EM Algorithm for Latent Variable Models

David Rosenberg

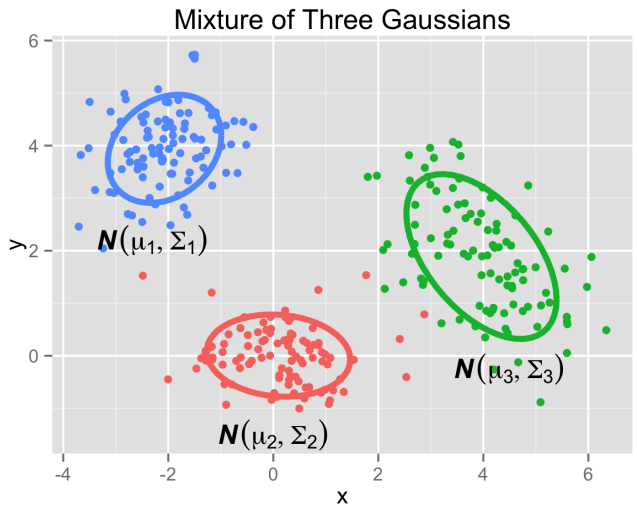
New York University

April 28, 2016

Gaussian Mixture Model: Review

Gaussian Mixture Model Parameters (k Components)

$$\text{Cluster probabilities}: \quad \pi = (\pi_1, \ldots, \pi_k)$$
$$\text{Cluster means}: \quad \mu = (\mu_1, \ldots, \mu_k)$$
$$\text{Cluster covariance matrices}: \quad \Sigma = (\Sigma_1, \ldots \Sigma_k)$$



GMM: The Joint and the Marginal Likelihood

- **Generative model** description:

$$\begin{aligned} z &\sim \text{Categorical}(\pi_1, \ldots, \pi_k) && \textbf{Cluster assignment} \\ x \mid z &\sim \mathcal{N}(\mu_z, \Sigma_z) && \textbf{Choose point from cluster distribution} \end{aligned}$$

- **Joint distribution** (includes observed x and unobserved z):

$$\begin{aligned} p(x, z) &= p(x \mid z) p(z) \\ &= \mathcal{N}(x \mid \mu_z, \Sigma_z) \pi_z \end{aligned}$$

- **Marginal distribution** (just observed variable x):

$$p(x) = \sum_{z=1}^{k} p(x, z) = \sum_{z=1}^{k} \pi_z \mathcal{N}(x \mid \mu_z, \Sigma_z)$$

Maximum Likelihood for the Gaussian Mixture Model

- Find parameters that give **observed data** the **highest likelihood**.
- The model likelihood for $\mathcal{D} = \{x_1, \ldots, x_n\}$ is

$$p(\mathcal{D}) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \left[\sum_{z=1}^{k} \pi_z \mathcal{N}(x_i \mid \mu_z, \Sigma_z) \right].$$

- The log-likelihood objective function:

$$J(\pi, \mu, \Sigma) = \sum_{i=1}^{n} \log \left[\sum_{z=1}^{k} \pi_z \mathcal{N}(x_i \mid \mu_z, \Sigma_z) \right]$$

- MLE is $\left(\hat{\pi}, \hat{\mu}, \hat{\Sigma}\right) = \arg\max_{\pi, \mu, \Sigma} J(\pi, \mu, \Sigma)$. EM algorithm to find it...

The EM Algorithm for GMM

Cluster Probabilities and Expected Cluster Sizes

- Probability that observed value x_i comes from cluster c:

$$\gamma_i^c := \mathbb{P}(z_i = c \mid x_i).$$

- The vector $(\gamma_i^1, \ldots, \gamma_i^k)$ gives the **soft cluster assignments** for x_i.
- Let n_c be the **expected number** of points in cluster c:

$$\begin{aligned} n_c &= \mathbb{E}\left[\sum_i 1(z_i = c) \mid x_1, \ldots, x_n\right] \\ &= \sum_i \mathbb{P}(z_i = c \mid x_i) \\ &= \sum_{i=1}^n \gamma_i^c \end{aligned}$$

.

EM Algorithm for GMM

1. Initialize parameters μ, Σ, π.
2. "E step". Evaluate the **responsibilities** using current parameters:

$$\gamma_i^j = \mathbb{P}(z_i = j \mid x_i) = \frac{\pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}{\sum_{c=1}^k \pi_c \mathcal{N}(x_i \mid \mu_c, \Sigma_c)},$$

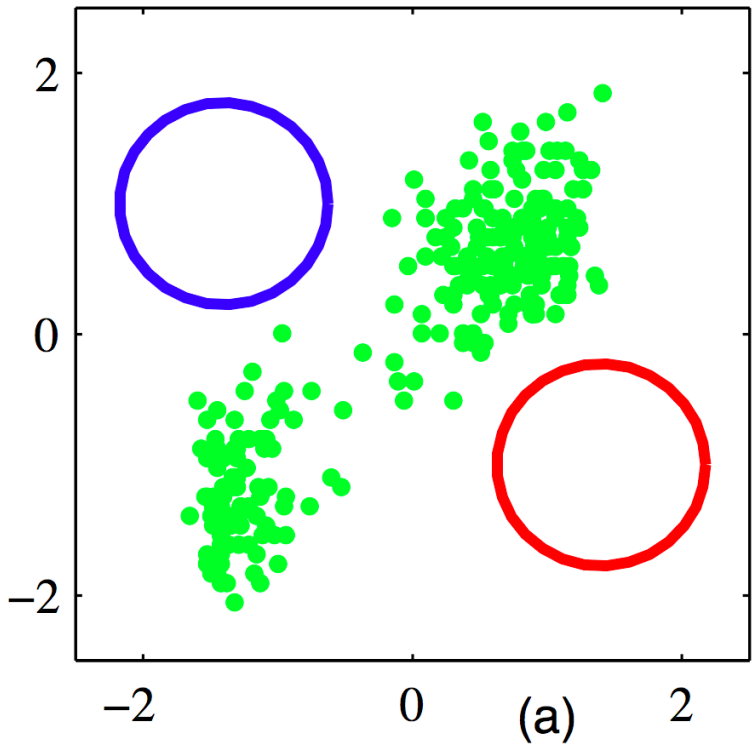
   for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.
3. "M step". Re-estimate the parameters using responsibilities:

$$\begin{aligned}
\mu_c^{\text{new}} &= \frac{1}{n_c} \sum_{i=1}^n \gamma_i^c x_i \\
\Sigma_c^{\text{new}} &= \frac{1}{n_c} \sum_{i=1}^n \gamma_i^c (x_i - \mu_{\text{MLE}})(x_i - \mu_{\text{MLE}})^T \\
\pi_c^{\text{new}} &= \frac{n_c}{n},
\end{aligned}$$

4. Repeat from Step 2, until log-likelihood converges.

EM for GMM

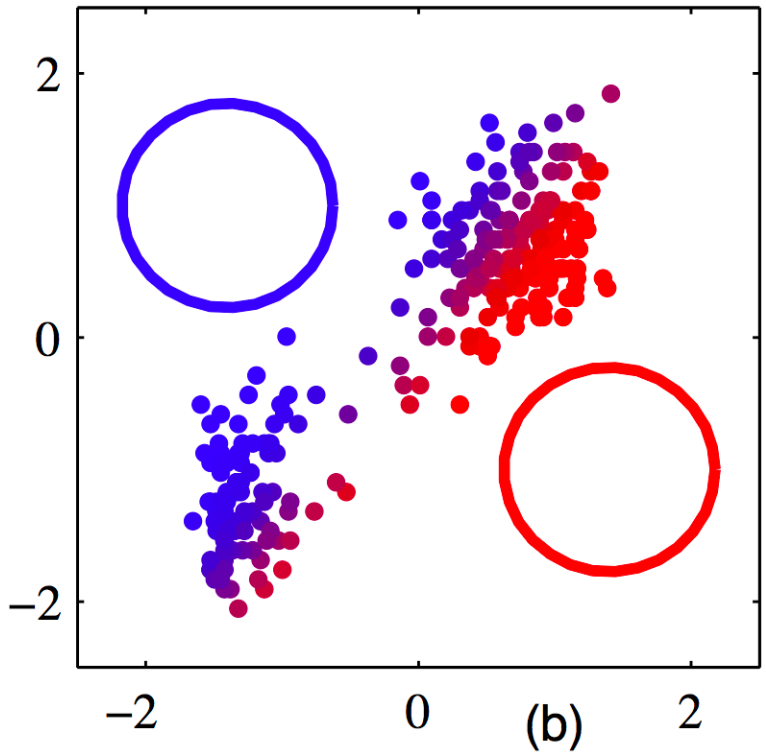
- Initialization



From Bishop's *Pattern recognition and machine learning*, Figure 9.8.

EM for GMM

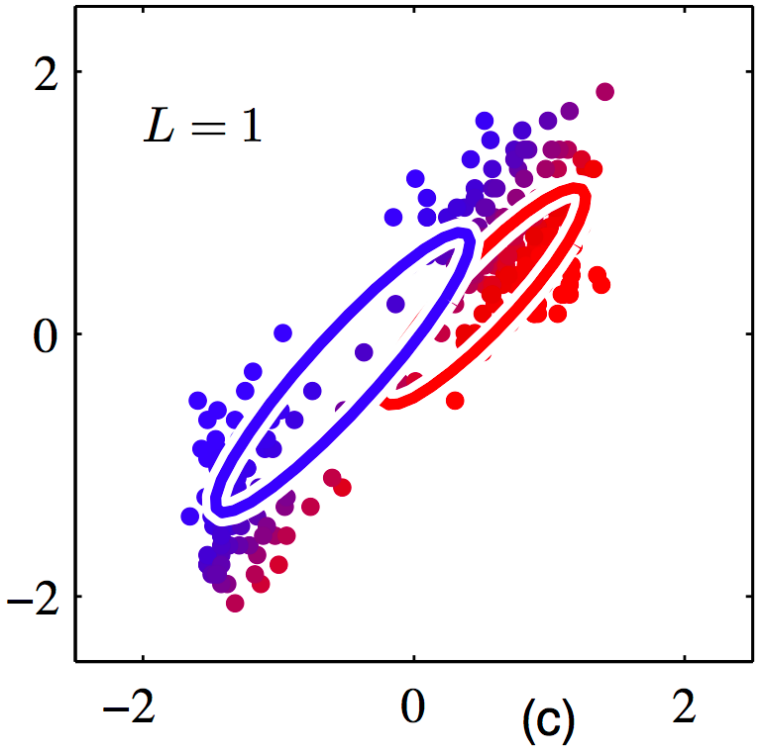
- First soft assignment:



From Bishop's *Pattern recognition and machine learning*, Figure 9.8.

EM for GMM

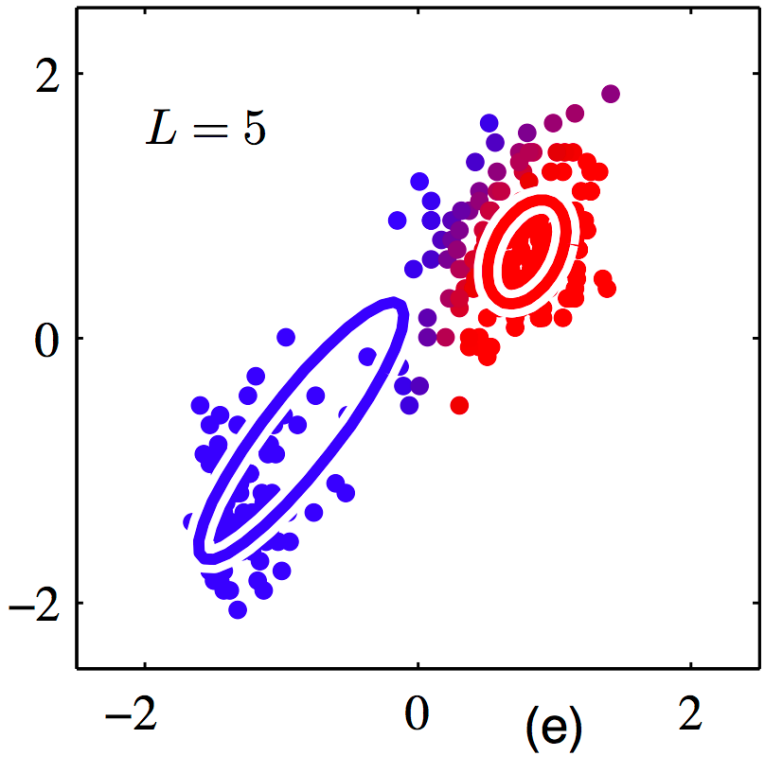
- First soft assignment:



From Bishop's *Pattern recognition and machine learning*, Figure 9.8.

EM for GMM

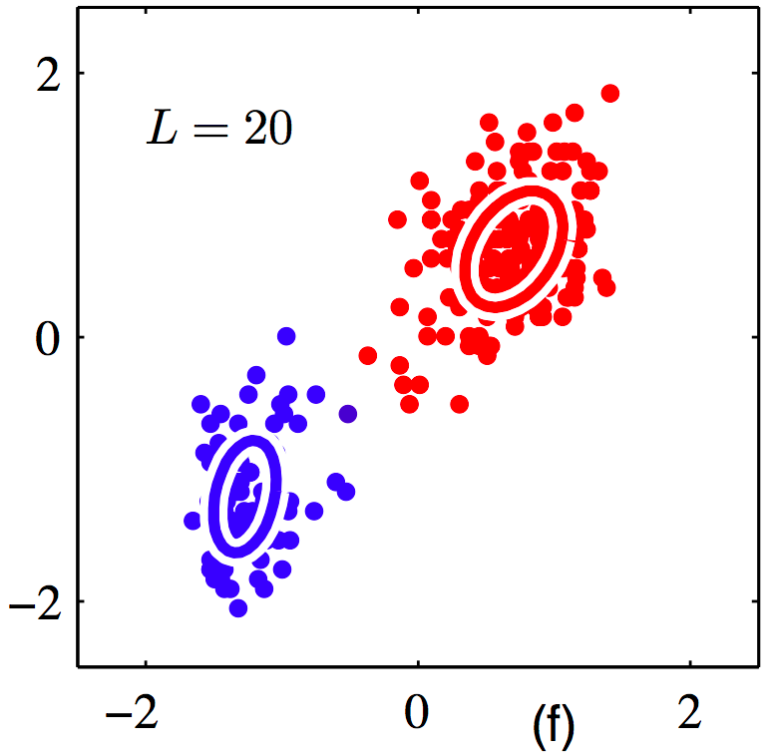
- After 5 rounds of EM:



From Bishop's *Pattern recognition and machine learning*, Figure 9.8.

EM for GMM

- After 20 rounds of EM:



From Bishop's *Pattern recognition and machine learning*, Figure 9.8.

Relation to K-Means

- EM for GMM seems a little like k-means.
- In fact, there is a precise correspondence.
- First, fix each cluster covariance matrix to be $\sigma^2 I$.
- As we take $\sigma^2 \to 0$, the update equations converge to doing k-means.
- If you do a quick experiment yourself, you'll find
  - Soft assignments converge to hard assignments.
  - Has to do with the tail behavior (exponential decay) of Gaussian.

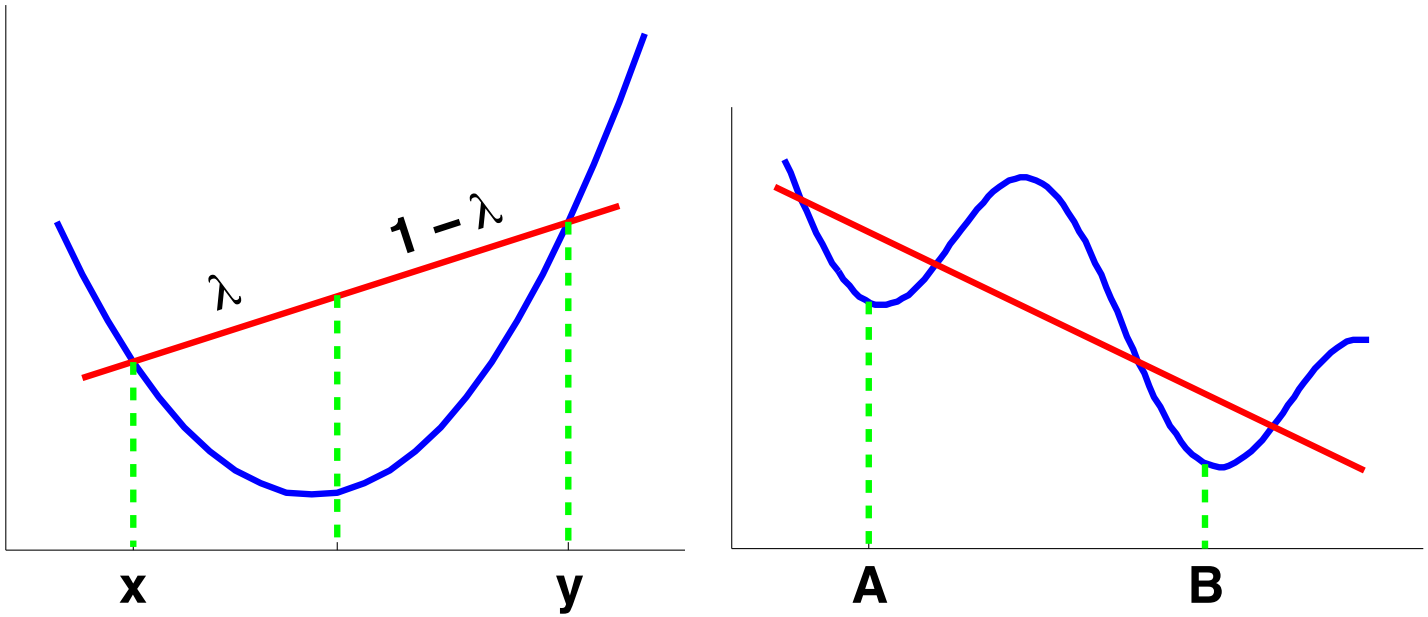
Math Prerequisites

Convex and Concave Functions

Definition

A function $f : \mathbf{R}^n \to \mathbf{R}$ is **convex** if for all $x, y \in \mathbf{R}^n$ and $0 \leqslant \theta \leqslant 1$, we have

$$f(\theta x + (1-\theta)y) \leqslant \theta f(x) + (1-\theta) f(y).$$



Jensen's Inequality

Theorem (Jensen's Inequality)

*If $f : \mathbf{R} \to \mathbf{R}$ is a **convex** function, and x is a random variable, then*

$$\mathbb{E}f(x) \geqslant f(\mathbb{E}x).$$

*Moreover, if f is **strictly convex**, then equality implies that $x = \mathbb{E}x$ with probability 1 (i.e. x is a constant).*

- e.g. $f(x) = x^2$ is convex. So $\mathbb{E}x^2 \geqslant (\mathbb{E}x)^2$. Thus

$$\mathrm{Var}x = \mathbb{E}x^2 - (\mathbb{E}x)^2 \geqslant 0.$$

Kullback-Leibler Divergence

- Let $p(x)$ and $q(x)$ be probability mass functions (PMFs) on $\mathcal{X}$.
- How can we measure how "different" p and q are?
- The **Kullback-Leibler** or **"KL" Divergence** is defined by

$$\mathrm{KL}(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}.$$

(Assumes $q(x)=0$ implies $p(x)=0$.)

- Can also write this as

$$\mathrm{KL}(p\|q) = \mathbb{E}_{x\sim p}\log\frac{p(x)}{q(x)}.$$

Gibbs Inequality ($\mathrm{KL}(p\|q) \geqslant 0$ and $\mathrm{KL}(p\|p) = 0$)

Theorem (Gibbs Inequality)

Let $p(x)$ and $q(x)$ be PMFs on $\mathcal{X}$. Then

$$KL(p\|q) \geqslant 0,$$

with equality iff $p(x) = q(x)$ for all $x \in \mathcal{X}$.

- KL divergence measures the "distance" between distributions.
- Note:
  - KL divergence **not a metric**.
  - KL divergence is **not symmetric**.

Gibbs Inequality: Proof

$$
\begin{aligned}
\mathrm{KL}(p\|q) &= \mathbb{E}_p\left[-\log\left(\frac{q(x)}{p(x)}\right)\right] \\
&\geqslant -\log\left[\mathbb{E}_p\left(\frac{q(x)}{p(x)}\right)\right] \qquad \text{(Jensen's)} \\
&= -\log\left[\sum_{\{x|p(x)>0\}} p(x)\frac{q(x)}{p(x)}\right] \\
&= -\log\left[\sum_{x\in\mathcal{X}} q(x)\right] \\
&= -\log 1 = 0.
\end{aligned}
$$

- Since $-\log$ is strictly convex, we have strict equality iff $q(x)/p(x)$ is a constant, which implies $q = p$.

EM Algorithm for Latent Variable Models

General Latent Variable Model

- Two sets of random variables: z and x.
- z consists of unobserved **hidden variables**.
- x consists of **observed variables**.
- Joint probability model parameterized by $\theta \in \Theta$:

$$p(x, z \mid \theta)$$

Notation abuse

Notation $p(x, z \mid \theta)$ suggests a Bayesian setting, in which θ is a r.v. However we are **not** assuming a Bayesian setting. $p(x, z \mid \theta)$ is just easier to read than $p_\theta(x, z)$, once θ gets more complicated.

Complete and Incomplete Data

- An observation of x is called an **incomplete data set**.
- An observation (x, z) is called a **complete data set**.
  - We never have a complete data set for latent variable models.
  - But it's a useful construct.
- Suppose we have an incomplete data set $\mathcal{D} = (x_1, \ldots, x_n)$.
- To simplify notation, take x to represent the entire dataset

$$x = (x_1, \ldots, x_n),$$

and Z to represent the corresponding unobserved variables

$$z = (z_1, \ldots, z_n).$$

The EM Algorithm **Key Idea**

- Marginal log-likelihood is hard to optimize:

$$\max_{\theta} \log \left\{ \sum_{z} p(x, z \mid \theta) \right\}$$

- **Assume that** complete data log-likelihood would be easy to optimize:

$$\max_{\theta} \log p(x, z \mid \theta)$$

- What if we had a **distribution** $q(z)$ for the latent variables z?
- Then maximize the **expected complete data log-likelihood**:

$$\max_{\theta} \sum_{z} q(z) \log p(x, z \mid \theta)$$

- EM **assumes** this maximization is feasible.

Lower Bound for Likelihood

- Let $q(z)$ be any PMF on $\mathcal{Z}$, the support of Z:

$$\begin{aligned}
\log p(x \mid \theta) &= \log \left[\sum_z p(x, z \mid \theta) \right] \\
&= \log \left[\sum_z q(z) \left(\frac{p(x, z \mid \theta)}{q(z)} \right) \right] \quad \text{(log of an expectation)} \\
&\geqslant \sum_z q(z) \log \left(\frac{p(x, z \mid \theta)}{q(z)} \right) \quad \text{(expectation of log)} \\
&=: \mathcal{L}(q, \theta).
\end{aligned}$$

- The inequality is by Jensen's, by concavity of the log.

This is the **key step** for "variational methods".

Lower Bound and Expected Complete Log-Likelihood

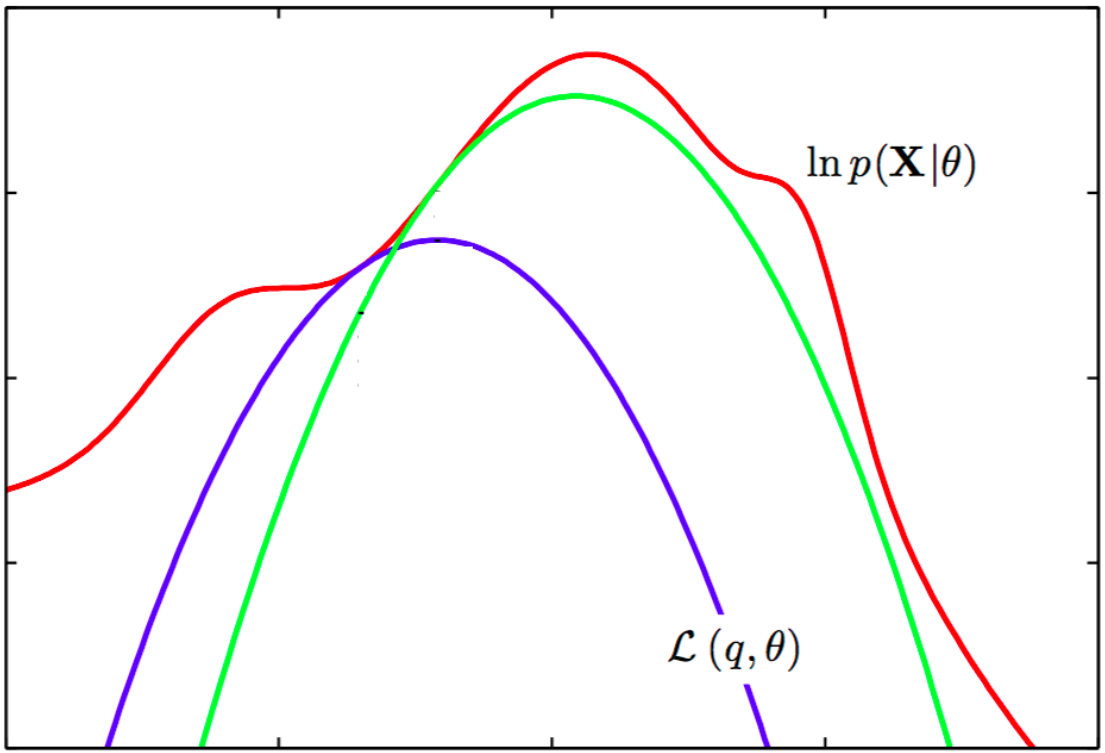
- Consider maximizing the lower bound $\mathcal{L}(q, \theta)$:

$$\begin{aligned}\mathcal{L}(q,\theta) &= \sum_z q(z) \log \left(\frac{p(x,z \mid \theta)}{q(z)} \right) \\ &= \underbrace{\sum_z q(z) \log p(x,z \mid \theta)}_{\mathbb{E}[\text{complete data log-likelihood}]} \underbrace{- \sum_z q(z) \log q(z)}_{\text{no } \theta \text{ here}}\end{aligned}$$

- Maximizing $\mathcal{L}(q, \theta)$ equivalent to maximizing $\mathbb{E}$[complete data log-likelihood].

A Family of Lower Bounds

- Each q gives a different lower bound: $\log p(x \mid \theta) \geqslant \mathcal{L}(q, \theta)$
- Two lower bounds, as functions of θ:



From Bishop's *Pattern recognition and machine learning*, Figure 9.14.

EM: Big Picture Idea

- The following inequality holds for all θ and q:

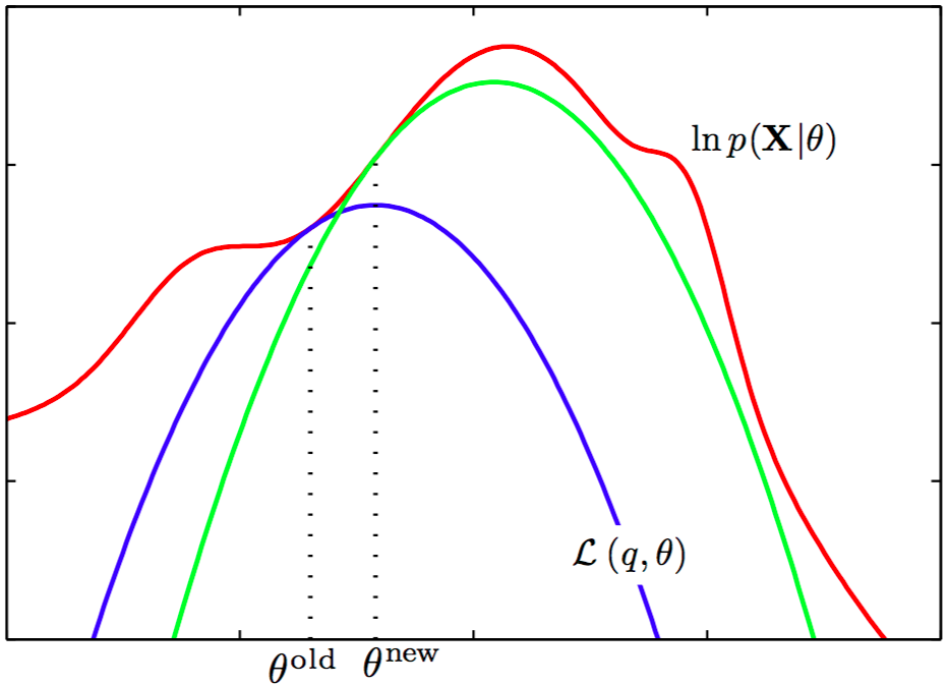
$$\log p(x \mid \theta) \geqslant \mathcal{L}(q, \theta).$$

- We want to find θ that maximizes $\log p(x \mid \theta)$.
- $\log p(x \mid \theta)$ is hard to maximize directly.
- Two step version of the EM algorithm:
  1. We vary q and θ, searching for the biggest $\mathcal{L}(q, \theta)$ we can find.
  2. Final result is $\hat{\theta}$ corresponding to the largest $\mathcal{L}(q, \theta)$ we found.
- Often this is a local maximum of the likelihood.
- One question left: How to choose the sequence of q's and θ's we try?

EM: Coordinate Ascent on Lower Bound

- Choose sequence of q's and θ's by "**coordinate ascent**".
- EM Algorithm (high level):
  1. Choose initial θ^{old}.
  2. Let $q^* = \arg\max_q \mathcal{L}(q, \theta^{\text{old}})$
  3. Let $\theta^{\text{new}} = \arg\max_\theta \mathcal{L}(q^*, \theta^{\text{old}})$.
  4. Go to step 2, until converged.
- Will show: $p(x \mid \theta^{\textbf{new}}) \geqslant p(x \mid \theta^{\textbf{old}})$
- **Get sequence of θ's with monotonically increasing likelihood.**

EM: Coordinate Ascent on Lower Bound



1. Start at θ^{old}.
2. Find q giving best lower bound at $\theta^{\text{old}} \implies \mathcal{L}(q,\theta)$.
3. $\theta^{\text{new}} = \arg\max_\theta \mathcal{L}(q,\theta)$.

From Bishop's *Pattern recognition and machine learning*, Figure 9.14.

The Lower Bound

- Let's investigate the lower bound:

$$\begin{aligned}
\mathcal{L}(q,\theta) &= \sum_z q(z) \log\left(\frac{p(x,z \mid \theta)}{q(z)}\right) \\
&= \sum_z q(z) \log\left(\frac{p(z \mid x,\theta)p(x \mid \theta)}{q(z)}\right) \\
&= \sum_z q(z) \log\left(\frac{p(z \mid x,\theta)}{q(z)}\right) + \sum_z q(z) \log p(x \mid \theta) \\
&= -\mathrm{KL}[q(z), p(z \mid x,\theta)] + \log p(x \mid \theta)
\end{aligned}$$

- Amazing! We get back an equality for the marginal likelihood:

$$\log p(x \mid \theta) = \mathcal{L}(q,\theta) + \mathrm{KL}[q(z), p(z \mid x,\theta)]$$

The Best Lower Bound

- Find q maximizing

$$\mathcal{L}(q, \theta^{\text{old}}) = -\mathrm{KL}[q(z), p(z \mid x, \theta^{\text{old}})] + \underbrace{\log p(x \mid \theta^{\text{old}})}_{\text{no } q \text{ here}}$$

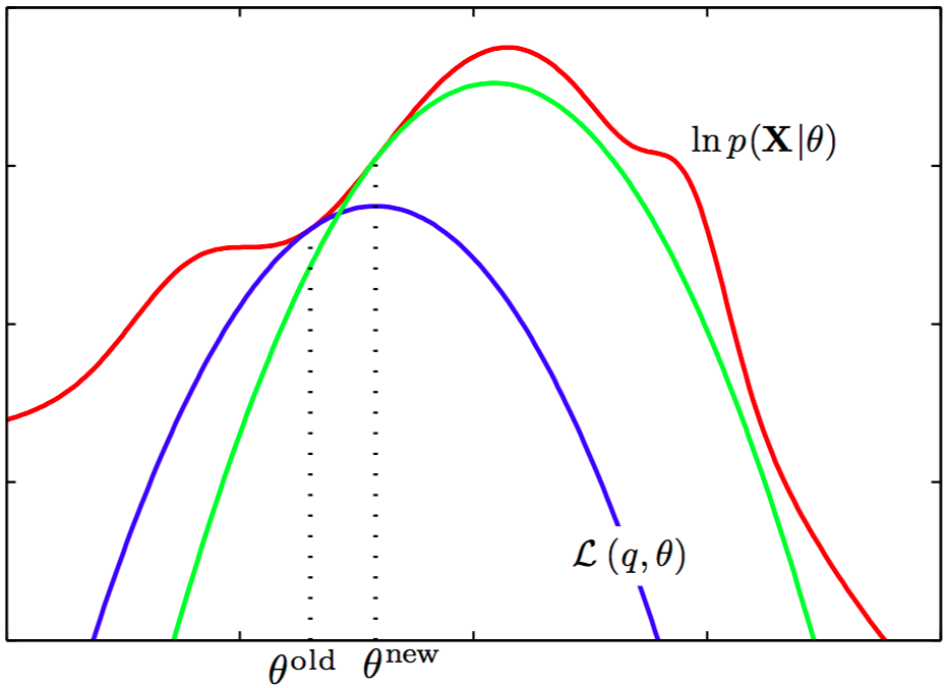
- Recall $\mathrm{KL}(p\|q) \geqslant 0$, and $\mathrm{KL}(p\|p) = 0$.
- Best q is $q^*(z) = p(z \mid x, \theta^{\text{old}})$. Proof:

$$\mathcal{L}(q^*, \theta^{\text{old}}) = -\underbrace{\mathrm{KL}[p(z \mid x, \theta^{\text{old}}), p(z \mid x, \theta^{\text{old}})]}_{=0} + \log p(x \mid \theta^{\text{old}})$$

- Summary:

$$\begin{aligned} \log p(x \mid \theta^{\text{old}}) &= \mathcal{L}(q^*, \theta^{\text{old}}) \quad \text{(tangent at } \theta^{\text{old}}\text{)}. \\ \log p(x \mid \theta) &\geqslant \mathcal{L}(q^*, \theta) \quad \forall \theta \end{aligned}$$

Tight lower bound for any chosen θ



Fix any θ' and take $q'(z) = p(z \mid x, \theta')$. Then

1. $\log p(x \mid \theta) \geqslant \mathcal{L}(q', \theta)$ $\forall \theta$. [Global lower bound].
2. $\log p(x \mid \theta') = \mathcal{L}(q', \theta')$. [Lower bound is **tight** at θ'.]

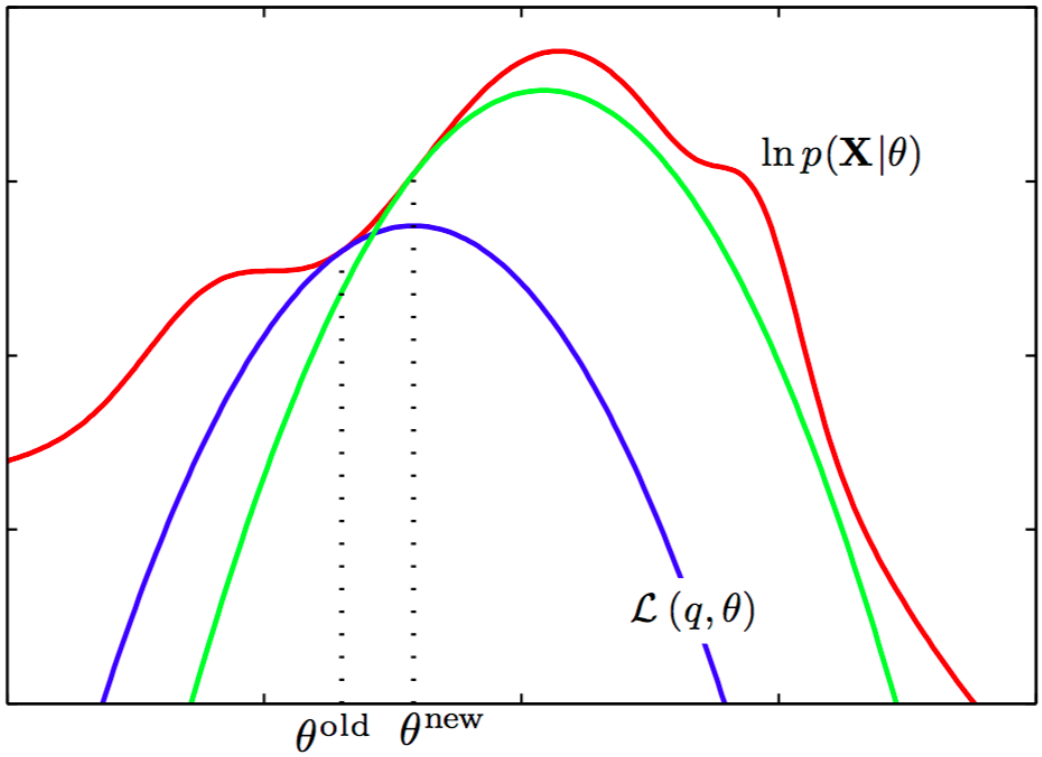
From Bishop's *Pattern recognition and machine learning*, Figure 9.14.

General EM Algorithm

1. Choose initial θ^{old}.
2. **Expectation Step**
   - Let $q^*(z) = p(z \mid x, \theta^{\text{old}})$. [$q^*$ gives best lower bound at θ^{old}]
   - Let
   $$J(\theta) := \mathcal{L}(q^*, \theta) = \sum_z q^*(z) \log \left(\frac{p(x, z \mid \theta)}{q^*(z)} \right)$$
   - Note that $J(\theta)$ is an **expectation** w.r.t. $z \sim q^*(z)$.
3. **Maximization Step**
   $$\theta^{\text{new}} = \arg\max_\theta J(\theta).$$
   [Equivalent to maximizing expected complete log-likelihood.]
4. Go to step 2, until converged.

EM Monotonically Increases Likelihood

EM Gives Monotonically Increasing Likelihood: By Picture



From Bishop's *Pattern recognition and machine learning*, Figure 9.14.

EM Gives Monotonically Increasing Likelihood: By Math

1. Start at θ^{old}.
2. Choose $q^*(z) = \arg\max_q \mathcal{L}(q, \theta^{\text{old}})$. We've shown

$$\log p(x \mid \theta^{\text{old}}) = \mathcal{L}(q^*, \theta^{\text{old}})$$

3. Choose $\theta^{\text{new}} = \arg\max_\theta \mathcal{L}(q^*, \theta^{\text{old}})$. So

$$\mathcal{L}(q^*, \theta^{\text{new}}) \geqslant \mathcal{L}(q^*, \theta^{\text{old}}).$$

Putting it together, we get

$$\begin{aligned}
\log p(x \mid \theta^{\text{new}}) &\geqslant \mathcal{L}(q^*, \theta^{\text{new}}) && \mathcal{L} \text{ is a lower bound} \\
&\geqslant \mathcal{L}(q^*, \theta^{\text{old}}) && \text{By definition of } \theta^{\text{new}} \\
&= \log p(x \mid \theta^{\text{old}}) && \text{Bound is tight at } \theta^{\text{old}}.
\end{aligned}$$

Suppose We Maximize the Lower Bound...

- Suppose we have found a global maximum of $\mathcal{L}(q, \theta)$:

$$L(q^*, \theta^*) \geqslant L(q, \theta) \; \forall q, \theta,$$

where of course

$$q^*(z) = p(z \mid x, \theta^*).$$

- Claim: θ^* is a global maximum of $\log p(x \mid \theta^*)$.
- Proof: For any θ', we showed that for $q'(z) = p(z \mid x, \theta')$ we have

$$\begin{aligned} \log p(x \mid \theta') &= \mathcal{L}(q', \theta') + \mathrm{KL}[q', p(z \mid x, \theta')] \\ &= \mathcal{L}(q', \theta') \\ &\leqslant \mathcal{L}(q^*, \theta^*) \\ &= \log p(x \mid \theta^*) \end{aligned}$$

Convergence of EM

- Let θ_n be value of EM algorithm after n steps.
- Define "transition function" $M(\cdot)$ such that $\theta_{n+1} = M(\theta_n)$.
- Suppose log-likelihood function $\ell(\theta) = \log p(x \mid \theta)$ is differentiable.
- Let S be the set of stationary points of $\ell(\theta)$. (i.e. $\nabla_\theta \ell(\theta) = 0$)

Theorem

Under mild regularity conditions[a]*, for any starting point* θ_0*,*

- $\lim_{n\to\infty} \theta_n = \theta^*$ *for some stationary point* $\theta^* \in S$ *and*
- θ^* *is a fixed point of the EM algorithm, i.e.* $M(\theta^*) = \theta^*$*. Moreover,*
- $\ell(\theta_n)$ *strictly increases to* $\ell(\theta^*)$ *as* $n \to \infty$*, unless* $\theta_n \equiv \theta^*$*.*

[a] For details, see "Parameter Convergence for EM and MM Algorithms" by Florin Vaida in *Statistica Sinica* (2005).
`http://www3.stat.sinica.edu.tw/statistica/oldpdf/a15n316.pdf`

Variations on EM

EM Gives Us Two New Problems

- The "E" Step: Computing

$$J(\theta) := \mathcal{L}(q^*, \theta) = \sum_z q^*(z) \log \left(\frac{p(x, z \mid \theta)}{q^*(z)} \right)$$

- The "M" Step: Computing

$$\theta^{\text{new}} = \arg\max_\theta J(\theta).$$

- Either of these can be too hard to do in practice.

Generalized EM (GEM)

- Addresses the problem of a difficult "M" step.
- Rather than finding

$$\theta^{\text{new}} = \arg\max_{\theta} J(\theta),$$

find **any** θ^{new} for which

$$J(\theta^{\text{new}}) > J(\theta^{\text{old}}).$$

- Can use a standard nonlinear optimization strategy
  - e.g. take a gradient step on J.
- We still get monotonically increasing likelihood.

EM and More General Variational Methods

- Suppose "E" step is difficult:
  - Hard to take expectation w.r.t. $q^*(z) = p(z \mid x, \theta^{\text{old}})$.
- Solution: Restrict to distributions $\mathcal{Q}$ that are easy to work with.
- Lower bound now looser:

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathbf{KL}[q(z), p(z \mid x, \theta^{\text{old}})]$$

EM in Bayesian Setting

- Suppose we have a prior $p(\theta)$.
- Want to find MAP estimate: $\hat{\theta}_{\text{MAP}} = \arg\max_\theta p(\theta \mid x)$:

$$\begin{aligned} p(\theta \mid x) &= p(x \mid \theta) p(\theta) / p(x) \\ \log p(\theta \mid x) &= \log p(x \mid \theta) + \log p(\theta) - \log p(x) \end{aligned}$$

.

- Still can use our lower bound on $\log p(x, \theta)$.

$$J(\theta) := \mathcal{L}(q^*, \theta) = \sum_z q^*(z) \log \left(\frac{p(x, z \mid \theta)}{q^*(z)} \right)$$

- Maximization step becomes

$$\theta^{\text{new}} = \arg\max_\theta [J(\theta) + \log p(\theta)]$$

- Homework: Convince yourself our lower bound is still tight at θ.

Homework: Gaussian Mixture Model (Hints)

Homework: Derive EM for GMM from General EM Algorithm

- Subsequent slides may help set things up.
- Key skills:
  - MLE for multivariate Gaussian distributions.
  - Lagrange multipliers

Gaussian Mixture Model (k Components)

- GMM Parameters

$$\begin{aligned} \text{Cluster probabilities}: &\quad \pi = (\pi_1, \ldots, \pi_k) \\ \text{Cluster means}: &\quad \mu = (\mu_1, \ldots, \mu_k) \\ \text{Cluster covariance matrices}: &\quad \Sigma = (\Sigma_1, \ldots \Sigma_k) \end{aligned}$$

- Let $\theta = (\pi, \mu, \Sigma)$.
- Marginal log-likelihood

$$\log p(x \mid \theta) \quad = \quad \log \left\{ \sum_{z=1}^{k} \pi_z \mathcal{N}(x \mid \mu_z, \Sigma_z) \right\}$$

$q^*(z)$ are "Soft Assignments"

- Suppose we observe n points: $X = (x_1, \ldots, x_n) \in \mathbf{R}^{n \times d}$.
- Let $z_1, \ldots, z_n \in \{1, \ldots, k\}$ be corresponding hidden variables.
- Optimal distribution q^* is:

$$q^*(z) \quad = \quad p(z \mid x, \theta).$$

- Convenient to define the conditional distribution for z_i given x_i as

$$\begin{aligned} \gamma_i^j \quad &:= \quad p\left(z = j \mid x_i\right) \\ &= \quad \frac{\pi_j \mathcal{N}\left(x_i \mid \mu_j, \Sigma_j\right)}{\sum_{c=1}^k \pi_c \mathcal{N}\left(x_i \mid \mu_c, \Sigma_c\right)} \end{aligned}$$

Expectation Step

- The complete log-likelihood is

$$\begin{aligned}\log p(x, z \mid \theta) &= \sum_{i=1}^{n} \log\left[\pi_z \mathcal{N}(x_i \mid \mu_z, \Sigma_z)\right] \\ &= \sum_{i=1}^{n} \left(\log \pi_z + \underbrace{\log \mathcal{N}(x_i \mid \mu_z, \Sigma_z)}_{\text{simplifies nicely}} \right)\end{aligned}$$

- Take the expected complete log-likelihood w.r.t. q^*:

$$\begin{aligned}J(\theta) &= \sum_{z} q^*(z) \log p(x, z \mid \theta) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_i^j \left[\log \pi_j + \log \mathcal{N}(x_i \mid \mu_j, \Sigma_j)\right]\end{aligned}$$

Maximization Step

- Find θ^* maximizing $J(\theta)$:

$$\begin{aligned}
\mu_c^{\text{new}} &= \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c x_i \\
\Sigma_c^{\text{new}} &= \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c \left(x_i - \mu_{\text{MLE}}\right)\left(x_i - \mu_{\text{MLE}}\right)^T \\
\pi_c^{\text{new}} &= \frac{n_c}{n},
\end{aligned}$$

for each $c = 1, \ldots, k$.