taking on the value of 1:

$$\Pr[\Delta = 1] = \pi$$

Let $\mathcal{N}_{\mu,\sigma^2}$ denote the normal density with mean μ and variance σ^2. Then the density of x is

$$p(x) = (1-\pi)\mathcal{N}_{\mu_1,\sigma_1^2}(x) + \pi\mathcal{N}_{\mu_2,\sigma_2^2}(x)$$

Probabilistic Machine Learning, Section 8.7
Elements of Statistical Learning, Section 8.5

Can we find the parameters through direct maximization?

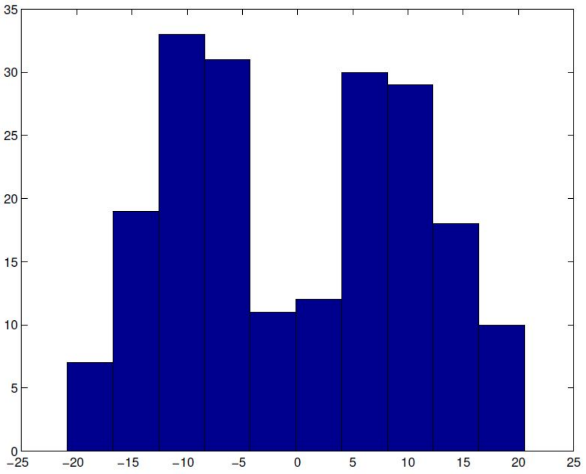
$$p(x) = (1-\pi)\mathcal{N}_{\mu_1,\sigma_1^2}(x) + \pi\mathcal{N}_{\mu_2,\sigma_2^2}(x)$$

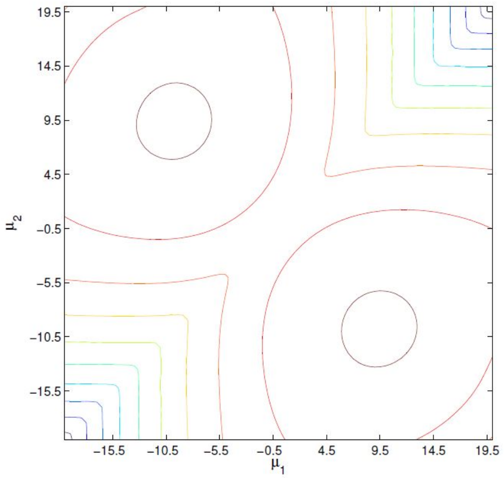
$$\mathcal{L}(\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2 \mid \boldsymbol{X}) = \prod_{i=1}^{n}\left[(1-\pi)\mathcal{N}_{\mu_1,\sigma_1^2}(x) + \pi\mathcal{N}_{\mu_2,\sigma_2^2}(x)\right]$$

$$\log\mathcal{L}(\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2 \mid \boldsymbol{X}) = \sum_{i=1}^{n}\log\left[(1-\pi)\mathcal{N}_{\mu_1,\sigma_1^2}(x) + \pi\mathcal{N}_{\mu_2,\sigma_2^2}(x)\right]$$

$$\frac{\partial \log\mathcal{L}}{\partial \sigma_2} = ?? \qquad \frac{\partial \log\mathcal{L}}{\partial \mu_2} = ?? \qquad \frac{\partial \log\mathcal{L}}{\partial \sigma_1} = ?? \qquad \frac{\partial \log\mathcal{L}}{\partial \mu_1} = ?? \qquad \frac{\partial \log\mathcal{L}}{\partial \pi} = ??$$

The likelihood function for a mixture model is nonconvex





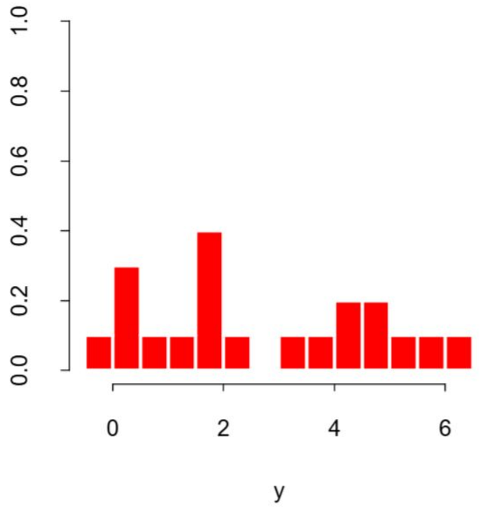
Label-switching problem:

- Parameters are unidentifiable because likelihood surface has two symmetric modes
- Even with mixing weight π, and variances σ_1^2, σ_2^2 known!

Probabilistic Machine Learning, Section 8.7 thought experiments ...

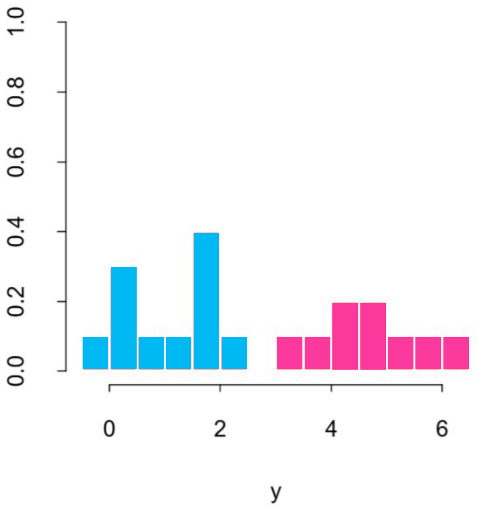
Thought experiment 1: If we knew the sample assignments...

$$
\begin{aligned}
&\log\mathcal{L}(\pi, \mu_1, \sigma^2, \mu_2, \sigma_2 \mid \boldsymbol{X}) \\
&\quad = \sum_{i=1}^{n} \log\left[(1-\pi)\mathcal{N}_{\mu_1,\sigma_1^2}(x) + \pi\mathcal{N}_{\mu_2,\sigma_2^2}(x)\right] \\
&\quad = \sum_{i=1}^{n} \left[(1-\Delta_i)\log\mathcal{N}_{\mu_1,\sigma_1^2}(x) + \Delta_i\log\mathcal{N}_{\mu_2,\sigma_2^2}(x)\right] \\
&\quad + \sum_{i=1}^{n} \left[(1-\Delta_i)\log(1-\pi) + \Delta_i\log\pi)\right] \\
&\quad = \begin{cases} \sum_{i=1}^{n} \log\mathcal{N}_{\mu_1,\sigma_1^2}(x) + \sum_{i=1}^{n} \log(1-\pi) & \text{if } \Delta = 0 \\ \sum_{i=1}^{n} \log\mathcal{N}_{\mu_2,\sigma_2^2}(x) + \sum_{i=1}^{n} \log\pi & \text{if } \Delta = 1 \end{cases}
\end{aligned}
$$



Thought experiment 1: If we knew the sample assignments...

$$= \begin{cases} \sum_{i=1}^{n} \log \mathcal{N}_{\mu_1,\sigma_1^2}(x) + \sum_{i=1}^{n} \log(1-\pi) & \text{if } \Delta = 0 \\ \sum_{i=1}^{n} \log \mathcal{N}_{\mu_2,\sigma_2^2}(x) + \sum_{i=1}^{n} \log \pi & \text{if } \Delta = 1 \end{cases}$$



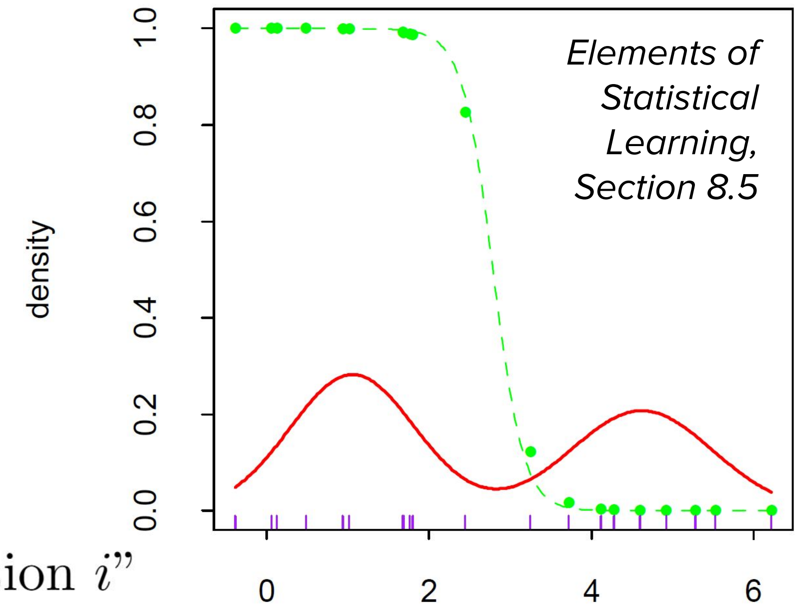
$$\hat{\mu}_1 = \frac{1}{|\Delta_0|} \sum_{i \in \Delta_0} x_i \qquad \hat{\mu}_2 = \frac{1}{|\Delta_1|} \sum_{i \in \Delta_1} x_i$$

$$\hat{\sigma}_1^2 = \frac{1}{|\Delta_0|} \sum_{i \in \Delta_0} (x_i - \hat{\mu}_1)^2 \qquad \hat{\sigma}_2^2 = \frac{1}{|\Delta_1|} \sum_{i \in \Delta_1} (x_i - \hat{\mu}_2)^2 \qquad \hat{\pi} = \frac{1}{N} \sum_{i=1}^{n} \Delta_i$$

... we could compute the parameters empirically

Thought experiment 2: If we knew the parameters...

$$p(x) = (1-\pi)\mathcal{N}_{\mu_1,\sigma_1^2}(x) + \pi\mathcal{N}_{\mu_2,\sigma_2^2}(x)$$

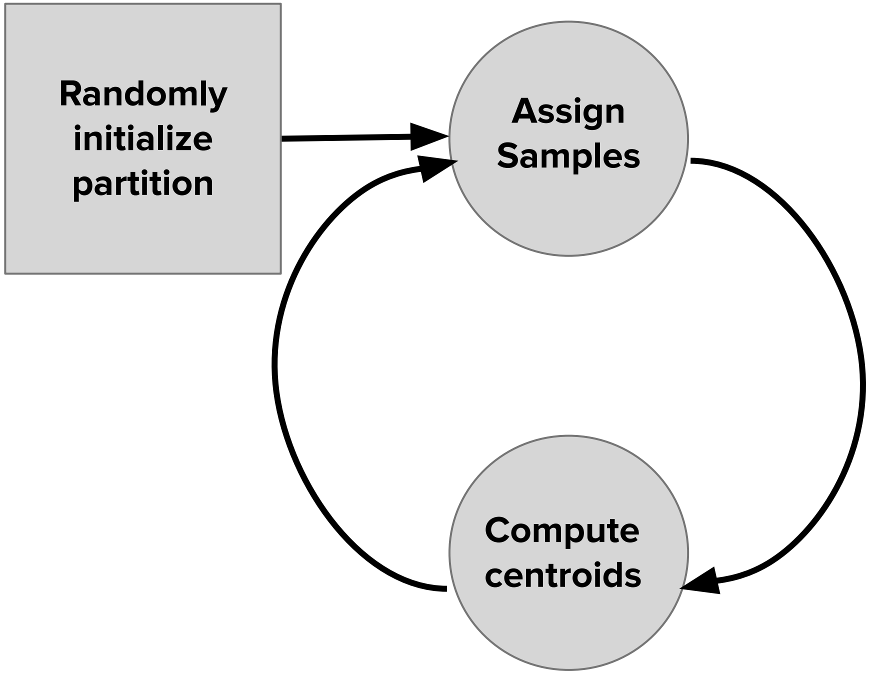


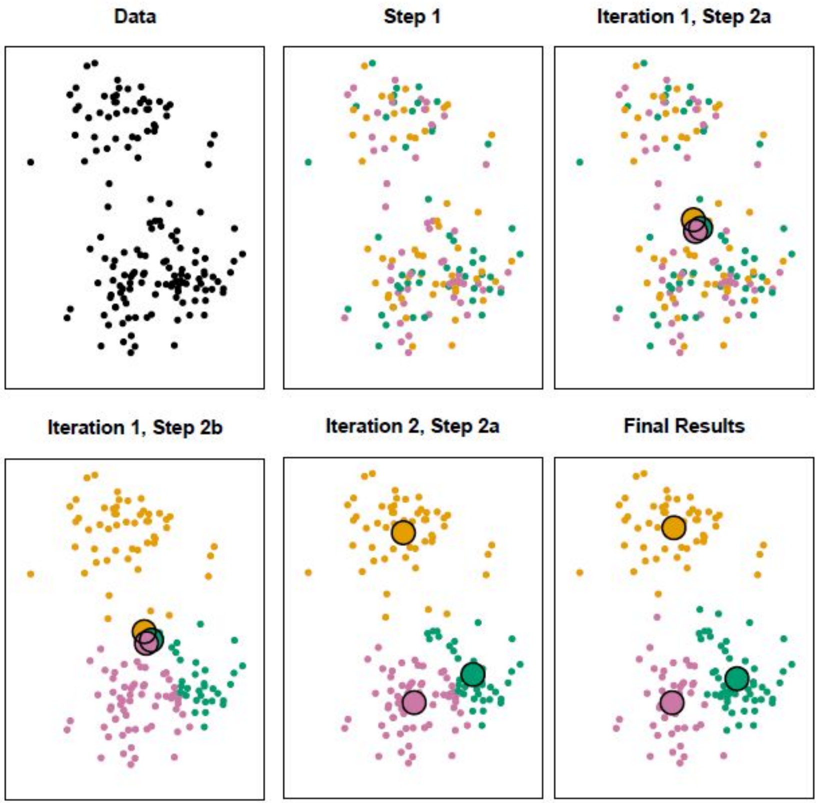
$$\Pr[\Delta_i = 1 \mid \pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \boldsymbol{X}]$$
$$= \frac{\pi\mathcal{N}_{\mu_2,\sigma_2^2}(x_i)}{(1-\pi)\mathcal{N}_{\mu_1,\sigma_1^2}(x_i) + \pi\mathcal{N}_{\mu_2,\sigma_2^2}(x_i)}$$
$$= \gamma_i : \text{ “responsibility of mode 2 for observation } i\text{”}$$
$$= \mathbb{E}[\Delta_i \mid \pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \boldsymbol{X}] :$$
$$\text{“expectation of } \Delta_i \text{ given parameters and data”}$$

...we could compute the probability of a sample assignment

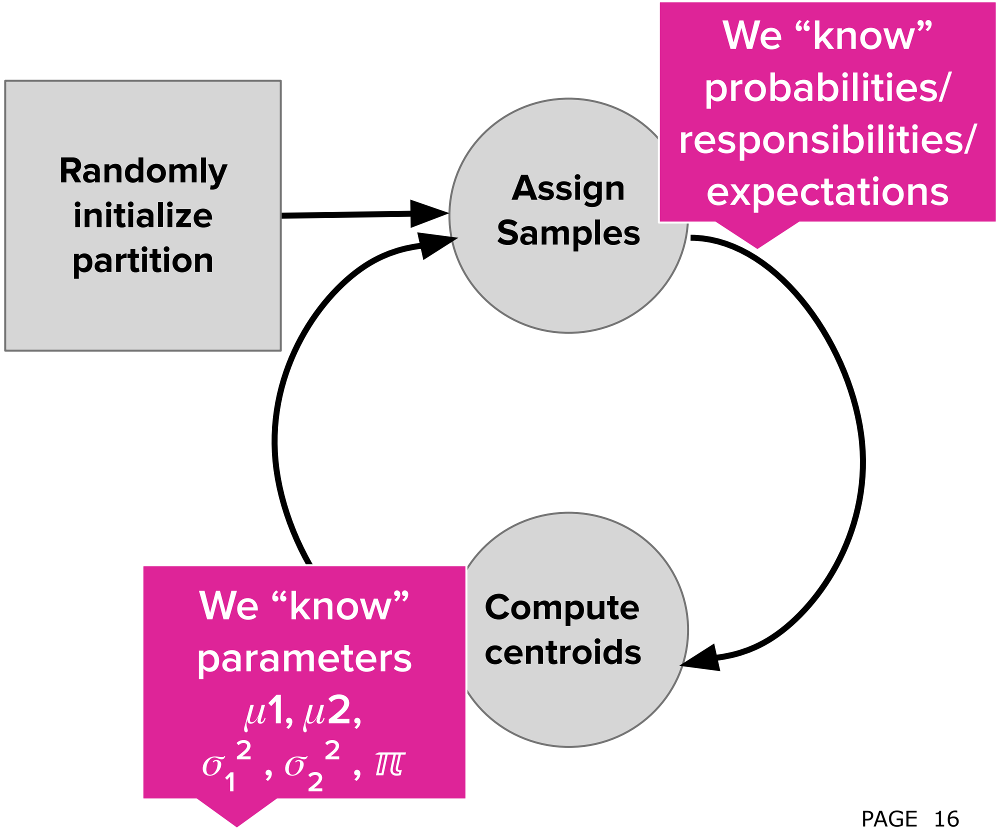
Could we combine these two somehow?

Recall Lloyd’s algorithm for K-Means clustering



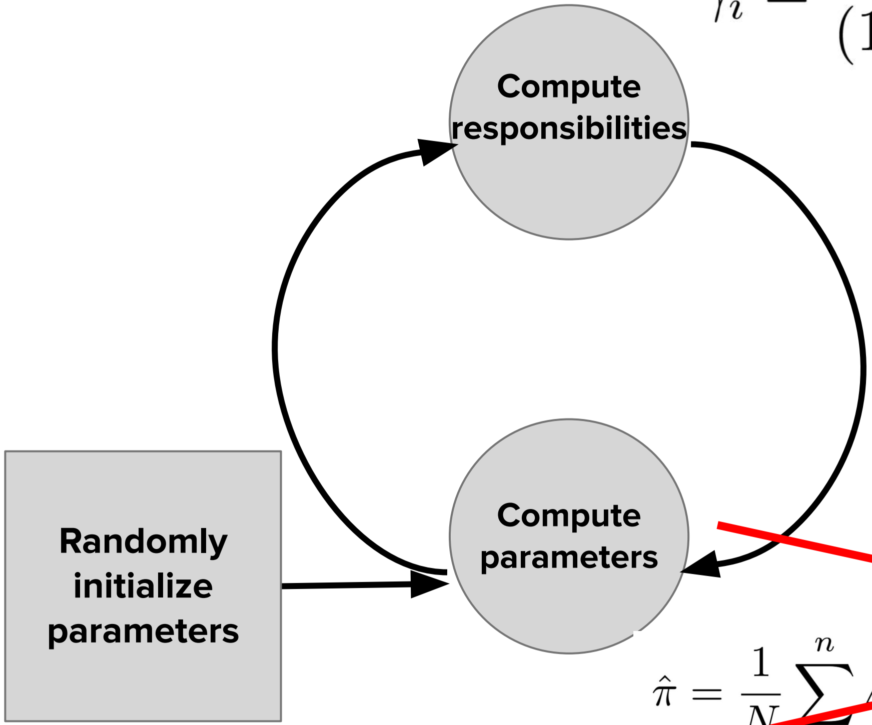


Comparing Lloyd's algorithm to GMM parameter estimation



Adapting Lloyd's algorithm for GMM parameter estimation?

$$\hat{\gamma}_i = \frac{\pi \mathcal{N}_{\mu_2, \sigma_2^2}(x_i)}{(1-\pi)\mathcal{N}_{\mu_1, \sigma_1^2}(x_i) + \pi \mathcal{N}_{\mu_2, \sigma_2^2}(x_i)}$$

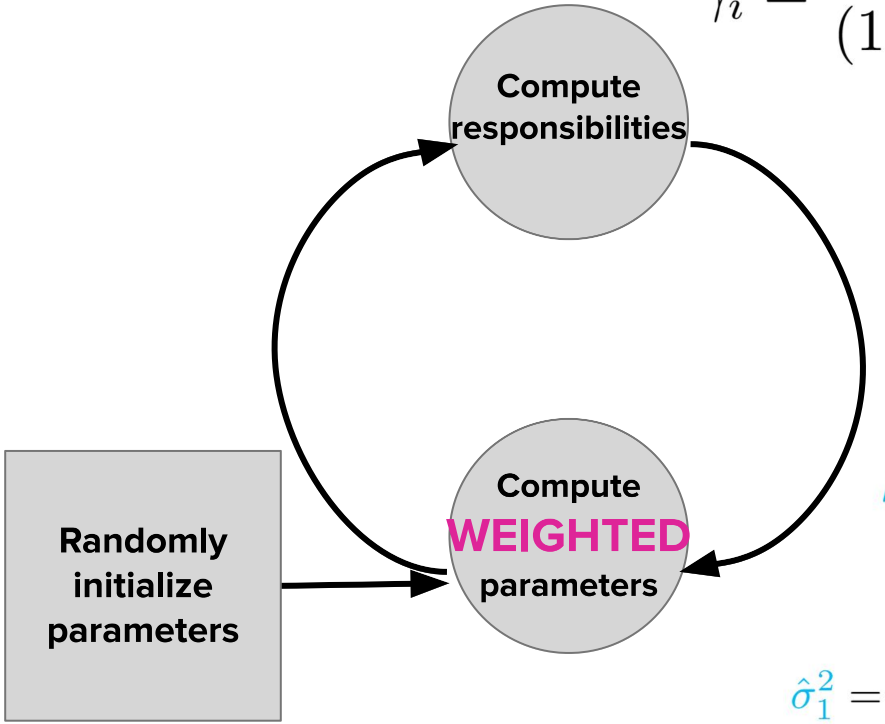


$$\hat{\mu}_1 = \frac{1}{|\Delta_0|}\sum_{i \in \Delta_0} x_i \qquad \hat{\sigma}_1^2 = \frac{1}{|\Delta_0|}\sum_{i \in \Delta_0} (x_i - \hat{\mu}_1)^2$$

$$\hat{\pi} = \frac{1}{N}\sum_{i=1}^{n} \Delta_i \qquad \hat{\mu}_2 = \frac{1}{|\Delta_1|}\sum_{i \in \Delta_1} x_i \qquad \hat{\sigma}_2^2 = \frac{1}{|\Delta_1|}\sum_{i \in \Delta_1} (x_i - \hat{\mu}_2)^2$$

Adapting Lloyd's algorithm for GMM parameter estimation?

$$\hat{\gamma}_i = \frac{\pi \mathcal{N}_{\mu_2,\sigma_2^2}(x_i)}{(1-\pi)\mathcal{N}_{\mu_1,\sigma_1^2}(x_i) + \pi \mathcal{N}_{\mu_2,\sigma_2^2}(x_i)}$$

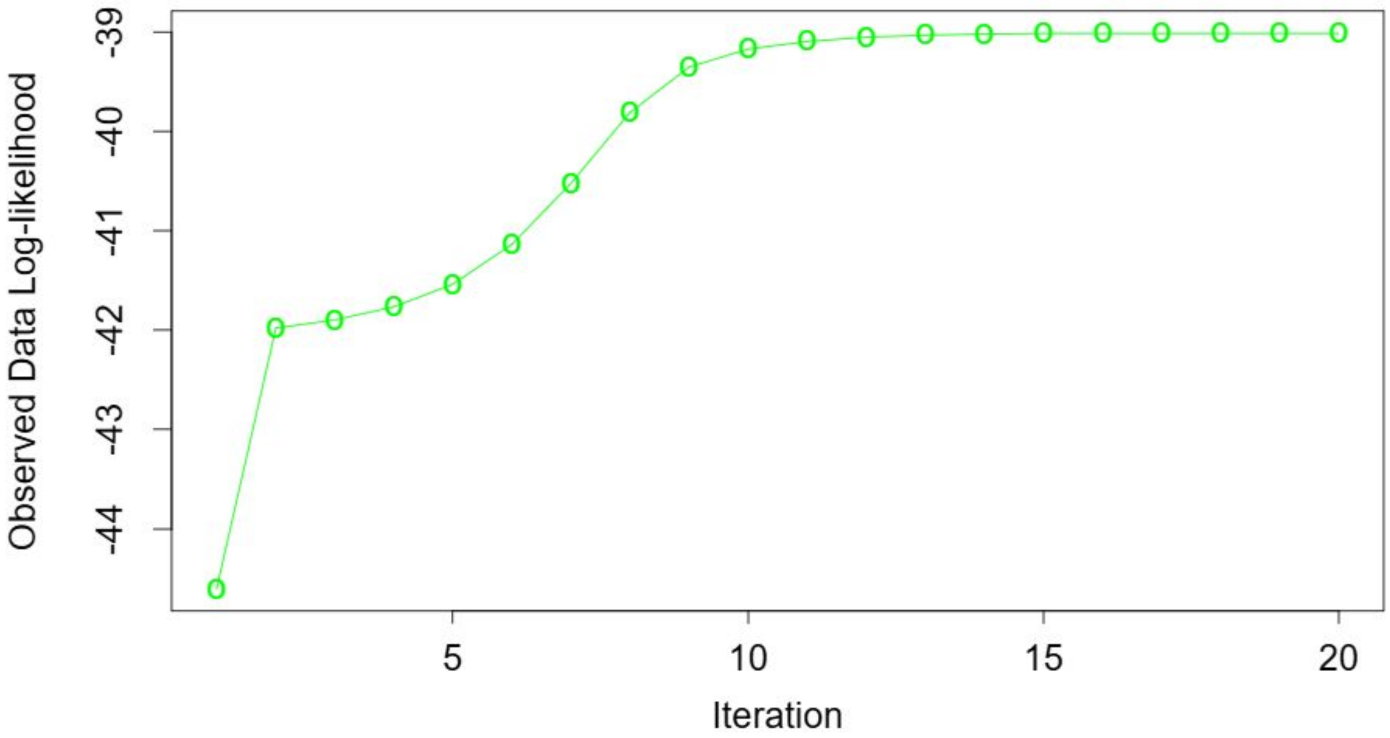


$$\hat{\pi} = \frac{\sum_{i=1}^{N} \hat{\gamma}_i}{N}$$

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{N} (1-\hat{\gamma}_i)\, y_i}{\sum_{i=1}^{N} (1-\hat{\gamma}_i)} \qquad \hat{\mu}_2 = \frac{\sum_{i=1}^{N} \hat{\gamma}_i\, y_i}{\sum_{i=1}^{N} \hat{\gamma}_i}$$

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{N} (1-\hat{\gamma}_i)\, (y_i - \hat{\mu}_1)^2}{\sum_{i=1}^{N} (1-\hat{\gamma}_i)} \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^{N} \hat{\gamma}_i\, (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^{N} \hat{\gamma}_i}$$

Iterative procedure convergences on the given dataset



Algorithm 8.1 *EM Algorithm for Two-component Gaussian Mixture.*

1. Take initial guesses for the parameters $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\pi}$ (see text).

2. *Expectation Step*: compute the responsibilities

$$\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\hat{\theta}_2}(y_i)}, \quad i = 1, 2, \ldots, N. \tag{8.42}$$

3. *Maximization Step*: compute the weighted means and variances:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i) y_i}{\sum_{i=1}^N (1-\hat{\gamma}_i)}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1-\hat{\gamma}_i)},$$

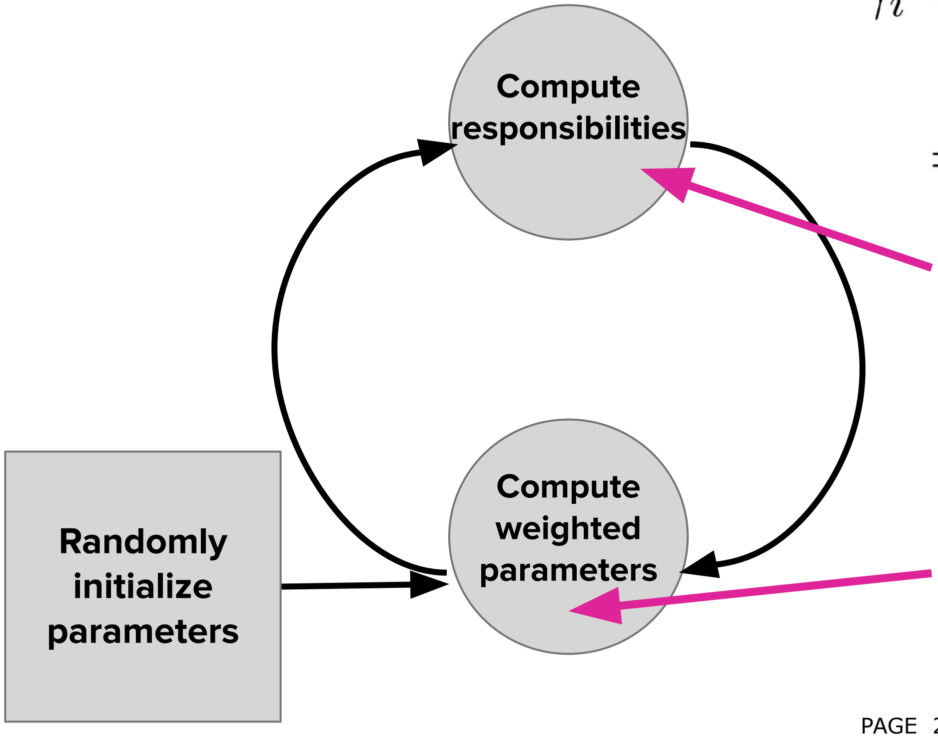
$$\hat{\mu}_2 = \frac{\sum_{i=1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^N \hat{\gamma}_i},$$

   and the mixing probability $\hat{\pi} = \sum_{i=1}^N \hat{\gamma}_i / N$.

4. Iterate steps 2 and 3 until convergence.

Elements of Statistical Learning, Section 8.5

Why is it called Expectation-Maximization (EM)?



$$\hat{\gamma}_i = \frac{\pi \mathcal{N}_{\mu_2,\sigma_2^2}(x_i)}{(1-\pi)\mathcal{N}_{\mu_1,\sigma_1^2}(x_i) + \pi \mathcal{N}_{\mu_2,\sigma_2^2}(x_i)}$$

$$= \mathbb{E}[\Delta_i \mid \pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \boldsymbol{X}]$$

Compute **expectation** of latent variables
"E" step

Compute parameters that
maximize the weighted log likelihood
"M" step

Gaussian Mixture Models

The probability density for a point x is determined by the sum of densities of independent Gaussian distributions

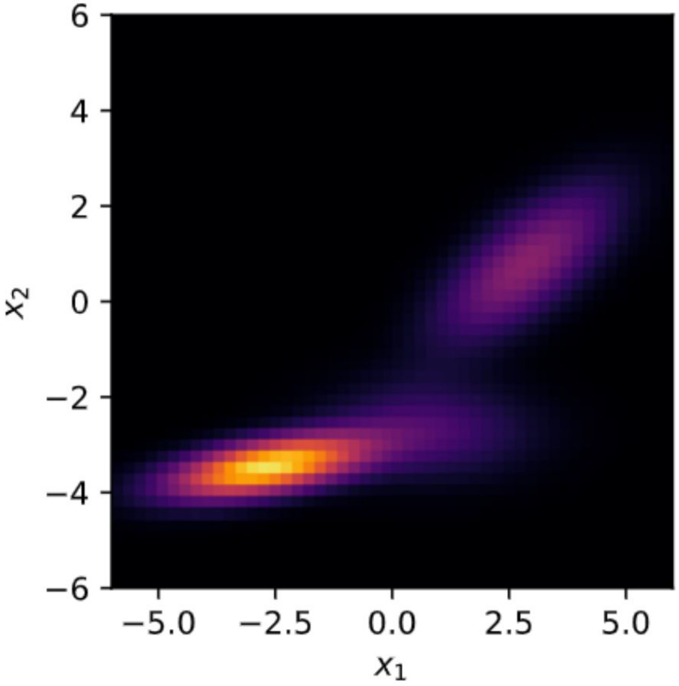
$$p(x) = \sum_{j=1}^{k} \pi_j \mathcal{N}(\mu_j, \Sigma_j, x)$$

Where:

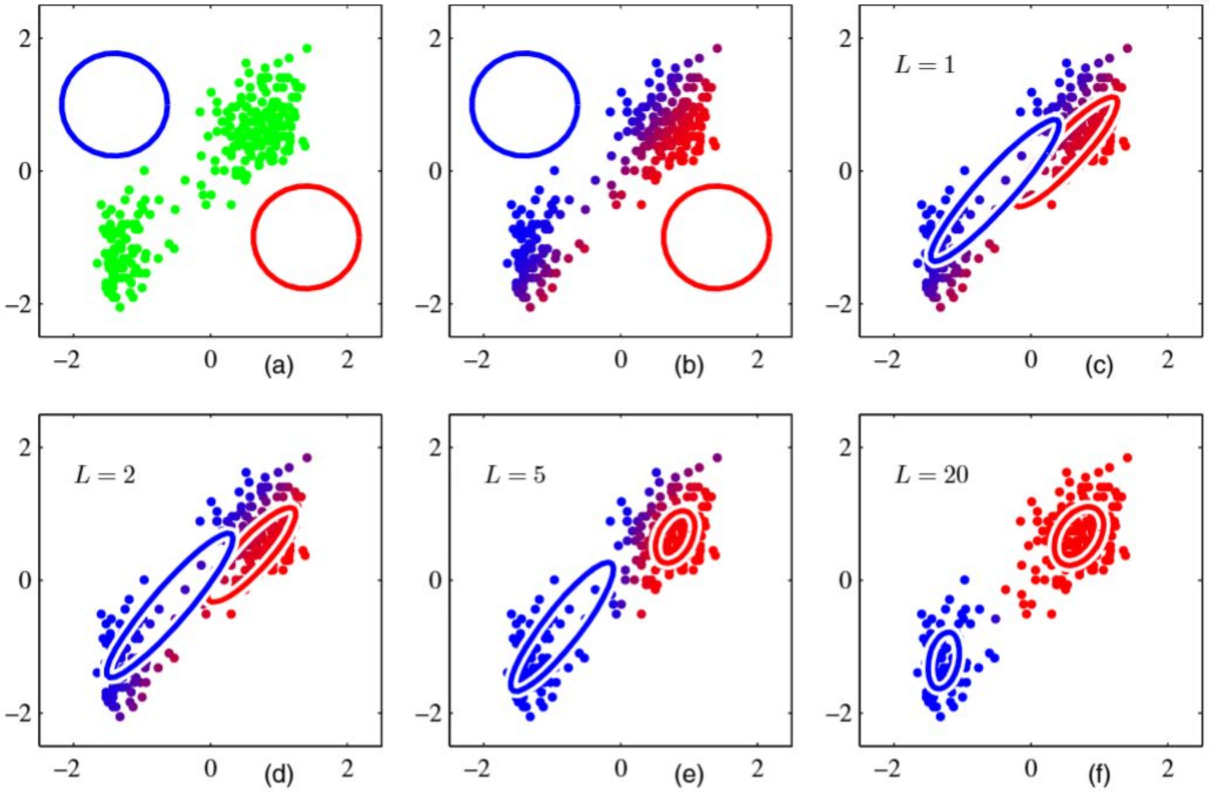
μ_j, Σ_j: mean vector and covariance matrix of j^{th} Gaussian, for $x \in \mathbb{R}^{\mathrm{d}}, \mathrm{d} > 1$ each Gaussian is multivariate

k: number of Gaussians in the model,

π_j: mixing weight associated with with the j^{th} Gaussian; $\pi_j \in [0, 1]$and $\sum_{j=1}^{k} = 1$



EM for mixtures of multivariate Gaussians



Pattern Recognition and Machine Learning, Section 9.2

Lecture Outline

I. How does the algorithm work in a common special case?

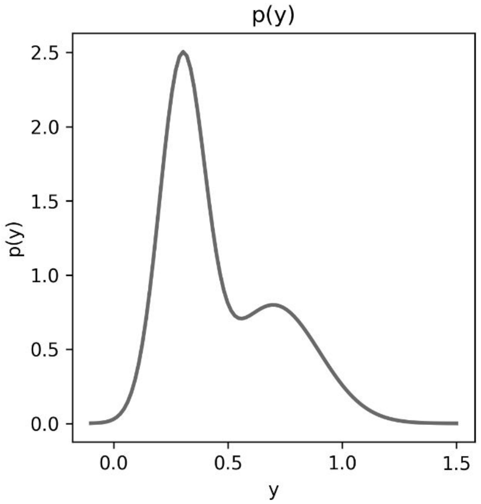
II. **How does the algorithm work in general?**

General Expectation Maximization

$$\ell(\theta) = \sum_{n=1}^{N} \log p(y_n \mid \theta)$$

y_n : observed data

θ : parameters to estimate



General Expectation Maximization

$$\ell(\theta) = \sum_{n=1}^{N} \log p(y_n \mid \theta)$$

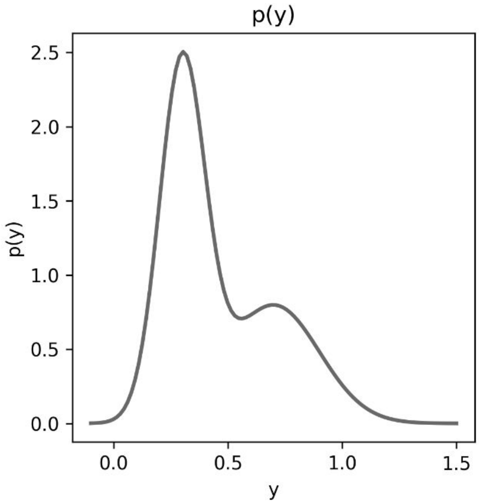
$$\ell(\theta) = \sum_{n=1}^{N} \log \left[\sum_{z_n} p(y_n, z_n \mid \theta) \right]$$

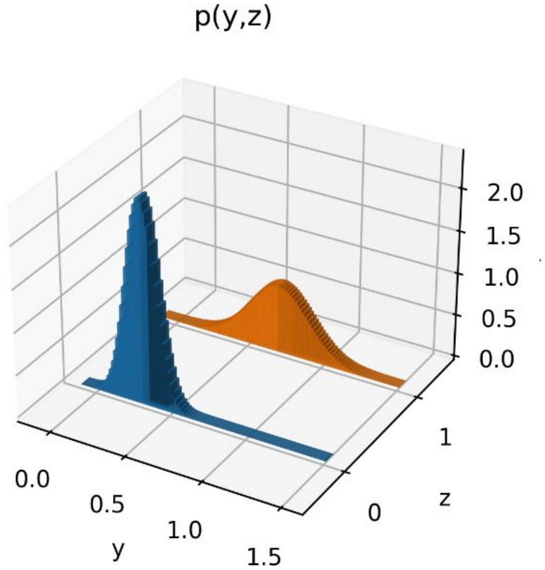
y_n : observed data

θ : parameters to estimate

z_n : hidden variables

$p(y_n, z_n \mid \theta)$: joint distribution of y_n and z_n





General Expectation Maximization

$$\ell(\theta) = \sum_{n=1}^{N} \log p(y_n \mid \theta)$$

$$\ell(\theta) = \sum_{n=1}^{N} \log \left[\sum_{z_n} p(y_n, z_n \mid \theta) \right]$$

$$\ell(\theta) = \sum_{n=1}^{N} \log \left[\sum_{z_n} p(y_n, z_n \mid \theta) \frac{q_n(z_n)}{q_n(z_n)} \right]$$

$$\ell(\theta) = \sum_{n=1}^{N} \log \left[\sum_{z_n} q_n(z_n) \frac{p(y_n, z_n \mid \theta)}{q_n(z_n)} \right]$$

$$\begin{aligned} y_n &: \text{ observed data} \\ \theta &: \text{ parameters to estimate} \\ z_n &: \text{ hidden variables} \\ p(y_n, z_n \mid \theta) &: \text{joint distribution of } y_n \text{ and } z_n \end{aligned}$$

General Expectation Maximization

$$\ell(\theta) = \sum_{n=1}^{N} \log p(y_n \mid \theta)$$

$$\ell(\theta) = \sum_{n=1}^{N} \log \left[\sum_{z_n} p(y_n, z_n \mid \theta) \right]$$

$$\ell(\theta) = \sum_{n=1}^{N} \log \left[\sum_{z_n} p(y_n, z_n \mid \theta) \frac{q_n(z_n)}{q_n(z_n)} \right]$$

$$\ell(\theta) = \sum_{n=1}^{N} \log \left[\sum_{z_n} q_n(z_n) \frac{p(y_n, z_n \mid \theta)}{q_n(z_n)} \right]$$

$$\ell(\theta) \geq \sum_{n} \sum_{z_n} q_n(z_n) \log \frac{p(y_n, z_n \mid \theta)}{q_n(z_n)}$$

y_n : observed data

θ : parameters to estimate

z_n : hidden variables

$p(y_n, z_n \mid \theta)$: joint distribution of y_n and z_n

Jensen's Inequality:

$$\log \mathbb{E}_{q_n}[Z] \geq \mathbb{E}_{q_n}[\log Z]$$

$$\log \sum_{z_n} q_n(z_n) \frac{p(y_n, z_n \mid \theta)}{q_n(z_n)} \geq \sum_{z_n} q_n(z_n) \log \frac{p(y_n, z_n \mid \theta)}{q_n(z_n)}$$

How can we maximize $\ell(\theta)$?

$$\ell(\theta) \geq \sum_n \sum_{z_n} q_n(z_n) \log \frac{p(z_n \mid y_n, \theta) p(y_n \mid \theta)}{q_n(z_n)}$$

$$\geq \sum_n \sum_{z_n} q_n(z_n) \log \frac{p(z_n \mid y_n, \theta)}{q_n(z_n)} p(y_n \mid \theta)$$

$$\geq \sum_n \left[\sum_{z_n} q_n(z_n) \log \frac{p(z_n \mid y_n, \theta)}{q_n(z_n)} + \sum_{z_n} q_n(z_n) \log p(y_n \mid \theta) \right]$$

$$\geq \sum_n \left[-D_{\mathrm{KL}}\left(q_n(z_n) \,\|\, p(z_n \mid y_n, \theta)\right) + \log p(y_n \mid \theta) \right]$$

Select: $q_n^* = p(z_n \mid y_n, \theta)$

$$\implies \ell(\theta) = \sum_n \log p(y_n \mid \theta)$$

Kullback-Leibler divergence

$$D_{\mathrm{KL}}(q \,\|\, p) \triangleq \sum_z q(z) \log \frac{q(z)}{p(z)}$$

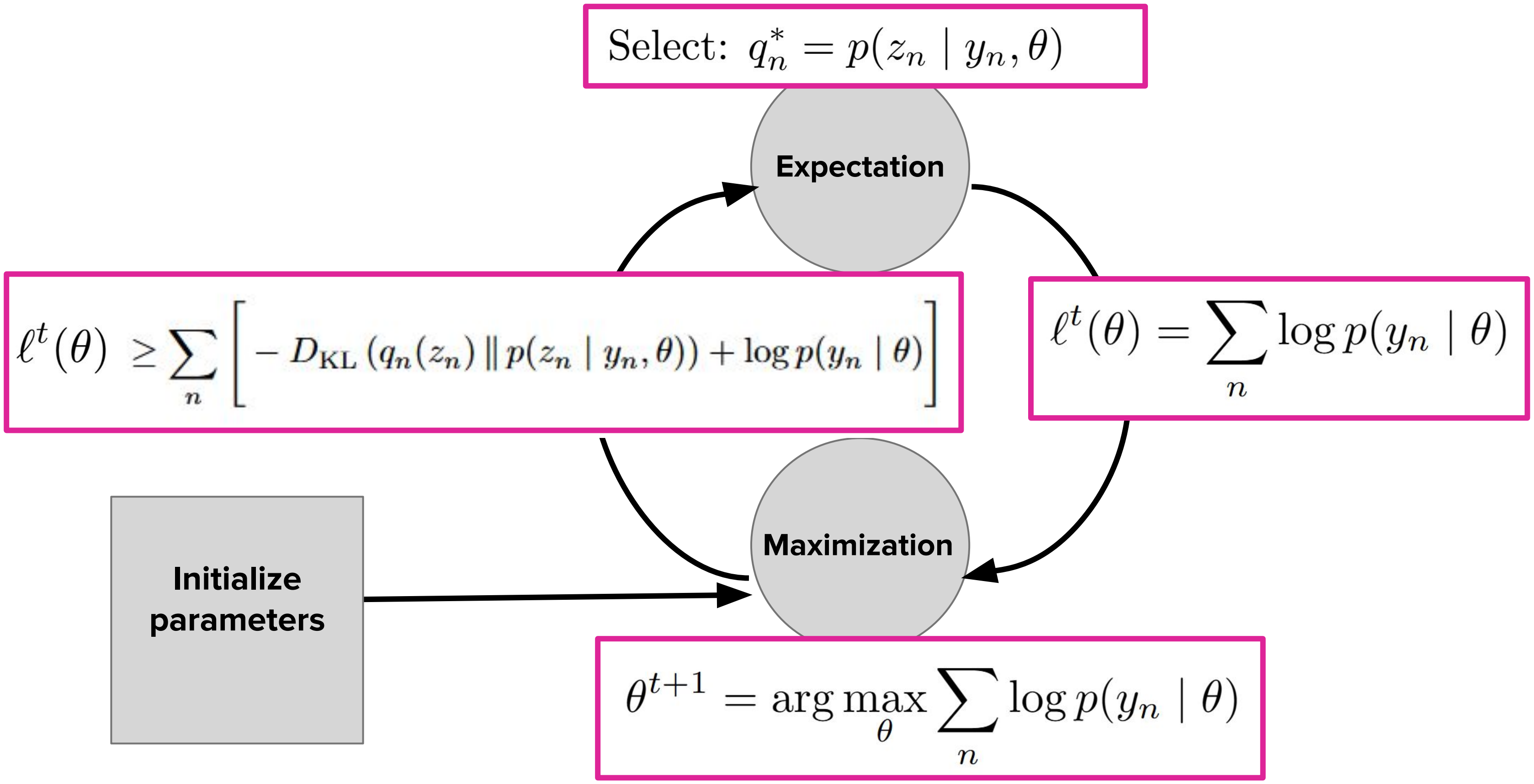
$$D_{\mathrm{KL}}(q \,\|\, p) \geq 0$$

$$D_{\mathrm{KL}}(q \,\|\, p) = 0 \quad \text{iff} \quad q = p$$

How can we maximize $\ell(\theta)$?

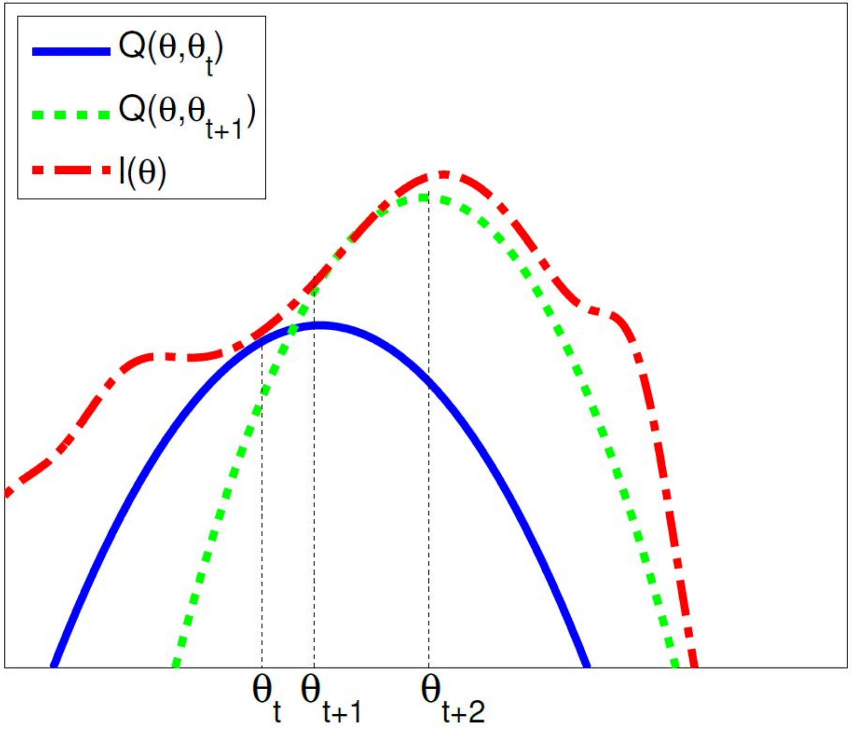
$$\ell^t(\theta) = \sum_n \log p(y_n \mid \theta)$$

$$\theta^{t+1} = \arg\max_\theta \sum_n \log p(y_n \mid \theta)$$



Select: $q_n^* = p(z_n \mid y_n, \theta)$
Expectation
$\ell^t(\theta) \geq \sum_n \left[-D_{\mathrm{KL}}\left(q_n(z_n) \,\|\, p(z_n \mid y_n, \theta)\right) + \log p(y_n \mid \theta) \right]$
$\ell^t(\theta) = \sum_n \log p(y_n \mid \theta)$
Maximization
Initialize parameters
$\theta^{t+1} = \arg\max_\theta \sum_n \log p(y_n \mid \theta)$

EM as bound optimization



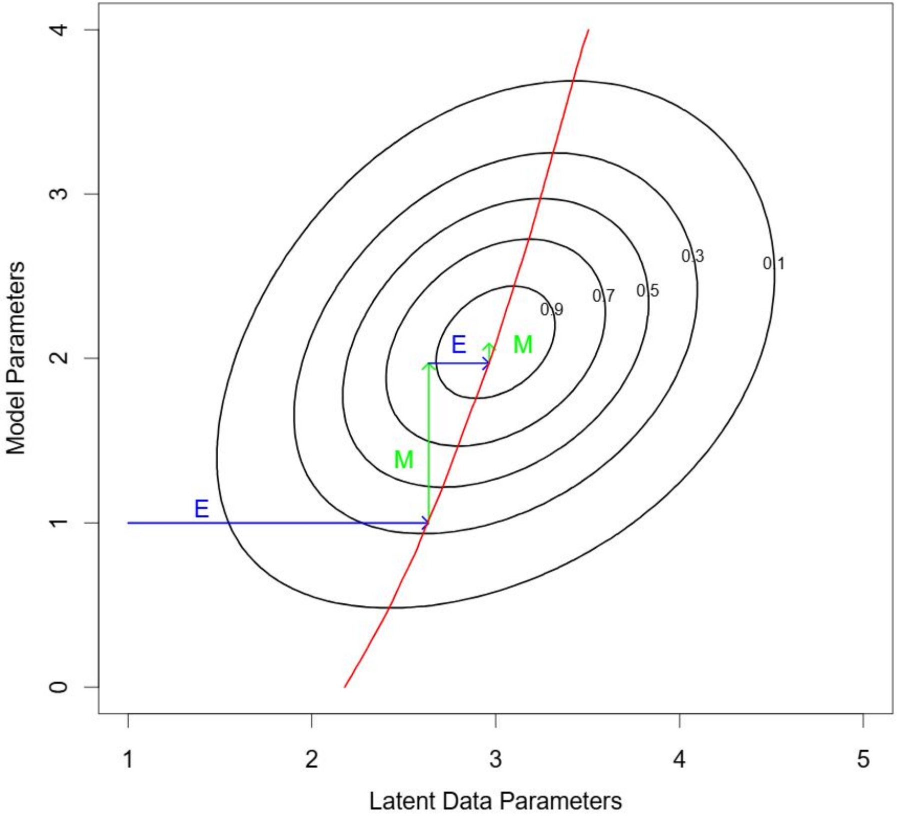
$$\ell(\theta) \geq -D_{\mathrm{KL}}\left(q_n(z_n) \,\|\, p(z_n \mid y_n, \theta)\right) + \log p(y_n \mid \theta)$$

$$\ell(\theta) \geq Q(\theta, \theta^t)$$

$$\ell(\theta^t) = Q(\theta^t, \theta^t)$$

Probabilistic Machine Learning, Section 8.7

EM as Maximization-Maximization



Elements of Statistical Learning, Section 8.5