Assignment 3: Variational Autoencoders

- Student Name:
- Student #:
- Collaborators:

Background

In this assignment we will implement and investigate a Variational Autoencoder as introduced by Kingma and Welling in **Auto-Encoding Variational Bayes**.

Data: Binarized MNIST

In this assignment we will consider an MNIST dataset of 28×28 pixel images where each pixel is **either on or off**.

The binary variable $x_i \in \{0, 1\}$ indicates whether the i-th pixel is off or on.

Additionally, we also have a digit label $y \in \{0, \ldots, 9\}$. Note that we will not use these labels for our generative model. We will, however, use them for our investigation to assist with visualization.

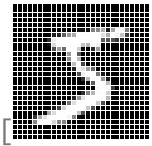
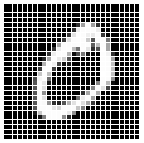
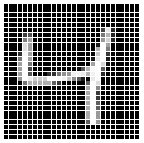
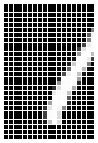
Tools

In previous assignments you were required to implement a simple neural network and gradient descent manually. In this assignment you are permitted to use a machine learning library for convenience functions such as optimizers, neural network layers, initialization, dataloaders.

However, you **may not use any probabilistic modelling elements** implemented in these frameworks. You cannot use `Distributions.jl` or any similar software. In particular, sampling from and evaluating probability densities under distributions must be written explicitly by code written by you or provided in starter code.

```
using Flux
```

```
train_digits =
```

Matrix{ColorTypes.Gray{FixedPointNumbers.N0f8}}[ ,  ,  , ]

```
# load the original greyscale digits
train_digits = Flux.Data.MNIST.images(:train)
```

```
greyscale_MNIST =
784×60000 Matrix{Float64}:
 0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
 0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
 0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
 0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
 0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
 0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
 0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
 ⋮                        ⋮              ⋱                 ⋮
 0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
 0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
 0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
 0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
 0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
 0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
```

```
# convert from tuple of (28,28) digits to vector (784,N)
greyscale_MNIST = hcat(float.(reshape.(train_digits,:))...)
```

```
binarized_MNIST =
784×60000 BitMatrix:
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 ⋮                 ⋮                 ⋮              ⋱              ⋮                 ⋮
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
```

```
# binarize digits
binarized_MNIST = greyscale_MNIST .> 0.5
```

```
BS = 200
```

```
# partition the data into batches of size BS
BS = 200
```

```
batches =
```

```
Flux.Data.DataLoader{BitMatrix, Random._GLOBAL_RNG}(784×60000 BitMatrix:
 0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0
 ⋮                 ⋮                 ⋮
 0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0
```

```
# batch the data into minibatches of size BS
batches = Flux.Data.DataLoader(binarized_MNIST, batchsize=BS)
```

```
(784, 200)
```

```
# confirm dimensions are as expected (D,BS)
size(first(batches))
```

Model Definition

Each element in the data $x \in D$ is a vector of 784 pixels. Each pixel x_d is either on, $x_d = 1$ or off $x_d = 0$.

Each element corresponds to a handwritten digit $\{0, \ldots, 9\}$. Note that we do not observe these labels, we are *not* training a supervised classifier.

We will introduce a latent variable $z \in \mathbb{R}^2$ to represent the digit. The dimensionality of this latent space is chosen so we can easily visualize the learned features. A larger dimensionality would allow a more powerful model.

- **Prior**: The prior over a digit's latent representation is a multivariate standard normal distribution. $p(z) = \mathcal{N}(z \mid \mathbf{0}, \mathbf{1})$
- **Likelihood**: Given a latent representation z we model the distribution over all 784 pixels as the product of independent Bernoulli distributions parametrized by the output of the "decoder" neural network $f_\theta(z)$.

$$p_\theta(x \mid z) = \prod_{d=1}^{784} \text{Ber}(x_d \mid f_\theta(z)_d)$$

Model Parameters

Learning the model will involve optimizing the parameters θ of the "decoder" neural network, f_θ.

You may also use library provided layers such as `Dense` **as described in the documentation**.

Note that, like many neural network libraries, Flux avoids explicitly providing parameters as function arguments, i.e. `neural_net(z)` instead of `neural_net(z, params)`.

You can access the model parameters `params(neural_net)` for taking gradients `gradient(()->loss(data), params(neural_net))` and updating the parameters with an **Optimiser**.

However, if all this is too fancy feel free to continue using your implementations of simple neural networks and gradient descent from previous assignments.

Numerical Stability

The Bernoulli distribution $\mathrm{Ber}(x \mid \mu)$ where $\mu \in [0, 1]$ is difficult to optimize for a few reasons.

We prefer unconstrained parameters for gradient optimization. This suggests we might want to transform our parameters into an unconstrained domain, e.g. by parameterizing the `log` parameter.

We also should consider the behaviour of the gradients with respect to our parameter, even under the transformation to unconstrained domain. For instance a poor transformation might encourage optimization into regions where gradient magnitude vanishes. This is often called "saturation".

For this reasons we should use a numerically stable transformation of the Bernoulli parameters. One solution is to parameterize the "logit-means": $y = \log(\frac{\mu}{1-\mu})$.

We can exploit further numerical stability, e.g. in computing $\log(1 + exp(x))$, using library provided functions `log1pexp`

```
using StatsFuns: log1pexp #log(1 + exp(x))
```

```
bernoulli_log_density (generic function with 1 method)
```

```
# Numerically stable bernoulli density, why do we do this?
function bernoulli_log_density(x, logit_means)
  """Numerically stable log_likelihood under bernoulli by accepting μ/(1-μ)"""
end
```

Model Implementation

- `log_prior` that computes the log-density of a latent representation under the prior distribution.
- `decoder` that takes a latent representation z and produces a 784-dimensional vector y. This will be a simple neural network with the following architecture: a fully connected layer with 500

hidden units and `tanh` non-linearity, a fully connected output layer with 784-dimensions. The output will be unconstrained, no activation function.

- `log_likelihood` that given an array binary pixels x and the output from the decoder, y corresponding to "logit-means" of the pixel Bernoullis $y = log(\frac{\mu}{1-\mu})$ compute the **log-**likelihood under our model.
- `joint_log_density` that uses the `log_prior` and `log_likelihood` and gives the log-density of their joint distribution under our model $\log p_\theta(x, z)$.

Note that these functions should accept a batch of digits and representations, an array with elements concatenated along the last dimension.

UndefVarError: log_prior not defined

1. top-level scope @ *Local: 1*

```
log_prior(z)
```

(2, 500, 784)

```
Dz, Dh, Ddata = 2, 500, 28^2
```

UndefVarError: decoder not defined

1. top-level scope @ *Local: 1*

```
decoder # You can use Flux's Chain and Dense here
```

log_likelihood (generic function with 1 method)

```
function log_likelihood(x,z)
  """ Compute log likelihood log_p(x|z)"""
    # use numerically stable bernoulli
end
```

UndefVarError: joint_log_density not defined

1. top-level scope @ *Local: 1*

```
joint_log_density(x,z)
```

Amortized Approximate Inference with Learned Variational Distribution

Now that we have set up a model, we would like to learn the model parameters θ. Notice that the only indication for *how* our model should represent digits in $z \in \mathbb{R}^2$ is that they should look like our prior $\mathcal{N}(0, 1)$.

How should our model learn to represent digits by 2D latent codes? We want to maximize the likelihood of the data under our model $p_\theta(x) = \int p_\theta(x, z)dz = \int p_\theta(x \mid z)p(z)dz$.

We have learned a few techniques to approximate these integrals, such as sampling via MCMC. Also, 2D is a low enough latent dimension, we could numerically integrate, e.g. with a quadrature.

Instead, we will use variational inference and find an approximation $q_\phi(z) \approx p_\theta(z \mid x)$. This approximation will allow us to efficiently estimate our objective, the data likelihood under our model. Further, we will be able to use this estimate to update our model parameters via gradient optimization.

Following the motivating paper, we will define our variational distribution as q_ϕ also using a neural network. The variational parameters, ϕ are the weights and biases of this "encoder" network.

This encoder network q_ϕ will take an element of the data x and give a variational distribution over latent representations. In our case we will assume this output variational distribution is a fully-factorized Gaussian. So our network should output the $(\mu, \log \sigma)$.

To train our model parameters θ we will need also train variational parameters ϕ. We can do both of these optimization tasks at once, propagating gradients of the loss to update both sets of parameters.

The loss, in this case, no longer being the data likelihood, but the Evidence Lower BOund (ELBO).

1. Implement `log_q` that accepts a representation z and parameters $\mu, \log \sigma$ and computes the logdensity under our variational family of fully factorized guassians.
2. Implement `encoder` that accepts input in data domain x and outputs parameters to a fully-factorized guassian $\mu, \log \sigma$. This will be a neural network with fully-connected architecture, a single hidden layer with 500 units and `tanh` nonlinearity and fully-connected output layer to the parameter space.
3. Implement `elbo` which computes an unbiased estimate of the Evidence Lower BOund (using simple monte carlo and the variational distribution). This function should take the model p_θ, the variational model q_ϕ, and a batch of inputs x and return a single scalar averaging the ELBO estimates over the entire batch.
4. Implement simple loss function `loss` that we can use to optimize the parameters θ and ϕ with `gradient`. We want to maximize the lower bound, with gradient descent. (This is already implemented)

UndefVarError: log_q not defined

1. **top-level scope** @ *Local: 1*

```
log_q(z, q_μ, q_logσ)
```

```
elbo (generic function with 1 method)
```

```
function elbo(x)
  q_μ, q_logσ#TODO variational parameters from data
  z #TODO: sample from variational distribution
  joint_ll #TODO: joint likelihood of z and x under model
```

```
  log_q_z #TODO: likelihood of z under variational distribution
  elbo_estimate  #TODO: Scalar value, mean variational evidence lower bound over batch
  return elbo_estimate
end
```

```
loss (generic function with 1 method)
```

```
function loss(x)
  return -elbo(x)
end
```

Optimize the model and amortized variational parameters

If the above are implemented correctly, stable numerically, and differentiable automatically then we can train both the `encoder` and `decoder` networks with graident optimzation.

We can compute `gradient` s of our `loss` with respect to the `encoder` and `decoder` parameters `theta` and `phi` .

We can use a `Flux.Optimise` provided optimizer such as `ADAM` or our own implementation of gradient descent to `update!` the model and variational parameters.

Use the training data to learn the model and variational networks.

```
train! (generic function with 1 method)
```

```
function train!(enc, dec, data; nepochs=100)
    params = Flux.params(enc, dec)
    opt = ADAM()

    for epoch in 1:nepochs
        for batch in data
            # compute gradient wrt loss
            # update parameters
        end
        # Optional: log loss using @info "Epoch $epoch: loss:..."
        # Optional: visualize training progress with plot of loss
        # Optional: save trained parameters to avoid retraining later
    end
    # return nothing, this mutates the parameters of enc and dec!
end
```

```
UndefVarError: encoder not defined

  1. top-level scope  @  Local: 1
```

```
train!(encoder, decoder, batches, nepochs=100)
```

Visualizing the Model Learned Representation

We will use the model and variational networks to visualize the latent representations of our data learned by the model.

We will use a variatety of qualitative techniques to get a sense for our model by generating distributions over our data, sampling from them, and interpolating in the latent space.

```
# using Plots ##
```

1. Latent Distribution of Batch

1. Use `encoder` to produce a batch of latent parameters $\mu, \log \sigma$
2. Take the 2D mean vector μ for each latent parameter in the batch.
3. Plot these mene vectors in the 2D latent space with a scatterplot
4. Colour each point according to its "digit class label" 0 to 9.
5. Display a single colourful scatterplot

```
Enter cell code...
```

```
Enter cell code...
```

```
Enter cell code...
```

```
Enter cell code...
```

2. Visualizing Generative Model of Data

1. Sample 10 z from the prior $p(z)$.
2. Use the model to decode each z to the distribution logit-means over x.
3. Transform the logit-means to the Bernoulli means μ. (this does not need to be efficient)
4. For each z, visualize the μ as a 28×28 greyscale images.
5. For each z, sample 3 examples from the Bernoulli likelihood $x \sim \text{Bern}(x \mid \mu(z))$.
6. Display all plots in a single 10 x 4 grid. Each row corresponding to a sample z. Do not include any axis labels.

```
Enter cell code...
```

```
Enter cell code...
```

```
Enter cell code...
```

```
Enter cell code...
```

3. Visualizing Regenerative Model and Reconstruction

1. Sample 4 digits from the data $x \sim \mathcal{D}$
2. Encode each digit to a latent distribution $q_\phi(z)$
3. For each latent distribution, sample 2 representations $z \sim q_\phi$
4. Decode each z and transform to the Bernoulli means μ
5. For each μ, sample 1 "reconstruction" $\hat{x} \sim \text{Bern}(x \mid \mu)$
6. For each digit x display (28x28) greyscale images of $x, \mu, \hat{x}$

```
• Enter cell code...
```

```
• Enter cell code...
```

4. Latent Interpolation Along Lattice

1. Produce a 50×50 "lattice" or collection of cartesian coordinates $z = (z_x, z_y) \in \mathbb{R}^2$.
2. For each z, decode and transform to a 28x28 greyscale image of the Bernoulli means μ
3. Each point in the `50x50` latent lattice corresponds now to a `28x28` greyscale image. Concatenate all these images appropriately.
4. Display a single `1400x1400` pixel greyscale image corresponding to the learned latent space.

```
• Enter cell code...
```

```
• Enter cell code...
```

```
• Enter cell code...
```

```
• Enter cell code...
```