Deep Generative Models: Variational Auto Encoders

Fall Semester 2026

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

Penn
University of Pennsylvania

The story up till now

- **Step 1**: We set out with our original goal of learning a model p_θ that gives maximum likelihood to our datapoints $\boldsymbol{x}_i$.

- **Step 2**: We introduced latent variables $\boldsymbol{z}$ such that $\boldsymbol{z} \sim p(\boldsymbol{z})$ and $\boldsymbol{x} \mid \boldsymbol{z} \sim p(\boldsymbol{x} \mid \boldsymbol{z})$, which gave us the marginalization $p(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z})\, d\boldsymbol{z}$
  - Step 2a: When we assumed $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$, and $p(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}, \sigma^2 \boldsymbol{I})$, we could solve for $(\boldsymbol{W}, \boldsymbol{b}, \sigma^2)$ 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:

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

  - Step 3a: When the integral is easy to evaluate, we can alternate between optimizing w.r.t. θ with q 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

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

$$\geq \max_{\theta,\psi} \sum_i \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}|\boldsymbol{x}_i)} \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z})}{q_\psi(\boldsymbol{z}|\boldsymbol{x}_i)}$$

Evidence Lower Bound (ELBO)

$$= \max_{\theta,\psi} \sum_i \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}|\boldsymbol{x}_i)} \log p_\theta(\boldsymbol{x}_i|\boldsymbol{z}) \frac{p(\boldsymbol{z})}{q_\psi(\boldsymbol{z}|\boldsymbol{x}_i)}$$

$$= \max_{\theta,\psi} \sum_i \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}|\boldsymbol{x}_i)} \log p_\theta(\boldsymbol{x}_i|\boldsymbol{z}) + \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}|\boldsymbol{x}_i)} \log \frac{p(\boldsymbol{z})}{q_\psi(\boldsymbol{z}|\boldsymbol{x}_i)}$$

Given $\boldsymbol{x}$, sample latent $\boldsymbol{z}$ from q.

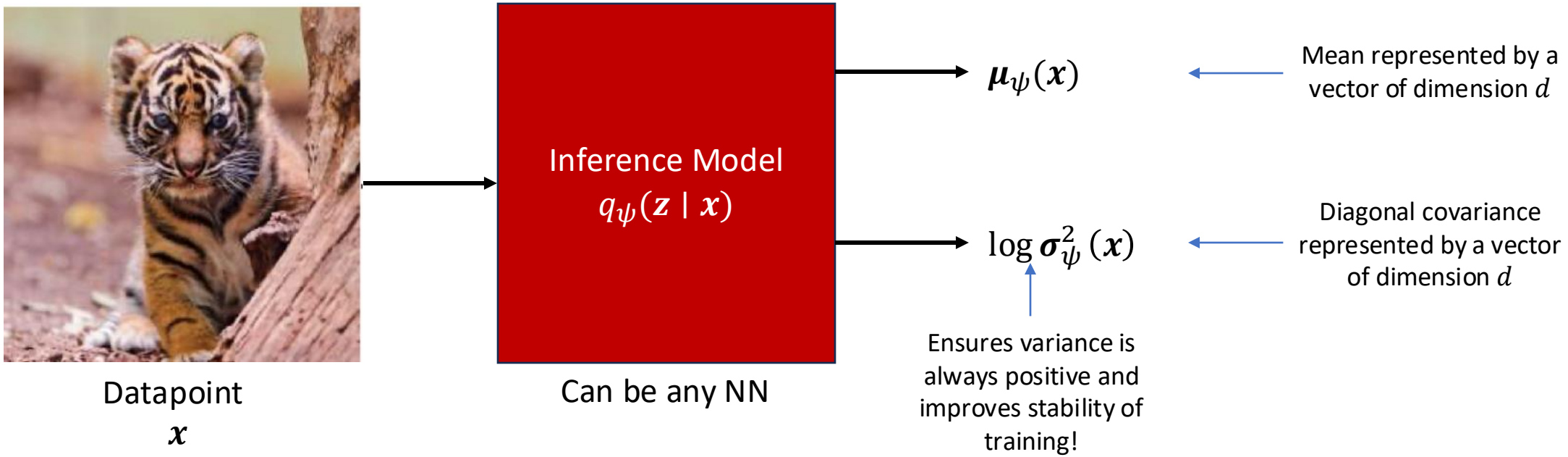
Given latent $\boldsymbol{z}$, sample $\boldsymbol{x}$ from p.

Measure average likelihood of $\boldsymbol{x}$ across all choices of $\boldsymbol{z}$

$-\mathrm{KL}[q_\psi(\boldsymbol{z}|\boldsymbol{x}_i) \,||\, p(\boldsymbol{z})]$

Variational AutoEncoders (VAEs): Setup

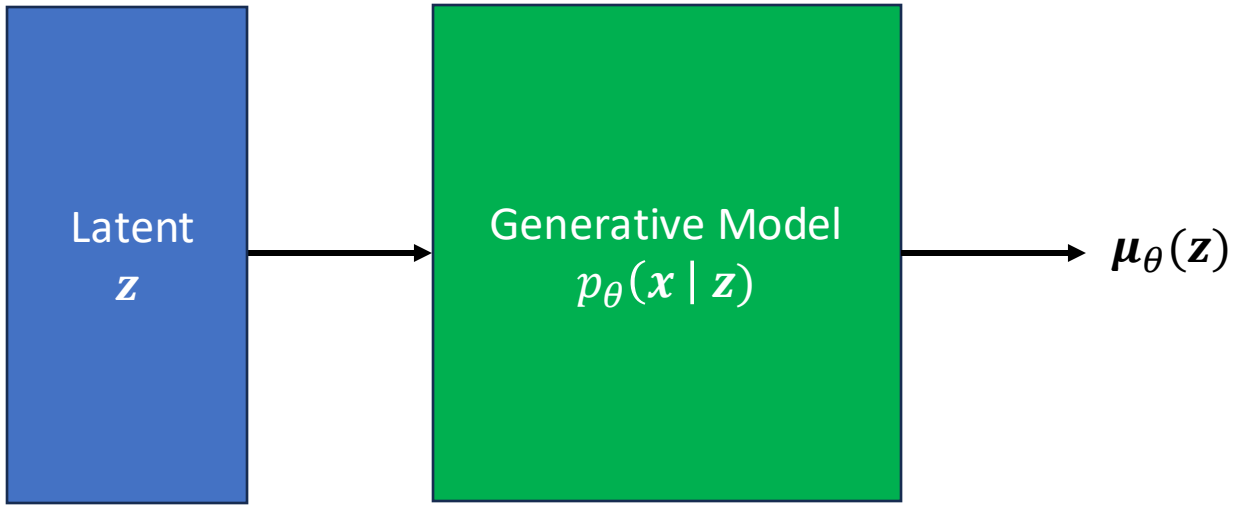
- We have three models we need to define for VAE model
  1. Inference model $q_\psi(\boldsymbol{z} \mid \boldsymbol{x})$: We will define it as $q_\psi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_\psi(\boldsymbol{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_\psi(\boldsymbol{x})))$, i.e., a Gaussian distribution with learned mean and covariance



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

Variational AutoEncoders (VAEs): Setup

- We have three models we need to define for VAE model
  3. Generative model $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$: We will define it as
     - $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_\theta(\boldsymbol{z}), \eta^2 \boldsymbol{I})$, i.e., a Gaussian distribution with learned mean and variance
     - $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) = \text{Cat}(\boldsymbol{x}; \boldsymbol{\pi}_\theta(\boldsymbol{z}))$, 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 256 values of each pixel)

Variational AutoEncoders: Training



Datapoint $\boldsymbol{x}$

Inference Model
$q_\psi(\boldsymbol{z} \mid \boldsymbol{x})$

$\boldsymbol{\mu}_\psi(\boldsymbol{x})$

$\log \boldsymbol{\sigma}^2_\psi(\boldsymbol{x})$

Sample
$\boldsymbol{z} \mid \boldsymbol{x}$
$\sim \mathcal{N}(\boldsymbol{\mu}_\psi(\boldsymbol{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_\psi(\boldsymbol{x})))$

$\boldsymbol{z}$

Generative Model
$p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$

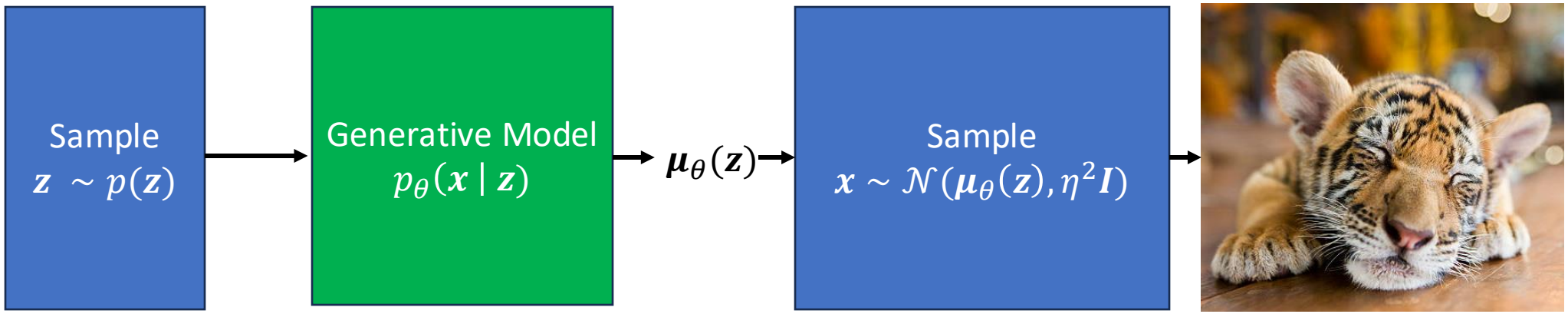
$\boldsymbol{\mu}_\theta(\boldsymbol{z})$

ELBO Objective

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

Variational AutoEncoders: Sampling

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



Computing the ELBO Loss

$$L_{\theta,\psi}(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}\,|\boldsymbol{x})}\,\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z}) + \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}\mid\boldsymbol{x})}\log\frac{p(\boldsymbol{z})}{q_\psi(\boldsymbol{z}\mid\boldsymbol{x})}$$

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

$$\mathbb{E}_{\boldsymbol{z}\sim p}[f(\boldsymbol{z})] = \int_{\boldsymbol{z}} f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \approx \frac{1}{M}\sum_j f(\boldsymbol{z}_j)$$

  - Applying the above formula to M i.i.d. samples $\boldsymbol{z}_j\sim q_\psi(\boldsymbol{z}\mid\boldsymbol{x})$ we get reconstruction error

$$\mathbb{E}_{\boldsymbol{z}\sim q_\psi}[\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z})] \approx \frac{1}{M}\sum_j \log p_\theta(\boldsymbol{x}\mid\boldsymbol{z}_j) = -\frac{1}{2\eta^2 M}\sum_j ||\boldsymbol{x}-\boldsymbol{\mu}_\theta(\boldsymbol{z}_j)||_2^2 + \text{const.}$$

Computing the ELBO Loss

$$L_{\theta,\psi}(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}\,|\boldsymbol{x})}\,\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z}) + \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}\mid\boldsymbol{x})}\log\frac{p(\boldsymbol{z})}{q_\psi(\boldsymbol{z}\mid\boldsymbol{x})}$$

- Term 2 (Regularization to Prior)
  - The second term is a KL divergence between two Gaussians, which can be computed as

$$\mathrm{KL}[\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)] = \frac{1}{2}\left(\log\left(\frac{\det\boldsymbol{\Sigma}_2}{\det\boldsymbol{\Sigma}_1}\right) - d + \mathrm{trace}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right)$$

  - Applying this formula to $q_\psi(\boldsymbol{z}\mid\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_\psi(\boldsymbol{x}),\mathrm{diag}(\boldsymbol{\sigma}_\psi^2(\boldsymbol{x})))$ and $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})$ yields

$$\mathrm{KL}\left[q_\psi(\boldsymbol{z}\mid\boldsymbol{x})||p(\boldsymbol{z})\right] = -\sum_{l=1}^{d}\log\left(\sigma_{l,\psi}(\boldsymbol{x})\right) + \frac{||\boldsymbol{\sigma}_\psi(\boldsymbol{x})||_2^2 + ||\boldsymbol{\mu}_\psi(\boldsymbol{x})||_2^2}{2} + \mathrm{const}$$

  - Note this term does not require any sampling w.r.t. $\boldsymbol{z}$ 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:

$$\max_{\theta,\psi} \sum_i L_{\theta,\psi}(\boldsymbol{x}_i) = \max_{\theta,\psi} \sum_i \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}|\boldsymbol{x}_i)} \log(p_\theta(\boldsymbol{x}_i \mid \boldsymbol{z})) - \mathrm{KL}\left[q_\psi(\boldsymbol{z} \mid \boldsymbol{x}_i)||p(\boldsymbol{z})\right]$$

$$= \max_{\theta,\psi} \sum_{i=1}^{N} -\frac{1}{2\eta^2 M} \sum_{j=1}^{M} [||\boldsymbol{x}_i - \boldsymbol{\mu}_\theta(\boldsymbol{z}_j)||_2^2] + \sum_{l=1}^{d} \log\left(\sigma_{l,\psi}(\boldsymbol{x}_i)\right) - \frac{||\boldsymbol{\sigma}_\psi(\boldsymbol{x}_i)||_2^2 + ||\boldsymbol{\mu}_\psi(\boldsymbol{x}_i)||_2^2}{2} + \text{const}$$

- **Simple idea**: Just alternate gradient ascent w.r.t. θ, ψ on objective function

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

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

Stochastic Optimization of ELBO w.r.t. θ

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

$$\nabla_\theta L_{\theta,\psi}(\boldsymbol{x}) = \nabla_\theta \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}|\boldsymbol{x})}\left[\log(p_\theta(\boldsymbol{x} \mid \boldsymbol{z}))\right] - \nabla_\theta \mathrm{KL}\left[q_\psi(\boldsymbol{z} \mid \boldsymbol{x})||p(\boldsymbol{z})\right]$$

$$= \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}|\boldsymbol{x})}[\nabla_\theta \log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})]$$

$$= -\frac{1}{2\eta^2}\mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}|\boldsymbol{x})}\nabla_\theta \left|\left|\boldsymbol{x} - \boldsymbol{\mu}_\theta(\boldsymbol{z})\right|\right|_2^2$$

- We can take sample averages to compute an unbiased estimator
  - For each datapoint $\boldsymbol{x}$, compute $q_\psi(\boldsymbol{z} \mid \boldsymbol{x})$ through encoder, which gives $\boldsymbol{\mu}_\psi(\boldsymbol{x}), \boldsymbol{\sigma}_\psi^2(\boldsymbol{x})$
  - Sample $\boldsymbol{z}_j \sim \mathcal{N}\left(\boldsymbol{\mu}_\psi(\boldsymbol{x}), \mathrm{diag}(\boldsymbol{\sigma}_\psi^2(\boldsymbol{x}))\right)$, find $\boldsymbol{\mu}_\theta(\boldsymbol{z}_j)$ through decoder
  - Estimate gradient from M samples: $\nabla_\theta L_{\theta,\psi}(\boldsymbol{x}) \approx -\frac{1}{2\eta^2 M}\sum_{j=1}\nabla_\theta ||\boldsymbol{x} - \boldsymbol{\mu}_\theta(\boldsymbol{z}_j)||_2^2$

Stochastic Optimization of ELBO w.r.t. ψ

- Here, we cannot just switch gradient and expectation because both are wrt to ψ

$$\nabla_\psi \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] \neq \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}|\boldsymbol{x})} \nabla_\psi [\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})]$$

- Reparameterization trick
  - Because $q_\psi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_\psi(\boldsymbol{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_\psi(\boldsymbol{x})))$, we can rewrite samples $\boldsymbol{z} \sim q_\psi(\boldsymbol{z} \mid \boldsymbol{x})$ as

$$\boldsymbol{z}_\psi = \boldsymbol{\mu}_\psi(\boldsymbol{x}) + \mathrm{diag}\left(\boldsymbol{\sigma}_\psi(\boldsymbol{x})\right)\boldsymbol{\epsilon} \quad \text{for} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

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

$$\nabla_\psi \mathbb{E}_{\boldsymbol{z}\sim q_\psi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] = \nabla_\psi \mathbb{E}_{\boldsymbol{\epsilon}\sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}_\psi)] = \mathbb{E}_{\boldsymbol{\epsilon}\sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I})} \nabla_\psi [\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}_\psi)]$$

Stochastic Optimization of ELBO w.r.t. ψ

- Reparameterization trick

$$\nabla_\psi L_{\theta,\psi}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\nabla_\psi\left[-\frac{1}{2\eta^2}||\boldsymbol{x}-\boldsymbol{\mu}_\theta(\boldsymbol{z}_\psi)||_2^2\right]+\nabla_\psi\left[\sum_{l=1}^{d}\log\left(\sigma_{l,\psi}(\boldsymbol{x})\right)-\frac{||\boldsymbol{\sigma}_\psi(\boldsymbol{x})||_2^2+||\boldsymbol{\mu}_\psi(\boldsymbol{x})||_2^2}{2}\right]$$

- The computation of the gradient requires gradients of the encoder and decoder networks with respect to the network weights + chain rule.
- Once the gradient has been computed, we can compute the expectation via Monte Carlo sampling

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



Datapoint $\boldsymbol{x}$

Inference Model
$q_\psi(\boldsymbol{z} \mid \boldsymbol{x})$

$\boldsymbol{\mu}_\psi(\boldsymbol{x})$

$\log \boldsymbol{\sigma}^2_\psi(\boldsymbol{x})$

Sample
$\boldsymbol{z} \mid \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_\psi(\boldsymbol{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_\psi(\boldsymbol{x})))$

$\boldsymbol{z}$

Generative Model
$p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$

$\boldsymbol{\mu}_\theta(\boldsymbol{z})$

ELBO Objective

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

VAEs in Action

- $q_\psi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_\psi(\boldsymbol{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_\psi(\boldsymbol{x})))$
- $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \boldsymbol{I})$
- $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) = \mathrm{Cat}(\boldsymbol{x}; \boldsymbol{\pi}_\theta(\boldsymbol{z}))$

Note this is different from the model considered up till now!

- Encoder network: $\boldsymbol{x} \in \mathbb{R}^D \rightarrow$ Linear(D, 256) $\rightarrow$ LeakyReLU $\rightarrow$ Linear (256, $2d$) $\rightarrow$ split into $\boldsymbol{\mu} \in \mathbb{R}^d, \log \boldsymbol{\sigma}^2 \in \mathbb{R}^d$

- Decoder network: $\boldsymbol{z} \in \mathbb{R}^d \rightarrow$ Linear(d, 256) $\rightarrow$ LeakyReLU $\rightarrow$ Linear(256, D) $\rightarrow$ $\mathrm{softmax}$

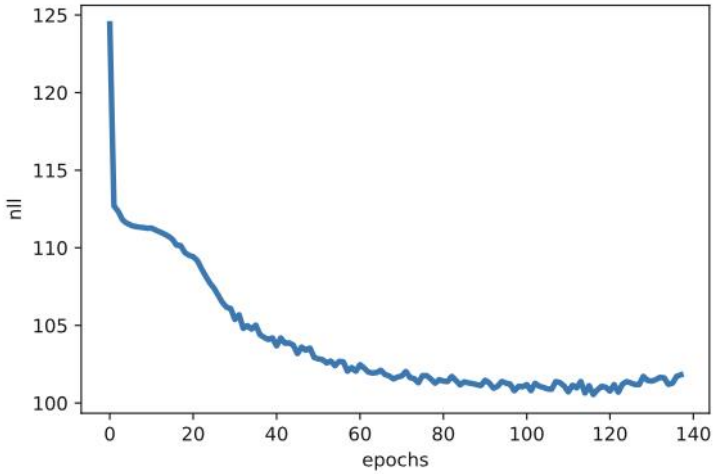
VAE's for Generation of MNIST Digits

nll stands for Negative Log Likelihood.



A

B



C

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

VAEs: extensions

- So far, we have seen how to model $p(\boldsymbol{x})$ via VAEs.

- **Question**: What if we want conditional generation, i.e., sampling from $p(\boldsymbol{x}|\boldsymbol{y})$?
- **Answer**: Model $p(\boldsymbol{x}|\boldsymbol{y})$ via conditional VAE

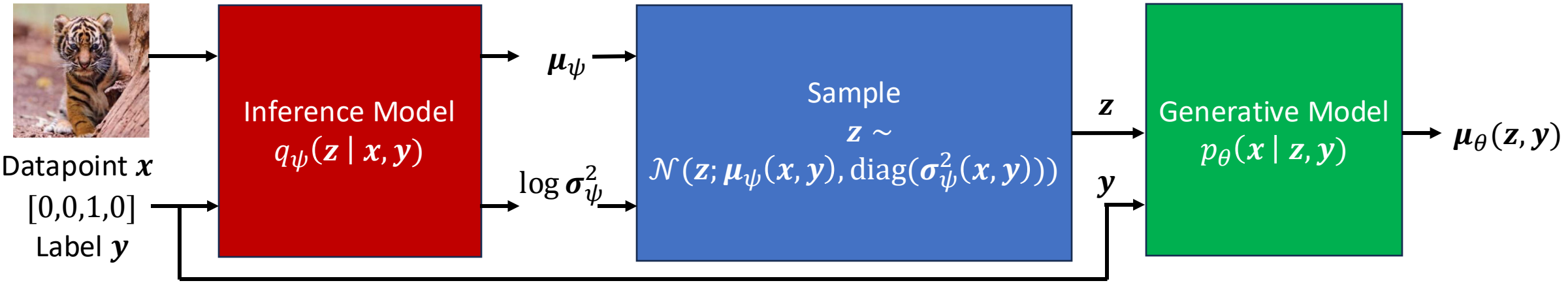
- **Question**: What if we want both conditional generation $p(\boldsymbol{x}|\boldsymbol{y})$ and classification $p(\boldsymbol{y}|\boldsymbol{x})$
- **Answer**: Model $p(\boldsymbol{x}, \boldsymbol{y})$ via joint VAE

Conditional VAE

- Regular VAE: start from $\log p_\theta(\boldsymbol{x})$ and derive its ELBO $\int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z}$
- Here, since we want to model $p_\theta(\boldsymbol{x} \mid \boldsymbol{y})$, let us derive ELBO for $\log p_\theta(\boldsymbol{x} \mid \boldsymbol{y})$

$$\max_\theta \log p_\theta(\boldsymbol{x} \mid \boldsymbol{y}) = \max_\theta \max_\psi \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} \log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y}) - \mathrm{KL}[q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \,||\, p_\theta(\boldsymbol{z} \mid \boldsymbol{y})]$$

  - Details in the next slide. Spoiler alert: the ELBO derivation is almost the same as before
- Architecture:



Conditional VAE: ELBO

- Let $q_\psi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ be the variational distribution. Observe that

$$\log p_\theta(\boldsymbol{x} \mid \boldsymbol{y}) = \int q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \log p_\theta(\boldsymbol{x} \mid \boldsymbol{y}) d\boldsymbol{z} = \int q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{y})}{p_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} d\boldsymbol{z}$$

$$= \int q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{y})}{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} \frac{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}{p_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} d\boldsymbol{z}$$

$$= \boxed{\int q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{y})}{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} d\boldsymbol{z}} + \boxed{\int q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \log \frac{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}{p_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} d\boldsymbol{z}}$$

Evidence **L**ower **Bo**und (ELBO)

$\mathrm{KL}[q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}) \,||\, p_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})]$

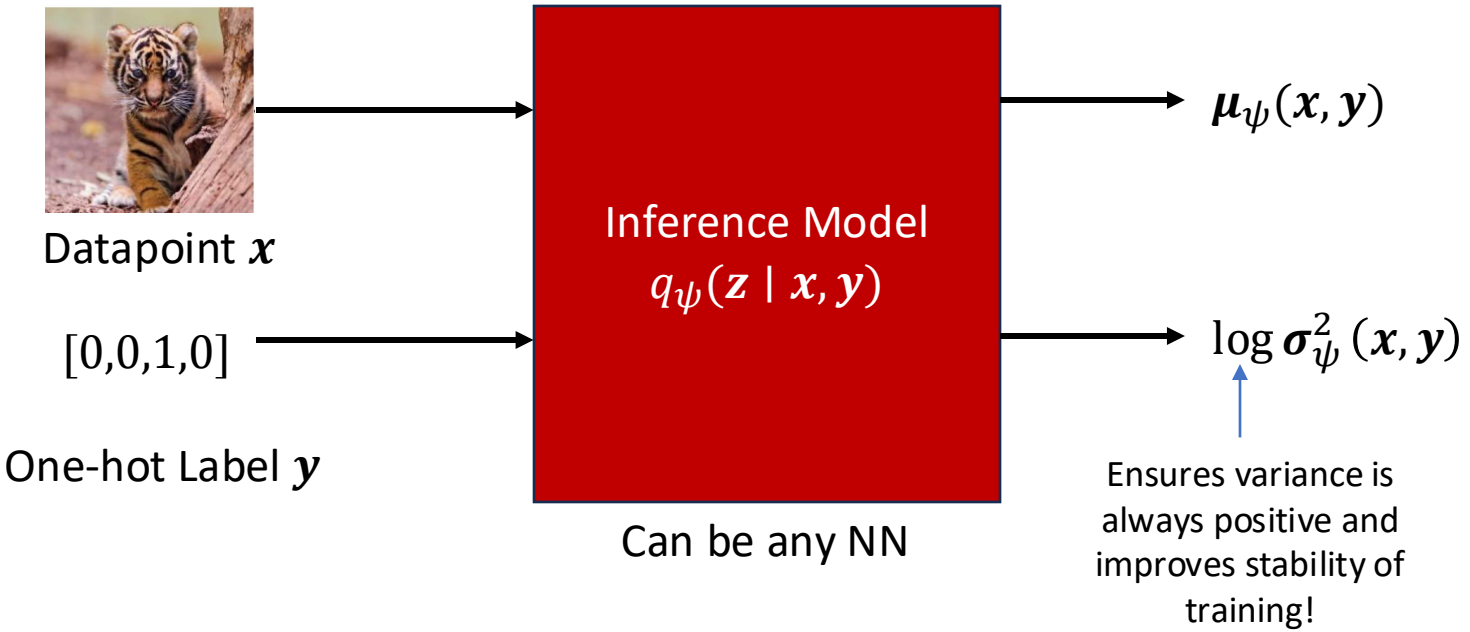
Therefore, $\max_\theta \log p_\theta(\boldsymbol{x} \mid \boldsymbol{y}) \geq \max_\theta \max_\psi \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{y})}{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}$

$$= \max_\theta \max_\psi \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} \log \frac{p_\theta(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y}) p_\theta(\boldsymbol{z}|\boldsymbol{y})}{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}$$

$$= \max_\theta \max_\psi \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} \log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y}) - \mathrm{KL}[q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \,||\, p_\theta(\boldsymbol{z} \mid \boldsymbol{y})]$$

Conditional VAE: Setup

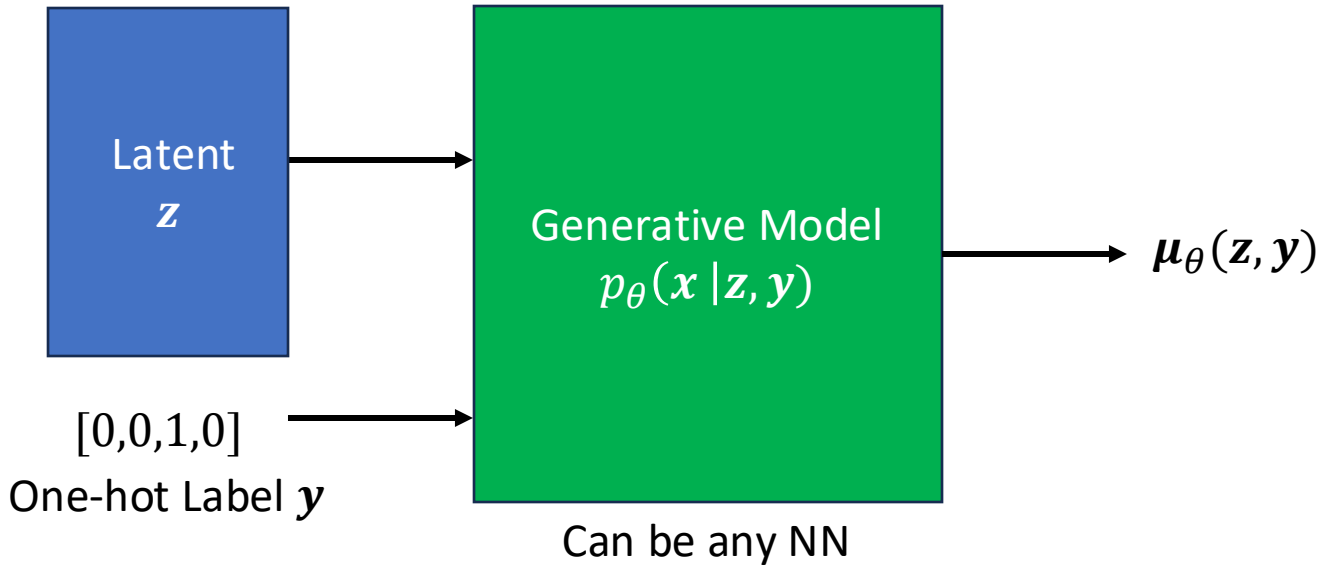
- We have three models we need to define for VAE model
  1. Inference model $q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$: Defined as $q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_\psi(\boldsymbol{x}, \boldsymbol{y}), \mathrm{diag}(\boldsymbol{\sigma}^2_\psi(\boldsymbol{x}, \boldsymbol{y})))$, i.e., a Gaussian distribution with learned mean and covariance



  2. Prior model $p(\boldsymbol{z} \mid \boldsymbol{y})$: $p(\boldsymbol{z} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$ or $p_\theta(\boldsymbol{z} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_\theta(\boldsymbol{y}), \mathrm{diag}(\boldsymbol{\sigma}^2_\theta(\boldsymbol{y})))$

Conditional VAE: Setup

- We have three models we need to define for VAE model
  3. Generative model $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y})$: Defined as
     - $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_\theta(\boldsymbol{z}, \boldsymbol{y}), \eta^2 \boldsymbol{I})$, i.e., a Gaussian distribution with learned mean and variance
     - $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y}) = \text{Cat}(\boldsymbol{x}; \boldsymbol{\pi}_\theta(\boldsymbol{z}, \boldsymbol{y}))$, 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 256 values of each pixel)

Joint VAE

- Regular VAE: start from $\log p_\theta(\boldsymbol{x})$ and derive its ELBO $\int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z}$
- Here, since we want to model $p_\theta(\boldsymbol{x}, \boldsymbol{y})$, let us derive ELBO for $\log p_\theta(\boldsymbol{x}, \boldsymbol{y})$
- Let $q_\psi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ be the variational distribution. Observe that

$$\log p_\theta(\boldsymbol{x}, \boldsymbol{y}) = \int q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \log p_\theta(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{z} = \int q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} d\boldsymbol{z}$$

$$= \int q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} \frac{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}{p_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} d\boldsymbol{z}$$

$$= \boxed{\int q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} d\boldsymbol{z}} + \boxed{\int q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \log \frac{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}{p_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} d\boldsymbol{z}}$$

Evidence Lower Bound (ELBO)

$\mathrm{KL}[q_\psi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}) \,||\, p_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})]$

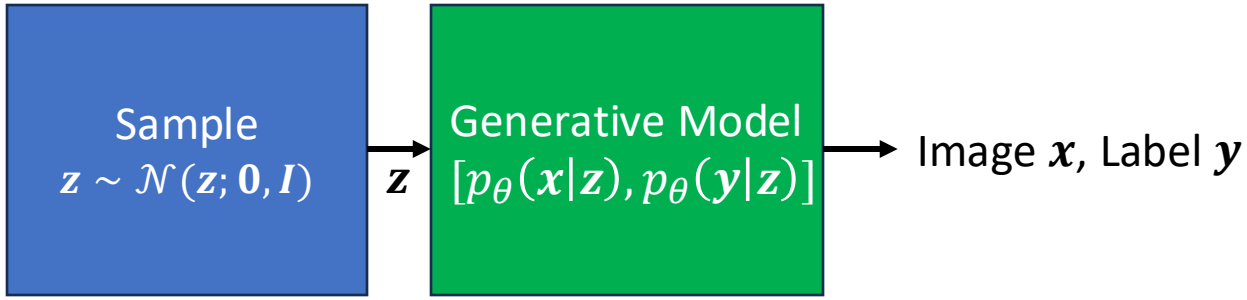
Therefore, $\log p_\theta(\boldsymbol{x}, \boldsymbol{y}) = \max_\psi \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}$ *joint probability*

- Now we turn to how to factorize and parameterize $p_\theta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$.

Joint VAE: joint probability factorization

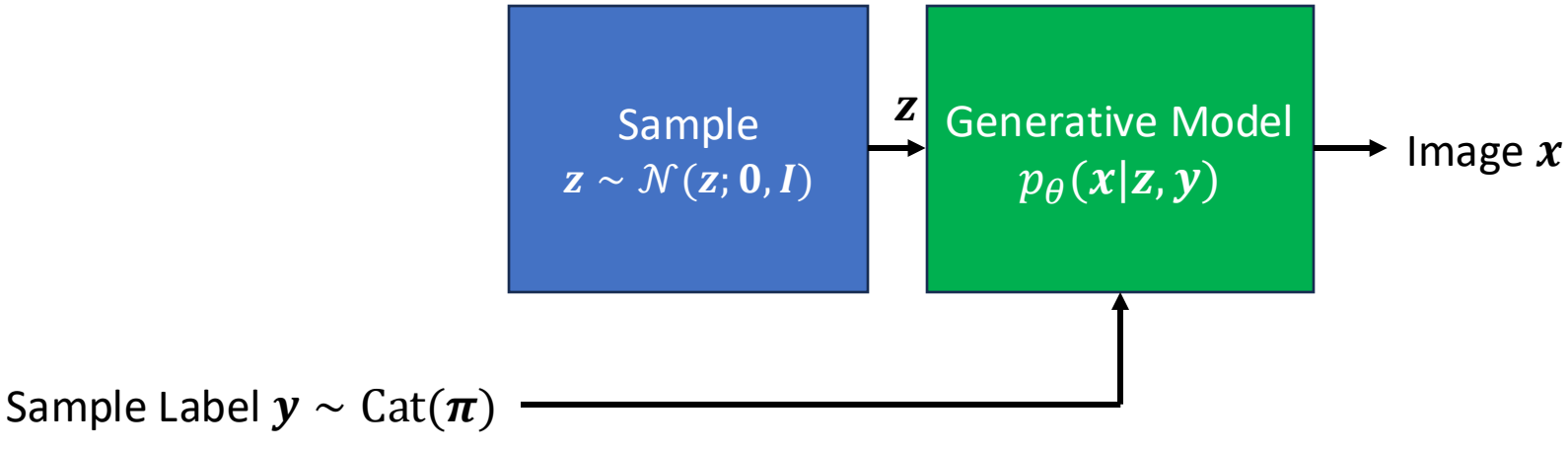
- If $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent given $\boldsymbol{z}$

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



- If $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent

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



Joint VAE: ELBO

- We had that the ELBO is $\max_{\psi} \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}$
- If $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent given $\boldsymbol{z}$: $p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$

$$\text{ELBO} = \max_{\psi} \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} \left[\log \frac{p(\boldsymbol{z})}{q_\psi(\boldsymbol{z} \mid \boldsymbol{x},\boldsymbol{y})} + \log p_\theta(\boldsymbol{y}|\boldsymbol{z}) + \log p_\theta(\boldsymbol{x}|\boldsymbol{z}) \right]$$

$$= \max_{\psi} -\mathrm{KL}\left[q_\psi(\boldsymbol{z} \mid \boldsymbol{x},\boldsymbol{y}) \,||\, p(\boldsymbol{z}) \right] + \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} [\log p_\theta(\boldsymbol{y}|\boldsymbol{z}) + \log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$$

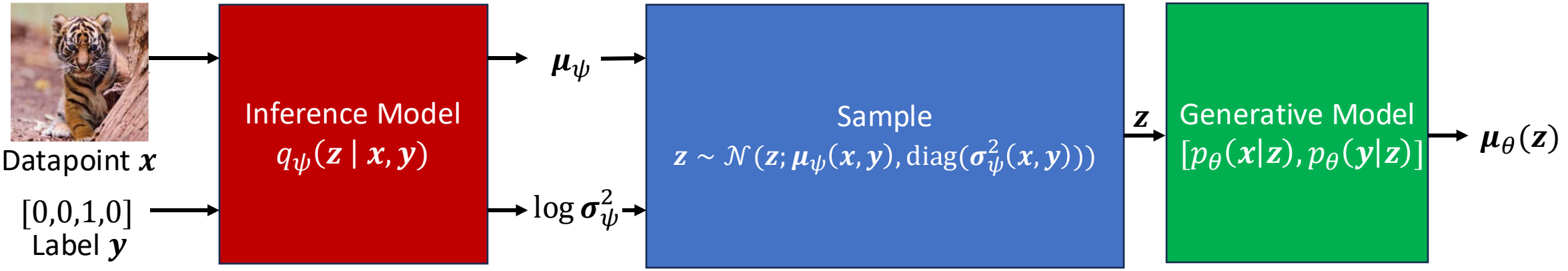
- If $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent: $p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})$

$$\text{ELBO} = \max_{\psi} \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} \left[\log \frac{p(\boldsymbol{z})}{q_\psi(\boldsymbol{z} \mid \boldsymbol{x},\boldsymbol{y})} + \log p_\theta(\boldsymbol{y}) + \log p_\theta(\boldsymbol{x} \mid \boldsymbol{y},\boldsymbol{z}) \right]$$

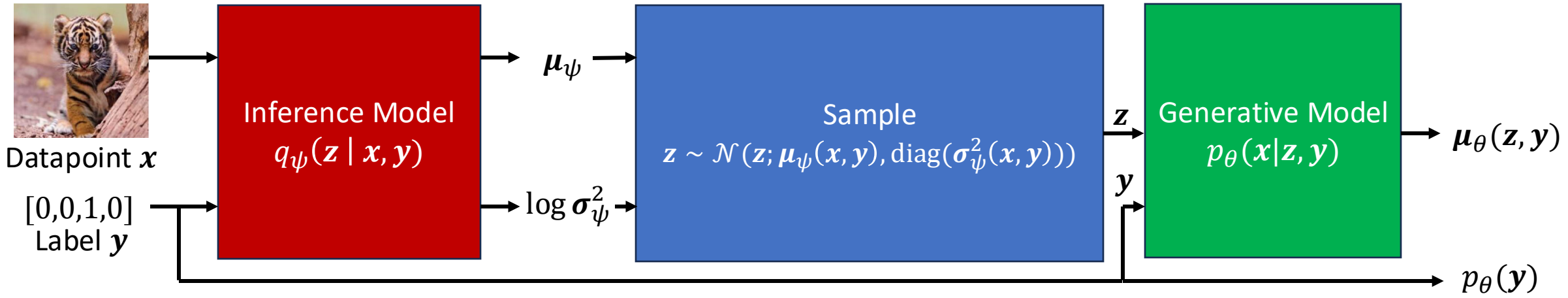
$$= \max_{\psi} -\mathrm{KL}\left[q_\psi(\boldsymbol{z} \mid \boldsymbol{x},\boldsymbol{y}) \,||\, p(\boldsymbol{z}) \right] + \log p_\theta(\boldsymbol{y}) + \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} [\log p_\theta(\boldsymbol{x} \mid \boldsymbol{y},\boldsymbol{z})]$$

Joint VAE: architecture for training

- If $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent given $\boldsymbol{z}$, same process as with regular VAE



- If $\boldsymbol{y}$ is independent of $\boldsymbol{z}$, we also learn a label prior $p_\theta(\boldsymbol{y})$, which we discuss next.



Joint VAE with label prior: Training objective

- The training objective is given by

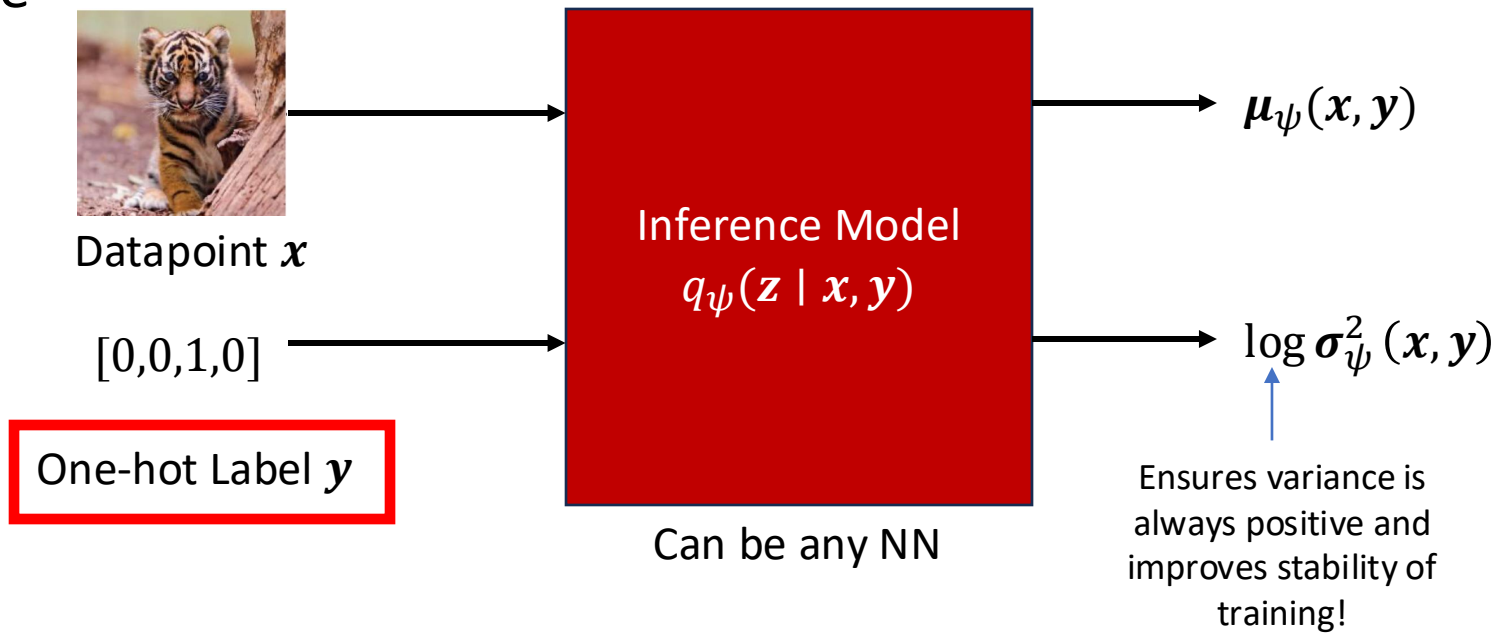
$$\max_{\theta} \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_{\text{data}}} \log p_\theta(\boldsymbol{x},\boldsymbol{y})$$

$$= \max_{\theta}\max_{\psi} \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_{\text{data}}}\left[\mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} \log p_\theta(\boldsymbol{x} \mid \boldsymbol{z},\boldsymbol{y}) - \text{KL}[q_\psi(\boldsymbol{z} \mid \boldsymbol{x},\boldsymbol{y}) \,||\, p(\boldsymbol{z})] + \log p_\theta(\boldsymbol{y})\right]$$

- What are $q_\psi(\boldsymbol{z} \mid \boldsymbol{x},\boldsymbol{y})$, $p_\theta(\boldsymbol{x} \mid \boldsymbol{z},\boldsymbol{y})$?
  - They are similar to what we needed in regular VAEs, with an additional input $\boldsymbol{y}$
- What is $p_\theta(\boldsymbol{y})$?
  - This is a new term denoting the prior probability on class labels
- Answer: next two slides

Joint VAE with label prior: Setup

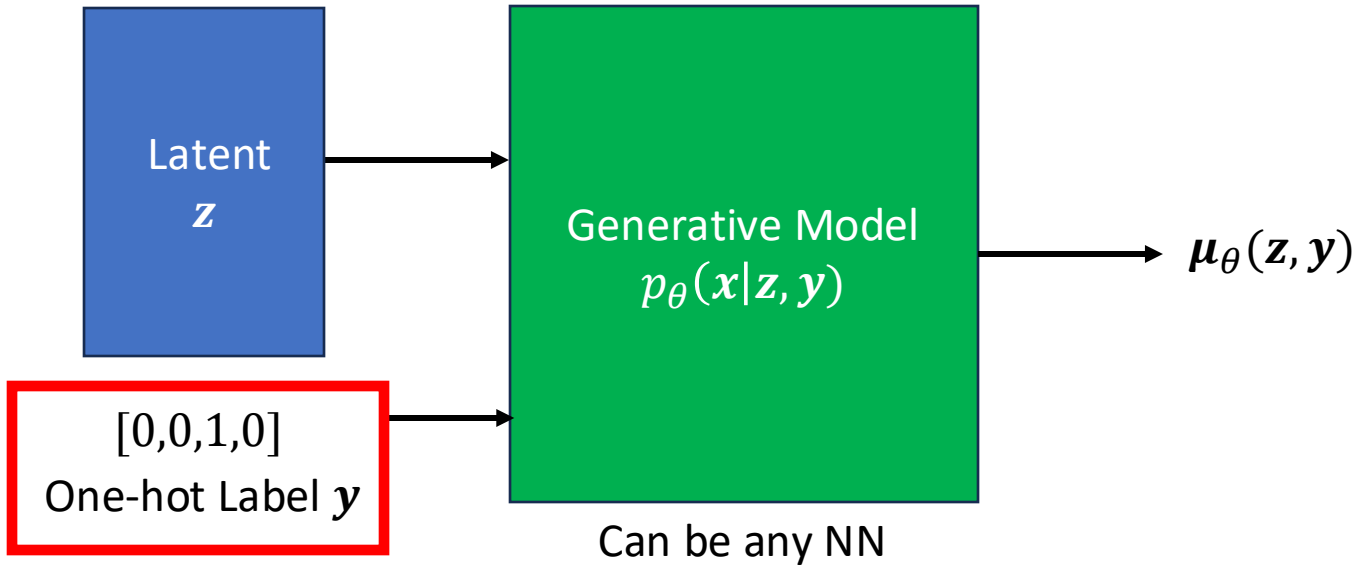
- We have four models we need to define for VAE model
  1. Inference model $q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$: We will define as $q_\psi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{z};\ \boldsymbol{\mu}_\psi(\boldsymbol{x}, \boldsymbol{y}), \mathrm{diag}(\boldsymbol{\sigma}^2_\psi(\boldsymbol{x}, \boldsymbol{y})))$, i.e., a Gaussian distribution with learned mean and covariance



  2. Latent prior: $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$
  3. Class prior: $p_\theta(\boldsymbol{y}) = \mathrm{Cat}(\boldsymbol{\pi})$

Joint VAE with label prior: Setup

- We have four models we need to define for VAE model
  4. Generative model $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y})$: We will define as
     - $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_\theta(\boldsymbol{z}, \boldsymbol{y}), \eta^2 \boldsymbol{I})$, i.e., a Gaussian distribution with learned mean and variance
     - Or $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y}) = \text{Cat}(\boldsymbol{x}; \boldsymbol{\pi}_\theta(\boldsymbol{z}, \boldsymbol{y}))$, 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 256 values of each pixel)

Joint VAE with label prior: Generation

- At inference time, we can do either:
  - Conditional Generation (as in Conditional VAE):
    - Specify a label $\boldsymbol{y}$
    - Sample from $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$
    - Sample from $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y})$
  - Joint Generation:
    - Sample a label $\boldsymbol{y} \sim p_\theta(\boldsymbol{y}) = \text{Cat}(\boldsymbol{\pi})$
    - Sample from $p(\boldsymbol{z})$, which is a Gaussian $\mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$
    - Sample from $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y})$

