EM Algorithm for Latent Variable Models

David S. Rosenberg

New York University

April 24, 2018
1 Latent Variable Models

2 EM Algorithm (and Variational Methods) – The Big Picture

3 Math Prerequisites

4 The ELBO: Family of Lower Bounds on $\log p(x | \theta)$

5 Does EM Work?

6 Variations on EM

7 Summer Homework: Gaussian Mixture Model (Hints)
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 $\theta \in \Theta$:

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

Definition

A latent variable model is a probability model for which certain variables are never observed.

e.g. The Gaussian mixture model is a latent variable model.
Suppose we observe some data \((x_1, \ldots, x_n)\).

To simplify notation, take \(x\) to represent the entire dataset

\[
x = (x_1, \ldots, x_n),
\]

and \(z\) to represent the corresponding unobserved variables

\[
z = (z_1, \ldots, z_n).
\]

An observation of \(x\) is called an **incomplete data set**.
An observation \((x, z)\) is called a **complete data set**.
Our Objectives

- **Learning problem**: Given incomplete dataset \( x \), find MLE
  \[
  \hat{\theta} = \arg \max_{\theta} p(x \mid \theta).
  \]

- **Inference problem**: Given \( x \), find conditional distribution over \( z \):
  \[
  p(z \mid x, \theta).
  \]

- For Gaussian mixture model, learning is hard, inference is easy.
- For more complicated models, inference can also be hard. (See DSGA-1005)
Log-Likelihood and Terminology

- Note that
  \[
  \arg\max_{\theta} p(x \mid \theta) = \arg\max_{\theta} \left[ \log p(x \mid \theta) \right].
  \]
- Often easier to work with this “log-likelihood”.
- We often call \( p(x) \) the **marginal likelihood**, because it is \( p(x, z) \) with \( z \) “marginalized out”:
  \[
  p(x) = \sum_z p(x, z)
  \]
- We often call \( p(x, z) \) the **joint**. (for “joint distribution”)
- Similarly, \( \log p(x) \) is the **marginal log-likelihood**.
EM Algorithm (and Variational Methods) – The Big Picture
Want to find $\theta$ by maximizing the likelihood of the observed data $x$:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \log p(x | \theta)$$

Unfortunately this may be hard to do directly.

Approach: Generate a family of lower bounds on $\theta \mapsto \log p(x | \theta)$.

For every $q \in \Omega$, we will have a lower bound:

$$\log p(x | \theta) \geq \mathcal{L}_q(\theta) \quad \forall \theta \in \Theta$$

We will try to find the maximum over all lower bounds:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \left[ \sup_{q \in \Omega} \mathcal{L}_q(\theta) \right]$$
The Marginal Log-Likelihood Function

\[ \log p(x|\theta) \]
The Maximum Likelihood Estimator

\[ \theta^* = \arg \max_{\theta} \left[ \log p(x|\theta) \right] \]
Lower Bounds on Marginal Log-Likelihood

\[ \log p(x | \theta) \]

\[ \mathcal{L}_{q_1}(\theta) \]
\[ \mathcal{L}_{q_2}(\theta) \]
\[ \mathcal{L}_{q_3}(\theta) \]
\[ \mathcal{L}_{q_4}(\theta) \]
Supremum over Lower Bounds is a Lower Bound

\[ \log p(x|\theta) \]

\[ \sup_{\theta} \mathcal{L}_q(\theta) \]

\[ \mathcal{L}_{q_1}(\theta) \]

\[ \mathcal{L}_{q_2}(\theta) \]

\[ \mathcal{L}_{q_3}(\theta) \]

\[ \mathcal{L}_{q_4}(\theta) \]
Parameter Estimate: Max over all lower bounds

\[ \hat{\Theta} = \arg\max_{\Theta} \left[ \sup_{q} \mathcal{L}_q(\Theta) \right] \]
The Expected Complete Data Log-Likelihood

- Marginal log-likelihood is hard to optimize:
  \[
  \max_{\theta} \log p(x | \theta)
  \]

- Typically the complete data log-likelihood is easy to optimize:
  \[
  \max_{\theta} \log p(x, z | \theta)
  \]

- What if we had a distribution \( q(z) \) for the latent variables \( z \)?
The Expected Complete Data Log-Likelihood

- Suppose we have a distribution \( q(z) \) on latent variable \( z \).
- Then maximize the expected complete data log-likelihood:

\[
\max_{\theta} \sum_z q(z) \log p(x, z | \theta)
\]

- If \( q \) puts lots of weight on actual \( z \), this could be a good approximation to MLE
- EM assumes this maximization is relatively easy.
- (This is true for GMM.)
Math Prerequisites
Jensen’s Inequality

Theorem (Jensen’s Inequality)

If \( f: \mathbb{R} \rightarrow \mathbb{R} \) is a convex function, and \( x \) is a random variable, then

\[
E f(x) \geq f(E x).
\]

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

- e.g. \( f(x) = x^2 \) is convex. So \( E x^2 \geq (E x)^2 \). Thus

\[
\text{Var}(x) = E x^2 - (E x)^2 \geq 0.
\]
Kullback-Leibler Divergence

- Let \( p(x) \) and \( q(x) \) be probability mass functions (PMFs) on \( X \).
- How can we measure how “different” \( p \) and \( q \) are?

- The **Kullback-Leibler** or “KL” **Divergence** is defined by

\[
KL(p∥q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}.
\]

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

- Can also write this as

\[
KL(p∥q) = \mathbb{E}_{x \sim p} \log \frac{p(x)}{q(x)}.
\]
Gibbs Inequality \((\text{KL}(p\|q) \geq 0 \text{ and } \text{KL}(p\|p) = 0)\)

Theorem (Gibbs Inequality)

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

\[
\text{KL}(p\|q) \geq 0,
\]

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

\[
\text{KL}(p\|q) = \mathbb{E}_p \left[ -\log \left( \frac{q(x)}{p(x)} \right) \right] \\
\geq -\log \left[ \mathbb{E}_p \left( \frac{q(x)}{p(x)} \right) \right] \quad \text{(Jensen’s)} \\
= -\log \left[ \sum_{\{x|p(x)>0\}} p(x) \frac{q(x)}{p(x)} \right] \\
= -\log \left[ \sum_{x\in X} q(x) \right] \\
= -\log 1 = 0.
\]

Since \(-\log\) is strictly convex, we have strict equality iff \(q(x)/p(x)\) is a constant, which implies \(q = p\).
The ELBO: Family of Lower Bounds on $\log p(x | \theta)$
Lower Bound for Marginal Log-Likelihood

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

$$\log p(x | \theta) = \log \left[ \sum_z p(x, z | \theta) \right]$$

$$= \log \left[ \sum_z q(z) \left( \frac{p(x, z | \theta)}{q(z)} \right) \right] \quad \text{(log of an expectation)}$$

$$\geq \sum_z q(z) \log \left( \frac{p(x, z | \theta)}{q(z)} \right) \quad \text{(expectation of log)}$$

- Inequality is by Jensen’s, by concavity of the log.

This inequality is the basis for “variational methods”, of which EM is a basic example.
The ELBO

- For any PMF $q(z)$, we have a lower bound on the marginal log-likelihood

$$\log p(x | \theta) \geq \sum_z q(z) \log \left( \frac{p(x, z | \theta)}{q(z)} \right)$$

- Marginal log likelihood $\log p(x | \theta)$ also called the evidence.

- $\mathcal{L}(q, \theta)$ is the evidence lower bound, or “ELBO”.

In EM algorithm (and variational methods more generally), we maximize $\mathcal{L}(q, \theta)$ over $q$ and $\theta$. 
For any PMF $q(z)$, we have a lower bound on the marginal log-likelihood

$$\log p(x | \theta) \geq \mathcal{L}(q, \theta).$$

The MLE is defined as a maximum over $\theta$:

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \log p(x | \theta).$$

In EM algorithm, we maximize the lower bound (ELBO) over $\theta$ and $q$:

$$\hat{\theta}_{\text{EM}} \approx \arg\max_\theta \max_q \mathcal{L}(q, \theta).$$

In EM algorithm, $q$ ranges over all distributions on $z$. 
A Family of Lower Bounds

- For each $q$, we get a lower bound function: $\log p(x \mid \theta) \geq \mathcal{L}(q, \theta) \forall \theta$.
- Two lower bounds (blue and green curves), as functions of $\theta$:

Ideally, we’d find the maximum of the red curve. Maximum of green is close.

From Bishop’s *Pattern recognition and machine learning*, Figure 9.14.
Choose sequence of $q$'s and $\theta$'s by "coordinate ascent" on $\mathcal{L}(q, \theta)$.

**EM Algorithm (high level):**

1. Choose initial $\theta^{\text{old}}$.
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: $p(x | \theta^{\text{new}}) \geq p(x | \theta^{\text{old}})$

Get sequence of $\theta$'s with monotonically increasing likelihood.
EM: Coordinate Ascent on Lower Bound

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

From Bishop’s *Pattern recognition and machine learning*, Figure 9.14.
In EM algorithm, we need to repeatedly solve the following steps:

- \( \text{arg max}_q \mathcal{L}(q, \theta) \), for a given \( \theta \), and
- \( \text{arg max}_\theta \mathcal{L}(q, \theta) \), for a given \( q \).

We now give two re-expressions of ELBO \( \mathcal{L}(q, \theta) \) that make these easy to compute...
ELBO in Terms of KL Divergence and Entropy

Let’s investigate the lower bound:

\[ \mathcal{L}(q, \theta) = \sum_z q(z) \log \left( \frac{p(x, z | \theta)}{q(z)} \right) \]

\[ = \sum_z q(z) \log \left( \frac{p(z | x, \theta)p(x | \theta)}{q(z)} \right) \]

\[ = \sum_z q(z) \log \left( \frac{p(z | x, \theta)}{q(z)} \right) + \sum_z q(z) \log p(x | \theta) \]

\[ = -\text{KL}[q(z), p(z | x, \theta)] + \log p(x | \theta) \]

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

\[ \log p(x | \theta) = \mathcal{L}(q, \theta) + \text{KL}[q(z), p(z | x, \theta)] \]
Maximizing over $q$ for fixed $\theta$.

- Find $q$ maximizing

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

- Recall $\text{KL}(p\|q) \geq 0$, and $\text{KL}(p\|p) = 0$.

- Best $q$ is $q^*(z) = p(z \mid x, \theta)$ and

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

- Summary:

$$\log p(x \mid \theta) = \sup_q \mathcal{L}(q, \theta) \quad \forall \theta$$

- For any $\theta$, $\sup$ is attained at $q(z) = p(z \mid x, \theta)$. 
Marginal Log-Likelihood is the Supremum over Lower Bounds

$$\log p(x|\theta) = \sup_q \mathcal{L}(q, \theta)$$

sup is over all distributions on z
Suppose we find a maximum of $\mathcal{L}(q, \theta)$ over all distributions $q$ on $z$ and all $\theta \in \Theta$:

$$\mathcal{L}(q^*, \theta^*) = \sup_{\theta} \sup_{q} \mathcal{L}(q, \theta).$$

(where of course $q^*(z) = p(z | x, \theta^*)$.)

Claim: $\theta^*$ is a maximizes $\log p(x | \theta)$.

Proof: Trivial, since $\log p(x | \theta) = \sup_{q} \mathcal{L}(q, \theta)$. 
Summary: Maximizing over $q$ for fixed $\theta = \theta^{\text{old}}$.

- At given $\theta = \theta^{\text{old}}$, want to find $q$ giving best lower bound.
- Answer is $q^* = p(z | x, \theta^{\text{old}})$.
- This gives lower bound $\mathcal{L}(q^*, \theta)$ that is tight (equality) at $\theta^{\text{old}}$

$$\log p(x | \theta^{\text{old}}) = \mathcal{L}(q^*, \theta^{\text{old}}) \quad \text{(tangent at } \theta^{\text{old}}).$$

- And elsewhere, of course, $\mathcal{L}(q^*, \theta)$ is just a lower bound:

$$\log p(x | \theta) \geq \mathcal{L}(q^*, \theta) \quad \forall \theta$$
Tight lower bound for any chosen $\theta$

For $\theta^{\text{old}}$, take $q(z) = p(z \mid x, \theta^{\text{old}})$. Then

1. $\log p(x \mid \theta^{\text{old}}) = \mathcal{L}(q, \theta^{\text{old}})$. [Lower bound is tight at $\theta^{\text{old}}$.]

2. $\log p(x \mid \theta) \geq \mathcal{L}(q, \theta) \forall \theta$. [Global lower bound].

From Bishop’s *Pattern recognition and machine learning*, Figure 9.14.
Consider maximizing the lower bound $\mathcal{L}(q, \theta)$:

$$\mathcal{L}(q, \theta) = \sum_z q(z) \log \left( \frac{p(x, z | \theta)}{q(z)} \right)$$

$$= \sum_z q(z) \log p(x, z | \theta) - \sum_z q(z) \log q(z)$$

$\mathbb{E}[$complete data log-likelihood$] -$ no $\theta$ here

Maximizing $\mathcal{L}(q, \theta)$ equivalent to maximizing $\mathbb{E}[$complete data log-likelihood$]$ (for fixed $q$).
Choose initial $\theta^{\text{old}}$. 

2. **Expectation Step**
   - Let $q^*(z) = p(z \mid x, \theta^{\text{old}})$. [{$q^*$ gives best lower bound at $\theta^{\text{old}}$}]
   - Let 
     \[ J(\theta) := \mathcal{L}(q^*, \theta) = \sum_z q^*(z) \log \left( \frac{p(x, z \mid \theta)}{q^*(z)} \right) \]
     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.
Does EM Work?
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 $\theta^{\text{old}}$.

2. Choose $q^*(z) = \arg\max_q L(q, \theta^{\text{old}})$. We’ve shown

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

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

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

Putting it together, we get

$$\log p(x \mid \theta^{\text{new}}) \geq L(q^*, \theta^{\text{new}}) \quad \text{L is a lower bound}$$

$$\geq L(q^*, \theta^{\text{old}}) \quad \text{By definition of } \theta^{\text{new}}$$

$$= \log p(x \mid \theta^{\text{old}}) \quad \text{Bound is tight at } \theta^{\text{old}}.$$
Convergence of EM

- Let $\theta_n$ 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 $\ell(\theta) = \log p(x | \theta)$ 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$^3$, for any starting point $\theta_0$,

- $\lim_{n \to \infty} \theta_n = \theta^*$ for some stationary point $\theta^* \in S$ and
- $\theta^*$ 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 $\theta_n \equiv \theta^*$.

---

Variations on EM
EM Gives Us Two New Problems

- The “E” Step: Computing

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

- 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 \( \theta^{\text{new}} \) 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.
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 $Q$ that are easy to work with.

Lower bound now looser:

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

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

\[
p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}
\]

\[
\log p(\theta \mid x) = \log p(x \mid \theta) + \log p(\theta) - \log p(x)
\]

- 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

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

- Homework: Convince yourself our lower bound is still tight at \( \theta \).
Summer 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**

  Cluster probabilities: \( \pi = (\pi_1, \ldots, \pi_k) \)
  
  Cluster means: \( \mu = (\mu_1, \ldots, \mu_k) \)

  Cluster covariance matrices: \( \Sigma = (\Sigma_1, \ldots, \Sigma_k) \)

- Let \( \theta = (\pi, \mu, \Sigma) \).

- Marginal log-likelihood

  \[
  \log p(x \mid \theta) = \log \left\{ \sum_{z=1}^{k} \pi_z N(x \mid \mu_z, \Sigma_z) \right\}
  \]
\( q^*(z) \) are “Soft Assignments”

- Suppose we observe \( n \) points: \( X = (x_1, \ldots, x_n) \in \mathbb{R}^{n \times d} \).
- Let \( z_1, \ldots, z_n \in \{1, \ldots, k\} \) be corresponding hidden variables.
- Optimal distribution \( q^* \) is:
  \[
  q^*(z) = p(z \mid x, \theta).
  \]
- Convenient to define the conditional distribution for \( z_i \) given \( x_i \) as
  \[
  \gamma_i^j := p(z = j \mid x_i)
  = \frac{\pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}{\sum_{c=1}^{k} \pi_c \mathcal{N}(x_i \mid \mu_c, \Sigma_c)}
  \]
Expectation Step

- The complete log-likelihood is

\[
\log p(x, z | \theta) = \sum_{i=1}^{n} \log [\pi_z N(x_i | \mu_z, \Sigma_z)]
\]

\[
= \sum_{i=1}^{n} \left( \log \pi_z + \log N(x_i | \mu_z, \Sigma_z) \right)
\]

simplifies nicely

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

\[
J(\theta) = \sum_z q^*(z) \log p(x, z | \theta)
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_i^j \left[ \log \pi_j + \log N(x_i | \mu_j, \Sigma_j) \right]
\]
Maximization Step

- Find $\theta^*$ maximizing $J(\theta)$:

$$
\mu_c^{new} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c x_i
$$

$$
\Sigma_c^{new} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c (x_i - \mu_{MLE}) (x_i - \mu_{MLE})^T
$$

$$
\pi_c^{new} = \frac{n_c}{n},
$$

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