Deep Generative Models: Latent Variable Models

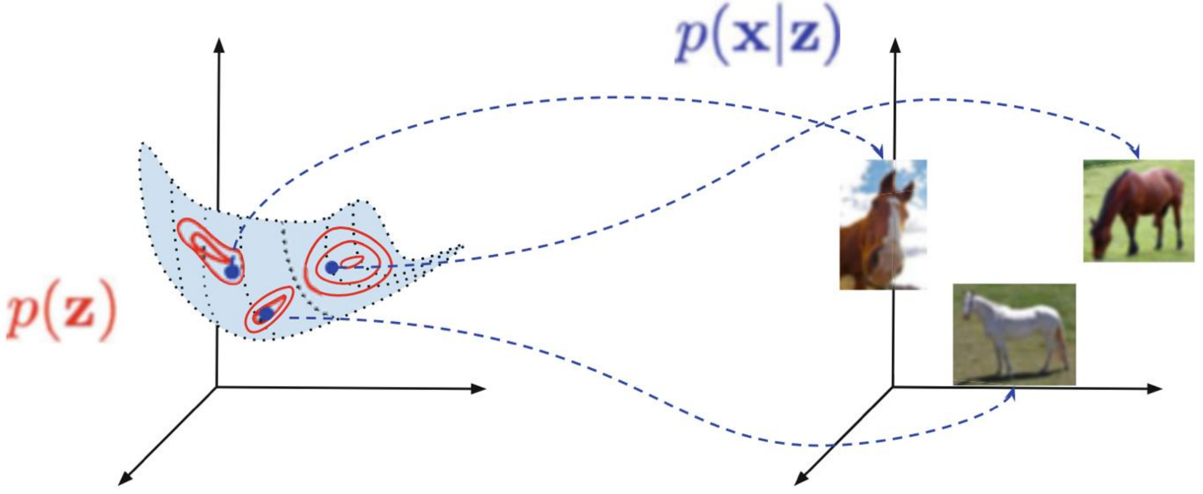
Fall Semester 2024

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

Latent Variable Models

- X = observed variable
- Z = latent variable

- $\mathbf{z} \sim p(\mathbf{z})$
- $\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})$



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

- Factorization of the joint model

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

- Marginalization of the model

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

Latent Variable Models

- Latent Variable Model $p(x, z) = p(z)p(x \mid z)$
- To sample $p(x, z)$, we have to first
  - Sample $p(z)$
  - Then sample $p(x \mid z)$
- How to learn the parameters θ of latent variable models?
  - Let's try directly applying maximum log likelihood

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

need many samples of z for each x_i to approximate this integral when dimension is high

  - **Variational Inference** is our best friend here, which we will describe next

Variational Inference

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

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

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

and the maximizing distribution is given by $p_\theta(z|x)$

- **New ML learning objective**:

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

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

Variational Inference

- **New ML learning objective**:

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

- If $p_\theta(z|x)$ is "accessible", then we can alternate between optimizing w.r.t. θ with $q(\cdot\ |x_i)$'s fixed and vice versa, leading to the **Expectation Maximization** algorithm
  - Promise: in many cases we will get closed form solutions in each step

- Else, parameterize $q(\cdot\ |x_i)$ with a NN that takes x_i and outputs a distribution $q_\phi(\cdot\ |x_i)$, where ϕ contains the parameters of the NN
  - Promise: the output posterior typically has a small variance => MC is a good approximation
  - Finding ϕ will be done by gradient descent

Variational Inference

- **New ML learning objective**:

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

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

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

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

Evidence **L**ower **Bo**und (ELBO) $\mathrm{KL}[q(z|x)\ ||\ p_\theta(z|x)]$

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

- To complete the argument, it suffices to show that

$$\min_{q:q(z)\geq 0, \int q(z)dz=1} \mathrm{KL}[q(z|x)\ ||\ p_\theta(z|x)] = 0$$

- Needs to dive a bit into optimization: first-order optimality conditions

Expectation Maximization

$$\max_{\theta} \sum_{i=1}^{N} \log p_\theta(x_i) = \max_{\theta} \max_{q(z|x_i),\forall i} \sum_{i=1}^{N} \int_z q(z|x_i) \log \frac{p_\theta(x_i, z)}{q(z|x_i)} dz$$

- Expectation Maximization alternates between two steps (k: iteration)

- E-step: $q^k(z|x_i) = p_{\theta_k}(z|x_i)$ maximizing w.r.t. w with θ fixed
- M-step: $\theta_{k+1} = \mathrm{argmax}_\theta \sum_{i=1}^{N} \int_z q^k(z|x_i) \log p_\theta(x_i, z)\, dz$

  maximizing w.r.t. θ with w fixed

- Examples
  - For a mixture of Gaussians, E & M steps are closed-form (next slide)
  - Often E-step can be done by sampling (MCMC) and M-step can be done by optimization (SGD)

E.g.: EM for Gaussian Mixture Model

- Consider a mixture of Gaussians $p_\theta(\boldsymbol{x}) = \pi_1 p_{\theta_1}(\boldsymbol{x}) + \pi_2 p_{\theta_2}(\boldsymbol{x}) + \cdots + \pi_k p_{\theta_k}(\boldsymbol{x})$
  - $\pi_i > 0$: prior probability of drawing a point from the i-th model; $\sum_{i=1}^{k} \pi_i = 1$
  - $p_{\theta_i} = \mathcal{N}(\mu_i, \Sigma_\mathrm{i})$. $\theta_i = (\mu_i, \Sigma_\mathrm{i})$: mean and covariance of the i-th Gaussian distribution
  - $\theta = (\theta_1, \ldots, \theta_k, \pi_1, \ldots, \pi_k)$: the parameters of the mixture model
- Goal: estimate θ from N i.i.d. samples $x_1, \ldots, x_N$ from p_θ using EM
- E-step: compute $q_{ij}^k = p_{\theta^k}\left(\boldsymbol{z}_j = i \mid \boldsymbol{x}_j\right) = \frac{p_{\theta^k}(\boldsymbol{x}_j | z_j = i) p_{\theta^k}(\boldsymbol{z}_j = i)}{p_{\theta^k}(\boldsymbol{x}_j)} = \frac{p_{\theta_i^k}(\boldsymbol{x}_j) \pi_i^k}{\sum_{i=1}^{n} p_{\theta_i^k}(\boldsymbol{x}_j) \pi_i^k}$
- M-step:
  - $\pi_i^{k+1} = \arg\max_{\pi_i} \sum_{j=1}^{N} q_{ij}^k \log(\pi_i) = \frac{\sum_{j=1}^{N} q_{ij}^k}{\sum_{j=1}^{N} \sum_{i=1}^{n} q_{ij}^k}$
  - $\theta_i^{k+1} = \arg\max_{\theta_i} \sum_{j=1}^{N} q_{ij}^k \left(-\frac{1}{2}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^\top \Sigma_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right) - \frac{1}{2} \operatorname{logdet}(\Sigma_i)\right)$
    - $\boldsymbol{\mu}_i^{k+1} = \frac{\sum_{j=1}^{N} q_{ij}^k \boldsymbol{x}_j}{\sum_{j=1}^{N} q_{ij}^k}$ and $\Sigma_i^{k+1} = \frac{\sum_{j=1}^{N} q_{ij}^k \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i^{k+1}\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i^{k+1}\right)^\top}{\sum_{j=1}^{N} q_{ij}^k}$