# Gaussian Mixture Model

He He
Slides based on Lecture 13b from David Rosenberg's course materials
(https://github.com/davidrosenberg/mlcourse)

CDS, NYU

April 27, 2021

# 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.

# Complete and Incomplete Data

- 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

$$\hat{\theta} = \underset{\theta}{\arg\max}\, p(x \mid \theta).$$

- **Inference problem**: Given $x$, find conditional distribution over $z$:

$$p\left(z \mid x, \theta\right).$$

- 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
$$\underset{\theta}{\arg\max}\, p(x \mid \theta) = \underset{\theta}{\arg\max}\, [\log p(x \mid \theta)] .$$
- 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

# Intuition

Problem: marginal log-likelihood $\log p(x;\theta)$ is hard to optimize (observing only $x$)

Observation: complete data log-likelihood $\log p(x,z;\theta)$ is easy to optimize (observing both $x$ and $z$)

Idea: guess a distribution of the latent variables $q(z)$ (soft assignments)

Maximize the **expected complete data log-likelihood**:

$$\max_{\theta} \sum_{z \in \mathcal{Z}} q(z) \log p(x,z;\theta)$$

EM assumption: the expected complete data log-likelihood is easy to optimize

Why should this work?

# Math Prerequisites

# 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

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

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

- Can also write this as

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

# 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**.

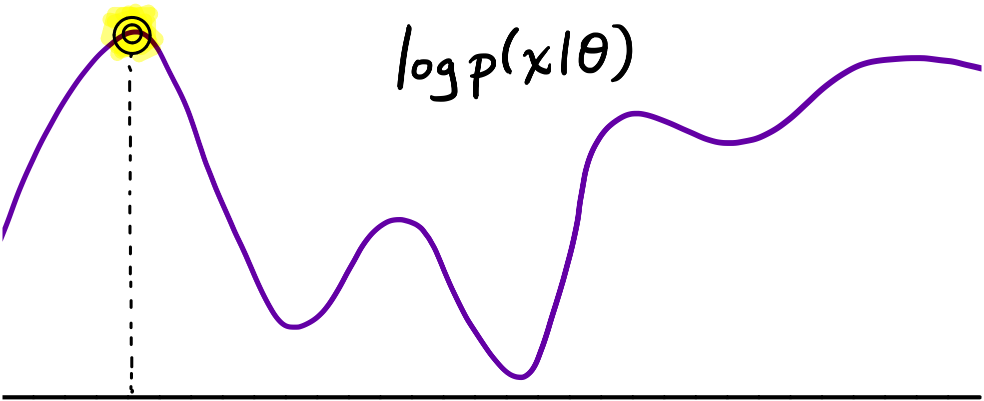
# Gibbs Inequality: Proof

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

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

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

# The Maximum Likelihood Estimator




$$\theta^* = \underset{\theta}{\operatorname{argmax}} \left[ \log p(x \mid \theta) \right]$$

# Lower bound of the marginal log-likelihood

$$
\begin{aligned}
\log p(x;\theta) &= \log \sum_{z\in\mathcal{Z}} p(x,z;\theta) \\
&= \log \sum_{z\in\mathcal{Z}} q(z)\frac{p(x,z;\theta)}{q(z)} \\
&\geqslant \sum_{z\in\mathcal{Z}} q(z)\log\frac{p(x,z;\theta)}{q(z)} \\
&\stackrel{\text{def}}{=} \mathcal{L}(q,\theta)
\end{aligned}
$$

- **Evidence**: $\log p(x;\theta)$
- **Evidence lower bound (ELBO)**: $\mathcal{L}(q,\theta)$
- $q$: chosen to be a family of tractable distributions
- Idea: *maximize the ELBO* instead of $\log p(x;\theta)$

# MLE, EM, and the ELBO

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

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\arg\max} \left[\log p(x \mid \theta)\right].$$

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

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

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

$$\hat{\theta}_{\text{EM}} \approx \underset{\theta}{\arg\max} \left[\max_{q} \mathcal{L}(q, \theta)\right]$$

- In EM algorithm, $q$ ranges over all distributions on $z$.

# EM: Coordinate Ascent on Lower Bound

- 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: $p(x \mid \theta^{\textbf{new}}) \geqslant p(x \mid \theta^{\textbf{old}})$
- **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 $\theta^{\text{old}} \Longrightarrow \mathcal{L}(q, \theta)$.
3. $\theta^{\text{new}} = \arg\max_{\theta} \mathcal{L}(q, \theta)$.

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

# Justification for maximizing ELBO

$$
\begin{aligned}
\mathcal{L}(q,\theta) &= \sum_{z\in\mathcal{Z}} q(z)\log\frac{p(x,z;\theta)}{q(z)} \\
&= \sum_{z\in\mathcal{Z}} q(z)\log\frac{p(z\mid x;\theta)p(x;\theta)}{q(z)} \\
&= -\sum_{z\in\mathcal{Z}} q(z)\log\frac{q(z)}{p(z\mid x;\theta)} + \sum_{z\in\mathcal{Z}} q(z)\log p(x;\theta) \\
&= -\mathrm{KL}\left(q(z)\|p(z\mid x;\theta)\right) + \underbrace{\log p(x;\theta)}_{\text{evidence}}
\end{aligned}
$$

- **KL divergence**: measures "distance" between two distributions (not symmetric!)
- $\mathrm{KL}\left(q\|p\right) \geqslant 0$ with equality iff $q(z) = p(z\mid x)$.
- ELBO = evidence - KL $\leqslant$ evidence

# [discussion]Justification for maximizing ELBO

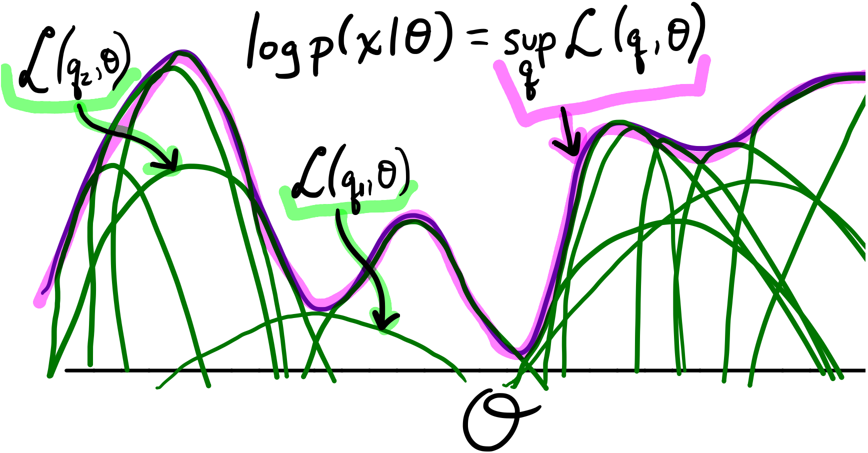
$\mathcal{L}(q,\theta) = -\mathrm{KL}\left(q(z) \| p(z \mid x;\theta)\right) + \log p(x;\theta)$

Fix $\theta = \theta_0$ and $\max_q \mathcal{L}(q,\theta_0)$: $q^* = p(z \mid x;\theta_0)$

Let $\theta^*, q^*$ be the global optimzer of $\mathcal{L}(q,\theta)$, then $\theta^*$ is the global optimizer of $\log p(x;\theta)$.
(Proof: exercise)

# Marginal Log-Likelihood **IS** the Supremum over Lower Bounds

**sup is over all distributions on z**



# Summary

**Latent variable models**: clustering, latent structure, missing lables etc.

Parameter estimation: maximum marginal log-likelihood

Challenge: directly maximize the **evidence** $\log p(x;\theta)$ is hard

Solution: maximize the **evidence lower bound**:

$$\text{ELBO} = \mathcal{L}(q,\theta) = -\text{KL}\left(q(z) \| p(z \mid x;\theta)\right) + \log p(x;\theta)$$

Why does it work?

$$\begin{aligned} q^*(z) &= p(z \mid x;\theta) \quad \forall \theta \in \Theta \\ \mathcal{L}(q^*,\theta^*) &= \max_\theta \log p(x;\theta) \end{aligned}$$

# EM algorithm

*Coordinate ascent on* $\mathcal{L}(q, \theta)$

1. Random initialization: $\theta^{\text{old}} \leftarrow \theta_0$
2. Repeat until convergence
   1. $q(z) \leftarrow \arg\max_q \mathcal{L}(q, \theta^{\text{old}})$

      **Expectation** (the E-step): $$q^*(z) = p(z \mid x; \theta^{\text{old}})$$ $$J(\theta) = \mathcal{L}(q^*, \theta)$$

   2. $\theta^{\text{new}} \leftarrow \arg\max_\theta \mathcal{L}(q^*, \theta)$

      **Maximization** (the M-step): $$\theta^{\text{new}} \leftarrow \arg\max_\theta J(\theta)$$

# EM Algorithm

1. **Expectation Step**
   - Let $q^*(z) = p(z \mid x, \theta^{\text{old}})$. [$q^*$ gives best lower bound at $\theta^{\text{old}}$]
   - Let

$$J(\theta) := \mathcal{L}(q^*, \theta) = \underbrace{\sum_z q^*(z) \log\left(\frac{p(x, z \mid \theta)}{q^*(z)}\right)}_{\textbf{expectation } \text{w.r.t. } z \sim q^*(z)}$$

2. **Maximization Step**

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

   [Equivalent to maximizing expected complete log-likelihood.]

EM puts no constraint on $q$ in the E-step and assumes the M-step is easy. In general, both steps can be hard.

# [discussion]Monotonically increasing likelihood

Exercise: prove that EM increases the marginal likelihood monotonically

$$\log p(x; \theta^{\text{new}}) \geqslant \log p(x; \theta^{\text{old}}) .$$

Does EM converge to a global maximum?

# 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

$$\theta^{\text{new}} = \underset{\theta}{\arg\max}\, J(\theta).$$

- 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.

# 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:

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