Deep Generative Models: Variational Inference

Fall Semester 2025

René Vidal

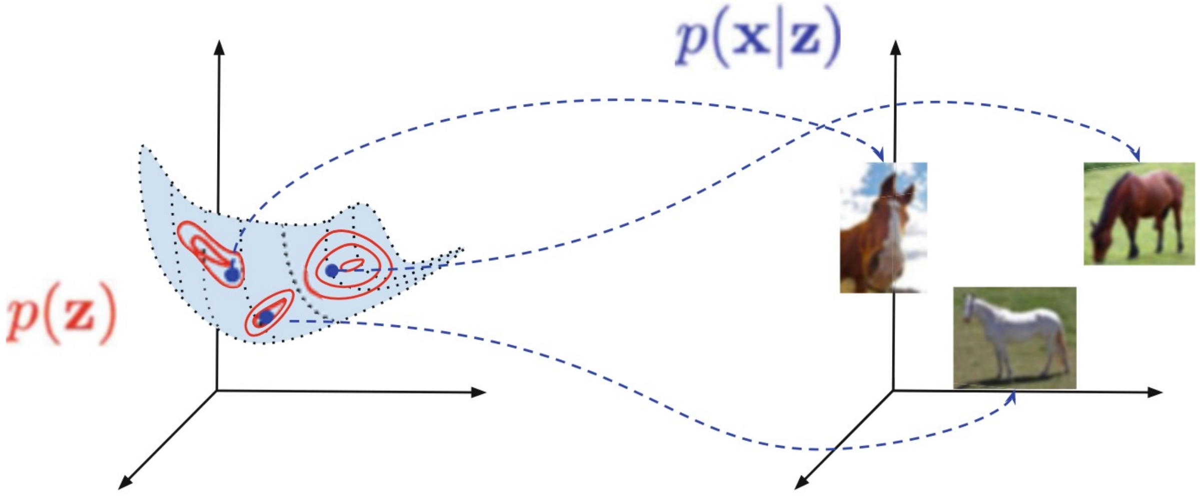
Director of the Center for Innovation in Data Engineering and Science (IDEAS),
Rachleff University Professor, University of Pennsylvania
Amazon Scholar & Chief Scientist at NORCE

Outline

- Latent Variable Models
  - Probabilistic PCA
  - Beyond PPCA?
- Variational Inference
  - Principle
  - Derivation
- Expectation Maximization
  - Derivation
  - EM for a Mixture of Gaussians

Latent Variable Models

- X = observed variable
- Z = latent variable



A latent variable model and a generative process. Note the low-dimensional manifold (here 2D) embedded in the high-dimensional space (here 3D)

- $\boldsymbol{z} \sim p(\boldsymbol{z})$
- $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{z})$

- Factorization of the joint model

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})$$

- Marginalization of the model

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

Probabilistic Principal Component Analysis (PPCA)

- Let $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon} \mid \boldsymbol{0}, \sigma^2 \boldsymbol{I})$ be independent.
- In PPCA, the dependency between $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{z} \in \mathbb{R}^d$, $d \ll D$, is defined by a linear Gaussian additive model

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\varepsilon}$$

- **Theorem.** Let $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_\mathrm{N}$ be, respectively, the ML estimates for the mean and the covariance of the data. Let U_1 be the matrix with the top d eigenvectors of $\boldsymbol{\Sigma}_\mathrm{N}$, Λ_1 be the matrix with the corresponding top d eigenvalues, and λ_i be the ith largest eigenvalue of $\boldsymbol{\Sigma}_\mathrm{N}$. The ML estimates for the PPCA parameters $(\boldsymbol{b}, \boldsymbol{W}, \sigma)$ is given by

$$\boldsymbol{b} = \boldsymbol{\mu}_N,\ \boldsymbol{W} = \boldsymbol{U}_1(\boldsymbol{\Lambda}_1 - \sigma^2 I)^{1/2} R \ \text{ and } \ \sigma^2 = \frac{1}{D-d} \sum_{i=d+1}^{D} \lambda_i$$

where $R \in \mathbb{R}^{d \times d}$ is an arbitrary orthogonal matrix.

What about Latent Variable Models other than PPCA

- We would like to learn the parameters of the model via Maximum Likelihood.
- Since $\boldsymbol{z}$ is latent, we need to marginalize $p_\theta(\boldsymbol{x}) = \int p_\theta(\, \boldsymbol{x} \mid \boldsymbol{z}\,) p(\boldsymbol{z}) d\boldsymbol{z}$, i.e.

$$\max_\theta \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}_i) = \max_\theta \sum_{i=1}^{N} \log \int p_\theta(\boldsymbol{x}_i \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}$$

- For PPCA, we could compute $p(\boldsymbol{x})$ in closed form and solve for θ analytically.
- In general, we need many samples of z for each x_i to approximate the integral.

$$\max_\theta \sum_{i=1}^{N} \log \sum_{j} p_\theta\left(\, \boldsymbol{x}_i \mid \boldsymbol{z}_j \,\right)$$

- We address this challenge using **Variational Inference**, which we describe next.

Variational Inference

- **Old ML learning objective**: $\max_\theta \sum_{i=1}^{N} \log p_\theta(x_i) = \sum_{i=1}^{N} \log \int p_\theta(x_i \mid z) p(z) dz$

- **Theorem**: the log likelihood can be written as

$$\log p_\theta(\boldsymbol{x}) = \max_{q(\cdot|\boldsymbol{x}):q(\cdot|x)\geq 0, \int q(z|x)dz=1} \int q(z \mid x) \log \frac{p_\theta(x,z)}{q(z|x)} dz.$$

and the maximizing distribution is given by $q^*(\boldsymbol{z} \mid \boldsymbol{x}) = p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$

- **New ML learning objective**:

$$\max_\theta \max_{q(\cdot|\boldsymbol{x}_i),\forall i} \sum_{i=1}^{N} \int q(\boldsymbol{z} \mid \boldsymbol{x}_i) \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x}_i)} dz$$

- Before going through the derivation, what is the gain here?

Variational Inference

- **New ML learning objective**:

$$\max_{\theta} \max_{q(\cdot|\boldsymbol{x}_i),\forall i} \sum_{i=1}^{N} \int q(\boldsymbol{z} \mid \boldsymbol{x}_i) \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x}_i)} d\boldsymbol{z}$$

- **Expectation Maximization**:
  - If finding q^* given θ is easy, then we can alternate between finding q given θ and vice versa
  - Here $q^*(\boldsymbol{z} \mid \boldsymbol{x}) = p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$. Thus, if the integral w.r.t. $\boldsymbol{z}$ is easy to evaluate for a fixed θ, we can alternate between computing the integral (E-step) and maximizing w.r.t. θ (M-step).

- **Variational AutoEncoders:** parameterize $q(\cdot \mid x_i)$ with a NN with parameters ψ that takes x_i and outputs a distribution $q_\psi(\cdot \mid x_i)$, and find (θ, ψ) via SGD

$$\max_{\theta} \max_{\psi} \sum_{i=1}^{N} \int q_\psi(\boldsymbol{z} \mid \boldsymbol{x}_i) \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z})}{q_\psi(\boldsymbol{z} \mid \boldsymbol{x}_i)} d\boldsymbol{z}$$

Variational Inference

- **New ML learning objective**:

$$\max_{\theta} \max_{q(\cdot|\boldsymbol{x}_i),\forall i} \sum_{i=1}^{N} \int q(\boldsymbol{z} \mid \boldsymbol{x}_i) \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x}_i)} d\boldsymbol{z}$$

- We will use VI for many latent variable models
  - Mixtures of Gaussians (a.k.a. Gaussian Mixture Models) -> EM
  - Probabilistic Principal Component Analysis (PPCA) -> EM
  - Mixtures of PPCA -> EM
  - Variational Auto-Encoders (VAE) -> VI
  - Diffusion models -> VI
  - …

Variational Inference: Derivation

- Proof: Let $q(z|x)$ be the variational distribution. Observe that

$$\log p_\theta(x) = \int q(\boldsymbol{z} \mid \boldsymbol{x}) \log p_\theta(\boldsymbol{x}) dz = \int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} dz$$

$$= \int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} \frac{q(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} dz$$

$$= \underbrace{\boxed{\int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} dz}}_{\text{Evidence Lower Bound (ELBO)}} + \underbrace{\boxed{\int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{q(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} dz}}_{\mathrm{KL}[q(z|x) \,||\, p_\theta(z|x)]}$$

- Thus, $\max_{q(\cdot|X)} \int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} dz = \max_{q(\cdot|X)} \log p_\theta(\boldsymbol{x}) - \mathrm{KL}[q(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p_\theta(\boldsymbol{z} \mid \boldsymbol{x})]$

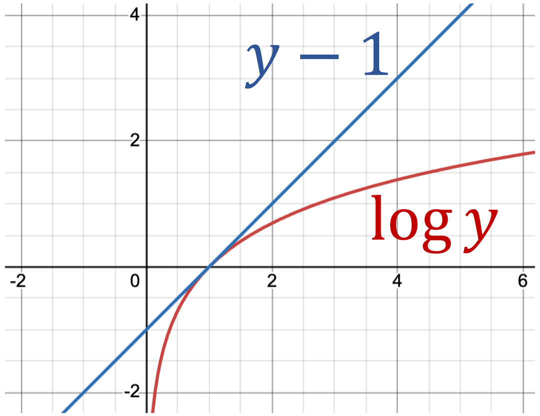
$$= \log p_\theta(\boldsymbol{x}) - \min_{q(\cdot|X)} \mathrm{KL}[q(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p_\theta(\boldsymbol{z} \mid \boldsymbol{x})] = \log p_\theta(\boldsymbol{x})$$

- The last step follows because $\mathrm{KL}(q \,||\, p) \geq 0$ and $KL(q \,||\, p) = 0$ iff $p = q$.

Variational Inference: Derivation

- **Proposition**: $\mathrm{KL}(q \,||\, p) \geq 0$. Further, $\mathrm{KL}(q \,||\, p) = 0$ if and only if $p = q$.

- **Lemma**: $\log y \leq y - 1$, equality holds if and only if $y = 1$
  - This is by the concavity of $\log(\cdot)$



- **Proof of the Proposition**: $\mathrm{KL}(q, p) = \int q(x) \log \frac{q(x)}{p(x)} dx$

$$= -\int q(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}$$

$$\geq -\int q(\boldsymbol{x}) \left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - 1 \right) d\boldsymbol{x} \qquad \text{(by the lemma)}$$

$$= -\int p(\boldsymbol{x}) d\boldsymbol{x} + \int q(\boldsymbol{x})\, d\boldsymbol{x}$$

$$= -1 + 1 = 0$$

Deep Generative Models: Expectation Maximization

Fall Semester 2025

René Vidal

Director of the Center for Innovation in Data Engineering and Science (IDEAS),
Rachleff University Professor, University of Pennsylvania
Amazon Scholar & Chief Scientist at NORCE

Outline

- Latent Variable Models
  - Probabilistic PCA
  - Beyond PPCA?
- Variational Inference
  - Principle
  - Derivation
- Expectation Maximization
  - Derivation
  - EM for a Mixture of Gaussians

Variational Inference

- **Theorem**: the log likelihood criterion can be written in variational form as

$$\max_{\theta} \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}_i) \equiv \max_{\theta} \max_{q(\cdot|\boldsymbol{x}_i),\forall i} \sum_{i=1}^{N} \int q(\boldsymbol{z} \mid \boldsymbol{x}_i) \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x}_i)} d\boldsymbol{z}$$

and the maximizing distribution is given by $q^*(\boldsymbol{z} \mid \boldsymbol{x}) = p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$

- **Expectation Maximization**: since we know q^*, if the integral w.r.t. $\boldsymbol{z}$ is easy to compute for a fixed θ, we can alternate between computing the integral (E-step) and maximizing w.r.t. θ (M-step).

- **Variational AutoEncoders:** parameterize $q(\cdot \mid x_i)$ with a NN with parameters ψ that takes x_i and outputs a distribution $q_\psi(\cdot \mid x_i)$, and find (θ, ψ) via SGD

$$\max_{\theta} \max_{\psi} \sum_{i=1}^{N} \int q_\psi(z|x_i) \log \frac{p_\theta(x_i, z)}{q_\psi(z|x_i)} dz$$

Expectation Maximization

- ML objective

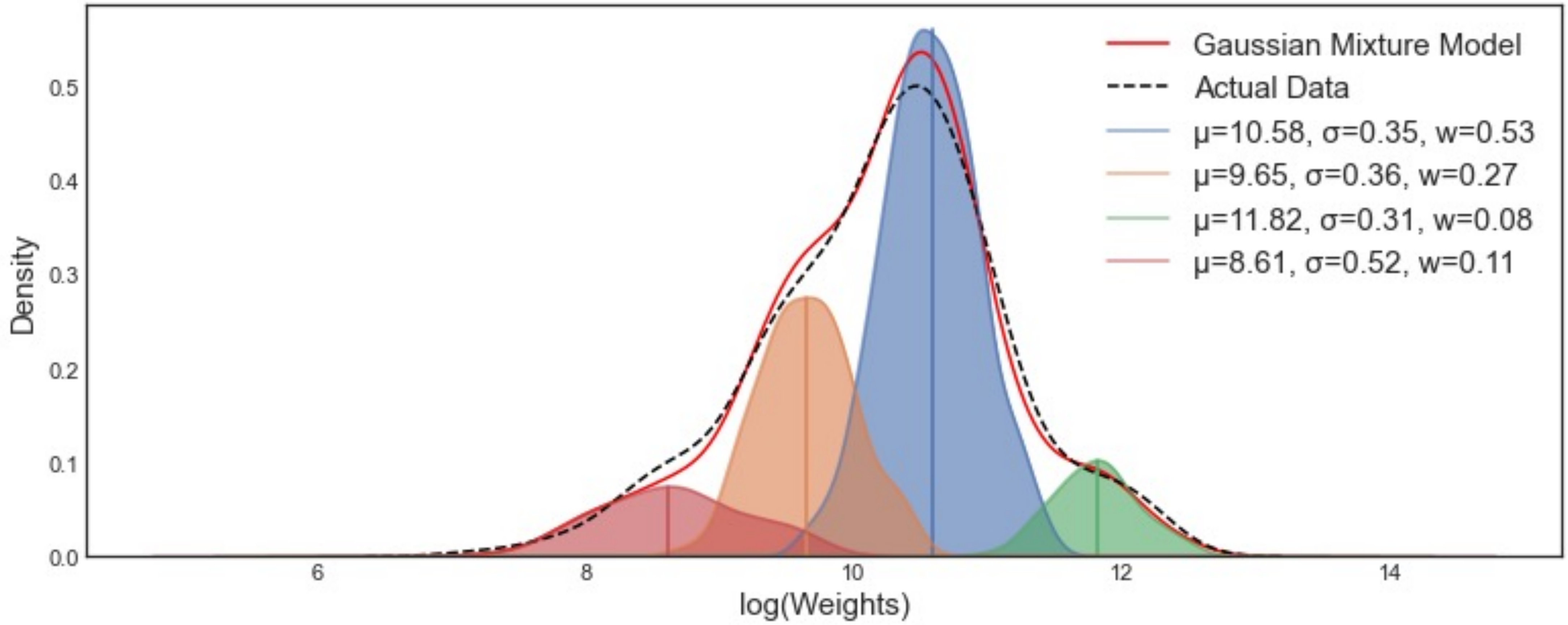
$$\max_{\theta} \sum_{i=1}^{N} \log p_{\theta}(\boldsymbol{x}_i) \equiv \max_{\theta} \max_{q(\cdot|\boldsymbol{x}_i),\forall i} \sum_{i=1}^{N} \int q(\boldsymbol{z} \mid \boldsymbol{x}_i) \log \frac{p_{\theta}(\boldsymbol{x}_i, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x}_i)} d\boldsymbol{z}$$

- Expectation Maximization alternates between two steps (k: iteration)
    - **E-step**: $q^k(\boldsymbol{z} \mid \boldsymbol{x}_i) = p_{\theta_k}(\boldsymbol{z} \mid \boldsymbol{x}_i)$ maximize w.r.t. q with θ fixed
    - **E-step**: $Q(\theta \mid \theta_k) = \sum_{i=1}^{N} \int_{z} q^k(\boldsymbol{z} \mid \boldsymbol{x}_i) \log p_{\theta}(\boldsymbol{x}_i, \boldsymbol{z})\, dz$ integrate w.r.t. q with θ fixed
    - **M-step**: $\theta^{k+1} = \mathrm{argmax}_{\theta}\, Q(\theta \mid \theta_k)$ maximize w.r.t. θ with q fixed

- Examples
    - For PPCA and for a mixture of Gaussians, E & M steps are closed-form (HW1, next slide).
    - E-step often done by sampling (MCMC) and M-step often done by optimization (SGD).

Example: EM for Gaussian Mixture Model (GMM)

- **GMM**: $p_\theta(\boldsymbol{x}) = \sum_{\mathrm{j}=1}^{n} p(x \mid z=j)p(z=j) = \pi_1 p_{\theta_1}(\boldsymbol{x}) + \cdots + \pi_n p_{\theta_n}(\boldsymbol{x})$
  - $\pi_j > 0$: prior probability of drawing a point from the j-th model; $\sum_{\mathrm{j}=1}^{n} \pi_j = 1$
  - $p_{\theta_j}(x) = \mathcal{N}(x; \mu_j, \Sigma_\mathrm{j})$, $\theta_i = (\mu_j, \Sigma_\mathrm{j})$ are the mean and covariance of the j-th Gaussian
  - $\theta = (\theta_1, \ldots, \theta_n, \pi_1, \ldots, \pi_n)$ are the parameters of the mixture model



Example: EM for Gaussian Mixture Model (GMM)

- **GMM**: $p_\theta(\boldsymbol{x}) = \sum_{\mathrm{j}=1}^{n} p(x \mid z = j)p(z = j) = \pi_1 p_{\theta_1}(\boldsymbol{x}) + \cdots + \pi_n p_{\theta_n}(\boldsymbol{x})$
  - $\pi_j > 0$: prior probability of drawing a point from the j-th model; $\sum_{\mathrm{j}=1}^{n} \pi_j = 1$
  - $p_{\theta_j}(x) = \mathcal{N}(x; \mu_j, \Sigma_{\mathrm{j}})$, $\theta_i = (\mu_j, \Sigma_{\mathrm{j}})$ are the mean and covariance of the j-th Gaussian
  - $\theta = (\theta_1, \dots, \theta_n, \pi_1, \dots, \pi_n)$ are the parameters of the mixture model

- **Goal**: estimate θ from N i.i.d. samples $x_1, \dots, x_N$ from p_θ using EM
  - For $i = 1, \dots N$, let $z_i = j$ if x_i belongs to class j , and let $q_{ij} = q(z_i = j \mid x_i)$

$$\max_{\theta} \sum_{i=1}^{N} \log p_\theta(x_i) = \max_{\theta} \max_{q_{ij} \geq 0 : \sum_j q_{ij} = 1, \forall i} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij} \log \frac{p_\theta(x_i, z_i = j\,)}{q_{ij}}$$

- **E-step**: compute $q_{ij}^k = p_{\theta^k}(\boldsymbol{z}_i = j \mid \boldsymbol{x}_i) = \dfrac{p_{\theta^k}(\boldsymbol{x}_i | z_i = j) p_{\theta^k}(\boldsymbol{z}_i = j)}{p_{\theta^k}(\boldsymbol{x}_i)} = \boxed{\dfrac{p_{\theta_j^k}(\boldsymbol{x}_i)\pi_j^k}{\sum_{\mathrm{j}=1}^{n} p_{\theta_j^k}(\boldsymbol{x}_i)\pi_j^k}}$

Example: EM for Gaussian Mixture Model (GMM)

- **GMM**: $p_\theta(\boldsymbol{x}) = \sum_{\mathrm{j}=1}^{n} p(x \mid z=j)p(z=j) = \pi_1 p_{\theta_1}(\boldsymbol{x}) + \cdots + \pi_n p_{\theta_n}(\boldsymbol{x})$
  - $\pi_j > 0$: prior probability of drawing a point from the j-th model; $\sum_{\mathrm{j}=1}^{n} \pi_j = 1$
  - $p_{\theta_j}(x) = \mathcal{N}(x; \mu_j, \Sigma_\mathrm{j})$, $\theta_i = (\mu_j, \Sigma_\mathrm{j})$ are the mean and covariance of the j-th Gaussian
  - $\theta = (\theta_1, \dots, \theta_n, \pi_1, \dots, \pi_n)$ are the parameters of the mixture model

- **M-step**:

$$\theta^{k+1} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^k \log \frac{p_\theta(x_i, z_i = j)}{q_{ij}^k} = \underset{\{\theta_j\}_{j=1}^n, \{\pi_j\}_{j=1}^n}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^k \log\left(\pi_j \cdot p_{\theta_j}(\boldsymbol{x}_\mathrm{i})\right)$$

  - For π_j's: $\max_{\{\pi_j\}_{j=1}^n : \sum_{\mathrm{j}=1}^{n} \pi_j = 1} \sum_{\mathrm{i}=1}^{N} q_{ij}^k \log(\pi_j)$. The solution is given by $\boxed{\pi_j^{k+1} = \frac{\sum_{\mathrm{i}=1}^{N} q_{ij}^k}{\sum_{\mathrm{i}=1}^{N} \sum_{\mathrm{j}=1}^{n} q_{ij}^k}}$

  - For θ_j's: $\max_{\mu_j, \Sigma_\mathrm{j}} \sum_{\mathrm{i}=1}^{N} q_{ij}^k \left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j) - \frac{1}{2}\operatorname{logdet}(\Sigma_j)\right)$. The solution is

$$\boxed{\boldsymbol{\mu}_j^{k+1} = \frac{\sum_{\mathrm{i}=1}^{N} q_{ij}^k \boldsymbol{x}_i}{\sum_{\mathrm{i}=1}^{N} q_{ij}^k}} \text{ and } \boxed{\Sigma_j^{k+1} = \frac{\sum_{i=1}^{N} q_{ij}^k \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{k+1}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{k+1}\right)^\top}{\sum_{i=1}^{N} q_{ij}^k}}$$

E.g.: EM for Gaussian Mixture Model

- **M-step**: $\theta^{k+1} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k} \log \frac{p_\theta(x_i, z_i = j)}{q_{ij}^{k}} = \underset{\{\theta_j\}_{j=1}^{n}, \{\pi_j\}_{j=1}^{n}}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k} \log\left(\pi_j \cdot p_{\theta_j}(\boldsymbol{x}_{\mathrm{i}})\right)$

  - For π_j's: $\max_{\{\pi_j\}_{j=1}^{n} : \sum_{\mathrm{j}=1}^{n} \pi_j = 1} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k} \log(\pi_j)$

  - To analyze the solution of a constrained optimization problem, we use the method of Lagrange multipliers:

$$L\left(\{\pi_j\}_{j}, \lambda\right) = \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k} \log(\pi_j) - \lambda\left(\sum_{\mathrm{j}=1}^{n} \pi_j = 1\right)$$

  - $\frac{\partial L}{\partial \lambda} = 0 \Leftrightarrow \sum_{\mathrm{j}=1}^{n} \pi_j = 1$
  - $\forall j$: $\frac{\partial L}{\partial \pi_j} = 0 \Leftrightarrow \sum_{i=1}^{N} q_{ij}^{k} \frac{1}{\pi_j} - \lambda = 0$

  - The solution is $\lambda = \sum_{\mathrm{i}=1}^{N} \sum_{\mathrm{j}=1}^{n} q_{ij}^{k}$, $\pi_j = \frac{\sum_{\mathrm{i}=1}^{N} q_{ij}^{k}}{\sum_{\mathrm{i}=1}^{N} \sum_{\mathrm{j}=1}^{n} q_{ij}^{k}} =: \pi_j^{k+1}$

E.g.: EM for Gaussian Mixture Model

- **M-step**: $\theta^{k+1} = \underset{\theta}{\text{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k} \log \frac{p_\theta(x_i, z_i = j)}{q_{ij}^{k}} = \underset{\theta}{\text{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k} \log p_\theta(x_i, z_i = j)$
  - If we spell out θ and $p_\theta(x_i, z_i = j)$, the above becomes $\underset{\{\theta_j\}_{j=1}^{n}, \{\pi_j\}_{j=1}^{n}}{\text{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k} \log \left(\pi_j \cdot p_{\theta_j}(\boldsymbol{x}_\mathrm{i})\right)$
  - For θ_j's: $\max_{\mu_j, \Sigma_\mathrm{j}} \boxed{\sum_{\mathrm{i}=1}^{N} q_{ij}^{k} \left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j) - \frac{1}{2} \operatorname{logdet}(\Sigma_j)\right)} =: \boldsymbol{\mathcal{L}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$
  - $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_j} = \sum_{i=1}^{N} q_{ij}^{k} \Sigma_j^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j) = \Sigma_j^{-1} \left(\sum_{i=1}^{N} q_{ij}^{k} \boldsymbol{x}_i - \left(\sum_{i=1}^{N} q_{ij}^{k}\right) \boldsymbol{\mu}_j\right).$
  - Setting it to 0 $\Rightarrow \boldsymbol{\mu}_j = \frac{\sum_{\mathrm{i}=1}^{N} q_{ij}^{k} \boldsymbol{x}_i}{\sum_{\mathrm{i}=1}^{N} q_{ij}^{k}} =: \boldsymbol{\mu}_j^{k+1}$
  - $\boldsymbol{\mathcal{L}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = -\frac{1}{2} \operatorname{tr}\left(\Sigma_j^{-1} \sum_{i=1}^{N} q_{ij}^{k} (x_i - \mu_j)(x_i - \mu_j)^\top\right) - \frac{1}{2} \left(\sum_{i=1}^{N} q_{ij}^{k}\right) \operatorname{logdet} \Sigma_j$
  - Reusing our derivation in MLE for Gaussian: $\Sigma_j = \frac{\sum_{i=1}^{N} q_{ij}^{k} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{k+1}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{k+1}\right)^\top}{\sum_{i=1}^{N} q_{ij}^{k}} =: \Sigma_j^{k+1}$