Course: **Generative AI Models (2AMU20)** *Responsible Lecturer:* Jakub Tomczak

Report Assignment 2

Name: Luca Mainardi, *Student ID:* 2014602

Please provide (concise) answers to the questions below. If you do not know an answer, please leave it blank. If necessary, please provide a (relevant) code snippet. If relevant, please remember to support your claims with data/figures.

Question 1 (1pt)

Please explain the reparameterization trick and provide a mathematical formula for it.

Answer The reparameterization trick is a technique used in variational autoencoders (VAEs) to allow the gradient of a stochastic variable to be propagated through a stochastic sampling process during backpropagation. This is crucial for optimizing the parameters of the model using gradient descent because, without this trick, it would be difficult to update the model weights using backpropagation, since, the sampling process is non-differentiable. In VAEs, we aim to approximate a posterior distribution $q(z|x)$, which is often modelled as a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ Kingma and Welling (2013). However, direct sampling from this distribution does not allow gradient propagation through the network. The Reparameterization Trick reformulates the sampling process to make it differentiable.

Mathematical Formulation: The Reparameterization Trick rewrites this sampling as:

$$z = \mu + \sigma \cdot \epsilon$$

where $\epsilon \sim \mathcal{N}(0, 1)$ is standard Gaussian noise Rezende et al. (2014).

In this formulation, the random variable z follows a normal distribution $\mathcal{N}(\mu, \sigma^2)$ parameterized by mean μ and standard deviation σ which are deterministic and differentiable, while ϵ is an additional independent random variable. This allows us to rewrite the sampling as a differentiable function of the parameters μ and σ, enabling the application of gradient-based optimisation methods.

Details

- **Latent Space Representation**: In a VAE, the encoder network maps input data x to the parameters of the latent distribution, typically the mean μ and the standard deviation σ. Mathematically, this can be represented as:
$$\mu, \log(\sigma^2) = f_{\text{encoder}}(x)$$
where f_{encoder} is the encoder neural network.
- **Sampling**: Instead of sampling z directly from $\mathcal{N}(\mu, \sigma^2)$, we sample ϵ from a standard normal distribution $\mathcal{N}(0, 1)$. Then, we reparameterize z as:
$$z = \mu + \sigma \cdot \epsilon$$
- **Differentiability**: This reparameterization allows the gradient to propagate through μ and σ, as the sampling process is now a differentiable operation.

By using this trick, the gradients can now flow through μ and σ during backpropagation, allowing us to optimize the encoder network parameters Kingma and Welling (2019).

Question 2 (2pts)

Please write down mathematically the log-probability of the encoder (variational posterior) for a Gaussian distribution with a diagonal covariance matrix.

Answer In a Variational Autoencoder (VAE), we approximate the true posterior $p(z|x)$ with a variational posterior $q(z|x)$, which we assume to be a Gaussian distribution with a diagonal covariance matrix. Mathematically, this can be expressed as:

$$q(z|x) = \mathcal{N}(z; \mu(x), \Sigma(x))$$

where $\mu(x)$ is the mean vector and $\Sigma(x)$ is the diagonal covariance matrix Bishop (2006).

The log-probability of the encoder (variational posterior) is given by the logarithm of the probability density function of the multivariate Gaussian distribution with a diagonal covariance matrix Tomczak (2022):

$$\mathcal{N}(z; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right)$$

Since Σ is diagonal, let $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$, where d is the dimensionality of the latent space. The determinant $|\Sigma|$ is the product of its diagonal elements, and the inverse Σ^{-1} is the diagonal matrix with elements $1/\sigma_i^2$. Thus, the log-probability is:

$$\log q(z|x) = \log\left(\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right)\right)$$

Breaking this down, we get:

$$\log q(z|x) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)$$

Since Σ is diagonal:

$$\log|\Sigma| = \log\left(\prod_{i=1}^{d}\sigma_i^2\right) = \sum_{i=1}^{d}\log(\sigma_i^2) \quad \text{and} \quad (z-\mu)^T\Sigma^{-1}(z-\mu) = \sum_{i=1}^{d}\frac{(z_i-\mu_i)^2}{\sigma_i^2}$$

Putting these together, we have:

$$\log q(z|x) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{d}\log(\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{d}\frac{(z_i-\mu_i)^2}{\sigma_i^2}$$

Question 3 (4pts)

Please do the following:

1. *(1pt) Please explain your choice of the distribution (the conditional likelihood) for image data used in this assignment. Please remember to motivate it properly!*
2. *(1pt) Please write the conditional likelihood down mathematically (i.e., present it as the log-probability).*
3. *(2pts) Please explain how one can sample from the distribution chosen by you. Please provide a mathematical formula and a code snippet.*

Answer

1. In the context of generative models, specifically Variational Autoencoders (VAEs), conditional likelihood is a crucial component. It represents the probability of the observed data given the latent variables. Mathematically, the conditional likelihood of a data point $\mathbf{x}$ given a latent variable $\mathbf{z}$ is denoted as $p(\mathbf{x}|\mathbf{z})$. We often use the log-probability form to work with this in optimization frameworks. In this assignment, the chosen distribution for the conditional likelihood $p(x|z)$ when dealing with image data is a Gaussian distribution. This choice is based on several key motivations outlined below:

   (a) **Natural Fit for Continuous Data**: Image pixel values are continuous, ranging from 0 to 255 in standard 8-bit grayscale images or 0.0 to 1.0 when normalized. The Gaussian distribution is naturally suited to model continuous data due to its ability to represent data spread (variance) and central tendency (mean).

   (b) **Flexibility**: The Gaussian distribution is highly flexible, capable of modelling a wide range of data distributions by adjusting its mean (μ) and variance (σ^2). This flexibility allows the model to learn the underlying data distribution more effectively.

(c) **Reparameterization Trick**: The Gaussian distribution facilitates the use of the reparameterization trick, essential for backpropagation in VAEs.

(d) **Mathematical Tractability**: The Gaussian distribution is mathematically tractable, allowing for analytical solutions to the Kullback-Leibler (KL) divergence term in the VAE loss function. This tractability simplifies the implementation and optimization of the VAE model.

2. The conditional likelihood in the context of Variational Autoencoders (VAEs) for continuous image data, modelled as a Gaussian distribution, can be written down mathematically as follows:

   Given:

   - x: the observed data (image data).
   - z: the latent variable.
   - $\mu(z)$: the mean of the Gaussian distribution, parameterized by the decoder network.
   - σ^2: the variance of the Gaussian distribution, often assumed to be a constant or a diagonal covariance matrix for simplicity.

   The conditional likelihood $p(x|z)$ is:

   $$p(x|z) = \mathcal{N}(x; \mu(z), \sigma^2 I)$$

   The log-likelihood for a single data point x given the latent variable z is:

   $$\log p(x|z) = -\frac{1}{2}\left(\log(2\pi\sigma^2) + \frac{(x-\mu(z))^2}{\sigma^2}\right)$$

   When dealing with a dataset $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ consisting of N samples, we sum the log-probabilities over all data points to obtain the total log-likelihood is:

   $$\log p(X|Z) = \sum_{i=1}^{N} \log p(x_i|z_i)$$

   where X and Z represent the sets of all data points and their corresponding latent variables, respectively.

   This summation is a key term in the Evidence Lower Bound (ELBO) objective, which VAEs maximize during training. Maximizing the ELBO ensures that the generative model produces data that closely matches the true data distribution while maintaining a well-structured latent space.

3. To sample from the learned latent distribution in a Variational Autoencoder (VAE), we typically assume a Gaussian distribution for the latent variables $\mathbf{z}$. The encoder network maps an input $\mathbf{x}$ to the parameters of this Gaussian distribution: the mean μ and the standard deviation σ.

   Mathematical Formulation: In a Variational Autoencoder (VAE), the variational posterior $q(z|x)$ is modeled as a Gaussian distribution with a mean vector $\mu(x)$ and a diagonal covariance matrix represented by the standard deviation vector $\sigma(x)$. Mathematically, this is expressed as:

   $$q(z|x) = \mathcal{N}(z; \mu(x), \text{diag}(\sigma^2(x))).$$

   To sample from this distribution, we use the reparameterization trick, which allows us to backpropagate through the sampling operation. The reparameterization trick expresses the sampling operation as:

   $$z = \mu(x) + \sigma(x) \odot \epsilon,$$

   where $\epsilon \sim \mathcal{N}(0, I)$ (a standard normal distribution), and $\odot$ denotes element-wise multiplication.

Encorder Class

```
    class Encoder(nn.Module):
    def __init__(self, encoder_net):
        super(Encoder, self).__init__()

        # The init of the encoder network.
        self.encoder = encoder_net

    # The reparameterization trick for Gaussians.
    @staticmethod
    def reparameterization(mu, log_var):
        # The formulat is the following:
        # z = mu + std * epsilon
        # epsilon ~ Normal(0,1)

        # First, we need to get std from log-variance.
        std = torch.exp(0.5*log_var)

        # Second, we sample epsilon from Normal(0,1).
        eps = torch.randn_like(std)

        # The final output
        return mu + std * eps

    # This function implements the output of the encoder network (i.e., parameters
    of a Gaussian).
    def encode(self, x):
        # First, we calculate the output of the encoder netowork of size 2M.
        h_e = self.encoder(x)
        # Second, we must divide the output to the mean and the log-variance.
        mu_e, log_var_e = torch.chunk(h_e, 2, dim=1)

        return mu_e, log_var_e

    # Sampling procedure.
    def sample(self, x=None, mu_e=None, log_var_e=None):
        #If we don't provide a mean and a log-variance, we must first calcuate it:
        if (mu_e is None) and (log_var_e is None):
            mu_e, log_var_e = self.encode(x)
        # Or the final sample
        else:
        # Otherwise, we can simply apply the reparameterization trick!
            if (mu_e is None) or (log_var_e is None):
                raise ValueError('mu and log-var can't be None!')
        z = self.reparameterization(mu_e, log_var_e)
        return z

    # This function calculates the log-probability that is later used for
    calculating the ELBO.
    def log_prob(self, x=None, mu_e=None, log_var_e=None, z=None):
        # If we provide x alone, then we can calculate a corresponsing sample:
        if x is not None:
            mu_e, log_var_e = self.encode(x)
            z = self.sample(mu_e=mu_e, log_var_e=log_var_e)
        else:
        # Otherwise, we should provide mu, log-var and z!
            if (mu_e is None) or (log_var_e is None) or (z is None):
                raise ValueError('mu, log-var and z can't be None!')

        return log_normal_diag(z, mu_e, log_var_e)

    # PyTorch forward pass: it is either log-probability (by default) or sampling.
    def forward(self, x, type='log_prob'):
        assert type in ['encode', 'log_prob'], 'Type could be either encode or
    log_prob'
        if type == 'log_prob':
            return self.log_prob(x)
        else:
            return self.sample(x)
```

Explanation

(a) **Compute the Standard Deviation**: Since the encoder outputs the log-variance ($\log \sigma^2$), we compute the standard deviation by taking the exponential of half the log-variance.

(b) **Sample from Standard Normal**: ϵ is sampled from a standard normal distribution using `torch.randn_like(std)`, which generates a tensor of random values with the same shape as `std`.

(c) **Reparameterization**: The sample $\mathbf{z}$ is obtained using the reparameterization trick: $\mathbf{z} = \mu + \sigma \cdot \epsilon$.

This approach ensures that the sampling process is differentiable, allowing gradients to propagate through the sampling operation during backpropagation. This is crucial for training the VAE using gradient-based optimization techniques Tomczak (2022).

Question 4 (3pts)

Please do the following:

1. *(1pt) Please explain your prior and write it down mathematically.*
2. *(1pt) Please write down its sampling procedure (incl. a code snippet).*
3. *(1pt) Please write down its log-probability (a mathematical formula).*

Answer

1. In this assignment, the prior distribution $p(z)$ for the latent variable z is modelled as a Mixture of Gaussians (MoG). This prior is more flexible than a standard Gaussian, allowing the model to capture more complex latent structures, indeed, it uses K different components to model a wider range of distributions, including multi-modal distributions. In addition, the MoG prior can act as a regularizer that encourages the latent representations to be well-separated according to the modes of the mixture, this helps prevent overfitting, improve generalization and consequently the generative capabilities of the VAE Tomczak (2022).

   Mathematical Formulation:

   The MoG prior can be mathematically expressed as a weighted sum of K Gaussian components. Each component k has its own mean μ_k and variance σ_k^2, and the mixture is weighted by the mixing coefficients ω_k. Formally, the probability density function of the MoG prior is:

$$p_\lambda(z) = \sum_{k=1}^{K} \omega_k \mathcal{N}(z|\mu_k, \sigma_k^2)$$

   where:

   - K: Number of components
   - μ_k: Means of the Gaussian components
   - σ_k^2: Variances of the Gaussian components (computed as exp(logvars))
   - ω_k: Mixing weights (computed as softmax of `self.w`)
   - $\lambda = \{\{\omega_k\}, \{\mu_k\}, \{\sigma_k^2\}\}$ are trainable parameters.

2. The function `sample` samples a batch of latent variables z_i from the MoG prior. It first samples component indices based on the mixture weights ω_k, and then samples from the corresponding Gaussian distributions.

   (a) **Initialization**: The `means` and `logvars` for each Gaussian component are initialized. While the `w` parameter is used to compute the mixture weights ω_k.

   (b) **Compute Mixture Probabilities**: The mixture weights are computed by applying the softmax function to `self.w` to get the probabilities π_k.

   (c) **Sample Component Indices**: Based on the mixture weights π_k, sample indices to determine from which Gaussian component each sample should be drawn.

   (d) **Sample from the Selected Gaussian Components**: For each selected component, it samples from the corresponding Gaussian distribution using its mean and variance.

Mathematically:

$$z_i \sim \mathcal{N}(\mu_{c_i}, \sigma^2_{c_i})$$

where c_i is the component index sampled from $\{1, \ldots, K\}$ with probability ω_k.

Sampling function

```
def sample(self, batch_size):
    # initialization
    means, logvars = self.get_params()

    # compute mixing probabilities
    w = F.softmax(self.w, dim=0)
    w = w.squeeze()

    # sample component indices
    indexes = torch.multinomial(w, batch_size, replacement=True)

    # Initialize the noise epsilon for sampling
    eps = torch.randn(batch_size, self.L)
    for i in range(batch_size):
        indx = indexes[i]
        m = means[[indx]].to('cuda')
        e = eps[[i]].to('cuda')
        l = torch.exp(logvars[[indx]]).to('cuda')
        #sample from the selected gaussian components
        if i == 0:
            z = m + e * l
        else:
            z = torch.cat((z, m + e * l), 0)
    return z
```

Where `z` represents the sample from the Gaussian component indexed by `indx`.

3. The `log_prob` function calculates the log-probability of a given latent variable z under the MoG prior. This involves calculating the log-probability under each Gaussian component and then combining them using the log-sum-exp trick to ensure numerical stability.

   Mathematically:

$$\log p_\lambda(z) = \log \left(\sum_{k=1}^{K} \omega_k \exp \left(\log \mathcal{N}(z|\mu_k, \sigma_k^2) \right) \right)$$

   where:

$$\log \mathcal{N}(z|\mu_k, \sigma_k^2) = -\frac{1}{2} \left(\log(2\pi) + \log(\sigma_k^2) + \frac{(z-\mu_k)^2}{\sigma_k^2} \right)$$

Question 5 (1pt)

Please derive the Negative ELBO including intermediate steps (be as specific as possible).

Answer In a Variational Autoencoder (VAE), we approximate the true posterior $p(z|x)$ with a variational posterior $q(z|x)$, which we assume to be a Gaussian distribution with a diagonal covariance matrix. The goal is to maximize the log-likelihood of the observed data x. This log-likelihood can be decomposed using the Evidence Lower Bound (ELBO) Doersch (2016); Rezende et al. (2014).

1. **Marginal Log-Likelihood**

$$\log p(x) = \log \int p(x, z)\, dz, \tag{0.1}$$

   where z is the latent variable.

2. **Introducing the Variational Distribution** $q(z|x)$

$$\log p(x) = \log \int \frac{p(x, z)}{q(z|x)} q(z|x)\, dz. \tag{0.2}$$

3. **Applying Jensen's Inequality**

$$\log p(x) \geq \mathbb{E}_{q(z|x)} \left[\log \frac{p(x,z)}{q(z|x)} \right]. \tag{0.3}$$

4. **Rewriting the Joint Distribution**

$$\log p(x) \geq \mathbb{E}_{q(z|x)}[\log p(x|z) + \log p(z) - \log q(z|x)]. \tag{0.4}$$

5. **Separating the Expectation Terms**

$$\text{ELBO} = \mathbb{E}_{q(z|x)}[\log p(x|z)] + \mathbb{E}_{q(z|x)}[\log p(z) - \log q(z|x)]. \tag{0.5}$$

6. **Recognizing the KL Divergence**

$$\text{ELBO} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - \text{KL}(q(z|x)\|p(z)), \tag{0.6}$$

where the KL divergence is defined as:

$$\text{KL}(q(z|x)\|p(z)) = \mathbb{E}_{q(z|x)} \left[\log \frac{q(z|x)}{p(z)} \right]. \tag{0.7}$$

7. **Final Form of the Negative ELBO**:

$$-\text{ELBO} = -\mathbb{E}_{q(z|x)}[\log p(x|z)] + \text{KL}(q(z|x)\|p(z)). \tag{0.8}$$

The first term represents the reconstruction error which measures how well the VAE can reconstruct the data x from the latent representation z, and the second term is a regularization term that ensures the learned distribution $q(z|x)$ is close to the prior $p(z)$.

Question 6 (1pt)

Please explain your choice of the optimizer, and comment on the choice of the hyperparameters (e.g., the learning rate value).

Answer For training our Variational Autoencoder (VAE), we selected the `Adamax` optimizer from the `torch.optim` library. This choice is motivated by the following reasons:

- **Adaptive Learning Rate:** Like Adam, `Adamax` incorporates adaptive learning rate adjustments for each parameter. This feature is particularly advantageous when dealing with high-dimensional data or models where different parameters exhibit gradients of varying magnitudes. The adaptive nature helps in stabilizing the training process and speeds up convergence Kingma and Ba (2014).
- **Handling Sparse Gradients:** `Adamax` performs well with sparse gradients, a common occurrence in large-scale neural networks. This robustness arises from its use of the infinity norm ($\|\cdot\|_\infty$) in the denominator of the update rule, which can effectively manage the sparsity and ensure more consistent updates across all parameters Reddi et al. (2018).
- **Empirical Performance:** Empirical studies and past research have demonstrated that `Adamax` can outperform other optimizers in terms of both convergence speed and final model performance for certain types of neural networks, including VAEs. Its stability and reliability make it a suitable choice for complex models.

Optimizer

```
lr = 1e-3 # learning rate
optimizer = torch.optim.Adamax([p for p in model.parameters() if p.requires_grad == True
    ], lr=lr)
```

The only hyperparameter we focused on was the learning rate. For `Adamax`, we set the learning rate (`lr`) to 1×10^{-3}. This value was chosen based on theoretical considerations and empirical tuning.

- **Theoretical Considerations:** The default learning rate for `Adamax` in the literature and many practical implementations is 0.002. However, we opted for a slightly lower value to ensure more stable updates during the initial training epochs, which can be crucial for the convergence of VAEs. Indeed, the theoretical guidelines suggesting that the learning rate should be small enough to ensure convergence but large enough to escape local minima efficiently
- **Empirical Tuning:** Through cross-validation and experimentation, we observed that a learning rate of 1×10^{-3} provided a good balance between convergence speed and model stability. Higher learning rates tended to cause divergence or unstable training dynamics, while lower learning rates significantly increased the time to convergence without noticeable improvements in model performance.

Question 7 (1pt)

Please show 16 real images and the final 16 generated images from a fully-trained model. By looking at the generations, was the model properly trained? Please motivate your answer well.

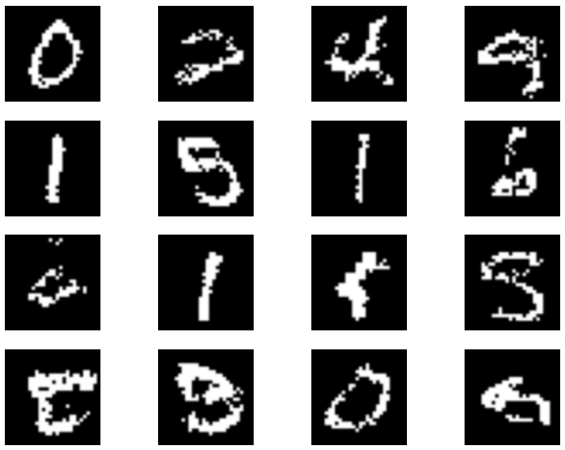

Figure 1: Generated Images

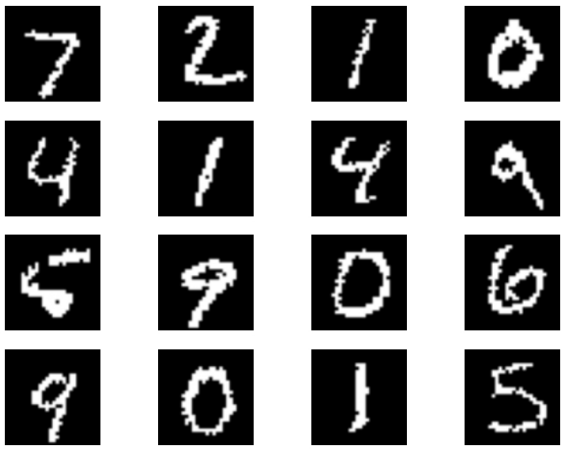

Figure 2: Real Images

```
FINAL LOSS: nll=92.0731965576172
```

Answer By looking at the generations, we can easily deduce that the model wasn't properly trained. Indeed, after comparing the generated images with the real ones, we can make some considerations to support our claim. First of all, it emerges that some of the digits generated by the VAE model are recognizable and visually similar to hand-written digits, however, this does not apply to all images, in fact only about 8 out of 16 show an acceptable quality. This indicates that the model has a partial ability to learn and generate realistic figures, but has failed to generalize completely. Moreover, the generated images do not show a wide variety of styles and shapes, suggesting a limited and less diversified latent representation. Further confirmation of our assessment is the final negative log-likelihood (NLL) of 92.0731965576172, which is relatively high, indicating significant inaccuracies in the reconstruction of images and a discrepancy between the distribution of the original data and that learned from the model. These findings suggest that the model, while capable of generating some realistic images, fails to maintain sufficient consistency with the original input and does not adequately capture the variety of digits present in the MNIST dataset. This may be due to problems in model structure, insufficient training, or latent representation that is not adequately constrained.

Question 8 (3pts)

Please provide answers to the following questions:

1. *(1pt) What are the potential problems with evaluating generative models by looking at generated data?*
2. *(2pts) How can we evaluate the perceptual quality of generative models (NOTE: ELBO or NLL do not count as answers)? Please provide two specific quantitative metrics (mathematical formulas and explanations). At most two, 1pt per metric.*

Answer

1. Evaluating generative models solely based on the inspection of generated data can present several significant challenges:

   - **Subjectivity:** Visual inspection relies heavily on human judgment, which is inherently subjective. Different observers may have varying opinions on what constitutes a "good" sample.
   - **Bias:** Evaluators may possess inherent biases based on their expectations or previous experiences with similar data, leading to inconsistent evaluations.
   - **Time-Consuming:** Evaluating large datasets by visual inspection is impractical and time-consuming. It is not feasible to assess thousands or millions of samples manually.
   - **Limited Sample Size:** Human evaluators can only assess a limited number of samples, which may not be representative of the overall performance of the model.
   - **Lack of Standard Metrics:** Visual inspection does not provide a quantitative metric that can be used to objectively compare different models. This makes it difficult to benchmark performance across different models or improvements over time.
   - **Reproducibility:** Quantitative metrics are essential for reproducibility in scientific research. Without them, it is challenging to replicate results and verify claims.
   - **Mode Collapse:** Generative models can suffer from mode collapse, where the model generates a limited variety of samples, missing out on other modes in the data distribution. Visual inspection might not detect this issue if the evaluator does not see the full range of diversity in the data.
   - **Misleading Samples:** A small set of visually inspected samples might look good, but the overall distribution might still be flawed.
   - **Overfitting to Human Preferences:** If models are tuned based on visual inspection, there is a risk of overfitting to human preferences rather than improving the underlying generative process. This can lead to models that produce visually pleasing but statistically invalid samples.

2. To evaluate the perceptual quality of generative models, two commonly used quantitative metrics are the Inception Score (IS) and the Fréchet Inception Distance (FID). These metrics focus on the visual quality and diversity of the generated samples.

   Inception Score (IS) The Inception Score uses a pre-trained Inception v3 model to assess the quality of generated images. The score considers two main aspects: the clarity of individual samples and the diversity across multiple samples Barratt and Sharma (2018). It is defined as follows:

   $$IS(G) = \exp\left(\mathbb{E}_{\mathbf{x}\sim p_g}\left[\mathrm{KL}(p(y|\mathbf{x})\|p(y))\right]\right)$$

   where:

   - G is the generative model.
   - $\mathbf{x}$ represents the generated samples.
   - p_g is the distribution of generated samples.
   - $p(y|\mathbf{x})$ is the conditional label distribution given by the Inception v3 network for a generated image $\mathbf{x}$.
   - $p(y) = \int p(y|\mathbf{x})p_g(\mathbf{x})d\mathbf{x}$ is the marginal distribution over all generated images.
   - $\mathrm{KL}(p\|q)$ represents the Kullback-Leibler divergence between distributions p and q.

   Explanation:

   - The Inception Score rewards generative models that produce images which are easily classifiable into a single category (high $p(y|\mathbf{x})$) and that also generate a diverse set of images covering all possible categories (high entropy of $p(y)$).
   - A high IS indicates both high quality and high diversity in the generated images.

Fréchet Inception Distance (FID) The Fréchet Inception Distance measures the distance between the distributions of real and generated images in the feature space of a pre-trained Inception v3 model Heusel et al. (2017). It is defined as follows:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$$

where:

- μ_r and μ_g are the means of the feature vectors of the real and generated samples, respectively.
- Σ_r and Σ_g are the covariance matrices of the feature vectors of the real and generated samples, respectively.
- $\|\cdot\|_2$ represents the Euclidean norm.
- $\text{Tr}(\cdot)$ denotes the trace of a matrix.

Explanation:

- The FID score measures the distance between the multivariate Gaussian distributions fitted to the real and generated data in the feature space of the Inception network.
- A lower FID indicates that the distributions of real and generated images are more similar, implying that the generated images are of higher perceptual quality and more closely resemble the real data distribution.

Both Inception Score and Fréchet Inception Distance provide quantitative ways to evaluate the perceptual quality of generative models, focusing on the clarity, diversity, and similarity to real data. The IS captures the quality and diversity of individual images, while the FID provides a holistic measure of the overall similarity between real and generated data distributions.

Question 9 (4pt)

After training, a validation loss curve is plotted. Please comment on the following:

1. *(1pt) Please add the plot. Based on that plot, can you say that the training of your VAE is stable or unstable? Why?*
2. *(3pts) Please run your model with three different sets of hyperparameters. Are the hyperparameter values of the optimizer important and how do they influence the training? Motivate well your answer (e.g., run the script with three values of the learning rate, present three plots here and provide a discussion).*

Answer

1. Observing the trend of the negative log-likelihood (NLL) during the training of the model we can observe a clear tendency to decrease rapidly in the first 50 epochs where the NLL decreases rapidly from about 135 to about 100, This indicates that the model is learning effectively and is rapidly improving its ability to reconstruct input images. Then, it is followed by a period of slow and constant reduction, until stabilizing around a value of about 92 after about 200 epochs but without significant fluctuations or sudden increases, the NLL remains relatively stable throughout the training. This is a positive indicator of stability since it means that the model is not developing numerical instability or difficulty, therefore, we can conclude that the VAE training is stable.

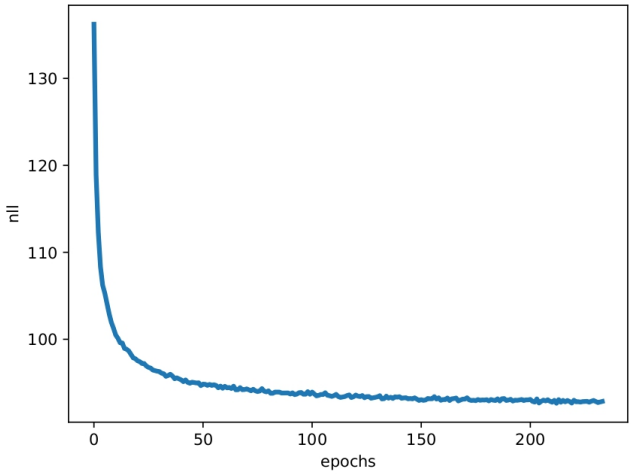

Figure 3: `learning_rate = 1e-3`

2. The hyperparameters values of the optimizer, especially the learning rate, play a crucial role in training a machine learning model. The learning rate determines the size of the steps the optimizer takes during gradient descent, directly influencing the quality, speed and stability of the model's convergence. Therefore, the selection of a balanced learning rate is essential to ensure that the model can generate good results but above all learn effectively and stably.

   Here, we evaluate the influence of the learning rate values on training a Variational Autoencoder (VAE), presenting and comparing the negative log-likelihood (NLL) plots obtained after the training of two new models using respectively learning rates equal to 1×10^{-5}, and 1×10^{-2} with the initial model in terms of *Speed, Stability and Convergence*.

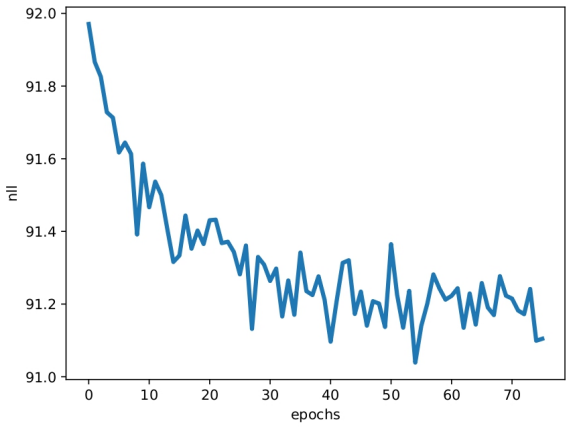

Figure 4: `learning_rate = 1e-5`

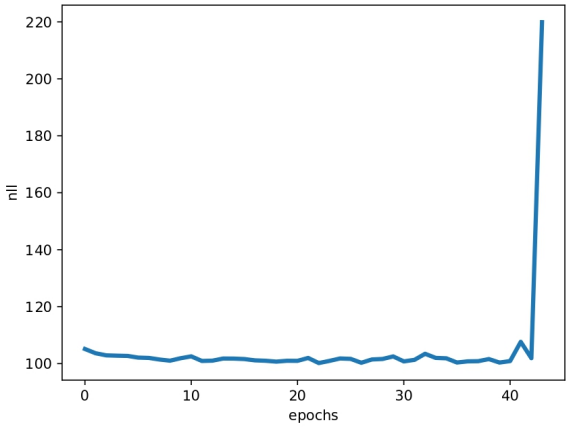

Figure 5: `learning_rate = 1e-2`

Influence of Learning Rate on training

- **Learning Rate** $= 1 \times 10^{-3}$:
  - **Speed**: A learning rate of 1×10^{-3} allows the model to learn quickly initially, as indicated by the rapid drop in NLL during the first 50 epochs.
  - **Stability**: After the initial phase, the NLL decreases slowly and stabilizes. This suggests the model can make fine adjustments without significant fluctuations, indicating good training stability.
  - **Convergence**: The stabilization of NLL indicates that the model is converging towards an optimal solution.
- **Learning Rate** $= 1 \times 10^{-5}$:

  - **Speed**: A learning rate of 1×10^{-5} is too low, resulting in extremely slow learning. This can be observed from the fact that the NLL does not significantly decrease even after many epochs.
  - **Stability**: Although the training appears stable, the slow learning rate is counterproductive, as it requires an excessive number of epochs to reach a good solution.
  - **Convergence**: The risk with such a low learning rate is that the model may get stuck in local minima or simply not reach an optimal solution in a reasonable time.
- **Learning Rate** $= 1 \times 10^{-2}$:
  - **Speed**: A learning rate of 1×10^{-2} allows very large updates to the model weights, which may initially seem advantageous for a rapid drop in NLL.
  - **Stability**: However, the plot clearly shows significant oscillations and a sudden spike in NLL, indicating instability. The model fails to stabilize, leading to potential continuous oscillations and lack of convergence.
  - **Convergence**: The presence of spikes and oscillations in NLL suggests that the model may not be able to converge to an optimal solution, risking overshooting the minima of the cost function.

References

Barratt, S. and Sharma, R. (2018). A note on the inception score. *arXiv preprint arXiv:1801.01973*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Doersch, C. (2016). Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Reddi, S. J., Kale, S., and Kumar, S. (2018). On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286. PMLR.

Tomczak, J. M. (2022). *Deep Generative Modeling*. Springer Nature.