

# Deep Generative Models: Variational Auto Encoders

Fall Semester 2025

René Vidal

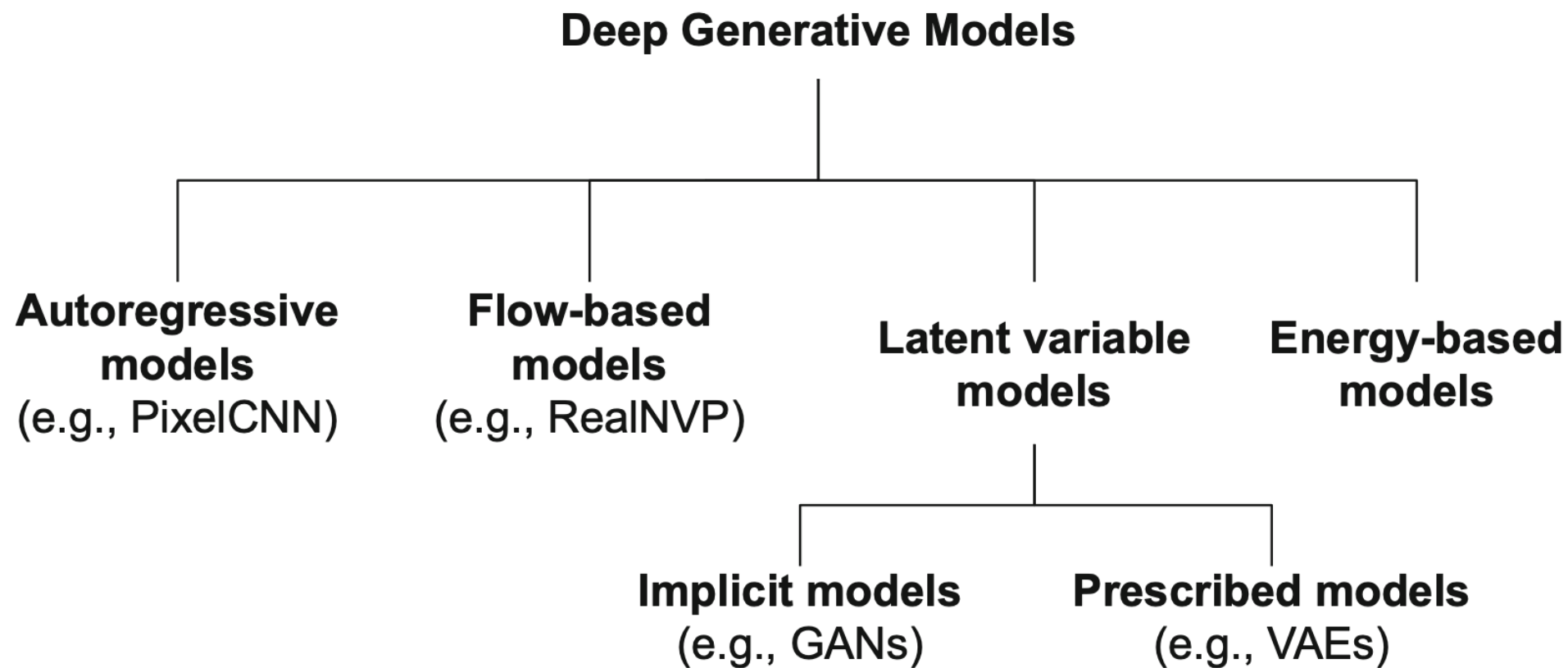
Director of the Center for Innovation in Data Engineering and Science (IDEAS)

Rachleff University Professor, University of Pennsylvania

Amazon Scholar & Chief Scientist at NORCE



# Taxonomy of Generative Models



# The story up till now

- **Step 1:** We set out with our original goal of learning a model  $p_\theta$  that gives maximum likelihood to our datapoints  $\mathbf{x}_i$
- **Step 2:** We introduced latent variables  $\mathbf{z}$  such that  $\mathbf{z} \sim p(\mathbf{z})$  and  $\mathbf{x} | \mathbf{z} \sim p(\mathbf{x} | \mathbf{z})$ , which gave us the marginalization  $p(\mathbf{x}) = \int p(\mathbf{x} | \mathbf{z})p(\mathbf{z}) d\mathbf{z}$ 
  - Step 2a: When we assumed  $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, I)$ , and  $p(\mathbf{x} | \mathbf{z}) = \mathcal{N}(\mathbf{x}; W\mathbf{z} + b, \sigma^2 I)$ , we could solve for  $(W, b, \sigma^2)$  in closed form! This gave us PPCA.
- **Step 3:** We set up variational inference because sadly, not everything in life is Gaussian and linear. This gave us a new Evidence Lower Bound (ELBO) objective:

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

- Step 3a: When the integral is easy to evaluate, we can alternate between optimizing w.r.t.  $\theta$  with  $q(\mathbf{x} | \mathbf{z})$  fixed and vice versa, leading to the Expectation Maximization (EM) algorithm.

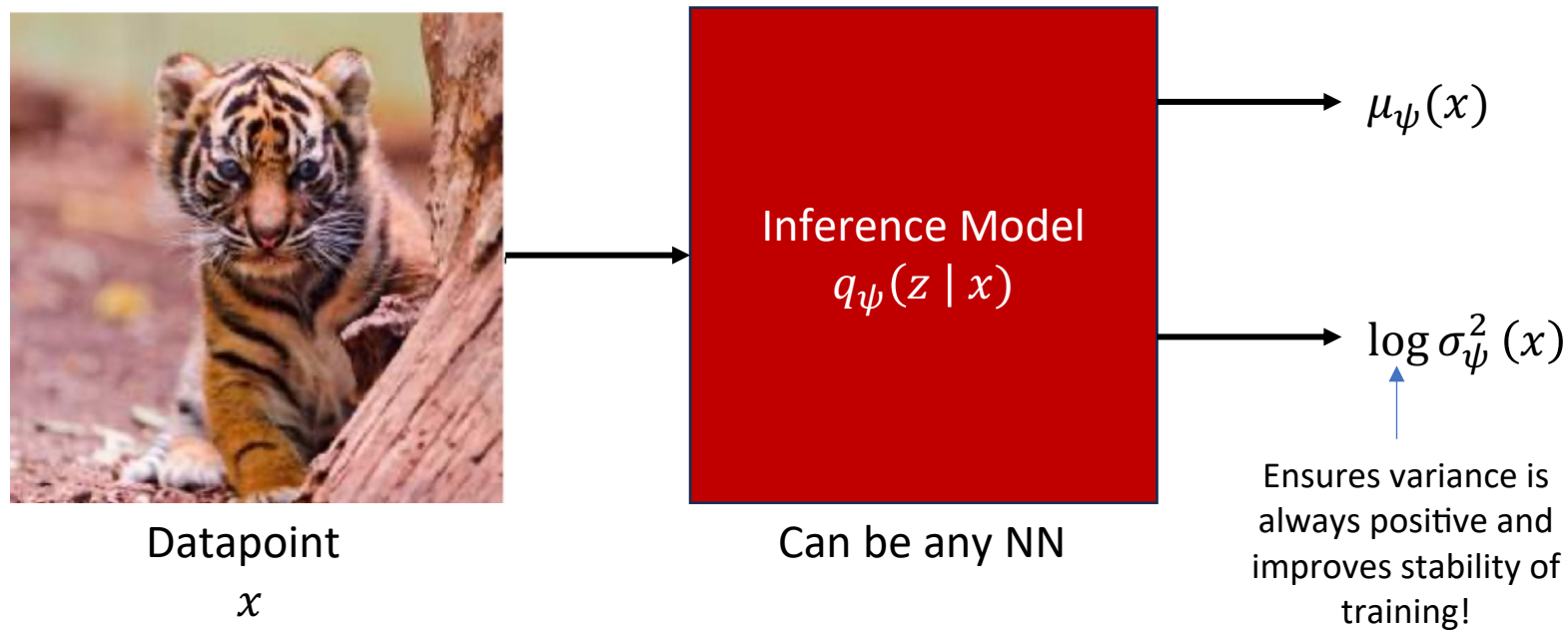
# The story continues with Variational Autoencoders

- Before introducing VAEs formally, let us decompose the ELBO further

$$\begin{aligned}\max_{\theta} \log p_{\theta}(x_i) &= \max_{\theta} \max_q \sum_{i=1}^N \int q(z|x_i) \log \frac{p_{\theta}(x_i, z)}{q(z|x_i)} dz \\ &\geq \max_{\theta, \psi} \sum_i E_{z \sim q_{\psi}(z|x_i)} \log \frac{p_{\theta}(x_i, z)}{q_{\psi}(z|x_i)} \quad \text{Evidence Lower Bound (ELBO)} \\ &= \max_{\theta, \psi} \sum_i E_{z \sim q_{\psi}(z|x_i)} \log p_{\theta}(x_i|z) \frac{p(z)}{q_{\psi}(z|x_i)} \\ &= \max_{\theta, \psi} \sum_i E_{z \sim q_{\psi}(z|x_i)} \log p_{\theta}(x_i|z) + E_{z \sim q_{\psi}(z|x_i)} \log \frac{p(z)}{q_{\psi}(z|x_i)} \\ &\quad \quad \quad \uparrow \\ &\quad \quad \quad -D_{KL}(q_{\psi}(z|x)||p(z))\end{aligned}$$

# Variational AutoEncoders (VAEs): Setup

- We have three models we need to define for VAE model
  1. **Inference model  $q_{\psi}(z | x)$** : We will define as  $q_{\psi}(z | x) = \mathcal{N}(z; \mu_{\psi}(x), \sigma_{\psi}^2(x)I)$ , i.e., a normal distribution with learned mean and covariance



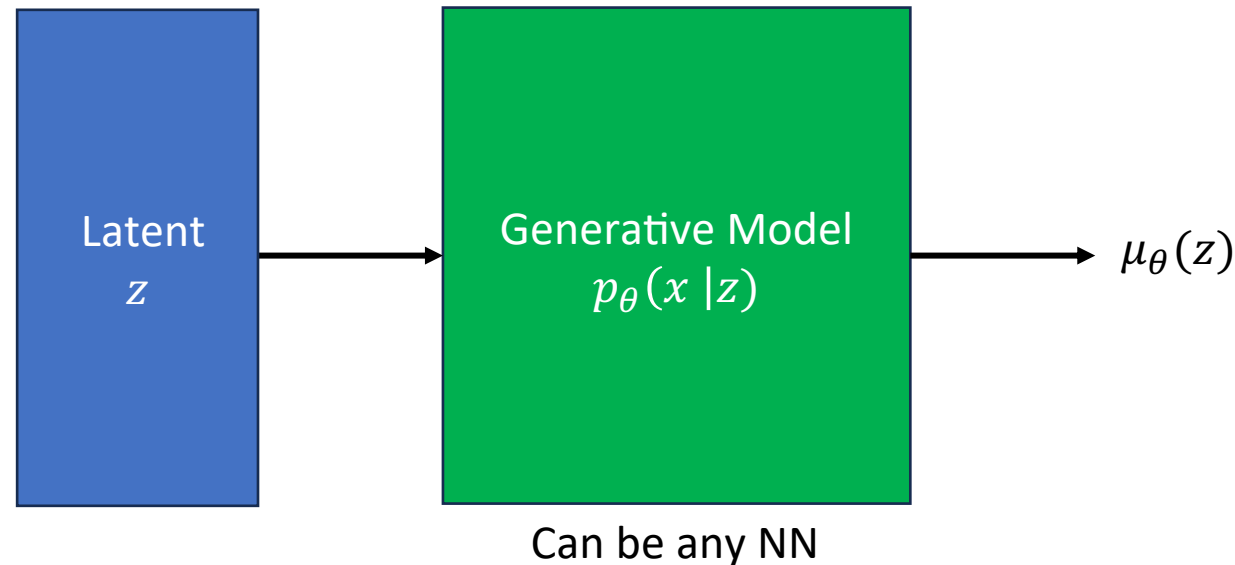
2. **Prior model  $p(z)$** : We will define prior for latent variables as  $p(z) = \mathcal{N}(z; 0, I)$

# Variational AutoEncoders (VAEs): Setup

- We have three models we need to define for VAE model

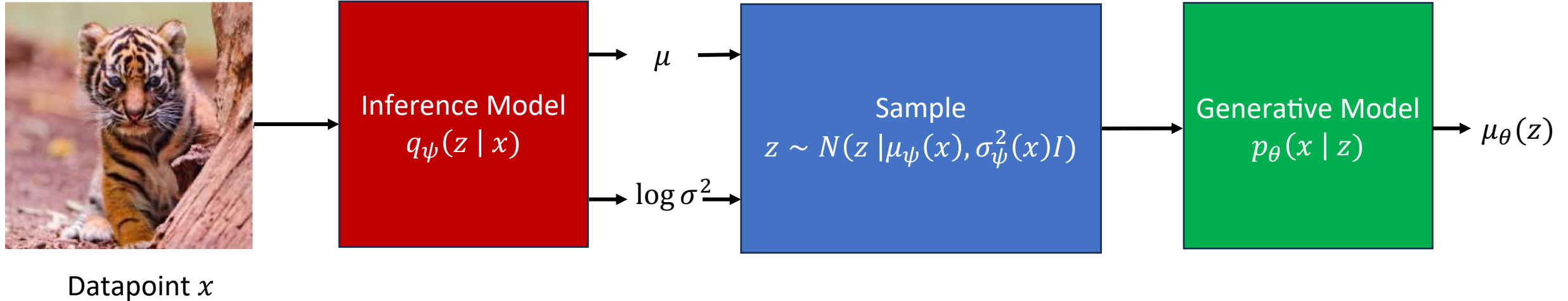
3. **Generative model  $p_{\theta}(x | z)$** : We will define as

- $p_{\theta}(x | z) = \mathcal{N}(z; \mu_{\theta}(z), \eta^2 I)$ , i.e., a normal distribution with learned mean and variance
- $p_{\theta}(x | z) = \text{Cat}(z; \pi_{\theta}(z))$ , i.e., a categorical distribution with learned class probabilities



- Note this can be defined in many different ways, yielding different models (such as a categorical distribution over 255 values of each pixel)

# Variational Autoencoders: Training

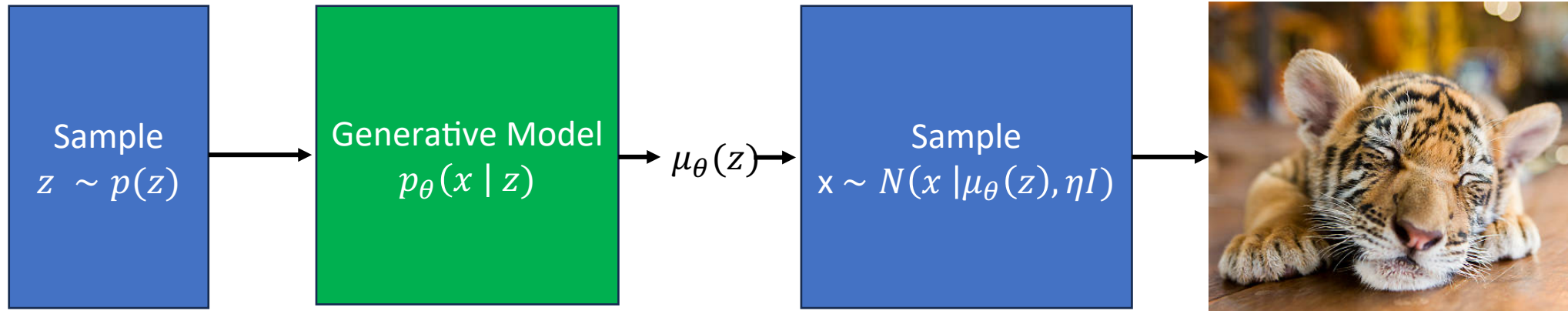


ELBO Objective

$$\mathbb{E}_{z \sim q_{\psi}(z|x)} [\log p_{\theta}(x | z) - KL(q_{\psi}(z | x) || p(z))]$$

# Variational Autoencoders after Training

- Suppose we have learned VAE using the ELBO loss (details to follow).
- Then, as a generative model, we just sample  $z \sim p(z)$  and use the fixed generative model  $p_{\theta}(x | z)$





# Computing the ELBO Loss

$$L_{\theta,\psi}(x) := \underbrace{\mathbb{E}_{z \sim q_{\psi}(z|x)} \log p_{\theta}(x|z)}_{\text{Term 1 (Reconstruction Error)}} + \underbrace{\mathbb{E}_{z \sim q_{\psi}(z|x)} \log \frac{p(z)}{q_{\psi}(z|x)}}_{\text{KL Divergence}}$$

- **Term 1 (Reconstruction Error)**

- Because  $p_{\theta}(x|z) = \mathcal{N}(x|\mu_{\theta}(z), \eta I)$ , we have  $\log p_{\theta}(x|z) = -\frac{1}{2\eta} \|x - \mu_{\theta}(z)\|_2^2 + \text{const.}$
- We can approximate the expectation over  $z \sim q_{\psi}(z|x)$  by an average over  $q_{\psi}(z|x)$
- Recall that the expectation of a function  $f(z)$  w.r.t. a random variable  $z \sim p$  can be approximated from i.i.d. samples  $z_j \sim p, j = 1, \dots, M$ , via Monte Carlo averages

$$\mathbb{E}_{z \sim p}[f(z)] = \int_z f(z)p(z)dx \approx \frac{1}{M} \sum_j f(z_j)$$

- Applying the above formula to  $M$  i.i.d. samples  $z_j \sim q_{\psi}(Z|x)$  we get **reconstruction error**

$$\mathbb{E}_{z \sim q_{\psi}}[\log p_{\theta}(x|z)] \approx \frac{1}{M} \sum_j \log p_{\theta}(x|z_j) = \frac{1}{2\eta M} \sum_j \|x - \mu_{\theta}(z_j)\|_2^2$$

# Computing the ELBO Loss

$$L_{\theta, \psi}(x) := \underbrace{\mathbb{E}_{z \sim q_{\psi}(z | x)} \log p_{\theta}(x | z)}_{\text{Term 1}} + \underbrace{\mathbb{E}_{z \sim q_{\psi}(z | x)} \log \frac{p(z)}{q_{\psi}(z | x)}}_{\text{Term 2}}$$

- **Term 2 (Regularization to Prior)**

- Because  $\mathbb{E}_{z \sim q_{\psi}(z|x)} \log \frac{p(z)}{q_{\psi}(z|x)} = -D_{KL}(q_{\psi}(z|x) || p(z))$ ,  $q_{\psi}(z|x) = \mathcal{N}(z | \mu_{\psi}(x), \sigma_{\psi}^2(x)I)$ ,  $p(z) = \mathcal{N}(0, I)$ , the second term is a KL divergence between two d-dimensional Gaussians, which has a closed form solution

$$KL(\mathcal{N}(\mu_1, \sigma_1^2 I) || \mathcal{N}(\mu_2, \sigma_2^2 I)) = \log \left( \frac{\sigma_2}{\sigma_1} \right) - \frac{d}{2} + \frac{d\sigma_1^2 + ||\mu_1 - \mu_2||_2^2}{2\sigma_2^2}$$

- Applying the above formula to our VAE model yields,

$$KL(q_{\psi}(z | x) || p(z)) = -\log(\sigma_{\psi}(x)) + \frac{d\sigma_{\psi}^2(x) + ||\mu_{\psi}(x)||_2^2}{2} + constant$$

- Note this term does not require any sampling w.r.t.  $z$  because the expectation is computed in closed form thanks for the KL formula for Gaussians.

# Maximizing ELBO: How to Optimize?

- **ELBO objective:** We want to solve the following optimization problem:

$$\max_{\theta, \psi} \sum_i L_{\theta, \psi}(x_i) = \sum_i E_{z \sim q_{\psi}(z | x_i)} \log \frac{p_{\theta}(x_i, z)}{q_{\psi}(z | x_i)}$$

- **Simple idea:** Just alternate gradient ascent wrt  $\theta, \psi$  on objective function

$$\theta_{k+1} = \theta_k + \alpha \frac{\delta L}{\delta \theta}(\theta_k, \psi_k)$$

$$\psi_{k+1} = \psi_k + \alpha \frac{\delta L}{\delta \psi}(\theta_k, \psi_k)$$

# Stochastic Optimization of ELBO wrt $\theta$

- **Issue:** Computing  $\nabla_{\theta}$  (ELBO) is not easy because of expectation w.r.t. latent  $z$ .
- **Solution:** Compute an unbiased estimator

$$\begin{aligned}\nabla_{\theta} L_{\theta, \psi}(x) &= \nabla_{\theta} E_{z \sim q_{\psi}(z | x)} [\log(p_{\theta}(x | z))] + \nabla_{\theta} E_{z \sim q_{\psi}(z | x)} \left[ \log \left( \frac{p(z)}{q_{\psi}(z | x)} \right) \right] \\ &= E_{z \sim q_{\psi}(z | x)} [\nabla_{\theta} \log p_{\theta}(x | z)] \\ &= -\frac{1}{2\eta} E_{z \sim q_{\psi}(z | x)} \nabla_{\theta} \|x - \mu_{\theta}(z)\|_2^2\end{aligned}$$

- We can take sample averages to compute an unbiased estimator
  - For each datapoint  $x$ , compute  $q_{\psi}(z | x)$  through encoder, which gives  $\mu_{\psi}(x), \sigma_{\psi}^2(x)$
  - Sample  $z_j \sim \mathcal{N}(\mu_{\psi}(x), \sigma_{\psi}^2(x)I)$ , find  $\mu_{\theta}(z_j)$  through decoder
  - Estimate gradient from  $M$  samples:  $\nabla_{\theta} L_{\theta, \psi}(x) \approx -\frac{1}{2\eta M} \sum_j \nabla_{\theta} \|x - \mu_{\theta}(z_j)\|_2^2$

# Stochastic Optimization of ELBO wrt $\psi$

- Here, we cannot just switch gradient and expectation because both are wrt to  $\psi$

$$\nabla_{\psi} E_{z \sim q_{\psi}(z|x)} [\log p_{\theta}(x, z) - \log q_{\psi}(z | x)] \neq E_{z \sim q_{\psi}(z|x)} \nabla_{\psi} [\log p_{\theta}(x, z) - \log q_{\psi}(z | x)]$$

- Reparameterization trick**

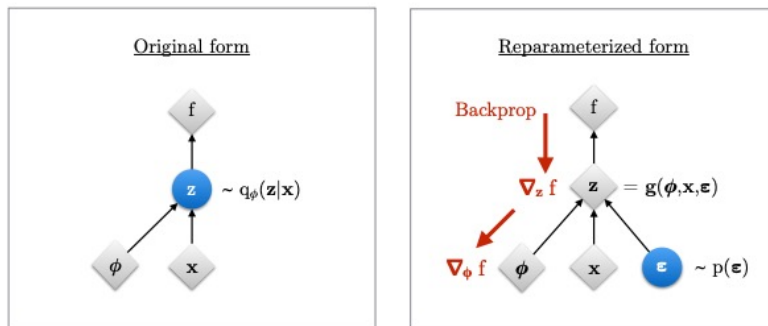
- Because  $q_{\psi}(z|x) = N(z; \mu_{\psi}(x), \sigma_{\psi}(x)I)$ , we can rewrite samples  $z \sim q_{\psi}(z | x)$  as

$$z_{\psi} = g(\epsilon, \psi, x) = \mu_{\psi}(x) + \sigma_{\psi}(x)\epsilon \quad \text{for } \epsilon \sim N(0, I)$$

- With this change of variables, we can rewrite the gradient of the loss as:

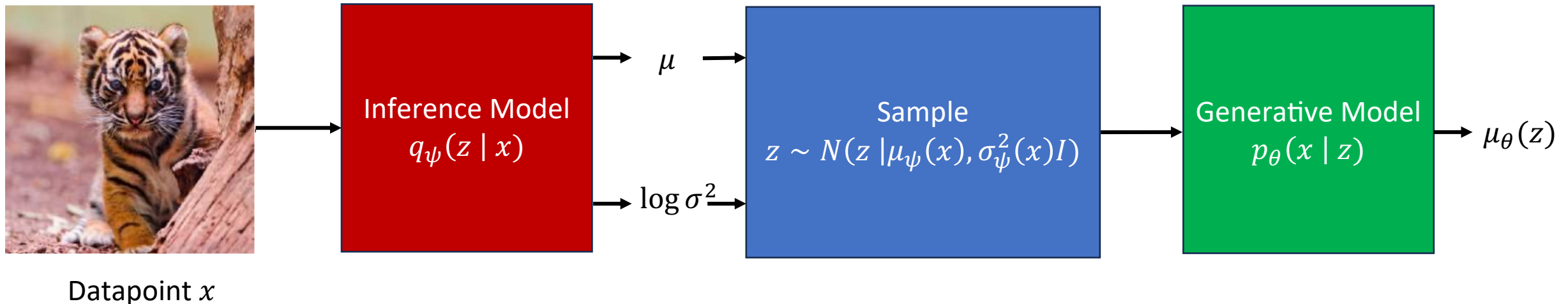
$$\begin{aligned} \nabla_{\psi} L_{\theta, \psi}(x) &= \nabla_{\psi} E_{z \sim q_{\psi}(z | x)} [\log p_{\theta}(x, z) - \log q_{\psi}(z | x)] \\ &= \nabla_{\psi} E_{\epsilon \sim N(0, I)} [\log p_{\theta}(x, z_{\psi}) - \log q_{\psi}(z_{\psi} | x)] \end{aligned}$$

Gradient and expectation can now be switched! So as before, we compute an unbiased estimator by sampling many  $\epsilon$



# Putting it all together

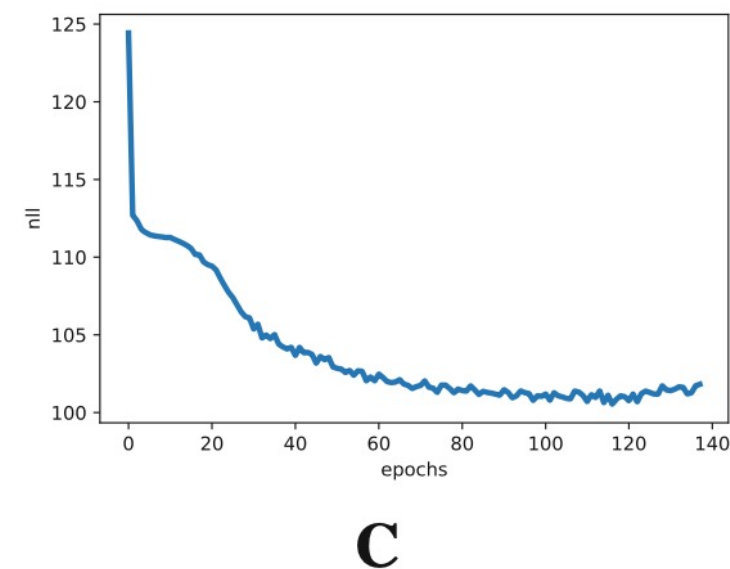
- Variational Autoencoder
  - We modelled inference and generative model as deep networks
  - We interpreted ELBO as an expected reconstruction error plus a KL-regularization to prior
  - Then, we rewrote the sampling in the latent space using the reparameterization trick
  - Finally, we derived stochastic gradient estimates to optimize the ELBO and learn a VAE



# VAE's in Action

- $q_{\psi}(z | x) = N(z | \mu_{\psi}(x), \sigma_{\psi}^2(x))$
- $p(z) = N(z | 0, I)$
- $p_{\theta}(x | z) = \textit{Categorical}(x | \pi_{\theta}(z))$  Note this is different from the model considered up till now!
- Encoder network:  $x \in R^D \rightarrow \text{Linear}(D, 256) \rightarrow \text{LeakyReLU} \rightarrow \text{Linear}(256, 2d) \rightarrow$   
split into  $\mu \in R^d, \log \sigma^2 \in R^d$
- Decoder network:  $z \in R^d \rightarrow \text{Linear}(d, 256) \rightarrow \text{LeakyReLU} \rightarrow \text{Linear}(256, D) \rightarrow$   
softmax

# VAE's for Generation of MNIST Digits



**Fig. 4.4** An example of outcomes after the training: (a) Randomly selected real images. (b) Unconditional generations from the VAE. (c) The validation curve during training

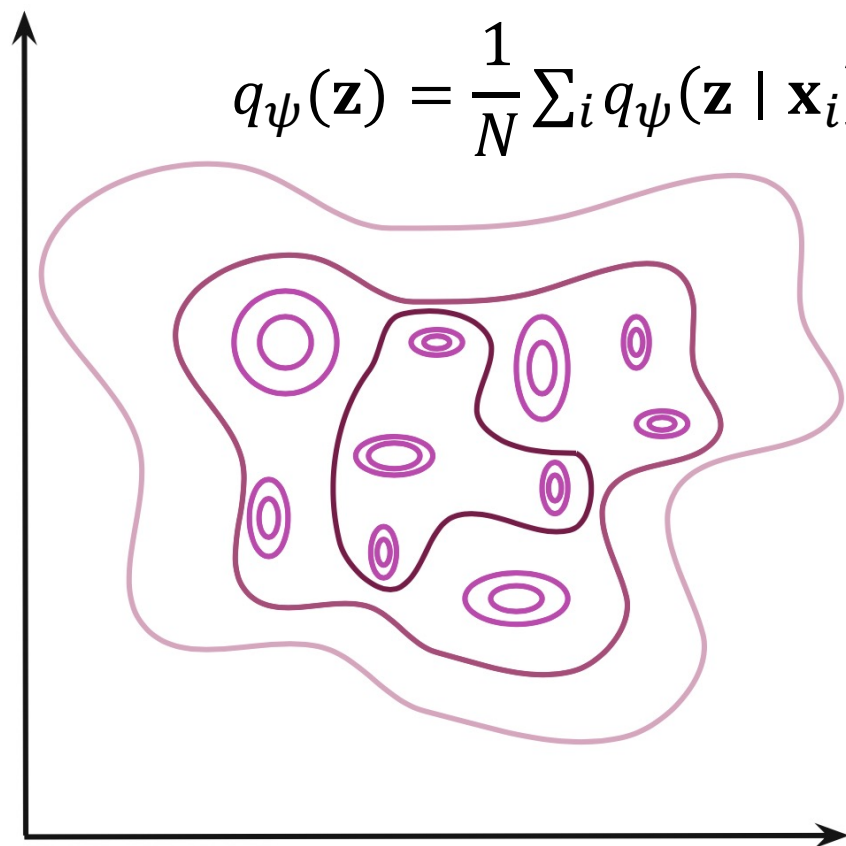


# Typical Issues with VAEs

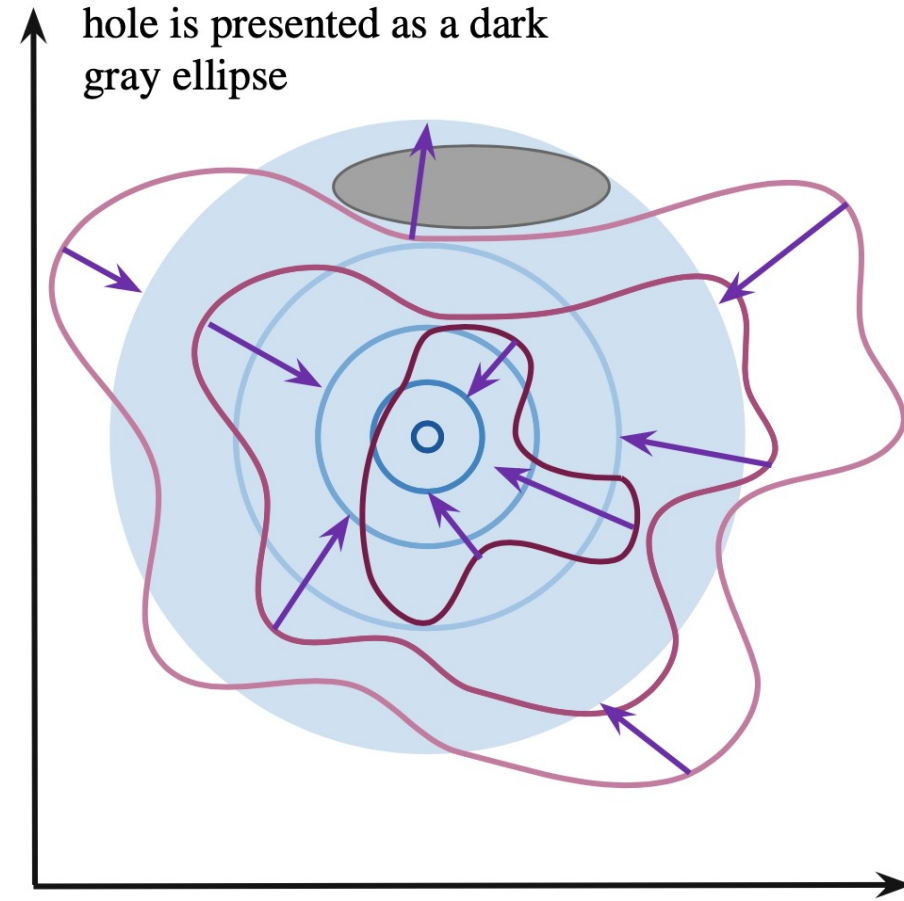
- ELBO: 
$$\ln p(\mathbf{x}) = \underbrace{\mathbb{E}_{\mathbf{z} \sim q_{\psi}(\mathbf{z}|\mathbf{x})} [\ln p_{\theta}(\mathbf{x} | \mathbf{z})] - KL[q_{\psi}(\mathbf{z} | \mathbf{x}) \parallel p(\mathbf{z})]}_{ELBO} + \underbrace{KL[q_{\psi}(\mathbf{z} | \mathbf{x}) \parallel p_{\theta}(\mathbf{z} | \mathbf{x})]}_{\geq 0}$$
- Posterior collapse
  - If the decoder is so powerful that it treats  $z$  as noise, then  $\forall_{\mathbf{x}} q_{\phi}(\mathbf{z} | \mathbf{x}) = p(\mathbf{z})$
- Mismatch between prior  $p(\mathbf{z})$  and aggregated posterior  $q_{\phi}(\mathbf{z}) = \frac{1}{N} \sum_i q_{\psi}(\mathbf{z} | \mathbf{x}_i)$ 
  - Prior assigns high probability but aggregated posterior assigns low probability, or other way around.
  - Sampling from such regions provides unrealistic latent values and the decoder produces images of very low quality.
- Out-of-distribution samples

# Aggregated posterior

**Fig. 4.5** An example of the aggregated posterior. Individual points are encoded as Gaussians in the 2D latent space (magenta), and the mixture of variational posteriors (the aggregated posterior) is presented by contours



**Fig. 4.6** An example of the effect of the cross-entropy optimization with a non-learnable prior. The aggregated posterior (purple contours) tries to match the non-learnable prior (in blue). The purple arrows indicate the change of the aggregated posterior. An example of a hole is presented as a dark gray ellipse



# Improving VAEs

- The ELBO consists of two parts: first, the reconstruction error

$$RE \triangleq \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} \left[ \mathbb{E}_{q_{\psi}(\mathbf{z}|\mathbf{x})} [\ln p_{\theta}(\mathbf{x} | \mathbf{z})] \right]$$

- Then the regularization term between the encoder and the prior

$$\begin{aligned} \Omega &\triangleq \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} \left[ \mathbb{E}_{q_{\psi}(\mathbf{z}|\mathbf{x})} [\ln p(\mathbf{z}) - \ln q_{\psi}(\mathbf{z} | \mathbf{x})] \right] \\ &= -KL[q_{\psi}(\mathbf{z}) \parallel p(\mathbf{z})] + \mathbb{H}[q_{\psi}(\mathbf{z} | \mathbf{x})] \end{aligned}$$

- For a Gaussian, the entropy is maximized when sigma -> infinity