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

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 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 $p(x) \ = \int p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z})\, d\boldsymbol{z}$
  - Step 2a: When we assumed $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, I)$, and $p(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; Wz + b, \sigma^2 I)$, we could solve for (W, b, σ^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:

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

  - Step 3a: When the integral is easy to evaluate, we can alternate between optimizing w.r.t. θ with $q(\boldsymbol{x} \mid \boldsymbol{z})$ 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

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

$$\geq \max_{\theta,\psi} \sum_{i} E_{z\sim q_{\psi}(z\,|x_i)} \log \frac{p_{\theta}(x_i, z)}{q_{\psi}(z\,|x_i)}$$

Evidence Lower Bound (ELBO)

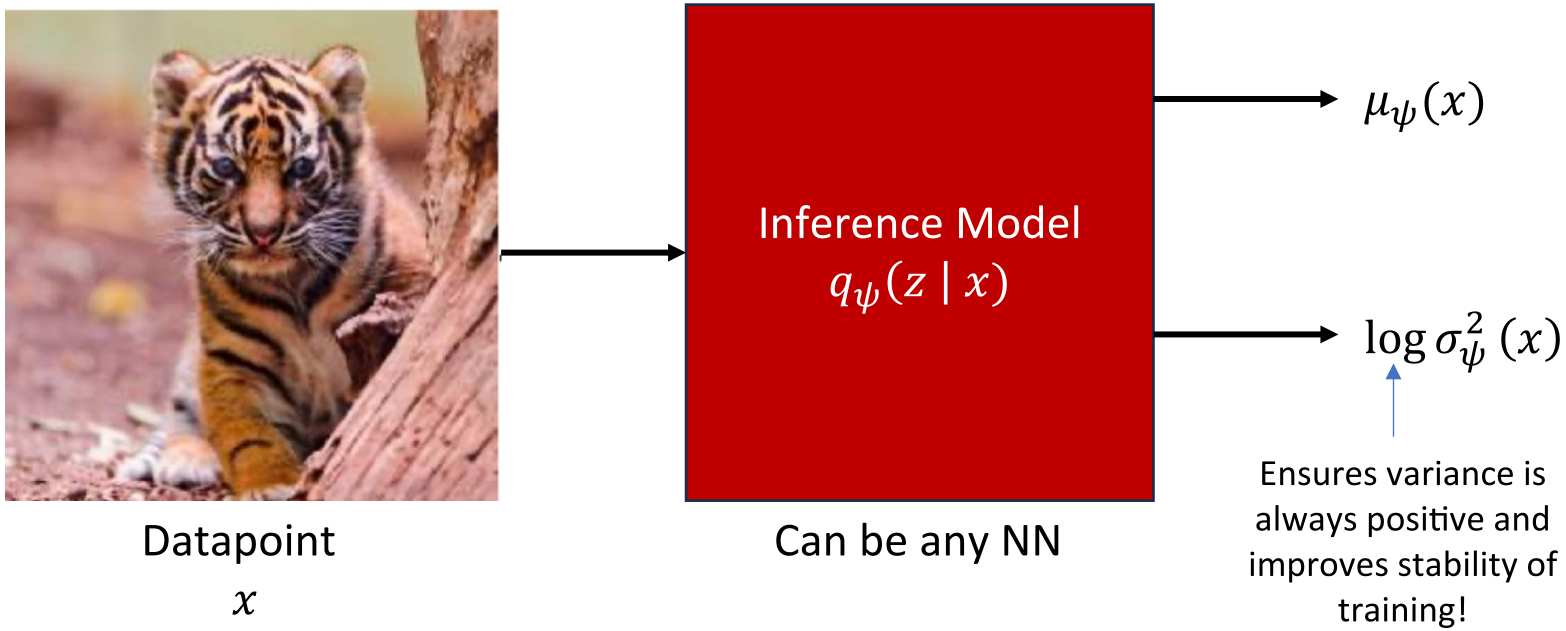
$$= \max_{\theta,\psi} \sum_{i} E_{z\sim q_{\psi}(z\,|x_i)} \log p_{\theta}(x_i|z) \frac{p(z)}{q_{\psi}(z\,|x_i)}$$

$$= \max_{\theta,\psi} \sum_{i} E_{z\sim q_{\psi}(z\,|x_i)} \log p_{\theta}(x_i|z) + E_{z\sim q_{\psi}(z|x_i)} \log \frac{p(z)}{q_{\psi}(z|x_i)}$$

$$-D_{KL}(q_{\psi}(z|x)||p(z))$$

Variational AutoEncoders (VAEs): Setup

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

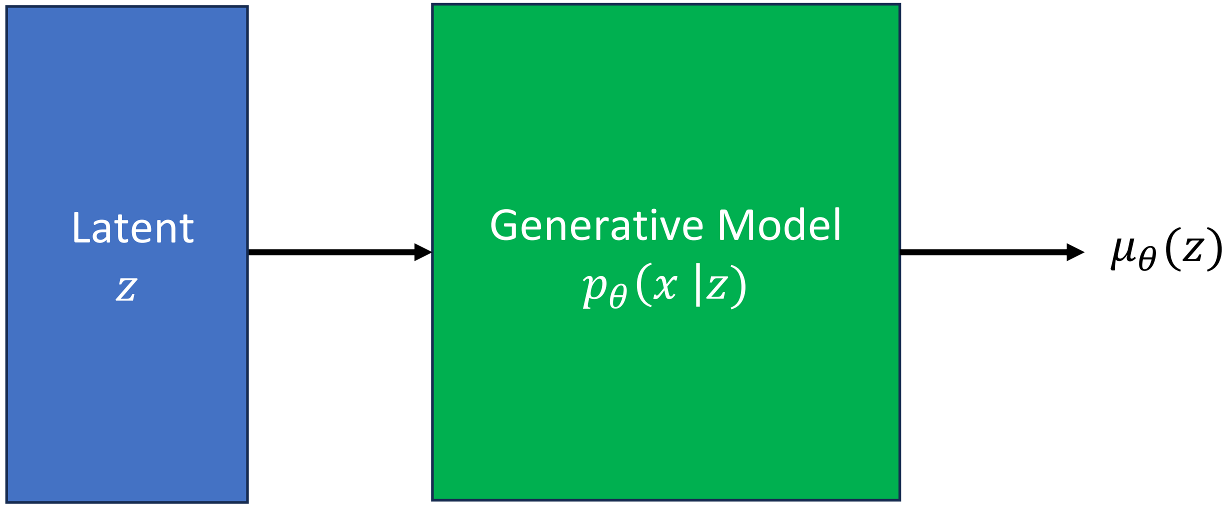


  2. Prior model $p(z)$: We will define prior for latent variables as $p(z) = \mathcal{N}(z; 0, I)$

Variational AutoEncoders (VAEs): Setup

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

3. Generative model $p_\theta(x \mid z)$: We will define as
   - $p_\theta(x \mid z) = \mathcal{N}(z; \mu_\theta(z), \eta^2 I)$, i.e., a normal distribution with learned mean and variance
   - $p_\theta(x \mid z) = Cat(z; \pi_\theta(z))$, i.e., a categorical distribution with learned class probabilities



Can be any NN

- 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

Datapoint x

Inference Model
$q_\psi(z \mid x)$

μ

$\log \sigma^2$

Sample
$z \sim N(z \mid \mu_\psi(x), \sigma_\psi^2(x) I)$

Generative Model
$p_\theta(x \mid z)$

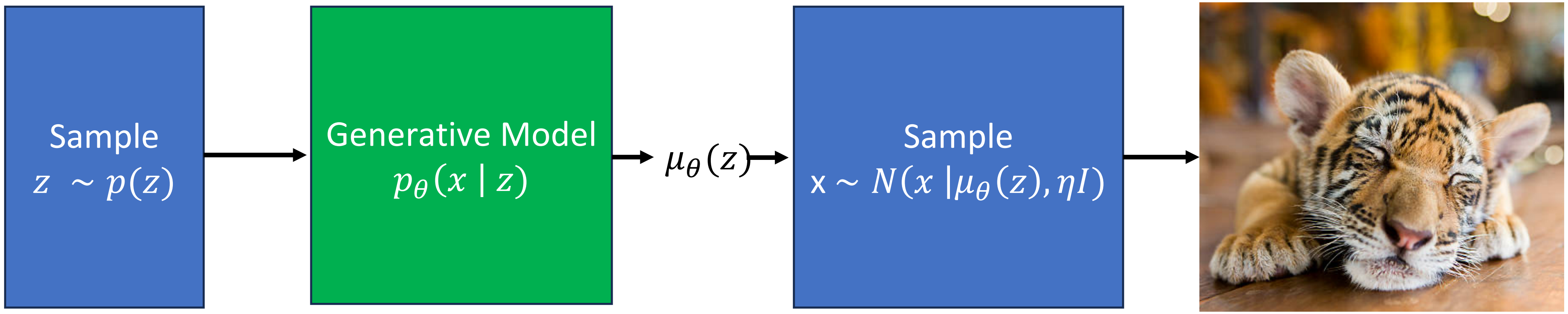
$\mu_\theta(z)$

ELBO Objective

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

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 $p_\theta(x \mid z)$



Computing the ELBO Loss

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

- Term 1 (Reconstruction Error)
  - Because $p_\theta(x|z) = \mathcal{N}(x|\mu_\theta(z), \eta I)$, we have $\log p_\theta(x|z) = -\frac{1}{2\eta}\left||x - \mu_\theta(z)|\right|_2^2 + const.$
  - 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 $z_j \sim q_\psi(Z\mid x)$ we get reconstruction error

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

Computing the ELBO Loss

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

- 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,

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

  - 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 θ, ψ 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 θ

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

$$\nabla_\theta L_{\theta,\psi}(x) = \nabla_\theta E_{z\sim q_\psi(z\,|x)}\left[\log(p_\theta(x\,|z))\right] + \nabla_\theta E_{z\sim q_\psi(z\,|\,x)}\left[\log\left(\frac{p(z)}{q_\psi(z\,|x)}\right)\right]$$

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

$$= -\frac{1}{2\eta} E_{z\sim q_\psi(z\,|x)} \nabla_\theta \left|\left|x - \mu_\theta(z)\right|\right|_2^2$$

- 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 $z_j \sim \mathcal{N}\left(\mu_\psi(x), \sigma_\psi^2(x) I\right)$, find $\mu_\theta\left(z_j\right)$ through decoder
  - Estimate gradient from M samples: $\nabla_\theta L_{\theta,\psi}(x) \approx -\frac{1}{2\eta M}\sum_j \nabla_\theta \left|\left|x - \mu_\theta\left(z_j\right)\right|\right|_2^2$

Stochastic Optimization of ELBO wrt ψ

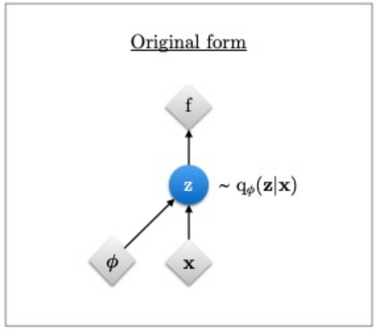
- Here, we cannot just switch gradient and expectation because both are wrt to ψ

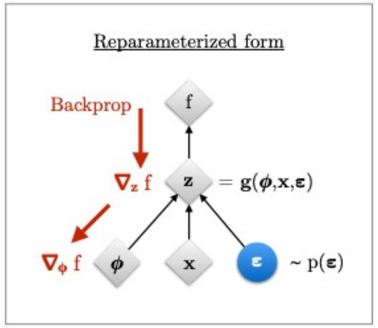
$$\nabla_\psi E_{z\sim q_\psi(z|x)}[\log p_\theta(x\,,z) - \log q_\psi(z\,|x)] \neq E_{z\sim q_\psi(z|x)}\nabla_\psi[\log p_\theta(x,z) - \log q_\psi(z\,|x)]$$

- 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

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

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





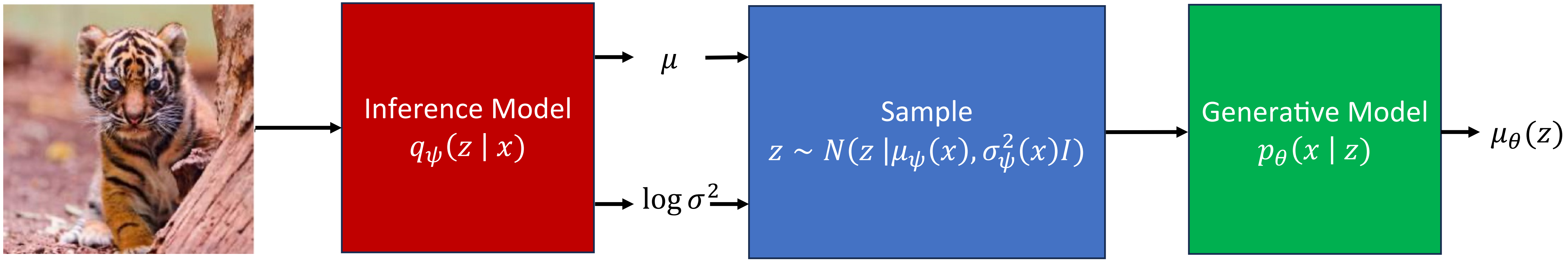
$$\nabla_\psi L_{\theta,\psi}(x) = \nabla_\psi E_{z\sim q_\psi(z\,|x)}[\log p_\theta(x,z) - \log q_\psi(z \mid x)]$$

$$= \nabla_\psi E_{\epsilon\sim N(0,I)}[\log p_\theta\big(x,z_\psi\big) - \log q_\psi(z_\psi \mid x)]$$

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

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 x

VAE's in Action

- $q_\psi(z\,|x) = N(z\,|\mu_\psi(x), \sigma_\psi^2(x))$
- $p(z) = N(z|0, I)$
- $p_\theta(x\,|z) = Categorical\;(x\,|\pi_\theta\;(z))$ Note this is different from the model considered up till now!

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

- Decoder network: $z \in R^d$ -> Linear(d, 256) -> LeakyReLU -> Linear(256, D) -> softmax

VAE’s for Generation of MNIST Digits

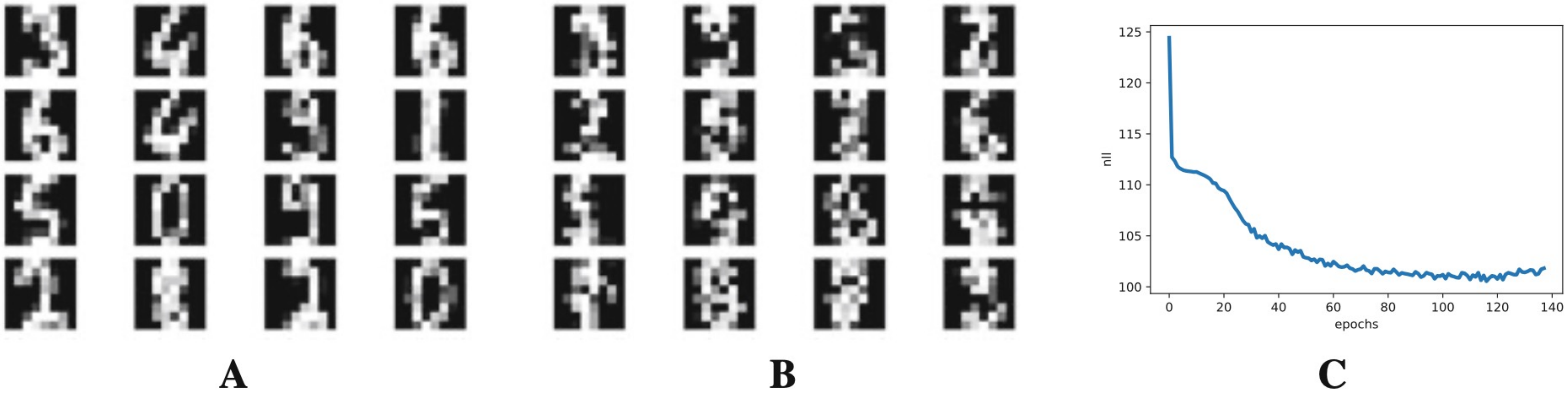


Fig. 4.4 An example of outcomes after the training: (**a**) Randomly selected real images. (**b**) Unconditional generations from the VAE. (**c**) The validation curve during training